Credit Card Marketing Study

Advised by: Dr. Kurt Cogswell

Senior Capstone - Data Science

Department of Mathematics and Statistics

South Dakota State University

Brianna Humphries

Fall 2020

Abstract

To understand why customers accept or refuse credit card offers, a bank designed a marketing study with current bank customers who had been offered a credit card, some of whom accepted an offer and some of whom rejected it. The survey asked for input on various demographic factors, which was combined with other marketing and financial data, the result being an 18,000 customer record data set.

The data set will be used to compare demographics against the variable in which the customer accepted or declined the credit card to determine different classification models. These models will show which variables are important when determining which customers are more likely to accept a credit card. The models will be compared based on the criteria of accuracy and explainability in order to produce an overall model that will predict who should be a targeted credit card customer.

CREDIT CARD MARKETING STUDY

1 Summary of Data Set

To understand why customers accept or refuse credit card offers, a bank designed a marketing study with current bank customers who had been offered a credit card, some of whom accepted an offer and some of whom rejected it. The survey asked for input on various demographic factors, which was combined with other marketing and financial data, the result being an 18,000 customer record data set.

The data set includes 16 different variables for each customer. These variables are:

- Offer Accepted (Yes/No)
- Reward plan (Air Miles/Cash Back/Points)
- Mailer Type (Letter/Postcard)
- Income Level (High/Medium/Low)
- Number of Bank Accounts Open
- Overdraft Protection (Yes/No)
- Credit Rating Level (High/Medium/Low)
- Number of Credit Cards Held
- Number of Homes Owned
- Household Size
- Home Ownership (Yes/No)

- Average Bank Balance (Across all accounts over time)
- Q1, Q2, Q3, and Q4 Balances (In the last year)

1.1 Data Preparation

The data set was imported, manipulated, and cleaned using RStudio software. All the categorical variables were already clean, meaning the variables did not include any duplicated or missing values. The continuous balance variables contained 24 missing observations. Since the data set has 18,000 observations, removing the missing values would most likely not affect the analysis of the data. Therefore, the 24 customers with missing balance data were removed from the data set. The final, clean data set includes 17,976 observations.

1.2 Variable Description

Before the relationships between the variables are compared, the distributions of each variable should be examined.

The overall focus of the study is the variable of credit card offer acceptance, where customers either accepted the card or did not. The distribution of this variable is shown in Figure 1. Note that each figure in this section is created using R. The study shows a large difference in the number of customers that declined the credit card versus the number of customers that accepted it, shown in Figure 1. Approximately 94.3%, of the customers in the study declined the credit card offer from the bank while only 5.7% of customers accepted the offer.

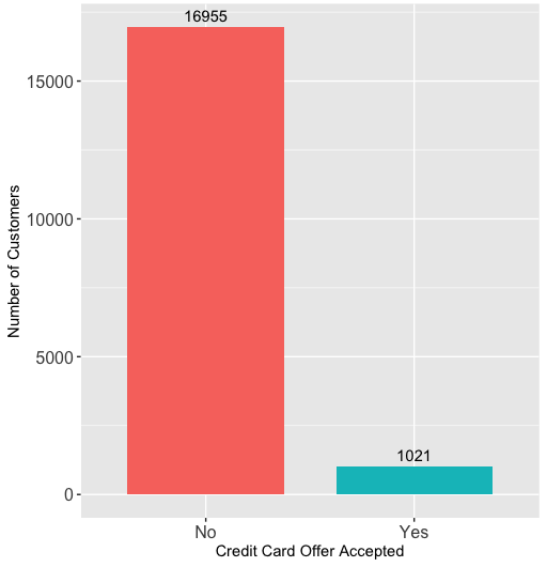

Figure 1: Bar graph of number of customers that accepted credit cards

Each credit card included a special reward type in the offer. There were 3 different reward types offered: Air Miles, Cash Back, and Points. The number of customers offered each reward type is shown in Figure 2. The distribution graph in Figure 2 shows each reward type was equally distributed among the customers.

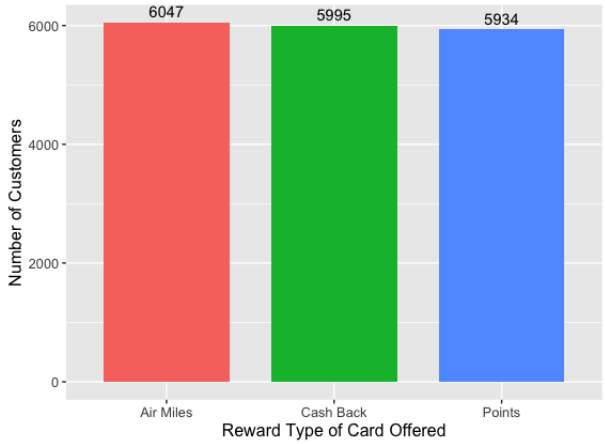

Figure 2: Bar graph of credit card reward type in offered card

The study was sent to the customers by either Letter or Postcard. The number of customers who received a letter or a postcard is shown in Figure 3. Figure 3 shows that mailer typers were also equally distributed among the customers.

Figure 3: Bar graph of number of customers with each mail type

The customer's income is represented in the data set by factors Low, Medium, and High. The number of customers at each income level is shown in Figure 4. The distribution of income level in Figure 4 show that more customers, about 50.1% , had a medium level income. The percent of customers with a low income is 24.8%, which is similar to the percent of customers with a high income at 25.1%.

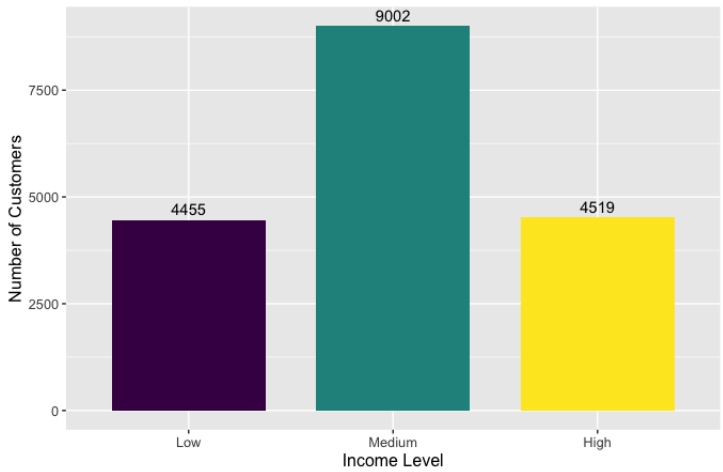

Figure 4: Bar graph of income level for each customer

The data set includes how many non-credit card bank accounts are held by each customer. The distribution of the variable is shown in Figure 5. Figure 5 shows that all the customers have 1-3 bank accounts, with the majority of customers, about 76.1%, having only one bank account.

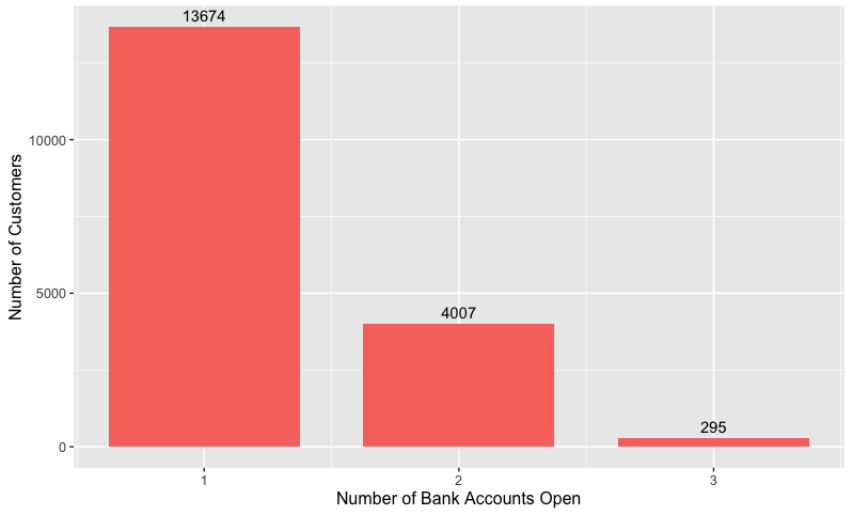

Figure 5: Bar graph of number of bank accounts each customer has open

Banks offer customers overdraft protection which provides customers an alternate payment method when their checking account does not have enough funds for a transaction. The data set contains whether a customer has overdraft protection on their checking accounts, shown in Figure 6. It appears that most of the customers, about 85.1% do not have overdraft protection on their checking account.

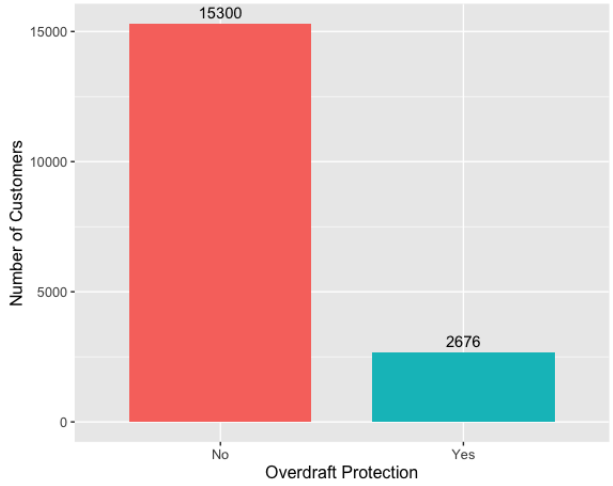

Figure 6: Bar graph of number of customers that have overdraft protection

The credit rating for each customer in the data is provided at factor levels Low, Medium, and High. The distributions of all the customers' credit ratings is shown in Figure 7. The percent of

customers in each credit level are relatively similar, around 33%.

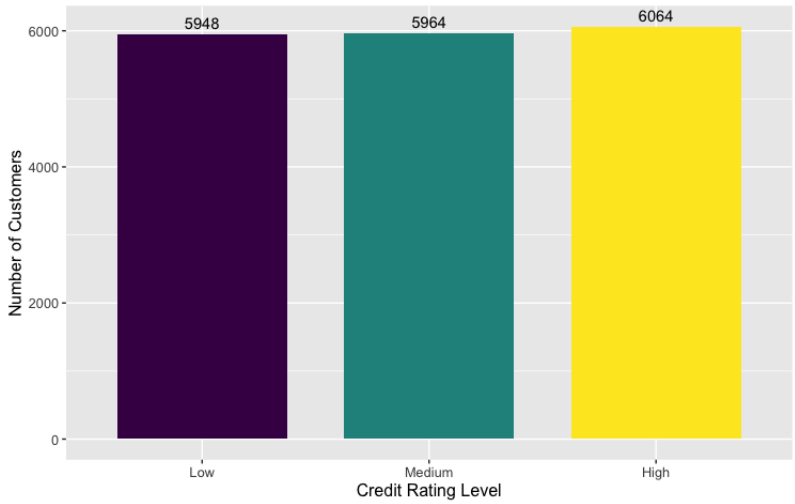

Figure 7: Bar graph of credit rating level for each customer

The data set provides the number of credit cards each customer holds at the bank, and Figure 8 shows how many customers have each number of credit cards. It can be seen from Figure 8 that each customer in the study has 1-4 credit cards held at the bank. About 44.1% of customers already held 2 credit cards at the bank, while only 2.9% of customers have 4 credit cards held at the bank.

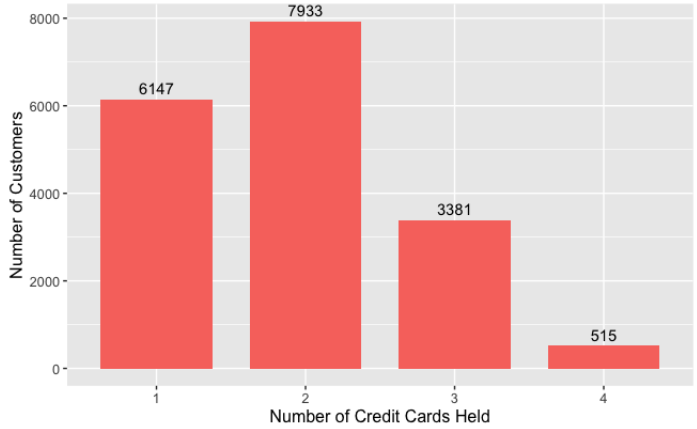

Figure 8: Bar graph of number of credit cards each customer has

The study asked each customer how many homes they own, which is shown in Figure 9. The graph shows that most of the customers in the study, about 80.7%, own one home. The

maximum number of homes owned by each customer is 3, which only accounts for about 1.0% of the customers.

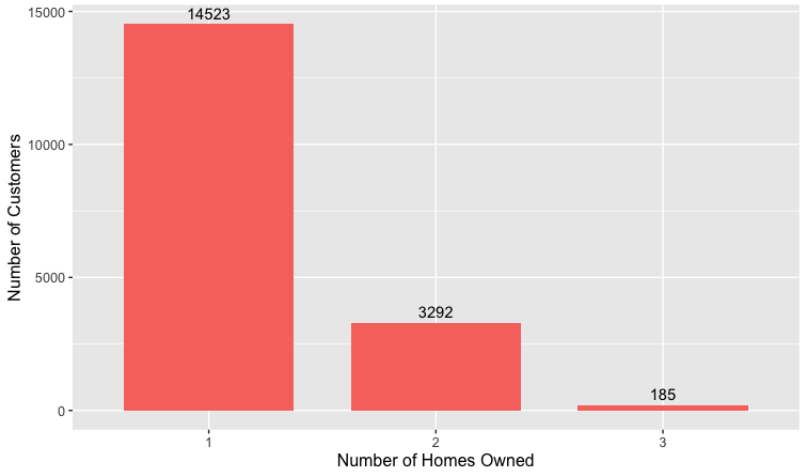

Figure 9: Bar graph of number of homes each customer owns

The household size of each customer is defined as the number of individuals in the family. The distribution is shown in Figure 10. Each customer shows to have a household size from 1-9 people, with only 1 customer having a household size of 8 and 9 people. While ignoring the household sizes of 8 and 9, the distribution appears symmetric across household sizes of 3 and 4 people.

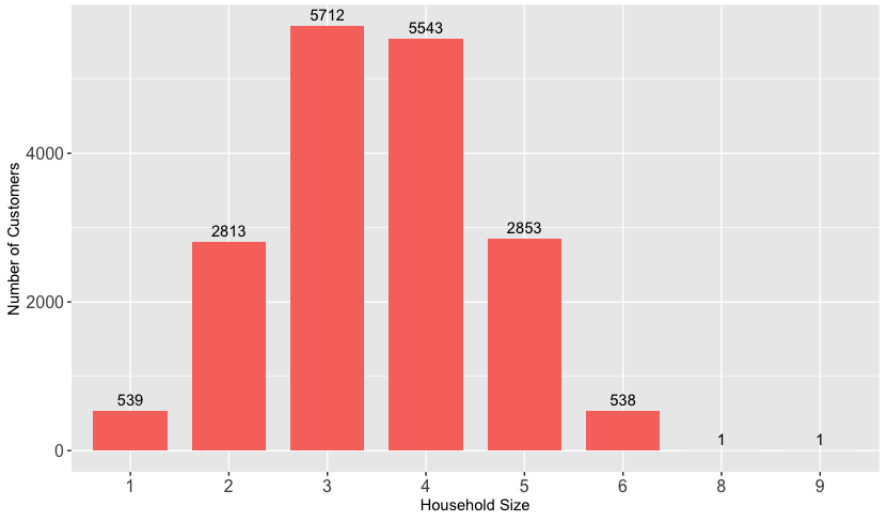

Figure 10: Bar graph of Household Size

Each customer was asked in the study if they own their own home, with answers "Yes" or

"No". The number of customers that own their home or do not own their home is shown in Figure 11. The graph shows that more customers, about 64.7%, own their own home,

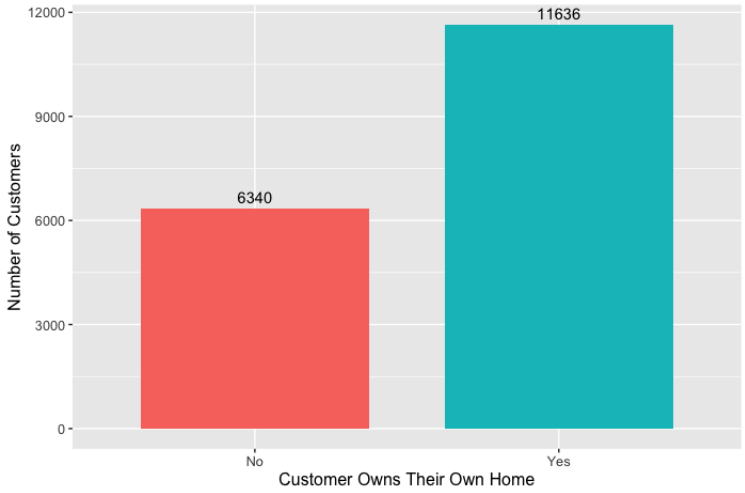

Figure 11: Bar graph of whether customer owns their own home

The average balance of every customer in the data represents the customer's average account balance across all of their bank accounts over time. The histogram in Figure 12 shows the distribution of average balances for all customers. The graph shows little to no data around $2,000 and after due to outliers. Upon further investigation, there are only 10 customers with an average balance above $2,000. The histogram also shows a large decrease at $500, meaning there are little to no customers with an average balance around $500. The graph maximizes around a balance of $1,000 with the number of customers at around 1,750.

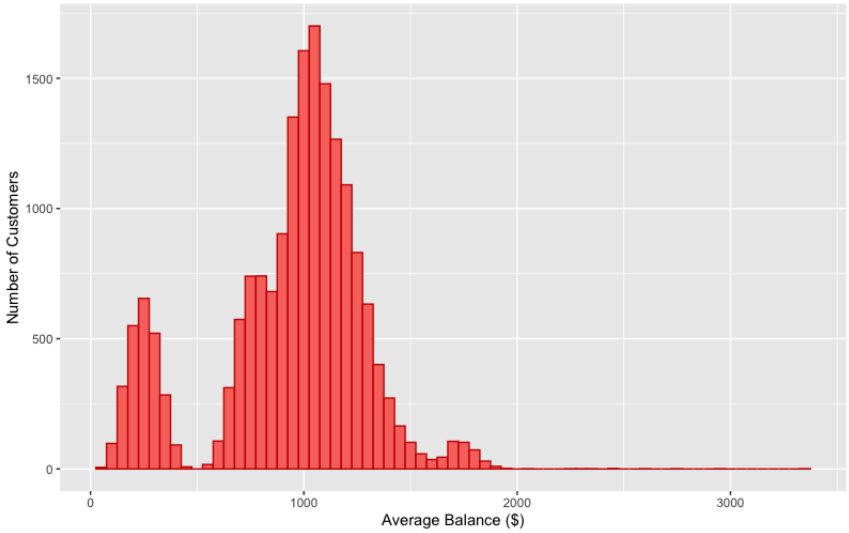

Figure 12: Histograms of overall average balance for all customers

The overall distribution of each customer's average balance in Figure 12 exhibits two separate distribution curves. Therefore, it can be assumed that there are two sub-populations of the data. The first sub-population includes customers with an average balance below $500, and the second sub-population includes customers with an average balance above $500.

The data set also provides the average balances for each quarter in the last year. The months respective to each quarter are as follows:

- Quarter 1 - January, February, March
- Quarter 2 - April, May, June
- Quarter 3 - July, August, September
- Quarter 4 - October, November, December.

The histograms for each quarter are all provided in Figure 13. Similar to the overall average balance for each customer, all the histograms cut off around $2,000 - $2,300. However, the quarter 1 balance also appears to have a bimodal distribution, where the first sub-population

includes customers with a quarter 1 balance below $1,150, and the second sub-population includes customers with a quarter 1 balance above $1,150.

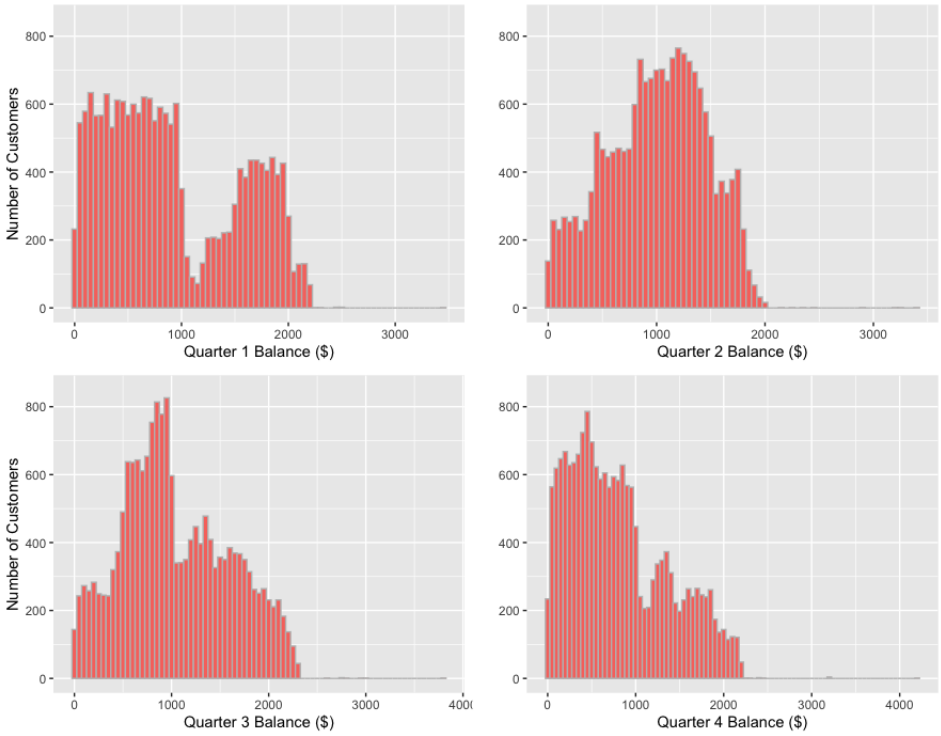

Figure 13: Histograms of each quarterly balance for all customers

1.3 Data Exploration

Certain features in the variable distributions suggest that the bank implemented a controlled study. First, the variables including each customer's reward type, mailer type, and credit rating all display equal factor levels in their distributions. This alone implies that the bank deliberately issued each credit card offer equally among equal types of customers.

Another feature of the data is if each factor level of any categorical variable is subsetted into its own population, then the distributions of the other variables still remain the same. For example, recall that the two factor levels of the mailer type variable are letter and postcard. If a sub-population of the customers who received a letter is created and distributed among the other

variables, then the distributions will remain the same as the overall distribution with only slight variance. The same idea also applies to the continuous variables in the data. Recall that the distribution graph of the customers' average balances exhibits two separate populations, where the change in populations occurs at $500. If the population consisting of below $500 is distributed among all the variables, the distributions of all the graphs portray a similar pattern to those of the entire data set's population. Thus, we can assume the bank carefully chose which customers to target based on all the variables in the study.

The purpose of this study is to understand why customers accept or refuse credit card offers. In order to find the relationship, we must answer three questions:

1. Which variables are significant to why customers accept or refuse credit card offers?

2. Which classification model of credit card offer acceptance is the most effective?

3. Is it possible to create a predictive model of offer acceptance?

A thorough analysis of each question will reveal which customers are more inclined to accept credit card offers.

2 Which variables are significant to why customers accept or refuse credit card offers?

In order to understand why customers accept or refuse credit card offers, we must first establish which customer demographics have the most significant relationship with credit card acceptance. Each demographic will be graphed with offer acceptance in R to visualize their relationship.

The relationship between reward type and offer acceptance is shown in Figure 14. Recall that the distribution of reward type was similar for each factor level. Figure 14 shows that more

customers, about 45.3%, accepted the credit card with an Air Miles reward. The Cash Back reward type had the least amount of customers accept the offer, with only about 20.2% of the customers that accepted the card.

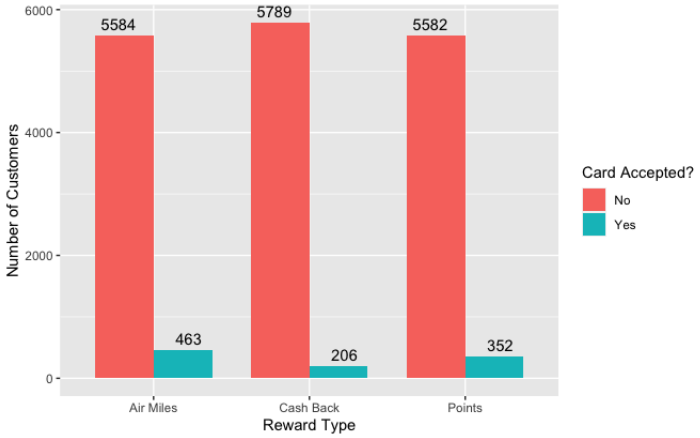

Figure 14: Reward Type vs. Offer Accepted

The Pearson's chi-squared test determines whether two categorical variables are independent or dependent of each other. The null hypothesis of the test is that the two variables are independent, while the alternate hypothesis is that the two variables are dependent. The null hypothesis is rejected if the p-value calculated from the test is less than the significance level of 0.05.

The p-value found in R from the Pearson's chi-squared test between variables reward type and offer accepted is less than 2.2e-16. Since the p-value is less than the significance level of 0.05, then the null hypothesis is rejected. Therefore, reward type and offer acceptance are dependent, indicating that reward type affects offer acceptance.

The relationship between mailer type and offer acceptance is displayed in Figure 15. Recollect that the distribution of mailer type is similar for both factors, with the Postcard factor having slightly more customers. About 7.89% of customers offered a card by postcard accepted the offer, while only 3.39% of customers offered by letter accepted the offer. While 50.8% of customers were offered a card through a postcard, about 70.6% of all the customers that accepted the offer were

mailed a Postcard. This indicates that there is a relationship between Reward Type and the Offer Accepted variable.

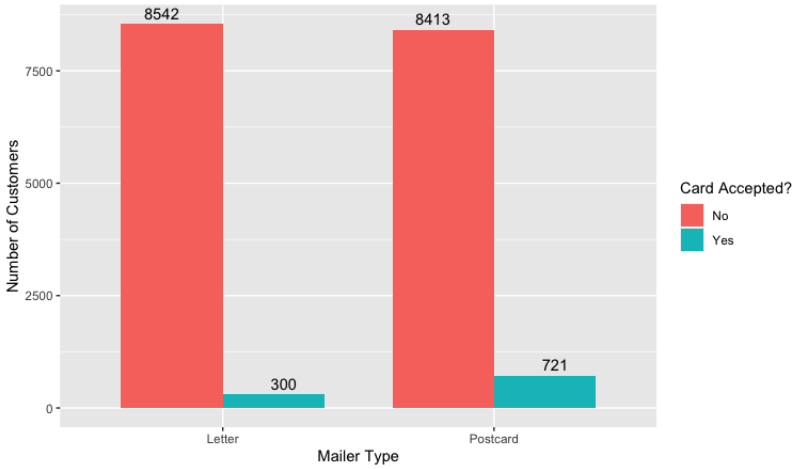

Figure 15: Mailer Type vs. Offer Accepted

The p-value calculated in R with the Pearson's chi-squared test is less than 2.2e-16 for the relationship between mailer type and offer accepted. Since the p-value is less than 0.05, then offer acceptance is dependent of mailer type. Therefore, the mailer type of the offer affects the customer's offer acceptance.

Figure 16 illustrates the relationship between income level and offer acceptance. It appears that 45.9% of customers that accepted the credit card offer have a medium income level. The distribution where the offer was accepted should appear symmetrical if there is no relationship between the variables, however, about twice the number of High income level customers that accepted the card have a low level income. Therefore the variables appear to have a relationship.

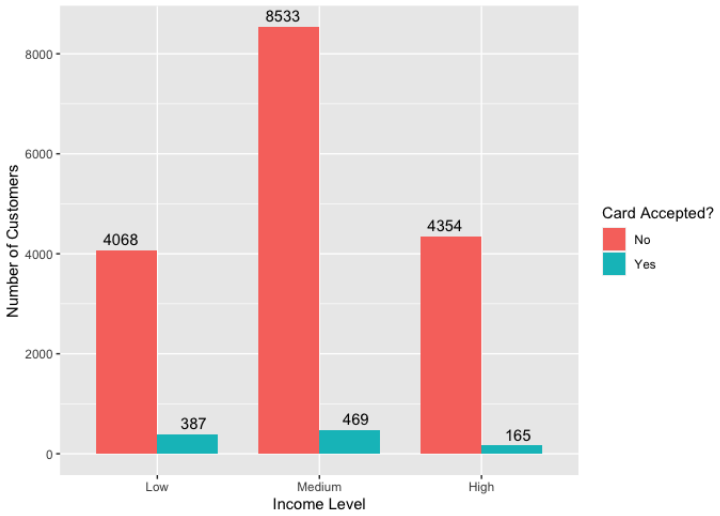

Figure 16: Income Level vs. Offer Accepted

The Pearson's chi-squared test for the relationship between income level and offer acceptance calculated a p-value of less than 2.2e-16. Therefore the offer acceptance is dependent on income level.

The relationship between the number of bank accounts each customer has and offer acceptance is shown in Figure 17. It appears that 75.9% of accepting customers have one bank account open, 22.9% of accepting customers have two bank accounts open, and only 1.2% of accepting customers have three bank accounts open. The distribution of accepting customers is similar to the overall distribution of the number of bank accounts open, therefore the variables do not appear to have a relationship.

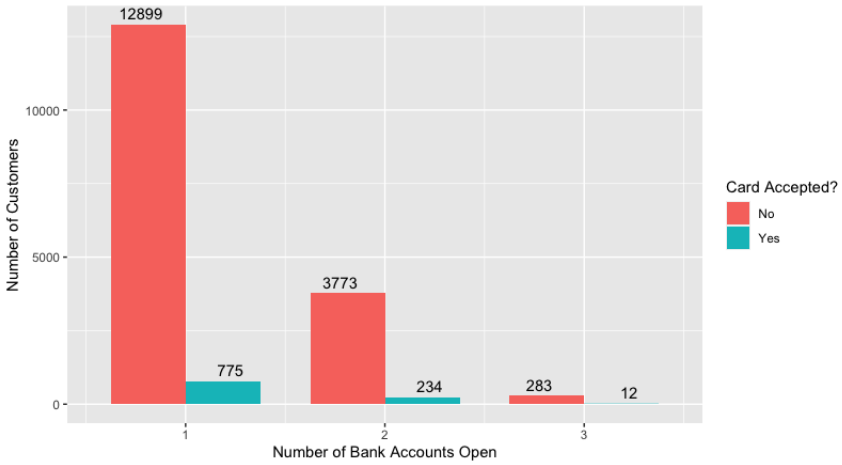

Figure 17: Number of Bank Accounts Open vs. Offer Accepted

The Pearson's chi-squared test evaluated a p-value equal to 0.44. Since the p-value is greater than 0.05, then the null hypothesis that the two variables are independent is not rejected. Therefore, offer acceptance and the number of bank accounts of each customer are independent, and offer acceptance is not affected by the number of bank accounts a customer has open.

Figure 18 shows the relationship between overdraft protection and offer acceptance. The graph shows that about 85.5% of accepting customers do not have overdraft protection, while 14.5% do. Since the distributions in Figure 18 are similar to the overall distribution of overdraft protection in Figure 6, then the variables do not appear to be associated.

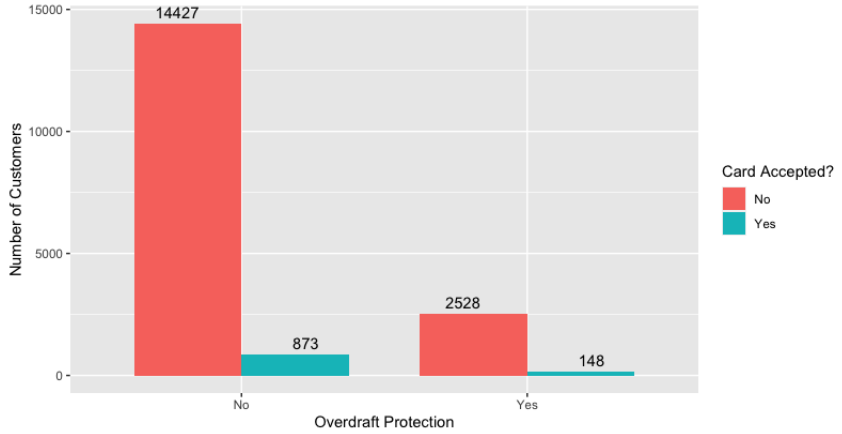

Figure 18: Overdraft Protection vs. Offer Accepted

The p-value from the Pearson's chi-squared test equals 0.752, which is greater than 0.05. Therefore the null hypothesis is not rejected, and the two variables are concluded to be independent and not related.

The association between credit rating level and offer acceptance is depicted in Figure 19. Recall that the distribution of credit rating level is similar for all levels. Figure 19 shows there is a slight relationship between the two variables, where the number of accepting customers decreases as the credit level increases. Specifically, 62.0% of accepting customers have a low credit level, while only 11.7% have a high credit level.

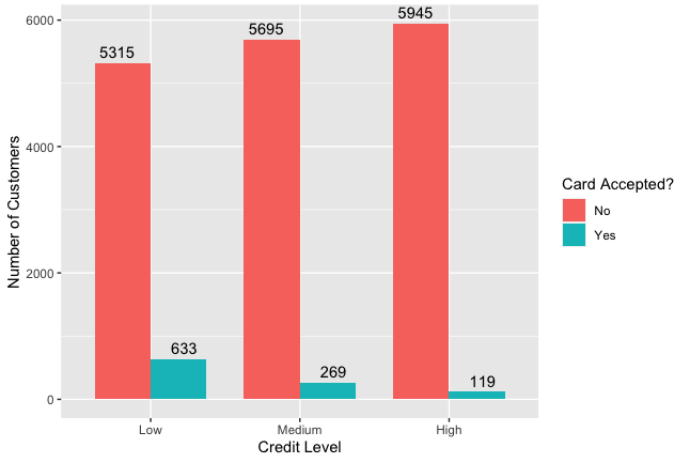

Figure 19: Credit Rating Level vs. Offer Accepted

The Pearson's chi-squared test between the variables has a p-value less than 2.2e-16. Therefore, credit level and offer acceptance are dependent, and credit rating affects offer acceptance.

The relationship between the number of cards each customer holds and offer acceptance is shown in Figure 20. The figure shows that the the customers who have two cards accepted a new card more often, which is also about 42.4% of accepting customers. The distribution of accepting customers in the graph is similar to the distribution of the number of cards held, therefore there does not appear to be a relationship between offer acceptance and the number of cards held.

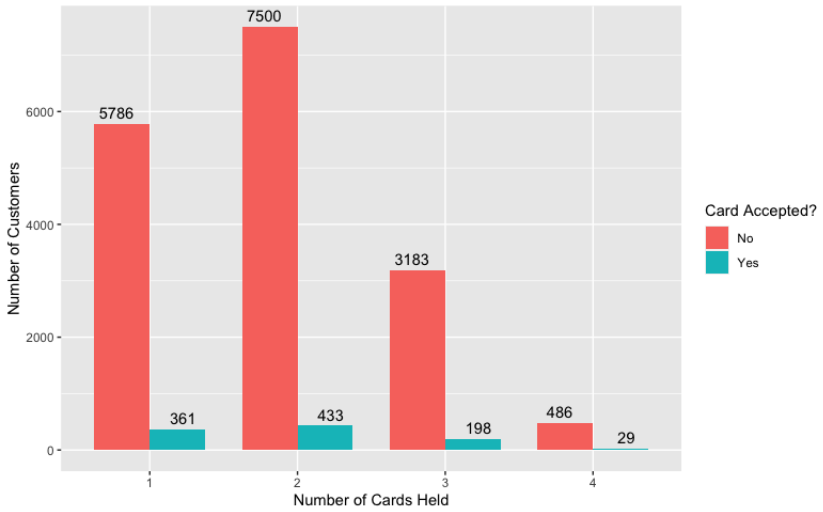

Figure 20: Number of Cards Held vs. Offer Accepted

The p-value from the Pearson's chi-squared test between offer acceptance and the number of cards held equals 0.7195. Since the p-value is greater than 0.05, then the variables are independent.

The relationship between number of homes each customer owns and offer acceptance is shown in Figure 21. It appears that 80.6% of accepting customers own one home, and the number of accepting customers decreases as the number of homes increases. However, since the overall distribution of the number of homes each customer owns is similar, then there is not an association between offer acceptance and number of homes owned.

The Pearson's chi-squared test between the two variables calculated a p-value of 0.8793. Therefore the offer acceptance and the number of homes owned are independent and not associated.

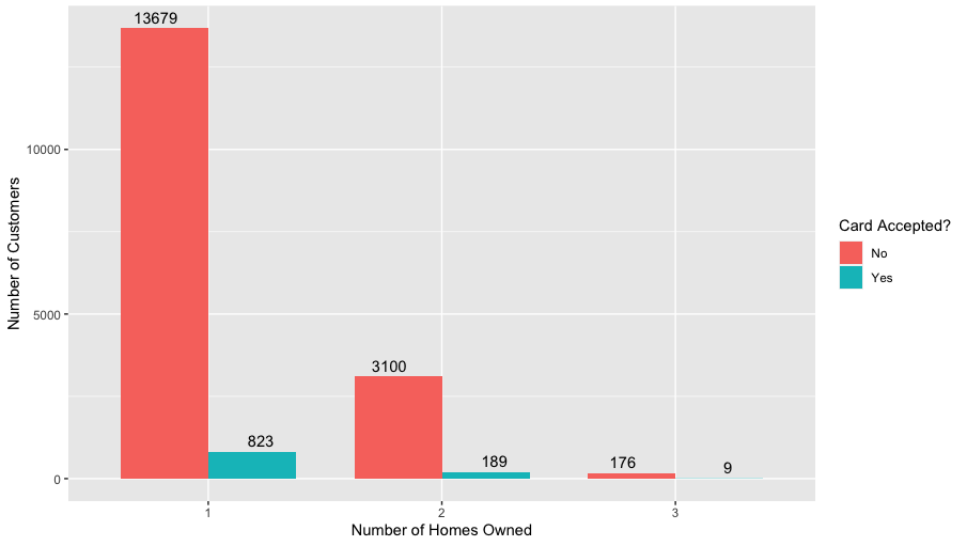

Figure 21: Number of Homes Owned vs. Offer Accepted

The relationship between household size and offer acceptance is shown in Figure 22. The graph shows that the distributions are relatively similar to the distributions of all the customers' household sizes. It appears that there are not any customers that have 7 members in their household, and only one customer has 8 or 9 household members.

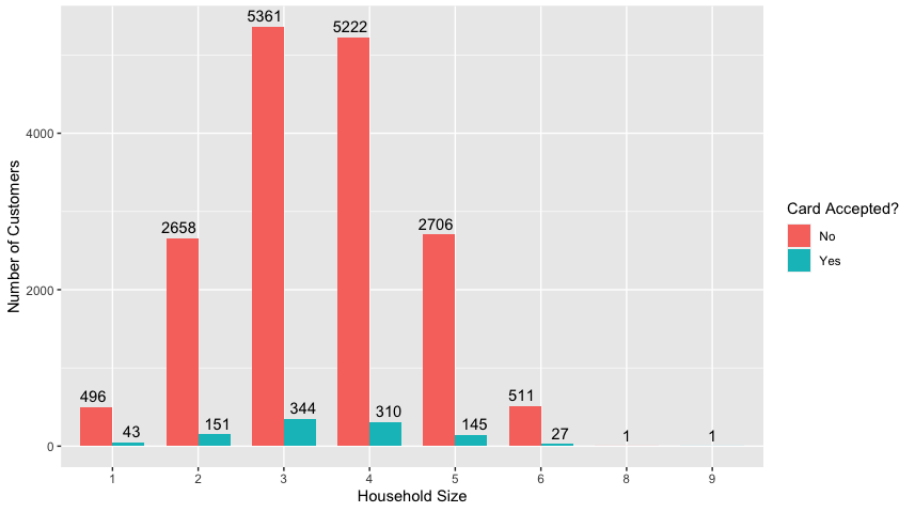

Figure 22: Household Size vs. Offer Accepted

The Pearson's chi-squared test for this relationship has a p-value of 0.0004745. Since the p-value is less than 0.05, then the variables are technically dependent of each other. However, if

household sizes 8 and 9 are considered outliers and removed, then the p-value of the chi-squared test becomes 0.0912. The value is now greater than 0.05, therefore household size is not significant to offer acceptance.

The relationship between offer acceptance and whether the customer owns their own home is shown in Figure 23. It appears that 64.7% of accepting customers own their own home, which is equal to the overall percentage, 64.7%, of customers that own their own home. Therefore, since the distributions did not change with the addition of offer acceptance, then the figure shows that there is not a relationship between offer acceptance and a customer that owns their own home.

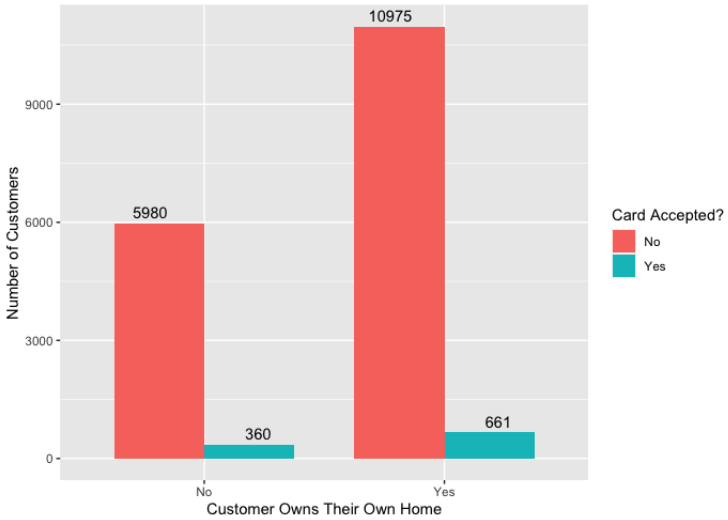

Figure 23: Home Owner vs. Offer Accepted

The Pearson's chi-squared test calculated p-value of 1, therefore the variables of offer acceptance and home ownership are independent. Note that since the percent of customers that own their own home equals the percent of accepting customers that own their own home, then the p-value is 1.

The next variable relationship is between average balance and offer acceptance. Since average balance is a continuous variable and offer acceptance is a categorical variable, then the relationship is graphed with a box plot, shown in Figure 24. The box plot of the accepting customers has an

average balance median of $1008.50 while the non-accepting customer box plot has a median of $1007.00. It also appears that the interquartile range, maximum, and minimum of both box plots are similar. Therefore, the box plot shows that average balance does not appear to affect offer acceptance.

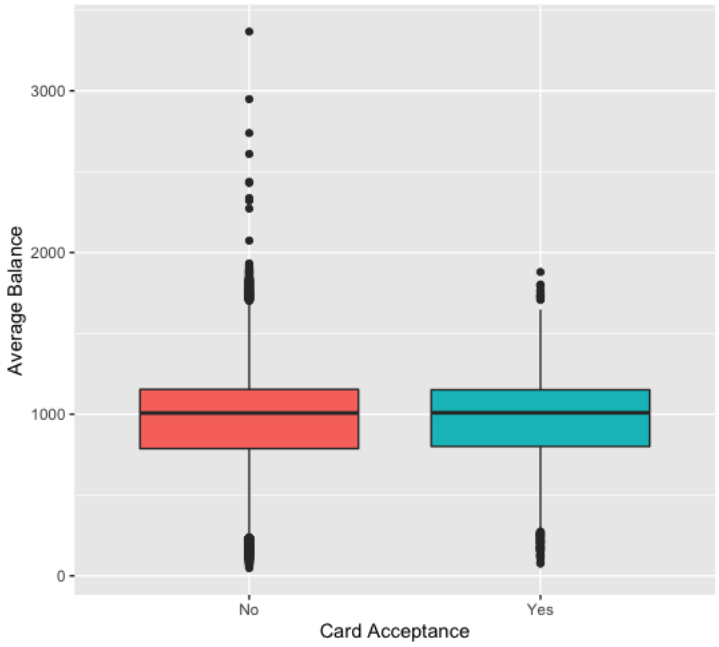

Figure 24: Average Balance vs. Offer Accepted

Since card acceptance is a dichotomous, or binary, variable, then the relationship with average balance can be tested by point-biserial correlation. Note that before the test was computed, the levels of offer acceptance were refactored to levels 0 and 1, since the test requires a numerical binary variable. After computing the test in R, the p-value is 0.6031. Since the p-value is greater than the significance level of 0.05, then we fail to reject the null hypothesis that the correlation coefficient is equal to 0. However, a requirement of the point-biserial correlation is that Average Balance must be normally distributed. Recall from Figure 12 that the distribution of average balance is bimodal and not normally distributed, which can affect the accuracy of the point-biserial test. To avoid this, the average balance was switched to a categorical variable with

factors Low and High. The low level represents the customers in the lower subpopulation of the bimodal distribution with average balances up to $500, and the high level represents the customers in the higher subpopulation with average balances above $500. Since average balance is now a categorical variable, then the Pearson's chi-squared test is used to test the significance. The test calculated a new p-value of 0.5946, which is a difference of only 0.0085 from the point-biserial test's p-value. Therefore, average balance remains to not be significant to offer acceptance.

The relationship between offer acceptance and each quarterly balance is shown in Figure 25. Similar to the relationship with average balance, the medians for each quarter's balances have minimal change between accepting and non-accepting customers. Therefore, at first glance, all of the quarterly balances appear to not have a relationship with offer acceptance.

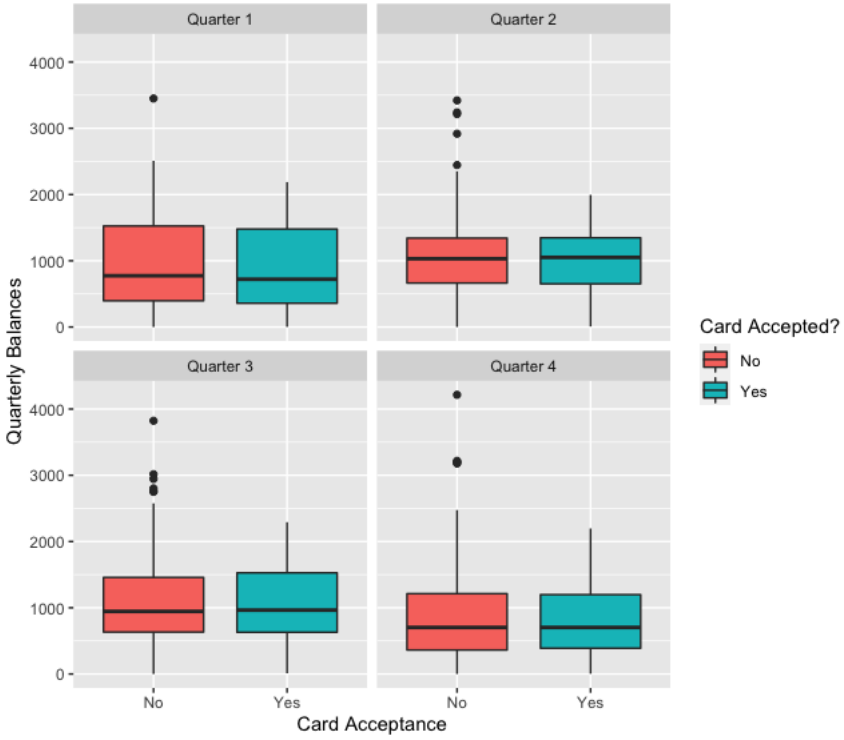

Figure 25: Quarterly Balances vs. Offer Accepted

The point-biserial test's p-value for each quarterly balance is shown below in Table 1. The only p-value in the table less than 0.05 is 0.0267 for the relationship of quarter 1 balance and offer

accepted, although the p-value is close to 0.05.

Quarter	P-Value
Quarter 1	0.0267
Quarter 2	0.9668
Quarter 3	0.2613
Quarter 4	0.9380

Table 1: P-Value of Point-Biserial Test of Each Quarter

However, similar to average balance, recall that the quarter 1 balance distribution appears bimodal, therefore the quarter 1 balance variable was converted to a categorical variable with factor levels Low and High. The low level represents the customers in the lower subpopulation of the bimodal distribution with quarter 1 balances up to $1,150, and the high level represents the customers in the higher subpopulation with quarter 1 balances above $500. The p-value now after calculating the Pearson chi-squared test is 0.1500, which indicates that quarter 1 balance is not significant to offer acceptance.Therefore every quarter balance variable does not have a statistically significant relationship with offer acceptance.

Each variable's relationship with offer acceptance has now been analyzed. Therefore whether a customer accepts a card is dependent on the variables reward type, mailer type, credit rating level, and income level. Note that reward type and mailer type are the only two variables in the study about the offer and not the customer. Then the only customer demographic variables that affect offer acceptance are income level and credit rating level.

3 Which is the most effective classification model of credit card offer acceptance?

In order to determine why the customers declined or accepted the credit card, the relationships between the demographics can be used to build classification models. Some classification models,

such as logistic regression, decision tree, and K-nearest neighbors, were examined in JMP. However, the R-squared values for each of these models all had a maximum at about 0.1, even with different variations of variables. Each model was also quickly reviewed for predictability using the confusion matrix, and the models either never or rarely ever predicted a customer to accept a credit card. Since only 5.7% of customers in the study actually accepted a credit card, the data is imbalanced, and it is difficult to build a model that will create accurate predictions.

However, there are methods to deal with imbalanced data. The techniques we will focus on are re-sampling methods called over-sampling, under-sampling, and a mixture of the two. Before each model is created, the data set was split into two datasets: training data and testing data. The training data includes of 70% of the data, while the testing data includes the remaining 30% of the data. Therefore the training data includes 12,583 observations and the testing data includes 5,393 observations. The training data will be used to create the models in JMP, and the models will be used on the testing data in Section 4 to determine if the models are accurate in predictions.

3.1 Re-Sampling Methods

The over-sampling technique involves duplicating the observations in which a customer accepted a credit card to make it balanced with the number of customers that did not accept a credit card. Since there are 11,877 non-accepting customers in the training data set, then the over-sampled data set duplicates the accepting customers to 11,877 customers, as well. Therefore, there are 23,766 observations in the over-sampled data set. However, the issue with this technique is that since the accepting customers are duplicated from 706 observations to 11,877 observations, the model created could be too over fitting. Therefore this re-sampling method would not execute well with the data and will not be used.

The under-sampling technique involves removing some of the observations in which a customer did not accept a credit card to make it balanced with the number of customers that did accept a credit card. Since there are 706 accepting customers in the data set, then the under-sampled data includes 706 non-accepting customers. However, the issue with this technique is that we are removing thousands of observations, which can cause a significant loss of data that could be essential in good predictive model. Therefore, this re-sampling method will not be used with the data.

A mixture of both over-sampling and under-sampling is also possible, which will be referred to as "both-sampling". This technique duplicates the accepting customer observations, but it also removes various non-accepting customer observations so that the two factors balance at a point between them. Recall that in the training data set, there are 11,877 non-accepting customers and 706 accepting customers, with a difference of 11,171 customers. The both-sampling method will be used to create models at three different combinations of over-sampling and under-sampling based on a quarter fraction of the difference between the accepting and non-accepting customers, shown in Figure 27.

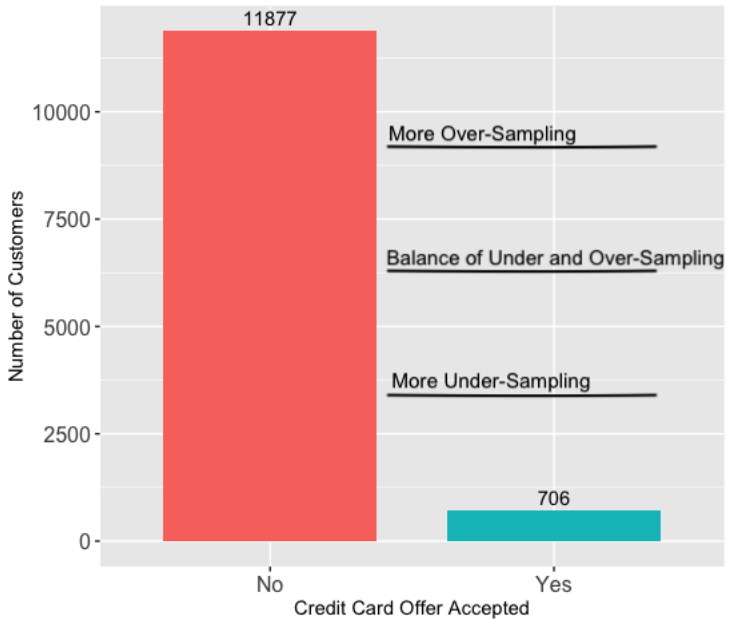

Figure 26: Variations of Both-Sampling Methods used on Training Data

The different levels in Figure 27 represent where the accepting and non-accepting customer bars will balance. The lowest level uses more under-sampling since the number of removed non-accepting customers is more than the number of duplicated accepting customers. The middle level represents where the amount of over-sampling and under-sampling is relatively even, since it balances right in the middle of the two bars in Figure 27. The highest level uses more over-sampling since the number of duplicated accepted customers is more than the number of removed non-accepting customers. The two models that we will focus on with the three variations of both-sampling are K-Nearest Neighbors (KNN) and Decision Tree.

The first variation of both-sampling implemented to create a model uses more under-sampling. After re-sampling, the data includes 2,844 non-accepting customers and 2,742 accepting customers, therefore there are 5,586 customers in the data set. Recall that the data was split into two datasets, where the training data set is used to create the models and the testing data set is used to test the model's predictability. When the models are created with the training data, the model

calculates a misclassification rate on the training data, providing the training misclassification rate. Therefore, every misclassification rate in this section is a training misclassification rate, and the misclassification rates in Section 4 are the testing, or validation, misclassification rates. For every K-Nearest Neighbors model created, K values 1-30 were implemented, and the lowest misclassification rate was used as the best-fitting model. The KNN model was computed with all the variables, and the lowest misclassification rate is 8.242% at K=1. The best-fitting Decision Tree model was also determined by finding the smallest misclassification rate. The "best" number of splits for the model was computed by JMP, where the model stops splitting when the split does not improve the model. The Decision Tree model including all the variables split 59 times with a misclassification rate of 24.31% and an R-squared value of 0.311.

The next variation of both-sampling implemented to create a model is the even mixture of under-sampling and over-sampling. After re-sampling, the data includes 5,523 non-accepting customers and 5,548 accepting customers, therefore there are 11,171 customers in the data set. The KNN model with all variables computed the lowest misclassification rate of 4.001% at K=1. The Decision Tree model with all the variables split 339 times with a misclassification rate of 7.23% and an R-squared value of 0.730.

The last variation of both-sampling implemented to create a model uses more over-sampling. After re-sampling, the data includes 8,435 non-accepting customers and 8,322 accepting customers, therefore there are 16,757 customers in this data set. The KNN model with all variables computed the lowest misclassification rate of 2.530% at K=1. The first Decision Tree model including all the variables split 566 times with a misclassification rate of 3.04% and an R-squared value of 0.861.

3.2 Comparison of Classification Models

In order to compare all the models created from the three variations of both-sampling, Table 2 displays each model's training misclassification rate.

	More Under-Sampling	Balanced Sampling	More Over-Sampling
Decision Tree	24.310	7.230	3.040
K-Nearest Neighbors	8.242	4.001	2.530

Table 2: Training Misclassification Rates of Decision Tree and KNN Models for Each Both-Sampling Method Variation

One observation from Table 2 is that for each both-sampling method, the KNN models have lower misclassification rates. Recall that for every KNN model, the best misclassification rate is at K=1. In other words, each observation was assigned as an accepting customer or non-accepting customer based on its one closest neighbor. The KNN models could then be over-fitting, since they allow noise in the data to possibly predict the class. The KNN algorithm could also perform better for the data set compared to the Decision Tree algorithm. These questions will be determined in Section 4.

Another distinction across the different both-sampling methods is that as the methods include more over-sampling, the misclassification rates decrease. As previously stated, over-sampling can cause an over-fitted model and under-sampling can cause important loss of information. This implies that while a model might have a low misclassification rate, it might not predict well. Technically speaking, the best-fitting classification model is a K-Nearest Neighbors model that uses more over-sampling. However, each model will need to be examined for predictability in the next section in order to determine if it is over-fitted or lacks important information.

4 Is it possible to create a predictive model of credit card offer acceptance?

For each model created in Section 3, the prediction formula will be used on the testing data that was set aside in the previous section to evaluate how each model performs on data it has not been exposed to. Table 3, shown below, displays the testing misclassification rates of each model when applied to the testing data.

	More Under-Sampling	Balanced Sampling	More Over-Sampling
Decision Tree	45.114	26.405	19.507
K-Nearest Neighbors	21.936	16.447	13.981

Table 3: Testing Misclassification Rates of Decision Tree and KNN Models for Each Both-Sampling Method Variation

Table 3 shows that every model created in Section 3 has a testing misclassification rate between 13.981% to 45.114%. Similar to Table 2, more over-sampling in the data and KNN models produced lower misclassification rates. Therefore the best model created is still a K-Nearest Neighbors model with more over-sampling. However, when the misclassification rates are compared to Table 2, there is a large difference between the training and testing misclassification rates, which implies that the models are overfitting.

Recall that every model was created using all the variables in the data set. In an attempt to improve predictability, models with only variables significant to offer acceptance were created. However, since the only significant variables were reward type, mailer type, credit rating level, and income level, the predictability of these models were even more insufficient. Models including household size and quarter 1 balance as significant variables were also created, since their p-values were close to 0.05. However, these models also did not improve predictability from the models with all variables listed in Table 3. Therefore, the models including all the variables remain as

the best models.

Each KNN model created in Section 3 had the lowest misclassification rate at K=1. As previously stated, KNN models at K=1 can be over-fitting, since they allow noise in the data to possibly predict the class. Therefore, the KNN models at K=3 and K=5 were also analyzed for predictability. Since the more over-sampled variation has the best misclassification rates, then this both-sampled variation was used to determine if higher K values can improve predictability. Recall that the training misclassification rate for the more over-sampled data at K=1 is 2.530%, and the testing misclassification rate is 13.981%. The KNN model of the more over-sampled data at K=3 has a training misclassification rate of 4.231% and a testing misclassification rate of 18.505%. The KNN model at K=5 has a training misclassification rate of 5.598% and a testing misclassification rate of 23.568%. Since the misclassification rates increase as K increases, then the KNN model at K=1 still has the best predictability.

The misclassification rates in Table 3 are also similar to those in Table 2 where the Decision tree rates are greater than the KNN rates. However, the difference between the Decision tree rates in Table 2 and Table 3 are higher than the difference between the KNN rates. This implies that the Decision Tree models are also over-fitting. In order to determine if a Decision tree with better predictability can be created, the number of splits in the Decision tree models were reduced. Recall that in the more over-sampled data, the Decision Tree model split 566 times with a 3.04% training misclassification rate and a 19.51% testing misclassification rate. The new Decision Tree model with 400 splits has a 5.39% training misclassification rate and a 20.93% testing misclassification rate. This implies that less splits in the model does not improve the predictibility. To verify this, another model with 300 splits was created. This model has a 8.53% training misclassification rate and a 25.03% testing misclassification rate. The testing misclassification rate increased even

more, therefore the original Decision Tree models are better predictive models than models with less splits.

Since each attempt at improving the predictability of the KNN and Decision Tree models failed, then the models with misclassification rates in Table 3 are the best predictive models of each division. The overall best model to predict whether a customer accepts a credit card offer is the K-Nearest Neighbors model that implements more over-sampling, which has a testing misclassification rate of 13.981%. Therefore, it is possible to create a reasonably predictive model for credit card offer acceptance.

5 Conclusion

In the analysis of the credit card marketing data set, the variables significant to which customers accept credit card offers were determined and different models were implemented to predict whether a customer will accept the credit card offer.

In order to determine which variables are significant to offer acceptance, each variable's relationship with offer acceptance was tested using either a Pearson's chi-squared test or point-biserial test, depending on whether the variable is continuous or categorical. Overall, the only significant variables to offer acceptance are the offered card's reward type, the mailer type of the card, the customer's credit rating level, and the customer's income level. This indicates that the distribution of the card is important to whether the customer accepts the card. Although the study distributed an equal amount of reward types among the customers, more customers accepted an Air Miles reward card versus a Points or Cash Back rewards card. Equivalently, more customers accepted an offer received by postcard than by letter. Therefore, customers would be most likely to accept Air Miles reward cards sent by postcard. Also, customers with low credit rating levels

and low income levels accepted cards more than customers with high levels. Then low level customers are more likely to accept a card and should be a more targeted group for credit card offers. From each significant variable, the "golden" offer would be an Air Miles reward card mailed by postcard to a low income and low level customer.

In order to predict which customers would accept a card offer, different classification models were created and analyzed for predictability. Since the percentage of customers that accepted the card is small, then the data was re-sampled using three different variations of both-sampling: more under sampling, more over sampling, and a balance of more and under sampling. K-Nearest Neighbors and Decision Tree models were created for each re-sampling method, and the training and testing misclassification rates for each model was calculated. Overall, the best predictive model is a K-Nearest Neighbors model that utilizes more over sampling with a testing misclassification rate of 13.981%.

The large difference the training and testing misclassifications rates suggest the models are over-fitting. Three methods were employed to attempt to increase the predictability of the models. First the K-Nearest Neighbors models at different K values were analyzed, they but did not increase predictability. Next, the number of splits in the Decision Tree models were decreased to adjust for over-fitting, but they also did not increase predictability. The last attempt involved creating models with only significant variables, but this also did not increase predictability. Therefore the K-Nearest Neighbors model with more over-sampled data prevailed as the best predictive model.

While the 13.981% misclassification rate of the best model isn't as low as desired, it is reasonable for the purpose of the study, which is determining which customers to send credit card offers to. If the study was regarding which customers are more likely to default on a loan, then a more

accurate predictive model is necessary because an inaccurate prediction would be more harmful to the bank. However, the model predicts which customers accept credit card offers, and the cost of sending an offer to a customer that doesn't accept it or not sending an offer to a customer that would accept it isn't significant. Therefore, it is possible to create a reasonably predictive model for credit card offer acceptance. In order to determine which customers to send offers to, the bank that implemented the study could either utilize the best predictive model or distribute offers based on the variables significant to offer acceptance.

Credit Card Marketing Study Formal Report

Brianna Humphries

Advised by Dr. Kurt Cogswell

Fall 2020

Summary of Study

The bank designed a credit card marketing study to determine why customers accept or refuse credit card offers. Offers were sent to 18,000 current bank customers, and 16 different variables were recorded. These variables are:

- Offer Acceptance (Yes/No)
- Reward plan (Air Miles/Cash Back/Points)
- Mailer Type (Letter/Postcard)
- Income Level (High/Medium/Low)
- Number of Non-Credit Card Bank Accounts Open
- Overdraft Protection on Checking Account(Yes/No)
- Credit Rating Level (High/Medium/Low)
- Number of Credit Cards Held at the Bank
- Number of Homes Owned
- Household Size
- Home Ownership (Yes/No)
- Average Bank Balance (Across all accounts over time)
- Q1, Q2, Q3, and Q4 Balances (In the last year)

The Offer Accepted variable is the focus of this study. The Reward Plan and Mailer Type variables describe the offers and how they were distributed. The customer demographic variables were collected from the bank's database.

Summary of Findings

The analysis of the study is split into two parts:

1. Which variables affect why customers accept or refuse credit card offers?

2. Can a model be created that predicts a customer's offer response?

In order to determine why customers accept or refuse credit card offers, each variable was compared to Offer Acceptance by using a statistical test. This test determined whether the variable affects whether a customer will accept a credit card offer. From the tests, the variables that do affect offer acceptance are the card's reward type, the offer's mailer type, the customer's income level, and the customer's credit rating level. The first two variables describe the offer distribution. For the reward type, customers are most likely to accept an Air Miles reward card and least likely to accept a Cash Back reward card. For the mailer type, customers are more likely to accept a card mailed by Postcard over a card mailed by letter. The last two variables are customer demographics. Low income level customers are more likely to accept a card offer than high income customers. Similarly, low credit rating customers are more likely to accept a card offer than high credit rating customers.

In the second part of the data analysis, several different models were created that predict whether a customer will accept a credit card offer based on their demographics and the offer distributions. It's important to note that only 5.7% of customers in the study accepted the card offer, therefore the data was imbalanced. After implementing various re-sampling methods and different types of models, the overall best model has a 13.98% misclassification rate. This indicates that the model will predict whether a customer will accept an credit card offer with only a 13.98% error rate. Since misclassification rates for this type of data often vary from 30% to 40%, this model is very actionable for the bank to use.

Recommendations

The analysis showed that the card's reward type, the offer's mailer type, the customer's income level, and the customer's credit rating level all affect whether a customer will accept a card offer. The "golden" offer based on these variables are an Air Miles reward card mailed by postcard to a low income and low credit rating customer. The bank could focus their offer distributions on these types of offers and customers. The bank could also use the predictive model to predict whether a customer will accept a credit card offer.

These methods are recommended, because they can help the bank cut costs. This can be done by cutting distribution to customers that are not likely to accept a card offer or discontinuing offer types that are not receptive. Since a Cash Back reward card is the least accepted, then this card should be discontinued, especially if it is more expensive. More receptive offers distributed could also increase the number of credit cards accepted and held at the bank. This could increase the bank's interest revenue, as well as the customer's loyalty, which inclines them to return to the bank for other financial needs, such as a loan. Therefore distributing offers using the methods listed above is highly recommended.

References

[1] JMP Pro, Version 14. SAS Institute Inc., Cary, NC, 1989-2019.

The JMP Pro software was used to create models in this paper.

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

The RStudio software was used to analyze data in this paper. Packages used were readxl, ggplot2, scales, ggpubr, tidyr, plyr, ROSE, and dplyr.

6 Biography

Brianna Humphries is currently an undergraduate at South Dakota State University, where she will complete her B.S. in Data Science in December 2020. She analyzed this data set and wrote this paper as part of the Senior Capstone requirements. While at SDSU, Brianna served as the Finance and Operations Vice President and Recruitment Data Director of Alpha Xi Delta Women's Fraternity.

Math has always been Brianna's favorite school subject. She also likes making spreadsheets for anything and everything possible, because she likes to be efficient and organized. Although Brianna is from a small town in South Dakota, she would love to live in a city, working with data at a large company. However, with the current state of the world, Brianna plans to move to her hometown after graduation.