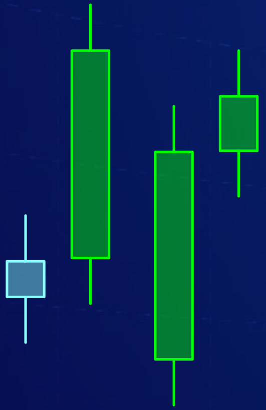
Stock Market Prediction & Web Scraping Analysis

Jatin Batta, Ramone Andrade, Kelly Tsai, Samuel Edwards,
Manaar Salama, Eduardo Hinojosa

Table of Contents



01

Introduction to the Project

Research Questions

1. Can we fit various machine learning models to perform stock predictions on prices?

2. How will these models be affected by sentiment analysis? Specifically, by adding sentiment analysis, will we be able to achieve more accurate predictions?

Stakeholders

Hedge Funds

Morgan Stanley
GUGGENHEIM
citi
EVERCORE
BANK OF AMERICA
Goldman Sachs
LAZARD
CENTERVIEW PARTNERS
J.P.Morgan
Rothschild & Co
Career Principles

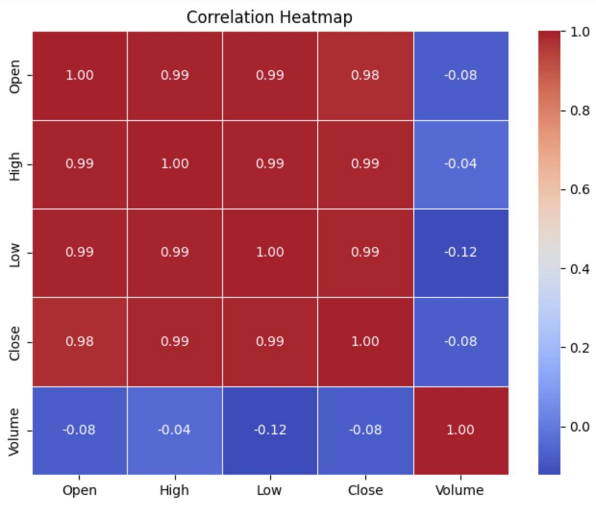
Investment Firms

02

Our Data

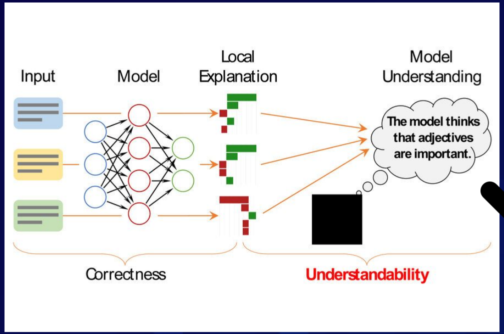
Data, Cleaning, Feature Engineering

Yahoo Finance: AAPL

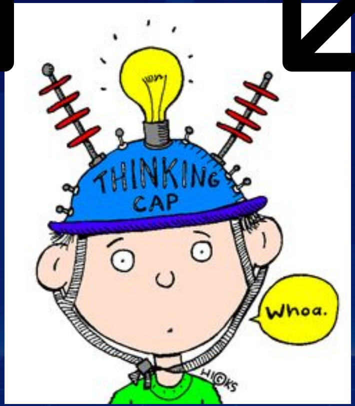


Features

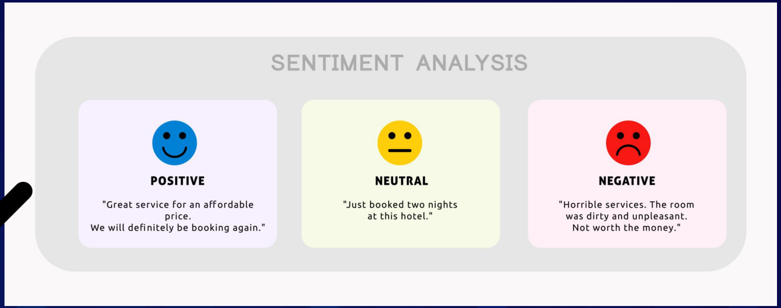
Experiments



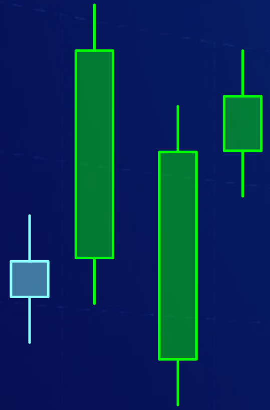
ML Models



Prediction



Sentiment Analysis



03

Methods

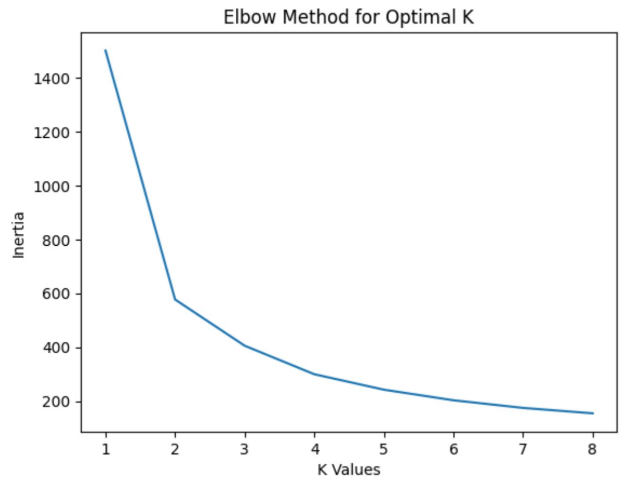
Methods

1. PCA
2. Linear Regression
3. XGBoost
4. GLMs
5. Neural Networks (particularly LSTMs and RNNs)
6. Transformers and NLP (for sentiment analysis)

Note: Prior to all of this we ran a principal component analysis, which we will discuss as our first step.

PCA

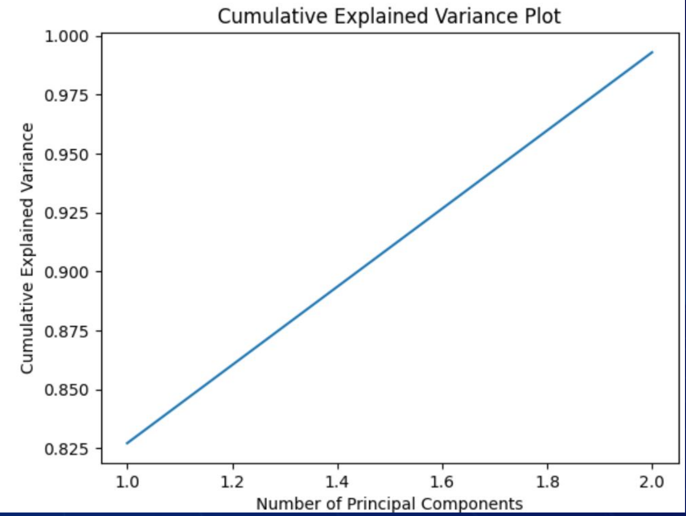
```
1: 1500.0000000000002
2: 577.1170505241802
3: 405.69909754794463
4: 300.02488096429425
5: 242.76113073559333
6: 203.4772465021124
7: 175.14410688205078
8: 155.16037787760254
```



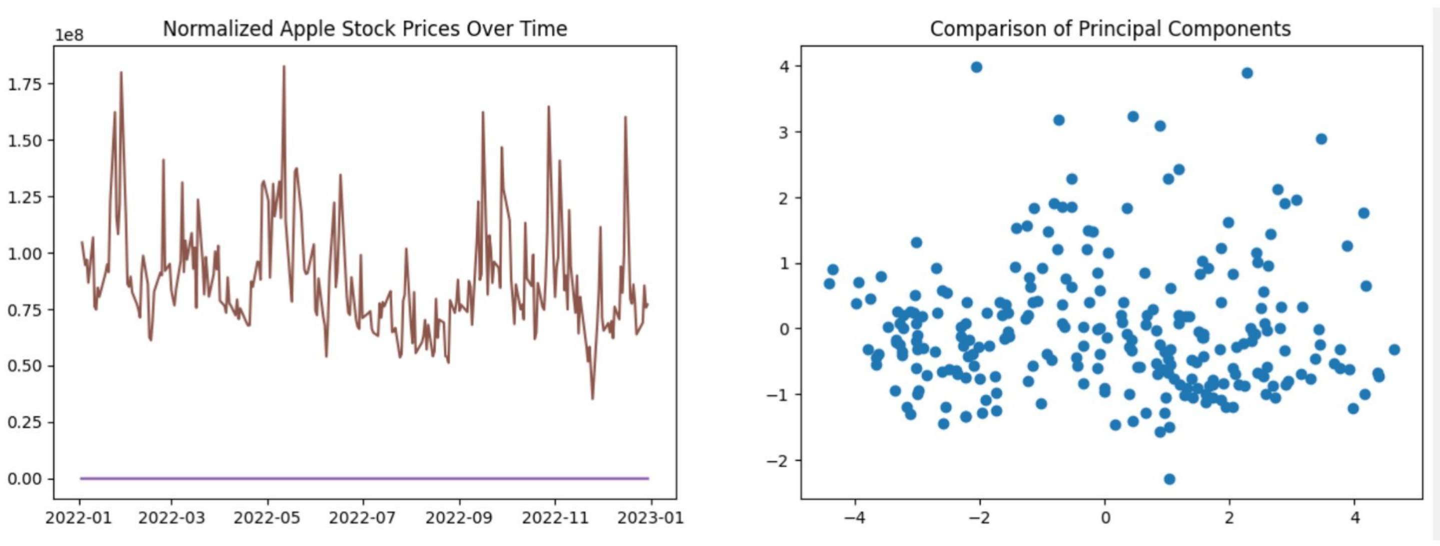
Elbow Method

```
2: 0.5139806559946883
3: 0.4144772076915389
4: 0.4230831963738426
5: 0.4002375679256944
6: 0.38984205558518775
7: 0.35254208291891487
8: 0.34182953044441144
The best choice for k is:2
```

Silhouette Score + Cumulative Explained Variance Plot



PCA



PCA Models

PCA

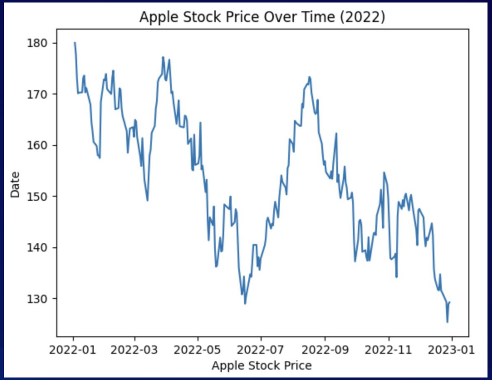
```
1  print("Principal Components:\n", pca.components_)
```

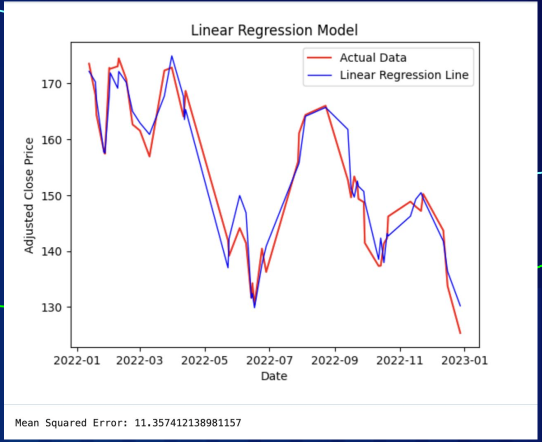
```
Principal Components:
 [[-0.44484027 -0.44720319 -0.44804396 -0.44682494 -0.44679314  0.0458957 ]
 [ 0.02233369  0.05848245 -0.02166819  0.02464018  0.01876218  0.99732259]]
```

- For the first principal component:
  - All features had negative influence on it except for last one, Volume
- For the second principal component:
  - Only the third feature, Low, had negative influence
  - Volume had the highest magnitude => strongest influence on the principal component
- **Conclusion: Volume** would be the best feature to use for the predictions

Linear Regression

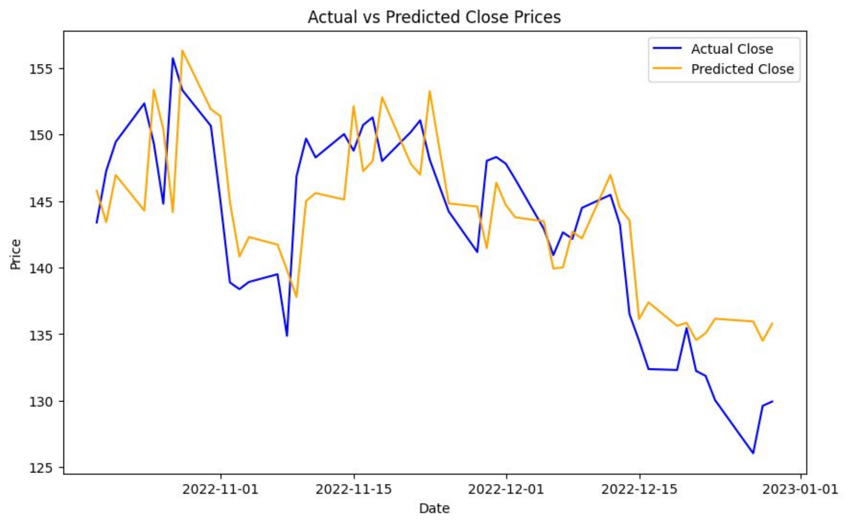
- Using Previous Adjusted Close to predict the Current Adjusted Close
  - MSE: 11.35
- More so this was because we were using the AAPL stock





XGBoost

- Trained XGBoost model on one year of Apple Stock data
- Target variable: Closing Stock Price
- MSE: 20.356



GLMs – Introduction and Motivation

- Prediction method most of our group learned this semester in Data 102.
- Essentially, just a more generalized form of linear regression, but allows for more flexibility with the distributions our data takes.
  - Predicting a continuous unknown variable as a linear combination of observed variables.
  - Using statsmodels package in Python to create these models.

Generalized Linear Models

Regression	Inverse link function	Likelihood
Linear	identity	Gaussian
Logistic	sigmoid	Bernoulli
Poisson	exponential	Poisson
Negative binomial	exponential	Negative binomial

GLMs

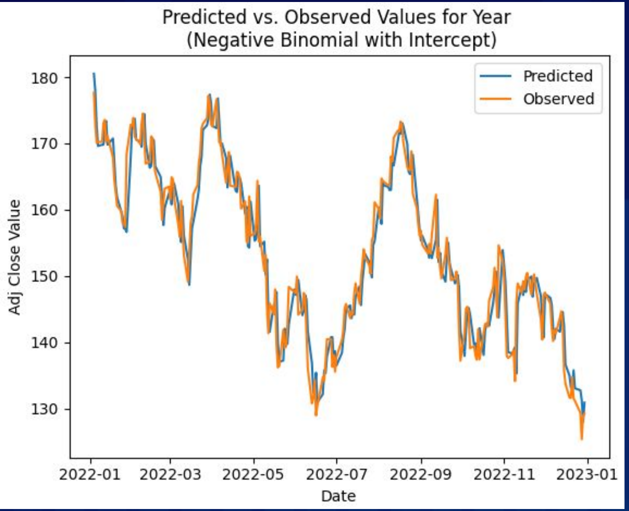
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Adj Close   No. Observations:                  250
Model:                            GLM   Df Residuals:                      248
Model Family:        NegativeBinomial   Df Model:                            1
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1508.3
Date:                Tue, 05 Dec 2023   Deviance:                      0.12398
Time:                        02:39:13   Pearson chi2:                    0.124
No. Iterations:                     3   Pseudo R-squ. (CS):           0.006299
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 4.0625      0.771      5.267      0.000       2.551       5.574
Previous Adj Close    0.0063      0.005      1.259      0.208      -0.004       0.016
=====================================================================================
```

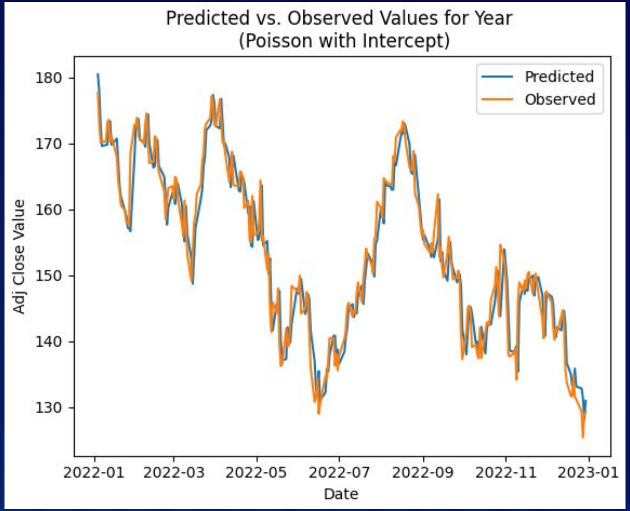
Negative Binomial MSE: 11.365

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Adj Close   No. Observations:                  250
Model:                            GLM   Df Residuals:                      248
Model Family:                 Poisson   Df Model:                            1
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -867.98
Date:                Tue, 05 Dec 2023   Deviance:                       18.776
Time:                        02:37:58   Pearson chi2:                     18.8
No. Iterations:                     3   Pseudo R-squ. (CS):             0.6219
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 4.0652      0.063     65.014      0.000       3.943       4.188
Previous Adj Close    0.0063      0.000     15.584      0.000       0.005       0.007
=====================================================================================
```

Poisson MSE: 11.362

GLMs



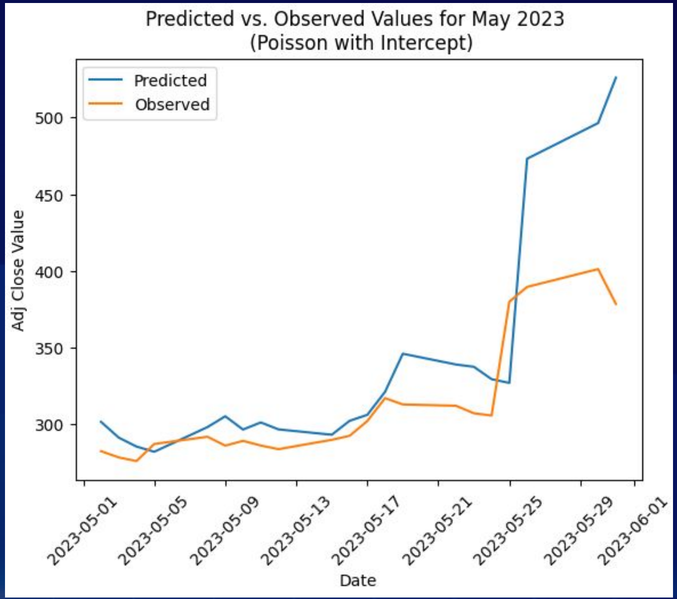


Negative Binomial MSE: 11.365

Poisson MSE: 11.362

Where GLMs Go Wrong

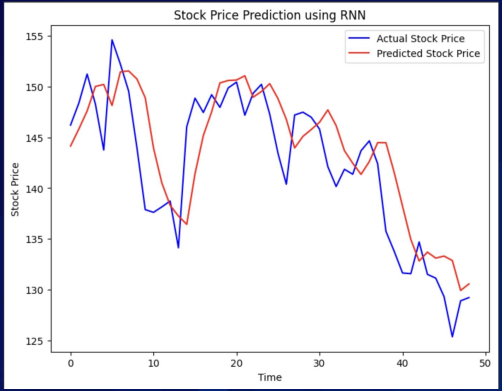
- These models performed well on our Apple stock data, which stayed consistent from 2022 - halfway through 2023.
- However, does not do nearly as well for more volatile stocks:
  - Right is Poisson model trained on NVIDIA's stock data from 2022-2023 tested on May 2023 data (stock went way up due to AI GPU market).
- Not entirely practical for stock prediction – only does well when change is consistent!



Poisson MSE: 2179.252

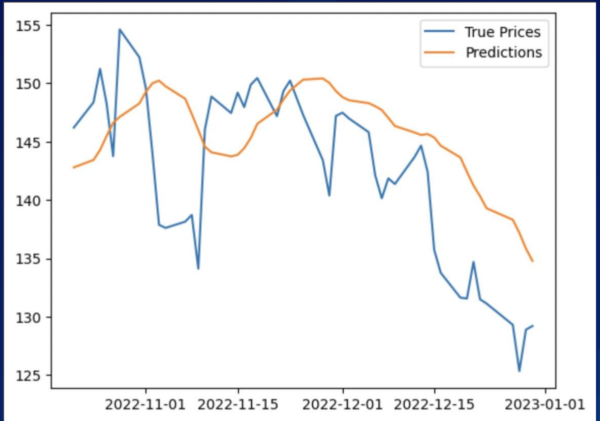
LSTMs, RNNs, and Ensemble Learning

MSE: 18.32



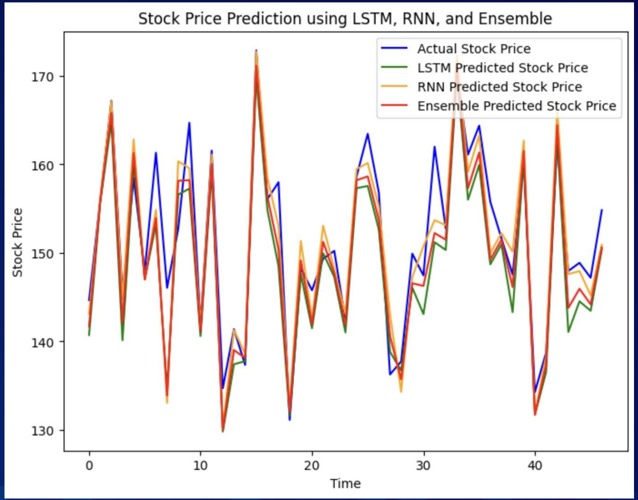
RNN

LSTM



MSE: 41.68

MSE: 15.26

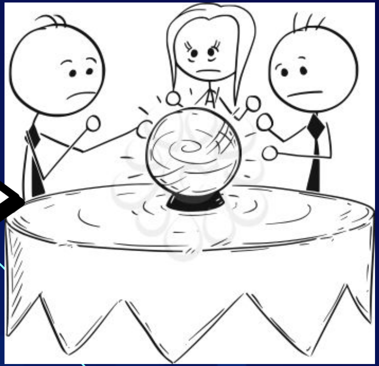


Ensemble Learning

Experiment: Article Analysis Effects on Prediction

Article Parsing

Transformer



Prediction

Article Analysis: Transformers and NLP

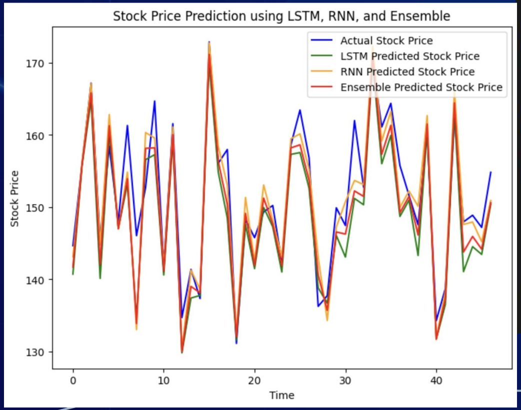
- Used NLTK library to tokenize and analyze sentiment of headlines, and use the BERT transformer to get the sentiment of the articles
- Retrieved sentiment scores and set thresholds for positive, neutral, and negative
  - ≥0.05 for positive, ≤-0.05 for negative, otherwise neutral
- Used new data frame to train models (LSTM, RNN)

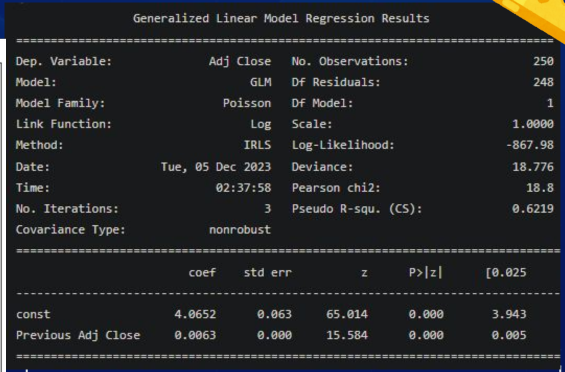
	Headline	Date	Sentimen Label	Sentiment Score
3668	Tech Stocks And FAANGS Strong Again To Start D...	2020-06-10	Positive	0.5574
3669	10 Biggest Price Target Changes For Wednesday	2020-06-10	Neutral	0.0000
3670	Benzinga Pro's Top 5 Stocks To Watch For Wed.,...	2020-06-10	Positive	0.2023
3671	Deutsche Bank Maintains Buy on Apple, Raises P...	2020-06-10	Neutral	0.0000
3672	Apple To Let Users Trade In Their Mac Computer...	2020-06-10	Positive	0.3818

04

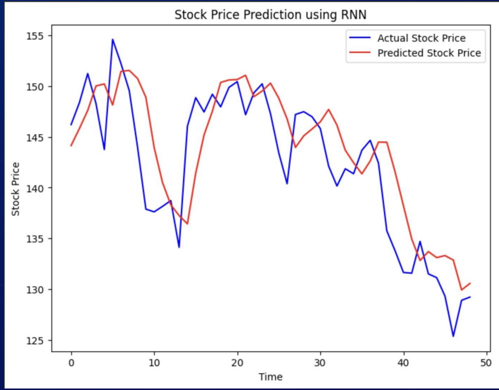
Results & Conclusions

MSE Champs





```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Adj Close   No. Observations:                  250
Model:                            GLM   Df Residuals:                      248
Model Family:                 Poisson   Df Model:                            1
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -867.98
Date:                Tue, 05 Dec 2023   Deviance:                       18.776
Time:                        02:37:58   Pearson chi2:                     18.8
No. Iterations:                     3   Pseudo R-squ. (CS):             0.6219
Covariance Type:            nonrobust
==============================================================================
                        coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------
const                 4.0652      0.063     65.014      0.000       3.943
Previous Adj Close    0.0063      0.000     15.584      0.000       0.005
==============================================================================
```



2

MSE: 15.26

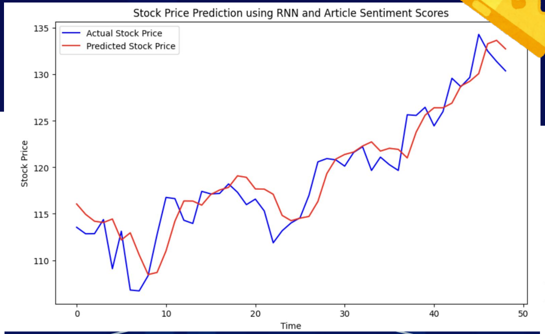
1

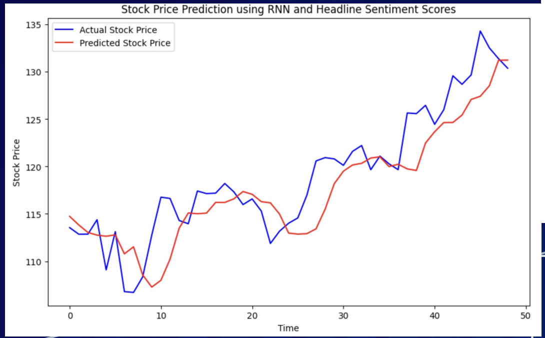
MSE: 11.362

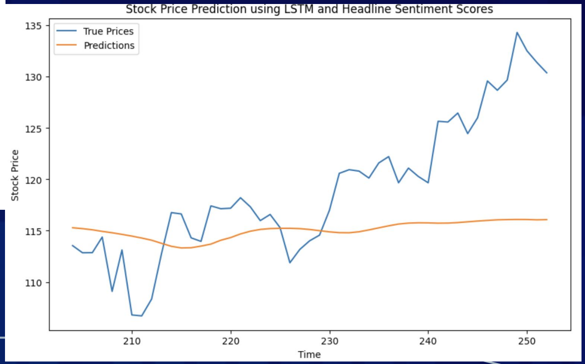
3

MSE: 18.32

Sentiment Analysis Results







2
MSE: 11.41

1
MSE: 6.61

3
MSE: 13049.71

Honorable Mention: LSTM w/ Article Score
MSE: 13052.92

Conclusions & Future Iterations

- Conclusions
  - Minor difference between headline and sampling sentences from articles
  - Semantic analysis improved overall models
  - Probably not ready for real-world usage yet; would need to fine tune incorporation of semantic analysis aspect to increase accuracy first
- If we had more time, or in future iterations, we could have:
  - Integrated more text data and properly integrated dates
  - Tested out additional data and arguments to increase accuracy
  - Account for overfitting with K-fold cross validation

Thank You

Questions?