MATH4995 Mini-Project 1: Model selection and regularization on prediction of survival on the Titanic

Tse Justin Chung Heng (20611518)

Introduction

In Chapter 6 of ISLR, we have dealt with the Hitter dataset on predicting the salary of baseball player. That is a regression problem. On the other hand, the goal of this project is to explore how the model selection and regularization method works in classification problem. More specifically, I will focus on using logistics regression model combine with various model selection method, information criteria and regularization method.

Data Preprocessing:

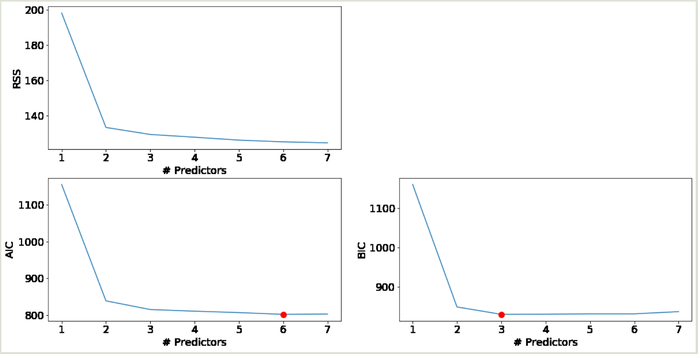
Several steps has been taken to clean the dataset. 1. Dropping features (Cabin and Ticket) 2. Extract title from name 3. Converting categorical features to numerical 4. Fill in null values 5. creating new feature from existing features .

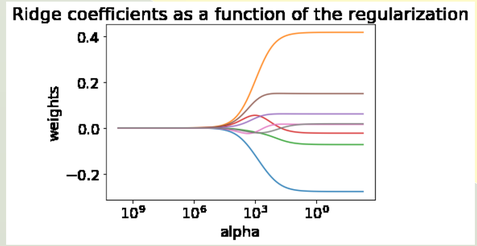
	Survived	Pclass	Sex	Age	Fare	Embarked	Title	IsAlone	Age*Class
0	0	3	0	1	0	0	1	0	3
1	1	1	1	2	3	1	3	0	2
2	1	3	1	1	1	0	2	1	3
3	1	1	1	2	3	0	3	0	2
4	0	3	0	2	1	0	1	1	6
5	0	3	0	1	1	2	1	1	3
6	0	1	0	3	3	0	1	1	3
7	0	3	0	0	2	0	4	0	0
8	1	3	1	1	1	0	3	0	3
9	1	2	1	0	2	1	3	0	0

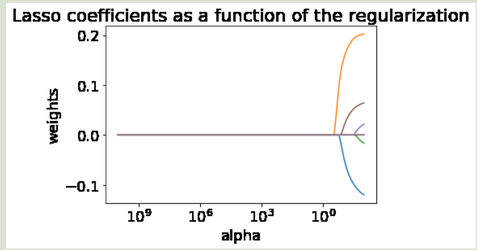
Model selection

In this project we will compare the result of 1. best subset selection with AIC 2. best subset selection with BIC 3. Ridge regression with CV 4. Lasso regression with CV.

	1. AIC	2. BIC	3. Ridge	4. Lasso
Number of parameters	6	3	8	5
Parameter	Pclass -0.7529 Sex 2.2610 Age 0.2535 Embarked 0.2845 Title 0.3444 Age*Class -0.2785	Pclass -0.9219 Sex 2.2909 Title 0.3840	Pclass -0.255532 Sex 0.397340 Age -0.056094 Fare -0.006139 Embarked 0.061012 Title 0.149729 IsAlone 0.016209 Age*Class 0.002501	Pclass -0.119751 Sex 0.202281 Age -0.016549 Embarked 0.021558 Title 0.064091
Accuracy	0.76794	0.76794	0.74162	0.77033







Analysis

The result is consistent with what we learnt from the lab of Chapter 6 of ISLR. For instance:

1. The BIC penalizes more heavily for additional parameters.
2. LASSO often shrinks coefficients to be identically 0.
3. All method yield similar result in term of accuracy and general importance of certain parameters.

Finding specific to this dataset:

1. Pclass, Sex and Title appear in all 4 models.
2. Sex is the highest positive coefficient in all 4 models, which may indicate that female(Sex =1) has higher likelihood to survive in the disaster.
3. Pclass is the highest negative coefficient in all 4 models, which may indicate that the lower the class you were, the less likely you would survive.

Conclusion

In this project, we have found that the 4 model selection and regularization methods perform similarly in classification task in term of the accuracy. However, how all these methods may perform differently is not in the scope of this project. Further investigation may be conducted in future project.

Reference

MANAV SEHGAL *Titanic Data Science Solutions*. Kaggle. Accessed on 1/10/2021

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An introduction to statistical learning: With applications in R*.