Financial Time Series Analysis and Prediction: Kaggle Case Study

LIANG Zhenlin

Supervisor: Prof. LI Xiang

Cognitive Science Lab, The University of Aizu

1.Introduction

Time series prediction is an area of interest to a lot of people. In recent years, time series prediction algorithm has been widely used in the fields of finance, sales forecast, energy and weather forecast. In the famous data modeling and data analysis competition platform **Kaggle**, many contests about time series forecasting are popular, such as Walmart Recruiter-Store Sales Forecasting, etc., which use historical Sales data to predict future Sales. The Recruit Restaurant Visitor Forecasting contest seeks to predict future customer numbers based on historical attendance. Zillow Prize: Zillow's Home Value Prediction seeks to use historical housing price information to predict future values. In these competitions, we can see that contestants are often not limited to models or algorithms, but to think and analyze the nature of the problem. **Exploratory data analysis**(EDA) is very important, it determines the framework of feature engineering, and **feature engineering** in turn affects the upper limit of a single model and thus affects the final performance of the solution. In this study, the **Jane Street Market Prediction Competition** of Kaggle will be used as an example to describe the generation of the whole solution.

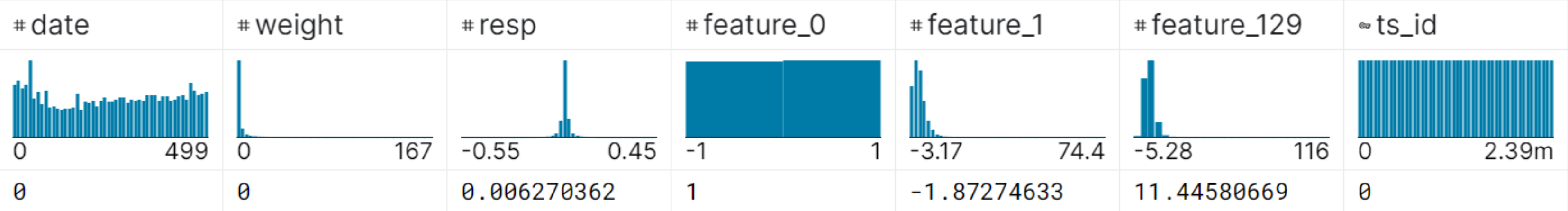
2.Jane Street Market Prediction Competition

- **Overview:**

Jane Street held the competition based on a real problem with the actual financial data. Each piece of data was based on a real transaction that had existed in the real world. The goal of the competition was to determine whether these transactions should be executed or abandoned in order to make a profit.

- **Data Description:**

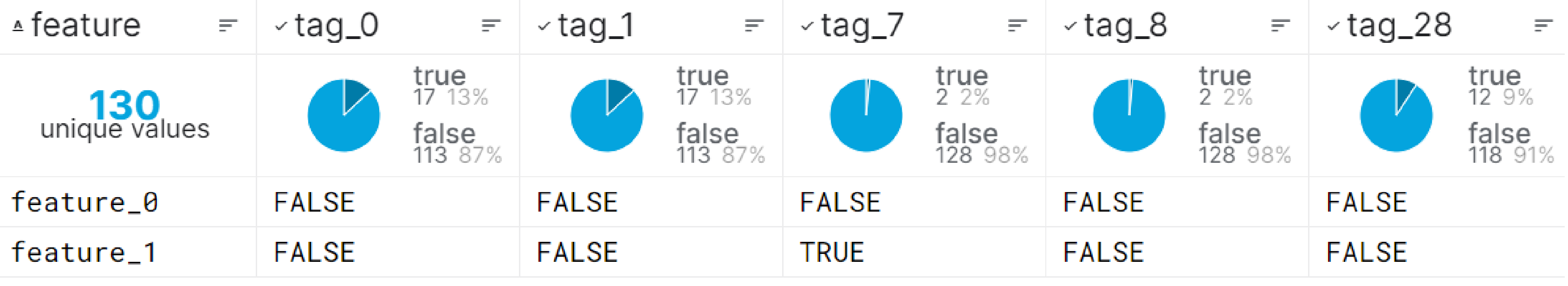
The key point of the data competition is the **cognition of the data set**. This dataset contains an anonymized set of features, feature_{0...129}, representing real stock market data. Each row in the dataset represents a trading opportunity, for which you will be predicting an action value: 1 to make the trade and 0 to pass on it. Each trade has an associated weight and resp, which together represents a return on the trade. The date column is an integer which represents the day of the trade, while ts_id represents a time ordering(See Fig.1).



# date	# weight	# resp	# feature_0	# feature_1	# feature_129	ts_id
0 499	0 167	-0.55 0.45	-1 1	-3.17 74.4	-5.28 116	0 2.39m
0	0	0.006270362	1	-1.87274633	11.44580669	0

Fig.1 Form and distribution of data sets

The competition does not explain the actual meaning of features, but it does give the corresponding meta-attribute for each feature(See Fig.2)



feature	tag_0	tag_1	tag_7	tag_8	tag_28
130 unique values	true 17 13% false 113 87%	true 17 13% false 113 87%	true 2 2% false 128 98%	true 2 2% false 128 98%	true 12 9% false 118 91%
feature_0	FALSE	FALSE	FALSE	FALSE	FALSE
feature_1	FALSE	FALSE	TRUE	FALSE	FALSE

Fig.2 Metadata tags for each feature

- **Evaluation:**

This competition is evaluated on a utility score. Each row in the test set represents a trading opportunity for which you will be predicting an action value, 1 to make the trade and 0 to pass on it. Each trade j has an associated weight and resp, which represents a return.

$$p_i = \sum_j \left(\text{weight}_{ij} * \text{resp}_{ij} * \text{action}_{ij} \right)$$

$$t = \frac{\sum p_i}{\sqrt{\sum p_i^2}} * \sqrt{\frac{250}{|i|}}$$

(Where i is date, j is the serial number of the trading activity of the day weight_{ij}, resp_{ij} and action_{ij} are the corresponding values in the data set)

3.Exploratory Data Analysis

Exploratory data analysis(EDA) is the very important part of the Kaggle competition. Generally, EDA is to explore existing data (especially the original data obtained from investigation) with as little prior assumptions as possible, and to explore the structure and rules of data by means of plot, tabulation, equation fitting, and calculation of characteristic quantities, so that the data can be better understood and build an intuition about data. From the results of EDA, the framework of feature engineering and model can be determined.

Here are some interesting and valuable EDA excerpts from the competition:

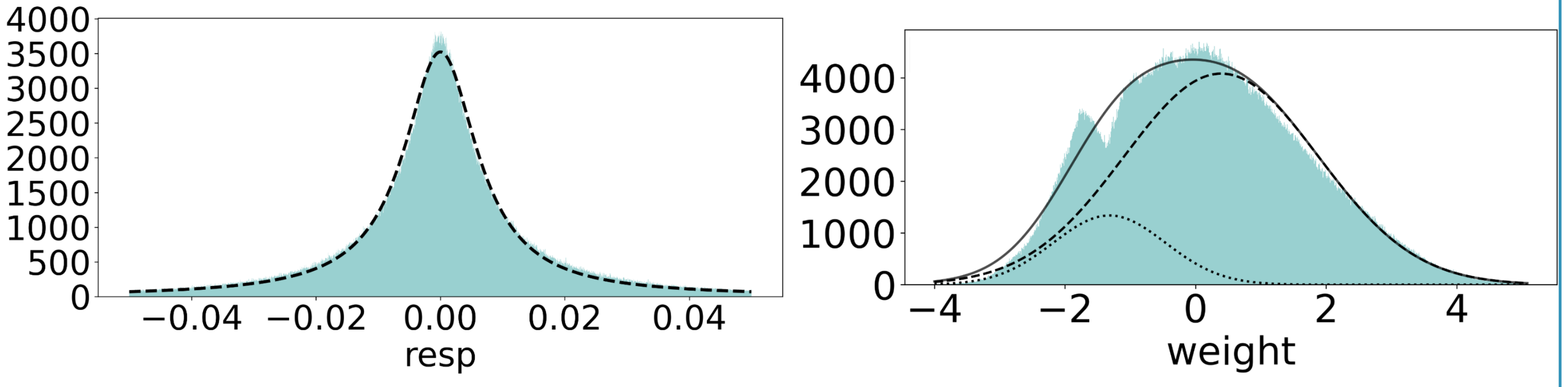


Fig 3. Distribution of resp and weight values

resp is a Cauchy distribution, and non-zero weight Gaussian distributions(See Fig3). In the anonymous feature data, the cumulative of feature_0 is distributed differently, and it corresponds to the values of return(resp*weight) and resp[1](See Fig.4) In addition, Random Forest algorithm is used to extract features and rank the importance of feature[2](See Fig.5)

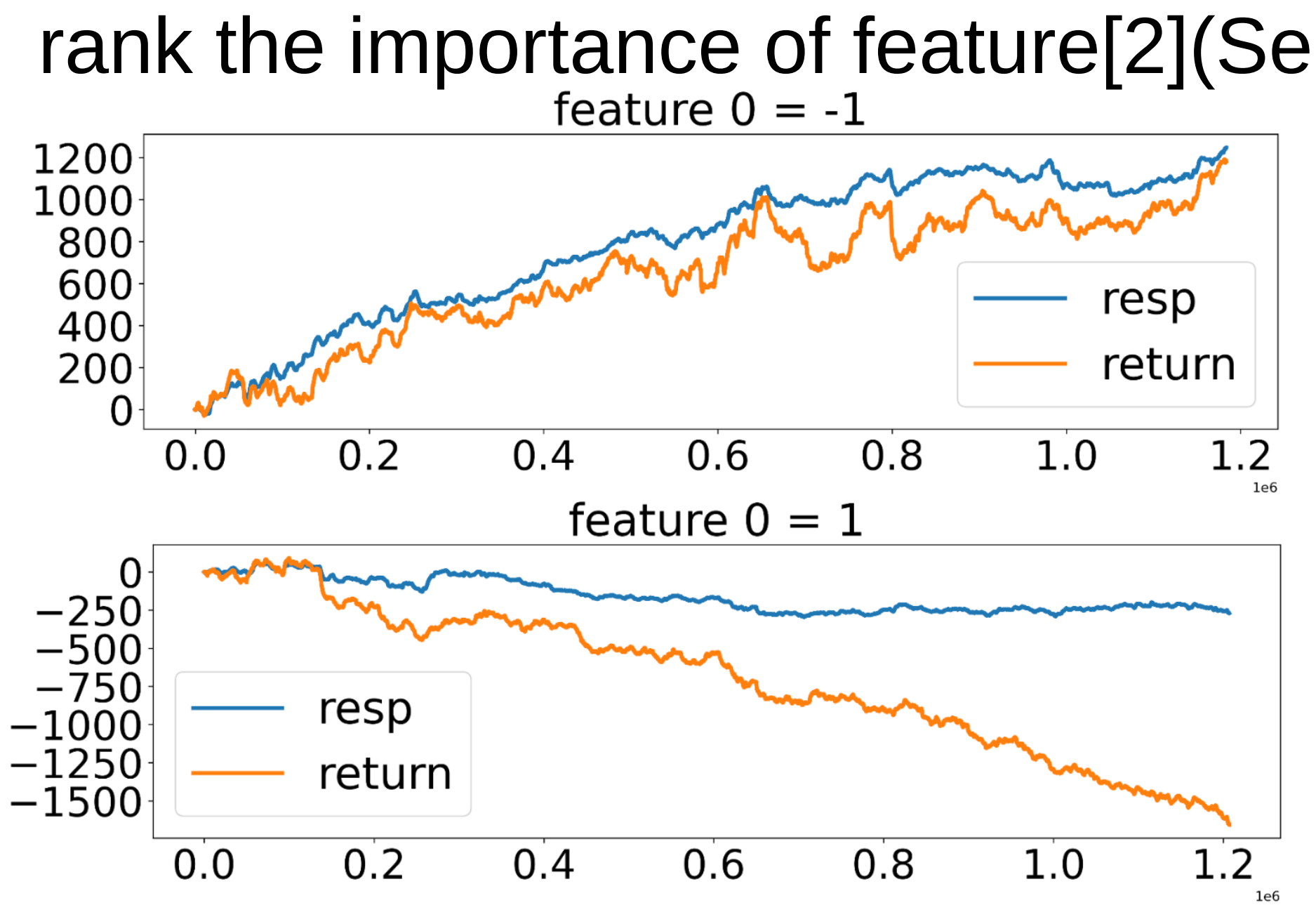


Fig.4 Cumulation of corresponding values for FEATURE_0 and RESP and RETURN

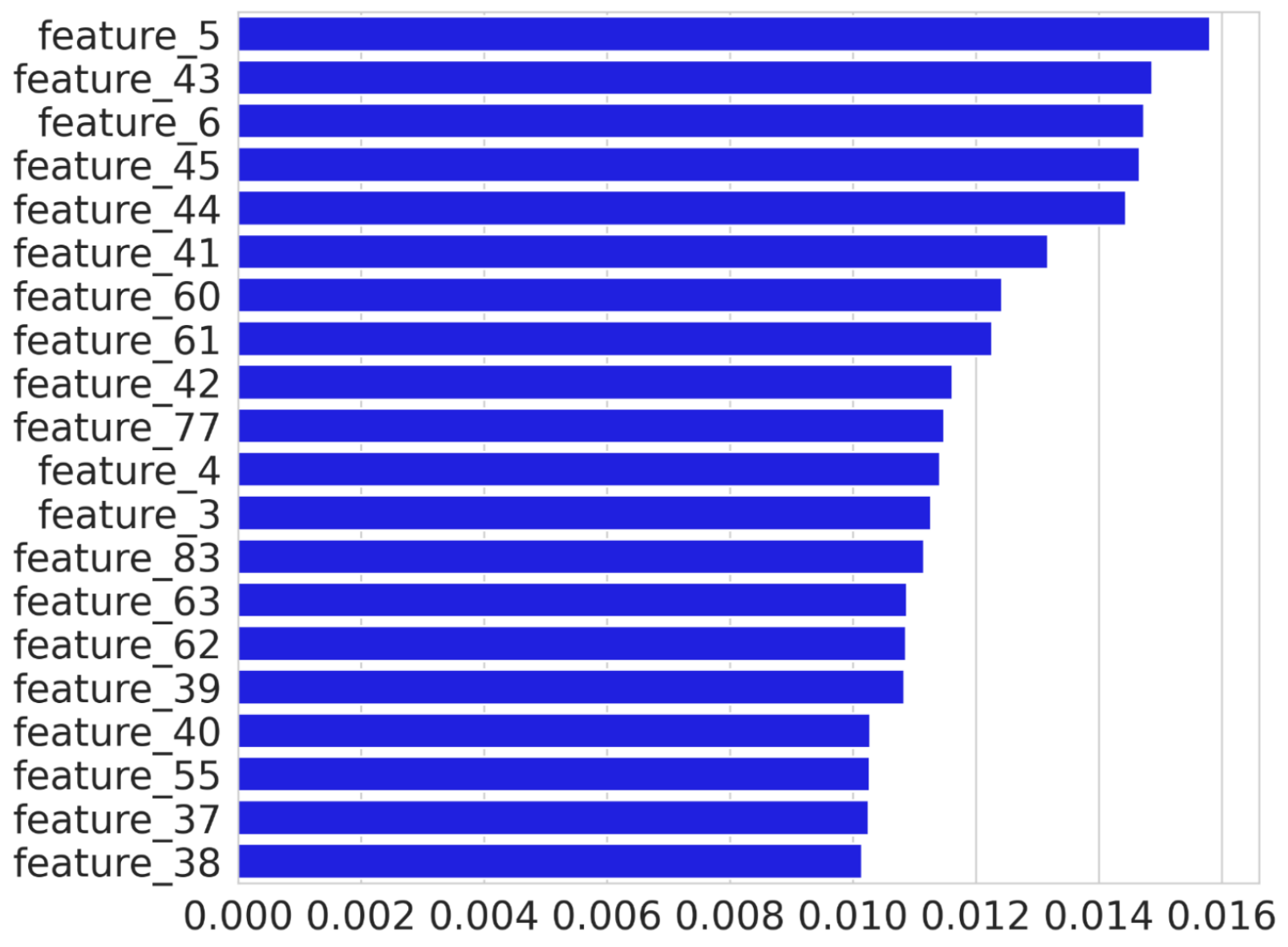


Fig.5 Feature importance ranking

4.Methods and Model

This study also **uses multiple model fusion as a solution**. In terms of feature engineering, logistic regression and correlation feature complementation are used to fill in the blank value. In addition, random forest is used to extract the feature importance and only the top 70% features are selected. In addition, for more EDA results, only the data after 85 days will be used as the training set. The final output will use a weighted average of the results of the neural network, XGBoost, and RNN with the embedding layer. The weights are 0.4, 0.41 and 0.19 respectively.

Result: accuracy of the scheme on the verification set is 0.57. (See Fig.6)

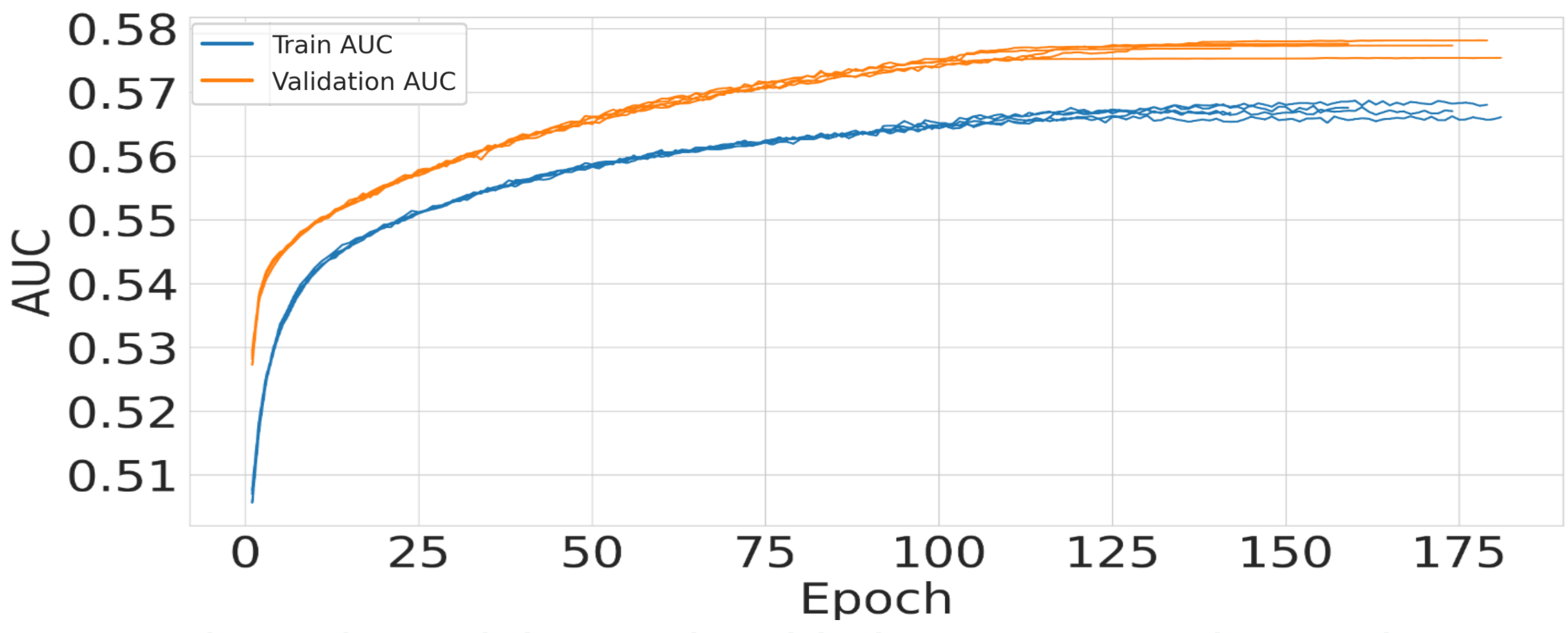


Fig.6 The training and validation curves of the model

References

[1] Carl McBride Ellis. Jane Street: EDA of day 0 and feature importance. https://www.kaggle.com/carlmcbrideellis/jane-street-eda-of-day-0-and-feature-importance

[2] Shahules. Jane : EDA & Feature Selection. https://www.kaggle.com/shahules/jane-eda-feature-selection