Data Mining Kaggle Competition

Team name: Jack & Ross

Team members: Anji Dong, Sunny Sun, Cloris Zhang, Yinglu Deng

The highest private rank and score:

The highest public rank and score:

Tie for second place, 0.83233

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

1. We first came up with a feature matrix that includes apparent useful features, such as age, fare, and gender. We threw away “ticket” and “cabin” initially because we could not think of any apparent relationship between these two features with survivals. Moreover, we left the numerical data as it was and converted categorical data into dummy variables. We also fill in the null values using mean. In this way, we came up with 8 features, and the accuracy was around 0.72.
2. In order to improve the model, we did a data exploration. We realized that age and fare(especially fare) have a wide range, and thus, we broke them into different numerical ranges (roughly by 10s as the gap between first, second and third quartiles are in intervals of 10). Moreover, we added more features. For example, by carefully extracting the letters from “cabin”, we realized that they could be converted into dummy variables as well, and thus we created such a feature. Moreover, we also extracted the numerical parts of the cabin, as we predicted the position of the cabin on the ship will result in a geographic advantage in terms of its closeness to lifeboats. Also, instead of filling the null values using the mean value of the whole training set, we grouped the set (group by gender and Pclass), and filled in the mean of the missing value with its group mean. We expanded the feature matrix to 13 columns. The accuracy increased to 0.80.
3. In the last step, we tried to improve our model. We used ensemble methods that averaged the predicted value from xgboost, catboost, logistic regression, and lgb. We also incorporated cross-validation and use Optuna to find the optimal hyperparameter for each model, and in this way, the accuracy increased to 0.83.

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?

We have tried many different models and made predictions separately. Then we also tried to combine the models' predictions and voted for the majority to get the final prediction. We found ensemble selection from libraries of models helps improve accuracy. In addition, methods such as grouping age and fares, filling null values with mean instead of 0 contribute to better accuracy.

From the model, we learned that Titanic survival outcome is highly depended on several predictors, such as sex, age and passenger class. In particular, while keeping other predicting variables constant, females are more likely to survive than male; older people are less likely to survive; children who have parents are more likely to survive. People from the First class have a higher chance to survive than the Second or the Third class people; people from a lower class are less likely to survive.

An interesting finding is that despite the numerical parts of Cabin having a wide range (from 0 to 100+) which could easily outweigh other variables, the numerical parts increase the model's accuracy the best when it is not altered. In other words, when trying to normalize and re-scale the numerical parts of Cabin, the model's accuracy would only decrease.