Analytathon 2: Credit Card Fraud Prediction

Evan Ganson Saldanha

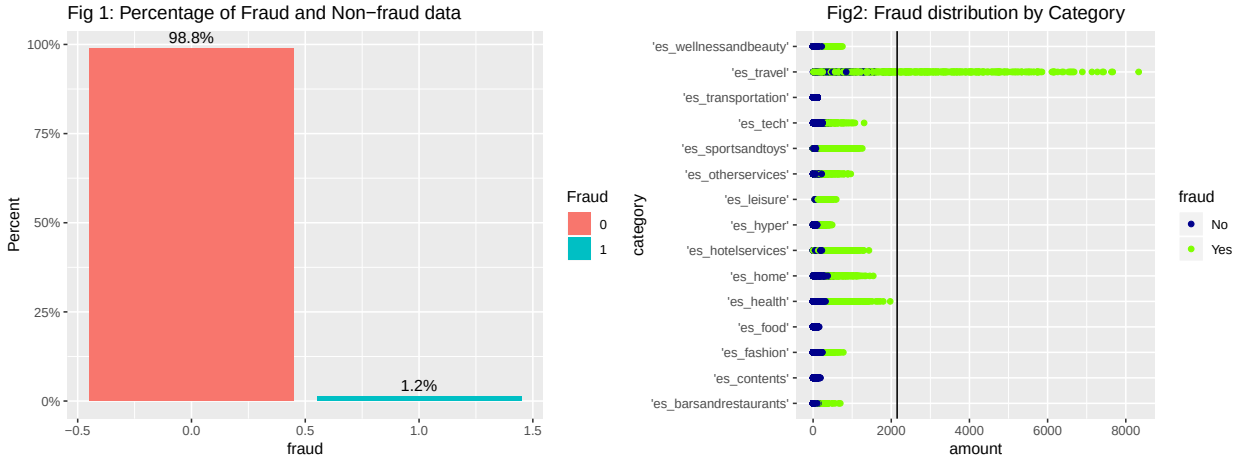
29/03/2019

Objective:

To explore the given credit card fraud data-set with the aim to predict the factors that lead to fraud

Exploring Dataset:

The data set consisted of 594,643 observation and 10 variable, of which 7 are factors, three are numerical variable. The data was highly imbalanced as the percentage of non-fraudulent data was 98.8% whereas the fraudulent data made just 1.2% of the total observations (as shown in Fig 1). Fig 2 shows the fraud distribution by each category and amount.



Exploratorty Data Analysis:

The analysis of data began with the removal of zero variance column such as 'zipcodeOri' and 'zipMerchant' and also the near-zero variance such as 'step' and 'customer'. Furthermore, the observation with the negligible amount(below 0.03) never had any frauds, hence such rows were filtered to reduce the data size. The response variable ('fraud'), was still a numeric variable which was transformed to factor where 0 and 1 resulted in 'No' and 'Yes' respectively.

Data Partitioning and Balancing:

The tidy data was further split into 80% and 20% termed as train and test data respectively. The proportion of fraudulent data was just 1.2% whereas the majority was of non-fraudulent data with 98.8%. Hence SMOTE (Synthetic Minority Over-sampling Technique) function which oversamples the train data using bootstrap and k-nearest neighbour approach was utilized for balancing, which resulted in the fraudulent data to be 48% and 52% of non-fradulent data.

Classification models:

A classification model attempts to make some inference from the information given for training. Since this is a the two-class classification problem and the prediction was on the response variable 'fraud', various models can be chosen under classification to predict the fraud. Among all the model, the following three models best suited this problem:

1. Decision Tree
2. Naïve Bayes
3. Random Forest

Having the balanced and cleaned train data in hand, the above models were fitted onto this train data set and best model was chosen based on ROC, sensitivity and specificity. Roc is the probability curve, sensitivity is a proportion of positive results(results that were truly positive) whereas the specificity is the proportion of negative result(results that were truly negative).

1.Decision tree:

A Decision tree is a graph that uses a branching method to divide the data based on a decision made. On trying the two instances of decision tree model namely, default and auto (tuneLength = 20), the accuracy of ROC/AUC (Area under the curve) for decision tree auto was 99.43% compared to default which had an accuracy of 98.99%. Figure 3 below, shows the representation of ROC along with sensitivity and specificity for both instances of the decision tree. The figure Fig 4, shows the important variable that influence this model.

Fig 3: Dotplot for Decision tree

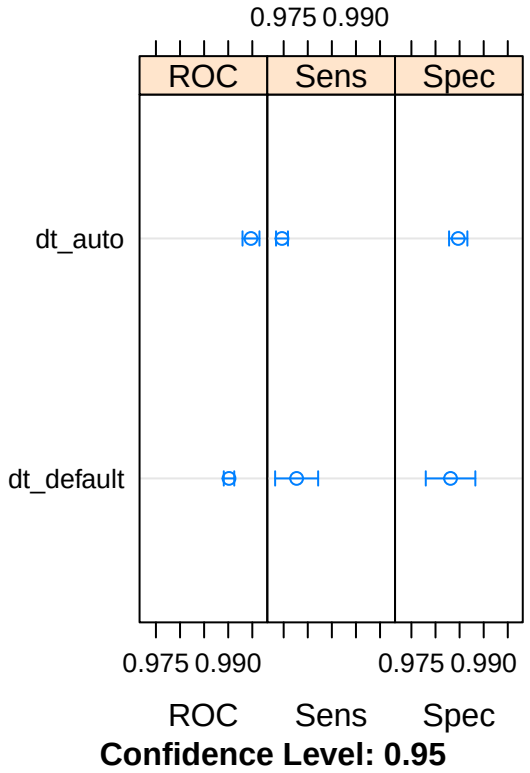
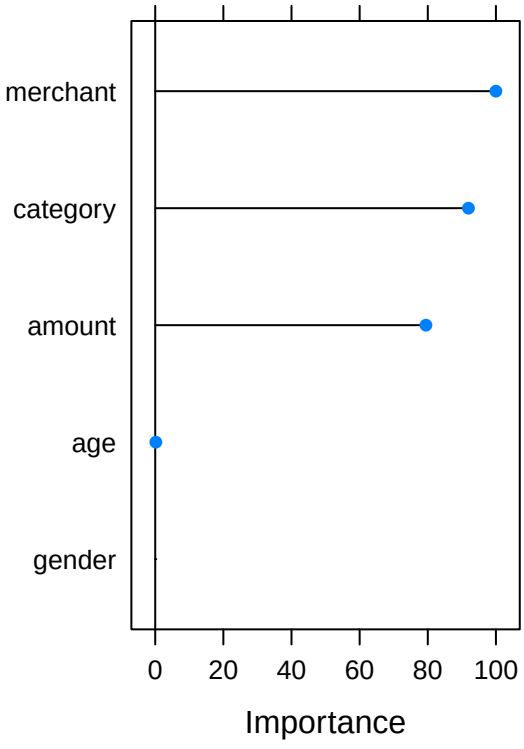


Fig 4: Important variables



2. Naïve Bayes:

Naïve Bayes is a group of "probabilistic classifiers" in light of applying Bayes' theorem with strong assumption among the features. Similar to the decision tree, on trying the two instances of Naïve Bayes model namely, default and manual(manually specifying the parameters), the accuracy of ROC/AUC (Area under the curve) for both the models seems to match each other which is 99.7%. Figure 5 below, shows the representation of ROC along with sensitivity and specificity for both instances of Naïve Bayes. Figure 6, shows the important variable that influences this model.

Fig 5: Dotplot Naïve Bayes

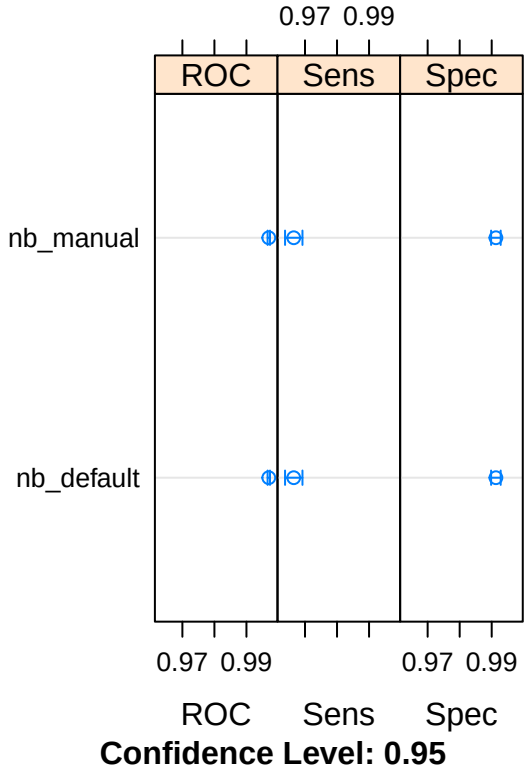
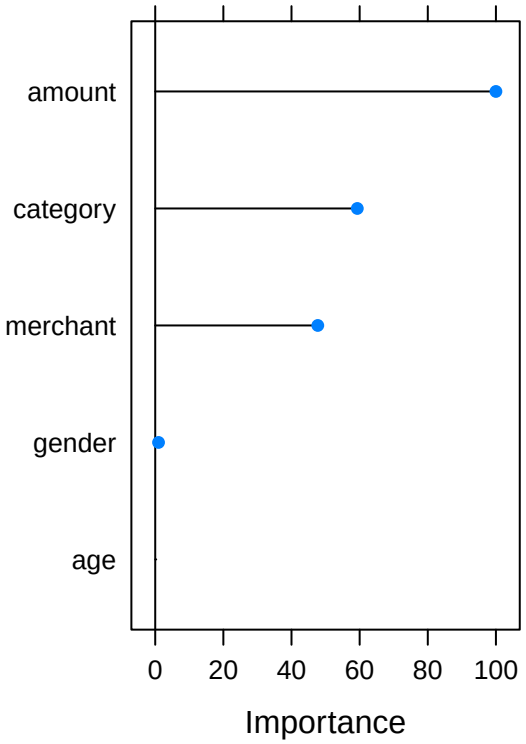


Fig 6: Important variables



3. Random Forest:

Random forest is a combination of many decision trees but unlike decision tree, there is no overfitting of data. Just like the previous models, on fitting the three variations of Random Forest model namely, default, auto (tuneLength = 20) and manual(manually specifying the parameters), the accuracy of ROC/AUC (Area under the curve) all three instances are as follows:
-> ranger default: 0.9972528
-> ranger auto : 0.9972885
-> ranger manual: 0.9970779

Figure 7 below, shows the representation of ROC along with sensitivity and specificity for all the instances of Random forest. The figure 8, shows the important variable that influences this model.

Fig 7: Dotplot: Random Forest

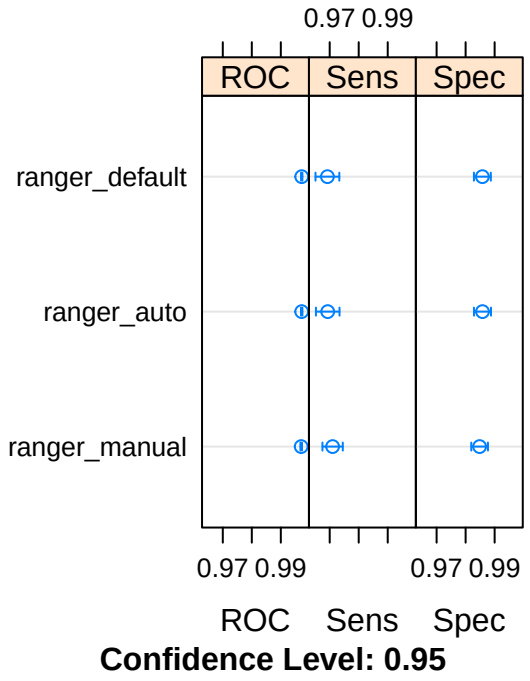
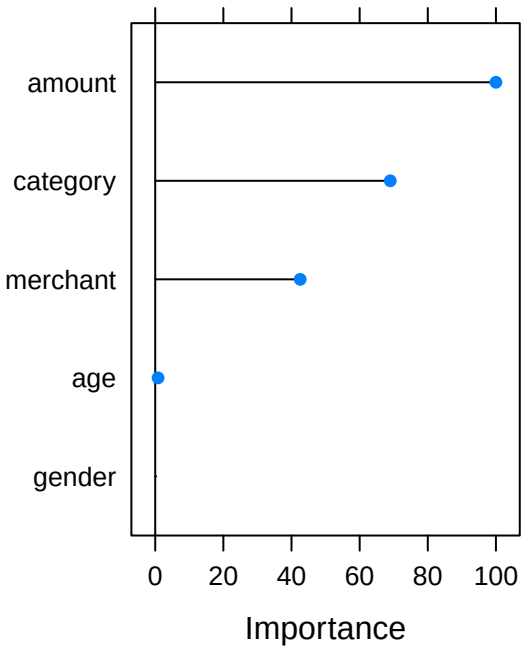


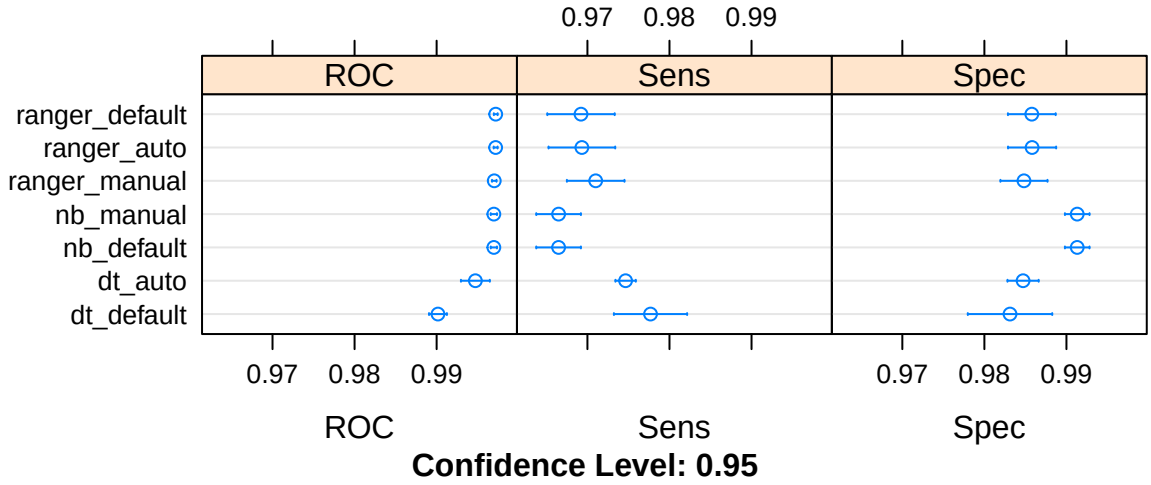
Fig 8: Important variables



Performance comparison

Comparing the performance of all the 3 model category by plotting them in a dot plot as shown in below figure:

Fig 9: Performance comparison of 3 models

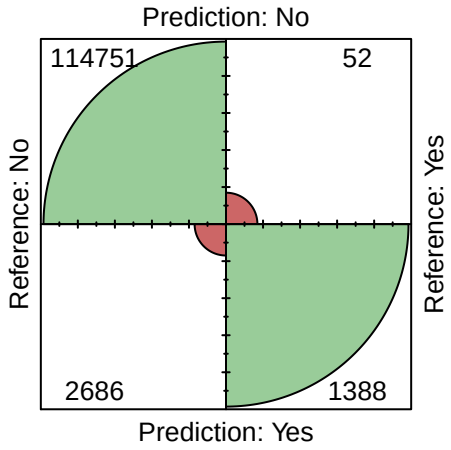


Taking ROC(probability curve) as the deciding factor the auto model of Random forest has the highest value of 0.9972885.

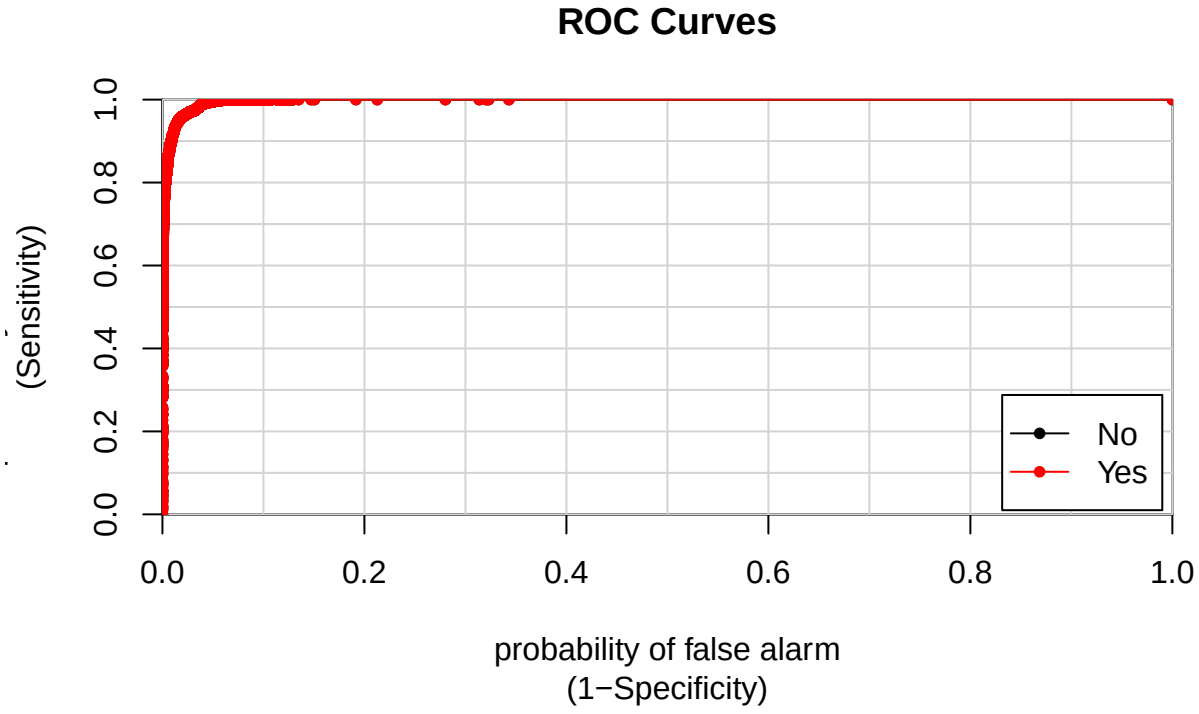
Prediction:

Random forest being an optimal model to do prediction on the unbalanced test data set, we obtain the accuracy of 97.7% through confusion matrix. To put in terms of response variable format, out of 1440 available frauds, the model accurately predicts 1338 frauds (as shown below).

Fig 10: confusion matrix



In a ROC curve (Fig below) the Sensitivity (true positive rate) is being plotted in the method of the 100-Specificity (false positive rate) for various cut-off points.



Conclusion:

The prediction model takes consideration mainly the three variables which are: amount, category and merchant. By focusing on the top candidates in each of these three categories, and taking some precautionary measures to avoid the customers or the bank to become a victim of such frauds would lower the fraud rate and also improve the trust of the customers on the Bank.