Data processing and modeling

dataset

The whole data set includes FX monthly trading data Including Underlying, Type, Strike, Expiry, Notional Curr Premium, Curr, Style, Code, Clr, Trd Time and minute-by-minute exchange rate change data, and is composed of 12 tables.

Data reading and processing

As a large amount of data is encoded in text, it is not conducive to model operation. Therefore, data processing is needed.

I processed the data in the following ways

1. Data encoding: two encoding methods are used. The distribution is encoded according to the number of occurrences and categories, and it is encoded into int or float type for easy model operation.
2. Detect missing values and duplicate values, delete missing values, and retain only the first duplicate value

In addition, I added four fields, namely the exchange rate data 1 minute, 30 minutes, 1 hour and 1 day after the transaction time. These data are calculated according to the month table and minute table.

methods

Converts a string encoding to a linux timestamp

Because the date data in the original table is represented by strings, it is difficult to calculate, and the date format in the Monthly table and the minute table is inconsistent.

I converted them into unix time stamps and matched them to get the exchange rate data of 1min,30min,1hour and 1day after the transaction time.

Range matching

Due to the lack of accurate period of exchange rate data, I used the range matching method to find the corresponding exchange rate data

Data analysis and feature processing

EDA

Preanalysis was performed using exploratory data analysis and visual toad library, as shown below.

	type	size	missing	unique	mean_or_top1	std_or_top2	min_or_top3
Underlying	int64	5500	0.00%	1	1.000000e+00	0.000000e+00	1.000000e+00
Type	int8	5500	0.00%	1	1.000000e+00	0.000000e+00	1.000000e+00
Strike	float64	5500	0.00%	605	1.031947e+00	5.382283e-02	0.000000e+00
Notional	int64	5500	0.00%	2746	4.288902e+07	5.414910e+07	1.010000e+02
Curr	int8	5500	0.00%	1	1.000000e+00	0.000000e+00	1.000000e+00
Premium	float64	5500	0.00%	4655	4.071870e+05	8.251832e+05	0.000000e+00
Curr.1	int8	5500	0.00%	2	1.752000e+00	4.318911e-01	1.000000e+00
Style	int8	5500	0.00%	2	9.989091e-01	3.301389e-02	0.000000e+00
Code	int8	5500	0.00%	3	1.969273e+00	2.342733e-01	0.000000e+00
Trd Time	int64	5500	0.00%	4339	1.667604e+12	7.885181e+08	1.666132e+12
1min	float64	5500	0.00%	627	1.002699e+00	7.466451e-02	-1.000000e+00
30min	float64	5500	0.00%	627	1.002699e+00	7.466451e-02	-1.000000e+00
1h	float64	5500	0.00%	627	1.002699e+00	7.466451e-02	-1.000000e+00
1d	float64	5500	0.00%	627	1.002699e+00	7.466451e-02	-1.000000e+00

Here you can see the quality and distribution of each initial indicator.

Then the correlation analysis of the index is carried out

heat map

Principle of thermal diagram

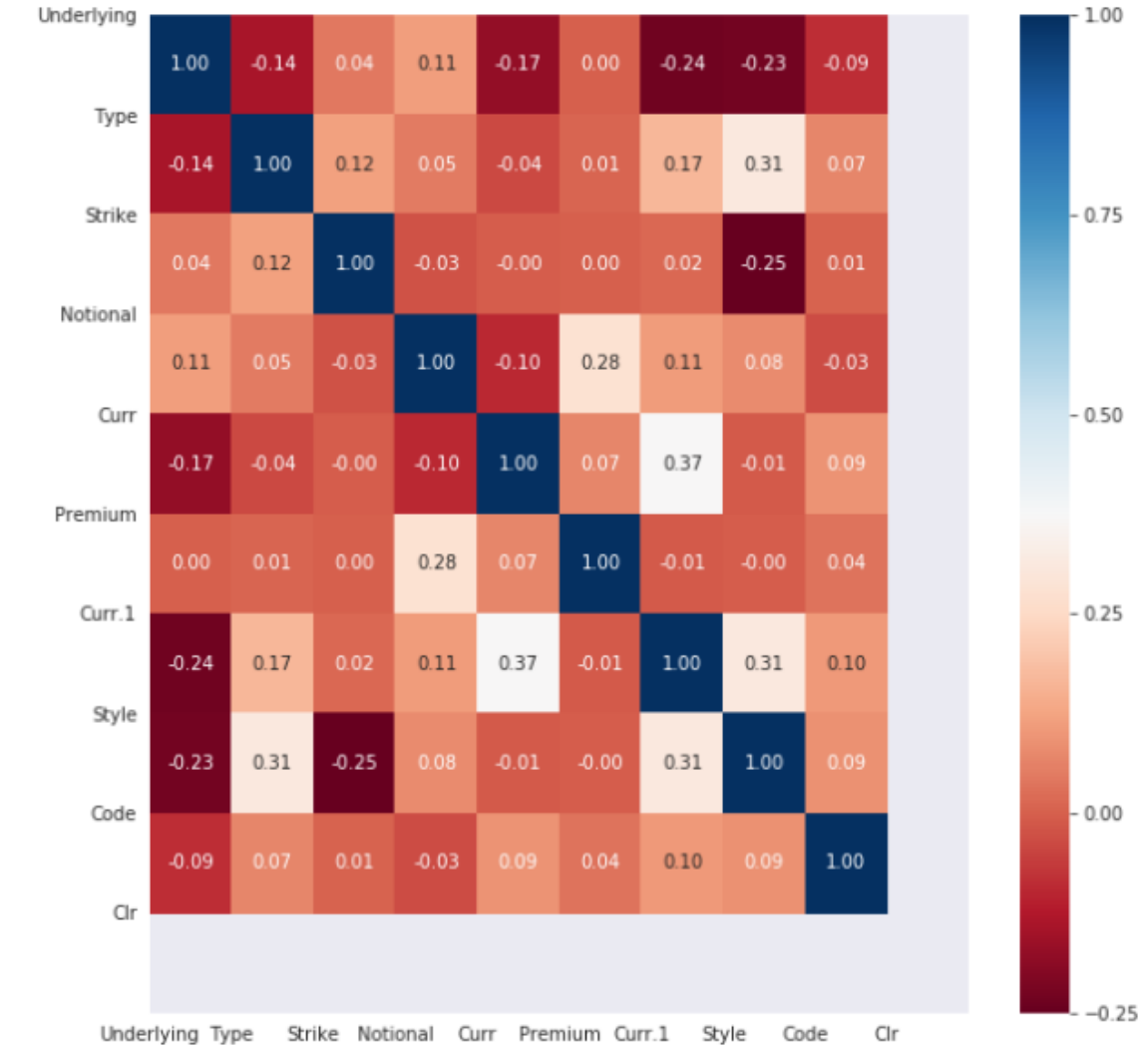
Thermal map, also known as correlation coefficient map. The correlation between variables can be judged according to the correlation coefficient of different square colors in the heat map. The calculation formula of correlation coefficient between two variables is as follows:

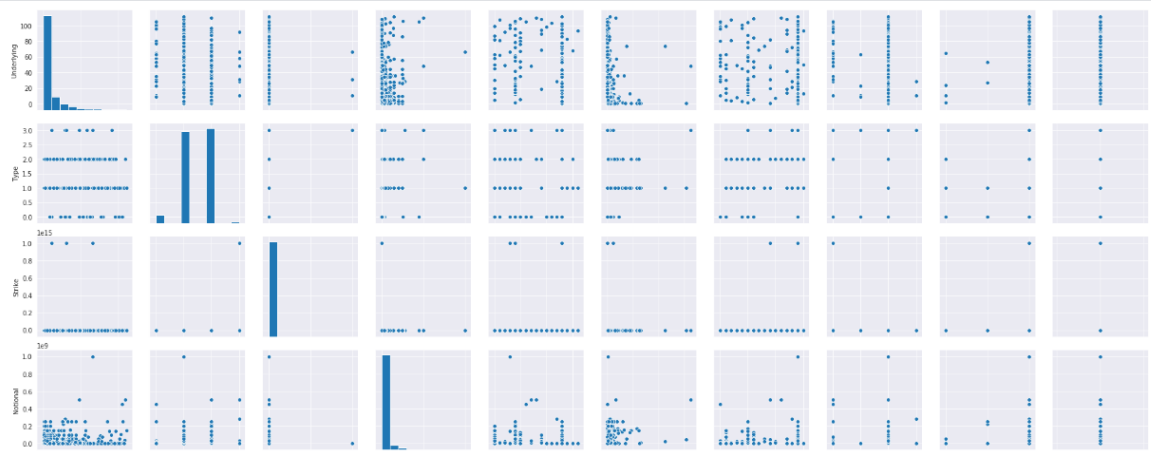
$$\rho_{X_1X_2} = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{DX_1, DX_2}} = \frac{EX_1X_2 - EX_1 * EX_2}{\sqrt{DX_1 * DX_2}}$$

In the formula, ρ represents the correlation coefficient, Cov represents the covariance, and E represents the mathematical expectation/mean

It is worth noting that the correlation coefficient can only measure the linear correlation between variables. In other words, the higher the correlation coefficient, the higher the degree of linear correlation between variables. For the two variables with small correlation coefficient, it can only show that the degree of linear correlation between the variables is weak, but it cannot show that there is no other correlation between the variables, such as curve relationship.

As shown in the results of the thermal map, no indexes with strong correlation were found, and there was no multidimensional collinearity.





Modeling

Four models were selected for prediction, namely SVM, KNN,RF and GDBT

SVM

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection

The advantages of support vector machines are:

> - Effective in high dimensional spaces.
> - Still effective in cases where number of dimensions is greater than the number of samples.
> - Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
> - Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

> - If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
> - SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

KNN

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

Random forests

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.However, data characteristics can affect their performance.

GDBT

GBDT(Gradient Boosting Decision Tree), also known as Multiple Additive Regression Tree (MART), is an iterative decision tree algorithm composed of multiple decision trees. The conclusions of all the trees are added up to give the final answer. When it was proposed, it was regarded as a generalization algorithm along with SVM. More recently, machine learning models for search ranking have attracted attention.

Introduction to Adoption Indicators

RMSE

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:

R2-SCORE

R2, also known as coefficient of determination, represents the degree of fitting of the model to the actual data. It is Peel

The square of the weak correlation coefficient, numerically speaking, R2 is between 0 and 1, the closer it is to 1, the better the fitting effect is, and above 0.7 indicates that

This figure shows that more than 70% of the data in the test set can be explained by the model, and the fitting effect is good, but there is still some space。

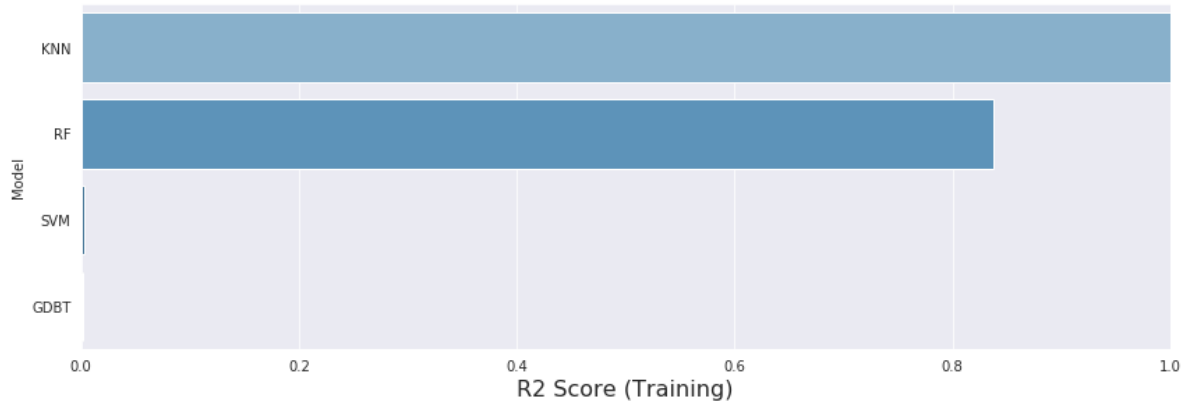
It is generally believed that the goodness-of-fit of the model exceeding 0.8 is relatively high, and the nonlinear model can be used for prediction。

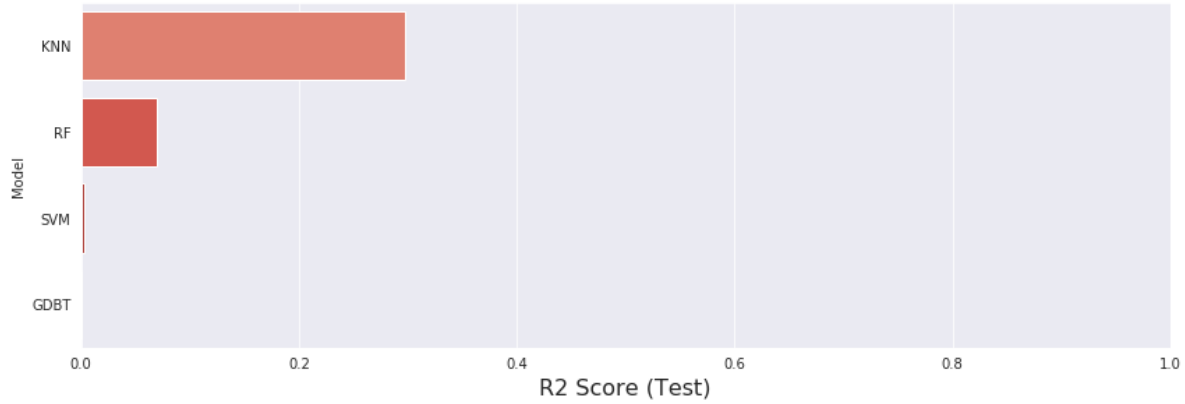
Process of training

I use Notional and Premium as training indexes, and conduct four sets of experiments with exchange rate data 1min, 30min, 1hour and 1day after the trading time as characteristic indexes.RMSE and R2-score were used as evaluation indicators.The data of usd-cad, usd-hkd, usd-jpy and other 6 tables were trained and tested.

results

Take EUR-USD data as an example.R2-score of four machine learning models is shown in the figure below. It can be seen that the R2-Score value of KNN and RF models is close to 1, indicating that the predicted value and the real value in the training set samples are almost exactly equal, and the error is small, indicating that the independent variable in the regression analysis has a better interpretation of the dependent variable. However, R2-score in SVM and GDBT models is close to 0, indicating that the numerator is close to the denominator, and each predicted value of the sample is close to the mean value.





More accurately, I combined the RMSE value to determine the performance of the model.The RMSE value of several models is relatively small. According to the rule of thumb, it can be said that the RMSE value is between 0 and 0.5, indicating that the model can predict the data more accurately。