Exploring the Efficacy of Machine Learning in Credit Card Fraud Prevention

Philip Kim

Abstract

This study investigates the presence of credit card fraud, and statistical methods in which to detect fraudulent transactions. A dataset containing 1,000,000 observations is reduced to 10,000 due to computational limitations, which represents a sample of the overall population (all credit card transactions in the corresponding location in which the data was collected). Data exploration reveals a heavily skewed relationship between some predictors and the response, which were transformed with a natural logarithm. Additionally, 3 machine learning algorithms, including Logistic Regression, k-Nearest Neighbors, and Random Forests are tested and evaluated based on accuracy and misclassification rates. Running the models reveal that Random Forests perform the best, with an accuracy of 99%. It turns out that some impactful variables in predicting fraudulent transactions are the amount ratio (ratio of purchase price to median purchase price of card holder), the presence of an online order, and the use of a PIN number. We can conclude that because the dataset is very noisy, random forests are expected to thrive in this environment. However, it is important to note that the other models performed at a high level as well. Future research should tackle the use of all observations, include several additional predictors, and consider alternative machine learning approaches to the data.

Exploring the Efficacy of Machine Learning in Credit Card Fraud Prevention

Introduction

With the technology sector growing faster than ever, new advances have made daily life more efficient. For example, one aspect that has been improved is the ease of purchases from online vendors. However, the consequences of this luxury are the many instances of online credit card fraud. Credit card fraud is essentially identity theft, where a thief impersonates the true credit card holder to make a purchase in their name (Legal Information Institute, 2022). While credit card fraud has been a prevalent issue since the integration of online purchases, the quantity of frauds has been at an all-time high in recent years. In fact, the Federal Trade Commission recorded approximately 390,000 instances of credit card fraud in 2021 (Egan, 2023). Additionally, the Nilson Report forecasts a loss of $165 billion for the United States over the next 10 years (Egan, 2023). Clearly, fraud is an issue that must be minimized in order to reduce losses. Therefore, credit card fraud has been a deeply researched topic, with various statistical models being utilized to identify fraudulent transactions. For instance, the paper *Hybrid Approaches For Detecting Credit Card Fraud* by Yiğit Kültür and Mehmet Ufuk Çağlayan details the use of an ensemble of machine learning models to tackle this issue (Kültür & Çağlayan, 2017). The approach that was discussed throughout the paper is called OPWEM, which is abbreviated for optimistic, pessimistic, and weighted voting in an ensemble of models, and provided a high detection rate (Kültür & Çağlayan, 2017).

Now, compared to the previously mentioned study, we will be answering a vastly simpler question, due to the scope of this paper. The main purpose of this paper is to apply a series of statistical learning models with the goal of concluding a singular model that most accurately

detects instances of credit card fraud. The models that will be tested are the following: logistic regression, K-nearest neighbors, naive bayes, decision trees, and random forest.

The justification of undertaking this study is simply to further protect potential victims of identity theft. Fraud objectively has a negative impact on society and the well-being of the people, so reducing this will help to improve quality of life. The results could be interesting, because we have quite a few candidates that are all classification models that will most likely provide similar results, but identifying the *best* model is a more intricate process. Additionally, although credit card fraud is a topic that has been studied many times, the results of this study with this particular data set could provide insights on how to apply these models to broader data. Therefore, perhaps utilizing a singular model could help firms such as banks achieve similar results to much more complex solutions, with the use of less computing power.

Literature Review

Article 1: Hybrid Approaches For Detecting Credit Card Fraud - Yiğit Kültür & Mehmet Ufuk Çağlayan

This article provided a very in depth and detailed approach to credit card fraud detection. A major focus on the first portion of the article was to describe what previous methods were used for fraud detection by banks, such as hiring fraud experts to manually create rules to identify fraud (Kültür & Çağlayan, 2017). However, the major disadvantage to this method was that those who perform fraudulent transactions were quick to adjust their own methods, and the time it took for experts to come up with new rules was too long (Kültür & Çağlayan, 2017). Therefore, AI and machine learning models have been used to be able to quickly adapt to new data and identify the underlying patterns that were present in fraudulent transactions (Kültür &

Çağlayan, 2017). The approach that was discussed throughout the paper is called OPWEM, which is abbreviated for optimistic, pessimistic, and weighted voting in an ensemble of models (Kültür & Çağlayan, 2017). The ensemble of models that are included in this overarching approach are decision tree, random forest, Bayesian network, Naive Bayes, support vector machine, and K* models (Kültür & Çağlayan, 2017). This means that the method used to identify fraud is not based on just one model, but rather based on the collective results from multiple models. Additionally, the acronym also includes optimistic, pessimistic, and weighted voting, which describe the method in making the final decision. An 'optimistic' approach means that even if just one of the above models (within the ensemble) concludes that a transaction is legitimate, then the final decision will be legitimate (Kültür & Çağlayan, 2017). Inversely, pessimistic means that if just one model concludes that the transaction is fraudulent, then the final decision will be 'fraudulent' (Kültür & Çağlayan, 2017). Weighted voting means that each model will have a different weight assigned to it that will help to determine the final decision (Kültür & Çağlayan, 2017). These weights are calculated using a multitude of equations that are included in the article. The results that the authors were able to conclude followed the above statements relatively well. They found that the optimistic approach had a lower detection rate and lower false alarm rate (Kültür & Çağlayan, 2017). The pessimistic approach had a much higher detection rate, but with a much higher false alarm rate as well (Kültür & Çağlayan, 2017). Finally, the weighted decision voting had values that were in between the optimistic and pessimistic rates (Kültür & Çağlayan, 2017).

Article 2: Credit Card Fraud Detection Using AdaBoost and Majority Voting - Kuldeep Randhaw et al.

Similar to the first article, the authors delve into the background of credit card fraud, and the detrimental losses that are incurred if they are not caught. The first half of this article also includes very detailed descriptions of various machine learning models, such as Naive Bayes, Logistic Regression, Linear Regression, etc. Additionally, the authors include some methods from other related studies that have approaches that are relevant to the topic. However, the main focus of this article is to examine the effects that Adaboost has on results and to further investigate majority voting. Adaboost is essentially a complementary addition that is used in conjunction with other models to improve their performance (Randhaw et al., 2018). To begin, some initial machine learning models were implemented without the addition of Adaboost (Randhaw et al., 2018). Then, the results are again examined once Adaboost has been implemented into the various models (Randhaw et al., 2018). The Matthews Correlation Coefficient was also included as a diagnostic tool to analyze the false/true positives/negatives (Randhaw et al., 2018). The conclusion that the authors reached was that using both Adaboost and majority voting provided the best results, according to the MCC (Randhaw et al., 2018). Overall, this article was shorter but a much tougher read, as it included a lot of dense information into sections that were not necessarily divided the best.

Article 3: Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning - Eunji Kim et al.

As with the other 2 articles, this article starts with the dangers of credit card fraud and why it is necessary to prevent it. The authors use the probability of a transaction being fraudulent determined by the model with the addition of a cutoff value that the bank determines to decide whether a transaction is fraudulent (Kim et al., 2019). Additionally, as we've seen in other studies, the false positive and false negative rates are utilized as a metric (Kim et al., 2019). The

main idea of the paper is to use the champion-challenger framework, where the ensemble model (collection of many models) is the champion, while the deep learning model is the challenger (Kim et al., 2019). After running the experiment on a data set from South Korea, the results showed that the deep learning model was better than the ensemble model (Kim et al., 2019).

Compare/Contrast: Overall, although the approaches that were taken in each article were very different, the structure of them were pretty similar. In general, all the articles included a summary or abstract outlining their findings, background information regarding the topic, the description of their experiment, a discussion of the statistical models that were used as well as the diagnostic metrics, and then a conclusion that included suggestions or contributions. In terms of how they all differed, the subsections that were present varied. For example, article 2 had a subsection titled “Machine Learning”, which discussed how the models worked, and was not in any other article. It is also difficult to say whether the conclusions of each article were similar, because the approaches were very different. That being said, Article 1 and Article 3 both considered an ensemble of models, but Article 3 decided that another model (deep learning) was better. With the response variables, Articles 1 and 2 used a clear binary classification problem, whereas Article 3 considered a more probabilistic approach with a cutoff determined by the company. The explanatory variables were not really discussed at all, as most of the data sets were confidential, and in one case, only the principle components were available, as it was a public data set.

Methodology

The data that we will be from Kaggle, a popular site that is widely utilized to download datasets for data science projects. According to the datacard, this dataset was sourced by some unnamed institute, and has been downloaded over 13,000 times (Narayanan, 2022). Additionally, there are over 90 notebooks that have used this data for projects, so overall, we can say that this dataset is acceptable for a study. As for the description of the data, the population that is being studied is all credit card transactions (corresponding to the location in which the data was collected). Naturally, the dataset (sample) is pertaining to a smaller proportion (1,000,000) of cases of credit card transactions that occurred in that location. When selecting this sample, there does not seem to be any disqualifying characteristics besides the case in which one of the 8 variables were not available to the entity that sourced the data. Furthermore, because there are millions of credit card transactions daily, there was most definitely enough data to fit the required variables and have a sufficient sample. As mentioned previously, the dataset contains 8 total variables and 1,000,000 total observations (rows). Clearly, there will be plenty of data that will be used in the analysis. However, due to sheer quantity of observations, this could cause computational issues when modeling (R takes too much time to run models). Therefore, a random subset of 10,000 observations will be used as the dataset going forward. It can most likely be concluded that the results will only get better as more data is added. As for the cleanliness of the data, there are zero missing values present, and the data is very clean overall, and therefore the data should require very minimal cleaning. The 8 variables that were included in the dataset are as follows:

- *Distance from Home*:
    - Unknown measurement for distance (most likely km or miles)

    - Quantitative variable
    - Distance in which the transaction occurred from the credit card owner's home
- *Distance from Last Transaction*:
    - Unknown measurement for distance (most likely km or miles)
    - Quantitative variable
    - Distance in which the transaction occurred from the holder's previous transaction
- *Transaction Amount to Holder's Median Purchase Price Amount (Amount Ratio)*:
    - Unknown measurement for currency
    - Quantitative variable
    - Ratio of the transaction amount to the median purchase amount of the holder
- *Repeat Retailer*
    - No unit of measurement
    - Qualitative variable
    - Coded as 1 if the transaction happened from a retailer the holder has already purchased from, and 0 if this is the holder's first transaction from the retailer
- *Used Chip*
    - No unit of measurement
    - Qualitative variable
    - Coded as 1 if the credit card chip was used for the transaction, and 0 if it was not used
- *Used PIN Number*
    - No unit of measurement
    - Qualitative variable

    - Coded as 1 if the pin number was used for the transaction, and 0 if it was not used
- *Online Order*
    - No unit of measurement
    - Qualitative variable
    - Coded as 1 if the transaction was completed online, and 0 if it was not
- *Fraud*
    - No unit of measurement
    - Qualitative variable
    - Coded as 1 if the transaction was fraudulent, and 0 if it was legitimate

For the analysis of the data, a multitude of models will be tested to see which one performs the best in predicting fraudulent transactions. Specifically, the logistic regression, random forest, and K-nearest neighbors algorithms will be examined. As for the predictors that will be included in each model, all 3 models will include the 7 predictors that are defined above and will be utilized to predict the *fraud* response variable.

Model 1: Logistic Regression

For Logistic Regression, consider the following model:

$$logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

where $Y = Fraud,\ X_1 = \log(Distance\ from\ Home), X_2 = \log(Distance\ from\ Last\ Transaction), X_3 = \log(Amount\ Ratio), X_4 = Repeat\ Retailer, X_5 = Used\ Chip, X_6 = Used\ PIN\ Number, X_7 = Online\ Order.$

For this model, we look to train the logistic regression on a set of training data and test its accuracy on a set of test data, and obtain measures of validity. From the coefficient

results, we can determine which variables cause the probability of a fraudulent transaction to increase/decrease, and by which magnitudes as well.

Model 2: K-Nearest Neighbors

For the K-Nearest Neighbors model, the model will need to be tuned so that the optimal number of neighbors are being considered when running the algorithm. This version of K-Nearest Neighbors will utilize the Euclidian Distance formula to calculate the closest neighbors to each observation. The formula for the Euclidean Distance is as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

The K-Nearest Neighbors model will contribute another model in which we can compare the other two models with. The model will also allow for the fine tuning of the k parameter, which controls the flexibility of the model, and can help with our predictions.

Model 3: Random Forests

The Random Forest algorithm consists of a collection of decision trees and utilizes the majority vote method to make predictions. Like logistic regression, we will be using a training and test set to measure the predictive power and accuracy of the model and compare its results to the two other models. Similarly to kNN, we can fine tune the number of variables and trees considered for our model, which can help to improve the model further.

Results

Credit card fraud is a dangerous form of identity theft that has plagued card holders for decades. Identifying instances of fraud is crucial in the safety and wellbeing of card holders, as well as companies who seek to provide quality credit card services. The key objective of this

paper is to accurately predict whether a transaction is fraudulent or legitimate, in hopes of aiding firms in identifying fraud and helping customers stay protected. A series of models will be tested, and the best performing model will be determined via accuracy and misclassification.

Descriptive Statistics

Figure 1 displays the number of cases of fraud vs legitimate transactions that were found in the data set (the response variable). Clearly, there is an overwhelming number of legitimate transactions, and a smaller fraction of fraudulent transactions. In fact, only 8.74% of all observations are classified as "1", which corresponds to a fraudulent transaction. In terms of quantity, out of 1,000,000 observations, 87,403 were fraudulent, whereas the other 912,597 transactions were legitimate, which seems reasonable when comparing this to reality.

Figure 1

Count of Total Fraudulent and Legitimate Transactions

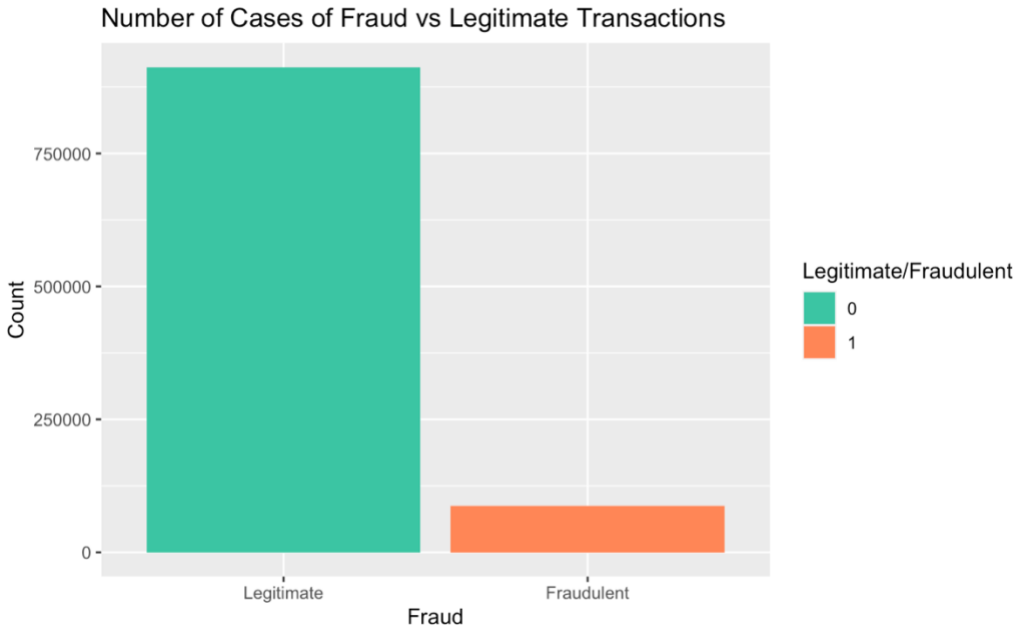


Table 1

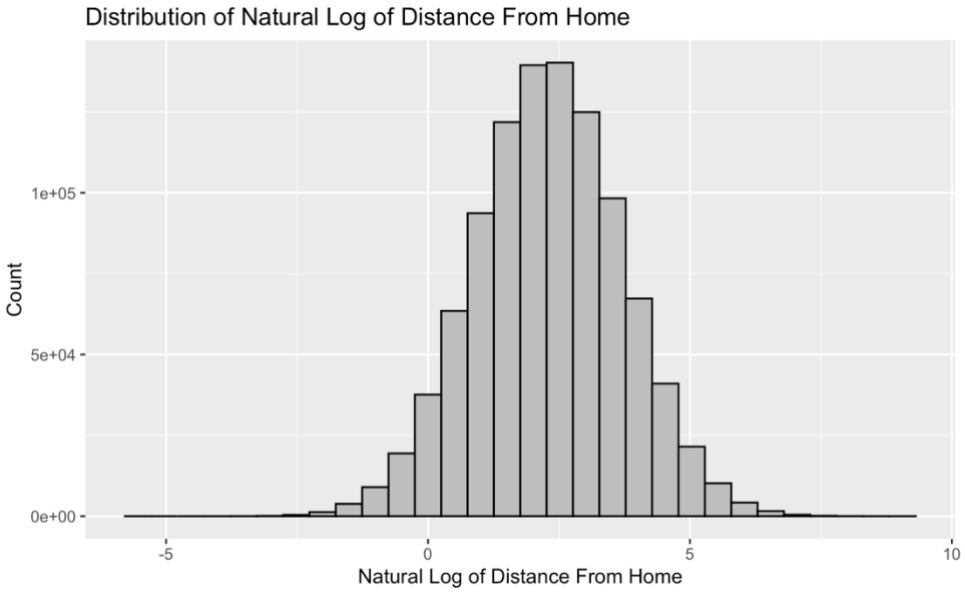
Summary Statistics of Quantitative and Transformed Quantitative Variables

Variable Name	*Min*	*Median*	*Mean*	*Max*	*Skewness*
Dist. from Home	0.005	9.968	26.629	10632.724	20.240
Log(Dist. from Home)	-5.324	2.299	2.301	9.272	-0.0002
Dist. from Last Trans.	0.000	0.999	5.037	11851.105	125.921
Log(Dist. from Last Trans.)	-9.042	-0.001	-0.003	9.380	-0.0005
Amount Ratio	0.004	1.000	1.824	267.803	8.915
Log(Amount Ratio)	-5.426	-0.002	-0.002	5.590	-0.002

Plotting the 3 quantitative variables revealed severe right skewness in their distributions due to a high number of outliers. Furthermore, the original data histograms did not reveal anything about the variables' behaviors, and thus are not included in this paper. Due to the severe skewness, a natural logarithm transformation was applied to these predictors, and the summary statistics of both the original and transformed variables are shown above in Table 1. The 3 original quantitative variables had very low minimum values and extremely large maximum values, as well as very differing medians and means. More specifically, the means of all 3 original variables are much greater than their medians. This again suggests that these predictors have very high value outliers that lift the mean to a higher value and skews the data. Thus, applying the transformation does well to give the data a nice bell curve shape and reduces the values of the skewness by a dramatic amount compared to the original skewness values. For instance, Distance from Last Transactions had an original skewness of 125.921, and applying the natural logarithm transformation was able to reduce this to essentially zero.

Figure 2

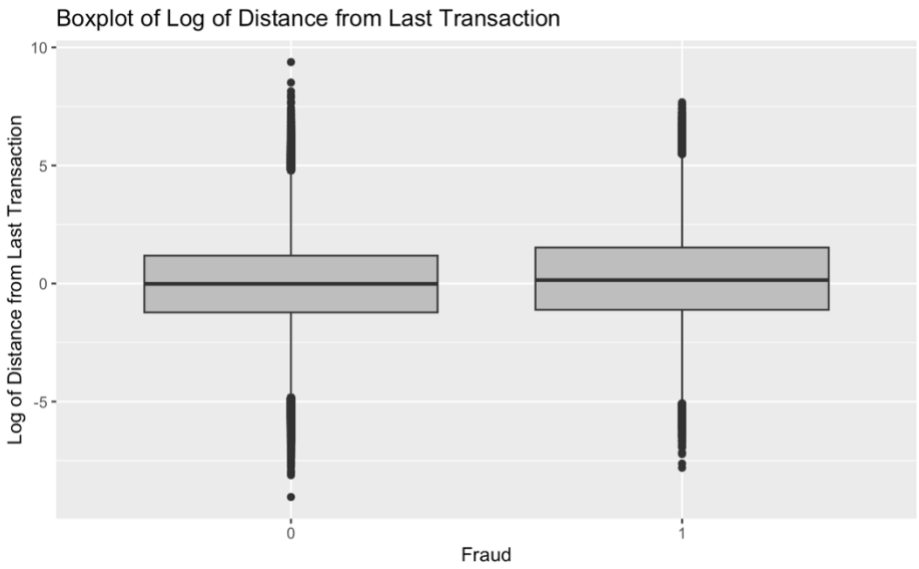
Distribution of the Transformed Distance from Home Variable



Here is an example of one of the heavily skewed quantitative predictors. The histogram of the original data provided very little to no insight on the distribution of the variable, but the natural log transformation helps to normalize the data, which can be seen in Figure 2. Additionally, this behavior can be seen in Distance from Last Transaction and Amount Ratio as well. For instance, the below side-by-side boxplot shown in Figure 3 displays the transformed Distance from Last Transaction variable. Clearly, the transformation does a nice job of normalizing the data for both cases of the response variable. Therefore, the transformed quantitative predictors will be used in the modeling process.

Figure 3

Distribution of Transformed Distance from Last Transaction Variable Across Response Values



Moving on to the categorical variables, bar charts (Figure 4) and proportions (Table 2) were utilized to analyze their behaviors. Firstly, we see an equal proportion of both frauds and legitimate transactions where the transaction was with a repeat retailer, as both settled at around 88%. For Used Chip, the proportion for legitimate transactions were greater than that of the fraudulent transactions by around 10%, indicating that given the transaction is fraudulent, there is a higher likelihood that a chip was not used. As for the PIN Number, almost none of the fraudulent transactions involved a PIN Number, whereas 11% of the legitimate transactions utilized the PIN Number. This suggests that fraudsters are less likely to use a PIN Number when completing a fraudulent transaction. Finally, almost all (~95%) of fraudulent transactions were done so online, whereas only ~62% of legitimate transactions were completed online. All in all, we can see some differences between the two responses in relation to the qualitative predictors.

Table 2

Proportions of Frauds/Legitimate Transactions where Qualitative Variables are True

Variable Name	% of Frauds where Variable is True	% of Legitimate where Variable is True
Repeat Retailer	88.01%	88.17%
Used Chip	25.64%	35.94%
Used PIN Number	0.31%	11.00%
Online Orders	94.63%	62.22%

Figure 4

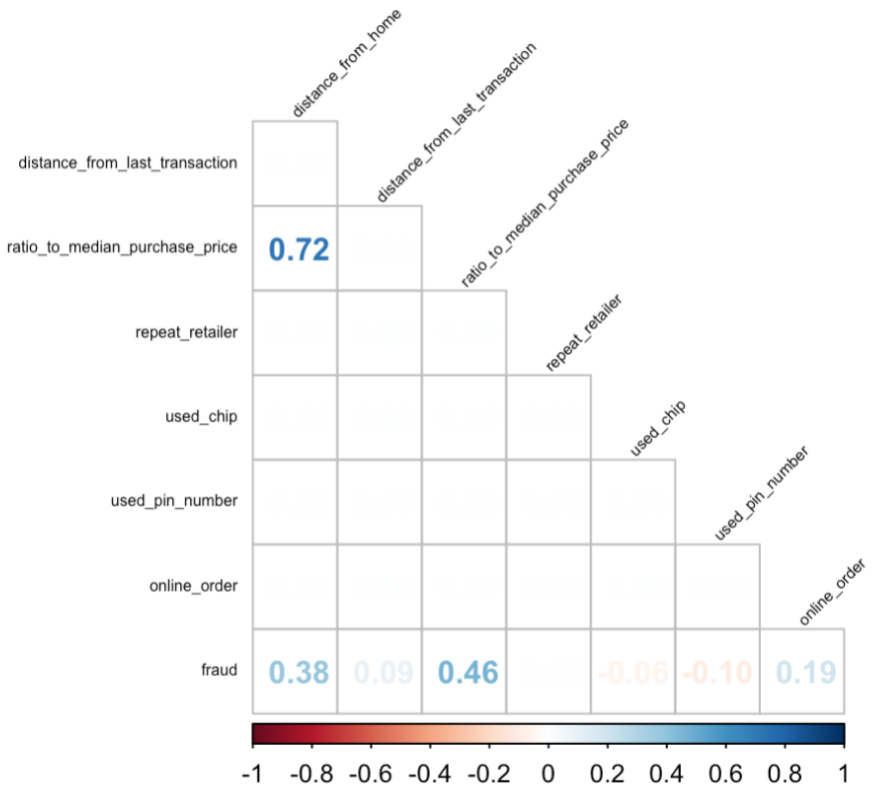
Count of Qualitative Variables Across Response Values

Lastly, let's explore the correlations between all the predictors. The correlation heatmap in Figure 5 displays the response, qualitative, and (transformed) quantitative variables that will

potentially be included in the final models. One thing to note about the correlation heatmap is the high correlation between log(Amount Ratio) and log(Distance from Last Transaction). This relationship will be investigated in our inferential statistics. Additionally, Fraud has very slight positive correlations with log(Distance from Home) and log(Amount Ratio), which can give us a general idea on the predictive power of these independent variables.

Figure 5

Correlation Heatmap of all Predictors (Quantitative Variables are Transformed)



Inferential Statistics

As mentioned, 3 total models were run on the data (Logistic Regression, k-Nearest Neighbors, and Random Forests). The first model was logistic regression, where it predicts the probability that a test observation is fraudulent. The regression included all variables in the dataset, and table 3 displays the estimates, standard error, z-values, and p-values from the output. Looking at the output, the intercept, repeat retailer, and used pin number have negative coefficients. For distance from home, a coefficient of 0.840 indicates that when all other predictors are held constant, a one increase in the log of the distance from home results in a $e^{0.840}$ = 2.32 times more likelihood that the transaction is fraudulent. Similarly, a one increase in the log of distance from last transaction results in a $e^{0.194}$ = 1.21 times more likelihood, and a one increase in the log of the amount ratio results in a $e^{2.280}$ = 9.78 more likelihood. This same reasoning can be applied to the negative coefficients as well. For example, if a transaction is at a repeat

retailer, the likelihood that the transaction is fraudulent is 2.382 times less likely than if the transaction was not with a repeat retailer. The coefficients reveal that an increase in the amount ratio tends to increase the likelihood of fraud the most, whereas the use of a pin number tends to decrease the likelihood the most. Finally, all p-values were essentially zero, indicating high statistical significance. As for performance metrics, logistic regression had an accuracy of 94.25% and a misclassification rate of 5.75%. A confusion matrix reveals that of the 2,000 test observations, the model correctly predicted 1,885 correctly and 115 incorrectly.

Table 3

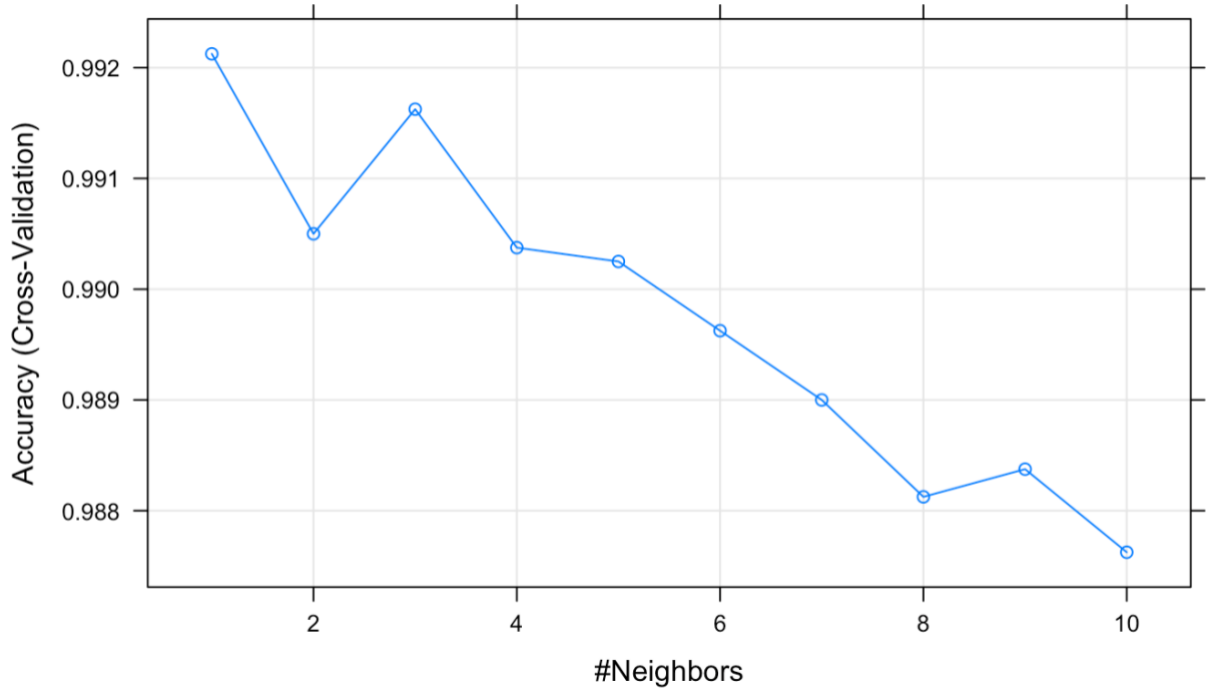
Coefficient Estimates, Standard Error, Z-Value, and P-Value from Logistic Regression

Variable Name	Coefficient Estimate	Standard Error	Z Value	P-Value
Intercept	-6.710	0.029	-233.20	<2e-16
Log(Distance from Home)	0.840	0.005	164.27	<2e-16
Log(Dist. from Last Trans.)	0.194	0.003	63.84	<2e-16
Log(Amount Ratio)	2.280	0.008	278.61	<2e-16
Repeat Retailer	-2.382	0.022	-107.11	<2e-16
Used Chip	-0.865	0.012	-71.39	<2e-16
Used PIN Number	-5.278	0.074	-71.11	<2e-16
Online Order	3.693	0.216	170.69	<2e-16

The second considered model was k-Nearest Neighbors. To find the best possible model, 5-fold cross validation was performed to identify the optimal k (the number of neighbors to consider). The cross validation reveals that k = 1 is the optimal number of neighbors and provides the best accuracy score. Figure 6 shows that as the number of neighbors considered increases, the accuracy score decreases. In fact, the model with k = 1 had an accuracy of 93.80% and a misclassification rate of 6.20%. The confusion matrix shows that the KNN model predicted 1,876 observations correctly and 124 observations incorrectly.

Figure 6

Cross Validation Accuracy for kNN Across Various Considerations of Neighbors (k)



The third considered model is Random Forest. Like k-Nearest Neighbors, 5-fold cross validation was performed to identify the value of mtry, which is the number of variables to consider at each split of the decision trees. The cross validation reveals that an mtry of 7 maximizes the accuracy, but further investigation revealed that an mtry that is greater than 2 did not improve the model significantly. Therefore, an mtry of 2 was used in the final model. The random forest model performed the best by far, producing an accuracy of 99.55% and misclassification rate of only 0.45%. The confusion matrix reveals that the random forest model predicted 1,991 transactions correctly, and only misclassified 9 observations. Random forest also provides relative variable importance from the model (refer to Table 4). From the predictors, the amount ratio seems to be the most impactful, followed by online order and pin number. Moreover, whether the transaction was from a repeat retailer seems to have the least impact on whether a transaction is fraudulent.

Table 4

Relative Variable Importance from Random Forest Model

Variable Name	Relative Importance
Amount Ratio	43.548
Online Order	35.638
Used PIN Number	25.623
Distance from Home	25.286
Used Chip	20.479
Dist. from Last Trans.	14.724
Repeat Retailer	7.117

Conclusion

As with any statistical study, it is crucial to explore the relationships that were discovered in the results. In particular, it is key for banks to understand which variables impact fraud the most, and more importantly, why they are impactful. The following subsections will dive deeper into these topics.

Interpretation of Results

While the k-Nearest Neighbors model is not too helpful in the importance of variables, logistic regression and random forest was able to offer insight on these predictors. In fact, both models revealed that when predicting fraudulent transactions, the amount ratio was the most impactful. This means that if the amount of the transaction was much greater than the median purchase price of the card holder, this was a major tip off that the fraud was more likely to be present. Logically, this makes sense, as for many card holders, making large purchases on a credit card is riskier. Larger purchases, such as a car, may have high interest rates and more expensive payments, so using a credit card probably isn't a great idea. Usually, these types of

purchases are made directly with bank loans, so having a high purchase price is a good indicator of a fraudulent transaction. Another shared high-importance predictor is if the transaction was completed online (online order). The explanation for the high importance is very simple, as fraudsters take less risk with an online order. In other words, completing a fraudulent transaction in person leaves room for physical risk, such as having to enter a PIN number or signature. Completing the order online allows for the fraudster to commit the crime without an actual physical trace, and therefore is an indication of fraud. Furthermore, we can see that in Table 3, PIN number has a negative coefficient, meaning if a PIN was used during a transaction, it is less likely that it was fraudulent. As for the distance measures (distance from home and distance from last transaction), as the location of the transaction gets further away from the card holder's home/last transaction, there is a higher likelihood of fraud. Logically, most of a person's transactions are typically confined within a certain distance of their residence. Also, most of the transactions probably occur close to each other, given that the card holder does not travel at high rates. However, it is also common that people go on vacation or travel for work and make purchases, which is why the distance measures are not as important as the previously discussed predictors. So why did random forest have the highest accuracy of the 3 considered models? The main reason why it performed at the highest level is that random forests thrive when the data is very noisy. As seen from the descriptive statistics, many of the variables had high variance in relation with the response. Therefore, because random forest is an ensemble method (utilizing many decision trees), it was able to become more robust and reduce the effects of the high variance. Additionally, the number of variables considered at each split (mtry) was able to be tuned, and even though only 2 were considered, it is enough to both increase the accuracy and reduce the variance slightly. On the flip side, KNN is more sensitive to outliers, which were

ever-present in the dataset. Logistic regression performed in the middle of the other two models, doing better than KNN but worse than random forest. That said, all 3 models produced some great results, and in application, having any of these accuracies would be very good.

Limitations

As with many other studies, there were some limitations to the research. One limitation was that due to the resources available, all 1,000,000 observations were not able to be considered in the models. This was due to the very intense computational power required to cross validate data and tune parameters for an extremely large dataset. If a more powerful machine were able to crunch all that data, the results could have potentially been even better than the results obtained in the current study. Additionally, although the results were very accurate, it also seems "too good to be true". Meaning, although the dataset seemed very valid, the extremely high accuracies may be because some of the predictors were much too correlated with the response (may have overfit). Another explanation could be that this may have been a simulated dataset that is engineered for the results to be very accurate. In practical applications, I would assume that the results would be much poorer than of those produced from this study. Nonetheless, the methods utilized, and the analysis could still very much be used in real life situations, just perhaps with lower metrics.

Suggestions for Future Research

One of the major implications of this study is for the protection of credit card holders. Any little improvement in the models to detect credit card fraud is helping people stay shielded against fraudsters and identity theft, and streamlining this process is positive. Also, it has also been shown that random forests are once again, very useful in real life applications, where data is extremely noisy. This can help lead other researchers into the right direction when considering a

singular or an ensemble of models to use in this topic. Specifically, for future research, I would recommend considering more predictors that tackle unexplored areas. For example, perhaps recording the industry of the purchase could help with identifying fraud. If a transaction occurs on a gambling site or through an international seller, we expect a higher probability of fraud. In addition, only 3 models were considered out of the vast space of machine learning. A future researcher could consider other models such as boosting, naïve bayes, and bagging to improve on the results. All in all, while this study steps into many areas of the field, there is certainly room for improvement, and much more to investigate in the world of fraud.

References

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Alboukadel Kassambara (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. https://CRAN.R-project.org/package=ggpubr

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Dhanush Narayanan R. (2022). *Credit Card Fraud* [Data Set]. https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud

Egan, J. (2023). *Credit Card Fraud Statistics*. Bankrate. https://www.bankrate.com/finance/credit-cards/credit-card-fraud-statistics/

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller and Davis Vaughan (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.1. https://CRAN.R-project.org/package=dplyr

Jarek Tuszynski (2021). caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.2. https://CRAN.R-project.org/package=caTools

Kim, Eunji, Lee, Jehyuk, Shin, Hunsik, Yang, Hoseong , Cho, Sungzoon, Nam, Seung-kwan, Song, Youngmi, Yoon, Jeong-a, & Kim, Jong-il (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications*, 128, 214-224. doi: 10.1016/j.eswa.2019.03.042

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kultur, Yiit, & Calayan, Mehmet Ufuk (2017). Hybrid approaches for detecting credit card fraud. *Expert Systems*, 34(2), p.np-n/a. doi: 10.1111/exsy.12191

Legal Information Institute. (2022) *Credit card fraud*. https://www.law.cornell.edu/wex/credit_card_fraud

Lukasz Komsta and Frederick Novomestky (2022). moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests. R package version 0.14.1. https://CRAN.R project.org/package=moments

Randhawa, Kuldeep, Chu Kiong Loo, Seera, Manjeevan, Chee Peng Lim, Nandi, Asoke K (2018). Credit Card Fraud Detection Using AdaBoost and Majority Voting. *IEEE Access*, 6, 14277-14284. doi: 10.1109/ACCESS.2018.2806420

Taiyun Wei and Viliam Simko (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.