Titanic-Machine-Learning-from-Disaster

Internship Project Report,
Based On Machine Learning & Data Science.

Objective :- Predict the survival of the passengers aboard RMS Titanic.

Conducted by :-

"Abhiyantrix & Sapience Academy Mysuru"

Under Supervision Of :
Mr.Krishna

Submitted By :
Preethi Barman (4AD16CS058)
Vishnu Tej k (4AD16CS097)

Acknowledgments

We are pleased to acknowledge **Mr.Krishna** for his invaluable guidance during the course of this project work.

We extend our sincere thanks to **Miss.Anjana Shashi Kiran** and **Mr.Harshith** who continuously helped us throughout the internship.

We would also be grateful to other members and team of **“Abhiyantrix & Sapience Academy Mysuru”**

Contents

Introduction

Overview

Titanic Data Set, As the name suggests (no points for guessing), this data set provides the data on all the passengers who were aboard the RMS Titanic when it sank on 15 April 1912 after colliding with an iceberg in the North Atlantic ocean. It is the most commonly used and referred to data set for beginners in data science. With 891 rows and 12 columns, this data set provides a combination of variables based on personal characteristics such as age, class of ticket and sex, and tests one's classification skills.

Background and Motivation

Data science can add value to any business who can use their data well. From statistics and insights across workflows and hiring new candidates, to helping senior staff make better-informed decisions, data science is valuable to any company in any industry.

Machine Learning is the core subarea of artificial intelligence. It makes computers get into a self-learning mode without explicit programming.

Objective

Predict the survival of the passengers aboard RMS Titanic.

Methodology

1. **Data visualization**: data analysis to understand missing values, data relations and usefulness of features
2. **Preprocessing**: with the knowledge acquired with the preceding step, apply preprocessing of data including dealing with missing values, drop unuseful features and build new features
3. **Classifier**: build classifiers based on the preprocessed data using a variety of techniques

Project Description

This project has been made in Python v3.4. It uses various data processing, visualisation and machine learning packages such as numpy, pandas, matplotlib, scikit-learn.
The project uses a 5 step process (general procedure) for it's predicting task which is as follows [2]:

1. Perform a statistical analysis of the data and look over it's characteristics such as data type of columns, number of instances, correlation of each attribute with the output variable, finding mean and other information about data, correlation matrix etc.
2. After performing statistical analysis, do a visual analysis by plotting the data. Do analyse the scatter_matrix, plot box plots etc. so as to know which attributes are relevant and which are not. Remove irrelevant attributes from the dataset for further analysis.
3. Make a list of all machine learning algorithms that can give good prediction results and spot check each one of them (apply each one of them on the dataset) to find which one is better for prediction.
4. Take some of the good performing algorithms and perform a grid search/ randomised search over it's hyperparameters to find the optimal hyperparameters for the prediction task.
5. Use an ensemble or Voting Classifier on the above selected algorithms to achieve better performance or use any one of the above algorithm directly to perform predictions.
6. Store the predicted values in a new file.

Keep iterating over the above steps again and again and tune them according to the need so as to achieve better performance.

Analysis of the Titanic Dataset

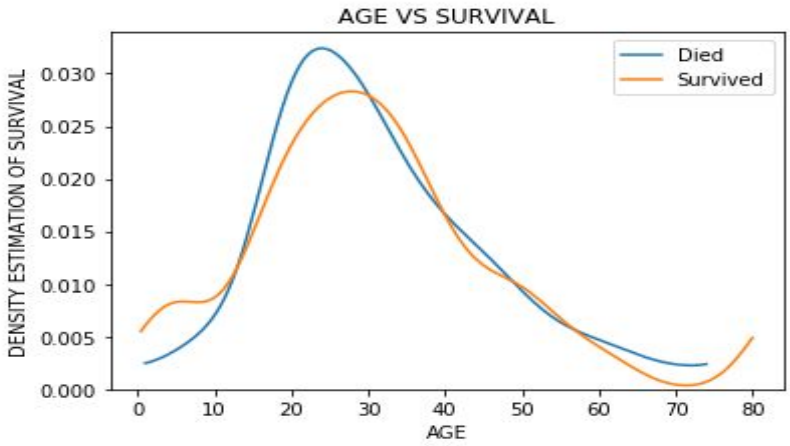
Did individual characteristics played any role in the survival rate?

1. Age
2. Gender
3. Adult or child

Let's Consider one by one.

Age : -

Based on the analysis, we observer the following graph.



> Analysis says lesser the age, survival rate is high.

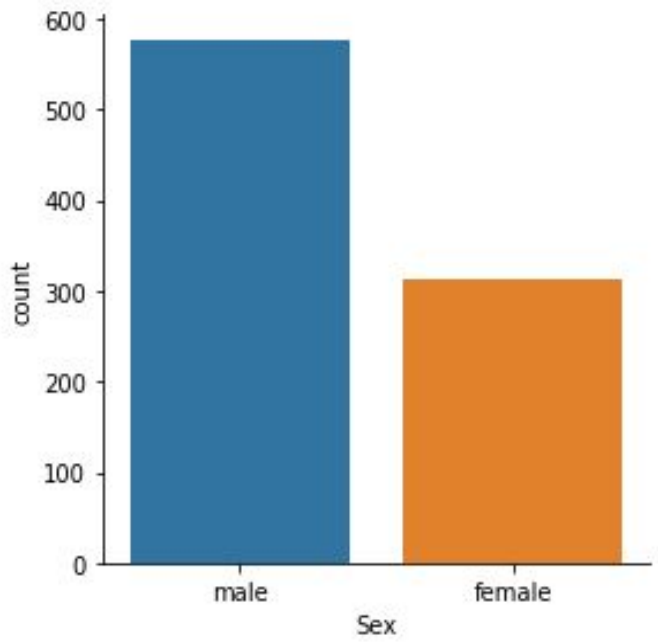
With reference to the above analysis we observe that the age had slight role, to decide the survival probability.

We can observe that lesser the age, more the probability of survival.

We can't decide or predict the actual probability based on this on factor.

Gender :-

Based on the analysis, we observer the following graph.

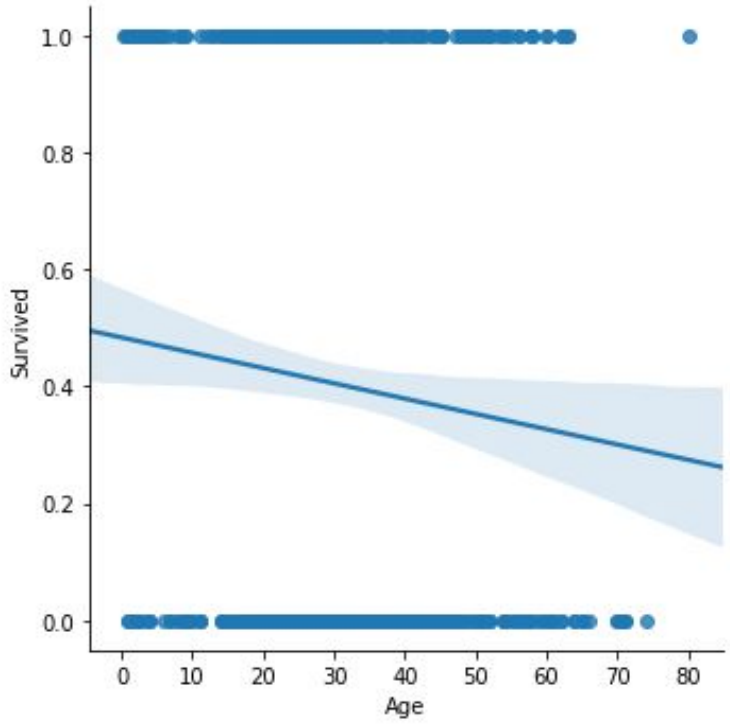


> Analysis says male were more in number.

The above plot, represents the statistical analysis of the ratio of sex.
Based on the graph, we can observe that the presence of male dominance.

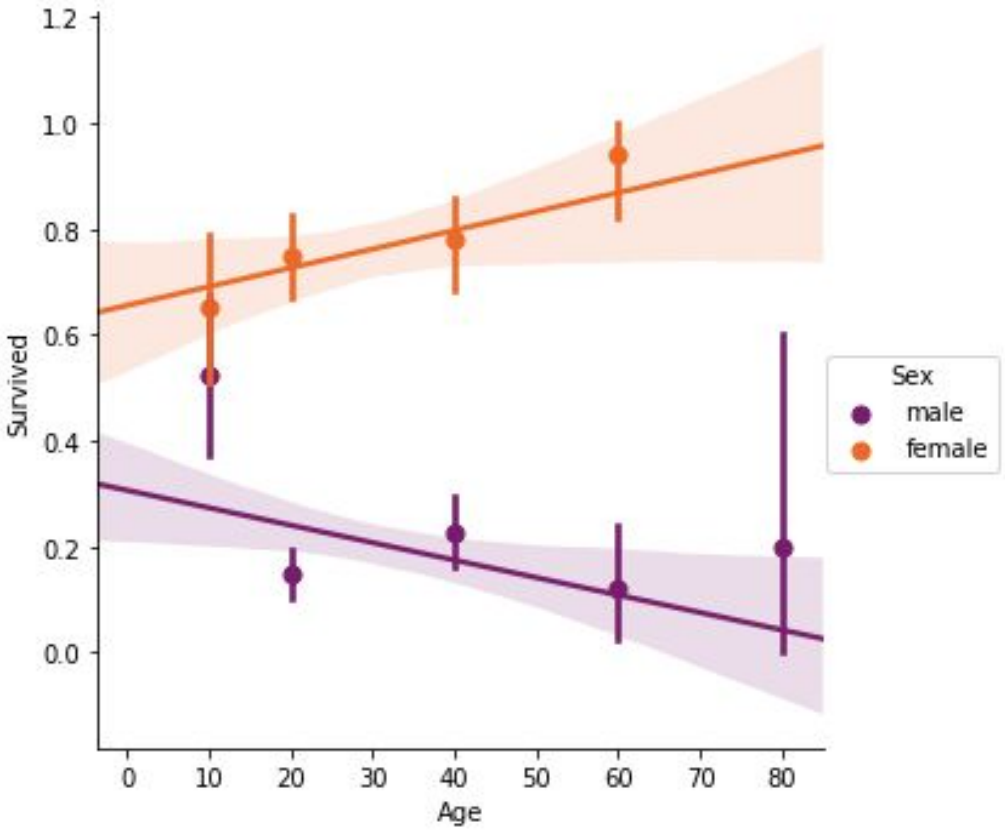
Linear plot of Analysis.

Based on Survival rates.



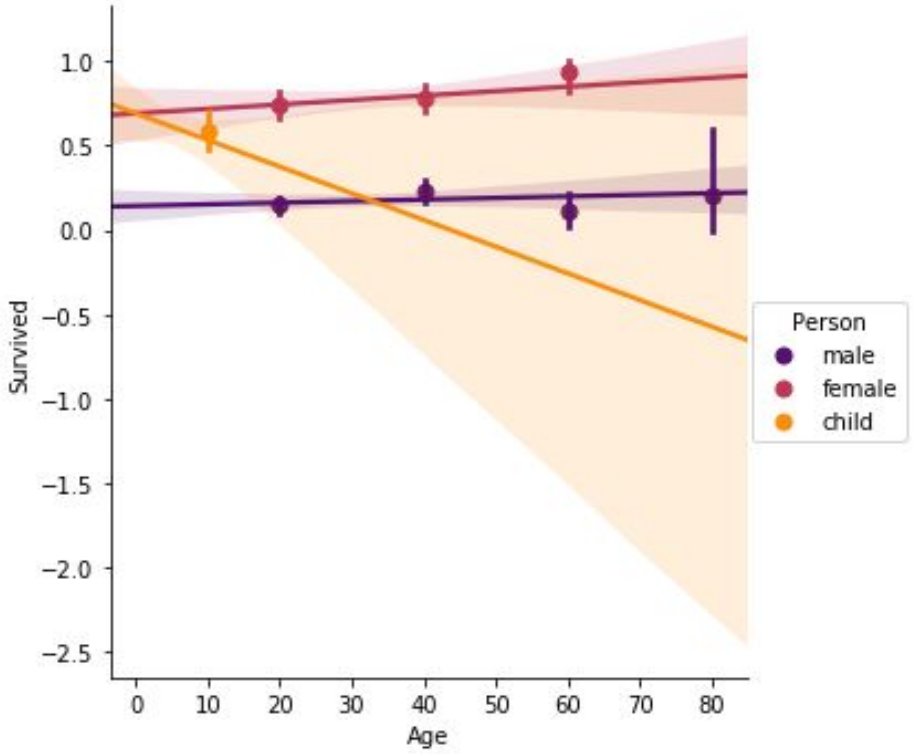
We see that when we considered Age and Sex, Survival Rate increased with increase in Age for Females but it decreased for Males. When we delved deeper, we saw that only for the females in the First Class, Survival Rate increased with Age, but for the other two classes, it decreased with Age. The outliers can actually affect the overall shape of the curves here. In the Age vs Survival stacked Histogram, we saw that younger people and older people actually had a higher survival rate than middle aged people which confirms to the protocol maintained. In the box plot **'Distribution of Age by Survival'**, we saw that the

Died group's IQR pertains to older people while the **Survived** group's IQR pertains to younger people. So except the two extreme ends, we see that younger people had a better chance of Survival.



Gender played another vital role in the Survival Rate of the passenger. We see from general trends that Females and Children were given more priority towards the lifeboats than the Males. When we look at the Box Plot **'Distribution of Age by Gender and Survival'**, we see that for the Male group, most of the **Not Survived** group IQR favoured the group of older people while most of the **Survived** group's IQR pertained to the Younger crowd. But when we see the same data for the Female crowd, we observe something interesting. We see that for the Female population, older age actually favoured the Survival Rate than Younger age. So maybe here, the older females were given more attention than younger females.

When we see the overall proabability of survival based on the Gender, we saw that Female population had a much higher chances of Survival**(74%)** than that of the Male population **(19%)**.

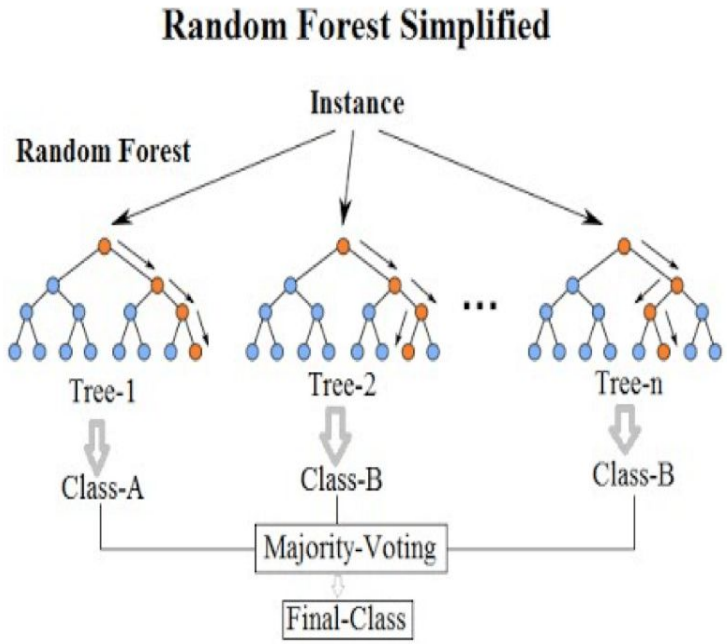


We saw from our analysis, that being a children helped in an increase of the Survival Rate. In almost every 'Passenger Class', we see that the children have a higher proabability of survival.

They are many more other factors to be considered and should be analyzed,
But these three factors had a major role in predicting.

Prediction Algorithm used is Random forest regressor.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.



(image source: YouTube)

Conclusion:

In the given **Titanic** Dataset, we divided it into different groups like *Gender*, *Age* and calculated different statistics on these groups based on the **Survival Rate**. We also arrived at a few derived columns in our dataframe like *Adult-Child* to get more information on whether the **Survival Rate** depended on the analysis.
For these parameters, we calculated basic statistics and also investigated Survival Statistics along with change in one of the factors or a combination of them.
And predicted the survival rate.

References

[1] Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)
[2] Titanic: Link to the Data Set: https://www.kaggle.com/c/titanic/data
[3] Algorithm referred :
(https://machinelearningmastery.com/implement-random-forest-scratch-python/)

Git-Link

Preethi Barman : https://github.com/preethibarman99/DataScience_Project
Vishnu Tej K : https://github.com/Vishnutejk98/DataScience_Project

Project Files

data

- train.csv : training data.
- test.csv : test data.
- program : project.ipynb

submission

- summit.csv : the prediction of the solution.