Highly Screwed Data Analysis – Based on Credit Card Fraud Detection

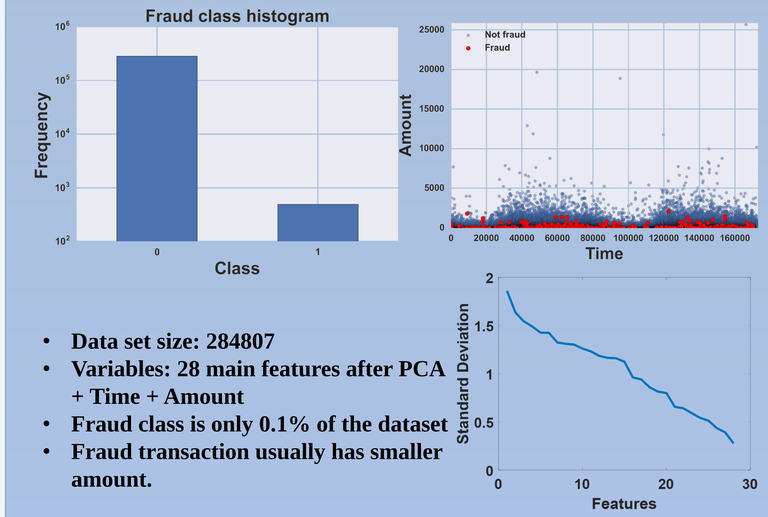
Michelle Shu[1], Minwei Xu[1], Xiang Li[1], Xinyu Huang[1], Shuyu Liu[2]

1 Department of Computer Science, 2 Department of Mechanical Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD

Introduction

In an ideal data set for analysis, training set is perfectly balanced, and there is no missing data. This situation, on the other hand, rarely happens in real life. Sometimes, we could only get skewed data set, which makes the minority insignificant during training. Then, we may wonder, how to give a right prediction if the data set is hugely skewed? For instance, if we want to predict the chance of snow in California tomorrow, we may want to investigate in the weather conditions in California for the past ten years. The data we collected, however, will be very unbalanced since snow rarely happens there. One result of such extremely skewed datasets is that although the overall accuracy is relatively high, predictions are almost always wrong on days when California snowed. In this research, we explore ways to tackle this skewed data problem. We investigate in a simple classification problem—credit card fraud detection—in which only two types of result can happen: either the transaction is a fraud (YES), or it is not (NO). In this data set, 99.9% of them are nonfraud transactions. In our analysis, to test the influence of skewness and different classification models, we have resampled the data set to unskewed and profoundly skewed groups and compared the effectiveness of various models based on the correctness of prediction to the original data.

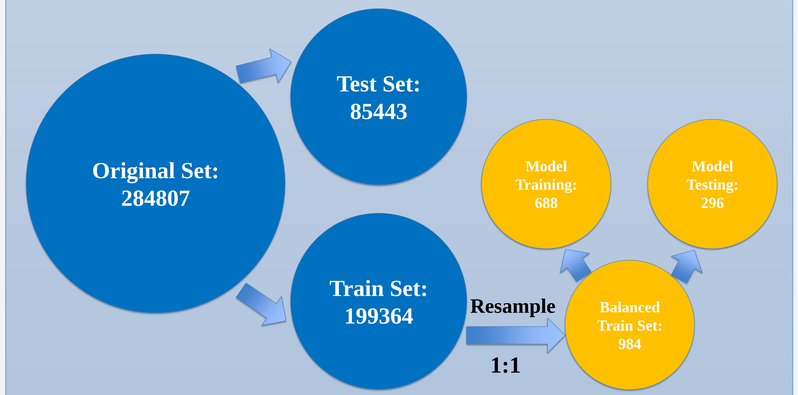
Data Set



- Data set size: 284807
- Variables: 28 main features after PCA + Time + Amount
- Fraud class is only 0.1% of the dataset
- Fraud transaction usually has smaller amount.

Method

- **Data Set Preprocessing**



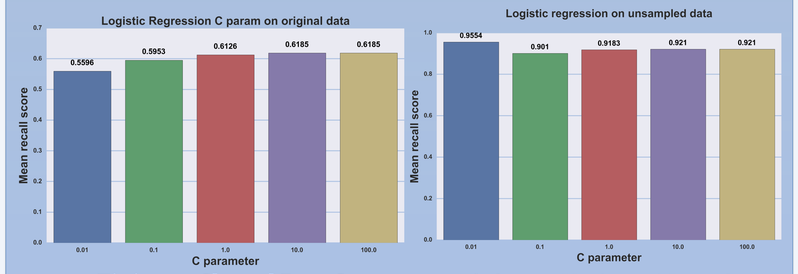
- **Classification Model**

- **Evaluation**

Recall Score: Detected Fraud / All Fraud

Results

- **Fitting on Skewed Set Gives Poor Prediction**

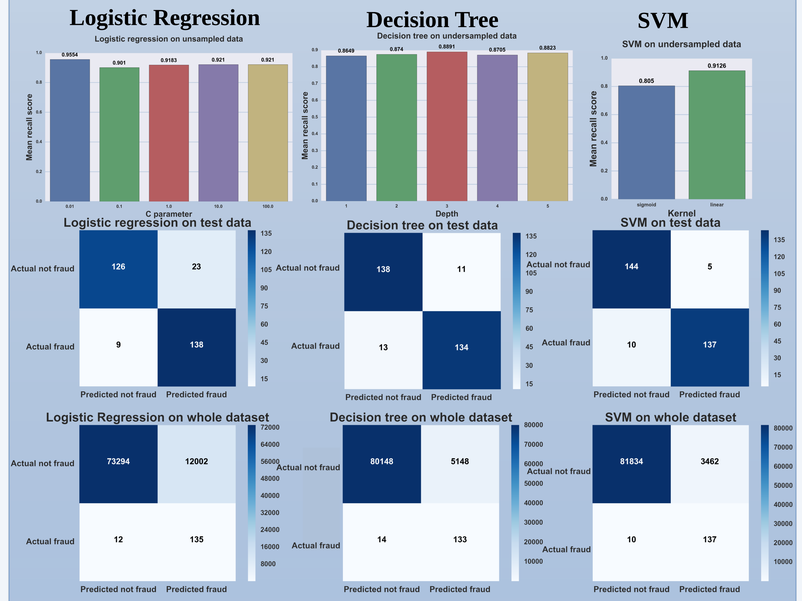


Fitting on Skewed Data Set | Fitting on Unskewed Data Set

*C Parameter: Penalty parameter to avoid over fitting

Results

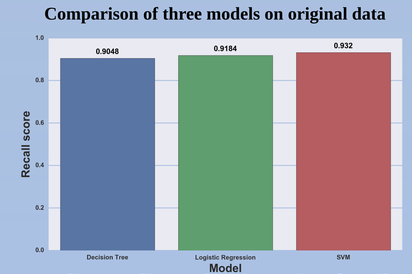
- **Linear SVM is the Best**



- In logistic regression model, large penalty gives better prediction
- In decision tree model, 3 trees are better
- In SVM model, linear model is more accurate than sigmoid model

Conclusion & Discussion

- Resample method largely saves prediction to skewed data set. (From 60% to 90% accuracy)
- Based on the three classification model, all of them give reliable prediction (scored higher than 90), and linear SVM method gives the best prediction.



- Recall score is a good method to predict credit card transaction fraud, because we want the model to be over protective. We weigh the detection of fraud transactions more than normal cases.