Data Mining Kaggle Competition

Team name: Ayo

Team members:

- Adam Wu, Amber Shao, Chloe Wang, Randy Li, Rui Zhou

The highest public rank and score:

- Highest public rank: 45
- Score: 0.79640

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

1. We first look through the data and identify several promising predicting features for our model. They are Pclass, sex, parch, and age columns. For column sex, we map males and females to 1 and 0. For variables with nan, we replace these null values with the column's average. Then, we fit this train set into our neural network model and get 0.77245 accuracy.
2. Then, we find out that the feature fare could be helpful. Therefore, we replace its null values with the column's average and add this feature into the train set. Also, arranging people of different ages into 4 subcategories: <18, 18-40, 40-60, 60< could be useful. Therefore, by implementing the method mentioned above and using the same neural network model, we boost the test accuracy to 0.77844.
3. After changing several hyperparameters for the neural network model, we decide to try more models and ensemble them through a simple combination. After trying random forests, neural networks, and decision trees, we choose to use these three models' predictions and take the average of their predictions with a threshold of ⅔. That is, we will mark a passenger as survived if two of three classifiers mark him as survived. By adopting this method, we reach 0.79041 accuracy.
4. Testing the performance of different classifiers, we try the logistic regression classifier and incorporate it into our combination. We still take the mean of these classifiers' predictions but increase our threshold to ¾. This means that we will only mark a passenger as survived if three of four classifiers mark him as survived. By adopting this method, we improve our accuracy to 0.79640.
5. Eventually, we choose to ensemble the four of our best predictions together to generate the last shot prediction. We still take the mean of these classifiers' predictions. However, we try remaining our threshold at ¾ once and change our threshold to ½ once. The accuracy still remains at 0.79640.

Description of what methods and what kind of features most improved accuracy. Did you learn anything about the nature of who survives from your models?

1. There are several key steps towards a higher accuracy of our models:
    a. First of all, incorporating meaningful features into the neural network model helps a lot in improving accuracy. Some features that we found useful are sex, siblings and parents (we combined these two into a feature called “relatives”), and fare.
    b. Next, we applied the k-fold method to our model to avoid our model being too biased for certain data. This is not for improving our current accuracy for the test set that we have, but to make sure that we are not overfitting and the model will also work fine under an unseen test data set.
    c. Then, we tried to add several other classifiers including random forests, decision trees, and logistic regression to see how different models performed. There is no single model that outperformed others a lot, but this step is essential for us to improve the accuracy as we can use the different models to apply the ensemble method.
    d. Last, we applied the ensemble method using all of the models used above: random forests, neural networks, logistic regression, and decision trees. For a single passenger, we predict the result using the four classifiers, and then decide the final result using the dominant result of the four classifiers. Say like three of the four models predict the passenger as survived, then the passenger is classified as survived in our final result. Under this method, our accuracy shows some major improvement.
2. From the data, we see that people that are easier to survive have the following features.
    a. First, if a passenger paid more for the ticket, that is, paid a higher fare and has a higher passenger class, he/she has a higher rate of survival. This makes sense as these people usually have more fortune than others and so are more likely to have higher authority and approaches to get out of danger. They are likely prioritized in the whole process as well.
    b. Next, females have a higher rate of survival. This also makes sense because females and kids are those who are prioritized in a disaster like this as women have the ability to give birth and kids have a longer life to live after. So from a social perspective, they are assets and are prioritized in saving
    c. In addition, passengers with fewer relatives (siblings/parents/kids) are a bit more likely to survive under the data. Although not obvious, the reason could be passengers with fewer relatives can focus on themselves instead of thinking about helping others. This could make them have a higher survival rate.
    d. Last but not least, different age groups show different survival rates. We categorized our passengers into four categories: <18, 18-40, 40-60, and >60 years old. From the data, we can see that passengers younger than 40 years old have a higher survival rate. This makes sense because escaping from this kind of disaster requires stamina and good body condition, and people above 40 years old could be lacking the physical strength to survive.