MATH 4995 FINAL PROJECT: TITANIC

Tsui Ying Tsz, Harjono Natasha Valerie

INTRODUCTION

We have chosen the 'Titanic: Machine Learning for Disaster' as our final project. In this project we will use machine learning techniques to create a model that predicts which passengers survived the Titanic shipwreck. In this report, we will discuss how we process the given data and choose a model to solve this prediction problem.

DATA OVERVIEW

> There are two datasets in csv format which is, the training set and the test set.

> The training set has **891** rows

> **12** columns with one observation identifier (PassengerId), one target variable (Survived)

> 10 predictor variables.

> test data has **418** rows with **11** columns, without the target variable column.

DATA ANALYSIS

> To show the relationship between each feature and the survival rate also we want to figure out if there are any correlation between one feature to another.

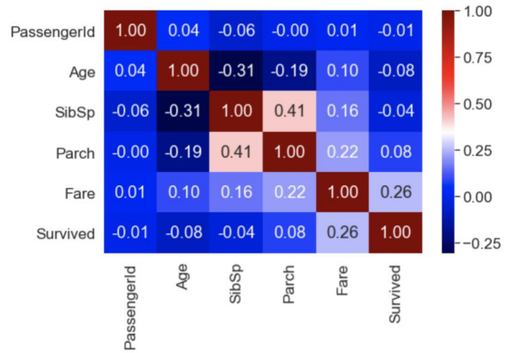
> We can see that people from Pclass 1 are very likely to survive with over 0.6 chance which is then followed by Pclass 2 with less than 0.5 and Pclass 3 around 0.2.

> Females are very likely to survive with over 0.7 probability rate.

> Children below the age of 10 have a high survival rate

> By visualizing the correlation between features, Fare is an important feature for Survived.

CORRELATION MAP OF FEATURES



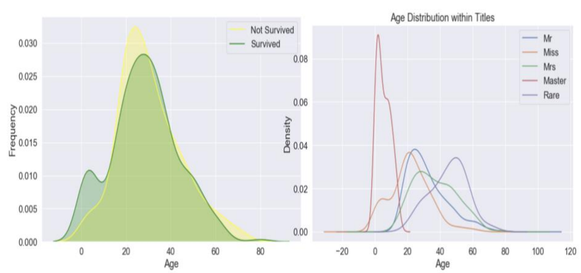
FEATURE ENGINEERING

> **Fare:** There are 1 missing value in Fare. Since the number is very small, we use the column's median to fill it. Then we used qcut to devide it into 5 intervals and depending on which interval it is. The column will display corresponding number.

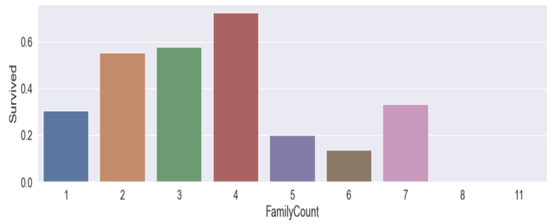
> **Age**: There are total of 233 missing values from both sets. We found that age influences the result easier when the passengers are young. We filled it with the median of the person's Title. Then we used qcut to be grouped it into 5.

> **FamilySurvival:** By checking the ticket number we could identify whether the passengers belong to a family. We create a new feature represents whether a person's relative survives.

AGE DISTRIBUTION WITHIN TITLES/ SURVIVAL



SURVIVED VS FAMILYCOUNT



FINALIZING

> We transform all categorical variable into indicator variable.

> To decided which feature to use, we have tried many combinations of features to be fed into our model. We found that using: Sex, Pclass, Fare, Age, and FamilySurvival will yeild the best result.

RANDOM FOREST

> We have chosen to use the Random Forest Classifier this time in accordance to the results we get from the first project.

> In the first project, we had found that the Random Forest model yields better results and are produce a more general or stable results compared to others such as Linear Regression, Logistic Regression, Decisions Tree, etc.

KNN

> We also decided to try using KNN. First, in order to find the right parameters so that we can get the maximum results, we use grid search cv.

> From that, we get the maximum accuracy of 89% for the training set when the n_neighbors are 14 and leaf_size are 16.

XGBOOST

> We also decided to try using Xgboost. Because Xgboost method added regularization term in the loss function. Not only using the first derivative, the second derivative is also included.

> With parameter tuning, the best performance score we got is 84% and when submitted online it is 79%.

MODEL EVALUATION

We have fed the data into the 3 models mentioned. Now, we want to explore further which model best fit our data. We use K-fold Cross Validation and compare the result of each model.

Model	Mean Accuracy	Std Accuracy
Random Forest	84.2859%	3.7654
KNN	83.9488%	3.9962
XGBoost	83.3845%	3.4548

From the table, we conclude that Random Forest model will produce the best results for our prediction.

CONCLUSION

> We feed the test data to the finalized model and finally obtain an accuracy score of 81.1% with an error rate of 19.9% on Kaggle website. Comparing to our first project, the accuracy has improved by 3.2%

CONTRIBUTION

> Tsui, Ying Tsz

Data Cleaning
Data Visualization
Model Selection
Model Evaluation

> Harjono, Natasha Valerie

Feature Engineering
Data Finalization
Model Selection
Hyperparameter Tuning