GS Quantify - Bond Liquidity Prediction Model

Team: threemusketeers3
Bharat Khanna
Ayush Kedia
Ishank Arora
Indian Institute of Technology Varanasi

1 Introduction

In the given problem, predicting the buying and selling volume of the corporate bonds can be treated as a time series problem.

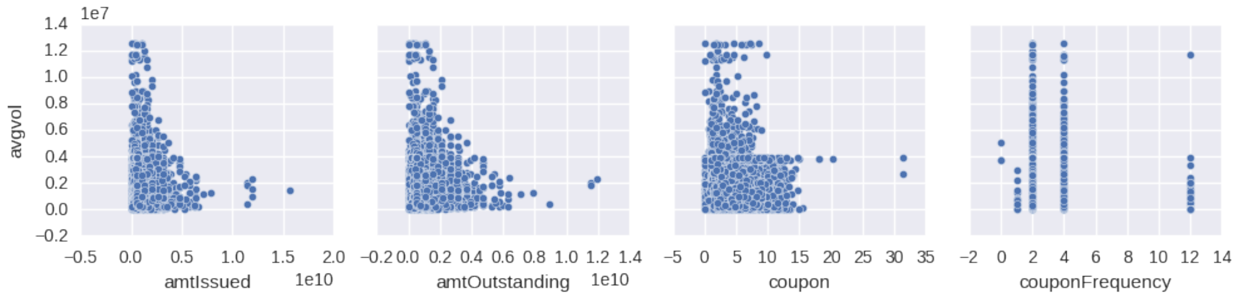
Firstly, the features of the metadata were converted to corresponding integer and float values for numeric computation. Generating the correlation between the average volume traded for a bond with its characteristics, it is evident from the attached graphs that for the features 'amtOutstanding' and 'amtIssued', there was a clear inverse dependence of the average volume on these two. Hence, these two features were reinforced by making new columns with their squared values.

As far as the other features were concerned, the columns having the date data type were ignored. The feature 'coupon' was also seen to influence the average volume, hence its weightage to the prediction was also increased by squaring its values.

2 Mean Reversion

The well-known concept of Mean Reversion was applied in all the different ideas and models for the problem. Mean reversion is the theory suggesting that prices and returns eventually move back toward the mean or average. This mean or average can be the historical average of the price or return, or another relevant average such as the growth in the economy or the average return of an industry.

So, for the bonds where the bought volume was higher than the volume sold, it was assumed that it would revert back to the mean, and so, for this case, the predictions as



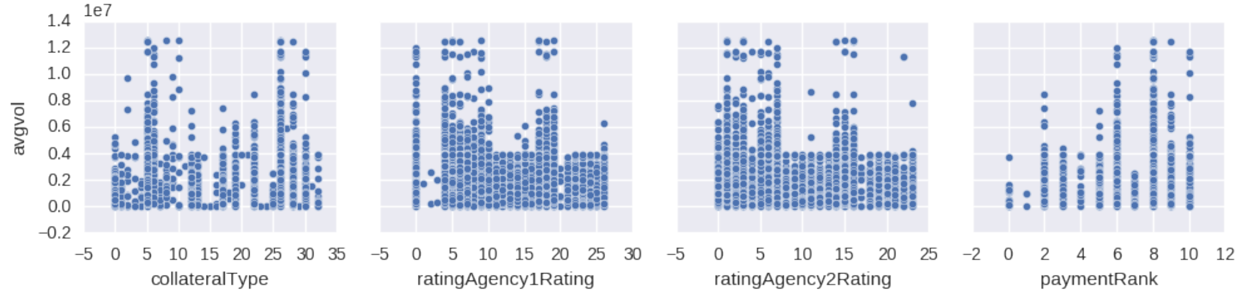


Figure 1: Dependence of average traded volume with other characteristics.

obtained from the statistical and machine learning methodologies, were reversed for the buy and sell cases.

3 Time series

The computation of future predictions was done by considering the dataset in the form of a time series, i.e., for predicting the function value $f(t)$ at a particular iteration t, $f(t-1), f(t-2)$... etc. were considered as input samples. So, each of the bond, 17261 in total, served as a training sample. The meta data characteristics along with the bond volume data from 16th March to 8th June, served as the training samples and the bond volume for 9th June was the corresponding training output.

The models were developed in IPython notebooks which have been submitted. An ensemble of Random Forest Regressors (n_estimators=200, min_samples_split=2, min_samples_leaf=50) and Linear Regression (with weights in the ratio 1:2) using the publicly available scikit-learn library was trained using the above mentioned modified dataset for buy and sell value accuracies of 41.1% and 40.7% respectively.

One point worth noticing is that the expected volume is predicted only for 10th June 2016, and the same is multiplied by 3 to obtain the total transactional volume for three days. One approach which was also implemented was that after predicting the volume for 10th June from the data of 16th March to 8th June, this predicted value would in turn serve as a feature for the prediction of 13th June volume, i.e.,

- For 10th June prediction, 16th March to 8th June data is used.
- For 13th June prediction, 17th March to 10th June data is used.
- For 14th June prediction, 18th March to 13th June data is used.

However, this approach of reinforcement learning resulted in a decrease in the accuracy and overall score. Hence, only one prediction was made in the final model.

This model incorporates the effect of the fluctuations in other bond instruments on a particular bond.

4 Additional ideas

One thing that was missed int the model was the use of issue and maturity dates in the regression model. Similar bonds such as those having the same issuer or security type could

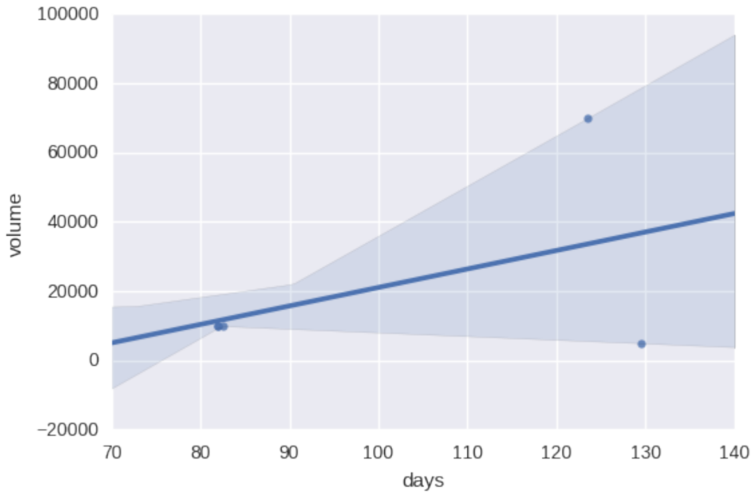


Figure 2: Buy Regression line prediction for bond 'isin0'.

have been clubbed together and the direct effect of changes in their volumes could have been further heightened.

Implementing the Dickey-Fuller test to check for stationarity in the time series was also something we read about but couldn't do due to time constaints. Decomposing the dataset into Trends, Seasonality and other components, and implementing statistical models like Auto-Regressive Integrated Moving Averages, would have provided further opportunities for analysis. This is definitely something that we would work on implementing as soon as possible.