MONEY TALKS

Money makes the world go round.

Presented by: Marcus Barnes, Hidaya Ismail, Faria Khandaker, Asel Kushkeyeva, Ngone Lo, & Shabrina J. Mardevi

01 INTRODUCTION

Project context

02 DATASET

A brief description of the datasets used.

03 RESEARCH Q'S

The research questions.

04 RESULTS

Our results

05 DISCUSSION

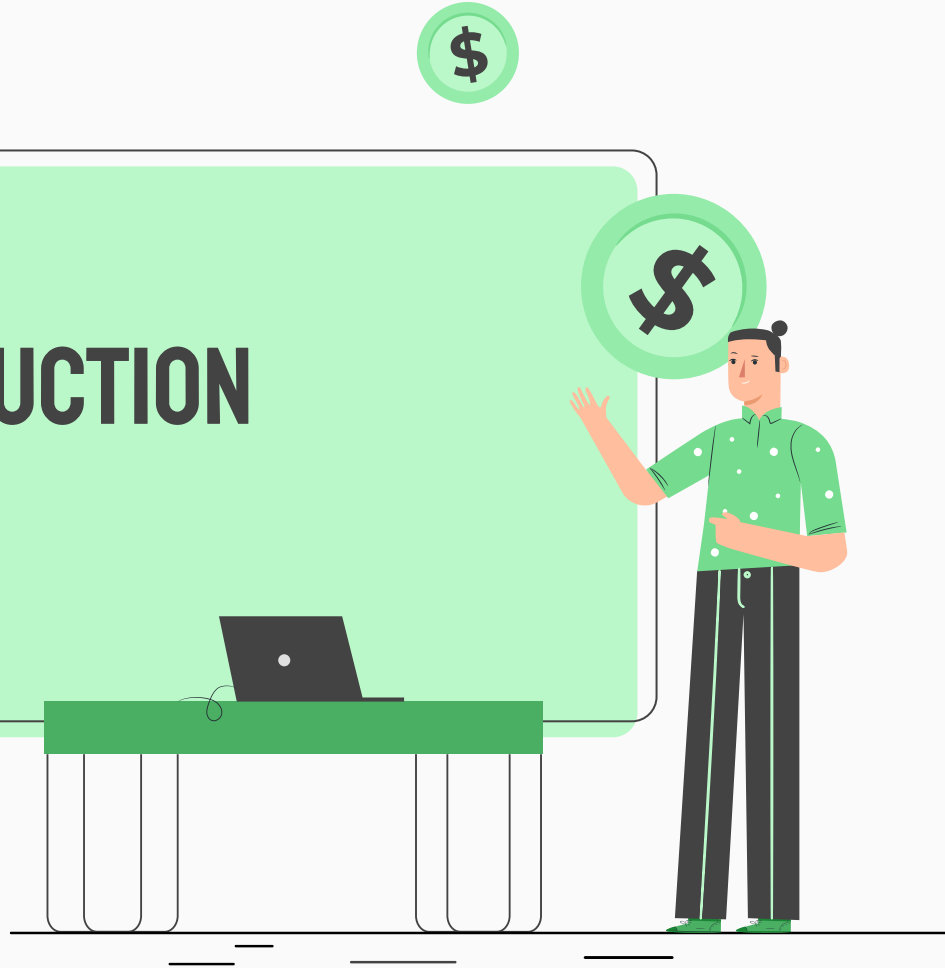
Reflections on our findings and the limitations of our research.

06 CONCLUSION

Concluding summary

Speaker: Marcus

INTRODUCTION



Speaker: Marcus

INTRODUCTION

- Stock Price prediction is hot topic in finance
- Curious to see how well traditional ML models perform since much of the literature uses Neural Networks
- We were interested in two main things
  - Whether link between politics and finance could be sensed by ML algorithms
  - Can we find a “best time” to invest in stocks
- Based on the datasets we found, we focused mainly on US stocks and the pharmaceutical industry

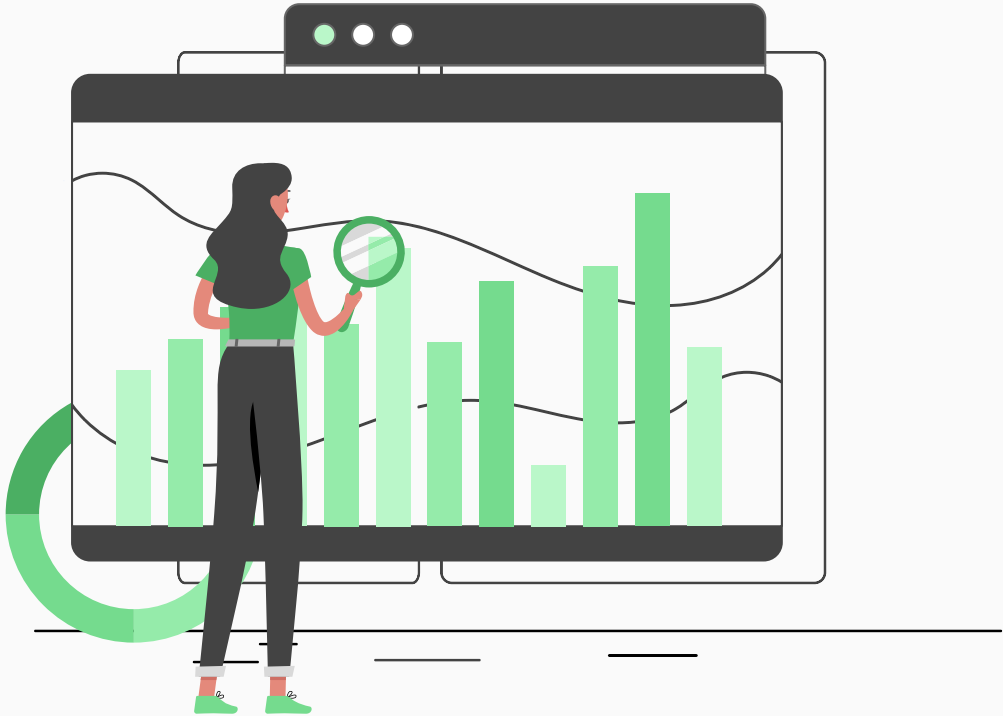
Speaker: Marcus

DATASET DESCRIPTION

1. Pharma sales data: Six years data (2014-2019) on sales of drugs classified in 8 ATC categories
   - 2,016 tuples & 13 variables

2. Stock Market Dataset: Historical daily prices of all stocks and ETFs.
   - Subset to 7 biggest pharma from 2014 to 2018
   - 8,805 tuples & 8 variables

3. 200+ Financial Indicators of US Stocks (2014-2018)
   - A repository; 3,800 to 5,000 tuples each dataset & 224 variables



Speaker: Sasja

Given today's stock and pharmaceutical sales data, what is the best stock value to buy tomorrow?

QUESTION ONE

When is the best time to generally invest? Buy when the stock value hits the predicted low best value!

Speaker: Sasja

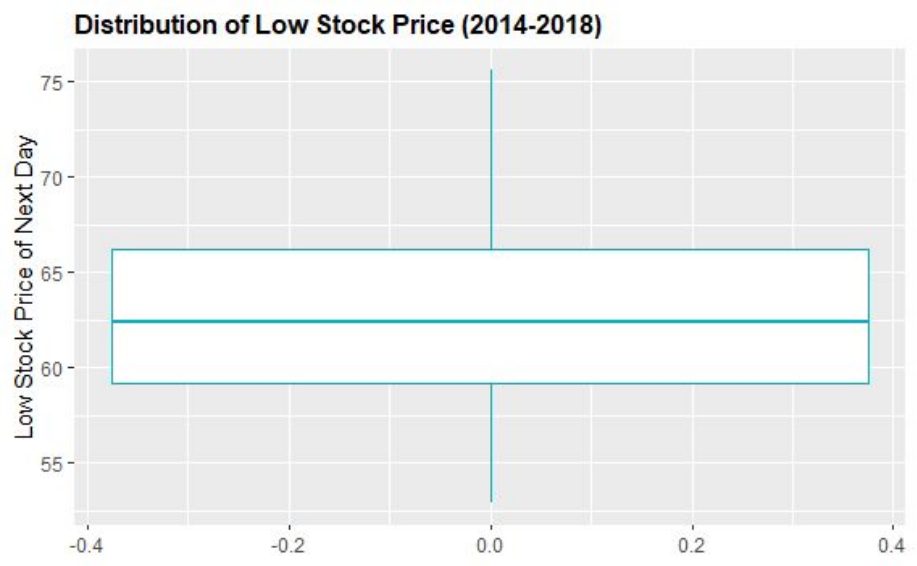
DATA PREPARATION & CLEANING

- Merged stock market and pharmaceutical datasets.
- Created new variable of low_price_next_dat
- Removal of highly correlated numeric variables.
- Normalization of numeric independent variables.
- Backward elimination and split training-test sets
  - Training: 2014-2017 data; testing: 2018 data

Speaker: Sasja

EDA OF TARGET VARIABLE

- Low Stock Price of Next Day: shift up of the variable low stock price by one day.
- Skewness: 0.37



The target variable is fairly symmetrical/normally distributed.

Speaker: Ngone

RESULTS: MODELS TESTED

- Cross validation: 5 folds, 3 repeats
- Evaluation Metrics:
  - RMSE
  - % with less than 25% error (Pred25%)
  - % with less than 10% error (Pred10%)
  - % with less than 5% error (Pred5%)
  - % with less than 1% error (Pred1%)

LINEAR REGRESSION

KNN REGRESSION

RANDOM FOREST
NTREES=425

SVM

ENSEMBLE MODEL
RF&LR

SIMPLE MOVING AVERAGE

AUTO-ARIMA

AVERAGE OF PREVIOUS 5 DAYS

Models	RMSE	Pred25%	Pred10%	Pred5%	Pred1%
Linear Regression	1.36	100.0%	100.0%	99.6%	34.8%
KNN Regression	2.01	100.0%	100.0%	94.4%	24.4%
Random Forest	1.68	100.0%	100.0%	97.2%	22.4%
SVM	1.38	100.0%	100.0%	98.8%	33.6%
Ensemble Model	1.70	100.0%	100.0%	97.6%	28.4%
Simple Moving Average	3.69	100.0%	97.2%	60.4%	16.4%
Auto-Arima	2.76	100.0%	100.0%	79.6%	13.6%
Average of Previous 5 Days	0.89	100.0%	99.9%	99.6%	60.2%

Speaker: Ngone

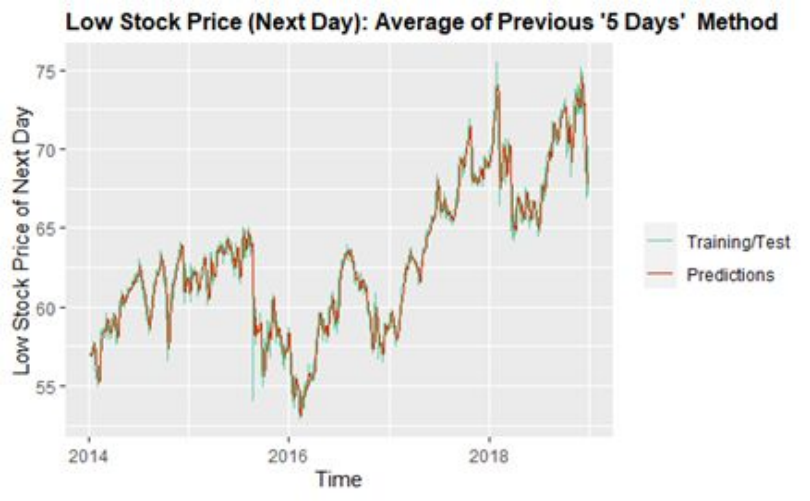
RESULTS: TOP PERFORMING MODEL

RMSE	Pred25%	Pred10%	Pred5%	Pred1%
1.36	100.0%	100.0%	99.6%	34.8%

LINEAR REGRESSION

RMSE	Pred25%	Pred10%	Pred5%	Pred1%
0.89	100.0%	99.9%	99.6%	60.2%

AVERAGE OF PREVIOUS 5 DAYS

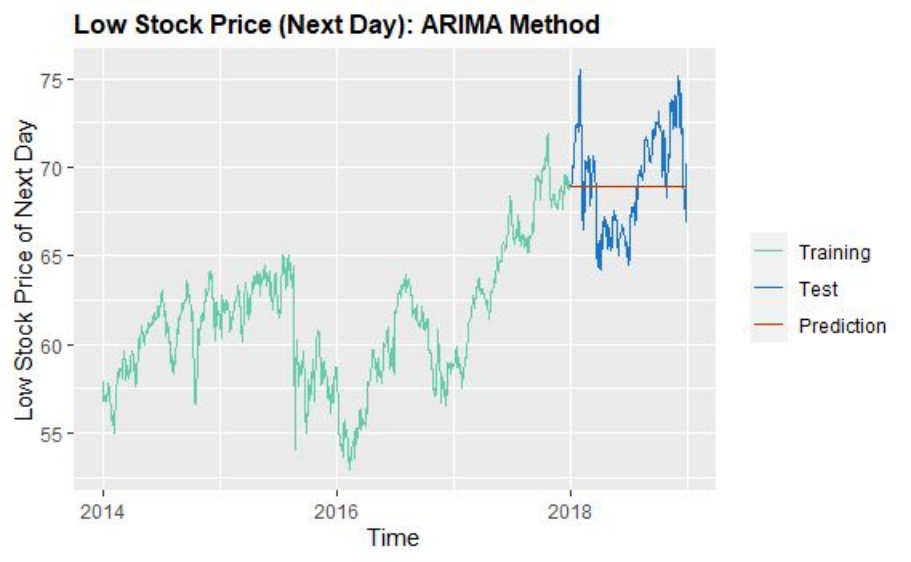


Speaker: Ngone

RESULTS: WORST PERFORMING MODEL

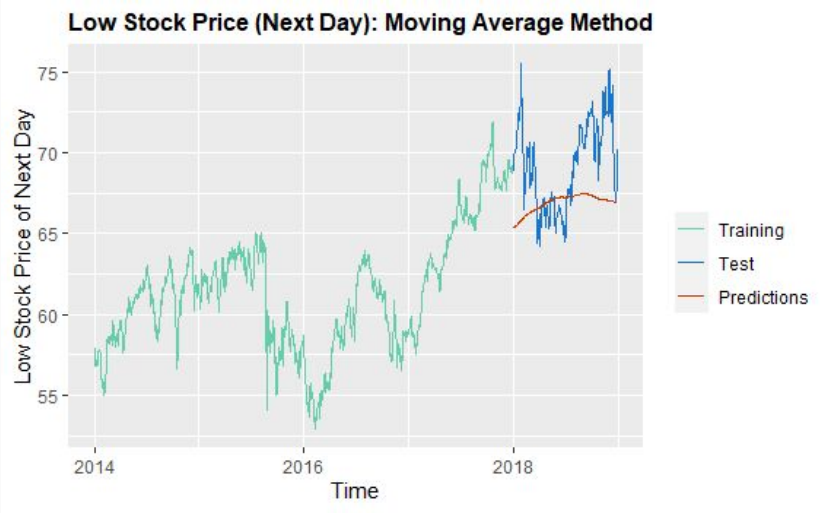
RMSE	Pred25%	Pred10%	Pred5%	Pred1%
2.76	100.0%	100.0%	79.6%	13.6%

AUTO-ARIMA



RMSE	Pred25%	Pred10%	Pred5%	Pred1%
3.69	100.0%	97.2%	60.4%	16.4%

SIMPLE MOVING AVERAGE



Speaker: Ngone

Are there changes in the types of drugs sold in the years after Trump was elected? Additionally, are there general changes in the overall financial standing of pharmaceutical industry between 2014 and 2018?

QUESTION TWO

Speaker: Asel

DATA PREPARATION & CLEANING

- Normalized Pharma Sales Data and changed medicines codes into specific groups and names
- For 200+ Financial Indicators: removed correlated predictors; replaced missing values with mean values; normalized the values; subsetted to pharmaceutical companies

Speaker: Asel

Results: Model Testing

Are there general changes in the overall financial standing of pharmaceutical industry between 2014 and 2018?

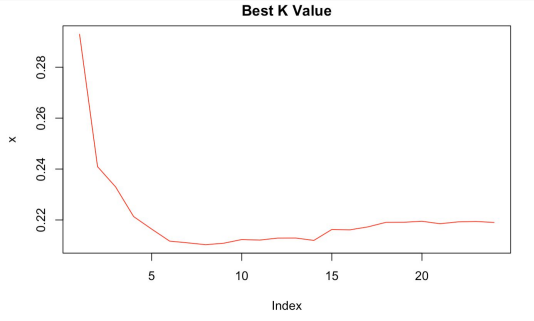
Regression Models

MULTIPLE LINEAR REGRESSION

RMSE on 'Obama' data, 70:30 split: 0.1642

RMSE on 'Obama'/'Trump' data, 50:50 split: 0.1697

KNN MODEL



RMSE on 'Obama' data split 70:30: 0.2103

RMSE on 'Obama'/'Trump' data, 50:50 split: 0.2057

Speaker: Asel

Results: Classification Models

Are there general changes in the overall financial standing of pharmaceutical industry between 2014-2018

ENSEMBLE MODEL

An ensemble of Random Forest, Support Vector Machines and Logistic Regression Models were trained on 2014/2015 stock data to classify increase or decrease in stock price. Data from 2017/2018 stock prices were tested against the model. The model performed moderately with an accuracy of 51%

```
Confusion Matrix and Statistics

          Reference
Prediction dec inc
       dec 137  67
       inc  21  15

               Accuracy : 0.6333
                 95% CI : (0.5689, 0.6944)
    No Information Rate : 0.6583
    P-Value [Acc > NIR] : 0.8123

                  Kappa : 0.0578

 Mcnemar's Test P-Value : 1.61e-06

            Sensitivity : 0.1829
            Specificity : 0.8671
         Pos Pred Value : 0.4167
         Neg Pred Value : 0.6716
              Precision : 0.4167
                 Recall : 0.1829
                     F1 : 0.2542
             Prevalence : 0.3417
         Detection Rate : 0.0625
   Detection Prevalence : 0.1500
      Balanced Accuracy : 0.5250

       'Positive' Class : inc
```

Classification metrics Obama 70/30 split

```
Confusion Matrix and Statistics

          Reference
Prediction dec inc
       dec 330 208
       inc 253 155

               Accuracy : 0.5127
                 95% CI : (0.4803, 0.545)
    No Information Rate : 0.6163
    P-Value [Acc > NIR] : 1.00000

                  Kappa : -0.0068

 Mcnemar's Test P-Value : 0.04043

            Sensitivity : 0.4270
            Specificity : 0.5660
         Pos Pred Value : 0.3799
         Neg Pred Value : 0.6134
              Precision : 0.3799
                 Recall : 0.4270
                     F1 : 0.4021
             Prevalence : 0.3837
         Detection Rate : 0.1638
   Detection Prevalence : 0.4313
      Balanced Accuracy : 0.4965

       'Positive' Class : inc
```

Classification metrics Obama trained model with Trump test data

Speaker: Faria

Results: Regression Model

Are there changes in the types of drugs sold in the years after Trump was elected?

LINEAR REGRESSION

```
[1] "2014-2015 model for Rheumatoid Arthritis Meds"
[1] "RMSE: 0.207611965685788"
[1] "PRED(10): 0.88"
[1] "PRED(25): 1"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.237   3.287   3.379   3.427   3.575   3.673
```

```
[1] "TrumpTest: Rheumatoid Athritis Med"
[1] "RMSE: 0.644350094321413"
[1] "PRED(10): 0.19"
[1] "PRED(25): 0.86"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.922   4.034   4.147   4.147   4.259   4.371
```

```
[1] "2014-2015 model for Aspirin"
[1] "RMSE: 0.23488553992087"
[1] "PRED(10): 0.88"
[1] "PRED(25): 1"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.428   3.431   3.437   3.439   3.448   3.454
```

```
[1] "TrumpTest: Aspirin"
[1] "RMSE: 0.383906036858621"
[1] "PRED(10): 0.6"
[1] "PRED(25): 0.91"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.350   3.361   3.372   3.372   3.383   3.393
```

```
[1] "2014-2015 model for Ibuprofen"
[1] "RMSE: 0.294812540818012"
[1] "PRED(10): 0.88"
[1] "PRED(25): 1"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.034   5.093   5.198   5.253   5.423   5.536
```

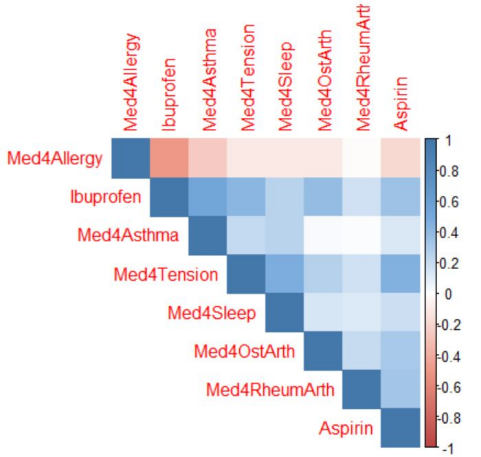
```
[1] "TrumpTest: Ibuprofen"
[1] "RMSE: 0.809521412261535"
[1] "PRED(10): 0.26"
[1] "PRED(25): 0.91"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.727   5.841   5.954   5.954   6.067   6.180
```

```
[1] "2014-2015 model for Sleep Meds"
[1] "RMSE: 0.534654841947156"
[1] "PRED(10): 0.22"
[1] "PRED(25): 0.66"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.065   1.212   1.506   1.434   1.643   1.720
```

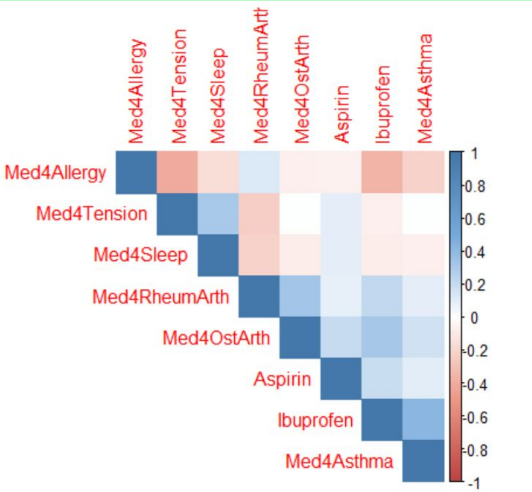
```
[1] "TrumpTest: Sleep Medicine"
[1] "RMSE: 1.25650813866716"
[1] "PRED(10): NA"
[1] "PRED(25): 0.02"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04665 0.22182 0.39700 0.39700 0.57218 0.74735
```

```
[1] "2014-2015 model for Allergy Meds"
[1] "RMSE: 0.485796867272014"
[1] "PRED(10): 0.34"
[1] "PRED(25): 0.88"
[1] "Summary of Prediction"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.624   2.649   2.694   2.718   2.791   2.840
```

```
[1] "TrumpTest: Allergy Medicine"
[1] "RMSE: 0.554797612893333"
[1] "PRED(10): 0.39"
[1] "PRED(25): 0.81"
[1] "Summary of Predictions"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.955   2.998   3.041   3.041   3.085   3.128
```



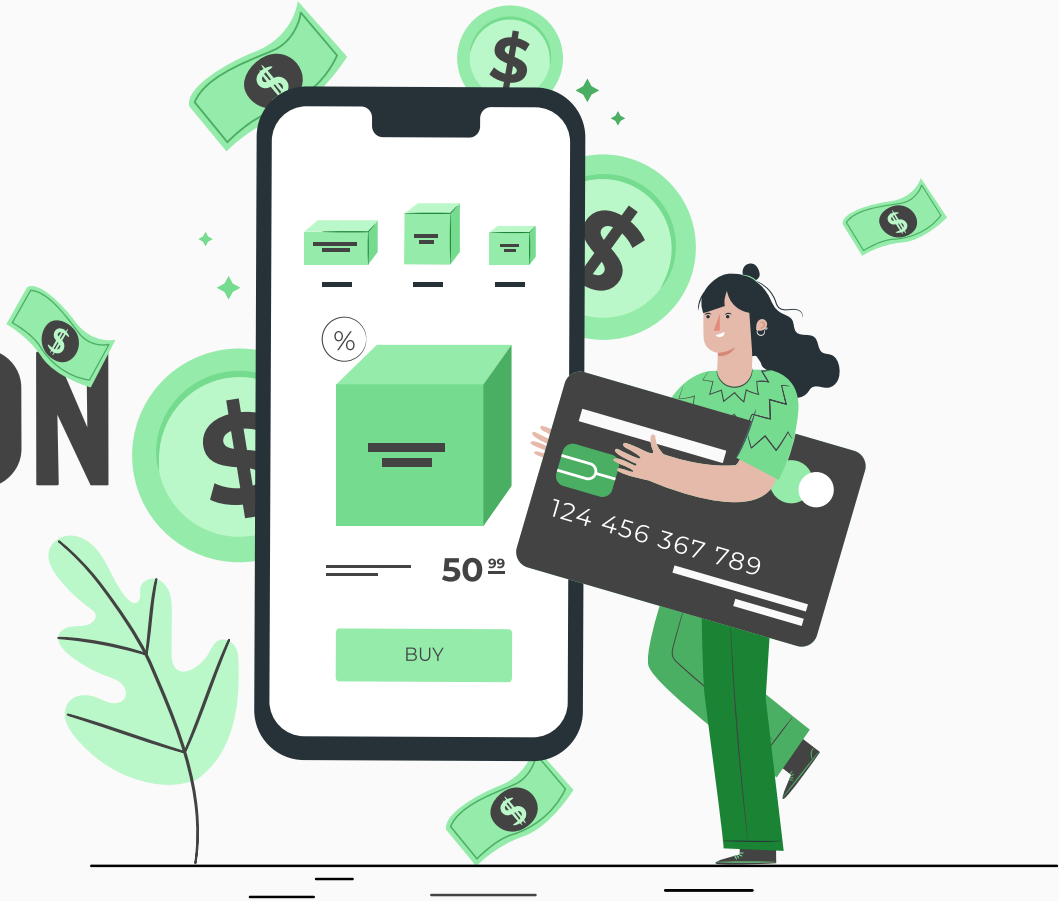
Trump era



Obama era

Speaker: Faria

DISCUSSION



Speaker: Hidaya

DISCUSSION POINTS

01
RESEARCH QUESTION ONE

Best model is Linear regression.

02
RESEARCH QUESTION TWO

Multi-linear Regression & Linear Regression

03
LIMITATIONS

Time-series data was not processed. This could have explained low scores. Knowledge gap issue. Unique terminology used in the stock market that required additional research.

04
ETHICS

Pharma companies may use manipulative techniques that influences the economic market. These companies also have a long history with cheap labour, human and animal testing.

Speaker: Hidaya

CONCLUSION

Speaker: Hidaya

CONCLUSIONS

A

It is not enough to compare financial indicators or which political party is in power to predict stock prices.

B

Regression models are best performing for stock data.

C

For future work, we would like to investigate Artificial Neural Networks and Long-Short Memory RNNs

Speaker: Hidaya

THANKS FOR LISTENING

Speaker: Hidaya