**Artificial**

**Intelligence and Machine Learning**

Project Report

Semester-IV (Batch-2022)

**Stock price prediction**



**Supervised By: Submitted By:**

**Mohammad Talib Shubham Jain 2210992365**

**Shubham 2210992360**

**Sudeep 2210992412**

**Sneha Rani 2210992391**

**Department of Computer Science and Engineering**

**Chitkara University Institute of Engineering & Technology,**

**Chitkara University, Punjab**

**Table of Content**

|  |  |  |
| --- | --- | --- |
| Sl. No | Name | Page No |
| 1 | Introduction | 3-5 |
| 2 | Exploratory Data Analysis | 6-7 |
| 3 | Model Building | 8-9 |
| 4 | Model Evaluation and Comparison | 10-11 |
| 5 | Results | 11-12 |
| 6 | Conclusions | 13 |
| 7 | References | 13 |

**1.Introduction**

**1.1 Project Objectives**

1. **Forecast Future Stock Prices:** Develop models to accurately predict future stock prices based on historical data and other relevant financial indicators, aiming to provide insights for investors and traders.
2. **Improve Investment Strategies:** Enhance the decision-making process in financial markets by using predictive analytics to optimize investment portfolios, reduce risks, and maximize returns.
3. **Leverage Machine Learning Techniques:** Utilize advanced AI and machine learning algorithms to analyze vast amounts of financial data, identify patterns, and generate reliable predictions that traditional statistical methods might miss.

**1.2 Scope of the Project**

1. **Data Preparation:** Gathering and preprocessing financial data for model input.
2. **Model Building:** Training and validating machine learning models to predict stock prices.
3. **Real-World Application:** Deploying models for practical use in trading and investment.

**1.3 Overview of Techniques**

 **Data Collection:** Aggregating historical stock prices and relevant financial indicators.

 **Algorithm Selection:** Choosing suitable machine learning techniques for prediction.

 **Model Deployment:** Implementing and testing the prediction models in real-world scenarios

**1.4 Data Features Overview**

1. **Historical Prices:** Past stock prices, including open, high, low, and close values.
2. **Volume:** Number of shares traded during a given period.
3. **Moving Averages:** Average stock prices over specified time frames (e.g., 50-day, 200-day).
4. **Technical Indicators:** Metrics like RSI, MACD, and Bollinger Bands.
5. **Economic Indicators:** Interest rates, inflation rates, and GDP growth.
6. **Sentiment Analysis:** Public sentiment derived from news articles and social media.
7. **Company Financials:** Earnings reports, revenue, profit margins, and debt levels.
8. **Industry Trends:** Performance and trends within the stock’s industry sector.
9. **Market Indices:** Indices like S&P 500 or NASDAQ for broader market context.
10. **Corporate Actions:** Information on stock splits, dividends, and mergers.

**2. Exploratory Data Analysis (EDA)**

**2.1 Dataset Description**

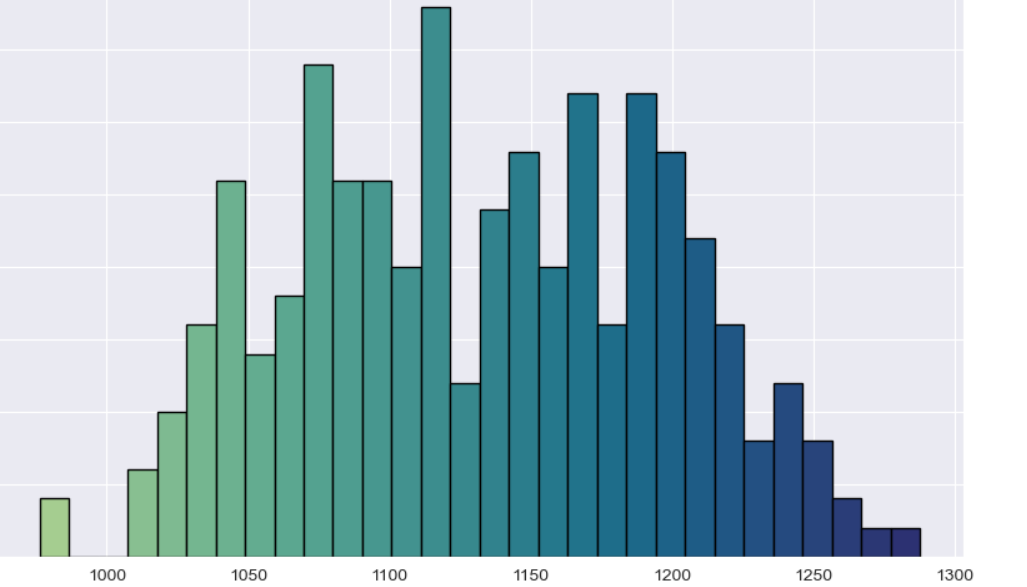
1. **Descriptive Statistics: Summary of central tendency, dispersion, and shape of dataset distributions.**
2. **Time Series Analysis: Examination of stock price trends and patterns over time.**
3. **Correlation Analysis: Identifying relationships between different financial indicators and stock prices.**
4. **Missing Data Handling: Techniques for dealing with incomplete data entries.**
5. **Visualization: Graphical representations of data, such as line charts and candlestick plots**

**2.2 Data Preprocessing**

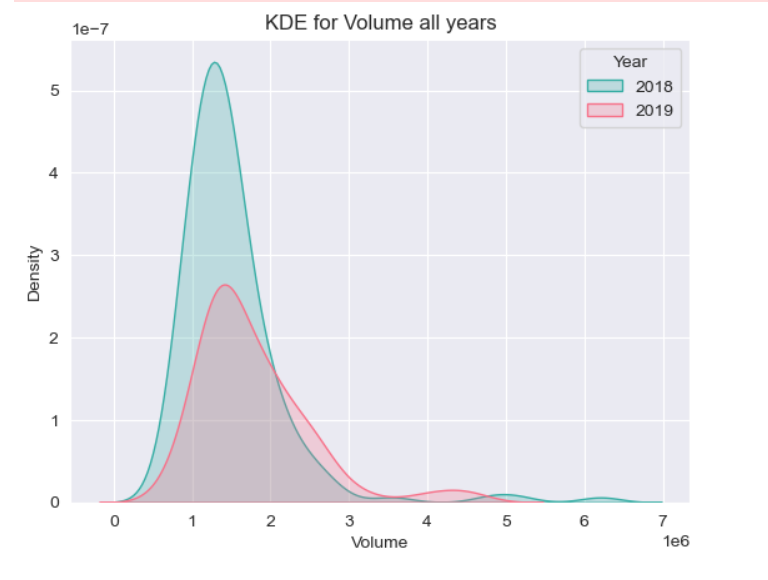
1. **Handling Missing Values: Imputing or removing missing data points to ensure dataset completeness.**
2. **Normalization or Scaling: Standardizing numerical features to a common scale for model compatibility.**
3. **Feature Engineering: Creating new features or transforming existing ones to improve model performance.**
4. **Removing Outliers: Identifying and filtering out data points that deviate significantly from the majority.**
5. **Splitting Data: Partitioning the dataset into training, validation, and test sets for model evaluation.**

**2.3 Data Visualization**

1. **Time Series Plots:** Visualizing stock prices over time to identify trends and patterns.
2. **Histograms:** Displaying the distribution of stock price data to understand its spread and shape.
3. **Correlation Heatmaps:** Illustrating the relationships between different financial indicators and stock prices.
4. **Candlestick Charts:** Representing open, high, low, and close prices for a given period, aiding in pattern recognition.
5. **Scatter Plots:** Examining the relationship between two variables, such as volume traded and stock price







**3 Introduction to Machine Learning Models**

1. **Linear Regression:** Predicting stock prices based on linear relationships with input features.
2. **Support Vector Regression (SVR):** Utilizing support vector machines to forecast stock prices while considering non-linear relationships.
3. **Random Forest Regression:** Employing an ensemble of decision trees to capture complex interactions and make accurate predictions
4. **Decision trees** : They are hierarchical models that recursively partition data based on feature values to make predictions.

**3.1 Linear Regression**

 **Linear Relationship:** Linear regression models the relationship between the independent variables and the dependent variable using a straight line.

 **Least Squares Optimization:** Linear regression minimizes the sum of squared differences between observed and predicted values to find the best-fitting line.

 **Interpretability:** Linear regression provides interpretable coefficients, allowing easy understanding of the impact of each independent variable on the dependent variable.

**3.2 SVM (support vector machine)**

* Margin Maximization: SVM aims to find the hyperplane that maximizes the margin, the distance between the hyperplane and the nearest data points of each class.
* Kernel Trick: SVM can efficiently handle non-linearly separable data by mapping input features into a higher-dimensional space using kernel functions.
* Regularization: SVM incorporates regularization parameters to control the trade-off between maximizing the margin and minimizing classification errors, preventing overfitting

**3.3 Random Forest Model**

* We first split the dataset into training and testing sets using the train\_test\_split function.
* We then initialize the Random Forest Regressor with 100 trees (n\_estimators=100) and set We first split the dataset into training and testing sets using the train\_test\_split function.
* We then initialize the Random Forest Regressor with 100 trees (n\_estimators=100) and set random\_state=42 for reproducibility.
* Next, we train the model using the training data with the fit method.

**3.4 Decision Tree Model**

* Decision trees are a non-linear algorithm that partitions the feature space into regions based on the values of the input features. Each internal node of the tree represents a decision based on a feature, and each leaf node represents a predicted quality. Decision trees are known for their interpretability and ability to capture relationships in the data.
* Hyperparameter tuning is essential in decision tree modeling to prevent overfitting and improve model performance. By adjusting parameters such as the maximum depth of the tree, minimum samples per leaf, and splitting criteria, we can optimize the model's accuracy and generalization ability.

**4. Model Evaluation and Comparison**

**4.1 Evaluation Metrics**

In this section, we discuss the evaluation metrics used to assess the performance of the models. These metrics provide insights into how well the models are performing and help us identify areas for improvement.

* **ACCURACY :** Accuracy measures the proportion of correctly classified instances out of the total number of instances. It's calculated as the ratio of the number of correct predictions to the total number of predictions made by the model.
* **RECALL:** Recall measures the ability of the model to correctly identify all positive instances. It calculates the ratio of true positives to the sum of true positives and false negatives.
* **F1 SCORE :** score is the harmonic mean of precision and recall. It provides a balance between precision (the ability of the model to correctly identify positive instances) and recall (the ability of the model to capture all positive instances).

**4.2 Model Comparison**

* We compare the performance of the different models based on the evaluation metrics discussed earlier. This allows us to identify the strengths and weaknesses of each model and determine which model is best suited for house price prediction.
* Decision trees are versatile models that can capture nonlinear relationships and interactions between features. They are easy to interpret and visualize, making them suitable for exploratory analysis. However, decision trees are prone to overfitting, especially with complex datasets.
* Random forests address the overfitting issue of decision trees by aggregating the predictions of multiple trees. They are robust and perform well on a wide range of datasets. Random forests also provide a measure of feature importance.
* Overall, each model has its advantages and limitations, and the choice of model depends on the specific requirements of the problem at hand. In the next section, we discuss the implications of
* our findings and provide recommendations for further research.

**5. Results**

In this section, we present the evaluation metrics for each machine learning model used in the quality prediction task. The evaluation metrics provide insights into the performance of each model and help assess their accuracy and generalization ability.

**5.1 Logisitc Regression Evaluation Metrics:**

Logisitc Regression Evaluation Metrics:

Accuracy: 0.93125

Precision: 0.9533678756476683

Recall: 0.908641975308642

F1 Score: 0.9304677623261693

With an accuracy of 0.93125, the model correctly predicts 93.125% of instances. Precision, measuring the proportion of true positive predictions among all positive predictions, is 0.953, indicating a high rate of correctly identified positive instances. The recall of 0.909 suggests that the model captures 90.9% of all true positive instances. The F1 score, which balances precision and recall, is 0.930, indicating overall good performance.

**5.2 Decision Tree Evaluation Metrics:**

* gini impurity- 0.91375
* entrophy- 0.92375

The decision tree model achieves gini impurity of 0.92375, indicating a slightly lower average deviation from the actual values compared to the logistic regression model.

**5.4 Random Forest Evaluation Metrics:**

Train data accuracy = 1.0

Test data accuracy = 0.95125

**6. Conclusion**

* In this study, we explored the task of quality prediction using machine learning algorithms. We investigated four different models: logistic regression, decision tree and random forest. Through extensive evaluation, we assessed their performance using various metrics.
* The results reveal that the logistic regression model outperformed the other models in terms of precision and F1 score. However, all models exhibited limitations indicating potential challenges in accurately predicting quality.
* Despite these limitations, our study underscores the significance of machine learning in determining quality. By leveraging advanced algorithms and techniques, stakeholders can gain valuable insights aiding in informed decision-making.
* In future research, enhancing model performance and addressing limitations such as underfitting or overfitting will be crucial. Additionally, exploring ensemble methods and incorporating more sophisticated feature engineering techniques could further improve predictive accuracy.

**7. References**

* ChatGPT. (n.d.). OpenAI. Retrieved from <https://openai.com/chatgpt/>
* Kaggle. (n.d.). Kaggle: Your Home for Data Science. Retrieved from <https://www.kaggle.com/>