Data_Science_2_Midterm

Qimin Zhang, qz2392

3/13/2020

Introduction

Financial market prediction has always been a field under heat. Over the recent years, researchers, investors, and managers have dedicated in developing models for forecasting the stock market behavior. With the emerging of big data and the increase in computing power, the trend continues. One of the main challenges of stock price prediction is that they are affected by highly correlated factors, while there could be hundreds of different financial indicators. Moreover, factors such as politics, psychology, and government interference are hard to be quantified and used in the existing models. Another challenge is the incompleteness of the data, where the missing parts were denoted as NA's or 0s.

The dataset used in this project is from the 2017 US stock market price with more than 4000 stocks and over 200 commonly used financial indicators. The price var [%] will be used as the response. Variable class indicates if the stock is worth-buying or not. To clarify, the reponse represents the stock price variation of year 2018: if positive, it means that the price is higher at the end of year 2018, so a buyer should consider buy the stock at the begining of 2018 and sell it for profit at the end of the year.

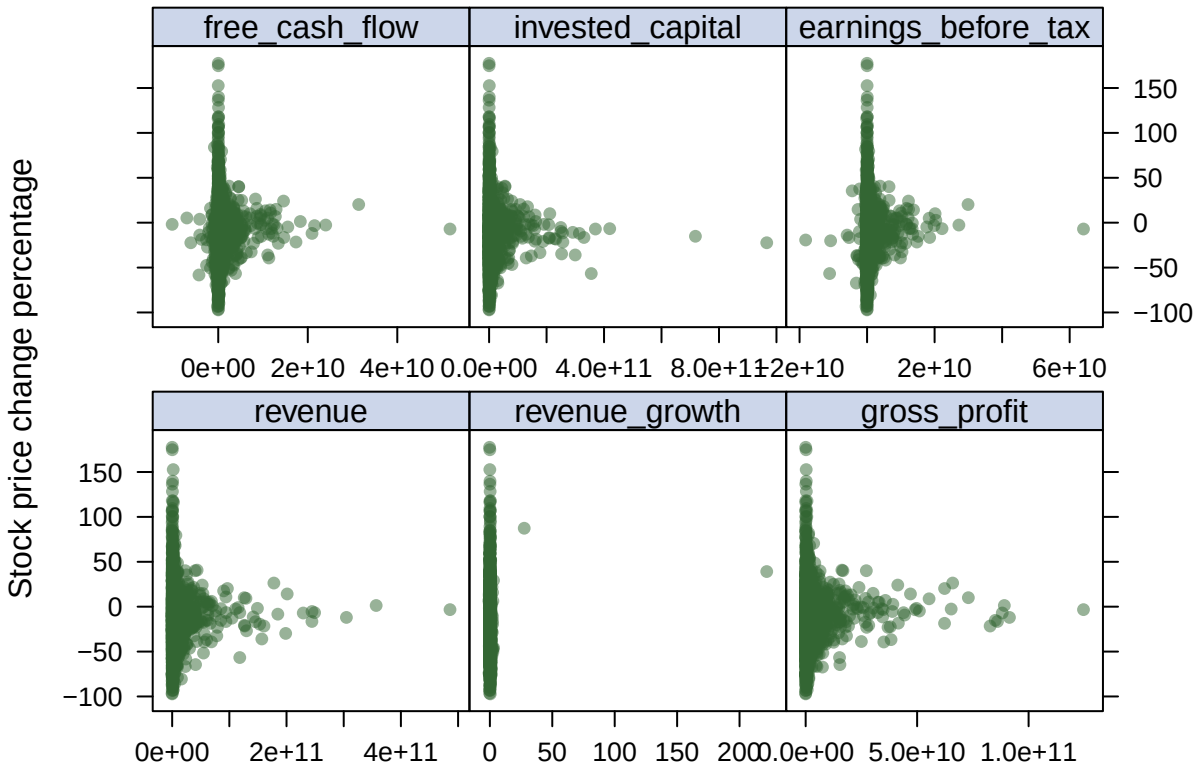
For the highly correlated factors, we use a series of models with tuning parameters to fit the data, and adjust the parameters with cross-valiation. To deal with NA's and 0's, we romove a row if there are more than 20% NA's. After the romoval of those rows, we fill all NA's with 0's, and remove a row if there are more than 20% 0s.

Exploratory analysis/visualization

Deal with NA's. Delete a row if it contains more than 20% NA's. Fill NA's with zeros, and delete a row with more than 20% 0s.

Now there are only 2123 rows in the dataset.

Since there are too many variables, we select some to find interesting structure present in the data



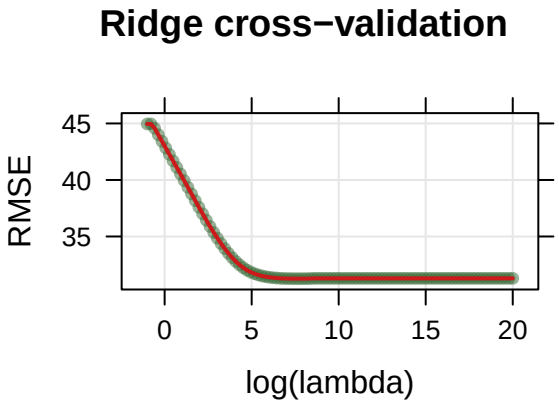
We can see that there is no clear pattern between price change and the variables selected. Many of the variables still contain lots of 0s.

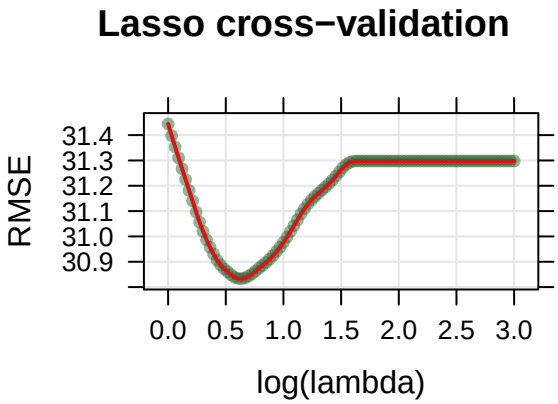
Models

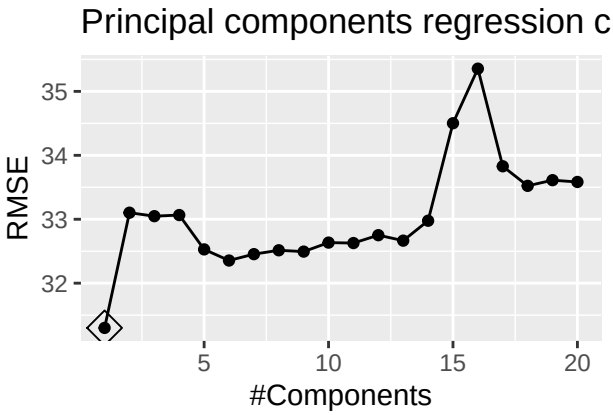
Split the dataset into training set and test set. Take 10% as test set.

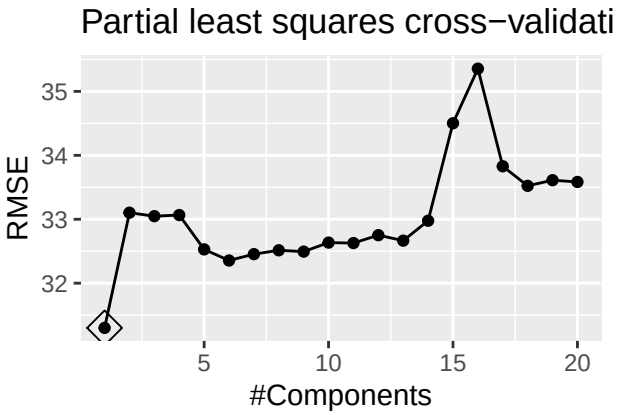
We use ridge regression, lasso regression, elastic net, principle component regression, partial least squares and multivariate adaptive regression splines. For the model fitting, all variables are included. Tuning parameters are chosen by 10-fold cross-validation, repeated for 5 times.

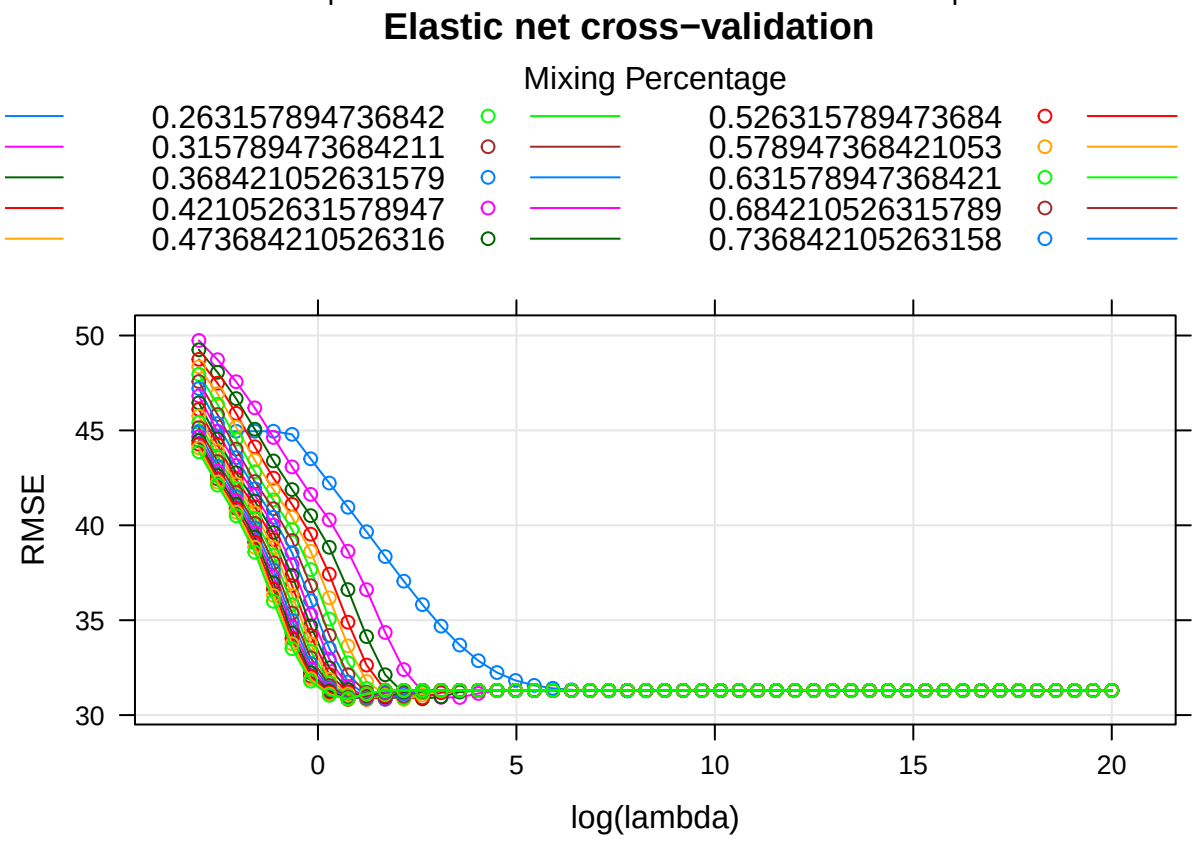
Results

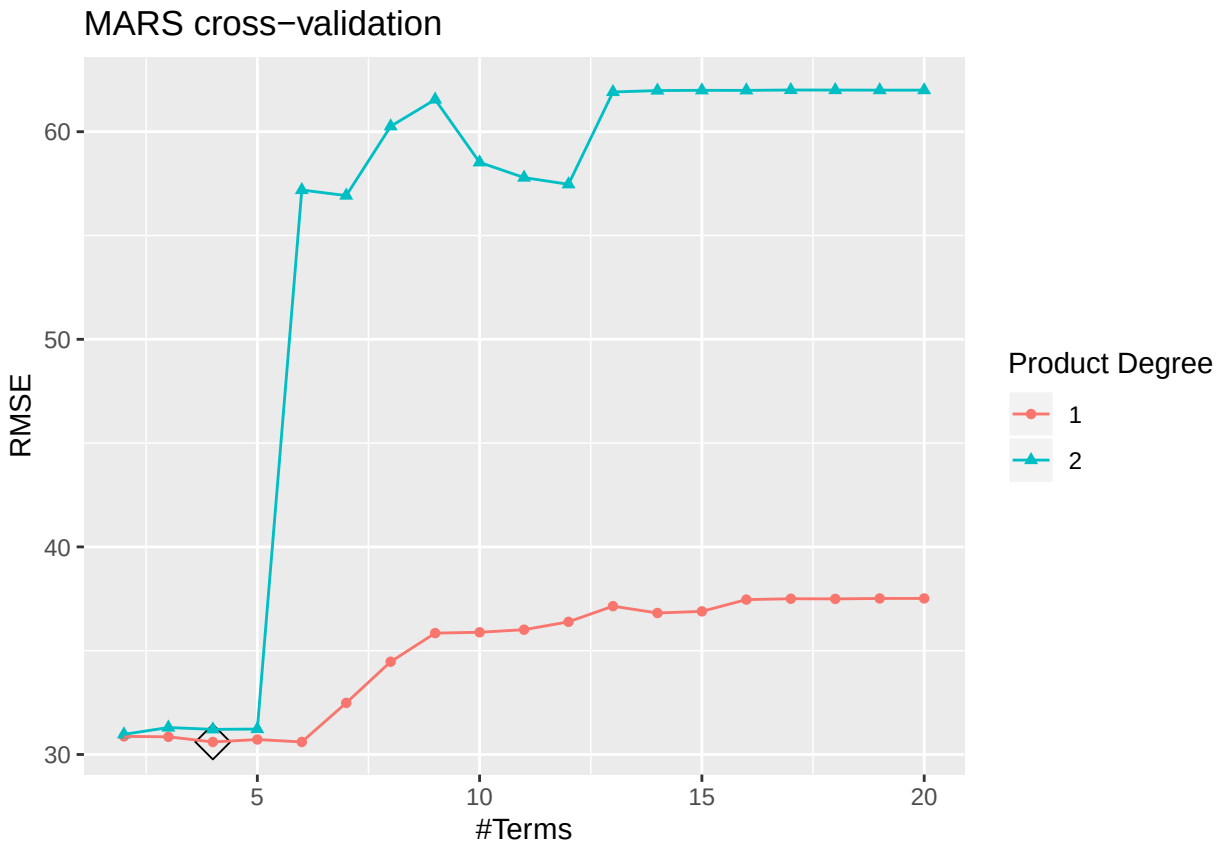

Ridge cross−validation
RMSE
45
40
35
0
5
10
15
20
log(lambda)

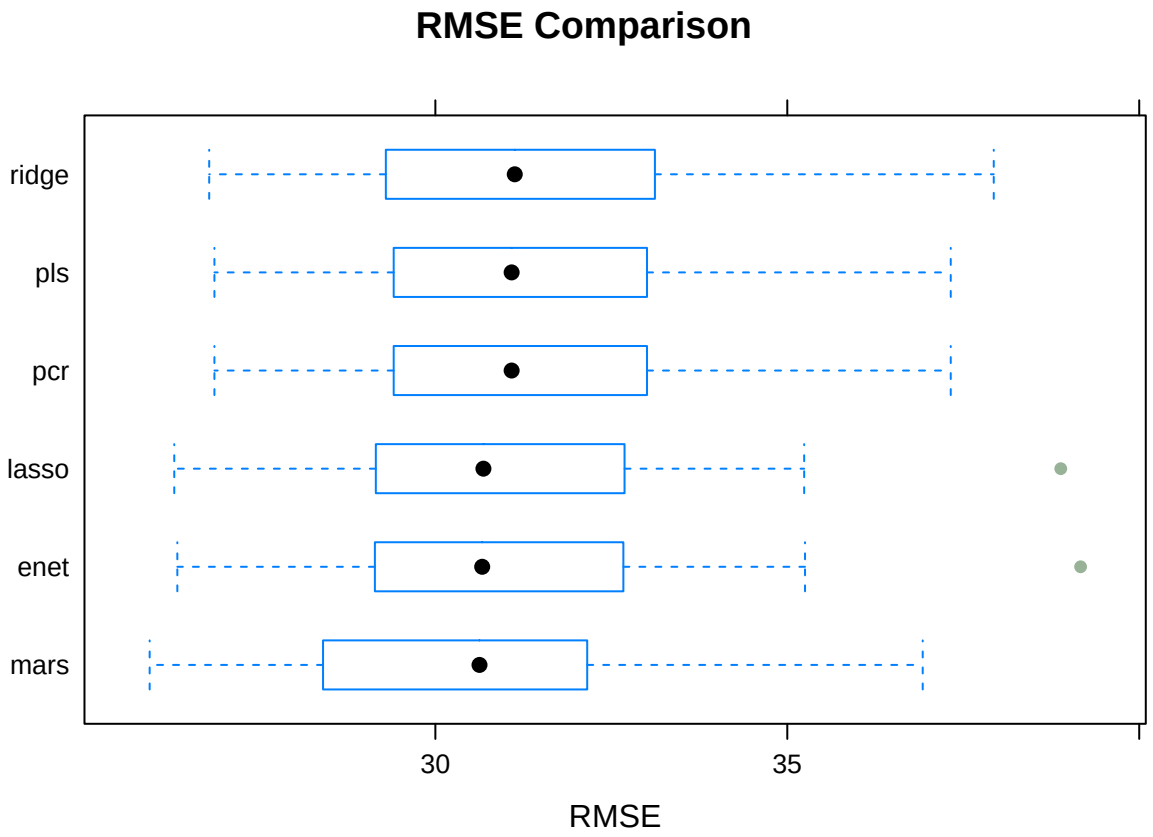

Lasso cross−validation
RMSE
31.4
31.3
31.2
31.1
31.0
30.9
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(lambda)


Principal components regression cr
RMSE
35
34
33
32
5
10
15
20
#Components


Partial least squares cross−validati
RMSE
35
34
33
32
5
10
15
20
#Components


Elastic net cross−validation
Mixing Percentage
0.263157894736842
0.315789473684211
0.368421052631579
0.421052631578947
0.473684210526316
0.526315789473684
0.578947368421053
0.631578947368421
0.684210526315789
0.736842105263158
RMSE
50
45
40
35
30
0
5
10
15
20
log(lambda)





In terms of RMSE, ridge, PCR and PLS share similar results, and lasso and elastics net have similar results

which are better than the previous 3 models. MARS has the best result so it's chosen to be the final model. Here is the model details:

```
##                               (Intercept)                  h(0.6574-gross_margin)
##                                 2.5103097                             -25.3946661
##  h(0.0643-x3y_revenue_growth_per_share) h(19.6039-enterprise_value_over_ebitda)
##                               -41.4469027                              -0.6751367
```

So the final model is:

Price change in percentage = 2.51 - 25.39 * h(0.6574 - gross margin) - 41.45 * h(0.0643 - x3y revenue growth per share - 0.68 * h(19.6039 - enterprise value over ebitda)

```
## [1] 26.07695
```

The RMSE on test set is 26.0769491, meaning that there is 26.0769491% RMSE in stock price change.

Conclusions

After cross-validation, we choose MARS as our final model, while there is 26.0769491% RMSE in stock price change on test set with MARS, which is not a good result.

Intuitively, we expect variables like 'revenue', 'revenue growth' or 'gross profit' remain in the model, but according to the results, these variables seem not so contributive to stock price change. We can see that all the models tend to have few viriables than 200+, or shrink the parameters to very small values. This indicates that the available predictors may not contribute a lot in the prediction of stock price change, maybe because there are too many 0's, or the predictors are highly correlated. Moreover, we can see a clear 'U-shape' in the lasso plot.

In general, the prediction of stock price change with machine learning methods still remain challenging without complete data with some key factors involved, which may not be easily collected due to highly unstable human behaviors and emotions.