Data Mining Kaggle Competition

Team name: Neptune

Team members: Bridget Silver and Davis Urlich

The highest private score: .77990

The highest public score: .75598

Please describe how you improved the accuracy of your model step by step and what the accuracy was after each optimization:

1. Filling NANs
    a. replaced missing Fares with the mean fare
    b. replaced missing ages with '999.0' to keep the missing ages distinguishable from other ages
    c. replaced missing Embarked with 'NA'
2. One-hot encoding features / preprocessing features
    a. changed the Fare feature to a float
    b. we one-hot encoded sex (Female and Male)
    c. also one-hot encoded embarked
3. Adding other features
    a. length of name
    b. number of relatives
    c. fare per person
    d. if the person was alone on the boat or not
    e. name titles
4. Model Tweaking:
    a. We made a train-validation split in order to have an accurate depiction of generalizability
    b. We kept a list of features and tried different combinations / orders of these features to see what worked best
    c. We made a decision tree model and a neural network model
    d. We first tried a decision tree with only sex as a feature, which interestingly got the highest private score of .77990, and a public score of .75119
    e. We then tried all of our features / data with a decision tree, and we got a lower accuracy of .727, (but the private score was high .775)
    f. Then, we tried smaller combinations of features and hyperparameters such as tree depth, but still could not improve our score

g. We then created a NN model with all of our features, and this received a higher public accuracy of .75598
h. We played around with the order we inputted our features into the NN model, which changed the accuracy and made it lower (.67942, 74641, .73684)
i. We incorporated a kfold split in order to have a more accurate accuracy score
j. We attempted to design an ensemble learning model but ran out of time to use it on the testing data

What sorts of people were more likely to survive?

- passengers of gender = female were more likely to survive
- of people who had long names (names greater than 36), 71% survived
- We also saw that people who had a name were more likely to survive (not a nan)
- People embarking on 'Q' had a higher survival rate than C and S
- People who had relatives were more likely to survive than people with no relatives
- people who were not alone on the boat were more likely to survive
- People with the title "Mrs" had the highest survival rate, while "Miss" had the second highest survival rate
- People who had a fare greater than the mean fare per person (21.35) were more likely to survive than people who paid a fare below the mean