COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

IEOR Deep Learning

Final Project

Portfolio Management based on EPS Prediction using Ensemble Gradient Boosting Models

Jintao Zheng
Rushabh Vinchhi
Xiaoyu Chen
Zequan Zhang

Supervised by
Prof. Ali Hirsa, Department of Industrial Engineering and Operations Research, Columbia University
&
Satyan Malhotra, Managing Partner, FlexStone Partners

May 2020

Contents

Abstract

The project aims to continue the work done by the previous group in predicting Earnings Surprise. To further extend the project, we have used a Machine Learning model to predict not just the earnings for the quarter but other possible important fundamental indicators that could add to the confidence of our predicted earnings and then subsequently be used to generate weights of each stock. The weight generation technique forms a unique aspect of our project and helps the portfolio achieve better returns.

We have made use of a small universe of stocks. At every step of the project we have tried to optimise and tune various parameters of the Machine Learning model so as to generate the best possible returns for the universe of stocks. We have also added a transaction cost module, which can also account for the possible transactions costs related to trading the portfolio.

Objective

The objectives that this group has set out and achieved are listed below:

1. Optimise the prediction of the Earnings per Share (EPS): We have implemented a GridSearch module for the prediction of Earnings per Share indicator. This module allows us to give the best possible prediction of Earnings per Share. This is important in reducing errors, since our trading strategy is mainly based on Earnings per Share.
2. Prediction of indicators in addition to EPS: This is an exploratory phase in our project where we try to find which other indicators besides EPS can we predict well using a Machine Learning algorithm called XGBoost.
3. Strategy: The prediction needs to be translated to a good trading strategy that makes use of predictions.
4. Evaluation: Evaluation of portfolio performance versus the benchmark index of S&P500 with transaction costs and without.

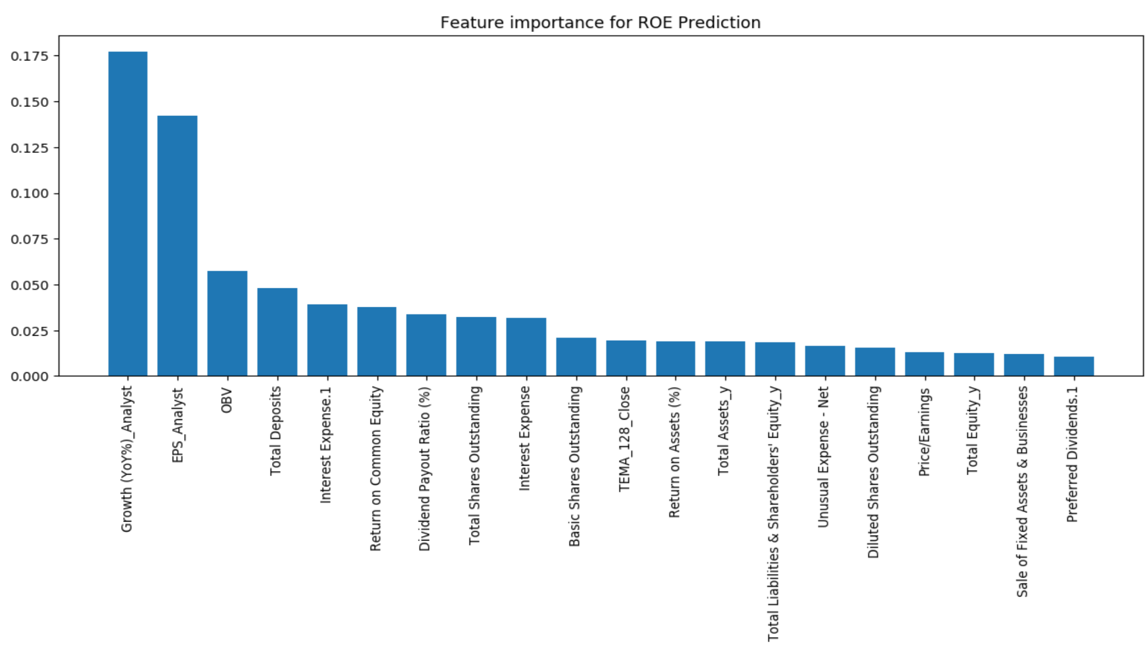
Data

The data consists of a limited universe of stocks from the banking and financial industry. The data collection still remains an extremely complex procedure of this project. Despite availability of data terminals and vendors, a very complex data processing pipeline is needed to assemble the data from these various sources. This aspect remains beyond the scope of the project.

Each stock has its own data that is primarily technical and fundamental in nature. The time series on the data runs from 1999 to 2019. Further, we have also tried to normalise the data wherever possible. However, this did not provide any material contribution to accuracy of the predictions.

The two data sets, namely, technical and fundamental are used as predictors. These different indicators add up to give a total of around 300 predictors. In Machine Learning, such a high number of predictors can lead to what is commonly known as the Curse of Dimensionality and can severely impact accuracy of predictions.

The number of predictors are reduced by using the top 40 predictors that contribute the most in predicting a value. The other predictors are dropped while making a prediction. This task is automated as a module before the prediction step. An example of predictor/feature importance in predicting ROE can be seen below:



Data is then divided into a training set and a testing set. Our results will be based on the test set.

Architecture

The architecture consists of the three parts: Prediction Phase, Exploratory Phase and Portfolio Construction phase:

1. Prediction Phase

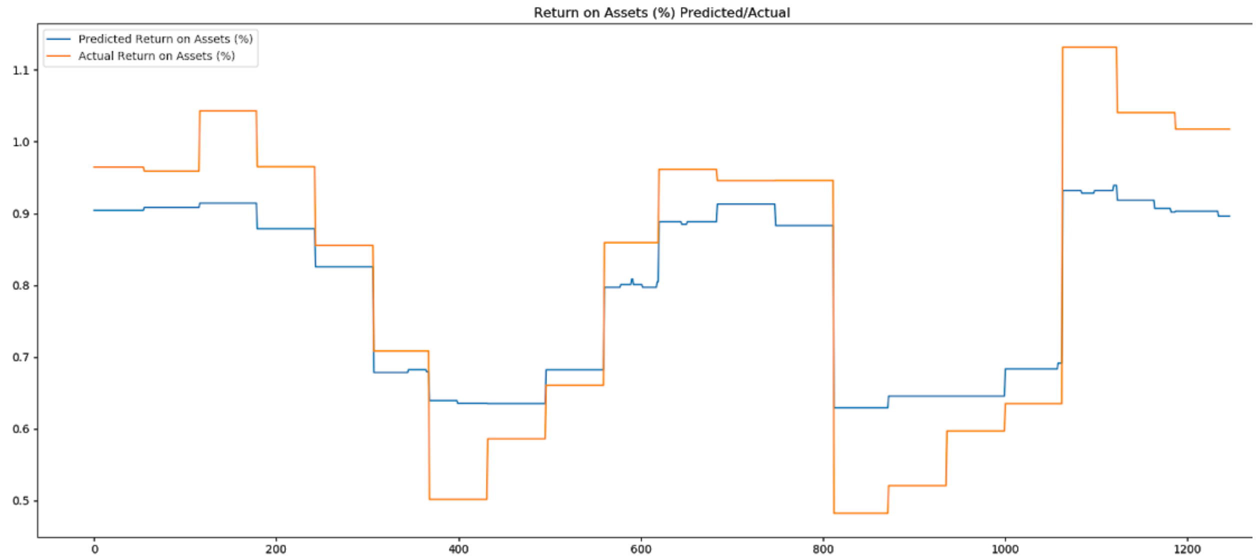
   This phase was almost entirely done by the previous group. However, we have added a GridSearch module to this phase. This module basically tries all the combinations of parameters of the Machine Learning Model that gives us the best validation error in prediction. Because of this module, we were able to improve the R squared metric of the prediction from the existing model. This metric basically tells us how well our ML model is doing in its predicting task.

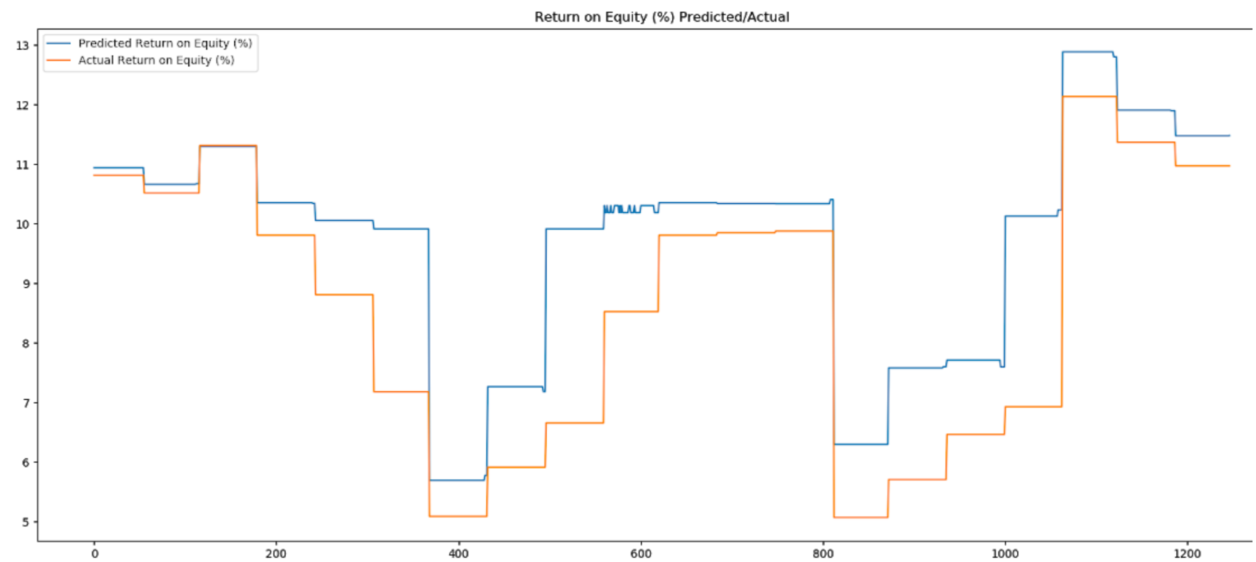
2. Exploratory Phase

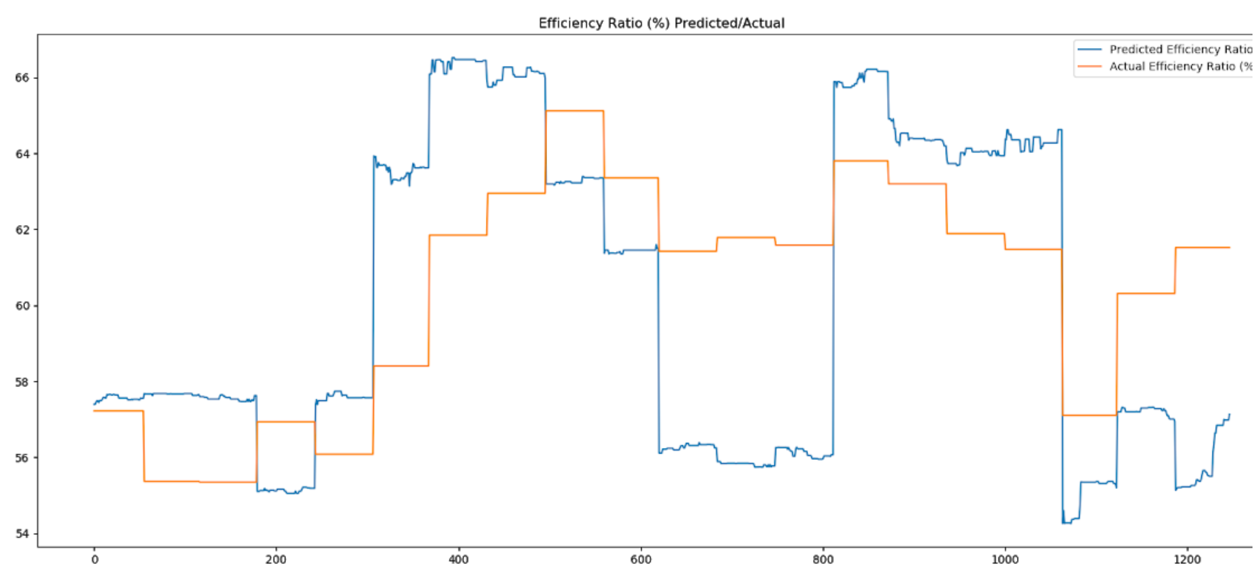
   This phase was newly implemented by our group to obtain more fundamental indicators that we can successfully predict. For this, we started off with a list of candidate fundamental indicators for each stock. Using the existing data, we used Machine Learning to predict each of these candidates. By a combination of eye balling and the R squared metric score, we selected the final three fundamental indicators that the Machine Learning model was able to predict with a very high accuracy.

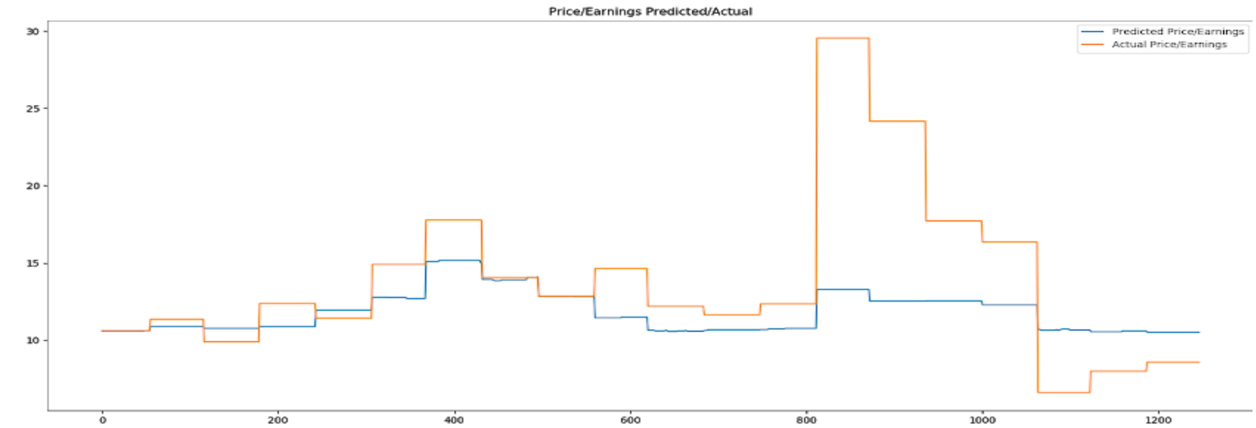
   The candidate list: Return on Assets (ROA), Return on Equity (ROE), Efficiency Ratio, Dividend Yield, Cash Deposits, Free Cash Flow (FCF) etc.

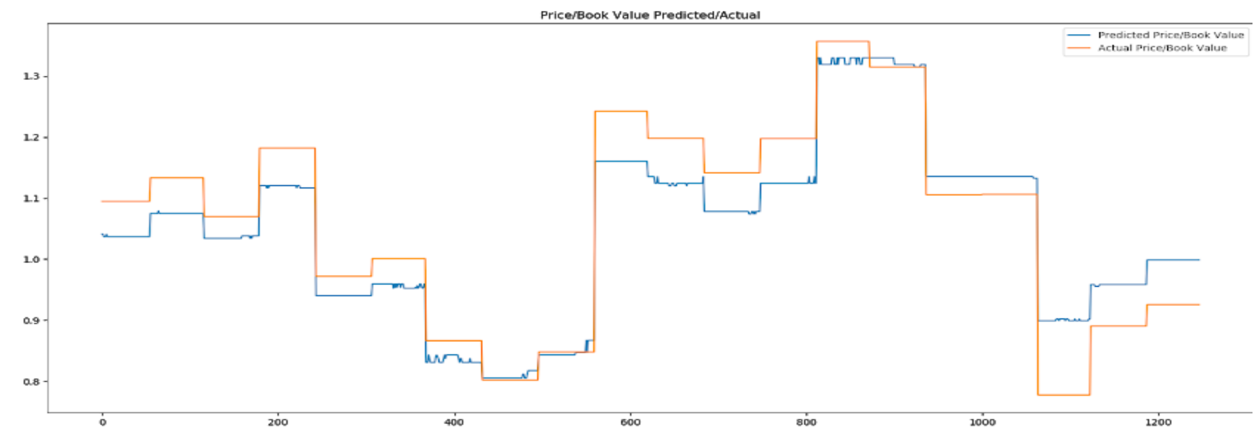
   Below are graphs that help us understand how well our ML model has been able to perform in predicting each of the above candidates.

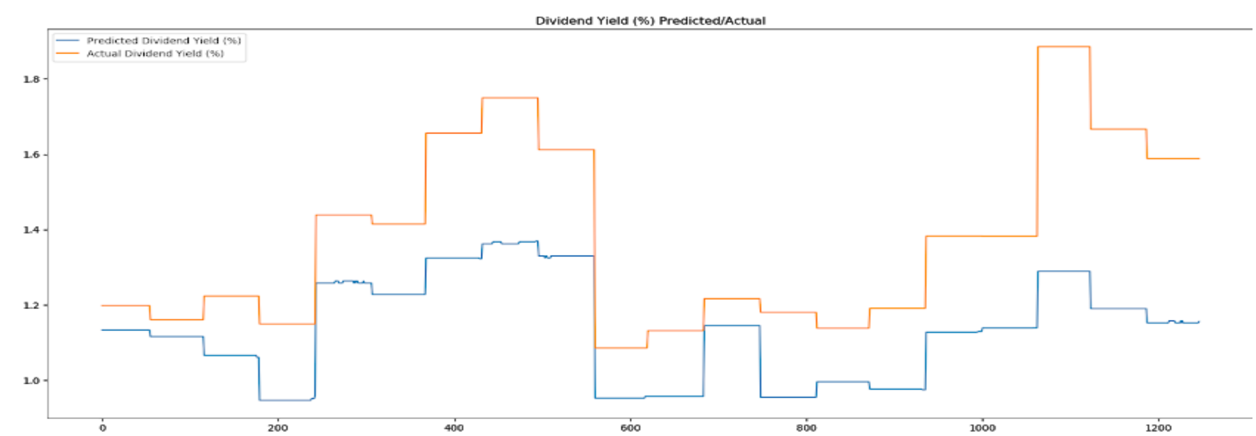

Return on Assets (%) Predicted/Actual
Predicted Return on Assets (%)
Actual Return on Assets (%)

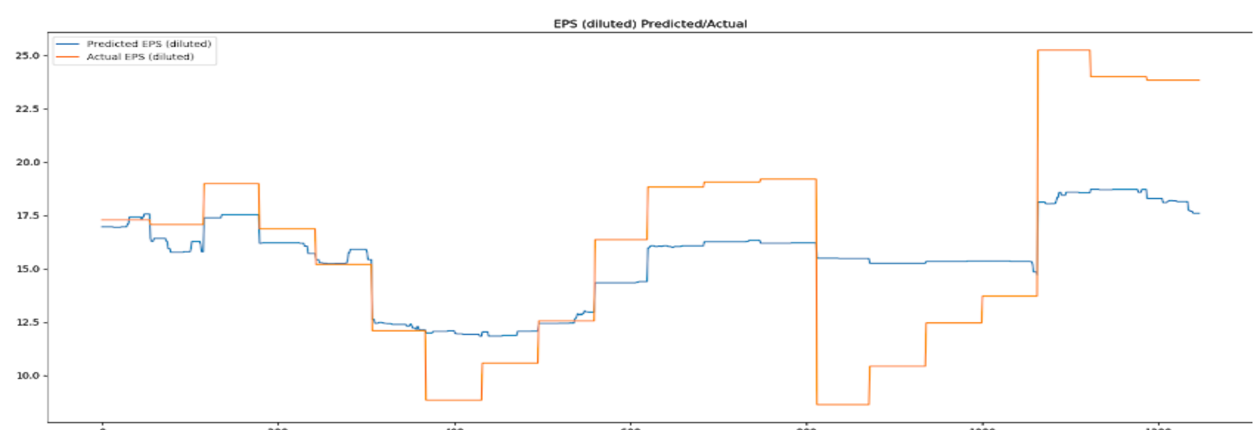

Return on Equity (%) Predicted/Actual
Predicted Return on Equity (%)
Actual Return on Equity (%)

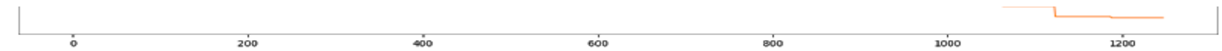

Efficiency Ratio (%) Predicted/Actual

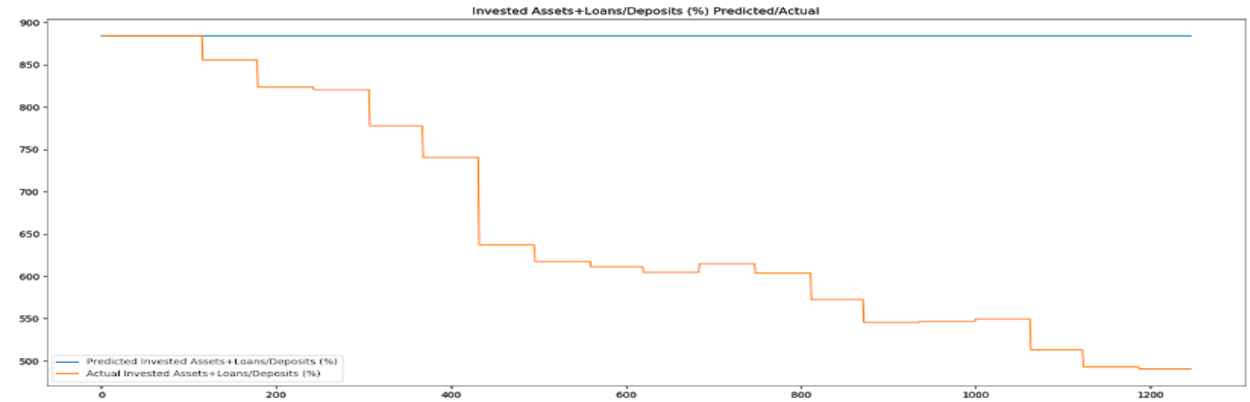

Price/Earnings Predicted/Actual
Predicted Price/Earnings
Actual Price/Earnings

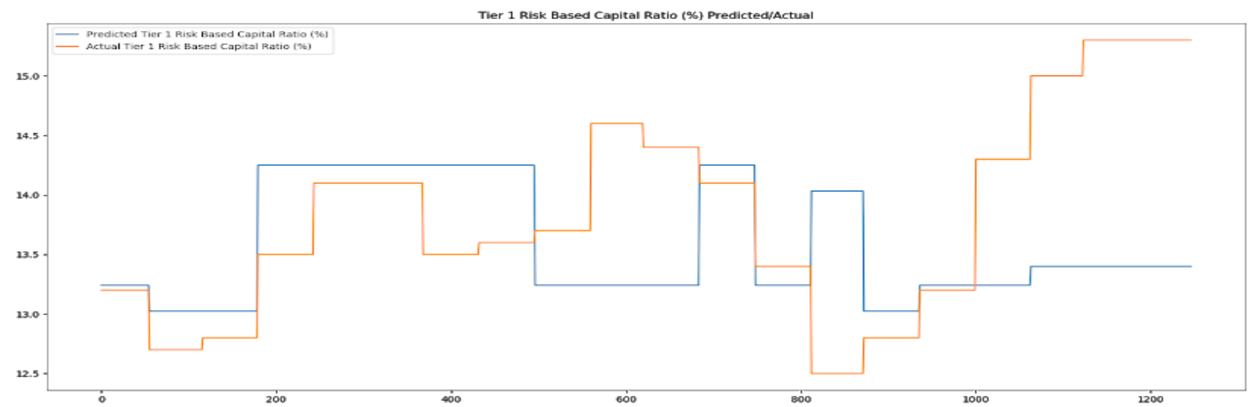

Price/Book Value Predicted/Actual
Predicted Price/Book Value
Actual Price/Book Value

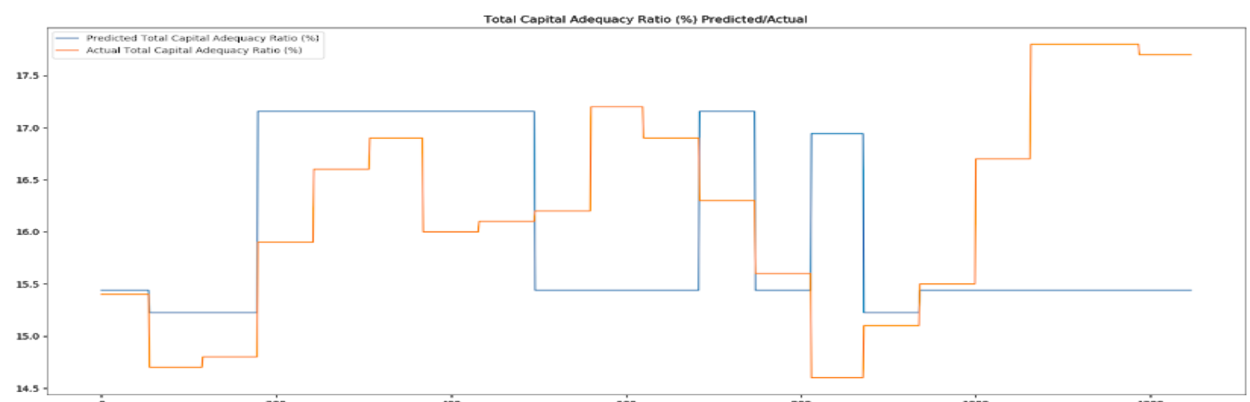

Dividend Yield (%) Predicted/Actual
Predicted Dividend Yield (%)
Actual Dividend Yield (%)

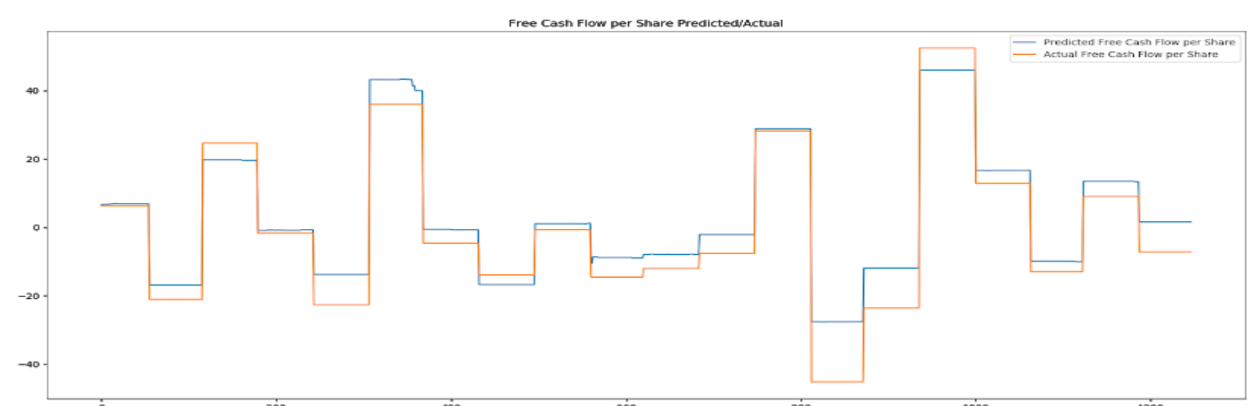

EPS (diluted) Predicted/Actual
Predicted EPS (diluted)
Actual EPS (diluted)




Invested Assets+Loans/Deposits (%) Predicted/Actual
Predicted Invested Assets+Loans/Deposits (%)
Actual Invested Assets+Loans/Deposits (%)


Tier 1 Risk Based Capital Ratio (%) Predicted/Actual
Predicted Tier 1 Risk Based Capital Ratio (%)
Actual Tier 1 Risk Based Capital Ratio (%)


Total Capital Adequacy Ratio (%) Predicted/Actual
Predicted Total Capital Adequacy Ratio (%)
Actual Total Capital Adequacy Ratio (%)


Free Cash Flow per Share Predicted/Actual
Predicted Free Cash Flow per Share
Actual Free Cash Flow per Share

Even by mere eye-balling one can see that the ML model is able to do a very good job with ROA, ROE and FCF.

So these are the three fundamentals in addition to EPS, we wish to predict.

3. Portfolio Construction

Now that we have predictions on EPS, ROA, FCF and ROE we wish to construct a portfolio of the universe of stocks that can make use of this information to maximise returns in the stock market.

For this, we included another feature in this phase called the recency ratio[3]. According to this theory, analysts overweight recent low or high earnings. We can use this recency bias to add more weight to a stock that has recently posted low earnings because all the analysts would tend to give a conservative forecast. However, our Machine Learning model is free from such bias and thus can exploit this opportunity to make substantial gains from the surprise in earnings.

The recency ratio, according to the study can be quantified as:

$$rr = 1 - (number\ of\ days\ since\ 52\ week\ high\ /\ 2520)$$

The higher the ratio, more the stock weight should be.

Having all this information, we came up with the following simple model to construct weights.

For the i th stock we have:

$$M_i = eps_i + (roa_i/10) + (roe_i/10) + (fcf_i/10) + rr$$

Here eps, roa, $roe\ and\ fcf$ denote surprises in the respective fundamentals that we have predicted in contrast to existing perceptions in the market (analyst predictions).

After we have M_i we can construct the weight of each stock W_i as:

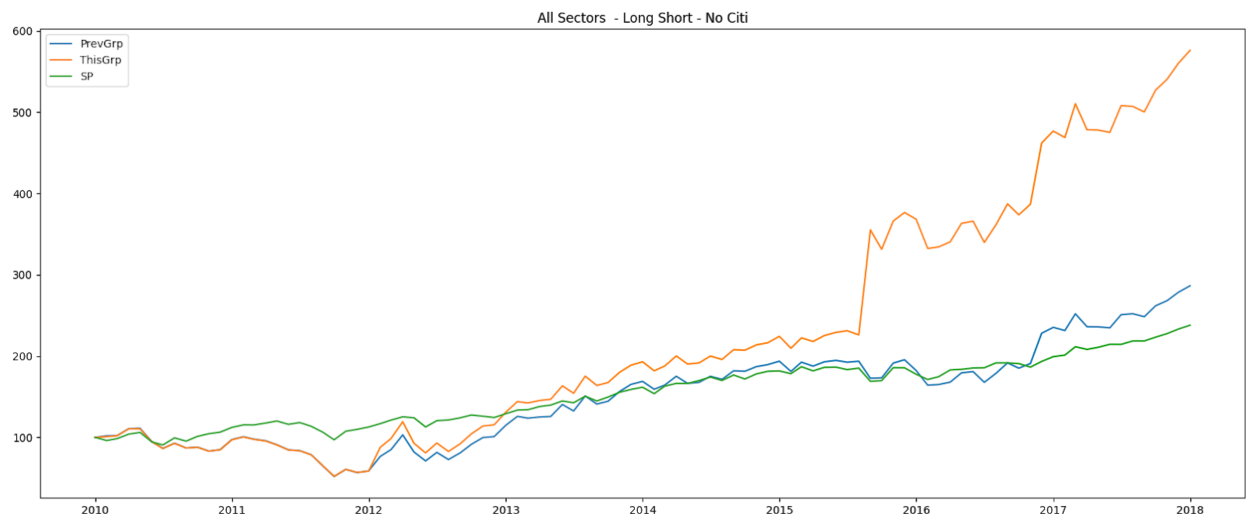
$$W_i = M_i / \sum M_i$$

This method of portfolio construction is clearly an extension to what the previous group had done. Since we had a robust machine learning model, the same model was then used to predict more fundamentals and get increased confidence in assigning weights to each stock for trading.

An interpretation of the above equation could be: Both stocks P and Q are in line to post earnings surprises. To present our case let the surprise of the earnings be almost comparable. Now when we consider additional surprises in ROA, ROE and FCF, we are able to favor the stock that beats all three expectations and thus would intuitively generate larger returns.

To validate our strategy we need to see how our portfolio built with the aforementioned weight distribution compares with the previous group and the benchmark S&P.

Below we can find some plots for returns for the period we traded that validate our claim:



As we can see, there is a clear improvement from the previous group and this is because we harness more information using other fundamental indicators.

Having done this, we also included a module that would handle transaction fees. This fee can be changed suitably as what is applicable at the time.

Conclusions and Future Work

- The model takes advantage of more information by predicting multiple fundamental indicators and not just relying on earnings per share.
- This additional information certainly helps in establishing higher returns as is seen above.
- However, the model needs to be appended further. More complex relationships need to be discovered between weights and the surprises so as to capture market dynamics perfectly.
- The portfolio, while seems to do better in returns, however suffers from a lower sharpe ratio. This particular feature might make it unattractive to many investors.
- Further work should include finetuning the Sharpe and excavating more complex relationships to find stock weights.
- Also work can be taken up on automated data collection specifically tailored for this project needs.

References

[1] Tom Y. Chang*, Samuel M. Hartzmark†, David H. Solomon* and Eugene F. Soltes‡, Being Surprised by the Unsurprising Seasonality and Stock Returns

[2] Ma, Qingzhong and Whidbee, David A. and Zhang, Wei Athena, Recency Bias and Post-Earnings Announcement Drift (July 19, 2014). Available at SSRN: https://ssrn.com/abstract=2469308 or http://dx.doi.org/10.2139/ssrn.2469308

[3] Brandt, Kishore, Santa-Clara, Venkatachalam: Earnings Announcements are Full of Surprises