Financial Data Engineers

风险预警师

刘天泽-南加州大学-Master of Science-2025.May
洪奕诗-南加州大学-Master of Financial Engineering
王楠-南加州大学-Master of Science-2025.May

Abstract

There exist four different datasets in CSV form namely swap-rates, market volatilities, trade information and the vegas.

Our envisioned goal is to train a generalized model to predict the value of the Vegas, which is in the Vegas CSV file.

The models can be splitted into different types including Generative models such as Fast Fourier Series, the traditional machine learning model like XGboosting, the deep learning model such as LSTM which performs pretty well on the dataset.

Contents

Data Preprocessing - For Swap Rates & Vols

For swap rates, we apply the Exponential Moving Average to integrate the results of them. For each start date, it has different swap rates on different date. We want to fetch all the swap rates to form the final transformed swap rate used in the final model training procedure.

For volatilities, we simply combined the vols under different strikes values to the average of them.

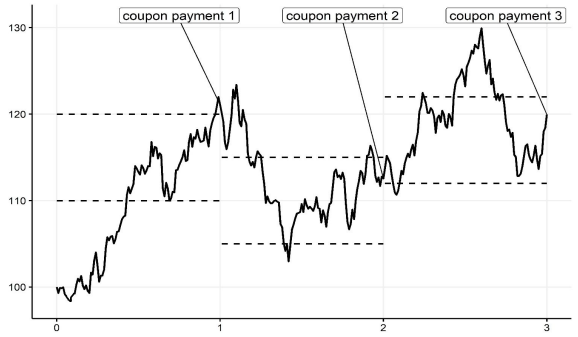
After these procedures, now we have the transformed vols and swap rates on each of the dates.

Data Preprocessing - For Trade Information

To simulate product characteristics, we have designed various aspects of the trade information feature architecture.

Underlying:

1. Time features: We calculate STD for interest spreads across different time intervals (daily&weekly&monthly) to further estimate the short-term, medium-term, and long-term volatility of underlying assets;
2. Statistic features: We apply mean, STD and minimum, maximum of CMS to describe the central tendency and dispersion of the underlying assets;



Pay Frequency: We transform the frequency string into the number of payments per year;

Maturity: We simply transform the maturity string into the expiration term (the number of years) of each dummy trade;

Lower bound & Upper bound: We introduce a new feature to handle bounds, i.e. the proportion of time the underlying asset's interest rates remain within upper and lower bounds, which is closely tied to each payment occurrence.

Models - XGboosting and Optuna

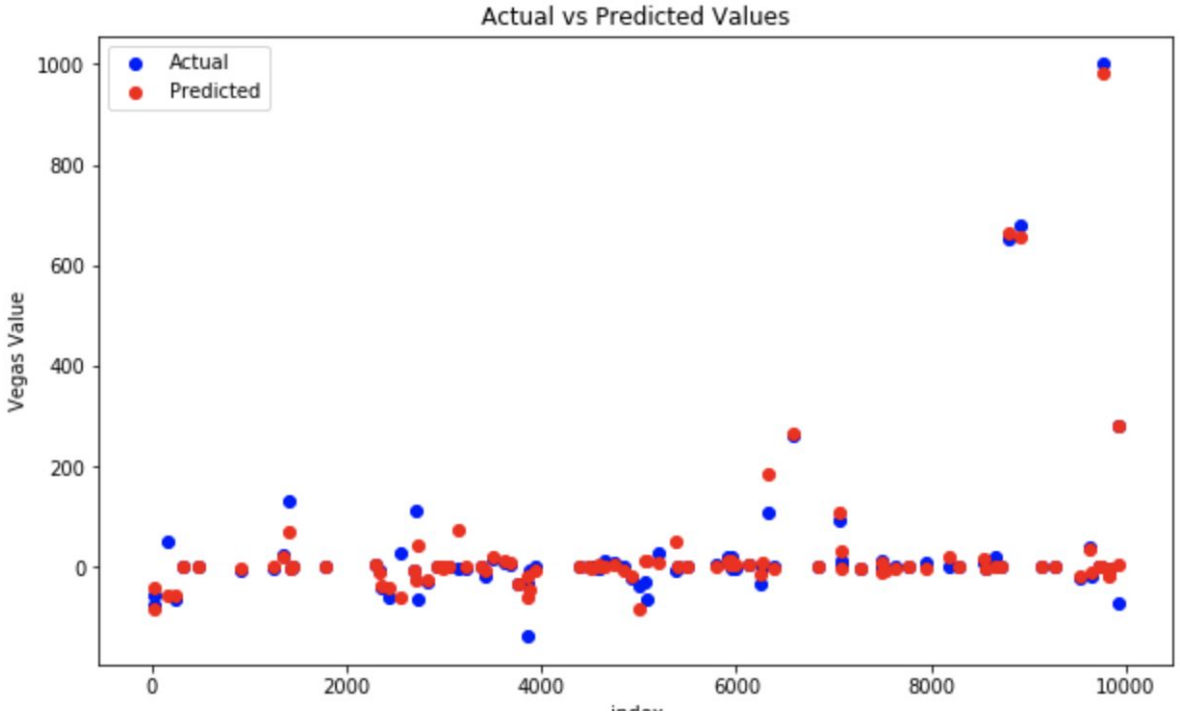
XGboosting is a widely-used algorithm in the financial engineering field. We combined it with the Optuna Package which is used to automatically search the best hyperparameters for the model.

After training, the RMSE could be lower as the 26.33. The corresponding hyperparameters can be found on the right.

```
Best trial:
  Value: 26.333607904481678
  Params:
    n_estimators: 1366
    max_depth: 10
    learning_rate: 0.011532067255216237
    subsample: 0.8491331872122521
    colsample_bytree: 0.7522507149450635
    reg_alpha: 6.735984595335068
    reg_lambda: 7.820908582924565
Test RMSE: 26.333607904481678
```

Models - XGboosting and Optuna

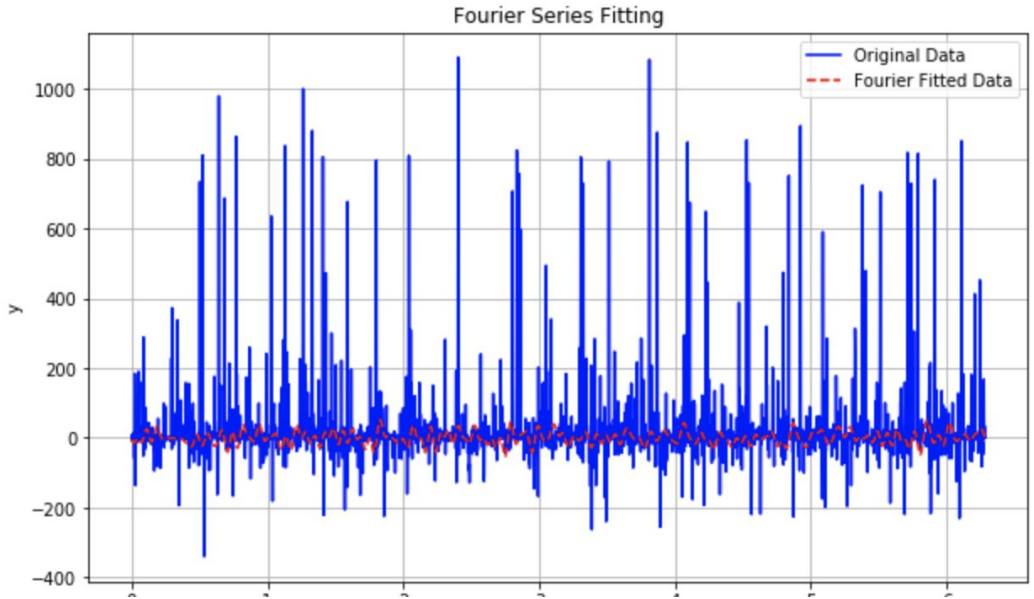
Data Visualization



Models - Fast Fourier Analysis

Fast Fourier Analysis is commonly used in the signal recovering field. It reconstruct the signal from the original value and represent them with the sin and cos functions.

It performs not well. The MSE is linear to the sample size. That is highly because the market is not a periodical phenomenon in the real world.



Models - ARIMAX & Lasso

We apply this approach by combining the feature selection capability of Lasso regression with the time series forecasting ability of the ARIMAX model:

$$Y_{t+1} = \sum_{i=0}^{P} \beta_i Y_{t-i} + \sum_{j=1}^{J} \theta_j X_{t,j} + \sum_{k=1}^{K} \Phi_k Z_{t,k} + \varepsilon_{t+1}$$

1. **Lasso Regression for Feature Selection:** Lasso regression selects multiple categorical variables like zero rate shock, tenor bucket, and expiry bucket.
2. **Integration of ARIMAX Model with Exogenous Variables:** We use the selected variables as exogenous variables to build an ARIMAX model.
3. **Multiple Time Series Models under Different Conditions:** ARIMAX are trained based on dummy trades in various scenarios to ensure better adaptation to different market conditions and trading situations
4. **Model Fusion and Generalization:** Coefficients from ARIMAX are averaged to generalize the entire model.
5. **Model Performance Evaluation:** Model performance is evaluated using Mean Squared Error (MSE) averaging.

Models - ARIMAX & Lasso

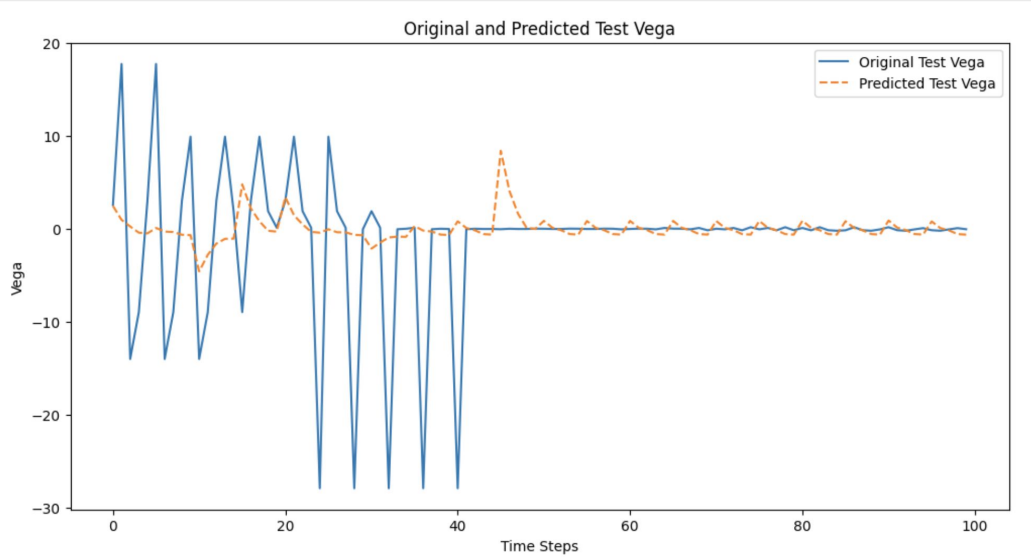
The final result is shown on the right, but it doesn't perform well since the vega does not exhibit obvious time series characteristics:

```
Average coefficients for ARIMAX model:
Zero Rate Shock     1.777650e-01
TV                  1.514279e-03
Expiry Days        -2.919429e+00
Tenor Bucket        5.853549e-03
Vols               -4.240321e+02
pay_frequency       8.609721e-03
maturity            2.269707e-02
Min_CMS             1.230542e-03
ar.L1               4.413422e-02
ma.L1              -6.376421e-01
sigma2              1.296168e+07
dtype: float64
Average MSE for ARIMAX model: 3506079.558920936
```

Models - LSTM

LSTM is a well-known deep recursive neural network in the sequential data predicting task. It can fetch the information throughout a long time arena. The further data will have smaller impact on current data and vice versa.

The model performs well and the MSE reach only 59 on the time scope while having a good performance on the generalization as well.



Results & Metrics

Model	Mean Square Error (Small Scope)
XGBoosting	676.12
Fast Fourier	Linear to the dataset size
ARIMA+ Lasso	3506079
LSTM	59.12

Conclusion

For inspection of the model accuracy, we use MSE — Mean Squared Error as an indicator. From the above statistics, LSTM model appears to be more steady in the scope of Time Series prediction models. However, other models do not appear as expected. It may stem from several reasons:

- Complexity of data structure, feature formation, and design of our algorithms
- Diversity of the market, undiversifiable risks, and the complexity of risk components

Limitations & Future Work

In modern AI field, the method of stacking the layers to form a powerful model is going popular. The appearance of the Transformer forwards the work to the complicated models time.

Here from our perspective, the Stockformer which performs well on stock prediction on the market is likely to have a good performance as well.

At the same time, due to the diversity of the conditions of the market, predicting the vegas requires plenty of financial knowledge as well. Therefore, working with the expert in financial field is necessary.