Predicting Survival on Titanic Disaster

Department of Computer Science and Technology
School of Engineering and Applied Science
Bennett University, Greater Noida
Uttar Pradesh (India), 201310

Mentors:
Ms. Shambhavi Mishra
Ms. Surbhi

Submitted By:
Akhil Chandail
Namitha Mariyam George
Shibin Varghese
Yash Chaudary

Predicting Survival on Titanic Disaster

Akhil Chandail, Yash Chaudhary, Namitha George, Shibin Varghese
Ms.Shambhavi, Ms.Surbhi

ABSTRACT—

In this project, we attempt prediction of survival of the passengers in Titanic disaster using different machine-learning techniques and feature engineering techniques. The dataset comprises of 891 passengers having features like age, gender, cabin, ticket fare, etc. and we try to predict the survival of 418 individuals. The first objective is to find unknown information and to fill empty fields of various factors like age, cabin and fare by preprocessing of available dataset by feature engineering techniques. And to find the correlations between various features of passengers which has impact on individual's survival. And after preprocessing of dataset, we apply different machine learning algorithms to predict whether a passenger was survived or not. After obtained results, machine learning/deep learning algorithms are compared and analyzed based on their accuracy. Here, we propose to apply five machine learning models including Dense net model, SVM, Random Forests, Logistic Regression and Decision Tree.

I. INTRODUCTION

The Titanic disaster resulted in death of many crews and passengers in the North Atlantic on April 15, 1912 due to sinking of ship [1]. Even after many years, the researchers are attracted by the research on understanding what impacts individual's survival or death. When we observe this problem we find some people have more chance of surviving than others. With the available datasets from kaggle, we take a look at the classification of passengers which have a relationship on people who survived. We create correlation between various features of passengers like age, sex, cabin, etc. which would had impact on passenger's survival. On the preprocessed dataset we apply different machine/deep learning models for the prediction of the passenger's survival.

The features of each passenger like family size, gender, age, class, embarkation are used to predict his/her survival. We use various feature engineering techniques to compare the machine learning algorithms. The preprocessed dataset is then used for comparing and analyzing the various factors that are impacting the survival rate.

II. RELATED WORK

In the past, many researchers have worked on the Titanic disaster problem to contrast and compare different machine learning algorithms. They analysed algorithms in terms of the efficiency of the machine learning model. Researchers have attempted to trade off between different features of the available dataset to obtain the accurate results of prediction. Lam and Tang has solved the Titanic problem to compare and contrast between three machine-learning algorithms which are Naive Bayes, Decision tree and Support Vector Machine [2]. According to Shawn Cicoria and John Sherlock, gender of passenger is the most significant feature in correctly predicting the survival of passengers as compared to other features available in dataset. They solved this problem using Decision tree classification and Cluster analysis [3]. Kunal Vyas and Lin suggest that dimensionality reduction and preprocessing the dataset could improve the accuracy of the algorithms [4].Many researchers concluded that more features utilized in the models do not necessarily make results better.

Although many researchers have worked hard to determine the survival of passengers, we also attempt to get better results using various combination of features and different machine learning methods.

III. METHODOLOGY

Machine/Deep Learning Models-

We implemented different machine/deep learning algorithms for prediction of the survival of passengers. The different algorithms which are applied on the Titanic dataset for preprocessing of raw data and prediction of survival are Random Forest, SVM, Dense Model, Decision Trees and Logistic Regression. These algorithms are described in a short notes as below:

A. Random Forest:

This algorithm is supervised classification algorithm. The various classification problems and regression problems can be solve by Random Forest. This algorithm is about making a forest using dataset with large number of tress that can help in binary classification [5]. The large number of trees can help in increasing the accuracy of the specific problem. For example, if it takes 200 samples randomly from the dataset, this algorithm will choose randomly about 10 initial variables which are uncorrelated with each other, so to predict classification accurately. After the classification and regression trees, the final result of prediction is obtained by calculating the mean of each prediction.

B. Support vector machine:

SVM is a supervised machine learning algorithm, which is also used to solve regression and classification problems. This algorithm classifies the dataset by plotting each data point or feature in a n-dimensional space, where 'n' is number of features available in dataset. Then it constructs a hyper plane which classifies the different class labels. The categorical dataset is classified into values as either 1 or 0.

C. Decision Trees:

This algorithm classifies the data as like random forest. This algorithm also arranges the data variables in a tree , from root node to the terminal node. And each data point at each instance in a tree is tested by the attribute specified by that particular node. For each sub-tree which is rooted at the new node, this same process gets repeated. In this tree-like graph, each node represents a point of decision.

D. DenseNet Model:

DenseNet is a convolution network. This model comprises of many dense blocks which consists of many dense layers. Each dense layer gets the input from previous dense layers and passes its own output feature maps to next subsequent dense layers. Each dense block receives the collective information from all previous blocks. The advantage of using this algorithm is that it reduces the vanishing gradient problem.

E. Logistic Regression:

Logistic Regression is a machine learning algorithm which measures the relations between categorical dependent feature and other independent features available in the dataset. It uses a logistic function i.e. cumulative logistic distribution for calculating probabilities.

IV. EXPERIMENTS

A. Dataset:

We get the whole dataset for our Titanic problem from the website of Kaggle [6]. The training dataset is in the form of CSV file, which consist of 891 samples of passengers and 12 columns that are describing the multiple features. The test data is the same format which consists of about 418 samples of passengers and having same number of columns as training dataset has. From dataset, it is seen that there are some empty fields of various feature columns especially age, embarkation, cabin. And we figured out that we can extract some of hidden features like family size, Title, etc from the dataset. To fill the missing fields, some of the features are grouped to find the mean of that specific group and then empty fields are filled with median values. The structure of the dataset has shown below:

TABLE 1. NUMBER OF FEATURES IN THE DATASET

Feature	Value of Feature	Characteristic of feature
Passenger Id	1-891	Integer
Survived	0 or 1	Integer
Pclass	1-3	Integer
Name	Passenger's name	String
Gender	Male or Female	String
Age	0-80	Real
SibSp	0-8	Integer
Parch	0-6	Integer
Ticket	Ticket	String
Fare	0-512	Real
Cabin	Number of Cabin	String
Embarked	S, C, Q	String

B. Feature Engineering:

The features that are used for training and making prediction are selected using feature engineering. All the unused or redundant features are filtered out. Then all these features are converted to numerical equivalent values.

Based on the analysis, the following features like sex, age, title, Pclass, cabin, family size (parch plus sibsp columns), embarked, fare. The survival column is chosen as response column for the dataset. These features are selected because their values have effect on survival rate.

The missing values of Age and Fare features are filled by median values of these features. The "Embarked" values are replaced by "C". The PassengerId, Name and Ticket features are removed from the feature set as there may not be a correlation between them and survival.

i. Sex: When we consider the distribution of the "Sex" feature, there are 314 female and 577 male passengers. 233 of female passengers have been rescued and others have lost their lives. On the other hand, 109 of male passengers have been rescued and others have lost their lives. From the analysis, we realize that the rate of survival of women is higher than men. It has been concluded that the effect of this feature on predicting the class label is significant.

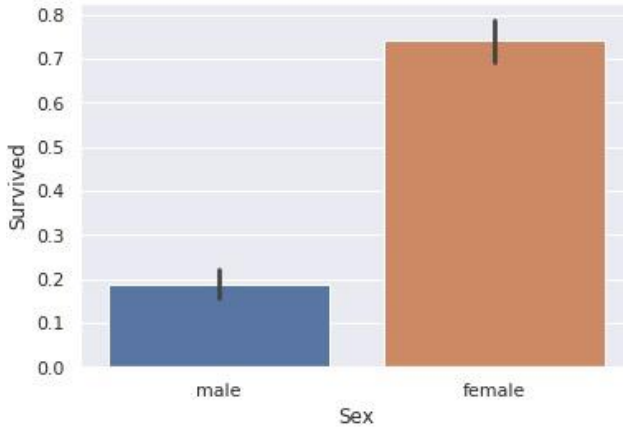

Fig 1. Comparison of male and female survivors

ii. Embarked: When we consider the distribution of the "Embarked" feature, there are 644, 168, 77 passengers boarding from the port "S", "C" and "Q" on the ship respectively. The survival rate of passengers boarding from port "C" has the highest.

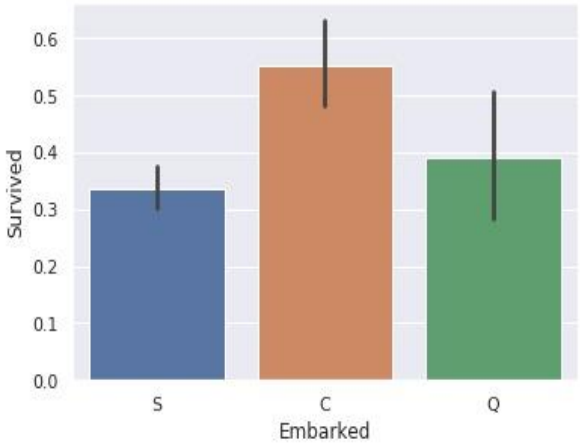

Fig 2. Comparison of survivors from different ports.

iii. Pclass: "Pclass" feature describes three different classes of passengers. There are 216 passengers belong to the class 1, 184 passengers in class2 and finally 491 passengers in class 3. The passengers with the highest

survival rates are the first class passengers. This ratio also shows that wealthy people are alive.

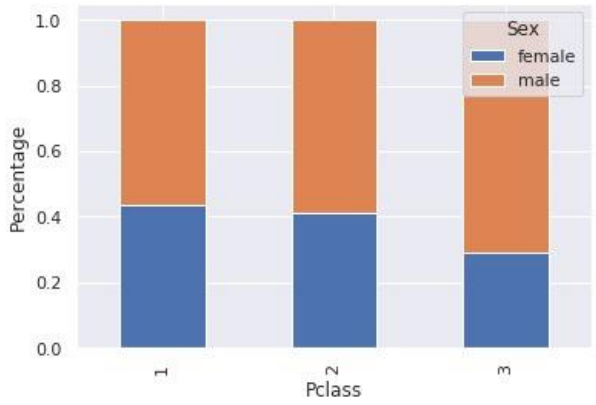

Fig 3. Comparison of survivors from different Passenger Class w.r.t Gender

iv. Age: We created new feature for Age bands, which will turn the continuous numerical values into an ordinary categorical values. The range of passenger's age is from 0 to 80. So, we group the passengers by specific age ranges such as 0-15, 15-25, 25-35, 35-60 and 60- above, then we realized that most of the passengers in the 0-15 age group are survived and a large majority of passengers in the age group 60-80 lost their lives. And we also determined the missing age of passengers by applying KNN algorithm on the 'survival', 'Family size', 'Sibsp' and 'Parch' features. This statistical information shows that the first children were rescued from sinking of Titanic.

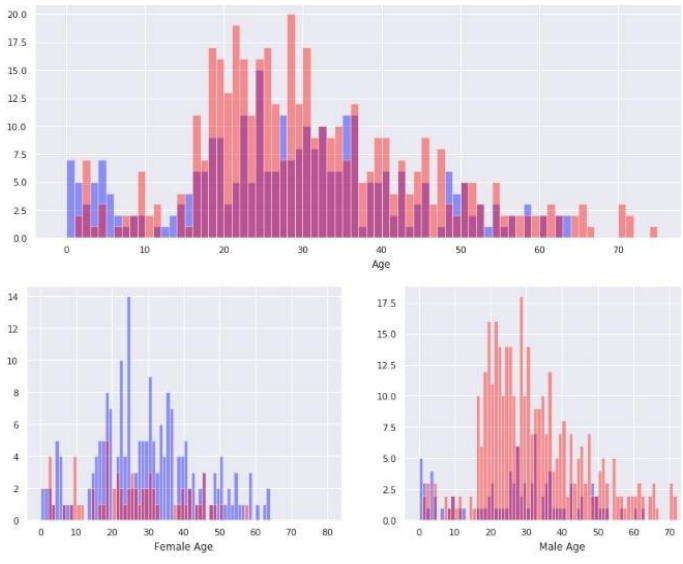

Fig 4. Comparison of survivors from different age groups

v. Fare: The "fare" feature specifies the fare paid by the passenger and it changes between 0-512. If we distinguish this feature with groups as 0-20, 20-30, 30-100 and above 100, then it is seen that majority of the passengers who paid between 100 and above survived.

vi. Family size: "Family size" is a new feature that is created which specifies the total number of family members on board based on SibSp and ParCh. We added the total number of siblings, spouse, children and parents of peassengers to get the total size of his/her family. After that, we have distinguished this feature with several groups. This results show that the number of family members strengthens the possibility of survival.

vii. Title: We created a feature named 'Title' which also specifies the age by extracting the title from name of passengers. As we assume that passengers having same title are having the same age or in the same age group. So, we replaced the missing field in age feature by the mean(average) of that particular group of title. For example, if we assume that Mrs X is elder than Ms X and if there is a missing age for a woman with the title specifying Mrs, then it is filled the age by the average of the all women having title with 'Mrs'. And we also assume that passengers having titles like 'Dr', 'Col', etc. are having the more chance of survival as it shows that these passengers are important and respectable to society.

C. Experimental Results:

By preprocessing the titanic dataset, we got a new dataset containing 8 feature columns. When applying algorithms to this Titanic dataset, we have seen that to make the algorithm accurate, some more adjustments on some model parameters are required.

Algorithms are evaluated according to accuracy and Kaggle score. We compare our accuracy scores with accuracy scores obtained from Kaggle.

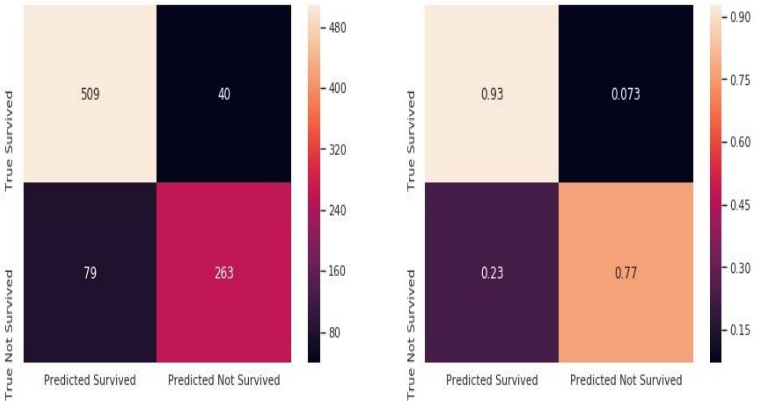

Fig 5.Confusion Matrix

TABLE 2. COMPARISON OF ACCURACY AND KAGGLE SCORES OF ALGORITHMS

Algorithm	Accuracy	Kaggle Benchmark
Random Forest	0.86	0.81
Support Vector Machine	0.83	0.78
Decision Tree	0.86	0.79
Logistic Regression	0.84	0.79
DenseNet	0.85	0.80

V. CONCLUSIONS AND FUTURE WORK

We obtained valuable results from the raw and missing dataset of titanic by using machine learning and feature engineering algorithms. We tried to attempt various different models which can predict the survival of passengers and it is seen that, there were not significant differences in accuracy between the algorithms we implemented on Titanic problem. The highest Kaggle score (0.81) is obtained with Random Forest. As a conclusion, this presents a comparative study on different machine learning models to analyze Titanic dataset and to know what features effect the results of prediction.

This project work can be used as reference to learn implementation of machine-learning models from very basic. In future, the idea can be extended by making advanced graphical user interface with the help of newer libraries. An interactive page can be made. We can also draw much better conclusions by combining the results we obtained. It would be interesting to continue this analysis with other possible features or with other machine learning algorithms like Naive Bayes, AdaBoost, XGBoost.

REFERENCES

[1] RMS Titanic – From Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/RMS_Titanic

[2] Eric Lam, Chongxuan Tang. 'Titanic–Machine Learning From Disaster'. Available FTP:cs229.stanford.edu

[3] Cicoria S., Sherlock J., Muniswamaiah M. and Clarke L., 'Classification of Titanic Passenger Data and Chances of Surviving the Disaster', 2014

[4] Vyas KE., Zheng Z. and Lil, 'Titanic-Machine Learning From Disaster', 2015.

[5] L. Breiman, 'Random Forests', *Machine Learning,* vol. 45, 2001.

[6] Kaggle, Titanic: Machine Learning form Disaster [Online-2019]. Available: http://www.kaggle.com/