Industry Stocks Returns Prediction

Course: Big Data in Finance
Xiaoyan Zhou
CID:01384662

2018-04-23

1 Introduction

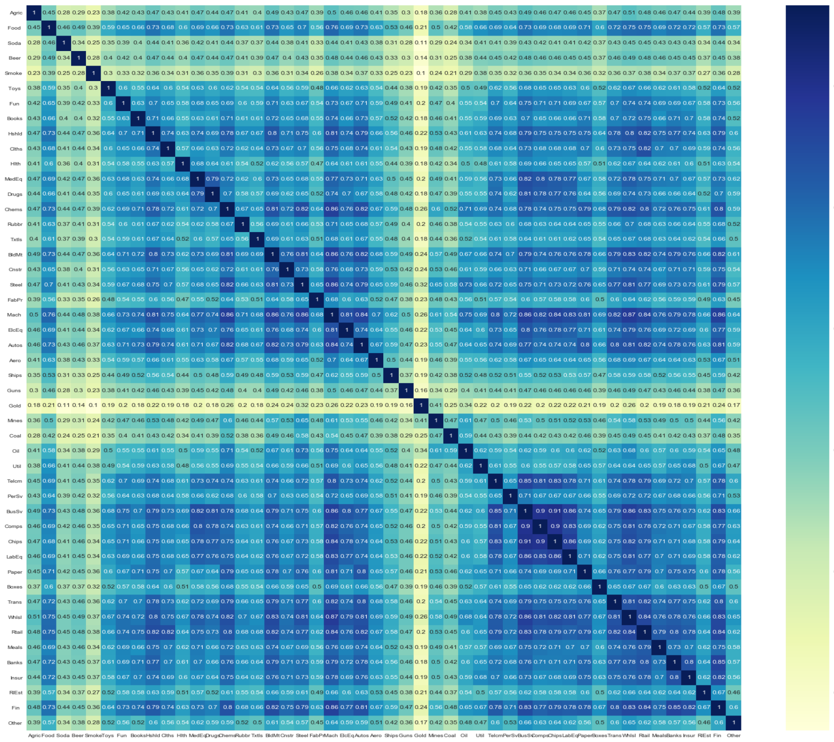
This report aims in comparing the outcome of 5 methods, including historical mean methods, linear regression method, lasso regression method, random forest method and support vector regression method to predict the change of stock returns in 3 industries. Besides, to find out whether lag effect shows up ins different industries and different methods, this report also uses different previous days' returns–lag1 day, lag 30 days and lag 360 days, of all industry to predict the stocks returns of 3 industries.

2 Data Description

The industry portfolios dataset analyzed contains 48 industries and daily returns of industry stocks from July 1, 1926, to February 28, 2018. The total number of observations it provides is 12196.

This report selects three target industry, including Agriculture, Pharmaceutical Products, and Computers, to do further analysis. The reason why choosing these 3 industries is to find out whether the lag effect is significant in different sector, considering computer industry may have higher lag effect as it may stockpile production material in advance, pharmaceutical product industry may have little lag effect as it mostly relies on technology, and Agriculture industry may be in between of this two industries.

By looking into the heat map indicating correlation between different industry,if we say the correlation above 0.7 is high, then it can be found that Agriculture has week correlation with other industries, while the correlation between drugs and other industry is moderate, being strongly correlated with health industry, business services, computers and electronic equipment, and computer industry is highly correlated with even more industries, including business services, chips, machinery, etc.



3 Prediction Method and Outcome

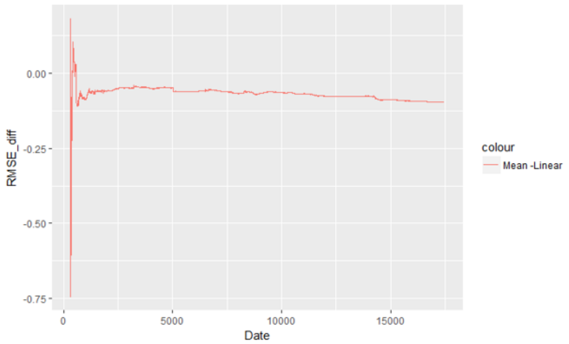
This report uses five methods to make industry stock return prediction, in which historical mean method is applied to all other method show whether that method of performance historical mean method. For all model, this report uses RMSE to show whether a model is good or not, the higher the RMSE, the worse the performance of that model.

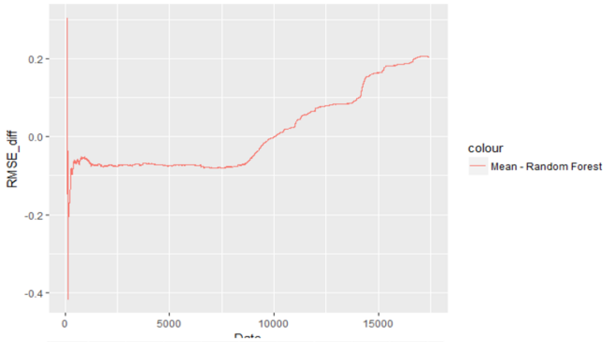
For each model, this report uses cross-validation to find optimal window size and model specific parameters. For lasso regression, this report uses cross-validation to find the optimal lambda, which leads to the lowest mean square error. For the random forest, this report also uses cross-validation to find optimal minimum node leaf, while for support vector regression method, this report uses Gaussian kernel, and use cross-validation to find the optimal cost.

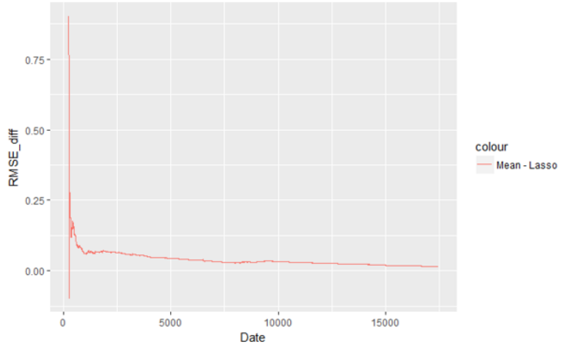
For linear regression, the window size sequence to test is from 140 to 360, by 40. For lasso regression, the window size to test is from 100 to 400, by 50. For the random forest, the testing window size is formed 100 to 350, by 50, and the leaf node to test is from 10 to 30, by 10, and each time grows only five trees. For support vector machine, the window size to test is from 300 to 400, by 30, and in the tuning process, the cost set to test contains 0.1, 0.5 and 1.

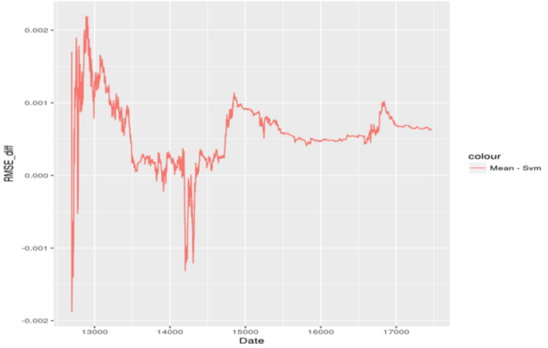
To compare the outcome of different methods in different industries, the last 1 day's return of the industry and last 1 day's return of other industries are used as independent variables. Besides, linear regression, lasso regression, and random forest methods have been applied to the whole data set, while support vector regression method has been used to the recent ten years'data to reduce the computational complexity.

3.1 Agriculture Industry



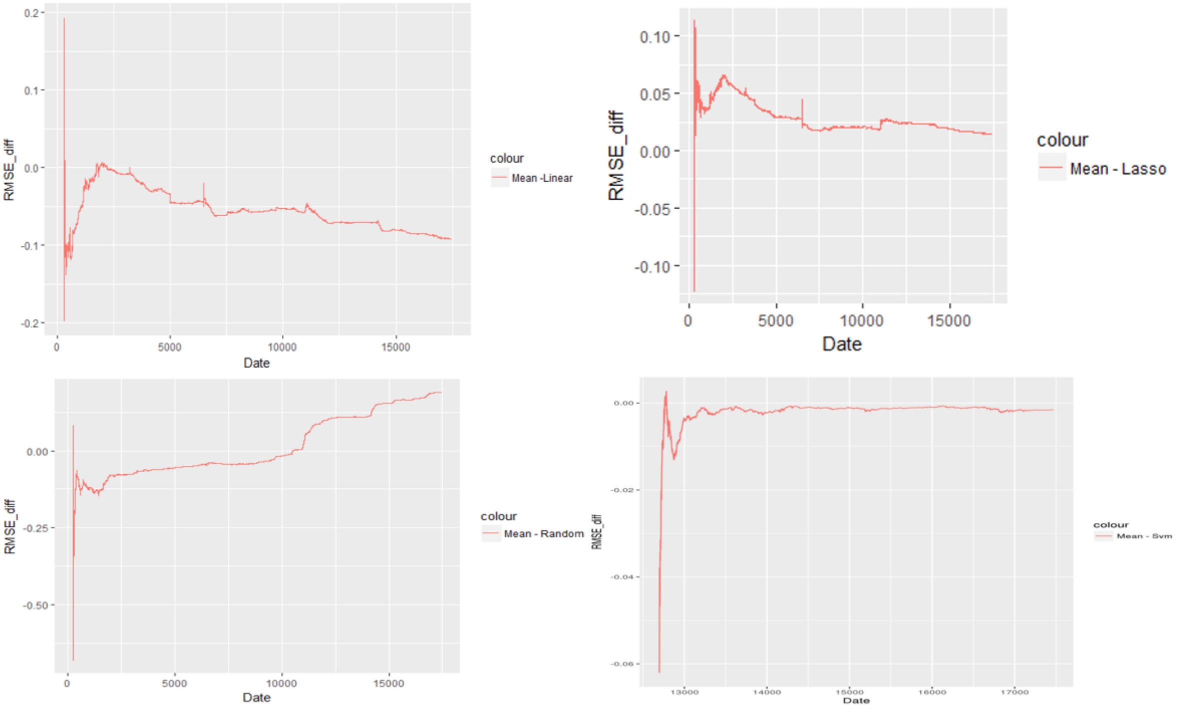






For Agriculture industry, when only lag one day, the prediction RMSE difference is most significant when applying random forest method, while in the same time, the RMSE (1.227) of using random forest is lowest too. The support vector regression performance is the worst, with RMSE being 1.596.

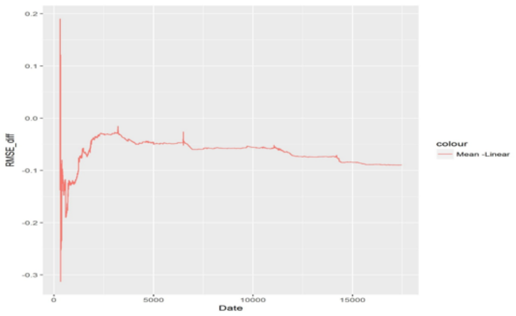
3.2 Pharmaceutical Products Industry

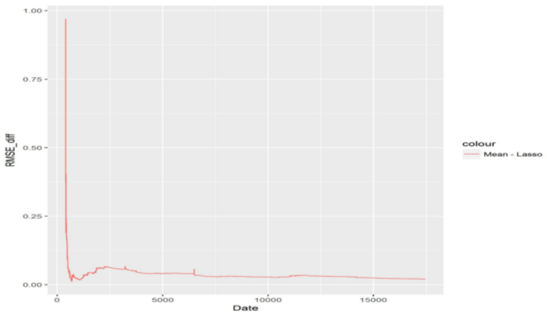


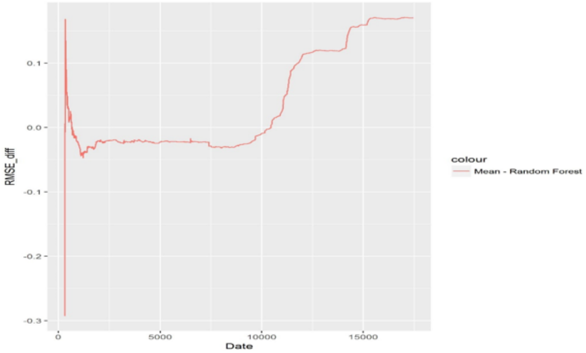
In Drugs industries, random forest method performs the best, with the RMSE being 1.0739, and the RMSE difference is also the highest among all four models, which is 0.1887. Support vector regression model's performance is the worst in this case, with the RMSE being 1.432, and the RMSE difference is also the lowest, which is -0.0016.

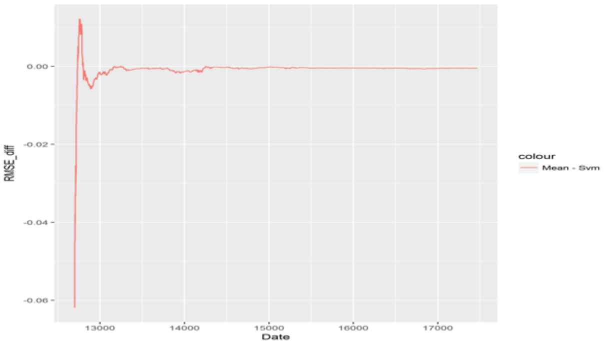
Lasso regression model's performance is better than historical mean model too, and it's RMSE, 1.2467, is the second lowest. Linear regression model's performance is worse than historical mean method.

3.3 Computers Industry









In the computer industry, random forest model is also the most useful model among this five models, with the RMSE difference being 0.17, and the RMSE being 1.111. Lasso regression model also beat the historical mean model in this industry, with the RMSE difference being 0.02.

The worse performance model in this industry is linear regression model, and the support vector regression model ranks the second.

3.4 The Best Model in Lag 1 Day Dataset

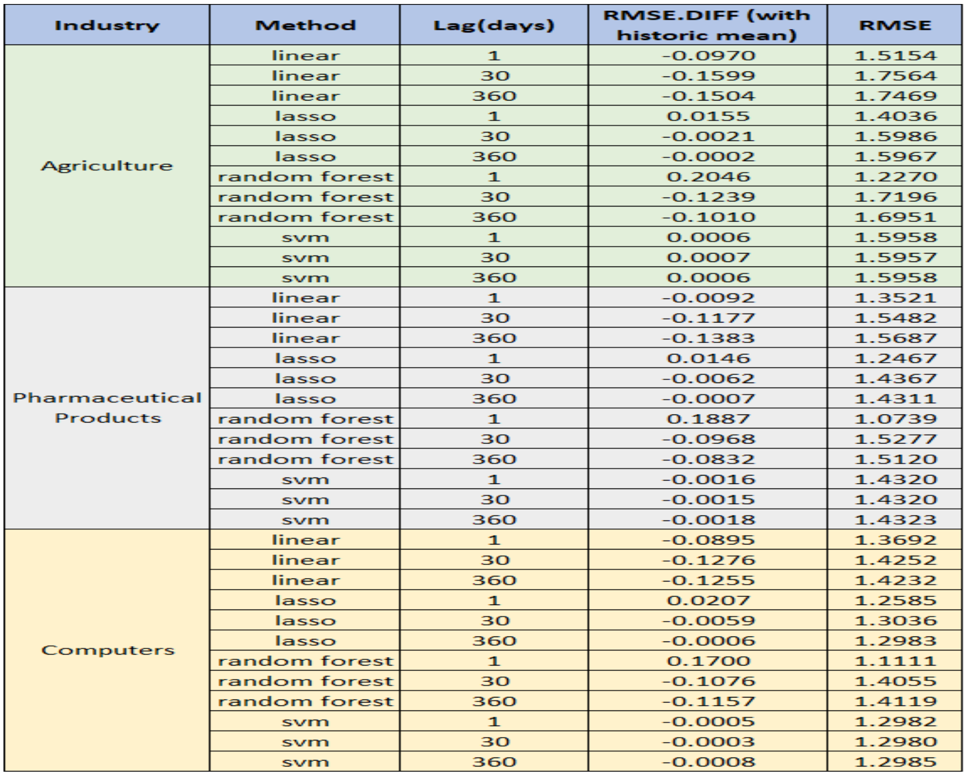
The analysis above shows that the random forest method is the best among all five model, and it can be applied to all three industries chosen. The optimal window size for different sectors used in random forest method is 200,300, and 350 for agriculture, pharmaceutical products and computers industries respectively, and the minimum leaf node are all 30, chosen by cross-validation.

Different industries use different random forest window sizes; it can be a result of that different industry have different circle period, during which the daily stocks return changes pattern is similar.

4 Lag effects Comparision

Using how many previous days to predict stocks return change is an important issue to decide, as the influence of other industry may have a substantial lag effect to some specific industries, while to other industries, the impact is not significant even mild. Consequently, this report compares the lag one day, lay 30 days, and lag 360 days dataset to predict future industries stocks return, to see whether for the chosen three industries, whether lag effect from other sectors is significant or not.

The final result of the analysis is presented in the following table, in which it can be concluded that for all three industries, and for each model, the performance of lag one day is the best. However, it is worth noticing that for all three industries, the performance of lasso regression model is significantly improved when changing dataset form lag 30 days to lag 360 days. It can result from that lasso model selects essential industries as variables to do prediction, so it tends to pick variables with a long-term effect, which can be shown in 1 years. The reason why lag one days performance is still better than lag 360 days is that lasso is selecting different variables in different lag datasets. In lag one day's dataset, lasso model selected the industries to have the most strong contemporaneous effect toward targeting industries, while in lag 360 dataset, lasso model chooses the industries to have the strongest lag effect on the targeting industries. However, using the contemporaneous impact to predict industry stocks return changes can be more effective in this three industries.



Industry	Method	Lag(days)	RMSE.DIFF (with historic mean)	RMSE
Agriculture	linear	1	-0.0970	1.5154
	linear	30	-0.1599	1.7564
	linear	360	-0.1504	1.7469
	lasso	1	0.0155	1.4036
	lasso	30	-0.0021	1.5986
	lasso	360	-0.0002	1.5967
	random forest	1	0.2046	1.2270
	random forest	30	-0.1239	1.7196
	random forest	360	-0.1010	1.6951
	svm	1	0.0006	1.5958
	svm	30	0.0007	1.5957
	svm	360	0.0006	1.5958
Pharmaceutical Products	linear	1	-0.0092	1.3521
	linear	30	-0.1177	1.5482
	linear	360	-0.1383	1.5687
	lasso	1	0.0146	1.2467
	lasso	30	-0.0062	1.4367
	lasso	360	-0.0007	1.4311
	random forest	1	0.1887	1.0739
	random forest	30	-0.0968	1.5277
	random forest	360	-0.0832	1.5120
	svm	1	-0.0016	1.4320
	svm	30	-0.0015	1.4320
	svm	360	-0.0018	1.4323
Computers	linear	1	-0.0895	1.3692
	linear	30	-0.1276	1.4252
	linear	360	-0.1255	1.4232
	lasso	1	0.0207	1.2585
	lasso	30	-0.0059	1.3036
	lasso	360	-0.0006	1.2983
	random forest	1	0.1700	1.1111
	random forest	30	-0.1076	1.4055
	random forest	360	-0.1157	1.4119
	svm	1	-0.0005	1.2982
	svm	30	-0.0003	1.2980
	svm	360	-0.0008	1.2985

5 Conclusion and Application

For all three industries chosen, the optimal model is random forest, which has the lowest RMSE and large RMSE.DIFF. More importantly, in each RMSE.DIFF graph shown for all three chosen industries, random forest has a better performance in the recent period, which can be concluded from the upper trend of the RMSE.DIFF line.

For different lag dataset in all three chosen industries, the lag one dataset gives the optimal prediction outcome. This shows that using the contemporaneous effect to predict industry stocks return is preferable for agriculture, pharmaceutical products, and computer industries.

The investment manager can use random forest and lag 1-day dataset to do a prediction for tomorrow's stocks changes in this three industries. Besides, lasso regression can be used to chose important variables during the model training process.

6 limitation and Future Research

Firstly, due to the large data set and huge computational complexity, it is hard to cross-validate with more window size and some important parameters, including minimum leaf node for the random forest, cost, and gamma for support vector regression. In the future application of those model, it is to test more set of this parameters and chose the optimal value.

Secondly, to test the lag effect, this report only used lag one day, lag 30 days and lag 360 days dataset. However, for different industries, the optimal lag period may be different, so it is recommended to do cross-validation to find the optimal lag period for different sectors when predicting different target industry stocks return changes.

Lastly, a more integrated model building can be tested in the future. Lasso regression is an ideal method can be used to chose essential variables, and then applied those chosen variables to support vector regression to make further prediction may increase the accuracy of this model.