ICS635 Homework 3

Lambert Leong

March 11, 2019

1 Introduction

I chose to participate in the Santander Customer Transaction competition. A quick survey of the raw training data revealed that there are 200 features which contain positive or negative float values. My initial thought was to perform a principal components analysis (PCA) to try and reduce the dimensionality of the data. Several visualization kernerls showed that the distribution for each class with respect to each field, i.e. 'var_0', 'var_1',etc..., is not very different. In fact, there is a significant overlap of the two classes for almost all features and determining a decision boundary appears to be non-trivial. This lead me to believe that mapping the data into a new space with PCA would not be useful for the purposes of reducing data dimensionality.

I first looked at an ensembling methods and used gradient boosting to classify the data as is, with all 200 features. I then used the same gradient boosting classifier on training data which had the dimensionality reduced via PCA. Comparing the results of the original data, with 200 features, to the PCA transformed data, with less than 200 features, indicated that I should not proceeds with PCA and models generated with all 200 features were able to generalize to the validation set better.

The instructions on the competetion website stated,"...identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted." From these instruction I hypothesized that each feature could represent a transaciton from an individual's transaction history and I looked into treating the data as sequence data. I explored two neural network architectures which include long short term memory (LSTM) recurrent neural network and a 1D convolutional neural network (CNN). The motivation behind implementing these neural networks was to try to explore patterns in the sequence of transacitons that may correlate to a particular class.

The instruction also noted that the transaction amount is not really relavent. Each of the 200 feature fields contained either a positve or negative number which could indicate money coming in and money going out, respectively. Under this assumption, I sought to capture the amount of times money came in and the number of times money went out for a particular individual to see if it had any correlation to a particular class. In other words, I captures the total amount of postive values and the total amount of negative values for each individual in the dataset.

Although it is stated that the ammount of the transaction is not relavent I added features that indirectly correlate to the amount and these features include mean, standard deviation, skew, and kurtosis. I applied the previously mention models, gradiant boosting, LSTM, and 1D CNN, to the new dataset with new features to see if it would lead to better classification.

2 Methods

2.1 PCA

2.2 Gradient Boosting

2.3 Neural Networks

2.4 Feature Engineering

3 Results

4 Conclusion