Data Science Career Track

Capstone 1 -

Milestone Report

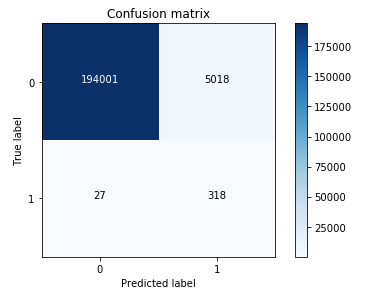
by Edward Franke

05/01/2019

Credit Card Fraud Detection Project – Milestone Report

**EXECUTIVE SUMMARY:**

The purpose for this project is to find a correlation connected to fraudulent credit card transactions that can separate them in real time compared to legitimate transactions. The dataset is a cleaned dataset from Kaggle.com. It has been discovered there is a correlation connected to fraudulent transactions. Most models fail to accurate predict which transactions are fraudulent because the dataset is severely unbalanced with over 99% of the transactions being legitimate. The solution was to run SMOTE first to rebalance the dataset then logistic regression with these results:



**CONCEPTS:**

IDEA: A model to detect fraud in credit card transactions. (problem to solve)

CLIENT: Credit Card Companies

REASON: If they don’t detect and stop the fraud as it happens, their customer don’t pay the cost, they do. This project is intended to reduce their expenses and increase their profits.

DATA: From Kaggle, a cleaned dataset with 284,807 transactions from 2013 with 492 frauds in total.

SOLUTION: Create a model or analysis to discover what makes fraudulent transaction similar to detect this relationship as it happens.

DETAILS: See Detail Section Below.

DELIVERABLES: Working code that detect fraud as it receives data and a presentation outlining the discoveries and explain the methods used to reach the discoveries.

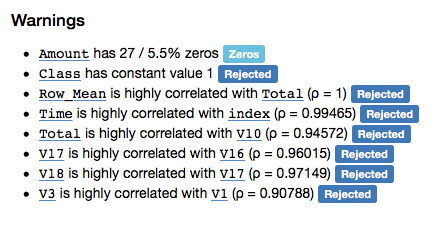
**INITIAL FINDINGS FROM EXPLORATORY ANALYSIS:**



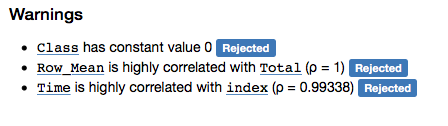
This is the code that was used to separate the initial dataframe into separate dataframes only containing fraud or legit transactions.

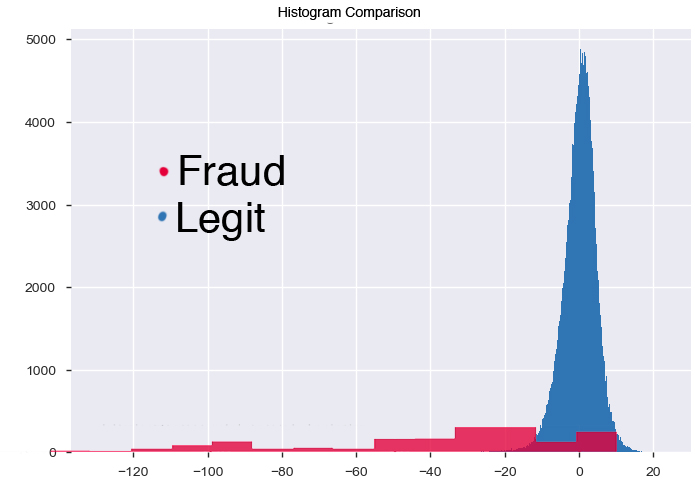
While comparing the transactions, I noticed fraudulent transactions have a greater histogram area that legitimate transactions. I was under the impression that I could create code to determine any transaction with a total V set as below -17 as fraudulent (see Appendix for more details). However, as this overlayed histogram comparison shows, that would fail the purpose of the project. A solution is still being sought.

Fraudulent Transactions (Class 1 indicated fraud)

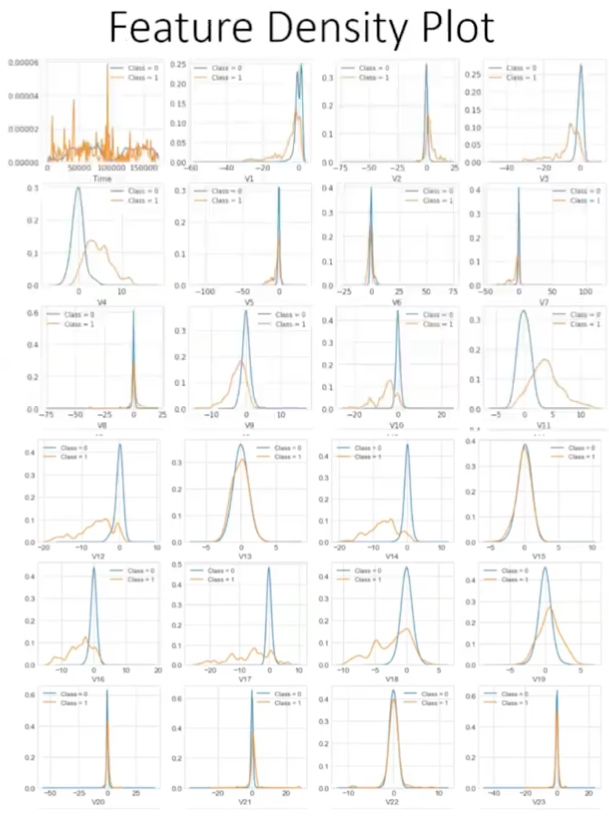


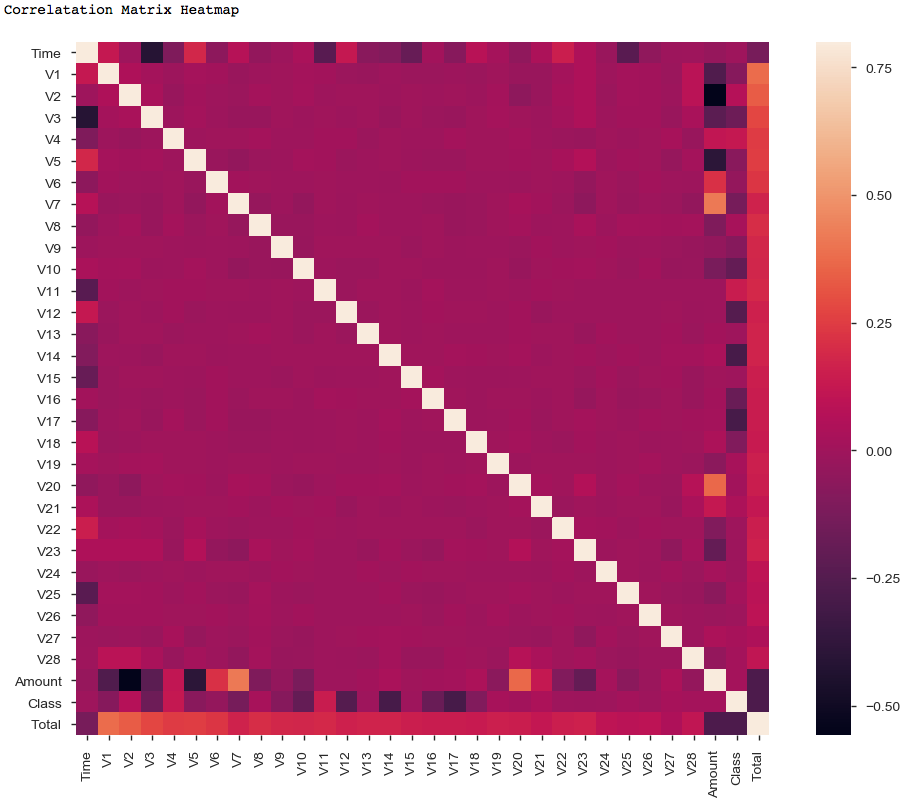
Legitimate Transactions (Class 0 indicates legit)



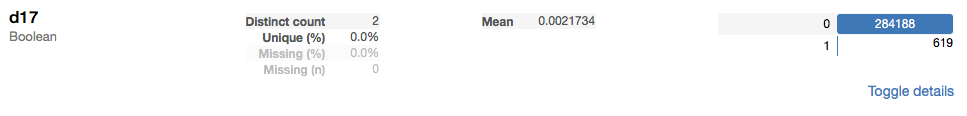


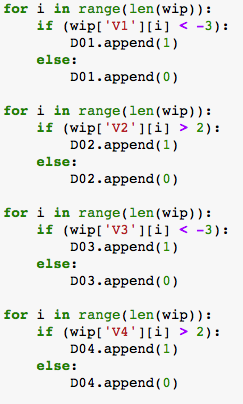
Because the row totals did not present useful data, I also looked at each V column category to determine what segments would more easily separate potential fraudulent transactions for the legitimate transactions.





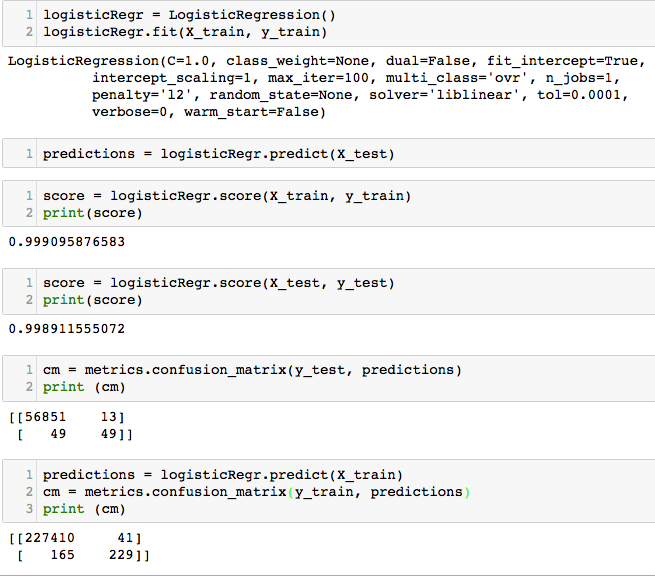
I then wrote code to attempt to isolate the fraudulent transactions using for loops.



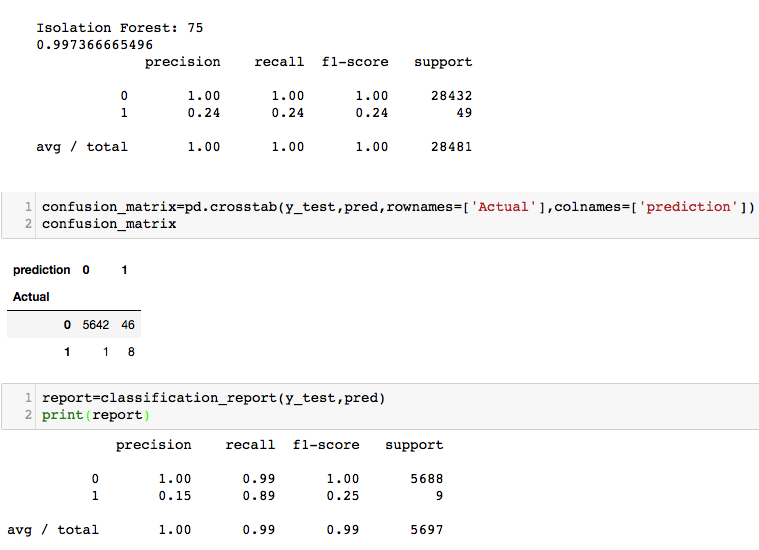


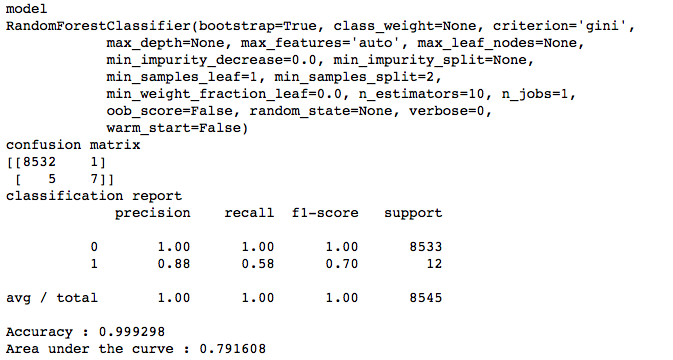
V17 appeared to have the best promise with no tweaking of the code. However, this isn’t machine learning techniques so this method was scrapped.

A straight logistic regression appears to give promising results on the surface.



The scores look good. However, when running the confusion matrix and comparing both legitimate transactions and fraudulent transactions, there is an issue. Legitimate transactions have great results. However, fraudulent transactions at best have a 50% score. This number is totally unacceptable.





Isolation Forest and Random Forest classifiers/models fair even worse.

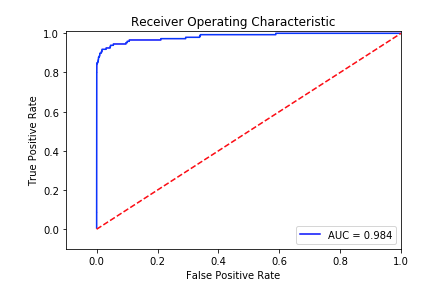
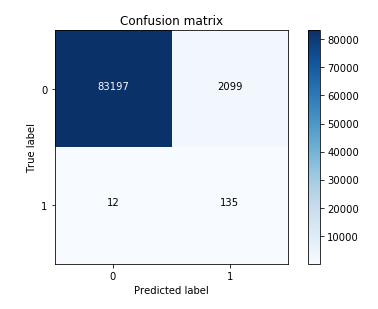
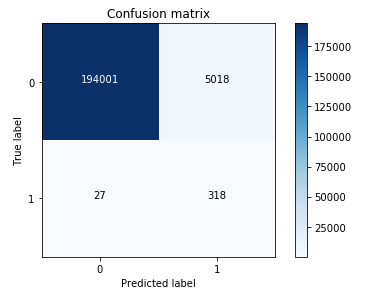
This is because the dataset has over 99% legitimate transactions and these models are unable to analyze unbalanced datasets properly.

**THE SOLUTION:**

The learning phase and the subsequent prediction of machine learning algorithms can be affected by the problem of unbalanced datasets. The balancing issue corresponds to the difference of the number of samples in the different classes. The decision function of the linear SVM is highly impacted. With a greater unbalanced ratio, the decision function favor the class with the larger number of samples, usually referred as the majority class.

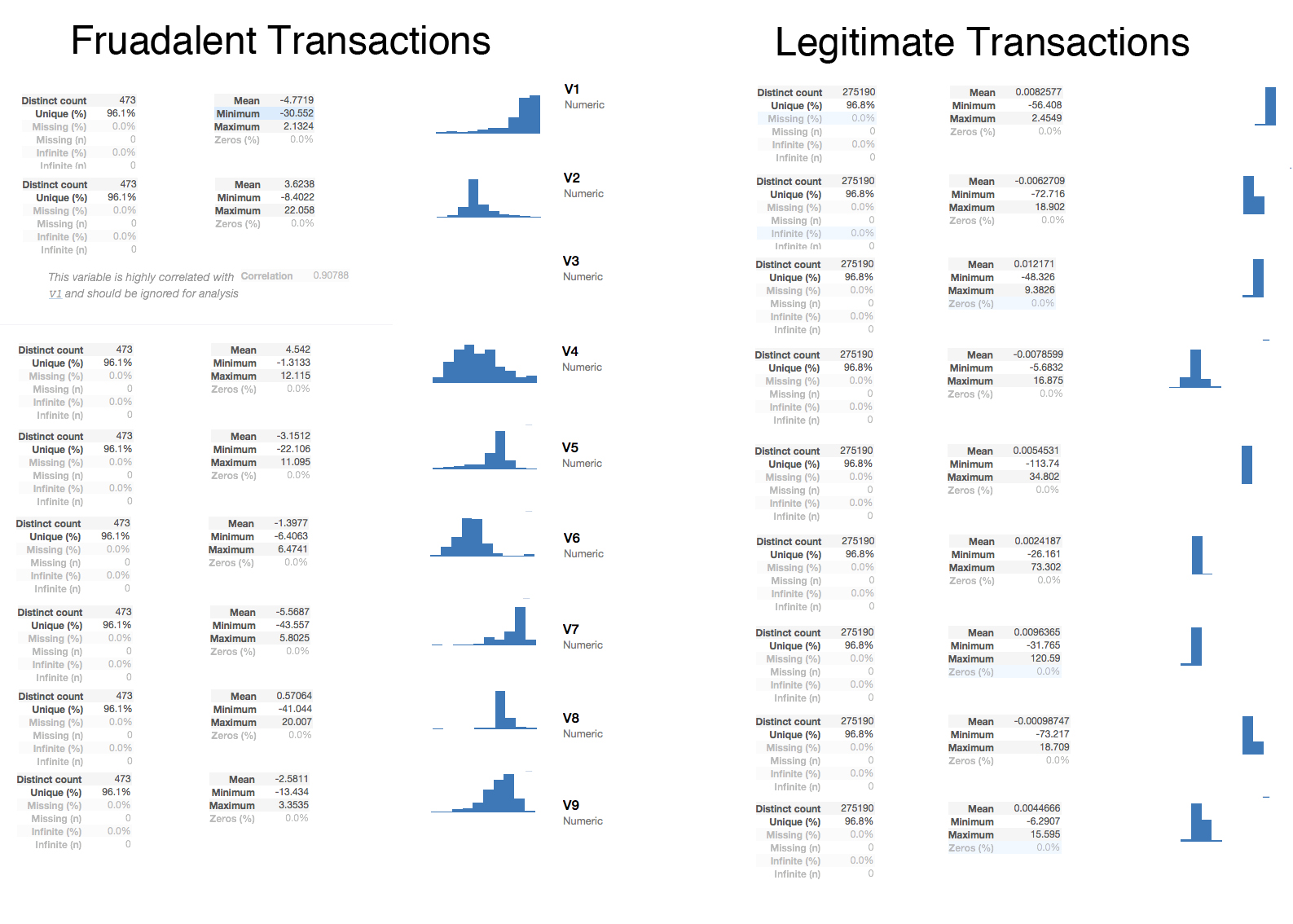
One way to fight this issue is to generate new samples in the classes which are under-represented. The most naive strategy is to generate new samples by randomly sampling with replacement the current available samples. The RandomOverSampler offers such scheme

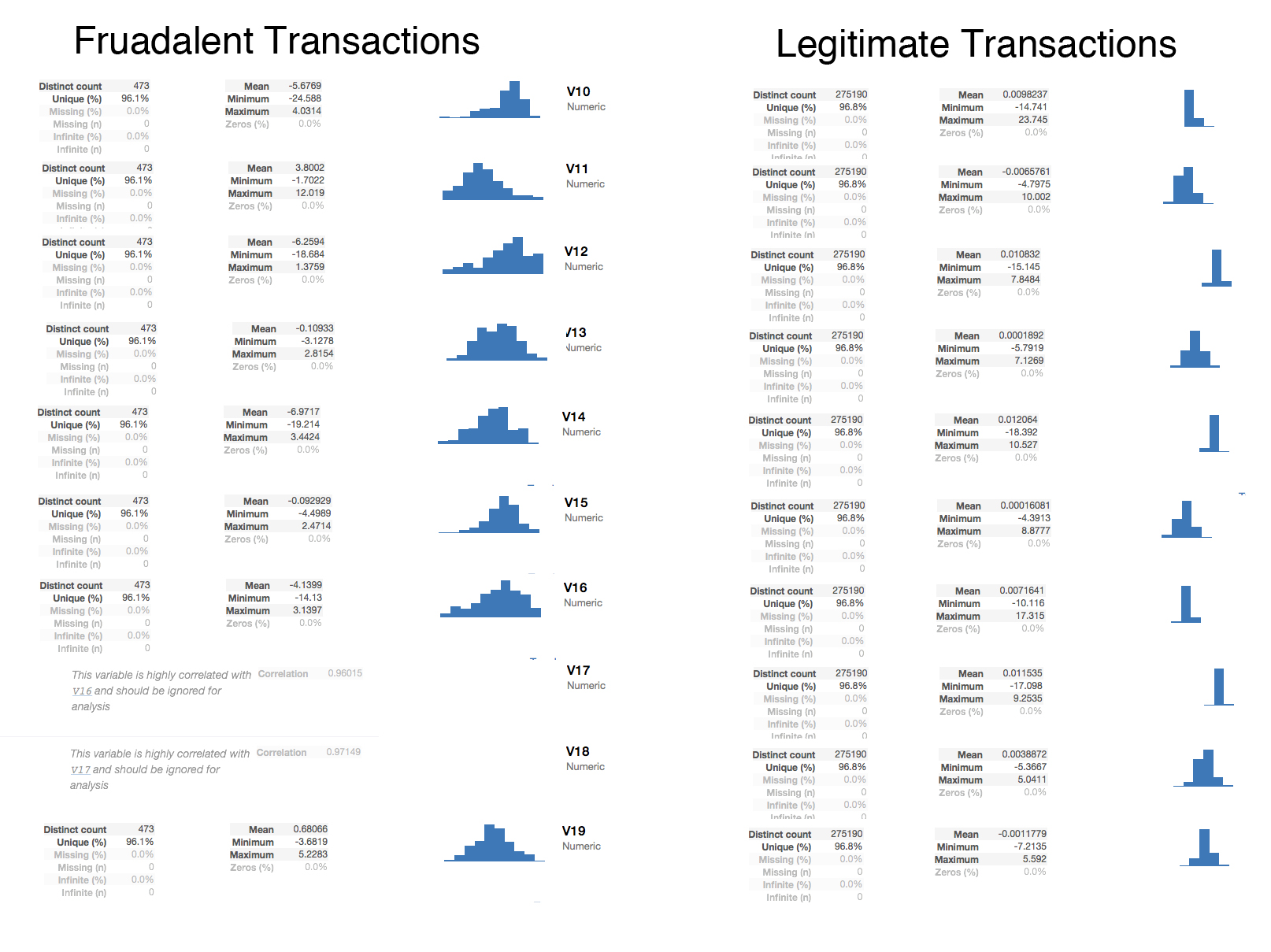
Apart from the random sampling with replacement, there are two popular methods to over-sample minority classes: (i) the Synthetic Minority Oversampling Technique (SMOTE) [CBHK2002] and (ii) the Adaptive Synthetic (ADASYN) [HBGL2008] sampling method. The solution for this dataset was to use the SMOTE to resample the data and logistic regression to perform the analysis with these results:

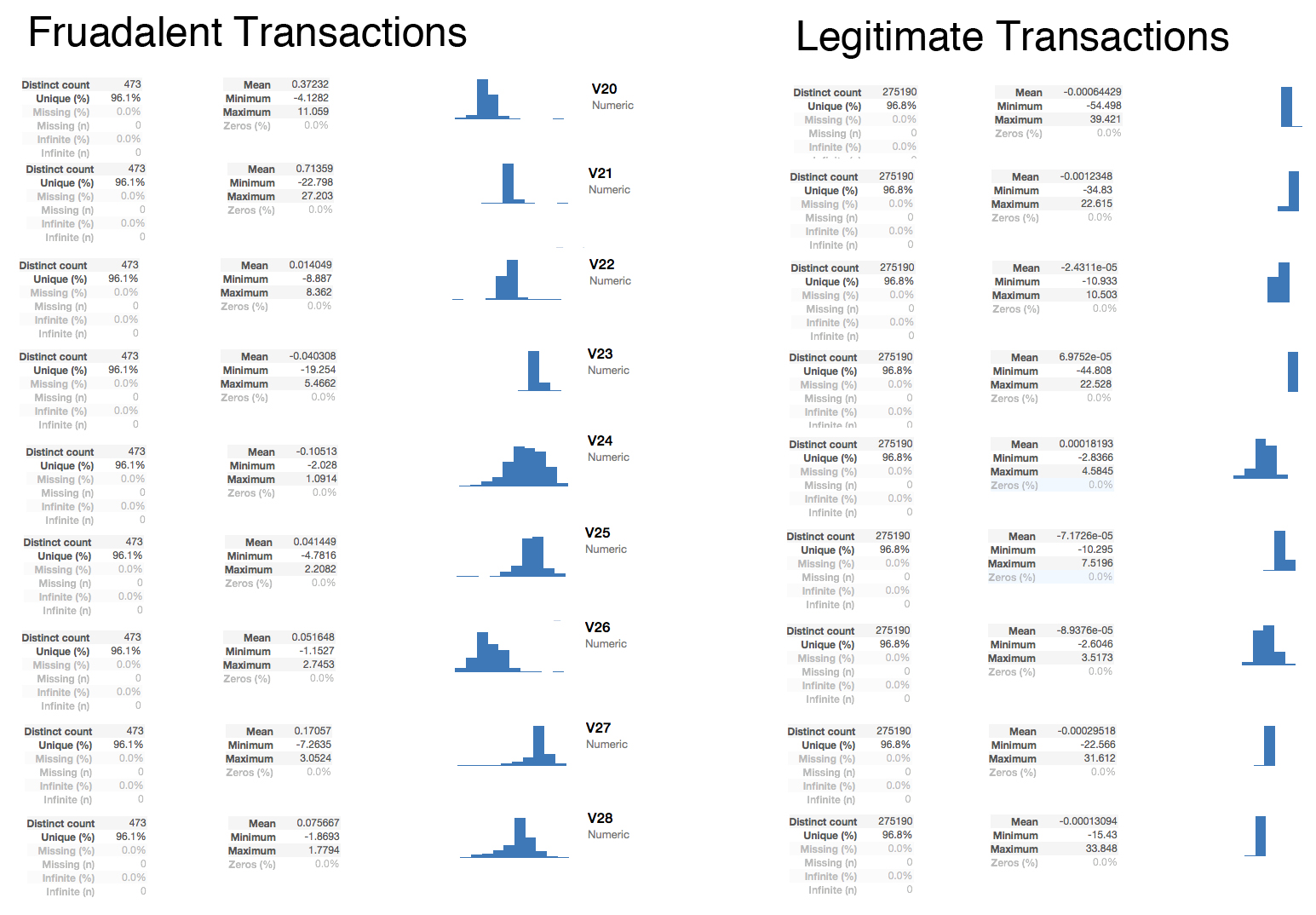


WHAT’S NEXT

APPENDIX – V Numbers compared







APPENDIX – Kaggle Details about the dataset.

Context

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Content

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Inspiration

Identify fraudulent credit card transactions.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

Acknowledgements

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on http://mlg.ulb.ac.be/BruFence and http://mlg.ulb.ac.be/ARTML

Please cite: Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015