CAPSTONE
PROJECT

by Joyce Ooi

PROBLEM STATEMENT

- There is difficulty in selecting stocks with consistent Return on Investment (ROI) due to market volatility.
- To identify stocks that allow us to predict consistent returns to allow portfolio diversification, by identifying stocks that correlate to each other or correlate to economic indicators.
- To train a classifier to predict whether a stock will generate positive or negative intraday returns. Success was evaluated via F1 score as well as Accuracy and Specificity.

TECHNOLOGY STOCKS

FANNG

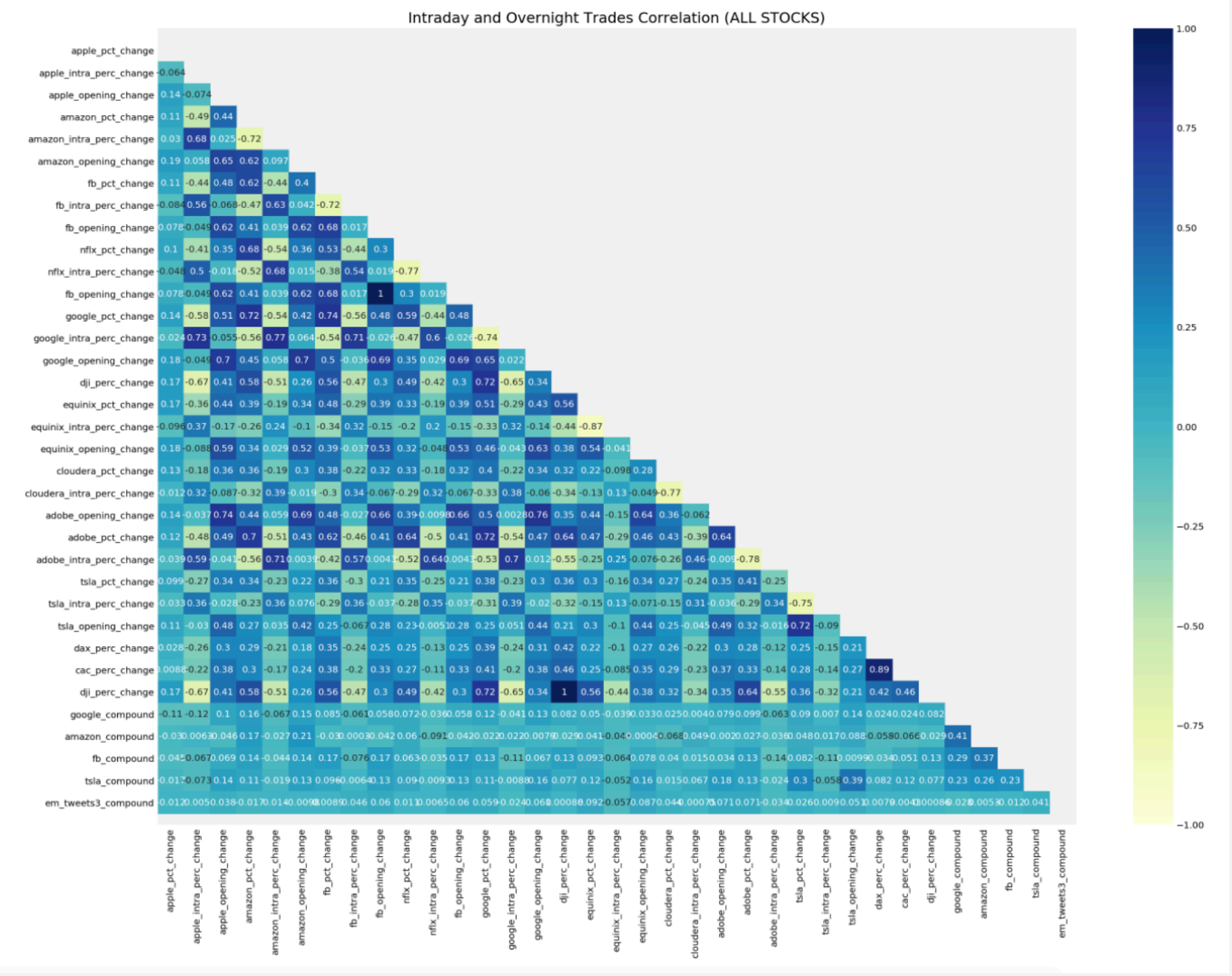
Facebook
Apple
Amazon
Netflix
Google

Cloud

Adobe
Cloudera
Equinix

Other

Tesla
Microsoft



- FAANG stocks are highly correlated
- Strong negative correlation between ‘pct_change’ and ‘intra_perc_change’,
- Apple stock does not correlate with other stocks
- Regional Indexes show low correlation to stocks price changes in general
- Sentiment shows low correlation to stock price changes in general (Tesla, Amazon and Facebook show slightly more correlation)
- Tesla stock correlates less with other stocks

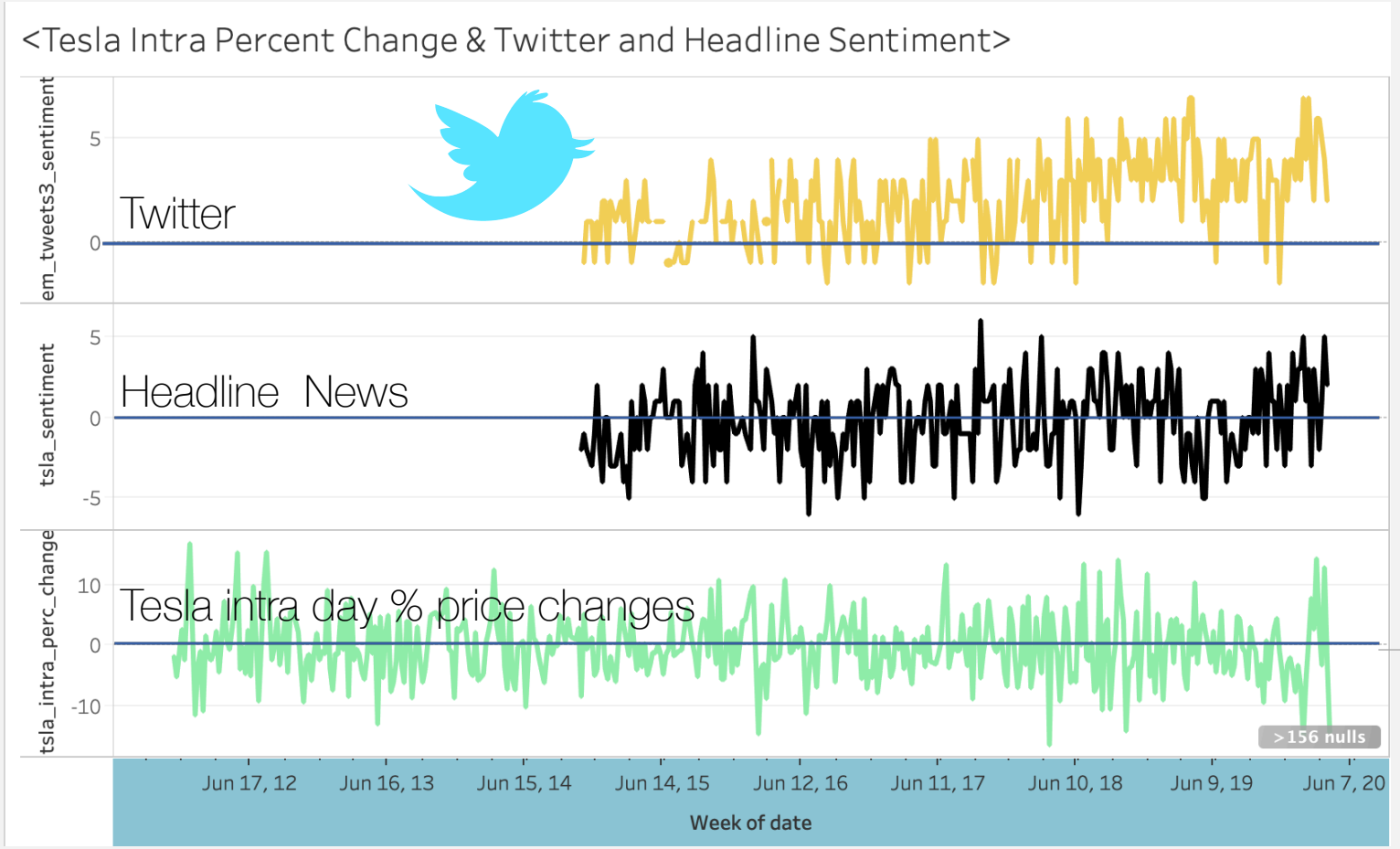
- Tweet frequency and Twitter following
- Relatively higher correlation of headline sentiment to stock price changes
- Opportunity for portfolio diversification due to low correlation with other stocks

STOCK

SELECTION : **TESLA**

VADAR

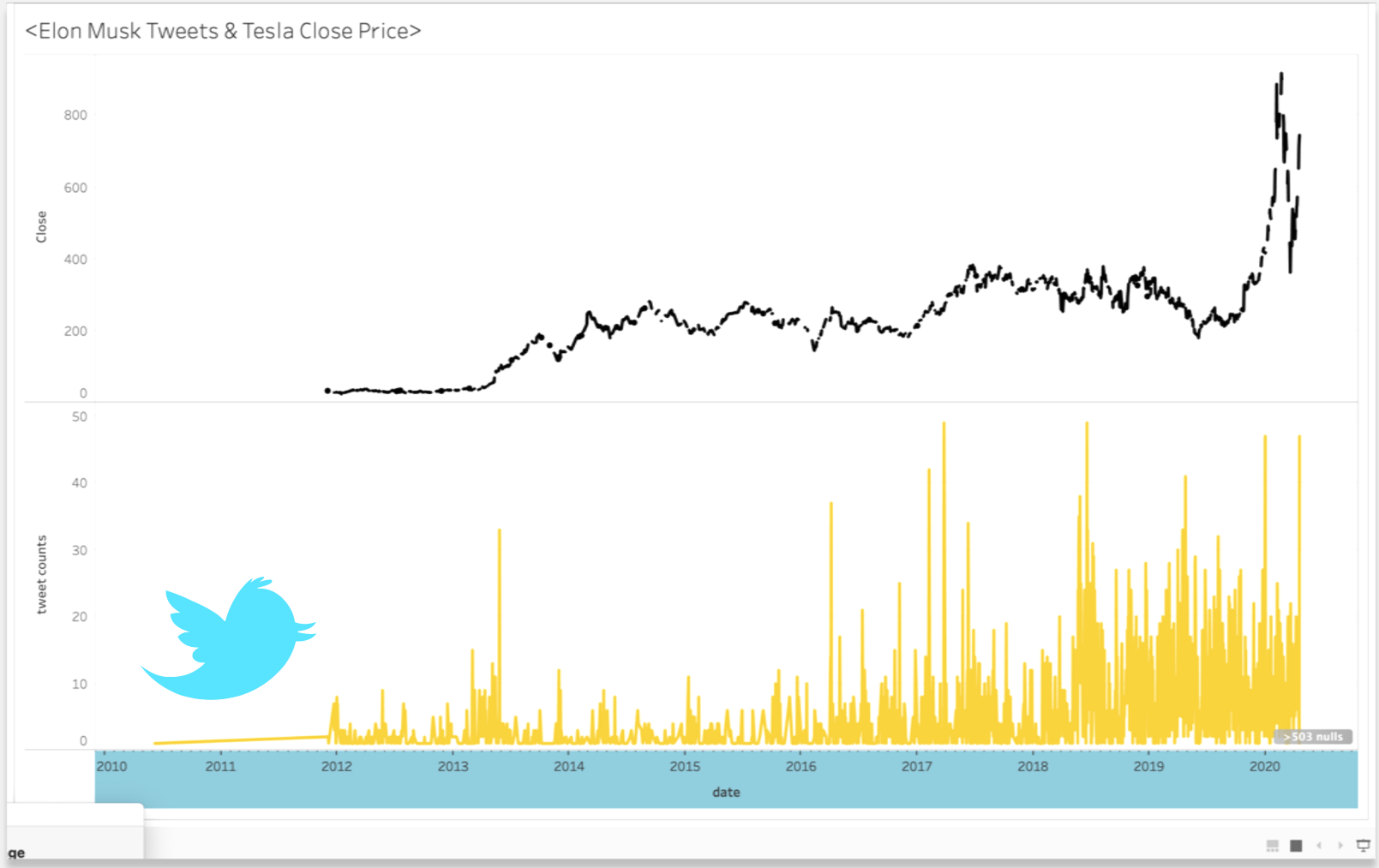
SENTIMENT SCORES



Relatively weak correlation of twitter and headline sentiment to share price changes

VADAR

TWEET COUNTS vs SHARE PRICE



713	2020-01-23 15:27:51+00:00	We should have a base on the moon, a city on Mars
714	2020-01-23 06:44:49+00:00	I love that Heart of Gold is moved by Infinite...
715	2020-01-23 06:12:10+00:00	We should strive to extend the light of consci...
716	2020-01-23 05:25:05+00:00	Star Peace
717	2020-01-22 13:04:04+00:00	We're working on it
718	2020-01-22 12:13:30+00:00	Agreed, v important
719	2020-01-22 12:02:54+00:00	Blazing Saddles
720	2020-01-22 07:23:38+00:00	Working on it
721	2020-01-22 07:19:46+00:00	It's on the list
722	2020-01-22 07:15:20+00:00	Want to play The Witcher game on your Tesla? (...
723	2020-01-22 07:02:27+00:00	Explains the sad lack of progress in candy res...
724	2020-01-22 05:52:01+00:00	Yeah, doors are 40 ft wide
767	2020-01-16 22:41:18+00:00	Densifying hydrogen is difficult, as its liqui...
768	2020-01-16 20:12:34+00:00	Sorry, migh have brought the site down
769	2020-01-16 09:35:55+00:00	Starship orbital vehicle SN1, liquid oxygen he...
770	2020-01-15 17:09:01+00:00	Just saw this today. Tesla refunds in general ...
771	2020-01-14 20:20:57+00:00	Very true. What's really mindblowing is how mu...
772	2020-01-14 20:14:59+00:00	Advancing humanity's understanding of the Univ...
773	2020-01-14 20:08:58+00:00	One person's MRI machine is another's railgun!...
774	2020-01-14 20:04:32+00:00	Exactly. We've had good discussions with leadi...
775	2020-01-14 18:23:23+00:00	Great song
776	2020-01-14 17:59:16+00:00	T-shirt is bulletproof & makes u buff! https:/...
777	2020-01-14 02:01:23+00:00	Thanks for mentioning! We should've done this ...
778	2020-01-13 18:47:40+00:00	Good analysis, although a bit conservative imo...

MODELS

Decision Tree

Random Forest

Extra Trees

K Nearest Neighbor

AdaBoost

Logistic Regression

DATASET

1st ATTEMPT

Use of RandomizedSearchCV to randomly find the optimal parameters for each of the models. Ran all models on unseen data.

Summary Table

Model	AUC Score	Accuracy	Precision	Recall	F1 Score
KNN:	0.5498824152893246	0.5254582484725051	0.4883720930232558	0.5550660792951542	0.5195876288659794
RF:	0.5071792918419397	0.5437881873727087	0.5057471264367817	0.5814977973568282	0.5409836065573772
ADA:	0.5559603855453612	0.5356415478615071	0.4978902953586498	0.5198237885462555	0.5086206896551724
DTREE:	0.4734357921234805	0.4786150712830957	0.4403292181069959	0.4713656387665198	0.45531914893617026
ETREE:	0.5217448908615151	0.5641547861507128	0.5261044176706827	0.5770925110132159	0.5504201680672268
LR:	0.5547265741446126	0.5356415478615071	0.4979757085020243	0.5418502202643172	0.5189873417721519

2nd Attempt: IMPROVEMENT ON EARLIER ATTEMPT

+ Added 5, 10, 20 , 50 day moving averages
+ Added + np.log (current close/previous close) + shift effect
+ Volatility variable (std deviation of 21 day MA) + shift effect

MODELLING

	RF	Dtree	Etree	Logistic Regression	AdaBoost	KNN
Accuracy	**0.789699571**	0.78111588	0.78111588	0.76824034	0.763948498	0.611587983
Precision	0.755458515	0.746724891	0.74248927	0.71314741	0.741935484	0.572649573
Sensitivity	0.623255814	0.623255814	0.623255814	0.83255814	0.623255814	0.623255814
Specificity	0.77689243	0.768924303	0.760956175	0.71314741	0.77689243	0.601593625
Recall	0.804651163	0.795348837	0.804651163	0.83255814	0.748837209	0.623255814
F1 Score	**0.779279279**	0.77027027	0.772321429	0.76824034	0.74537037	0.59688196

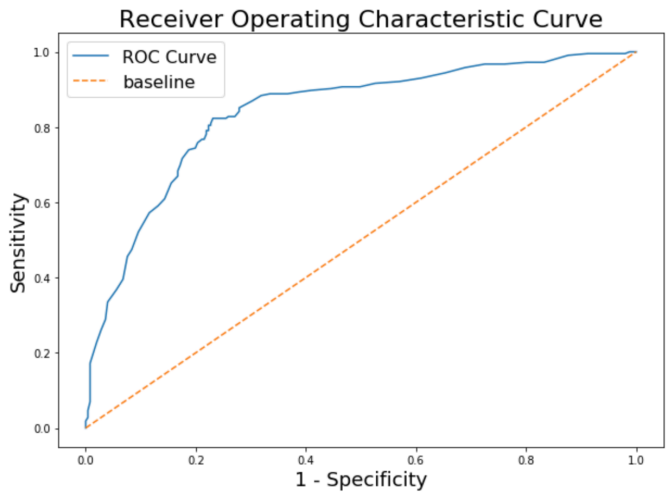
MODEL EVALUATION

1

Baseline: 51% probability of stock closing up on any given day

1.0	0.508463
0.0	0.491537

AUC Score for RF: 0.790771796534791



2

3

	RF
Accuracy	**0.789699571**
Precision	0.755458515
Sensitiity	0.623255814
Specificity	0.77689243
Recall	0.804651163
F1 Score	**0.779279279**

CONCLUSION

LIMITATION

- Financial market forecasting is one of the most difficult practical applications and especially given market volatility. An even better score may have been attained by employing a more sophisticated model and deeper learning approach. One consideration could have been the use of LSTM, RNN and ARIMA which could have been combined in a Feedforward Neural Network to give a more accurate prediction. An approach of combining different methods through Ensemble learning could also have been adopted.

ADVANTAGES

- Combines both endogenous and exogenous approach which is more holistic.
- Scoring was reasonable although improvements could certainly be made.