Forecasting Stock Prices, Predicting Earnings Per Share and Classifying Bankruptcy

Pratik Sahoo and Swetha Kallam
Instructor: David Belanger

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

INTRODUCTION

Stock market is very dynamic in nature and it is a constant struggle to determine the stock value in near future. With the advancement of data science techniques, the challenges of predicting stock prices has been better addressed. This is our approach in forecasting stock prices, predicting earnings per share and classifying bankruptcy

TOOLS

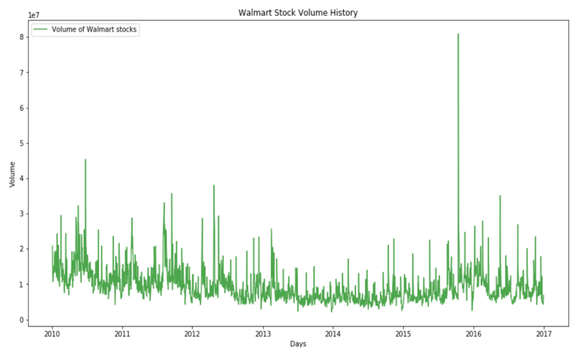
* Python (Sckit-learn, Pandas, Numpy, Matplotlib, Random Forests, Linear Regression, K-means)
* Tableau

FORECASTING CLOSING STOCK PRICES

Data Understanding and Exploring

The NYSE prices dataset is a time series data set of 851264 records depicting the daily prices of stocks for 501 companies, spanning from years 2010 to 2016.

We choose to forecast closing stock price of Walmart. It had 1762 records with 7 columns.



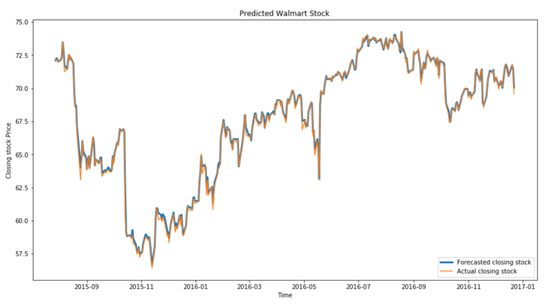
We couldn't do the regular predictions on this kind of data, as the stock market rapidly changes every day, it makes no sense if we are predicting todays stock value based on stocks of last 5 years .

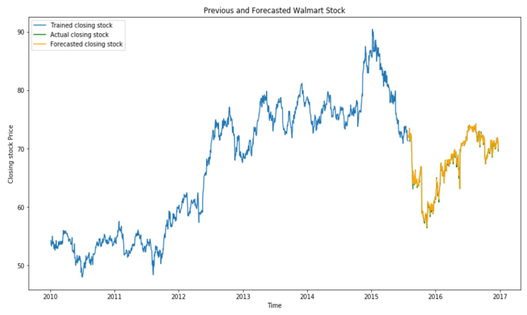
So we are predicting a particular day's closing stock value based on past five days.

date	open	close	low	high	volume
2010-01-04	53.740002	54.230000	53.669998	54.669998	20753100.0
2010-01-05	54.090000	53.689999	53.570000	54.189999	15648400.0
2010-01-06	53.500000	53.570000	53.419998	53.830002	12517200.0
2010-01-07	53.720001	53.599998	53.259998	53.750000	10662700.0
2010-01-08	53.430000	53.330002	53.020000	53.529999	11363200.0
2010-01-11	53.330002	54.209999	53.099998	54.439999	13987700.0
2010-01-12	54.000000	54.730000	53.860001	54.750000	15117000.0
2010-01-13	54.790001	55.009998	54.410000	55.200001	13290700.0
2010-01-14	54.730000	54.209999	54.160000	54.840000	13772000.0
2010-01-15	54.340000	53.680000	53.599998	54.549999	19087500.0
2010-01-19	53.910000	54.029999	53.500000	54.259998	14618500.0
2010-01-20	53.919998	53.860001	53.230000	53.939999	13308600.0
2010-01-21	54.070000	52.919998	52.840000	54.070000	16028700.0

Forecast Modelling

We used Linear Regression to train the model and it was trained with an accuracy of **83.97 %** .





Evaluating the Forecasting model

Performance Metrics of the Forecasting Model	
Metrics	**Results**
Mean Absolute Error	1.2513082
Mean Squared Error	3.4253527
Root Mean Squared Error	1.8507708
Mean Absolute Percentage Error	0.20973468
Accuracy	79.27 %

PREDICTING EARNINGS PER SHARE

Data Understanding and Exploring

The NYSE fundamentals data set has 1781 records with 77 feature metrics from annual SEC 10K fillings (2012-2016) .

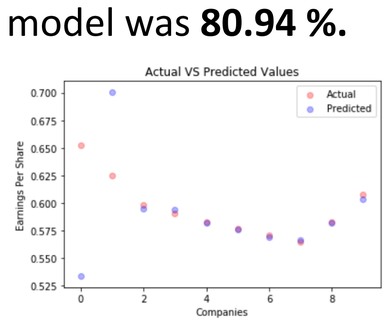
We started of with Min-Max scaling and since there were huge number of features, we had to scale them down.

Feature Selection

We used Radom Forests classifier for feature selection, since random forests considers correlated features as well, to over come this, we used co-relational matrix to trim out highly correlated features. We successfully got the 77 features down to 31 features.

Predictive Modelling

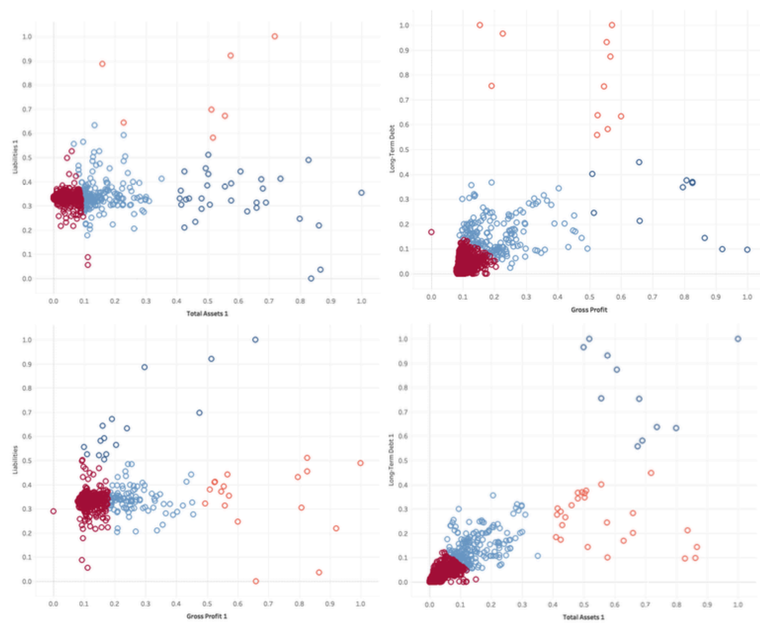
We used Random Forest Regressor for predicting earnings per share and training model was **80.94 %.**



Evaluating the Predictive model

Performance Metrics of the Forecasting Model	
Metrics	**Results**
Mean Absolute Error	0.02112
Mean Squared Error	0.00119
Root Mean Squared Error	0.04438
Mean Absolute Percentage Error	0.03317
Accuracy	96.66%

CLASSIFYING BANKRUPTCY



Using K Means Clustering, we classified 2 companies F and VZ that might go bankrupt in near future, using features like Total Assets, Gross Profit, Net Income, Total Equity, Long Term Debt.

CONCLUSIONS AND FUTURE WORK

* Linear Regression in our dataset has performed well for forecasting the shares. Nonetheless, we would like to delve into LSTM.
* Though, our predictive model using random forests has performed extremely well in predicting earnings per share in our scenario, it would be interesting to see how it would perform if we have enough data for each company.
* Using K-Means Clustering we classified companies that might go bankrupt in future as per the last 4 years trend, nonetheless the classification can be improved if we have more data for the companies, which would give us better results.