TITANIC-MACHINE LEARNING FROM DISASTER

Tanuj Raghav

19/11/EC/027

Pragyan Jaiminy

19/11/EC/028

Pushpak Prateek

19/11/EC/039

Abstract:

In April,1912 the "RMS" Titanic claimed to be unsinkable, went for it's maiden voyage from Southampton, UK to New York City. Unfortunately, it met with an accident with an iceberg and sank, claiming lives of over 1500 people out of 2204 passengers aboard. While luck was an important factor in surviving the ship wreck, it was later found that some groups had better chances of surviving than other.

This research aims to use machine learning techniques on the given Kaggle dataset to predict the survival rate of the passengers aboard using several data-mining models including Linear Regression, Random Forest etc. The computing work was done by the help of Google Collab.

I. Introduction

The Kaggle dataset given to us, provided in comma separated value(csv) format, consists of 3 sub parts i.e., train.csv, test.csv, gender submission.csv. The train.csv file had 891 passenger ids whereas the other two files had 418 ids each.

In the training set, we had used several features of the data like the passenger's gender, class, profession etc. Further, as we knew the outcomes (often called as 'Ground truth'), we tried to engineer some new features in our dataset in order to make it simpler and of course to train our model in a better way. Our best results were obtained using the linear regression method, giving an accuracy of more than 80 %.

Our Contributions:

We had used the train.csv file to train our data set and then the test.csv file to check how accurately our model is performing. The submission.csv file helped us understand how we have to make the submissions in order to get our scores.

First of all, we deeply studied and understood the dataset. The data had several variables already in use like the pclass, sex, age, cabin number, fare etc., we added some more features which were helpful while we were training the model.

We had first tried out KNN then Random Forest and then Linear Regression to proceed in our project. This helped us utilize the knowledge we had gained in real life.

Further, we have created a website for regular people to check whether an arbitrary person would have survived the wreck on the basis of our trained model.

II. Discussions

To accurately predict who would have survived the incident only on the basis of the class, cabin numbers, ticket price etc. is a difficult task. The survival rate had been affected by several other factors like luck, position where they're at the time of the vessel colliding with the iceberg or to how soon they got the information of abandoning the ship.

Also, as the general notion assumes, female, elderly and children would have survived the crash wasn't entirely true, several men from the richer class or better professions had a good rate of survival.

In our project, we have tried to provide several graphs for masses to better understand the statistics related to the survival data.

Dataset:

The dataset had 891 entries of the passengers with sort of idea of how the demography was of the 'HMS Titanic'. There were 65% males and 35% females in the passengers of which the 16-24 age group people were in the majority when compared to other age groups.
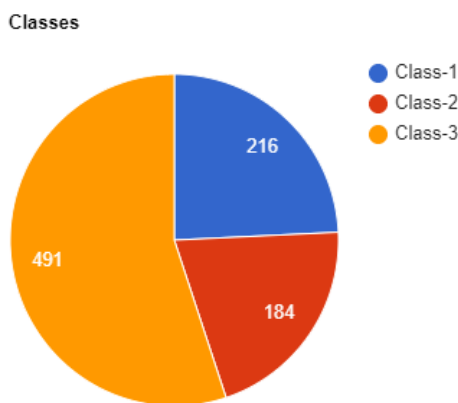
Figure-1 and 2 give a rough idea of how the data was:
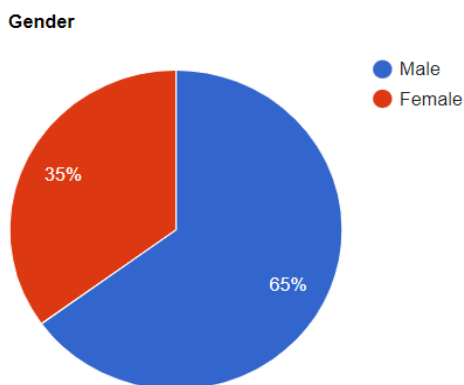

Fig-1

Fig-2

The train.csv file's data had to be cleaned and missing entries had to be checked. For example: The 'Age' feature had ~20% entries missing which was creating issue while training the model.

III.   Methodology
1.   Data Analysis and Preprocessing:

Before working on the dataset given to us by Kaggle, we sat down to make the dataset more precise and smaller. After that, we had analyzed the data so as to better understand and decide our approach towards this project problem.

2.   Data Cleaning

Right after deciding our approach, we had started by cleaning the data. The false/repeated/empty entries were removed/modified from the dataset so as to make our dataset ready to work upon.

3.   Feature Engineering

After getting the model eligible to be worked upon, we started to decide upon the features which have to be included in the data training and testing.

4.   Base Models

We had used several models including the Linear Regression, KNN, Random Forest to train our models and noted down the relevant scores of them. The linear regression looked a natural choice due to its easy nature and better scores.

5.   Linear Regression Model

The Linear Regression method is based on the simple mathematics of plotting the straight-line graph between dependent and independent variables. This would provide us a slopped line graph with datapoints lying close to/on the slopped line. For reference, see the below figure:
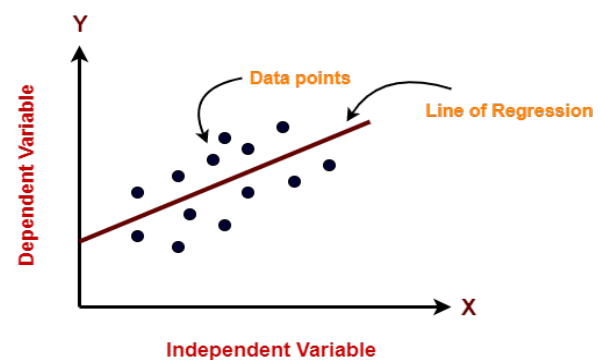
For the best predictive model, we have to find the best slopped line i.e., the error between the actual and predicted values must be minimized.

Mathematically, the linear regression model can be easily represented by a straight-line equation:

$$\Rightarrow \quad y = a_0 + a_1 x + \varepsilon$$

IV. Results and Analysis

The major known algorithms (9) including the famous Linear Regression (~80% accuracy), KNN (~72% accuracy), Random Forest (~76% accuracy) were used upon the data.

The results achieved through our models were of ~80% accuracy, through this we could conclude several things like :
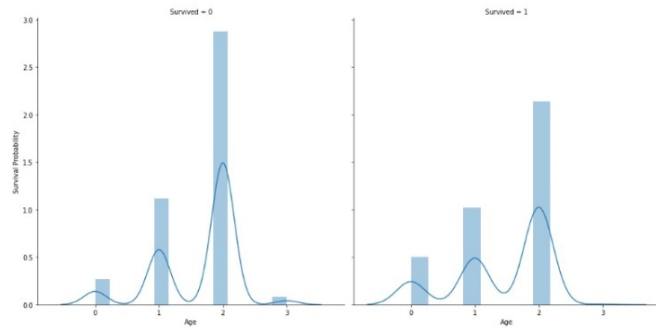
1. Some classes had a better rate of survival than the other classes. (Class-1 had the best rate)
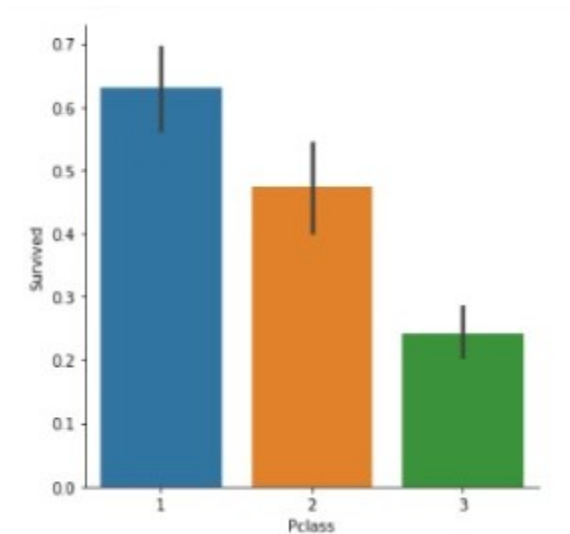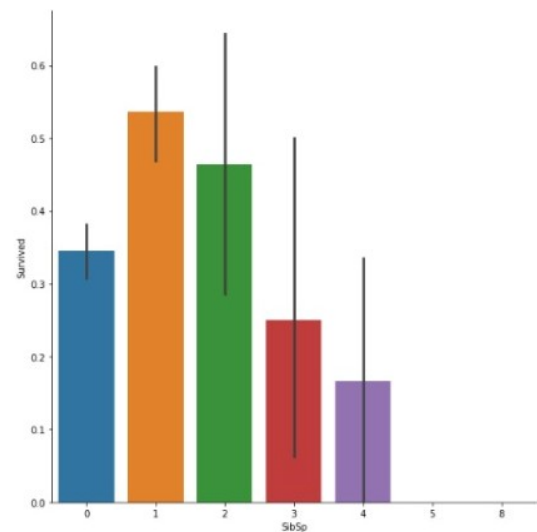

2. Females had better rate of survival than Men.
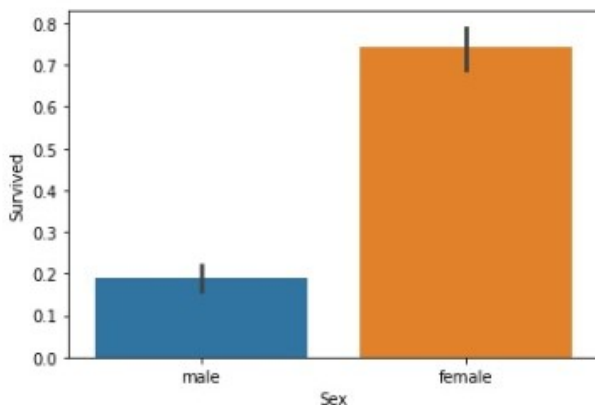
3. Some Age groups had better survival rates than others.

4. People who had families had better survival rate:

V. Conclusion

After implementing various models on the given datasets, we could improve the RMSE score and improve our accuracy. Through this we could infer that Linear Regression model worked best in our case.

VI. Acknowledgement

We would like to thank our instructor Dr. Prenana Ma'am and Prof. TV Vijay for their constant support and help.

VII. References

1. Kaggle Website- Titanic Challenge
2. Wikipedia
3. Data Research Papers available on ACM Digital Library
4. IEEE Website.