UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

MASTER OF DATA SCIENCE

Twitter Sentiment Analysis on DogeCoin Stock Performance

Group Member:

- Yufeng (Adam) Xing
- Xinming Wang
- Nancy Ruan
- Tina Liu
- Summer Zhang

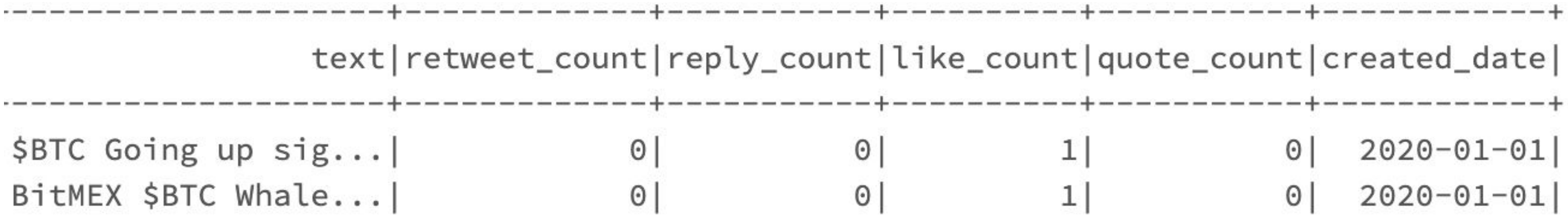
Course Name: Distributed Data System (MSDS697)

Table of Content

- Description of data and data collection process
- Analytical goal
- Models included in the analysis
- Preprocessing algorithm & Time Efficiency
- Findings (one slide for each objective and model)
- Conclusion

Description of Data and Data Collection

- Data Size : 24036
- Date Range: 2020-01-01 – 2020-05-09
- Data Collection:
    - Pull Tweets related to DogeCoin
    - Pull DogeCoin Stock Price
    - Merge two dataframe



```
+--------------------+-------------+-----------+----------+-----------+------------+
|                text|retweet_count|reply_count|like_count|quote_count|created_date|
+--------------------+-------------+-----------+----------+-----------+------------+
|$BTC Going up sig...|            0|          0|         1|          0|  2020-01-01|
|BitMEX $BTC Whale...|            0|          0|         1|          0|  2020-01-01|
```

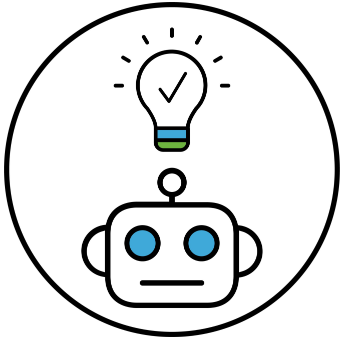
Analytical Goal

- Predict the daily stock price given the number of tweets on a given date (tweets volume)
- Add retweet, reply, comment, like counts as features and predict price (y = daily stock price)
- Add text sentiment score into the model and predict the daily stock price

Possible Models

Since the target variable is the stock price, our problem is a regression question, thus following models are chosen for the evaluation:

- Linear Regression
- RandomForest
- H2O autoML



Data Preprocessing, Model - Time Efficiency

Analytical Goals	Model	Model Run Time	Data Preprocessing	Person Responsible
Objective 1 - Tweet Count	Linear Regression	74.8 ms	<6s	Tina
Objective 1 - Tweet Count	Random Forest	25.75 m	<6s	Tina
Objective 2 - Multi-features	Linear Regression	1.82 s	23.99s(data cleaning) + 849 ms+(standardization)	Adam
Objective 2 - Multi-features	Random Forest	4.31 s	23.99s(data cleaning)	Xinming
Objective 3 - Sentiment	H2O	2 min 2s	2 min 43s (data cleaning) + 5 min 10s(toH2OFrame)	Nancy
Objective 3 - Sentiment	RandomForestRegressor	3 min 50s	2 min 43s (data cleaning) + 14 min 19s (Log/FV/Normalize)	Summer

Objective 1 Findings

- Analytical Goal: Predict the daily stock price given the number of tweets on a given date (tweets volume)
- Findings: Both the R^2 of test set of and Linear Regression and Random Forest < 0.1 - Tweets Counts and DogeCoin stock price are not significantly correlated.   (R2_RF : -0.11 ; R2_LR : 0.06)
- Possible Explanation:
    - The data we collected is not enough to generate some results
    - The fluctuation of stock price depends on so many factors - and tweets count is not important enough to explain this fluctuation.

The model runs on a cluster of 30.5 GB Memory, 4 Cores for both driver and workers, with # worker of 8

Objective 2 Findings

- Analytical Goal: use retweet count, reply count, like count, quote count as features to predict close price of stock
- Data preprocessing:
  - Enlarge close price 1000x (avoid underflow & float accuracy problem) - 18.1 s
  - Take the average of counts per day as features - 5.89 s
  - Standardized the features with standardScaler
- Linear regression model
  - R2 score: 0.5115
  - MSE: 0.05533
- The model runs on a cluster of 30.5 GB Memory, 4 Cores for both driver and workers, with # worker of 8

Objective 2 Findings

Features	Model	Time	R^2	MSE
Avg counts per day	Random Forest	4.31 s	0.931	0.0094
Avg counts per day	Random Forest with cv=5	3min4s	0.939	0.0094
Log (avg counts per day)	Random Forest	36.6 s	-0.160	0.1155
Truncate 95% (avg counts per day)	Random Forest	25.1	0.925	0.0074

- Data Exploration

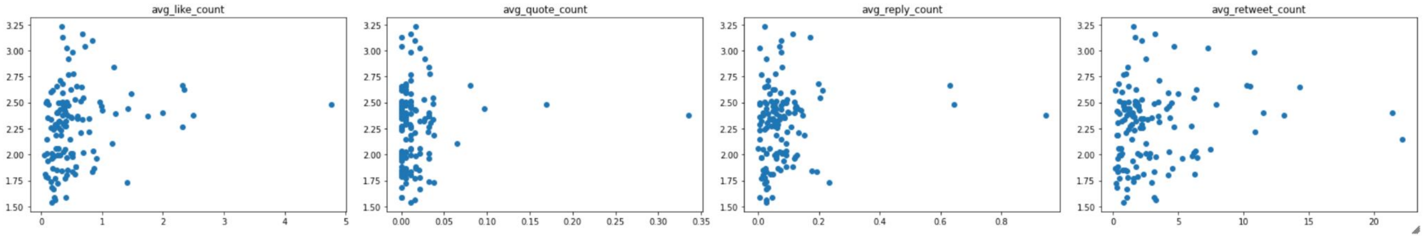
By looking at the distribution of features and the scatter plots, we find that our data was right skewed, which may lead to poor prediction result.

- Feature Engineering

We trained four random forest models to predict the stock price. The naive one without feature transformation got an extremely high R^2, which is counter-intuition. After taking the log, R^2 becomes negative.

- Possible explanation

It's hard to tell why our data is right skewed. if it's due to the rare market shock, then the high metric value from original model makes sense; if it's due to outliers, the truncate one makes sense; if it's due to lack of data, then the metric is not reliable



Objective 3 Findings

Objective: add sentiment score into the model and predict the daily stock price (see how closely the sentiment is related to the stock performance)

Sentiment Score Conversion Package:

- We used SentimentIntensityAnalyzer to convert cleaned text to a score
- converted score to a number between -1 and 1 (negative and positive)

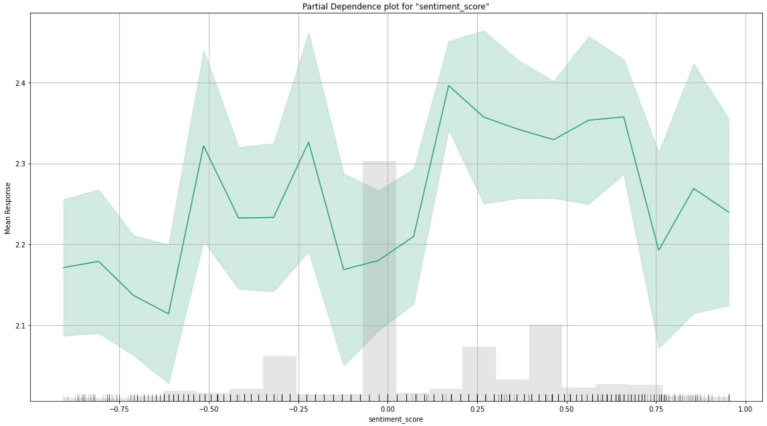
RandomForestRegressor

- **RMSE:** 0.3640
- **MSE:** ~0.1325
- **R2:** ~-0.0271

H2O:

- **Best Model:** StackedEnsemble_AllModels_4_AutoML_3_20220309_
- **MSE:** ~0.1165
- **R2:** ~0.1608
- **Partial Dependence Plot:** with many neutral sentiments, we still see negative sentiment tweets would yield to low stock closing price, vice versa.

The model runs on a cluster of 30.5 GB Memory, 4 Cores for both driver and workers, with # worker of 8



Lessons Learned + Conclusion

- Different models could be used to explore different analytical goals; when we have more features, more time is required to do data cleaning and model fitting.
- Tweets Counts and DogeCoin stock price are not significantly correlated. (R2_RF : -0.11 ; R2_LR : 0.06)
- We notice some extremely validation metrics results when we add more features (reply count, like count, quote count) to the model. Some of the possible reasons have been discussed before.
- When we take **tweet sentiment** into considerations, with the H2O model, only 16.1% of the variation was explained. The sentiment score is making some influences on the stock price prediction, though limited. H2O would more likely to select the best model comparing to we individually implement one single model, as it could ensemble good models together.
- There might be many possible limitations in our dataset, e.g. right skewed data, limited amount, etc. Further investigations are required to build a better model. For future steps, we would like to gather more data and also include other possible features should be introduced.