The Story Behind a Fraudulent Transaction

Sherveer Dhillon, Hanif Kolahdoozan

Data Description and Preparation

The cybersecurity space is rapidly emerging as dimensionality of data increases in the modern world the threat of hackers and misuse of data has increases exponentially. However with the ongoing collection of credit card transactional data we see that what constitutes a fraudulent transaction has a correlation with many other factors at the point of transaction. We work with the data set "creditcardcsvpresent" which has 3075 transactions recorded. Each transaction record records weather a given transcaction is fraudulent or not alongside a record of various other variables recorded at the point of transaction which can be seen below in Table 1. After importing the dataset into R, the variables are not all in the correct format that we need to build our model. Hence we proceed with ensuring that each variable is in our desired type and after this conversion we have the final list of variable type in Table 1. We proceed our data wrangling procedure by checking for any missing values as this will be a problem for the application of our data to a logistic regression model. The variable Transaction_date is missing for every recorded transaction, hence we proceed by removing the variable as it is consistently missing. We proceed our initial analysis by checking for outliers, looking below at Figure 1 and 2 we visually see that some points visually are much larger and are by definition 1.5 times the IQR. For example for the variable Total.number.of.declines.day the point pertaining to 20 declines that occurred for a fraudulent transaction visually stands out however further researching and applying our judgement as a data scientist we see that this completely normal behavior for someone who has partaken in a fraudulent transaction. This behaviour has been studied on many accounts and comes from the fact that the fraudulent buyer will continue to try multiple transactions and keep getting rejected until they finally complete their fraudulent transaction. Hence we conclude this point is not an outlier. For the average.amount.transaction and transaction_amount plots in Figure 2 we also flag two points that are much larger visually. However once again using our judgement and the same rationale as before we conclude these are NOT outliers.

Table 1: Description of Variables in Dataset

Variable Name	Type	Missing Values	Outliers	Output Variable Classes
Merchant_id	Numerical	No	No	Our data is imbalanced wit h 2627 non-fraudulent case a nd 448 fraudulent cases.
Transaction date	Numerical	All Missing	-	
Average Amount/transaction/day	Numerical	No	No	
Transaction_amount	Numerical	No	No	
Is declined	Categorical	No	No	
Total Number of declines/day	Integer	No	No	
isForeignTransaction	Categorical	No	No	
isHighRiskCountry	Numerical	No	No	
Daily_chargeback_avg_amt	Numerical	No	No	
6_month_avg_chbk_amt	Numerical	No	No	
6-month_chbk_freq	Numerical	No	No	
isFradulent	Categorical	No	No	

In pursuit for optimally building our classifier, we must ensure that the data set is balanced with respect to the number of non-fraudulent and fraudulent transactions. In our data set there was a large class imbalance in which the fraudulent transactions are a minority class and due to this there is a larger likelihood of our classifier being bias and "ignoring" the minority class. Due to the scope of this report we have investigated three common ways to correct the class imbalance of the data set: Random Oversampling, Random Undersampling and SMOTE (Syntethic Minority Over-Sampling). Random Undersampling removes some of our valuable data and will not allow our classifier to have the broadest amount of training data. We are investigating a broad spectrum activity with many variations and removing data randomly may cause us to miss out on important information. Furthermore Random Oversampling may result to overfitting our model to our training data and does not give the broadest spectrum of training data to our classifier. It is important to understand that fraudulent and non-fraudulent behaviour is a wide spectrum of behaviour and it is important to try to account for the largest range of behaviours we can account for. Avoiding the two problems mentioned above, we use SMOTE in our report alongside the "ROSE" package in R to balance our data set. By using SMOTE we ensure that we do not loose any of our valuable data and overfit our model. However we may be introducing

noise into our data by having overlapping classes for some of the points. However as a data scientist the judgement of going ahead with SMOTE is valid because the noise introduced is a smaller problem to have then a overfitted model or the loss of information since as a data scientist we aim for generalizability of our model as a guideline. Once we now have a balanced dataset we proceed by splitting our dataset into a training and testing segments. Doing so allows us to critically asses our model and understand it's behaviour on non-trained data.

Planning

We wish to build a classifier which can model the type of transaction (fraudulent or not fraudulent) based on several characteristics that make up a transaction. We start our analysis by visually looking at Figure 1 and Figure 2 below which allows us to look at the individual relationships between each transactional characteristic versus the outcome variable and in turn generate a hypothesis from this. From both Figure 1 and 2, **we see our data is telling us a story**! In Figure 1 and 2 we see the larger the visual difference between the fraudulent and non-fraudulent transaction there is for an given variable, the greater effect that particular variable has in distinguishing a fraudulent transaction from a non-fraudulent transaction. Because of this certain transaction variables have a greater effect on differentiating a fraudulent from a non-fraudulent transaction. For example looking at the transaction variable "Transaction_amount" we see that on average transactions that have a larger transaction amount value correspond to fraudulent transactions. Hence we hypothesis that the Transaction_amount of a transaction effects weather a transaction is fraudulent or not. Repeating the same visual exploration of all the boxplots in Figure 1 and Figure 2 we hypothesis that Is.declined, IsHighRiskCountry, X6.month_chbk_freq, Average.Amount.transaction.day, Transaction_amount and Total.Numer.of.declines.day. Our hypothesis is:

H_0:	Weather a transaction is fraudulent or not fraudulent depends on Is.declined, IsHighRiskCountry, X6.month_chbk_freq, X6_month_avg_chbk_amt,Average. Amount.transaction.day, Transaction_amount and Total.Numer.of.declines.day
H_a:	Weather a transaction is fraudulent or not fraudulent depends on at least one of Is.declined, IsHighRiskCountry, X6.month_chbk_freq, Average.Amount.transaction.day, Transaction_amount and Total.Numer.of.declines.day

Figure 1: Relationship Between Categorical Variables and Transaction Type

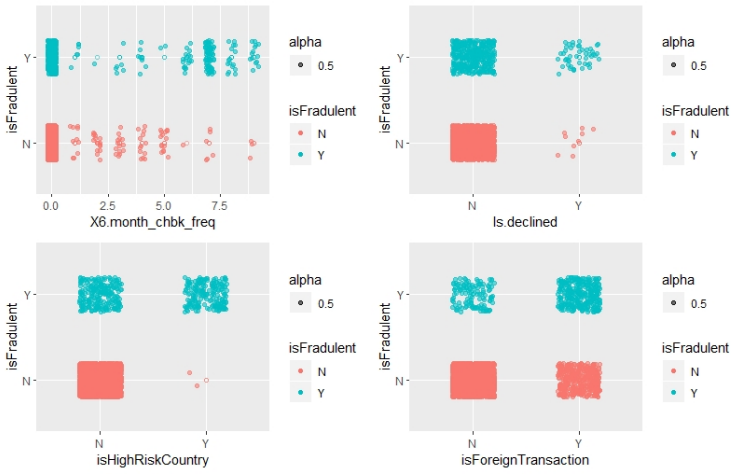
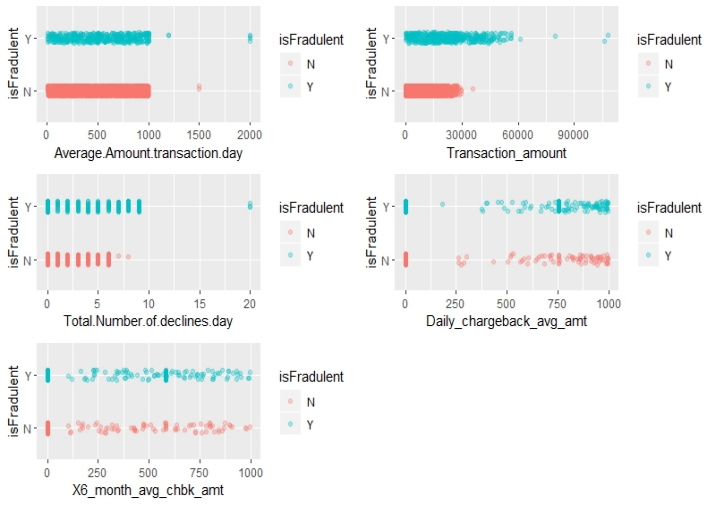


Figure 2: Relationship Between Numerical Variables and Transaction Type



Creating the Model

We will employ multinomial logistic regression to develop a classifier that produces a binary output (Fraudulent or Not-Fraudulent) based on several transaction characteristics as the input. We construct our model based on the hypothesis we generated above which allows us to build our model on the transaction characteristics that have biggest effect on differentiating between fraudulent and non-fraudulent transactions. Note we do not use the AIC to test the strength of our model and predictions, we will use a confusion matrix for this later on. In our first iteration of building our model we fit all of the variables we hypothesized to have an effect on type of transaction. Once our model is fit and built on these variables we further analyze the output of the model in R and remove any coefficients of variables that are insignificant in the model (based on the coefficient's respective p-value). After removing the variable with the insignificant co-efficient we build our model once again with the variables left. After building our new model, we reiterate the above process of removing insignificant variables and altering the model until we minimize our AIC value and make sure all of the variables in our model have a co-efficient in the R output that is significant. We went from building our initial model with the variables: Is.declined, Transaction_amount, Total.Number.of.declines.day, isHighRiskCountryY, X6.month_chbk_freq, isForeignTransactionY, X6_month_avg_chbk_amt and Average.Amount.transaction.day **to having only the variables**: Transaction_amount, Total.Number.of.declines.day, isHighRiskCountryY, X6.month_chbk_freq, isForeignTransactionY, X6_month_avg_chbk_amt and Average.Amount.transaction.day in our final model. Our final model had the lowest AIC value out all of the previous models. Furthermore to make sure our model is not effected by an Influential points we ensured that none of points had a cook's distances greater than 1. Looking at appendix Table B, we that there are no influential points because the highest cooks distance value is 0.12. The output of our final model is summarized neatly can be seen below in appendix Table B.

Model Validation

We initially segmented our data into both training and testing sets so that in our model validation procedure we test the model on the testing data set. Furthermore as data scientists it is important to not just report "numbers" and force a meaning on to them hence this is the reason we do not measure our model using prediction accuracy. A classifier with 99.99% accuracy can cost a company billons of dollars more than a classifier with 96% accuracy. Why? This is because accuracy does not give us an understanding on the type of inaccuracy or wrong predictions the model is making. For example it is less costly for a company if the classifier wrongly flags non-fraudulent transaction as fraudulent rather than the organization flagging fraudulent transactions as non-fraudulent. We as data scientists must critically ask questions such as "Given the transaction was truly fraudulent how many times did our model predict that the transaction is not-fraudulent?". The numeric answer to this question is known as "False-Negative". The very question asked previously is what saves companies millions and millions of dollars. We assess our model by looking at the "confusion matrix" which tells us the false negatives, false positives, true positives and true negatives of a models prediction. The interpretation of these numbers and the output of the confusion matrix for our testing data set can be seen in table 2 below.

In the financial technology world we are more concerned about false negatives then false positives as false negatives can cost companies a lot of money. Our model has predicted 95 false negatives, meaning that the 95 fraudulent transactions were not flagged as fraudulent. 95 false negatives is very small relative to the size of the dataset as it reflects only 3% of all cases. Therefore our model has performed very well in predicting false negatives. To further investigate the performance of our model we use the AUC metric which is area under the ROC curve which plots the True Positive Rate versus the False positive rate of our classifier. Refering to the ROC curve in Appendix Figure C, the ROC curve has an AUC value of 0.9868785 which is very close to 1. This means are classifier has performed outstandingly! To further interpret an AUC value of 0.9868785 let's dig deeper into values of the AUC. A model with an AUC of 0 means that 100% of the models predictions are wrong, in contrast a model with an AUC of 1 means that 100% of the models predictions are correct. Furthermore in binary classification we can achieve an AUC of 0.5 as the "baseline" from randomly guessing. Therefore we see that a AUC of 0.986975 means that our model is performing great. Because we are more worried about false negatives then false positives in financial technology we pay larger attention to the "Recall" statistic rather than the "Precision" statistic calculated below. The recall for our model is 0.9159 which means that there is 91.59% of cases that are fraudulent are predicted as fraudulent by our model which is also

outstanding. Therefore measuring our model performance on multiple metrics alongside the **context** of our real life scenario, we conclude that our model has performed outstandingly.

Table 2: Description of Confusion Matrix and Advanced Metrics

Metric	Interpretation	Metric Value
False Negative	95 times the model predicted a transaction to be non-fraudulent when the transaction was actually fraudulent.	95
False Positive	57 times the model predicted a transaction to be fraudulent when the transaction was actually non-fraudulent.	57
True Positive	1035 times the model predicted a transaction to be fraudulent when the transaction was actually fraudulent.	1035
True Negatives	1120 times the model predicted a transaction to be non-fraudulent when the transaction was actually non-fraudulent.	1120
$Precison$	The number of correctly predicted fraudulent transaction divided by the number of predicted fraudulent transactions is 0.9478.	0.9478
$Recall$	The number of correctly predicted fraudulent transaction divided by the number of actual fraudulent cases is 0.9159.	0.9159

Testing of Assumptions

To ensure our model can be generalized to other data sets other then the training data set we used we must make sure all of the assumptions of logistic regression are met.

Table 2: Assumptions of Multinomial Logistic Regression Model.

Assumption	Test	Is the assumption met?
Linearity	We must ensure that there is a linear relationship between the logit of the transaction type and all the continuous predictor variables.	**Yes**, this assumption is met. A linear relationship can be clearly seen when each continuous predictor variable is plotted against the logit of the transaction type variable. Please refer to the appendix Figure A for this plot.
Multicollinearity	We want to make sure that there is not a strong correlation between two or more predictor variables. This can be measured the VIF which tells is a metric that we can use to assess whether a predictor has a strong correlation with other predictors. The VIF should be below 10 for us to ensure that multicollinearity will not create a problem in our model.	**Yes**, all the VIF values for each variable are all way below our cutoff. The highest VIF is 2.081990. We conclude there is no collinearity in our data. Please refer to the appendix Table A to look at the VIF for each variable.
Independence of Errors	We must use the Durbin Watson test to check for the independence of residuals. We must make sure	**Yes**, at 5% level of significance the Durbin-Watson test returned d=1.43 which is very close to 2 which ensures

	the value of the durbin Watson test is around 2 to ensure independence of errors.	us autocorrelation does not occur. We validate this conclusion of independent errors by taking a look at a plot of the residuals and ensure that there is NO autocorrelation.
Incomplete Information	We must check if there are any missing values in the dataset.	**Yes,** there are no missing values in our dataset. This can be seen from Table 1 above.
Complete Separation	We must check if the data for all the variables overlaps. If there is no overlap of data in our model our logistic regression model will "fail".	**Yes**, looking at Figure 1 and Figure 2 above we see that all of the variables have data that overlaps and there is NO complete separation.

Interpretation of our Model

Our calculations in this report are based of an 95% confidence interval therefore referring to Appendix Table B and looking at the model coefficient's we are 95% confident that none of the coefficients in our model overlap with positive and negative values which means that the outcome of our model clearly supports our alternative hypothesis generated above. When interpreting logistic regression model coefficients we are looking at an increase or decrease in **probabilities** with respect to a unit change in predictor variable. Hence we can see a unit increase in any of the predictor variables with a negative coefficient results in a decrease in the probability of our outcome variable being close to 1(fraudulent transaction) and corresponds to increasing the probability that the transaction will be non-fraudulent. On the other hand we see that a unit increase in any of the predictor variables with a positive coefficient results in the probability being closer to 1, which corresponds to an increased probability of the transaction being fraudulent. Due to space constraints, we cannot interpret every model coefficient in this report however taking a look at the variable Total.Number.of.declines.day. The co-efficient of rthis variable is 6.695e-01, since this has a positive co-efficient this tells us that a unit increase in Total.Number.of.declines.day results in higher probability of the sample being fraudulent. Furthermore we can understand this change mathematically by looking at the "odds", the odds of a transaction being fradulent is equal to the probability of the transaction being fraudulent divided by the probability of the transaction being non-fraudulent. The odds ratio which allows us to mathematically interpret the change in odds do the a change in the predictor value. Continuing with the Total.Number.of.declines.day example we see that the odds ratio of this variable is 1.95 which means a unit increase in Total.Number.of.declines.day variable for a given transaction results in a 1.95 increase in the odds of the transaction being fraudulent.

Business Significance of Model

Speaking from an economic point of view a firms purpose is an economy is to maximize it's profit and a firms profit is revenue minus it's cost. Therefore we see that a firm making a decision of investing in financial technology must yield a positive outcome and maximize profit. After using the large data set collected we are very confident and impressed by our model performance as this model can be used in real time to detect fraudulent transactions and save the company millons of dollars. However interpreting our model from a business point of view and looking at the co-efficients of our logistic regression model we see that having an unit increase in the variables Transaction_amount, Total.Number.of.declines.day, isHighRiskCountryY, X6.month_chbk_freq, isForeignTransactionY, X6_month_avg_chbk_amt all increase the odds of a transaction being fraudulent. This strong result from our analysis makes sense as this goes hand in hand with the behaviour of many fraudulent purchasers. I proceed by giving real life meaning to our mathematical findings in Table 3 below:

Table 3: Business Significance of Model Coefficient

Varibale Name	Significance
Transaction_amount	A unit increase in transaction_amount increases the odds of transaction being fraudulent. This is because financial crime is done to maximize the money stolen

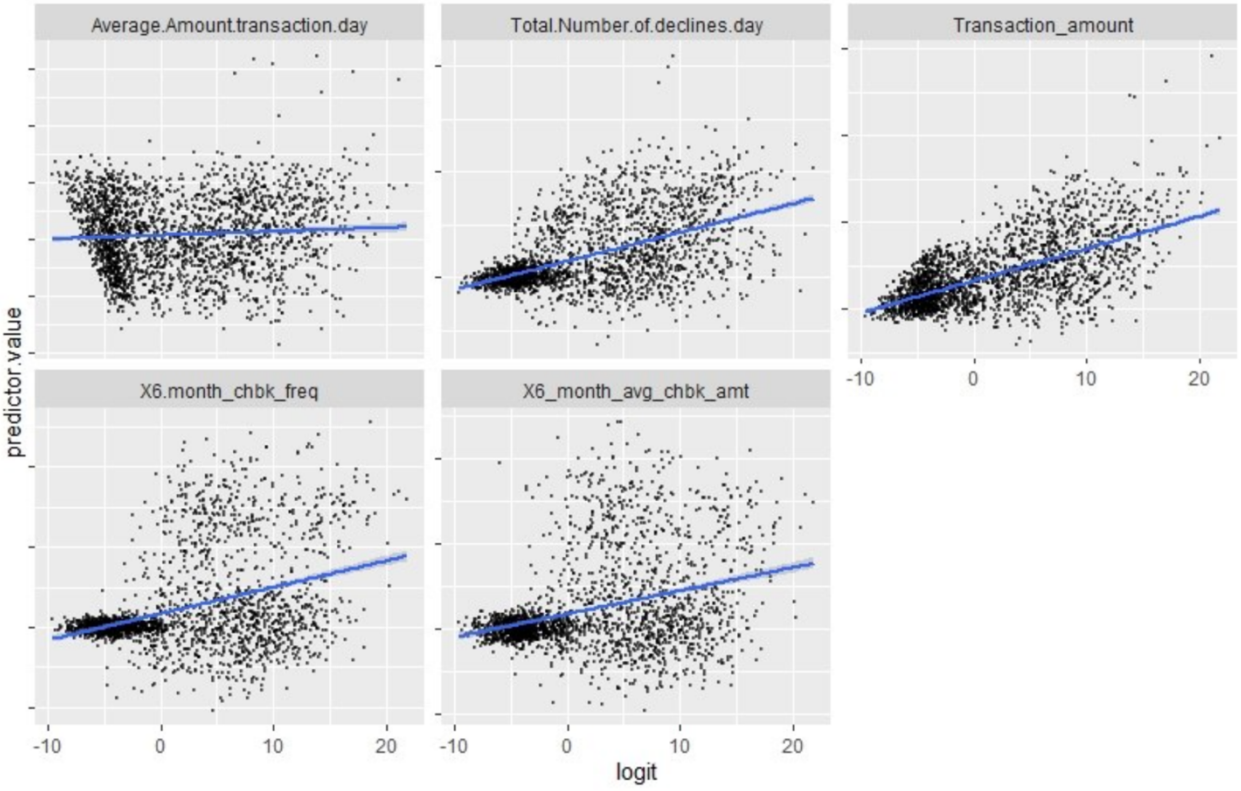
	from the innocent, therefore larger transactions are preferred by people committing financial crimes. As they get a much larger marginal benefit with the same marginal cost of committing financial crime.
Total.Number.of.declines.day	A unit increase in Total.Number.of.declines.day increases the odds of transaction being fraudulent. This is because these criminals attempt to do multiple transactions with multiple different companies in a single day. Some cybersecurity websites catch them and some do not. However someone who contains to keep getting decined without calling there bank is suspect to a financial crime.
isHighRiskCountryY	Many large crime organizations commit financial fraud in other countries as it is easier for them to get rid of there money this way and steal money from numerous countries. One reason researched is because the money invested in other countries to commit financial crimes are much more and therefore there is more action taken to steal. Therefore certain countries have a higher risk.
X6.month_chbk_freq	A unit increase in X6.month_chbk_freq increases the odds of transaction being fraudulent. This is because people committing financial crime will continue to check in a try to maximize the money they take out. The more they check in the more money they can take out if they do not get declined.
isForeignTransactionY	Many foreign transactions are done as a way to cover up money laundering and criminal activities therefore we must question in a global world what reason is the foreign transaction being taken place in this country.
X6_month_avg_chbk_amt	A unit increase in X6_month_avg_chbk_amt increases the odds of transaction being fraudulent. This is because as mentioned above criminals will log on multiple times in order to attempt to get the money out. Therefore they aim to get the maximum possible money out.
Average.Amount.transaction.day	A unit increase in Average.Amount.transaction.day decreases the odds of transaction being fraudulent. This is because criminal behaviour gets caught very quickly at one point and has a high turnover ratio. Therefore what we see in high credit card crime is days with large peaks of money stolen, followed by 0 money. Therefore the average is lower. Someone who lives a normal life will have a larger average as they are constantly spending on a daily basis.

Conclusion

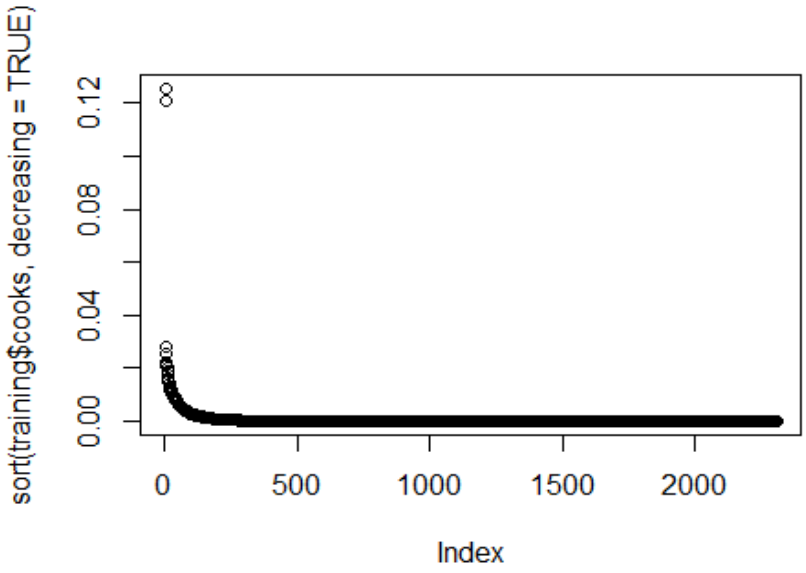
In conclusion from our analysis we see that our model performed outstandingly well on the testing data with a AUC very close to 1 and had a low rate of false negatives. Because of this we conclude that the generalizability of our model is present as it has been assessed on numerous metrics and we see that our model has the ability to identify patterns very well after viewing the ROC curve! We also conclude from our model and the meeting of all the model assumptions that Transaction_amount, Total.Number.of.declines.day, isHighRiskCountryY, X6.month_chbk_freq, isForeignTransactionY, X6_month_avg_chbk_amt and Average.Amount.transaction.day variables all have statistically significant effect on differentiating a fraudulent from a non-fraudulent transaction. Hence **we reject our null hypothesis** in favor of our alternative hypothesis at a level of significance of $\alpha = 0.05$. In addition we conclude that the variable Is.declined did not have an statistically significant effect on fraudulent transaction because anyones card can get declined but it's the frequency that matters. For further investigation, the financial institution should experiment using different cut off points or "maximum cap" for certain variables such as Transaction_amount and Total.Number.of.declines.day and use ANOVA analysis to see if this results if different cap control levels effect the mean fraudulent transactions. In future reports I will use more sophisticated balancing methods which will be made specifically for the financial technology fraud detection space. As clearly shown in this report, having a balanced data set is vital in binary classification.

Appendix

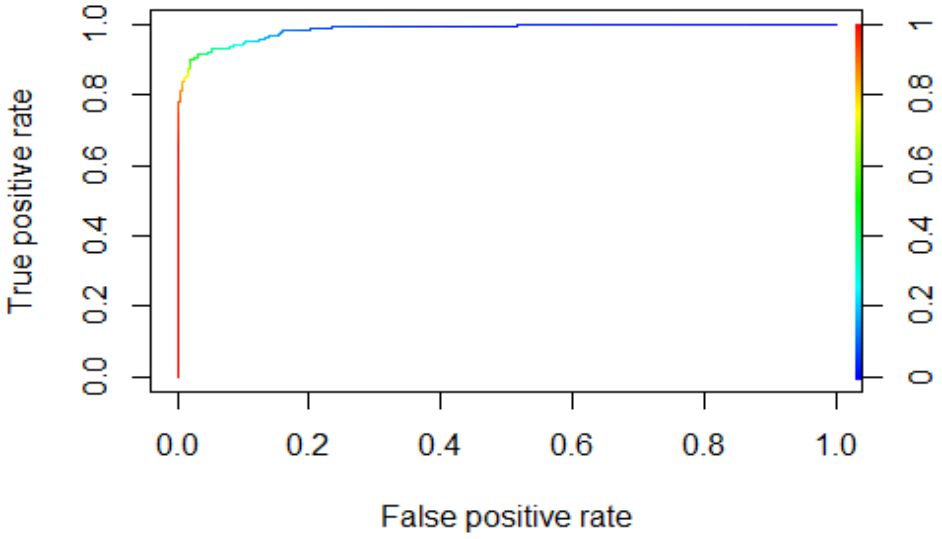
Appendix Figure A: Relationship between predictor variables and logit of the wine color variable



Appendix Figure B: Cooks Distance for points in data set



Appendix Figure C: ROC Curve for Multinomial Logistic Regression Model



Transaction_amount	Total.Number.of.declines.day	isHighRiskCountry	X6.month_chbk_freq	isForeignTransaction	X6_month_avg_chbk_amt	Average.Amount.transaction.day
2.081990	1.561443	1.119333	1.915404	1.377167	1.788145	1.851796

Table A: VIF for Each Predictor Variable

Table B: Logistic Regression Model Output

Coefficients	**Estimate**	**Odds Ratio**	**Std.Error**	**Z value**	**Pr(>|z|)**	**2.5%**	**97.5%**
(Intercept)	-4.436e+00	-	2.889e-01	-15.356	< 2e-16	6.543065e-03	2.034228e-02
Transaction_amount	1.841e-04	1.0001841	1.536e-05	11.983	< 2e-16	1.000155e+00	1.000216e+00
Total.Number.of.declines.day	6.695e-01	1.9532310	4.406e-02	15.196	< 2e-16	1.797786e+00	2.137269e+00
isHighRiskCountryY	6.842e+00	936.4306750	8.125e-01	8.421	< 2e-16	2.361630e+02	6.503976e+03
X6.month_chbk_freq	4.853e-01	1.6245821	5.683e-02	8.539	< 2e-16	1.458087e+00	1.822939e+00
isForeignTransactionY	2.895e+00	18.0782808	2.508e-01	11.540	< 2e-16	1.120399e+01	3.001378e+01
X6_month_avg_chbk_amt	3.335e-03	1.0033409	5.555e-04	6.005	1.92e-09	1.002268e+00	1.004459e+00
Average.Amount.transaction.day	-3.641e-03	0.9963652	4.627e-04	-7.86	3.57e-15	9.954349e-01	9.972446e-01
$\boldsymbol{AIC = 635.9}$							