Credit Card Fraud

Abe Durrant

11/29/2021

Abstract:

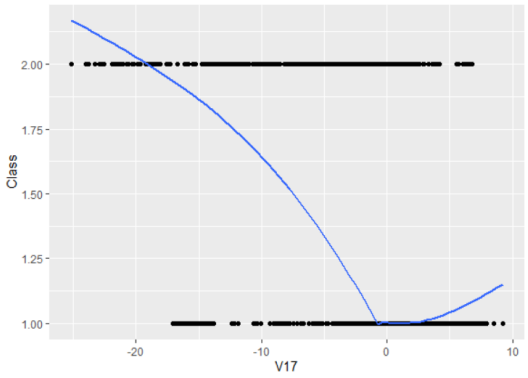
Credit card fraud is an issue costing people billions of dollars annually. This analysis takes a dataset of approximately 300,000 credit card transactions and builds a model which can predict what transactions are fraudulent. A random forest model was developed which can identify about 85% of fraudulent transactions when they happen. It was used to identify among new data which transactions may be fraudulent.

Introduction

Credit card fraud is a very important issue that costs consumers around $22 billion annually. For this reason, it is important to identify which transactions are fraudulent so that a company can either quickly deal with them or can prevent them from going through. This analysis will focus on analyzing a dataset that includes about 300,000 credit card transactions and attempting to create a model that can identify fraudulent transactions.

While the dataset in question contains approximately 300,000 transactions, only 492 of them are known to be fraudulent. This means that we can't simply look at how accurate the overall predictions are for a model, as if it predicts every transaction to be non-fraudulent then it will be correct 99.9% of the time. We will instead have to look at specific metrics to decide how our model is doing. The most important metric will be the sensitivity of our model, or rather the percent of true positives. That is, of all the transactions that are labeled as fraudulent, what percent can our model identify? We are also interested in the specificity, although not as much. The specificity will tell us the percent of non-fraudulent transactions we labeled as such. While we mostly want to identify fraudulent transactions at a high rate, we also don't want to label much more than a very small percent of non-fraudulent transactions as fraudulent as this will take time and resources. Finally, we can also look a little at positive predictive value, or rather what percent of the transactions that we labeled as fraudulent are actually fraudulent.

A further look at our data shows we have 29 different variables, 28 unlabeled ones as well as the amount of the transaction. A quick look through the data shows it is hard to determine if there are any strong relationships between the variables and whether or not there is fraud. A look at the variable V17 and whether or not there is fraud shows a potential weak relationship:



Here the 2 represents fraudulent transactions and the 1 is non-fraudulent transactions. We can see that when V17 is higher that the proportion that are fraudulent goes down. However, it seems to start going up again as it goes above 5. This is a potential issue of monotonicity if we are using a logistic regression.

There are a couple of potential other issues we can run into with the data. We have a lot of observations as well as 29 variables, so it is hard to check through any assumptions we may have such as monotonicity. Performing variable selection through many algorithms will also take a long time which means that it will be harder to narrow down which variables are important to keep in a model. We also simply have a large dataset so the more computational heavy the model we use the longer it will take to run and tune. If we don't account for these issues, then we may have a model that is not as optimal as it could be in identifying fraud. While we aren't really interested in statistical inference with this dataset, if we were to use a logistic regression model with violated assumptions then we may draw conclusions that aren't completely accurate, and a lack of monotonicity means our model is both doing poor, and the coefficients can't be interpreted.

The goals of this analysis are very simple. We want to be able to identify fraudulent transactions as best as possible. We aren't as interested in how a model does this as we are in how well it does this. As stated earlier, we are most interested in what percent of the fraudulent transactions we can identify. However, we also want a model that is not identifying large numbers of normal transactions as fraud.

Methods and Models

To begin developing a model many different methods were explored. After splitting the data into a train set to develop the models and a test set to test them on, multiple models were tested. Logistic regression, a simple neural network, a K-nearest neighbors' model, a gradient-boosting model, a random forest model, discriminant analysis, and support vector machines were all tested on this data. After this first test, random forest, discriminant analysis and support vector machines did the best, so they were explored and tested more thoroughly to find the best model. This are the three models that will be discussed including the strengths and weaknesses of each in regard to this analysis.

A random forest model can be a strong model for this analysis as it doesn't rely on any strict assumptions. Thus, we don't have to worry about any collinearity or non-monotonic variables. It also performs variables selection on its own. However, it can take a fairly long time to run if a large number of trees are used and it can often overfit the data if not tuned and cross-validated properly. Discriminant analysis can be strong for many of the same reasons such as no assumptions of monotonicity. It runs much faster even on a large dataset which is something that must be considered. Quadratic discriminant analysis is the specific type that we are using and can sometimes have a weakness if categorical variables or used or if the data can't be modeled in a quadratic way very well. Finally, support vector machines also have many of the same strengths that account for the issues of this data. They can be relatively flexible if we change the basis function (in this case we used polynomial) and don't rely on strict assumptions. However, they do take a while to run. All of the models being considered aren't the best choice if we were wanting to gain insight into which variables indicate fraud the most.

All of the models can help accomplish the goals as they can be very strong at predicting data. They also can be tuned to increase the specific metrics we are interested in. None of the models really on any explicit assumptions although our support vector machine will attempt to use a polynomial kernel to classify the data. We are assuming that this type of fit would be similar across any new data. It is reasonable to assume this.

Model Selection and Justification

First, to see how well each model fit the dataset as a whole, they were fit to the entire dataset and then predicted the entire dataset. From these predictions we can look at their specificity, sensitivity, and positive predictive value. We can also look at the area under the roc curve which is a good overall metric for evaluating fit. These numbers can be seen in the following table:

Model	**Sensitivity**	**Specificity**	**PPV**	**AUC**
Random Forest	85.34%	99.91%	54.77%	.9261
QDA	86.15%	98.41%	8.57%	.9228
SVM	86.76%	100%	100%	.9338

Of the models the support vector machine fits the data best across all metrics. A sensitivity of 86.76% means that it is identifying 86.76% of the fraudulent transactions correctly and its specificity and positive predictive value of 100% means that it never identifies a non-fraudulent activity as fraudulent. This is pretty great; however, we must cross-validate on new data to make sure it isn't simply overfitting on the dataset. The random forest does pretty well overall with a similar sensitivity, although from the other metrics we see that it occasionally identifies a non-fraudulent transaction as fraud. The QDA does well identifying fraud, but it is too often predicting non-fraudulent transactions as fraud.

To analyze how well the models predict new data we can look at the sensitivity, specificity, and positive predictive value after performing cross-validation. Due to the large dataset, 5 k-fold cross validation was performed on each along with some tuning of the parameters such as the number of trees on the random forest model. The following table shows the results:

Model	Sensitivity	Specificity	PPV
Random Forest	87.59%	99.86%	54.91%
QDA	84.78%	98.56%	8.76%
SVM	72.98%	99.98%	92.23%

As we can see when it comes to predicting new data, the random forest model does the best. It actually had a slightly higher sensitivity on new data and roughly the same metrics overall. QDA performed similarly and was very poor in identifying only the true fraudulent transactions. The support vector machine model did much worse on new data meaning that it was likely overfitting the original dataset.

Since the random forest model is performing the best across the metrics on predicting new data, which is what we care about most, we will use it to answer the questions and to predict whether transactions are fraudulent.

The random forest model is specified as the following:

$$\hat{p}(x_0) = \frac{1}{B}\sum_{b=1}^{B} \hat{p}^{b}(x_0)$$

B represents the number of bootstrapped samples that we take from our model, and the p-hat of each b is the probability obtained from that particular classification tree of whether or not the transaction was fraudulent. The average of these estimated probabilities is taken which makes our random forest rather robust. It is then compared to a threshold value (0.04 in this case determined from cross-validation) and if the probability is above that then the transaction is labeled as fraudulent. Our particular random forest model that is being used has 1000 different trees. This was determined in cross validation when looking at 50, 100, 500 or 1000 trees since it took so long to run.

The random forest model runs under no assumptions of the data and will determine through tree methods which variables it should use. The parameter of 1000 trees was determined in cross validation.

Results

The following are the 5 most important variables in our model:

Variable	Importance
V14	.00093
V1	.00070
V12	.00058
V17	.00053
V10	.00044

These importance values don't have a strict interpretation in context of the problem, we simply interpret it as the higher the importance value the more that variable was used in determining whether a transaction was fraudulent. We can see that there isn't one specific variable that is much more important than the rest. The random forest model doesn't return uncertainties with these importance values, but they can vary depending on the trees used so they shouldn't be taken as hard values.

To answer the first question of how well we can identify fraudulent transactions, we can look at the sensitivity of the random forest model. As listed in the model selection section we see that we were able to identify 87.59% of fraudulent transactions in our test data in our cross validation. From the overall dataset we identified 85.34%. So, given that a data is fraudulent, we should be able to identify it around 85-87% of the time. The range of the sensitivity across the 5 cross validation studies was 85.8% to 89% so we can expect any new data to fall in that range as long as it was collected in the same way. The standard deviation was about 1.6% across the five sensitivities meaning they were typically about 1.6% away from the mean of 87.59%.

Using the model on the five unknown transactions that we were tasked with assessing; we predict that the 3rd transaction is fraudulent. While the model gives some probability of the 1st and 5th being fraudulent, it is below the threshold determined in cross validation of 0.04.

From our results we find overall that we are able to identify fraudulent transactions a high percentage of the time. Of the transactions that we do think are fraudulent, about 55% are on average so we can reasonably take action to resolve the problem. We also identified around 85% of actual fraudulent transactions. Overall, our results from our model give us reasonable confidence that we were able to correctly identify a fraudulent transaction (the third) among the 5 transactions given to us.

Conclusions

Overall, we were able to meet the goals of the analysis by creating a model that identifies fraudulent transactions among credit cards a high percentage of the time. This model can be used to save consumers money and quickly identify fraud when the given variables are available.

Our model had some shortcomings in that it will sometimes identify non-fraudulent activities as fraud. However, this is only about 0.1% of the normal transactions, so it isn't a common occurrence. It also only identifies around 85% of actual fraudulent transactions meaning that it will sometimes miss a fraudulent transaction which could end up costing a consumer and potentially the company money.

If the company wishes to take further steps in identifying the transactions that are fraudulent, then committing more computer power and time to the project would be very important. Some methods were not fully explored as it would have taken too much time and computational power to fully tune them. A deep neural network might outperform the chosen model; however, a personal laptop and a week of time limit the amount of tuning that can be done to a neural network or some of the other more computationally heavy models.