1. Introduction

1.1 Background

As we know the famous incident of Titanic which is called as the ‘Unsinkable Ship’. But the ship collides with the iceberg and in result 1502 people dead out of 2224 people. We will analyse and study of person who survived in this tragedy.

1.2 Problem

The problem is to predict whether the person will survive or not based on some attribute like (name, age, sex etc). Consider such challenge we will make a predictive model which is used to train on our train data and classify the test data whether a person will survive or not.

2. Data

The data used in this challenge is taken from Kaggle site which is available in form of train and test. This train data is used for training and building our model which is used to predict whether a person will survive or not on test data.

2.1. Data Understanding

There are 891 rows and 9 columns or variables in train data in which 8 variables are the features used to train our model and one ‘Survived’ is a target variable. Variables and their data types shown as:

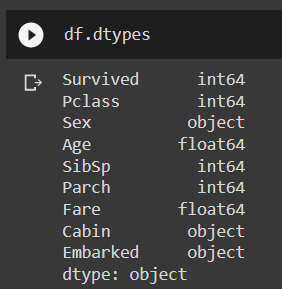


Fig1: Variable types

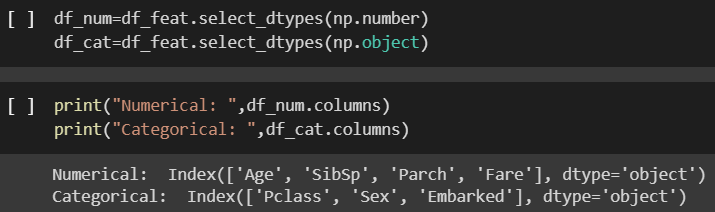


Fig2: Numerical and Categorical Variables

2.2. Exploratory Data Analysis

In this part the proper analysis of the type and origin of data is performed. The relationship between variable is explored. The relationship between the attributes is examined via the frequency. We see the summary of data. And dropped some variable which are not appropriate. Descriptions of each variable shown as:

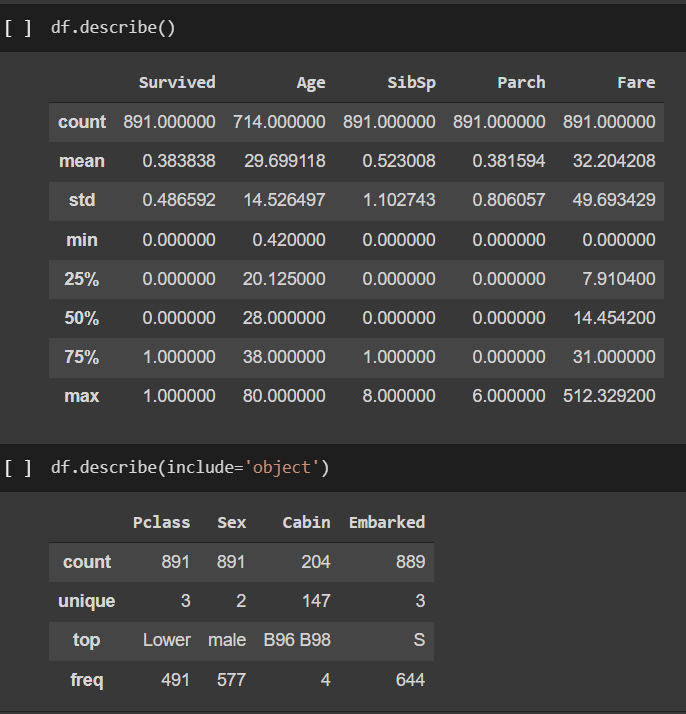


Fig3: Description of Numerical and Categorical Variables

2.3. Data Preparation

- Here we are dealing with the null or missing values in the data, If the missing values is more than 70%, we dropped that variable and if it is less than 70%, we will do the imputation on them. Null values in our data shown as:

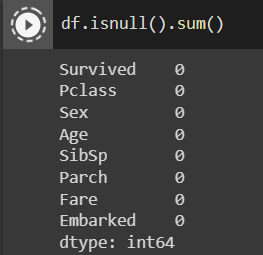
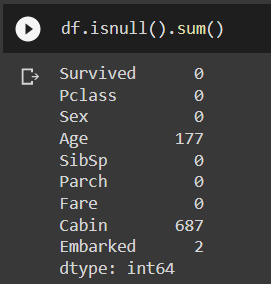


Fig4: Before removing null (left side) and After removing null (right side)

- As we can see from the above fig that Age missing values is imputed with mean and Cabin variable is dropped because of high percent of missing values.

- After that we will deal with the outliers and remove the extreme values through inter quartile range.

- Then, we do the preparation of data for modelling in which we do the dummy encoding on the categorical variable and split the data in 80:20 ratio.

3. Model

For this problem I use many classifications algorithm to classify whether a person will survive or not when the Titanic sinks. List of algorithms which is applied on our train data are as follows:

- First, I build a model using Logistic Regression.

- Second, I build a model using K Nearest Neighbour.

- Third, I build a model using Random Forest.

- Fourth, I build a model using Xtreme Boosting.

- And finally, I build a model using Support Vector Machine.

4. Result

This is our final phase where we will predict whether a person will survive or not from our test data and calculate the accuracy score of each model, based on this accuracy scores of each model we will decide which is a good model as for now. Here is the table of models with their accuracy score shown as:

|  |  |
| --- | --- |
| Model | Accuracy Score |
| Logistic Regression | 76.55 % |
| K Nearest Neighbour | 74.88 % |
| Random Forest | 77.99 % |
| Xtreme Boosting | 76.07 % |
| Support Vector Machine | 77.27 % |

5. Conclusion

From above table we can conclude that the Random Forest and Support Vector Machine model giving good accuracy score. Here we will select Random Forest model for our predictions as it is best among all. Further we can improve our model by tuning the hyperparameters and also we can improve our model by applying some feature engineering techniques to select appropriate features.