Data Science Career Track

Capstone 1 -

Milestone Report

by Edward Franke

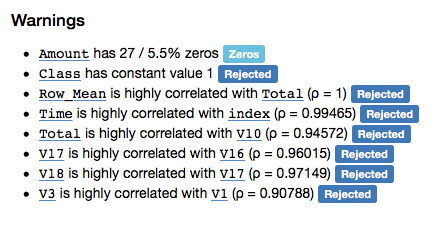
04/17/2019

Credit Card Fraud Detection Project – Milestone Report

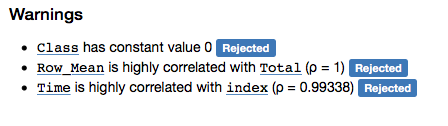
EXECUTIVE SUMMARY:

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, but the code required to separate it from the legitimate transactions has not yet occurred. The most promising results comes from pandas profiling function which gives the following data:

Fraudulent Transactions (Class 1 indicated fraud)



Legitimate Transactions (Class 0 indicates legit)



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.

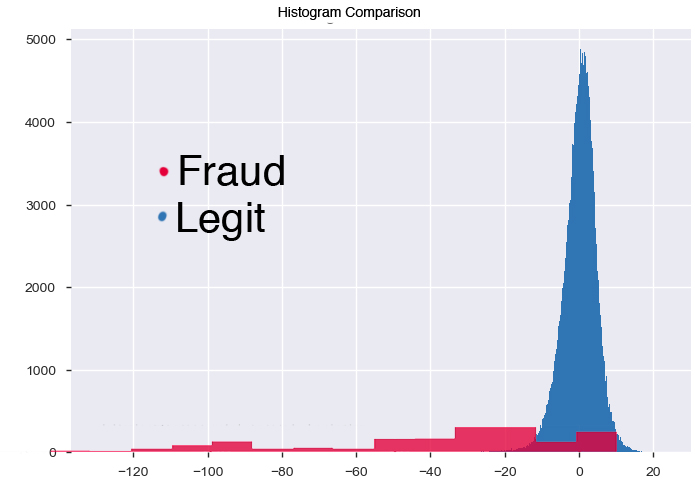
DELIVERABLES: Working code that detect fraud as it receives data and a presentation outlining the discoveries and explain the methods used to reach the discoveries.

**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.



WHAT’S NEXT

Deep Dive

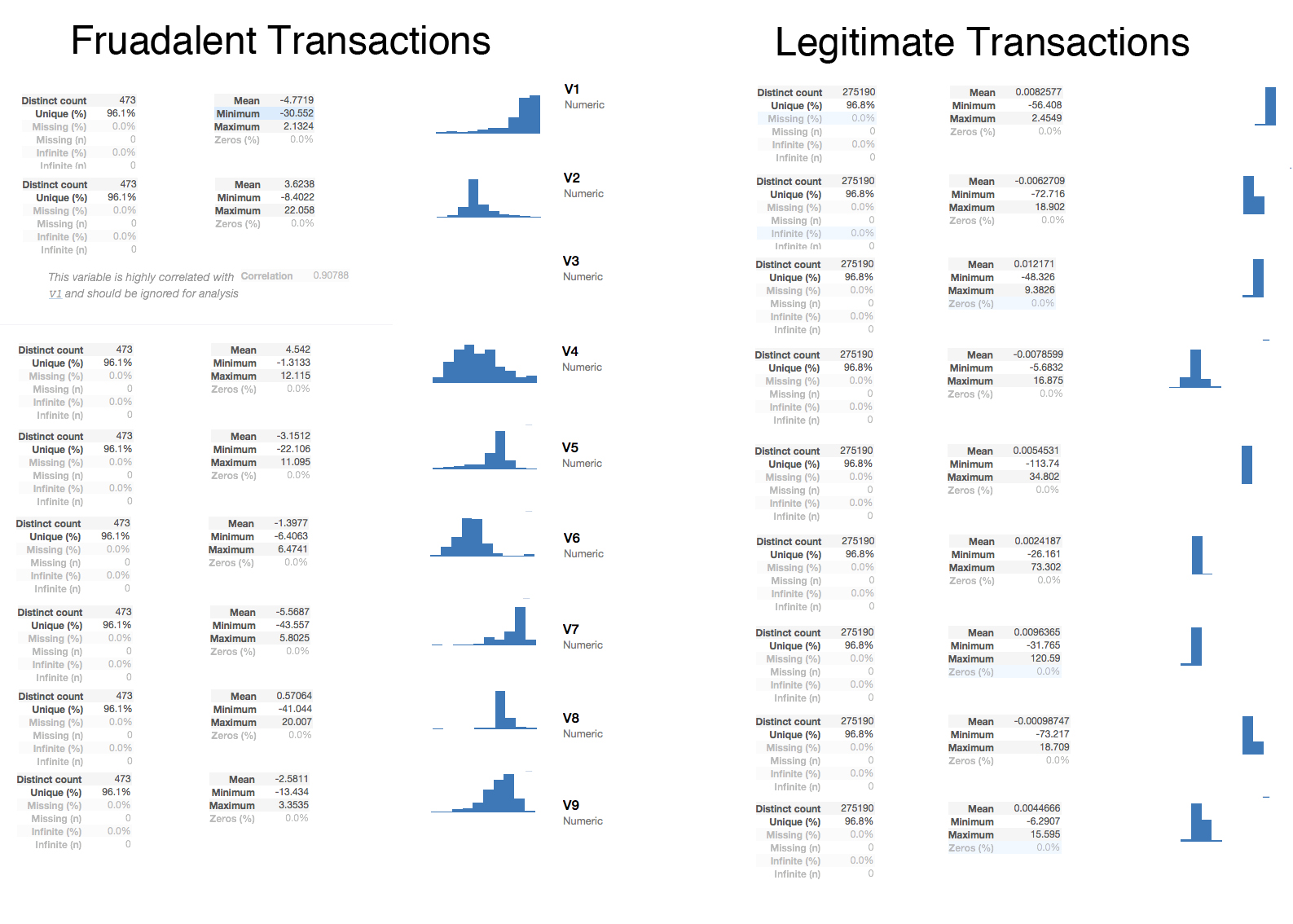
Discover what connects fraud transactions (V1 and V3, V16 and V17 and V18)

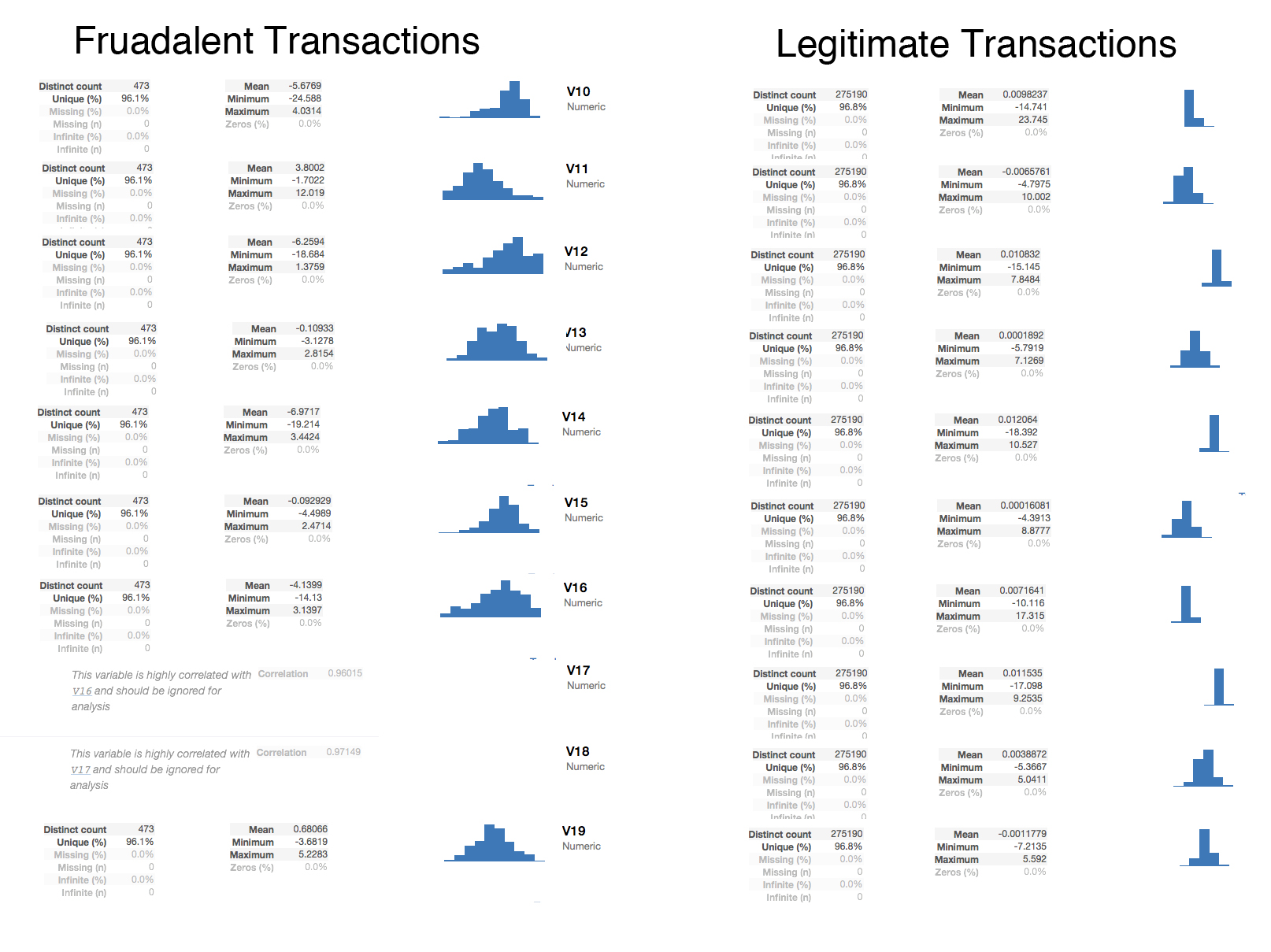
Write code that detects this correlation as the transactions come in

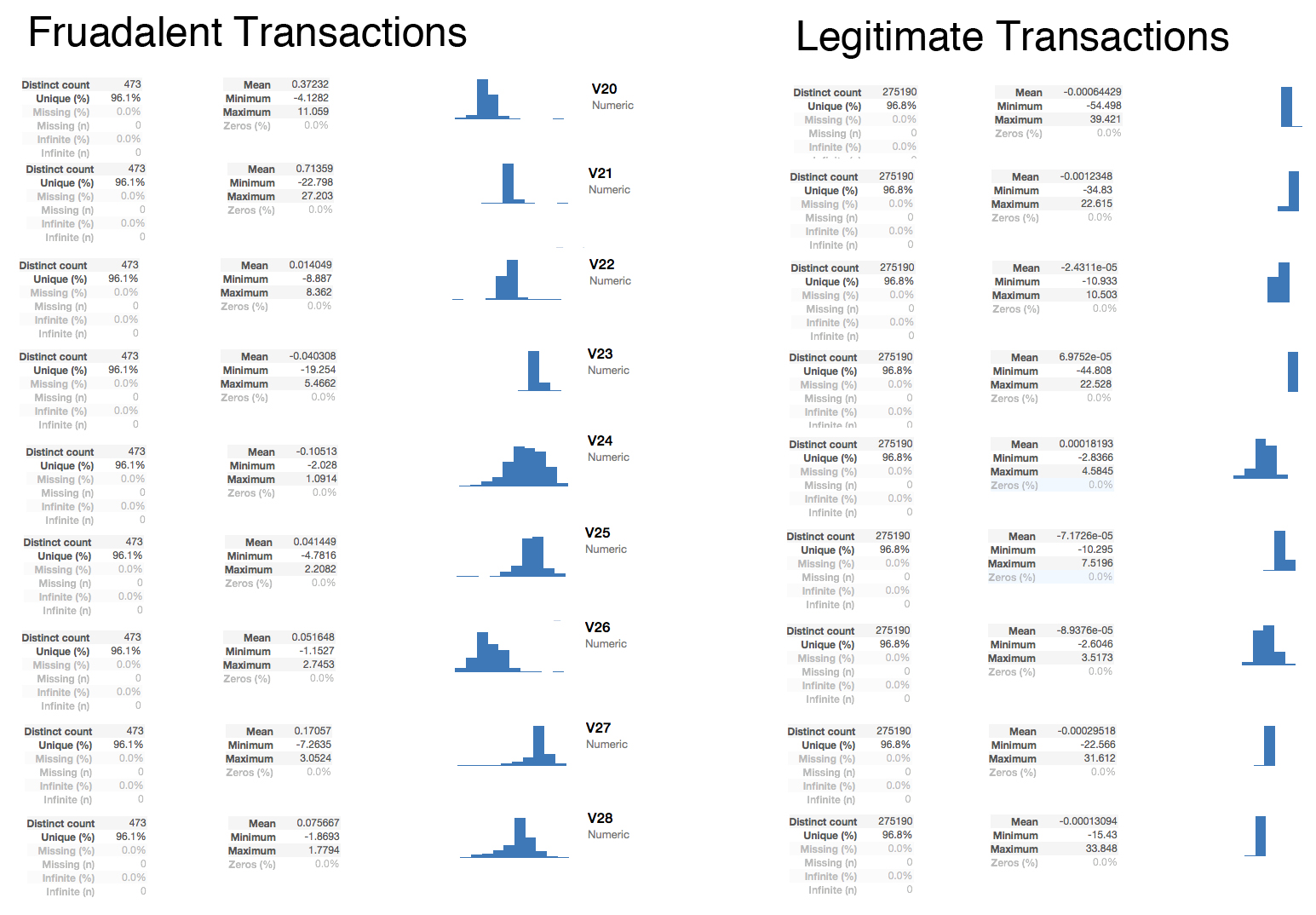
Run Statistical Analysis to determine error levels

Final Report and Presentation

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