Prediction of survival on the Titanic

Antar Karmakar 2016-2-60-113,
MD.Rabbi Hasan 2016-2-60-111,
Rubayat Ara Jannat 2016-1-63-004

May 15, 2020

1 Introduction

In April 15,1912 the most infamous disaster occurred,which is very well known as sinking of "The Titanic". The RMS Titanic was known as the unsinkable ship and was the largest, most luxurious passenger ship of its time.The sinking of the RMS Titanic is one of the most shocking shipwrecks in history. The collision with the iceberg ripped of many parts of the titanic. The dead included a large number of men whose place was given to the many women and children on board as there were shortage of life boats to rescue. Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking of the Titanic. Features like age,sex,ticket fare, class will be used to make the prediction. In this challenge,the analysis of what sorts of people were likely to survive will be completed. Using the Machine learning algorithms, survival is predicted on different combinations of features.

1.1 Objectives

- The main objective of this project is to build a classification model that can successfully determine whether a Titanic passenger can live or die.
- Predicting survival on the Titanic using multiple machine learning techniques.
- Analyzing the Titanic dataset to obtain useful insights.

- By applying algorithms the data analysis will be done and accuracy will also be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions.

1.2 Motivation

The purpose of this project is to document the process we go through to create the predictions for Titanic Survivor Prediction. Machine learning is used for predicting the outcome which is interesting. Using the machine learning algorithms it is possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results. This research aims to use machine learning techniques on the Titanic data to analyze the data for classification and to predict the survival of the Titanic passengers by using Machine Learning algorithms. The prediction and efficiency of these algorithms depend greatly on data analysis and the model. Our project is about the prediction of survival on the titanic. The titanic dataset is one of the most widely used datasets in the field of machine learning. From our model, we will able to determine who was survived or who was died and we will be able to know that which classification can give more accurate result. The algorithm which can give more accuracy will be the best model for prediction of survival on the titanic. By knowing the survivals from the Titanic tragedy we can define that the people in between which age and the rate of which sex are more to survive.

1.3 Existing works

Eric Lam and Tang used the Titanic problem to compare and contrast between three algorithms- Naive Bayes, Decision tree analysis and SVM. They concluded that sex was the most dominant feature in accurately predicting the survival. They also suggested that choosing important features for obtaining better results is important. There are no significant differences in accuracy between the three methods they used [1].

Akriiti Singh, Shipra Saraswat and Neetu Faujdar used Naive Bayes, Decision Tree, Random Forest have been implemented to predict the survival of passengers. In particular, this research work compares the algorithm on the basis of the percentage of accuracy on a test dataset[2]

1.4 Necessity

The necessity of this project is to find the model which gives the more accuracy. The goal is to predict if a passenger of Titanic will survive or not. This

problem can be solved by using many Machine Learning algorithms. Here any of the models to predict survival of test sample can be chosen. Since we have evaluated all algorithms, we will predict by using model which has highest accuracy. After developing this project we get the finest algorithm with the highest accuracy. So for this reason someone should invest his time to develop this project. If the project is not done properly then some problems may occurs. Such as, peoplecannot get the actual result of prediction and the wrong prediction can change the accuracy of the algorithm.

2 Methodology

At first, we imported the Pandas library to perform the operations on the collected dataset. Then we read the CSV file of the dataset to display. After that, we analyzed the data with some features such as 'Survive',' Sex' and 'P-class'. Then we dropped unwanted columns such as PassengerId, Name, SibSp, Parch, Ticket, Cabin, Embarked. Because these columns could affect the results. We will use the updated dataset which has rest 5 columns now such as Survived, Pclass, sex, age and Fare. After that, we labeled the 'Survived' column as output and the rest of the updated dataset as inputs. These inputs were provided to our machine and our machine was able to determine whether the passengers survived or died. In the 'Survived' column we see numeric data 0 and 1. These are actually binary classes. Those who died were put in the '0' class and those who survived were put in '1' class. We assigned the variable x for inputs and y for output. In the column 'Sex' there are male and female which are non-numeric. So we mapped the male as 1 and female as 2 using the map function. Because It became helpful to label encode the non-numeric data in the column. And then, we imported train_test_split from sklearn library to split arrays into random train and test subsets. With the train data, we trained our machine and with the test data, we tested our machine if our machine was able to find the accurate result. We took 20% of the values for testing. The length of training data is 712 and the length of the test data is 179. We performed machine learning algorithms to find accuracy and predict the survival of the titanic. .

3 Implementation

We used programming language Python and its libraries NumPy (to perform matrix operations), Pandas and SciKit-Learn (to apply machine learning algorithms). Then we applied several machine learning algorithms (decision

tree, random forests, SVM, naive bias) For data analyzing we used Matplotlib Library.

3.1 Data Collection

We collected the dataset from Kaggle.com. Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals. The dataset has 12 columns

pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children traveled only with a nanny, therefore parch=0 for them[1].These are the columns of the dataset.

3.2 Data Processing

1. We dropped unwanted columns such as PassengerId, Name, SibSp, Parch, Ticket, Cabin, Embarked. Because these columns could affect the results. We will use the updated dataset which has rest 5 columns now such as Survived, Pclass, sex, age and Fare

2. We mapped the male as 1 and female as 2 using the map function. Because It became helpful to label encode the non-numeric data in the column.

3. We checked in the column 'Age' whether there was any missing value or not. Then we found 177 missing values in that column. After that, we filled the missing values with the mean of all values in the column 'Age'.

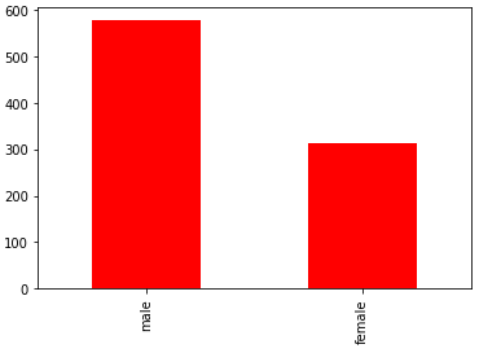

Figure 1: male and female

3.3 Model Development

We used programming language Python and its libraries NumPy (to perform matrix operations), Pandas and SciKit-Learn (to apply machine learning algorithms). Then we applied several machine learning algorithms (decision tree, random forests, SVM, naive bias) For data analyzing we used Matplotlib Library. Then we analyzed the data of Survival, Sex and P-class. we applied the decision tree algorithm. For that, we imported tree from sklearn library. After that, we trained our data with this algorithm. Then we checked the accuracy with our test data. Then we predicted the result of survival or died with given inputs.

3.4 Results

There were 577 males and 314 females of the passenger in the Titanic. We implemented a bar chart using matplotlib library to view this.

There were 3 types of passenger class. Using python and matplotlib library we sorted by index because we wanted to make sure the classes 1st, 2nd, and 3rd are displayed in the correct order. From this plot we see most of the passengers were from class 3rd.

And then we analyzed the data from 'Survived' feature. 0 means died and 1

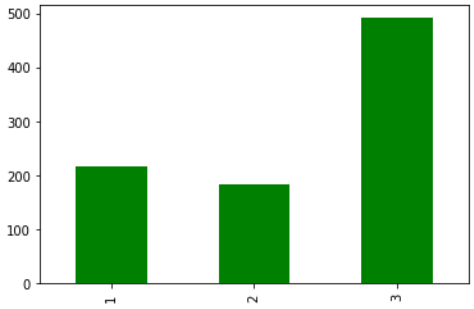

Figure 2: types of passenger

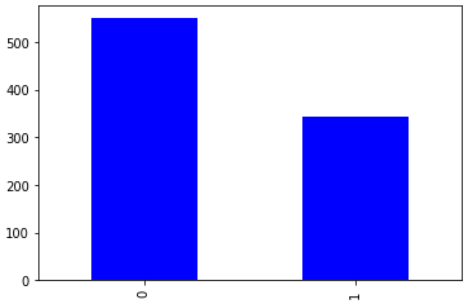

Figure 3: Survived chart

means survived. So we can see most of the passengers were died.

After that, we applied the decision tree algorithm. For that, we imported tree from sklearn library. After that, we trained our data with this algorithm. Then we checked the accuracy with our test data. Then we predicted the result of survival or died with given inputs. Secondly, we applied the random forest classification. The accuracy of the random forest was the highest. And then We applied Support Vector Machine and Naive Bias. We saw the accuracy of SVM and naive bias are same but lower compared to the decision tree and random forests.

We considered Pclass, Sex, Age and Fair as inputs. We considered output or target is Survival. We took input randomly from our dataset to predict. We used Decision Tree, Random Forests, Support Vector Machine and Naive Bias classification.

Classification	Classification
Decision Tree	76.536%
Random Forests	83.240%
Support Vector Machine	63.687%
Naive Bias	63.687%

We considered inputs,

P-class = 1

Sex = Male (male:1 , female: 2)

Age = 37

Fair = 29.7

If we look at the actual dataset according to these inputs, we see the person is died. Our model also predicted the accurate result - The person is died.

4 Conclusions

In this project, we came to the conclusion that Random forest classification gives the best accuracy and SVM classification and Naive bais classification gives the lowest and same accuracy. We analyzed the data correctly by using matplotlib library.

4.1 Limitations

Sometimes there was a wrong prediction of Random forest and Decision tree classification. But most of the time all classification that we used in our project predicted correctly.

4.2 Future Directions

This model can be improved more by working on it and we can apply our model in other problems so that our model can give the best accuracy and predict correctly. From our model, we can find out the more mystery on Titanic with the dataset.

References

[1] article, *Analyzing Titanic Disaster using Machine Learning Algorithms*, Nair, Drs, Volume-2, International Journal of Trend in Scientific Research and Development, page[410-416], 2017.

[2] A. Singh, S. Saraswat and N. Faujdar, *2017 International Conference on Computing, Communication and Automation (ICCCA)*, title=Analyzing Titanic disaster using machine learning algorithms, year=2017 pages=406-411