DEPARTMENT OF
COMPUTER AND
INFORMATION SYSTEMS
ENGINEERING NED

OPEN ENDED LAB

MACHINE LEARNING (CS-324)

PREPARED BY

Faizan Soomro CS-21090
Subhan Sohair CS-21141
Shehryar Nasir CS-21143

PRESENTED TO

Ms. Mahnoor Malik

Cryptocurrency Price Prediction System

Introduction

In the rapidly evolving world of cryptocurrency, accurate price prediction is crucial for investors and traders. Our project aims to develop a robust cryptocurrency prediction model that leverages historical market data to predict future prices. By using advanced machine learning techniques, we provide insights that can help in making informed investment decisions. This project encompasses data collection, preprocessing, exploratory data analysis, feature engineering, model building, and evaluation, ultimately resulting in a user- friendly interface for real-time analysis.

Data Collection

For our cryptocurrency prediction system, we utilized data from the yfinance API, which provides dynamically updating data for various cryptocurrencies. Our dataset includes the following labels: Open, High, Low, Close, Adj Close, and Volume. The dataset was chosen for two main reasons:

1. **Dynamic Updates:** The yfinance API ensures that our dataset is always up-to-date, reflecting the latest market trends and prices.
2. **Preprocessed Data:** The data from yfinance is preprocessed and has no null values, allowing us to focus on analysis and model building without extensive data cleaning.

Data Collection

Since the yfinance data is already clean and has no missing values, our preprocessing efforts were minimal. We performed scaling using the StandardScaler from a built-in library to normalize the data, ensuring that our models perform optimally. This step is crucial to ensure that features are on the same scale, preventing any single feature from dominating the model.

Exploratory Data Analysis (EDA)

Our EDA revealed several key insights about the cryptocurrency market:

- **Trading Volume:** The heat map shows the overall trading volume for each year for a cryptocurrency. For example, for Bitcoin (BTC), the heat map indicated that the highest trading volume occurred between January 2021 and May 2021, coinciding with a significant market bull run.
- **Price Trends:** Visualizations such as open vs. close charts and high vs. low charts helped us understand daily price movements and trends over time.
- **Volatility Analysis:** Candle charts provided insights into the volatility and price movements within specific time frames.

We used various visualizations to gain insights into the data:

- **Open vs. Close Chart:** Showed the relationship between opening and closing prices.
- **High vs. Low Chart:** Highlighted the daily price range.
- **Basic Line Chart for Yearly Analysis:** Provided an overview of price trends over each year.
- **Candle Chart:** Illustrated price movements and volatility.
- **Volume Heatmap:** Highlighted periods of high trading activity.

Feature Engineering

To enhance our predictive model, we engineered several new features that are commonly used in financial analysis:

- **MACD (Moving Average Convergence Divergence):** Used to identify changes in the strength, direction, momentum, and duration of a trend.
- **RSI (Relative Strength Index):** A momentum oscillator that measures the speed and change of price movements.
- **SMA (Simple Moving Average):** The average of a selected range of prices, typically closing prices, by the number of periods in that range.
- **EMA (Exponential Moving Average):** Similar to the SMA, but gives more weight to recent prices, making it more responsive to new information.

These features were chosen because they capture essential market dynamics and trends, which are crucial for accurate price prediction. Our final model used the following features: Close, MACD, RSI, SMA, and EMA.

Model Building

We chose three machine learning algorithms for our predictive models: Linear Regression, Support Vector Machine (SVM) Regressor, and Random Forest Regressor, each offering unique strengths:

1. **Linear Regression:**
    - **Selection:** Provides a baseline for predictive performance with simplicity and interpretability.
    - **Benefits:** Helps understand the linear relationship between input features and the target variable.
2. **Support Vector Machine (SVM) Regressor:**
    - **Selection:** Handles non-linear relationships through kernel functions, suitable for complex financial data.
    - **Benefits:** Finds the optimal hyperplane to minimize prediction error, capturing intricate data patterns effectively.
3. **Random Forest Regressor:**
    - **Selection:** An ensemble method that reduces overfitting and improves generalization.
    - **Benefits:** Handles large datasets with high dimensionality, balancing bias and variance, making it reliable for time-series predictions.

Using the sklearn library facilitated quick implementation and testing, while implementing these models without packages was challenging but deepened our understanding of their mechanics.

Model Evaluation

Model	MSE	RMSE	R^2	MAE
SVR (with package)	0.015999	0.126489	0.979852	0.086614
SVR (w/o package)	0.031633	0.177856	0.960165	0.134469
RF (with package)	0.060916	0.246812	0.923288	0.200458
RF (w/o package)	0.060909	0.246797	0.923297	0.197739
LR (with package)	0.016220	0.127359	0.979574	0.085073
LR (w/o package)	0.016220	0.127359	0.979574	0.085073

Best Performing Model

The Linear Regression (LR) performed the best among the implemented models, achieving the highest R^2 score and lowest MSE and RMSE. This indicates that the LR model was most effective at capturing the relationships in the data and making accurate predictions. The simplicity and interpretability of LR in modeling linear relationships likely contributed to its superior performance.

Conclusion

Our analysis revealed that incorporating technical indicators such as MACD, RSI, SMA, and EMA significantly improved the prediction accuracy of cryptocurrency prices. The Linear Regression emerged as the best-performing model, suggesting its effectiveness in financial time series prediction. However, the models without package implementations highlighted the importance of using well-established libraries for practical applications.

Limitations:

- The model’s accuracy is highly dependent on the quality and recency of the data.
- Market anomalies and external factors (e.g., regulatory changes, macroeconomic events) are not accounted for.

Suggestions for Improvement:

- Incorporate additional features such as social media sentiment or macroeconomic indicators.
- Explore more advanced machine learning techniques like LSTM or GRU for time series prediction.

User Interface

We developed a user-friendly interface using Streamlit, allowing users to interact with the dataset and view the results of our analysis and models. Users can select the cryptocurrency of their choice and analyze graphs for a specific time period, enhancing the usability and accessibility of our predictive model.