Apple Stock Price Prediction using NLP and Time Series Models

Group 1:
Yao Zhang - yz4481
Catherine Liu - xl3239
Shuobofang Yang - sy3113
Xianglin Lu - xl3258
Justin Ingardia - ji2278

Source code: https://github.com/yaoyzz/Financial-NLP-Hybrid-Model

Project Overview

03 Natural Language Processing

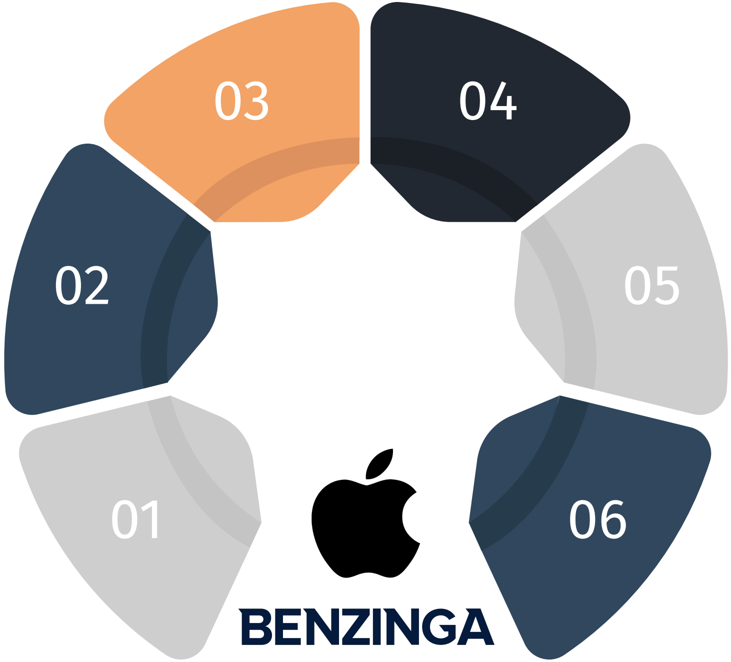
Techniques include
- Term frequency
- Word-level vectorization
- Character-level vectorization
- Sentiment classification

02 Preprocess

- Select relevant fields
- Combine dataset
- Adding features

01 Prepare Data

Sources:
- News data from Benzinga REST API
- Stock trading data from Yahoo Finance



04 Regression Models

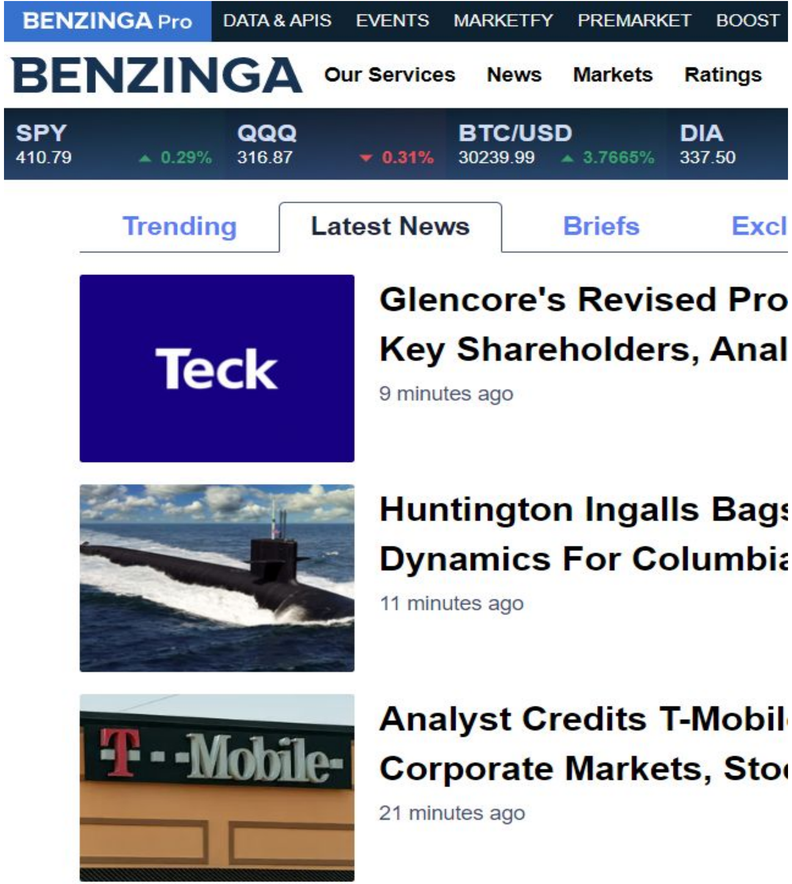
- Train, tune and validate the supervised learning model using text features as predictors and price movement as dependent variable

05 Time Series Model

- Train, tune and validate the models on the dates and the responsive price movement

06 Projection and Inference

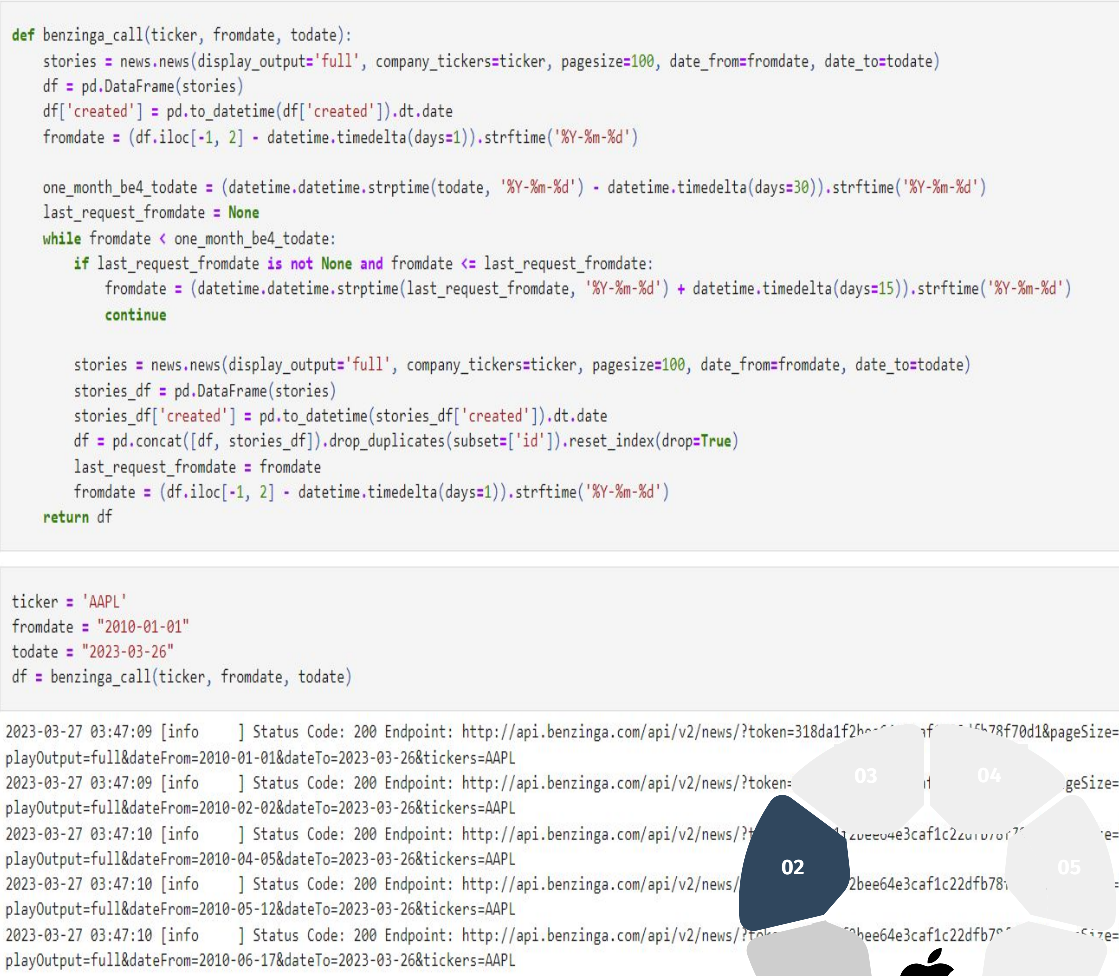
- Plot and compare the predictions created by different models and render the outcomes



Glencore's Revised Pro
Key Shareholders, Anal
9 minutes ago

Huntington Ingalls Bags
Dynamics For Columbia
11 minutes ago

Analyst Credits T-Mobil
Corporate Markets, Stoc
21 minutes ago

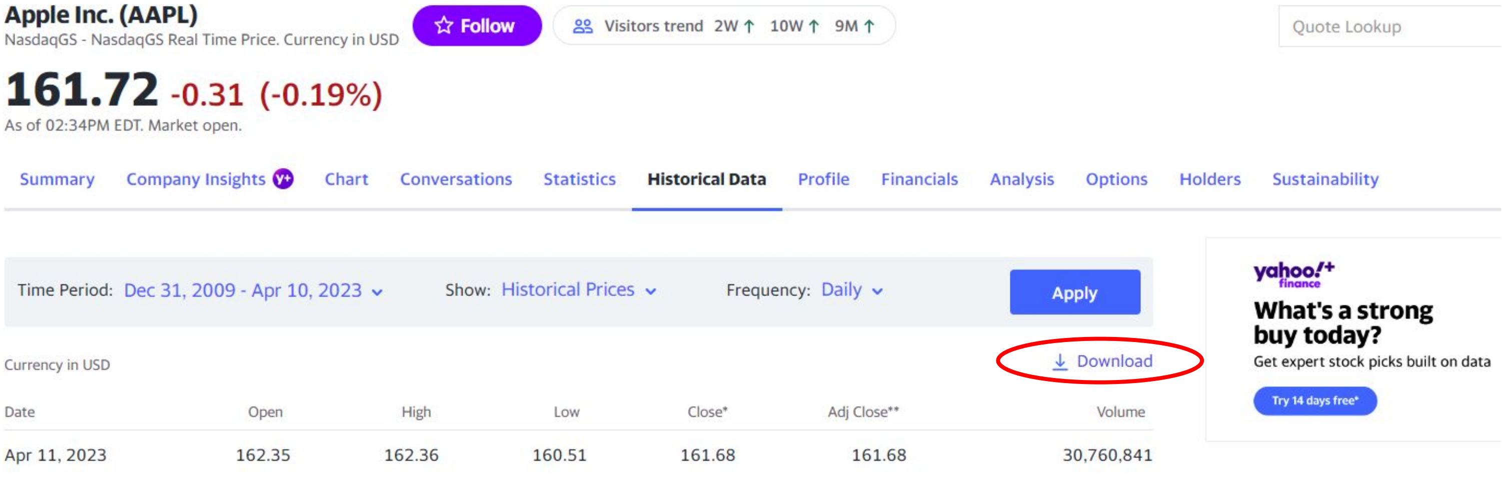


```
def benzinga_call(ticker, fromdate, todate):
    stories = news.news(display_output='full', company_tickers=ticker, pagesize=100, date_from=fromdate, date_to=todate)
    df = pd.DataFrame(stories)
    df['created'] = pd.to_datetime(df['created']).dt.date
    fromdate = (df.iloc[-1, 2] - datetime.timedelta(days=1)).strftime('%Y-%m-%d')

    one_month_be4_todate = (datetime.datetime.strptime(todate, '%Y-%m-%d') - datetime.timedelta(days=30)).strftime('%Y-%m-%d')
    last_request_fromdate = None
    while fromdate < one_month_be4_todate:
        if last_request_fromdate is not None and fromdate <= last_request_fromdate:
            fromdate = (datetime.datetime.strptime(last_request_fromdate, '%Y-%m-%d') + datetime.timedelta(days=15)).strftime('%Y-%m-%d')
            continue

        stories = news.news(display_output='full', company_tickers=ticker, pagesize=100, date_from=fromdate, date_to=todate)
        stories_df = pd.DataFrame(stories)
        stories_df['created'] = pd.to_datetime(stories_df['created']).dt.date
        df = pd.concat([df, stories_df]).drop_duplicates(subset=['id']).reset_index(drop=True)
        last_request_fromdate = fromdate
        fromdate = (df.iloc[-1, 2] - datetime.timedelta(days=1)).strftime('%Y-%m-%d')
    return df
```

```
ticker = 'AAPL'
fromdate = "2010-01-01"
todate = "2023-03-26"
df = benzinga_call(ticker, fromdate, todate)
```

```
2023-03-27 03:47:09 [info      ] Status Code: 200 Endpoint: http://api.benzinga.com/api/v2/news/?token=318da1f2b[illegible]f[illegible]fb78f70d1&pageSize=
playOutput=full&dateFrom=2010-01-01&dateTo=2023-03-26&tickers=AAPL
2023-03-27 03:47:09 [info      ] Status Code: 200 Endpoint: http://api.benzinga.com/api/v2/news/?token=[illegible]geSize=
playOutput=full&dateFrom=2010-02-02&dateTo=2023-03-26&tickers=AAPL
2023-03-27 03:47:10 [info      ] Status Code: 200 Endpoint: http://api.benzinga.com/api/v2/news/?t[illegible]2bee64e3caf1c22dfb78[illegible]e=
playOutput=full&dateFrom=2010-04-05&dateTo=2023-03-26&tickers=AAPL
2023-03-27 03:47:10 [info      ] Status Code: 200 Endpoint: http://api.benzinga.com/api/v2/news/[illegible]2bee64e3caf1c22dfb78[illegible]=
playOutput=full&dateFrom=2010-05-12&dateTo=2023-03-26&tickers=AAPL
2023-03-27 03:47:10 [info      ] Status Code: 200 Endpoint: http://api.benzinga.com/api/v2/news/?to[illegible]2bee64e3caf1c22dfb7[illegible]Size=
playOutput=full&dateFrom=2010-06-17&dateTo=2023-03-26&tickers=AAPL
```

1 & 2 Prepare and Preprocess Data



Apple Inc. (AAPL)
NasdaqGS - NasdaqGS Real Time Price. Currency in USD

☆ Follow | Visitors trend 2W ↑ 10W ↑ 9M ↑ | Quote Lookup

161.72 -0.31 (-0.19%)
As of 02:34PM EDT. Market open.

Summary | Company Insights | Chart | Conversations | Statistics | **Historical Data** | Profile | Financials | Analysis | Options | Holders | Sustainability

Time Period: Dec 31, 2009 - Apr 10, 2023 ⌄ Show: Historical Prices ⌄ Frequency: Daily ⌄ Apply

Currency in USD

Download

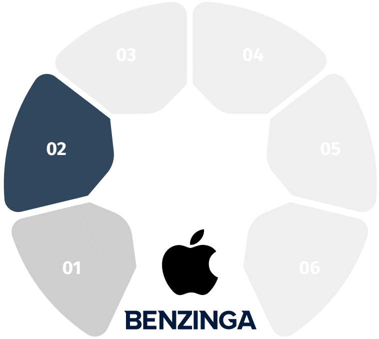
Date	Open	High	Low	Close*	Adj Close**	Volume
Apr 11, 2023	162.35	162.36	160.51	161.68	161.68	30,760,841

1. Select **features**
   ➢ close, volume, date, body, title

2. Join the tables by date

3. Impute missing variables

4. Generate **dependent variables**
   ➢ 1 - 30 days Future price change



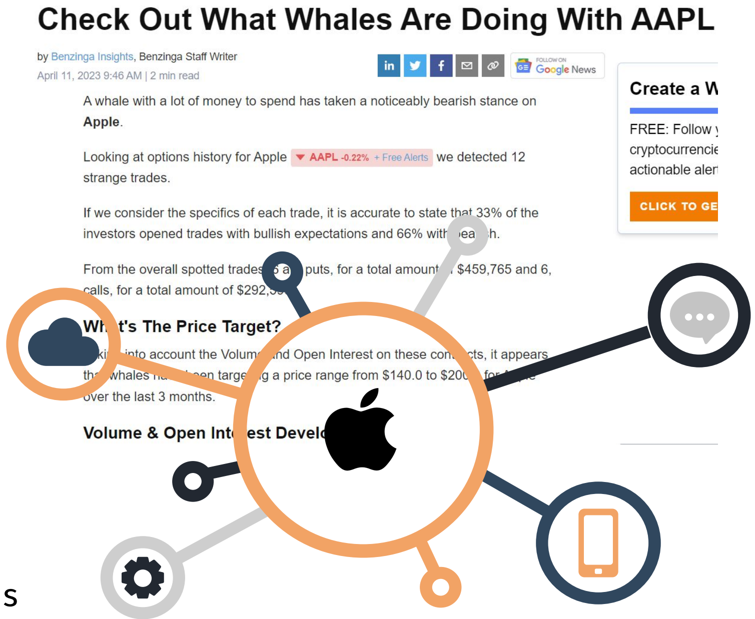
1 & 2 Prepare and Preprocess Data

1. TF-IDF

- ➢ Get TF for each word for each article.
- ➢ Get IDF for each word for each article.
- ➢ Get the sum of TF x IDF for each article.

2. Word Embeddings

- ➢ Created a vocabulary of words
- ➢ Trained word embeddings using **GloVe** (unsupervised learning algorithm)
- ➢ Calculated mean word embeddings for each article
- ➢ Calculated the max, sum, and mean of the embeddings (vectors)



Check Out What Whales Are Doing With AAPL

by Benzinga Insights, Benzinga Staff Writer
April 11, 2023 9:46 AM | 2 min read

A whale with a lot of money to spend has taken a noticeably bearish stance on **Apple**.

Looking at options history for Apple ▼ AAPL -0.22% + Free Alerts we detected 12 strange trades.

If we consider the specifics of each trade, it is accurate to state that 33% of the investors opened trades with bullish expectations and 66% with ea h.

From the overall spotted trades 6 a puts, for a total amount $459,765 and 6, calls, for a total amount of $292,

Wh t's The Price Target?

ki into account the Volum nd Open Interest on these con cts, it appears th whales h en targe g a price range from $140.0 to $200 for A over the last 3 months.

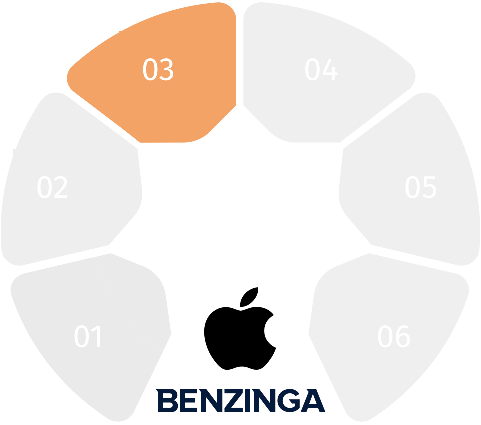
Volume & Open Int est Devel

3. Character-level Embeddings

- ➢ Converted text data to sequences
- ➢ Padded sequences to a fixed length
- ➢ Created an embedding layer that maps the integer-encoded characters to dense vectors.

4. Sentiment Inference

- ➢ Calculated sentiment scores using different lexicons
- ➢ Calculated **AFINN, Bing, and NRC** sentiment scores per article
- ➢ Replaced NA values with 0 in the sentiment scores data frame.



03 Natural Language Processing

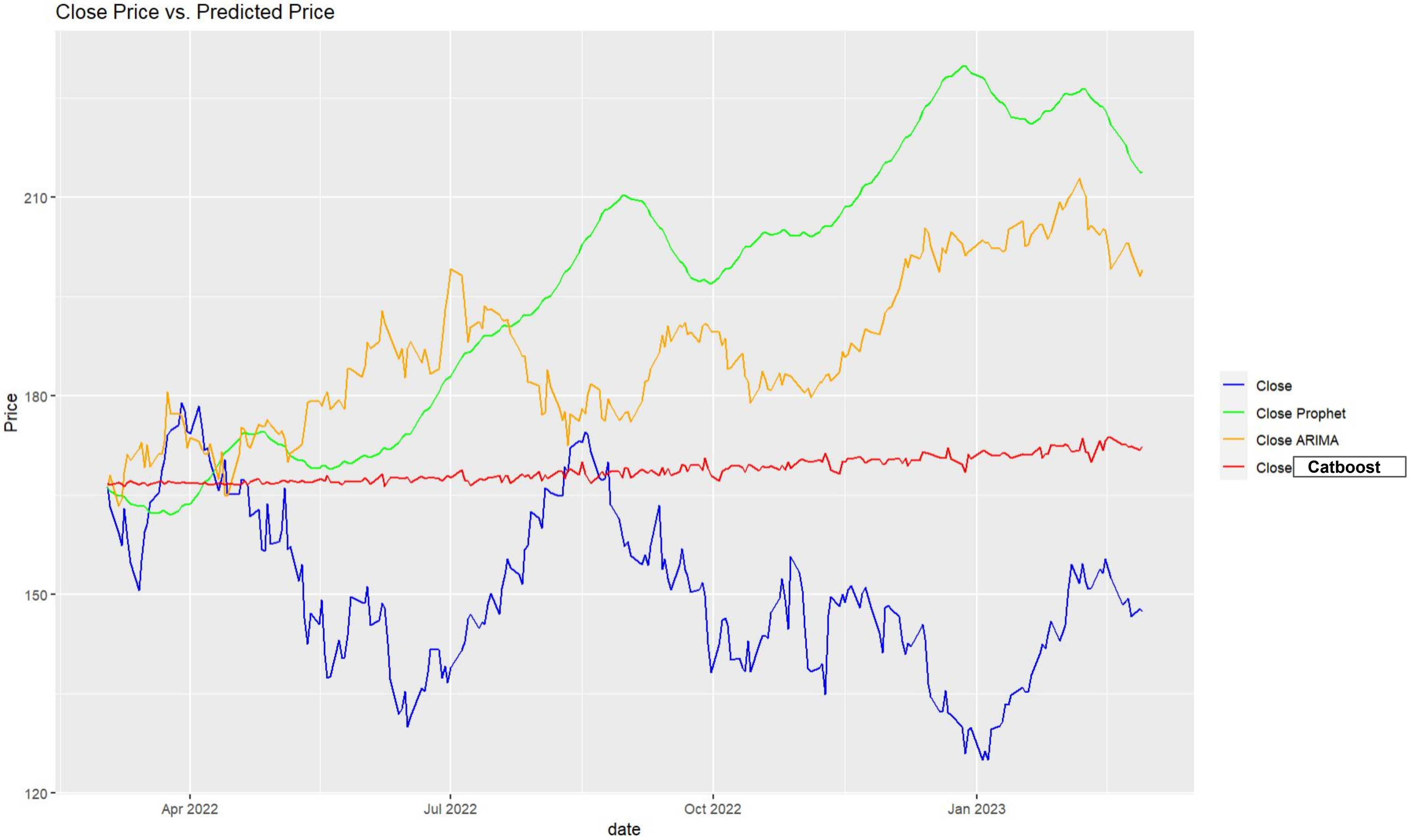
Price prediction

Regression Models

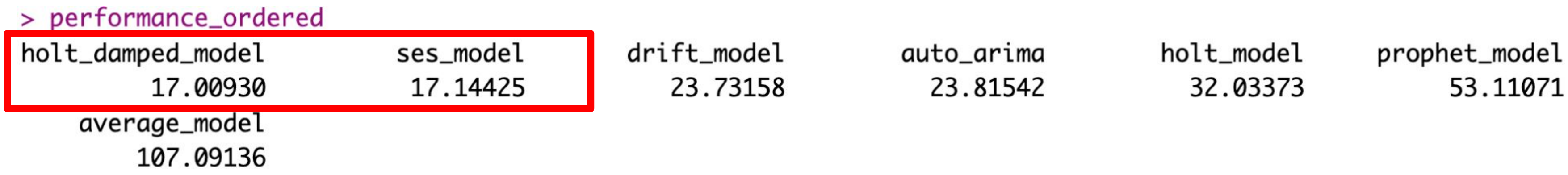
- Train the **Linear Regression** and **CatBoost** model
- Use dependent variables including future change in one day to 30 days

Time Series Models

- Auto Arima
- Prophet
- Average
- Exponential smoothing
- Holt Model
- Random Walk (Drift Model)

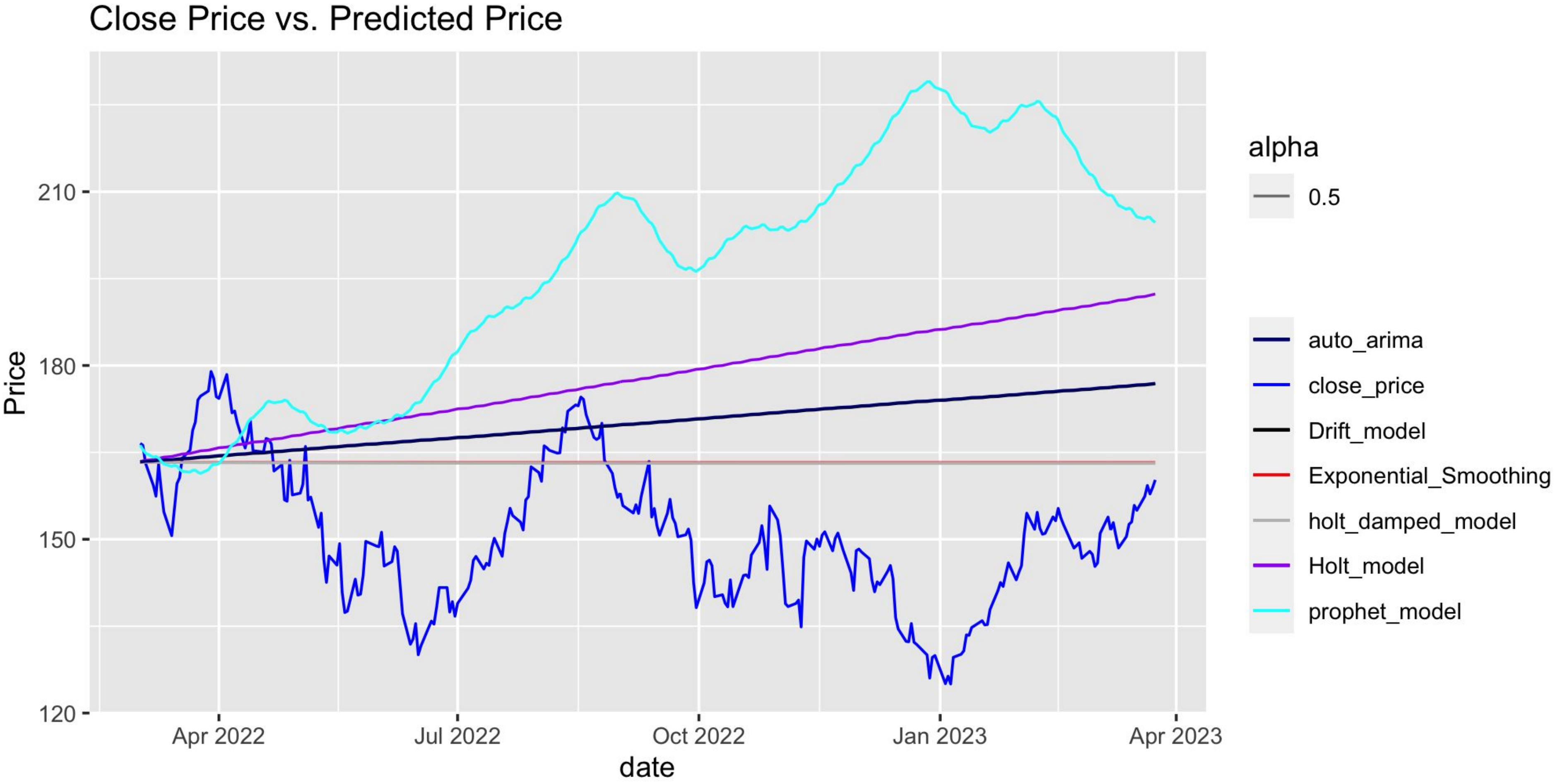

Close Price vs. Predicted Price
Price
210
180
150
120
Apr 2022
Jul 2022
Oct 2022
Jan 2023
date
Close
Close Prophet
Close ARIMA
Close Catboost

Model Performance



```
> performance_ordered
holt_damped_model         ses_model       drift_model        auto_arima        holt_model     prophet_model
         17.00930          17.14425          23.73158          23.81542          32.03373          53.11071
    average_model
        107.09136
```

All indicates that **Holt’s damped model** and **exponential smoothing model** are the best.


Close Price vs. Predicted Price
Price
210
180
150
120
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023
date
alpha
0.5
auto_arima
close_price
Drift_model
Exponential_Smoothing
holt_damped_model
Holt_model
prophet_model

Use Cases

➢ Use Time Series models to predict mid-term and long-term trends and generate profits.

➢ Use NLP supervised models for short-term directional prediction in volatile market to generate profits

Future Development

➢ Add more technical indicators as features such as moving average, MACD, RSI and KDJ to boost the supervised learning model.

➢ Automate the ETL process and deploy the models on stocks or other securities in various sectors to compare the performance under different market conditions.

THANK YOU