Cyber Data Analytics
Credit Card Fraud Detection
Lab 1
Github URL: `https://github.com/PedroCaldeira/CDA2018_G35`

Liliana Oliveira , 4767306
Pedro Caldeira, 4768477
Lab Date: 13/05/2018

Visualization

By having an overview on the dataset it was possible to conclude that 81.5% of the dataset are benign transactions, 18.4% are refused and 0.1% are fraudulent transactions. However, all refused transactions were not taken into consideration in this study due to its unknown labeling (one cannot say if they are benign or fraudulent.

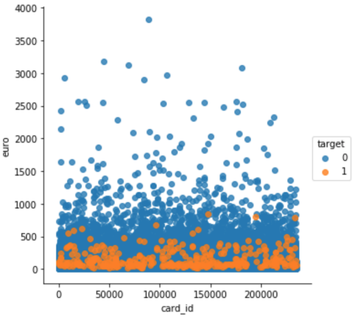

Figure 1: Amount spent per card

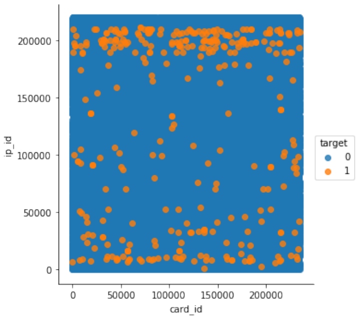

Figure 2: Cards per Ip

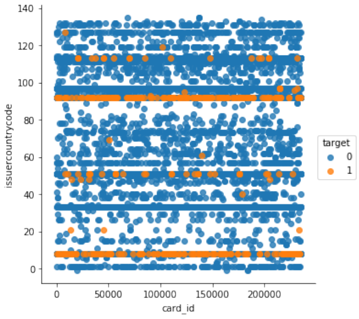

Figure 3: Issuer Country per card id

Through Figure 1 one can conclude that all of the fraudulent transactions had were executed with an amount lower than 1000. By making transaction with such low value nature, they often are hard to detect. By doing this, the attacker tries to remain undetected by undergoing a normal behavior pattern.

In Figure 2 it is possible to visualize that a lot of fraudulent transactions are made from the same IP address. One can see that attackers tend to try multiple cards on their transactions. More often than not, the fraudulent cases that are detected are those where the attacker does not attempt to change its IP Address (Through the use of a Proxy/VPN).

Through Figure 3 it is possible to conclude that transaction with cards from some issuer countries are more often detected as fraudulent than others. This could be due to the fact that cards from some countries cards are more easily forged than others, and such cards are the ones sold in the black market.

Imbalance Task

To deal with imbalance data different procedures were considered. We tested three classifiers: Logistic Regression, Decision Tree and Random Forests. To attend to the imbalanced dataset issue the following methods were applied and tested:

- Oversampling with SMOTE
- Oversampling with SMOTE + Tomek
- Random undersampling the majority class (Settled transactions)

As a baseline we also ran these classifiers without any kind of sampling algorithm. cutoffs were tried through trial and error picked until the seeming best performance was obtained.

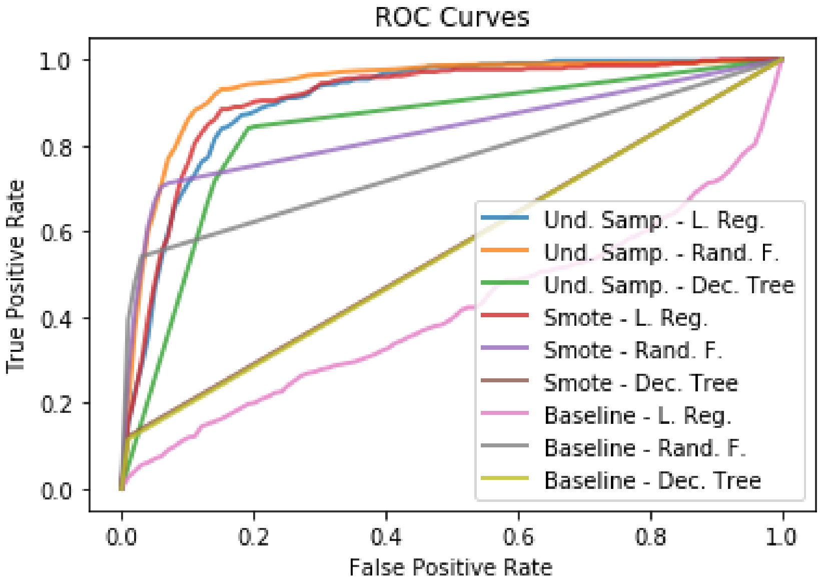

Figure 4: Amount spent per card

Smote-TOMEK Curves were removed due to its similarity to SMOTE curves

Classification

To begin we detect missing values and treat them by giving them -1 except to *cardverificationcodesupplied* where we took in consideration the value in *cvcresponsecode* column where we assumed that if the cvc code was a match, then it was provided. The date columns were formatted to an adequate type of date. Two columns were added with only the date and the time,separately, from the *creationdate* field. The variables were converted into numerical values. The amounts of all the transactions were converted to Euro currency. After this we only took in consideration the transactions that were benign and fraudulent, excluding all the refused transactions do to its unknown source. We first start by considering only the transactions individually and then we added features that took in consideration the costumer history, namely:

- 'SuspShopInt' - if the current ShopInteraction is *ContAuth* and the previous was of the type *Ecommerce*
- 'SuspCardVerif' - if the customer gave the cvc in the previous transaction and it does not provide one in the current transaction
- 'SuspCVCCode' . if the cvc provided in the previous transaction was a *match* and in the current transaction it is not.
- 'IPChange' - if the IP changed
- 'AmountChange' - the amount changed from the last transaction
- 'LastTransactionin2hrs'- if the last transaction happened in the last 2 hours

Even though not presented, Precision-Recall (PR) curves have an impact on deciding the best algorithm and they were implemented for each cross validation fold. ROC and PR curves are both used for binary classification. In ROC spaces, algorithms that might look comparable end up having PR values that invalidate such algorithm. This occurs, in this case, because the number of benign transactions greatly exceeds the number of fraudulent transactions. Consequently, a large change in the number of false positives can lead to a small change in the false positive ratio used in ROC. PR curves on the other hand, by comparing false positives to true positives rather than true negatives, captures the effect of the large number of negative examples on the algorithm's performance. So, we say that a curve has relevant information in ROC space if it has relevant information in PR space.[5]

Bonus

A single transaction information is not sufficient to efficiently detect a fraudulent transaction since it ignores the consumer spending behavior. By analyzing the dataset it was observed that a lot of fraudulent transactions appeared after a user made a legitimate transaction of the type *Ecommerce* which was followed by a *ContAuth* transaction considered fraudulent. To take this in consideration, we created new features that were based on the card behavior in the previous transactions. It was not considered a large window of transactions due to the limited number of transactions per customer/card, if this was considered it was needed to be handle carefully due to the fact that consumers not always behave in the same way and that a normal behavior changes over time [3].

References

[1] Tom Fawcett *An introduction to ROC analysis.*

[2] Haibo He and Edwardo A. Garcia *Learning from Imbalanced Data.*

[3] Varun Chandola, Arindam Banerjee and Vipin Kumar *Anomaly Detection: A Survey.*

[4] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, J. Christopher Westland *Data mining for credit card fraud: A comparative study.*

[5] Jesse Davis, Mark Goadrich *The relationship between precision-recall and ROC Curves.*

[6] Philip K.Chan, Salvatore J. Stolfo *Toward Scalable Learning with Non-uniform class and cost distributions: A case study in Credit card fraud detection.*