Title
Application of Machine Learning Algorithms in Predicting Social-Planning Platform Donations

Author
Kim, Daehyun

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Application of Machine Learning Algorithms in Predicting Social-Planning Platform Donations

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Applied Statistics

by

Daehyun Kim

2020

ABSTRACT OF THE THESIS

Application of Machine Learning Algorithms in Predicting Social-Planning Platform Donations

by

Daehyun Kim
Master of Science in Applied Statistics
University of California, Los Angeles, 2020
Professor Yingnian Wu, Chair

Pledgeling is a platform that powers corporate giving and social impact programs for businesses of all sizes by integrating donation features in the backend. One of Pledgeling's partners is Evite, the world's leading digital platform for bringing people together with event invitations. In this study, the data from Pledgeling is used to train a logistic regression model to determine which aspects of RSVP events lead to hosts turning on the donation features. From the first study, it was statistically significant that categories of events that fall into Organizations, Weddings, and Animals were more likely to add donation features. Also events from the West and Northeast regions during Q1 and Q4 during the calender year were more likely to add donation features. Additionally, data is used to train a linear regression model to study which features lead to users donating more for each event. From the second study, there were statistically significant results that categories of events that fall into Organizations, Weddings, and Get Togethers were donating more money. Also, results indicated that events in West and Northeast regions were donating more from the beginning of summer to the end of the year. Lastly, events donated more to Science-related as well as Public Health-related causes compared to causes for Animals organizations.

The thesis of Daehyun Kim is approved.

Vivian Lew

Qing Zhou

Yingnian Wu, Committee Chair

University of California, Los Angeles

2020

This thesis is dedicated to my parents ...

For their endless love, support and encouragement

Table of Contents

List of Figures . . . . . **vii**

List of Tables . . . . . **ix**

1 Introduction **1**

1.1 Background 1

1.2 Data Description 2

1.3 Outline 3

2 Methodology **4**

2.1 Logistic Regression 4

2.2 Synthetic Minority Over-Sampling Technique (SMOTE) 6

2.3 Linear Regression 7

3 Donation Feature Usage Analysis **10**

3.1 Data Preprocessing 10

3.2 Exploratory Data Analysis 12

3.3 Model analysis and Performance Evaluation 16

4 Donation Amount Analysis **21**

4.1 Data Preprocessing 21

4.2 Exploratory Data Analysis 22

4.3 Model analysis and Performance Evaluation 27

5 Future Enhancement and Conclusion **32**

5.1 Future Enhancement 32
5.2 Conclusion 32

References **34**

List of Figures

2.1 Logistic Function 5

2.2 Before and After SMOTE applied 7

2.3 Cost Function 8

2.4 Gradient Descent Algorithm with Different Learning Rates 9

3.1 Events Table Mapping Diagram 12

3.2 Distribution of Events With/Without Donation Features 12

3.3 Frequency of Grouped Categories 13

3.4 Percentage to Add Donations for Grouped Categories 13

3.5 Frequency of Grouped Regions 14

3.6 Percentage to Add Donations for Grouped Regions 14

3.7 Average Wage for Grouped Regions 15

3.8 Frequency of Quarters 15

3.9 Percentage to Add Donations for Quarters 16

3.10 Initial Logitic Regression Summary 17

3.11 Final Logitic Regression Summary 18

3.12 Donation Feature Confusion Matrix 19

3.13 ROC Curve 20

4.1 Non-Profit Cause Mapping 22

4.2 Histogram for Donation Amount 22

4.3 Donation Amount vs. Donation Goal from an event 23

4.4 Donation for Grouped Categories over time 24

4.5 Donation for Grouped Causes over time 24
4.6 Average Donation for Grouped Categories 25
4.7 Average Donation for Grouped Regions 25
4.8 Average Donation for Grouped Causes 26
4.9 Average Donation for Months 26
4.10 Initial Linear Regression Summary 28
4.11 Final Linear Regression Summary 29
4.12 Errors of Prediction 30

List of Tables

1.1 Events Data 2

1.2 Donations Data 3

1.3 Organizations Data 3

3.1 Number of NA Values in Events Data 11

CHAPTER 1

Introduction

1.1 Background

As many physical activities have shifted to online for more convenience in the recent years, event invitation industry has not been an exception. Most people nowadays forgo RSVP cards for events like wedding and have their guests do it all online. This trend has seen a surge in online RSVPs and upon popular demand, online RSVP platforms like Evite have emerged as popular business for the past decade. Evite's CEO, Victor Cho, at the Wharton Customer Analytics Initiative conference back in 2018, has mentioned how a company should not just focus on serving three stakeholders — customers, shareholders, and employees [6]. He placed an emphasis on the fourth stakeholder: society. Since 2015, Evite has partnered with Pledgeling to let users donate to a list of charities on the site. Evite Donations, enabled by Pledgeling's API, has benefited thousands of nonprofits and personal causes — from children's health to helping survivors of the 2018 California wildfires. With Evite Donations, event hosts have the option to select their favorite nonprofit from 1 million+ organizations in the Pledgeling network, recommend a new nonprofit, or crowdfund for a personal cause. In 2019, Evite and Pledgeling together raised over $10 million for non profit organizations and personal causes across the United States. There are multiple business questions raised from Pledgeling for boosting event's donations even further. In this study, events and donations are analyzed with regression models to present some insights into these questions.

1.2 Data Description

The dataset used in this study is from Pledgeling and Evite. There are three tables that contain relevant information: Events, Donations, and Organizations (Non-Profits). Events include 10115798 rows of entire events on Evite between 2015 and 2020 regardless of the donation feature. Donations include 389116 rows of donations between 2015 and 2020. Organizations include 1721326 rows of non-profit organizations in the Pledgeling networks to which an event host can choose to donate. Additionally, public data was pulled from online for average wages for states in the U.S. and states data for each zip code for the mapping purpose. Details on the data descriptions are shown on tables below.

Events	
Feature	Description
Beneficiary Type	Type of donation events: Organization for non-profit donations; Crowdfund for self-fundraising; N/A for events without donation
Category	Category of events
Event Date	Date of events
Zip Code	Zip code of events location
Total Raised	Total amount raised from an event
Total Donors	Total number of donors for an event
Goal	Donation goal set from an event host

Table 1.1: Events Data

Donations	
Feature	Description
Donation Amount	Donation amount donated from an order
Tip Amount	Tip amount donated from an order
Date	Date of donations
Postal Code	Zip code of donations location
Region Code	State/Region of donations location
Country Code	Country of donations location

Table 1.2: Donations Data

Organizations (Non-Profit)	
Feature	Description
Name	Name of non-profit organizations
State	State of non-profit organizations location
Cause	Cause (Type) of donation

Table 1.3: Organizations Data

1.3 Outline

In this thesis, there are two questions to be answered. One is to find out what types of an event are more likely to add donation features. Another question to be answered is what types of a donation order lead to more donation amounts. For both of these questions, initial data processing is applied to check and handle missing values, join the tables accordingly, and filter out unnecessary columns. Exploratory data analysis is then performed to observe and visualize correlations and patterns amongst features. Additionally, logistic and linear regression models are trained to predict for the corresponding dependent variables. Each model will be evaluated via performance measures and meeting assumptions. Conclusions and further improvements are discussed in the final step.

CHAPTER 2

Methodology

2.1 Logistic Regression

Supervised machine learning algorithms are divided into two major groups: classification and regression. Classification predicts a dependent variable with class labels, while regression predicts a continuous output variable. Logistic regression is the most simple classification algorithm that predicts a dichotomous (binary) output variable. In this study, the output is whether or not an event host will add a donation feature. There are multiple assumptions for logistic regression [7].

- Dependent variable is binary. (0 or 1, True or False)
- Predictor variables should be independent of each other. No multicollinearity.
- Independent variables are linearly related to the log odds.
- Logistic regression requires relatively large sample sizes.

Logistic regression is named for the logistic function, a core function in the methodology. The logistic function, also known as sigmoid function, was developed to describe population growth, that rises quickly and saturates at the carrying capacity [2]. It's an s-shaped curve that takes any number and map it to a value between 0 and 1. The logistic function is represented by the following equation:

$$1/(1 + e^{-z})$$

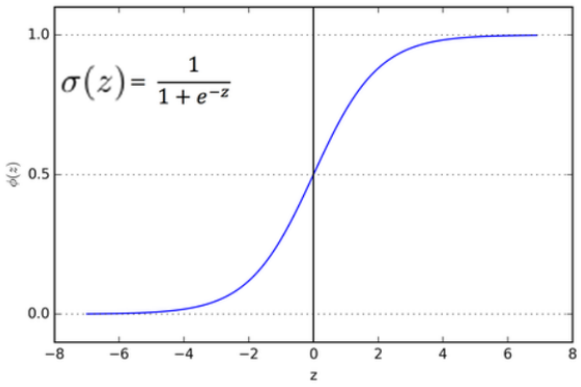


Figure 2.1: Logistic Function

where e is the base of the natural logarithms and z is the numeric value to be transformed. Figure 2.1 is a plot of numbers transformed to those between 0 and 1 using the sigmoid function.

Logistic regression forms an equation with input values x combined linearly with coefficients predicting an output variable y. An example is as follows:

$$y = p(x) = e^{b_0+b_1 x}/(1 + e^{b_0+b_1 * x})$$

where y is the predicted output, b_0 is an intercept and b_1 is the coefficient for a feature input x. Each feature in the input data has coefficient b trained from the model. Logistic regression models the probability of the class output. To give logistic regression more interpretability, an equation should be transformed using natural logarithm. Above example equation can be transformed into:

$$ln(p(X)/1–p(X)) = ln(odds) = b_0 + b_1 X.$$

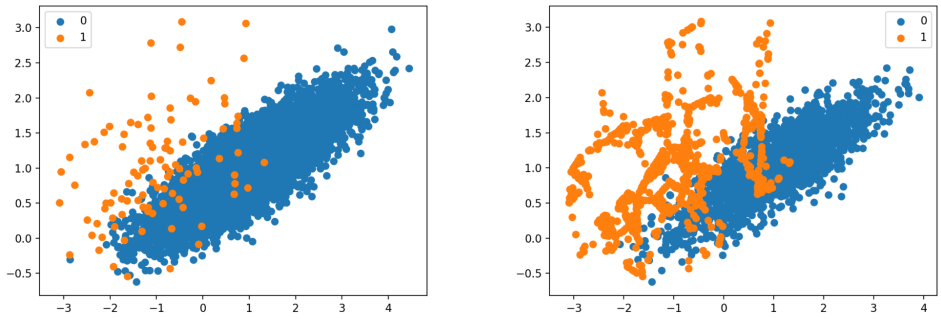
The new transformed equation is now linear and the term on the left is called the odds of the default class. Odds are calculated as a ratio of the probability of an event occurring over the probability of an event not occurring. The final form of an equation after simplifying is:

$$odds = e^{b_0+b_1 X}.$$

Now, we can interpret a coefficient by stating as X increases by 1, odds increases by the coefficient. To check the performance of logistic regression, there are measures of confusion matrix and AUC – ROC Curve.

2.2 Synthetic Minority Over-Sampling Technique (SMOTE)

When classification algorithm is applied to imbalanced data with very small size of event compared to non-event sample (further to be discussed in data processing), data-preprocessing technique such as the Synthetic Minority Over-Sampling Technique (SMOTE) is used to create synthetic minority class samples [3]. This technique was described by Nitesh Chawla, et al. in their 2002 paper. SMOTE selects examples that are clustered in the feature space, "drawing a line between the examples" in the feature space and a new sample at a point along the line. First, choose a random point from the minority class. Then amongst k (typically 5) nearest neighbors of the point, another random point is chosen, and a new synthetic example is created at a random location between the two examples. This step is continuously applied to create many synthetic examples for the minority class to balance the class distribution. This technique has an advantage of creating plausible examples for the minority class by choosing random samples relatively close in the feature space. The scatter plots below represent the minority class samples before and after the SMOTE technique is applied.



(a) Before SMOTE

(b) After SMOTE

Figure 2.2: Before and After SMOTE applied

2.3 Linear Regression

Linear regression is used for finding linear relationship between output variable and predictor variables. Multiple linear regression is linear regression between two or more predictor variables and output variable. In this thesis, multiple linear regression is trained to predict a donation amount from an order for events that added donation features. There are several assumptions for multiple linear regression:

- Normality or residuals: residuals from the regression model follow a normal distribution.
- Independence of variables: predictor variables should be independent of each other. No multicollinearity.
- Linearity: Predictor variables are linearly related to output variable.
- Homogeneity of variance (homoscedasticity): residuals have constant variance across the values of the independent variable.

The equation for a multiple linear regression is:

$$y = b_0 + b_1 x_1 + ... + b_n x_n + \varepsilon$$

where y is dependent variable, b_0 is the intercept, b_i is the regression coefficient and x_i is each feature/predictor variable, ε is model residuals. The linear regression calculates three factors: the regression coefficients, the t-statistic of the overall model, and the associated p-value (how likely it is that the t-statistic would have occurred by chance if the null hypothesis of no relationship between target and predictor variables was true). While training the model, it calculates the cost function which measures the Root Mean Squared error between predicted value and true value. The model aims to minimize the cost function as follows:

$$minimize(\frac{1}{n})\sum_{i=1}^{n}(pred_i - y_i)^2.$$

Gradient descent algorithm is used to minimize the cost function by initially selecting estimates and iteratively update these values until the cost function reaches the minimum [5]. The equation and mathematics for gradient descent algorithm is illustrated in the Figure 3.1.

Cost Function

$$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y_i]^2$$

Predicted Value; True Value

Gradient Descent

$$\Theta_j = \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta_0, \Theta_1)$$

Learning Rate

Now,

$$\frac{\partial}{\partial\Theta}J_\Theta = \frac{\partial}{\partial\Theta}\frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y]^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x_i) - y)\frac{\partial}{\partial\Theta_j}(\Theta x_i - y)$$

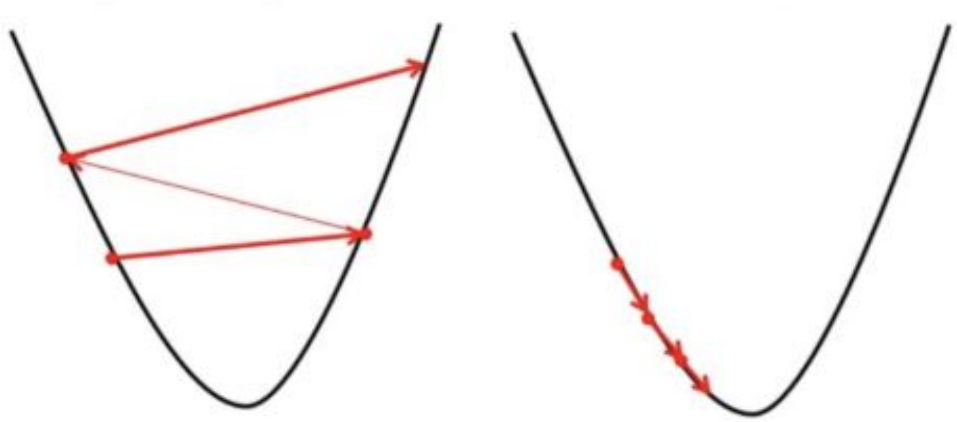
$$= \frac{1}{m}(h_\Theta(x_i) - y)x_i$$

Therefore,

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

Figure 2.3: Cost Function

If this cost function is graphed as a function of parameter estimates, the gradient descent algorithm moves downward towards the minimum point in the curve with iterations. It moves toward the minimum with each step determined by learning rate α.



(a) Big Learning Rate (b) Small Learning Rate

Figure 2.4: Gradient Descent Algorithm with Different Learning Rates

The model evaluation metrics for linear regression are as below:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - x_i| .$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$RMSE = \sqrt{(\frac{1}{n}) \sum_{i=1}^{n} (y_i - x_i)^2}.$$

The lower the residual errors, the better the model fits the data.

R-squared Statistic is the proportion of variability in the target that can be explained using a feature:

$$R^2 = 1 - (\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}).$$

A good model has the R-squared value close to 1.

CHAPTER 3

Donation Feature Usage Analysis

3.1 Data Preprocessing

Before jumping into building models, data should be examined and cleaned so that model can take in proper data. In the reality, data almost always has missing values unlike some cleaned version in Kaggle competitions. There can be multiple reasons for these missing values: users not filling in some information, or sometimes missing values representing some information. Depending on the reasons and circumstances behind the missing values, appropriate means of processing should be conducted to handle them. The first approach is to exclude the null values from the dataset. This approach can be simple and concise when there are enough samples in the data, but can be a problem when it leads to losing important information with small sample size. Another approach is to impute the missing values using feature means or medians. The last approach is a rather sophisticated one by applying k-nearest-neighbors to the missing values and impute a value that is close to the values in the similar feature space.

In the scope of donation feature usage analysis, we are only interested in whether an event leads to adding a donation feature or not, so we do not need donations or organizations tables, but only events table. In the events table, we deem 'Event Date' and 'Zip code' as essential information for our modeling, so handling these missing values are key to successful modeling (number of missing values shown in below table). Since there is a large set of samples, we can safely exclude these missing values from the data.

Additionally, 'Beneficiary Type' has a lot of null values because the ones without this feature are the events without donation features. Target variable 'Donation' is generated by mapping data with missing Beneficiary Type to 0 (no donation feature) and the rest to 1

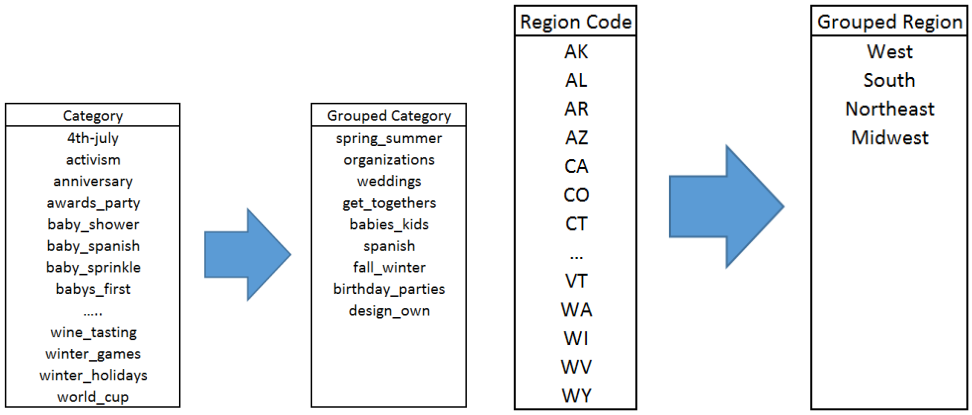
NA Values	
Feature	Number of NA
Beneficiary Type	9948566
Category	477015
Event Date	32220
Zip Code	2497197
Total Raised	0
Total Donors	0
Goal	9944542

Table 3.1: Number of NA Values in Events Data

(added donation feature).

To focus more on the recent events on the domain, events that were created before 2017 were excluded. Also, since this table only provides zip code for each event instead of state or country, publicly available zipcode data was pulled to map each zip code to state and country, and country was filtered to be only US.

The only available features for this table were categorical variables such as state, category, and event date so grouping categories and transforming into dummy variables were needed to be able to input into models. 215 granular categories were mapped into 9 bigger categories and 51 states were grouped into 4 big regions in the US. Also each event dates were mapped into four different quarters from Q1 to Q4. Following tables demonstrate the mapping.



(a) Events Category Mapping (b) Events State Mapping

Figure 3.1: Events Table Mapping Diagram

3.2 Exploratory Data Analysis

Pre-modeling data analysis is essential in modeling projects because it generates insights in the data and expectations which features models should focus on. For a binary classification, the first thing to observe is whether data has a balanced target variable distribution. From the pie chart below, events that added donations were only 1.73% of total events from 2017 to 2020. This creates an imbalanced target variable problem in binary classification.

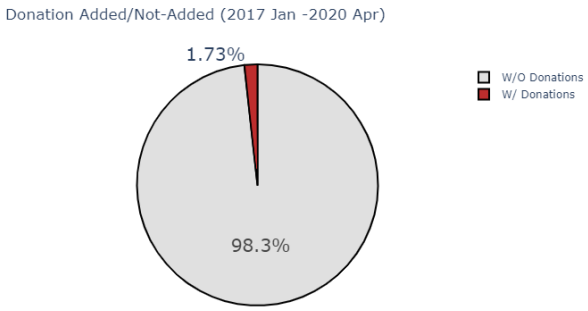


Figure 3.2: Distribution of Events With/Without Donation Features

To resolve this problem, SMOTE algorithm as discussed in Chapter 2 was applied to create 50% of events with no donation features, and other 50% of events with donation

features. Since there are multiple categorical variables in this dataset, exploratory analysis was performed to see frequency and percent of adding donation features distributions for these groupings.

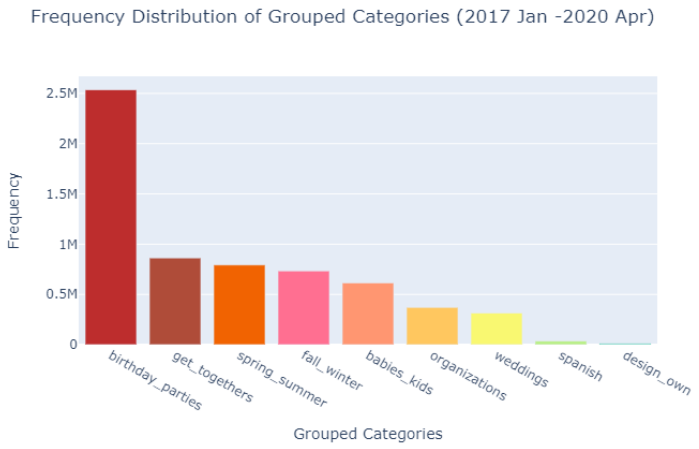


Figure 3.3: Frequency of Grouped Categories

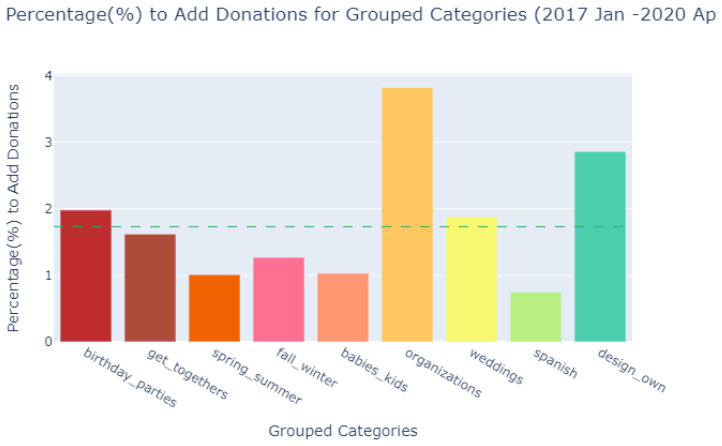


Figure 3.4: Percentage to Add Donations for Grouped Categories

For grouped categories, birthday parties were noticeably dominant in terms of number of events created. When percentage to add donation features was calculated for each category, there was an interesting trend that organizations and design own had much higher ratio of adding donations compared to other categories. Green dashed line indicates the percentage

of total events that added donation features, which was 1.73%. Organizations are mostly businesses and club meetings including fundraising events, which can explain the high ratio.

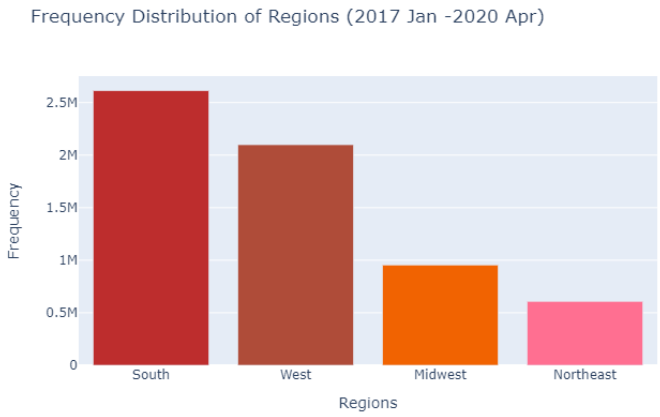


Figure 3.5: Frequency of Grouped Regions

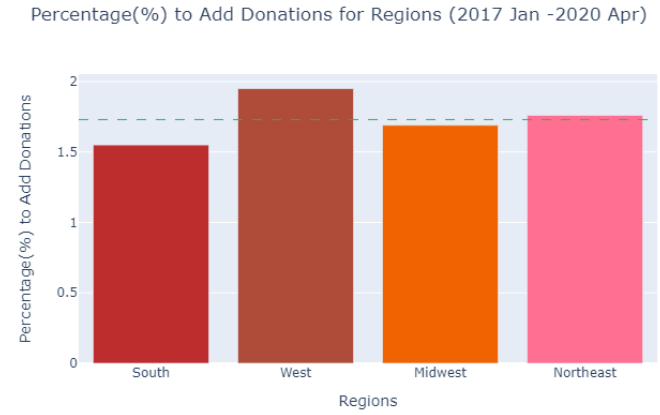


Figure 3.6: Percentage to Add Donations for Grouped Regions

South and West Region make up most of locations of total events in the frequency graphs, but West and Northeast demonstrate higher percentage of adding donation features. This can be explained by West and Northeast include states that have higher average wages such as California, Washington, New York, and Massachusetts. The hypothesis is supported from the average wage for those regions in 2018. People who are relatively well-off will more likely to consider donation features so that they can contribute to social causes.

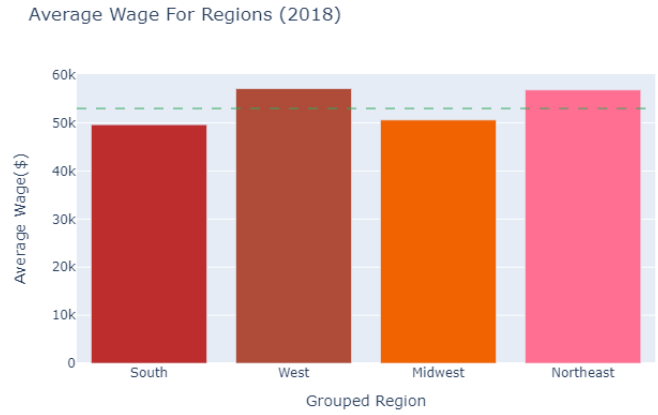


Figure 3.7: Average Wage for Grouped Regions

For quarterly divided groups, Quarter 1 showed the highest ratio of adding donations but the differences are not big enough to discuss whether the quarters return any meaningful results.

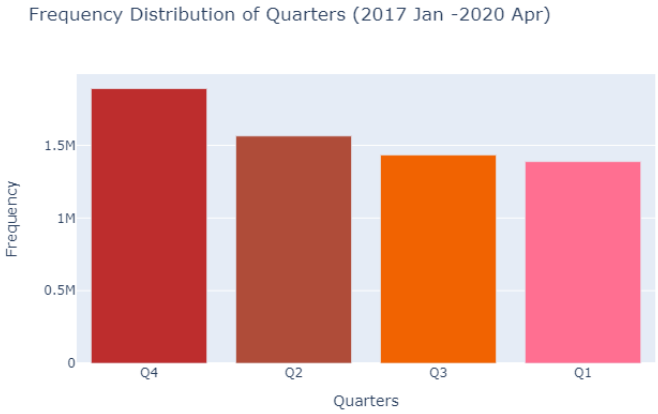


Figure 3.8: Frequency of Quarters

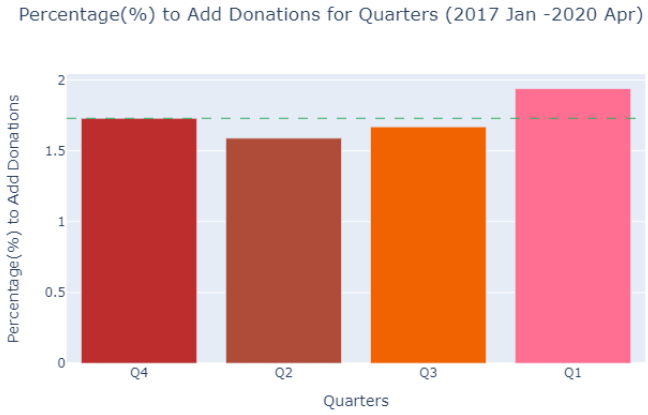


Figure 3.9: Percentage to Add Donations for Quarters

3.3 Model analysis and Performance Evaluation

In order to validate the training model results and prevent overfitting, diving dataset into training-test groups is essential. For this study, 75% of the dataset was randomly selected as training data, while 25% of the dataset was left out to be used as test validation data. When the logistic regression fits the model, not all the features that go into the model will have coefficients statistically significant. The process to eliminate features that are not statistically significant (that has a p-value higher than 0.05) is necessary to have a robust model. Initial logistic regression model from all the features in the data is below:

Model:	Logit	Pseudo R-squared:	0.025
Dependent Variable:	Donation	AIC:	12501176.4456
Date:	2020-05-22 15:26	BIC:	12501415.1322
No. Observations:	9252360	Log-Likelihood:	-6.2506e+06
Df Model:	16	LL-Null:	-6.4132e+06
Df Residuals:	9252343	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Quarter_Q1	0.1345	0.0023	58.4338	0.0000	0.1300	0.1390
Quarter_Q2	0.0034	0.0012	2.8171	0.0048	0.0010	0.0057
Quarter_Q3	0.0146	0.0023	6.3511	0.0000	0.0101	0.0191
Quarter_Q4	0.0932	0.0023	40.3245	0.0000	0.0887	0.0977
Grouped Category_babies_kids	-0.5544	0.0038	-146.8750	0.0000	-0.5618	-0.5470
Grouped Category_birthday_parties	0.1015	0.0031	33.1558	0.0000	0.0955	0.1075
Grouped Category_design_own	0.4851	0.0136	35.6693	0.0000	0.4585	0.5118
Grouped Category_fall_winter	-0.3768	0.0038	-98.2640	0.0000	-0.3843	-0.3693
Grouped Category_get_togethers	-0.1010	0.0034	-29.4818	0.0000	-0.1078	-0.0943
Grouped Category_organizations	0.7884	0.0037	212.2976	0.0000	0.7811	0.7957
Grouped Category_spanish	0.0069	0.0010	6.6220	0.0000	0.0048	0.0089
Grouped Category_spring_summer	-0.5537	0.0035	-156.7639	0.0000	-0.5606	-0.5467
Grouped Category_weddings	0.0594	0.0041	14.5204	0.0000	0.0514	0.0674
Grouped Region_Midwest	-0.0629	0.0030	-20.7844	0.0000	-0.0689	-0.0570
Grouped Region_Northeast	0.0018	0.0014	1.2636	0.2064	-0.0010	0.0047
Grouped Region_South	-0.1390	0.0027	-51.1936	0.0000	-0.1444	-0.1337
Grouped Region_West	0.0954	0.0027	34.8376	0.0000	0.0900	0.1008

Figure 3.10: Initial Logitic Regression Summary

P-value for Grouped Region Northeast is 0.2064 (greater than 0.05) so it's necessary to eliminate this feature from the model because this coefficient for Northeast is not statistically significant to reject the null hypothesis that the coefficient is zero. In other words, the high p-value indicates changes in the predictor are not associated with changes in the target variable. After eliminate the feature, the final summary of the model is below:

Results: Logit

Model:	Logit	Pseudo R-squared:	0.025
Dependent Variable:	Donation	AIC:	12501176.0765
Date:	2020-05-22 15:27	BIC:	12501400.7228
No. Observations:	9252360	Log-Likelihood:	-6.2506e+06
Df Model:	15	LL-Null:	-6.4132e+06
Df Residuals:	9252344	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Quarter_Q1	0.1346	0.0023	58.4746	0.0000	0.1301	0.1391
Quarter_Q2	0.0034	0.0012	2.8420	0.0045	0.0011	0.0057
Quarter_Q3	0.0146	0.0023	6.3833	0.0000	0.0101	0.0191
Quarter_Q4	0.0933	0.0023	40.3704	0.0000	0.0888	0.0978
Grouped Category_babies_kids	-0.5528	0.0035	-156.5459	0.0000	-0.5597	-0.5458
Grouped Category_birthday_parties	0.1032	0.0028	37.4559	0.0000	0.0978	0.1086
Grouped Category_design_own	0.4868	0.0135	35.9677	0.0000	0.4603	0.5134
Grouped Category_fall_winter	-0.3752	0.0036	-104.2081	0.0000	-0.3822	-0.3681
Grouped Category_get_togethers	-0.0994	0.0032	-31.4695	0.0000	-0.1055	-0.0932
Grouped Category_organizations	0.7901	0.0035	227.9730	0.0000	0.7833	0.7969
Grouped Category_spanish	0.0069	0.0010	6.6475	0.0000	0.0049	0.0089
Grouped Category_spring_summer	-0.5520	0.0033	-168.9498	0.0000	-0.5584	-0.5456
Grouped Category_weddings	0.0611	0.0039	15.7958	0.0000	0.0535	0.0687
Grouped Region_Midwest	-0.0647	0.0027	-23.9619	0.0000	-0.0700	-0.0594
Grouped Region_South	-0.1408	0.0023	-60.1587	0.0000	-0.1454	-0.1362
Grouped Region_West	0.0937	0.0024	39.5751	0.0000	0.0890	0.0983

Figure 3.11: Final Logitic Regression Summary

To interpret the coefficients in the logistic regression model, odds are used as discussed in the Chapter 2. The coefficients correspond to the log odds of the probability of adding donation feature to an event. After exponentiating the coefficient, odds are easily interpretable. For example, an event categorized in organizations has $e^{0.7901} = 2.204$ odds of adding donation feature. In other words, an "organization" event is 2.2 times likely to add donation feature. On the other hand, if an event is about babies and kids, it is $e^{-0.5528} = 0.575$, 43% less likely to add donation feature.

The model has statistically significant features, but it is also essential to check the performance of the model in terms of predictability. This study will use confusion matrix and ROC curve for the performance measures. From the below confusion matrix, there are 701497 true positive values and 628903 true negative values. This means the accuracy for the test dataset is 0.58. The model accurately predicted the outcome of 58% of the test features whether or not an event would add a donation feature.

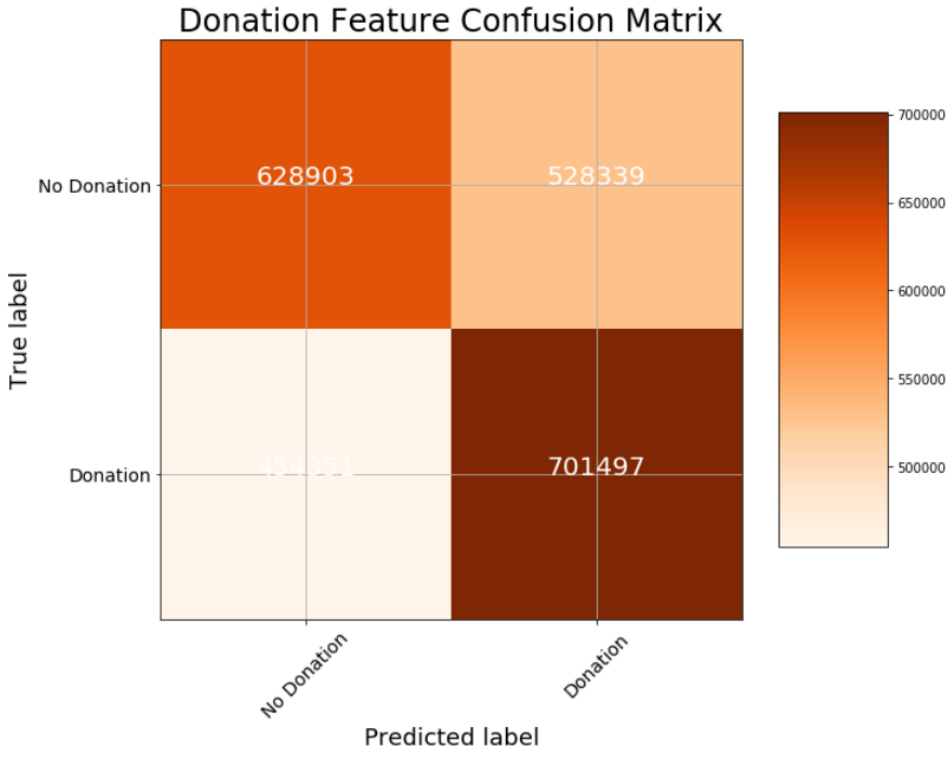


Figure 3.12: Donation Feature Confusion Matrix

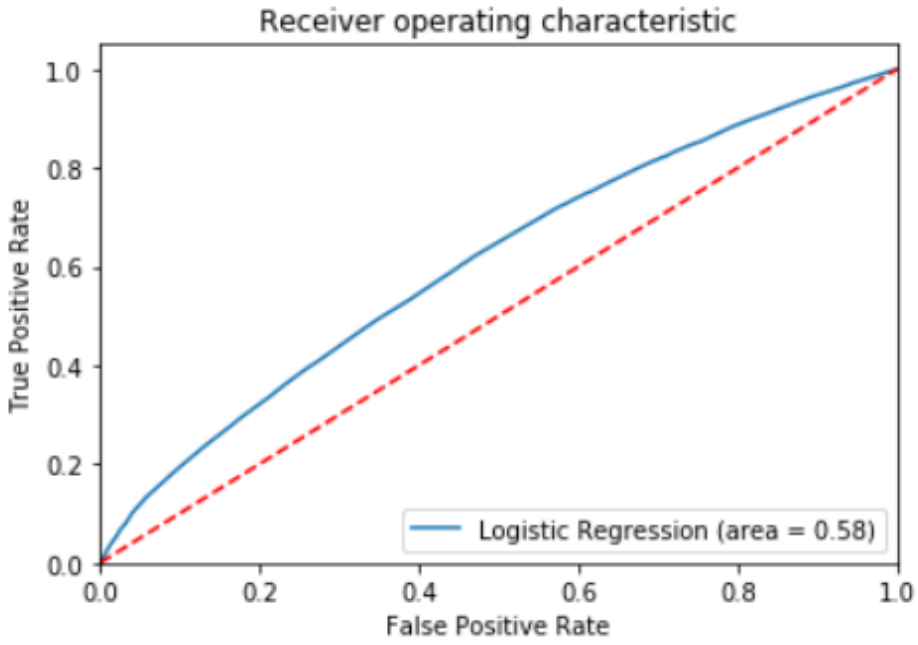


Figure 3.13: ROC Curve

A robust model has the ROC plot that scews to the top left because smaller x-value indicates lower false positive values, and larger y-value indicates higher true positive values.

The model is not strongly accurate with 58% accruacy, which makes sense that the ROC curve is not curved to the upper-left corner. The reasons behind the lower accuracy will be discussed in Chapter 5 Future Enhancement.

CHAPTER 4

Donation Amount Analysis

4.1 Data Preprocessing

Another question raised from the Pledgeling stakeholders was what leads to more donations from event participants. To answer this question, joining available data tables — Events, Donations, and Organizations — was needed to take all aspects into account.

With the joined table, I followed similar methodology of handling missing values in 'Date' and 'Region Code' (State) and 'Cause' (Non-Profit Cause) by excluding data with missing values in these essential columns since there is a large set of samples already. To focus more on the recent events on the domain, donations that were created before 2017 July were excluded. Additionally, country was filtered to be only US and beneficary type 'Crowdfund' was excluded because this type is for self-fundraising events with no social cause.

Most features for the data were categorical variables at the highly granular level such as state, category, and cause so grouping categories and transforming into dummy variables had to be done to be able to input into models. 215 granular categories were mapped into 9 bigger categories and 51 states were grouped into 4 big regions in the US. 39 different granular causes were mapped into 5 bigger causes as well. Also each event dates were mapped into each month and four different quarters from Q1 to Q4. Following table demonstrate the mapping for cause, and the other mappings are the same as those done in Chapter 3.

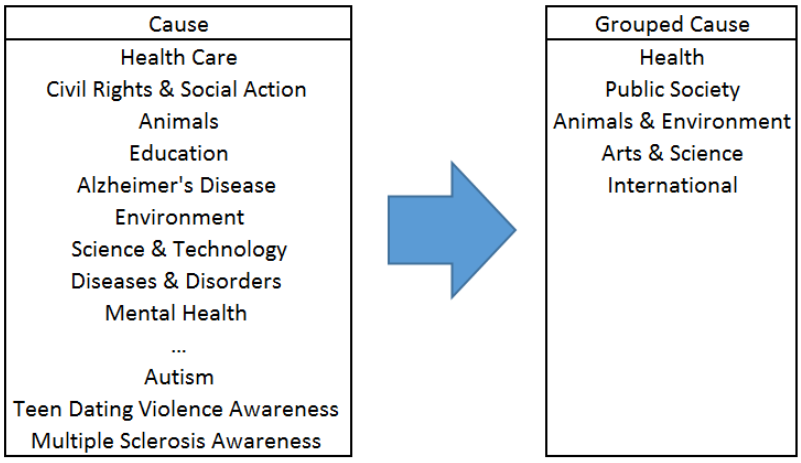


Figure 4.1: Non-Profit Cause Mapping

4.2 Exploratory Data Analysis

For continuous target variable in this case being the donation amount from each order, the first visual to check is the distribution of the variable. As expected, there were a few donations with very high numbers such as $25,000 and $10,000, which can be deemed as outliers, while most donations fell in the range of $10 to $50. Most people would not donate more than $100 for public causes.

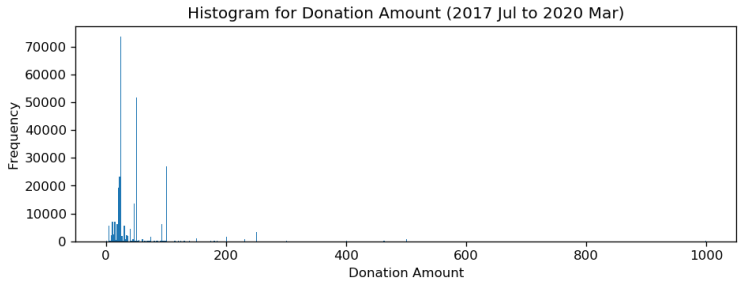


Figure 4.2: Histogram for Donation Amount

Donation goal was only available continuous predictor that an event host set, and it was expected to see a linear relationship between Donation Amount and Goal. However, from the below figure, I could not find a clear linear pattern between Goal and Donation

Amount. This could be because a lot of event hosts do not set the goal amount, or they set an unreasonably high amount on the goal.

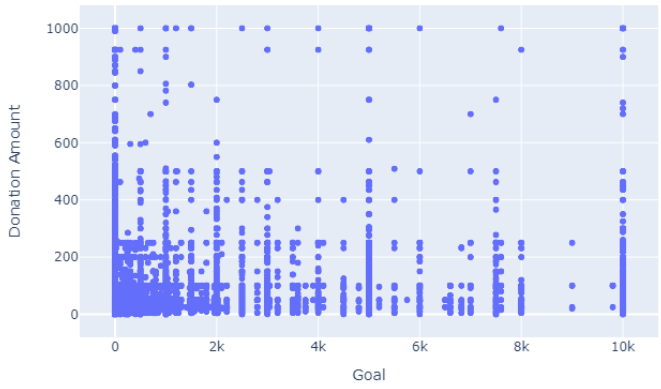


Figure 4.3: Donation Amount vs. Donation Goal from an event

Next, we looked at total donations over time for different categories of events as well as different social causes the donations of events went to. There were interesting seasonality patterns for the categories. Weddings show peaks during late spring to summer, while birthday parties showed clear peaks during September and October, while declining in November and December. These lower donation amounts during holiday times are related to lower frequency of events during these times, given that people do not want to throw birthday parties during holidays like Thanksgiving and Christmas breaks. Fall winter events showed a huge peak in December of 2019.

As for social causes, public society and health causes indicated significant increase in donation amounts in 2019 especially in Q4, while other causes relatively stated at the constant level. These are correlated to the noticeable peaks for fall winter and organizations events during the same weeks. Events during the holiday times seem to lead to more donations for Public Society and Public Health related organizations.

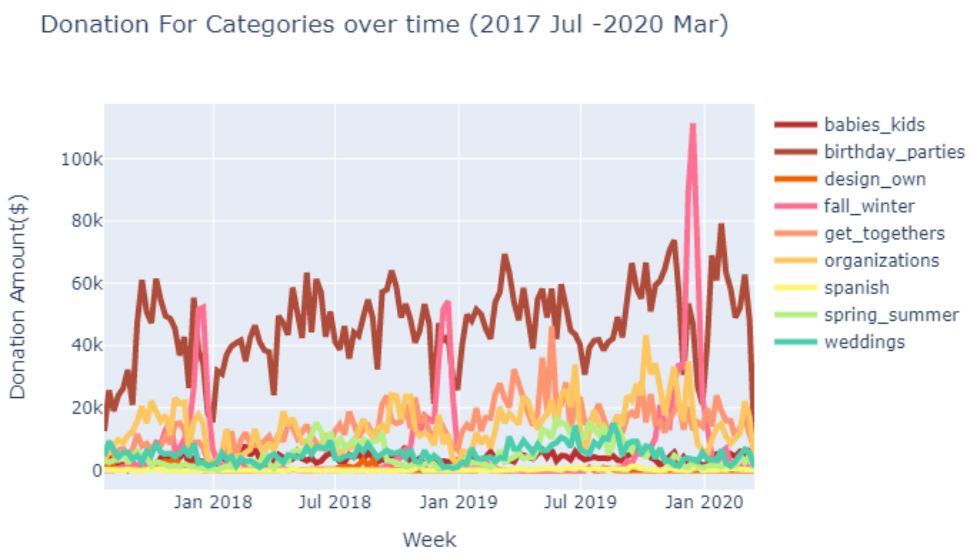


Figure 4.4: Donation for Grouped Categories over time

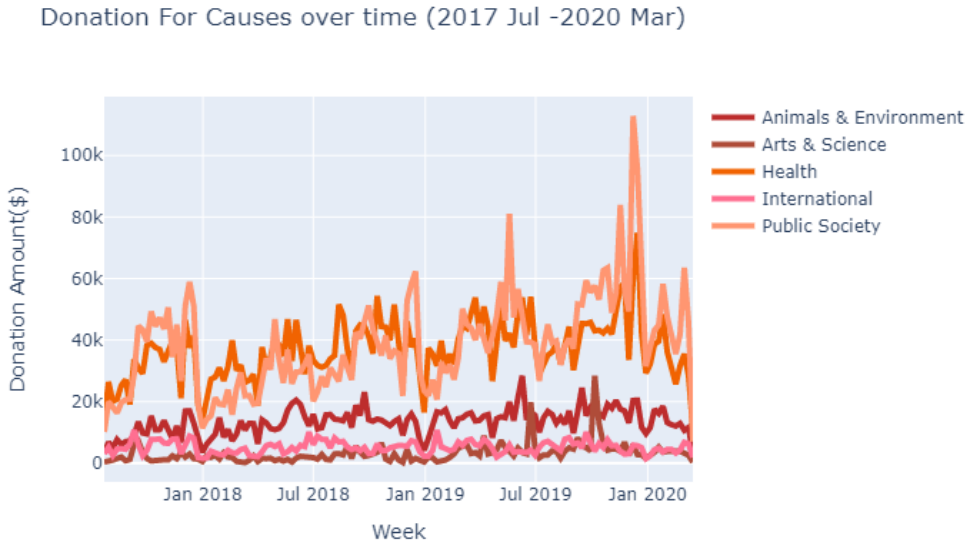


Figure 4.5: Donation for Grouped Causes over time

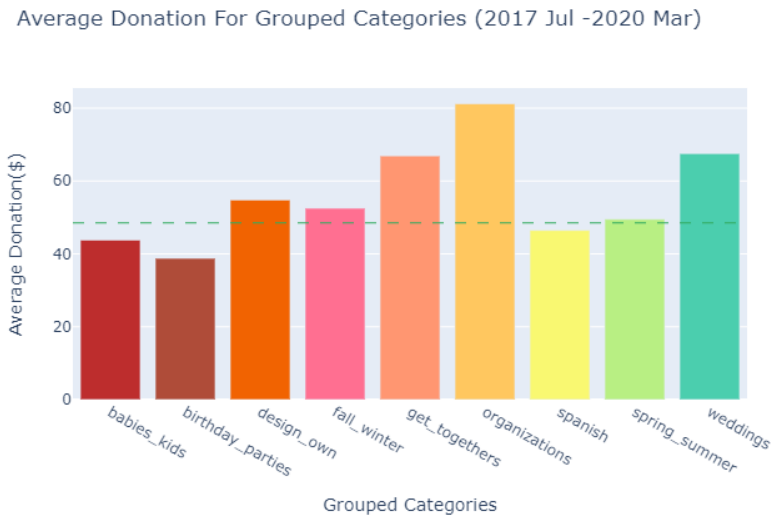


Figure 4.6: Average Donation for Grouped Categories

The green dashed line is the average donation amount amongst the entire data at $49. The average donation amount graph for an order for each grouped category is somewhat similar to the percentage of adding donation feature in Chapter 3. Organizations showed significantly higher average donation amount compared to other categories. The difference with this visual is that weddings and get togethers have above-average donation amounts and birthday parties had really low average donations.

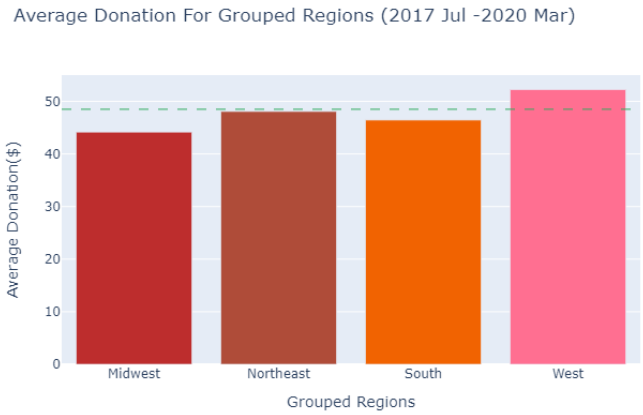


Figure 4.7: Average Donation for Grouped Regions

Average donations for grouped regions are similar to Chapter 3. West and northeast show higher average donation amounts, which is supported by the average wages for these regions. Average wages are high in states that are in these regions such as CA, WA, and NY.

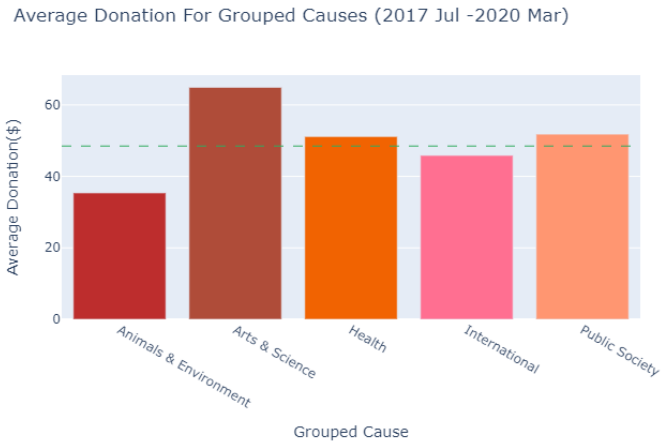


Figure 4.8: Average Donation for Grouped Causes

Average donation graph for grouped causes indicates that people tend to donate more to arts & science causes followed by health and public society. There were a lot more events for animals but average donations for animals-related non-profits were much lower than average.

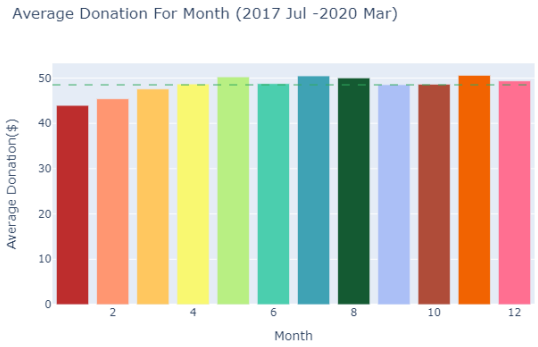


Figure 4.9: Average Donation for Months

Average donation graph for each month contradicts what we saw in Chapter 3. There were higher chance of adding donation features in Q1 and Q4, but average donations, on

the other hand, were much lower in Q1. This is probably due to lower donation amounts for birthday parties and fall winter events in Q1 which make up most of the events donations.

4.3 Model analysis and Performance Evaluation

Similar to Donation Feature Addition Analysis, it is necessary to divide dataset into training-test groups. For this study, 75% of the dataset was randomly selected as training data, while 25% of the dataset was used to measure the predictability with the trained model. When the linear regression fits the model, not all the features that go into the model will have coefficients statistically significant. The process to eliminate features that are not statistically significant (that has a p-value higher than 0.05) is necessary to have a robust model.

Initial linear regression model from all the features in the data is below:

OLS Regression Results

Dep. Variable:	Donation Amount	R-squared:	0.027
Model:	OLS	Adj. R-squared:	0.027
Method:	Least Squares	F-statistic:	404.8
Date:	Wed, 27 May 2020	Prob (F-statistic):	0.00
Time:	04:26:42	Log-Likelihood:	-1.6576e+06
No. Observations:	278792	AIC:	3.315e+06
Df Residuals:	278772	BIC:	3.315e+06
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Goal	7.674e-05	8.06e-06	9.524	0.000	6.09e-05	9.25e-05
Quarter_Q1	15.2308	1.435	10.611	0.000	12.418	18.044
Quarter_Q2	17.4770	1.435	12.179	0.000	14.664	20.290
Quarter_Q3	17.0281	1.429	11.917	0.000	14.228	19.829
Quarter_Q4	16.4558	1.432	11.490	0.000	13.649	19.263
Grouped Category_babies_kids	-3.9257	3.998	-0.982	0.326	-11.761	3.910
Grouped Category_birthday_parties	-7.1477	3.917	-1.825	0.068	-14.825	0.529
Grouped Category_design_own	7.3632	4.891	1.505	0.132	-2.224	16.950
Grouped Category_fall_winter	5.0437	3.973	1.270	0.204	-2.743	12.830
Grouped Category_get_togethers	19.2811	3.947	4.885	0.000	11.544	27.018
Grouped Category_organizations	32.8573	3.955	8.309	0.000	25.107	40.608
Grouped Category_spring_summer	2.3907	3.988	0.600	0.549	-5.425	10.206
Grouped Category_weddings	20.0566	4.009	5.004	0.000	12.200	27.913
Grouped Region_Midwest	13.0518	1.449	9.009	0.000	10.212	15.891
Grouped Region_Northeast	18.1768	1.437	12.652	0.000	15.361	20.993
Grouped Region_South	14.7808	1.427	10.355	0.000	11.983	17.578
Grouped Region_West	20.1823	1.424	14.178	0.000	17.392	22.972
Grouped Cause_Animals & Environment	3.2472	1.192	2.725	0.006	0.912	5.583
Grouped Cause_Arts & Science	21.5417	1.487	14.485	0.000	18.627	24.456
Grouped Cause_Health	15.3019	1.167	13.108	0.000	13.014	17.590
Grouped Cause_International	10.7174	1.283	8.356	0.000	8.203	13.231
Grouped Cause_Public Society	15.3835	1.163	13.230	0.000	13.104	17.663

Omnibus:	1083812.892	Durbin-Watson:	1.843
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4872748081285.626
Skew:	92.890	Prob(JB):	0.00
Kurtosis:	20483.218	Cond. No.	4.56e+17

Figure 4.10: Initial Linear Regression Summary

P-value for Grouped Category — babies kids, design own, fall winter, and spring summer — are 0.325, 0.132, 0.204, 0.549 respectively (greater than 0.05) so it's necessary to eliminate these features from the model because the coefficient for these features are not statistically significant to reject the null hypothesis that the coefficient is zero. The relatively high p-values indicate changes in these predictor varaibles are not associated with changes in the target variable. After eliminating the features, the final summary of the model is below:

OLS Regression Results

Dep. Variable:	Donation Amount	R-squared:	0.027
Model:	OLS	Adj. R-squared:	0.027
Method:	Least Squares	F-statistic:	476.3
Date:	Wed, 27 May 2020	Prob (F-statistic):	0.00
Time:	04:26:43	Log-Likelihood:	-1.6577e+06
No. Observations:	278792	AIC:	3.315e+06
Df Residuals:	278775	BIC:	3.316e+06
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Goal	7.693e-05	8.06e-06	9.546	0.000	6.11e-05	9.27e-05
Quarter_Q1	15.6384	0.371	42.109	0.000	14.910	16.366
Quarter_Q2	17.8817	0.387	46.152	0.000	17.122	18.641
Quarter_Q3	17.4951	0.358	48.865	0.000	16.793	18.197
Quarter_Q4	17.6007	0.312	56.483	0.000	16.990	18.211
Grouped Category_birthday_parties	-8.8771	0.494	-17.978	0.000	-9.845	-7.909
Grouped Category_design_own	5.5524	2.970	1.870	0.062	-0.269	11.373
Grouped Category_get_togethers	17.5560	0.691	25.397	0.000	16.201	18.911
Grouped Category_organizations	31.0919	0.735	42.282	0.000	29.651	32.533
Grouped Category_weddings	18.3694	0.989	18.571	0.000	16.431	20.308
Grouped Region_Midwest	13.6461	0.437	31.195	0.000	12.789	14.503
Grouped Region_Northeast	18.7857	0.368	51.004	0.000	18.064	19.508
Grouped Region_South	15.4297	0.336	45.878	0.000	14.771	16.089
Grouped Region_West	20.7543	0.319	64.983	0.000	20.128	21.380
Grouped Cause_Animals & Environment	3.8103	0.428	8.899	0.000	2.971	4.649
Grouped Cause_Arts & Science	22.1813	0.990	22.411	0.000	20.241	24.121
Grouped Cause_Health	15.7364	0.343	45.830	0.000	15.063	16.409
Grouped Cause_International	10.9484	0.668	16.380	0.000	9.638	12.259
Grouped Cause_Public Society	15.9395	0.339	47.043	0.000	15.275	16.604

Omnibus:	1083701.799	Durbin-Watson:	1.843
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4868202405369.915
Skew:	92.860	Prob(JB):	0.00
Kurtosis:	20473.663	Cond. No.	5.91e+17

Figure 4.11: Final Linear Regression Summary

The coefficients in the linear regression models can be interpreted as how much the continuous target variable increases with the increase in the continuous predictor variable by 1 or with the inclusion of the categorical variable. For example, an event categorized in organizations leads to increase of donation amount by $31 in average. On the other hand,

if an event is for birthday parties, it will decrease the donation amount by $8.9 in average. Goal does not have much impact in the donation amount since the coefficient for Goal is near to zero.

One noticeable aspect of the linear regression summary is that R-squared for this model is very low with 0.027. The low R-squared means data is very noisy with high variability. Since the model has all the features with low p-values and low R-squared, the summary indicates that the predictor variables still provide relevant information about the response even if the data points fall further from the regression line. The model with low R-squared is not able to provide precise predictions, but low p-values still indicate a real relationship between the significant predictors and the response variable [4].

The predictability of linear regression models can be also checked with mean errors as discussed in Chapter 2. This study will use Mean Absolute Error and Root Mean Squared Error. Mean absolute error for the test data was around 29.7, while root mean square error was around 79. Since both errors indicate high residuals in the prediction, we can conclude that the model does not have good predictability even if it provides useful information about relationships between predictors and the output variable.

```
Mean Absolute Error: 29.676701371295902
Root Mean Squared Error: 79.42434048613873
```

Figure 4.12: Errors of Prediction

To discuss a low R-squared value again, it indicates that predictor variable is not explaining much in the variation of the dependent variable, in this case the donation amount. The predictor variables, even though significant, are not accounting for much of the mean of the dependent variable. There can be several reasons why R-squared is low from the regression model. One could be that the predictor variables have non-linear relationships with the dependent variable. Another reason could be that the existing predictor variables are not enough to explain the variance of the dependent variable. From the limited dataset of merely several categorical variables without any relevant continuous predictor variables, it could be concluded that the model needs more relavent continous variables to be able to

explain the variance in the predictor variable. On the other hand, since the initial objective of the study was to extract important features that invite donations from events, as long as these contributions of categorical variables are statistically significant, the study provided a lot of relevant information about what kind of events generated more donation feature considerations as well as donation amounts themselves.

CHAPTER 5

Future Enhancement and Conclusion

5.1 Future Enhancement

In this thesis project, there are multiple areas where we can observe to make improvements. For example, algorithms that can explain non-linear relationship between predictors and output were not considered in the modeling. These black-box algorithms can improve predictability and fit of the model in some cases. There were, however, several reasons why I considered linear and logistic regressions for the scope of this study. The stakeholders' main focus was to see what kind of relationships features had with the output, and for this interpretability, linear and logistic regressions were the best choices. Even if decision tree algorithms provide feature importances, they are still not directly interpretable like linear/logistic regression's coefficients. On top of this factor, the data itself was not useful for predictability because there were only three or four categorical variables that were helpful as predictors. Even if the large sample size helped the training phase, more continuous features are needed to give better prediction results. Goal and State Average Wages were considered as predictor features, but they were not successful in providing any useful insights in the model. Adding more features will definitely help reduce the high variance in the model.

5.2 Conclusion

Plegeling and the stakeholders were interested in insights from their available data in Evite's events and donations. The questions I tried to answer for the stakeholders were what kind of Evite events are more likely to add the donation feature provided by Pledgeling network,

and what kind of Evite events with the added donation feature will make participants donate more. With the insights provided by this study, the business stakeholders can make actions to focus marketing and make promotions in certain regions or in certain time periods as well as make their recommendations for non-profit causes in smarter ways.

Over the course of the study, the dataset was initially prepared, joined and cleaned. For the preprocessing phase, proper filtering and handling missing values as well as mapping features into grouped features were covered. The exploratory data analysis was extremely important for this study since the graphs and visuals generate many useful insights regarding the relationships between features and the target variable. There were noticeable trends in the data such as events for organizations and weddings category generated higher average donations versus birthday parties. Additionally, causes for Science and Public Society/Health resulted in higher average donation amounts compared to Animals. Then we utilized these pre-modeling insights to predict if an event will add a donation feature as well as how much a participant will donate to an event. For modeling donation feature classification, logistic regression was implemented to result in predictability of 58%. For donation amount regression, linear regression was developed to indicate very low predictability with high errors, but it still demonstrated statistically significant relationships between features and the target variable.

All codes for this study can be found on Github [1].

References

[1] Github code:. https://github.com/daehyunk927/event-donation.

[2] Jason Brownlee. Logistic regression for machine learning, Apr 2016. https://machinelearningmastery.com/logistic-regression-for-machine-learning.

[3] Jason Brownlee. Smote for imbalanced classification with python, Jan 2020. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification.

[4] Minitab Blog Editor. How to interpret a regression model with low r-squared and low p values, Jun 2014. https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values.

[5] Mohit Gupta. Gradient descent in linear regression, May 2019. https://www.geeksforgeeks.org/gradient-descent-in-linear-regression.

[6] Knowledge@Wharton. How evite avoided becoming another social network flop, Jul 2018. https://knowledge.wharton.upenn.edu/article/evite-avoided-becoming-another-social-network-flop.

[7] Susan Li. Building a logistic regression in python, step by step, Sep 2017. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8.