UNIVERSITY OF CALIFORNIA
SANTA CRUZ

A COMPUTATIONAL APPROACH TOWARDS ONLINE CONSUMER TYPE CLASSIFICATION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Master of Science

in

TECHNOLOGY INFORMATION MANAGEMENT

by

Yuting Huang

December 2014

The Dissertation of Yuting Huang
is approved:

Professor Yi Zhang, Chair

Professor Subhas Desa

Professor Daniel Friedman

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

A Computational Approach towards Online Consumer Type Classification

by

Yuting Huang

Categorizing online consumers into Hedonic or Utilitarian (H/U) has long been known helpful for improving shopping experience and this problem has been investigated by scholars from economics and marketing domains over the last decade. Existing work focuses on identifying characteristics that best represent H/U at a qualitative level and there is a dearth of effort in the study of computational approach for the categorization problem. We are motivated to solve the categorization problem by utilizing methods developed in data science field. The present paper employs a machine learning based approach which categorizes online consumers by leveraging available consumer behavioral data. We evaluated the proposed approach on a real world e-commerce data set. The experimental result demonstrates the feasibility of classifying consumers just based on small number of high leveled behavioral data. In the future, better classification accuracy could be achieved by incorporating more sophisticated information about consumer.

Keywords:

hedonic, utilitarian, classification, online shopping, crowdsourcing

Acknowledgments

This thesis paper can not be possible without financial support, guidance and advices from my parents, teachers and friends. I would like express gratitude toward the following significant advisors:

I sincerely thank Professor Yi Zhang who guided me into the world of data science, passing me knowledge of data mining and machine learning. She also encouraged me to apply data science techniques to solve a real-world problem and instructed all the way along from detection of a problem, to development of treatments, until the completion of the present paper.

I sincerely thank Professor Daniel Friedman who is most insightful regards experiment design. He offered valuable suggestions on improving performances of experiments.

I sincerely thank Professor Subhas Desa who expertises in product design. He provided industrial thoughts into my methodologies.

I also sincerely thank Qi Zhao who has answered various of questions along my way to learn new techniques. His enriched experiences in data analysis, applied machine learning helped me with much less detour to accomplish goal.

Chapter 1

Introduction

With the rapid development of the Internet technology and logistic, online shopping nowadays has become an indispensable part of hundreds of millions of people's daily lives. Online consumers are exposed to much larger number of products than ever before and become tireless of seeking deals which they are interested in. On the other hand, online shopping presents merchants great opportunities of reaching ever largest consumer base. One of the most important problems with online shopping is how to improve the shopping experience through better personalization, namely, tailoring the delivery of product information on an individual base. This problem has drawn significant attention from both industry and academia as better personalization brings benefit for both consumers and merchants.

There is large volume of literature on studying online shopping by researchers from different domains. One of the interesting findings from previous researches in marketing and economics reveals that consumers assess their shopping experience not

just by considering the quality of products or services, but also by emotional costs and benefits [3]. Hedonic and Utilitarian are two most important dimensions measuring emotional outcomes. Generally, Utilitarian consumers are ergic, task-related and rational and consider shopping a work mentality. Efficiently find targets and accomplish tasks would relieve Utilitarian consumers from "dark side of shopping" [4]. On the other side, however, Hedonic consumers perceives, other than product value, entertainment and emotional worth from "festive" shopping, with or even without purchases [4].

Business people have been making good use of such knowledge in order to explore potential shopping abilities of customers. Marketers are especially interested in targeting Hedonic consumers for Hedonics are known of impulsive buyers and being easily influenced [11]. Nowadays, it is common to have such sections as "You might also like" or "Customer who viewed this item also viewed" on one's Amazon web pages. Macy's gives out free makeup samples with purchases of skincare products. No doubt is there that these recommendation approaches, to a certain extent, stimulate customers' shopping desire [4]. Nevertheless, in the meanwhile, same recommendation approaches might make shopping experiences of Utilitarian consumers even more "miserable", because additional recommended products result in heavier "workload", lower efficiency and slower to accomplish tasks. In fact, Utilitarians can be more involved in "overnight shipping", "express checkout" or specs chart clearly comparing price, material/ingredient, or financial option etc. of several products.

We recognized that accurate classification of online social consumer would be beneficial to both merchants and individual consumers. On one hand, merchants would

have increased sales by providing Hedonic consumers with relevant and interesting recommendations and Utilitarian ones with express shopping; on the other hand, consumers would receive much more pleasant shopping experiences in the sense of either discovering more interesting products or finding the best deal in an efficient and accurate manner. However, it is non trivial task to classify the online shoppers automatically due to the enormous number of consumers and products.

While many pioneer researchers have studied shopping values, motivations, design artifacts about Hedonic and Utilitarian consumers, very few of them have investigated classification models for online consumers from the perspective of data science. Inspired by the discoveries and insights of existing researches, in this paper, we present a principled approach which classifies consumers into Hedonic or Utilitarian by building machine learning models on top of consumer shopping behavioral data. To the best of our knowledge, our work is the first attempt to approach the H/U type consumer classification from a computational perspective.

In the remainder of this paper, we first reviewed related literature and described characteristic of Hedonic and Utilitarian. Secondly, we discussed approaches to conduct experiments as well as to collect result, including experiments design, incentive strategy, website development etc. After processing feature engineering to generate training data set, we built multiple classification models by applying various machine learning algorithms and evaluated those classifiers on such characteristic as accuracy etc. Later on, we also talked about limitation of present paper and proposed approaches to make improvement in future.

Chapter 2

Related Work

We investigated characteristics of the two types of consumers, Hedonic and Utilitarian, by looking into their shopping values, motivation etc. Summarized comparison of characteristics of H/U are presented in Table 2.1

Hedonic

Hedonic consumers enjoy the experience of shopping itself very much, much more than the need to purchase products.

.

Hedonic	**Utilitarian**
Shop for fun.	Narrow, goal-focused user.
Curious, creative. Explore popular, innovative products.	Informativeness. Weight products' pros and cons.
Impulsive buyers. Make unplanned purchases	Rational. Only access products fitting needs.
Less price driven. But shopping for discounts and bargains	Efficient. Quickly find, buy and leave.

Table 2.1: Characteristic of Hedonic and Utilitarian

- Shopping stimulates Hedonic consumer's feeling [4]. Increased arousal, heightened involvement, perceived freedom, fantasy fulfillment, and escapism all may indicate entertainment and emotional worth [5].

- Hedonic consumers communicate, shop, and explore interesting, popular, innovative products with other shoppers [11]. So they keep up with trends and fashions [4].

- Hedonic consumers are likely to make unplanned shopping for solely satisfaction of emotional needs such as stress relief or the sense of achievement out of finding the perfect deal [11].

- The "bargains" process is to pursue transaction utility which is the difference between a product's selling price and a consumer's internal reference price [14]. Hedonic consumers perceive increased sensory involvement and excitement from "bargains" [4].

Utilitarian

- Utilitarian consumers only perceive satisfaction through completion of purchases and minimization of time and effort expenditures [4].

- Many Utilitarian respondents in previous experiments find shopping tedious and hard. They tend to consider shopping as a task, a mission to accomplish [16].

- Utilitarian consumers are more motivated to pursue the right product. So they

prefer to reach sufficient, reliable information about candidate products, which helps them to identify the right one.

- Utilitarian consumers are task-oriented and rational, so they tend to avoid making impulsive, unplanned purchases [11]. The difficulty of finding satisfied products may make them irritated [4].

- Utilitarian consumers expect quick, efficient shopping approaches [11] such as direct access to products, express checkout which save their effort.

Chapter 3

Proposed Approach

We believe that accurate and efficient consumer categorization provides merchants opportunities of gaining large profit margin by offering customers personalized products and services. Having studied characteristics that most precisely define the two types of consumers, we are motivated to implement computational approaches for consumer categorization in the context of a real-world typical modern e-commerce site. On our way to implement consumer type categorization, we proposed methodologies to construct training data set, to compare and analyze modeling algorithms as well as to discuss possible channels for improvement.

Step 1:

Besides manipulate massive data set, it is necessary to acquire labels for consumers through crowdsourcing. As the dataset itself does not come with consumer types, it demands to assign labels manually. We resorted to principles and experience from

crowdsourcing for this task.

Step 2:

Examine a variety of machine learning algorithms on a sample of real world e-commerce dataset.

Step 3:

Gain insights out of the resulted machine learning models and discussing the possible directions of further improvement in the future.

Chapter 4

Experiment

The experiments were conducted to acquire labels on consumer types through crowdsourcing. We first extracted structured data from Apache raw logs which carries consumer shopping history. Then, we designed rules, incentive strategy to hire people for working on labeling. In the end, we analyzed and refined experiment results, preparing for modeling afterwards.

4.1 Structured data

The raw data of consumer shopping records are from Apache raw logs of a real-world typical modern e-commerce site. The shopping site provides products and services covering departments of clothing, beauty, home, electronics, baby, travel etc. to people from as many as nine countries/areas, including US. The Apache logs store consumer's shopping behavioral history, recording behaviors like login, sign up, browsing, searching, clicking, purchase etc. In summary, the raw data renders as a collection of

diverse consumer behavioral records. As mentioned earlier that the dataset comprises raw Apache log, further process steps are required to make it ready for machine learning modeling. Specifically, the following steps are taken,

Step 1: Organize user information

There are approximately 5.5 million records in total with around 2 million unique consumers (UserID) and 2.6 million sessions, indicating averagely 2.8 records per consumer and 1.3 sessions per consumer. This rate was much lower than what we expected since we were trying to learn from multiple records. As a result, we decided to eliminate noisy data, which were of little value to learning procedure. We processed elimination by,

- Discarding consumers of less than 5 records. We deem that it is challenging labels for consumers of few records.
- Discarding consumers of more than 100 records. We think that it would involve intensive human effort to examine a large number of records.

We organized user activity in the following format. Note that some of the fields might be missing and this phenomenon poses challenges to our later machine learning procedure.

⟨UserID, Time, Behavior, ProductID, ProductName, SessionID, price, category⟩

Step 2: Sampling

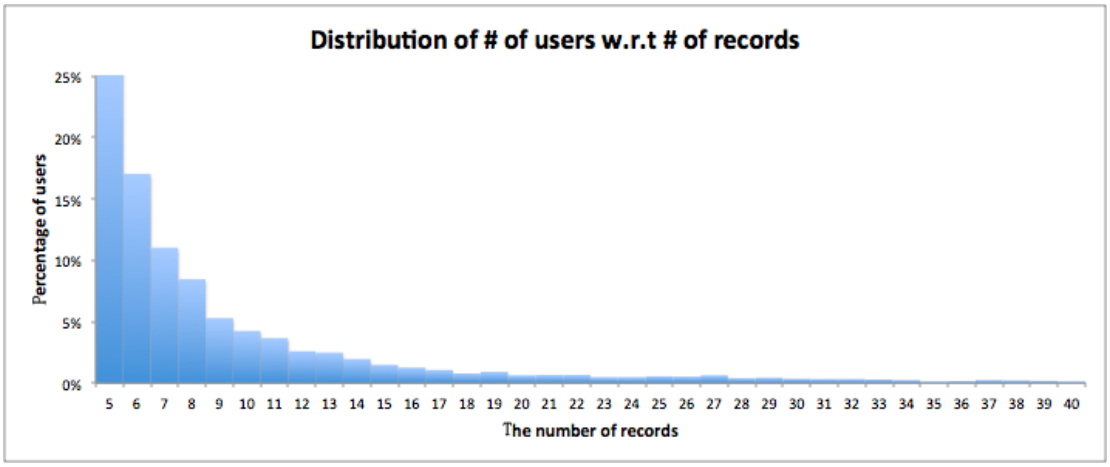

Figure 4.1: distribution of # of users w.r.t # of records

As the data set size is considerably large and the type of each user is unknown, it would be costly to label all users manually. Hence, instead of working with the whole dataset, we strategically sampled the dataset by,

- From the rest of the consumers, we randomly selected 2.5% - 4602 consumers out of the whole, as our experiment dataset. We assume that the 4602 samples are representative of the whole.

We assume that the methodology of randomization ensures representativeness of samples. Figure 4.1 presents the distribution of number of consumers with respect to number of their own records. We found that of the selected 4602 samples the majority ranks between 5 to 15 records.

Step 3: Convert data set into JSON files

We generated 4602 JSON files in total, of which each corresponds to one con-

sumer in sample data set, including all shopping records of the consumer. JSON files were all stored on server and we performed consumer behavioral data in user readable format on website, one file per page.

4.2 Acquire label through crowdsourcing

Crowdsourcing was employed to obtaining H/U labels for each user in the dataset sample. We formulated the labeling task as question which asks agents to choose the appropriate label based on the shopping history about the consumer. We are concerned that labeling result from an individual worker might biased, and researchers from crowdsourcing area claim that approximately 4 non-expert labels lead to expert-level label quality [17]. As a result, we had each consumer/question to be labeled by 5 different workers and used the most voted label for machine learning modeling.

Crowdsourcing experts also claim that general performance of workers can be improved by cheap but well designed incentive mechanism [17]. We planned to pay workers based on the quantity and quality of their answers. In practice, our reward strategy includes BASE bonus (quantity purpose) and EXTRA bonus (quality purpose) if the agent's label receives no less than 40% vote of the total.

In order to find the optimal number for the bonus, we initiated two small scaled experiments with different amount of bonus: group A with \$0.02(BASE) + \$0.03(EXTRA) per question and group B with \$0.01(BASE) + \$0.02(EXTRA) per question. Much to our surprise, the performance from Group B was better than that from group

A while at a much lower cost! A proper explanation might be the agents are rational and motivated to act to max their financial gain at lowest (time) cost. Thus, agents in group A are motivated to answer as many questions as possible without worrying about the quality of their labels. As a result, we adopted the \$0.01(BASE) + \$0.02(EXTRA) per question strategy for the rest of experiments. Afterwards, we also evaluated performances of workers and blocked bad quality worker from further participation.

Rules to assign questions:

- Questions are initially assigned equal possibility.
- Questions with more labels have a higher priority to be assigned.
- Agent answers each question exactly once.

Rules for workers participation:

- Each question invites 5 agents to label. Agents do not know others' labeling results.
- Agents are paid by quantity and quality of their labels.
- Every worker receives a BASE bonus of \$0.01/question and an EXTRA \$0.02/question if the label receives at least 40% votes.

Block bad quality workers:

We determined workers' eligibility of doing labeling job by their responding Match Rate. If one worker's Match Rate is no more than 40%, we will block the worker

Label value	Number	Percentage
Hedonic (H)	4984	51.62%
Utilitarian (U)	4198	43.48%
Unknown (K)	474	4.9%
Total	9656	100%

Table 4.1: Experiment results before refinement

Label value	Number	Percentage
Hedonic (H)	898	48.78%
Utilitarian (U)	943	51.22%
Total	1841	100%

Table 4.2: Experiment results after refinement

from further participation. If one worker's label is the same with majority, we say it matches. Otherwise, it does not match.

4.3 Data analysis and processing

We ended up with a total number of 1853 answered questions and 9656 labels - approximately 5 labels per question. Table 4.1 reports the numbers and percentage of the three types of labels respectively.

There were slightly more labels of "Hedonic (H)" than those of "Utilitarian (U)", making ratio of "H" to "U" a little bit over 1.18. We then claimed that the experiment data were fairly balanced, while unbalanced training data set may result in a classifier that is biased towards this majority class. Besides, a small number of questions (474) were claimed difficult to label ("Unknown"). This makes sense because, for example, some Hedonic consumer might act differently when in urgent purchasing

of some product. In order to better prepare training data set, we took two following approaches to refine experiment data.

- Discard "Unknown" records. Now that "Unknown (K)" indicates unable to tell whether "Hedonic (H)" or "Utilitarian (U)", we think they are of less value for prediction. Therefore, we eliminated those 474 records labeled "Unknown (K)".
- Resolve each user an unique label. Realizing that each user (question) were labeled multiple times, we decided to adopt the value, which was labeled most times, as the unique label value to the respective user.

Table 4.2 shows labeling results after refinement. We found that ratio of H/U is much closer to 1, as performed in Figure 4.2, H/U distribution of experiment results after refinement. This means that refined data set was even more balanced, and better for modeling later on.

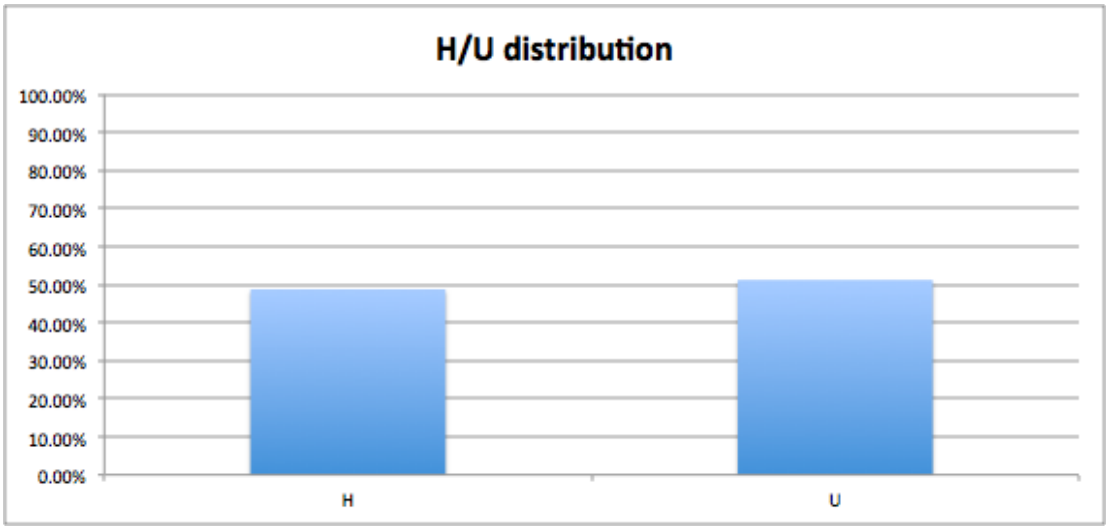

Figure 4.2: H/U distribution of experiment results after refinement

Chapter 5

Modeling

Having completed training data set, we are ready to apply machine learning algorithms for modeling. Firstly, we perform feature engineering, a process in which we proposed as many features as possible. Secondly, we imported the training data prepared in previous step into WEKA - a data mining software widely used by machine learning practitioners [2]. Thirdly, we closely examined 4 machine learning algorithms on the same training data. At last, we compared the performance of all classifiers and briefly discussed the reason behind the difference.

5.1 Feature engineering

Unfortunately, our present available structured data, a set of characteristics reflecting the two types of consumers, are not quite "understandable" to machines, as gap exists between user requirement domain and machine solution domain [15]. For example, machine would learn very little from data that someone shops at cate-

#	**Features**	**Motivation**
1	Average "click" price. Average price of all "click: records for each user.	The feature needs comparing with feature-4 on whether the two prices match each other.
2	Average "click" times needed for purchase. Average # of "click" prior to first "purchase".	The feature indicates whether the present user is a find-buy-go consumer or someone that likes spending time shopping around.
3	Average "purchase" price. Average price of all "purchase" records for each user.	The feature needs comparing with feature-2 on whether the two prices match each other.
4	Average duration. The time difference between the first records (any) to the first ?purchase? records for each user (unit: second)	The feature tells whether the consumer is a quick decider or a no-hurry one.
5	Conversion rate. The percentage that # of sessions ?click? leading to ?purchase? over # of sessions in total for each user.	The higher the conversion rate is, the more likely consumer?s willingness to purchase can be.
6	Frequency of "purchase"s. How many purchases does each user make per month.	The feature reflects consumer's potential consuming ability.
7	Diversity of categories. Average # of categories in each session for each user.	The feature indicates whether the consumer is interested in exploring different products.
8	Purchase percentage. The percentage that # of "purchase" records over # of all records for each user.	The feature reflects consumer's overall tendency to purchase.

Table 5.1: Features for training data set

#	**Attributes**	**Description**
1	UID	The primary key of table. No missing.
2	Average “click” price.	23% missing.
3	Average “click” times needed for purchase.	23% missing.
4	Average “purchase” price.	No missing.
5	Average duration.	23% missing.
6	Conversion rate.	23% missing.
7	Frequency of “purchase”s.	23% missing.
8	Diversity of categories.	49% missing.
9	Purchase percentage.	No missing.
10	label	The Y-value. No missing.

Table 5.2: preliminary analysis

gory of “Business|Calendars, Organizers & Address Books|Wall Calendars” and “Pet Supplies|Dogs|Clean-Up & Odor Control”. However, we can make data more informative by converting above into visiting 2 different categories per session. Therefore, it is necessary to perform feature engineering in order to guarantee training data set is informative to machine. Table 5.1 presents features to be used for model training and the motivation behind them.

5.2 Data Statistics

We generated training data set by combing features from feature engineering and labels from crowdsourcing. Table 5.2 reports a preliminary analysis of each of those 10 attributes from training data set.

#	Aspect	Bagging	Random Forest	Lostistic Regression	Gradient Boosting Trees
1	Accuracy	65.1276%	66.4856%	65.3992%	67.19%
2	TP rate	0.651	0.665	0.654	/
3	FP rate	0.351	0.336	0.348	/
4	Precision	0.652	0.665	0.654	/
5	Recall	0.651	0.665	0.654	/
6	F-Measure	0.650	0.665	0.653	/
7	Precision	0.302	0.329	0.307	/
8	Recall	0.695	0.705	0.700	/
9	F-Measure	0.683	0.684	0.673	/

Table 5.3: classification models

5.3 Algorithms and results

Weka provides a handful of machine learning algorithms which we can tinker with and we examined 3 representative algorithms, namely, Random Forest, Bagging and Logistic Regression. Besides, we also included Gradient Boosting Trees algorithm which is implemented in gbm package [1]. All of the chosen algorithms are considered to be decent without much parameter teaking.

5 fold cross validation is adopted for evaluating the performance of each algorithm and the results are summarized in Table 5.3. A brief description of each algorithm is also included.

5.3.1 Bagging

Bagging is an ensemble learning algorithm for classification which constructs multiple versions of sub-classifiers and obtains its class by aggregating outputs from

#	**Parameters of model**	**Value**
1	Test mode	5-fold cross-validation
2	bag size percentage	100
3	classifier	REP tree
4	num execution slots	1
5	num Iterators	12
6	seed	1

Table 5.4: Parameters for Bagging

#	**Parameters of model**	**Value**
1	initial Count	0.0
2	max Depth	-1
3	min Num	2.0
4	min Variance Prop	0.001
5	num Fold	3
6	seed	1

Table 5.5: Parameters for REP Trees (Bagging)

individual sub-classifiers [6]. Each sub-classifier is formed by applying a different bootstrap sample of the data set, a random sample with replacement of the training data set [7], and then fitting trees to the samples. Every single sub-classifiers is independent from others [6]. The aggregation does a simple majority vote to predict a class [13].

The present paper adopted REP tree making bootstrap replicates of data set as new learning set. Parameter settings for Bagging and REP Trees can be found in Table 5.4, Table 5.5 respectively. Via averaging outputs from multiple version of individual trees, Bagging reduces radiance and helps with avoiding overfitting (Wikipedia). We see Bagging has made notacible improvement in term of accuracy comparing with single REP Tree (Table 5.6).

#	**Algorithm**	**Accuracy**
1	Bagging (REP Tree)	65.1276%
2	REP Tree	63.3351%

Table 5.6: Accuracy for Bagging and REP Tree

#	**Parameters of model**	**Value**
1	Test mode	5-fold cross-validation
2	max Depth	3
3	num execution slots	1
4	num Features	4
5	num Trees	5
6	seed	1

Table 5.7: Parameters for Random Forest

5.3.2 Random Forest

Besides using different bootstrap sample of the data set when constructing individual trees, Random Forest differs only in one way from the above bagging approach. The algorithm uses a modified tree learning algorithm, selecting a random subset of the features other than the best subset during splitting process [13]. The reason to select subset randomly is to avoid one or a few features of significant importance being chosen repeatedly in many subsets . Correlation among sub-classifiers might bring about negative effect for improving accuracy. Parameter settings for Random Forest please see Table 5.7.

Random Forest resulted a slightly better classification model than Bagging did. This might be because none of features is of overwhelming importance.

#	**Parameters of model**	**Value**
1	Test mode	5-fold cross-validation
2	maxIts	-1
3	ridge	1.0E-8

Table 5.8: Parameters for Logistic Regression

#	**Parameters of model**	**Value**
1	Test mode	5-fold cross-validation
2	depth	4
3	fraction	0.5
4	shrinkage	0.001
5	distribution	multinomial
6	num of trees	5000

Table 5.9: Parameters for Gradient Boosting Trees

5.3.3 Logistic Regression

Logistic Regression is a probabilistic statistical classification model which makes use of both continuous and categorical variables and predicts binary outcomes [8]. This makes it a perfect algorithm to learn from our training data set. Parameter settings for Logistic Regression please find Table 5.8.

5.3.4 Gradient Boosting Tree

Gradient Boosting Trees is also an ensembling machine learning algorithm. It works by combining the result of multiple weak learners [9] [10]. Compared to other boosting algorithms like AdaBoost, it features by fitting subsequent learning function to the residual error of existing function. This algorithm performs well for dataset of sufficient samples and relatively small number of features. The usage of decision tree as

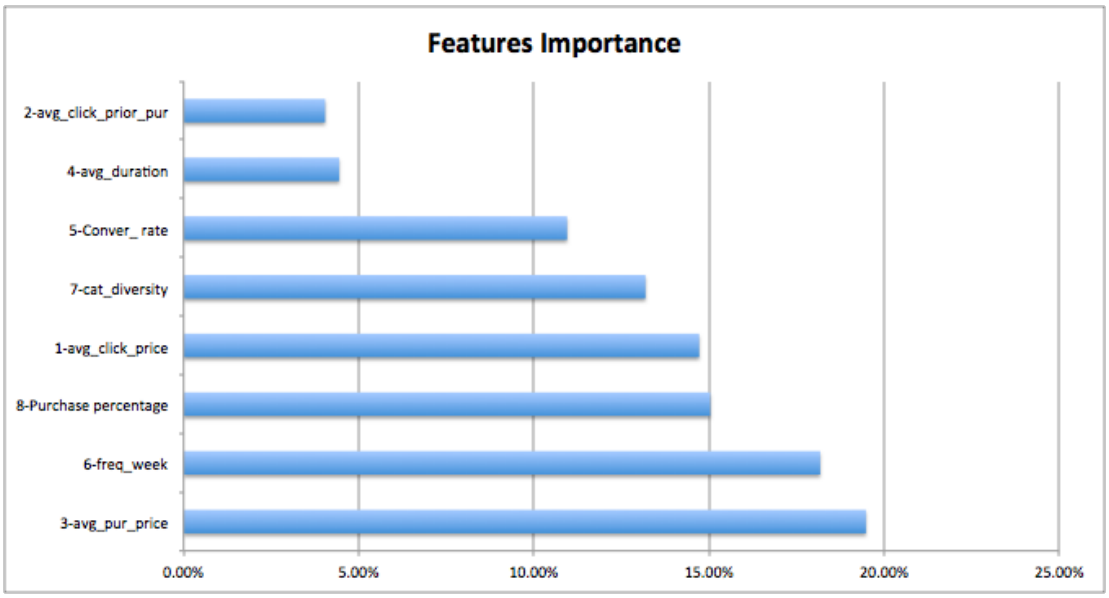

Figure 5.1: Significance of features

weak learner can well capture the interaction between features. Parameter settings for Gradient Boosting Trees please find Table 5.9.

The Gradient Boosting trees model also tells us the importance of features. Figure 5.1 shows that "average purchase price" and "purchase frequency per week" are two most influential factors while "average click times prior to first purchase" and "average duration" affect least.

5.4 Further analysis: label ambiguity

From experiment result statistic, we found that there were quite a few "disagreeing" labeling results, approximately 47%, ambiguous to tell what type the objects belong to. In sense of "disagreement", for example, of the 5 labels for one object (ques-

Type	**Number**	**Percentage**
Positive	10	41.67%
Negative	14	58.33%
Total	24	100%

Table 5.10: Distribution of "Positive" and "Negative" objects

tion), 3 workers labeled "Hedonic" while the other 2 labeled "Utilitarian". Via observing records of ambiguity questions, we induced that consumer type kept on varying from session to session, as people are more likely to act Hedonic on a Saturday morning than during weekdays. In order to verify the hypothesis, we ran one additional experiment in a small scale.

We selected 24 ambiguity customers which have at least two sessions and got H/U label for each session individually, instead of by each consumer. To ensure it comparable with previous experiments, we adopted same rules: incentive strategy, 5 different label, using mostly voted as result, etc.

We got back labeling results of 24 objects. We define the object "positive" if all sessions of the object belonged to a same label and "negative" if at least one session was labeled differently from other sessions. Table 5.10 shows that around 58.33% ambiguity objects results from sessions do NOT "agree" with each other. This indicates that earlier hypothesis, consumer acting differently from time to time, is verified.

We assumed that consumers classification do not depend on the time their shopping conducted for now. We understand that the assumption are strong and would affect model accuracy considerably. However, due to time limitation, we would like to

leave it as future work.

Chapter 6

Future work

We acknowledged that quite a few "disagree" phenomenon existed among experiment results and they were responsible for lowering upper bound of models accuracy. We then plan further exploration via partitioning consumer shopping history into more pieces, not limited to partition by session. We expect an improvement of model accuracy resulting from more sensible experiment labeling results.

It is also needed to enrich influential features for generating more precise prediction models. In addition to shopping history, consumers' role on Hedonic or Utilitarian depends largely on many other characteristics as well. One most important dimension we would like to propose for future improvement is to involve personal profiles of online consumers, such as: gender, age, education, work, interest etc. For example, men usually describe shopping as "women's job" as female are assumed to perceive more joy than male would. A teenager student never gets tired of browsing clothing, skin care products even for hours but might change her preference to express shopping when the

elder she becomes an extremely busy business woman. A rock climbing enthusiast is most Hedonic at Sports Authority and tends to be less patient at Office Depot. We believe that exploring additional features on personal profiles would be of great use towarding the end of improving accuracy of classification models.

Another improvement we are considering is to optimize experiment mechanism. As the experiment requires to choose between two categories, it is possible for workers to guess that ratio of Hedonic over Utilitarian should be around 1:1. Therefore, workers might tend to label same numbers of H and U and this tendency can have negative effect on enhancing reliable experiment results.

Chapter 7

Conclusion

In the paper, we studied an economic concept regarding categorizing online consumers as Hedonic or Utilitarian and discussed the benefits of implementing such concepts in real world e-commerce application. Our attempt to apply crowdsourcing techniques to construct training data set turned out to be a success as the proposed incentive strategy encouraged workers to label honestly. We closely examined a variety of machine learning algorithms on the dataset generated from raw web log and crowdsourcing. The experiment results suggested that online consumers are classifiable based on their shopping history. Meanwhile, we are intrigued by other interesting findings. For example, higher performances came from experiments with lower BONUS. Many features such as "category diversity", "product purchase price" are most influential on determining the type of customer.

The present paper contributes in following aspects. First of all, our work is the first attempt to apply data science techniques to solve the problem of online shop-

ping type categorization. Secondly, the findings from our investigation are of practical value to industry. Admittedly, our classification models are not quite there for direct commercialization. That is also why we would like to improve the model performance by incorporating more consumer behavioral information.

Bibliography

[1] gbm. `http://cran.r-project.org/web/packages/gbm/index.html`. Accessed: 2014-11-28.

[2] weka. `http://www.cs.waikato.ac.nz/ml/weka/`. Accessed: 2014-12-01.

[3] Mark J Arnold and Kristy E Reynolds. Hedonic shopping motivations. *Journal of retailing*, 79(2):77–95, 2003.

[4] Barry J Babin, William R Darden, and Mitch Griffin. Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of consumer research*, pages 644–656, 1994.

[5] Peter H Bloch and Marsha L Richins. A theoretical model for the study of product importance perceptions. *The Journal of Marketing*, pages 69–81, 1983.

[6] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[7] Bradley Efron and B Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

[8] David Freedman. *Statistical models: theory and practice*. Cambridge University Press, 2009.

[9] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[10] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[11] Camille Grange and Izak Benbasat. Online social shopping: The functions and symbols of design artifacts. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[12] Zhao Huang and Morad Benyoucef. From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications*, 12(4):246–259, 2013.

[13] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[14] Kent B Monroe and Joseph D Chapman. Framing effects on buyers? subjective product evaluations. *Advances in consumer research*, 14(1):193–197, 1987.

[15] C Reid Turner, Alfonso Fuggetta, Luigi Lavazza, and Alexander L Wolf. A conceptual basis for feature engineering. *Journal of Systems and Software*, 49(1):3–15, 1999.

[16] John F Sherry Jr. A sociocultural analysis of a midwestern american flea market. *Journal of Consumer Research*, pages 13–30, 1990.

[17] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A survey of crowdsourcing systems. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 766–773. IEEE, 2011.

Appendix A

Engineering & Implementation

Figure A.1 is where people we hired to work on labeling.

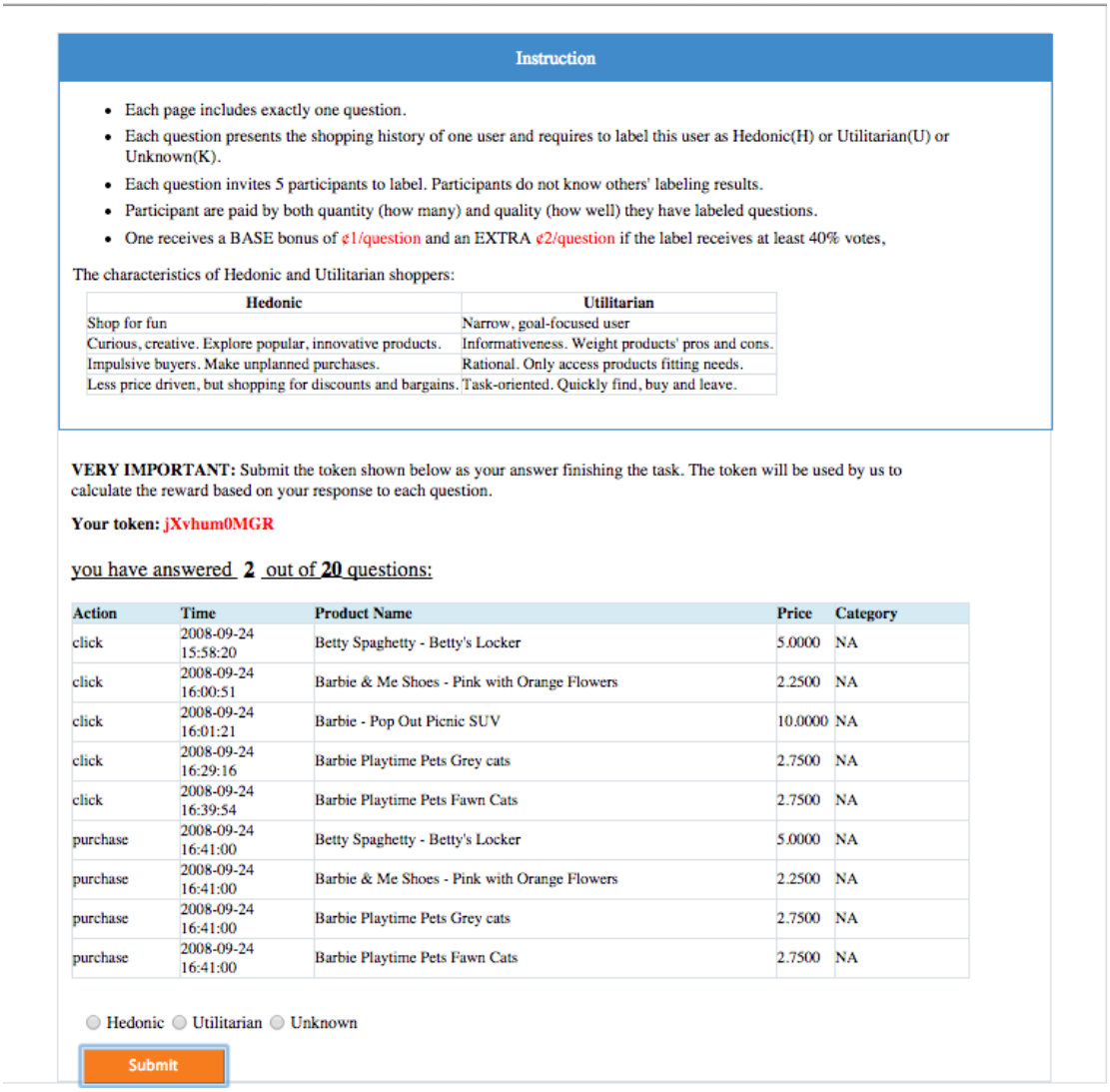

Figure A.1: Website for labeling