Title
Machine Learning for the Developing World using Mobile Communication Metadata

Author
Khan, Muhammad R

Publication Date
2018

Peer reviewed|Thesis/dissertation

Machine Learning for the Developing World using Mobile Communication Metadata

by

Muhammad R. Khan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Management and Systems

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joshua Evan Blumenstock, Chair
Professor John Chuang
Professor David Bamman
Professor Dennis Feehan

Fall 2018

Machine Learning for the Developing World using Mobile Communication Metadata

Abstract

Machine Learning for the Developing World using Mobile Communication Metadata

by

Muhammad R. Khan

Doctor of Philosophy in Information Management and Systems

University of California, Berkeley

Professor Joshua Evan Blumenstock, Chair

Machine learning algorithms have started having an unprecedented impact on human society due to their improved accuracy. The ability to collect and analyze information at a large scale has enabled the researchers to develop novel algorithms that can beat the best of the human experts in quite a few cases. The size of the data and the quality of the data has been the primary factor behind the success of the machine learning algorithms. However, when it comes to the data related to human behavior a digital divide still exists. An indirect consequence of the popularity of the social networks and the proliferation of sensors in the developed world is that the researchers in industry and academia have been able to fine tune their findings and algorithms using these huge behavioral datasets providing accurate and deep insights about human behavior in the developed world. However, the same is not true about the developing world where until recently the surveys have been the primary way of collecting information about individuals in the society.

Social networks and digital sensors have not been that common in the developing world as compared to the developed world with one big exception, i.e., the mobile phones. More than 95% of the world population today has mobile phone coverage and event in some of the most under-developed places of the earth the penetration of mobile phones is much higher as compared to other measures of human development like literacy or access to the financial infrastructure. As a result, researchers have been using the meta-data collected by the mobile phone companies in these developing countries as an alternative to the more conventional data sources. However, the raw mobile phone data may not be very well suited for the machine learning algorithms. In other words, there is a need for algorithms to convert the raw mobile communication meta-data into features suited for the machine learning algorithms. Developing novel ways to extract features from the mobile phone meta-data has been the central question of my research.

In this dissertation, I am going to describe my work on extracting features from mobile communication logs using techniques like Deterministic Finite Automata (DFA). I will also show that how this approach outperforms other methods for problems like product adoption and churn prediction. I further show that by using DFA based features and spectral anal-

ysis of the multi-view nature of mobile communication networks, advanced neural network training algorithms can be developed that beat the current state of the art methods for the problems like poverty prediction and gender prediction. Last part of this dissertation describes the value of communication networks data for research questions related to social networks analysis like what are the salient differences between the behavioral patterns of men and women in the developing world as exhibited in the communication networks data.

I dedicate this dissertation to my mother, Niaz Fatima, to whom I owe every achievement and success in life. She passed away even before I started my Ph.D., but her love and words have been the most significant source of motivation for me.
The desire for excellence and knowledge that she inculcated in me kept me going through the most challenging phases of research.

Contents

Acknowledgments

I will first like to thank Allah Almighty for His blessings. My Ph.D. research has had many ups and downs but even in the deep moments of despair life of Prophet Muhammad (Peace be upon him) has been the beacon of hope and optimism.

The content presented in this dissertation has benefited immensely from the guidance and advice of my mentor and committee chair Joshua Evan Blumenstock. His continuous support, guidance and commitment to excellence has enabled me to overcome all the challenges faced throughout my Ph.D..

I would also like to thank my co-authors Philip Reed, Anikate Singh and Joshua Cherian Manoj for their help and collaboration in different phases of my research.

I am also grateful to other members of my thesis committee members: Dr. John Chuang, Dr. David Bamman and Dr. Dennis Feehan for their guidance and comments. I am also grateful to Dr. Bapu Vaitla for his extensive feedback on the work on the analysis of gender dynamics and disparities using call detail records (Chapter 6)

I would also like to thank my colleagues and friends in the Data Intensive Development Lab especially Guanghua Chi and Niall Kelleher for technical and non-technical discussions that have kept me going forward in this journey.

This journey would not have been complete without the support and encouragement of friends especially Muhammad Bilal Anwer, Wasif Latif and Qasim Javed. The continuous encouragement of all of you has worked like an energizer for me.

Finally, this dissertation would not have been possible without the continuous support and encouragement of my family. First of all, a special thank you to my mother who inculcated a love of reading and desire for knowledge in me. I wish, she was here in this world to see this achievement of "hers".
Last, but not the least, I want to thank my wife Mahrukh for her continuous support, encouragement, love and above all patience throughout this journey.

Chapter 1

Introduction

> "At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. .... Often, the raw data is not in a form that is amenable to learning, but you can construct features from it that are. This is typically where most of the effort in a machine learning project goes". (Domingos, 2012).

Over the last decade, machine learning has transformed human society. From health care to object recognition, machine learning algorithms have started to outperform human experts (Rajpurkar et al., 2017; He et al., 2015; Assael et al., 2016; Gatys et al., 2015; Gebru et al., 2017; Kelly, 2017). However, most of the machine learning algorithms perform well in the presence of huge structured datasets. As a result, one of the first problem that machine learning researchers have to resolve is the availability of data. Though a sub-domain of machine learning is emerging, that intends to design strategies and algorithms to handle shortage of labeled training data but this sub-domain is still in the early stages of research.

The lack of comprehensive datasets for the analysis of human behavior is a much bigger problem in developing countries than in developed countries. Thus, researchers have been exploring non-conventional data sources in their research. One such non-conventional data source is mobile phone communication meta-data (or Call Detail Records (CDR) data). CDR datasets are a good alternative to traditional behavioral datasets in developing countries as the penetration of mobile phones in developing countries is quite high. CDR datasets have been used to model problems like poverty mapping (J. Blumenstock, Cadamuro, et al., 2015; Smith-Clarke et al., 2014) and spread of diseases (Buckee et al., 2013), in developing world with varying success. CDR datasets provide a good alternative to traditional data sources in developing countries but these raw datasets have to be converted into useful metrics and features to more accurately model the underlying machine learning problems.

Feature engineering, defined as "the process of converting raw data into a form that can support accurate machine learning" is a critical phase of every data science project. The focus of my research has been to show that the process of feature engineering can be automated in a way that the human expert involvement is minimum without compromising

on the accuracy of the results. This research has many different underlying questions: How can one reduce the role of human experts in feature engineering? How can one make sure that the generated feature set is comprehensive and highly predictive? How can the feature generation process be made scalable? And last but not the least, how to make the extracted features more interpretable for nontechnical audience? The success of a machine learning project is heavily dependent on the quality of features used. Pedro Domingos (one of the experts in the field) notes on feature engineering, "it is also one of the most interesting parts where intuition, creativity and "black art" are as important as the technical stuff" (Domingos, 2012). The automation of feature engineering process is one of the holy grails of machine learning and even with all the progress made in the field of machine learning the issues of feature representation, interpretation and scalability are open issues. Machine learning researchers have focused mostly on developing or fine tuning learning algorithms, but in many cases, data preprocessing and features learning ends up taking more time than actual machine learning. All of this emphasizes the importance of feature engineering, and automated methods of feature engineering would be helpful in almost all machine learning projects. This is precisely the motivation of my research.
The main topics of research that I have described in this thesis are as follows:

- **Semi-automated feature engineering:**
  This part of my study (Chapters 2, 3 and 4) corresponds to the automation of feature design using techniques likes Deterministic Finite Automaton (DFA) (McCulloch and Pitts, 1943). The semi-automatic nature of the process implies that some human involvement is required, but such participation is minimal and does not go beyond specifying the aggregate operations to summarize the raw data.
  My research papers related to semi-automated feature engineering are as follows:

  - Khan, Muhammad R. and Joshua E. Blumenstock (2016). "Predictors Without Borders: Behavioral Modeling of Product Adoption in Three Developing Countries". *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016 (Chapter 2)
  - Khan, Muhammad Raza, Joshua Manoj, et al. (2015). "Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty". *Big Data (BigData Congress), 2015 IEEE International Congress on*. IEEE. 2015, 677–680 (Chapter 4)
  - Khan, Muhammad Raza and Joshua E Blumenstock (2016). "Machine Learning Across Cultures: Modeling the Adoption of Financial Services for the Poor". *Proceedings of the 2016 ICML Workshop on Data4Good: Machine Learning in Social Good Applications*. 2016 (Chapter 2)
  - Khan, Muhammad R. and Joshua E. Blumenstock (2017). "Determinants of Mobile Money Adoption in Pakistan". *31st Annual Conference on Neural Infor-*

mation Processing Systems, Workshop on Machine Learning for the Developing World. 2017 (Chapter 3)

- **Unsupervised feature engineering:**.
  The second half of my research corresponds to unsupervised feature learning over CDR data. Only recently, convolutional neural networks have been applied to the graph structured data or the network data but most of the work in this area does not handle the multi-view nature of the social networks.

  As I describe in Chapter 5, my approach of merging multi-view networks over CDR data outperforms many state of the art network embedding algorithms on problems like product adoption, gender prediction and poverty level prediction. However, my approach is not limited to the CDR data and as it is shown in Chapter 5, this approach can be applied to other type of network datasets (e.g. citation network datasets) as well.
  My research on unsupervised feature engineering is described in the following paper:

  – Khan, Muhammad R. and Joshua E. Blumenstock (2018b). "Multi-GCN: Graph Convolutional Networks for Multi-View Networks with Applications to Global Poverty". Forthcoming in AAAI 2019. 2018 (Chapter 5)

- **Social networks analysis using CDR data**.
  The CDR datasets that I have been using in my research on feature engineering can be used for social network analysis as well. One interesting question that I have explored is how the social networks of men and women as extracted from the CDR data differ in developing countries. Chapter 6 describes my work on the differences between the social networks of men and women in a developing country in detail.
  My research papers related to social network analysis are as follows

  – Reed, Philip J et al. (2016). "Observing gender dynamics and disparities with mobile phone metadata". *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. ACM. 2016, 48 (Chapter 6)

  – Khan, Muhammad R. and Joshua E. Blumenstock (2018a). "Gender Disparity Signals. Analyzing gender disparities with mobile phone metadata". To be submitted to ICWSM 2019. 2018 (Chapter 6)

The remainder of this document is organized as follows: Chapter 2 describes the deterministic finite automata (DFA) based approach for feature engineering and its application on the problem of product adoption in three different developing countries. Chapter 3 attempts a deeper analysis of the features generated by the DFA based approach to see how different determinants play different roles across different sections of the society. Chapter 4 applies features generated through DFA for modeling the churn of subscribers of a mobile phone operator in a developing country.

Chapter 5 incorporates latest trends of deep learning and spectral graph theory to develop a novel method of merging multiple views of an underlying network and applying graph convolution networks over the merged network.

Chapter 6 analyzes the gender disparities and dynamic using social network features extracted from mobile phone metadata.

Chapter 2

Adoption of Mobile Money

Abstract

Billions of people around the world live without access to banks or other formal financial institutions. In the past several years, many mobile operators have launched "Mobile Money" platforms that deliver basic financial services over the mobile phone network. While many believe that these services can improve the lives of the poor, in many countries adoption of Mobile Money still remains anemic. In this chapter, we develop a predictive model of Mobile Money adoption that uses billions of mobile phone communications records to understand the behavioral determinants of adoption. We describe a novel approach to feature engineering that uses a Deterministic Finite Automaton to construct thousands of behavioral metrics of phone use from a concise set of recursive rules. These features provide the foundation for a predictive model that is tested on mobile phone operators logs from Ghana, Pakistan, and Zambia, three very different developing-country contexts. The results highlight the key correlates of Mobile Money use in each country, as well as the potential for such methods to predict and drive adoption. More generally, our analysis provides insight into the extent to which homogenized supervised learning methods can generalize across geographic contexts. We find that without careful tuning, a model that performs very well in one country frequently does not generalize to another[1].

[1]The content presented in this chapter is based on joint work with Joshua Blumenstock originally published in 2016. See (Muhammad R. Khan and Joshua E. Blumenstock, 2016) for more details

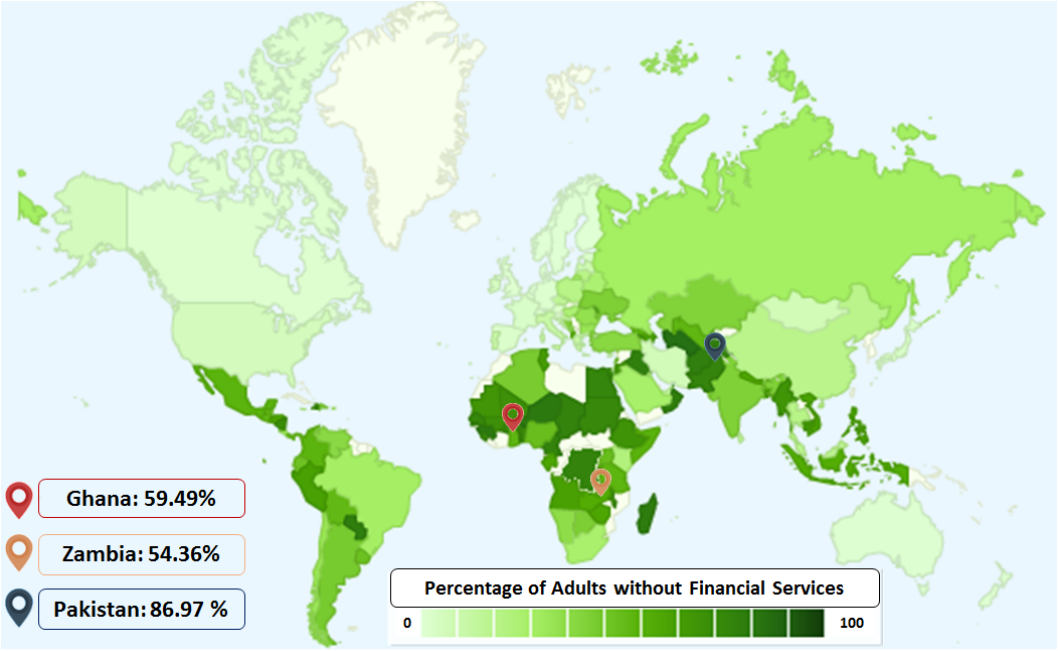


Figure 2.1: Worldwide access to formal financial services constructed using data from the Global Financial Inclusion Database Demirguc-Kunt et al., 2015. Study locations are identified by pins.

2.1 Introduction

The rapid penetration of mobile phones in developing countries is creating new opportunities to provide basic financial services to billions of individuals who have never before had access to banks or other formal financial institutions (Figure 2.1). In particular, over the last several years, mobile phone operators across the globe have launched "Mobile Money" platforms, which make it possible for mobile phone subscribers to conduct basic financial transactions from inexpensive feature phones. In several countries, these platforms have been wildly successful: two thirds of all Kenyan adults are active subscribers on the dominant Kenyan Mobile Money system Safaricom, 2014; in Bangladesh and Tanzania the corresponding usage rates are 40% and 50% Chen and Rasmussen, 2014; Di Castri and Gidvani, 2014. Globally, there are 255 live Mobile Money deployments in 89 countries, with an additional 102 planned deployments in the near future. With industry group GSMA estimating that 1 billion individuals currently own a phone but do not have a bank account Scharwatt et al., 2014, this presents massive potential to provide useful services to poor customers.

However, outside of the countries mentioned above and a few others, worldwide adoption of Mobile Money has been extremely anemic. The vast majority of deployments have struggled to promote sustained product adoption, and an industry report from 2014 estimates that 66% of registered customers were inactive Scharwatt et al., 2014. An open and important question thus revolves around understanding what drives customers to adopt and use

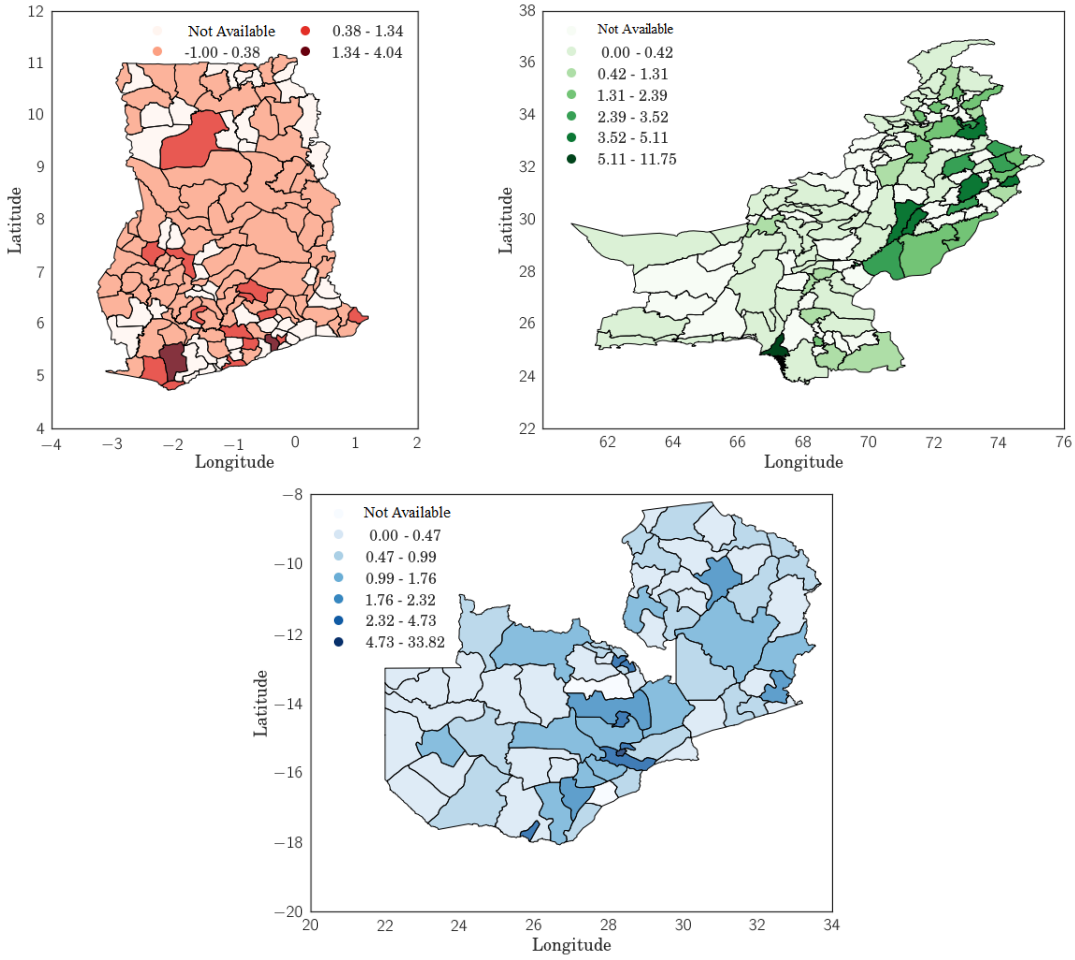


Figure 2.2: Geographic distribution of registered Mobile Money users in Ghana, Pakistan, and Zambia. Cells are colored according to the fraction of Mobile Money users in each region

Mobile Money, and whether patterns observed in one country will generalize to another.

In this chapter, we use spatio-temporal transactions data on mobile phone activity to model Mobile Money adoption in three developing countries. Our focus is on mining the Call Detail Records (CDR) collected by mobile phone operators, which contain detailed metadata on all events that transpire on the mobile phone network, including phone calls, text messages, and Mobile Money transactions. For thousands of unique individuals in each country, we can thus infer a wealth of information about the structure of their social networks, their daily movements about the country, patterns of communication, and several other behaviors that we discuss in greater detail below. We also know whether each subscribers eventually signs up for Mobile Money, and if so, whether he or she remains an active user on the system. In the three countries we study - Ghana, Pakistan, and Zambia - each Mobile Money platform is owned and operated by a separate, independent mobile phone op-

erator, and the subscriber population in each country has very different social and economic characteristics.

There are three substantive and one methodological contributions of this study. Substantively, we (1) develop a richer understanding of what drives the adoption of Mobile Money, by mining several large databases of transactions data; (2) construct a supervised learner that can predict, to varying degrees of accuracy depending on the prediction task and country context, the likelihood than an individual subscriber will use Mobile Money; and (3) explore the possibility that transfer learning could be used to train models in one country or context and apply them in another. To our knowledge, this is the first study to train and evaluate models of product adoption in three very different contexts. Since Mobile Money adoption is notoriously idiosyncratic, we hope this "cross-cultural" comparison can provide insight into the generalizability of our results, and increase their broader relevance to the policy and business communities working in developing countries.

Methodologically, we develop a novel framework for extracting behavioral metrics from transaction logs, which produces interpretable features that can provide the input data into standard supervised learning algorithms. This framework extends previous efforts described in J. Blumenstock, Cadamuro, et al., 2015, which used a simplified approach to predict poverty and wealth from mobile phone data. The core of this approach is formalized as a Deterministic Finite Automaton, which provides a structured, recursive grammar that relies on relatively few degrees of freedom to generate a comprehensive and interpretable set of "dense" features from sparse log data. This approach is sufficiently generalizable that we hope it can be further extended to a much broader range of contexts where researchers and data scientists wish to extract interpretable knowledge from transaction log data.

Related Work

Our work builds on several distinct strands in the academic literature. The first is concerned with understanding the determinants of mobile money adoption. This literature has historically been the domain of development researchers, and includes both macroeconomic and ethnographic work. The macro-scale work is concerned with the national and regulatory forces that can promote and hinder the spread of mobile money, such as interoperability regulations, barriers to customer registration, and the need for a robust network of mobile money agents Mas and Radcliffe, 2011; Dermish et al., 2011; Donovan, 2012. The ethnographic work has focused primarily on qualitative studies of how mobile money can be integrated into the daily lives of the poor Morawczynski, 2009; Medhi et al., 2009; Etim, 2014.[2]

A second strand of literature seeks to derive general insights from patterns revealed in mobile phone transactions logs. This encompasses a wide array of applications, including

[2]A closely related body of work explores the welfare consequences of the spread of Mobile Money Aker et al., 2014; Jack and Suri, 2014; Joshua E Blumenstock, Callen, et al., 2015; Joshua E Blumenstock, Eagle, et al., 2016, though relatively few studies provide rigorous evidence that mobile money has a positive impact on the lives of the poor.

Country	Ghana	Pakistan	Zambia
Panel A: National statistics (Source: World Bank)			
Population	25.90 Million	185.00 Million	15.72 Million
Percent with bank accounts	40.51	13.02	45.64
GDP per capita (PPP adjusted)	$4081.70	$4811.4	$3904.00
Mobile phone subscriptions (per 100 people)	115	73	67
Mobile phone operators	6	6	3
Panel B: Mobile phone use (Source: Call Detail Records)			
Calls per subscriber per day	6.53 (6.99)	7.76 (10.25)	10.26 (102.86)
SMS per subscriber per day	3.10 (100.36)	38.71 (80.83)	10.88 (262.81)
Number of unique contacts	21.66 (24.91)	46.93 (139.67)	17.63 (328.63)
Number of unique Towers	12.98 (16.07)	24.15 (57.30)	7.35 (17.56)

Notes: Standard deviations reported in parenthesis.

Table 2.1: Summary statistics by country: national indicators and sample CDR metrics

predicting the socioeconomic status J. Blumenstock, Cadamuro, et al., 2015, gender V. Frias-Martinez, E. Frias-Martinez, et al., 2010b, and age Y. Dong et al., 2014 of individual mobile phone subscribers. These studies illustrate the rich signal latent in mobile operator data, which reflects social phenomena including the structure of social networks, patterns of mobility and migration, diurnal rhythms of daily activity, and expenditures on communication and airtime.

A third area of prior work, and the one most relevant to our study, contains several papers that use transactions data from mobile operators to study product adoption.[3] For instance, Khan et al. Muhammad Raza Khan, Manoj, et al., 2015 use billing data to predict customer churn in an Afghan telecom, using a brute-force approach to feature generation. Sundsoy et al. Sundsøy et al., 2014 construct 350 features from call detail records and compare the performance of basic machine learning algorithms to that of a marketing department in predicting uptake of data plans. Finally, an industry report by CGAP compares the relative influence of different types of mobile phone metrics on the adoption of Mobile Money in Africa, using 180 metrics derived from call data CGAP, 2013. Relative to these studies, our study moves this literature forward by (a) innovating in the method used to generate features, thereby providing a systematic and comprehensive approach to feature engineering; (b) leveraging data from three different contexts to calibrate the external validity and generalizability of our results; and (c) carefully articulating the experimental protocols and algorithms in a way that will enable other researchers to replicate and extend these methods.

[3]A much broader literature, which we do not review here, studies the role of social networks in the adoption of new technologies Ugander et al., 2012; Leskovec et al., 2007.

2.2 Data and Context

For this study, we worked in collaboration with three mobile phone operators in Ghana, Pakistan, and Zambia. All three countries rank in the bottom third of the Human Development Index, a metric developed by the United Nations to capture a broad range of welfare outcomes such as income, education, inequality, and life expectancy. As can be seen in Figure 2.1, penetration of financial services is very low in each country. Geographic patterns of Mobile Money adoption also vary greatly within each country, as shown in Figure 2.2. Additional information on each country is provided in Table 2.1. Of course, these country-level statistics mask enormous diversity between and within nations in social and demographic characteristics, religious and political attitudes, and general ways of living.

From each mobile phone operator, we obtained the anonymized Call Detail Records (CDR) and Mobile Money Transaction Records (MMTR) of every subscriber on the network. The CDR and MMTR contain basic metadata on every event that occurs on the mobile phone network, including phone calls, text messages, and any form of Mobile Money activity. CDR typically consists of tuples containing {`callerID, recipientID, date, time, duration, callerLocation`}, where the two ID's are anonymized phone numbers, the date and time indicate when the event transpired, the duration of the call is recorded in seconds, and the location field indicates the cellular tower through which the call was routed, which can be used to pinpoint the approximate location of the individual at the time of the call.[4] For the mobile money platforms we study, the MMTR contain similar metadata for basic financial transactions, such as deposits, withdrawals, purchases, balance checks, and so forth.

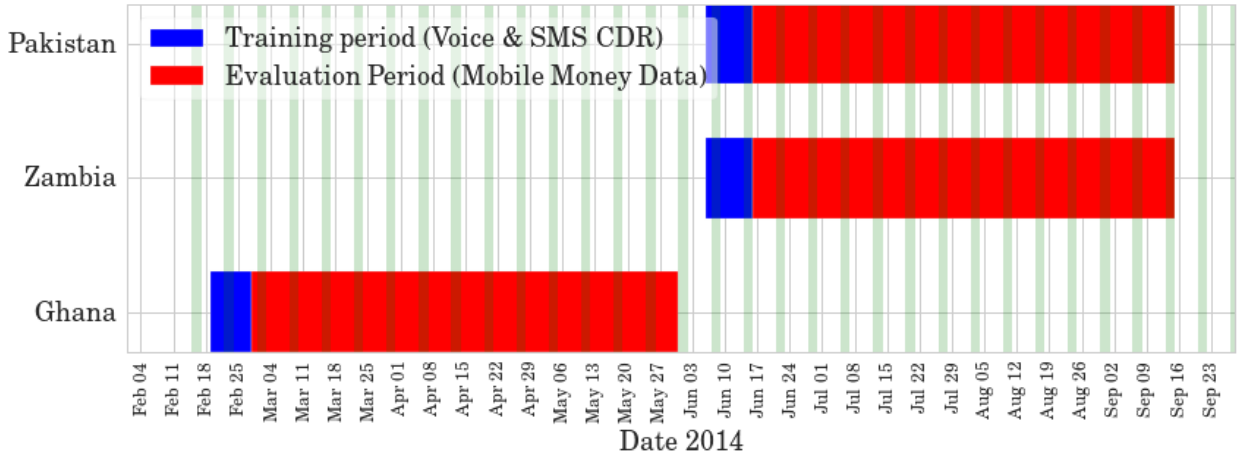


Figure 2.3: Training and evaluation periods

[4]In practice, the cell tower is accurate within several hundred meters in urban areas, and tens of kilometers in rural regions. We do not observe the contents of text messages. CDR are only generated when an individual initiates a transaction on the network, so we do not observe, for instance, the individual's location when she is not using her phone.

In total, the original data contains billions of transactions conducted by tens of millions of unique individuals. Each dataset spans several months of activity, which we divide into a "training" period and an "evaluation" period. CDR from a 10-day training period was used to engineer features and fit a predictive model, where the target variables (based on Mobile Money activity) were measured in a subsequent 3-month evaluation period. The timing of these periods is depicted in Figure 2.3.[5]

Using data from the evaluation period, each subscriber in each sample was labelled as either a "Voice Only" User or a "Registered Mobile Money" user, where Registered Mobile Money users could also be labelled as "Active Mobile Money" users according to the following criteria:

- **Voice Only Users:** If the user did not make any Mobile Money transactions during the evaluation period.
- **Registered Mobile Money User:** If the user made one or more Mobile Money transactions during the evaluation period.
- **Active Mobile Money User:** If the user made at least one Mobile Money transaction in each month of the evaluation period. Note that all Active Mobile Money users are also Registered Mobile Money Users.

2.3 Feature Engineering with Deterministic Finite Automata

Approach

As highlighted in the introduction, the CDR contain a wealth of latent information about how people communicate, with whom they interact, the locations they visit, and many other social and behavioral characteristics. Our eventual goal is to leverage this information to better understand why people use Mobile Money, and to develop a predictive model of Mobile Money adoption. However, the raw CDR are not natural inputs to most machine learning algorithms, and interpretable metrics must first be derived from the CDR before inferences can be made.

In the prior literature, the vast majority of studies take a rather ad hoc approach to constructing interpretable metrics ("features") from the phone data. The most common approach is to hand-craft a small number of features that correspond to some intuition of the researcher. For instance, Y. Dong et al., 2014 focus on 5 topological properties of the static social network; Gutierrez et al., 2013 use two metrics that quantify airtime purchases; and V. Frias-Martinez, Virseda, et al., 2012 construct 6 measures of physical mobility. Even the more ambitious approaches, such as Sundsøy et al., 2014 and CGAP, 2013, which respectively use

[5]Our intent was to exactly align the training and evaluation periods across countries, but implementation constraints made this impossible.

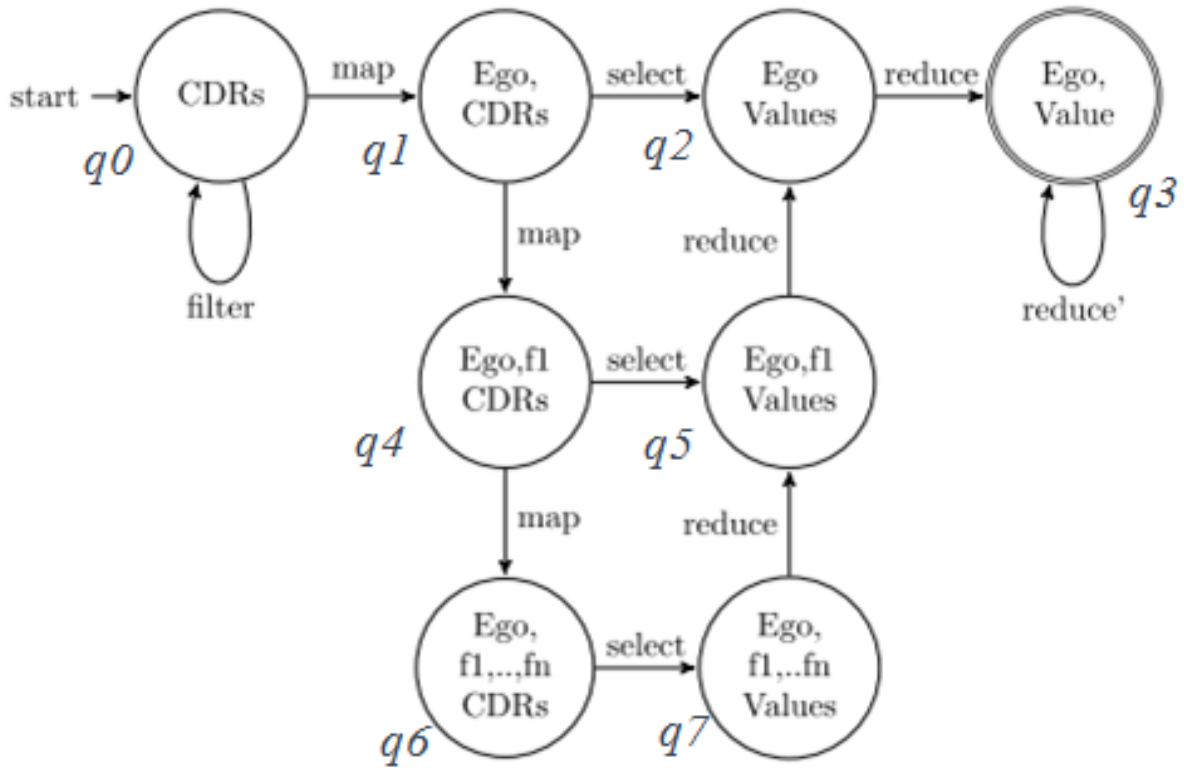


Figure 2.4: Deterministic Finite Automaton

350 and 180 CDR-based metrics, employ a large number of idiosyncratic rules to determine which features should be considered by the learning algorithm. These approaches have the advantage of producing metrics that are convenient to interpret, but they may systematically overlook non-intuitive features, mis-attribute relationships (if, for instance, feature A is weakly correlated with the target variable only because an omitted feature B is strongly correlated with both A and the target variable), or fail to maximize the predictive power of a classifier that would perform better with a more comprehensive set of features.

Our approach is different. We develop a method for feature engineering from transactional data that is designed to construct a large and comprehensive set of features from a small number of recursive operations. While the application is to CDR, we believe this method could be used to engineer features from a more diverse class of data including IP logs, social media data, and financial transaction records.

Deterministic Finite Automaton

We employ a deterministic finite automaton (DFA), a model of computation from automata theory also referred to as a deterministic finite state machine, to formalize the feature gener-

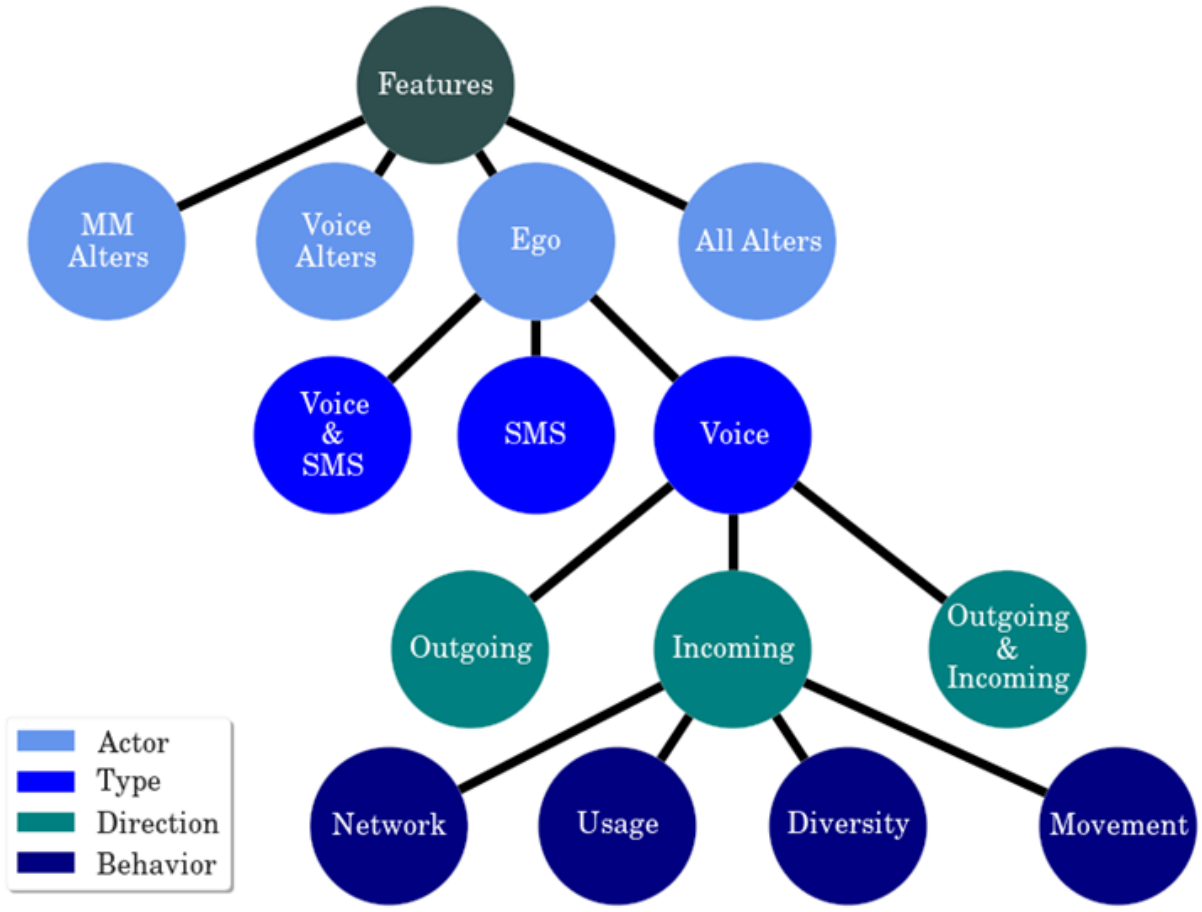


Figure 2.5: Tree-based feature classification

ation process McCulloch and Pitts, 1943. DFA's are typically used in more formal settings to determine whether an expression can be computed, to design circuits, or to operate simple devices. In the abstract, however, DFA's simply define a sequence of legal operations. We appropriate this concept to specify a set of legal operations that can be recursively applied to raw transactional data in order to produce valid features.

Example

As an example, say we are interested in constructing a feature for each individual i that corresponds to, "the variance in the average duration of outgoing calls made by i on different days of the week." We allow for the construction of this feature through the following set of recursive rules:

1. filter *outgoing calls*
2. filter transactions *initiated by* i

3. group by *day of week*
4. focus on *call duration*
5. aggregate by *average* (duration per day of week)
6. aggregate using *variance* (over average daily durations)

By using different filter criteria (or difference group-by and aggregation operations), by adding and removing rules, or by applying the rules in a different order, we produce different features. It is important to note, however, that not all combinations of operations are valid. For instance, it does not make sense to take the variance of a categorical variable (such as `recipientID`), nor does it make sense to group by day of week if a day of week filter has already been applied. The power of the DFA is that it allows us to formalize the set of valid features using a relatively parsimonious specification.

Formalization

The DFA we use to generate features from CDR is shown in Figure 2.4. In the figure, each circle represents a state, and for convenience we note the data structure expected for each state inside the circle. Valid features are constructed through traversals of the state machine, which start at the start state (`q0`) and end at the end state (`q3`) and follow only legal transitions between states (denoted by arrows). For instance, the feature described above would begin with the full CDR in `q0`, filter outgoing calls and return to `q0`, map (group by) "ego" i and proceed to `q1`, map by day of week to `q4`, select duration and proceed to `q5`, reduce (aggregate) by average - this produces average call duration for each day of week for each i - and proceed to `q2`, aggregate by variance and exit at `q3`.

Formally, the DFA is specified by:

- **Legal states:** `Q` = {`q0,q1,q2,q3,q4,q5,q6,q7`}
- **Start State:** `q0` $\in$ `Q`
- **End State:** `q3` $\in$ `Q`
- **Alphabet:** Σ = {`CDRs, Ego-CDRs, Ego-f-CDRs, Ego-Values, Ego-value`}
- **Transition Functions** ($\delta : Q \times \Sigma \rightarrow Q$)**:** `filter, map, select, reduce`. The set of legal transitions is described in greater detail in Appendix 2.3

In total, there are several thousand valid traversals of the DFA, each of which produces a different feature. Together, the resulting set of features covers almost all of the hand-crafted metrics used in prior work, as well as many, many more.

An additional advantage of the DFA is that it can be efficiently and elegantly implemented.

Feature classification and tree structure

The DFA is a convenient abstraction for generating a very large number of features from a small set of rules. To interpret the set of features produced by the DFA, we label each gener-

ated feature with interpretable tags. These tags are determined by the path taken through the automaton, and indicate whether each feature captures information on, for example, incoming vs. outgoing communications, calls vs. text messages, variance vs. volume, and so on. Specifically, we map each feature onto the tree structure shown in Figure 2.5, which is designed to encapsulate the substantive behavior captured by each feature. Each level of the tree corresponds to a different partition of the feature space:

1. **Actor:** Whether the feature relates to activity of the individual i (the "ego"), or the activity of i's first degree network of connections (the "alters"). We separately look at i's full set of alters, the set who have previously used Mobile Money, and the set that have never used Mobile Money. A simple aggregation operation (such as `mean` or `SD`) is applied to the alter network to produce a feature for i.
2. **Type:** Whether the feature relates to phone calls or text messages (SMS).
3. **Direction:** Whether the feature relates to incoming (e.g., call received) or outgoing (e.g., call placed) activity.
4. **Behavior:** Whether the feature relates to movement (e.g., number of unique cell towers used), network structure (e.g., number of unique contacts in network), phone usage (e.g., number of calls made), or Diversity (e.g., geographic spread of social network). This tag is determined by the data type of the field over which aggregated is performed (e.g., continuous vs. discrete data) and the actual statistical function used in aggregation (`count, unique, min, max, mean, median, SD, variance, radius of gyration`)

Figure 2.5 is simplified to show only a single expansion along the ego-voice-incoming path. In practice, all nodes on a given level can be expanded analagously to the path shown in Figure 2.5. For instance, the example feature described in Section 2.3 would be a leaf on the branch of ego-voice-outgoing-usage.

2.4 Models and Methods

The DFA-based process of feature engineering described above generates thousands of features that quantify patterns of mobile phone use. Armed with these features, our goals are to (a) use these features to understand the determinants of Mobile Money use, (b) build a predictive model that can be used to identify likely adopters, and (c) determine the extent to which models and features from one context can generalize to another.

Experimental Design

To facilitate our supervised learning experiments, we drew a stratified random sample of 10,000 subscribers from each of the three categories (Voice Only, Registered Mobile Money, Active Mobile Money) from each country. We elected to draw a balanced sample since, as

Algorithm 1: Feature Generation Algorithm

Data: cdr,Call Detail Records of all users
Data: opmap,Dictionary of possible operations
$cdrtypes \leftarrow$ [Voice, SMS, Voice and SMS]
$direction \leftarrow$ [In, Out , In plus Out]
$featuresArray \leftarrow []$

Result: Features

Step 1: Perform reduce by grouping only on ego

foreach *type, dir*1 *in cdrtypes,direction* **do**
  $filteredCDR \leftarrow cdr$.filter(*type*,*dir*1)
  foreach *field in cdr* **do**
    $groupeddata \leftarrow filteredCDR$.map([*ego*]+[combinations($field$])
    foreach *op in opmap[field]* **do**
      $reduceddata \leftarrow groupeddata$.reduce(*op*) insert $reduceddata$ in $featuresArray$
      $reducedata2 \leftarrow reduceddata$.map(*ego*,*alters*).reduce(*op*) insert $reduceddata2$ in $featuresArray$
    end
  end
end

can be seen in Figure 2.6, the vast majority of subscribers in each country fall into the "Voice Only" category.[6]

Classification and Model Selection

We then use a variety of supervised learning algorithms to tackle two classification tasks. First, we seek to differentiate between Voice Only subscribers and Registered Mobile Money Users (one or more Mobile Money transactions); second, we attempt to differentiate between Voice Only and Active Mobile Money users (at least one transaction per month). In all cases, we report the average accuracy across testing sets from 10-fold cross validation.

Since our data has a large number of features relative to observations, we focus on learners that are robust to overfitting, such as regularized and elastic net logistic regression Zou and Hastie, 2005, gradient boosting Friedman, 2001, and Extremely Randomized Trees Geurts

[6]To protect the commercial interests of the operators, we show only the fraction of users of each type, rather than the raw numbers, which are in the millions.

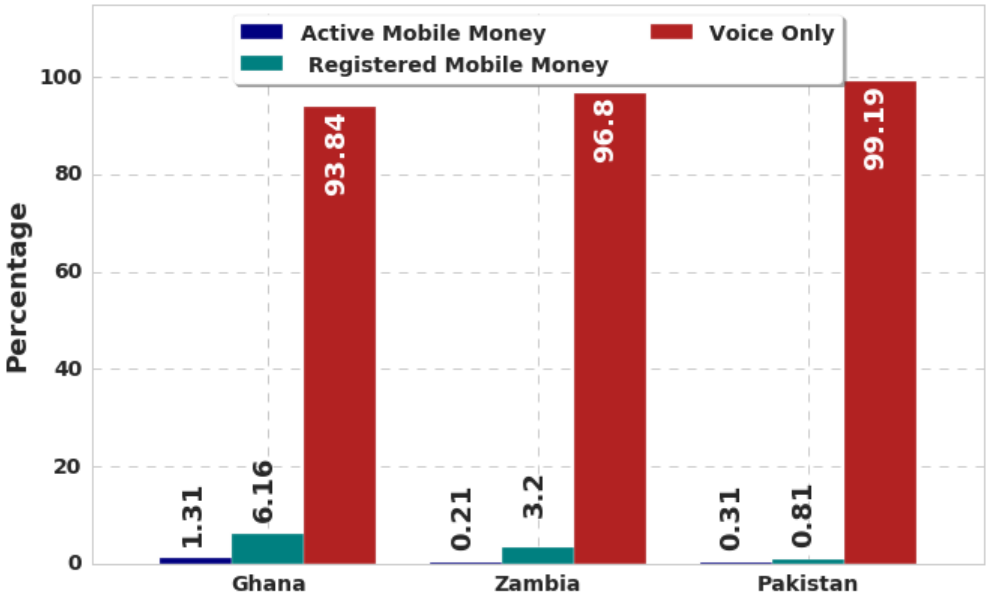


Figure 2.6: Distribution of user types by country

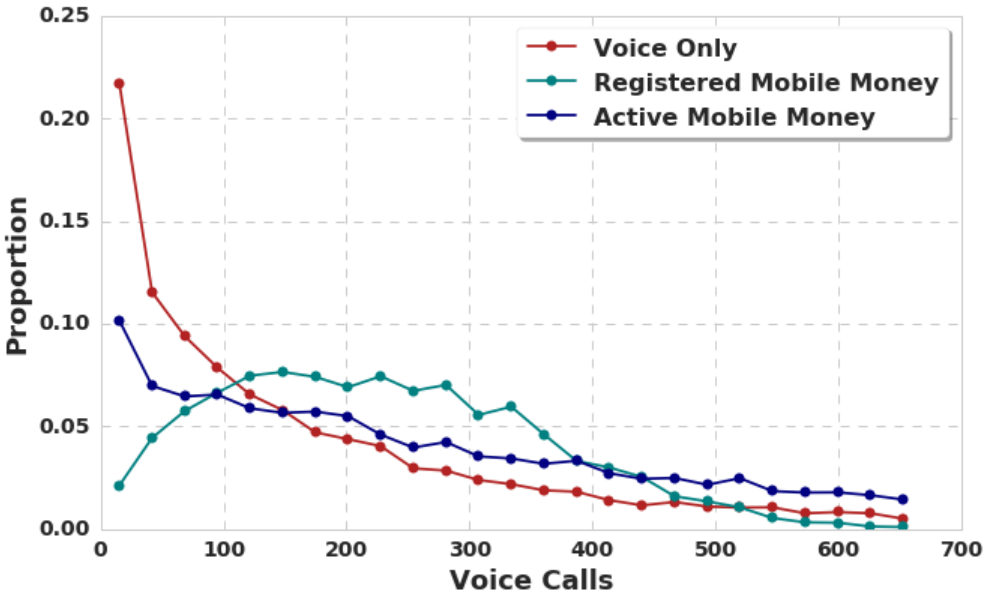


Figure 2.7: Distribution of calls per subscriber, Ghana

et al., 2006. Performance was comparable across these classifiers, although as expected these methods generally performed better than unregularized alternatives. To streamline the analysis that follows, we report only the results from gradient boosting, which outperformed

the other classifiers by a small margin.

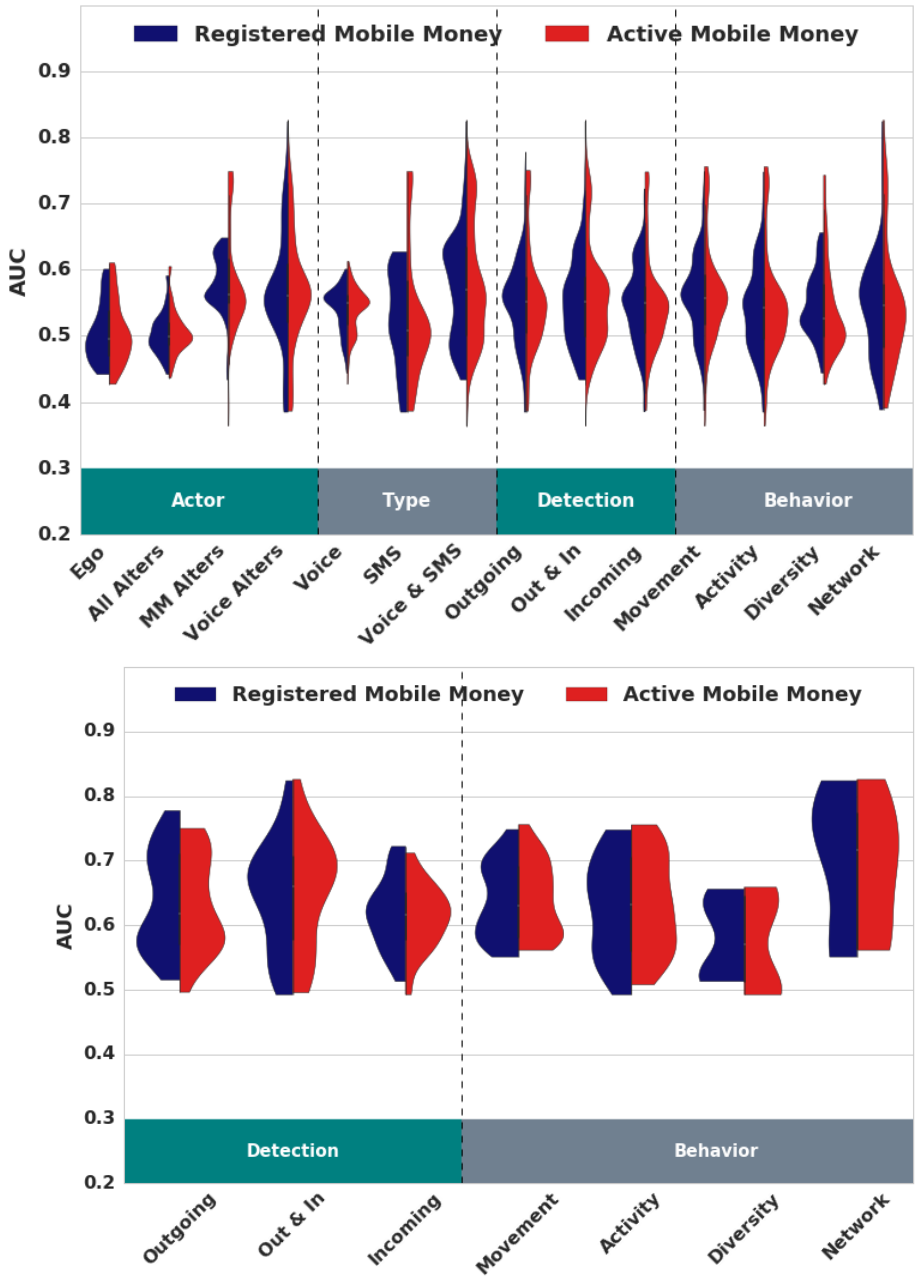


Figure 2.8: Distribution of AUC values for each feature category. Top figure shows all features in Ghana; Bottom figure shows the subset of features in Ghana where Actor='Voice Alters' and Type='All'

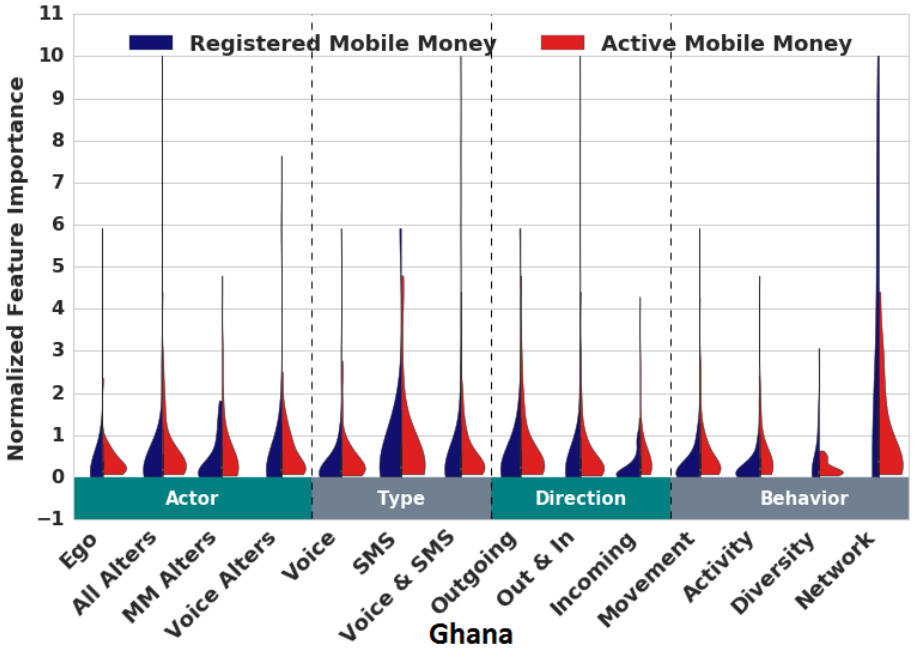


Figure 2.9: Normalized feature importance for Ghana

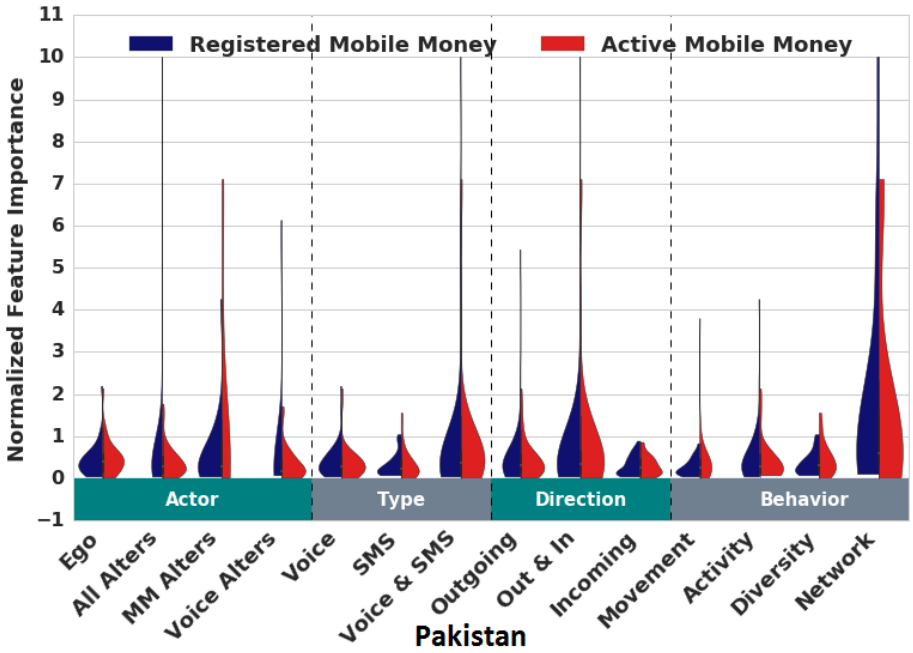


Figure 2.10: Normalized feature importance for Pakistan

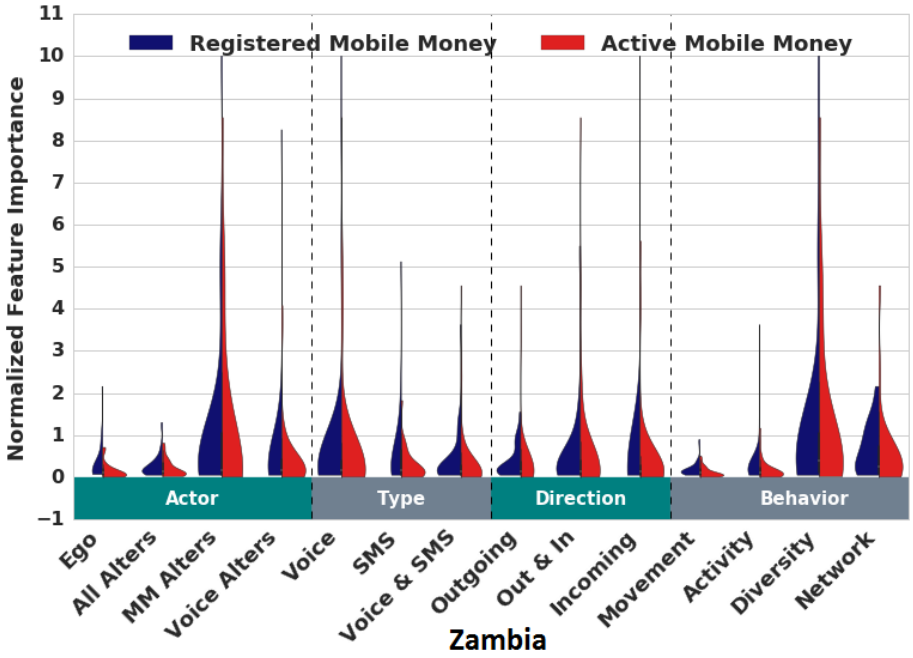


Figure 2.11: Normalized feature importance for Zambia

Feature selection and importance

To understand which CDR-based features are related to Mobile Money use, we calculate two metrics:

1. **(Unconditional) AUC**: We run a (cross-validated) bivariate logistic regression of the response variable (one of the above definitions of Mobile Money use) on each feature separately. The provides an indication of the unconditional correlation between each feature and the response variable.
2. **(Conditional) Normalized feature importance**: We calculate the importance of each feature to the final gradient boosting classifier. As we are primarily interested in the *relative* importance, the set of feature importances is standardized to be comparable across countries and classification tasks. Following Friedman, 2001, we denote the relative influence of feature x_j in tree T as

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 1(v_t = j)$$

where $\hat{i}^2$ is the improvement in (squared) error achieved by splitting feature v_t at node t, summed over all non-terminal nodes $J \in T$. In a collection of gradient-boosted trees,

the average feature importance $\bar{I}_j$ is the arithmetic mean of $\hat{I}_j^2(T)$ across all trees, and the *normalized feature importance* is the z-score obtained by subtracting the mean (of all $\bar{I}_j$) and dividing by the standard deviation (of all $\bar{I}_j$) for each $\bar{I}_j$.

2.5 Results

Determinants of Mobile Money Adoption

The DFA described in Section 2.3 produces roughly 3,000 unique features. As one example, Figure 2.7 shows the distribution of total calls made per subscriber in Ghana, for each of the three subscriber types. There are clear differences between the three user types in this distribution, with Voice Only users making the fewest calls, Registered Mobile Money users concentrated in the range from 100-300 calls (in the 10-day training period), and Active Mobile Money users more evenly distributed across the full range from 100-700 calls.

The distribution of unconditional AUC values for each of the 3,000 features is shown in Figure 2.8 (left panel), using Ghana as a test case. To construct this figure, we use the feature classification schema from Figure 2.5 to label each feature with four tags corresponding to the Actor, Type, Direction, and Behavior of the feature. Each violin plot then shows the distribution of AUC values for all features of a given type - such as all "ego" features, or all "movement" features. The left (blue) half of each violin plot indicates the distribution of AUC values for features when discriminating between Voice Only and Registered Mobile Money; the right (red) half shows the distribution when discriminating between Voice Only and Active Mobile Money.

While a large number of features have AUC values near 0.5, indicating they contain little information about the distinction between Voice Only and Mobile Money users, several noteworthy patterns emerge. First, when feature types are defined by the coarse classification tree in Figure 2.5, no single type dominates; rather, most types of features have a large number of uninformative features and a small number of highly predictive features with AUC$\geq$0.75. At the same time, feature classes do matter. The right panel of Figure 2.5 shows the distribution of AUC values for the subset of features where Actor="Voice Alters" and Type="All", a subset that are generally more predictive of Mobile Money use in Ghana. Here, the range of AUC values is significantly higher than in the full set of features, and some sub-classes such as "Network" have uniformly high predictive power.[7] Finally, the usefulness of each class of features depends on whether the goal is to identify Registered or Active Mobile Money users. For instance, the "MM Alter" features, which capture information about the characteristics of i's network who have previously adopted Mobile Money, are bimodally distributed and on average more useful in predicting Registered users than Active

[7]This particular class, where Actor="Voice Alters," Type="All," and Behavior="Network", corresponds to information about the network structure of i's network; in other words, 2nd degree properties of i's network.

users. However, that same class contains a small number of features that are extremely good predictors of Active Mobile Money use.

A similar approach is taken to construct Figures Figures 2.9, 2.10 and 2.11, except here we show the distribution of normalized feature importance values obtained through gradient boosting. The difference between the values in Figure 2.9 and Figure 2.8 is that the former are conditional on all features present in the final classification model, which includes several hundred features, whereas the latter are unconditional, i.e., they indicate performance in a univariate model with no other features. As in the unconditional ranking, each class of features in the conditional ranking contains a mass of features with low predictive power, but closer inspection reveals interpretable patterns.

Perhaps most striking in Figures 2.9, 2.10 and 2.11 are the differences between countries in the relative importance of each class of features. For instance, we see that in Ghana and Pakistan the "Network" features are in general more important to the classification model than the other types of Behavior, whereas in Zambia "Diversity" is most important. Zambia is also unique in the higher importance placed on voice calls relative to SMS activity, and in the fact that more signal exists in incoming calls than in outgoing calls. As we discuss below, these cross-country differences imply that models trained in one context may not generalize well to others.

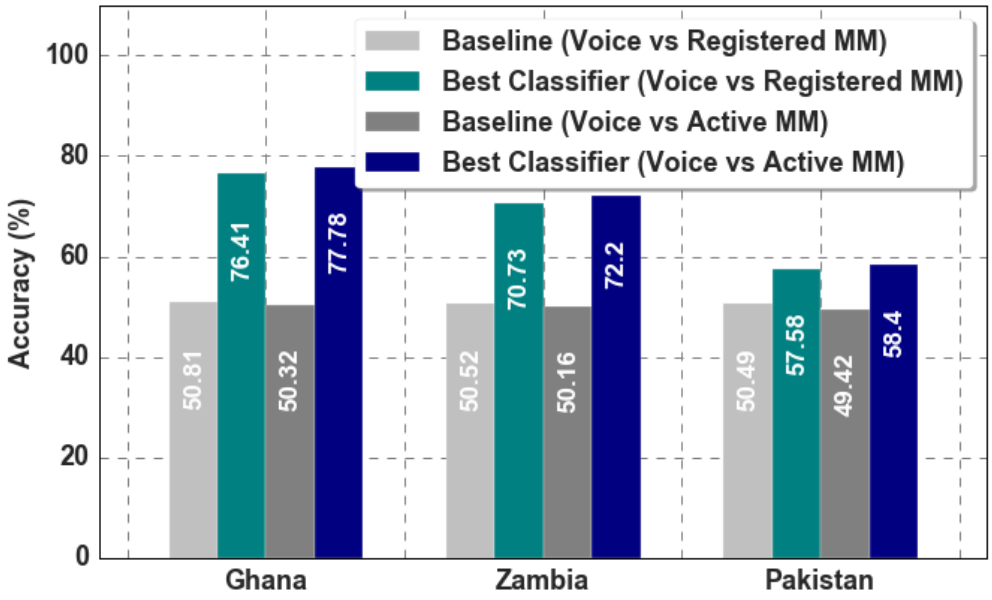


Figure 2.12: Accuracy in identifying Mobile Money users within each country, using gradient boosting

Predicting Mobile Money Use

As discussed in Section 2.4, we test the ability of several supervised classification models to discriminate between Voice Only and Mobile Money users, using the CDR-based features constructed from the DFA. Cross-validated results from gradient boosting are reported separately for each country in Figure 2.12. The best-performing models included several hundred features, but in practice there was little difference in performance between models in the range of 50-1000 features. We also include results from a baseline classifier, which uses the same model trained on a single "intuitive" feature – the total number of outgoing calls made by the subscriber, which is the feature shown in Figure 2.7.

In each country, the DFA-based classifier significantly outperforms the naive baseline, and in all countries, we achieve marginally better success in identifying Active Mobile Money users than Registered Mobile Money users. Across countries, however, there is a great degree of variability in classifier performance, with classification accuracy between 71% and 78% in Ghana and Zambia, but only 58%-59% in Pakistan. We discuss several possible explanations for these results in Section 2.6.

"Transfer Learning"

In the proceeding analysis, we have been careful to standardize the methods and analysis performed across all three countries. In each instance, we use the exact same source data, DFA specification, classification algorithm, experimental sample size, and so forth. In some cases, this meant that we knowingly discarded data that might have improved the performance of the classifier in a single country. For example, in some countries we had several months of CDR that could be used for training, additional fields in the CDR metadata, or much larger samples of Mobile Money users available for training and cross-validation. However, our approach reduced everything to the lowest common denominator in order to maintain comparability across contexts.

A key advantage of this approach is that it makes it possible to answer a question that has been elusive in prior studies of the adoption of new technologies in developing countries: *Do the behavioral determinants of adoption identified in one context generalize to another?* Based on the analysis we have performed, our short answer to this question appears to be, "No."

Figure 2.13 shows the performance of a classifier trained in one country and evaluated in another. Thus, the first set of six bars shows that the classifiers trained in Ghana perform well in Ghana (the first two grey bars), essentially replicating the results in Figure 2.12. However, that same Ghana model does quite poorly when it is evaluated in Zambia (the next two blue bars) and Pakistan (the final two green bars). While it is almost certain that a more sophisticated approach to transductive transfer learning would perform better Pan et al., 2011, the naive application of a model out of context is quite ineffective. We return to these ideas in the discussion that follows.

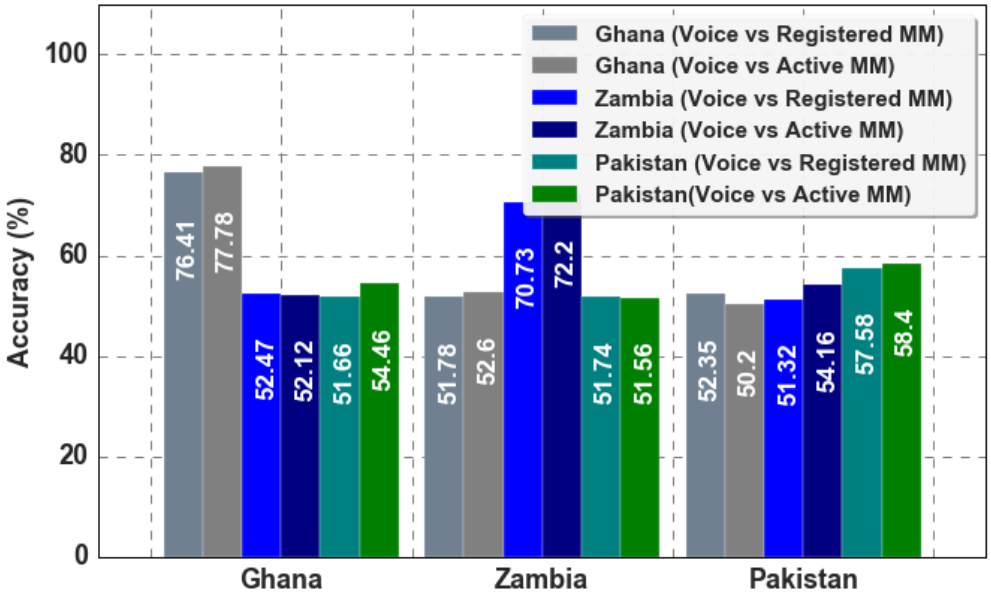


Figure 2.13: Accuracy when model is trained in one country and evaluated in another

2.6 Interpretation and Discussion

Taken in the broader context of research into the determinants of Mobile Money adoption and use, the preceding results uncover several unexpected patterns. Superficially, it is not surprising that CDR-based metrics can be used to construct classifiers that predict Mobile Money use, though to our knowledge this is the first study to publish performance metrics that can serve as benchmarks in future work in this area. However, in our analysis we were surprised to find that in a given country, the supervised model was only marginally better able to identify Active Mobile Money users (who make at least one transaction per month) than Registered Mobile Money users (who make at least one transaction ever). By contrast, our expectation was that active users, who are quite rare in all three countries, would have distinct patterns of phone use that would make them easier to detect. Since most policymakers agree that true financial inclusion requires active use, this remains an open topic for future work.

Also interesting are the differences in performance of the same modeling approach applied in different contexts (Figure 2.12). Most striking here is the relatively poor performance in Pakistan, where the 18% improvement over the baseline is dwarfed by the 55% improvement over the baseline achieved in Ghana.[8] At face value, this finding implies that Mobile Money

[8]This result is also unexpected, given the evidence in Figures 2.9, 2.10 and 2.11, which indicates that Ghana and Pakistan have very similar profiles in terms of relative feature importance. If anything, this figure might lead one to suspect that Zambia would be an outlier in the analysis, since Zambia's feature profile is

users in Pakistan are very similar to non-Mobile Money users, or at least that the two groups have similar patterns of mobile phone user. However, looking more carefully at the data, we believe this may also in part be an artifact of the "one size fits all" approach we have taken to standardizing definitions and methods across countries. In particular, there is one type of Mobile Money transaction that is extremely common in Pakistan, which allows a subscriber to add prepaid phone credit to her phone account using Mobile Money. Anecdotally, it is common practice in Pakistan for the retailers of phone credit to perform this Mobile Money transaction on behalf of the subscriber. Thus, a subscriber might appear to be using Mobile Money, though in practice she was not responsible for the transaction. This potential source of bias highlights the brittle nature of the cross-country analysis, which in its current form does not allow for country-specific adaptation.

Perhaps most importantly, our results suggest that across different countries and cultures in developing world, no single set of behavioral features is likely to consistently predict Mobile Money adoption and use. This is most clearly evident in Figure 2.13, which shows that a classifier trained in one country performs very poorly when tested in another country. But the same conclusion may also be drawn from Figures 2.9, 2.10 and 2.11, where we see that the same model, when trained in different countries, selects different features and attaches different weights to to those selected features. In results not shown, we further inspect the list of top-ranked features for each country, using both (unconditional) AUC and (conditional) normalized feature importance, and note very few features that appear consistently across countries. However, even though there may not be a "golden" list of features that always predict Mobile Money use, we are optimistic that more generalized insights can be extracted from one context and applied in another. In ongoing work, we are exploring methods for transfer learning that may strike this balance.

More concretely, over the past several months our partner in Ghana has been using the methods we describe to generate "Adoption Scores" that indicate the likelihood that any given mobile subscriber will adopt and use Mobile Money. They recently reported that when using these scores to target promotions, response rates were roughly 30% higher than promotions targeted with traditional methods. Such estimates are notoriously unreliable and subject to many possible sources of bias, but their optimism provides an indication of the potential for this line of research. At the same time, it should be noted that if the end goal is to increase financial inclusion of the poor, further methodological innovation is needed beyond what identifies the "low hanging fruit" subscribers whose behavior indicates that they are likely to adopt of their own volition.

2.7 Conclusion

In this chapter, we present a new approach to feature engineering that uses deterministic finite automata to construct a very large number of features from a concise set of rules. In applying this technique to mobile phone data from Ghana, Pakistan, and Zambia, we show

distinct from the other two countries.

that the resultant metrics correlate with, and can be used to predict, both active and passive Mobile Money use in three very different contexts. In so doing, we discover several previously undocumented patterns related to the adoption and use of Mobile Money. Superficially, the analysis makes it possible to highlight specific correlates of Mobile Money use, such as the relative importance of network structure in Ghana and Pakistan, and the relative importance of geographic diversity in Zambia. More fundamentally, the results provide insight into the extent to which standard predictive models can generalize across contexts. Here, it is clear that each population has a unique signature in terms of what metrics are good predictors of adoption, and as a result, models trained in one location do not perform well in another. Retraining the model helps, but does not solve, the underlying issue. Despite the fact that the data structures, experimental design, and Mobile Money products are nearly identical in the three countries, the performance of each country-specific model varies greatly.

2.8 Acknowledgments

Financial support for this project was provided by The MasterCard Foundation - IFC Partnership for Financial Inclusion in Sub-Saharan Africa. We are grateful to Sven Harten for providing thoughtful feedback throughout the research process, and to Gabriel Cadamuro and Robert On for help in developing an earlier version of the DFA.

Chapter 3

Determinants of Mobile Money Adoption

Abstract

A better understanding of the consumer behavior when it comes to adopting a new product is a key research problem for researchers working in the social sciences and behavioral economics domain. However, this problem is quite complex in nature. Part of the complexity is due to the fact that comprehensive data about the consumer performances may not be available while part of it is due to the lack of models that can accurately analyze the behavior of individuals. The promotion of products in developing countries is even more complex as the socio-economic disparity, social and traditional norms towards different genders may all play a role in determining the behavior of individuals. In this work, we analyze the problem of adoption of mobile money or digital financial services in Pakistan. We use the Call Detail Records of a major mobile communication company as our input and the social demographics data to see that how the trends of adoption of mobile money vary across different demographics. Our results show that machine learning and social network analysis can help in better understanding of the individuals' propensity to adopt mobile money services. At the same time, our results highlight the fact that different sections of the society have different determinants of adoption of digital financial services. More generally, the knowledge of determinants of mobile money adoption in developing countries can help the development agencies to align their marketing and user engagement schemes in a better way, hence resulting in improved financial inclusion. Our findings show that for most of the sections of the Pakistani society, the user mobility related features are the most important one when it comes to adopting and using mobile money services. Though our current findings are based on the data in Pakistan but our models and frameworks are generic and can be easily applied to other countries as well[1].

[1]This chapter is based on the joint work with Joshua Blumenstock published originally in 2017 (Muhammad R. Khan and Joshua E. Blumenstock, 2017).

3.1 Introduction

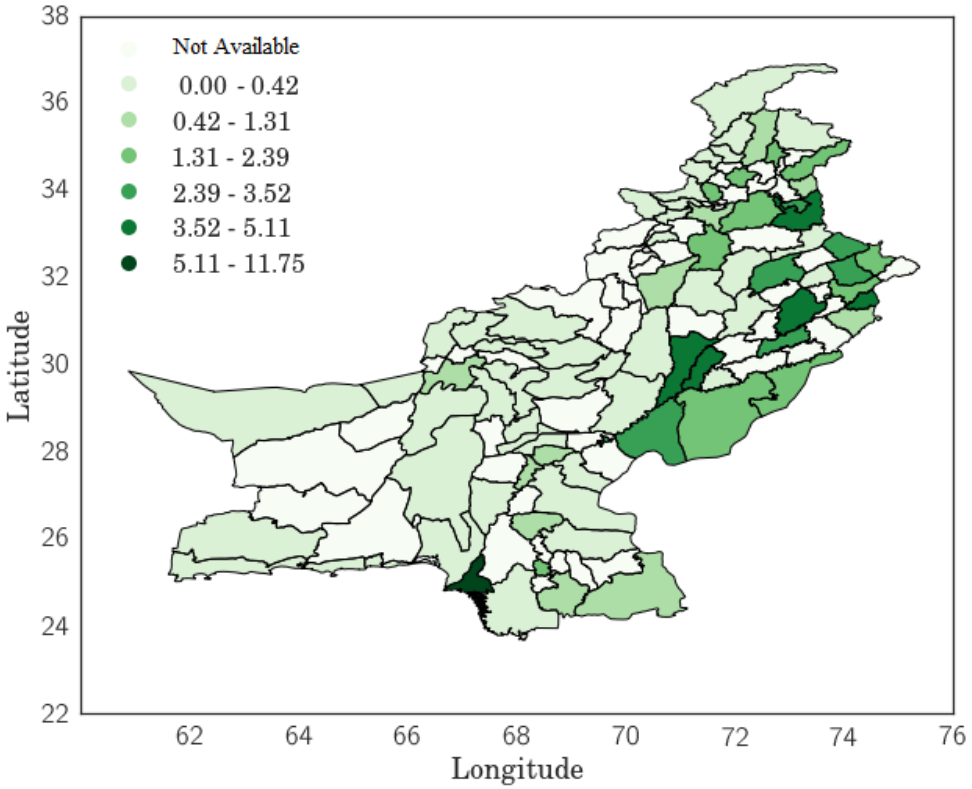


Figure 3.1: Geographic distribution of registered mobile money users in Pakistan.

There are many countries in developing world where the majority of the population does not have access to financial services like banks and credit institutes. One of the primary reason behind the lack of the financial institutes and services in these countries is the lack of infrastructure and resources. However, even in the poorest of the countries, mobile communication services have seen broad adoption. According to estimates by the industry group GSMA, more than 1 billion individuals owned a mobile phone but did not have a bank account in 2014Scharwatt et al., 2014. The penetration of the mobile phone services in these countries is providing alternate ways to provide services related to health, education, and finance in these countries. Use of mobile phone to provide financial services or "Mobile money" services enable mobile phone users to perform basic financial transactions using even the low-end mobile phones. These services have seen good penetration in countries like Kenya, Bangladesh and Tanzania Safaricom, 2014; Chen and Rasmussen, 2014; Di Castri and Gidvani, 2014.

Considering the lower cost of additional infrastructure and massive popularity of mobile money services in countries like Kenya, government, mobile operators and development agencies all around the world have been trying to promote mobile money services. As a

result, almost all of developing countries today have one or more mobile money services. However, except for a few handful of countries, only a few have seen widespread adoption of mobile money. Most of the services around the world have been unable to achieve a critical user mass, and as many as 66% of the registered customers worldwide were determined to be inactive customers Scharwatt et al., 2014.

Country	**Pakistan**
Panel A: National statistics	
Population	185.00 Million
Percent with bank accounts	13.02
GDP per capita (PPP adjusted)	$4811.4
Mobile phone subscriptions (per 100 people)	73
Mobile phone operators	6
Panel B: Mobile phone use	
Calls per subscriber per day	16.58 (21.44)
SMS per subscriber per day	23.23 (113.29)
Number of unique contacts	7.8 (62.73)
Number of unique Towers	6.15 (13.14)

Notes: Standard deviations reported in parenthesis.

Table 3.1: Summary statistics: national indicators(Source: World Bank) and CDR metrics

An important question thus revolves around understanding what drives customers to adopt and use mobile money. In other words, how do the patterns of adoption of mobile money differ for different sections of society like people belonging to different genders and different socio-economic conditions? This paper is focused on the following questions.

- **How do patterns of adoption of mobile money vary across different genders?**
- **How do patterns of adoption of mobile money vary across urban and rural areas; rich and poor areas?**
- **How accurately can we predict adoption of mobile money for different sections of the society mentioned in the last two questions?**

In this chapter, we use the mobile communication meta-data to model the adoption of mobile money services in Pakistan. Our input data contains all the events that transpire the network of a major telecom operator for a span of two weeks. This data in total contains millions of users and billions of voice and SMS transactions. By augmenting this data with the ground truth information (like the gender of the users and the urban density & poverty index of the districts) and using comprehensive feature generation and machine learning

algorithms, we have explored that how does the adoption of mobile money vary across different sections of the society.

3.2 Related Work

Our work builds on previous work in different strands of literature involving adoption of digital financial services, mining of call detail records, analysis of adoption of products and behaviors using social networks, and analysis of digital divide
The work on the adoption of digital financial services has been dominated by macroeconomic and ethnographic work. The focus of macro-scale work has been on the regulatory issues around interoperability, barriers to customer registration and the logistics of mobile money agents Mas and Radcliffe, 2011; Dermish et al., 2011; Donovan, 2012. The ethnographic work consists of qualitative studies on incorporating the digital financial services into the lives of the poor Morawczynski, 2009; Medhi et al., 2009; Etim, 2014.

Our research methodology is quite similar to some other research works related to the mining of insights from mobile communication meta-data. This has been a popular area of research and examples of work in this area include predicting the socioeconomic status Blondel et al., 2015, gender V. Frias-Martinez, E. Frias-Martinez, et al., 2010b, age Y. Dong et al., 2014 of individual mobile phone subscribers, customer churn behavior Muhammad Raza Khan, Manoj, et al., 2015 and analysis of gender disparities using social networks extracted from mobile communication logs Philip J. Reed et al., 2016.

The third area of prior work, and the one most relevant to this chapter studies the role of social networks in the adoption of new technologies Ugander et al., 2012; Leskovec et al., 2007.There have been some studies that have analyzed the social networks of mobile money users like Kusimba et al., 2013 and Murendo et al., 2017, however, most of these studies use limited data sets while our focus in this study is to use big data sets to determine prominent determinants of adoption of mobile money.

We have used deterministic finite automata (DFA) based feature engineering over CDR data which is quite similar to the approach used in Muhammad R. Khan and Joshua E. Blumenstock, 2016 and J. Blumenstock, Cadamuro, et al., 2015. The DFA based approach incorporates many simple and complex transactional and social network features. One important feature that we use in our analysis is the diversity of users' communication as described by Eagle et al. Eagle et al., 2010a.

Lastly, the differences in the usage of technology across men and women, poor & rich, and the urban & rural population has been a popular theme of research in the ICTD domain Jackson et al., 2008, J. Blumenstock and Eagle, 2010,V. Frias-Martinez, E. Frias-Martinez, et al., 2010b,etc. However, the analysis of adoption of mobile money on such lines has been limited to some blog posts and case studies in most of the cases McKay and Kaffenberger, 2013; Intermedia, n.d.; Minischetti, 2017; Minischetti, 2016.

In comparison to these studies, our work is unique as we build on the DFA based feature engineering over big country level CDR dataset to evaluate the role of the different type of features in the adoption of mobile money for people with different demographic backgrounds. To the best of our knowledge, this work is the first one to analyze the adoption of mobile money across different genders, urban/rural regions and socio-economic conditions using country level datasets.

3.3 Data and Context

For this study, we have used the mobile money adoption data from a major telecom operator in Pakistan. The data that we have can cover around 100 of Pakistan. When it comes to the human development, quite a significant population of Pakistan lives below the poverty line. Furthermore, more than 80% of the adult population of Pakistan does not have access to financial services, hence increasing the adoption of these services can have a real impact on the life of poor people in Pakistan. Some of the statistics about Pakistan are mentioned in the Table 3.1

Our input data consists of anonymized call detail records (CDR) and mobile money transaction information (MMTR) from the mobile operator. Each row of the CDR typically consists of the tuple containing {`callerID, recipientID, date, time, duration, callerLocation, receiverLocation`}, where the two ID's are anonymized phone numbers, the date and time indicate when the event transpired, the duration of the call is recorded in seconds. The callerLocation and receiverLocation specify the location of the cellular tower from which the call was made and location of the tower in which the receiver was present respectively. The MMTR contain similar metadata for basic financial transactions, such as deposits, withdrawals balance checks, and so forth.

In total, the original data contains billions of transactions conducted by tens of millions of unique individuals. We use this data as input to our feature generation framework, and the generated features are used to classify users into different types. We categorize each user either as a "Voice only user","Registered mobile money user" or "P2P mobile money user" depending on the usage of mobile money services as explained below.

- **Voice only users:** Voice only users are the ones who do not use mobile money services at all.
- **Registered mobile money users:** The users who have registered for the mobile money services.
- **P2P mobile money users:** The users who have signed up for mobile money and have either sent or received money to another mobile money users.

P2P mobile money users can be considered to be more advanced users of mobile money as they are using the mobile money services to full potential.

Feature	**Voice Only vs Registered Mobile Money**	**Voice Only vs P2P**
Panel A: Gender: Female		
Number of registered mobile money users	6570	3325
Outgoing calls per user	58.59 (215.55)	62.12 (213.76)
Outgoing calls' network size per user	30.8 (41.09)	9.2 (22.25)
Distinct outgoing calls locations per user	32.8 (46.73)	12.0 (27.94)
Panel A: Gender: Male		
Number of registered mobile money users	60050	28585
Outgoing calls per user	111.26 (294.04)	117.07 (300.92)
Outgoing calls' network size per user	16 (48)	11.6 (24.86)
Distinct outgoing calls locations per user	11.2 (33.6)	13.8 (27.63)
Panel C: Rural		
Number of registered mobile money users	17393	7512
Outgoing calls per user	128.23 (313.78)	139.04 (328.99)
Outgoing calls' network size per user	20.4 (33.83)	27.4 (41.59)
Distinct outgoing calls locations per user	17.0 (23.56)	24.2 (37.9)
Panel D: Urban		
Number of registered mobile money users	45960	22230
Outgoing calls per user	97.16 (275.80)	102.28 (280.14)
Outgoing calls' network size per user	20.0 (43.46)	7.8 (23.4)
Distinct outgoing calls locations per user	17.4 (35.62)	7.2 (21.6)
Panel E: Poor		
Number of registered mobile money users	9657	4086
Outgoing calls per user	125.49 (315.09)	138.36 (326.77)
Outgoing calls' network size per user	34.2 (46.09)	19.2 (38.66)
Distinct outgoing calls locations per user	42.8 (55.03)	21.0 (37.79)
Panel F: Rich		
Number of registered mobile money users	54664	26273
Outgoing calls per user	101.65 (280.75)	107.03 (286.06)
Outgoing calls' network size per user	36.0 (59.60)	7.8 (23.4)
Distinct outgoing calls locations per user	28.6 (44.81)	7.2 (21.6)

Notes: Standard deviation is shown in parenthesis.

Table 3.2: Sample statistics for Voice Only Users vs Registered Mobile Money Users (Column 2) and Voice Only Users vs Peer-to-peer Mobile Money Users

In addition to the CDR and mobile money usage related information explained earlier, we have also used three additional datasets as the ground truth. To analyze the patterns of adoption for men and women, we have used the gender information for each of the users as provided by the mobile operator. To explore that how the patterns of adoption vary for urban and rural population, we have specified each of the districts either as urban or rural

based on the urban density according to the last census of Pakistan [2]. To analyze differences in adoption and usage of mobile money services for districts with different socio-economic conditions we have used Multidimensional Poverty Index (MPI) estimates of 2016. MPI measures poverty for each individual as a composite of educational health and living related deprivations. A person is categorized as poor if he or she is facing deprivations in three or more of the ten different indicators.[3].
We have used the district level MPI measurements to specify the districts as rich or poor depends on whether the MPI is lower than the mean value (richer districts) or higher than the mean value (poor districts). We assign each of the users to either urban/rural district or rich/poor districts based on the tower location to which they are most frequently associated.

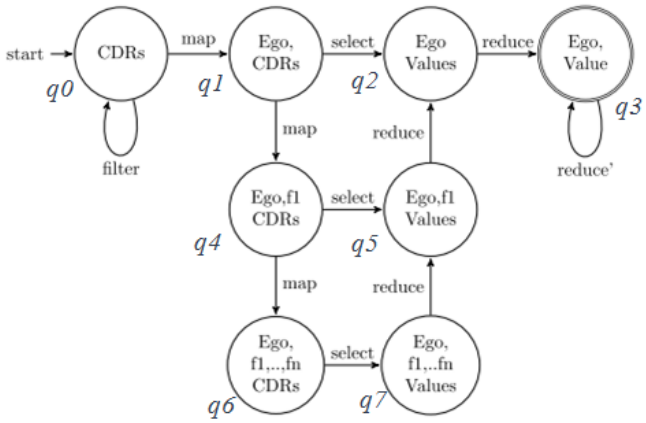


Figure 3.2: Deterministic Finite AutomatonMuhammad R. Khan and Joshua E. Blumenstock, 2016

3.4 Methods

Feature Engineering

Most of the prior work related to the mining of CDR data uses an ad-hoc approach of feature generation where a few hand-picked features are used to analyze the user behavior. Examples of such works include the work by Dong et al. Y. Dong et al., 2014, where the authors have focused on five topological properties of the static social network. Similarly, Gutierrez et al. use two metrics that quantify airtime purchases Gutierrez et al., 2013. Even the more aggressive approaches, like Sundsøy et al., 2014 and CGAP, 2013, which respectively use 350 and 180 CDR-based metrics, do not follow a systematic approach towards feature

[2]Last census was held in Pakistan in 1998 and the updated information about urban, rural mapping is not available

[3]http://www.ophi.org.uk/wp-content/uploads/Multidimensional-Poverty-in-Pakistan.pdf

generation.These approaches though accurate in some cases and interpretable in most of the cases may not be able to handle the complexity of human behavior in many cases. Instead of relying on a few handpicked features, we wanted to use a comprehensive set of features for which we have used the DFA based feature generation algorithm as described in Muhammad R. Khan and Joshua E. Blumenstock, 2016. The DFA-based process of feature engineering (explained briefly in the following section) generates thousands of features that quantify patterns of mobile phone use. Using the features generated through the DFA, our goals were to (a) use these features to understand the determinants of Mobile Money use and (b) build a predictive model that can be used to identify likely adopters

Deterministic Finite Automaton

The DFA we use to generate features from CDR is shown in Figure 3.2. The details of DFA based feature engineering over CDR data is beyond the scope of this chapter, and interested users are referred to Muhammad R. Khan and Joshua E. Blumenstock, 2016 and J. Blumenstock, Cadamuro, et al., 2015 for more details. DFA, in essence, provides a useful grammar for generating the different type of features related to different dimensions of user behavior. For the purpose of better understanding of our approach, some brief examples of DFA based feature engineering are provided here.
As an example, say we are interested in constructing a feature for each individual i that corresponds to, "the variance in the average duration of outgoing calls made by i on different days of the week". We allow for the construction of this feature through the following set of recursive rules:

1. filter *outgoing calls*
2. filter transactions *initiated by i*
3. group by *day of week*
4. focus on *call duration*
5. aggregate by *average* (duration per day of week)
6. aggregate using *variance* (over average daily durations)

Similarly, "the number of unique locations for outgoing calls" can be generated through filter operation to select our calls made by the user; focusing on the cell tower field in the raw data and using count as the aggregate operation. The comprehensiveness of the DFA based feature generation algorithms enables the researcher to compute the network related features as well. For instance, the "size of the incoming SMS network" can be generated by first filtering the incoming SMS messages, focusing on the sender id and then applying the unique count as the aggregate operation.

Features Categorization

Another advantage of DFA based feature engineering is that each feature can be mapped to a particular type or dimension of user behavior. For the purpose of this work, we categorize

the features either as usage related features, mobility related features or network related features,

1. **Usage:** All the features related to the usage of the mobile services (e.g. number of calls made, number of messages sent, number of active days etc). are categorized as the usage related features."Variance in the average duration of outgoing calls made by i on different days of the week" as explained in the last section is also a usage related feature.
2. **Mobility:** The features related to the movement patterns of the users(e.g. the number of different cell towers visited and geographic spread of the network etc.) correspond to mobility related features. "The number of unique locations for outgoing calls" as explained in the last section is an example of mobility features.
3. **Network:** The features corresponding to the structural properties of the network of each of the ego node (e.g. Size of the incoming call network, diversity of network, etc.)

Some of the features can be categorized under more than one category. For example, diversity of communication is dependent both on usage and the network-related features and hence is categorized under both usage and network types.

Experimental Design

We have performed two different type of supervised learning experiments:

- **Voice Only Users vs. Registered Mobile Money Users**
- **Voice Only Users vs. Peer-to-peer Mobile Money Users**

For each of these experiments, we further drew a stratified random balanced sample of adopters/peer-to-peer users and non-adopters for each of the six categories (Males, Females; Urban, Rural; and Rich, Poor). The decision of balanced sample selection is made as most of the users are nonadopters. Summary statistics for each of the samples are shown in the Table 3.2. For each of these experiments, we use gradient boosting algorithm Friedman, 2001 to train the models to classify the users either as voice only or registered mobile money user, and voice only users or peer-to-peer mobile money users.
To understand which CDR-based features are related to Mobile Money use, we calculate the importance of each feature using gradient boosting classifier. The importance of features is the average of importance in each of the individual trees in the gradient boosting. Each of the trees calculates the importance of features using the information gain with intent to favor pure subtrees at each level of the tree.

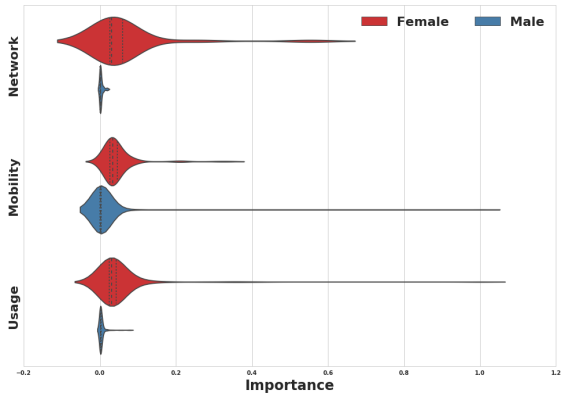


Figure 3.3: Feature importance for different genders

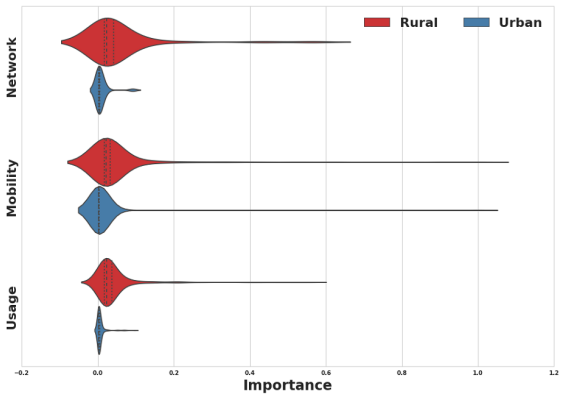


Figure 3.4: Feature importance for urban and rural districts

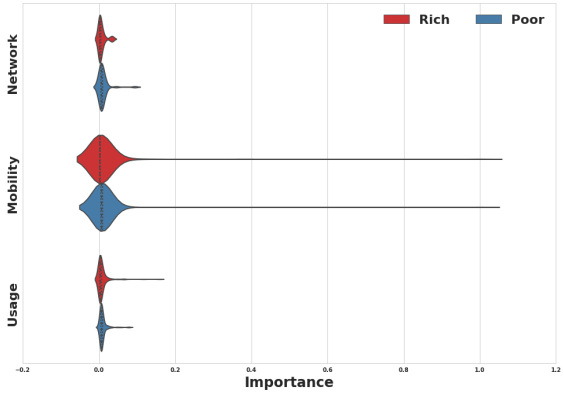


Figure 3.5: Feature importance for rich and poor districts (determined through MPI)

3.5 Results

Determinants of Mobile Money Adoption

Voice only users vs. Registered mobile money users

The DFA described in Section 3.4 produces roughly 16,000 unique features for each of the experiments. However, many of the features can have a strong correlation with each other. The distribution of conditional feature importance values for each of the 16,000 features for different cases is shown in the Figures 3.3, 3.4, 3.5. To construct these figure, we use the feature categorization scheme as explained in the section 3.4 to label each feature as Network, Usage or Mobility related feature. Each violin plot then shows the distribution

of feature importance values for all features of a given type - such as all "usage" features, or all "movement or mobility" features. The usage-related features are simple to map and understand as they mostly correspond to either strength of ties or the differences in usage of the network across different days. Mobility related features include metrics like the number of unique locations visited, the radius of gyration and diversity of places as well. The network related measures correspond more to the structural properties of the network like the size of the network etc. It must be noted that there can be some overlap in features mapping as the strength of ties calculated through number of communication between two individuals can be considered as a network feature as well but in this work I have mapped the tie strength to the usage of the network as many of the existing theories suggest that the more active users are more likely to adopt new services. First thing to note in the Figures 3.3, 3.4, 3.5 is that most of the features are not important, however, different category of features have different aggregate feature importance for different cases. The analysis of features importance for different genders reveals some interesting patterns. Both network and usage related features are a more likely indicator of adoption for females while mobility related features have more predictive power for male users. Usage related features are the most important category for the female users while the mobility related features are the most important one for male users.

Similarly, the analysis of urban and rural districts shows that mobility is the most important category of features for both urban and rural districts. However, compared to rural districts, the network, and usage related features of the users in the urban districts are not that important. Network related features contain features like the size of the network as well as the number of users already using the mobile money services. The importance of the network related features for the rural users indicates that the awareness of the mobile money services, endorsement effects from other mobile money users and the size of the social network all play an important role in the adoption of mobile money services. The trends for the rich vs. poor districts are quite similar to the urban and rural districts with the only difference being that the importance of mobility related features is further pronounced in this case.

Voice only users vs Peer-to-peer mobile money users

The number of registered mobile money users who make peer-to-peer transactions is almost half of the registered mobile money users as shown in the Table 3.2. The features importance for each of the experiments for voice only users vs peer to peer mobile money users is shown in the figures 3.6, 3.7 and 3.8.

Usage related features are the top determinants of peer-to-peer transactions for the female users, while the mobility related features are the top determinant for the male users as shown in the Figure 3.6. Mobility related features are still the top features for the urban and rural districts (3.7) while in comparison to the Figure 3.7 the importance of network related features is lower while the importance of the usage related is higher for the rural district's users. In comparison to the Figure 3.5, Figure 3.8 shows that the mobility related features

are the most important one for the users in the rich districts while the usage related features are the most important determinant for peer to peer transactions in the poor districts. Top features for each of the experiment conducted to analyze the determinants of adoption of mobile money are shown in the Table 3.3.

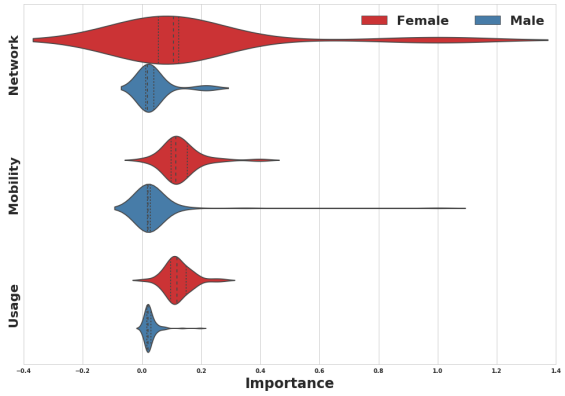


Figure 3.6: Feature importance for different genders

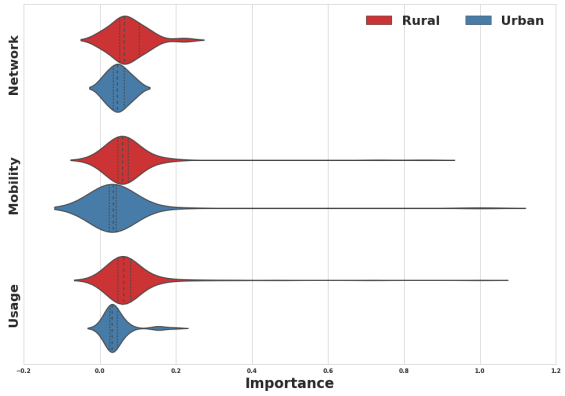


Figure 3.7: Feature importance for urban and rural districts

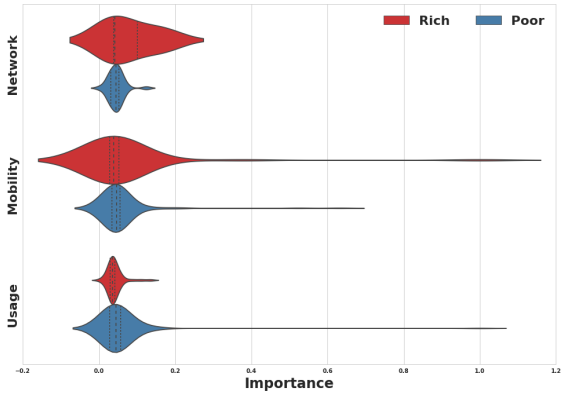


Figure 3.8: Feature importance for rich and poor districts (determined through MPI)

Top features across different experiments are shown in the Table 3.3.

Predicting Mobile Money Use

As discussed in Section 4.4, we used Gradient Boosting based supervised classification models to discriminate between Voice Only and Mobile Money users, using the CDR-based features constructed from the DFA. Cross-validated results from gradient boosting are reported separately for each case in Figure 3.9. The best-performing models included several hundred features, but in practice, there was little difference in performance between models in the

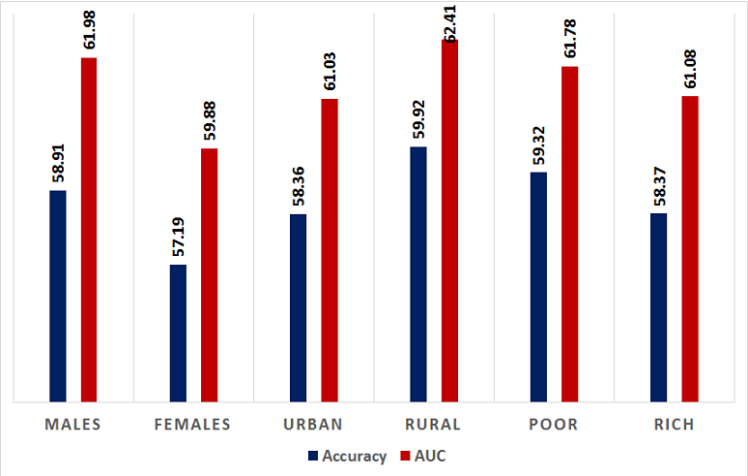


Figure 3.9: Predictive accuracy for mobile money adoption

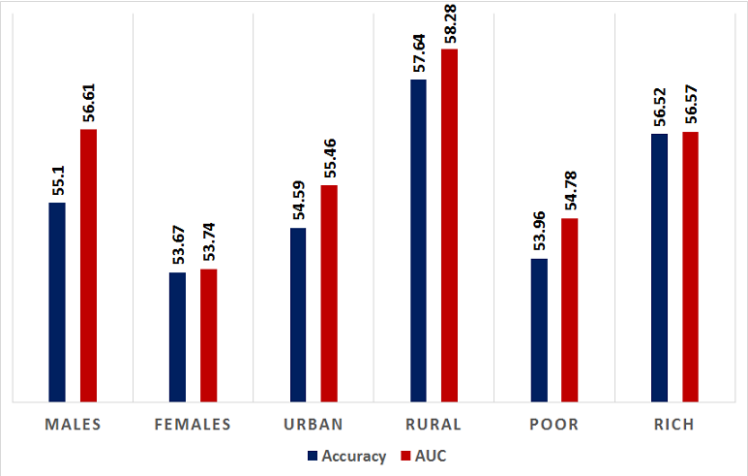


Figure 3.10: Predictive accuracy for peer-to-peer usage

range of 50-1000 features. Figure 3.9 also shows some interesting trends as the performance of our model is higher for males as compared to females. However, the performance of our model to predict the adoption for the rural population and poor districts is higher as compared to the urban population and rich districts respectively. While the trends for the males & female users, urban & rural districts are the same for the peer-to-peer users, the predictive accuracy for the usage of peer-to-peer financial transactions in contrast to the mobile money adoption case is higher in rich districts as compared to the poor districts. As we are using a balanced sample, the improvement of 7-8% over the baseline may seem to be a bit low but first of all it is consistent with the trends found in the earlier work Muhammad R. Khan

Index	Feature	Category
Panel A: Gender: Female		
1	Variance in the number of active days with incoming SMS	Usage
2	Rank Percentile of the Size of the incoming SMS contact network	Network
3	Rank Percentile of the active days with incoming SMS	Usage
Panel A: Gender: Male		
1	Number of unique contact locations on weekdays	Mobility
2	Variance in the contact locations per day	Mobility
3	Number of unique contact locations on weekend	Mobility
Panel C: Rural		
1	Number of unique contact locations on weekend	Mobility
2	Percentile of the network size on weekend sms	Network
3	Variance in the number of sms per active day	Usage
Panel D: Urban		
1	Number of unique contact locations on weekdays	Mobility
2	Variance in the duration of the weekend calls to each contact locations	Usage
3	Size of the outgoing calls network on the weekdays	Network
Panel E: Poor		
1	Number of unique contact locations on week days	Mobility
2	Size of the network per contact locations per day	Network
3	Variance in the number of contact locations per active day	Mobility
Panel F: Rich		
1	Number of unique contact locations on the weekdays	Mobility
2	Number of unique contact locations on the weekend	Mobility
3	Number of SMS on the weekdays	Usage

Notes: Rank percentile for a feature indicate the percentile of the feature value for the subscriber as compared to all the contacts.

Table 3.3: Top features for mobile money adoption

and Joshua E. Blumenstock, 2016 and secondly, even this margin of improvement in targeting customers likely to use mobile money can result in improvement of lives of hundreds of thousands of users.

3.6 Discussion

Taken in the broader context of research into the determinants of Mobile Money adoption and use, the preceding results uncover several unexpected patterns. Superficially, it is not surprising that CDR-based metrics can be used to construct classifiers that predict Mobile Money use, though to our knowledge this is the first study to explore the dynamics of adoption across different demographics in a developing country. The baseline or random

prediction accuracy would have been 50 % as we are using a balanced sample so at first glance at the Figures 3.9 and 3.10 may indicate that we do not have a lot more improvement from the baseline. But as the ultimate aim is to use this model to predict adoption of mobile money for all the customers for the telecom operators which has the subscriber base of around 30 million users, even a 10% increase in modeling the likelihood of adoption can result in hundred of thousands of users adopting financial services (and in turn having better life).

Also interesting are the differences in performance of the different type of features. The importance of features casts some light on the social norms especially in the case of female users. For example, the usage category was the top one for the female users but not for the male users. This indicates that male who may be working remotely or have to commute to work are more likely to use the mobile money like services, whereas the women who are more active or have higher technology literacy are more likely to user mobile money. Similarly, the tendency of more mobile users in both urban & rural and rich & poor districts indicates that mobility of the users is the prime determinant when it comes to the adoption of mobile money. However, the people in rural districts with a bigger network are also more likely to use mobile money. This indicates that network effects can be much more important in the rural areas.

The decrease in predictive accuracy for female users from 57.19% for mobile money adoption to 53.67% for peer-to-peer usage shows that there can be much more complex patterns of peer to peer usage. Same is true for other segments of the society. Table 3.2 shows that the number of registered mobile money users who make peer to peer transactions is roughly half the number of registered mobile money users. This indicates that even the users who are using mobile money may not be using it to its full potential and this highlights the required improvement in the marketing strategies of the operator as well.

There is a general tendency in the marketing section of the telecom operators to target more active users for more innovative services. However, our results show that across different sections of the society different trends may have more explanatory power when it comes to predicting adoption and usage of mobile money. This also highlights the fact that these telecom companies need to user intelligent marketing solutions aligned with the needs and behaviors of people of different demographic backgrounds. The dominance of mobility related features for male users as compared to usage related features for female users indicates the prevalent social norms of the society as well.

3.7 Conclusion

To our knowledge, this is the first attempt at analyzing the determinants of adoption of mobile money across different sections of the society using large data sets. And there are some interesting things that can be concluded from this work. First of all, the machine learning based approaches are becoming ever more important for promoting the adoption of different products in both the developed and undeveloped countries. The difference in relative impor-

tance of the different type of features highlights the fact that integrating the demographic profiles of the users in the predictive model can further increase the performance of the model. Furthermore, this work also shows that digital financial services need to take care of the requirement of different segments of the society and align their marketing schemes according to these requirements as well. This work also opens up many interesting questions for future research as why mobility related features are the most important determinant of mobile money adoption for the male users, etc.

Looking forward, it would be interesting to see that how do patterns of adoption and usage of mobile money services vary for people of different age groups and working status. Additionally, people in different countries may have different requirements and patterns of adoption and we intend to apply our current methods on data from other developing countries to see how much of the patterns of adoption are similar to those in Pakistan.

Chapter 4

Churn Prediction

Abstract

Churn prediction, or the task of identifying customers who are likely to discontinue use of a service, is an important and lucrative concern of firms in many different industries. As these firms collect an increasing amount of large-scale, heterogeneous data on the characteristics and behaviors of customers, new methods become possible for predicting churn. In this chapter, we present a unified analytic framework for detecting the early warning signs of churn, and assigning a "Churn Score" to each customer that indicates the likelihood that the particular individual will churn within a predefined amount of time. This framework employs a brute force approach to feature engineering, then winnows the set of relevant attributes via feature selection, before feeding the final feature-set into a suite of supervised learning algorithms. Using several terabytes of data from a large mobile phone network, our method identifies several intuitive - and a few surprising - early warning signs of churn, and our best model predicts whether a subscriber will churn with 89.4% accuracy[1].

[1]The material in this chapter is based on the joint work with Anikate Singh, Joshua Cherian Manoj and Joshua Evan Blumenstock. See (Muhammad Raza Khan, Manoj, et al., 2015) for more details

4.1 Introduction

A common expression in industry is that it costs five times more to acquire a new customer than it does to retain an existing one Hart et al., 1990. In sectors ranging from finance and insurance to cable service and internet dating, significant sums of money are spent in reducing customer attrition and retaining hard-won customers. To prevent such attrition – also commonly referred to as *churn* – it is critical to be able to identify the early warning signs of churn, and to accurately target those customers who are most likely to attrite.

In this chapter, we describe a quantitative and computational framework that can be used to find leading indicators of churn, and identify individuals with a high likelihood of churning. Our approach is data-centric, in the sense that the intent is to "let the data speak for themselves." By using simple machine learning algorithms to mine historical transaction records, this method discovers behavioral patterns that are empirically correlated with the propensity to churn. This is in contrast to the more traditional approach taken by many companies, where churn strategies are determined by the intuition of key individuals, or through direct feedback from customers.

The framework has two primary components. The first is designed to identify early warning signs of churn by isolating specific and easily-measured behavioral patterns that are highly correlated with churn. One such pattern we find, which is indeed highly correlated with churn, is the total amount of observed activity of a customer. It should come as no surprise that customers with low levels of activity are more likely to churn than customers with high levels of activity. However, this particular metric is one among thousands of other metrics that are correlated with churn, and the aim of the framework is to zero in on the most predictive metrics. The method we develop is a semi-supervised, brute force approach to feature engineering, in which our algorithm first constructs tens of thousands of features through combinatoric feature generation, then uses established techniques for feature selection to prune the long list down to a manageable and interpretable subset of the most predictive behavioral traits.

The second component in our framework is designed to construct a "Churn Score" that assigns a probability to each customer indicating the predicted likelihood that the particular subscriber will churn within a predefined period of time. Whereas the focus of the first stage is interpretability, in the sense that the behavioral traits identified can be easily understood, the focus of this second stage is on predictive accuracy. The goal is to create a metric that could, for instance, be used by a marketing department to target promotions to customers who are on the brink of defecting to a competitor. To accomplish this task, we utilize a "kitchen-sink" approach to supervised learning in which the full set of thousands of features are passed through basic classifiers, with parameters selected through cross-validation, in order to optimize standard metrics of predictive performance. While the models themselves are difficult to interpret, the Churn Scores they produce are highly accurate, and as we show they are remarkably robust to the set of features used and the machine learning algorithms implemented.

We test and calibrate this framework on a large dataset from a mobile phone operator

in South Asia. Starting with a raw dataset of several billion transactions, spanning roughly ten million prepaid mobile phone subscribers over a period of multiple years, we extract a calibration dataset consisting of all network-based communication for roughly 100,000 subscribers over 6 months. On this dataset, where the natural churn rate during our evaluation period is roughly 24 percent, our method is able to predict customer churn with just under 90 percent accuracy.

The remainder of the paper is organized as follows: in the following section, we discuss related research on churn prediction. Section 4.3 describes the data, and section 4.4 describes the methods in greater detail. Results are presented in section 4.5, and we conclude with a discussion of the strengths and weaknesses of our approach in section 4.6.

4.2 Related work

Understanding why customers terminate relationships has been a focus of marketing research for several decades Jain and Singh, 2002. In recent years, as data on customer activities and characteristics becomes increasingly available to companies, more sophisticated metrics have evolved to describe customer behavior and better understand how behavioral traits can be linked to customer retention and firm performance Gupta and Zeithaml, 2006. Spurred in part by a competition associated with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, churn prediction has received recent attention from the applied machine learning community. These approaches have tested a battery of models including expert systems Wei and Chiu, 2002, support vector machines Archaux et al., 2004, and bagging and boosting Lemmens and Croux, 2006, to name just a few. They further vary in terms of the approach to the data, with some focusing on customer profiles and features Qian et al., 2006, and others concerned primarily with the importance of social ties and social structure Dasgupta et al., 2008; Bonchi et al., 2011; Karnstedt et al., 2011.

Zhang et al., 2012 provide a recent overview of the different types of subscriber attributes used to model and predict customer churn in prior work, and Verbeke et al., 2012 benchmark several classification techniques for prediction. Neslin et al., 2006 discuss the importance of different methods for predicting churn in the context of a public tournament between 33 different competitors. In these and related studies, the behavioral traits are pre-computed – for instance in the contest described by Neslin et al., 2006, each contestant was given a curated dataset with 171 predictor variables. By contrast, a focal point of our study is on the process of generating these predictor variables from the raw transactional records. Specifically, as discussed in Section 4.4, we describe an "accordian" method for first engineering a large set of features and then collapsing this large feature spaces into a smaller set of relevant and predictive metrics.

The empirical setting for our study – a prepaid mobile network in South Asia – is further distinct from almost all previous work on churn prediction in two ways. First, in the network we study, churn is a relatively common event. As we discuss in greater detail, we observe an approximate churn rate of 25 percent every two months, which is an order of magnitude

higher than the "rare-event" churn discussed by most previous research.[2] This necessitates a slightly different empirical framework, and one where neither precision nor recall is of higher priority ex ante. Second, the vast majority of studies of churn in telecommunications focus on a more canonical formulation of the problem in the context of a post-paid network. In this traditional post-paid setting, churn is equivalent to the cancellation of a contract, and requires that the customer take action. By contrast, our empirical setting is one of pre-paid accounts, where churn simply means that the subscriber discontinues use of the network, and must be inferred by a prolonged period of (passive) inactivity. This distinction has important implications for how we quantify churn and measure the accuracy of our Churn Score.[3] These subtleties are discussed in greater detail in Section 4.4.

4.3 Data and Preprocessing

We study customer churn in the context of a mobile phone use in South Asia, using data from one of the largest mobile network operators in the country.[4] The operator has roughly ten million active subscribers, and has provided us with access to an anonymized database containing several years of transactional communication histories. The transactions on which we focus are the Call Detail Records (CDR), simple metadata logs that record entries for every event that transpires on the network. Thus, for every call or text message (SMS) received, we observe the anonymized identity of the initiating and receiving party, a date and timestamp, information on the approximate physical location of the two parties, and additional metatdata describing the communication event.

The unprocessed CDR that we receive capture hundreds of billions of transactions, and is several terabytes in size. It is worth noting, however, that while the CDR contain a very large number of observations, each individual observation contains very limited information. Thus, a critical portion of the analyiss, which we describe in detail below, involves transforming the original transactional histories into a set of interpretable metrics that can be used to understand and predict customer behavior. To perform such transformations, we use a Map-Reduce framework that leverages open source distributed computing environments including Apache Hadoop, Apache Spark, and Hive.[5]

[2]For instance, Lemmens and Croux, 2006 report a churn rate of 1.8% per month for a major U.S. wireless telecommunications carrier. Verbeke et al., 2012 provide an excellent summary of 11 other datasets used in churn prediction research, all but one of which have churn rates of less than 7 percent.

[3]Dasgupta et al., 2008 provide a noteworthy exception, focusing on a prepaid operator with annual churn rates between 50 and 70 percent.

[4]To protect the competitive interests of our data partner, specific information on the context and subscriber base is kept confidential.

[5]Available at `http://hadoop.apache.org/`, `https://spark.apache.org/`, and `https://hive.apache.org/`.

Property	Value
Number of Unique Subscribers	98,514
Total Number of Days	183
Date Range	Jan. 2011 - Apr. 2012
Total Number of Calls	69,804,577
Total Number of SMS	137,174,182
Average (Standard Deviation) of calls per subscriber	20.3 (317.2)
Average (Standard Deviation) of SMS per subscriber	386.5 (1729.12)

Table 4.1: **Summary Statistics**. Values in table are calculated for the subset of Call Detail Records used for training and testing as described in Figure 4.1.

4.4 Methods

An overview of the framework we develop, which allows us transform the raw transactional records into a set of early warning signs of churn, as well as a subscriber-specific Churn Score, is shown in Figure 4.1. We begin by randomly selecting a subsample of roughly 100,000 subscribers from the full mobile phone subscriber base, and extract all transactions in which they are involved. This sample is the primary dataset we use for development and benchmarking (labeled as *Random Sample Selection* in the figure). Basic summary statistics for this subset of the data are given in Table 4.1. For this subset of subscribers, we then generate a large number of aggregated metrics that describe a wide range of inferred behavioral characteristics (*Feature Engineering*). The resultant dataset, which we use for training and testing the core algorithms, is a balanced rectangular matrix containing approximately 10,000 features for approximately 100,000 individuals. With this matrix, we then separately isolate the handful of metrics that are most predictive of churn (*Feature Selection*), and develop a Churn Score that indicates the likelihood that a subscriber will churn in a 3-month period (*Machine Learning*). We describe each of these steps in detail below.

Feature Engineering

In constructing a set of behavioral metrics from the raw transactional data, we employ a data-centric approach that is intended to minimize the role of the analyst in determining which features are relevant to the task of modeling customer churn. To this end, we use a combinatoric "brute force" technique that defines a feature along K different axes (where the i^{th} axis is of dimension D_i), then creates all possible combinations of such dimensions. This results in a total of $\prod_{i=1}^{K} D_i$ unique features.

In our case, we use eight different axes, each of which has between two and seven dimensions. A schematic of the feature space covered by these axes is given in Figure 4.2. Most of these axes are intuitive: we divide communication by total activity, which captures the total

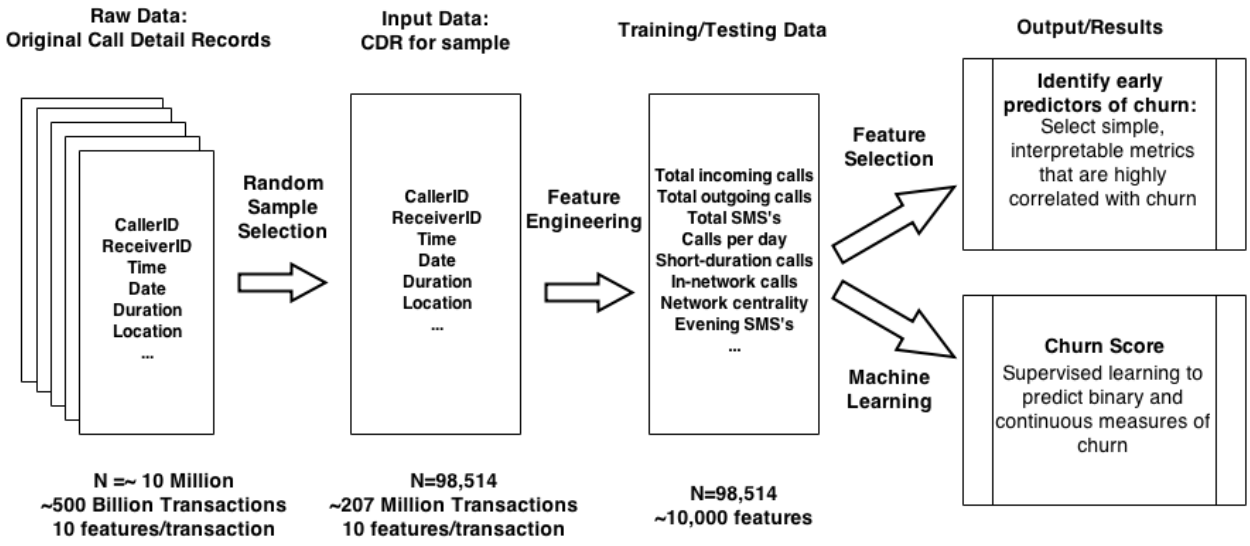


Figure 4.1: **Overview of approach.** Full details of each step in the processing pipeline are provided in Section 4.4.

number of communication events of a certain type, and by the 'degree', which indicates the number of unique alters with which a given ego communicates. We further divide by the type of communication (call vs. SMS), the time of day, the direction of communication, the status of the alter (for instance whether they are an international number, a local number, a special number such as the operator, a number on a different operator's network, etc.), and the duration of the event (only relevant for calls). Additionally, we segment features based on the total amount of activity over the entire observation window, as well as by the trend in activity from month to month, looking at the average monthly change in activity, as well as the maximum and minimum monthly change. Finally, we allow for each feature to be normalized with different denominator: "1" indicates an un-normalized feature, while the other dimensions of the "Denominator" axis indicate that the metric is normalized by the total number of unique days on which the subscriber is active on the network, by the total number of calls made, etc.

This brute force approach is, of course, not entirely unsupervised, as it is still the responsibility of the analyst to define the axes of importance and the dimensions of each axis. In our case, we chose a set of axes that were both easy to compute and which we felt provided reasonable coverage of the space of features relevant to churn. There is, however, a much larger feature space that one could cover, time permitting, which would capture a much richer set of metrics including measures of social network structure, mobility and migration patterns, and more subtle temporal dynamics.

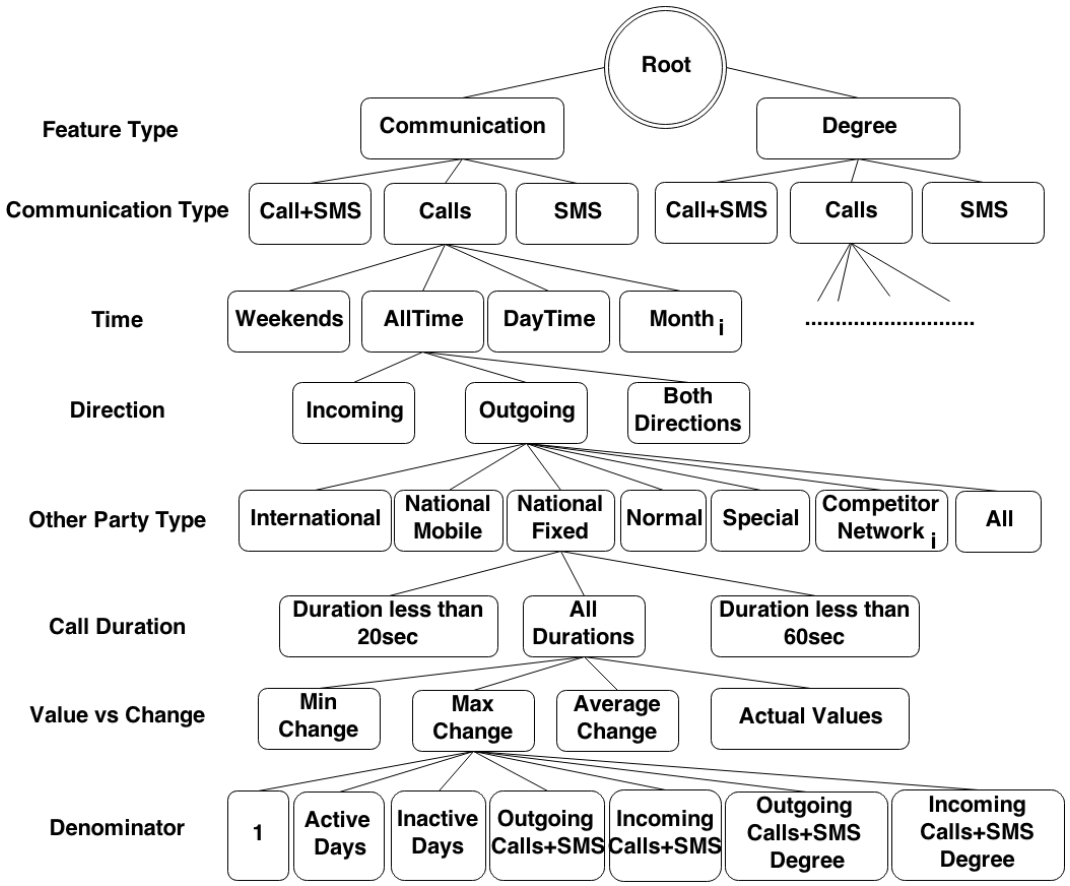


Figure 4.2: **Tree used for combinatoric feature generation**. A total of 12,914 features are generated following the method described in Section 4.4. Here, we show an example expansion of one path in the tree. Each node can be expanded downward in a manner analagous to the way the sample node at the same level is expanded.

Feature Selection

Given the large number of features generated through the process describe above, we then employ standard methods of feature selection that determine which features are most correlated with customer churn. We do this in two ways: We first use compute a t-test separately for each individual feature, which indicates the extent to which that single feature can accurately differentiate between people who churn and people remain active. Separately, we use a tree-based method for feature selection that allows us to estimate the conditional ability of each additional feature to improve the overall accuracy of a joint classifier Geurts et al., 2006; Hastie et al., 2009. In Section 4.5, we will compare and contrast the results, as well as the advantages and disadvantages of these two methods for feature selection.

Churn Prediction and Churn Scoring

The process of feature selection describe above is useful in producing an interpretable list of metrics that are correlated with customer churn. Our second goal is to use these features to predict churn. Specifically, we seek to assign each subscriber a Churn Score between zero and one that indicates the likelihood that the subscriber will churn. To this end, we set up a simple supervised learning experiment where we divide our data into a training period and an evaluation period, derive features using the above process during the training period, and use these features to predict churn during the evaluation period. Using k-fold cross-validation, we then train several different supervised learning algorithms on randomized subsets of individuals. The specific algorithms we test are linear and logistic regression, Support Vector Machines, K-Nearest Neighbors, Random Forests, and AdaBoost (with a decision tree classifier);[6] as we will see below, the choice of learning algorithm has only a marginal impact on the accuracy of the Churn Score.

This experimental setup raises one empirical subtlety: How do we define "churn" when no cancellation events are directly observed? In other contexts, churn is defined by an action taken by a customer, such as the cancellation of a membership or the failure to renew by an established deadline. In our case, however, all of the mobile phone subscribers use prepaid accounts, so "churn" simply means that the subscriber has stopped using the network. A common assumption made in industry is that a customer is considered lost after 2-3 months of continuous inactivity; thus, in some of the experiments described below we construct a binary measure of churn that approximates this condition. Alternative formulations of churn (sometimes reffered to as the "always a share" retention model Dwyer, 1997) consider the inactivity as transient; to capture this we will separately create a continuous measure of inactivity that reflects relative likelihood of churn.

4.5 Results

We now present the results from testing the churn analysis framework on the training dataset of 98,514 unique prepaid mobile phone subscribers. We set up our experiment using a 6-month period of data, where we treat the first 4 months as a training period used to construct features, and use the final 2 months as an evaluation period. We quantify churn in two ways: first, as a binary condition that is true if the subscriber is completely inactive for the entire 2-month window. In total, 26 percent of our subscribers fit this stringent definition of churn. Second, we define a more flexible version of churn as the percentage of days on which no activity is observed. The distribution of inactivity for our sample can be seen in Figure 4.3. While there are a large number of individuals who are completely inactive, and roughly 13 percent are active on every day, the majority of subscribers fall in between the two extremes.

[6] See Verbeke et al., 2012 for a more thorough exposition of the impact of different learning algorithms on churn prediction accuracy. Additional details on the algorithms and their implementation can be found in Hastie et al., 2009.

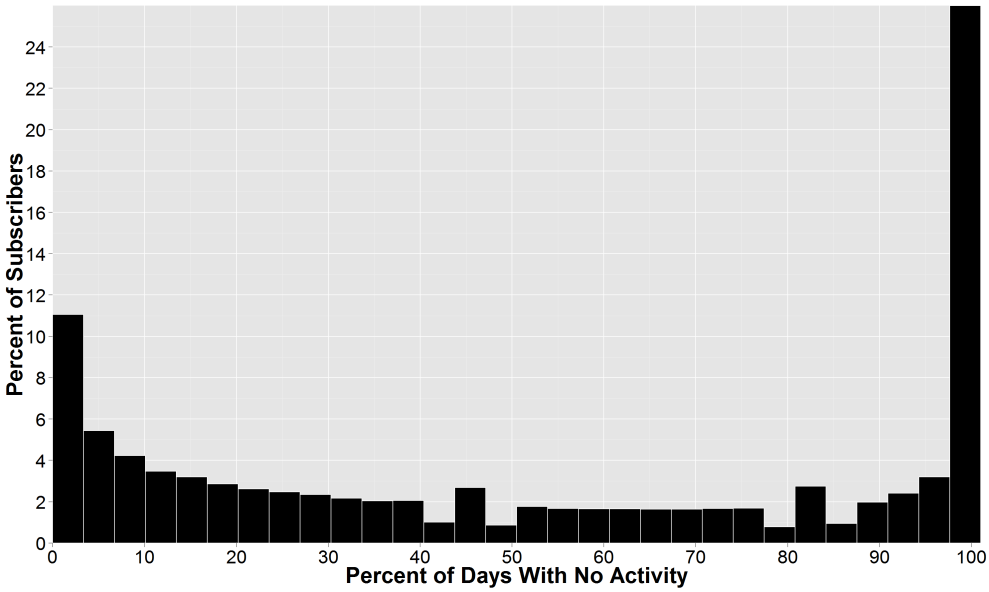


Figure 4.3: **Distribution of Inactivity.** In the evaluation period, there are a large number of individuals who are completely inactive, a mass of people with very high levels of activity, and an even distribution between the two peaks.

Results: Early predictors of mobile subscriber churn

Using the combinatoric approach described in Section 4.4, we generate 12,914 unique features, each of which quantifies a different type of mobile phone-based behavior. In Panel A of Table 4.3, we list the 10 features that, taken in isolation, are most predictive of churn. For the ease of exposition, we focus initially on the continuous definition of churn, and indicate in the right column the R^2 that results from separately regressing the proportion of days with no activity on each feature. Unsurprisingly, we find that "Percent of inactive days" – a feature indicating the fraction of days during the training period when the subscriber had zero transactions – is highly predictive of future churn. People who are inactive in one period are likely to be inactive in the next period. The next highest ranked feature is perhaps less obvious: we find that high variance in call activity (measured as the maximum month-to-month change in the ratio of incoming to outgoing calls), is strongly predictive of churn.

While the set of features listed in Table 4.3 are all unconditionally highly correlated with churn, these features are also correlated with one another. Thus, while the "Percent of inactive days" feature may be the best single linear discriminant between churners and non-churners, it is not necessarily the case that the ensemble of 10 features in Panel A will be the best *joint* predictor of churn. Thus, in Panel B of Table 4.3, we provide the 10 features

Model	Accuracy	Precision	Recall	F-Score	AUC
SVM	89.4	0.89	0.89	0.89	0.91
Random Forest	88.4	0.88	0.88	0.88	0.92
KNN	88.2	0.88	0.88	0.89	0.89
AdaBoost	89.2	0.89	0.89	0.89	0.94
Logistic Regression	89.3	0.89	0.89	0.89	0.93
Baseline Model	83.9	0.83	0.84	0.83	0.89

Table 4.2: **Performance.** Different metrics of performance in predicting binary churn for the baseline model and five learning algorithms.

that are, taken together, the best joint predictors of subscriber churn. For comparison with Panel A, we also list the R^2 from the unconditional (univariate) regression for each of the features, though it is important to note that this is the unconditional R^2 and is different from the criteria used to rank-order features in Panel B. The full distribution of R^2 values for all 12,914 features is shown in Figure 4.4.

While it is perhaps perilous to read too much into the list of the top ranked features in Table 4.3, as these are sampled from a full list of over 10,000 features, we make two observations. First, the unconditionally predictive features (Panel A) seem quite similar and generally reflect aggregate metrics of activity. On the other hand, the conditionally predictive features (Panel B) tend to be micro-aggregates; the final set is thus less redundant and less mechanically correlated. Second, several of the most predictive features are non-obvious. For instance, the 9th feature in Panel B is not one we would have expected ex ante; it is not immediately obvious why the ratio of incoming international text messages on weekends should be a better joint predictor than any other feature on the list, let alone the same metric normalized by the total incoming network degree of the individual. Ex post, it may be possible to concoct a story that explains the predictive power of this statistic, but it is sufficiently idiosyncratic to likely elude even the sharpest marketing director.

Results: Predicting churn

In order to evaluate the performance of our predictive framework, we first construct a simple and intuitive baseline that builds on the results of Section 4.5. Namely, we select the single best unconditional predictor of churn, "Percent of inactive days", and build a linear discriminant model based on that feature. Namely, we first calculate this metric for each of 98,514 unique mobile subscribers during the 4-month training period. We then test all possible threshold values of this metric to determine the value that most accurately separates churners from non-churners, where we now define churn as a binary indicator that takes the value one for churners and zero for non-churners. As can be seen in Figure 4.5, a maximum

accuracy of 83.9% for this metric is obtained at a threshold of 76. In other words, when we classify as churners all subscribers who were inactive on more than 76 percent of days during the training period, our prediction is correct in 83.9% percent of cases. As a relatively naive baseline, this simple linear discriminant actually performs quite well. This is in part due to the fact that we have an unbalanced sample, in the sense that 76.6 percent of our sample do not churn, and a very simple model that just predicted the majority class (i.e. "not churn") for all subscribers would achieve 76.6 percent accuracy.

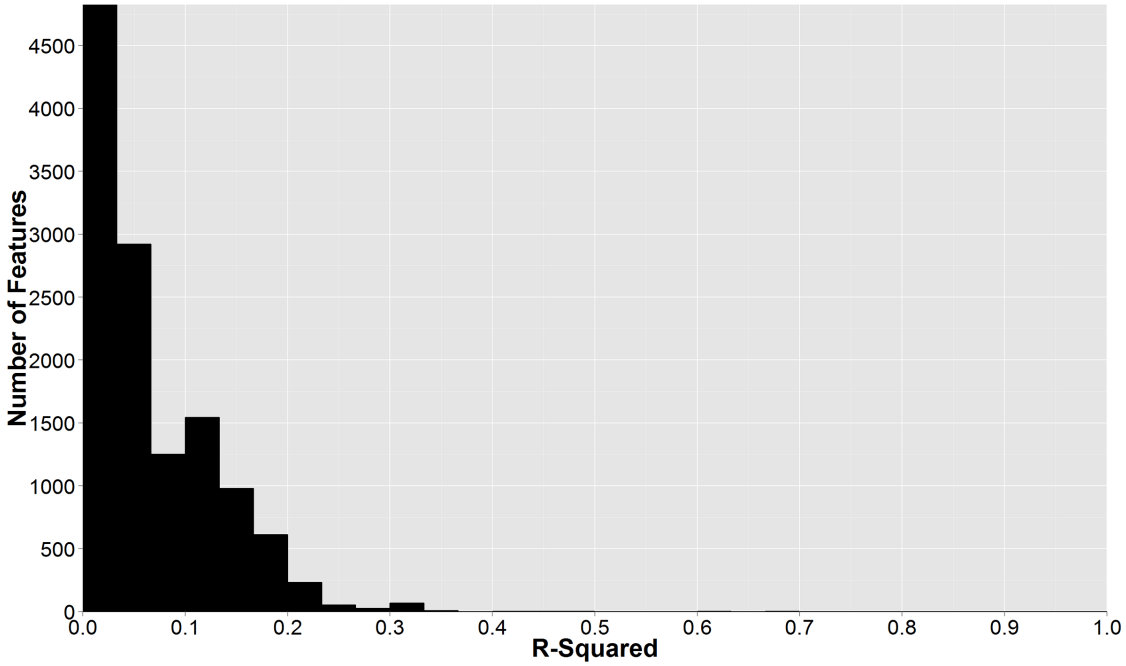


Figure 4.4: figure

Histogram of R^2 values. Distribution of R^2 values resulting from univariate regression of continuous churn metric (percent of inactive days) on each of 12,914 different features.

Depending on the algorithm used to predict churn, we achieve accuracy rates of roughly 88.5-89.5 percent. This represents a modest improvement of roughly 6 percent over the single-feature baseline, or approximately 14 percent over the majority-class baseline. A variety of performance characteristics of each model, as well as the linear discriminant baseline, are given in Table 4.2; the ROC curves for these models is presented in Figure 4.6. Area under the curve is consistently high, and the ROC curves show improvements over the baseline, particularly at low thresholds where the baseline has trouble identifying true positives. For each model, we use the set of 100 features selected via the bagging approach described in Section 4.4 (including the 10 features listed in Panel B of Table 4.3). We experimented with using a different number of features, but as can be seen in Figure 4.9, performance was actually quite robust to the number of features used. While modest improvements in

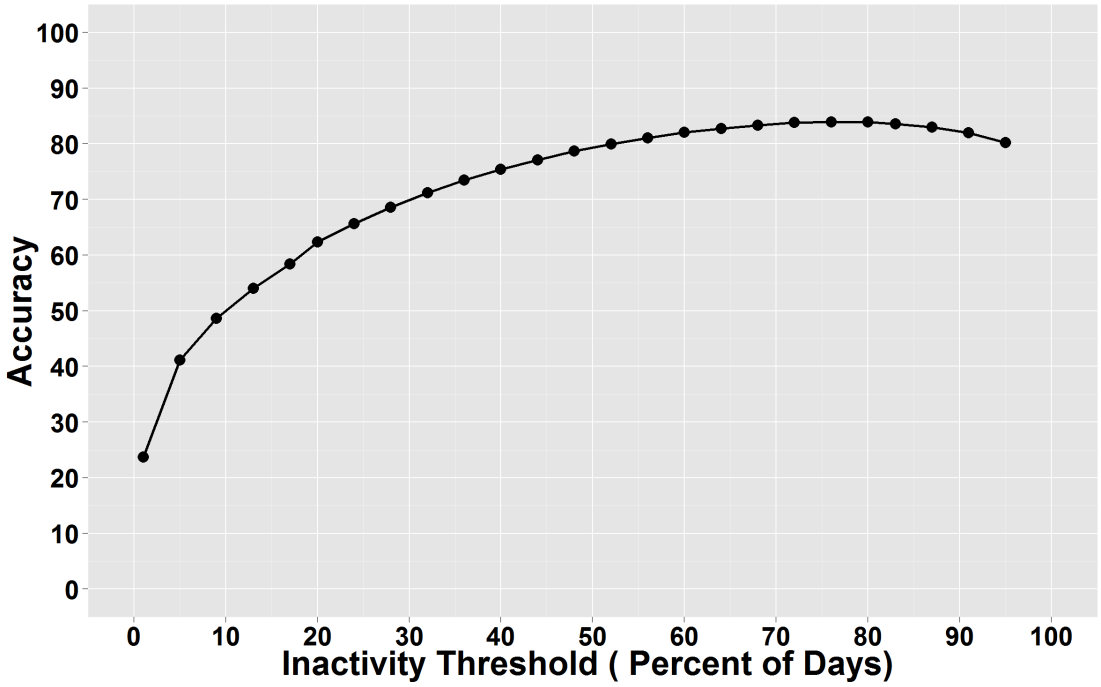


Figure 4.5: **Churn prediction with a single linear discriminant.** Predictive accuracy depends on the threshold used to classify subscribers. For the "Percent of inactive days" feature, maximum accuracy of 83.9% is obtained at a threshold of 76%.

accuracy result until roughly 50 features are incorporated into the model, no additional improvements result from adding additional features.

Similar results obtain when considering churn as a continuous measure of inactivity. This metric, which we denote as the "Churn Score", indicates the percentage of days in which the subscriber is predicted to be inactive. Individuals with high Churn Scores are the likely churners. Figures 4.7 and 4.8 show prediction accuracy, measured as the absolute difference between the predicted percentage of inactive days and the actual percentage of inactive days, measured for each subscriber during the evaluation period. Across all five learning algorithms, the Churn Score based on 100 features is highly predictive of actual subscriber churn.

4.6 Discussion and Conclusion

In this chapter, we have presented a simple framework for detecting the early warning signs of churn, and for developing a Churn Score to identify subscribers who are likely to end their relationship with the company. The approach uses a brute force approach to feature

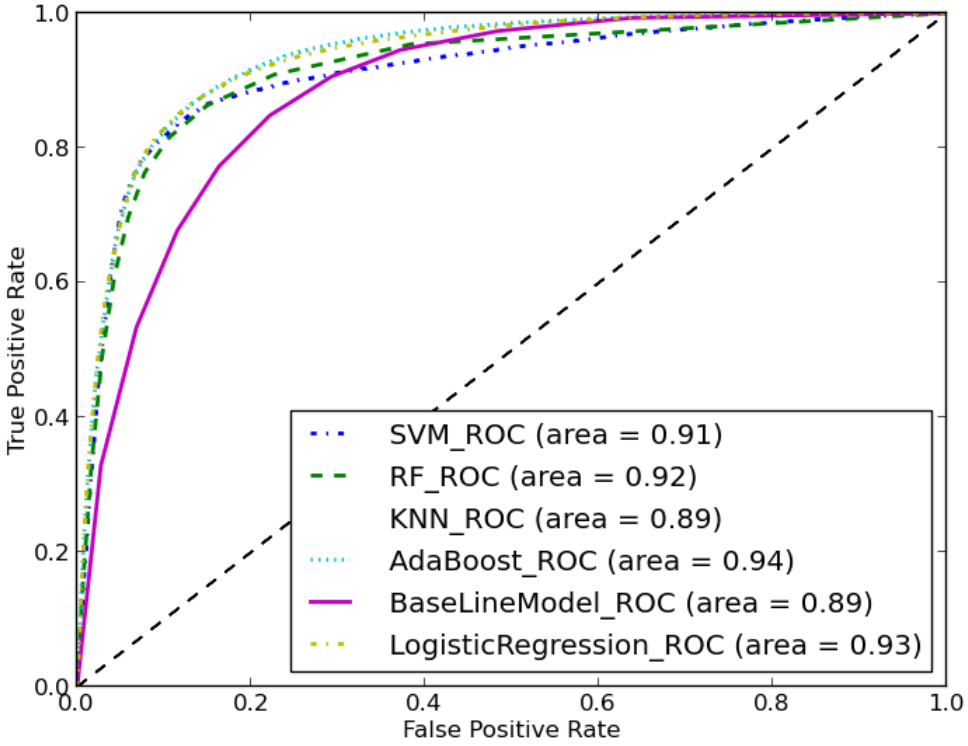


Figure 4.6: **ROC curves**. Performance of five different binary classifiers and one baseline at predicting churn.

engineering that generates a large number of overlapping features from customer transaction logs, then uses two related techniques to identify the features and metrics that are most predictive of customer churn. These features are then fed into a series of supervised learning algorithms that can accurately predict subscriber churn. Testing this approach on several terabytes of data from a South Asian mobile phone operator, we highlight a few of the early predictors of churn, and show that relatively simple models predict churn with close to 90 percent accuracy.

While these initial empirical results are promising, we see the primary contribution of this chapter being the description of a systematic framework that can be used to generate interpretable features and predict customer outcomes. Several of the modeling assumptions we have made, such as the axes and dimensions used to generate features, are quite arbitrary and it is likely that more careful design of these behavioral metrics could yield more intuitive predictors and more accurate predictions. Additionally, the two methods for feature selection which we employed, while effective, leave considerable room for improvement. For instance,

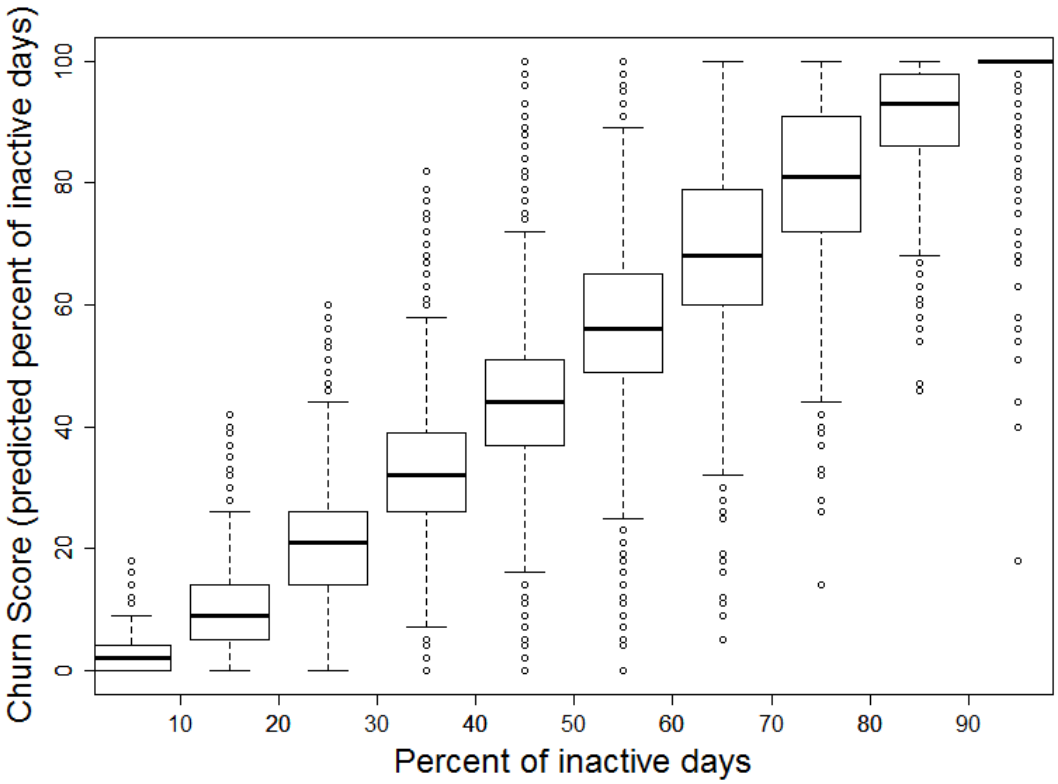


Figure 4.7: **Actual vs. Predicted percentage of inactive days.** For each decile of subscriber activity, we show the distribution of predicted inactivity – the Churn Score – as predicted by Random Forest algorithm.

there could be profitable gains from constraining the search of the feature space to predictors that might more easily be recognized or utilized by a marketing department for targeting and segmentation.

In future work, we are most actively interested in methods that more systematically explore the possible feature space, and which require even less guidance from the analyst to generate a set of relevant and intuitive features. Because the large set of features generated by our brute force approach were somewhat contrived to show the predictive power of the combinatoric algorithm, and also because of confidentiality agreements with the data provider, we have largely refrained from interpreting the set of top-ranked features. However, as these early warning signs have the potential to provide considerable insight into the nuances of customer behavior that are fundamentally related to churn, we believe a more systematic exploration of these results would provide fertile ground for future research.

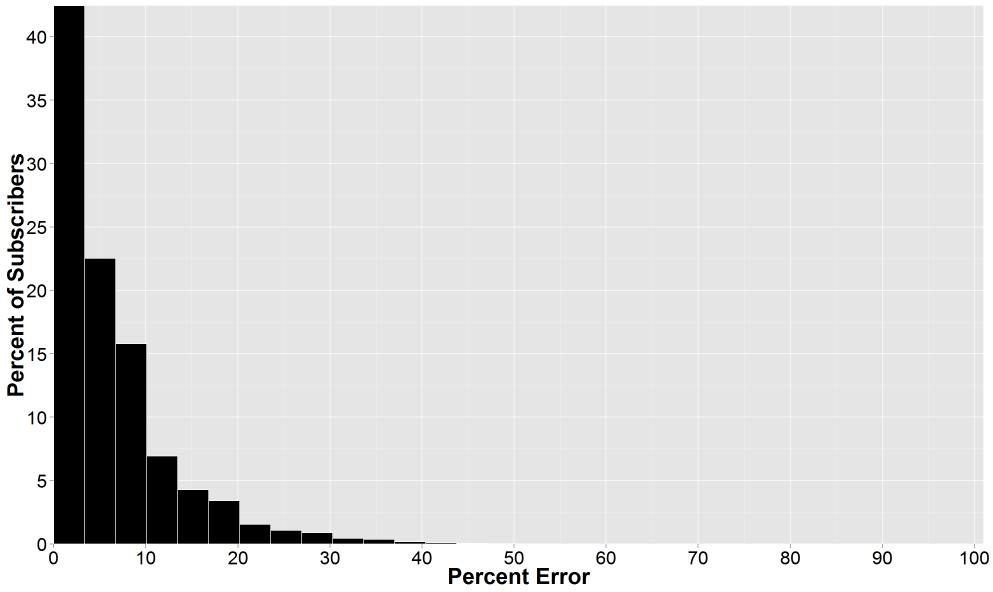


Figure 4.8: **Distribution of prediction errors**. Distribution of errors (absolute value of Actual inactivity minus Predicted inactivity) using the Random Forest algorithm.

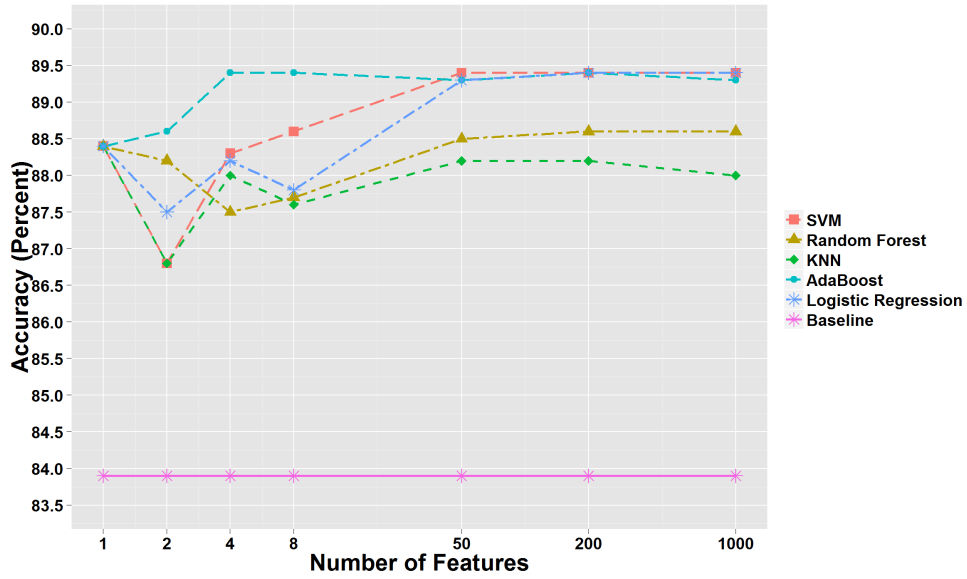


Figure 4.9: **Sensitivity to size of feature set.** While additional features generally improve the accuracy of predictions, diminishing returns are achieved after roughly 50 features are used.

Rank	Feature	R^2
	Panel A: Tested individually	
1.	Percent of inactive days (during training period)	0.66
2.	Maximum monthly Δ in incoming calls / Total incoming calls	0.44
3.	Maximum monthly Δ in incoming calls / Total outgoing calls	0.42
4.	Outbound network degree (most recent month)	0.35
5.	Incoming text messages received from competitor's network	0.33
6.	Average number of calls to Information Portal per active day	0.33
7.	Unique weekend contacts per active day	0.33
8.	Average daily text messages received from competitor's network	0.33
9.	Number of Inactive Days in the First Month of the Training Period	0.31
10.	Daytime degree (voice calls)	0.25
	Panel B: Tested jointly	
1.	Outgoing degree (most recent month)	0.35
2.	Outgoing degree (first month) / SMS Degree	0.09
3.	Incoming degree (second month) / Total incoming calls	0.14
4.	SMS to Mobile Money Service / Total days of inactivity	0.01
5.	Short-duration calls (first month) / Total incoming calls	0.24
6.	Incoming calls / Incoming events	0.12
7.	Calls to Mobile Money servce (first month) / Total days of inactivity	0.32
8.	Outgoing events (first month) / Call degree	0.22
9.	Total incoming int'l SMS (weekends) / Incoming degree	0.18
10.	Total outgoing degree (first month) / Call degree	0.22

Table 4.3: **Top predictors of churn.** *Panel A:* Individual features that are most predictive of churn. The ranked list of features is determined by individually assessing the ability of each feature to predict churn, absent other features. *Panel B:* The ranked list of features is determined through a joint feature selection process that computes the attribute importance from a set of decision trees, using the algorithm describe by Hastie et al (2009).

Chapter 5

Multi-view Graph Convolution Networks

Abstract

With the rapid expansion of mobile phone networks in developing countries, large-scale graph machine learning has gained sudden relevance in the study of global poverty. Recent applications range from humanitarian response and poverty estimation to urban planning and epidemic containment. Yet the vast majority of computational tools and algorithms used in these applications do not account for the multi-view nature of social networks: people are related in myriad ways, but most graph learning models treat relations as binary. In this chapter, we develop a graph-based convolutional network for sparse multi-view networks. We show that this method outperforms state-of-the-art semi-supervised learning algorithms on three different prediction tasks using mobile phone datasets from three different developing countries. We also show that, while designed specifically for use in poverty research, the algorithm also outperforms existing benchmarks on more traditional node labelling tasks [1] .

[1]The content presented in this chapter is based on joint work with Joshua E. Blumenstock (forthcoming in AAAI2019)

5.1 Introduction

Over the past several years, large-scale graph machine learning has gained increasing relevance in the domain of international poverty research. Driven largely by the expansion of mobile phone networks throughout developing countries – roughly 95% of the world population now has mobile phone coverageGSMA, 2016 – vast quantities of network data are constantly being generated by people living in even extremely poor and marginalized communities. Recent work has shown how such data can be used to inform critical policy decisions, including the measurement of living conditions J. Blumenstock, Cadamuro, et al., 2015, the spread of infectious diseases Wesolowski et al., 2015, and the management of humanitarian crises Lu et al., 2012. Private companies are also taking advantage of this new source of data, for instance by using data from mobile phones to generate credit scores that can expand credit to millions of people historically shut out of the formal banking ecosystem Francis et al., 2017.

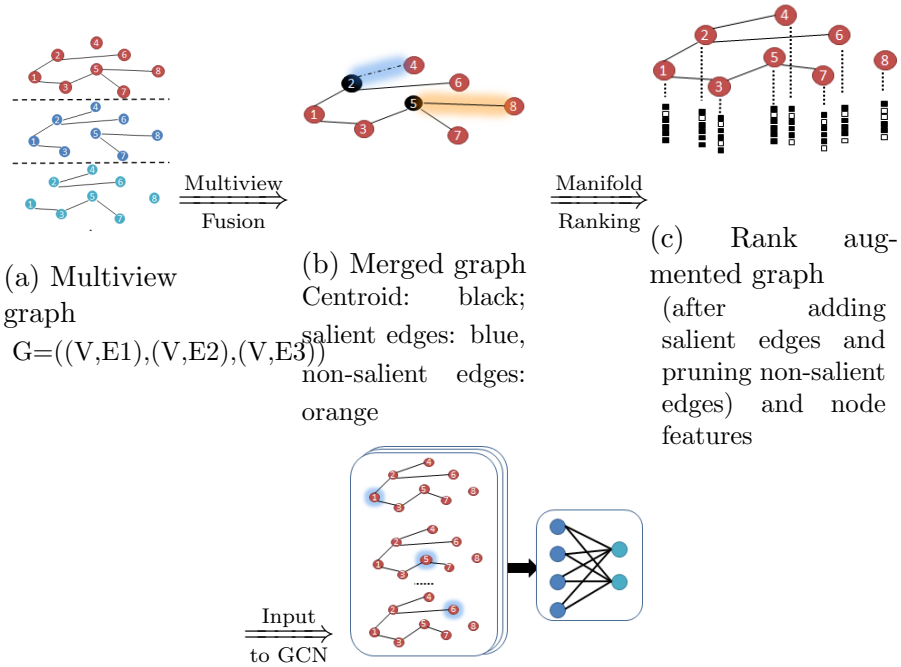
However, a critical constraint to the use of these data in settings related to economic development is the lack of scalable algorithms for performing prediction tasks on sparse multi-view networks. Multi-view networks (also referred to as multiplex and multi-modal networks), are networks in which nodes can be related in multiple ways, and are the natural abstraction for mobile phone networks, where different individuals have different types of relationships and can interact using different modalities (such as phone calls, text messages, money transfers, and app-based activity). Yet, the vast majority of applied research using mobile phone data — in developing and developed countries alike — ignores the multi-view nature of phone networks (see Blondel et al., 2015 for a survey).

This paper develops a novel approach for learning on multi-view networks, which bridges two different strands in the research literature. The first strand involves methods for efficient analysis of multi-view networks; the second explores algorithms for semi-supervised graph learning (see Related Work, below). The method we develop provides an efficient approach for applying convolutional neural networks to multi-view graph-structured data. We benchmark this new method, which we call Multi-GCN (short for Multi-View Graph Convolutional Networks), on three different mobile network datasets, on three different prediction tasks relevant to the international development community: (1) predicting the adoption of a new "financial inclusion" technology in a West African country; (2) predicting whether an individual is living below the poverty line in an East African country; (3) predicting the gender of mobile phone subscribers in a South Asian country. In all cases, we find that Multi-GCN outperforms state-of-the-art benchmarks, including standard Graph Convolutional Networks Kipf and Welling, 2017, Node2Vec Grover and Leskovec, 2016, Deepwalk Perozzi et al., 2014, and Line Tang et al., 2015.

We find that Multi-GCN works well when networks are relatively densely connected, as well as when they are extremely sparse. While designed specifically with the developing-country context in mind (where the sparsity and multi-view properties of networks are very salient), we further show that Multi-GCN performs remarkably well on more traditional node label prediction tasks. In particular, we replicate the experimental design of two recent

papers that provide state-of-the-art performance baselines for predicting class labels on two citation network datasets from Citeseer and Cora Z. Yang et al., 2016; Kipf and Welling, 2017, and find that Multi-GCN outperforms both baselines on both tasks.

5.2 Related Work



(a) Multiview graph G=((V,E1),(V,E2),(V,E3))

(b) Merged graph Centroid: black; salient edges: blue, non-salient edges: orange

(c) Rank augmented graph (after adding salient edges and pruning non-salient edges) and node features

(d) Graph Convolution Network Hidden layers→Dense layers

Figure 5.1: Overview of our Multi-view GCN approach

Technical Related Work

Our goal is to develop an efficient method for node-level transductive semi-supervised learning over multi-view (or multi-modal) graphs. Here, we begin with a general overview of semi-supervised learning, then focus on various approaches to graph-based semi-supervised learning, and finally discuss related work on multi-view networks.

Graph-Based Semi-Supervised Learning

One of the biggest issue with applying supervised learning algorithms in a developing country environment is that it is often costly to collect labels for training. For instance, when using mobile phone data to predict the wealth of subscribers, J. Blumenstock, Cadamuro, et al., 2015 manually conducted a survey of roughly 1,000 subscribers. Semi-supervised learning tries to solve this problem by using unlabeled data along with the labeled data to train better classifiers (see X. Zhu, 2005 for a survey). Our focus is on transductive semi-supervised learning, which assumes that all the unlabeled data is available at the training time and does not attempt to generalize to data unseen during training.

Graph-based semi-supervised learning (GSSL) is a popular approach for semi-supervised learning that treats labeled and unlabeled instances as graph vertices, and relationships between instances as edges Liu et al., 2012. GSSL algorithms try to learn a classifier that is consistent with the labeled data while making sure that the prediction for similar nodes is also similar. This is achieved by minimizing a loss function with two factors: a) supervised loss over the labeled instances, and b) a graph-based regularization term. Different GSSL algorithms use different functions for graph regularization. Label propagation-based approaches, for instance, use a constrained label lookup function (e.g., Zhou, Bousquet, et al., 2004). Related, kernel-based approaches parameterize regularization term in the Reproducing Kernel Hilbert Space (RKHS).

Learning Over Graphs

The success of word embedding algorithms like Word2Vec Mikolov et al., 2013 has inspired similar algorithms for graphs. For instance, DeepWalk Perozzi et al., 2014 learns embeddings by predicting the neighborhood of nodes based on random walks over the graphs, while LINE Tang et al., 2015 and Node2vec Grover and Leskovec, 2016 improve DeepWalk using advanced sampling schemes. Such models first use a random walk to generate embeddings, then separately use such embeddings to train classifiers. More recently, neural network-based approaches have been proposed to perform learning over graphs. These have been extended to the task of semi-supervised learning Bruna et al., 2013; Defferrard et al., 2016, including recent work by Kipf and Welling, 2017 that proposes a Graph Convolutional Network (GCN), which we take as a starting point for our approach.

Learning Over Multi-View Graphs

The key distinction between our approach and prior work is our desire to handle graphs with multiple views, i.e., graphs where vertices can be connected in more than one way. In recent years, many different algorithms have been proposed for learning on multi-view graphs. These algorithms can be broadly divided into three main categories: 1) co-training algorithms, 2) learning with multiple kernels, and 3) subspace learning (See Xu et al., 2013 for a survey). Recent work by X. Dong et al., 2014 show that subspace approaches — which find a latent subspace shared by multiple views — perform well relative to co-training and

kernelized approaches on a range of tasks. We therefore focus our attention on integrating subspace learning approaches with recent innovations in graph convolutional networks.

Comparison with existing work

Our main contribution is to propose an efficient method for adapting GSSL to multi-view contexts. Existing approaches to GSSL cannot be readily implemented on such data; those algorithms that do handle multiple views generally treat views and vertices equally. We show that the current "state of the art" methods like Graph Convolutional Networks Kipf and Welling, 2017 can be enhanced by augmenting the input graph using subspace analysis over Grassman manifolds. Efficiency comes from ranking edges in the merged graph to enable edge pruning and induction.

Empirical Related Work

Our experimental results focus on three prediction tasks of relevance to the international development community:

Predicting poverty.

A large number of humanitarian applications — from poverty targeting to program monitoring — require accurate estimates of the welfare for beneficiary populations. Recently, several paper have shown how digital trace data can be used to estimate the socioeconomic status of individuals, households, and villages. For instance, Jean et al., 2016 show that daytime satellite imagery can be used to estimate village wealth; Quercia et al., 2012 find that Twitter data can be used to estimate levels of deprivation, and J. Blumenstock, Cadamuro, et al. (2015) shows that mobile phone metadata can be used to estimate the welfare of individuals and regions, respectively.

Product adoption.

We focus is on the adoption of "mobile money", a suite of phone-based financial services that are designed to promote financial inclusion among those traditionally shut out of the formal banking ecosystem Suri, 2017. Within this literature, our work relates most closely to Muhammad R. Khan and Joshua E. Blumenstock (2016), who analyze the predictors of mobile money adoption in three different developing countries. More broadly, our approach adapts insights from models designed for more generic network adoption prediction tasks — these are discussed in Section 5.2 below.

Gender prediction.

Gender equality and women's empowerment one of the Sustainable Development Goals, and recent work explores how digital trace data can be used to assess progress toward this goal

Fatehkia et al., 2018. Prior work has shown that both social media data Mislove et al., 2011 and phone data V. Frias-Martinez, E. Frias-Martinez, et al., 2010b can be used to predict the gender of users.

Broadly, these prior studies demonstrate a proof of concept: that digital trace data can be used to predict the characteristics and outcomes of individuals. But the analyses rely heavily on off-the-shelf machine learning algorithms, and rarely, if ever, account for the multi-view nature of real-world social networks. Our study is focused on technical innovation, and we show that a novel approach yields significant improvements on these real-world prediction tasks.

5.3 *Multi-GCN*: Multi-View Graph Convolutional Networks

Our approach to semi-supervised learning on multi-view graphs integrates three steps, depicted in Figure 5.1. First, we use methods from subspace analysis on a Grassmann manifold to efficiently merge multiple views (or layers) of the same graph. Second, we use a manifold ranking procedure to identify the most informative sub-components of the graph and to prune the graph upon which learning is performed. Finally, we apply a convolutional neural network, adapted to graph-structured data, to allow for node classification in a semi-supervised setting.

Merging Subspace Representations

Given an undirected multilayer graph with M layers $G = {G_i}_{i=1}^{M}$ such that each layer G_i has the same vertex set V but same or different edges set E_i, we first calculate the graph Laplacian for each of the individual layers. If D_i and W_i represent the degree matrix and the adjacency matrix respectively for the i^{th} view of the graph, then the normalized graph Laplacian is defined as

$$L_i = D_i^{-1/2}(D_i - W_i)D_i^{-1/2} \tag{5.1}$$

Given the graph Laplacian L_i for each layer of the graph, we next calculate the spectral embedding matrix U_i through the trace minimization problem:

$$\min_{U_i \in \mathbb{R}^{n*k}} tr(U_i' L_i U_i), \quad \text{s.t. } U_i' U_i = 1 \tag{5.2}$$

This trace minimization problem can be solved by the Rayleigh-Ritz theorem Von Luxburg, 2007. The solution U_i contains the first k eigenvectors corresponding to the k smallest eigenvalues of L_i.

A Grassman manifold $\mathcal{G}(k, n)$ can be considered as a set of k-dimensional linear subspaces in $\mathbb{R}^n$ where each unique subspace is mapped to a unique point on the manifold. Each point

on the manifold can be represented by an orthonormal matrix $Y \in \mathbb{R}^{n*k}$ whose columns span the corresponding k-dimensional subspace in $\mathbb{R}^{n*k}$ and the distance between the subspaces can be calculated as a set of principal angles $\{\theta_i\}_{i=1}^{k}$ between these subspaces. X. Dong et al., 2014 show that the projection distance between two subspaces Y_1 and Y_2 can be represented as a separate trace minimization problem:

$$d^2_{proj}(Y_1, Y_2) = \sum_{i=1}^{k} \sin^2 \theta_i = k - tr(Y_1 Y_1' Y_2 Y_2') \tag{5.3}$$

where, based on Eq. 5.3, the projection distance between the target representative subspace U and the individual subspaces $U_{i i=1}^{M}$ can be calculated as:

$$\begin{aligned} d^2_{proj}(U, \{U_i\}_{i=1}^{M}) = & \sum_{i=1}^{M} d^2_{proj}(U, U_i) \\ = & \; kM - \sum_{i=1}^{M} tr(UU' U_i U_i') \end{aligned} \tag{5.4}$$

Minimization of Eq. 5.4 ensures that individual subspaces are close to the final representative subspace U.

Finally, to make sure that the original vertex connectivity in each graph layer is preserved, we include a separate term that minimizes the quadratic-form Laplacian (evaluated on the columns of U):

$$\begin{aligned} \min_{U \in \mathbb{R}^{n*k}} \sum_{i=1}^{M} tr(U' L_i U) + \alpha_i [kM - tr(UU' U_i U_i')], \\ \text{s.t. } U_i' U = 1 \end{aligned} \tag{5.5}$$

In Eq 5.5, α is the regularization parameter that balances the trade-off between the two terms in the objective function. Rearranging Eq. 5.5 and ignoring the constant terms yields

$$\min_{U \in \mathbb{R}^{n*k}} tr[U'(\sum_{i=1}^{M} L_i - \sum_{i=1}^{M} \alpha_i U_i U_i')U], \tag{5.6}$$

As before, the Rayleigh-Ritz theorem can be used to solve Eq 5.5. The solution is given by the fist k eigenvectors of the modified Laplacian:

$$L_{mod} = \sum_{i=1}^{M} L_i - \sum_{i=1}^{M} \alpha_i U_i U_i' \tag{5.7}$$

Graph-Based Manifold Ranking

Though the modified Laplacian calculated in the last section can be fed directly to the downstream graph convolutional networks, the accuracy of the models can be further increased by ranking the nodes in the manifold based on their saliency with respect to some critical nodes Zhou, Weston, et al., 2004. To rank points on the manifold, we use the closed form function,

$$f^* = (I - \beta * L_{mod})^{-1} q \tag{5.8}$$

Here, I represents the identity matrix, L_{mod} is the normalized Laplacian as calculated in Eq. 5.7, and β is the regularization parameter. Given a vector q containing the indices of the query nodes, Eq. 5.8 calculates the saliency of the other nodes with respect to the query nodes and saliency of these nodes can be used to induce or prune edges in the underlying graph. The use of manifold based ranking suits our approach as the modified Laplacian representing merged subspaces can be used directly for saliency detection. The query nodes can be selected as the centroids determined by any clustering algorithm over the manifold.

The algorithm for the subspace merging and subsequent manifold ranking is shown in Algorithm 1.

Dataset	**Nodes**	**Edges (view 1)**	**Edges (view 2)**	**Classes**	**Features**
Panel A: CDR Datasets					
Product Adoption	17,000	23,032	18,371	2	132
Poverty Prediction	422	544	1,799	2	1,709
Gender Prediction	958	992	978	2	821
Panel B: Citation Datasets					
Citeseer	3,327	4,732	3,492	6	3,703
Cora	2,708	5,429	2,846	7	1,433

Table 5.1: **Summary statistics.** The Label Rate indicates the ratio of the number of labeled instances used for training and overall number of instances in a dataset.

Graph Convolution Networks

The application of convolutional neural networks to irregular or non-Euclidean grids, such as graphs, is based on the fact that convolutions are multiplications in the Fourier domain, which implies that graph convolutions can be expressed as the multiplication of a signal $x \in \mathbb{R}^N$ with a filter $g(\theta)$ (see Bruna et al., 2013):

$$g_\theta * x = g_\theta(L)x = U g_\theta U^T x \tag{5.9}$$

Algorithm 2: Fusion of multiple views of a graph

Input: $\{A_i\}_{i=1}^{M}$: $n \times n$ adjacency matrices of individual graph layers $\{G_i\}_{i=1}^{M}$, with G_1 being the most informative layer
Input: $\{\alpha_i\}_{i=1}^{M}$,regularization parameters per subspace to be merged
Input: K, salient query points
Input: Y, number of salient edges per centroid to add
Input: Z, number of non-salient edges per centroid to prune
Input: β, manifold ranking regularizer

Output: L_{mod}:Merged Laplacian,A_{mod}: Merged Adjacency matrix, E_s:Salient Edges, E_{ns}: Non salient edges

---

Step 1: Compute normalized Laplacian matrix L_i for each layer of the graph
Step 2: Compute subspace representation U_i for each layer of the graph
Step 3: Compute the modified Laplacian matrix $L_{mod} = \sum_{i=1}^{M} L_i - \sum_{i=1}^{M} \alpha_i U_i U_i'$
Step 4: Perform clustering on the modified Laplacian to identify K salient points i.e. centroids $\{q_i\}_{i=1}^{K}$
Step 5: For each of the centroid rank other edges on the manifold $f^* = (\mathcal{I} - \beta * L_{mod})^{-1} q$
Step 6: For each centroid q_i add Y salient edges to the E_s and Z non-salient edges to the E_{ns}
Step 7: Add E_s to A_1 to form A_{mod}
Step 8: Remove E_{ns} from A_{mod}

Here, U represents the eigen decomposition of the normalized graph Laplacian $L = I - D^{-1/2}AD^{-1/2}$ and I, D, A represent the identity, degree and the adjacency matrix, respectively. Graph convolutions can be further expressed in terms of Chebyshev polynomials as

$$g_{\theta'} * x = \sum_{k=0}^{K} \theta'_k T_k(\tilde{L}) x \tag{5.10}$$

where $\tilde{L}$ is the rescaled Laplacian, T_k represents the Chebyshev polynomials, and θ' represents the vector of Chebyshev coefficients. Following Kipf and Welling, 2017, by approximating the maximum value of the largest eigenvalue and constraining the number of free parameters, the convolution operation can be represented as

$$g_\theta * x = \theta(I + \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2})x \tag{5.11}$$

where $\tilde{A} = A + I$ and $\tilde{D} = \sum \tilde{A}$ are the renormalized versions of A and D. This renormalization avoids numerical instabilities resulting from exploding/vanishing gradients Defferrard et al., 2016; Kipf and Welling, 2017.

The modified graph (A_{mod} in Algorithm 2) resulting from the merger of Laplacians using the subspace analysis and manifold ranking can be fed directly into the graph convolution networks defined above. The forward propagation model for a two layer network can then be represented as

$$Z = F(X, A) = softmax(\hat{A}\, ReLU(\hat{A}XW^0)W^1) \tag{5.12}$$

Here, $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ is calculated as a preprocessing step before giving the input to the neural network. W^0 and W^1 represent the input-to-hidden-layer and hidden-layer-to-output weight matrices for a two layer neural network, and can be trained using gradient descent.

5.4 Experiments and Data

Datasets

Our first set of experiments test Multi-GCN on three three prediction tasks relevant to international development. Each one uses a different dataset of mobile phone Call Detail Records (CDR), obtained from three different developing countries with GDP per capita less than $1,600 USD. These data contain meta-data on all communication events that are mediated by the mobile phone network, such as phone calls and text messages. Each CDR dataset contains multiple possible relationships between nodes (views); we extract one view corresponding to phone calls between users, and another corresponding to text messages. We separately construct a large set of features of each user (such as total call volume and degree centrality), using the combinatoric approach described in J. Blumenstock, Cadamuro, et al., 2015.

Table 4.1 presents summary statistics for each of these datasets. The spy plots, which provide an indication of the sparsity of the graph structure in the three mobile phone networks, are shown in Figures 5.2-5.3.

Product adoption dataset

The first dataset that we use is a sample of a dataset of mobile phone activity from a West African country. Here, the classification of interest is whether or not the user eventually adopts a new financial inclusion product. There are two possible classifications: (1) Did not adopt; (2) Adopted and used the product. Following the experimental setup described in Kipf and Welling, 2017, we randomly 20 users from each category (40 total) for the training dataset; the validation and the testing dataset consist of 500 and 1000 randomly selected users, respectively.

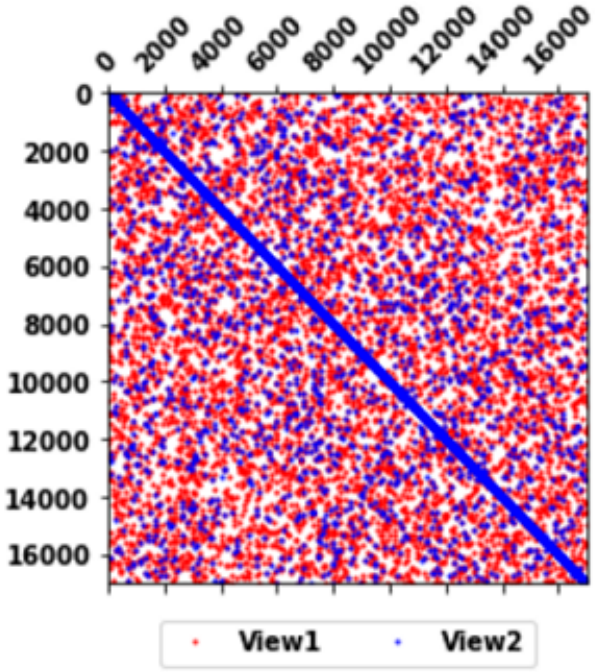


Figure 5.2: Product Adoption

Poverty prediction dataset

The wealth prediction dataset consists of several thousand transactions of different mobile phone users from an East African country. We attempt to classify users as poor or non-poor, where labels were obtained by J. Blumenstock, Cadamuro, et al., 2015 through a small set of phone surveys that were conducted with mobile phone subscribers.

Again, we randomly selected 20 users from each category as the training dataset, while the size of the validation dataset and the testing dataset is 100 and 200 respectively.

Gender prediction dataset

The gender prediction dataset originates from a developing country in South Asia. Here, the classification task is to predict the gender of the mobile phone users, where gender labels are provided by the operator for a small number of labeled instances.

We randomly select 20 users from each category for training; the size of the validation and the testing datasets are 100 and 800, respectively.

Citation classification datasets

A final set of experiments tests Multi-GCN on more standard node labelling tasks. Specif-

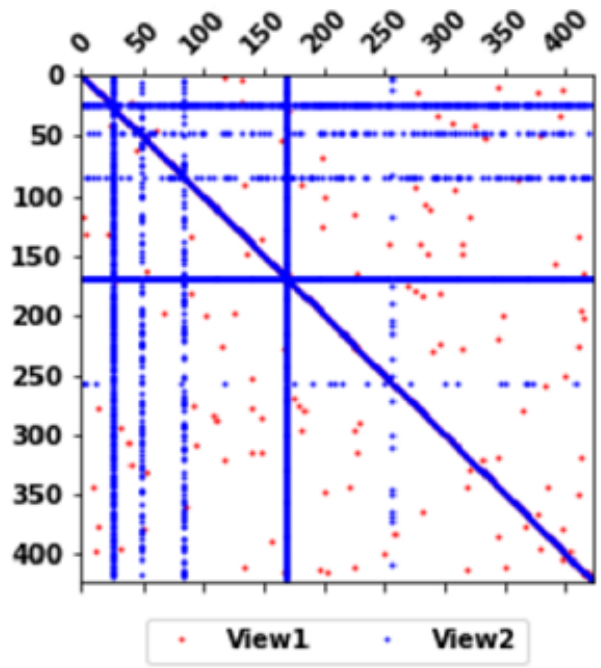


Figure 5.3: Wealth Prediction

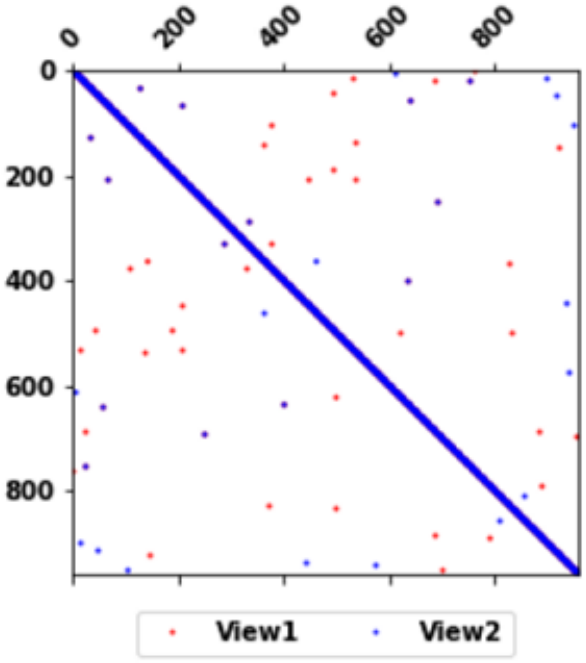


Figure 5.4: Gender Prediction

ically, we replicate the experimental design of recent work by Kipf and Welling, 2017, who benchmark a large number of algorithms' ability to correctly classify academic articles on

the Cora and Citeseer citation networks. In these datasets, nodes are documents and the first view corresponds to the citation links between the research papers. We construct the second view from the textual similarity of the research papers. Specifically, we use the bag of word vector representation of the document, and if the normalized cosine similarity between documents is greater than 0.8, say that an edge exists between the documents.

Experimental setup

In general, our goal is to correctly classify nodes in a network, where only a very small fraction of nodes are labeled. In the experiments, we start from a small sample of labeled nodes and test the ability of Multi-GCN, as well as several state-of-the-art algorithms from the literature, to correctly classify unlabeled nodes in the validation and testing sets. We use three popular node embedding algorithms (Node2vec, Deepwalk and LINE) as a first set of baselines. In addition, we provide three baselines based on graph convolutional networks Kipf and Welling, 2017. The first two, *GCN (first view)* and *GCN (second view)*, apply GCN over the two respective adjacency matrices from phone and text message activity. The third, *GCN (view union)*, operates on the union of the adjacency matrices of the first view and the second view. In each GCN baseline, the node features are constructed from the first view adjacency matrix.

After merging different views, we rank the interaction between nodes using Eq. 5.8 based on their salience with respect to the query points. The value of parameter α used as a regularizer for subspace merging (Eq. 5.7) is selected through 10-fold cross-validation. We similarly tune the hyper-parameters β to 0.99 and set the number of query points to ten times the number of unique ground truth labels.

After adding salient edges and eliminating non-salient edges through the ranking process, both the adjacency matrix of the modified graph and the node features are passed as input to a two-layer graph convolutional network as described in Section 5.3. All of the GCN-based baseline models (GCN(first view), GCN(second view), GCN(view union)) and our model are trained for a maximum of 200 training iterations, using *Adam* (Adaptive moment estimation extension to stochastic gradient descent, see Kingma and Ba, 2014) and a learning rate of 0.01. Other GCN hyper-parameters are the same as in Kipf and Welling, 2017: Dropout rate: 0.5; No. of neurons in the hidden layer: 16; L2 Regularization: 5.10^{-4}.

5.5 Results

Experimental results for the three developing-country datasets are shown in Table 5.2. Each row in this table indicates the average and standard error of the classification accuracy over 10 randomly drawn train-test splits of the same size for each dataset, constructed as described in Section 5.4. The last row in Table 5.2 shows the performance of Multi-GCN. In all four datasets, Multi-GCN outperforms existing state-of-the-art benchmarks, with the

Method	**Product Adoption**	**Poverty Prediction**	**Gender Prediction**
DeepWalk (first view)	56.43±0.187	51.91±0.62	53.18± 0.55
DeepWalk (second view)	51.97±0.112	50.34±0.36	50.84±0.64
DeepWalk (view union)	56.81± 0.114	50.87±0.95	52.34±0.5
Node2vec (first view)	53.87±0.2	52.26±0.58	50.12± 0.4
Node2vec (second view)	50.5±0.11	49.7±0.23	51.68±0.4
Node2vec (view union)	54.5±0.11	50.52±0.63	51.64±0.53
LINE (first view)	51.11±0.01	50.15±0.02	51.56± 0.001
LINE (second view)	50.83±0.01	52.29±0.001	50.00±0.001
LINE (view union)	56.26±0.003	50.18±0.001	51.33±0.002
GCN (first view)	70.74±2.2	55.19±2.33	63.97± 1.29
GCN (second view)	71.4±1.81	50.06±0.81	63.01±0.013
GCN (view union)	71.9±0.9	50.22±0.56	63.9±1.32
Multi-GCN (this paper)	**73.47±0.91**	**59.23±0.2**	**66.34± 1.03**

Table 5.2: Classification accuracy (percentage). Each entry of the table specifies mean classification accuracy and standard error over 10 randomly selected dataset splits of equal size.

margin of improvement greatest in the poverty prediction task and smallest in the gender prediction task.

The second set of experimental results, which test Multi-GCN against recent benchmarks on a more standard node classification task, are shown in Table 5.3. Since Kipf and Welling, 2017 previously showed GCN (first view) to outperform a large number of competing algorithms (including DeepWalk and Planetoid) on this exact prediction task, we omit those results from the table. In addition to performing a comparison over randomly drawn train-test splits (as in Table 5.2), we also compare the performance of Multi-GCN against a different set of randomized test-train splits, as used in the original tests by Kipf and Welling, 2017, with an additional validation set of 500 instances used for hyper-parameter tuning. In all cases, we observe modest improvements in predictive accuracy of Multi-GCN relative to existing approaches.

5.6 Discussion

Above, we have proposed an approach to semi-supervised learning on multi-view graphs. Through a series of experiments, we show that this approach improves upon state-of-the-art GCN- and embedding-based algorithms on a variety of prediction tasks, both related

	Predefined train-test splits		Randomized train-test splits	
Method	**Citeseer**	**Cora**	**Citeseer**	**Cora**
GCN (first view)	70.3	81.5	67.9± 0.5	80.1±0.5
GCN (second view)	50.7	53.6	53.56±0.1	56.93±0.32
GCN (view union)	70.7	80.38	67.98±0.32	78.49±0.1
Multi-GCN (this paper)	**71.28**	**82.5**	**70.5± 0.15**	**81.07±0.15**

Table 5.3: **Classification accuracy for the non-CDR datasets.** Left panel shows the average classification accuracy (percentage) for the pre-defined test-train splits Z. Yang et al., 2016. Right panel shows the classification accuracy (percentage) and standard error over 10 randomly selected dataset splits of equal size.

to poverty research and node labelling in general. Intuitively, this does not come as a surprise. In instances where graph structure is informative for node classification, it makes sense that allowing for heterogeneity in that structure might improve classification accuracy. Concretely, we might expect the social network induced by text messages to contain non-redundant information to the social network induced by phone calls; if both networks are informative in predicting the characteristics of an individual, we should expect to see gains from algorithms that account for both.

This intuition is also supported by a closer look at the results in Table 5.2. Here, we observe that while Multi-GCN provides the biggest gain relative to Deepwalk, Node2vec and LINE in the case of product adoption, but the gains relative to single-view GCN are more modest. By contrast, the performance gain on the poverty and gender prediction tasks is significantly higher for Multi-GCN, even relative to the other single-view GCN benchmarks. A closer inspection of the spy plots in Figures 5.2-5.4 provide additional insight. For instance, we can see that different views in the product adoption setting appear somewhat redundant, whereas for poverty and gender prediction the views appear more independent. However, it may be difficult ex-ante to identify which views are complementary in the context of the downstream learning task. Moreover, the extent to which the fusion of multiple views of the network can help increase classification performance depends on many different factors like sparsity of the views, mutual information contained in the views, etc.

5.7 Conclusion

Graph convolutional networks have recently achieved considerable success in a variety of learning tasks on irregular, graph-structured data. Leveraging insights from spectral graph theory, GCN's are beginning to replicate the success that CNN's have seen on more regular image and text data. For a wide variety of learning tasks relevant to graph-structured data—in contexts ranging from advertising in online networks to intervening in the spread of a

contagious disease—this is a promising recent development. In this chapter, we have shown that state-of-the-art GCNs can achieve even higher performance on a variety of classification tasks related to poverty research when the multi-view nature of the underlying network is incorporated into the learning process. We see Multi-GCN as an important first step in adapting neural network-based approaches to multi-view network data and hope that it can provide a foundation for future work in this space.

Chapter 6

Gender Disparity Analysis

Abstract

Ending gender discrimination in all its forms has long been a sustainable development goal of the United Nations. Despite initiatives and efforts of different governments across the world, however, gender discrimination still exists. One of the biggest problems has been the lack of accurate information about gender disparity. This project aims at modeling educational gender disparity at the district level in Pakistan using network data extracted from call detail records. Using survey-based educational gender disparity as ground truth data, we first explore the prominent network metrics and patterns that help in better understanding of gender disparities in the population. In the second half of this project, we use these network metrics to predict educational gender disparity at the district level. Our model uses a dataset of more than 30 million customers, advanced network features and prediction algorithms to model educational gender disparities accurately. Our findings show that call detail records can be effectively used to analyze social networks of men and women in developing world as an alternative data source to more expensive forms of data collection. Secondly, our analysis shows that the men and women of developing countries like Pakistan manifest significant differences in their social network activities and patterns. Lastly, the predictive model that we have developed enables a relatively accurate and cheaper way of estimating gender disparities. Our proposed model can be easily applied to other countries, provided gender-annotated call detail records are present [1].

[1]This work is based on yet to be published joint work with Joshua E. Blumenstock

6.1 Problem Statement

Getting accurate data about gender disparities can be a challenging task in many countries. There exist countries in developing world where there has been a gap of more than a decade between consecutive censuses. As a result, getting accurate demographic information about women and gender disparities can be the first hurdle for the researchers working on gender-related issues. Researchers working on gender disparities have primarily relied on the data collected through surveys. Survey data may fit the requirements in many cases, but population-level surveys can be expensive and hard to manage, and thus researchers have been actively using alternative data sources in their research. Internet-based social networks like Facebook, Twitter and Google+ can go a long way toward solving this problem in developed world, but in many developing countries the penetration of these social networks remains low. For instance, Magno and Weber Magno and Weber, 2014 have tried to analyze gender disparity using Twitter and Google+ datasets, but they found that in some countries like Pakistan there can be quite some disparity between the ratio of male and female users on the social network as compared to the ratio of males and females in the actual population Magno and Weber, 2014. Because of this issue, research based on these sources can result in analytic discrepancies, as women in developing countries having regular access to these social networks may already be more privileged than the random male counterparts of the society (Jackie Robinson effect Magno and Weber, 2014). Thus, there is a need for extensive research on gender issues at the population level using more comprehensive (conventional or non-conventional) data sources.

Mobile phones have seen good penetration in developing world. As a matter of fact, in many developing countries, mobile phone networks have higher penetration than financial institutes like banks and Internet-based social networks Muhammad R. Khan and Joshua E. Blumenstock, 2016. High penetration of mobile phones in developing countries makes them an ideal source to passively collect information about the mobility and behavioral patterns of individuals. These patterns have been used by researchers to analyze poverty J. Blumenstock, Cadamuro, et al., 2015; Smith-Clarke et al., 2014, unemployment Toole et al., 2015, and migration Joshua E Blumenstock, 2012. However, even with the popularity of the mobile phones in developing world, the analysis of gender disparities is complicated by two factors: 1) lack of gender information for each of the subscribers; and 2) lack of ground-truth gender disparities data at fine resolution.

In this research, we use gender-annotated call detail records (CDRs) data from a major operator in Pakistan to analyze how gender disparities manifest themselves in social networks. Furthermore, the mobility information present in the CDRs enables us to associate these social networks with district boundaries. We use educational gender parity data, available at the district level, to ground truth the CDR-derived conclusions. The overall goal is to ascertain how much of the educational gender disparity can be accurately inferred from the different type of features learned through the CDR data. This research builds on our exploratory work presented in the ICTD 2016 (Philip J. Reed et al., 2016).

6.2 Research Questions

The main questions that we intend to address in this research include:

- **Research Question1 (RQ1):** To what extent are gender disparities reflected in the social networks extracted from the call records?

  In other words, do special social network features exist which can tell us about the gender disparities in a district? A lot of research work has been done on the analysis of social networks of less privileged populations. Examples of this work include the analysis of social networks of migrants Ryan et al., 2008 and the impact of collaboration on the success of individuals Uzzi and Spiro, 2005, but similar analysis has not been done to compare the social networks of women and men. We explore through this question whether prevailing concepts in the literature on social network analysis are valid for gender-annotated networks extracted from the call detail records of developing countries.

- **Research Question2 (RQ2):** How accurately can educational gender parity be modeled using the features extracted from the call detail records?

  Our second goal in this research is to accurately model gender educational parity through the social network metrics extracted during the analysis for Research Question 1. The first research question is our attempt to better understand the correlations between the social network-related metrics (or features) and gender disparities in society. The second research question aims to develop an accurate predictive model that can help in predicting educational gender disparity in different districts of Pakistan. Both research questions are related in the sense that training the predictive model on high-quality features is a key to the high performance of the model. The predictive quality of the features can be measured in different ways, but simply the features for which males and females have significantly different patterns are expected to have higher predictive power as compared to other features.

6.3 Data Description

We used two main data sources for this research:

1. **Communication metadata**

   We extract social networks from the Call Detail Records (CDRs) of a major telecom provider in Pakistan. The data consists of more than 1 billion transactions (voice and text messages) around 30 million users and. In addition to the anonymized caller and recipient ids, CDRs also contain the timing of the activity and the location of the cell tower through which the call was made. Furthermore, gender and age of each of the subscribers are also provided by the telecom operator. The CDR data is quite rich in

information, as it can be used to collect metrics related to the usage of the network, mobility of individuals and temporal characteristics of the user's network. Summary statistics of the CDR data used in this project are shown in the Table 6.1

Property	**Value**
Panel A: CDR Data Characteristics	
Male Users	5.51 Million
Female Users	0.57 Million
Number of days	7
Total Calls + SMS	1.07 Billion
Total Districts Covered	93
Panel B: Pakistan's Demographic Indicators	
Population	185.00 Million
GDP per capita (PPP adjusted)	$4811.4
Human Development Index	0.58
Gender Gap Index	0.559
Mobile Subscribers	130 Million
Mobile phone subscriptions (per 100 people)	73
Mobile phone operators	6

Table 6.1: Summary statistics.

National indicators(Source: World Bank) and CDR metrics

The CDR data that we have is quite rich as, in addition to the information about the caller and the recipient IDs, it also contains information about the cell tower used by the individuals. This information can be used to calculate different mobility related metrics. Though the CDR data used in this research only spans seven days, it covers 93 out of 128 districts, capturing variation in population density and human development index 6.1. The time span of the CDR data does not coincide with any of the major national holidays or any weather-related catastrophe.

2. **Educational Gender Disparity Data**

To augment the CDR data with the ground truth data about educational gender parity, we use the data collected by the gender advocacy group Alif Ailaan . This dataset contains statistics about the district wise educational gender parity score calculated as a ratio of the net primary enrollment rate of girls to the net primary enrolment rate of boys. Net primary enrollment rate in primary education is the number of pupils of

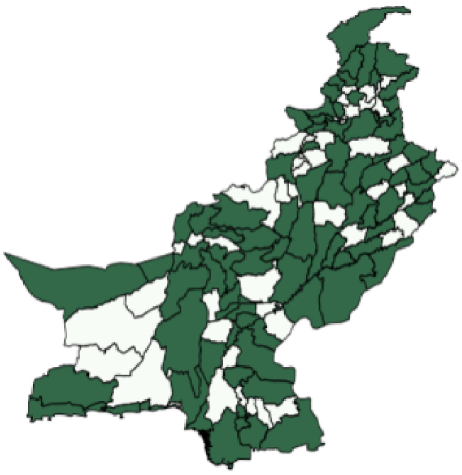


Figure 6.1: Mobile network penetration

official primary school age who are enrolled in primary education as a percentage of the total children of the official school-age population . Distribution of educational gender parity across different districts of the country is shown in 6.2. It is clear from 6.2 that most of the districts with higher gender parity scores are in the northwestern part of the country. The distribution of educational gender parity is positively correlated with the population density of the districts, as shown in 6.3.

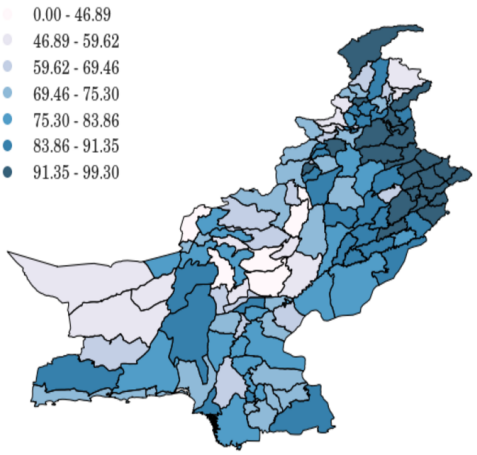


Figure 6.2: Educational Gender Disparity

Data Preprocessing

The CDR dataset contains the details of calls and SMS of millions of customers from a major telecom operator in Pakistan. The networks corresponding to calls and messages can

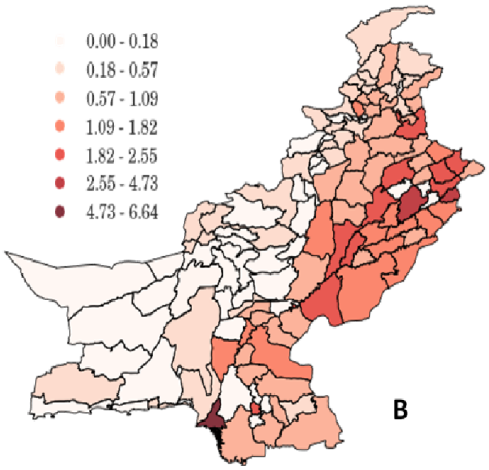


Figure 6.3: Urban Density

be analyzed separately or jointly; the rest of the analysis in this report corresponds to the joint network, such that the edge between two nodes or subscribers can either represent a call or a message. Furthermore, all the edges are considered undirected. We did explore the call and SMS networks separately with directed edges, but the results were inferior as compared to the network corresponding to the undirected calls and SMS. One possible reason for the better performance of the merged (calls + SMS) undirected network is that such a network is denser as compared to the calls only or the SMS only network. The location information present in the CDRS (Caller Cell and the Recipient Cell) is useful for calculating different mobility related features as described in the next section. In addition to the CDR transactions, we also had information about the gender and age of each of the subscriber. This information, coupled with movement patterns, helped us to analyze the social networks along different patterns related to the age, gender and mobility of the ego node (the user whose properties are being analyzed) and the alter (the set of users connected to the ego node).

6.4 Social Network Analysis

The popularity of social networks like Twitter and Facebook has resulted in an increased research focus on social network analysis, but there is no consensus around ways to generate features, or which features are most important. Many times, the choice of the features used in analysis depends on the personal preferences of the researchers. Instead of relying on a few hand-selected features, we have tried to analyze the social networks of men and women in Pakistan through the lens of as many features as possible, as explained in the following subsections.

Network Activity

Number of Calls and SMS

It is a general perception that underprivileged communities and individuals may be less active in call and SMS networks. However, different researchers have found evidence to support or contradict this assumption. Friaz-Martinez et al. (2010) found that women in developing country of their study were more active than men on mobile phone networks V. Frias-Martinez, E. Frias-Martinez, et al., 2010a. Our analysis also confirms this trend, as shown in Figure 6.4. In Figure 6.4, the y-axis shows the number of subscribers (men (green) or women (red)) making a given number of calls or SMS over the 7-day period, as shown by the x-axis. The average number of network transactions made by men and women is also shown through dotted green and red lines respectively. Figure 6.4 also shows that the percentage of women making a higher number of calls is greater than the percentage of men.

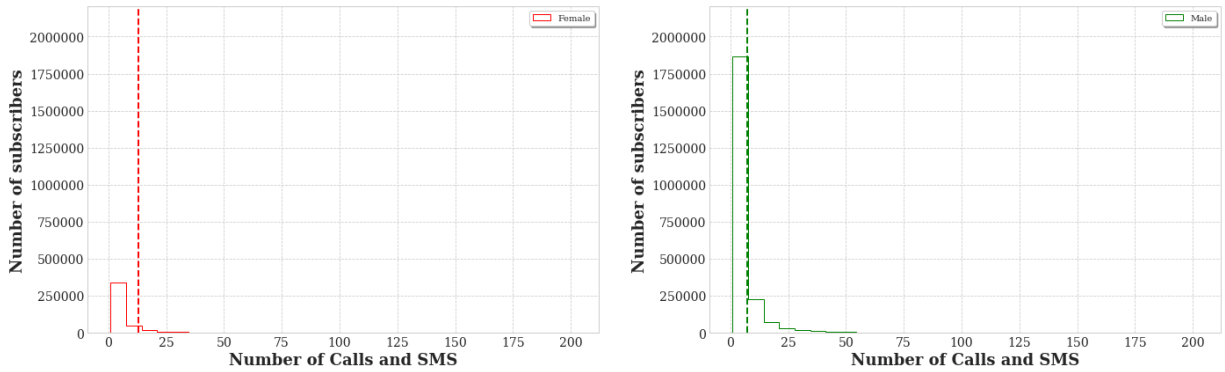


Figure 6.4: Number of Calls and SMS

Network Status

In addition to network activity, the status of the node in the network can be approximated through measures like network size, embeddedness and centrality.

Network Size (Degree Centrality)

The comparison of the network size for women and men is shown in Figure 6.5. Dotted lines in this figure showing the averages indicate that males in Pakistan have higher degrees on the average as compared to females. This observation is a contrast with the findings of Friaz-Martinez et al. (2010) Magno and Weber, 2014. This trend shows that women in Pakistan, in general, have a smaller number of contacts, but as shown in Figure 6.4 the average number of calls and messages for women is higher. Larger network size of males is consistent with social trends in Pakistan, as the number of working men is much higher as

compared to working women, and working individuals are expected to have a higher number of contacts.

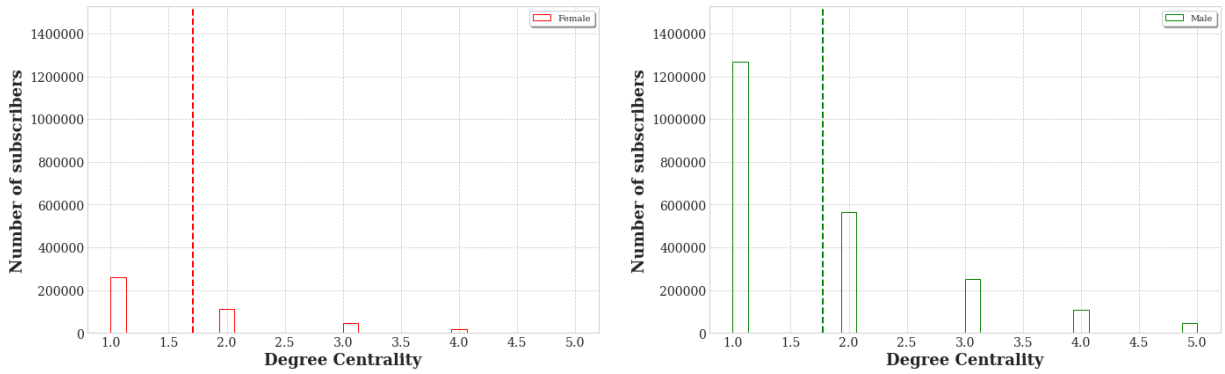


Figure 6.5: Degree Centrality

Network Embeddedness

Embeddedness describes the degree to which the ego nodes are enmeshed in their networks Y. Dong et al., 2014. In other words, embeddedness is the measure of the extent to which the contacts of a node i and the friend of the node i are connected to each other. Embeddedness of the ego node i can be defined as follows Y. Dong et al., 2014

$$\frac{\sum_{v \in V(i)} |(V(i) \cap V(v))| / |(V(i) \cup V(v))|}{V(i)}$$

Here, $V(i)$ represents the list of the neighboring nodes of the ego node i. Women in Pakistan have higher average embeddedness as compared to men as shown by the dotted lines in Figure 6.6. One possible explanation for this trend can be the fact that women have more responsibilities within the house and the family while the men have more responsibilities out of the home. In other words, a woman's network may be largely comprised of relatives, friends or acquaintances, many of whom are connected in ways independent of the woman. Men, however, many have commercial contacts that are not linked in ways independent of the man.

Network Constraints

The concept of structural holes [3] is a popular instrument in social network analysis research and has been used to assess the status of nodes in a network, diffusion of information and many other problems. Network constraints measure the extent to which the network does not span structural holes. If most of the neighbors of a node are connected to each other, then the node has higher constraints and vice versa. Disenfranchised communities like women

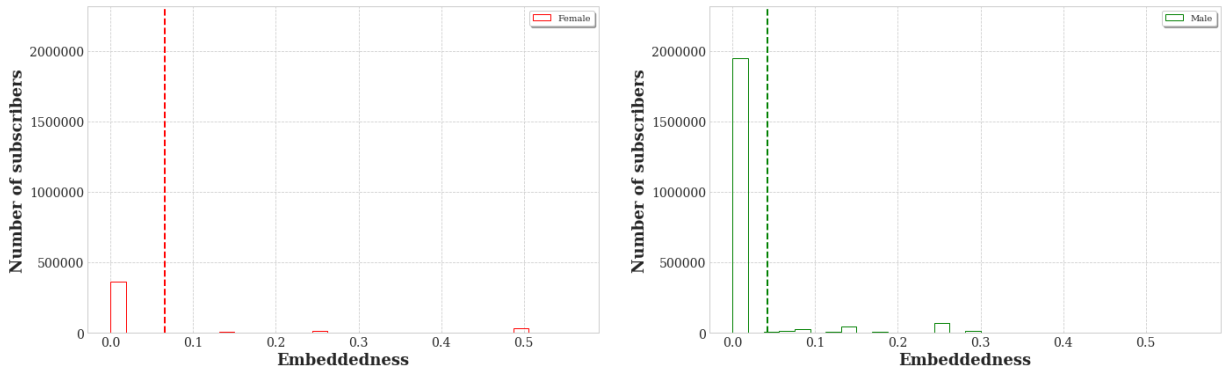


Figure 6.6: Embeddedness

in developing world are expected to have higher constraints or lower number of structural holes.

The constraint of a node i denoted as $C(i)$ is defined as follows

$$C(i) = \sum_{j \in V(i)} \left(p_{ij} + \sum_{q \in V(i)} p_{iq} + p_{qj}\right)^2$$

$V(i)$ represents the list of neighboring nodes of the node i, while the proportional tie strengths p_{ij} is based on the adjacency matrix A and is defined as follows

$$p_{ij} = \frac{a_{ij} + a_{ji}}{\sum_{k \in V(i)} \left(a_{ik} + a_{ki}\right)}$$

Average constraints for women in our network is slightly higher as compared to the constraints for men (0.67 and 0.64 respectively) as shown in Figure 6.7.

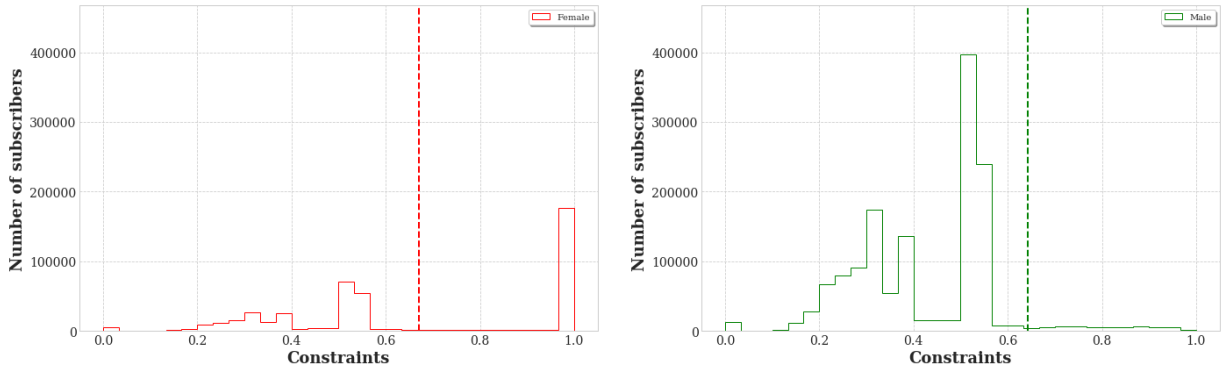


Figure 6.7: Network Constraints

Betweenness Centrality

Importance of a node in a network can also be quantified by calculating the betweenness centrality of the node, i.e., the number of shortest paths that go through the node.

More formally, the betweenness centrality of a node i is given by the following equation

$$g(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Where σ_{st} is the total number of shortest paths from node s to node t, while $\sigma_{st}(i)$ is the number of those paths that pass through i.
Comparison of betweenness centrality of males and females is shown in Figure 6.8. Figure 6.8 shows that the betweenness centrality of the men is consistently higher as compared to the betweenness centrality of women in Pakistan, which is a trend that we expected given that the number of men in the network is much higher than the number of women.

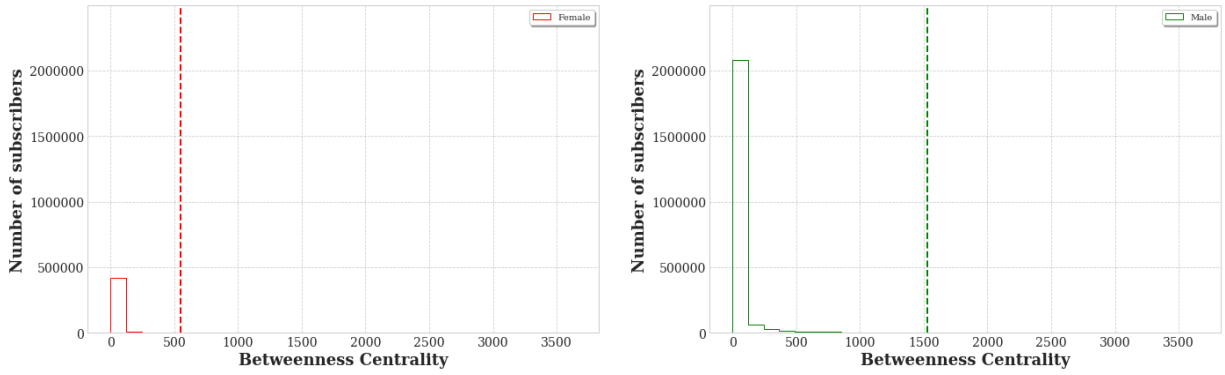


Figure 6.8: Betweenness Centrality

Network Formation

Gender Homophily

People belonging to different categories and communities can have different preferences for network formation, and these preferences can result in patterns of homophily regarding gender, age, and other characteristics. Figure 6.9 shows that male to male edges in the network are much more frequent as compared to female to female edges.
Homophily in a network tells us about the tendency of a node to form connections with similar nodes. However, plain homophily cannot accommodate the frequency of communication between the edges. Diversity-related measures can be used to measure the extent to which a node interacts with a particular type of node. Network diversity has been shown to be an important feature for predicting the socio-economic status of individuals Eagle et al., 2010a.

Network diversity is defined as a function of Shannon entropy, as shown in the following equation.

$$Diversity = \frac{\sum_{i=1}^{N} -p\,(i) * \log(p(i))}{logN}$$

Here N indicates the total number of possible groups across which the diversity is to be calculated while p(i) indicates the proportion of calls being made to the ith group. Based on this equation, we calculate different diversity related metrics, as explained in the following subsections.

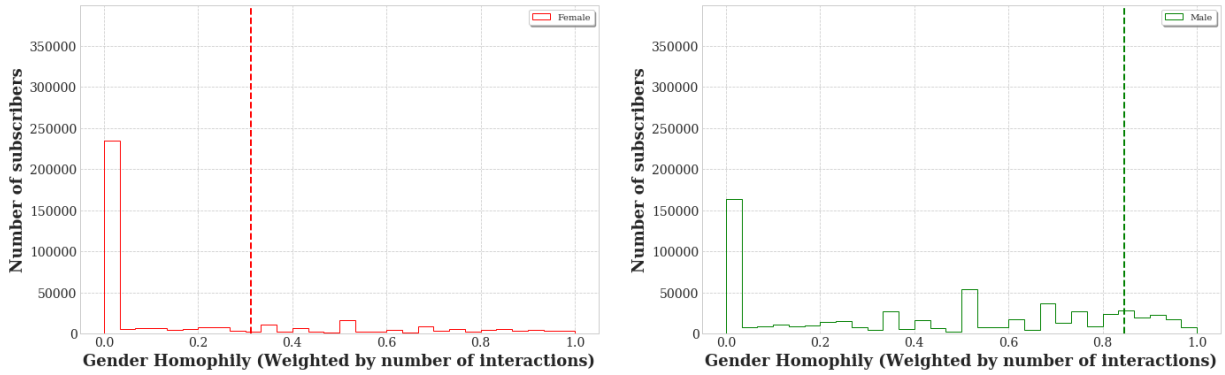


Figure 6.9: Gender Homophily

Gender Diversity

Gender Diversity calculates the proportion of calls being made by an individual to each gender. Women have higher gender diversity compared to males in our dataset, as shown in Figure 6.10. This trend seems to be a contradiction of the trend seen in Figure 6.9 at first, but the concept of gender homophily does not incorporate the volume of calls made to each of the group.

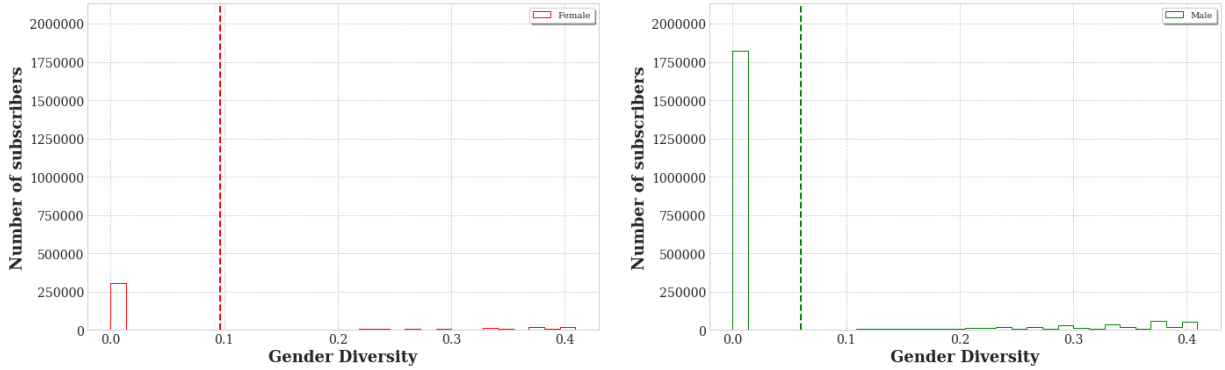


Figure 6.10: Gender Diversity

Age Diversity

Just like gender diversity, age diversity calculates what proportion of calls are being made by the individual to each age group. Based on the age distribution of the subscribers in our dataset, we have defined four different age groups as follows: Group 1 (25-24 years), Group 2 (25-39 years), Group 3 (40 – 54 years), Group 4 (55 and beyond).
The comparison of the age diversity of males and females in our dataset is shown in Figure 6.11. Males have slightly higher age diversity as compared to females on the average (0.057 vs. 0.049). Once again, the larger network size of the males on the average is one possible reason for the higher age diversity in the network.

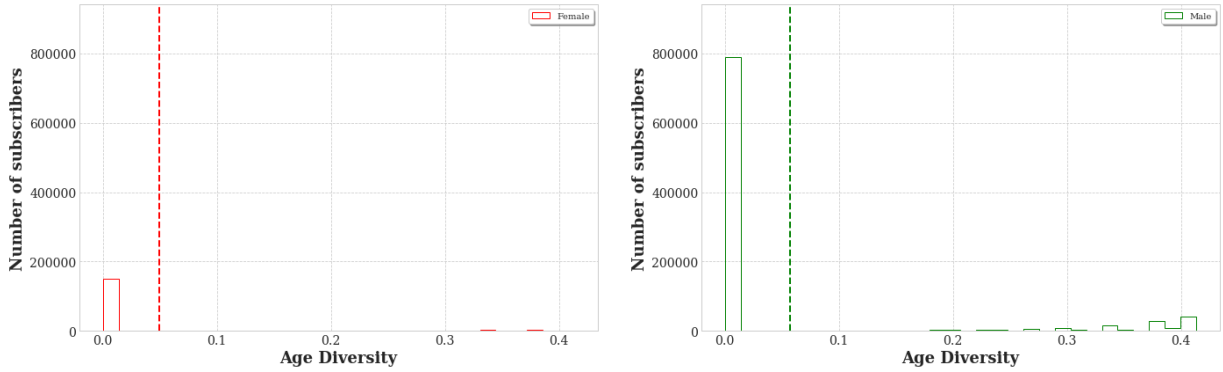


Figure 6.11: Age Diversity

Topological Diversity

Topological diversity analyzes the proportion of calls being made to each of the persons in an individual's network. The distribution of topological diversity for males and females in the dataset is shown in Figure 6.12. The topological diversity of men on the average is slightly higher as compared to the topological diversity of women (0.25 vs. 0.22).

As most of the working class in Pakistan constitutes men, the men are expected to be calling to different contacts in their network, and these contacts may have higher variation in their location as compared to the contacts of the women in general which explains the higher topological diversity of men on the average. But the difference between the average topological diversities between females and males is not very high which may be because of the fact that females are more active in communication with other family members in different locations.

Mobility Related Measures

The location information in the CDR data (Caller Cell Id and Recipient Cell Id) enables us to calculate many location-related or mobility- related features as well. These mobility

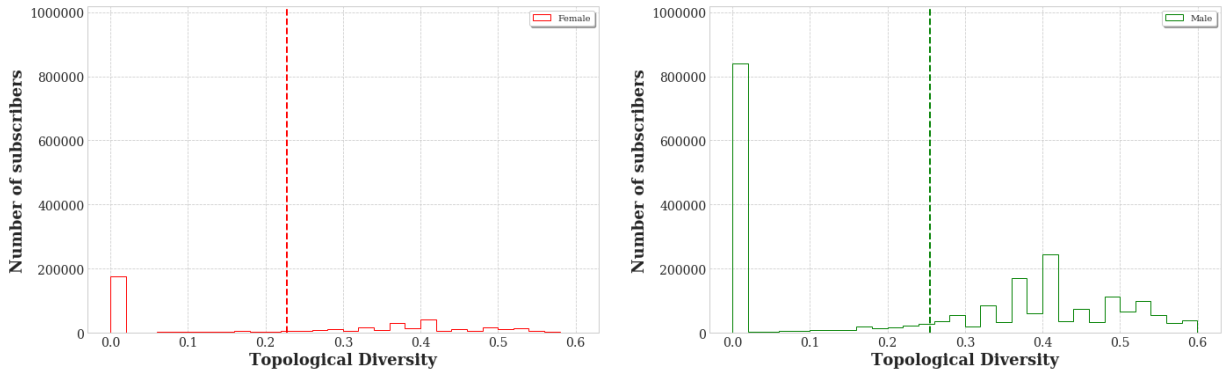


Figure 6.12: Topological Diversity

related features can be really important as it has been shown in different research papers that many of these diversity-related measures may be correlated with the socio-economic status of individuals at the micro level and different regions at the macro level (For example, see Eagle et al., 2010b and Guyon et al., 2002). Amongst these features, the distribution of two, namely location diversity and average geographical reach, of males and females is shown in Figure 6.13 and 6.14, respectively.

Location Diversity calculates the proportion of calls being made to each of the locations to which the user has been communicating. Location diversity of males is higher than the location diversity of females on the average, as shown in Figure 6.13.
Average geographical reach calculates the average geographical distance between the caller and the recipient. The average geographical reach of males is higher as compared to the average geographical reach of women, as shown in Figure 6.14.

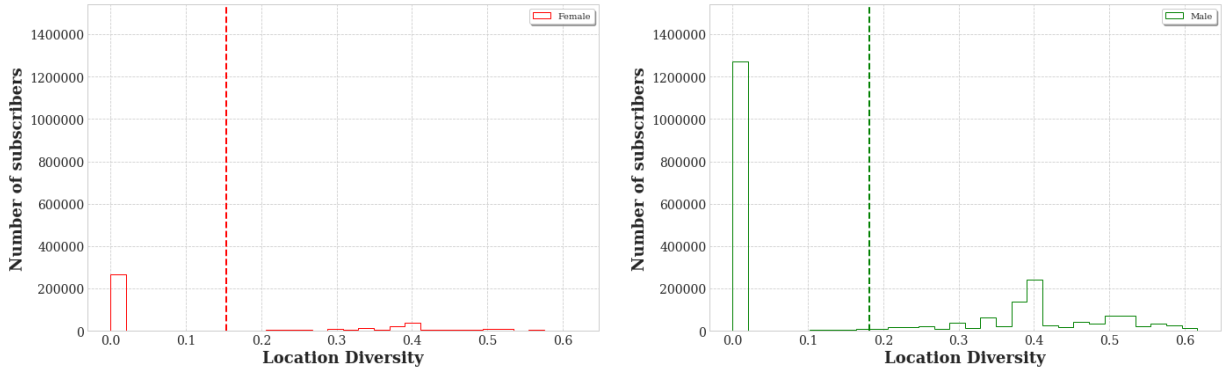


Figure 6.13: Location Diversity

From the distributions of different features highlighted in these sections, we can see three prominent trends.

- Women have a higher average value for the features, for example, network constraints and network embeddedness, which depend on the interconnections of the nodes in the

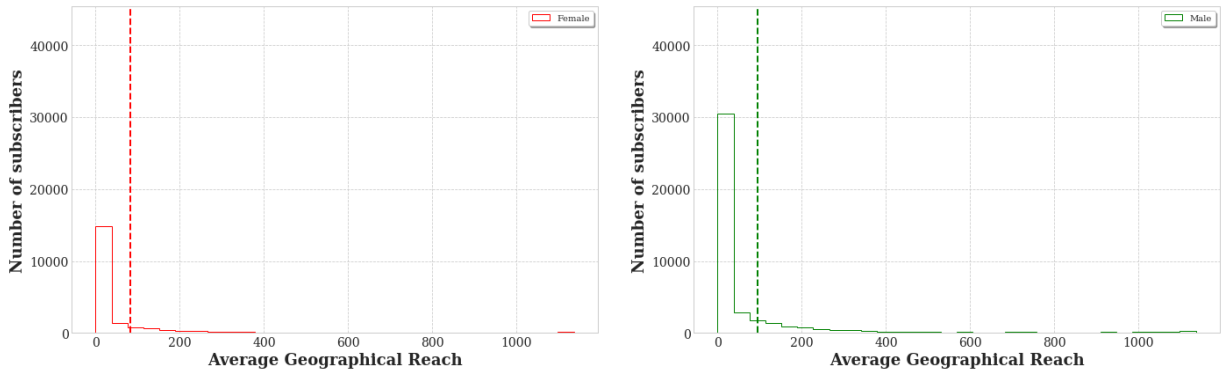


Figure 6.14: Average Geographical Reach

ego's network.

- Men have higher average values for the features which depend on the number of males in the network. Some of the examples of these features include gender homophily and betweenness centrality.
- Lastly, in Pakistan's culture, most of the men work outside of the home whereas women are expected to be the homemakers. This cultural norm implies that the network of the men is bigger on average as compared to the network of women. The higher average network size of men also results in higher topological diversity. The average value of the mobility-related features for the men is also higher as compared to that of women, as men's networks may contain many geographically scattered professional contacts.

Statistical Significance of Features

The statistical significance of different features used in our analysis is shown in Table 6.3. Avg(M) and Avg(F) indicate the average values of the feature for male and female subscribers in the dataset. The last column shows the difference between the average value of males and females along with the p-value calculated through a t-test. The distribution of each of these features has already been described in the last section; however, this table shows that age diversity and topological diversity are not as significant as compared to the other features.

Feature	Avg(M)	Avg(F)	Diff(F-M)
Number of Calls and SMS	10.54933	18.08716	7.54***
Degree Centrality	2.143208	1.990399	-0.15***
Network Embeddedness	0.051923	0.07631	0.025***
Betweenness Centrality	2035.78	544.2583	-1491.52***
Gender Diversity	0.075737	0.114932	0.04***
Age Diversity	0.066672	0.055562	-0.01*
Topological Diversity	0.310245	0.266003	-0.04*
Gender Homophily	0.843515	0.318313	-0.53***
Average Geographical Reach	92.50838	80.0252	-12.48***
Location Diversity	0.371363	0.358917	-0.012***

Table 6.3: Statistical significance of different features used for analyzing social networks of males and females.

*Avg(M) and Avg(F) indicate the average of the feature values for the males and females respectively. The last column shows the difference between the averages. ***: P-value <=0.001, **: 0.001<P-value<=0.01, *0.01<P-value<=0.1*

6.5 Predicting Educational Gender Disparity at the District Level

Our second goal in this project was to see how accurately we can predict the educational gender disparity at the district level given the CDR based features described in the last section. Given the individual level features, this is accomplished in three main steps.

Converting individual features to district level features

We first convert the individual level features calculated in the last step to district level features. Given the subscriber level features, we apply the mean, median and standard deviation operations to each of the features to get the district-level features. We further calculate the ratio features as the ratio of the district level average feature value of females to the same for males. Furthermore, we also calculate the proportion features as the ratio

of the district level average feature value of females to the same value for all subscribers in that district.

Top Feature Selection

Given the high number of district-level features that we have, it is important to eliminate redundant or useless features from the final model so that the accuracy of the model can be improved. Elimination of redundant features also helps in improving the interpretability of the model. We used a cross-validated Random Forest Classifier to rank features based on the value of R-squared, and the most optimal set consisting of 30 features was selected using Recursive Feature Elimination (RFE) (Guyon et al., 2002).

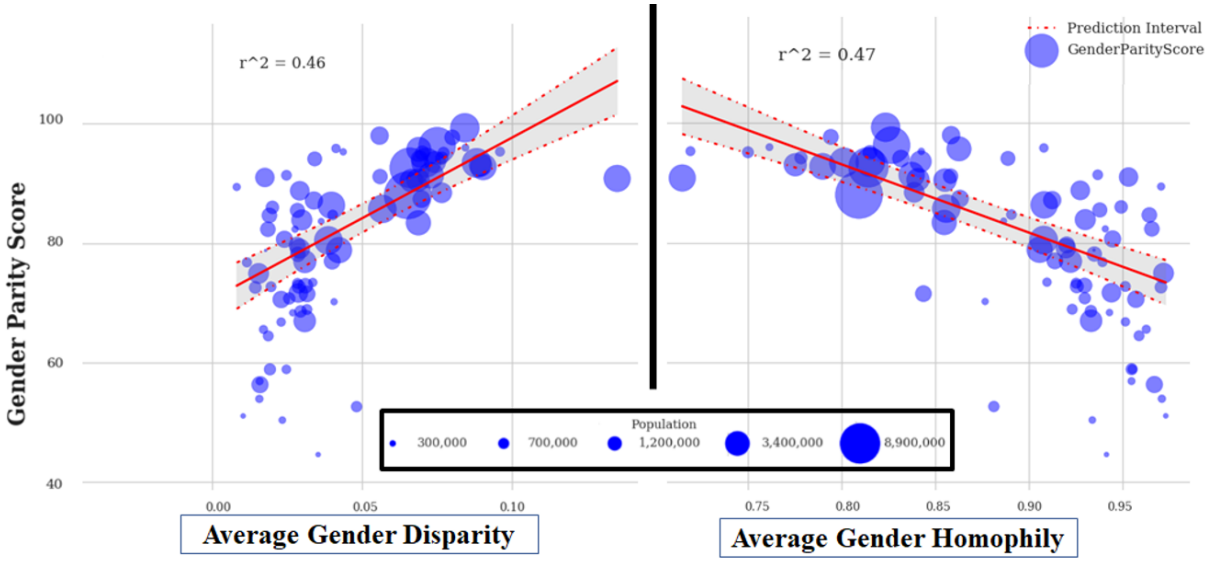


Figure 6.15: Top features selected through RFE. Left: Gender parity score vs Average gender disparity and Right: Average gender homophily of the males in a district

The top 4 features selected through RFE are listed below.

- Average gender diversity of males in a district
- Average gender homophily of males in a district
- Average embeddedness of all the users in a district
- Average geographical reach of all the users in a district

Figure 6.15 shows the relationship between the top 2 features selected through RFE and the gender parity score. The subfigure on the left shows the predictive performance of a linear least square regression model weighted by the district population and trained on the average gender diversity of males in a district. The subfigure on the right shows the predictive

performance of a linear least square regression model weighted by the district population and trained on the average gender homophily of males in a district. The R-squared of these models on the training dataset is 0.46 and 0.47 respectively.

Final Prediction Models

In the final prediction models, we use the top 30 features selected through the feature selection process described in the last step and build different machine learning models using these features. Gender disparity at the district level has not been a widely researched problem, so there is no consensus or existing baselines to serve as comparisons. We thus use network activity (Baseline Model 1), network size of the users (Baseline Model 2) and the ratio of female to male users in a district (Baseline Model 3) as the models against which to compare our models.

Given the top features selected through the feature selection process, we set up three different type of experiments. Detail of these experiments are as follows:

- **Experiment A: Prediction of district-level gender parity score**
  In this experiment, we use different regression models like Weighted Least Squares and Random Forest Regression to predict the gender parity score of each of the districts. In this experiment, the least square regression model weighted by the population of the districts outperformed other regression models. The R-squared for our best performing model in comparison to the baseline models is shown in Figure 6.16.

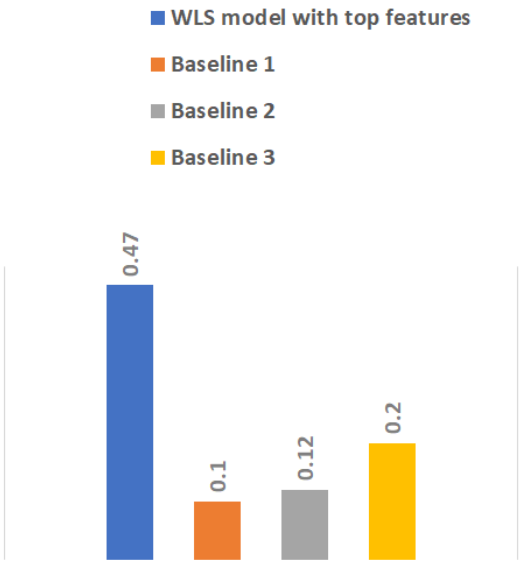


Figure 6.16: Results for Experiment A

- **Experiment B: Classifying districts into fine-grained categories based on the gender parity score**
  In this experiment, we have tried to classify the districts based on the gender parity score binned into seven different categories, based on the minimum and maximum value of the gender parity score. The intervals defining these categories are 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 and 90-99. The random forest classifier with top features selected through RFE outperforms other models for this experiment. The performance of random forest classifier and baseline models for this experiment is shown in Figure 6.17.

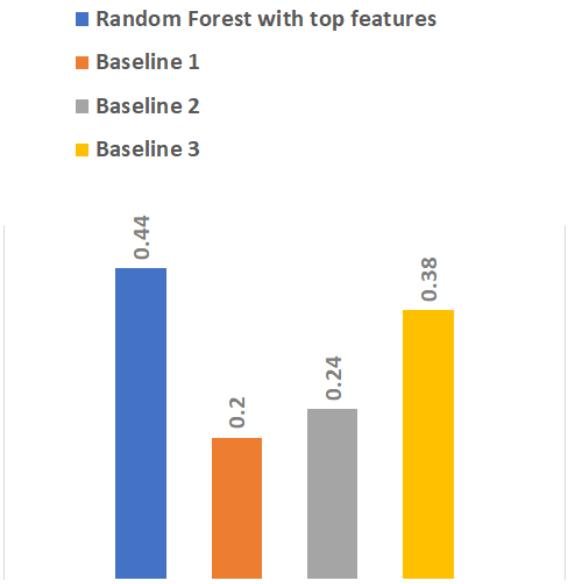


Figure 6.17: Results for Experiment B

- **Experiment C: Classifying districts as having low, medium or high gender parity score**
  Some of the categories in experiment B have very few districts. For instance, only two districts have the gender parity score between 30-40. In experiment C, the gender parity score is binned into three different categories: Low, Med and High such that each of these categories has an almost equal number of districts. In this experimental setting, the districts having gender parity score less than 75 are classified as low, the interval 75-89 corresponds to a medium gender parity score, and greater than 89 corresponds to the high category. Just like the case of experiment C, the random forest classifier with top features outperforms other models for this experiment. The performance of random forest classifier and baseline models for this experiment is shown in Figure 6.18.

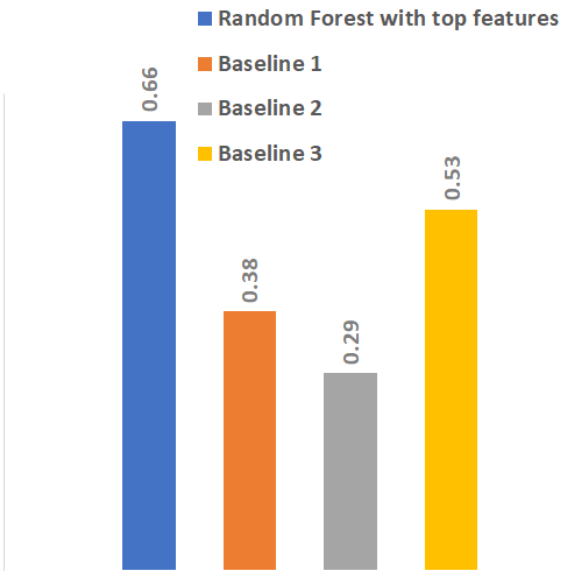


Figure 6.18: Results for Experiment C

For each of experiments A, B and C we have evaluated different machine learning models using 10-fold cross-validation with a train/test ratio of 80:20. It is clear from Figure 15 that our approach beats the baseline models by a significant margin. The weighted least squares regression model used in experiment A and trained on the top features selected through RFE beats the weighted least squares model trained on individual top features as well. Similarly, the Random Forest classifier with top 30 features selected using RFE outperforms baseline models for both experiments B and C (Figure 16 and 17). Performance of the Random Forest classifier with top features is much better for experiment C as compared to experiment B as the class distribution is balanced.

Discussion

In this project, we have analyzed the differences in social networks of men and women in a developing country using call detail records, with the aim of developing a predictive model to predict gender disparity at the district level. Our first contribution in this project is to demonstrate the suitability of CDR-based social networks for research on gender disparities. For large-scale studies like the one discussed here, CDR-based data can provide many advantages. Collecting CDR data does not require a great deal of additional investment; even the poorest of countries have seen a good penetration of mobile phones. Secondly, in comparison to other online social networks like Facebook and Twitter, mobile phones have seen wider adoption across all segments of the society, whereas networks like Facebook and Twitter are relatively more popular among youth, urban residents, and higher income classes. Many notable differentiating patterns can be seen between the social networks of men and women in Pakistan. The rich information contained in the call detail records en-

ables us to compare the social networks of men and women not only according to simpler features like size of the network and the activity on the network. We also have advanced measures related to network formation and network status, which can provide much more interesting information about the salient differences between the social networks of men and women in developing world. As explained in Section 4, the social network of men and women have statistically significant differences in terms of call volume, network size, embeddedness, gender homophily, and average geographical reach. Interestingly, women in Pakistan use the mobile phone network more actively but consistently have a smaller network size. This pattern indicates that women either prefer relatively stronger ties with fewer nodes in their network, as compared to men or are constrained to do so. However, as the percentage of women with higher embeddedness is greater than the corresponding percentage of men, this shows that women may have a more central position in the networks. Females show higher gender homophily in accordance with the prevalent social norms of the Pakistani society. Furthermore, the trends of gender and diversity are also in accordance with the prevalent social norms of the Pakistani society, wherein the females have more central roles in the families while the men are the breadwinners. As most of the professional workforce in Pakistani society consists of men, therefore, men, as expected, have higher network size and topological diversity. Furthermore, the network of the men is geographically more spread. Each of the behavioral features discussed in this report casts a different light on the social networks of men and women in developing countries. On the one hand, these differences highlight how men and women organize their social networks, while on the other hand, these differences highlight how the men and women can be susceptible to diffusion of information and opportunities through their networks. As machine learning models can play a helpful role in the spread of different initiatives (e.g., digital financial services, health services), the knowledge of the factors influencing the diffusion of information can help immensely in the success of these initiatives. A critical question that requires further exploration is whether the features selected by the feature selection process represent trends in the society or not. We intend to handle this question in our future research as our focus in this project has been on selecting the features which result in the best performance for the machine learning models. Some of the top features selected by the RFE algorithm are easier to interpret in the social context while some others are not easy to interpret. For example, among the top features, gender diversity is positively correlated with gender parity while gender homophily is negatively correlated with the gender parity. Higher gender diversity of males in a district indicates a relatively higher activity of females in the district which can show the better social status of the females in the district. On the contrary, higher gender homophily of males may be an indicator of lower network activity of females in the district. Similarly, the higher geographical reach of the individuals of a district on the average may also be positively correlated with higher HDI of the district. However, the relationship of higher average embeddedness of the users in a district with the gender parity may not be that obvious. Many of the features selected by the RFE algorithm have positive correlations with the network activity, but it is not obvious why these features are more important as compared to the network activity in general. The correlation of these features with the actual social trends is a research topic on

its own and needs much more attention, but it was not the focus of this project. Lastly, the predictive model we have developed beats the performance of other baseline models by quite some margin. Not only we can predict the raw gender parity score, but we are also able to relatively accurately classify the districts as having low, medium or high gender parity scores. From the perspective of the government and social work organizations, we think that this classification model will be useful for these organizations to align their resources and initiatives in the districts with lower gender disparity. In the absence of such classification models, organizations are either dependent on the statistics collected by the government organizations or the surveys conducted by the social welfare organizations. Government level data collection and analysis can demand a great deal of time and resources, while the surveys collected by the social welfare organizations are usually not comprehensive. Our study opens some other interesting questions for future exploration as well. For example, do the country-level social network findings correlate with the provincial-level social networks, or do the more progressive provinces have different patterns? Furthermore, to what extent do the patterns found in this society hold for countries with similar cultural and socio-economic background as Pakistan? We intend to handle these questions in future work. Similarly, the application of deep neural networks based models on the population level CDR data is another interesting area for future research.

6.6 Acknowledgements

The financial support for this project by Data2x under Big Data for Gender Challenge.

Chapter 7

Discussion

In this dissertation, I have shown that how machine learning over mobile communication metadata can be used to accurately model different problems related to poverty research in the developing world. My main contribution to the field of machine learning is the development of two algorithms for feature engineering or feature learning that are able to beat the existing state of the art for different problems related to poverty research in terms of prediction accuracy.

DFA based approach (an example of semi-automatic feature engineering) for feature engineering is able to combine domain expertise and general practices for feature engineering resulting in an extensible and scalable method for feature engineering. This method can be easily used by the researchers and the industry practitioners. I have used DFA based approach for feature engineering over CDR datasets from Ghana, Pakistan, Zambia, Afghanistan, and results have been quite impressive. International Finance Corporation, a subsidiary of World Bank, has used this approach for feature engineering over multiple CDR datasets from different years for modeling product adoption. In one trial of such methods for feature engineering, the reported improvement in product adoption was 30% more than the existing methods.

It is worth noting that the DFA based approach can be used on network datasets other than the CDR dataset as well. An additional advantage of DFA based approach is that the features it generates are highly interpretable.

Methods for feature engineering can improve the performance of machine learning models in almost every cases. In the second phase of my research I have shown that Multi-GCN (a neural network models that can incorporate the multi-layer nature of the CDR datasets) can beat the performance of the current state of the art models like Node2vec (Grover and Leskovec, 2016), Deepwalk (Perozzi et al., 2014), etc. not only for the problems related to poverty research but more traditional computer science problems like node classification in citation networks as well.

There are quite a few avenues for future research based on the content presented in this dissertation. DFA based approach though generally accurate and extensible can be made more efficient by combining feature evaluation with the feature generation process. In other

words, it would be better to prune those paths of DFA that lead to non-informative features.

Multi-GCN based approach for machine learning over CDR networks requires extensive hyper-parameter tuning. An end-to-end method approach for machine learning over CDR networks is the logical next step for research in this area.

Lastly, there are quite a few research questions related to the analysis of social networks in the developing world that we could not address due to time constraints. Identification and modeling of cascades in the CDR networks, structural characteristics of different types of events in the networks are some of the questions that can benefit from the content described in the Chapter 6 of this dissertation.

To conclude, this dissertation shows that machine learning over CDR networks can help the researchers to better model different phenomenon in the developing world. Commercial applications and research studies both can benefit from the methods of feature extraction presented in this thesis.

Bibliography

Aker, J.C., R. Boumnijel, A. McClelland, and N Tierney (2014). "How do electronic transfers compare? Evidence from a mobile money cash transfer experiment in Niger". *CGD Working Paper 268* Washington, D.C.

Archaux, Cedric, Arnaud Martin, and Ali Khenchaf (2004). "An SVM based churn detector in prepaid mobile telephony". *Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on*. IEEE, 2004, 459–460.

Assael, Yannis M, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas (2016). "Lipnet: End-to-end sentence-level lipreading". *arXiv preprint arXiv:1611.01599*.

Blondel, Vincent D, Adeline Decuyper, and Gautier Krings (2015). "A survey of results on mobile phone datasets analysis". *EPJ Data Science* 4, 10.

Blumenstock, Joshua E (2012). "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda". *Information Technology for Development* 18, 107–125.

Blumenstock, Joshua E, Michael Callen, Tarek Ghani, and Lucas Koepke (2015). "Promises and pitfalls of mobile money in Afghanistan: evidence from a randomized control trial". *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*. ACM. 2015, 15.

Blumenstock, Joshua E, Nathan Eagle, and Marcel Fafchamps (2016). "Airtime Transfers and Mobile Communications: Evidence in the aftermath of natural disasters". *Journal of Development Economics (forthcoming)*.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On (2015). "Predicting poverty and wealth from mobile phone metadata". *Science* 350, 1073–1076.

Blumenstock, Joshua and Nathan Eagle (2010). "Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda". *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. ACM. 2010, 6.

Bonchi, Francesco, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes (2011). "Social network analysis and mining for business applications". *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 22.

Bruna, Joan, Wojciech Zaremba, Arthur Szlam, and Yann LeCun (2013). "Spectral networks and locally connected networks on graphs". *arXiv preprint arXiv:1312.6203*.

Buckee, Caroline O, Amy Wesolowski, Nathan N Eagle, Elsa Hansen, and Robert W Snow (2013). "Mobile phones and malaria: modeling human and parasite travel". *Travel medicine and infectious disease* 11, 15–22.

CGAP (2013). *The Power of Social Networks to Drive Mobile Money Adoption*. Tech. rep. 2013.

Chen, Gregory and Stephen Rasmussen (2014). *bKash Bangladesh: A Fast Start for Mobile Financial Services*. Tech. rep. 2014.

Dasgupta, Koustuv, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A. Nanavati, and Anupam Joshi (2008). "Social ties and their relevance to churn in mobile telecom networks". *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, 2008, 668–677.

Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst (2016). "Convolutional neural networks on graphs with fast localized spectral filtering". *Advances in Neural Information Processing Systems*. 2016, 3844–3852.

Demirguc-Kunt, Asli, Leora F Klapper, Dorothe Singer, and Peter Van Oudheusden (2015). "The Global Findex Database 2014: measuring financial inclusion around the world". *World Bank Policy Research Working Paper*.

Dermish, Ahmed, Christoph Kneiding, Paul Leishman, and Ignacio Mas (Oct. 2011). "Branchless and Mobile Banking Solutions for the Poor: A Survey of the Literature". *Innovations: Technology, Governance, Globalization* 6, 81–98. doi: 10.1162/INOV˙a˙00103.

Di Castri, Simone and Lara Gidvani (Feb. 2014). *Enabling Mobile Money Policies in Tanzania: A "test and learn" approach to enabling market-led digital financial services*. Tech. rep. GSMA, 2014.

Domingos, Pedro (2012). "A few useful things to know about machine learning". *Communications of the ACM* 55, 78–87.

Dong, Xiaowen, Pascal Frossard, Pierre Vandergheynst, and Nikolai Nefedov (2014). "Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds". *IEEE Transactions on signal processing* 62, 905–918.

Dong, Yuxiao, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla (2014). "Inferring user demographics and social strategies in mobile social networks". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, 15–24.

Donovan, Kevin (2012). "Mobile Money for Financial Inclusion". Information and Communications for Development. The World Bank, 2012, 61–73.

Dwyer, F. Robert (Sept. 1997). "Customer lifetime valuation to support marketing decision making". en. *J. Direct Mark.* 11, 6–13.

Eagle, Nathan, Michael Macy, and Rob Claxton (2010a). "Network diversity and economic development". *Science* 328, 1029–1031.

— (2010b). "Network diversity and economic development". *Science* 328, 1029–1031.

Etim, Alice S (2014). "Mobile banking and mobile money adoption for financial inclusion". *Research in Business and Economics Journal* 9, 1.

Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber (2018). "Using Facebook ad data to track the global digital gender gap". *World Development* 107, 189–209.

Francis, Eilin, Joshua Blumenstock, and Jonathan Robinson (2017). "Digital Credit: A Snapshot of the Current Landscape and Open Research Questions". *CEGA White Paper*.

Frias-Martinez, Vanessa, Enrique Frias-Martinez, and Nuria Oliver (2010a). "A Gender-centric Analysis of Calling Behavior in a Developing Economy". *AAAI Symposium on Aritificial Intelligence and Development* Forthcoming.

— (2010b). "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records." 2010.

Frias-Martinez, Vanessa, Jesus Virseda, and Enrique Frias-Martinez (Feb. 2012). "Socio-Economic Levels and Human Mobility". *Jourhnal of Information Technology for Development*, 1–16.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". *Annals of statistics*, 1189–1232.

Gatys, Leon A, Alexander S Ecker, and Matthias Bethge (2015). "A neural algorithm of artistic style". *arXiv preprint arXiv:1508.06576*.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei (2017). "Using deep learning and google street view to estimate the demographic makeup of the us". *arXiv preprint arXiv:1702.06683*.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel (Apr. 2006). "Extremely Randomized Trees". *Mach. Learn.* 63, 3–42.

Grover, Aditya and Jure Leskovec (2016). "Node2Vec: Scalable Feature Learning for Networks". *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016. doi: 10.1145/2939672.2939754.

GSMA (2016). *Unlocking Rural Coverage: Enablers for commercially sustainable mobile network expansion*. Tech. rep. 2016.

Gupta, Sunil and Valarie Zeithaml (Nov. 2006). "Customer Metrics and Their Impact on Financial Performance". *Marketing Science* 25, 718–739.

Gutierrez, Thoralf, Gautier Krings, and Vincent D. Blondel (Sept. 2013). *Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets*. arXiv e-print 1309.4496. 2013.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik (2002). "Gene selection for cancer classification using support vector machines". *Machine learning* 46, 389–422.

Hart, Christopher, James Heskett, and W. Earl Sasser Jr. (July 1990). "The Profitable Art of Service Recovery". *Harvard Business Review*.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T. Hastie, J. Friedman, and R. Tibshirani (2009). *The elements of statistical learning*. Vol. 2. 1. Springer, 2009.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". *Proceedings of the IEEE international conference on computer vision*. 2015, 1026–1034.

Intermedia (n.d.). *Mobile Money Use and Gender: Less Disparity than Might be Expected*. N.d.

Jack, William and Tavneet Suri (2014). "Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution". *American Economic Review* 104, 183–223. doi: 10.1257/aer.104.1.183.

Jackson, Linda A, Yong Zhao, Anthony Kolenic III, Hiram E Fitzgerald, Rena Harold, and Alexander Von Eye (2008). "Race, gender, and information technology use: the new digital divide". *CyberPsychology & Behavior* 11, 437–442.

Jain, Dipak and Siddhartha S. Singh (Mar. 2002). "Customer lifetime value research in marketing: A review and future directions". en. *J. Interactive Mark.* 16, 34–46.

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon (2016). "Combining satellite imagery and machine learning to predict poverty". *Science* 353.

Karnstedt, Marcel, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes (2011). "The effect of user features on churn in social networks". *Proceedings of the 3rd International Web Science Conference*. ACM, 2011, 23.

Kelly, Heather (2017). "Google's Project Sunroof calculates solar cost". *CNN. Retrieved* 15.

Khan, Muhammad R. and Joshua E. Blumenstock (2016). "Predictors Without Borders: Behavioral Modeling of Product Adoption in Three Developing Countries". *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016.

— (2017). "Determinants of Mobile Money Adoption in Pakistan". *31st Annual Conference on Neural Information Processing Systems, Workshop on Machine Learning for the Developing World*. 2017.

— (2018a). "Gender Disparity Signals. Analyzing gender disparities with mobile phone metadata". To be submitted to ICWSM 2019. 2018.

— (2018b). "Multi-GCN: Graph Convolutional Networks for Multi-View Networks with Applications to Global Poverty". Forthcoming in AAAI 2019. 2018.

Khan, Muhammad Raza and Joshua E Blumenstock (2016). "Machine Learning Across Cultures: Modeling the Adoption of Financial Services for the Poor". *Proceedings of the 2016 ICML Workshop on Data4Good: Machine Learning in Social Good Applications*. 2016.

Khan, Muhammad Raza, Joshua Manoj, Anikate Singh, and Joshua Blumenstock (2015). "Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty". *Big Data (BigData Congress), 2015 IEEE International Congress on*. IEEE. 2015, 677–680.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*.

Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". *International Conference on Learning Representations (ICLR)*. 2017.

Kusimba, Sibel, Harpieth Chaggar, Elizabeth Gross, and Gabriel Kunyu (2013). "Social networks of mobile money in Kenya". *Institute for Money, Technology & Financial Inclusion (IMTFI), Working Paper* 1.

Lemmens, Aurélie and Christophe Croux (May 2006). "Bagging and Boosting Classification Trees to Predict Churn". *Journal of Marketing Research* 43, 276–286.

Leskovec, Jure, Lada A Adamic, and Bernardo A Huberman (2007). "The dynamics of viral marketing". *ACM Transactions on the Web (TWEB)* 1, 5.

Liu, Wei, Jun Wang, and Shih-Fu Chang (2012). "Robust and scalable graph-based semisupervised learning". *Proceedings of the IEEE* 100, 2624–2638.

Lu, Xin, Linus Bengtsson, and Petter Holme (2012). "Predictability of population displacement after the 2010 Haiti earthquake". *Proceedings of the National Academy of Sciences* 109, 11576–11581.

Magno, Gabriel and Ingmar Weber (2014). "International gender differences and gaps in online social networks". *International Conference on Social Informatics*. Springer. 2014, 121–138.

Mas, Ignacio and Dan Radcliffe (2011). "Scaling mobile money". *Journal of Payments Strategy & Systems* 5, 298–315.

McCulloch, Warren S. and Walter Pitts (Dec. 1943). "A logical calculus of the ideas immanent in nervous activity". en. *The bulletin of mathematical biophysics* 5, 115–133. doi: 10.1007/BF02478259.

McKay, Claudia and Michelle Kaffenberger (2013). *Rural Versus Urban Mobile Money Use: Insights from Demand-Side Data*. 2013.

Medhi, Indrani, Aishwarya Ratan, and Kentaro Toyama (2009). "Mobile-banking adoption and usage by low-literate, low-income users in the developing world". *Internationalization, Design and Global Development*. Springer, 2009, 485–494.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality". *Advances in neural information processing systems*. 2013, 3111–3119.

Minischetti, Elisa (2016). *Taking a look at women's financial inclusion via mobile money – Barriers and drivers to the mobile money gender gap in Rwanda*. 2016.

— (2017). *Examining the financial inclusion of women – the mobile money gender gap in Rwanda*. 2017.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist (2011). "Understanding the Demographics of Twitter Users." *ICWSM* 11, 25.

Morawczynski, Olga (Oct. 2009). "Exploring the usage and impact of "transformational" mobile financial services: the case of M-PESA in Kenya". *Journal of Eastern African Studies* 3, 509–525. doi: 10.1080/17531050903273768.

Murendo, Conrad, Meike Wollni, Alan De Brauw, and Nicholas Mugabi (2017). "Social network effects on mobile money adoption in Uganda". *The Journal of Development Studies*, 1–16.

Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason (2006). "Defection detection: Measuring and understanding the predictive accuracy of customer churn models". *Journal of marketing research* 43, 204–211.

Pan, Sinno Jialin, I.W. Tsang, J.T. Kwok, and Qiang Yang (Feb. 2011). "Domain Adaptation via Transfer Component Analysis". *IEEE Transactions on Neural Networks* 22, 199–210. doi: 10.1109/TNN.2010.2091281.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena (2014). "DeepWalk: Online Learning of Social Representations". *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: ACM, 2014. doi: 10.1145/2623330.2623732.

Qian, Zhiguang, Wei Jiang, and Kwok-Leung Tsui (2006). "Churn detection via customer profile modelling". *International Journal of Production Research* 44, 2913–2933.

Quercia, Daniele, Jonathan Ellis, Licia Capra, and Jon Crowcroft (2012). "Tracking gross community happiness from tweets". *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM. 2012, 965–968.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. (2017). "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". *arXiv preprint arXiv:1711.05225*.

Reed, Philip J., Muhammad Raza Khan, and Joshua Blumenstock (2016). "Observing Gender Dynamics and Disparities with Mobile Phone Metadata". *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. Ann Arbor, MI, USA: ACM, 2016. doi: 10.1145/2909609.2909632.

Reed, Philip J, Muhammad Raza Khan, and Joshua Blumenstock (2016). "Observing gender dynamics and disparities with mobile phone metadata". *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. ACM. 2016, 48.

Ryan, Louise, Rosemary Sales, Mary Tilki, and Bernadetta Siara (2008). "Social networks, social support and social capital: The experiences of recent Polish migrants in London". *Sociology* 42, 672–690.

Safaricom (2014). *FY14 Presentation*. 2014.

Scharwatt, Claire, Arunjay Katakam, Jennifer Frydrych, Alix Murphy, and Nika Naghavi (2014). *State of the Industry: Mobile Financial Services for the Unbanked*. Tech. rep. GSMA, 2014.

Smith-Clarke, Christopher, Afra Mashhadi, and Licia Capra (2014). "Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks". *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2014, 511–520.

Sundsøy, Pål, Johannes Bjelland, Asif M Iqbal, Yves-Alexandre de Montjoye, et al. (2014). "Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling". *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2014, 367–374.

Suri, Tavneet (2017). "Mobile Money". *Annual Review of Economics* 9, 497–520.

Tang, Jian, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei (2015). "Line: Large-scale information network embedding". *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, 1067–1077.

Toole, Jameson L, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González, and David Lazer (2015). "Tracking employment shocks using mobile phone data". *Journal of The Royal Society Interface* 12, 20150185.

Ugander, Johan, Lars Backstrom, Cameron Marlow, and Jon Kleinberg (2012). "Structural diversity in social contagion". *Proceedings of the National Academy of Sciences* 109, 5962–5966.

Uzzi, Brian and Jarrett Spiro (2005). "Collaboration and creativity: The small world problem". *American journal of sociology* 111, 447–504.

Verbeke, Wouter, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens (2012). "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach". *European Journal of Operational Research* 218, 211–229.

Von Luxburg, Ulrike (2007). "A tutorial on spectral clustering". *Statistics and computing* 17, 395–416.

Wei, Chih-Ping and I. Chiu (2002). "Turning telecommunications call details to churn prediction: a data mining approach". *Expert systems with applications* 23, 103–112.

Wesolowski, Amy, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O Buckee (2015). "Impact of human mobility on the emergence of dengue epidemics in Pakistan". *Proceedings of the National Academy of Sciences* 112, 11887–11892.

Xu, Chang, Dacheng Tao, and Chao Xu (2013). "A survey on multi-view learning". *arXiv preprint arXiv:1304.5634*.

Yang, Zhilin, William W Cohen, and Ruslan Salakhutdinov (2016). "Revisiting semi-supervised learning with graph embeddings". *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org. 2016, 40–48.

Zhang, Xiaohang, Ji Zhu, Shuhua Xu, and Yan Wan (2012). "Predicting customer churn through interpersonal influence". *Knowledge-Based Systems* 28, 97–104.

Zhou, Denny, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf (2004). "Learning with local and global consistency". *Advances in neural information processing systems*. 2004, 321–328.

Zhou, Denny, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf (2004). "Ranking on data manifolds". *Advances in neural information processing systems*. 2004, 169–176.

Zhu, Xiaojin (2005). *Semi-Supervised Learning Literature Survey*. Tech. rep. 1530. Computer Sciences, University of Wisconsin-Madison, 2005.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the Elastic Net". *Journal of the Royal Statistical Society, Series B* 67, 301–320.