Bitcoin Price Prediction

Introduction

This report details the process and results of predicting Bitcoin prices for December 2017 using chainlet data and historical price information. The goal was to build a predictive model, evaluate its performance, and explain its behavior using SHAP analysis.

Data Preprocessing

- **Data Loading:** The datasets pricedBitcoin2009-2018.csv, AmoChainletsInTime.txt, and OccChainletsInTime.txt were loaded using pandas.
- **Date Handling:** The 'date' column in pricedBitcoin2009-2018.csv was converted to datetime objects. Year and day columns were extracted.
- **Data Merging:** The chainlet data was merged with the price data using 'year' and 'day' as keys.
- **Lagged Price Features:** Lagged price features ('price_lag1' to 'price_lag5') were created to represent the price from the previous five days.
- **Data Filtering:** The data was filtered to include only the year 2017.
- **Feature Selection:** Specific chainlet features ranging from '1:1_amo' to '1:5_amo' and their corresponding occurrence features ('1:1_occ' to '1:5_occ') were selected. Other columns were dropped.
- **Data Splitting:** The data was split into training (January-November 2017) and testing (December 2017) sets.

Model Selection and Training

- A Ridge Regression model was chosen for its simplicity and effectiveness in handling linear relationships with regularization.
- The model was trained using the training data, with an alpha value of 1.0.

Evaluation and Results

- The model's performance was evaluated using Root Mean Squared Error (RMSE).
- The RMSE value was: 1164.39
- The predictions for each day of December 2017 were saved to a CSV file (price_predictions_ridge.csv).
- A plot comparing the actual and predicted Bitcoin prices was generated.

Explanation Findings (SHAP Analysis)

- SHAP decision plots were used to explain the model's predictions for the first five days of December 2017. The plots showed the progression of the model's prediction as features were considered, highlighting the impact of each feature.
- **On 2017-12-01:** The predicted price was 10262.94 USD (actual: 10883.91 USD, difference: -620.97 USD). The most influential positive feature was 'price_lag1' (+7452.28 USD), followed by 'price_lag5' (+1328.94 USD) and '1:4_amo' (+12.08 USD). Negative influences included 'price_lag4' (-1115.69 USD), 'price_lag3' (-735.43 USD), and '1:5_occ' (-51.25 USD). '1:1_amo' (+6.78 USD) and '1:2_occ' (+90.15 USD) also contributed positively.
- **On 2017-12-02:** The predicted price was 11040.01 USD (actual: 11071.37 USD, difference: -31.36 USD). 'price_lag1' (+8189.43 USD) was the strongest positive contributor, along with 'price_lag5' (+1415.82 USD) and 'price_lag2' (+534.71 USD). Negative influences included 'price_lag4' (-1152.67 USD) and 'price_lag3' (-728.04 USD). '1:4_amo' (+24.18 USD) had a noticeable positive impact, while '1:5_amo' (-12.34 USD) contributed negatively.
- **On 2017-12-03:** The predicted price was 11325.47 USD (actual: 11332.62 USD, difference: -7.15 USD). 'price_lag1' (+8377.04 USD) remained the most influential positive feature, followed by 'price_lag5' (+1462.70 USD) and 'price_lag2' (+587.42 USD). Negative impacts came from 'price_lag4' (-1141.11 USD) and 'price_lag3' (-755.10 USD). '1:4_amo' (+17.85 USD) contributed positively, while '1:5_amo' (-19.17 USD) had a negative effect.
- **On 2017-12-04:** The predicted price was 11438.64 USD (actual: 11584.83 USD, difference: -146.19 USD). 'price_lag1' (+8638.51 USD) was the largest positive contributor, along with 'price_lag5' (+1448.05 USD) and 'price_lag2' (+600.84 USD). Negative influences included 'price_lag4' (-1183.44 USD) and 'price_lag3' (-829.43 USD). '1:4_occ' (+39.71 USD) showed a positive impact, while '1:5_occ' (-114.66 USD) had a significant negative effect.
- **On 2017-12-05:** The predicted price was 11639.89 USD (actual: 11878.43 USD, difference: -238.54 USD). 'price_lag1' (+8890.93 USD) was the most influential positive feature, followed by 'price_lag5' (+1501.70 USD) and 'price_lag2' (+619.54 USD). Negative influences included 'price_lag4' (-1299.73 USD) and 'price_lag3' (-848.35 USD). '1:2_occ' (+102.27 USD) and '1:4_occ' (+58.62 USD) contributed positively, while '1:5_occ' (-138.74 USD) had a notable negative impact.

Challenges and Discussion

- **Data Integration Complexity:**
    - Merging the chainlet data with the price data required careful handling of date alignment. The "day of year" representation, while necessary, introduced potential complexities in ensuring accurate synchronization between datasets.
    - The different file formats added a layer of complexity to the initial data loading and parsing.
- **Feature Selection and Dimensionality Reduction:**
    - The chainlet datasets contained many features which posed a challenge in terms of feature selection and potential overfitting.
    - I chose to focus on the '1:1' to '1:5' chainlets for both amount and occurrence to capture relevant short-range interactions. Exploring the impact of other chainlet types or using more sophisticated feature selection techniques could be beneficial.
    - Dropping the year and day features increased the difficulty of the time series predictions.
- **Lagged Feature Creation:**
    - Creating the lagged price features introduced missing values requiring careful handling.
    - The choice to include five lagged price features was based on exploring a short-term historical window to capture recent price trends. The impact of considering a longer or shorter window of lagged prices could be further investigated.
- **Time Series Data Splitting:**
    - Splitting the time series data into training and testing sets required careful consideration to avoid data leakage. We chose a chronological split (January-November 2017 for training, December 2017 for testing) to maintain the temporal order of the data.
- **Model Selection and Parameter Tuning:**
    - While Ridge Regression was chosen for its simplicity and regularization, exploring other models could potentially yield better results.
    - The alpha parameter in Ridge Regression was set to 1.0, but further tuning could have optimized the model's performance.
- **Potential Data Noise:**
    - The chainlet data, while potentially informative, might contain noise or irrelevant information that could negatively impact the model's predictions.

Conclusion

The Ridge Regression model demonstrated moderate prediction success in predicting Bitcoin prices for December 2017. The SHAP analysis provided valuable insights into the model's decision-making process, highlighting the significant influence of recent price history and the varying impacts of specific chainlet interactions.

CSV File

- A separate CSV file (price_predictions_ridge.csv) containing the predicted prices for each day of December 2017 is included with this report.

GitHub Repository

- https://github.com/Jose-Ayala/Bitcoin-Price-Prediction