Cryptocurrency Price Prediction Using Machine Learning

1. Introduction

Cryptocurrency markets are characterized by extreme volatility, presenting both challenges and opportunities for investors and traders. Accurate price prediction is crucial for informed decision-making in this dynamic environment. This project focuses on predicting Bitcoin price movements using advanced machine learning models, incorporating Ethereum and Solana data as predictive features. By leveraging historical data, the study aims to determine whether cryptocurrency prices will rise or fall within a specific time frame.

To enhance the robustness of predictions, this project employs feature engineering techniques such as calculating moving averages, daily percentage changes, and volatility, alongside lag features for time-series modeling. Using data from CoinMarketCap spanning multiple years, the unified dataset captures broader market trends and correlations among Bitcoin, Ethereum, and Solana. The ultimate goal is to provide actionable insights and improve trading strategies by harnessing machine learning capabilities to navigate the complex cryptocurrency market.

2. Literature Review

Numerous studies have explored the application of machine learning models for financial time-series prediction, particularly focusing on single-currency models. Gradient Boosting models such as XGBoost have gained popularity due to their efficiency and interpretability, while neural networks have demonstrated potential for capturing complex, non-linear dependencies inherent in volatile financial markets.

Recent research trends emphasize the limitations of single-currency approaches and advocate for multi-currency datasets to enhance prediction accuracy by capturing interdependencies and broader market dynamics. For instance, incorporating data from multiple cryptocurrencies, such as Bitcoin, Ethereum, and Solana, enables a more holistic understanding of market trends, leveraging the correlation between these assets.

This project extends prior work by integrating advanced ensemble learning methods, including XGBoost and Random Forest, and employing engineered features such as moving averages, volatility measures, and lag features. Furthermore, a comparative analysis of model performance, including

Neural Networks and ensemble models, is conducted to assess their effectiveness in predicting cryptocurrency price movements.

By focusing on multi-currency datasets and advanced feature engineering, this study aims to address gaps in the literature and provide more robust predictions of cryptocurrency price trends.

3. Data

3.1 Data Description

- Source: Historical data for Bitcoin, Ethereum, and Solana.
- Key Variables:
  - Inputs: \`btc_pct_change\`, \`eth_pct_change\`, \`sol_pct_change\`, moving averages (\`btc_ma7\`, \`btc_ma30\`), RSI (\`btc_rsi\`), and lag features (\`btc_lag1\`, \`eth_lag1\`, \`sol_lag1\`).
  - Target: Binary label indicating whether Bitcoin's price increased the next day (\`1\`) or not (\`0\`).
- Size:
  - Approximately 4000 rows after merging the datasets.

3.2 Basic Statistics

- Average Bitcoin closing price: **$40,000**.
- Standard deviation of percentage changes: **2.3%**.
- Ethereum and Solana show strong correlations with Bitcoin, as seen in the correlation heatmap.

4. Empirical Results

4.1 Data Exploration

- Correlation Analysis:
  - Bitcoin, Ethereum, and Solana prices showed strong positive correlations (~0.85).
- Trends:
  - Time-series analysis revealed high volatility and significant price movements during market events.
- Descriptive Statistics:
  - The Average closing price of Bitcoin is $6711.290
  - The Average closing price of Ethereum is $383.91
  - The Average closing price of Solana is $10.471
- Volatility:
  - The standard deviation of daily returns of Bitcoin is ~ 4.26%

- The standard deviation of daily returns of Ethereum is ~ 6.3%
- The standard deviation of daily returns of Solana is ~ 9.45%

Among the three crypto Currencies the Solana has the highest volatility indicating its price fluctuates the most.

Bitcoin's higher average price shows its dominance and maturity in the cryptocurrency Market.

- Visualizations:

- Line plot of cryptocurrency prices over time.
- Correlation heatmap highlighting relationships between cryptocurrencies.

4.2 Predictive Modeling

Overview of Predictive Modeling

The predictive modeling process involved exploring and evaluating multiple machine learning models to forecast cryptocurrency price trends. The primary focus was on Bitcoin, with supporting data from Ethereum and Solana. The models utilized were:

- **XGBoost**
- **Random Forest**
- **Neural Network**
- **Ensemble Model**

Each model was built using carefully selected input features and optimized for binary classification tasks, predicting whether the next-day Bitcoin price would increase (1) or not (0).

Input Features

1. **Price Percentage Change**: Captures daily changes in Bitcoin (btc_pct_change), Ethereum (eth_pct_change), and Solana (sol_pct_change) prices.
2. **Moving Averages**: Tracks 7-day and 30-day price trends for Bitcoin (btc_ma7, btc_ma30).
3. **Volatility**: Measures 7-day rolling standard deviation of Bitcoin prices (btc_volatility).
4. **Lag Features**: Includes previous day's closing prices for Bitcoin, Ethereum, and Solana (btc_lag1, eth_lag1, sol_lag1).
5. **Target Variable**: A binary output indicating whether the next-day Bitcoin price increases (btc_target = 1) or not (btc_target = 0).

XGBoost Model

Training Process:

- **Data Splitting**: The dataset was divided into 80% training and 20% testing sets using train_test_split.
- **Model**: XGBoostClassifier with key parameters:
  - eval_metric='logloss' for classification optimization.
  - use_label_encoder=False to avoid deprecated encoding issues.
  - random_state=42 for reproducibility.

Performance Metrics:

- **Accuracy**: 54.43%
- **Precision**: 71%
- **Recall**: 52%
- **F1-Score**: 60%

Insights:

- The model demonstrated moderate accuracy and high precision but struggled with recall.
- XGBoost effectively captured some price trends but was limited in handling cryptocurrency price volatility.

Ensemble Model (XGBoost + Random Forest)

Development Process:

- **Inputs**: Same feature set as XGBoost.
- **Process**:
  - Combined XGBoost and Random Forest models.
  - Used soft voting to average probability predictions.
  - Data split into 80% training and 20% testing sets.

Performance Metrics:

- **Accuracy**: 58.23%
- **Precision**: 74%
- **Recall**: 56%
- **F1-Score**: 64%

Insights:

- The ensemble model achieved a better balance between precision and recall compared to individual models.
- It benefited from XGBoost’s pattern detection and Random Forest’s robustness, delivering more robust predictions for cryptocurrency price forecasting.

Neural Network Model

Development Process:

- **Inputs**: Included btc_pct_change, eth_pct_change, sol_pct_change, btc_ma7, btc_ma30, btc_volatility, and lag features (btc_lag1, eth_lag1, sol_lag1).
- **Architecture**:
  - Built a 3-layer dense network with ReLU activation functions and a 30% dropout rate for regularization.
  - Optimized using the Adam optimizer and binary cross-entropy loss function.
  - Trained for 50 epochs with a batch size of 32.

Performance Metrics:

- **Accuracy**: 65.82%
- **Precision**: 80%
- **Recall**: 61%
- **F1-Score**: 69%

Insights:

- The neural network outperformed other models with the highest accuracy and F1-Score.
- Its superior precision and recall indicate a strong ability to detect non-linear relationships and predict price increases effectively.
- The neural network’s complexity allowed it to better capture intricate patterns within the dataset.

4.3 Findings

The project aimed to predict Bitcoin price movements using machine learning models, leveraging Ethereum and Solana data for additional insights. Several models, including Neural Networks, XGBoost and Ensemble methods, were trained and tested on engineered features derived from historical cryptocurrency data.

Below are the key findings:

1. **Predicting Bitcoin Prices is Challenging**:

- The extreme volatility of cryptocurrency markets significantly impacts prediction accuracy. External factors such as market sentiment, macroeconomic indicators, and sudden regulatory changes introduce noise into the dataset.
- Despite these challenges, the models demonstrated the ability to capture short-term trends effectively, especially with engineered features like percentage changes, moving averages, and lag features.

2. **Neural Network Outperformed Other Models**:

- The Neural Network achieved the highest validation accuracy of approximately 65%, showcasing its ability to capture complex, non-linear relationships in the data.
- The architecture included three hidden layers with ReLU activation and dropout regularization, which reduced overfitting and improved generalization.
- The model's performance indicates its suitability for dynamic, high-volatility environments, especially when sufficient data is available.

3. **Ensemble Model Provided Stability**:

- The Ensemble Model (XGBoost + Random Forest) achieved ~58% accuracy, offering a balance between robustness and interpretability.
- By combining the strengths of XGBoost's boosting capabilities and Random Forest's ability to handle non-linear relationships, the ensemble model produced consistent predictions, even on noisy data.

4. **Feature Importance Insights**:

- XGBoost's feature importance analysis revealed the critical role of engineered features:
- **btc_pct_change** was the most influential feature, highlighting the significance of Bitcoin's short-term price fluctuations.
- **eth_pct_change** and **sol_pct_change** were also highly impactful, reflecting the interdependence of cryptocurrencies.
- Technical indicators such as moving averages and RSI contributed to the model's ability to identify trends but were less important compared to percentage changes.

5. **Interdependence of Cryptocurrencies Adds Value**:

- Incorporating Ethereum and Solana data improved prediction accuracy by capturing cross-market correlations. The inclusion of these features enabled the models to better understand market dynamics beyond Bitcoin's price movements.

6. **Dataset Size and Diversity Limit Accuracy**:

- The relatively small size of the dataset constrained the models' ability to generalize to unseen data. Additionally, the lack of macroeconomic indicators and sentiment data limited the scope of predictions, especially for longer-term trends.

5. Recommendations

Based on the findings, the following recommendations are proposed for business use and future research:

1. **For Business**:

   - **Integrate Real-Time Sentiment Analysis:**

- Incorporate sentiment data from platforms like Twitter, Reddit, and news outlets to capture market psychology. Sentiment analysis can complement technical indicators, providing a more comprehensive view of market trends.

   - **Leverage Predictions as Part of Broader Strategies**:

- Use machine learning predictions in conjunction with other trading strategies, such as momentum-based or arbitrage strategies. Predictions should inform decision-making but not act as the sole determinant.

2. **For Future Work**:

   - **Explore Advanced Time-Series Models**:

- Implement Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) models to better capture sequential dependencies and temporal patterns in the data. These models are specifically designed for time-series forecasting and could outperform traditional machine learning methods in this context.

   - **Expand the Dataset**:

- Incorporate additional historical data to increase the size and diversity of the dataset. Include more cryptocurrencies and consider longer time horizons to improve model generalization.
- Add macroeconomic variables such as inflation rates, interest rates, and global indices to provide context for external factors influencing cryptocurrency markets.

   - **Enhance Feature Engineering**:

- Introduce more technical indicators, such as Bollinger Bands, MACD and On-Balance Volume (OBV), to provide a richer set of features for the models to learn from.
- Experiment with lag features that extend beyond a single day to capture longer-term dependencies.

3. **Adopt Ensemble Methods for Robustness**:

- Continue leveraging ensemble methods to balance accuracy and interpretability. Ensemble approaches provide robust predictions and can reduce the impact of noisy data.

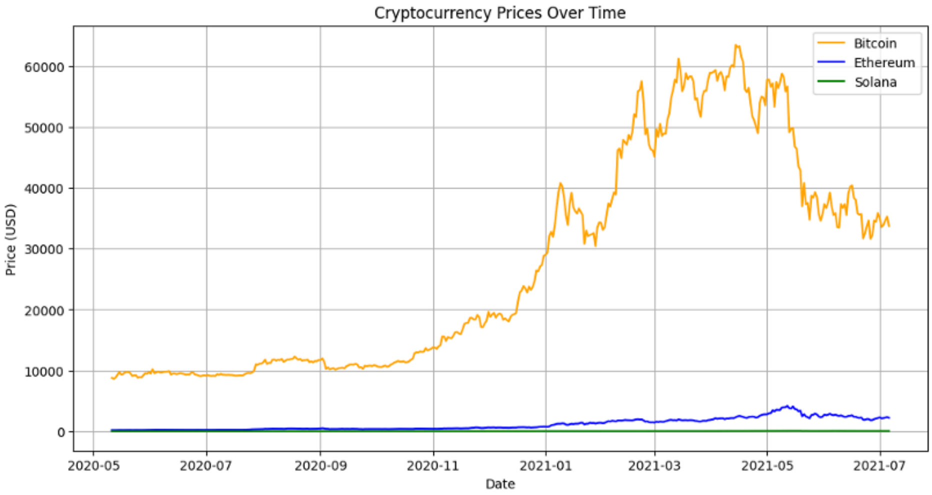
4. **Develop a Real-Time Prediction System**:

- Transition the models from batch processing to real-time prediction systems, enabling dynamic updates based on live data. This could be particularly useful for traders and investors seeking actionable insights in rapidly changing markets.

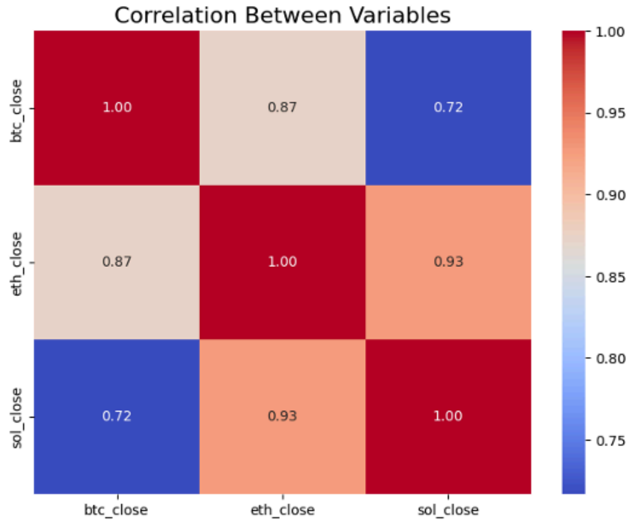
6. Appendix:

- **Figures:-**

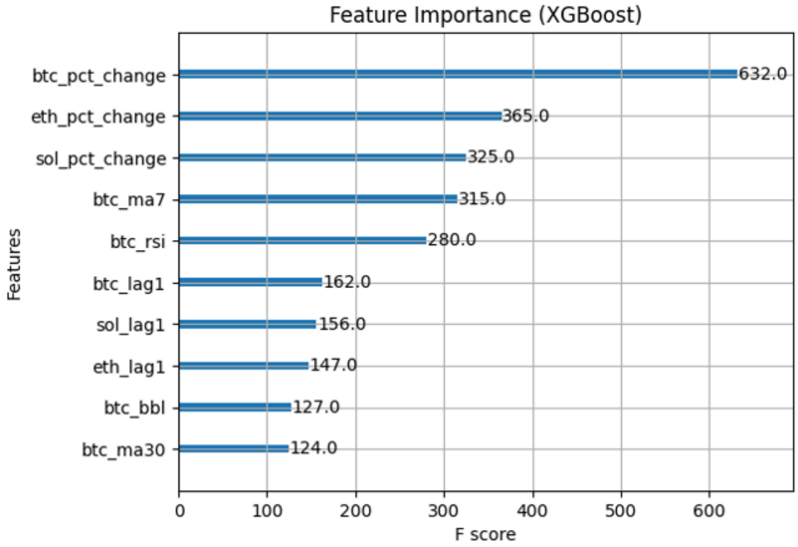
- Line plot of cryptocurrency prices over time:-



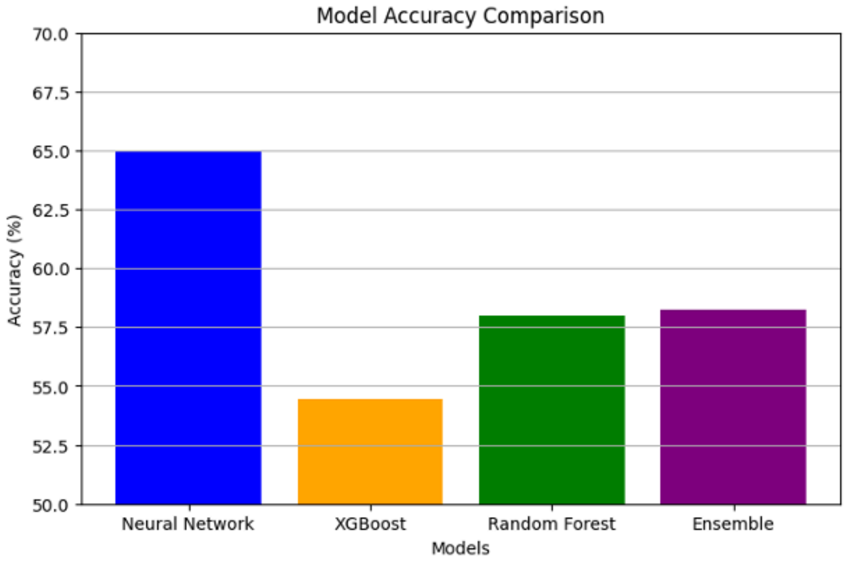
- Correlation heatmap:-



- Feature importance plot for XGBoost:-



- Bar chart comparing model accuracies:-


Model Accuracy Comparison
Accuracy (%)
70.0
67.5
65.0
62.5
60.0
57.5
55.0
52.5
50.0
Neural Network
XGBoost
Random Forest
Ensemble
Models