Peer Review of Kaggle Titanic Project

Jonathan Mi

1. Jeffrey:

One key strength that Jeffrey demonstrates in his Titanic research report is that he uses clear and concise language to illustrate the problems he had to face, and precisely explains his method of solving them. For example, in the section “Cleaning the Data”, Jeffrey divides each variable he had to clean into separate paragraphs. Next, in an organized fashion, he portrays his logical assumptions and conclusions to clean each of these variables. This can be seen when Jeffrey laid his step-by-step plan to clean missing values for the “Age” category, where he first says “The most basic method is to calculate the median of the age values I already have, and then just fill in every missing value using the median. This improved the accuracy, but the model could definitely still be improved”. From this sentence, it is clear that Jeffrey has a logical process in mind, and that he has a plan to effectively clean missing data values. Afterwards, in the next paragraph, he goes more in-depth to describe a more advanced model, which is creating a linear model using data that was filtered to exclude all rows with age as an outlying variable. Thus, as a reader, it is very easy to follow Jeffrey’s ideas and understand his reasoning.

However, one weakness that Jeffrey should definitely fix is is lack of attention to detail. In the third paragraph under the Problem Introduction section, Jeffrey says “In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy”. Cleary, Jeffrey did not filter his language, because when in his lab report, he should not be using the pronoun “we” because it is

his lab report, not someone else's as well. Also, in some parts in the report, Jeffrey did not use precise language. For example, in the section where he talked about creating a linear model for age, he said that Pclass and Sibsp had the lowest Pr value. What he should have wrote was that they have the lowest Pr(>|t|) value, because in data analysis, there is also a Pr(>|z|) value.

Overall, I would rate Jeffrey 3.5 out of 4 because he had a successful model and his report was adequate, but lacked the detail and precision that would have been necessary for a 4.

2. Brad Zhang:

Brad did a very good job of both explaining the lab and showing how he cleaned the data. For example, in section 3, titled "Data Cleaning", Brad divides each variable he had to clean into separate sections. Next, in a concise manner, he portrays his logical assumptions and conclusions to clean each of these variables. For instance, in the Fare section, he states that since there was only one missing value, he decided to just set it to the median Fare value, which was the easiest way to complete his model. In addition, Brad gives plenty of graphs and charts to support his analysis, which helps the reader visualize his thought process.

However, one weakness that Brad should address is his Standardized section. As a reader who has just started learning R, I was confused about what the term actually meant. In my opinion, a picture of his code, or a simple definition that explained what he did could have helped greatly for people to understand what Brad meant.

Overall, I would give Brad 4/4 stars, because his paper was precise and sufficiently explained his model so that readers can understand and follow his thought process. There are still some areas that need to be fixed, but he did a great job analyzing the data comprehensively throughout his report.

3. Kaitlin:

Kaitlin's most prominent strength is that she organized her sections very neatly in an easily navigable poster, which also makes her report seems more visually appealing. For example, in the Relationship between "Survived" and other Variables section, as a reader, I found the charts very effective to demonstrate the correlation between the variables. In addition, Kaitlin demonstrated that she had a clear logical thought process by how she first presented her "First Prediction", and then created a "Modification" section, and then finally displayed the "Further Predictions" section and the "Conclusion".

On the other hand, one weakness that Kaitlin should address is that in some sections, which occur very infrequently, she sometimes does not fully explain how she created her model. For example, when cleaning the "Age" column, Katilin says that she used PMM, but did not explain what that function is, or show a picture of her code.

In conclusion, Kaitlin did a great job with her report. Her poster was clear, effective, and visually appealing. There are some minor errors that could be fixed easily, but otherwise she did a great job. That is why I give her 4/4 stars on her Titanic report.

4. Jim:

One strength in Jim's report is that he uses scientific language and has a clear sense of what he is doing in his report. For example, he uses terms such as inductive bias, which is very advanced, and makes his report more professional. Also, Jim mentions terms such as SVM, which he used for his predictive Titanic model, which demonstrates that he thinks at a high-level in a logical manner.

Nonetheless, there are two major flaws in Jim's report. First, he does not adequately explain how he created his predictive model. For instance, when he cleaned the "Age" column, Jim mentions that he uses SVM, but does not go more in-depth to explain this model, which can greatly confuse readers. Second, Jim's report can be

organized a lot better, with clear headings and references to his charts, because for most of them, he just puts in pictures of bar graphs, but does not explain them.

In conclusion, I would give Jim 3/4 stars based on the fact that his report is adequate, but not sufficient enough to deserve a 4 because he could be a lot more organized and detailed in his writing. However, Jim does have a model that clearly works, which can be seen from his 83% accuracy rate, which is why I give his Titanic Project a 3.