Project Report: Predicting Amazon Prices Based on Market Data

HUBBLEMIND **Shashank R**

1. Introduction

This project aims to predict the price of Amazon stock using various market data features such as crude oil prices, Bitcoin prices, and other financial indicators. By leveraging historical data, we explore the relationships between different variables and develop a machine learning model to predict Amazon's stock price.

The project is divided into four phases:

- **Data exploration and preprocessing**
- **Model development**
- **Model validation and testing**
- **Documentation and submission**

2. Challenges

Throughout the project, several challenges arose:

- **Handling Non-Numeric Data**: The dataset contained non-numeric columns (e.g., Date), as well as numeric columns with commas, which required preprocessing.

- **Missing Data**: Some columns had missing values that needed to be handled properly to avoid biasing the model.

- **Feature Scaling**: Ensuring that features with varying scales were standardized to avoid disproportionate influence on the model.

- **Model Accuracy**: Tuning the model to balance prediction accuracy while avoiding overfitting or underfitting.

3. Solutions

To overcome the challenges:

- **Data Preprocessing**: Non-numeric columns were removed, and commas in numeric columns were handled by converting string values to float. Missing values were filled with the mean of the respective columns.

- **Feature Engineering**: StandardScaler was applied to standardize the features for better model performance.

- **Cross-Validation**: 5-fold cross-validation was applied to evaluate the model's generalization on unseen data.

- **Error Metrics**: We used multiple evaluation metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), to assess the model's performance.

4. Roadblocks

- **Data Preprocessing**: One roadblock involved converting columns with string values like '43,865.90' into numeric types, which required additional preprocessing steps.

- **ID Correlation Tracking**: Another challenge was to ensure that the ID relationships were handled properly when splitting the data for training and testing.

- **Model Evaluation on Unseen Data**: Validating the model on recent data presented challenges with ensuring robust performance over time.

5. Data Exploration and Preprocessing

During the first phase, I have explored the dataset and preprocessed it for model development:

Loading the dataset: We loaded the dataset and examined the first few rows to understand the structure.

Handling missing values: Missing values were filled using the mean of the respective columns.

Standardizing data: Features were standardized using StandardScaler to ensure consistent scales across variables.

Exploratory Data Analysis (EDA): We visualized the target variable distribution (Amazon_Price) and explored relationships between Amazon_Price and other features using scatter plots. A correlation matrix was generated to highlight key relationships between variables.

6. Model Development

I have developed a linear regression model to predict Amazon prices:

Splitting the Data: The dataset was split into training and testing sets using an 80-20 split.

Training the Model: We initialized a Linear Regression model and trained it on the training data.

Evaluation: The model was evaluated on the test data using MAE, MSE, and RMSE. These metrics helped assess the model's accuracy in predicting Amazon prices.

7. Model Validation and Testing

I have focused on model validation and further testing:

Cross-Validation: We performed 5-fold cross-validation to assess the model's performance more robustly. This helped identify potential overfitting or underfitting issues.

Testing on Unseen Data: A portion of recent data was used to validate the model on unseen data, giving a more realistic evaluation of model performance on future data.

Feature Importance Analysis: Coefficients from the linear regression model were analyzed to understand the importance of different features. A bar plot was used to visualize the contribution of each feature.

8. Documentation and Submission

Finally, the project was documented and prepared for submission:

Project Summary: A comprehensive summary of the project was written, covering the introduction, challenges, solutions, and key outcomes.

Jupyter Notebook Compilation: All code from Weeks 1 to 3 was compiled into a single Jupyter Notebook, with explanations provided for each code block.

GitHub Submission: The Jupyter Notebook and a README file were uploaded to GitHub, making it easily accessible for review.

GitHub Repository Link: GitHub Repository *(Replace with your actual repository link)*

Google Doc Link: Google Doc Summary *(Replace with your actual Google Doc link)*

9. Conclusion

This project successfully applied a linear regression model to predict Amazon prices based on market data. The model performed reasonably well, with key features like Crude Oil Price and Nasdaq 100 Price playing important roles in predictions. Cross-validation and testing on recent data confirmed the model's robustness, although improvements in feature selection or model choice could be explored for future work.