Project for the subject

ARTIFICIAL INTELLIGENCE

(UCS411)

Submitted by

Aayushi Puri (102103676)
Group No.: COE24

Project Title: Stock Price Prediction

Course Instructor:
Dr. Simran Setia

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY,
PATIALA SESSION: SUMMER SEMESTER 2024

(Date of submission: 17/07/2024)

CONTENTS

Introduction

Stock price prediction plays a crucial role in financial markets, enabling investors to make informed decisions and potentially achieve significant financial gains. This project focuses on developing models to predict stock prices using historical data, utilizing both Linear Regression and Decision Tree algorithms.

Motivation of Project

The motivation for this project stems from the following factors:

1. **Financial Gains**: Accurate stock price predictions can help investors maximize their returns by making timely and informed decisions.
2. **Market Efficiency**: Enhancing the efficiency of financial markets by reducing uncertainties and helping in better price discovery.
3. **Technological Advancement**: Utilizing advanced machine learning algorithms and big data analytics to solve complex financial problems.
4. **Educational Purpose**: Gaining a deeper understanding of financial markets, stock price movements, and the application of machine learning in real-world scenarios.

Dataset Details

For this project, we utilized historical Stock price data of HCLTECH stock at day level. The dataset includes:

1. **Time Period**: The dataset spans from January 2000 to April 2021.
2. **Frequency**: with 5300 entries
3. **Attributes**:
    - **Date**: The date of the stock price data point.
    - **Open**: The opening price of the stock on the given date.
    - **High**: The highest price of the stock on the given date.
    - **Low**: The lowest price of the stock on the given date.
    - **Close**: The closing price of the stock on the given date.
    - **Volume**: The number of shares traded on the given date.

Link of Dataset: https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data

	Date	Open	High	Low	Close	Volume	Name
0	11-01-2000	1550.0	1725.00	1492.00	1554.45	1192200	HCLTECH
1	12-01-2000	1560.0	1678.85	1560.00	1678.85	344850	HCLTECH
2	13-01-2000	1790.0	1813.20	1781.00	1813.20	53000	HCLTECH
3	14-01-2000	1958.3	1958.30	1835.00	1958.30	270950	HCLTECH
4	17-01-2000	2115.0	2115.00	1801.65	1801.65	428800	HCLTECH

Implementation

Data Preprocessing

1. **Data Cleaning**: Handling missing values, outliers, and erroneous data points.
2. **Feature Selection**: features include Open price, High price, Low price, Close price and Volume
3. **Normalization**: Scaling the data to ensure all features contribute equally to the model.

Model Selection

Decision Trees: A non-linear model to capture complex patterns in the data. Decision Tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction. Output is calculated on the basis of reduced variance which we can refer to impurity factor.

Linear Regression: Linear Regression is statistical approach for modelling relationship between a dependent variable with a given set of independent variables
Example:-
x as feature vector, i.e..... x=[x_1,x_2,.....x_n]
Y as response vector, i.e.....y=[y_1,y_2.....y_n]

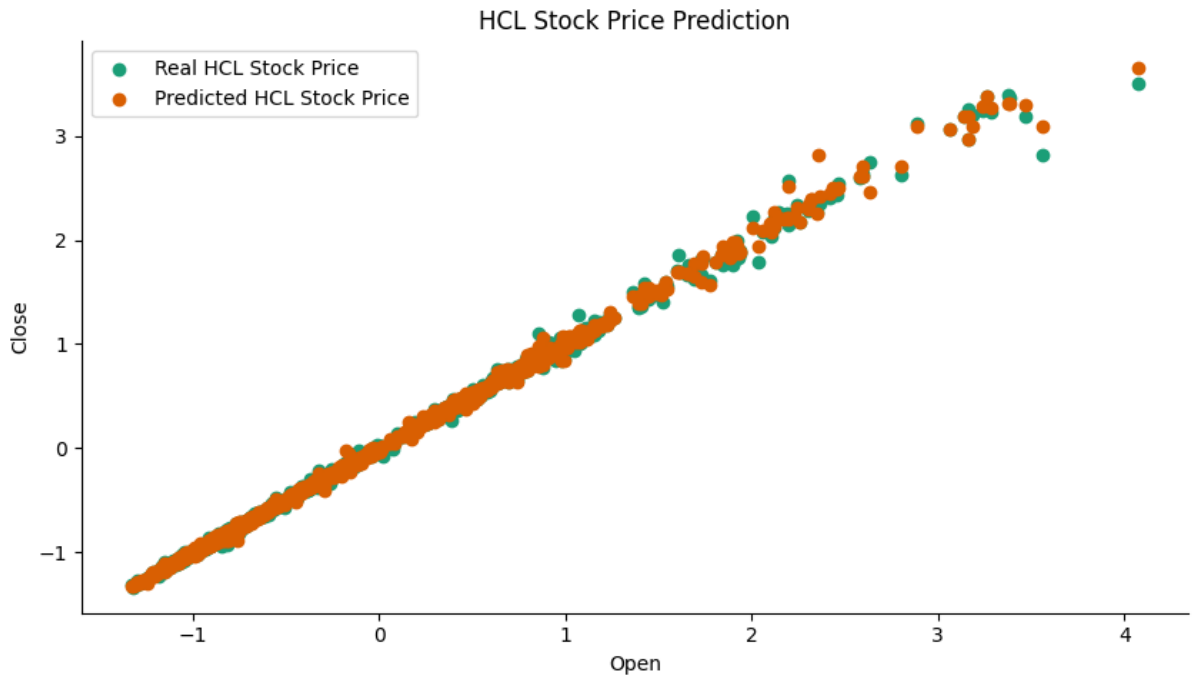
Model Training and Evaluation

1. **Training**: Splitting the dataset into training and testing sets, with 80% for training and 20% for testing.
2. **Hyperparameter Tuning**: Optimizing the model parameters using techniques such as Grid Search and Random Search.
3. **Evaluation Metrics**: Using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared to evaluate model performance.

Results

The performance of the models was evaluated using the testing dataset. Key results include:

1. **Decision Trees**:
    - Mean Squared Error: 0.0011
    - Mean Absolute Error: 0.0180
    - R-squared: 0.9988
2. **Linear Regressor**:
    - Mean Squared Error: 0.0005
    - Mean Absolute Error: 0.0111
    - R-squared: 0.9995



The Linear Regressor outperforms the Decision Tree with a lower MSE (0.0005 vs. 0.0011) and MAE (0.0111 vs. 0.018), and a higher R^2 (0.9995 vs. 0.9988), indicating better accuracy and predictive performance. Despite both models fitting the data well, the Linear Regressor is the recommended choice for this dataset.

Conclusion

Both models perform exceptionally well, with very high R^2 values indicating that they both fit the data very well. The Linear Regressor model outperforms the Decision Tree model in terms of both MSE and MAE, suggesting it is more accurate and has better predictive performance, while the higher R-squared value suggests it explains more variance in the data. The.

However, it's also important to consider the context and nature of the problem, as decision trees might still offer advantages in terms of interpretability and handling non-linear relationships.