Bitcoin Price Movement Prediction using Machine Learning

Research paper made by Dejan Simonovski 211039 & Elena Boshkoska 213040 for UKIM FCSE Skopje

1. Introduction

Financial markets, and especially cryptocurrencies such as Bitcoin, are highly volatile, complex, and influenced by a wide variety of factors ranging from global macroeconomics to investor sentiment.

In this project, we attempt to predict short-term Bitcoin price movements using supervised machine learning. We implement an **XGBoost classification** model combined with time-series cross-validation and feature engineering techniques. The goal is to understand how well such models can perform in this unpredictable environment and to learn which features contribute most to predictive performance.

2. Dataset

The dataset is sourced from Kaggle ('Bitcoin Historical Datasets 2018–2024'), covering daily Bitcoin price data from 2018 until 2025(the present), updated every day to include more data. It includes open, high, low, close prices, and trading volume for each day. After loading, timestamps were converted into datetime format, and the data was sorted by time. Columns irrelevant to the prediction task, such as ignored values, were removed. This time-ordered dataset forms the basis for all preprocessing and feature engineering steps.

3. Preprocessing

Preprocessing is crucial when dealing with financial time series. Missing values were handled, non-essential columns were dropped, and data was aligned sequentially to preserve the temporal structure. We also scaled features using robust normalization to limit the influence of extreme values, which are common in cryptocurrency markets. Preprocessing ensures that our model receives consistent, high-quality input data without information leakage across time.

4. Feature Engineering

A wide range of features was created to capture different aspects of Bitcoin's price behavior. These include:

- **Returns and Volatility**: Daily, 3-day, 7-day, and 14-day returns were calculated to capture short-term price changes. Volatility measures (rolling standard deviations) help the model identify periods of high uncertainty.
- **Moving Averages and EMA**: Simple and exponential moving averages (5, 10, 20, 50-day windows) smooth out noise and indicate long-term trends. Ratios between short and long windows (e.g., MA5/MA20) highlight trend shifts.
- **Bollinger Bands**: Upper and lower bands with a 20-day window provide insight into price

momentum and market extremes. Derived features include band width and relative price position.

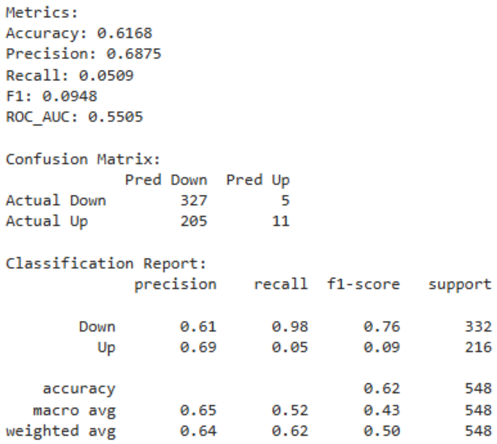
- **MACD and RSI**: The Moving Average Convergence Divergence (MACD) and Relative Strength Index (RSI) capture trend-following and momentum aspects of the market. MACD crossovers often signal buy/sell points, while RSI highlights overbought/oversold conditions.
- **ATR and Volume Indicators**: Average True Range (ATR) measures market volatility. Volume-based features such as volume change, moving averages, and On-Balance Volume (OBV) capture shifts in trading activity.
- **Price Ratios**: Ratios such as Close/Open and High/Low provide additional relative measures of price behavior.
- **Time Features**: Day of week, month, and indicators for month start/end help capture cyclical patterns in market activity.
- **Lagged Features**: Lagging returns, RSI, volume, and other signals up to 10 days provide temporal dependencies for the model to exploit.
- **Rolling Statistics**: Rolling means and standard deviations of returns and volume provide local trends and volatility patterns.

5. Model and Evaluation

The model of choice was **XGBoost**, a gradient boosting algorithm well-suited for structured data and capable of handling nonlinear relationships. We applied a time-series split cross-validation to avoid look-ahead bias and used **GridSearchCV** for hyperparameter tuning. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC AUC. Confusion matrices and ROC curves were plotted to visualize model performance.

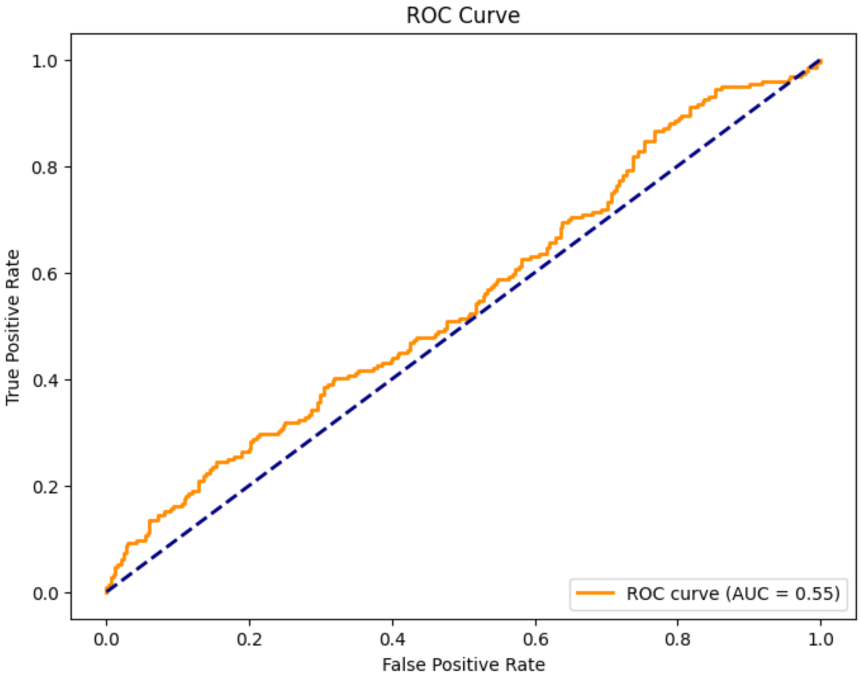
6. Results and Plots

The model achieved moderate accuracy, with strong recall for downward movements (around 98%) but poor recall for upward movements (approximately 5%). This means the classifier is heavily biased toward predicting declines, reflecting the imbalance and difficulty in distinguishing short-term bullish signals.



```
Metrics:
Accuracy: 0.6168
Precision: 0.6875
Recall: 0.0509
F1: 0.0948
ROC_AUC: 0.5505

Confusion Matrix:
             Pred Down  Pred Up
Actual Down        327        5
Actual Up          205       11

Classification Report:
              precision    recall  f1-score   support

        Down       0.61      0.98      0.76       332
          Up       0.69      0.05      0.09       216

    accuracy                           0.62       548
   macro avg       0.65      0.52      0.43       548
weighted avg       0.64      0.62      0.50       548
```

ROC curves indicated that the model performs only slightly above random chance for positive movements. Confusion matrices further confirmed the skewed predictions, with the majority of positive cases misclassified.



These results highlight a **critical challenge**: financial markets are influenced by countless hidden variables, news events, and investor behavior, making them inherently unpredictable. While our engineered features capture technical indicators, they are insufficient to fully explain or predict Bitcoin's highly volatile nature.

7. Conclusion

This project demonstrates the process of preparing, engineering, and modeling financial time series data with machine learning. Through extensive feature creation and careful preprocessing, we built an XGBoost model to predict Bitcoin's short-term price movements. The model, however, struggled to achieve balanced accuracy and failed to generalize bullish signals.

This underlines **an important lesson**: predicting markets is an exceptionally challenging task due to their stochastic and non-stationary nature. Nonetheless, the exercise provides valuable insights into data-driven analysis and highlights directions for future research, such as including sentiment data, news analysis, or alternative modeling approaches.