STA 4350 Final Project

Alan Perez

05/20/2022

1. Introduction

Credit card theft and fraud has become some of the most lucrative crimes to have come out of the 21st century. That being said, it has also become one of the most expensive and difficult for banks to combat. My goal is to implement a model that will help this bank with its fraud detection. Specifically, I will explore the accounts of credit card customers who have had a recent decline (at least one decline in the last five transactions) and try to determine which accounts have been compromised.

This problem is important for the company because our success relies on our ability to keep customers happy and safe. If we fail to be proactive, customers will certainly look for a different company that is actively working on solutions and we will lose business. Above all, we want to minimize company costs.

2. Data

We begin our approach by first examining and summarizing the sample data. There are 7 variables of interest: fraud, gender, age, college, score, amount, and declines.

The fraud variable is an indicator for whether the account information has been compromised. If there are signs of fraud, then the observation is recorded as a 1. If there are no signs of fraud, it is recorded as a 0. From Table 1, we see that 62% of our sample data is not fraudulent and the other 38% is fraudulent.

Not Fraudulent	Fraudulent
0.62	0.38

Table 1: Frequency Distribution of Fraud

The gender variable is another indicator variable where the variable is coded as gender = 0 for males and gender = 1 for females. From the frequency distribution table (Table 2), we see that 47.2% of our sample data are males and the remaining 52.8% are female.

Male	Female
0.472	0.528

Table 2: Frequency Distribution of Gender

The age variable describes how old in years the account holder is. The youngest and oldest account holders are 18 and 95 years old, respectively. Our sample data has an average age of about 47.6 years and a median age 47.50. Table 3 and Figure 1 show the summary statistics of age and display no outliers.

Minimum	1st quartile	Median	3rd Quartile	Maximum
17	32	47.50	62	95

Table 3: Five Number Summary of Age

Figure 1: Boxplot of Age

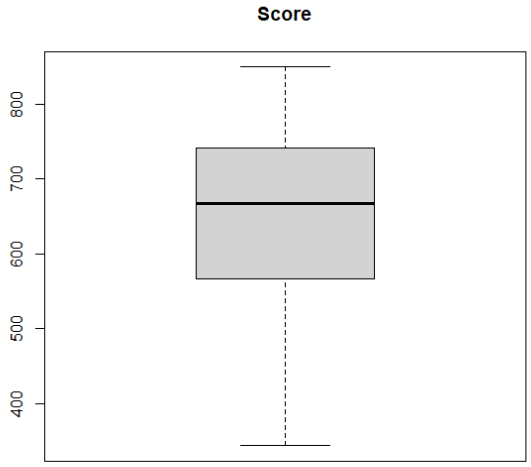

Figure 2: Boxplot of Score

The college variable tells us whether the account holder has obtained a bachelor's degree or higher (college = 1 for bachelors or higher and 0 otherwise). Table 4 shows that 54.3% of our customers in this sample have not achieved a bachelor's degree or higher while 45.7% have.

No Bachelors or Higher	Bachelors or Higher
0.543	0.457

Table 4: Frequency Distribution of Bachelor's Degree or Higher

The score variable tells us the account holder's credit score on a scale of 300 to 850, where higher values are better. The distribution of the scores in Figure 2 show no outliers beyond the minimum and maximum values of 344 and 849. The average score of the sample was 651.5 with a median of 667.

Minimum	1st quartile	Median	3rd Quartile	Maximum
344	566	667	741.2	849

Table 5: Five Number Summary of Score

The amount variable told us the average amount (in U.S. dollars) of the 5 most recent attempted charges. The minimum and maximum recorded amounts were $6.02 and $1761.90, respectively, and the average recorded amount was $430.33. Looking at the distribution in Figure 3, we see two obvious outliers at around the $1600 and $1700 mark. One of these outliers is our maximum recorded value.

Minimum	1st quartile	Median	3rd Quartile	Maximum
6.02	157.89	394.96	642.61	1761.90

Table 6: Five Number Summary of Amount

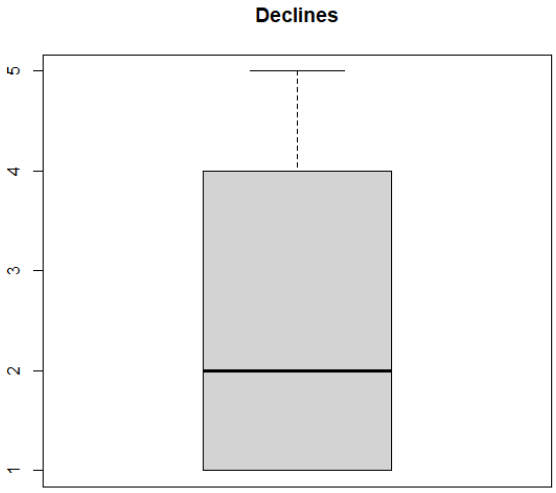

Figure 4: Boxplot of Declines

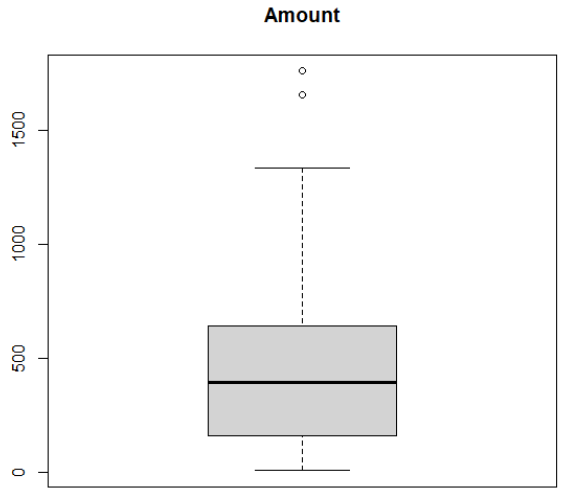

Figure 3: Boxplot of Amount

Lastly, our declines variable is the count of how many out of the 5 most recent charges were declined. Given that our values will always be between 1 and 5 (inclusive), it is no surprise that our minimum and maximum are 1 and 5 respectively. The average number of declines is 2.424, which we interpret as 2. There are no outliers in the distribution as seen in Figure 4.

Minimum	1st quartile	Median	3rd Quartile	Maximum
1	1	2	4	5

Table 7: Five Number Summary of Declines

Figure 4. Boxplot of Declines

3. Maximizing Prediction Accuracy

The overall goal is to find a way to predict whether a credit card account has been compromised by fraud as accurately as possible. To accomplish this, I used 6 different classification methods to see which one performed the best. The performance of each model was determined by the AUC value produced. To avoid overfitting, I implemented 10-fold cross validation in each method. In each of these models, the "fraud" is the response variable, and the others serve as predictor variables.

The first method I tried was a simple logistic regression model with all the predictor variables (gender, age, college, score, amount, and declines) as linear terms. Using cross validation and a threshold of 0.5, this model produced an overall accuracy of 66.8%. The AUC value obtained from the ROC curve in Figure 5 is 0.687.

The second classification method was a linear discriminant analysis (LDA) model. Using a threshold of 0.5, this model produces an overall accuracy of 0.669 – or 66.9%. The AUC value obtained from the ROC curve in Figure 6 is 0.688. s

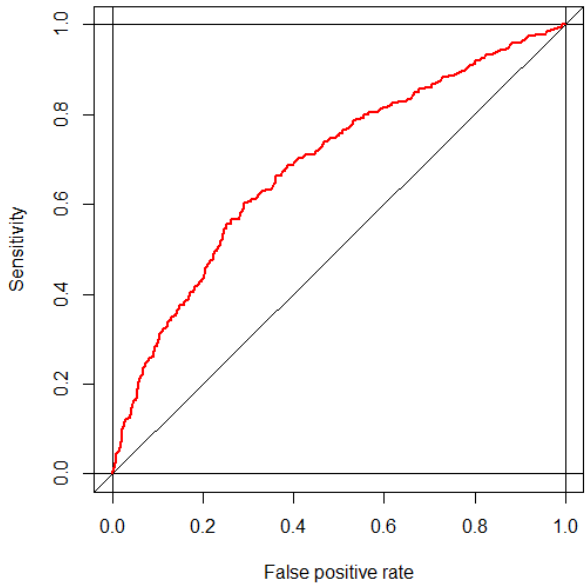

Figure 5. Logistic Regression Model ROC

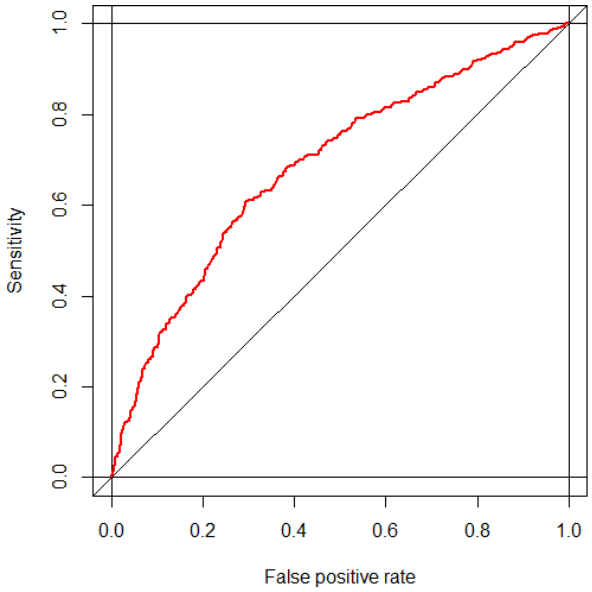

Figure 6. LDA ROC

The third method was a generalized additive model (GAM). The model accurately predicted 68.4% of the observations in the model using a threshold of 0.5. The AUC value produced by this model is 0.724 (Figure 7). The decision tree model with a threshold of 0.5 accurately predicted 62.6% of the observations and produced an ROC (Figure 8) with an AUC of 0.629.

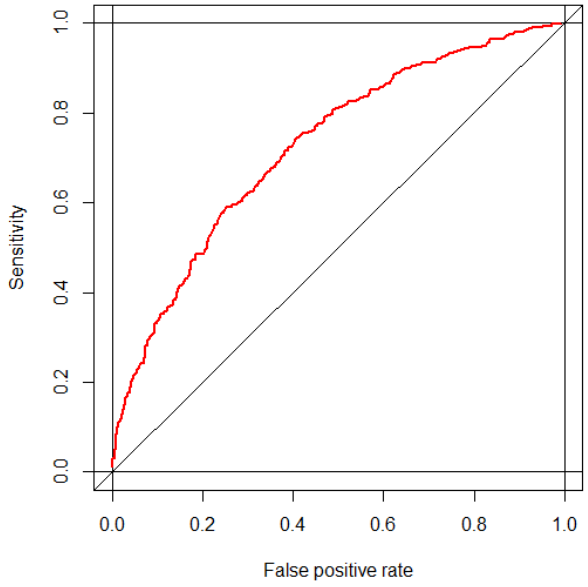

Figure 7. GAM ROC

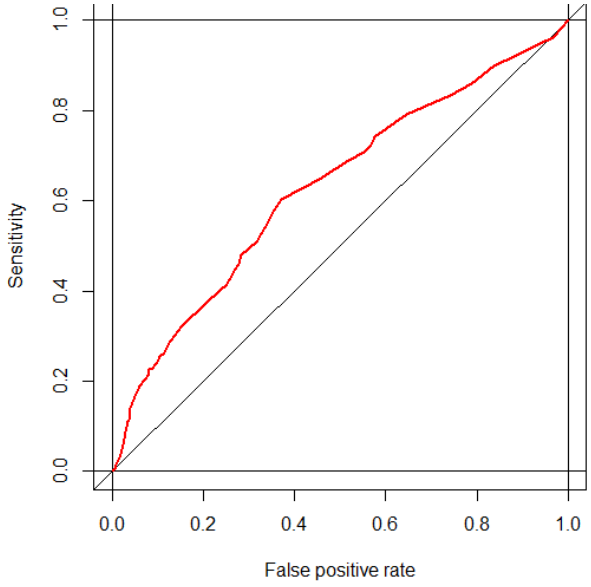

Figure 8. Decision Tree ROCs

The random forest model with a threshold of 0.5 accurately predicted 66.8% of the observations in the sample. It also produced and ROC curve (Figure 9) with an AUC of 0.674. The sixth and final method I implemented was an SVM model that accurately predicted 67% of the observations at a threshold of 0.5. The ROC curve it produced (Figure 10) had an AUC of 0.691.

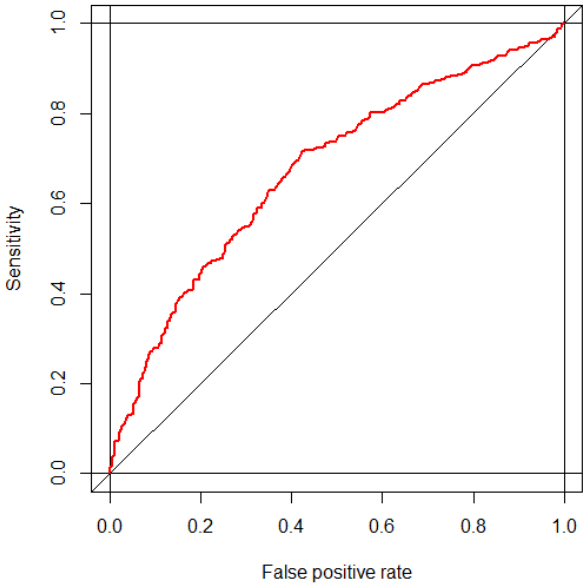

Figure 9. Random Forest ROC

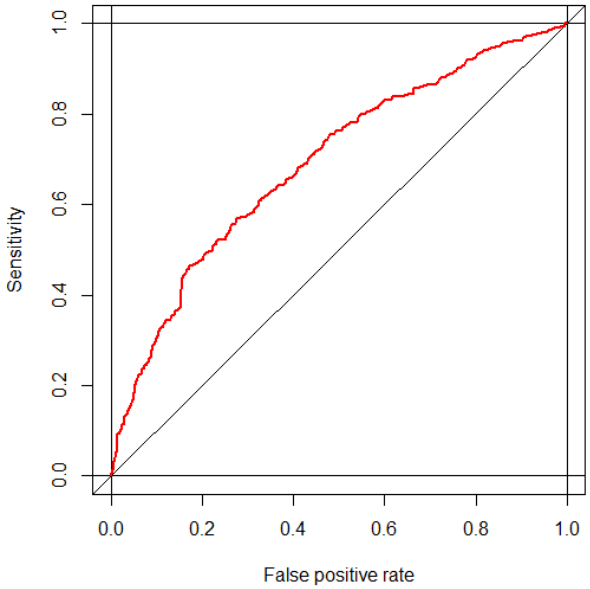

Figure 10. SVM ROC

Based on all the results presented above, the model that had the best predictive performance was the generalized additive model. At the 0.5 threshold, the overall accuracy of our model was 0.684; that is, our model accurately predicted 68.4% of the observations in our sample. The AUC value of 0.724 was the highest of all the models we looked at.

4. Minimizing Company Costs

As mentioned previously, the method that had the best predictive performance was the generalized additive model. Using this model, we will calculate the company's expected monetary loss based on the sample data using a few assumptions. First, we assume that suspending a customer's account that is not compromised by fraudsters will cost the company $25. The cost associated with failing to suspend a customer's account that has been compromised will be split into two different cases.

In our first case, we assume the company will be responsible for paying the sample mean transaction amount of $430.33 when we fail to suspend a fraudulent account. The first thing we must do is find a threshold value that will minimize company costs. To find this value, I tried every tau between 0 and 1 in increments of 0.01 and plotted the resulting cost per customer (Figure 11). As we can see in Figure 11, the lower tau values perform much better than the higher ones. Thus, since our results improve as tau approaches zero, I set our threshold value to

0. We do this to decrease the number of false negatives in our model since they cost the company significantly more money. In tandem with the first assumption, this method and threshold value produce an expected loss of about $15.50 per customer.

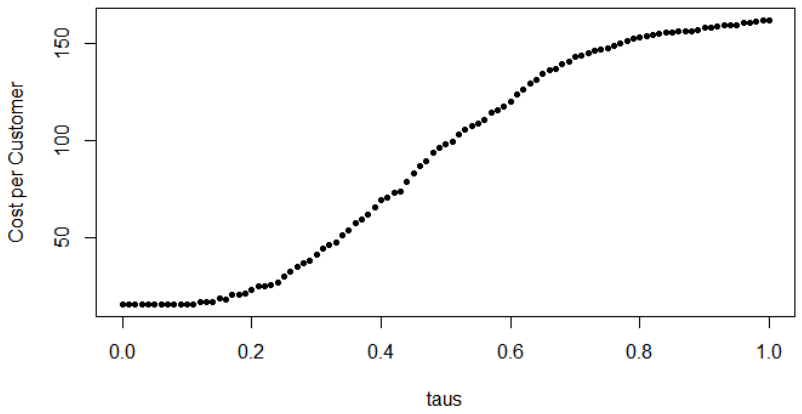

Figure 11. Costs Per Customer for Different Threshold Values (Assumption 1)

In our second case, we assume the cost for failing to suspend a fraudulent account will vary customer-to-customer and be equal to the mean transaction amount of that customer's 5 most recent transactions. Once again, we need to find the tau value that will minimize company costs, so we check the results for all tau values between 0 and 1 at increments of 0.01 again. Figure 12 shows us how different tau values perform and we see that lower tau values perform better overall. Once again, we set our tau value to 0 and we get an expected cost of $15.50 per customer.

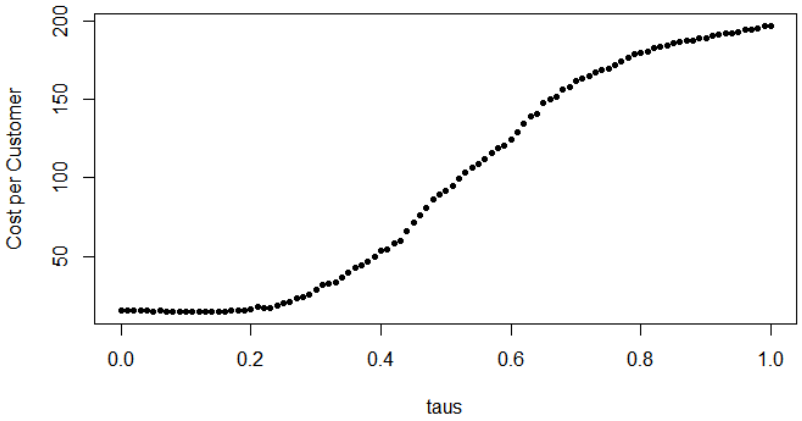

Figure 12. Cost Per Customer for Different Threshold Values (Assumption 2)

5. Summary

We successfully implemented and tested the predictive performance of 6 models. Of these, we found that the generalized additive model performed the best based on the overall accuracy and the AUC value it produced. Using that same model and the probabilities it produced, we calculated the company's expected monetary loss under two different circumstances. In both cases, one in which the company must pay $430.33 for each failed suspension and the other where the cost varies customer-to-customer, the company loses about $15.50 per customer. This makes sense since we are using a tau of 0 for both. That is, we are only coming up with false positives and thus are only adding $25 each time.

I must note something that was very interesting. It seems that in order to minimize company loss, it is actually better for the company to be overly protective of its customers. That is, the company will most likely perform several account shutdowns based on false alarms, but very few (if any) actual fraudulent accounts will remain open. Although this may come at the expense of the consumer's patience, it will both benefit the company and the consumer in the long run.

We can potentially improve predictions in the future by exploring the introduction of nonlinear terms in different models such as the logistic regression model. Some of the difficulties included choosing a proper threshold value for what we identify as a false positive and false negative. I solved this problem by testing a range of threshold values and plotting the performance of each one.

6. Implementation

If the boss wanted to manually check the results of a new data frame containing new observations called *new_obs*, they would simply run the code below:

```
set.seed(4350)
# K-fold
n <- nrow(cc_fraud)
k <- 10
fold_size <- n/k
folds <- matrix(sample(n, n, replace=FALSE), fold_size, k)

# Get the pi_hat values
pi_hat_gam <- numeric(n)

for (i in seq_len(k)) {
  # FIt the GAM excluding observations in fold i:
  my_gam <- gam(fraud ~ gender + s(age, 4) + college + s(score, 4) +
s(amount, 4) + s(declines, 4),
                subset=-folds[, i] )

  # Fill in phat values for observations in fold i:
  pi_hat_gam[folds[, i]] <- predict(my_gam, newdata=cc_fraud[folds[, i], ])
}
pi_hat_new_obs <- predict(my_gam, newdata=new_obs)
yhat <- as.numeric(pi_hat_new_obs > 0)
```

yhat