Assignment 1 CDA

Tim van Rossum, 4246306
Michiel Doesburg,

April 30, 2018

1 A visualization of the data

For the visualization part of this assignment, we first started out by making bar plots of different kinds, but eventually settled for the scatterplot as seen in figure 1. This scatterplot uses the amount of Eurocent spent per transaction, to allow for better comparisons to be made, as there were five different currencies in the dataset. Exchange rates of April 25, 2018 were used. As can be seen from the scatterplot, there are basically no fraudulent transactions where more than 800 Euro was spent, while there are benign transactions where more than 800 Euro was spent. Also, there are significantly more fraudulent transactions originating from Mexico and Australia than there are from the rest of the world.

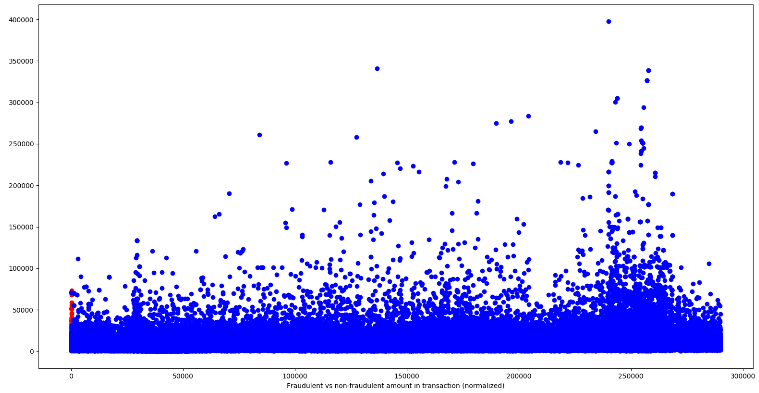

Figure 1: A scatterplot of the amount of money spent in the sampled transactions. A red dot indicates a fraudulent transaction (these are only at the very left of the plot due to very little fraudulent transactions existing), while a blue dot indicates a benign transaction.

2 Applying SMOTE to the data

Because we used Python for this assignment with `scikit-learn`, we used SMOTE as implemented in the package `imblearn`. The `fraud_detection.py` script already preprocessed the data in such a way that applying SMOTE to it was very easy, as the implementation only needed the data and the class labels, and both were already generated by the script. The general steps of preprocessing are: remove data with the "Refused" label (as that is data where we cannot be certain whether or not it is fraudulent), and transform the mail identifiers, IP identifiers etc. to simply the number that they use. The classifiers that we used were the random forest classifier, the 5-NN classifier, and the logistic classifier. The ROC curves are shown in figure 2.

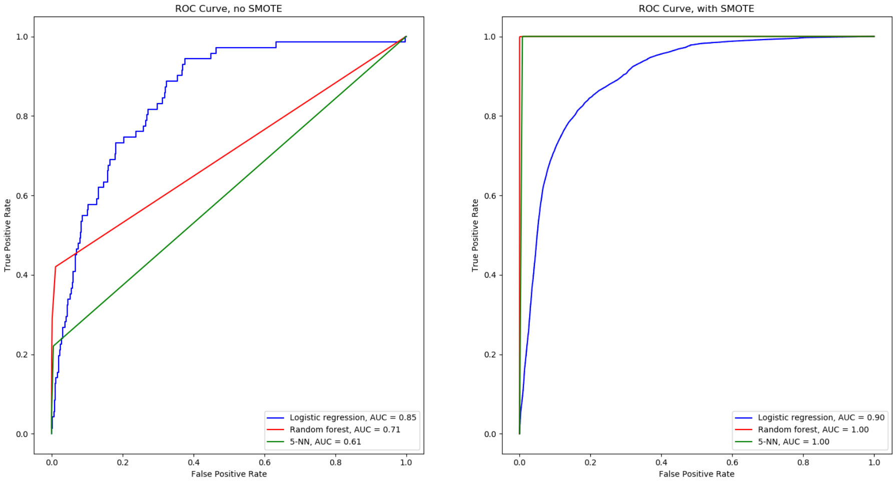

Figure 2: Caption of ROC curves.