Project 2b

Model Selection

For the stock price prediction of project 2b, the initial data I got were open, high, low, close, adj close price and volume, and I needed to predict the close price of the next day through these data within 60 days. I had noticed that these basic variables had high collinearity (e.g. close price and adj close price), which would lead to the failure of standard linear regression and polynomial regression, and seriously affected the analysis results. However, ridge regression solved the problem of collinearity better, it introduced a regular term. Although it introduces deviation, it greatly reduces the variance and is conducive to my prediction, so I finally choose ridge regression as my prediction model.

Data Preparation

I was provided with five kinds of data in 60 days to predict the close price of the next day, but if all the data were added to the regression, there would be many parameters and affected the prediction results, so I chose to use the basic data to generate four new factors, which were 60-days moving average, correlation, the difference between the open price of yesterday and today and the difference between close price of yesterday and the open price today. In addition, I also cleaned up the data to see if there were missing items and dropped some N.A. items.

Splitting Dataset

When dividing the training set and the validation set, I had made two attempts. The first was to use the function in sklearn which named train_ test_ split to divide two sets randomly, but I didn’t think this was reasonable, because the close price was a time series with strong order. If I randomly divided the training and validation set, I might use the later data to predict the previous close price, which was meaningless to me. Therefore, I chose the second conventional splitting method, taking the first 80% as the training set and the last 20% as the validation set, and took out all factors as X and closing price as y.

Regression

Since the hyperparameter λ required by ridge regression are unknown to me, I established an arithmetic sequence of 20 data from 0.01 to 5 as the value of hyperparameters λ. Then I used different λ to do the ridge regression. The predicted results were compared with the close price in the validation set, and the error level was represented by RMSE. The results were shown in Figure 1.

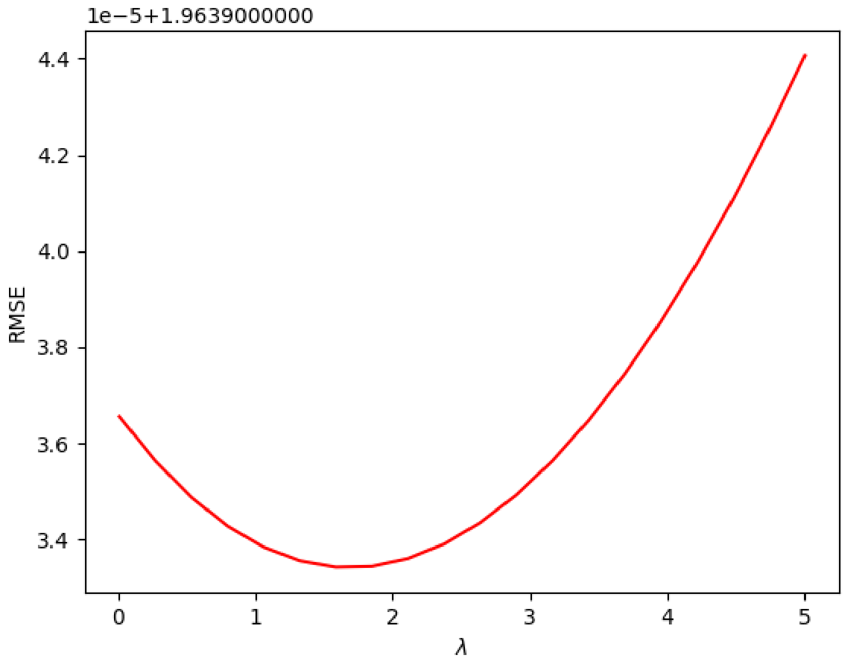

Figure 1 Ridge Regression for λ from 0.01 to 5

From Figure 1, we could see that RMSE showed a trend of first decline and then rise. The reason for the first decline was that with the increase of λ, the constraint of ridge regression on the result became smaller and smaller, so the error between the predicted value and the real value became smaller either. Then the reason for the rise was that the model was over-fitting, which would lead to the model being too similar to the training set, resulting in large prediction error, which should be abandoned.

Therefore, through observation, I finally selected the 7th point in the arithmetic sequence, where the RMSE was the smallest and represented the best prediction result, and λ = 1.59. The coefficients before each factor were shown in table 1.

Factors	**Coefficients**
Open	5.479820e-02
High	-3.015260e-01
Low	7.505161e-02
Adj Close	1.158554e+00
Volume	5.831755e-09
S_60	1.056031e-02
Corr	4.278974e-03
Open-Close	-4.727616e-02
Open-Open	-2.204034e-02

Table 1 The factors and their coefficients in Ridge regression(λ=1.59)

Conclusion and deficiency

Through this project, I used ridge expression to predict the close price of the stock tomorrow. But I still think there were some areas that need to be improved.

First, the ridge regression would minimize the coefficient of each factor to reduce the impact of a single factor on the prediction results, however, we could see from table 1 that the coefficient of adjusted close price was large, which might have a great impact on the prediction results.

Second, in terms of factor selection, 60 days moving average and coefficients might ignore the sharp fluctuation of close price in the short term, which might affect the prediction results.

Third, the selection of λ values was relatively rough. This project only selected 20 λ values, which may be different from the most suitable λ in Ridge regression.

Finally, it was worth thinking about the division of training set and validation set. In the future, we can try some cross-validation methods, such as k-fold cross-validation, leave one out cross-validation and time series cross-validation, which may have better prediction results.