**Amazon Stock Price Prediction Using Linear Regression**

**1. Introduction**

This project aims to predict Amazon stock prices using a linear regression model based on historical data. The analysis focuses on understanding the relationship between stock prices, commodity prices, and trading volumes to generate reliable predictions. The entire machine learning pipeline, from exploratory data analysis (EDA) to model evaluation, is covered.

Key machine learning techniques such as feature engineering, data pre-processing, and model validation are applied to ensure the robustness of the predictions. Although linear regression is used as a baseline, the project paves the way for incorporating more complex models for higher accuracy in future iterations.

**2. Dataset**

The dataset contains historical data on Amazon stock prices along with commodity prices and trading volumes. Handling missing data and feature standardization were critical pre-processing steps to ensure the integrity of the dataset for model training.

* **Missing Data Handling**: Missing values in the dataset were imputed using the mean of respective columns. This approach preserved the dataset's size while maintaining data accuracy.
* **Feature Standardization**: To prevent features with large ranges from dominating the model, all features were standardized using a StandardScaler. This ensured uniformity across the dataset and helped the linear regression model learn effectively.

**Data Visualization**

Visual representations of the dataset will provide insights into feature relationships:

* **Line Plots**: Showing the historical trend of stock prices.

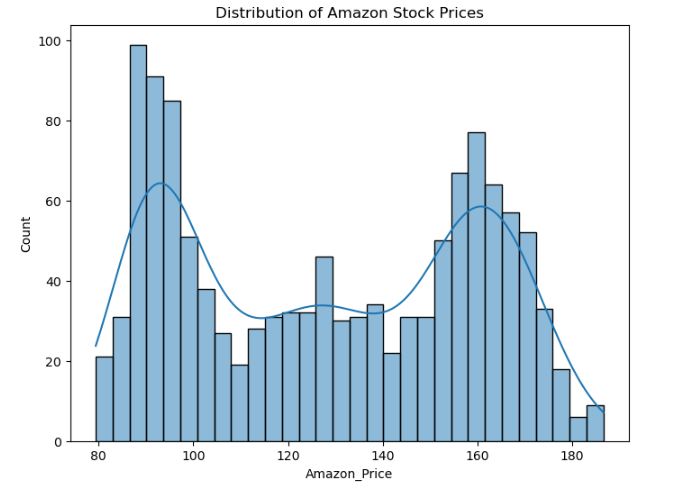


Figure 1

* **Scatter Plots**: Illustrating relationships between stock prices and commodity volumes.

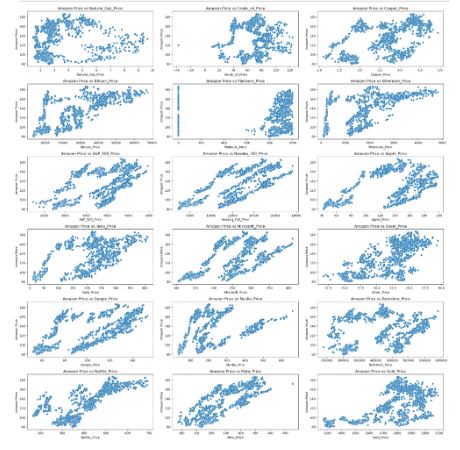


Figure 2

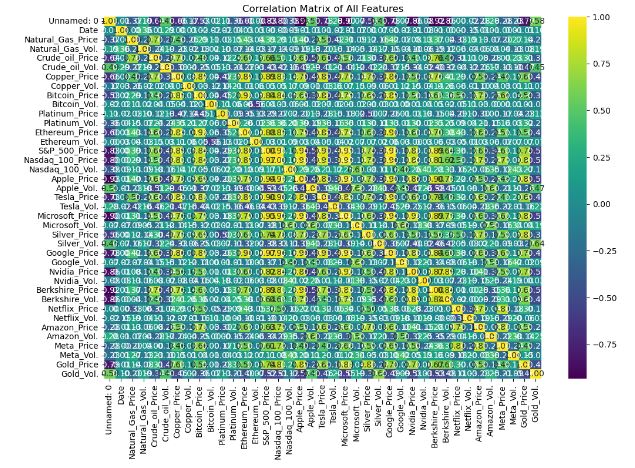
* **Correlation Heatmap**: Highlighting the degree of correlation between different features and stock prices.

Figure 3

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

EDA is a key step in understanding the dataset before model development. Key insights were gained from statistical summaries and visualizations, revealing correlations between Amazon stock prices and other relevant features like trading volumes and commodity prices.

**Key Observations from EDA:**

* Stock prices showed a strong correlation with trading volumes and specific commodities.
* There were identifiable patterns in historical stock price movements that suggest a linear relationship between features.

**Visualizations from the EDA** will include:

* **Price Trends**: Showing fluctuations in Amazon stock prices over time.
* **Correlation Matrix**: To visualize feature interdependencies.

**4. Feature Engineering**

After performing EDA, the next step was to transform and prepare features for model training. The following methods were used to handle the dataset:

* **Data Imputation**: Missing values were filled using column-wise means, ensuring no rows were discarded.
* **Feature Selection**: Only features with strong correlations to stock prices were selected to enhance prediction accuracy.

The dataset was then standardized to eliminate bias from features with larger numerical ranges.

**5. Model Development**

A linear regression model was chosen for this project, given its simplicity and ease of interpretation. The following steps were taken in the model development process:

1. **Data Split**: The dataset was split into 80% for training and 20% for testing to avoid data leakage and ensure robust model evaluation.
2. **Cross-Validation**: To prevent overfitting and assess the generalization ability of the model, 5-fold cross-validation was used.

**Model Training Pipeline**:

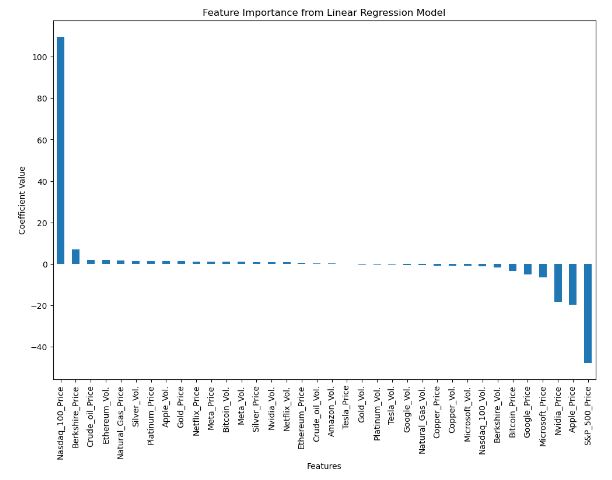
* **Exploratory Data Analysis (EDA)**: Initial visualizations and statistical summaries.
* **Feature Engineering**: Imputation of missing values, feature selection, and standardization.
* **Training**: The linear regression model was trained on the processed dataset.

Figure 4

**6. Model Evaluation**

The model was evaluated using three primary metrics:

* **Mean Absolute Error (MAE)**
* **Mean Squared Error (MSE)**
* **Root Mean Squared Error (RMSE)**

The evaluation results showed that the model provided reasonably accurate predictions of Amazon stock prices, though improvements could be made with more complex models.

**Visualizations for Model Results**:

* **Actual vs Predicted Prices**: A line plot showing the comparison between real stock prices and predicted values.

**7. Challenges and Solutions**

During the project, several challenges were faced:

* **Handling Missing Data**: This was resolved using mean imputation, ensuring the integrity of the dataset.
* **Feature Selection**: Determining which features were most relevant to stock price prediction required careful analysis.
* **Computational Resources**: Running cross-validation and hyperparameter tuning was resource-intensive. This was mitigated by running smaller batch sizes and testing on local machines.

**Real-Time Data Limitations**

Real-time stock data was not included, limiting the model’s ability to account for current market conditions.

**8. Conclusion**

This project demonstrated the successful application of a linear regression model to predict Amazon stock prices. The pipeline from data pre-processing to model evaluation emphasized the importance of clean data and careful model selection. Although linear regression is a good starting point, future iterations of the project can explore more sophisticated models such as Random Forest or Neural Networks to improve prediction accuracy.

**9. Future Work**

* **Incorporating Real-Time Data**: Implementing real-time stock data will make the model more relevant to current market conditions.
* **Exploring Complex Models**: In the future, advanced algorithms like XGBoost or LSTM could enhance the predictive performance of the model.

**10. GitHub Repository**

The project code and Jupyter Notebook can be accessed in the GitHub repository: [GitHub Link](https://github.com/ss1291/Stock-Price-Prediction-Project)

**THE END**