Machine Learning Assignment 97

Elijah Tarr

March 5, 2021

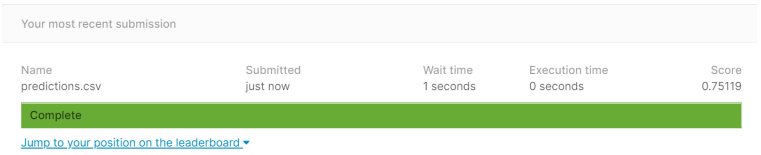
1 Feature Selection

- Survived: We won't include survival in the model, since that is what we are trying to predict.
- Ticket Class: We should include this because the staff might have prioritized higher paying customers for the life boats. (Yay capitalism!)
- Sex: We should include this because women were prioritized for getting onto the life boats.
- Age: We should include this because children were prioritized for getting onto the life boats.
- Siblings/Spouses: We should include this because families would likely have been prioritized when getting on life boats.
- Parents/Children: We should include this for the reasons explained above.
- Ticket: We shouldn't include the ticket number, because that's just an arbitrary value assigned to each passenger. Plus, it's difficult to parse.
- Fare: We shouldn't include the passenger fare, because it doesn't perfectly correspond to ticket class.
- Cabin: We shouldn't include the cabin because 1. It's difficult to parse, and 2. There are too many missing values for it to improve the accuracy. However, there could be different amounts of life boats pertaining to each cabin.
- Embarked: We shouldn't include this because everyone experienced the same crash, no matter where they came from.
- Name: We shouldn't include this because the name likely isn't a factor in if they boarded a life boat.
- Passenger ID: The passenger ID is just a number assigned to each passenger arbitrarily. So, we shouldn't include it.

2 Model Selection

Model	Training Accuracy	Testing Accuracy
Linear Regressor	80.11%	83.19%
Logistic Regressor	80.11%	83.19%
Gini (Depth 5)	59.38%	59.38%
Gini (Depth 10)	59.38%	59.38%
Random Forest (Depth 5)	59.38%	59.38%
Random Forest (Depth 10)	59.38%	59.38%
Naive Bayes	40.61%	40.61%
KNN ($k = 5$)	59.38%	59.38%
KNN ($k = 10$)	59.38%	59.38%

3 Submission



For these predictions, I used a standard linear regressor since it's what gave me the highest accuracy. Unfortunately, my predictions fell just below the objectives, but I have an idea as to why. While training my model, I remove all rows that have null data, for example if the age is left blank, I don't include that in my regression. Then, when I predict the data, I instead set the empty fields to 0, which definitely would skew my results if the corresponding coefficient was large enough. One way I could think to fix this is to train another model on every field except for the ones with blank values, and I could use that model to fill in the blank values, then I could use the other model to predict the survival. Otherwise, I could just not include the columns with the blank data. But, I want to include those columns since they still contain a lot of good data.