# Predictive Modeling in Finance: A Study of Weekly Stock Performance Using Machine Learning

**Chinmay Chandra**[1]

[1]Seidenberg School of CS, Pace University, New York City, USA

**Abstract:** This research endeavors to explore the relationship between financial market indicators and stock price movements using logistic regression and other modeling techniques. The primary objective is to assess the impact of lagged returns and trading volumes on predicting the direction of stock price movements. The study utilizes a dataset comprising historical stock data, including daily returns, trading volumes, and directional indicators. The analysis begins with data preprocessing, including encoding the directional variable, handling missing values, and converting relevant predictors to numeric types. Subsequently, Logistic Regression, k Nearest Neighbors and Naïve Bayes models are employed to investigate the significance of lagged returns and trading volumes in predicting the direction of stock price changes.

## INTRODUCTION

In the fast-paced realm of financial markets, the ability to predict stock price movements remains a critical pursuit for investors and analysts alike. This research aims to contribute to this ongoing dialogue by leveraging advanced statistical modeling techniques to examine the relationship between financial market indicators and stock price changes.

Our focus is to evaluate the predictive power of lagged returns and trading volumes in determining the direction of stock price movements. Utilizing a rich dataset encompassing historical stock data, including daily returns, and trading volumes, we employ logistic regression models as well as alternative classification algorithms like k-Nearest Neighbors (kNN) and Gaussian Naive Bayes.

This study not only seeks to unravel patterns within financial data but also to compare the efficacy of various modeling approaches. Our preliminary findings suggest promising results with logistic regression models and offer insights into the relative strengths of alternative strategies.

Beyond academic curiosity, our research holds practical implications for traders, investors, and financial analysts striving to enhance predictive models and make informed decisions in the dynamic financial landscape. The subsequent sections elaborate on our methodology,

present results, and engage in a discussion that contextualizes our findings within both academic and practical frameworks.

## UNDERSTANDING THE DATA

Our exploration begins with a meticulous examination of the dataset, encompassing historical stock data, lagged values, returns, trading volumes, and directional indicators. Through careful preprocessing, including encoding of directional variables and addressing missing values, we ensure the dataset's readiness for sophisticated modeling techniques. This foundational step sets the stage for a comprehensive analysis of the interplay between financial market indicators and stock price movements.

### Numerical Summary

| | Year | Lag1 | Lag2 | Lag3 | Lag4 | Lag5 | Volume | Today |
|---|---|---|---|---|---|---|---|---|
| mean | 2000.048669 | 0.150585 | 0.151079 | 0.147205 | 0.145818 | 0.139893 | 1.574618 | 0.149899 |
| std | 6.033182 | 2.357013 | 2.357254 | 2.360502 | 2.360279 | 2.361285 | 1.686636 | 2.356927 |
| min | 1990.000000 | -18.195000 | -18.195000 | -18.195000 | -18.195000 | -18.195000 | 0.087465 | -18.195000 |
| 25% | 1995.000000 | -1.154000 | -1.154000 | -1.158000 | -1.158000 | -1.166000 | 0.332022 | -1.154000 |
| 50% | 2000.000000 | 0.241000 | 0.241000 | 0.241000 | 0.238000 | 0.234000 | 1.002680 | 0.241000 |
| 75% | 2005.000000 | 1.405000 | 1.409000 | 1.409000 | 1.409000 | 1.405000 | 2.053727 | 1.405000 |
| max | 2010.000000 | 12.026000 | 12.026000 | 12.026000 | 12.026000 | 12.026000 | 9.328214 | 12.026000 |

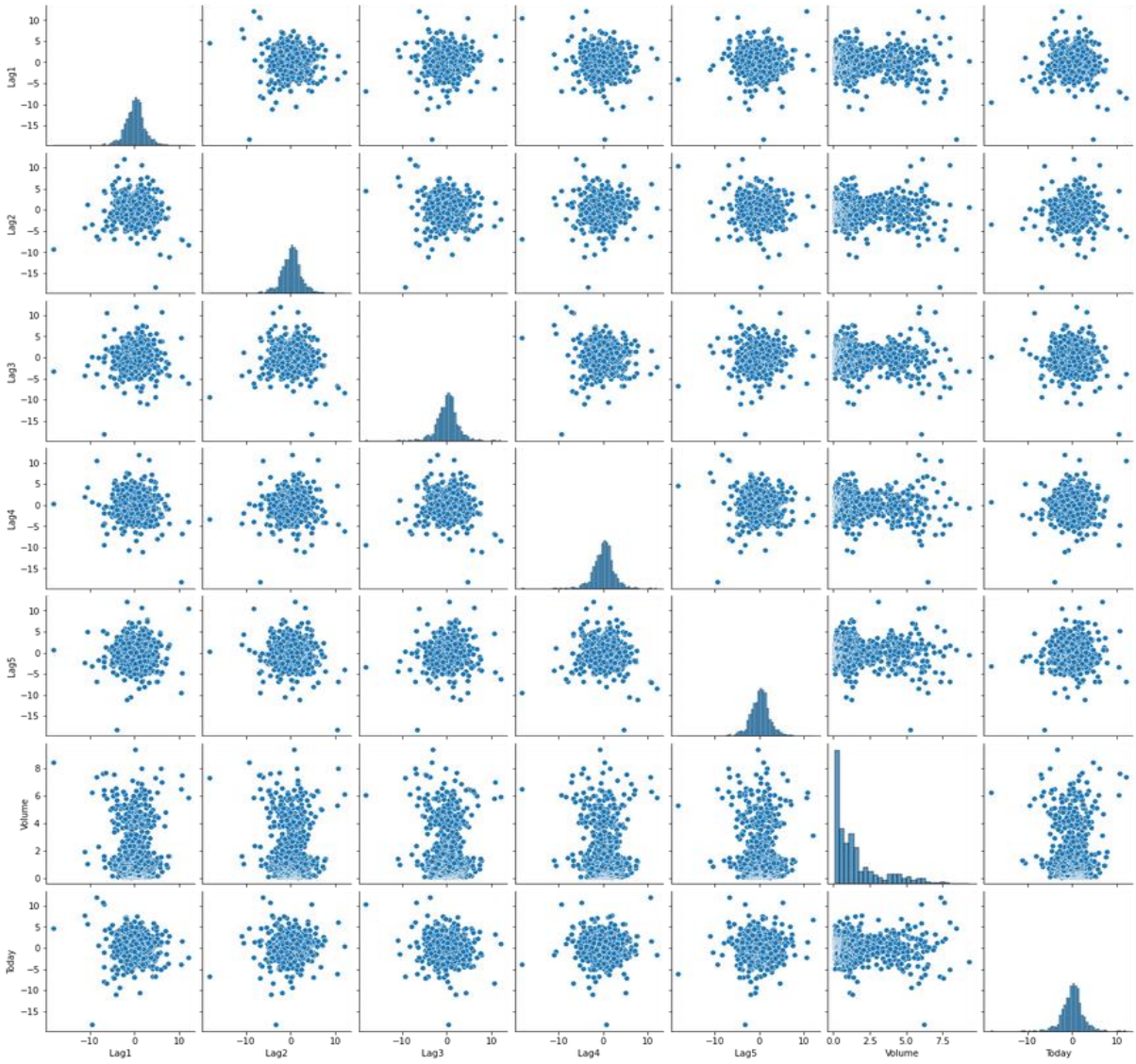
Descriptive statistics, encompassing measures like mean, median, and standard deviation, offer vital insights into the central tendency and variability within the dataset. These statistics provide a foundational understanding necessary for a robust and informed analysis of the data.

| | Year | Lag1 | Lag2 | Lag3 | Lag4 | Lag5 | Volume | Today |
|---|---|---|---|---|---|---|---|---|
| Year | 1.000000 | -0.032289 | -0.033390 | -0.030006 | -0.031128 | -0.030519 | 0.841942 | -0.032460 |
| Lag1 | -0.032289 | 1.000000 | -0.074853 | 0.058636 | -0.071274 | -0.008183 | -0.064951 | -0.075032 |
| Lag2 | -0.033390 | -0.074853 | 1.000000 | -0.075721 | 0.058382 | -0.072499 | -0.085513 | 0.059167 |
| Lag3 | -0.030006 | 0.058636 | -0.075721 | 1.000000 | -0.075396 | 0.060657 | -0.069288 | -0.071244 |
| Lag4 | -0.031128 | -0.071274 | 0.058382 | -0.075396 | 1.000000 | -0.075675 | -0.061075 | -0.007826 |
| Lag5 | -0.030519 | -0.008183 | -0.072499 | 0.060657 | -0.075675 | 1.000000 | -0.058517 | 0.011013 |
| Volume | 0.841942 | -0.064951 | -0.085513 | -0.069288 | -0.061075 | -0.058517 | 1.000000 | -0.033078 |
| Today | -0.032460 | -0.075032 | 0.059167 | -0.071244 | -0.007826 | 0.011013 | -0.033078 | 1.000000 |

A correlation matrix quantifies relationships between variables, indicating the strength and direction of associations. Correlation coefficients, ranging from -1 to 1, help identify patterns and dependencies, essential for informed statistical analyses and modeling decisions.

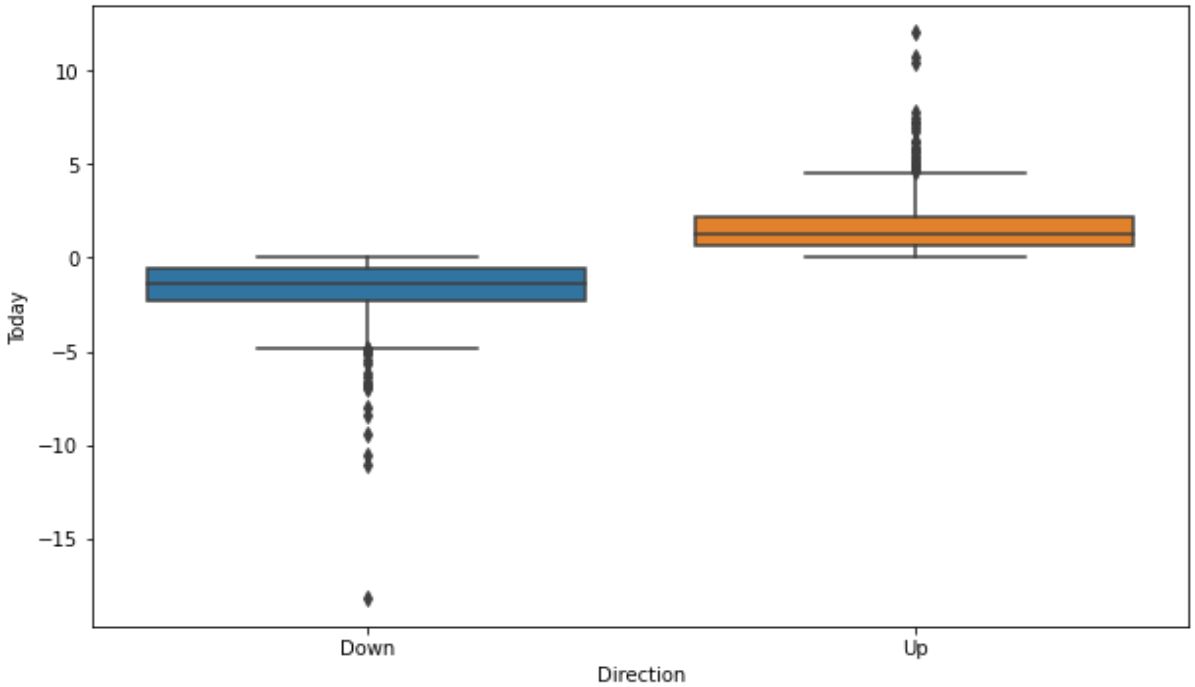
Graphical Summary

1. Pair Plot of Our Various Variables:



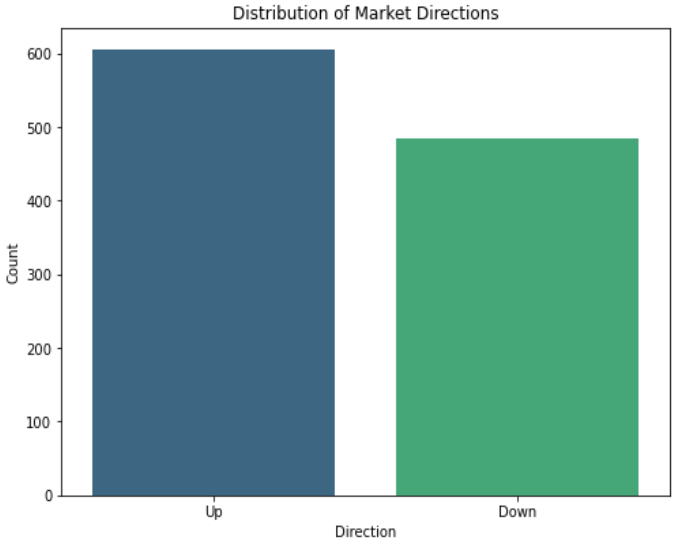

The pair plot visually encapsulates relationships between pairs of variables, showcasing correlations, distributions, and potential patterns. This concise representation aids in identifying trends, dependencies, and insights within the dataset, serving as a valuable preliminary exploration tool for further analysis.

2. Box Plot of Week's Percentage Return (Today) and direction indicators (Direction):



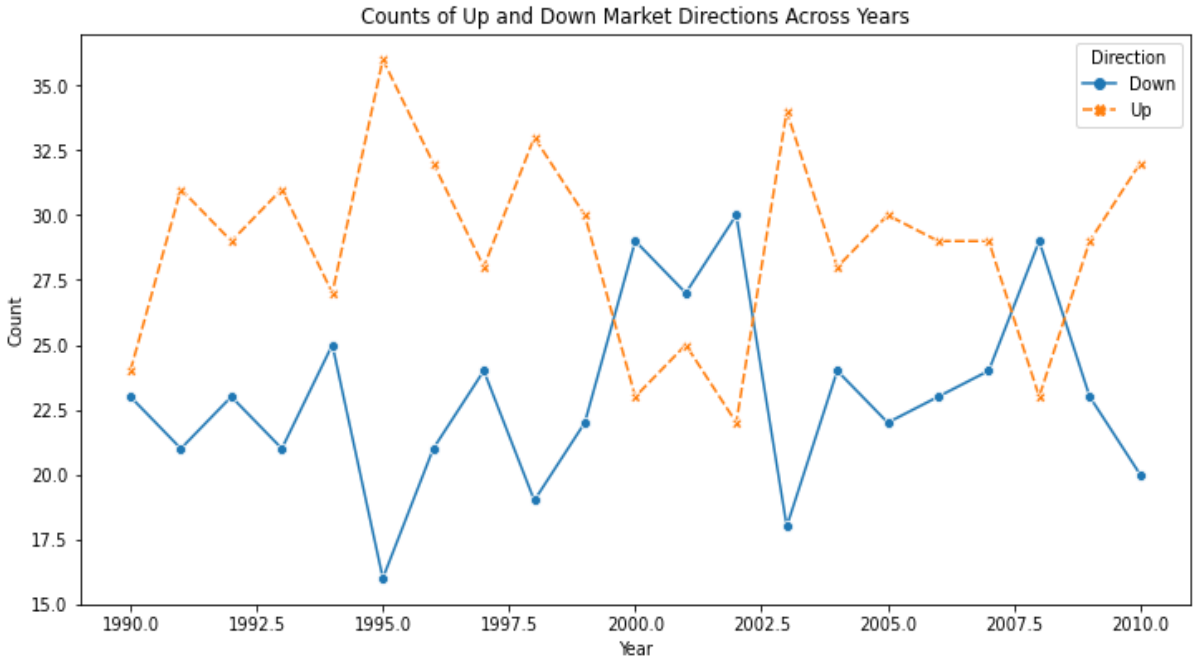

The box plot shows the distribution of a numerical variable called "Direction" over different weekly percentage returns. The variable is categorical, with two possible values: "Down" and "Up.

3. Bar Chart between Direction and Count:




This bar chart depicts the relationship between the Directions and Count. We have more records where the direction is "Up".

4. Line Chart between different years and count to depict the change in trends over time:




This chart gives us a deeper understanding of the trends that occurred in the market. For example, a shade of the 2008 market collapse can be seen here. In the year 2008 we notice that the count of down is high and count of up is low.

## METHODS

**1. Data Collection:**

- Obtain a comprehensive dataset comprising historical stock data, including daily returns, trading volumes, and directional indicators.

**2. Data Preprocessing:**

- Encode the directional variables ('Up' and 'Down') numerically for analysis.
- Handle missing values, ensuring the dataset's integrity.
- Convert relevant predictors to numeric types for compatibility with modeling algorithms.

**3. Descriptive Statistics:**

- Conduct exploratory data analysis using descriptive statistics to understand the central tendency, dispersion, and distribution shape of key variables.

**4. Logistic Regression Modeling:**

- Implement logistic regression models using both scikit-learn and statsmodels libraries.
- Determine effective predictors for analysis.
- Add a constant term and define lagged returns and trading volumes as predictors.
- Split the dataset into training and testing sets for model evaluation.

**5. Model Evaluation:**

- Evaluate logistic regression models using key metrics, including accuracy and confusion matrices, to assess predictive performance.
- Interpret coefficient estimates and associated p-values for individual predictors.
- Calculate the confusion matrix fraction of correct predictions.

**6. Alternative Model Exploration**:

- Explore alternative classification algorithms, such as k-Nearest Neighbors (kNN) and Gaussian Naive Bayes, to compare predictive capabilities.
- Implement these models using scikit-learn and calculate the confusion matrix fraction of correct predictions to assess their performance.

**7. Enhancing kNN model:**

- Further explore kNN with different metrics and value of k.

**8. Pair Plot Visualization:**

- Create a pair plot to visually inspect relationships between key variables, revealing potential patterns and correlations.

**9. Comparative Analysis:**

- Synthesize findings from logistic regression and alternative models, comparing their strengths and weaknesses in capturing directional trends in stock prices.

**10. Result Interpretation:**

- Interpret the results, drawing conclusions about the impact of lagged returns and trading volumes on predicting stock price direction.

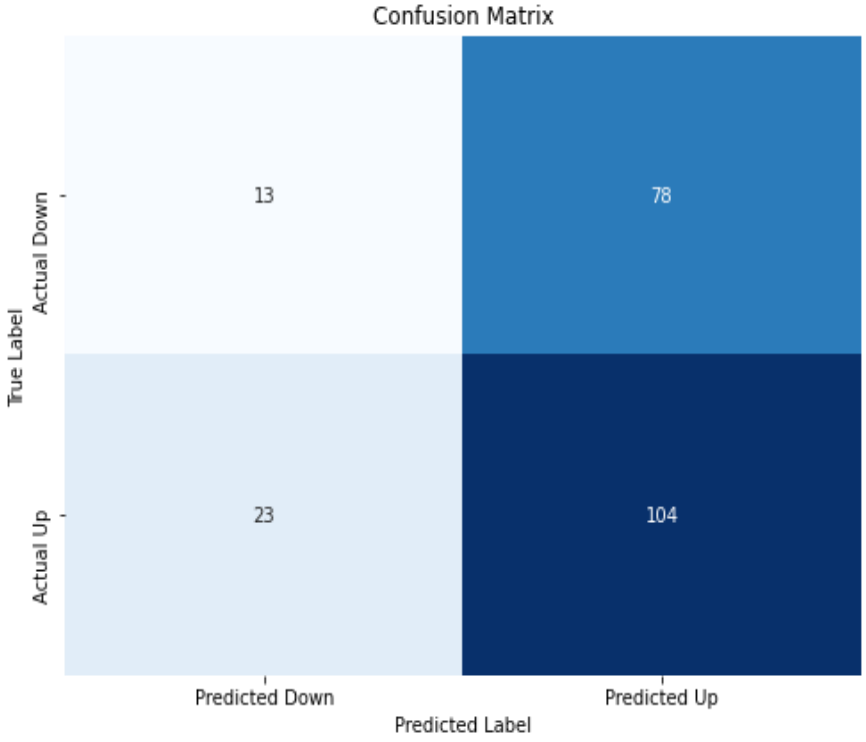
# RESULTS

**1. Logistic Regression (Multiple Predictors):**

- Logistic Regression with Multiple Predictors yielded an accuracy of 53.67%, which depicts that it is fairly accurate in predicting the direction.

```
         Current function value: 0.681414
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  871
Model:                          Logit   Df Residuals:                      864
Method:                           MLE   Df Model:                            6
Date:                Mon, 18 Dec 2023   Pseudo R-squ.:                 0.01012
Time:                        04:33:39   Log-Likelihood:                -593.51
converged:                       True   LL-Null:                       -599.58
Covariance Type:            nonrobust   LLR p-value:                   0.05910
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2513      0.096      2.620      0.009       0.063       0.439
x1            -0.0656      0.029     -2.249      0.025      -0.123      -0.008
x2             0.0651      0.030      2.166      0.030       0.006       0.124
x3            -0.0137      0.031     -0.441      0.659      -0.074       0.047
x4            -0.0296      0.029     -1.012      0.311      -0.087       0.028
x5            -0.0047      0.030     -0.159      0.874      -0.063       0.053
x6            -0.0290      0.041     -0.701      0.484      -0.110       0.052
==============================================================================
```

- The p-values of various variables depict that only x1 and x2 are statistically significant because their values are less than 0.05. Hence, you can reject the null hypothesis that the coefficient for that predictor is zero.



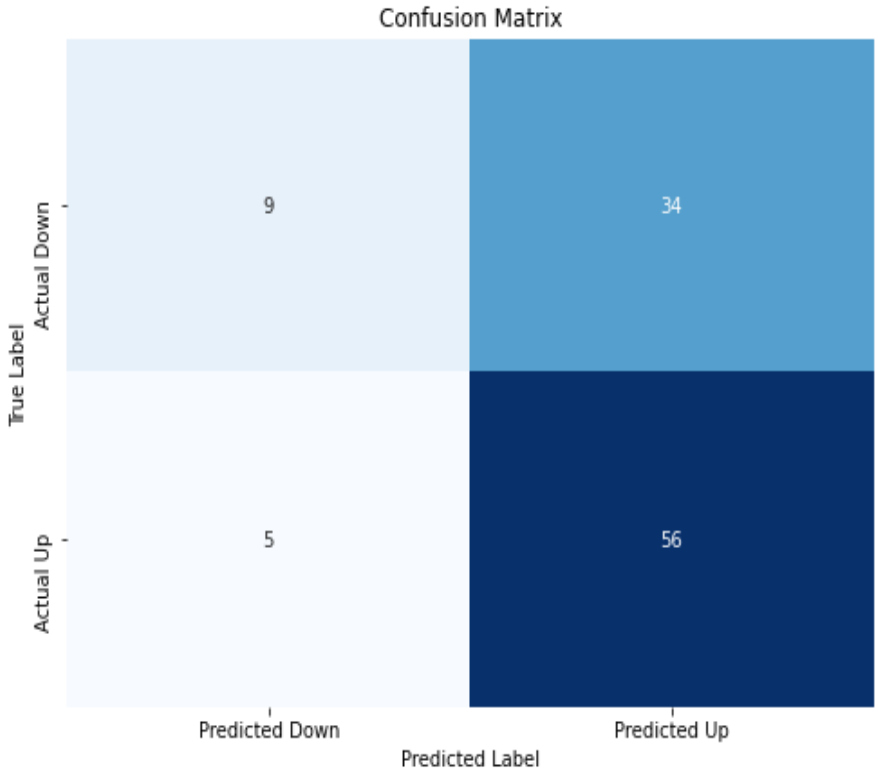

- In stock prediction, True Positives are correct predictions 'Up,' and True Negatives are correct predictions 'Down.' False Positives are incorrect predictions 'Up,' and False Negatives are incorrect predictions 'Down.'

**2. Logistic Regression (Single Predictor):**

- Logistic Regression with a Single Predictor yielded an accuracy of 62.50%, which depicts that it is the most accurate in predicting the direction.

```
Optimization terminated successfully.
         Current function value: 0.685555
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  985
Model:                          Logit   Df Residuals:                      983
Method:                           MLE   Df Model:                            1
Date:                Mon, 18 Dec 2023   Pseudo R-squ.:                0.003076
Time:                        03:58:20   Log-Likelihood:                -675.27
converged:                       True   LL-Null:                       -677.35
Covariance Type:            nonrobust   LLR p-value:                   0.04123
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2033      0.064      3.162      0.002       0.077       0.329
x1             0.0581      0.029      2.024      0.043       0.002       0.114
==============================================================================
```

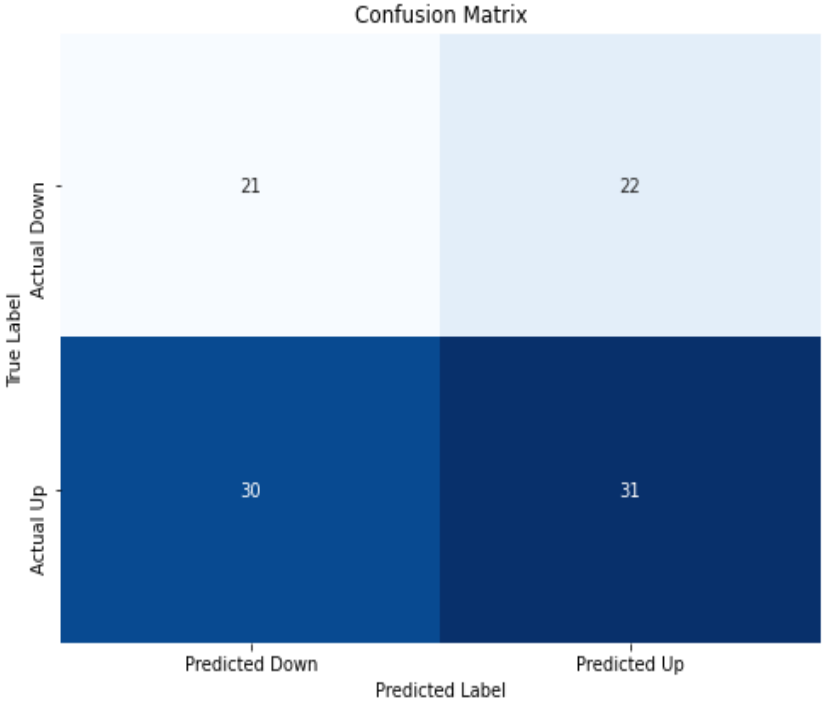
- The p-values of various variables depict that only x1 is a significant variable since it is the only variable used as a predictor.




- Similarly, here True Positives are correct predictions 'Up,' and True Negatives are correct predictions 'Down.' False Positives are incorrect predictions 'Up,' and False Negatives are incorrect predictions 'Down.'
- We can see that this has a very few False Negatives and True Negatives. This is probably the result of the test set being very small as it contains only the data from 2009 and 2010.

**3. k Nearest Neighbors (1 Neighbor):**

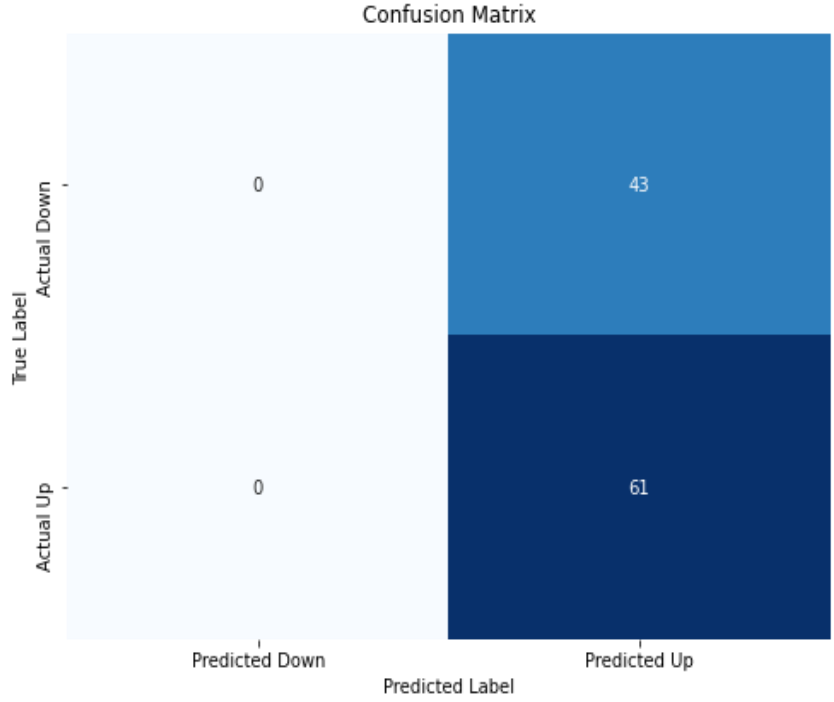
- The k Nearest Neighbors algorithm yielded an accuracy of 50%, which depicts that it is the least accurate in predicting the direction.




- The confusion matrix represents that the model does not do a very good job at prediction. It has a certain balance to its True Positives, False Positives, True Negative and False Negatives. The confusion matrix is a good indicator of why accuracy is precisely 50%.

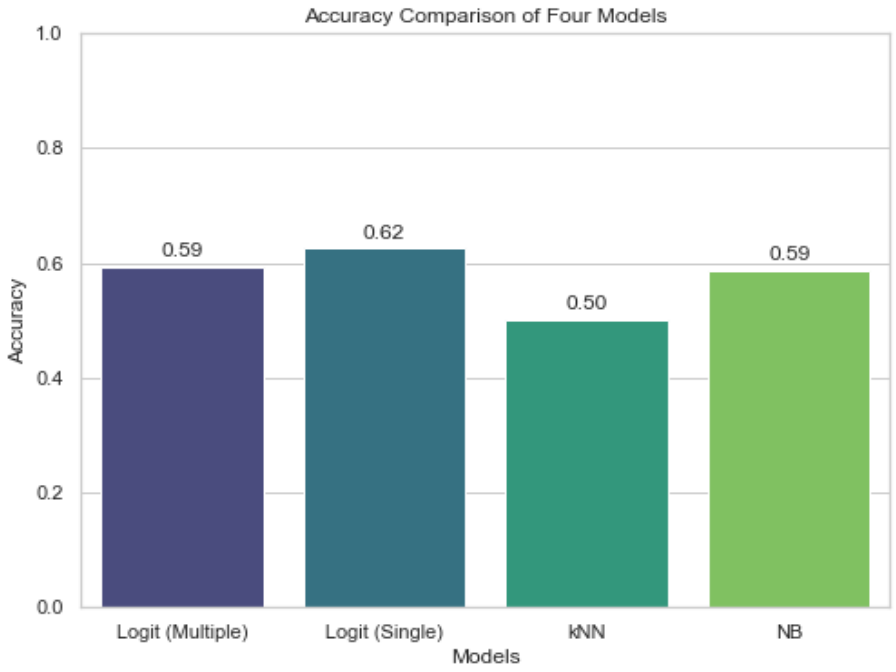
**4. Naïve Bayes:**

- Naïve Bayes yielded an accuracy of 58.65% which depicts that it does a fair job in predicting the direction.



- Naïve Bayes does a fairly well job, but it is not the best. I believe part of the reason is the fact that it has no False Negatives and True Negatives which takes a toll on the accuracy of the model.

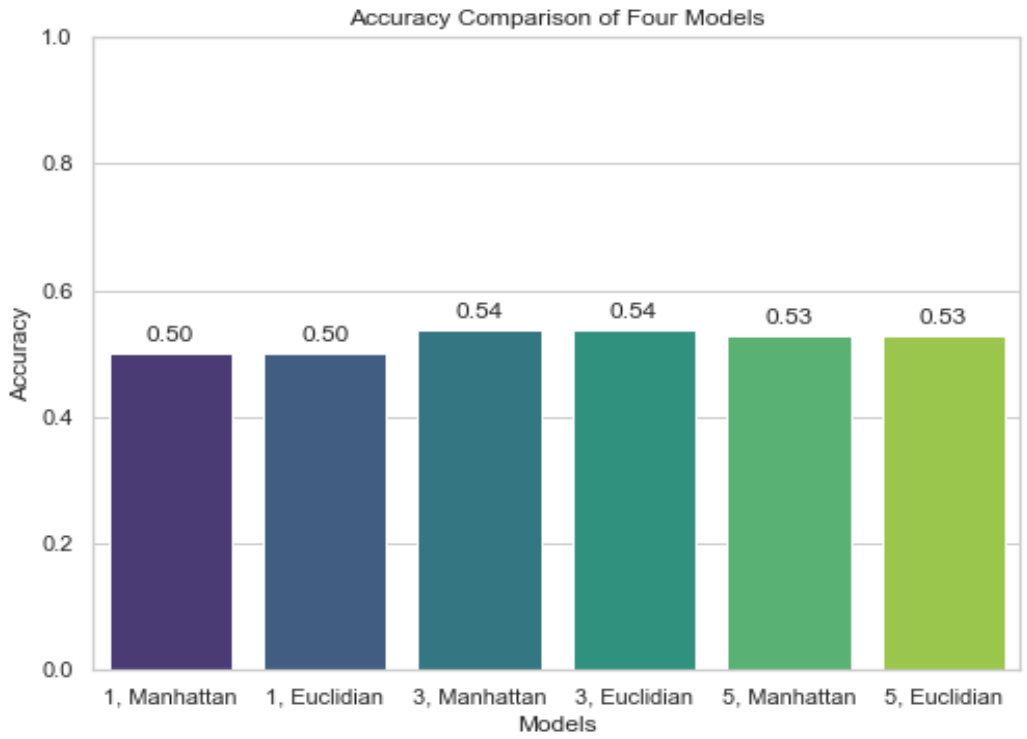
**Comparative Analysis of All the Models:**




- From the graph above it becomes clear that Logistic Regression with a single predictor is the most effective model of all. Part of the reason is that Logistic Regression is made for binary responses.
- Another reason could be the smaller size of the test set as it is apparent from the confusion matrix that it has very few False Negatives and True Negatives.

**5. Building kNN Models with Different Parameters:**

- kNN being the least accurate model compelled us to build models with different parameters. We built the model with a combination of {1, 3, 5} neighbors and {Euclidian, Manhattan} distances. We would have a total of 3 x 2 = 6 models.




- We discover that even after making changes to the parameters of kNN, we don't notice that much change in the accuracy of our models. It only increases by 3 or 4 percent in some cases. Which leads us to the conclusion that the kNN is not very effective for this dataset.

## CONCLUSION

This study explores predictive modeling in finance, focusing on the relationship between financial indicators and direction of stock prices. Logistic regression, k Nearest Neighbors, and Naïve Bayes models are employed to evaluate the impact of lagged returns and trading volumes.

Findings reveal Logistic Regression with a single predictor (Lag2) as the most effective model, emphasizing its suitability for binary responses. Comparative analysis highlights the importance of model simplicity, especially with smaller datasets. Despite parameter adjustments, k Nearest Neighbors exhibits limited effectiveness, while Naïve Bayes falls short of Logistic Regression's accuracy.

In conclusion, this research provides insights into predictive modeling for stock prices, emphasizing Logistic Regression's robustness. Future research may explore additional predictors for improved accuracy in the dynamic financial landscape.

# REFERENCES

[1] Barber, D. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press.