MATH4995 Final Report

Discovery after reducing number of features on Titanic Problem

Name: Lee Cheuk Yin

Review on Titanic Project

Previously, I have successfully use SVM on Titanic passengers' data to predict the survival of passengers. And obtain the following result:

Linear SVM Accuracy	0.7715355805243446
Linear SVM Precision	0.7474747474747475
Linear SVM Recall	0.6727272727272727
RBF SVM Accuracy	0.7940074906367042
RBF SVM Precision	0.7835051546391752
RBF SVM Recall	0.6909090909090909
Kaggle Score of Linear SVM	0.77272
Kaggle Score of RBF SVM	0.77511

From previous conclusion, I had discovered that if the centralization of data will favor the performance of SVM. Also,I had doubted that the fewer the number of features, the better performance of SVM. In this report, I will verify this hypothesis by reducing the number of features that will be fed to SVM model.

Data Preprocessing

Data Summary

After repeating the steps that are stated as the previous report, obtain the following data summary again.

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
PassengerId	1	-0.00502832	-0.0353298	-0.043136	0.0313186	-0.0576859	-0.00165658	0.0127032	0.0305551
Survived	-0.00502832	1	-0.335549	0.541585	-0.0698217	-0.03404	0.0831508	0.25529	-0.108669
Pclass	-0.0353298	-0.335549	1	-0.127741	-0.336512	0.0816556	0.0168245	-0.548193	-0.0438347
Sex	-0.043136	0.541585	-0.127741	1	-0.0865058	0.116348	0.247508	0.179958	-0.118593
Age	0.0313186	-0.0698217	-0.336512	-0.0865058	1	-0.232543	-0.171485	0.0937071	0.00716521
SibSp	-0.0576859	-0.03404	0.0816556	0.116348	-0.232543	1	0.414542	0.160887	0.0606061
Parch	-0.00165658	0.0831508	0.0168245	0.247508	-0.171485	0.414542	1	0.217532	0.0793198
Fare	0.0127032	0.25529	-0.548193	0.179958	0.0937071	0.160887	0.217532	1	-0.0634623
Embarked	0.0305551	-0.108669	-0.0438347	-0.118593	0.00716521	0.0606061	0.0793198	-0.0634623	1

Correlation matrix

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
count	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000
mean	446.000000	0.382452	2.311586	0.350956	29.315152	0.524184	0.382452	32.096681	1.637795
std	256.998173	0.486260	0.834700	0.477538	12.984932	1.103705	0.806761	49.697504	0.636157
min	1.000000	0.000000	1.000000	0.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	224.000000	0.000000	2.000000	0.000000	22.000000	0.000000	0.000000	7.895800	1.000000
50%	446.000000	0.000000	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200	2.000000
75%	668.000000	1.000000	3.000000	1.000000	35.000000	1.000000	0.000000	31.000000	2.000000
max	891.000000	1.000000	3.000000	1.000000	80.000000	8.000000	6.000000	512.329200	2.000000

Data description

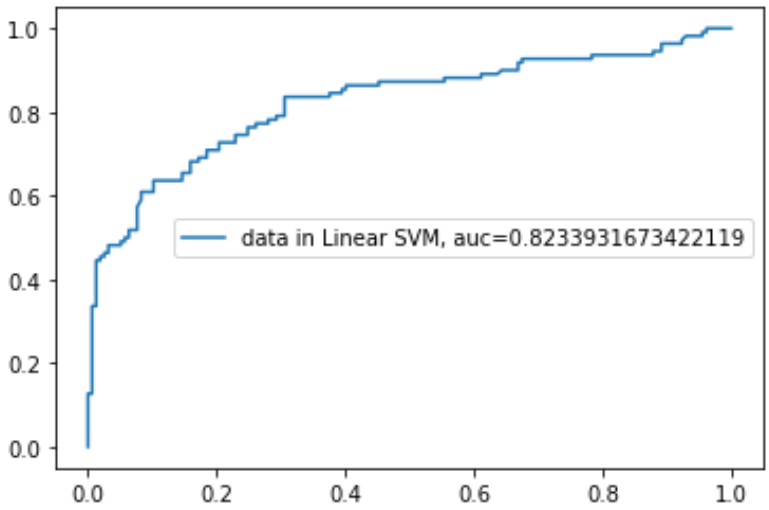
Key Part: Data Cleansing

From the conclusion inspired by past experience, the more centralized the data is, the higher the accuracy that the SVM can achieve. Besides feature “Embarked” needs to be removed because of unexpected weighting effect, from the data description table, it is shown that the feature “SibSp” and “Parch” are clearly skewed towards right-hand side. The first quartile and the median are both 0 for both “SibSp” and “Parch” and having the extreme value at the maximum. In order to fulfill the above criteria of centralized data (i.e. no extreme data features existed). The “SibSp” and “Parch” features needed to be removed. After data cleansing is done, the data table will be further simplified as follow:

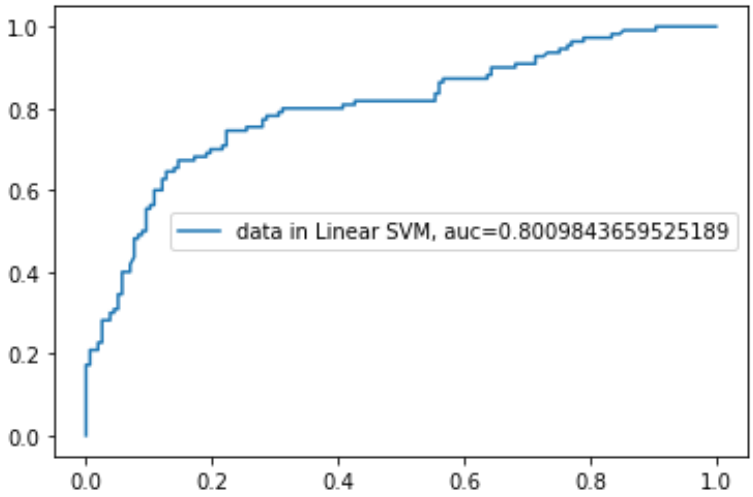
	Survived	Pclass	Sex	Age	Fare
0	0	3	0	22.0	7.2500
1	1	1	1	38.0	71.2833
2	1	3	1	26.0	7.9250
3	1	1	1	35.0	53.1000
4	0	3	0	35.0	8.0500
5	0	3	0	28.0	8.4583
6	0	1	0	54.0	51.8625
7	0	3	0	2.0	21.0750
8	1	3	1	27.0	11.1333
9	1	2	1	14.0	30.0708
10	1	3	1	4.0	16.7000
11	1	1	1	58.0	26.5500
12	0	3	0	20.0	8.0500
13	0	3	0	39.0	31.2750
14	0	3	1	14.0	7.8542

Model Performance compare with the previous data preprocessing procedure

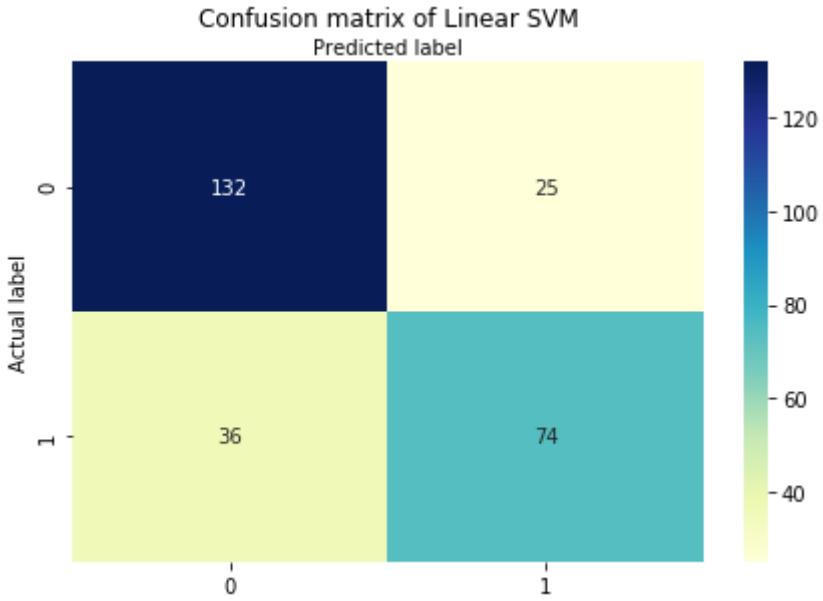
Linear SVM:



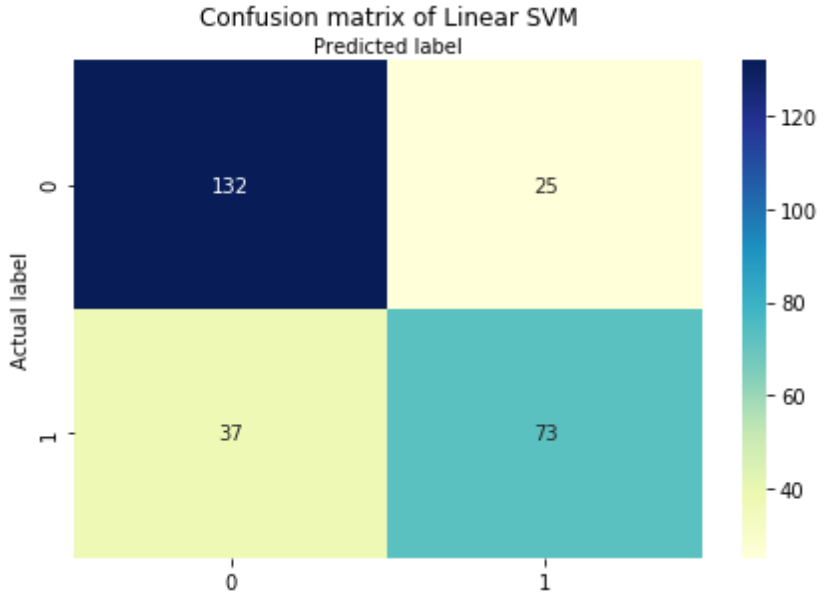
ROC curve after removing “Parch” and “SibSp”



ROC curve without removing “Parch” and “SibSp”

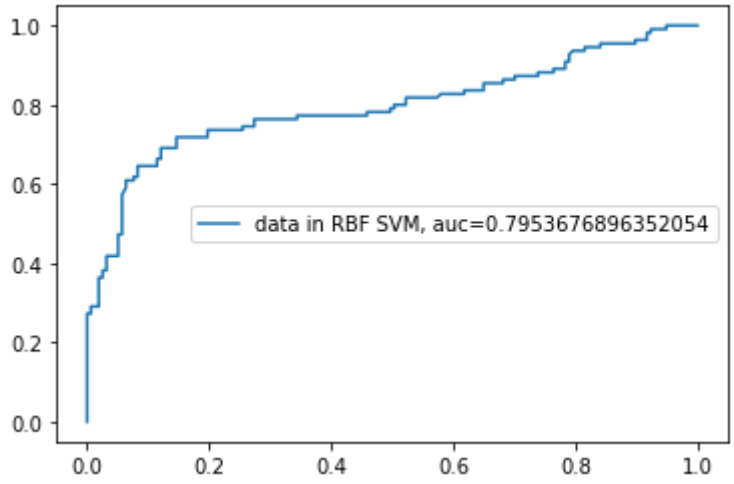


Confusion matrix before removing “SibSp” and “Parch”

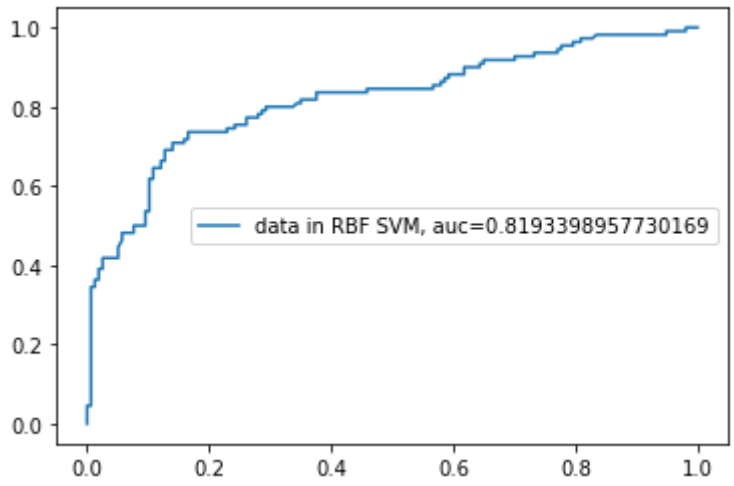


Confusion matrix after removing “SibSp” and “Parch”

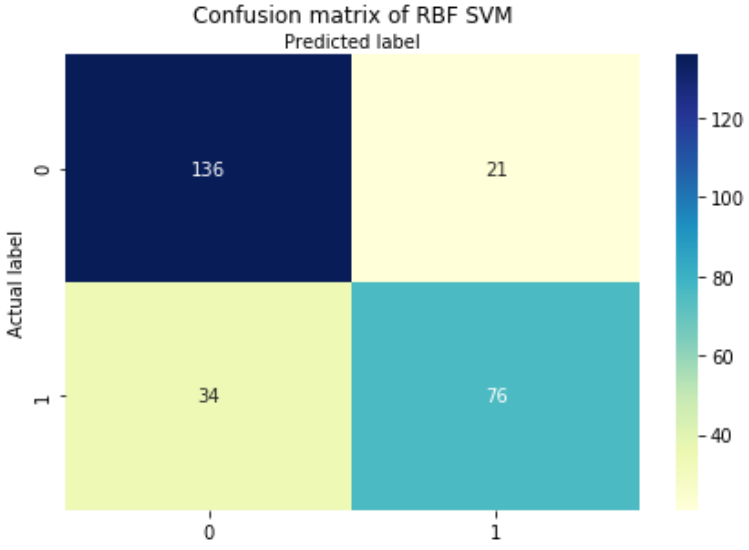
RBF SVM:



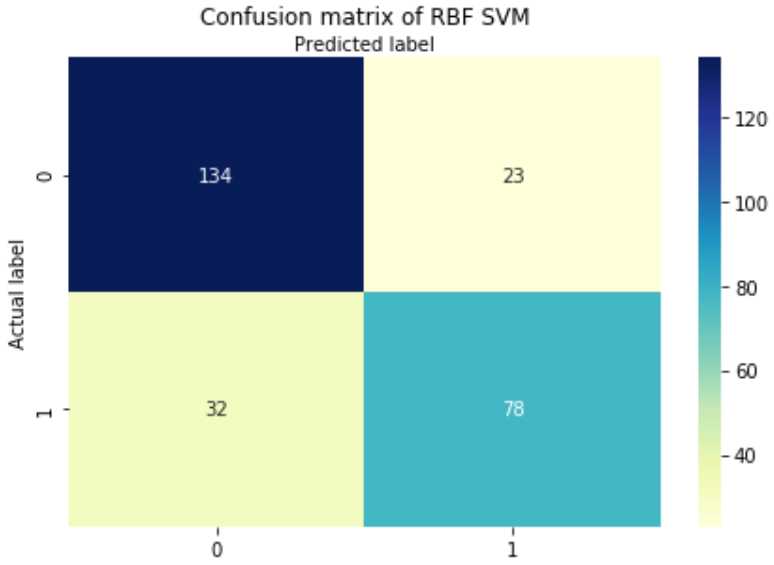
ROC curve after removing “Parch” and “SibSp”



ROC curve without removing “Parch” and “SibSp”



Confusion matrix before removing “SibSp” and “Parch”



Confusion matrix after removing “SibSp” and “Parch”

Scores comparison

	Before removing “Parch” and “SibSp”	After removing “Parch” and “SibSp”
Linear SVM Accuracy	0.7715355805243446	0.7677902621722846
Linear SVM Precision	0.7474747474747475	0.7448979591836735
Linear SVM Recall	0.6727272727272727	0.6636363636363637
RBF SVM Accuracy	0.7940074906367042	0.7940074906367042
RBF SVM Precision	0.7835051546391752	0.7722772277227723
RBF SVM Recall	0.6909090909090909	0.7090909090909091
Kaggle Score of Linear SVM	0.77272	0.76555
Kaggle Score of RBF SVM	0.77511	0.77751

Conclusion

From the above result, there is no significance difference on the performance of model after deleting features “SibSp” and “Parch”. In other words, features “Parch” and “SibSp” have a negligible effect on the model performance. As a result, the hypothesis is falsified. However, there is a new discovered fact: The SVM model based on the previous data preprocessing procedures can be further simplified to achieve a similar result. That means the old SVM model is indeed “overfitted” because the parameters “SibSp” and “Parch” may not have a huge correlation with the survival result. Moreover, simplifying model will favor the computation process. Although it maybe not very significant in this titanic problem because the test dataset and training dataset are in fact relatively small, it will have a huge effect on training time and computation time when the training and testing test are very large. In fact, the simplification of regression or classification model is essential in regression analysis or statistical analysis. Even the detailed overfit testing is not carried out in a very detailed way, this report has shown the general idea of the

process of simplification of SVM model in Titanic problem. For me, this is also an important reminder for me to do machine learning in future career if I have chance indeed.