Titanic Survival Classification

By team Love&Peace

Team Member:
Shuo Jin shj42@pitt.edu
Chen Ruo chr87@pitt.edu
Fenghao Shi fes18@pitt.edu
Runzhong Wang ruw29@pitt.edu
Dong Yang doy16@pitt.edu

Introduction

As is known to us, the sinking of RMS Titanic is one of the most famous shipwrecks in history, based on which, the aim to this project is to develop a relative precise model of predicting "Survival" variable of Titanic data through utilizing different kinds of models with tools of machine learning. In this report, the approaches and working progress are presented, a semester plan is developed, and further questions are indicated at the end.

Approach and Method

First of all, our group cleaned the train data by dropping irrelevant variables and filling all the null value. Since the "cabin" variable has too much null value, we decide to drop the variable.

We decide to run the KNN model to give us a general idea regarding the performance of data. We will run Logistic Regression model, Decision Tree model, and Random Forest classification model for the next step of our project.

Since the Titanic data can be stored and processed in a single PC, we do not have difficulties in storing and retrieving data.

Working Progress

Data Cleaning:

map "Sex" variable as binary (0 / 1);
fill the null value of the "embark" with the mod;
map the "embark" to 1-3;
fill the null value of the "fare" with the mean of existed fare;

fill the null value of the “age” with the random distribution on the 95% confidence area of the normal distribution of the “age” ;
extract the title(e.g. Mr., Miss, Master) from the “name” variable;
map the titles as number from 1-5;
Drop “cabin” and “ticket” variable.

Running KNN model:

The accuracy of the KNN model prediction is 0.739910 according to our model.

Semester Plan

Based on our goal of developing a relative precise model of predicting “Survival” variable of Titanic data, our group decide to run different kinds of models and evaluate their performance respectively. So far we have done the data cleaning, such as filling the null values and dropping off irrelevant variables.

For the next step, we want to check if there are apparent dependencies among the variables using customized Apriori algorithm to try improving performance. After this, our group will train the cleaned data with Logistic Regression model, Decision Tree model, and Random Forest classification model. Then we test our model with all the models that we have. For each of the models, we will evaluate the model based on the Area Under Curve (AUC). Our final result of the project will go with the model who has the highest AUC.

Further Questions

We find that the accuracy might increase if we map the “age” and “fare” to the ranges. We will try this method in the future. If the accuracy is increased, we will go with this way.