Team 2
Client Pitch

Predicting Realized Volatility


AXEL LIBRATA


BARRY YE


MARVEL NELWAN


WILLIAM MOOG


WYUN NG

UNSW SYDNEY

optiver

Executive Summary

optiver

Problem

How might we develop a predictive short-term stock volatility model that outperforms the Naïve model through Ensemble Learning?

Strategy

Hybrid Model – Bayesian Averaging

Exponentially Weighted Moving Averages (EWMA)

Light Gradient-Boosted Machine (LightGBM)

Risks

Heteroskedasticity

Computational Resources

Data Quality

Impact

R^2: 0.878

RMSPE: 0.19

Runtime: 16 Seconds

Project Overview

optiver

How might we develop a predictive short-term stock volatility model that outperforms the Naïve model through Ensemble Learning?

1. **Market Efficiency**
   Identify mispricing in options and other derivatives to hedge for profit
2. **Portfolio Optimization**
   Investors can create more efficient portfolios that generate higher returns
3. **Competition**
   Stay ahead of competitors who are relying on less sophisticated models

Project Overview

optiver

How might we develop a predictive short-term stock volatility model that outperforms the Naïve model through Ensemble Learning?

Aim

Develop a predictive model for **daily operations**

Baseline aim to outperform the **Naïve model**

Scope

Will not consider:
~~Black Swan events~~
~~Market Sentiments~~

Provides a Quantitative Approach on Simulated Data

Strategy Journey Mapped

optiver

1

Client Meeting 2 | Adrian and Virginia

- ARIMA and GARCH as the base estimator in Hybrid model
- Idea of Interpretability v Predictability
- Ensembles, encouraged us to pursue our hybrid model idea further

2

Client Meeting 3 | Greg and Virginia

- Alternative Model – EWMA
    - A faster, simpler, and, similar in accuracy to the ARIMA
- Explore other Evaluation Metrics:
    - RMSE and MAE

Strategy Journey Mapped

3

Client Meeting 4 | Adrian and Virginia

- Explore Diagnostic Analytics
  - Residual Plots
  - Heteroskedasticity Plots
- Explore Data Visualization
  - Scatter Plots
  - R2 v RMSPE Graph

4

Brock Sherlock | UNSW PhD Math & Stats

- Identified flaws in Bayesian calculations
- Idea of Bayesian Averaging
- Replace Random Forest base estimator with LightGBM model

Strategy Journey Mapped

1 Client Meeting 2

2 Client Meeting 3

3 Client Meeting 4

4 Brock Sherlock

Consolidated Stakeholder **Feedback Iteration**

Base Estimators: EWMA + LightGBM

- Refined Bayesian Averaging
- Significantly reduce runtimes
- Explore evaluation methods and metrics
- Mitigate risks and limitations

Base Estimator – EWMA

optiver

Exponentially Weighted Moving Average (EWMA)

EWMA Model

1. **Simple technique** to estimate volatility and **minimal** computational resources

2. Applies more **weight to recent** data points
   - Well-suited for capturing short-term trends

EWMA Formula

Calculates the exponentially weighted moving standard deviation of log returns

Adapts the annualized volatility equation for 30-minute intervals

```
ewm_vol = group['log_return'].ewm(span=10).std() * np.sqrt(len(group))
```

Annualized Volatility = Standard Deviation x SquareRoot of data size

Base Estimator – EWMA

Model Performance

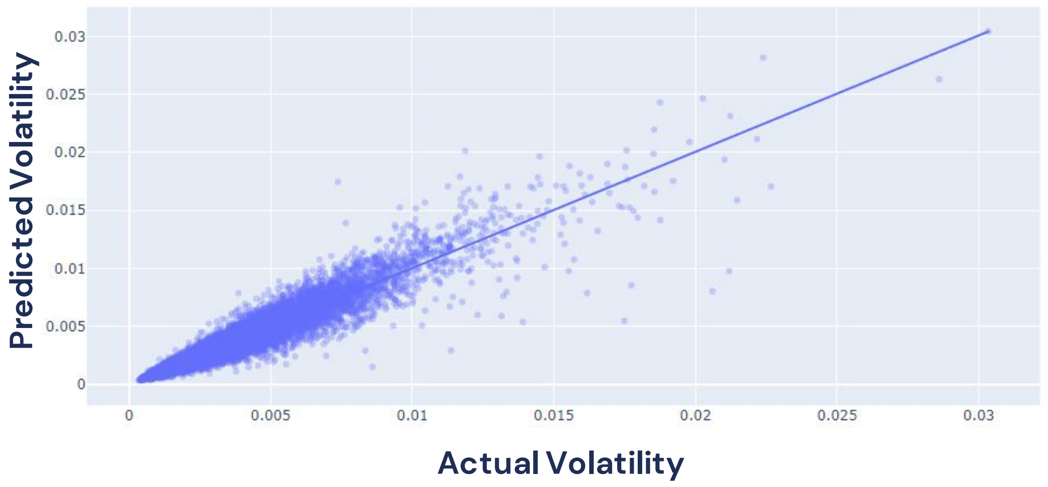
EWMA Model

Runtime: 0.133 mins ~ 8 secs
R^2: 0.872
RMSPE: 0.188
MAE: 0.0059
RMSE: 0.0009

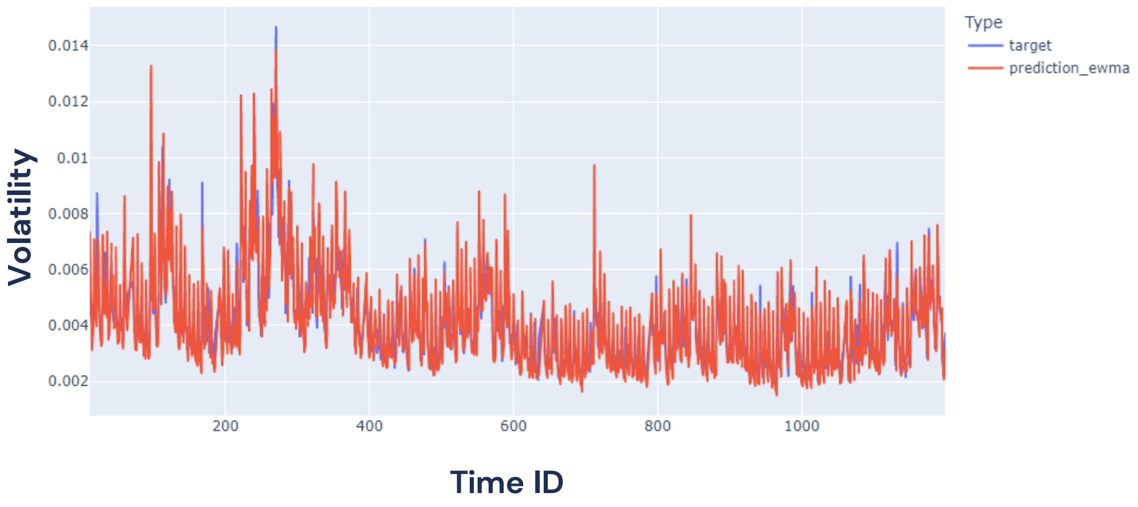
Effective in estimating short-term stock volatility

Solid baseline against other models

Actual vs Predicted Volatility for All Stocks (EWMA)



Actual vs Predicted Volatility for All Stocks (EWMA)



Base Estimator – LightGBM

optiver

Light Gradient-Boosting Machine (LightGBM)

LightGBM

1 **Gradient Boosted Decision Trees**
- ➢ Well-suited for short-term volatility predictions

2 Effective for **large datasets** and high-dimensional features

Data Preparation & Feature Engineering

1. Filtered data for each stock ID
2. Use log returns squared sum and count of log returns as feature
3. Realized volatility as the target variable

Model Training

Train-test split:
80% training | 20% testing

Evaluation Matric : RMSE

40 boosting rounds with early stopping

Base Estimator – LightGBM

Model Performance

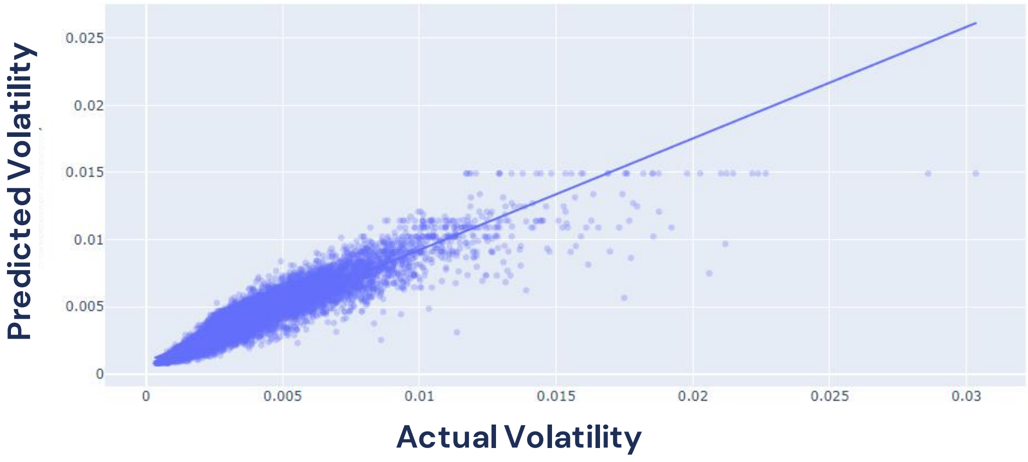
LightGBM

Runtime: 0.141 mins ~ 8 secs
R^2: 0.855
RMSPE: 0.261
MAE:
RMSE:

Effective for handling large datasets and complex features

Captures complex non-linear relationships

Actual vs Predicted Volatility for All Stocks (LightGBM)



Actual vs Predicted Volatility for All Stocks (LightGBM)

Bayesian Averaging – Hybrid Model Explored

optiver

Bayesian Averaging Formula

$$\frac{C_1 P_1 + C_2 P_2}{C_1 + C_2}$$

Basic weighted average formula which weighs each model's prediction based on how accurate the predictions are

$$\frac{(1 - \text{RMSE\_ewma}) * \text{pred\_EWMA} + (1 - \text{RMSE\_lgb}) * \text{pred\_lgb}}{((1-\text{RMSE\_ewma}) + (1 - \text{RMSE\_lgb}))}$$

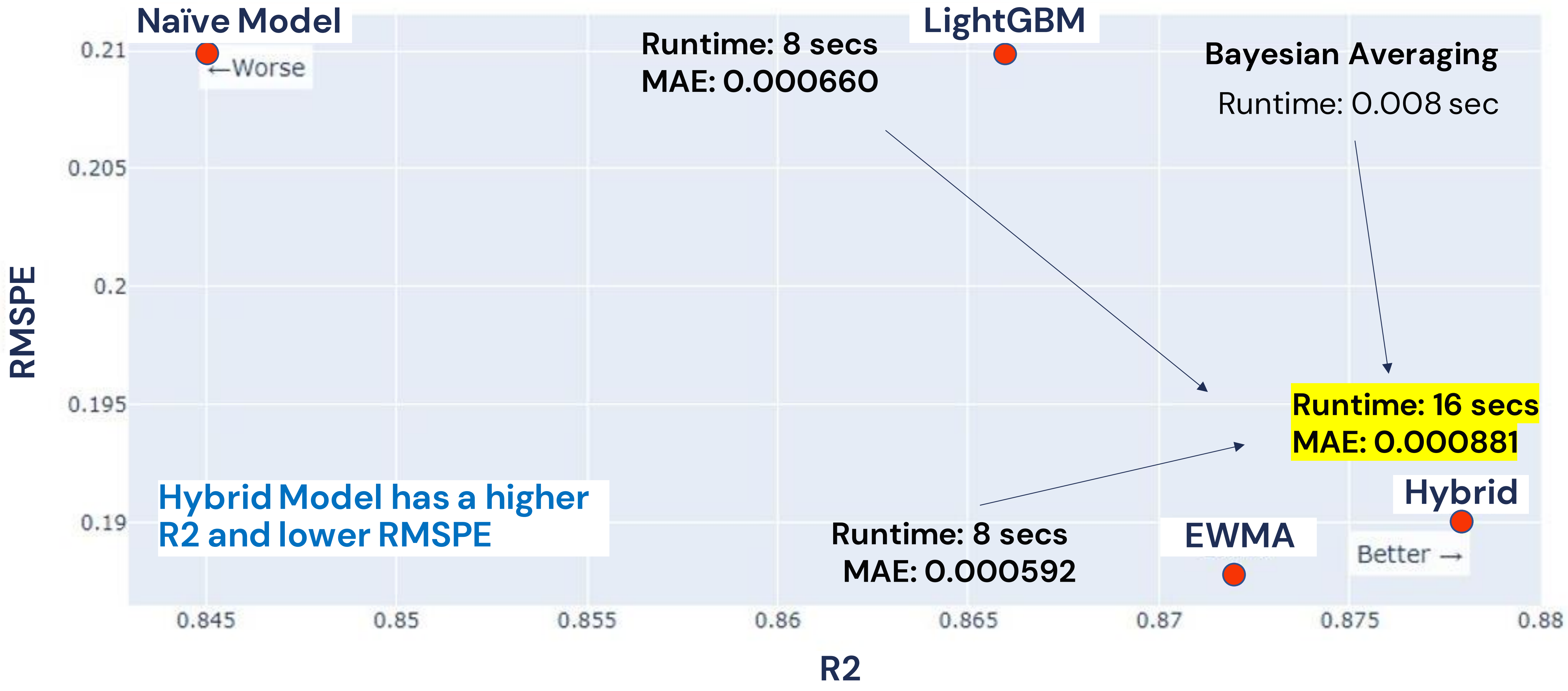
Note

C_1 = (1 – RMSE_ewma)
P_1 = ewma prediction
C_2 = (1 – RMSE_lgb)
P_2 = lightGBM prediction
These act as dependent values of the stock

Prediction Overview

stock_id	RMSE_ewma	RMSE_lgb
9323	0.000752	0.000709
22675	0.000793	0.000878
22951	0.000812	0.000845
22729	0.001092	0.001154
48219	0.001195	0.001130
22753	0.000647	0.000604
22771	0.001080	0.001113
104919	0.000453	0.000518
50200	0.000361	0.000417
8382	0.001321	0.001618

Hybrid Model Performance Overview



Evaluating the Hybrid Models

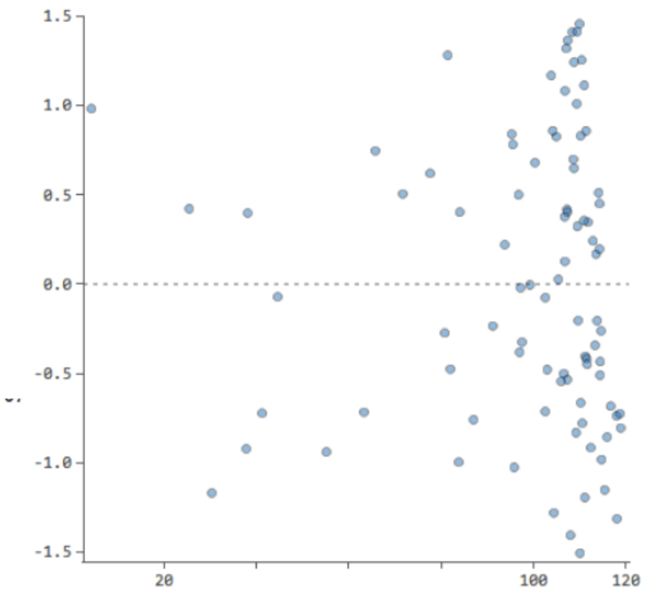
optiver

Diagnostic Plots

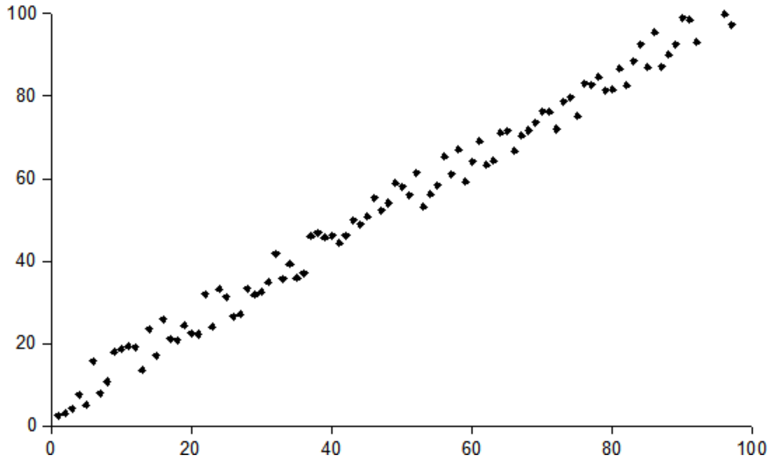
Evaluation Plots

Ideal plots:

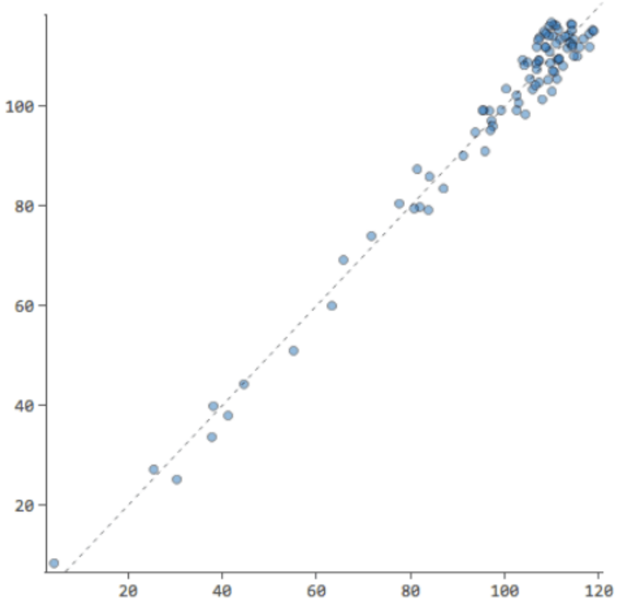
Residual Plots



Heteroskedasticity Plots

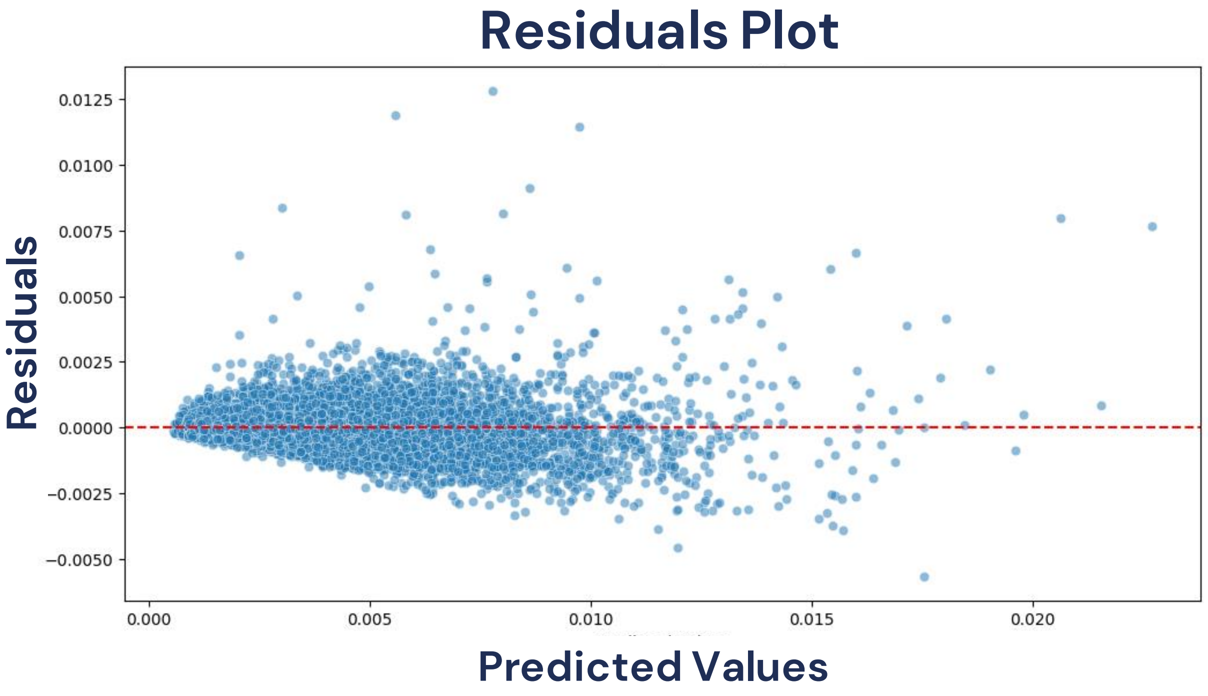


Scatter Plot



- [x] residuals randomly scattered
- [x] Homoskedasticity
- [x] linear pattern

Hybrid Model Diagnostic Plots



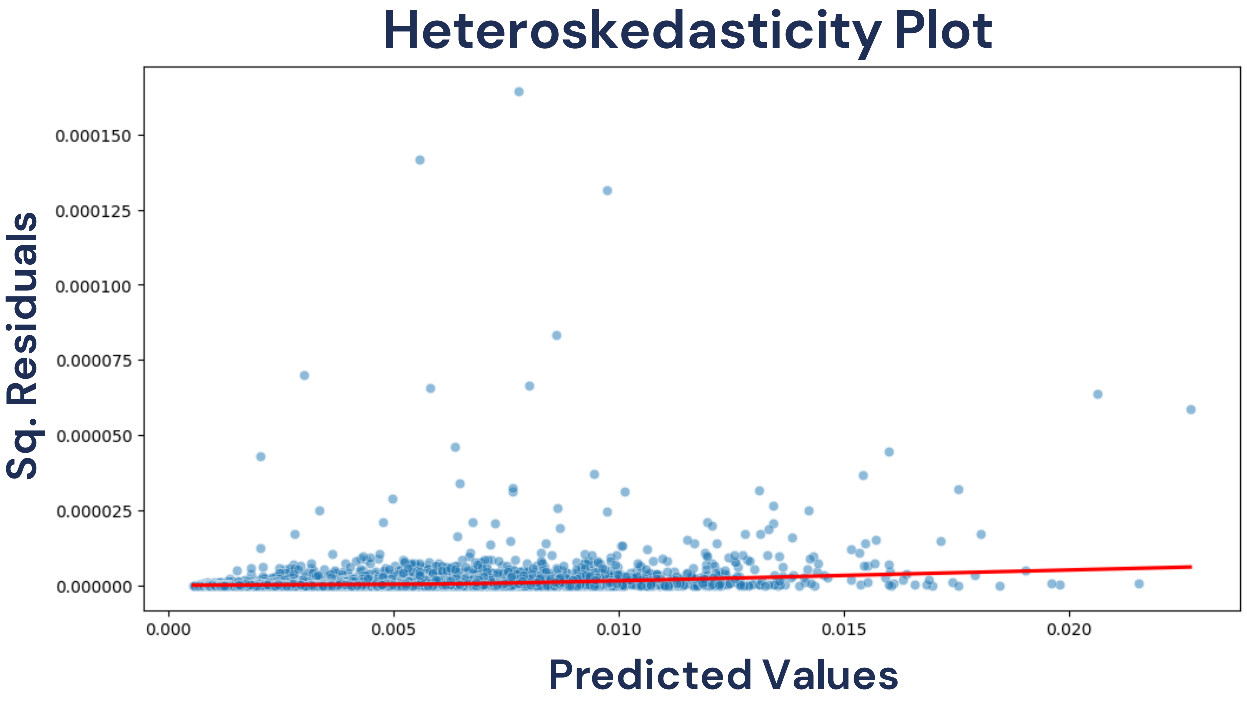
Good fit model, as:

- ☑ **Linearity**: The relationship between the independent and dependent variables should be linear.
- ☑ **Independence**: The residuals should be no correlation between them.
- ☑ **Homoscedasticity**: The variance of the residuals are constant.
- ☑ **Normality**: The plot shows a normal distribution.

However,

- ✗ **Outliers**: There are quite several outliers at specific data points.

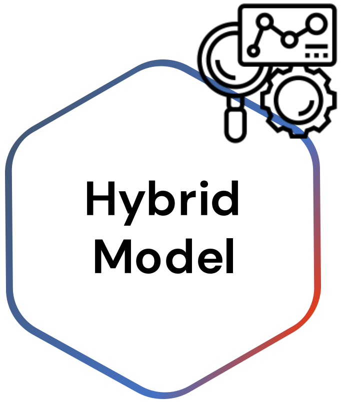
Hybrid Model Diagnostic Plots

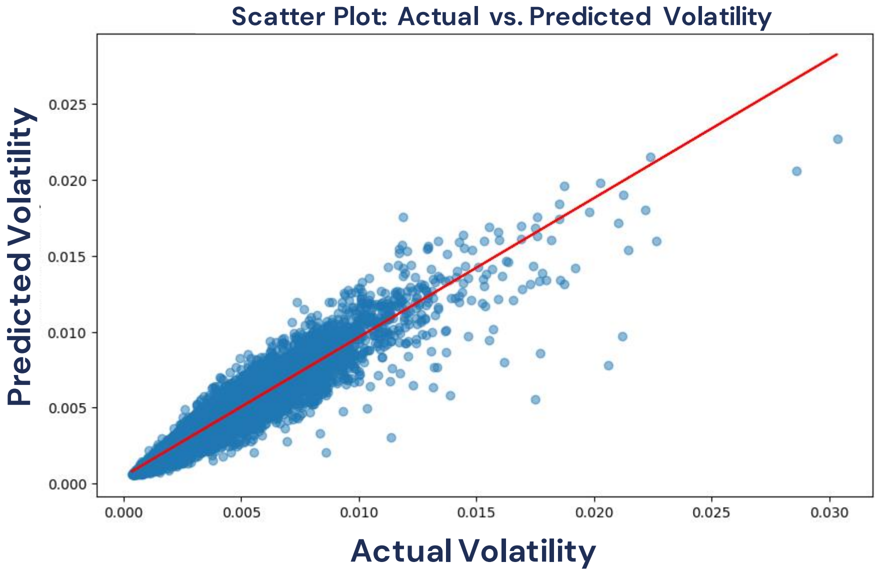


Homoscedasticity, as:

- **Linearity**: The relationship between the independent and dependent variables should be linear.
- **Independence**: The observations in the dataset should be independent of each other.
- **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variable(s).
- **Normally distributed errors**: The error term should be normally distributed, with a mean of zero.

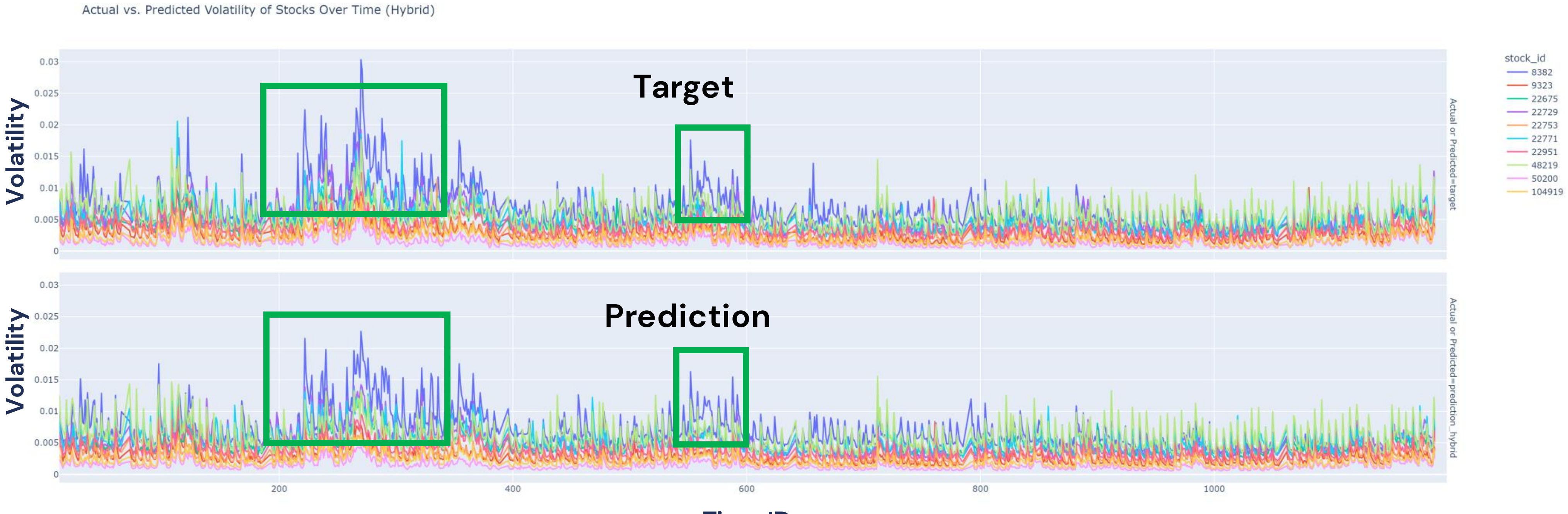
Hybrid Model Evaluation Plots





Good fit model:

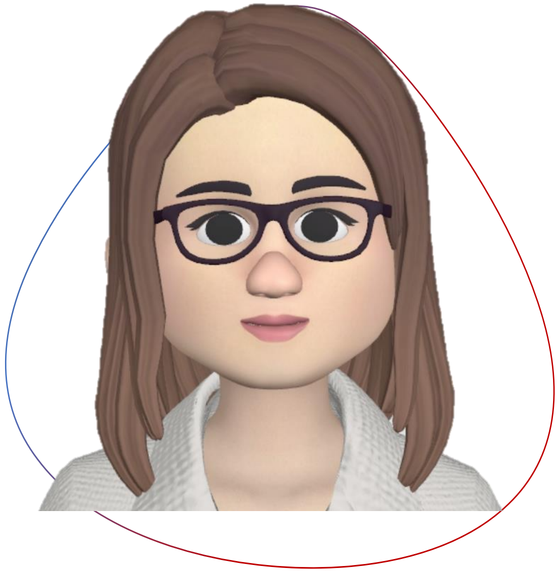
- Shows a linear pattern with a positive slope, indicating a strong positive correlation between the predicted values and the actual values.
- Only a few outliers as well.



Runtime: 16.008 sec
R^2: 0.878
RMSPE: 0.19
MAE: 0.000587
RMSE: 0.000881

Our Value Proposition Visualized

Implementing the Model



User/ Trader

1 Receives a code package and a dataset of our model

2 Plug in the code and dataset into their existing trading system

Define Python functions and data pre-processing steps

Hybrid Model

Load New Stock Dataset

```
train = pd.read_csv('train.csv')
book = pd.read_parquet('order_book_feature.parquet')
trade =  pd.read_parquet('trades.parquet')
```

Our Value Proposition Visualized

Implementing the Model

Technical Report Analyses

Outline

(1) Assumptions

- List the assumptions made while developing the model
- Explain their impact on the model's performance

(2) Methodology

Explain EWMA, LightGBM and Bayesian Averaging approach

(3) Findings

- Summarize the key insights
- Discuss the overall performance
- Highlight any unexpected findings

(4) Performance Evaluations

- Present performance metrics
- Include visualizations plots
- Discuss and compare the model's performance

(5) Risks and Limitations

- Identify potential limitations
- Discuss the limitations of the data used
- Explain how these limitations might affect the model's accuracy and reliability

Our Value Proposition Visualized

Implementing the Model → Technical Report Analyses → Forecast Volatility

Evaluating Results

Hybrid Model

Identify Trading Opportunities

- Help traders find opportunities in stocks with significant price movements

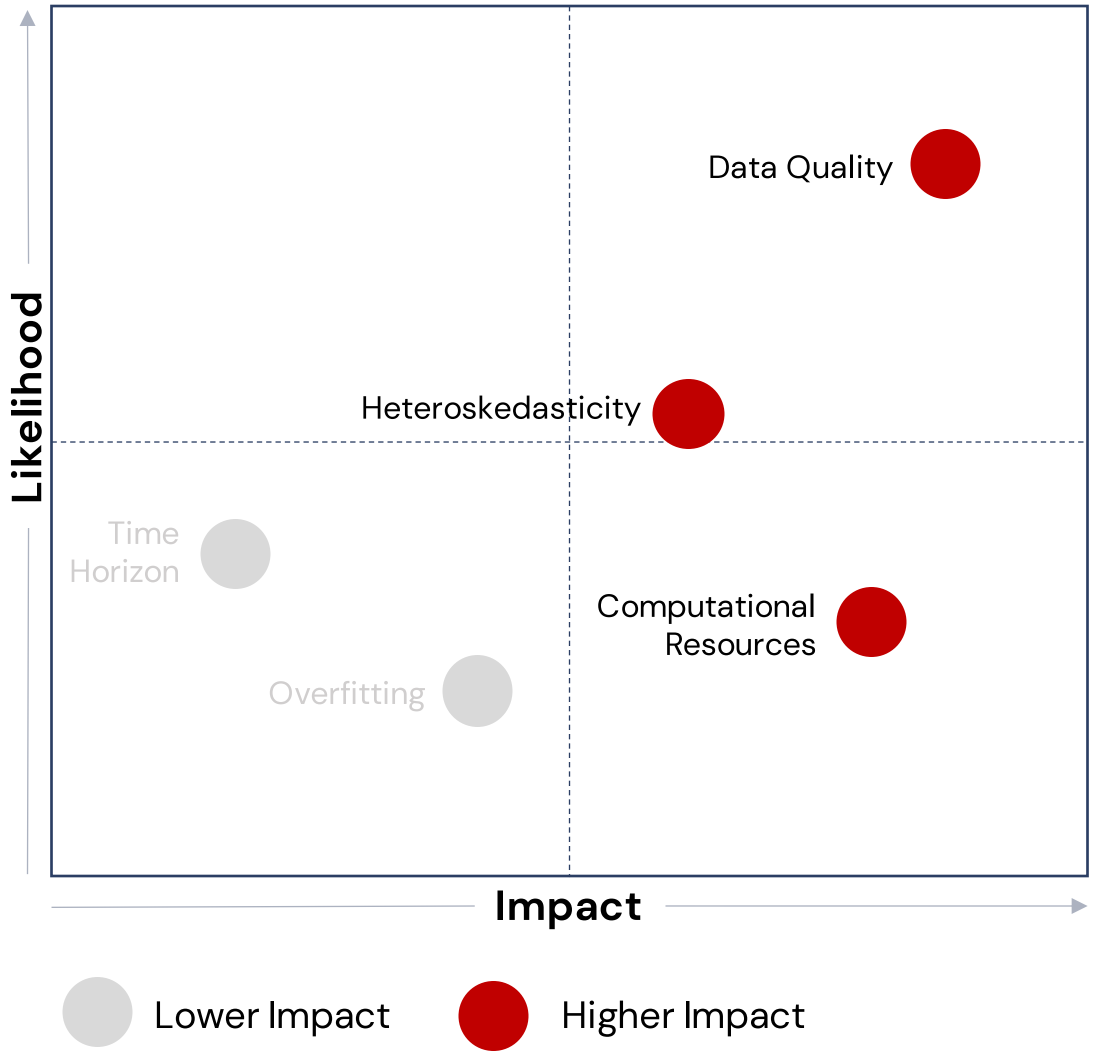
Optimize Portfolio Construction

- Aid investors in optimizing their portfolio by adjusting asset allocation, diversification, and risk management

Informed Trading Decisions

- Enable traders to make informed decisions regarding entry and exit points and strategy adjustments

Risks and Mitigations



1 **Data Quality**
- The accuracy of the prediction relies on dataset quality, with more errors leading to biased predictions
- Mitigation: We have **pre-processed the data beforehand**

2 **Computational Resources**
- Hybrid models are usually complex and slow
- Mitigation: We have **optimised our code** (e.g., reduced time taken to run EWMA from 9 minutes to 10 seconds)

3 **Heteroskedasticity**
- It is ideal to have constant variability in errors
- Mitigation: Bayesian Averaging **incorporates different model structures & parameterisations** to better capture variability

Limitations and Improvements

Improvements & suggestions…

Data Interpretability

The model is currently **limited to interpreting dataset format** that is the same as the book & train datasets

Improved model: **Interpret different dataset formats**

Bayesian Weighting Calculations

The model currently uses comparisons of model predictions and the target dataset (which is **not possible in real-word application**)

Suggestion: Use **first 15 minutes of training data to predict the next 15 minutes**. Then, use the results to calculate weightings for actual prediction.

Predictive Power

The model's prediction power **mainly relies on the provided dataset**.

Improved model: **Considers external factors and additional financial indicators**, and go in depth on provided data to consider

What Next? Recommendations

Data Interpretability

Bayesian Weighting Calculations

Predictive Power

Plans to improve...

1

Gathering External Information
Such as news article or other financial indicators as aforementioned

2

Model Reiteration
The model should then be reiterated to include the improvements. Modification can also be done on the baseline model to improve performance

3

Training the Improved Model
The model should then be trained with varying datasets to judge its performance and then reiterated accordingly

APPENDIX NETWORK

Presentation

Performance Metrics Overview

optiver

Naïve Model

Optiver

R^2: 0.845
RMSPE: 0.21
MAE: 0.000657979

~~ARIMA Model~~

Runtime: ~1200 mins
R^2: 0.874
RMSPE: 0.187
MAE:

Large runtime
Low explainability

EWMA Model

Runtime: 0.133 mins ~ 8 secs
R^2: 0.872
RMSPE: 0.188
MAE: 0.00059190

~~Gaussian Process Regression~~

Runtime: 160.78 mins
R^2: 0.841
RMSPE: 0.001
MAE: 0.149

Large runtime

~~Random Forest~~

Runtime: 0.153 mins ~ 9 secs
R^2: 0.846
RMSPE: 0.229
MAE: 0.00066182

Not ideal for hybrid model

LightGBM

Runtime: 0.141 mins ~ 8 secs
R^2: 0.855
RMSPE: 0.261
MAE: 0.000673140

Standardizing Testing Environment

optiver

0110
1001
1010

Data from book_data

stock_ids = 9323, 22675, 22951, 22729, 48219, 22753, 22771, 104919, 50200, 8382

time_id: from 12 to 1200

Evaluation Metrics

R^2: how well the model fits the data

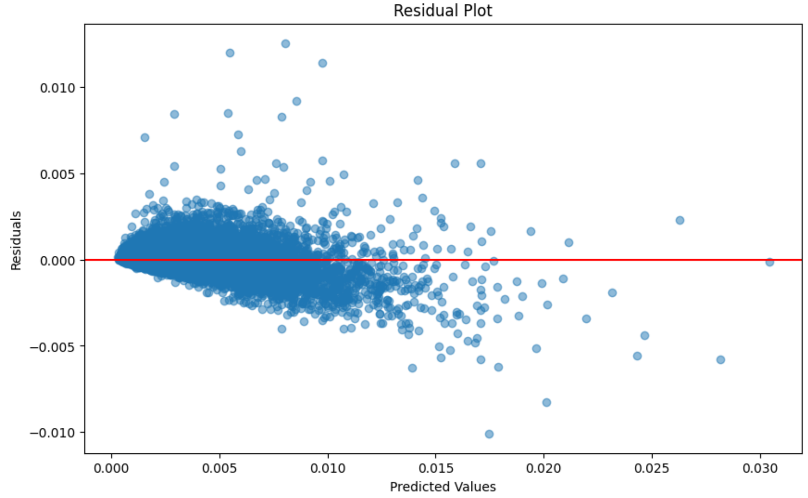
RMSPE: the accuracy of the model's predictions

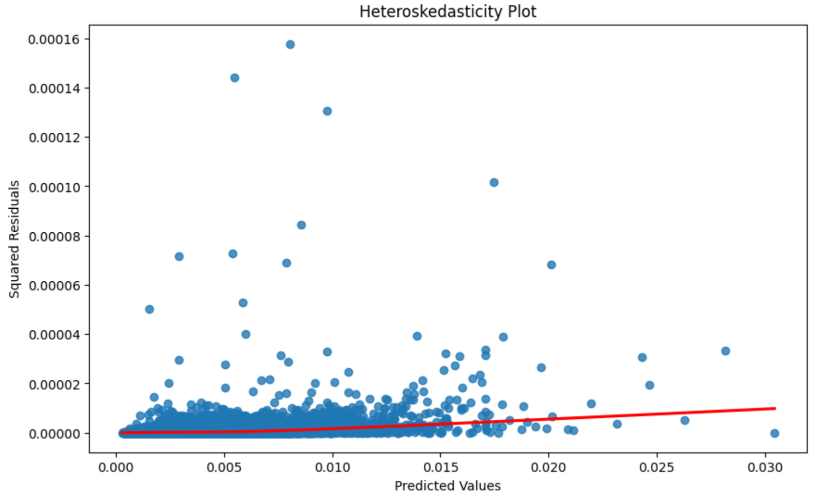
MAE: the average difference between prediction vs actual

Prediction **runtime** in mins

EWMA Model and LightGBM Diagnostic Plots

EWMA Model





Residuals plots:

- constant variance -> **Homoscedastic**

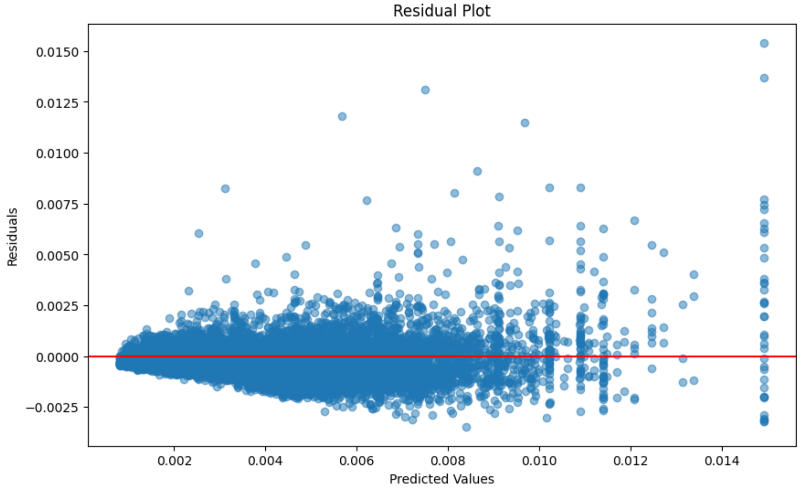
However

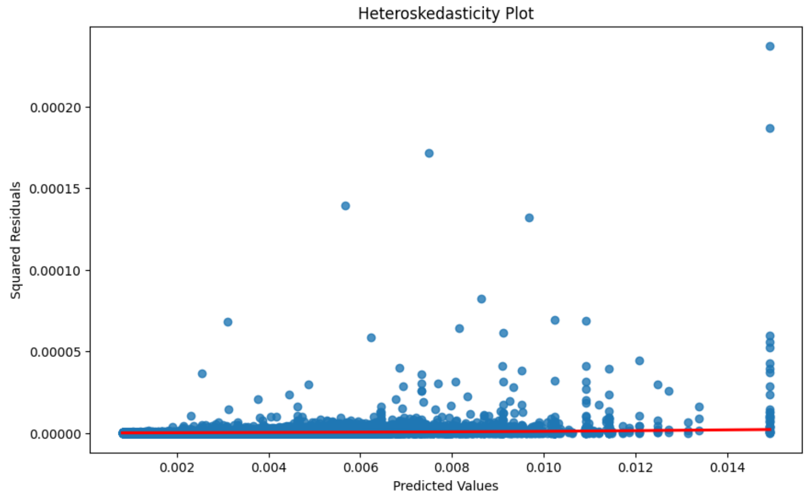
- decreasing trend -> not capturing all the information -> underestimating the trend in the data
- outliners

Heteroskedasticity plots:

- constant variance patterns -> shows homoscedasticity

LightGBM Model





Explaining the EWMA Equation

This code aims to calculate the volatility at each time point. It uses two key concepts: Exponential Weighted Moving Average (EWMA) and annualized volatility calculation.

Exponential Weighted Moving Average (EWMA):
EWMA is a method to calculate a moving average by weighting past data points according to their distance from the current observation. The closer a data point is to the current observation, the higher the weight it receives, and the further away it is, the lower the weight. This allows EWMA to capture recent volatility changes more effectively. The span parameter is used to control the speed of weight distribution. Larger span values will result in smoother weight distribution, while smaller span values will focus more on recent changes. In this code, we calculate the exponential weighted moving average standard deviation at each time point using ewm(span=10), which represents the volatility of past price changes.

Annualized volatility calculation:
This part of the code converts the EWMA standard deviation at each time point into annualized volatility. Annualized volatility is the standard of extending short-term volatility to a period. The np.sqrt(total number of time point) here represents converting daily volatility into annualized volatility.
Combining the two points above, the formula for this code is:
volatility(t) = EWMA_std(t) * sqrt(total number of time point)

where t represents each time point, EWMA_std(t) represents the exponential weighted moving average standard deviation at time point t. volatility(t) is the annualized volatility at time point t.

Further Risks, Limitations, and Improvements

Data quality risk

- In the future, we can improve this process by undergoing correlation analysis or principal component analysis (PCA) before running the model

Other risks:

- Time horizon
  - Low impact because the task we gave to do is to predict short-term volatility (not long-term, so it is unlikely that we have to change the time interval of our predictions)
  - This risk can be mitigated by training the model with different subsets of different time intervals

- Overfitting
  - Overfitting has a low impact as we have considered this well.
  - The Bayesian averaging technique helps mitigate overfitting by introducing a level of regularization technique through the averaging process.
  - However future regularization or feature selection process can still be done to further consider this