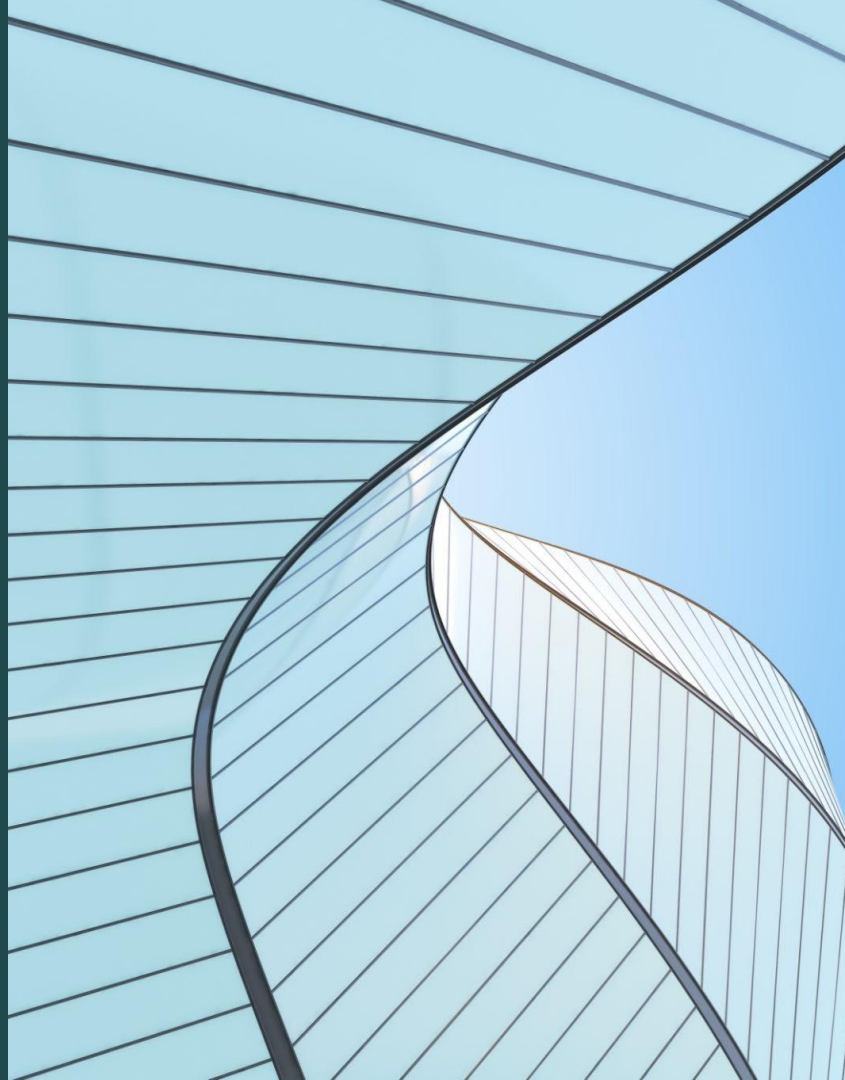
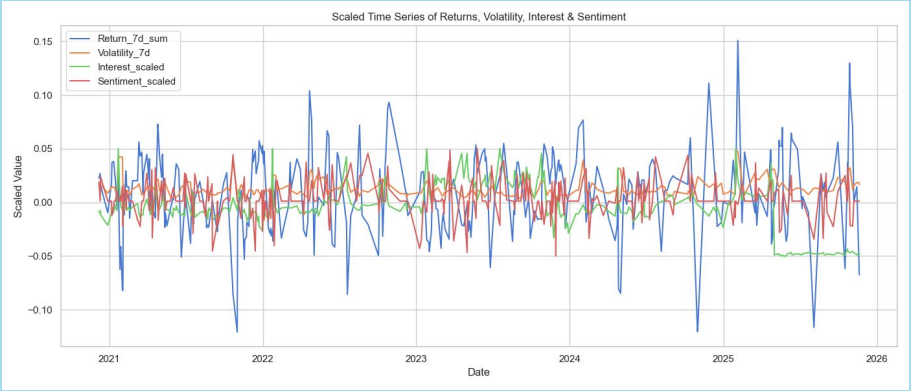
11/25/2025

# Modeling IBM Stock Movements with Public Interest and Sentiments

Principles of Programming: Final Project Presentation
DSCI 510 – Fall 2025 – Dr. Alexey Tregubov
University of Southern California
Madeleine Willson



# Project *Overview*




Goal: Predict daily stock returns

Motivation: Understand Market Behavior

Scope: Why IBM? Why 5 years?

Modeling: Binary Classification

Proxy for Interest: Search Volume

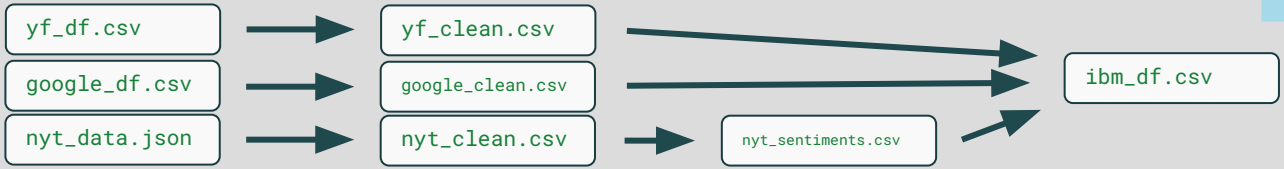
Proxy for Sentiment: Article Headlines

# Data Sources

| | DESCRIPTION | KEY FIELDS | RAW DATA SIZE | ACCESS METHOD |
|---|---|---|---|---|
| **Yahoo Finance** | Historical IBM Stock Prices | • Open<br>• Close<br>• Volume<br>• Return | ~1250 rows | • yfinance library<br>• Gives DataFrame |
| **Google Trends** | Search interest for “IBM” keyword | • Date<br>• Interest | ~1825 rows | • pytrends library<br>• Gives DataFrame |
| **NYT Articles** | Headlines mentioning “IBM” keyword | • Date<br>• Abstract<br>• Headline<br>• URL | ~600 articles | • NYT Article Search API<br>• Gives JSON |

# Summary of *Results*

## Data Cleaning & Integration




- Aligned on a common date range, cleaned missing values
- Created lagged & rolling features for returns, volatility, interest, sentiment
- Sentiment scored using VADER on NYT headlines
- Ended with 14 features for 359 days

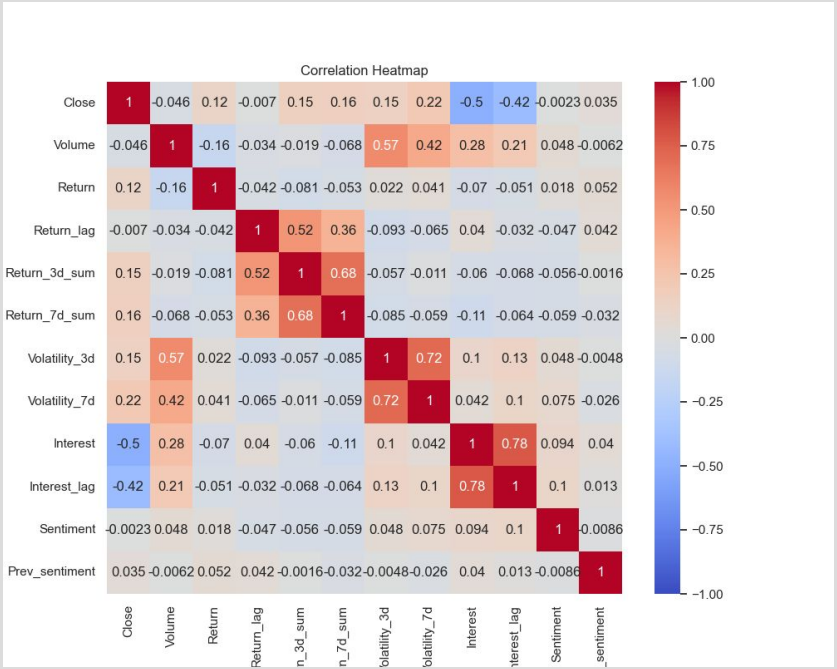
```
1  Date,Close,Volume,Return,Return_lag,Return_3d_sum
2  2020-11-23,114.8087921142578,5910318,,,,,,,Monday
3  2020-11-24,118.94837188720705,8109115,0.036056295
```

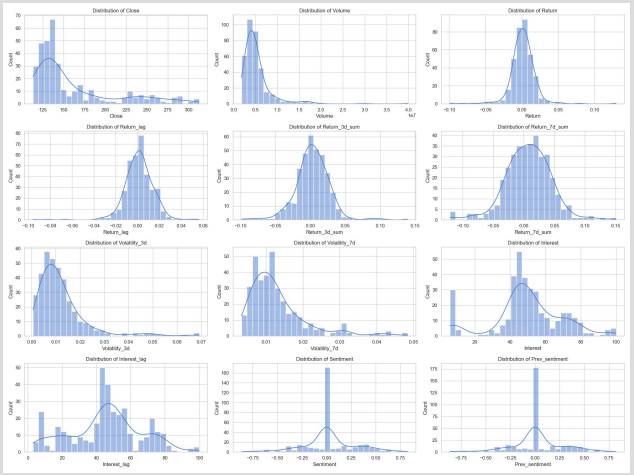
```
Date,Interest,Interest_lag
2020-11-23,44,
2020-11-24,51,44.0
2020-11-25,42,51.0
```

```
1  Date,Headline
2  2025-11-04,IBM to Cut Thousands of Workers Amid A.
3  2025-04-28,IBM Plans to Invest $150 Billion Domest
```

# Summary of *Results*

## Exploratory Data Analysis

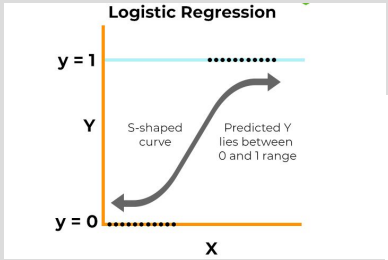


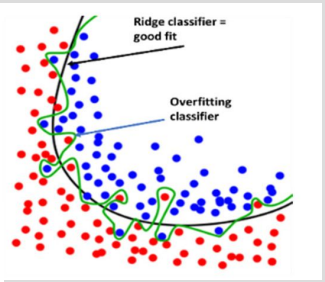



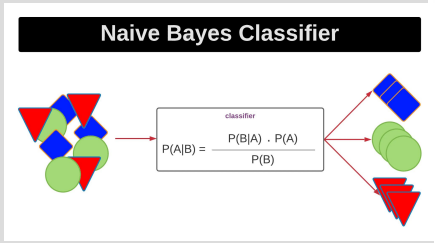

- Mean return (~0.001) positive
- Mean sentiment (~0.03) positive
- Mean return (~50.1) positive
- No missing values
- Distribution and correlation heatmaps to explore relationships among variables
- Day of the week & time series trends explored

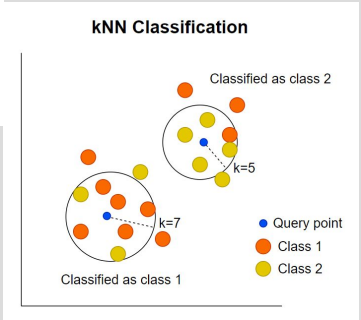
# Summary of *Results*

## Modeling Approach




### Binary classification

80% train, 20% test

### Class Imbalance

Handled via weighted training & extra .5%

### Feature Selection

Lagged return, rolling returns(3&7), rolling volatility (3&7), lagged interest, previous sentiment

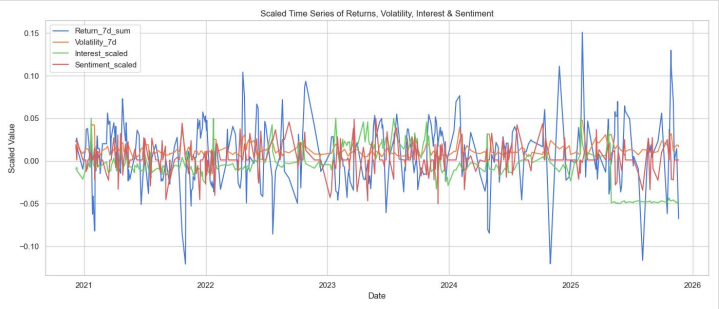
### Models Tested

Logistic Regression, Ridge Classifier, Naive Bayes, KNN

# Summary of *Results*

## Model Performance

- Evaluated with f1 score (minority class) & accuracy
- Logistic Regression and Ridge performed best
- Reasonable predictive power despite very noisy data

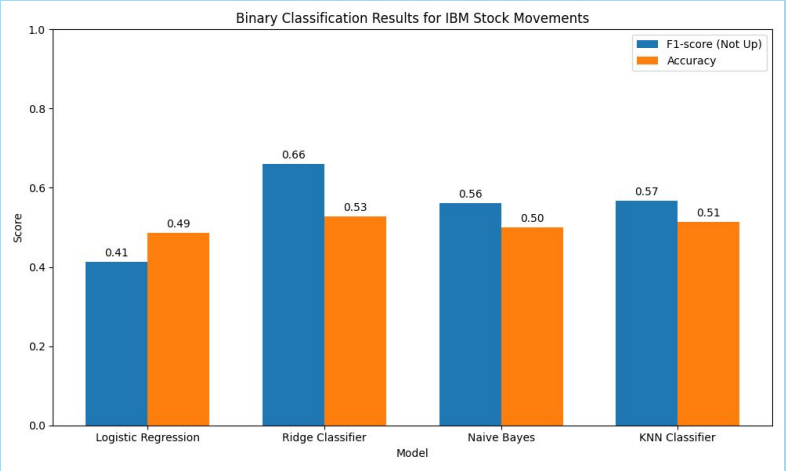



Logistic Regression:
f1-score (Stock not up): 0.4127
Preformance Metrics:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.4815 | 0.3611 | 0.4127 | 36 |
| 1 | 0.4889 | 0.6111 | 0.5432 | 36 |
| accuracy | | | 0.4861 | 72 |
| macro avg | 0.4852 | 0.4861 | 0.4780 | 72 |
| weighted avg | 0.4852 | 0.4861 | 0.4780 | 72 |

Naive Bayes:
f1-score (Stock not up): 0.5610
Preformance Metrics:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.5000 | 0.6389 | 0.5610 | 36 |
| 1 | 0.5000 | 0.3611 | 0.4194 | 36 |
| accuracy | | | 0.5000 | 72 |
| macro avg | 0.5000 | 0.5000 | 0.4902 | 72 |
| weighted avg | 0.5000 | 0.5000 | 0.4902 | 72 |




Ridge Classifier:
f1-score (Stock not up): 0.6600
Preformance Metrics:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.5156 | 0.9167 | 0.6600 | 36 |
| 1 | 0.6250 | 0.1389 | 0.2273 | 36 |
| accuracy | | | 0.5278 | 72 |
| macro avg | 0.5703 | 0.5278 | 0.4436 | 72 |
| weighted avg | 0.5703 | 0.5278 | 0.4436 | 72 |

KNN Classifier:
f1-score (Stock not up): 0.5679
Preformance Metrics:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.5111 | 0.6389 | 0.5679 | 36 |
| 1 | 0.5185 | 0.3889 | 0.4444 | 36 |
| accuracy | | | 0.5139 | 72 |
| macro avg | 0.5148 | 0.5139 | 0.5062 | 72 |
| weighted avg | 0.5148 | 0.5139 | 0.5062 | 72 |

# *Challenges*

First time API user, and overly optimistic about my proxies.

**01**

API limits & date window restrictions complicated data collection

**02**

Aligning datasets on datetime variables took a lot of time to resolve reactively

**03**

Handling noisy, small datasets to prevent overfitting

**04**

Experimenting with feature selection and model tuning to improve performance

# Thank you

Questions?