TO BUY OR NOT TO BUY

By: Chuan Fu, Yap

IDEA:
Make investing in stock (relatively) safe

- Investing in the stock market can be risky to some, and considered as gambling.
- What if the risk can be reduced?
- What if you don't have to deal with numbers?
- How?
- Let the machine find patterns human eyes cannot see and tell you what next!

APPROACH

- Dataset: Stock data of company of interest (case study here will use Microsoft's stock) and investor sentiment analysis data. Though the latter is a weekly data, hence all results are weekly (stock data is formatted to match it via averaging).

- Methods:
  i. Data transformation to find meaningful features for modeling purpose.
  ii. Apply machine learning methods to make predictions via Regression, and Classification.
  iii. Validation of the models from Step 2.

Microsoft Data:

- Microsoft Dataset provides, 5 different daily data of interest:
  - Opening Value
  - Closing Value
  - Highest Value
  - Lowest Value
  - Volume traded

Sentiment Data:

- Sentiment data which are polled from individual investors gives their sentiment on the stock market which includes:
  - Bullish (If the stock market will do well)
  - Neutral
  - Bearish (If the stock market will do badly)

Data Transformation

- In order to make future predictions using available data. They will be transformed like so:
  - Both stock data, and sentiment data will have a rolling window of mean, median, standard deviation and mean of gradient of the past 9 weeks as features of the 10^{th} week.

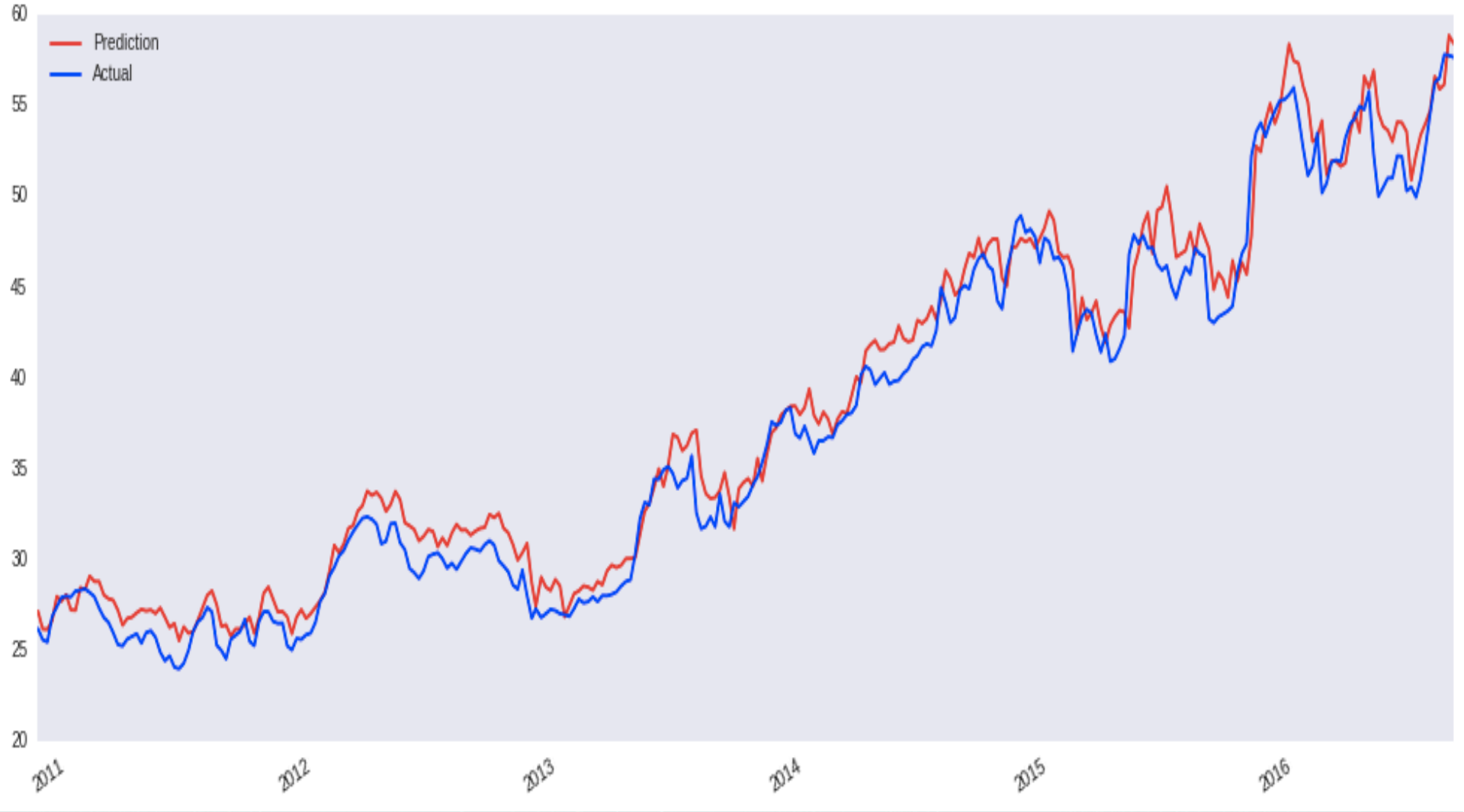
  Tldr;

  - 10^{th} week = Target
  - Past 9 week leading to the 10^{th}'s mean, median, etc = Features/Data for modelling

Regression Analysis

- Linear regression, Support Vector Machine and Random Forest modeling were all used on the stock's Opening value, however the best results was from Linear Regression.
- With the following score on model validation:
  - Pearson Correlation Coefficient: 0.992
  - Normalised RMSE: 0.785

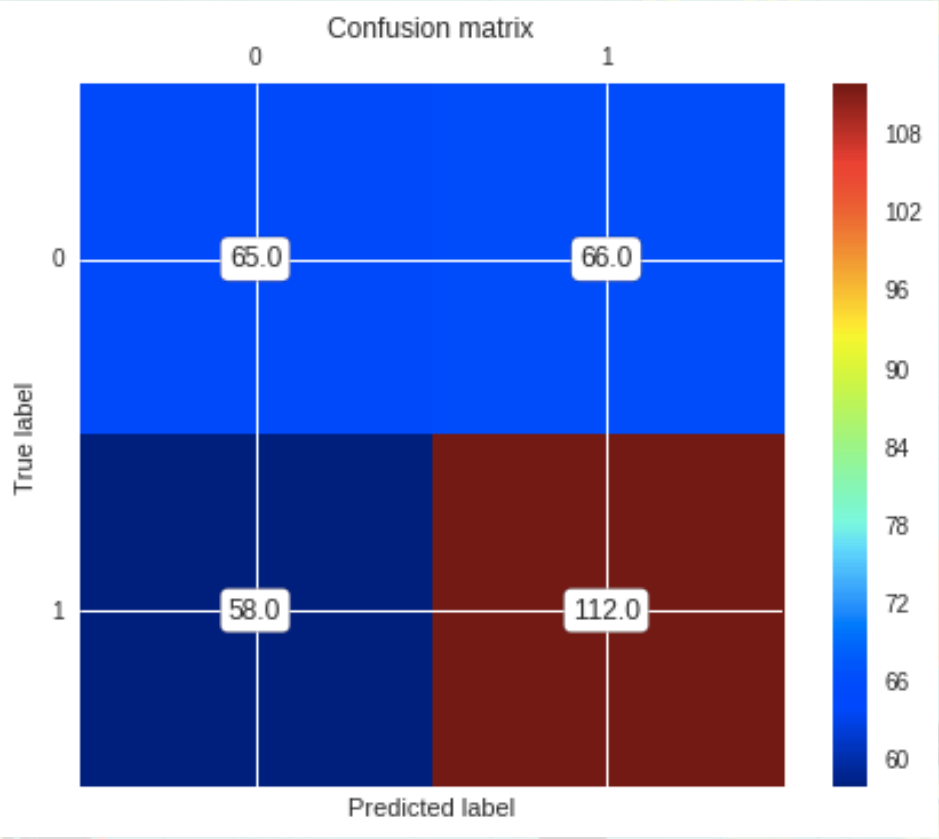
Linear Regression doing wonders



Is it going UP or DOWN?

- Graphs and numbers are nice and all, but what if all you want to know is if the stock will increase or decrease in value?
- Let's turn the values into 2 separates classes, up from the week before, down from the week before.
- For this, Random Forest Classifier was applied.

Classification Results:



	Precision	Recall	F1-Score	Support
Down	0.45	0.44	0.44	124
Up	0.62	0.63	0.63	178
Avg/Total	0.55	0.55	0.55	302

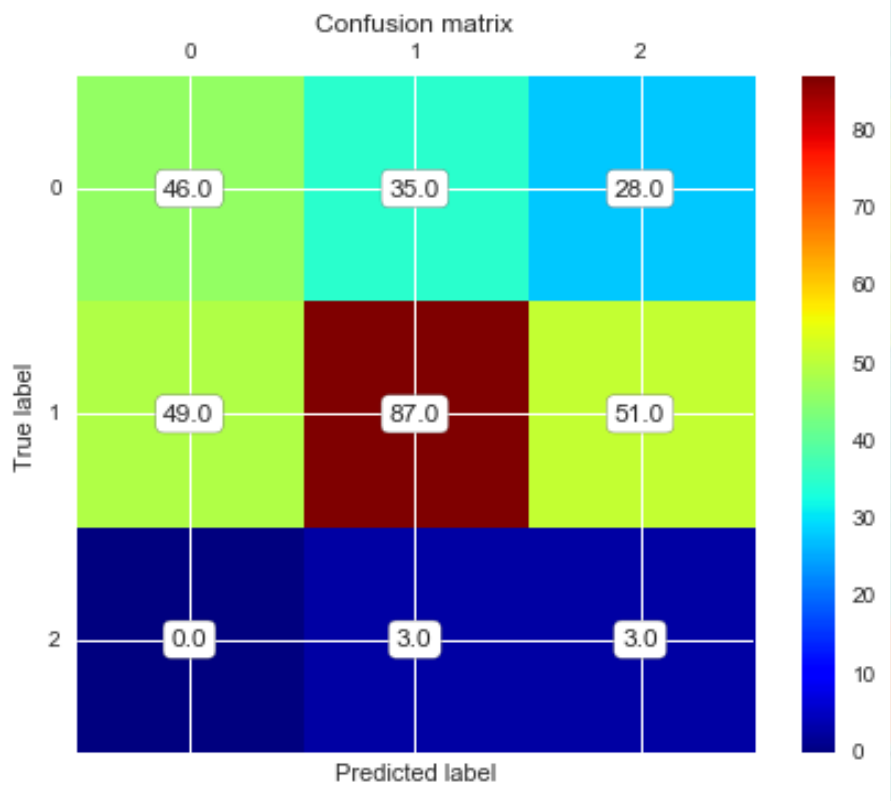
Analysis:

- Is it great? Not really.
- But is it bad? Well. . . I mean it could be better.
- Okay, knowing if it'll go up or down is not that helpful for clueless investor what if the machine can tell them whether to BUY, SELL or HOLD

BUY, SELL or HOLD

- How these classes are generated.
  - If Closing value > Opening value = BUY
  - If Closing value < Opening value = SELL
  - If stock future stock > current stock = HOLD
- Let’s see what Random Forest cooks up…

Results TWO:



	Precision	Recall	F1-Score	Support
BUY	0.50	0.04	0.07	82
SELL	0.42	0.48	0.45	85
HOLD	0.47	0.70	0.56	125
Avg/Total	0.46	0.45	0.39	302

Classification is hard

- In short, turning this continuous dataset into a classification problem will need to revisited.

- Details:
  - When predicting just two classes, neither predictions were all that good, with precision that is below 80%.
  - Increasing number of classes did not help, prediction were at most 50% accurate, while doing very poorly when it comes to BUY class.

What now?

- This pipeline of transforming stock and sentiment data then applying linear regression, can be used on other stocks.

- While, this may not have simplify the problem all that much, this can be used to inform would be investors on making safer choices.

What next

1. More data can be found to match the company of interest depending on the sector, which can be used to further inform the models.
2. Further understanding the stock market and higher quality data (hourly or daily) instead of weekly can be used to generate better classes.

Acknowledgements

I would like to thank for mentor, Shubhabrata Roy for his patience and guidance throughout this project.

…and of course Quandl.com for the free dataset.