Homework 5 Report

Shae Sims, RJ Cass

November 2025

Abstract

Credit card fraud causes billions of dollars in losses for consumers across the world. We used a data set with 300,000 credit card transactions to help identify what factors could be linked to fraudulent charges. Using over- and under-sampling and a bagging model we found that this model has a predictive sensitivity of ∼97.2%, and a predictive precision of ∼90.5%. Using this model, we predicted that transaction ID 3 in the provided dataset is fraudulent.

1 Introduction

Fraudulent credit card charges cost the financial industry billions of dollars in loses every year: being able to correctly identify fraudulent transactions is a vital aspect of this industry. As such, we want to use existing data to create a predictive model to indicate whether or not a transaction is fraudulent. The data we are using is a record of transactions that includes indicators of whether a transaction is fraudulent or not, the amount of the transaction, and the Principal Component Scores of 28 associated proprietary variables.

1.1 Research Questions

We want to use these data to answer the following questions:

How accurately can you identify the fraudulent transactions? In other words, given a transaction is fraudulent, how well do you identify it as such?

To answer this, we will create a predictive model and measure the sensitivity (given a transaction is fraudulent, how likely is it to be predicted as fraudulent) and also consider the precision (given the model predicts a transaction as being fraudulent, how likely is it to actually be fraudulent).

Which of the provided uncategorized transactions, if any, do you think are fraudulent?

To answer this, we will use the model created for the previous question and apply it to the uncategorized transactions and show whether the model indicates if they are fraudulent or not.

1.2 Data Exploration

The primary feature of this dataset is that the outcome variable (fraudulent yes/no) is a binary variable. Any model employed will need to work with binary outputs, and continuous inputs. Of the $\sim$ 300k transactions in the data set, less than 500 of them are known to be fraudulent, creating a large imbalance in the dataset. If the model does not account for that imbalance it will not predict accurately (it may just say everything is non-fraudulent, or vice-versa).

Another feature of the data is that the input variables are not all monotonic (constantly increasing or decreasing) as related to the proportion of values that are fraudulent. A sample of these variables is shown in Figure 1. As such, the model will either need to transform the variables, or not be dependent on monotonic relationships. If this is not accounted for, the relationships between the input and output variables will not be properly addressed and the predictions will be inaccurate.

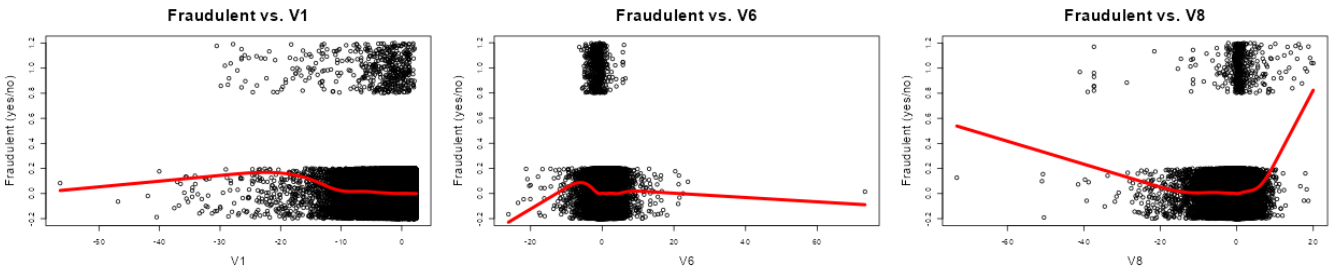

Figure 1: A sample of the monotonicity plots. Note that in each, the red line is not constantly increasing or decreasing

2 Methodology

2.1 Proposed Models

Model 1

The first model considered is a bagging random forest model. The basic unit of a bagging model is a decision tree, which is a model that reduces the data into binary decisions (ie. IF variable 1 is above X, AND variable 2 is below Y, AND ... then the transaction is fraudulent). An example tree is provided in Figure 2. The bagging model takes a random sample of the available data and generates a tree, and repeats that process many times to make a large number of separate trees. When predicting, the bagging model has each tree make its own prediction, then uses the average value across all trees to predict if that transaction is fraudulent.

As mentioned previously, the model also needs to account for the imbalance in the number of fraudulent responses. As such we conducted a small cross-validation study where we up-sampled (artificially increased the sample size of) the fraudulent responses and down-sampled (artificially reduced the sample size of) the non-fraudulent responses at varying ratios. In considering parameters for this model, we prioritized the balance of sensitivity and precision, so that we can identify as many fraudulent transactions as we can, without flagging an unnecessarily large number of non-fraudulent transactions. We tested various combinations of up/down-sampling quantities, as well as cutoff values (what percentage of trees need to vote 'fraudulent' for us to consider the transaction to be fraudulent). From these tests we determined the most balanced sensitivity and precision came from down-sampling non-fraudulent transactions to 60k, and up-sampling the fraudulent responses to 1.5k with a cutoff value .5 (ie. moving away from these values only marginally increased one metric while significantly decreasing the other). The resulting metrics are provided in Table 1.

This model will allow us to answer the research questions by predicting whether a given transaction is fraudulent. It will also allow us to measure how accurately it predicts. It does provide overfit protection (not prevention), but does not handle biased data very well. This model does not have any inherent assumptions so we do not need to worry about the monotonicity of the explanatory variables.

Model 2

For model 2, we used the Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalance in the dataset. This method helps reduce overfitting compared to simple oversampling. SMOTE works by creating synthetic minority-class data points using the k nearest neighbors. Some weaknesses of this approach are that it has difficulty handling high-dimensional data—since it relies on nearest neighbors—and it is sensitive to noise in the dataset.

After generating the new dataset, we applied boosting. Boosting is a modeling method that first fits an initial tree and then fits subsequent trees on the residuals of the previous one. It then averages over these trees, allowing the model to capture more of the nuances in the data. To prevent overfitting, we tested multiple parameter settings and monitored the model's performance using cross-validation. Specifically, we compared out-of-sample predictions to in-sample results to ensure that accuracy did not drop significantly when applied to unseen data. We ultimately selected 1,500 trees, an interaction depth of 5, and 10-fold cross-validation (these were the only parameters adjusted in this model). With this model, we can identify which features are most important for predicting fraud and evaluate

how well the model performs, with results shown in Table 1. We are also able to predict whether new observations are fraudulent, meaning this model answers both of our research questions.

This model does not rely on strict distributional assumptions, so monotonicity in the dataset is not a concern. Some strengths of boosting are that it focuses on misclassified points, which can help with imbalance. However, weaknesses include that it is sensitive to outliers, since it is focusing on the errors of the previous fits.

2.2 Model Evaluation

To evaluate the performance of each model, we examined their sensitivity and precision. We selected these metrics because they allow us to compare how effectively the models predict fraud both in and out of sample. For the values shown in the table below, the same cutoff of .5 was applied to the predicted probabilities from the bagging model. This threshold was chosen after testing multiple options, as it provided a reasonable balance between false positives and false negatives. However, this cut off can be adjusted to fit the companies need.

	In-Sample		Out-of-Sample	
Model	Precision	Sensitivity	Precision	Sensitivity
1	.9927	.9987	.9051	.9715
2	.9856	.9995	.8028	.9997

Table 1: Comparison of the precision and sensitivity of each model, both In-Sample and Out-of-Sample

Due to our research questions, we are primarily interested in the out-of-sample sensitivity of the model, as this best reflects how accurately the model identifies fraudulent transactions. At the same time, it is important that the model does not flag too many legitimate charges as fraudulent, since that could create issues for both the company and consumers. Considering both metrics, Model 1 was selected to address the research questions because it offers higher precision without sacrificing much sensitivity.

Model 1 is also more straightforward to interpret and the method of up- and down-sampling to handle class imbalance is simpler to explain. Despite its relative simplicity, it retains the predictive power needed to answer our questions. The model aggregates multiple decision trees and averages their predictions; an example of one such tree is shown in Figure 2. As noted earlier, this approach does not rely on any assumptions and incorporates all available variables.

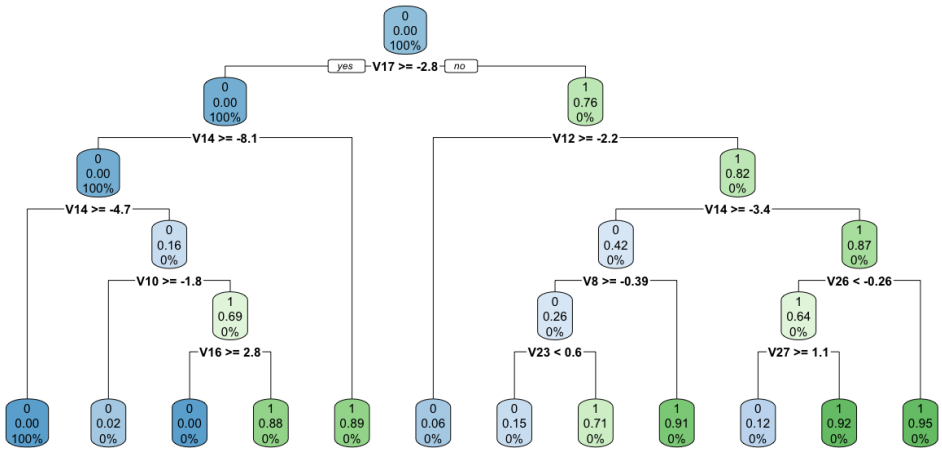

Figure 2: Example of a Tree that could be produced during the bagging process

3 Results

As shown in Table 1, Model 1 has a sensitivity of .9715 meaning that, given a set of fraudulent transactions, the model will predict ∼97.2% of them to be fraudulent. It is however important to note that the model has a precision of .9051; this means that if the model predicts a transaction to be fraudulent, there is a ∼9.5% chance that it is not actually fraudulent.

Using Model 1, we determined the variable importance for all variables included in the model. This analysis helps identify which variables are most influential in predicting fraudulent charges. The top 10 most important variables are shown in Figure 3, while the complete list is provided in Table 2 in the Appendix. It is important to note that variables not listed among the top 10 still contribute to the final model's predictions. Looking at this we can see that the variable with the most impact on if a charge is fraudulent is variable 14.

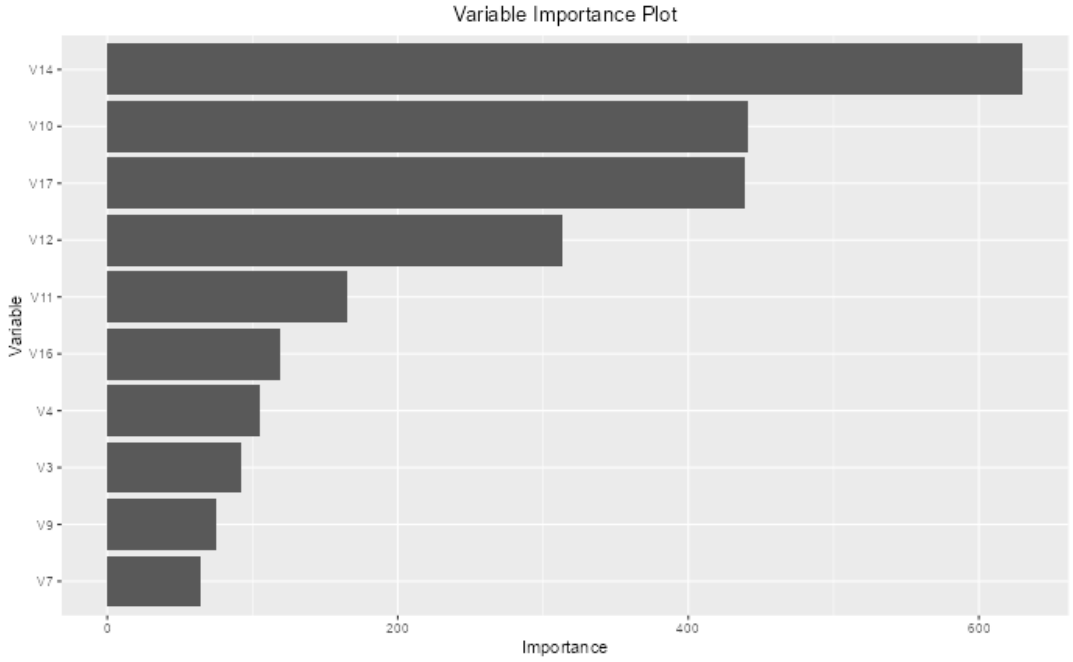

Figure 3: The top 10 variables in terms of importance in determining if a transaction is fraudulent

For predictions on unseen data, our model identified 1 out of 5 transactions as fraudulent. This was the third transaction (column X has the id '3') in the dataset, which involved a charge of $1.00. The plots below highlight the fraudulent transaction in red. As shown, this transaction is separated from the others when comparing variables 14 vs. 10 and variables 17 vs. 12.

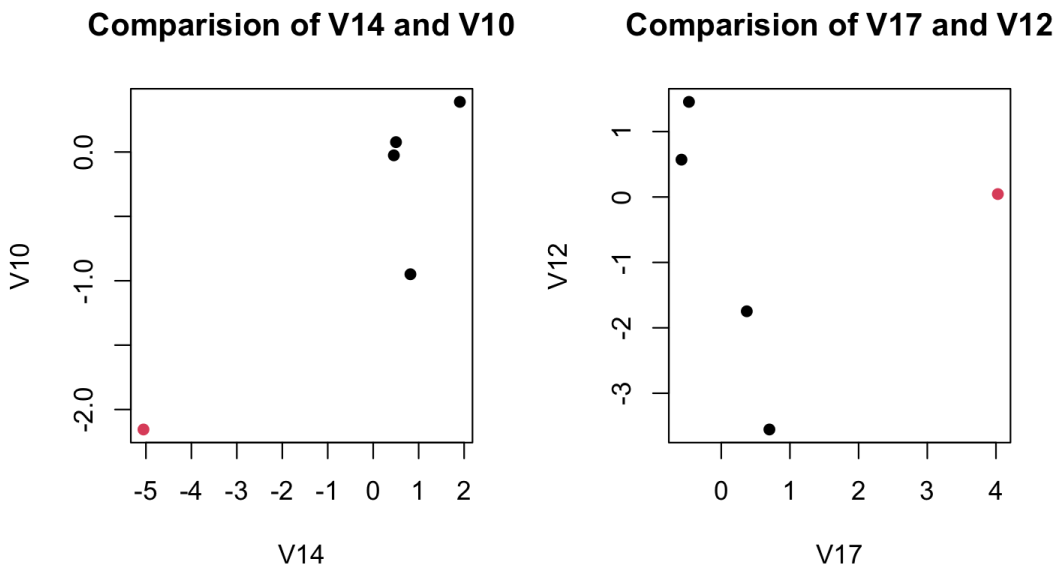

Figure 4: Comparisons of different variables and if the charge was marked as fraudulent

4 Conclusion

In this analysis we used the provided transaction dataset, consisting of 28 Primary Component Scores, the amount of each transaction, and whether it was fraudulent or not, to create a predictive model. The chosen model has a predictive sensitivity of ∼97.2% (given a transaction is fraudulent, the model has an ∼97.2% probability of predicting it to be fraudulent) and a predictive precision of ∼90.5% (given the model predicted a transaction is fraudulent, there is a ∼90.5% probability of it actually being fraudulent). Using this model, we predicted that of the 5 provided transactions, 1 transaction is fraudulent (transaction ID 3).

Moving forward, a way to improve this analysis is to get access to more transactions that are known to be fraudulent. These data points contain the information most needed to identify the distinction between fraudulent or not, and having more (though difficult to obtain) will improve the accuracy of the model. These models are also able to be tuned significantly (particularly the cutoff value), so conversations with the legal team to determine how to tune the model will help (ie. is it a priority to identify any transaction that might be fraudulent, or do we only want to flag those that have a high confidence of being fraudulent, etc.).

5 Teamwork

Shae: Abstract, Results, Model 2, Model Evaluation
RJ: Introduction, Conclusion, Model 1, Model Evaluation

6 Appendix

Variable	Importance
V14	630.18
V10	441.32
V17	438.56
V12	313.32
V11	165.42
V16	119.47
V4	105.34
V3	92.68
V9	74.75
V7	64.36
V18	61.86
V19	28.94
V1	27.74
Amount	27.72
V21	24.87
V5	24.27
V6	22.19
V2	21.79
V15	21.47
V13	20.10
V26	19.95
V25	17.94
V8	17.76
V20	17.29
V22	15.69
V28	15.28
V24	13.27
V27	12.70
V23	11.61

Table 2: The list of variables ordered by their importance in determining if a transaction is fraudulent