CoinCast: Bitcoin Price Forecasting Using CRISP-DM Methodology

In this project, CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is implemented to develop a robust solution for forecasting Bitcoin prices. CRISP-DM, a widely adopted framework in data science, guided our workflow through its structured stages from understanding business objectives to deploying the final solution. By leveraging advanced machine learning and deep learning models, we systematically applied CRISP-DM to address the challenges of Bitcoin price volatility and produce accurate 60-day forecasts.

The following sections outline how each stage of CRISP-DM was utilized in our project:

1. Business Understanding

The primary objective of this project was to develop a robust and accurate forecasting model for Bitcoin price predictions to support traders, investors, and financial analysts in making informed decisions. Given the high volatility and non-linear behavior of Bitcoin prices, traditional models struggle to capture the complex temporal patterns of cryptocurrency data. To address this, the project leveraged advanced machine learning and deep learning models to forecast Bitcoin closing prices, identifying the most suitable approach for achieving reliable predictions.

The business goals:

1. To forecast Bitcoin prices for the next 60 days.
2. To evaluate multiple forecasting models (ARIMA, Random Forest, XGBoost, and LSTM) and identify the best-performing model.
3. To ensure usability by deploying the solution via an interactive Gradio-based UI hosted on Hugging Face Spaces for broader accessibility.

2. Data Understanding

The dataset used for this project was sourced from Kaggle's "**Bitcoin Historical Market Data**" and consisted of minute-level time-series records from January 2012 to the present. Key features include:

- OHLC (Open, High, Low, Close) Prices: Price points during each interval.
- Volume: Trading volume during the interval.
- Timestamp: Date and time of the recorded data.

The closing price was chosen as the target variable for forecasting, as it represents the final price of Bitcoin within each interval and serves as a critical indicator of market sentiment.

Key observations during Exploratory Data Analysis (EDA):

- Significant volatility and upward price trends were observed.
- Outliers were identified using the Interquartile Range (IQR).

- Strong correlations were found among OHLC prices, while trading volume showed moderate correlation with price movements.

3. Data Preparation

To prepare the dataset for model implementation, the following steps were performed:

1. **Data Cleaning**:
    - Converted timestamps to readable date formats.
    - Aggregated minute-level data to a daily level by calculating mean OHLC prices and trading volumes.
    - Detected and removed outliers using IQR.
2. **Feature Engineering**:
    - Created new features such as year, month and day of the week from the timestamp to identify seasonal patterns.
3. **Data Scaling**:
    - Used MinMaxScaler to normalize the dataset, scaling values between 0 and 1 to improve model convergence for LSTM.
4. **Train-Test Split**:
    - Split the data into training (75%) and testing (25%) sets for model evaluation.
5. **Time-Series Analysis**:
    - Verified that the data was non-stationary using ADF and KPSS tests.
    - Applied transformations like differencing and log scaling to stabilize variance and trends.

4. Modeling

The following models were implemented to forecast Bitcoin prices:

1. **ARIMA (Auto-Regressive Integrated Moving Average)**:
    - Served as the baseline model to capture linear trends.
    - Parameters (p, d, q) were optimized to improve accuracy.
2. **Random Forest Regressor**:
    - Captured non-linear relationships by tuning the number of estimators and maximum depth.
3. **Gradient Boosting (XGBoost)**:
    - Fine-tuned hyperparameters like learning rate and tree depth for improved accuracy.
4. **Long Short-Term Memory (LSTM)**:
    - LSTM, a deep learning model designed for sequential data, was fine-tuned using Keras Tuner for hyperparameter optimization.
    - Multiple LSTM layers with dropout regularization and optimized learning rates improved the model's ability to capture complex temporal dependencies.

5. Evaluation

The models were evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **R-squared (R^2)**

Model	MSE	RMSE	R-Squared
ARIMA	347,173,971.42	18,630.65	0.72
Random Forest	37,444,453.09	6116.64	0.85
Gradient Boosting	52,355,774.68	7234.95	0.81
LSTM	14,347,100.00	3787.76	0.96

The **LSTM model** achieved the lowest RMSE and MAE, demonstrating superior performance in capturing both long-term trends and short-term fluctuations in Bitcoin prices.

6. Deployment

The final LSTM model, alongside other models, was deployed using the following pipeline:

1. Designed an interactive Gradio-based user interface allowing users to:
    - Select forecasting models (ARIMA, Random Forest, XGBoost, LSTM).
    - Specify the desired forecast duration (e.g., 30 or 60 days).
    - Visualize forecasts and display evaluation metrics interactively.
2. Automated the deployment process using GitHub Actions, which pushes updates to the Hugging Face repository.
3. Hosted the Gradio UI and trained models on Hugging Face Spaces, providing scalability and public access.

7. Conclusion

This project successfully implemented an end-to-end pipeline for Bitcoin price forecasting, integrating advanced models like ARIMA, Random Forest, XGBoost and LSTM. Through rigorous model evaluation, LSTM emerged as the best-performing model, offering reliable predictions for the next 60 days. The deployment of an interactive UI on Hugging Face Spaces ensures accessibility for both technical and non-technical users, bridging the gap between machine learning techniques and practical usability.