Midterm Project Report

wq2151

Contents


```
# import data
original_data = read_csv("./data/2018_Financial_Data.csv") %>%
  janitor::clean_names() %>%
  rename(price2019 = "x2019_price_var_percent",
         company = "x1")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   X1 = col_character(),
##   Sector = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

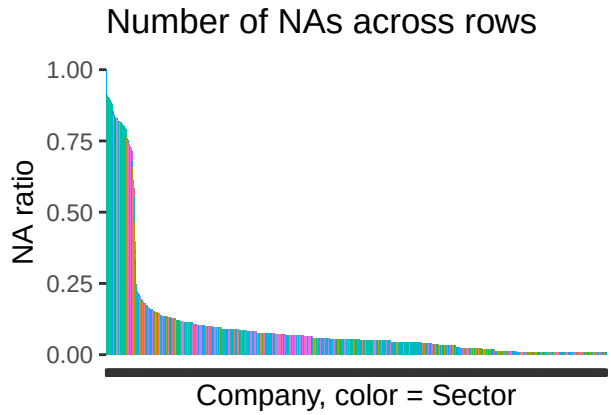
Introduction

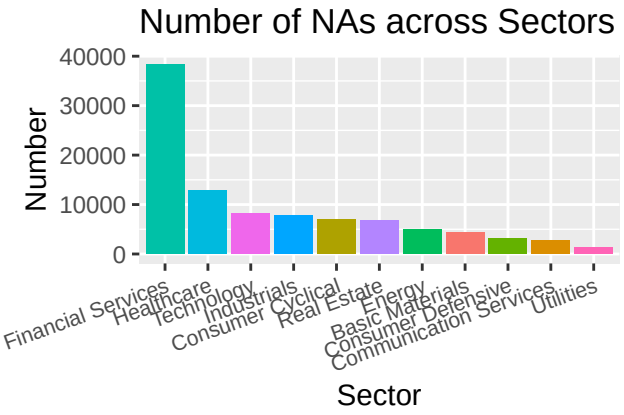
Financial market prediction has always been a field under heat. Over the recent years, researchers, investors, and managers have dedicated in developing models for forecasting the stock market behavior. With the emerging of big data and the increase in computing power, the trend continues. One of the 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.

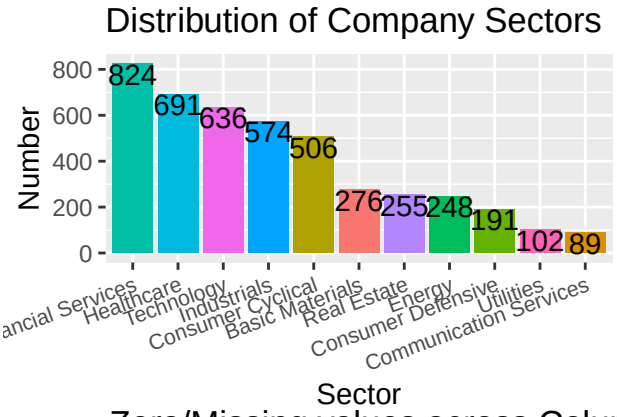
The dataset used in this project is from the 2018 US stock market price with more than 4392 stocks and 222 commonly used financial indicators. The `price var [%]` will be used as the response. Variable `class` indicates if the stock is worth-buying or not. To clarify, the reponse represents the stock price variation of year `2019`: if positeve, it means that the price is higher at the end of year `2019`, so a buyer should consider buy the stock at the begining of `2019` and sell it for profit at the end of the year.

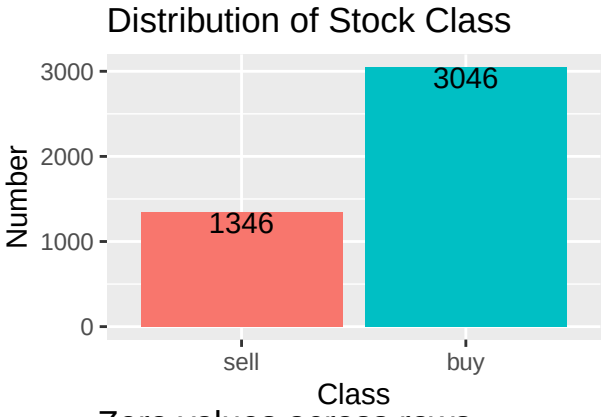
Exploratory Data Analysis

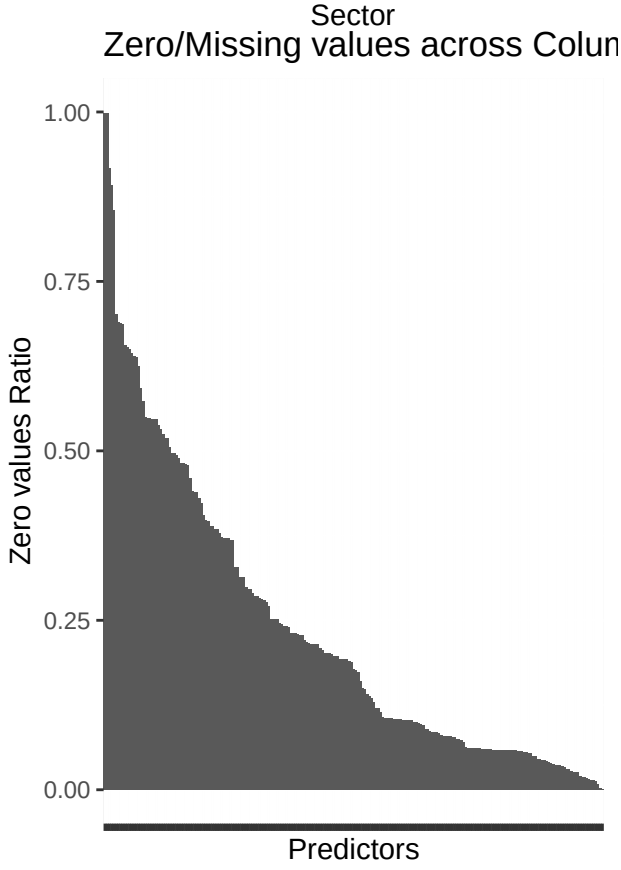
First, the dataset contains huge amount of NA valus, which should be removed or filled with 0 value; `Financial Service` companies has the most NA values; Most of the stock perform well from a trading perspective; Also, from the dataset we can see that some values are 0 where it couldn't be zero in the normal sense; for example, the R&D expense of GE is 0 (which is not correct). Therefore we assume that NA are the same as 0 in this dataset, produced by accounting/financial report errors.

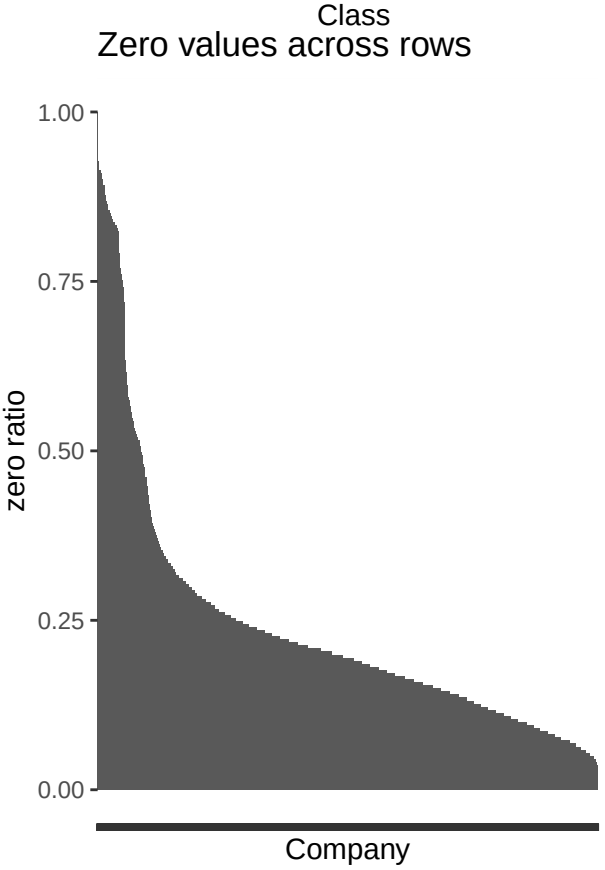

Number of NAs across rows
NA ratio
1.00
0.75
0.50
0.25
0.00
Company, color = Sector


Number of NAs across Sectors
Number
40000
30000
20000
10000
0
Financial Services
Healthcare
Technology
Industrials
Consumer Cyclical
Real Estate
Energy
Basic Materials
Consumer Defensive
Communication Services
Utilities
Sector


Distribution of Company Sectors
Number
800
600
400
200
0
824
691
636
574
506
276
255
248
191
102
89
Healthcare
Technology
Industrials
Consumer Cyclical
Basic Materials
Real Estate
Energy
Consumer Defensive
Utilities
Communication Services
Sector


Distribution of Stock Class
Number
3000
2000
1000
0
3046
1346
sell
buy
Class


Zero/Missing values across Colun
Zero values Ratio
1.00
0.75
0.50
0.25
0.00
Predictors


Zero values across rows
zero ratio
1.00
0.75
0.50
0.25
0.00
Company

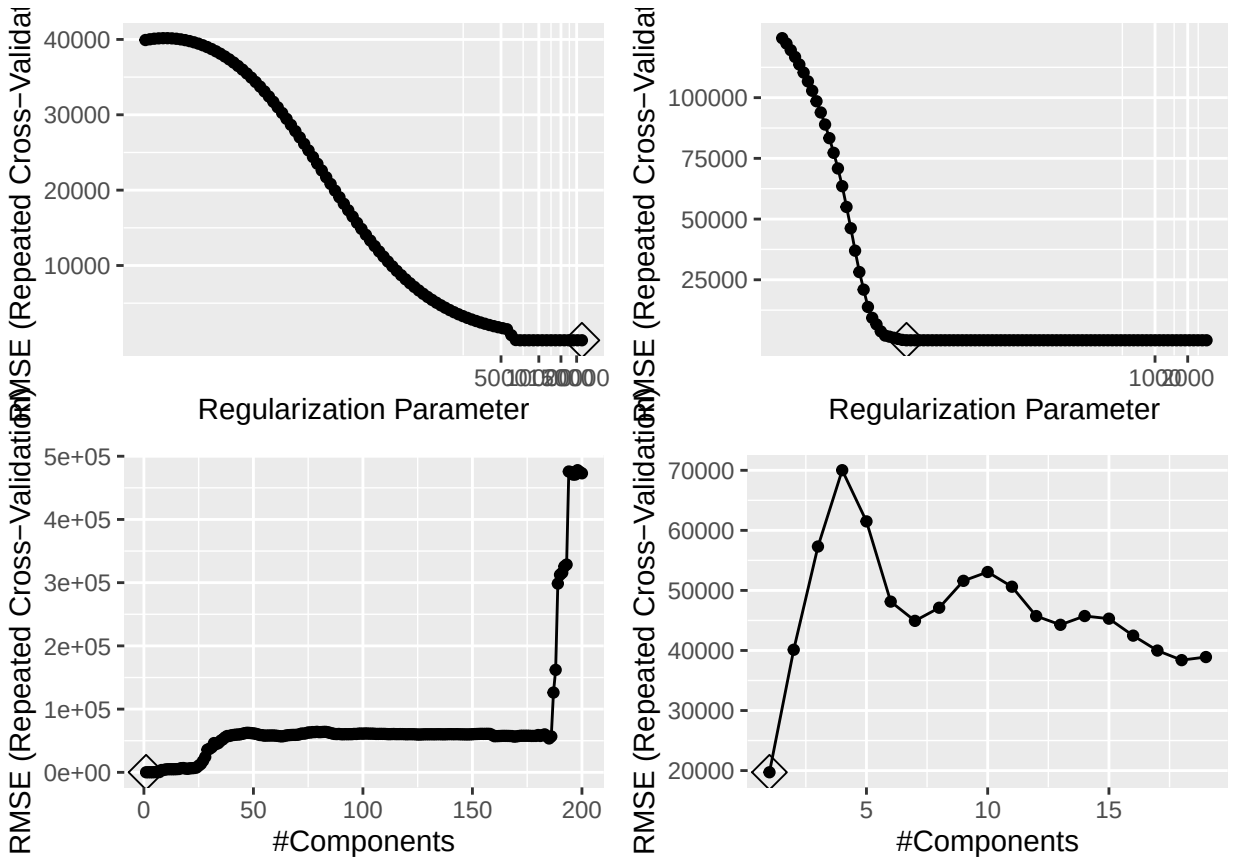
Methods

Several different regression models were used to predict the stock price variation. `Ridge` regression uses L2 penalty term while `Lasso` regression uses L1 penalty term, which might cause `lasso` regression has less coefficients than `ridge` regression model. Elastic net use lasso penalty to select deature but use regularization via ridge-type penalty. Principle component regression (PCR) and Partial least square (PLS) methods are also used to fit the data.

The data is split into 2 part, one for training (75%) and one for testing (25%). the test data will be later used for evaluating the model performance.

Results

Lasso, ridge, enet have close RMSE results. The `ridge` model basically shrink most of the coefficients towards zero, but not exacty zero; the lambda is not optimal as the range of lambda goes wider, the best lambda go up to the right bound of the range. It indicates that the bias of the ridge regression is significantly high, meaning that the model is underfitting the data. On the other hand, `lasso` produces a relatively small lambda 1, 5.13645748377035 as it forces some of the coefficients to be zero. As the lasso is more indifferenet to very correlated predictors, its better performance on highly correlated financial data is not unexpected. Overall speaking, `ridge`, `lasso`, and `elastic net` has close results, it is hard to determine which one is dominantly better than one another. Traditional linear model has the worst performance in predicting the response, which might be caused by the high correlation of predictors and a lot of 0 values. PLS can be viewed as a supervised learnign procedure while PCR is an unsupervised procedure. The PLS model has significant higher MSE than other models (excluding the linear model). To conclude, in order to better predict the stock market variation, it better to choose `lasso` or `elastic net` methods.



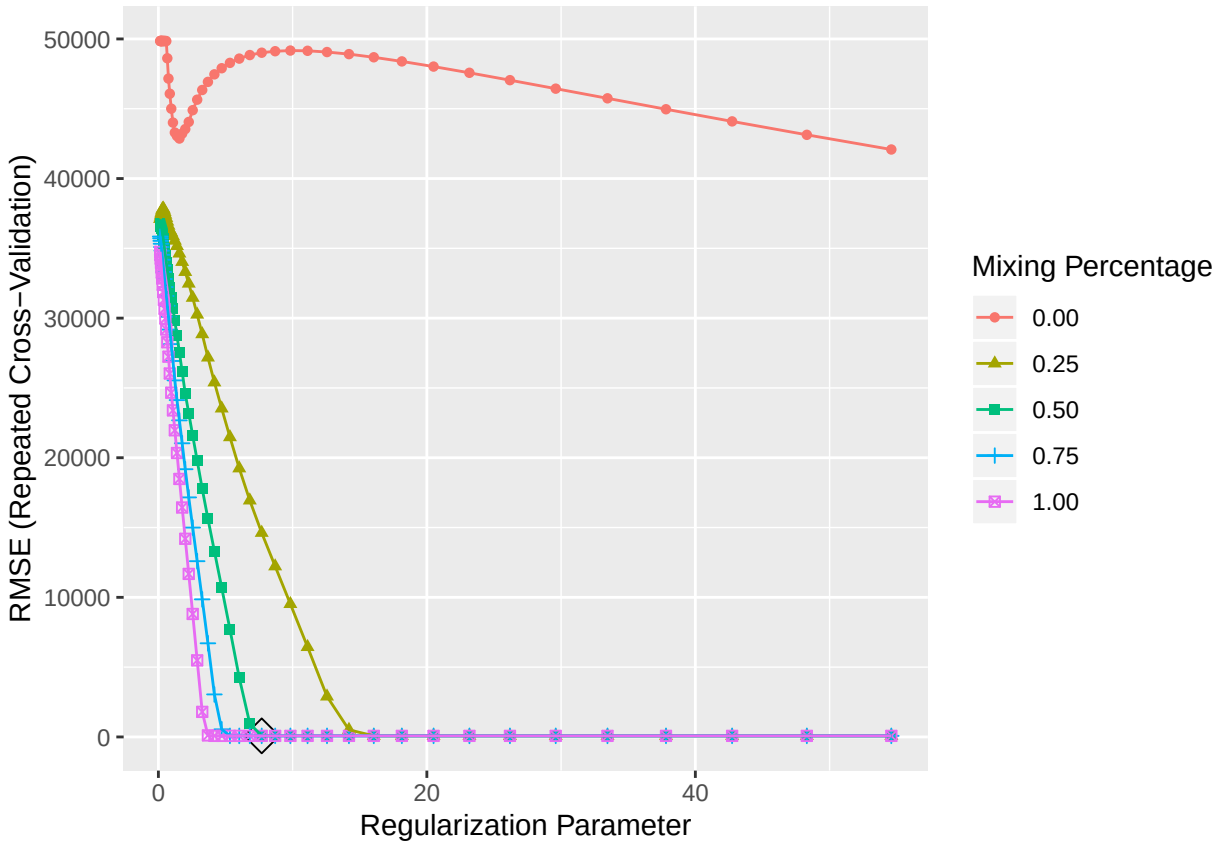
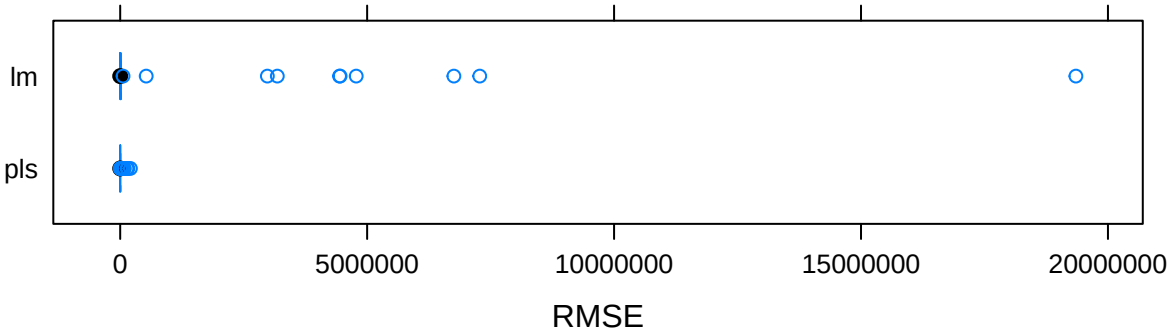
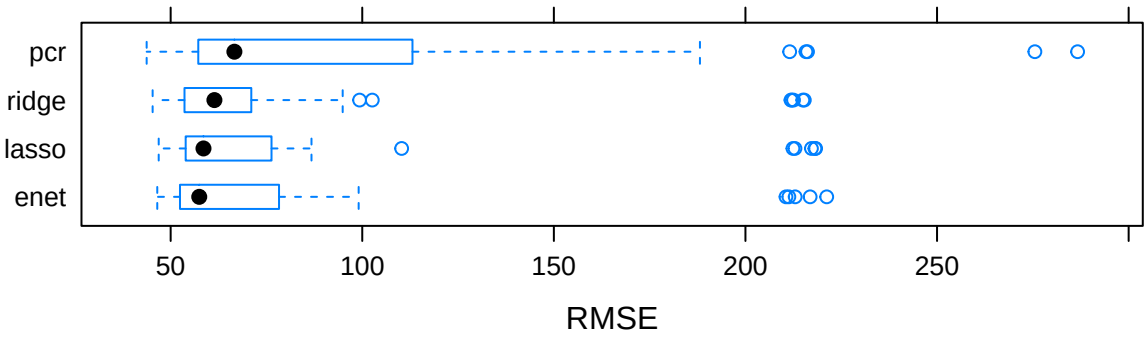


Table 1: MSE of different models

lm.mse	ridge.mse	lasso.mse	enet.mse	pcr.mse	pls.mse
7074569	2543.451	2529.431	2515.184	2597.527	104274.2





Limitations

The dataset has been shown to contains too many zero values and missing values. Some columns (predictors) are over 90 percent zero/NA dominant. It is a big challenge in dealing with these cells, as they will not be able to represent the real stock market data. Second, the MARS and GAM model are not used as all the RMSE metrics are missed in the model fitting process. Better approaches might exist in filling the zero or missing values to get a better analysis results and prediction.