Data Preprocessing

First, download the daily data of the stock (NVDA) from Jan 2,2015 to Dec 31,2019. Next, calculate the daily net return rate and convert it into a dummy variable. Finally, split the data set into two parts: 2015 to 2018, and 2019 alone. The first part is used for training and second part is for out-of-sample testing. The data description is shown in table1.

Table 1. Data Specification

Category	Variable Name	Variable Type
Dependent Variable	return_d	Binary variable
Independent Variable	return_lag1	Quantitative Variable
	returan_lag2	Quantitative Variable
	return_lag3	Quantitative Variable
	returan_lag4	Quantitative Variable
	return_lag5	Quantitative Variable
	volumn_lag1	Quantitative Variable
	volumn_lag2	Quantitative Variable
	volumn_lag3	Quantitative Variable
	volumn_lag4	Quantitative Variable
	volumn_lag5	Quantitative Variable

Model

a) Logistic Regression

Establish empty model and full model according to independent and dependent variables, and do analysis of variance for the two models. The statistic of the generalized likelihood ratio test is 7.8243, the overall significance level of the model is 0.6460, therefore the overall model is not significant. It shows that none of the ten explanator variables are significantly related stock net return changes.

Table 2. The Logistic Regression Model Output

Coefficients					
(Intercept)	return_lag1	return_lag2	return_lag3	return_lag4	return_lag5
0.1804	-0.0830	0.0795	0.0458	0.0336	-0.0320
Volume_lag1	Volume_lag2	Volume_lag3	Volume_lag4	Volume_lag5	
0.0628	0.0336	-0.0230	0.1142	-0.0922	

Degrees of Freedom: 999 Total (i.e. Null); 989 Residual
Null Deviance: 1378 Residual Deviance: 1370 AIC: 139

Use the train set to make predictions, the ROC curve is as follows, when the threshold is 0.538, the model obtains the maximum AUC value of 0.540.

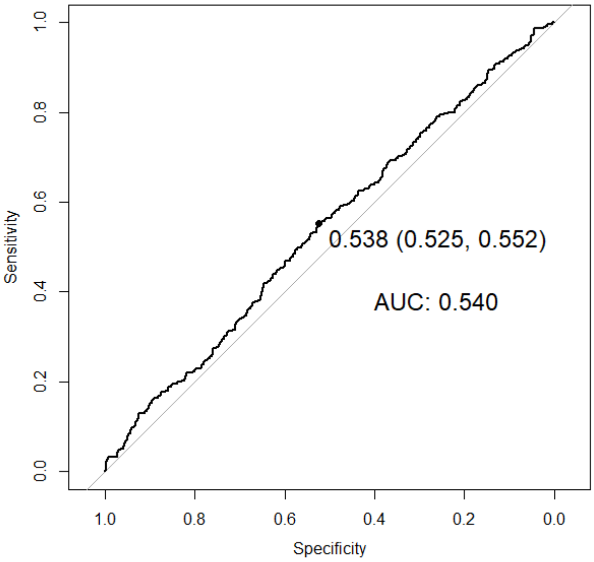

Figure1. The ROC Curve of Logistic Regression Model Train Set

From the confusion matrix, the overall error rate is 46.30%, true positive rate is 54.50%, false positive rate is 47.25% positive prediction value is 58.00%, negative prediction value 49.18%.

Table 3. The Logistic Regression Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	240	248
	1	215	297

Use the test set to make predictions, the ROC curve is as follows, when the threshold is 0.57, the model obtains the maximum AUC value of 0.584.

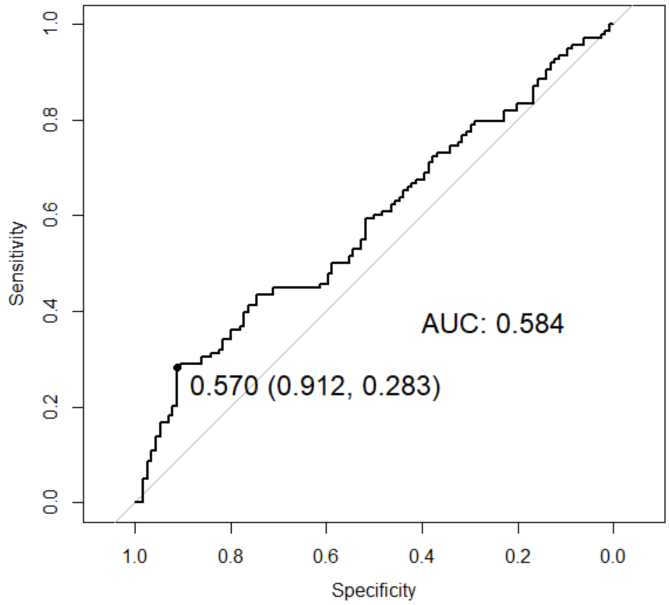

Figure2. The ROC Curve of Logistic Regression Model Test Set

From the confusion matrix, the overall error rate is 43.30%, true positive rate is 28.26%, false positive rate is 8.77% positive prediction value is 20.41%, negative prediction value 51.23%.

Table 4. The Logistic Regression Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	104	99
	1	10	39

b) LDA

Convert stock ups and downs from 0-1 variable to label variable (YES or NO) and establish LDA model. The output of the LDA model is as follows, the prior probabilities of groups are 0.455 and 0.545.

Table 5. The LDA Model Output

Prior probabilities of groups					
0			1		
0.455			0.545		
Group means					
	return_1	return_2	return_3	return_4	return_5
0	0.042	-0.046	-0.024	-0.021	0.010
1	-0.038	0.034	0.015	0.009	-0.015

	Volume_1	Volume_2	Volume_3	Volume_4	Volume_5
0	-0.022	-0.019	-0.014	-0.024	0 .015
1	0.068	0.063	0.057	0.062	0.026
Coefficients of linear discriminants					
	return_1	return_2	return_3	return_4	returnl5
LD1	-0.460	0.431	0.257	0.185	-0.180
	Volume_1	Volume_2	Volume_3	Volume_4	Volume_5
LD1	0.333	0.181	-0.121	0.630	-0.508

Use the train set to make predictions, the ROC curve is as follows, the model obtains the maximum AUC value of 0.54.

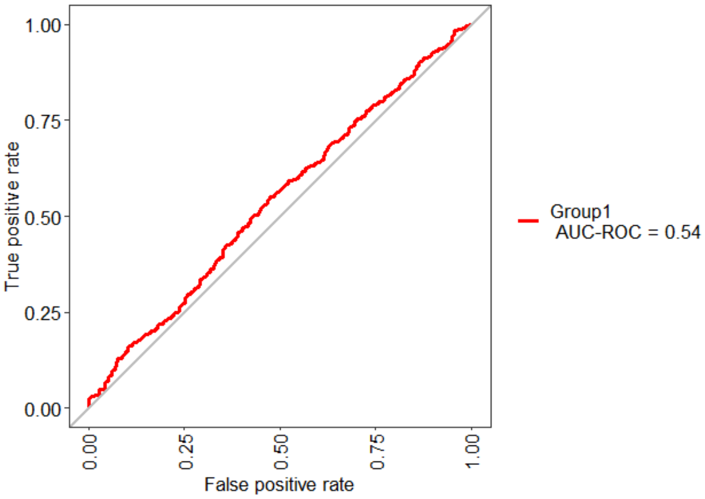

Figure3. The LDA Model (Train Set) ROC Curve

From the confusion matrix, the overall error rate is 45.10%, true positive rate is 93.39%, false positive rate is 91.21%, positive prediction value is 55.09%, negative prediction value 52.63%.

Table 6. The LDA Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	40	36
	1	415	509

Use the test set to make predictions, the ROC curve is as follows, the model obtains the maximum AUC value of 0.59.

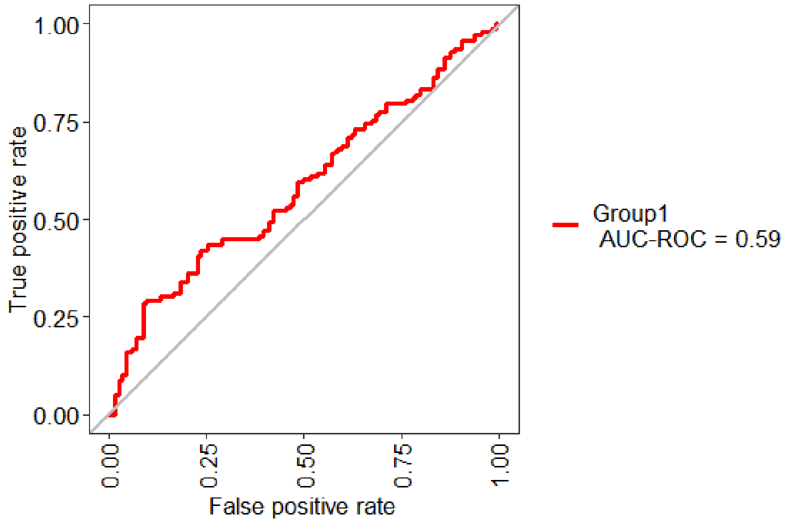

Figure4. The LDA Model (Test Set) ROC Curve

From the confusion matrix, the overall error rate is 43.65%, true positive rate is 92.75%, false positive rate is 87.72%, positive prediction value is 56.14%, negative prediction value 58.33%.

Table 7. The LDA Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	14	10
	1	100	128

c) QDA

The output of the QDA model is as follows, the prior probabilities of groups are 0.455 and 0.545.

Table 8. The QDA Model Output

Prior probabilities of groups					
0			1		
0.455			0.545		
Group means					
	return_1	return_2	return_3	return_4	return_5
0	0.042	-0.046	-0.024	-0.021	0.010
1	-0.038	0.034	0.015	0.009	-0.015

	Volume_1	Volume_2	Volume_3	Volume_4	Volume_5
0	-0.022	-0.019	-0.014	-0.024	0.016
1	0.068	0.063	0.057	0.062	0.027

Use the train set to make predictions, the ROC curve is as follows, the model obtains the maximum AUC value of 0.55.

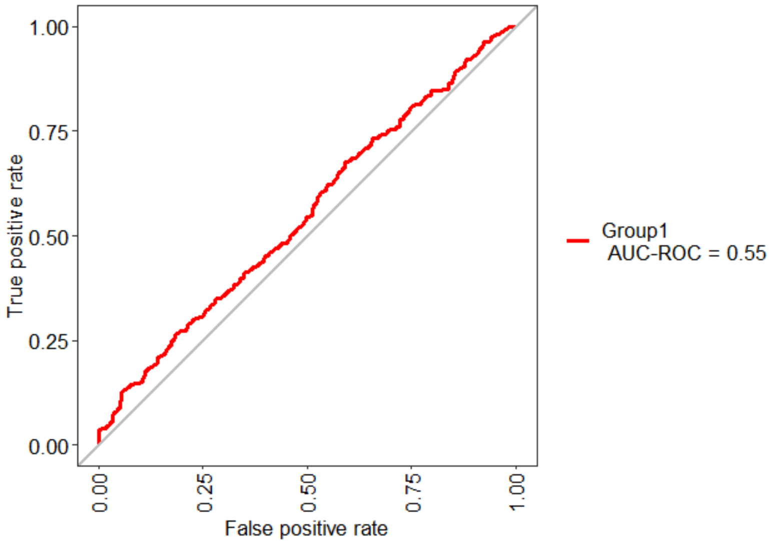

Figure 5. The QDA Model ROC Curve (Train Set)

From the confusion matrix, the overall error rate is 48.80%, true positive rate is 27.16%, false positive rate is 20.00%, positive prediction value is 61.92%, negative prediction value 47.83%.

Table 9. The QDA Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	364	397
	1	91	148

Use the test set to make predictions, the ROC curve is as follows, the model obtains the maximum AUC value of 0.51.

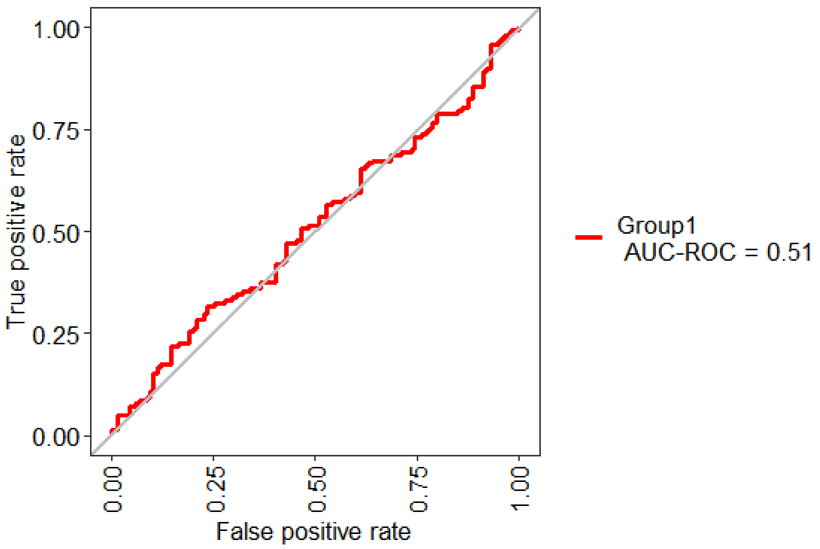

Figure 6. The QDA Model ROC Curve (Test Set)

From the confusion matrix, the overall error rate is 48.81%, true positive rate is 32.61%, false positive rate is 26.32%, positive prediction value is 60.00%, negative prediction value 47.46%.

Table 7. The QDA Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	84	93
	1	30	45

d) Decision Tree

Establish a decision tree model, but the model output shows that the effective variable cannot be found, so the decision tree model cannot be used for prediction.

Table 10. The Decision Tree Model Output

Variables actually used in tree construction	character(0)
Number of terminal nodes	1
Residual mean deviance	1.38 = 1378 / 999
Misclassification error rate	0.455 = 455 / 1000

e) Decision Tree – Bagging

Use Bootstrap aggregating to train a decision tree model, since the model has 10 predictors, the mtry parameter is set to 3. The OOB estimate of error rate is 49%.

View the importance of each variable, the results are as follows. Use the average of decrease accuracy as the most metric, the volume variable lagging 2 and 4 periods and the net return lagging 5 periods are more important.

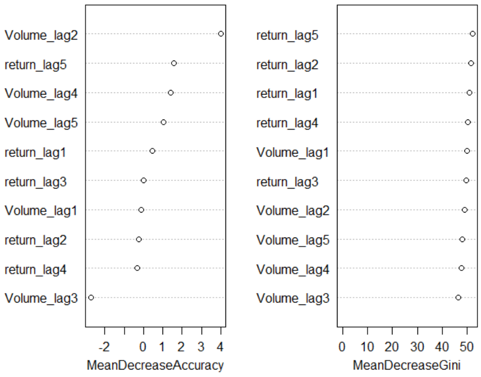

Figure 7. The Variables Importance

From the train set confusion matrix, the overall error rate is 49.20%, true positive rate is 64.95%, false positive rate is 65.71%, positive prediction value is 54.21%, negative prediction value 44.96%.

Table 11. The Decision Tree (Bagging) Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	156	191
	1	299	354

From the test set confusion matrix, the overall error rate is 46.43%, true positive rate is 68.84%, false positive rate is 64.91%, positive prediction value is 54.91%, negative prediction value 48.19%.

Table 12. The Decision Tree (Bagging) Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	40	43
	1	74	95

f) Decision Tree – Random Forest

Use random forest to train a decision tree model, the mtry parameter is set to 1. The OOB estimate of error rate is 48.4%.

View the importance of each variable, the results are as follows. Use the average of decrease accuracy as the most metric, the volume variable lagging 2 and 1 periods and the net return lagging 5 periods are more important.

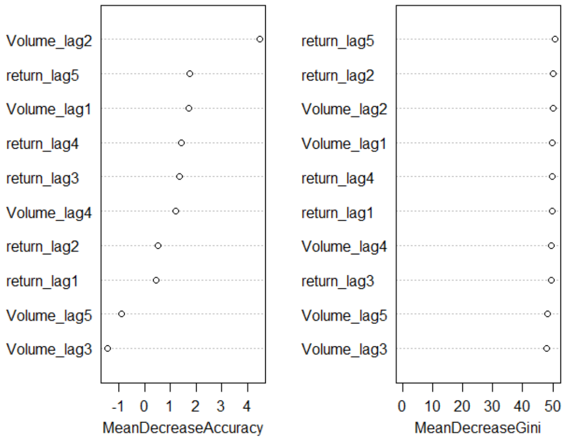

Figure 8. The Variables Importance

From the train set confusion matrix, the overall error rate is 48.40%, true positive rate is 67.52%, false positive rate is 67.47%, positive prediction value is 54.52%, negative prediction value 45.54%.

Table 13. The Decision Tree (Bagging) Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	148	177
	1	307	368

From the test set confusion matrix, the overall error rate is 46.83%, true positive rate is 71.74%, false positive rate is 69.30%, positive prediction value is 55.62%, negative prediction value 47.30%.

Table 14. The Decision Tree (Bagging) Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	35	39
	1	79	99

g) Decision Tree-Boosting

Use the Boosting method to train the decision tree model, the output is as follows. The best cross-validation iteration was 8. There were 10 predictors of which 5 had non-zero influence.

Table 15. The Decision Tree (Boosting) Model Confusion Matrix

var	rel.inf
return_lag3	12.229
return_lag2	11.845
return_lag5	11.766
Volume_lag3	11.712
return_lag4	11.653
return_lag1	10.728
Volume_lag1	10.537
Volume_lag4	7.413
Volume_lag5	6.109
Volume_lag2	6.003

The relative influence of predictors shows that the net return lagging 3/2/5 periods are more important.

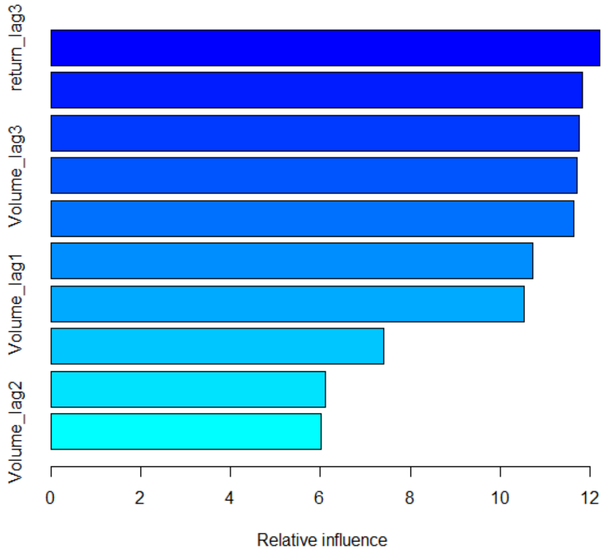

Figure 9. The Relative Influence of Predictors

From the train set confusion matrix, the overall error rate is 44.60%, true positive rate is 98.53%, false positive rate is 96.26%, positive prediction value is 55.08%, negative prediction value 68.00%.

Table 16. The Decision Tree (Boosting) Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	17	8
	1	438	537

From the test set confusion matrix, the overall error rate is 45.63%, true positive rate is 98.55%, false positive rate is 99.12%, positive prediction value is 54.62%, negative prediction value 50.00%.

Table 17. The Decision Tree (Boosting) Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	1	2
	1	113	136

h) SVM (linear)

Establish an SVM model, which tries to use 0.001, 0.01, 0.1, 1,5,10,100 as the cost parameter, and pick the best one among them, and finally use 0.001 as the cost parameter.

It can be clearly seen from the confusion matrix that the model predicts all responses as 1, so the model is invalid.

Table 18. The SVM (linear) Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	0	0
	1	455	545

Table 19. The SVM (linear) Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	0	0
	1	114	138

i) SVM (polynomial)

Establish an SVM model with polynomial kernel, from the train set confusion matrix, the overall error rate is 32.00%, true positive rate is 90.09%, false positive rate is 58.46%, positive prediction value is 64.86%, negative prediction value 77.78%.

Table 20. The SVM (polynomial) Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	189	54
	1	266	491

From the test set confusion matrix, the overall error rate is 47.22%, true positive rate is 72.46%, false positive rate is 71.05%, positive prediction value is 55.25%, negative prediction value 46.48%.

Table 21. The SVM (polynomial) Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	33	38
	1	81	100

j) SVM (radial)

Establish an SVM model with radial kernel, from the train set confusion matrix, the overall error rate is 16.30%, true positive rate is 87.89%, false positive rate is 21.32%, positive prediction value is 83.16%, negative prediction value 84.04%.

Table 22. The SVM (radial) Model Confusion Matrix (Train Set)

		Actual Class	
		0	1
Predict Class	0	358	66
	1	97	479

From the test set confusion matrix, the overall error rate is 48.80%, true positive rate is 60.14%, false positive rate is 59.65%, positive prediction value is 54.97%, negative prediction value 45.54%.

Table 23. The SVM (radial) Model Confusion Matrix (Test Set)

		Actual Class	
		0	1
Predict Class	0	46	55
	1	68	83

Conclusion

Based on the analysis of accuracy, precision, sensitivity and specificity of various models, the overall prediction of the model is not optimistic. Logistic model and decision tree model cannot even be established. Although some models have high prediction accuracy in a certain category, for example, the LDA model has good performance in sensitivity and specificity, but the overall situation of the model is not good.

Therefore, I think it is undesirable to predict the next day's stock changes through the lag period's rate of return and trading volume.