Title
Revisiting Prediction of Credit Card Chargebacks in the Live Events Ticket Industry Using an Updated Tidymodels Framework

Author
Sierra, Angel

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Revisiting Prediction of Credit Card Chargebacks in the Live Events Ticket Industry Using an Updated Tidymodels Framework

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

by

Angel Sierra

2023

ABSTRACT OF THE THESIS

Revisiting Prediction of Credit Card
Chargebacks in the Live Events Ticket Industry
Using an Updated Tidymodels Framework

by

Angel Sierra
Master of Applied Statistics in
University of California, Los Angeles, 2023
Professor Frederic R. Paik Schoenberg, Chair

Fraudulent credit card chargebacks continue to be an ongoing issue in the live event ticketing industry. Using past work in the field as a guide, logistic, random forest, and k-nearest neighbor models are trained and evaluated using a Tidymodels framework. To address the imbalanced nature of the data set, upsampling, downsampling, SMOTE, ADASYN, and ROSE resampling techniques were applied to the data set. Findings suggest that past results are consistent in that unsampled random forest models perform best for predicting chargeback fraud. The potential to streamline more machine learning models using a tideymodels framework seems possible and would have potential benefit for company use. Sales Amount associated with the order stands out as an influential variable in predicting chargeback fraud.

The thesis of Angel Sierra is approved.

Robert Gould

Ying Nian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2023

For my father, mother, sister, and brother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This thesis would not exist without the work done by Kjell Sawyer. Thank you.

CHAPTER 1

Introduction

1.1 What is a Chargeback?

The credit bureau Experian defines a chargeback as "a consumer protection tool you can use to dispute a charge and reverse a transaction" [1]. Chargebacks are meant to be used by customers to correct billing errors or potential fraud. In unpacking this definition provided by Experian, some realizations arise. The chargeback process is a "tool" for the customer in mind, suggesting that the concerns of the company are less important than those of the customer. This sentiment reveals that with respect to chargebacks, credit card companies prioritize customers over companies. The definition then states that chargebacks are a tool used to "dispute a charge." The chargeback process a lengthy one. One that begins with a dispute between the cardholder and the issuing bank. The merchant's payment processor then removes the transacted amount from the merchant's account and informs the merchant of the dispute claim. At this point the merchant is given the option to dispute that claim. The merchant must provide evidence to dispute this claim. This is where the previously mentioned sentiment has effect. It is the duty of the merchant to fight the chargeback, and it is often the case that the bank will not find the evidence provided by the merchant as sufficient to refund the amount removed from the merchant's account. Chargebacks tend to lean towards the customer. They are inevitable in every industry, but are particularly an issue for live event ticketing. Credit card companies do not want to do business with merchants with high rates of chargebacks. If chargeback rates exceed a certain percentage of orders for an extended period of time, credit card companies will close their account with

the merchant's payment processor. These accounts are integral to a company's survival.

There are three different types of chargebacks. There is merchant error, friendly fraud, and true fraud. Merchant error chargebacks occur because of mistakes made on the part of the merchant. For example, if a customer ordered five VIP tickets, but received five General Admission tickets, they are within their rights to begin a chargeback dispute. While these types of chargebacks are preventable and can be minimized, it is not the focus of this thesis. Friendly fire are chargebacks that result from outside the power of the customer. They are usually the result of an accidental purchase. These chargebacks will always occur and are difficult to minimize. It is impossible to stop any customer from accidentally making a purchase. True fraud are the chargebacks at the center of this thesis. These are the chargebacks that are the result of credit card or account theft. These are chargebacks that are difficult for companies to win in the chargeback process but can be minimized. And because of the penalty associated with having too many chargebacks associated with a payment processor's account, there is an importance to minimize these occurrences.

The occurrence of true fraud chargebacks are rare. Less than one percent of all orders are expected to be true fraud [2]. This is known as an imbalanced data set. There are known approaches to address this issue, as will be discussed. Looking for details or aspects in these true fraud chargebacks can be generalized so that when these details or aspects are seen in other purchases, consideration from a non machine learning perspective can be taken to decide whether these purchases may be fraudulent.

1.2 Past Work

In 2019, MAS alumni Kjell Sawyer implemented fifteen machine learning models on historic ticket order data to predict which transactions could become fraudulent [3]. As a result of the work, fraudulent chargebacks went down, and company time and money were saved. To address the imbalanced nature of the data set, Kjell utilized four different resampling tech-

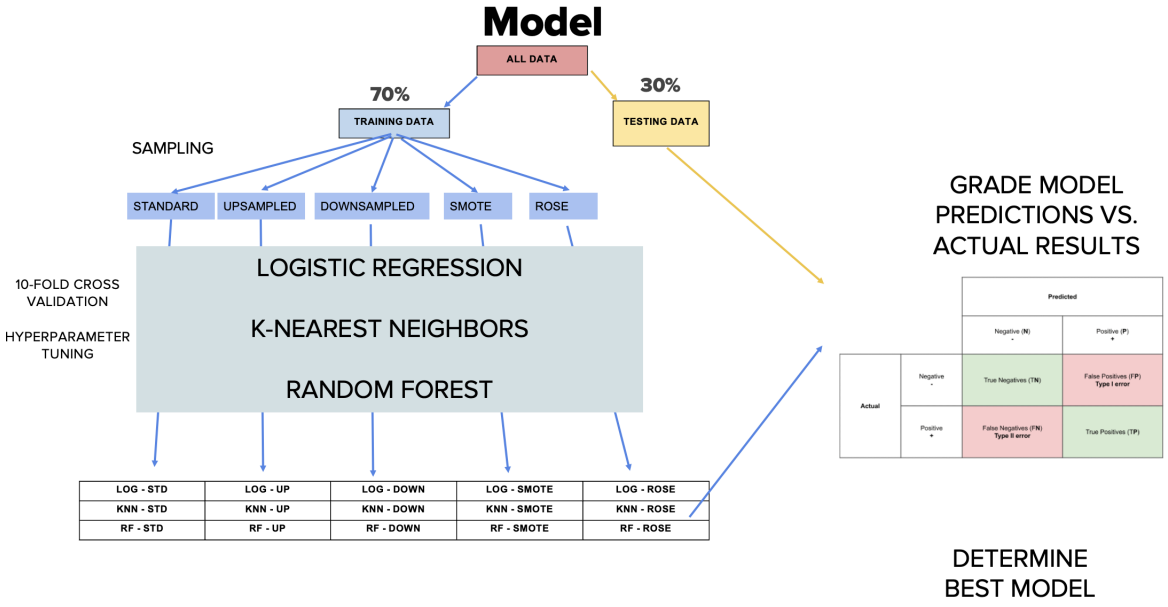

Figure 1.1: Modeling Technique Used by Kjell Sawyer

niques: upsampling, downsampling, Synthetic Minority Oversampling Technique (SMOTE) sampling and Random Over-Sampling Examples (ROSE) sampling. With these four re-sampled data sets and the unsampled data set, he applied three different models: logistic regression, random forest, and k-nearest neighbors. This resulted in 15 models, as outlined in Figure 1.1 [4]. 10-fold cross validation was used to help create metrics that could be used to evaluate model performance. In his conclusion, the unsampled random forest model performed the best.

1.3 Purpose

Since the completion of his thesis, there has been a change of payment processing accounts within the company. An important feature with this new payment processor is the existence of "risk score" associated with every purchase. This measure is designed to measure the "likelihood that a payment is fraudulent" [5]. This will be added to new models. Instead of

four resampling techniques, five will be utilized. Adaptive synthetic (ADASYN) sampling will be included with the other four resampling techniques. Logistic regression, random forest, and k-nearest neighbors were applied to all these samples, resulting in 18 different models. Implementation of these models were done using tidymodels packages, with the motivation to evaluate the ease of implementation. Easy implementation of these machine learning models would allow for the streamlining of these type of models to similar data sets. Recently, requests for checking orders for potential fraudulent transactions have gone up. While the implementation of machine learning models is possible here, identification of potentially fraudulent orders can also be done manually. This is done by examining details of the orders. Kjell's findings suggest that the quantity of tickets in an order is a significant indicator of potential fraudulent orders. By examining current data with similar models, past findings can be updated to determine what other factors may help signify chargeback fraud.

The purpose of this thesis is 3-fold: to train new models on current data with a tidymodels framework, to observe the effects of switching payment processors, and to attempt to update current strategies for countering chargebacks.

CHAPTER 2

Exploratory Data Analysis

2.1 Data Set

The data set is all orders made through an events ticketing company between the years 2021 through early 2023. It is important to note that COVID restrictions were lifted sometime in 2021 [6]. The data set is very similar to the data set used in Kjell's thesis. Both data sets come from orders of the same type live event from the same geographic location. The data set contains 59,581 orders, with 58,955 legitimate orders and 626 orders with fraudulent chargebacks. Email address is recorded for every order, along with user birthday, full name, and gender if filled in. Time and location of purchase are recorded and will be used as well. Type of credit card used to make the purchase is also recorded. Models were trained using 19 predictor variables, with one outcome variable. Table 2.1 describes all variables used.

The correlation plot shows minimal correlation between the predictors. No data was intentionally removed. This would include potential outliers. Because of how rare fraudulent orders are, and because fraudulent orders may have extreme characteristics, outliers were not removed from the data set.

Variable	Description
QUANTITY	Number of Tickets in the Order
SALES_AMOUNT	Total Sales Amount of the Order
HOURS	Difference between Purchase Time and Event Start Time
TIME	Hour of the Day (24) of Purchase Time
LATITUDE	User Latitude During Purchase
LONGITUDE	User Longitude During Purchase
NCHAR	Number of Characters in Email Address before @
EMAILNUMBERS	Count of Numeric Characters in Email Address before @
RATIO	Ratio of Numbers to Letters in Email Address before @
MONTH	User Birth Month
DAY	User Birth Day
YEAR	User Birth Year
NAMELENGTH	Number of Characters in User Full Name
PER.CONS	Percent of Consonant Letters in User Full Name
CAP	Number of Capital Letters in User Full Name
GENDER	Stated Gender of User
DOMAIN	Whether Email is .com Associated
OUTCOME_ RISK_ SCORE	Risk Score Provided by Payment Processor
CARD_ BRAND	Brand of Credit Card used by User
TYPE	Outcome: LEGITIMATE or CHARGEBACK order

Table 2.1: Variable Description

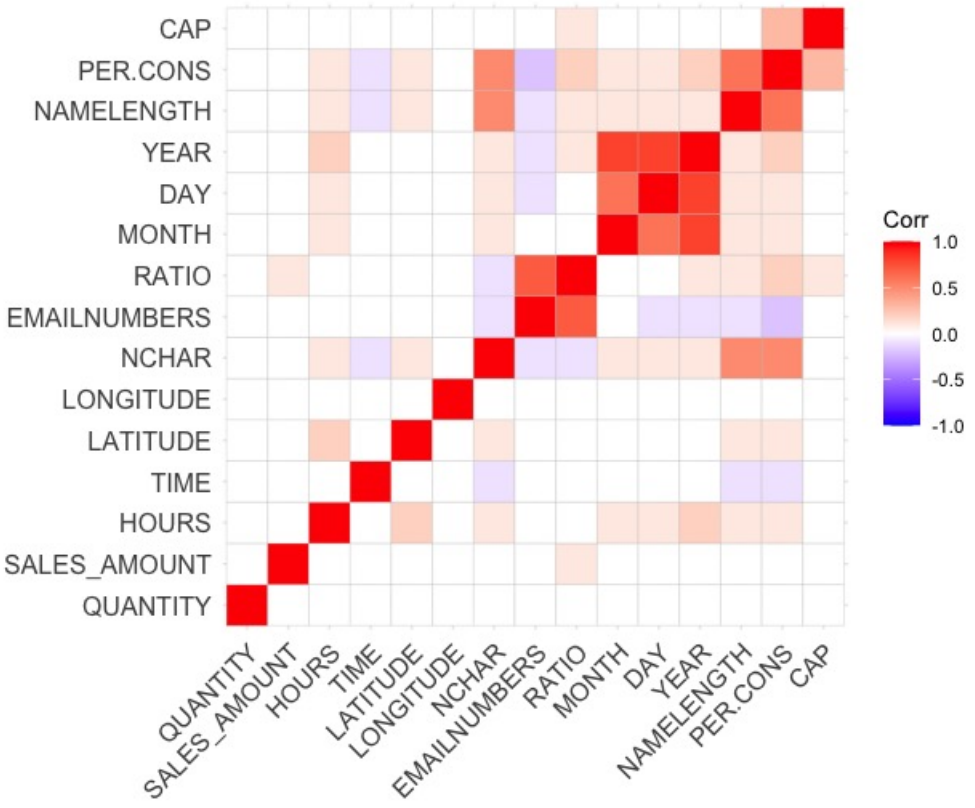

Figure 2.1: Correlation Plot of the Predictors

2.2 Descriptive Analysis

Smoothed density estimates of all variables were explored. QUANTITY represents the number of tickets in the order. Compared to the distribution of quantity of tickets between fraudulent and legitimate orders, the density of legitimate orders decrease as number of tickets increase, but with fraudulent chargebacks, the density increases slightly then decreases as number of tickets increase. SALES_ AMOUNT represents the total sales of the order. This differs from Kjell, which examined average price of tickets on the order. The average price of an order with an associated chargeback is higher than that of a legitimate order. HOURS represent the number of hours before the live event in which the order was placed. Chargeback orders seem to occur more often closer to the date of the event than otherwise. Less chargebacks seem to occur during the beginning and end of the 24 TIME hour cycle. LAT and LONG represent the latitude and longitude of the location where the order was made. There does not seem to a significant difference between legitimate and chargeback orders in terms of geography. The number of characters in a user's email, NCHAR, seems to be smaller with chargeback orders than with legitimate ones. There does not seem to be much of difference between the amount of numbers in the user email EMAILNUMBERS. User birthday is broken into MONTH, DAY, and YEAR. The difference between chargeback and legitimate densities seem to be similar. The densities of NAMELENGTH, the number of characters in first and last name, seem to be slightly different between chargebacks and legitimate orders. The same seems to be true for PER.CONS, the percent of consonant letters in the user full name. The number of capital letters in the user full name seems to skew towards having fewer for fraudulent orders. GENDER of user seems to be similar in density between both kinds of orders. This variable is split into NA – 0, Male – 1, Female – 2, Non-binary – 3, and Prefer-not-to-say – 4. DOMAIN is the domain of the email of the user. It seems like being a chargeback is a subset of both coming from ".com" and not. CREDIT_ CARD_ BRAND shows what brand of credit card was used for the purchase: Mastercard – 0, Visa – 1, American Express – 2, Other – 3. It seems that chargebacks are more likely to

come from specific credit card brands. OUTCOME_ RISK_ SCORE indicates the likelihood an order is fraudulent as provided by the payment processor, and it does seem that orders with higher scores tend to chargeback. It can be noticed that in some of the graphs, the chargeback curves are always smoother. This is in part due to the imbalanced nature of the data set. Higher occurrences of specific values at specific ticket prices cause spikes at those values, whereas for chargebacks, these value occurences do not occur at such frequencies, and are not as well defined. All together, these density plot give a sense as to what variables may be of interest in our machine learning models.

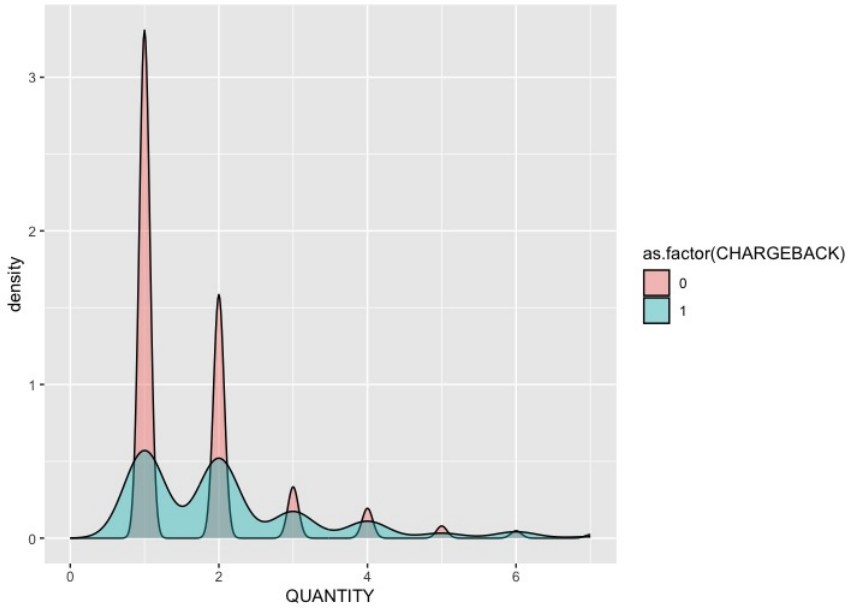

Figure 2.2: Smoothed Density Estimate of Number of Tickets per Order. Chargeback orders have a slightly higher mean number of tickets ordered.

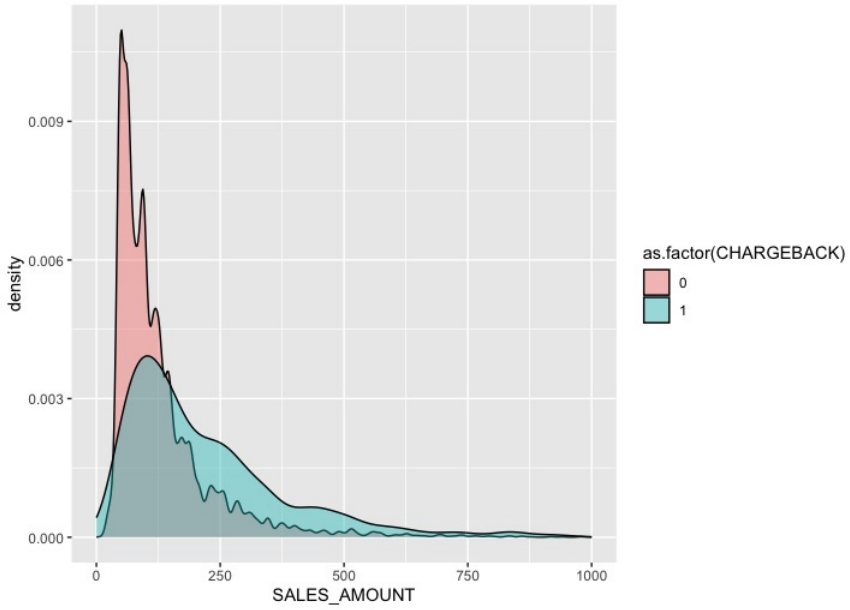

Figure 2.3: Smoothed Density Estimate of Total Sales of Order. Chargeback orders have a slightly higher mean total sales amount.

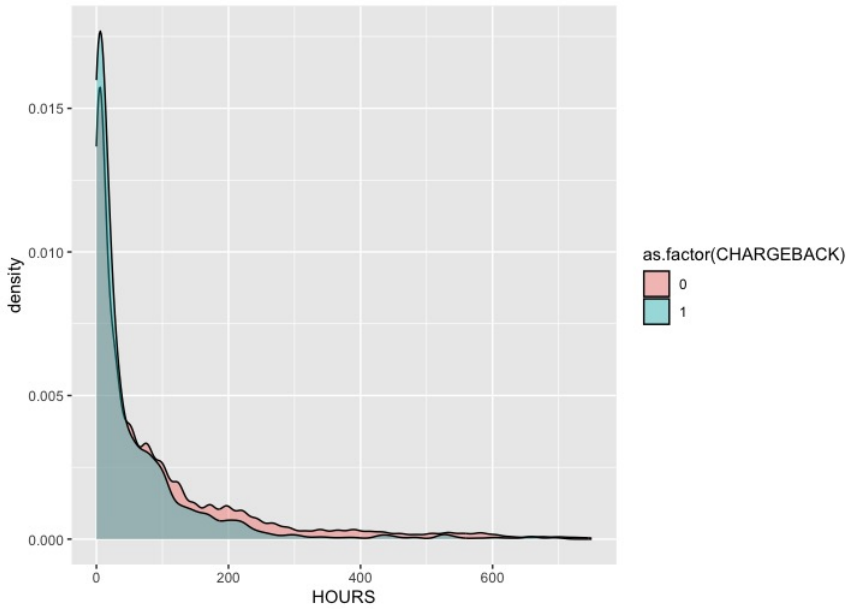

Figure 2.4: Smoothed Density Estimate of Time Between Purchase Date and Event. Legitimate orders seem to occur earlier than chargeback orders, which seem to take place right before the event.

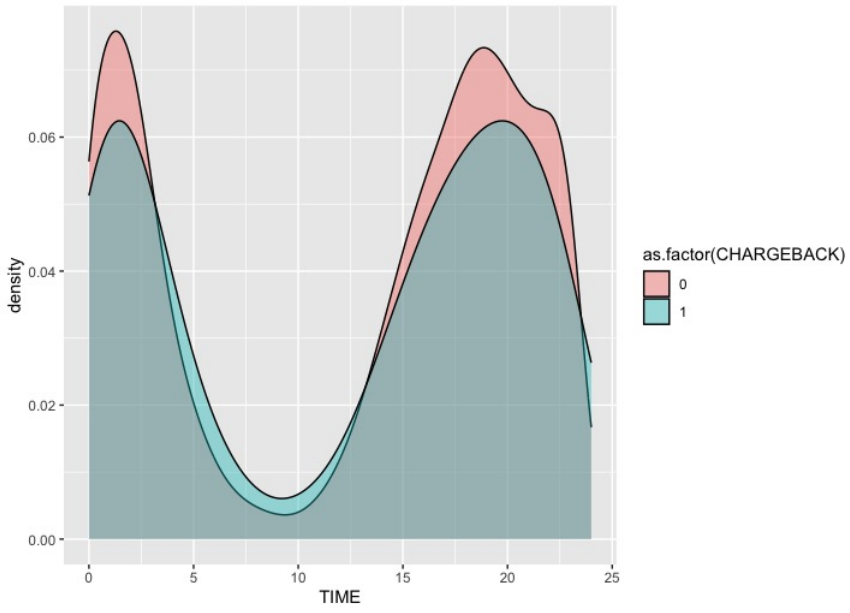

Figure 2.5: Smoothed Density Estimate of Hour of Day Purchase was Made. Chargebacks seem to occur at slightly different times of the day than legitimate orders.

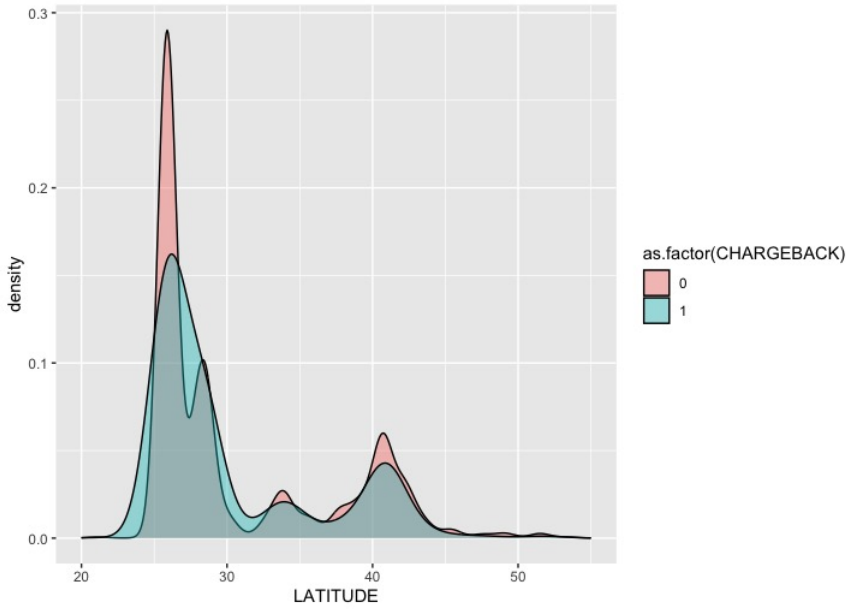

Figure 2.6: Smoothed Density Estimate of Latitude of Purchase Location. Densitites seem similar.

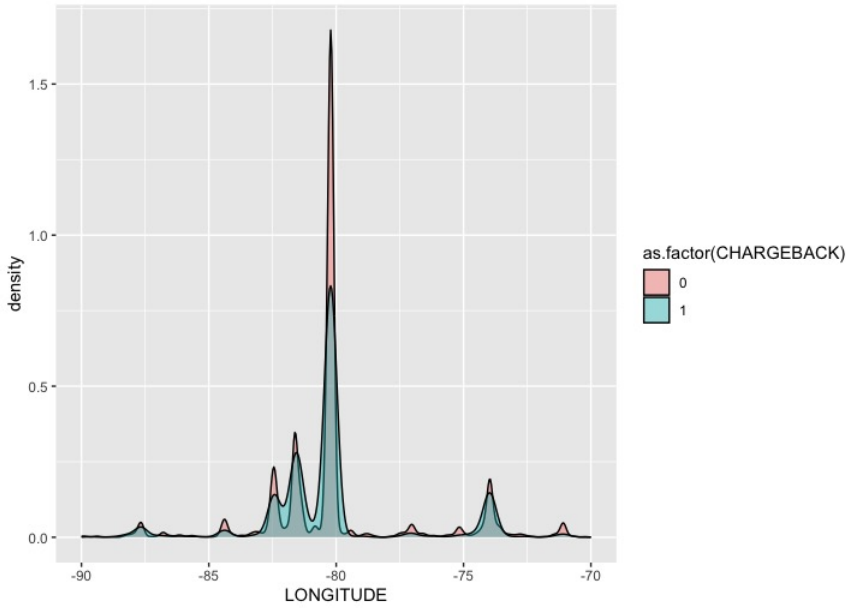

Figure 2.7: Smoothed Density Estimate of Longitude of Purchase Location. Densities seem similar.

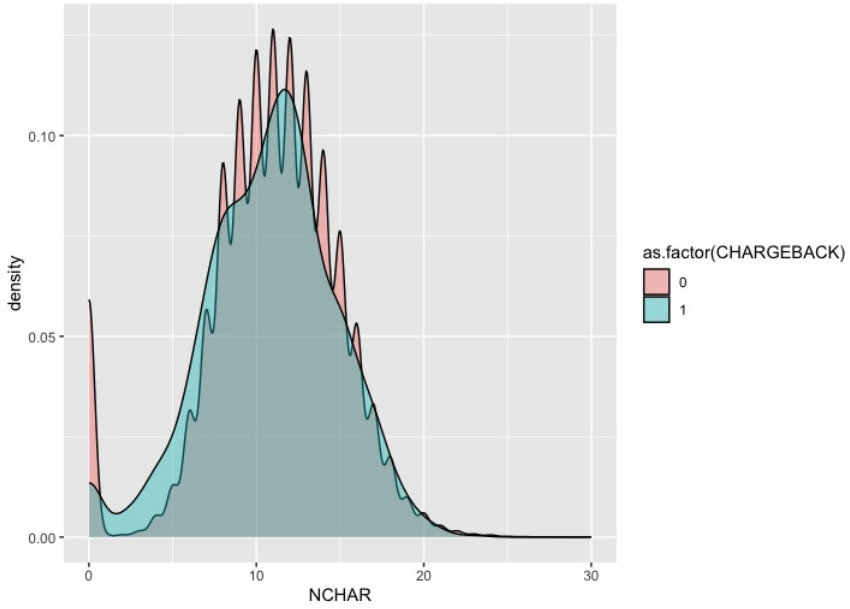

Figure 2.8: Smoothed Density Estimate of Email Length. Chargeback orders seem to have a fewer number of characters than legitimate orders.

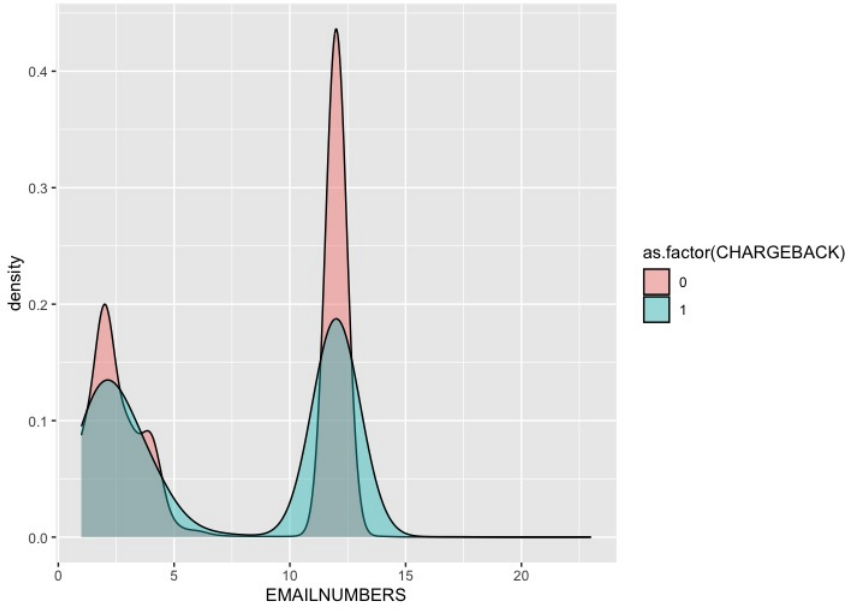

Figure 2.9: Smoothed Density Estimate of Numbers in User Email. Chrageback orders seem to have similar number of numbers in their emails.

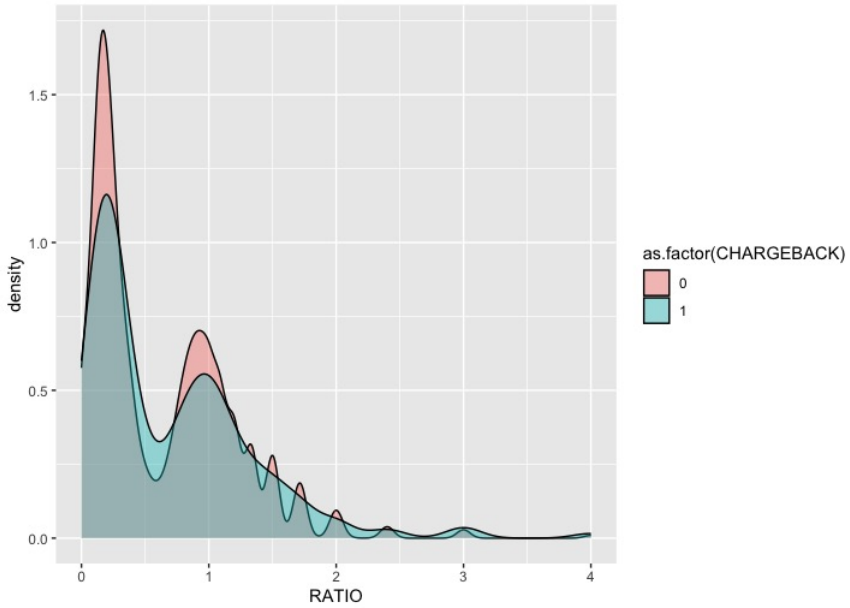

Figure 2.10: Smoothed Density Estimate of Email Address Number to Letter Ratio. Chargeback orders seem to have a higher Ratio of Numbers to letters than Legitimate orders.

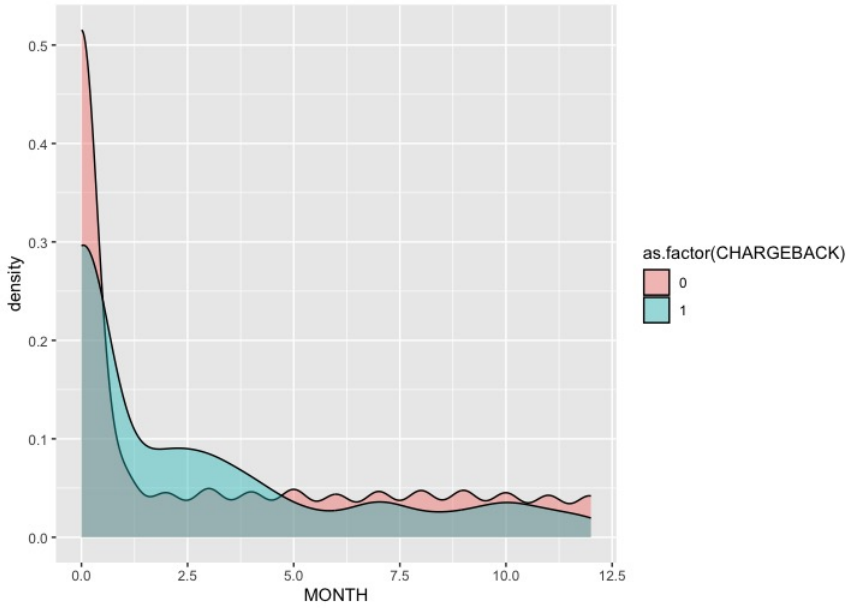

Figure 2.11: Smoothed Density Estimate of User Birth Month. Chargeback orders are not as evenly distributed as legitimate orders.

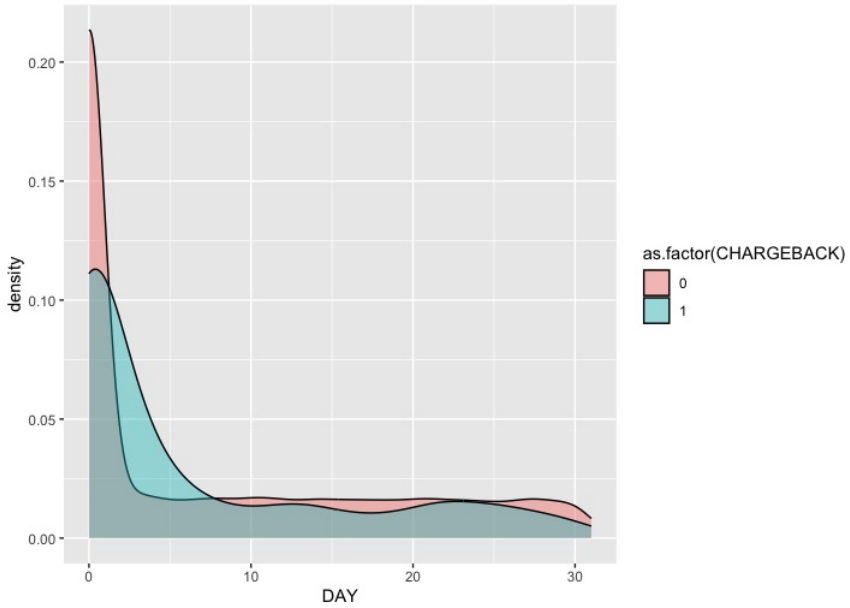

Figure 2.12: Smoothed Density Estimate of User Birth Day. There is more variation in User Birth Day for chargeback orders than for legitimate.

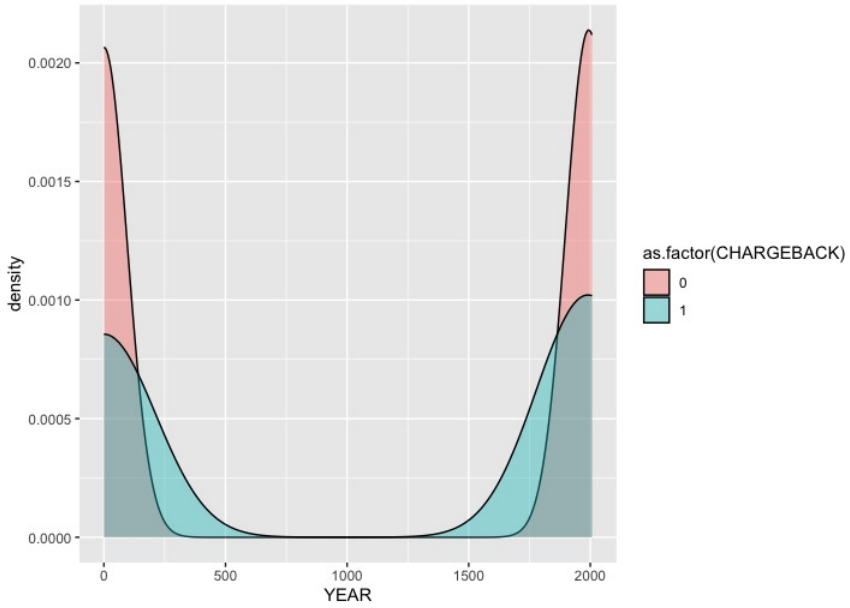

Figure 2.13: Smoothed Density Estimate of User Birth Year. There is more variation in User Birth Year for chargeback orders than for legitimate.

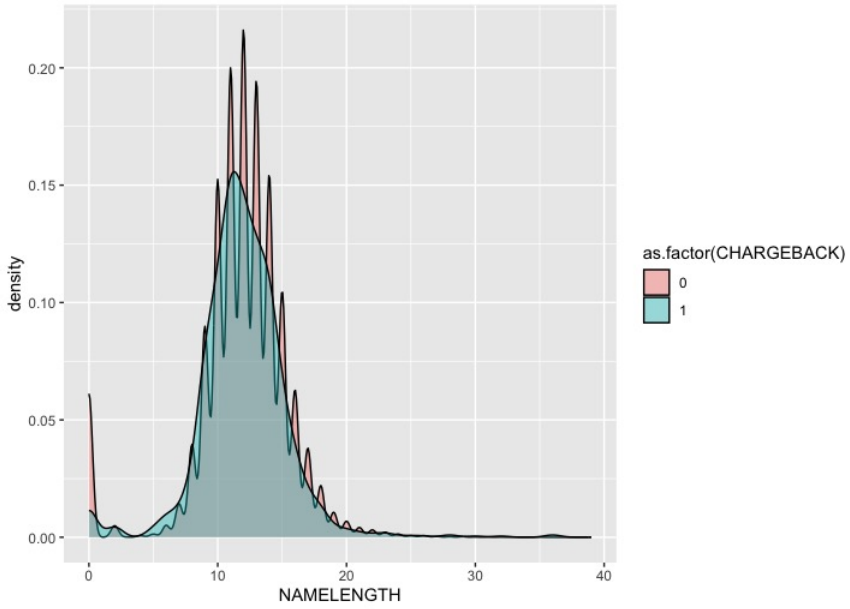

Figure 2.14: Smoothed Density Estimate of User Name Length. Distributions seem similar.

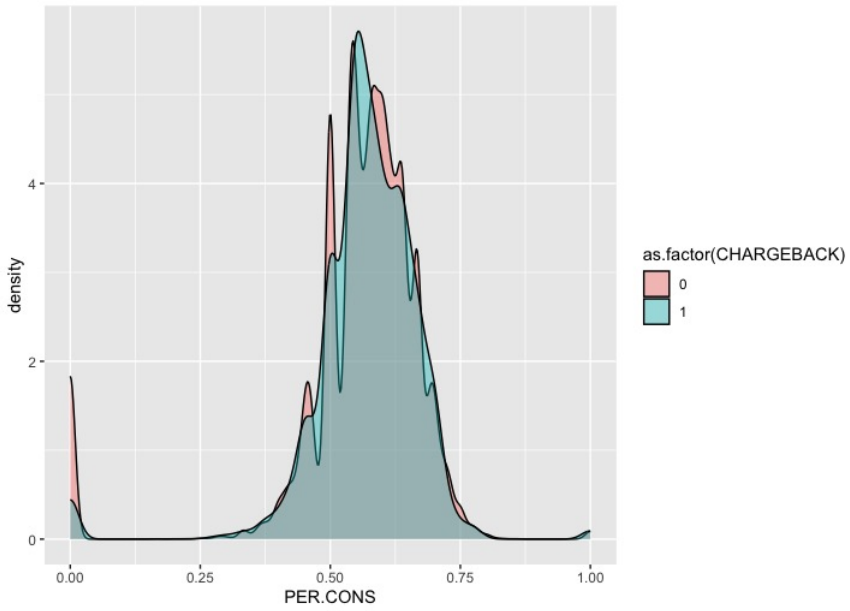

Figure 2.15: Smoothed Density Estimate of Percent Consonants in User Name. Distributions seem similar.

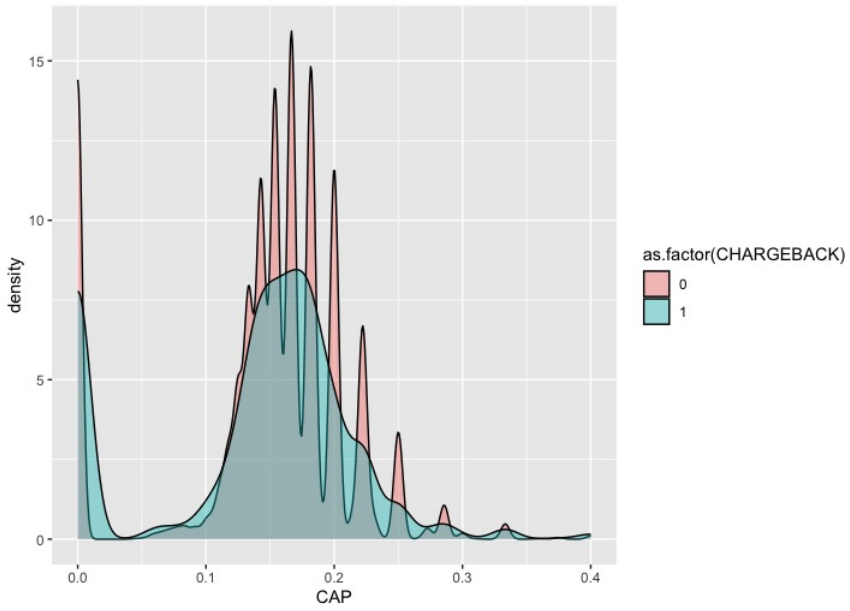

Figure 2.16: Smoothed Density Estimate of Number of Capital Letters in Email. Distributions seem similar.

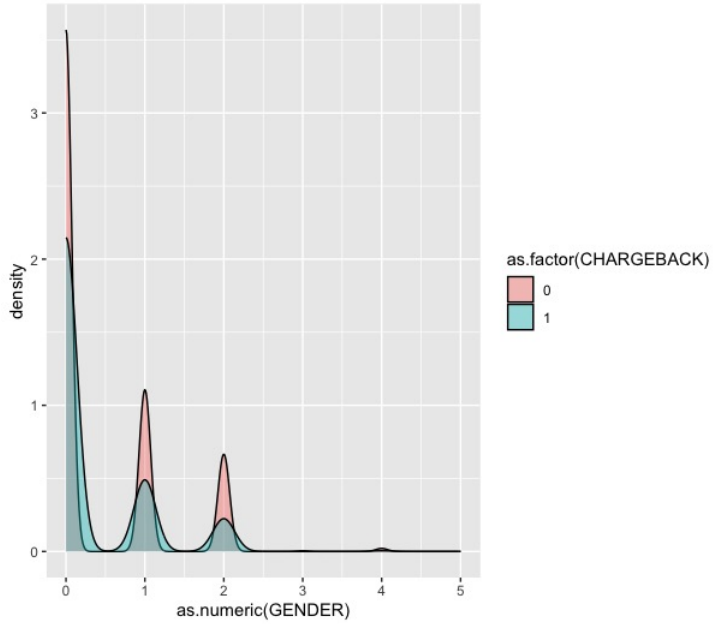

Figure 2.17: Smoothed Density Estimate of Reported User Gender. Distributions seem similar.

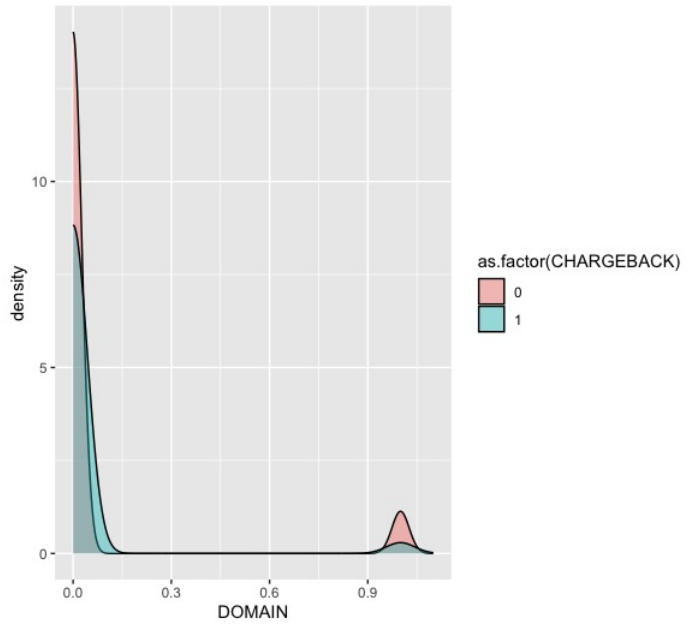

Figure 2.18: Smoothed Density Estimate of Domain. Densities seem similar.

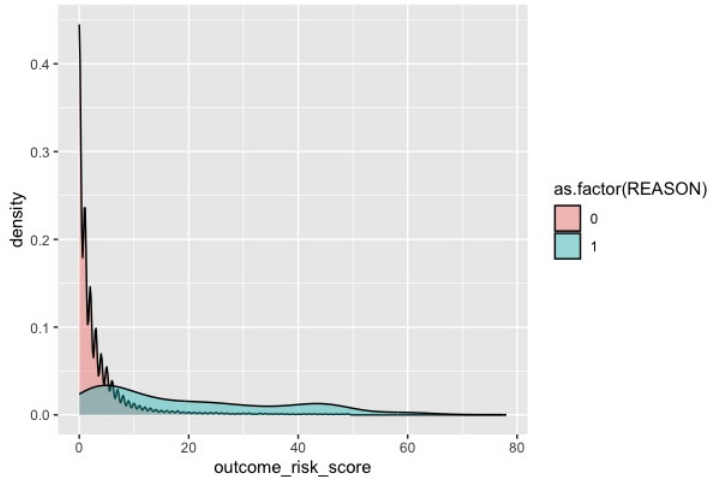

Figure 2.19: Smoothed Density Estimate of Risk Score. Outcome risk scores for chargeback orders seem higher on average than for legitimate orders.

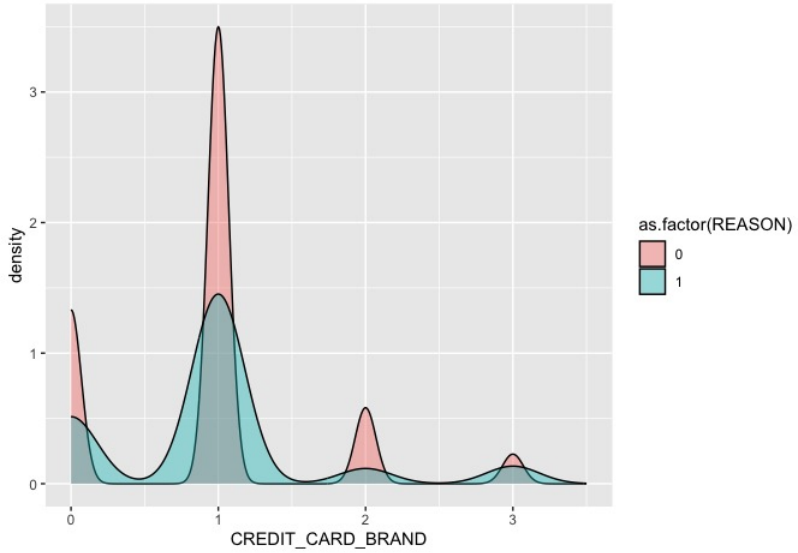

Figure 2.20: Smoothed Density Estimate of the Credit Card Brand. Chargebacks seem to occur at higher rates with non Visa, Mastercard, and American Express credit cards.

CHAPTER 3

Models

3.1 Logistic Model

When the predictor variable has a binary outcome such as 0 or 1, logistic regression is often a reasonable approach. This is essentially a linear model that can predict the natural logorithm of odds (log odds) which corresponds to a specific probability. Note that it is specifically the natural log of odds as to maintain all values between 0 and 1. The corresponding relationship between the log odds of any variable and the probability is shown in 3.1

$$p_x = \frac{\exp(x)}{1 + \exp(x)} \tag{3.1}$$

The inverse can be taken to find the log odds given any probability with the logit function as in 3.2

$$\text{logit}(p) = \ln \frac{p}{1-p} \tag{3.2}$$

The logistic regression model is a generalised linear model that uses the logit as a link function [7]. A linear equation is used to predict the logit of a probability.

The lostistic regression function is often written as

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \tag{3.3}$$

This linear regressive model is then fit on the training data and is then fed the testing data to output values between 0 and 1. Anything less than 0.5 would be classified as a legitimate order and anything greater than 0.5 would be classified as a chargeback fraud.

3.2 Random Forest

Random Forests are ensembles of decision trees. Many decision trees are made for this ensemble based on slight differences in the training data. The creation of these individual trees encourages for a variety of differences between trees [7]. The collection of all these decision trees results in a random forest. To classify with this random forest, votes are taken from each tree and then the average is taken to determine the final classification of the given sample. This can be seen in Equation 3.4

$$f_{prediction} = \frac{1}{B}\sum_{b=1}^{B} f_{decision\ tree\ b}(x) \quad \text{for } b = 1, ..., B \tag{3.4}$$

Because of the random nature of where the splits may occur for the decision trees, impurities may arise. For instance, if one tree splits on one variable and another tree splits on a different variable, the variable of more importance can be determined, depending on which variable allows for a more likely prediction. Given a total of T classes (the classifier is binary in this case) and $p(i)$ is the probability of picking a datapoint from class i, then the Gini Impurity is in Equation 3.5

$$G = \sum_{i=1}^{T} p(i)(1 - p(i)) \tag{3.5}$$

This impurity allows for determination as to which variables serve more importance in classifying the data.

3.3 k-nearest Neighbor

k-nearest neighbor takes in a training set knowing which orders are chargebacks and which are not. Each order represents it's own point in the model and the distance between the points are measured using the values of the predictor variables [7]. For example, two orders that were made by somebody who shares the same birthday will be close in distance in terms of birthday variables, but may be further away with email characteristics. Note that for this distance, chargeback orders are expected to be nearer to each other than to legitimate orders. So given a testing point, the k-nearest neighbors can be examined to see if they are chargebacks or not. If the majority are chargebacks, it is likely the test point is one as well. The distance is measured often with a Euclidean manner (3.6).˘

$$d(x, y) = \sqrt{\sum (y_i - x_i)^2} \tag{3.6}$$

Due to the number of dimensions in the data, Manhattan distance (3.7) is chosen. Default parameters were selected for this model.

$$d(x, y) = \sqrt{\sum |y_i - x_i|} \tag{3.7}$$

3.4 Accuracy Metrics

Table 3.1 refers to the confusion matrix. A is true positive, B is false positive, C is false negative, and D is true negative. In this thesis, A is legitimate orders, D legitimate chargebacks, C is legitimate orders classified as chargeback, and B is chargebacks but are classified as legitimate orders.

10-fold cross-validation was used to measure the performance of the models. The following metrics were used to evaluate the performance of a model: ROC accuracy, Accuracy, Sensitivity, and Specificity. ROC accuracy is the area under the ROC curve comparing the

Table 3.1: Confusion Matrix

		Actual	
		Legitimate	Chargeback
Predicted	Legitimate	A	B
	Chargeback	C	D

relation between true positive rate and false positive rate. Accuracy is the percentage of correct predictions made by the model. Sensitivity is the true positive rate (3.8), and here represents the number of correctly identified legitimate orders.

$$\text{Sensitivity} = \frac{A}{A + C} \tag{3.8}$$

and Specificity is the true negative rate (3.9), and here represents the number of correctly identified chargeback orders.

$$\text{Sensitivity} = \frac{D}{B + D} \tag{3.9}$$

10-fold cross-validation allows for the assessment of model performance without entirely predicting on the training set, as shown in Figure 3.1 [8]. A confusion matrix was also used to evaluate each model. Because of the imbalanced nature of the data set, the true negative rate is one of the most important factors to consider when measuring the performance of the model. This represents the number of true fraudulent chargebacks that the model was able to predict. Also worth noting, it is important to minimize false negative rates. Again, legitimate orders are positive, and chargebacks are negative in our confusion matrix. These would be orders that will not have any chargeback associated with them but will be predicted by the model to have been fraudulent. If these orders were believed to be fraudulent, and were stopped, they would represent lost money from legitimate orders.

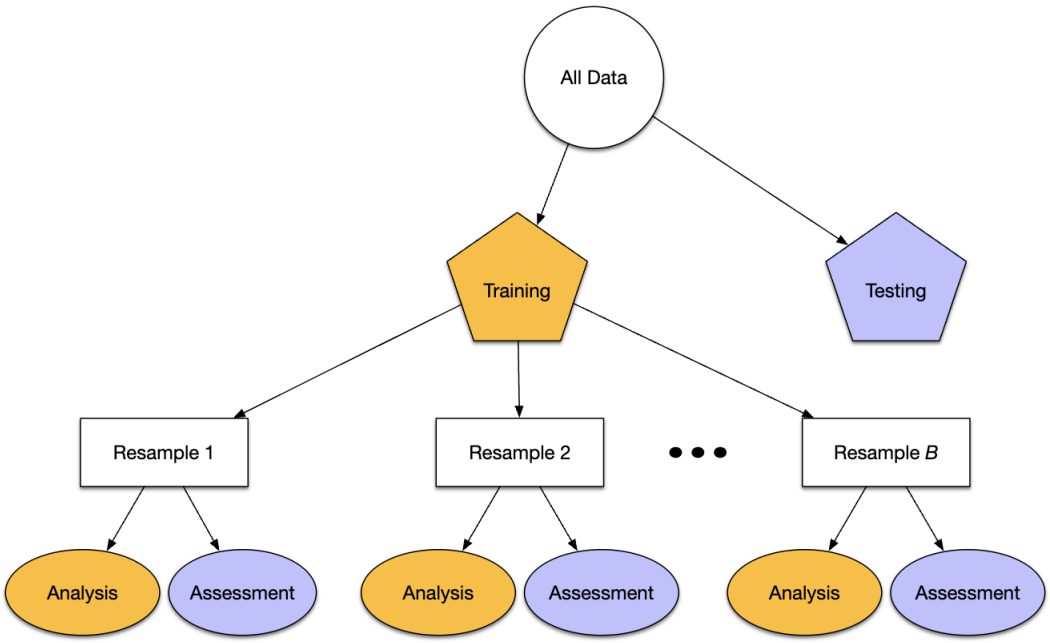

Figure 3.1: Resampling Procedure Schematic

CHAPTER 4

Model Implementation

4.1 Resampling Techniques

The unsampled data set is simply the raw data. The goal of upsampling is to address the imbalanced nature of the data set. Rows will be replicated in the data of the minority class until they match the number in the majority class. Downsampling has a similar goal, but will randomly remove rows of the majority class until they match the number of rows in the minority class.

The ROSE samples are created from a sample of synthetic data that is created with the aim of balancing the features of the minority and majority class. This technique based on a smoothed bootstrap re-sampling is followed as outlined in ROSE (Menardi and Torelli, 2014) [9].

SMOTE sampling creates more rows of the minority class by using a nearest neighbor algorithm. SMOTE sampling involves the over-sampling the minority class in a process that creates synthetic minority class examples. A random chargeback is selected and then a neighboring legitimate order is selected. The resulting synthetic data would be a value between these two [10].

ADASYN sampling applies an adaptive synthetic algorithm instead of a nearest neighbor algorithm to achieve the same feat [11]. The number of majority neighbors of each minority group determines the number of synthetic rows of data to be generated from the minority group. Essentially, for each non chargeback neighbor a chargeback has, a new synthetic

chargeback neighbor is added, entirely based on the pool of preexisting chargebacks.

4.2 Tidymodels Framework

Implementation of all the models were done using a tidymodels framework, and done using R version 4.2.1 in an RStudio IDE on an Apple M1 Pro[12]. The tidymodels framework is a collection of packages for machine learning and model building that are all done using tidyverse principles. The motivation for using this framework is to see how easy these models are to create and manage for possible streamlining purposes. The data set is first split between a training and testing set, with presence of chargebacks being stratified so that a proportional number of chargebacks were present in the training and testing sets. 10-fold cross-validation was performed on the training set to allow for resampling to be used for model performance [13].

The first big step in the tidymodels framework is to build a "recipe." This is a preprocessing step that helps transform data sets into objects that can be taken into and fitted by models. Data cleaning occurs in this step. Six different recipe objects were created at this point, one for each sampling technique (unsampled, upsampled, downsampled, ROSE, ADASYN, and SMOTE). In each of these recipes, the data was centered and scaled, and it is worth noting that conversion of qualitative variables to indicator variables would happen at this step.

Separately, model objects are created using certain specifications. Models are initiated in this step, and specifications are loaded in using tidyverse principles, indicating which "engine" to run – which package or system to be used to fit the model. Logistic models are initiated with the "glm" package. Random forest specifications were set to use the "ranger" package, and with 1000 trees. The nearest neighbor specifications were set the use the "knn" package. All were set to be classification models. It is also important to note that tuning parameters can be set at this step. The example in Appendix A helps outline tidyverse

principles.

To help put everything together, "workflow" helper objects need to be created. These helper objects help manage modeling pipelines that allow previously made pieces to be added in. A workflow is created for each sampling technique and each created recipe is added to these "workflow" objects using tidyverse syntax.

The final step for creating the models is to fit the resamples to the workflow helper objects. Resamples and performance metrics are specified here. Models were performed with parallel processing to speed run time. With six different workflow objects and three different types of models to be fitted, 18 models were fitted and evaluated in total. As an example, Appendix A contains code used to fit a random forest model on upsampled training data.

CHAPTER 5

Model Evaluation

5.1 Confusion Matrix and Accuracy Metrics

Accuracy, ROC accuracy, sensitivity, and specificity were all used to evaluate model performance. These estimates are pulled from the resampling results of the 10-fold cross-validation done on the training data. Because of this, confusion matrices are made with the resampled data sets. Accuracy is an important metric, but it does not encompass everything to be are examined. The goal is to not only find a model that accurately predicts which orders will become fraudulent, but to also not inaccurately predict too many orders to be fraudulent. Because of this, models will not not only look for high accuracy and ROC accuracy, but also high specificity. The evaulation metrics for the logistic models are shown in Figure 5.1, for the random forest models in Figure 5.2, and for the knn models in Figure 5.3.

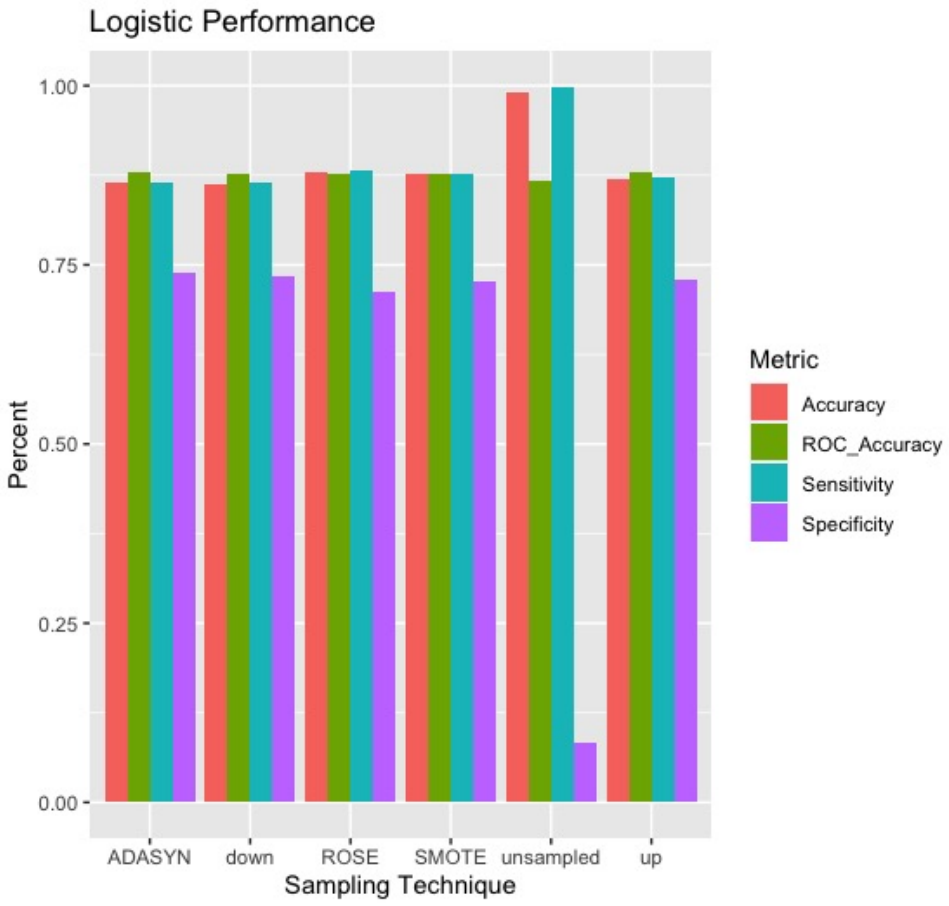

Figure 5.1: Accuracy, ROC Accuracy, Senitivity, and Specificity of Logistic Models

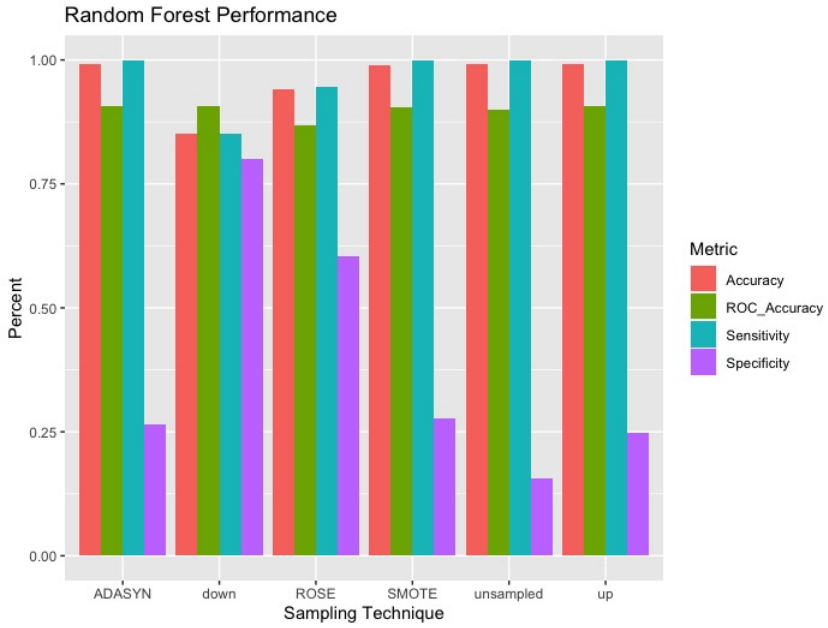

Figure 5.2: Accuracy, ROC Accuracy, Senitivity, and Specificity of Random Forest Models

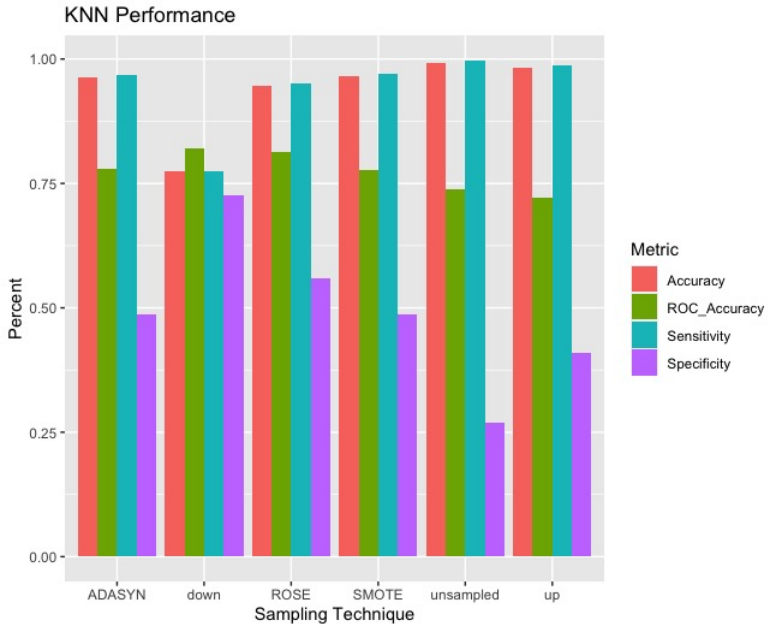

Figure 5.3: Accuracy, ROC Accuracy, Senitivity, and Specificity of knn Models

The visualizations of the confusion matrices provide an intuitive sense as to which models perform better. The models must not have too many false negatives - these being legitimate orders that are predicted to be chargebacks. The models that follow these metrics include the unsampled logistic regression model (Fig 5.4), the unsampled k-nearest neighbor model (Fig 5.6), and all but the ROSE sampled random forest model (Fig 5.5). In looking at the visualizations for the metrics used to evaluate the models, the unsampled logistic regression model outperformed the other logistic regression models in terms of ROC accuracy and accuracy. The unsampled k-nearest neighbor model also outperformed the other knn models with the same metrics. For the random forest models, all but the ROSE and downsampled models performed similarly. The unsampled models outperformed the other sampled models. This is striking to note, as the purpose of sampling the data set is to address the imbalanced nature of it. It is here that it is acknowledged that it seems that in trying to adjust for this imbalance, the models tend to make more false positives. This was also seen in Kjell's thesis, where in the end, he selected the unsampled random forest as the best model. Performance metrics and confusion matrices also suggest that the unsampled random forest model performed the best. Examining the ROC curves between the unsampled random forest model (Fig 5.7) and the upsampled random forest model (Fig 5.8) from the cross-validation resamples would also further this conclusion.

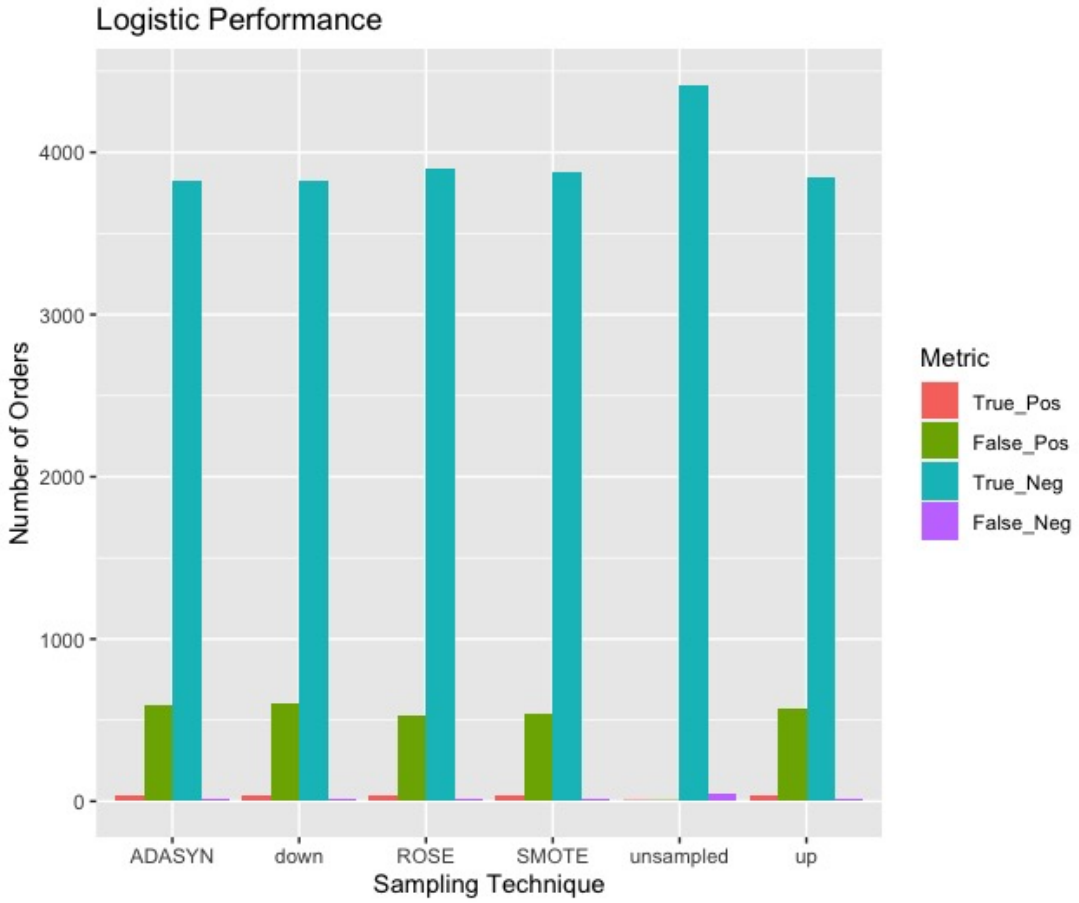

Figure 5.4: Visualization of Confusion Matrix for Logistic Models

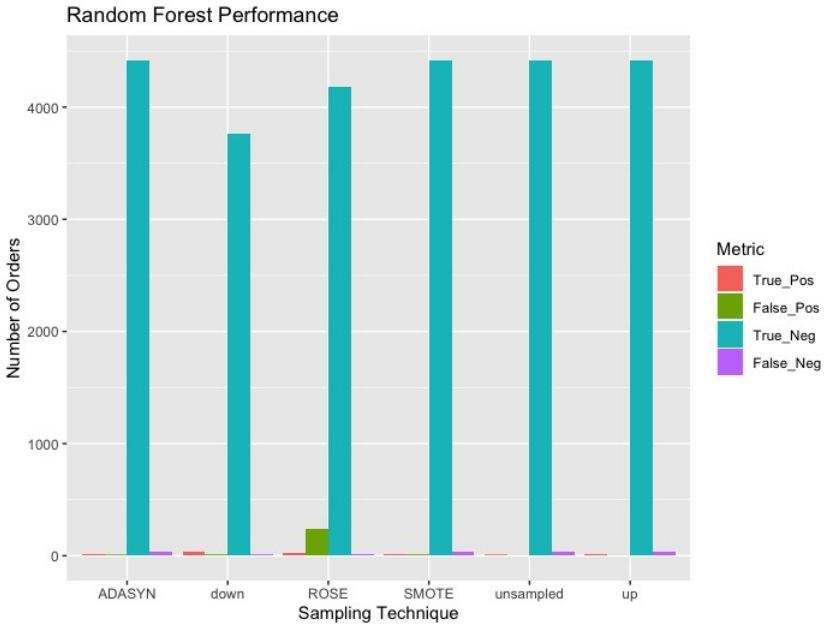

Figure 5.5: Visualization of Confusion Matrix for Random Forest Models

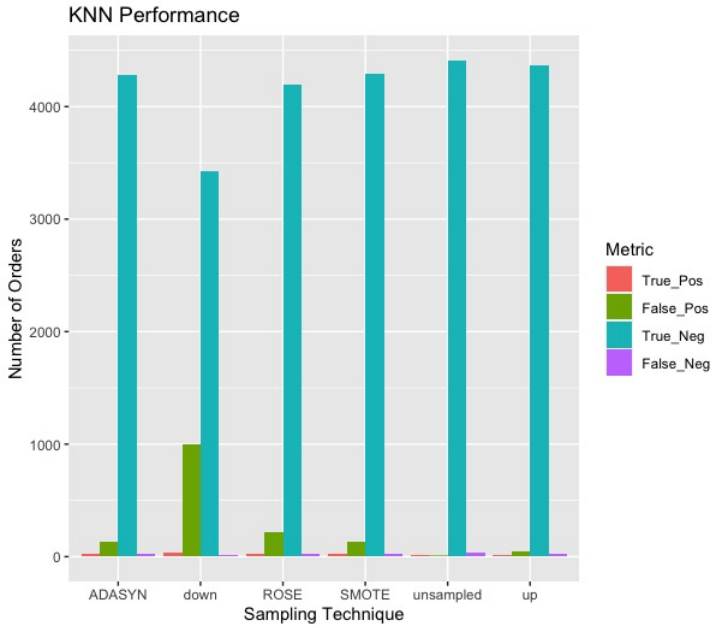

Figure 5.6: Visualization of Confusion Matrix for Random Forest Models

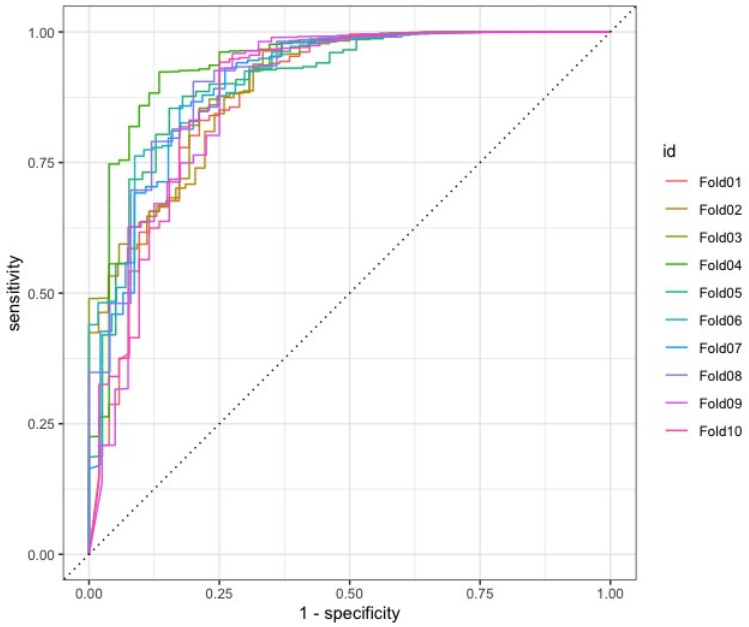

Figure 5.7: ROC Curves for Unsampled Random Forest Model

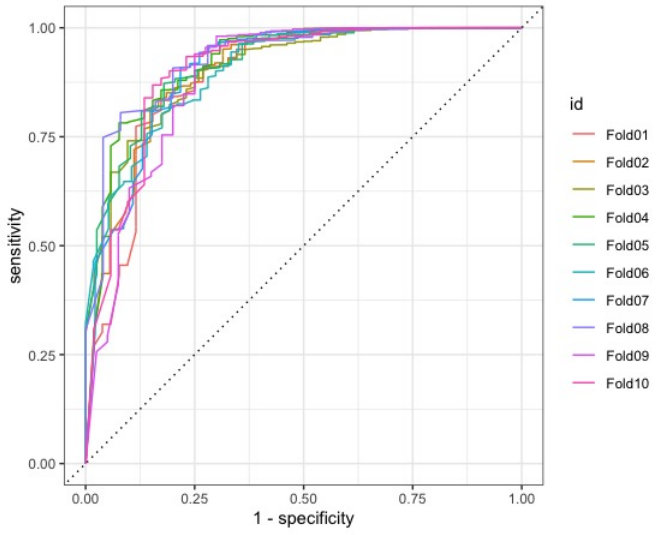

Figure 5.8: Visualization of Confusion Matrix for Random Forest Models

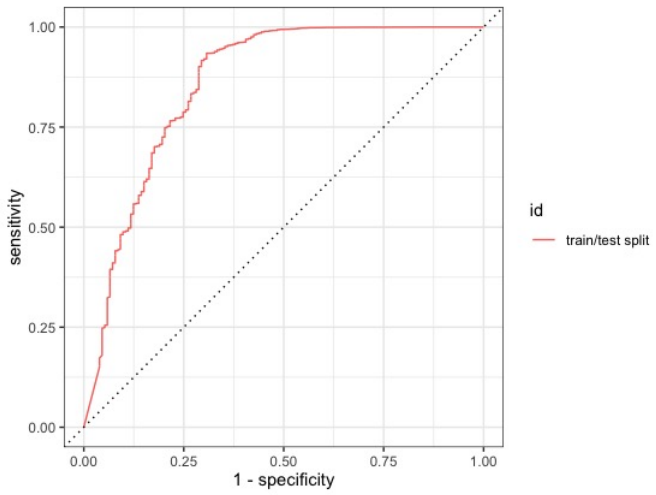

Figure 5.9: ROC Curve of Final Model Fitted on Testing Data

5.2 Final Fit

To assess the final model's performance, the model was used to make predictions on the original testing data. In tidymodels, this is done with the use of the "last_fit()" function that is added to the workflow object of interest. The model is fitted to the testing data and is evaluated. The testing data contained 14,896 orders, of which 153 were chargebacks. The model correctly identified 28 of these orders to be fraudulent, while incorrectly predicting 3 orders to be fraudulent. A ROC curve is provided of this model (Fig 5.9).

CHAPTER 6

Conclusion

6.1 Model Discussion

The Unsampled Random Forest Model's ability to correctly predict legitimate orders is very impressive, and its ability to predict true chargebacks is much to be desired. Only a bit more than 18% of fraudulent orders were detected with this model. It is important to note here that this percentage could have easily reached 80% instead of 18%. For instance, the downsampled random forest model correctly predicted 80% of chargeback orders as such, but because of the restriction to minimize the number of false negatives, the prior model was selected.

What then can be learned from this model? Recent inquiries for manual fraud checks for events susceptible to high chargeback rates have gone up. The process of these manual fraud checks involves examining ticket orders and finding patterns that could suggest possible fraudulent behavior among the purchases. These manual checks are done without the use of machine learning models. By knowing what variables are of importance to the model for these predictions, possible insight can be gained that could influence how future manual checks go.

The unsampled random forest model can accurately predict a subset of the chargebacks without too many false negatives. Feature importance is based on the mean decrease in impurity within each tree of the random forest. The importance of these features can be graphed and examined. Fig 6.1 shows that for this model, outcome risk score, latitude,

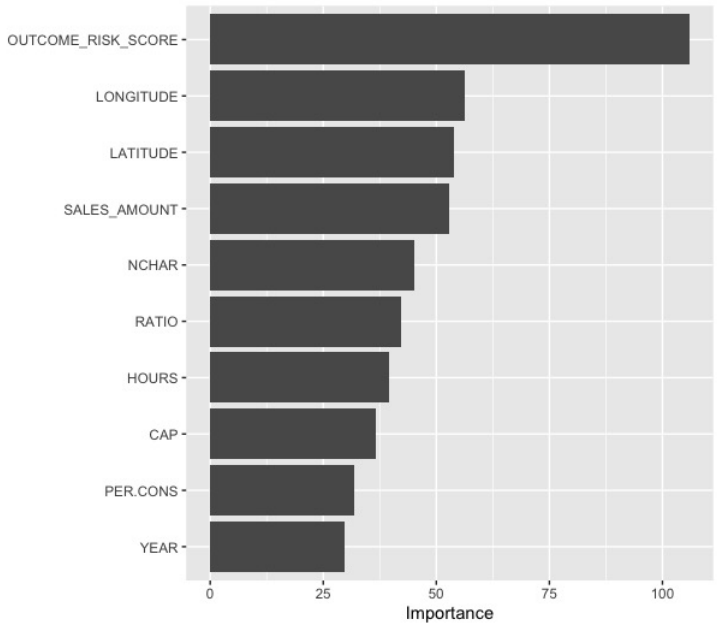

Figure 6.1: Most Important Features of Unsampled Random Forest Models

longitude, and sales amount of the order are significant factors in determining chargebacks. In the manual checks for fraudulent orders, outcome risk scores, latitude, and longitude are indeed factors that are considered, but not as much order sales amount. This suggests that it may be of interest in future manual checks to use sales order amount as an indicator of potential chargeback fraud.

The concern with selecting the unsampled random forest model as opposed to the down-sampled random forest model is that the latter would have too much of a false negative rate. In manual checks these are less of a concern, since doing these checks in a manual nature allow for more of a control on falsely misidentifying chargebacks. Given this, features were of importance can be exmanined in the downsampled random forest model and with caution, consider them as well for future manual checks. With a similar graph, the most important features are outcome risk score, sales amount, hours, and number of capital letters in email of user. This confirms that sales amount may be a significant factor that is not currently being used for manual checks. There may be some merit in checking the impact of when

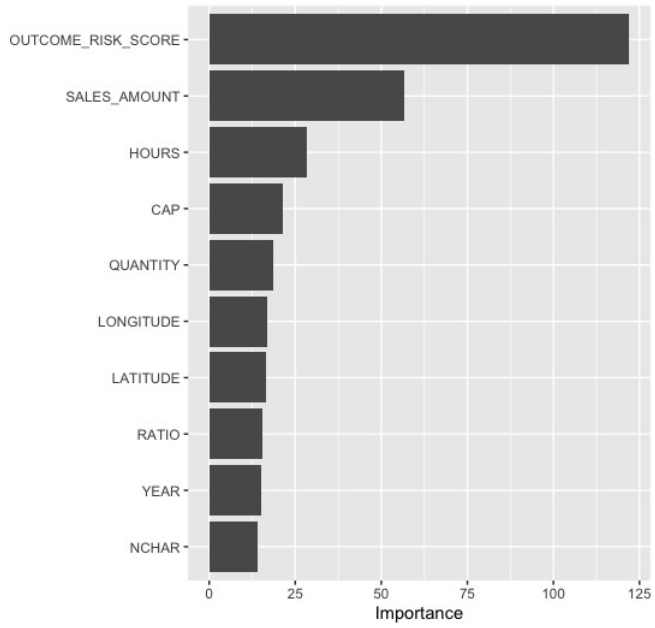

Figure 6.2: Most Important Features of Downsampled Random Forest Models

the order was placed, but again caution should be used. Number of capital letters in a user email is misleading. There could be many different reasons why an email may have capital letters and discerning a fake email from a legitimate one would involve more than just the number of capital letters in it.

6.2 Final Thoughts

In conclusion, the three main purposes of this thesis were accomplished. Firstly, new models were trained with current data utilizing the tidymodels framework. Creation of these models confirm the findings that unsampled random forests seem to perform best in predicting chargeback fraud. Secondly, with the switch of payment processors, the implementation of "outcome risk score" allows for significant aide in identifying and preventing potential chargeback fraud. This can be seen with how "outcome risk score" ranks the highest in terms of importance in both the unsampled random forest model and the downsampled

random forest model. Lastly, updating current strategies for countering chargebacks will include examining factors of interest that may be indicators of potential fraud that were not previously being used. These would include sales amount of the order and the number of hours the purchase occurred before the live event.

The tidymodels framework for implementing models does seem possible for streamlining similar models with other datasets. The use of tidyverse principles allow for an intuitive and straightforward implementation and evaluation of machine learning models. If one is familiar with tidyverse syntax and grammar, these models can be made with ease. Another advantage would be the ability to repeatedly use the same "recipe" objects for pre-processing data sets. One recipe that details how to prepare data for models that can be used for any data set is very convenient. The ability to use tidyverse principles also allow for changes to any workflow, permitting the possibility to iterate through a variety of model designs and parameters. The implementation of tidymodels for future model building seems like it would be a worthwhile investment, but would require further commitment and initiation.

APPENDIX A

R Code for Random Forest Model with Upsampled Data

```
library(rsample)
library(recipes)
library(themis)
library(parsnip)
library(workflows)
library(tune)

#Split the data into training and testing
set.seed(123)
cell_split <- initial_split(df, strata = CHARGEBACK)
cell_train <- training(cell_split)
cell_test  <- testing(cell_split)

#10 fold cross validation
set.seed(234)
cell_folds <- vfold_cv(cell_train)

#lets make the intial recipes
up_sample <- recipe(CHARGEBACK ~ ., data = cell_train) %>%
```

```
  step_upsample(CHARGEBACK)

#lets make some models to test
rf_spec <- rand_forest(trees = 1000) %>%
  set_engine("ranger",importance= "impurity") %>%
  set_mode("classification")

#Workflows
glm_wf_up <- workflow() %>%
  add_recipe(up_sample)

#Lets fit our models: random forests
doParallel::registerDoParallel()
up_rf <- glm_wf_up %>%
  add_model(rf_spec) %>%
  fit_resamples(
    resamples = cell_folds,
    metrics = metric_set(roc_auc, accuracy,
    sensitivity, specificity),
    control = control_resamples(save_pred = TRUE)
  )

#Evaluate Random Forests Model with Metrics
collect_metrics(up_rf)

#Confustion Matrix
up_rf %>%
```

```
  conf_mat_resampled()

#ROC curve
up_rf %>%
  collect_predictions() %>%
  group_by(id) %>%
  roc_curve(CHARGEBACK, .pred_0) %>%
  autoplot()

#Final fit
final_fit <- glm_wf_down %>%
  add_model(rf_spec) %>%
  last_fit(cell_split)

#Important features
final_fit %>%
  extract_fit_parsnip() %>%
  vip(num_features = 10)
```

REFERENCES

[1] Maxwell, Tim. "What Is a Chargeback?" Experian, January 24, 2023. https://www.experian.com/blogs/ask-experian/what-is-a-chargeback/.

[2] Chargeback Gurus. (2023, March 1). Visa Dispute Monitoring and Visa Fraud Monitoring Programs (VDMP & amp; VFMP).

[3] Sawyer, Kjell. (2019, June). Prediction of credit card chargebacks in the live events ticketing industry using machine learning algorithms (thesis).

[4] Sawyer, Kjell. (2023). Speaker Event: Using Machine Learning to Catch Fraud in Live Event Ticketing (DSML Speaker Series). YouTube. Retrieved June 2, 2023, from https://www.youtube.com/watch?v=ZQUU8qOyHT0&ab_channel=Codesmith.

[5] Risk evaluation (2023). Access the Stripe Radar risk evaluations in the Dashboard and the API.

[6] Press, A. P. (2021, November 8). The U.S. lifts the pandemic travel ban and opens the doors to international visitors. NPR. https://www.npr.org/2021/11/08/1053434232/the-u-s-lifts-the-pandemic-travel-ban-and-opens-the-doors-to-international-visit

[7] Rhys, H.I. (2020). Machine Learning with R, the tidyverse, and mlr. Manning Publications. https://books.google.com/books?id=BoeryQEACAAJ

[8] Johnson, Kjell. and Kuhn, Max. (2019, June 21). Feature Engineering and Selection: A Practical Approach for Predictive Models. Taylor & Francis Group, LLC.

[9] Menardi, Giovanna. Torelli, Nicola. (2012, October 30). Training and assessing classification rules with imbalanced data. Data Min Knowl Disc 28, 92–122

[10] Chawla, N.V. (2002, June 1). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research.

[11] Hvitfeldt Emil. (2023, April 15). Extra Recipes Steps for Dealing with Unbalanced Data. CRAN. https://github.com/tidymodels/themis

[12] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[13] Kuhn, Max. (2023, April 11). A Common API to Modeling and Analysis Functions. CRAN. https://github.com/tidymodels/parsnip