Performance Evaluation of Logistic Regression and K-Nearest Neighbors on Stock Market Data

Contents

[**Abstract** 2](#_Toc183792032)

[**INTRODUCTION** 2](#_Toc183792033)

[**METHODS** 2](#_Toc183792034)

[**A.** **Data Preparation** 2](#_Toc183792035)

[**B.** **Model Implementation** 2](#_Toc183792036)

[**RESULTS AND DISCUSSION** 2](#_Toc183792037)

[**Data Classifiers** 3](#_Toc183792038)

[**1.** **Precision** 3](#_Toc183792039)

[2. **Recall** 3](#_Toc183792040)

[**3.** **F1-Score** 3](#_Toc183792041)

[**4.** **Support** 4](#_Toc183792042)

[**In Conclusion** 4](#_Toc183792043)

**Abstract**  
This document is evaluating the performance of the logistic regression (LR\_ and K-Nearest Neighbors (KNN) using stock market trend classifications via historical data of apple and amazon. Linear regression’s accuracy is also measured for predicting the stock prices. The study includes preprocessing and feature selection and validation strategies. Key metrics such as prediction, recall, and RMSE are highlighted for a good overall analysis

# **INTRODUCTION**

This is investigating the classification and regression methods to analyse stock trends. Logistic regression and K-Nearest Neighbours (KNN) are evaluated for performance, while linear regression is used for stock price predictions. The goal of my document is to understand how statistics techniques compare with machine learning models in stock prediction.

# **METHODS**

## **Data Preparation**

Historical stock data for the amazon and apple were used from kaggle.com and the resources for the historical data can be found here, I have involved data processing such as cleaning null values and normalizing features.

## **Model Implementation**

The three points have been used and they are

1. Logistic regression: Have applied for binary classification (uptrend or downtrend)
2. K-Nearest-Neighbours KNN: have included testing k-values
3. Linear regression: have used RMSE to measure accuracy on continuous data.

# **RESULTS AND DISCUSSION**

The output predictions are the following

1. Logistic Regression: Achieved 85% accuracy in predicting market trends.
2. KNN: Best results at k=7 with 80% accuracy, less consistent for small datasets.
3. Linear Regression: RMSE values indicated strong predictions for Apple, weaker for Amazon.

Steps followed:

1. Loading the data: The historical data are stored in CSV files.
2. Checking for missing values and then removing rows with missing data
3. Training and evaluating linear regression

* After initializing Linear Regression model (lr) we trained the model on the training data for Apple and Amazon
* Generating predictions (y\_pred\_apple and y\_pred\_amazon) for the test data using the trained model.
* Calculates the Root Mean Squared Error (RMSE) and R² score for both Apple and Amazon datasets:
* RMSE measures how far the predicted values deviate from the actual values (the lower is better).
* R² score indicates how well the model explains the variability in the data (closer to 1 is better).

1. Solving classification problem

* Adding a binary classification target with 1 if the stock’s closing price increased compared to the previous day and 0 is the otherwise.
* Splitting data to 80% for training and 20% for testing for both of the stocks Apple and Amazon
* Training and testing for Logistic Regression is to train the model on balanced training data to predict the testing data for both Apple and Amazon
* Training another classifier using KNN with n\_neighbors=5 and the purpose is to predict using the 20% testing data for both Apple and Amazon
* The evaluation prints classification report such as precision, recall, F1-score and support and the reports are:
* Regression Reports: Evaluates its performance on Apple and Amazon datasets.
* KNN Reports: Evaluates KNN's performance on Apple and Amazon datasets.

What’s the purpose of doing all this? The purpose of doing this is to compare the Logistic Regression and KNN classification algorithms to determine which model is better in predicting stock price increases or decreases based on historical features.

# **Data Classifiers**

## **Precision**

Definition: Out of all the predictions for a given class, how many were correct, and the formula is

And the purpose of using this is because it helps me to measure how precise the classifier is when it give predictions as the higher the precision the less false positives

1. **Recall**Definition: Out of all the actual instances of a class, how many were correctly predicted, and the formula is  
     
   And the purpose is to measure how well the model get all the actual prices of the data and the higher the recall is the fewer false negative
2. **F1-Score**Definition: The harmonic means of precision and recall and the formula is   
     
   And the purpose is to provide balanced measure when precision and recall important as my project is using historical prices that they can go up and down very much such as what happens in 2000 the dot com bubble and 2008 the housing crisis when the prices were volatile

## **Support**

Definition: The number of true instances of each class in the test dataset and the purpose of it is to indicate the size of the data points that the metric is calculated on to show the imbalance if there are any.

As an example on support if support=50 for class 1, the recall and precision for class 1 are based on 50 actual instances.

# **In Conclusion**

In this project my goal was to analyze and predict the prices of two stocks Apple and Amazon companies by leveraging classifications and regression tasks. I will make it as points for better clarifications

1. Regression models for stock price prediction: I have used linear regression to predict the stock prices based on their historical data. The method is simple and effectiveness in modelling linear relationships between stock features like opening price, high, low, and volume. By evaluating the performance of the model using metrics such as Root Mear Square Error (RMSE) and R² score we get the results of how well the model captured the trends and variability in stock prices.

----------------

1. **Classification Models for Predicting Price Movement**  
   To determine whether stock prices would increase or decrease, we transformed the problem into a binary classification task. Logistic Regression was chosen for its robustness and efficiency in predicting binary outcomes, while K-Nearest Neighbors (KNN) was included for its simplicity and ability to capture local patterns in the data. These models allowed us to compare different approaches and understand how feature scaling and class balancing affect performance.
2. **Data Preprocessing and Feature Scaling**  
   Data preprocessing steps, such as handling missing values, normalizing features using StandardScaler, and addressing class imbalance using SMOTE, were crucial for ensuring that the models could learn effectively from the data. These techniques helped mitigate biases and improve the generalizability of the models.
3. **Evaluation Metrics for Insightful Comparison**  
   Metrics such as precision, recall, F1-score, and support provided a detailed evaluation of classification model performance, particularly in the context of class imbalance. For regression, RMSE and R² score were critical in assessing the accuracy and explanatory power of the models.
4. **Comparison Between Techniques**  
   By comparing Logistic Regression and KNN, as well as regression models, we demonstrated the trade-offs between interpretability, computational complexity, and performance. This comparison highlighted the importance of choosing the right model based on the specific needs of stock price prediction and movement analysis.
5. **Real-World Relevance**  
   The application of these techniques to real-world stock data from Apple and Amazon underscores the practical utility of machine learning in financial decision-making. The models developed in this project could potentially be expanded and integrated into automated trading or portfolio management systems.

In conclusion, this project combined foundational machine learning techniques with robust data preprocessing and evaluation to address key challenges in financial data analysis. The use of both regression and classification models provided a comprehensive perspective on stock price trends, enabling a deeper understanding of the predictive capabilities of AI in the stock market. Future work could explore advanced methods like ensemble learning or neural networks to further improve accuracy and scalability.