

Peer Review For Group 6

Mini-Project 1. MATH 4995

Group Number: 6

Title: Model Selection and Regularization on Prediction of Survival on the Titanic

Summary of the report

The report primarily tries to reproduce the result of Hitter's problem from ISLR chapter 6, and also strives to predict the survival of people in Titanic. Data Pre-processing is done to both reduce unnecessary parameter and to make it more convenient to fit the model into different machine learning functions. Multiple models is selected and regularization is used. The analysis is mostly related to how regularization can be implemented. From the result, it has been shown that the titanic question can indeed reproduce similar result to Hitter's problem.

Strengths of the report

The report has a very clear idea on what the author would like to do. The analysis did shown to be echoing with the problems stated in the introduction. Instead of focusing on prediction, the author take the approach on discussing the regularization method, which makes the report differ from many of the groups.

Weaknesses of the report

As some of the data are dropped, it might be better if those data can be explored a bit which helps to explain why the data is dropped. From the graph of the Lasso coefficients, it might be better if the graph of the alpha could extend a bit further so that we can understand the trend of the weights further. Moreover, from the report, the accuracy of both AIC and BIC is the same. It might be necessary to explore it further, as same accuracy may indicate potential same prediction result.

Evaluation on quality of writing: 3

Is the report clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.

The report is clear enough that most part of the project are written. However, I believe that it might be needed to explain a bit more on the result of ISLR's Hitter's problem. It might also be needed if ISLR is quoted in the reference. There is a graph related to feature selection, which might need to be clearer as it is unknown on what the y-axis is representing. Also, while those diagram are made for each of the model, it might be necessary to indicate why Lasso is missing. From the model selection, the meaning of values next to each parameter is also unclear.

Evaluation on presentation: 4

Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.

The flow of the presentation is clear as it help us to follow the flow of the investigations, and have a clear enough explanation on why several decisions are made in the project. Each steps explained are clear. Inside the presentation, it has also shown the score of different prediction models which can

help us to have a preliminary background knowledge on why the author would like to take the approach on reproducing the result.

Evaluation on creativity: 2

Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations or extensions of existing ideas are also valuable.

In terms of the issues proposed by the group, it differs from the traditional approach of solving a Kaggle contest, as it discusses the regularization model more than the prediction itself. While it is a reproduction of a result using a different dataset, it should still be considered as creative as the report could potentially report results that might be partially different from the textbook.

Confidence on your assessment: 2

(3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)

While I have read the entire report, presentation and the source code, there is one issue that makes my assessment less confident, which is on the source code, as I am not sure which part of the source code is from the authors' work and which is actually data from the public. Moreover, some of the contents appear in the source code while not in the report or presentation, which might make it "lose the spotlight" during reviewing.