International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Prediction of Survivors in the Titanic Cruise

Rajesh M

Abstract*: *On the 15th of April, 1912 the titanic witnessed a disaster resulting in the sinking of her passengers on the maiden voyage near North Atlantic. Even though it is a very long time since this maritime disaster took place, the idea behind what impacts each individual survival is still a great research attracting researcher's attention. The approach taken in this paper is to utilize the publically available data set from website called Kaggle. Kaggle is a popular data science webpage that put together information of people in the titanic into a data set for the data mining competition: "Titanic: Machine Learning from Disaster". The research and comparisons in this paper uses a few machine learning techniques and algorithms to analyse the data for classification and prediction of survivors. The prediction and efficiency of these algorithms depend greatly on data analysis and model. The techniques used to do so are Random Forest, Support Vector Machine, Gradient Boosting Machine.

Keywords*: *Machine Learning, Data Mining, Random Forest algorithm, Support Vector Machine, Gradient Boosting Machine.

I. INTRODUCTION

The development in technology has both benefitted our lifestyles and has also got us into a few problems from time to time. One of the advantages added by the technology is that a broad category of data can easily be requested. However, it is not that easily feasible to accumulate the proper statistics. Raw and plain records that is effortlessly accessed from the internet sources do not make sense and it has to be processed to serve a data retrieval system. On this regard, feature extraction methods and system machine learning algorithms are performing an essential function in this method.

The focus of their study at is to get as dependable results as efficiently as possible from the raw missing records by means of using machine learning algorithms and feature extraction techniques. Hence one of the most famous datasets in data mining, Titanic is used. This dataset represents diverse functions of passengers on the titanic, which includes who survived and who didn't. It's far realized that some missing and uncorrelated functions decreased the performance of prediction. For a more detailed records evaluation, the effect of various features has been investigated. Accordingly, some new functions are introduced to the dataset and some existing capabilities are eliminated from the dataset.

Manuscript published on 30 September 2019
* Correspondence Author
Rajesh M*, Computer Science and Engineering, Vellore Institute of Technology, Vellore, India. Email: rajesh.marudhachalam@gmail.com

Ekin Ekinci [1] has applied various machine learning algorithms and techniques to predict the survivors, which include, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, k-Nearest Neighbours, Support Vector Machine, Gradient Boosting, Artificial Neural Networks, Bagging, AdaBoost and Calibrated Gradient Boosting. He also obtained the F-measure score for each of the machine learning techniques and compared them with each other.

Chatterjee [2] implemented more than one logistic regression and logistic regression to test whether a passenger is survived. He said overall performance metrics throughout specific instances contrast and concluded that, the maximum accuracy obtained from multiple linear regression is 78.426%; the most accuracy is obtained from logistic regression is 80.756%.

Datla [3] compared the effects of choice tree and random forests algorithms for large dataset. Decision tree resulted 0.84% efficiently categorized times, while random forests resulted 0.81%. because the function engineering steps, they created new variables including "survived", "baby", "new_fare", "identify", "familysize", "familyidentity" which aren't covered in characteristic listing of massive dataset and additionally changed a missing value by means of the imply price of a given function. Li et al. [4] have used support vector machine as the main component classifier for Adaboost, they have used titanic dataset as the experimental data and obtained a minimum error of 21.8%.

The upcoming part of the paper is organized as follows:

Section 2 presents the methods and techniques that were employed,

Section 3 gives the overview of dataset, experimental setup and results, and

Section 4 concludes the paper.

II. METHODOLOGY

A. Random Forest

This is a classification algorithm that was developed by Breiman and Cutler. It is an algorithm that falls under the category of supervised learning which is flexible, easy for usage in machine learning. It is widely used algorithm because of its simplicity alongside the usability with both regression and classification. The forest that this algorithm builds is an ensemble of Decision Tress which is mostly trained using a bagged method. The concept behind the usage of a bagged method is that combining multiple algorithms improves is that the learning increases resulting in an overall performance increase. Random forest adds some more randomness to out model while in the growing stage.

Unlike other models this model searches for the best feature in the random subset of features other than the important features. The big concern in using this algorithm is how it tends to slow down and eb ineffective when there are large number of trees.

B. Support Vector Machine

SVM is a supervised data mining algorithm and machine learning technique that may be employed for both type and regression functions. Support vector machines are greater typically utilized in type problems and as such.

SVM's are based totally at the concept of locating a hyperplane that high-quality divides a dataset into classes. This algorithm preforms classification by finding hyperplanes that maximizes the margin between the two classes. The cases or vectors defining the hyperplanes are the support vectors. The main advantage of support vector machine is the linear data inseparability. Support vector machine maintains a zero slack variable while still maximizing the margin. The perpendicular distance from line to the only closest points is called margin. These points define the line and construction of classifier defining the hyperlane.

C. Gradient Boosting Machine

Boosting is a technique of changing susceptible newbies into robust rookies. In boosting, every new tree is fit onto the modified model of the unique facts set. the gradient boosting algorithm is always used along with the adaboost algorithm. The adaboost algorithm starts off evolved with the aid of a decision tree in which every observation is assigned an identical weight. after comparing the primary tree, we increase the weights of these observations which are tough to categorise and lower the weights for the ones which can be clean to categorise. the second tree is therefore grown in this weighted statistic. The idea is to enhance upon the predictions of the primary tree. Therefore, the new model obtained is a combination of tree 1 and tree 2. Then on we compute the class error from this newly obtained 2-tree ensemble model and develop a new third tree in order to predict the revised residuals. we repeat this procedure for a specific variety of iterations. subsequent bushes help us to classify observations that aren't properly categorised by way of the preceding timber. Predictions of the ensemble final model is hence the sum of the weighted predictions made through the previous tree fashions.

III. EXPERIMENTS

A. Dataset

Titanic dataset was provided by Kaggle. This is divided into two parts, first, the training dataset consisting of 891 passengers, and second, test dataset consisting of 418 passengers.

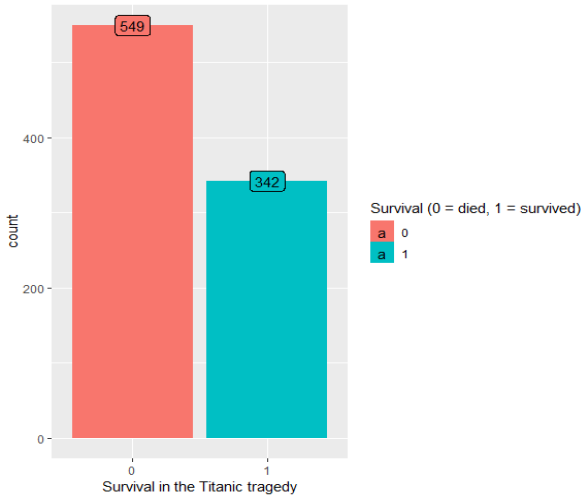

Figure 1: Histogram of survivors

Figure 1. is a histogram plot that gives us an idea about the number of passengers who survives that disaster were more than the deceased. The survival rate was 38.4%. A general overview of the dataset and description of features are given in the below tables.

Table 1: Overview of Titanic Dataset

File Name	Description	Significance
Train.csv	Contains 12 unique features Survived feature exists 891 rows each for one passenger	Ground truth is survived feature
Test.csv	Contains 12 unique features Survived feature does not exist 418 rows referring to 418 passengers	No ground truth
Output.csv	2 columns PassengerID and Survived 419 rows	Used to evaluate performance of algorithm used

Table 2: Features in dataset

Feature	Value	Characteristic	Definition
PassengerID	1-891	Integer	Unique ID given to each passenger
Survived	0,1	Integer	Survival
Pclass	1-3	Integer	Ticket class
Name	Name of passenger	Object	Name of passenger
Sex	Male, female	Object	Sex
Age	0-80	Real	Age in years
SibSp	0-8	Integer	Number of spouses or siblings aboard
Parch	0-6	Integer	No of children or parents aboard
Ticket	Ticket number	Object	Ticket number
Fare	0-512	Real	Passenger fare
Cabin	Cabin number	Object	Cabin number
Embarked	S, C, Q	Object	Port of Embarkation

PassengerID: It is a numeric feature of the dataset representing the number of passengers aboard. Since it does not make to include this in our predictive model, it is eliminated.

Survived and Pclass: These are a numeric feature 0 or 1 for Survived and 1 – 3 for Pclass, and these features are changed from numeric to categorical as it will not make sense to average numeric value for these columns.

Name and Sex: these are relevant strings that might be used for some text analysis to do some reputable extraction. Hence, converted to categorical from string.

Age: this numeric data might actually be highly relevant, so we categorize the passengers into three broad categories of children, adults and elderly. And this will be added to our predictive model. The problem with the missing values can be resolved in 2 different ways:

- Using average of the median of ages
- Using logical equation from general age difference between mother and child

SibSp and Parch: These features are used to represent family; they are numeric features as they represent the corresponding in number. These when consolidated together can be a relevant feature.

Ticket: this is a string feature containing lots of uncertainty and irrelevance so this is eliminated.

Cabin: this string feature tells us where the cabins are booked. This could have been an interesting feature had there been lesser missing values.

Embarked: this string represents where the ship has embarked before heading into Atlantic. This is converted to categorical and the missing two values are estimated to be Southampton

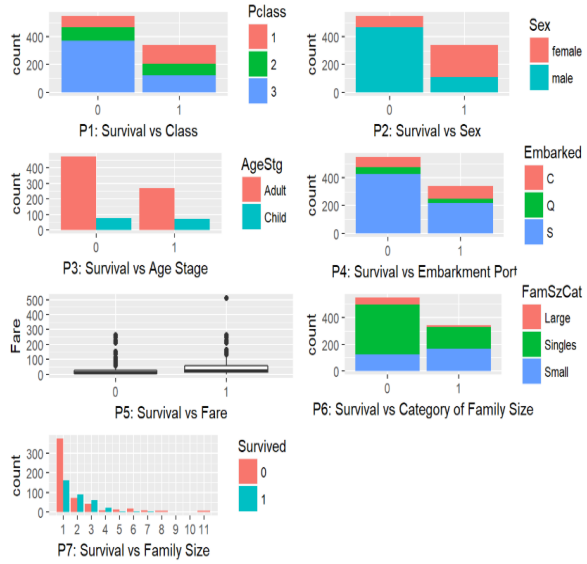

Figure 2: Plot to study features related to survival

Figure 2. Observations from the plots:

P1 = the deceased passengers mainly belonged to the lower class, while the distribution was even among the survived.

P2 = Males were the majority among those of the deceased and females were among the majority of the survivors. This was also the belief that females and children were given more priority.

P3 = Adults in the higher class were given more priority, whereas as a whole the number of children who survives were higher.

P4 = Distribution of survivors and deceased among various ports is almost similar making it difficult for prediction.

P5 = those with higher ticket fare made it out alive than the ones with the lower fare

P6 = Even though, titanic was filled with many singletons, the probability of them surviving were poor and small families survived better.

P7 = Families with more than 4 members had a poor rate of survival.

B. Pre-Processing

The pre-processing steps include data cleaning, data integration and data transformation. The PassengerID, Name, Ticket, Cabin features were removed from the feature set and the missing values in the columns of Age and Fare were filled using the median of values of these features, and the missing values in column Embarked were set to C upon categorical casting.

C. Experimental Results

The algorithms mentioned previously in this paper were implemented in order to predict the likelihood of survival of passengers aboard and the correlation between the crew and the passengers. Upon analysing the obtained results, we have learned that to make the algorithms more efficient some more corrections and adjustments to a few parameters are required through the previous tree fashions.

Table 3: Algorithm Performance

Algorithm		Random Forest	Gradient Boosting Machine	Support Vector Machine
Accuracy	Min	0.7014925	0.7014925	0.6818182
	1st quartile	0.7886703	0.7667112	0.7667112
	Median	0.8059701	0.7910448	0.8195389
Kappa	Min	0.3437806	0.3532819	0.3381089
	1st quartile	0.5372884	0.4955821	0.5079540
	median	0.5844765	0.5537583	0.6103324

Comparing the results among the 3 models depicts that they are very close in performance. Hence, the confusion matrix has been used to decide the highest accuracy that can be obtained

Table 4: Accuracy of models

	Accuracy	Error
Random Forest	82.5%	17.5%
Gradient Boosting Machine	77.6%	22.4%
Support Vector Machine	83.9%	16.1%

The above table of accuracy among the models evidently suggests that Support Vector Machine is marginally more accurate for this situation over the other two proposed models

IV. CONCLUSION

In today's knowledge-based world, obtaining results from raw and unprocessed data by using machine learning and feature extraction techniques has become very important. In this paper, there are three proposed models namely, Random Forest, Gradient Boosting Machine and Support Vector Machine for predicting the survival of passenger aboard on the maiden voyage of titanic. Primarily a study was conducted on various features of the dataset on whether they can be used for the process of correlation or if they are irrelevant to predictive model being built. In the pre-processing stage a few features such as Name, Ticket, Cabin etc., have been excluded and the missing values in the other features were filled in. Finally, the machine learning algorithms and classifications algorithms were used in order to predict the survival of crew members and passengers aboard during Titanic's maiden voyage. Hence, this paper gives a comparative study of the three-machine learning algorithm and techniques on the titanic dataset to learn the about features effecting the classification and also tells us the technique of "Support Vector Machine" is more robust and accurate with a highest accuracy of "83.9%".

REFERENCES

1. Ekin Ekinci, "A Comparative Study on Machine learning Techniques Using Titanic Dataset".
2. T. Chatterjee, "Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms," International Journal of Emerging Research in Management &Technology, vol. 6, pp. 1-5, June 2017.
3. M. V. Datla, "Bench Marking of Classification Algorithms: Decision Trees and Random Forests – A Case Study using R," in Proc. I-TACT-15, 2015, pp. 1-7.
4. X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," Engineering Applications of Artificial Intelligence, vol. 21, pp. 785-795, Aug. 2008.
5. Titanic – Machine Learning from Disaster. Eric Lam, Chongxuan Tang .
6. Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, Second Edition (Chapter 6.3.1).
7. Joseph Ottinger. CI-Bayes; https://ci-bayes.dev.java.net/ Open Source
8. Titanic Dataset downloaded from Kaggle, https://www.kaggle.com/pavlofesenko/titanic-extended/downloads/titanic-extended.zip/2

AUTHORS PROFILE

Rajesh M, is currently pursuing Bachelor of Technology in Computer Science and Engineering (4th year) at Vellore Institute of Technology, Vellore. Email Id: rajesh.marudhachalam@gmail.com .He is currently working on projects using concepts such as Natural Language Processing, Visual Cryptography and Statistics. His current domain of interests includes *Data Science, Artificial Intelligence and Machine Learning.*