Titanic survival Analysis

Nutheti Nikhila Priya
Computer Science & Engineering
PES University
Bangalore, India
nikhila.nutheti@gmail.com

Divya Gannanmaneni
Computer Science & Engineering
PES University
Bangalore, India
g13divya@gmail.com

Yasaswini Madineni
Computer Science & Engineering
PES University
Bangalore, India
yasawinim17@gmail.com

***Abstract*— The sinking of the RMS Titanic caused the death of thousands of passengers. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. With more than 2200 passengers on board, nearly half of them died after the unprecedented mishap. This research is aimed at achieving an exploratory data analysis and understanding the effect or parameters key to the survival of a person had they been on the ship. The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive. After this the results of applying machine learning models are compared and analyzed on the basis of accuracy.The survival prediction has been done by applying various algorithms like Logistic Regression and Support vector machines.**

Keywords—data analytics, titanic, prediction, logistic regression, kaggle dataset

I. Introduction

The Titanic disaster is one of the most famous shipwrecks in world history. Titanic is a British cruise liner that sank in the North Atlantic Ocean a few hours after colliding with an iceberg .Many categories of people of all ages and genders where he was on that fateful night, but luckily that was only a few lifeboats helped. The dead include a large number of men who were given their place by the majority women and children riding.

The inevitable development of technology has both facilitated our life and brought some difficulties with it. One of the benefits brought by the technology is that a wide range of data can be obtained easily when requested. However, it is not always possible to acquire the right information. Raw data that is easily accessed from internet sources alone does not make sense and it should be processed to serve an information retrieval system. In this regard, machine learning algorithms play an important role in this process.

The aim of this study is to get as reliable results as possible from the raw and missing data by using machine learning. Therefore one of the most popular datasets in data science, Titanic is used. This dataset records various features of passengers on the Titanic, including who survived and who didn't. It is realized that some missing and uncorrelated features decreased the performance of

prediction. For a detailed data analysis, the effect of the features has been investigated. Thus some new features are added to the dataset and some existing features are removed from the dataset.Over the years, data of survived as well as deceased passengers has been collected. This dataset has been studied and analyzed using various machine learning algorithms like logistic regression and decision tree.

Machine learning algorithms are used to make predictions on passengers who survived drowning The Titanic. Factors such as ticket boarding, age, gender, category will be used to make predictions. Predictive analysis is the process that including the use of calculation methods to find patterns are important and useful for large data.Using the machine learning algorithms, survival is predicted on different combinations of features.

The paper has been organised as per the following sections: II. Related Works/Literature Review, III Solution Approach; Dataset description, preprocessing, descriptive analysis, and proposed solution approach.

II. Related Works/Literature Review

The observations made by this analysis is that females and children were given more priority than men and the loss of lives was mostly caused due to shortage of lifeboats. In this research paper[1] Logistic Regression has been used for the classification. Before data classification, data pre- processing is done by replacing the missing values with the average of that column and transforming all the raw data into an understandable format.

Analysis of accuracy of the model and performance of the algorithm is done by the confusion matrix. Two attributes are used at a time for the confusion matrix plotting. The values shown in the confusion matrix are the probability of survival of the individual considering only those parameters. Based on the confusion matrix, the

logistic regression gave the accuracy of 95%. That's a good percentage. So, it's good to use logistic regression for classification of data. Finally it was concluded that it works better with the binary dependent variables. Here, ROC curve is also used to depict the performance of different algorithms and helps to decide which algorithm is best with their dataset.

In this research paper[2] the objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive. After applying they are compared and analysed on the basis of accuracy. At first the data cleaning was done by replacing the missing values with a random sample in the age column and using the median method to find out missing values in the fare column. Also performed exploratory data analysis to figure out the features which would influence the survival rate and use those features for analysis and exclude the remaining. Finding the relationship between each attribute and the survival rate. Feature engineering is also done to this dataset which helps in selecting correct features which are used in making predictions. Bad feature selection may lead to less accurate or poor predictive models.

Initially, it is realized that some missing and uncorrelated features decreased the performance of prediction.They applied multiple logistic regression and logistic regression to check whether a passenger is survived.Here they have used 14 different machine learning techniques, including Logistic Regression (LR), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Support Vector Machines, Decision Tree, Bagging, AdaBoost, Extra Trees, Random Forest (RF), Gradient Boosting (GB), Calibrated GB, Artificial Neural Networks (ANN), Voting (GB, ANN, kNN) and, Voting (GB, RF, NB, LR, kNN).

The reported performance metrics across different cases comparison and concluded that, the maximum accuracy obtained from Multiple Linear Regression is 78.426%; the maximum accuracy obtained from Logistic Regression is 80.756%, then compared the results of Decision tree and Random Forests algorithms for Titanic dataset. Decision tree is resulted 0.84% correctly classified instances, while Random Forests resulted 0.81%, and then compared SVM implementation to 16 classification algorithms and for titanic dataset they achieved 20.81% and 21.27% error rates with neural networks and SVM respectively as minimum errors, then compared Adaboost classifiers to SVM and RBF classifiers. For the titanic dataset, 22.4% error rate is obtained from SVM as the minimum error rate. Also used SVM as a component classifier for Adaboost. They used a titanic dataset as one of the experimental data and the minimum error rate they obtained is 21.8%.

For comparing these 14 algorithms accuracy, F-measure are used. Both the accuracy and F-measure are highest for Voting (GB, ANN, kNN). (Whereas accuracy was same for both Voting and Gradient boosting). So it is concluded that the proposed model can predict the survival of passengers and crew with 0.82 F-measure score with Voting (GB, ANN, kNN).

Here, four machine models are used for the analysis. They are 1) Logistic Regression 2) Decision Tree 3) Random Forest and 4) Support Vector Machine. After analysing the data in each model, the model evaluation is done using Confusion Matrix drawn and the accuracy is calculated for each machine learning model. The Logistic Regression machine model got the highest accuracy among all four. So, it is concluded that this model is best suitable for this dataset. The accuracy of the models may vary when the choice of feature modelling is different. Ideally logistic regression and support vector machines are the models which give a good level of accuracy in case of classification problems.

[3]Before building a model, data exploration is done to determine the factors that can be beneficial while creating the classifier for the prediction. Few plots are plotted to get an overall idea on the relations between the attributes. Here, they have used four machine learning algorithms to predict the model accuracy. They are 1) Logistic Regression 2) Decision Tree 3) Random Forest. All of these algorithms are compared to one another on the basis of accuracy. The observed accuracies of Logistic Regression, Decision Tree and Random Forest are 94.26% ,93.06%, 91.86% respectively.

For comparing the four algorithms two measures are used. Those are accuracy and false discovery rate. The observed false discovery rates of Logistic Regression, Decision Tree and Random Forest are 7.90%, 9.26%, 10.66% respectively. To consider an algorithm as the best, the accuracy score should be high and false discovery rate should be low. By these observations, it has been proven that Logistic Regression is the best algorithm with high accuracy and low false discovery rate.

III. Solution Approach

We performed the data analysis and visualisation on our dataset titanic_train.csv to understand and analyse the relationships that were existing among the attributes.

A. Dataset Description

The dataset used in this analysis is titanic_train.csv from Kaggle. The dataset contained the information about the passengers on the ship including name, gender, ticket, age, passenger class etc..along with the target variable 'survived', which tells if the particular passenger has survived or not.

B. Pre- processing and Descriptive Analysis

The dataset contained 7 numerical attributes and

5 categorical attributes.We checked for the missing data across the entire dataset and found that two attributes- ‘age’ and ‘cabin’ had many NaN values.

Roughly 20 percent of the Age data is missing. The proportion of Age missing is likely small enough for reasonable replacement with some form of imputation. Looking at the Cabin column, it looks like we are just missing too much of that data to do something useful with at a basic level.So we decide to drop the ‘Cabin’ column.

For ‘age’ we decide to check for the mean values for all the passenger classes and fill the missing data with the mean of the passenger classes accordingly. Because the age of the passengers follows normal distribution and many of them belonged to the age group 17-34.

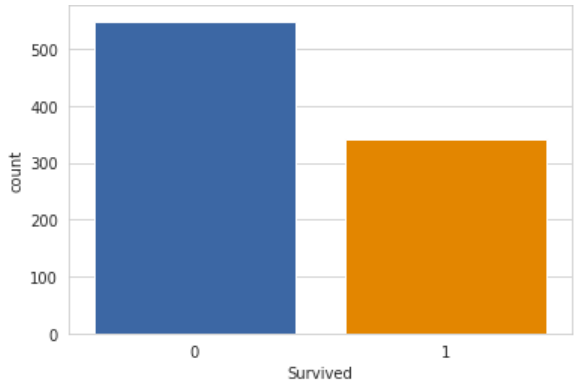

Fig. 1.countplot showing the count of survived and non survived

Survivors are less when compared to the survival ratio. This Tells many of them that they could not survive the Titanic.

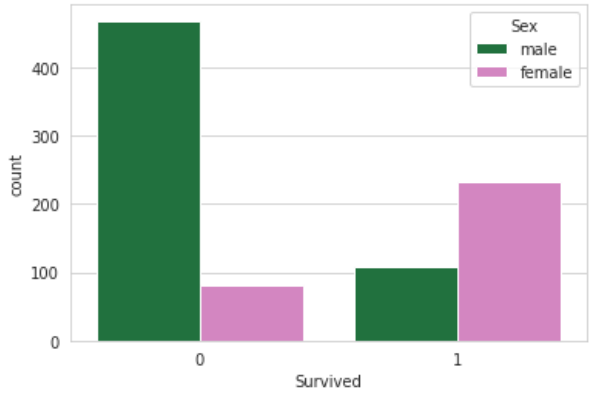

Fig. 2.countplot showing the deaths with hue varibale ‘sex’

Here we can obtain the insight that many of the non-survivors are males. Amongst the survived females are more than the males.Most of men could not survive, this can be due to the reason that male were trying to save females and children and lost lives.

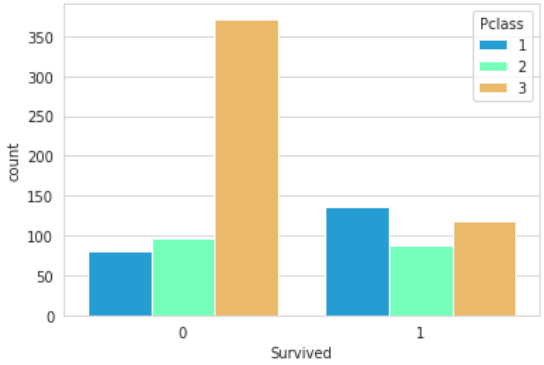

Fig. 3.countplot showing the deaths with hue variable as ‘Pclass’

The passengers from the passenger class 3 are more in number when compared to passenger class 1 and 2. It tells that many passengers on the ship belonged to poor and below middle class families. The death rate of the passenger class 3 people is enormously higher than 2 and 1.Most of the class 3 passengers could not survive and rich people from class 1 somehow managed to survive using their power or authority in getting the lifeboats. Class 3 passengers were probably given least priority to get life boats and hence many deaths.

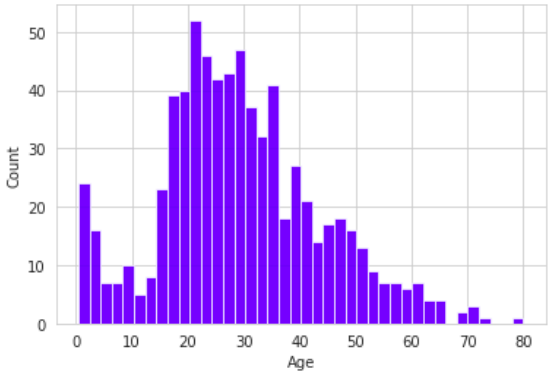

Fig. 4.histplot showing the normal distribution of age

It is clear that age follows normal distribution where most of the passengers belonged to the 17-34 age group.Most of the passengers were young.There were more children when compared to the older passengers.

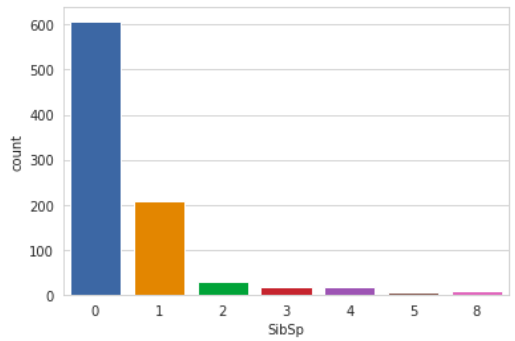

Fig. 5.countplot of siblings and spouse

This plot tells us many passengers did not travel with their families. SibSp gives the total count of siblings and spouses. Most of them had no spouse or siblings.Few of them had spouses and very few had siblings.

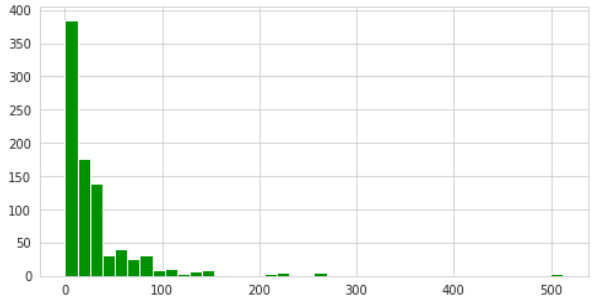

Fig. 6.histplot of fare of all the passengers

It is histplot of fare, as many passengers were from class 3 the fare is low. Hence the count of the fare is higher as many bought class 3 tickets.Also there is a passenger who had a fare of around 500 which seems to be an outlier.

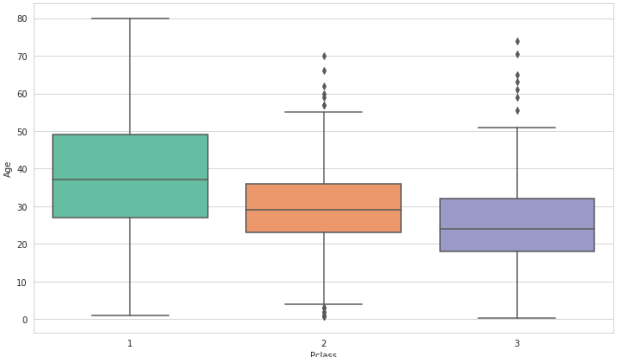

Fig. 7.boxplot of all three passenger classes

We want to fill in missing age data instead of just dropping the missing age data rows. One way to do this is by filling in the mean age of all the passengers (imputation). However we can be smarter about this and check the average age by passenger class.We used a function to replace the NaN values.For passenger class 1 age=37, class 2=29 and class 3=24.

Here in class 1 passengers most of the people belonged to all the age groups, wherein class 2 had age groups below 57 and class 3 passengers were younger with few older people.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	S
0	1	0	3	22.0	1	0	7.2500	1	0	1
1	2	1	1	38.0	1	0	71.2833	0	0	0
2	3	1	3	26.0	0	0	7.9250	0	0	1
3	4	1	1	35.0	1	0	53.1000	0	0	1
4	5	0	3	35.0	0	0	8.0500	1	0	1

Fig. 8.final dataset given to the model for training

We converted the categorical variable like age and embark into dummy variables so that it gets easier for the model to analyse.We also dropped the attributes 'Sex', 'Embarked', 'Name', 'Ticket' and appended the dummy variables S, Q and male. So this is our final dataset with the target variable as ‘survived’.

C. Plan for the next phase

We plan to perform a few more Visualization to further understand how to enhance the dataset to be more amenable for the model such that it can predict with even higher accuracy. Our solution includes the use of various data visualisation tools to find and retrieve relationships between them. Using these and more features we look at what variables affected the death of a passenger and come up with a prediction model which when given the data will be able to predict if the person survives or not. We plan to use the logistic regression as our prediction model as it gives the highest accuracy.

REFERENCES

[1] https://bit.ly/3vmzViN

[2] https://www.researchgate.net/profile/Yogesh-Kakde/publication/325228831_Predicting_Survival_on_Titanic_by_Applying_Exploratory_Data_Analytics_and_Machine_Learning_Techniques/links/5c068f63a6fdcc315f9c0bb9/Predicting-Survival-on-Titanic-by-Applying-Exploratory-Data-Analytics-and-Machine-Learning-Techniques.pdf

[3]]http://sajrest.com/Archives/vol4issue4_2019/v4i4p2.pdf

[4] https://www.researchgate.net/profile/Neytullah-Acun/publication/324909545_A_Comparative_Study_on_Machine_Learning_Techniques_Using_Titanic_Dataset/links/607533bc299bf1f56d51db20/A-Comparative-Study-on-Machine-Learning-Techniques-Using-Titanic-Dataset.pdf