Titanic Data Analysis

Akanksha Mutha
Computer Engineering Department
Rajiv Gandhi Institute of Technology, Mumbai, India.
akankshamutha2@gmail.com

Kanishka Patel
Computer Engineering Department
Rajiv Gandhi Institute of Technology, Mumbai, India.
kanishkapatel22@gmail.com

Shivani Rawat
Computer Engineering Department
Rajiv Gandhi Institute of Technology, Mumbai, India.
shivanirawat121@gmail.com

Abstract*—Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for decision-making. The sinking of the Titanic is one of the most historic shipwrecks of all time. The tragedy killed thousands and led many wondering what could have been done better. There were some groups of people that were more likely to survive than others. A data analytical study is conducted with the passenger's data from the Titanic from a data platform Kaggle to find out about this survival likelihood. For the data analytical approach, we use PySpark, Pandas ,Seaborn and other Python modules, to come up with models that can best predict what kinds of passengers are more likely to survive. Various machine learning algorithms namely Logistic Regression, Naïve Bayes, Decision Tree are implemented to predict the survival of passengers.

Keywords*—Data analysis, PySpark , Pandas, Seaborn , Logistic Regression, Naïve Bayes, Decision Tree

I. INTRODUCTION

The Big Data trend is quite noticeable lately. New theories and tools become available for many unsolved old questions such as the Titanic that sunk in 1912. While many studies explained the human and hydraulic cause of the sinking, questions remained regarding the chances of survival for the passengers. This renewed interests is mainly derived from a Kaggle competition. Being able to understand data is a key competence of many companies and analysts nowadays. Regression, classification and machine learning theories are identified as one of the foundational technologies and emerging data analytic.

This is focused around the analysis of a historical dataset from the Titanic tragedy . This dataset has been studied and analyzed using various machine learning algorithms like Random Forest, SVM etc. Our approach is centered on R and Python for executing algorithms- Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest. The prime objective of the research is to analyze Titanic disaster to determine a correlation between the survival of passengers and characteristics of the passengers using various machine learning algorithms. The dataset has been studied and analyzed using various Big data modules like Matplotlib, Pandas, seaborn etc.

II. LITERATURE REVIEW

The dataset used for the paper is provided by the Kaggle website. The data consists of 891 rows in the train set which is a passenger sample with their associated labels. For each passenger, the name of the passenger, sex, age, his or her passenger class, number of siblings or spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation were provided. The data is in the form of a CSV (Comma Separated Value) file. For the test data, the website provided a sample of 418 passengers in the same CSV format. Before building a model, data exploration is done to determine what all factors or attributes can prove beneficial while creating the classifier for prediction. [1]

The logistic regression gives the accuracy of 95% which is based on the confusion matrix. The parameters used here are accuracy and false discovery rate. This would prove dangerous as the prediction may go wrong and hampers the accuracy of the results. The attempts are being made to increase the accuracy rate and reduce the false discovery rates. It works better with binary dependent variable which means the variable has a binary value as its output like yes or no, true or false. [2]

Various algorithms were compared on the basis of accuracy. The four algorithms were Naïve Bayes, Logistic Regression, Decision Tree and Random Forest. Two metrics were used to compare the four classification techniques. First metric is accuracy and the second metric is false discovery rate. It was observed that Logistic Regression proved to be the best algorithm for the Titanic classification problem since the accuracy of Logistic Regression is the highest and the false discovery rate is the lowest as compared to all other implemented algorithms. It also determined the features that were the most significant for the prediction. Logistic regression suggested that Pclass, sex, age, children and SibSp are the features that are correlated to the survival of the passengers.[3]

III. METHODOLOGY

Before building a model, data exploration is done to determine what all factors or attributes can prove beneficial while creating the classifier for prediction. To start the exploration, few X-Y generic plots are made to get an overall idea for each attribute

.

Attributes	Description
PassengerID	Identification no. of the Passengers.
Pclass	Passenger class (1, 2 or 3)
Name	Name of the passengers
Sex	Gender of the passengers (male or female)
Age	Age of the passenger
SibSp	Number of siblings or spouse on the ship
Parch	Number of parents or children on the ship
Ticket	Ticket number
Fare	Price of the ticket
Cabin	Cabin number of the passenger
Embarked	Port of embarkation (Cherbourg, Queenstown or Southampton)
Survived	Target variable (values 0 for perished and 1 for survived)

Fig: Attribute Of the Dataset

- **PySpark**
  PySpark is the Python API written in python to support Apache Spark. Apache Spark is a distributed framework that can handle Big Data analysis. Apache Spark is written in Scala and can be integrated with Python, Scala, Java, R, SQL languages. Spark is basically a computational engine, that works with huge sets of data by processing them in parallel and batch systems.
  Advantages of PySpark are:
  - Easy Integration with other languages: PySpark framework supports other languages like Scala, Java, R.
  - RDD: PySpark basically helps data scientists to easily work with Resilient Distributed Datasets.
  - Speed: This framework is known for its greater speed compared with the other traditional data processing frameworks.
  - Caching and Disk persistence: This has a powerful caching and disk persistence mechanism for datasets that make it incredibly faster and better than others.

- **Pandas**
  In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.
  Features:
  DataFrame object for data manipulation with integrated indexing.Tools for reading and writing data between in-memory data structures and different file formats.Data alignment and integrated handling of missing data.Reshaping and pivoting of data sets.Label-based slicing, fancy indexing, and subsetting of large data sets.Data structure column insertion and deletion.Group by engine allowing split-apply-combine operations on data sets.

- **Seaborn**
  Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

  Here is some of the functionality that seaborn offers:

  - A dataset-oriented API for examining relationships between multiple variables
  - Specialized support for using categorical variables to show observations or aggregate statistics
  - Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data
  - Automatic estimation and plotting of linear regression models for different kinds dependent variables
  - Convenient views onto the overall structure of complex datasets
  - High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations
  - Concise control over matplotlib figure styling with several built-in themes
  - Tools for choosing color palettes that faithfully reveal patterns in your data

  Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

- **Matplotlib**
  Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

- **Tkinter**
  The Tkinter module (“Tk interface”) is the standard Python interface to the Tk GUI toolkit from Scriptics (formerly developed by Sun Labs).
  Both Tk and Tkinter are available on most Unix platforms, as well as on Windows and Macintosh systems. Starting with the 8.0 release, Tk offers native look and feel on all platforms.
  Tkinter consists of a number of modules. The Tk interface is provided by a binary extension module named _tkinter. This module contains the low-level

interface to Tk, and should never be used directly by application programmers. It is usually a shared library (or DLL), but might in some cases be statically linked with the Python interpreter.

IV. RESULT

Some of the generic plots have been shown below. The age plot suggested that maximum or majority of the passengers belonged to the age group of 20-40.

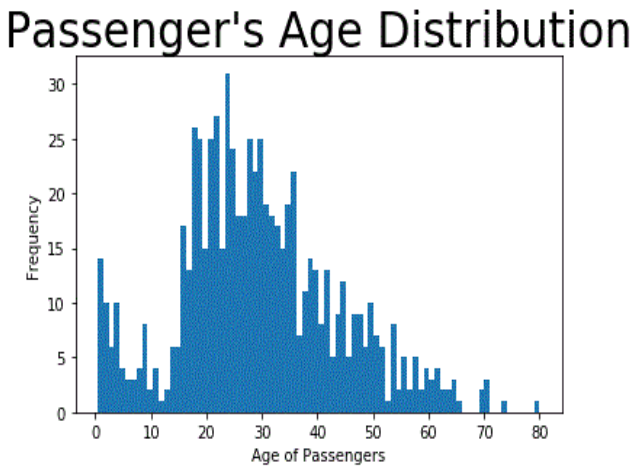

Fig: Passenger's Age Distribution

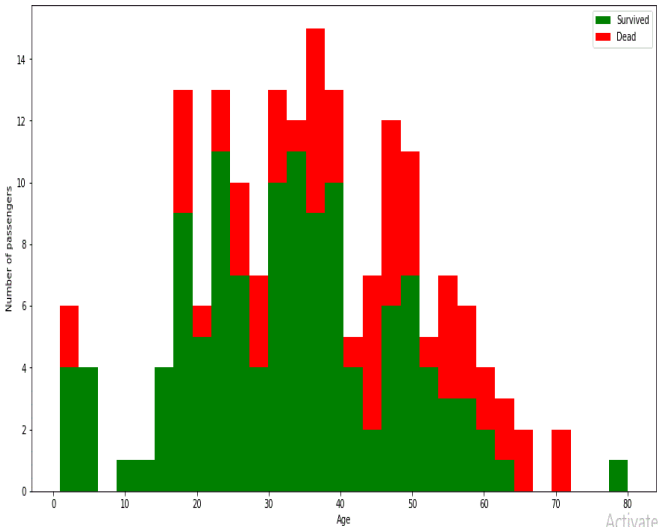

Fig: Survival Based on Age

Similarly, a graph is plotted and some calculations are performed for the sex attribute and the results suggested that the survival rate of the female is 25.67% higher to that of the male.

```
No of males: 577
No of females: 314
```

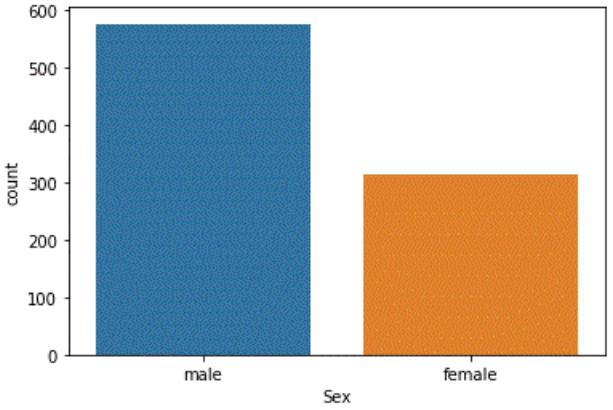

Fig: Distribution Based On Sex

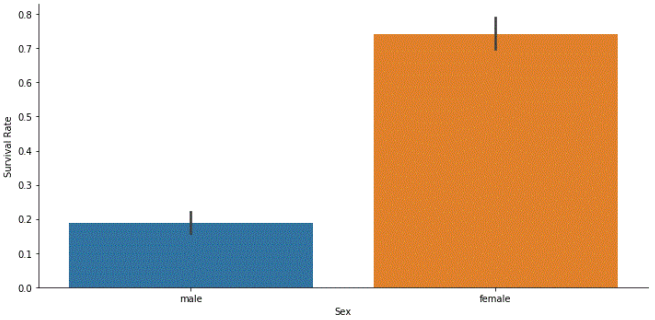

Fig: Survival Based on Sex

Similarly, each of the attribute are explored to extract those attribues or features which would be used later for ediction. A survival barplot is generated to determine the number of people survived vs. number of people who can not survive. From the barplot it is clear that the number of people who survived is less than the number of people who could not survive.

```
Survived
0    549
1    342
Name: PassengerId, dtype: int64
0= did not survive
1= survive
```

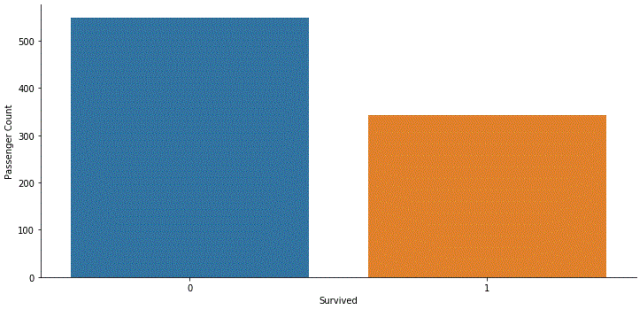

Fig: Survival of Passengers

Similarly, a graph is plotted for survival of passengers on the basis of class.

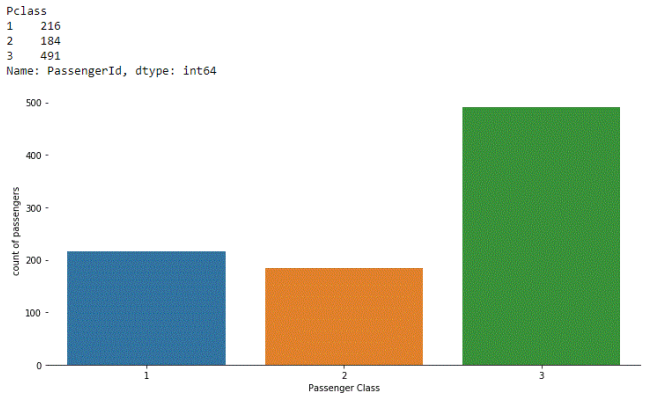

Fig: Distribution Based On Class

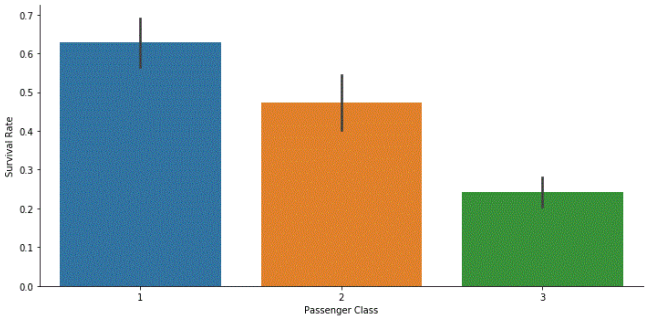

Fig: Survival Based On Class

A graph is plotted to keep a count of passengers based on embarkment

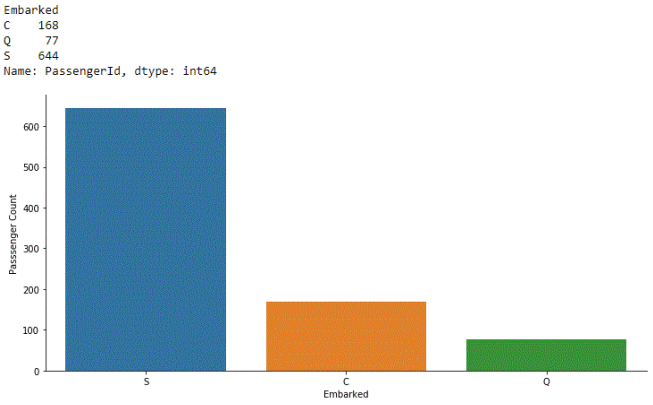

Fig: Distribution Based Embarkment

V. CONCLUSION & FUTURE SCOPE

Logistic Regression proved to be the best algorithm for the Titanic classification problem since the accuracy of Logistic Regression is the highest and the false discovery rate is the lowest as compared to all other implemented algorithms. The research also determined the features that are the most significant for the prediction. Logistic regression as well as Random forest suggested that Pclass, sex, age, children and SibSp are the features that are correlated to the survival of the passengers. We also used Python Libraries like PySpark and Pandas to read and analyse the data. Also Matplotlib and Seaborn were used for data visualization and plotting of the graphs.

Future work might include potentially validating more using pruning techniques that is to see if a shallower tree with same or improved accuracy can be achieved. Cross validation could also be used that is calculating accuracy based on different combinations of training and test data. It would be interesting to play more with dataset and introducing more attributes which might lead to good results. Various other machine learning techniques like SVM, K-NN classification can be used to solve the problem.

VI. ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Principal, Dr. Sanjay U. Bokade, and Dr. Satish. Y. Ket, H.O.D of Computer Department of Rajiv Gandhi Institute of Technology for providing us an opportunity to do our project work on “Titanic Data Analysis”.

This project bears an imprint of many people. We sincerely thank our project guide Ms. Anita Lahane for her guidance and encouragement in carrying out this project work. Finally, we would like to thank our colleagues and friends who helped us in completing the Project & Paper work successfully.

VII. REFERENCES

I. Aakriti Singh, Shipra Saraswat, Neetu Faujdar, Analyzing Titanic Disaster using Machine Learning Algorithms, International Conference on Computing, Communication and Automation (ICCCA2017)

II. Vaishnav Kshirsagar, Nahush Phalke, Titanic Survival Analysis using Logistic Regression, International Research Journal of Engineering and Technology(IRJET 2019)

III. Dr.Prabha Shreeraj Nair, Analyzing Titanic Disaster using Machine Learning Algorithms, International Journal of Trend in Scientific Research and Development (IJTSRD)

IV. https://www.researchgate.net/publication/330909610_Machine_Learning_from_Disaster_Predicting_the_Titanic_Survival_Rate_a_Random_Forest_approach