Author: Sabin Adhikari

Capstone project - Santander Value Prediction Challenge

Overview:

Recognizing that customers are more likely to do business with their banks if the banks provide personalized service, Santander Group is asking to help them identify the values of transactions for each potential customer. The problem here is to predict customer's future transaction values by using their past transaction values.

After performing some data cleaning and exploratory data analysis, I started modeling from simple multilinear regression, ridge, lasso and univariate feature selection. After that, I used LightGBM and XGB to make a regression model and ensambled them. I followed the steps listed below.

1. Preparation of data

2. Exploratory data analysis

3. Modeling

(a) Multilinear regression, Ridge and Lasso

(b) LightGBM

(c) XGB

4. Submission to Kaggle and Discussion

1. Preparation of data

The data provided has train and test sets. The train data has 4991 feature columns, one target column and 4459 rows. The test data has 4991 feature columns and 49342 rows. Some constant columns were found in both test and train data sets and were removed. We found a few columns which were constant in train but not in the test. We ignored these columns from the test while modeling.

2. Exploratory data analysis

Looking at some descriptive statistics of the train data shows that target is highly dispersed (shown by the standard deviation, min and max). Features values are even more highly dispersed than the target values and most of the features having the value 0 at 75 percentile shows that the data might be highly sparse. Features of test data set also seem to have similar distributions as features in train data set.

The histograms of both train and test data sets shows that the features are highly dispersed and sparse. The target column of the train data set is less disperse compared to the feature columns. Also, by finding the pearson's correlation coefficients, we found that several features are strongly correlated to each other which is generally true for data sets having large number of features.

3. Modeling

Multilinear regression

The train data set was splitted in to train and test set and multilinear regression was used to fit the train set. When tested against the test set, As expected, the R_squared value was very high.

Ridge regression

Ridge regression is regularized multilinear regression where the large coefficients are penalized by adding to the loss function a penalty proportional to the square of the coefficient. We found that the R_squared error decreases with increasing the value of hyper parameter (alpha) of ridge regression. The R_squared error stills seems to be very high.

Lasso

Lasso regression is another regularized multilinear regression which removes the unimportant features by shrinking their coefficients to zero. We fitted the train set with Lasso and found the R_squared error with the test set. It still seems to be very high.

Above models are only linear. Since there is no guarantee that the target and features are linearly related to each other, linear models tend to perform poorly in many situations. Tree based models are usually suitable for regression problems when there are a large number of features.

Light GBM and XG Boost

We used Light GBM and XG Boost with cross validation to fit the entire train data set. The predictions from these two models were ensembled to improve the score. However, we found that Light GBM alone performs slightly better than the ensemble of the two models.

4. Submission to Kaggle and Discussion

Our submission was made with the prediction of Light GBM. However, the score can be hugely improved if we exploit the data leakage present in this competition. We do not include data-leakage exploitation for this project.