Transaction Fraud Prevention Strategy

Bradley Hupf*, Yuqing Pan*, Wenjing Wang*, Peng Zhao*

April 1, 2016

1 Background

With the growing number of fraudulent credit card transactions, a stable, efficient and accurate prediction model is needed. Both to reduce the costs to banks associated with fraudulent charges and to improve customer experience by detecting fraudulent charges quickly. Given data covering multiple months of transactions, we were tasked with building a model which is able to predict the probability of a transaction being fraudulent and locate the best cutoff for this probability.

2 Approach and model building process

Given about 80,000,000 past transactions and 69 potential factors (variables), our basic strategy is:

- Remove factors which we believed were unrelated or factors which only obtained 1 level
- For the remaining factors, proceed with variable selection and decide the importance of each.
- Build a classifier and conduct parameter tuning to obtain the final model.

2.1 Approach

Under this outline, our approach uses one of the most popular machine learning methods, random forests, to classify the fraudulent and non-fraudulent transactions. Random forest models can include a lot of factors and can automatically find the best model. During the model fitting procedure, we re-sampled many times, making the model more stable, as it becomes the average of multiple models. Since random forests can give the importance of each parameter directly, we compared the importance of each factor with the factors obtain from parameter selection using Lasso and (weighted) logistic regression to assure our model included the most important factors.

A random forest is a collection of decision trees. Given an observation, each decision tree votes for a class and the class with the most votes is regarded as the final result. Each tree conducts a binary split, and after each split, variables are re-sampled. In this way, the most important variables and the best fitted model will eventually be obtained.

[1]* Department of Statistics, FSU. Email: yuqing.pan@stat.fsu.edu

[2]Advisor: Dr. Xufeng Niu

2.2 Model building process

- **Step 1.**

  The first step is to clean the data. Among the given variables, we converted categorical variables into factors and deleted a few variables we thought were unrelated to detecting fraud. All factors which were date related (V4, V10, V17, V22, V27, V28) were not used in the decision tree since it does not make sense to include them, but these factors were kept in Lasso and logistic regression. But most of these factors were not selected in either Lasso or logistic regression. We also deleted identification factors for transactions (V1), accounts (V2) and terminals (V67) as well as variables which contained the same type of information but were less informative than another. For example, we deleted the variable V56 (date and time when a card was requested) as it is overlapped with V53 (date and time when plastic was first used) since knowing when a credit card is first used is related to when it was requested, but a card cannot have fraudulent charges until it has been compromised in some way. Furthermore, factors V17, V49, V61, V63 were also removed since they only obtained 1 level.

- **Step 2.**

  After data cleaning, we realized that it was necessary to re-sample since the fraud rate in the original dataset is below 1%. Therefore, the best model would be bias toward recognizing all observations as non-fraud. Considering the validation set ranges from September to October, we choose the most recent month in the training data, August, as the new validation set (N-validation) in the training process. We selected all the fraudulent transactions in the first 9 training sets and defined a parameter to be the sampling ratio, representing the ratio of number of fraudulent transactions to non-fraudulent transactions in new sampled training dataset (N-training).

- **Step 3.**

  We tuned the number of trees in the random forest between 10, 20, and 30 then used the bisection method to identify the best ratio from 1:20, 1:5, 1:15, 1:9, 1:13, 1:10, 1:11. For each combination, we used cross-validation to get the best random forest model, determined by having the lowest cost on the N-validation dataset. We eventually set the number of trees to be 20 and the sampling ratio as 1:10. The model which learned from the N-training dataset and was tested on N-validation dataset is used to obtain the ROC curve and the total cost as a function of fraction of transactions declined. We also plotted the total cost vs. cutoff on the N-validation to obtain the optimum cutoff value.

3 Result

We used our model on the two given validation datasets and obtained the probabilities of being fraudulent for each transaction. See attachment for detailed probabilities.

Since the true response was not included in the given validation sets, we evaluated our model on the N-validation set. The receiver operating characteristic (ROC) curve we obtained is in Figure 1. The area under the ROC curve is 0.84 and the fitted shape of the ROC curve demonstrates the effectiveness of our proposed model.

The classification model assigned each transaction a probability of being fraudulent. We did a cost-benefit analysis to find the optimized cutoff.According to Figure 2, the optimum probability threshold is 0.65, which means when the probability obtained from the classification model for a transaction is above 0.65, we decline the transaction. We can also get a curve showing the total cost to Capital One as a function of the fraction of transactions declined, illustrated in Figure 3. The total cost is minimized to be $433032.92 when the fraction of transactions declined is 0.15325032%.

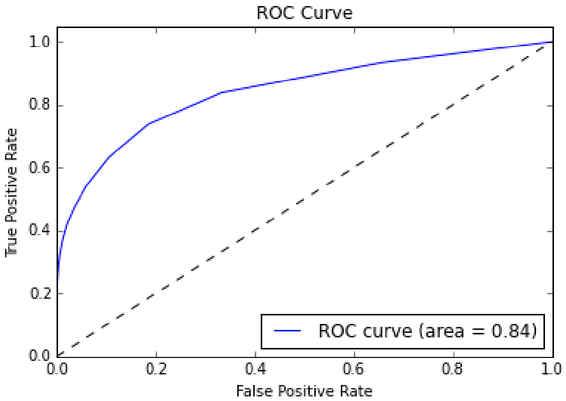

Figure 1: ROC curve. This ROC curve is obtained from the model on N-training and tested on N-validation.

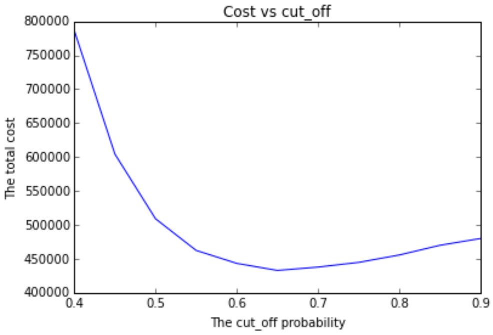

Figure 2: Cost vs. Cutoff.

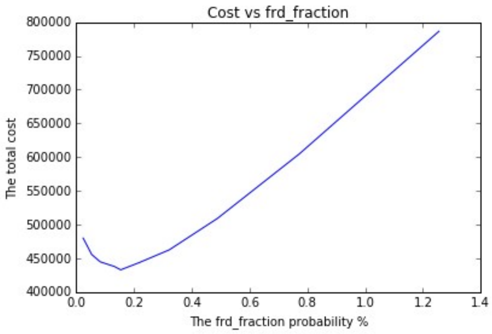

Figure 3: Cost vs. Fraction.

4 Texture interpretation

We can place all variables into one of three categories: card holder related, current transaction related, or card history. In the random forest model, we obtained the importance for each variable and our model selected the variables which represented these three categories. Among the 69 variables, as expected, many of them have low importance, which means only a few variables dominate the classification, as shown in Figure 4. We chose the top 11 variables which had the highest importance, which uses an importance cutoff close to 0.4. Detailed interprepation is listed in Table 1.

In feature selection conducted through Lasso and Logistic regression, V5, V6, V8, V15, V19, V54, V69 were also selected. These selected variables match our intuition and explain incidences of fraud well. They also give us a general idea on how to reduce fraudulent transactions from occurring, by educating customers about these important variables, which could further reduce costs associated with fraud.

Table 1: Significant variables

No.	Importance	Explanation
V5	0.0441	The cash available money on the account. Fraud can also occur due to a fraudulent cash withdraw.
V6	0.0628	The available money on the account.
V7	0.0392	Current credit limit on the account. Fraudulent transactions are more likely to happen to accounts with low credit limit.
V8	0.0569	Current balance of the account at the end of posting. Intuitively, if a large transaction takes place which leaves the account with a low balance then there is the threat that the transaction was fraudulent.
V11	0.1192	The product code. This verifies the Capital One unique product code works well, as some fraudulent transactions can be found through the product code.
V15	0.0394	Total number of authorizations. This is kind of related to V54 as older cards are more likely to have a higher number of total authorizations.
V19	0.0653	Transaction amount. Fraud may happen when the transaction amount is extreme.
V24	0.0616	Authorization outstanding amount. Outstanding amount can also be used to evaluate the credit level of a customer. Further, previous fraudulent transactions may lead to a high outstanding amount, which is helpful to identify future fraudulent charges within a short length of time.
V45	0.0457	A code which identifies a specific merchant. Since the original data contained many categories, we combined them into three categories, entertainment, traveling and other, by SIC code.
V54	0.0606	Duration in days since the plastic was issued. This seems intuitively important as a new card may not be as safe as an older one.
V69	0.0883	Approximate distance of customers home from merchant. Transaction which take place far away may have taken place through the phone or on the internet, this may result in a higher probability of fraud as no face-to-face contact was made. Further, if a transaction did occur in person the customer may be less aware of the transaction as they are in a foreign region.

5 Summary

We used a random forest algorithm to classify each transaction. Parameter tuning through the bisection method based on sampled training dataset and validation dataset, guarantees the minimization of the total cost.

Further, we also identified important variables used in the prediction of fraud, which may offer a new point of view on how to further reduce the occurrence of fraudulent transactions.

Random forest is simple but effective. Since it is simple, it is easy to understand and interpret which prevents the model from being applied in an incorrect manner. And this model is also resistent to overfitting given the natural property of random forest, which completed resample during training process.

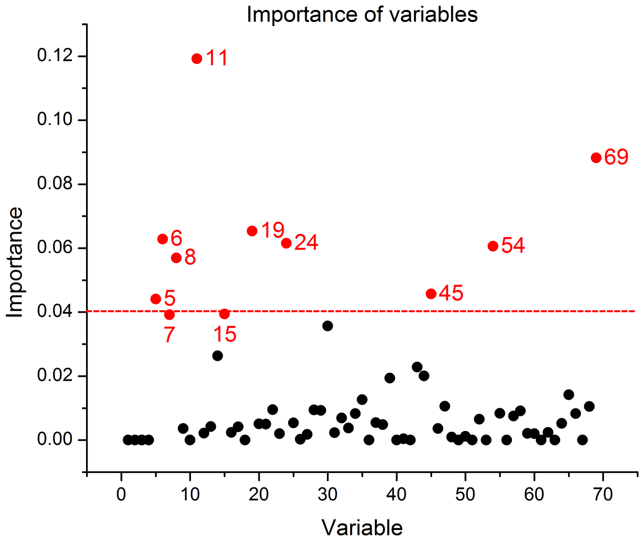

Figure 4: Importance of variables. We assumed the importance of removed variables is 0. The dashed line represents the importance level of 0.04. Red Variables are regarded as important.

Given a new transaction, we are able to predict the probability of it being fraudulent and apply the cutoff to decide if it should be declined or not. Since our optimized cutoff is obtained from the latest historical transactions, it can be updated between payment periods of credit cards. Further, we have found that we do not need to use all of the non-fraudulent transactions in the data to update the cutoff, which will save on data storage and computation time. Thus, our model can be easily and effectively updated to further reduce total cost over time.