Analysis and Prediction of the Survival of Titanic Passengers Using Machine Learning

Amer Tabbakh, Jitendra Kumar Rout, and Minakhi Rout

Abstract The Royal Mail Ship (RMS) Titanic is the largest liner built in 1912, of the estimated 2224 passengers and crew aboard, more than 1500 died after the ship struck an iceberg during her maiden voyage from Southampton to New York City. The dataset collected from Kaggle has information about the passengers and crew which are P-class, name, age, sex, etc., and can be used to predict if the person onboard has survived or not. In this paper, six different machine learning algorithms are used (i.e., logistic regression, k-nearest neighbors, SVM, naive Bayes, decision tree and random forest) to study this dataset and deduce useful information to know the knowledge of the reasons for the survival of some travelers and sinking the rest. Finally, the results have been analyzed and compared with and without cross-validation to evaluate the performance of each classifier.

Keywords Titanic dataset · Survival prediction · Machine learning · Cross-validation

1 Introduction

Machine learning is a subset of AI that mainly focuses on learning from past experiences and making predictions based on it. Titanic, one of the largest and most luxurious ocean liners ever built on April 15, 1912, was considered unsinkable at that time. The ship carried around 2200 passengers and crew, but it sunk and 1500 passengers and crew died. The dataset was collected from Kaggle regarding Titanic contains many information about the passengers and crew (such as ID, name, age).

A. Tabbakh · J. K. Rout (✉) · M. Rout
Kalinga Institute of Industrial Technology University, Bhubaneswar, Odisha, India
e-mail: jitu2rout@gmail.com

A. Tabbakh
e-mail: mr.amertabbakh@gmail.com

M. Rout
e-mail: minakhi.rout@gmail.com

A. K. Tripathy et al. (eds.), *Advances in Distributed Computing and Machine Learning*, Lecture Notes in Networks and Systems 127, https://doi.org/10.1007/978-981-15-4218-3_29

The information can be extracted, and machine learning can be used to predict who has survived as well as died during the disaster. In this paper, different machine learning algorithms such as logistic regression, k-NN, SVM and naive Bayes, decision tree and forest have been used for prediction of survival of passengers.

2 Related Work

Lam and Tang [2] used three algorithm, namely SVM, decision tree and naive Bayes for the Titanic problem. The authors have observed that sex was a significant feature in accurately predicting survival. The best result was 79.43% (in terms of accuracy) by the decision tree algorithm. In a similar way, Balakumar et al. [3] have used three algorithms for the Titanic prediction problem(LR, DT, and RF). It has been found that logistic regression was the best to provide an accuracy of 94.26% with selected features such as P-class, sex, age, children and SibSp. Farag and Hassan [4] focused more on feature selection and concluded that sex is highly correlated to survival and also passenger's ticket proved to be strongly correlated to survival as well. Out of the two algorithms (naive Bayes, decision trees) used for the Titanic prediction problem, naive Bayes was the best to provide an accuracy of 92.52%. On a different note, Kakde and Agrawal [5] have focused on the preprocessing of data. They have observed that the age was highly influential on the survival rate. Apart from these, $parch$ and $sibsp$ columns are combined to know the family size where if family size becomes greater than three, the survival rate decreases. Four different algorithms (logistic regression, SVM, random forest, decision trees) were used for the Titanic prediction problem, and logistic regression gave the best accuracy of 83.73%.

Singh et al. [6] have observed that the features (P-class, sex, age, children and SibSp) which are selected as more significant are highly correlated to the survival of the passengers. Out of four different algorithms (naive Bayes, logistic regression, decision tree, random forest) used for the titanic classification problem, logistic regression gave the best result in terms of accuracy as 94.26%. Ekinci and Acun [7] have tried with different feature combinations; i.e., some features are added to dataset (such as family size), and some of them were eliminated like (name, ticket and cabin). Different machine learning techniques were used, and the results were compared on the basis of F-measure on the dataset obtained from Kaggle. The best F-measure was 0.82% for voting(GB, ANN, KNN).

Nair [8] used four ML techniques(LR, NB, DT and Rf) to predict the survival of passengers. The survival rate of the female is observed to be higher than that of the male. Some features are added to dataset (mother, children, family, etc.), and two performance measures are used to compare the results (accuracy and false discovery rate) in which LR proved to be the best for this problem.

3 Methodology

3.1 Contribution

The major contribution of work is:

- Selection of appropriate features for the Titanic dataset and experimentally find out which combination will work better.
- Analysis and prediction of passenger survival using different classifiers.
- Analysis of results with and without cross-validation to avoid the effect of improper distribution of data (if any).

3.2 Classifiers Used

In this work, different machine learning algorithms (i.e., logistic regression, KNN, SVM, naive Bayes) are used to predict whether a passenger will survive or not. Most appropriate attributes selected for training and testing dataset are: P-class, sex, age, SibSp, ParCh, nickname. Here are the brief descriptions of different algorithms used:

- **Logistic Regression**: Logistic regression is used for the classification, and the target variable is categorical as well as a binary where 1 means survival and 0 means demise.
- **Naive Bayes(NB)**: NB applies Bayes theorem to build a prediction model of a classification problem. It assumes that features are independent of each other. It is called the conditional probability; by multiplying all the conditional probabilities, the probability of a class value is obtained, and the class which has the highest probability is assigned as the class of a given instance. Out of four different types of NB algorithms, the Gaussian NB algorithm is used in this work with the significant features (P-class, sex, age, SibSp, ParCh, nickname) to build the model, which are categorical (or numeric). The target value is surviving which takes two values (0 for demise and 1 for survival). Summary data included calculating the standard deviation and mean for each attribute, by class value.
- **KNN**: The KNN is a supervised learning algorithm that makes use of the class labels of training data during the learning phase as a target. It is an instance-based machine learning algorithm, where new data points are classified based on stored, labeled instances (data points). Though it can be used both for classification and regression, it is more widely used for classification purposes. The k(the number of nearest neighbors) in KNN is a crucial variable also known as a hyperparameter that helps in classifying a data point accurately. The value of k taken in our case is five.
- **SVM**: SVM is a supervised machine learning algorithm that is mostly used for classification problems. Support vector machine is a frontier that best segregates the two classes.

Table 1 Snapshot of the dataset before preprocessing

PID	Survived	P-class	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	Male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	Female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	Female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Female	35	1	0	113803	53.1	C123	S

- **Decision Tree**: The decision tree is one of the important classification models; it classifies the problem depending on making a flowchart; it has nodes and leaves, which every node is as a feature; this feature has branches related to the outcome of the test on it; through these tests and branches, we get the leaf which is the class label, where each leaf represents one class label.
- **Random Forest**: Random forest is a type of classification model based on a decision tree model with a multitude of them and outputting the class. Random forest is used to correct the DT outcome of over-fitting from the training set.

3.3 Dataset Used

The dataset was obtained from the Kaggle Website. It has three files:

- Titanic_train.csv: it is the dataset having 891 records for training the models, and its contents are information about passengers like (ID, name, age, etc., and the target attribute is survived). Table 1 shows the snapshot of the training dataset.
- Titanic_test.csv: it is the dataset having 418 records for testing the models, and its contents are information about passengers like training dataset without the target attribute survived.
- Titanic_gender_submission.csv: it is the dataset having 418 records which are actual values for the testing dataset of survived passengers.

3.4 Preprocessing

Data Cleaning and Feature Selection: The dataset which is there on the Kaggle Website has twelve features.

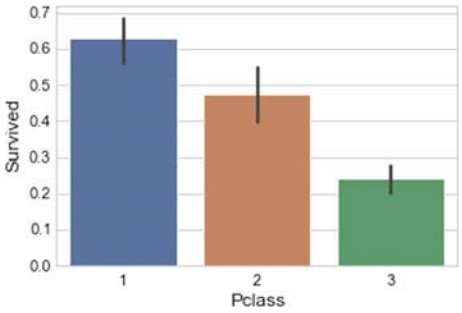

Fig. 1 Survival rate with respect to P-class

- *Passenger ID*: it represents a number of passengers; it is a counter, and it does not make a sense, so we can eliminate it.
- *Survived*: it represents the target which is needed to classify; it has two values (1,0) for survived or not survived in, respectively.
- *P-class*: it represents classes of seats and cabins for passengers; it has three numeric values (1,2,3) where the 1 for first class, 2 for second class and 3 for third class. Figure 1 shows the rate of survival of passengers for each class seats.
- *Name*: it represents names and nicknames of passengers; the nickname of the name is extracted by this feature, and those nickname (Master, Mr., Ms., Mlle, Miss, etc.) because the nickname of the name is important things in our reality.
- *Sex*: it represents the gender of passengers; the rate of survival of womens was more than that of men as at that time they used the policy of women children first. The same information is shown in Fig. 2.
- *Age*: it represents the age of passengers; in this feature, 177 instances have NA from training dataset; the missing values are handled by using the average mean of the ages.
- *SibSp and ParCh*: it represents a number of siblings, spouses, parents and children. They are relevant for the group of family units together.
- *Ticket*: it represents the ticketing system in that time, and it does not make a sense, so it can be eliminated.
- *Far*: it represents how much a passenger paid for his ticket; it does not make a sense, so it can be eliminated.
- *Cabin*: it represents seating area of the passenger; in this feature, there are lots of NA 687, and it cannot replace these by frequent value or other value because lots of cabins are not there in the dataset, and it cannot be filled by the same cabin.
- *Embarked*: it represents from which port the passenger booked his ticket; there were three ports of embarkation which are Southampton, Queenstown and Cherbourg. Two missing values are there, and it is handled by replacing them with a frequent value which is Southampton, but this feature does not make sense for our target (Survive on not Survive), so it can be eliminated and the selected features are shown in Table 2.

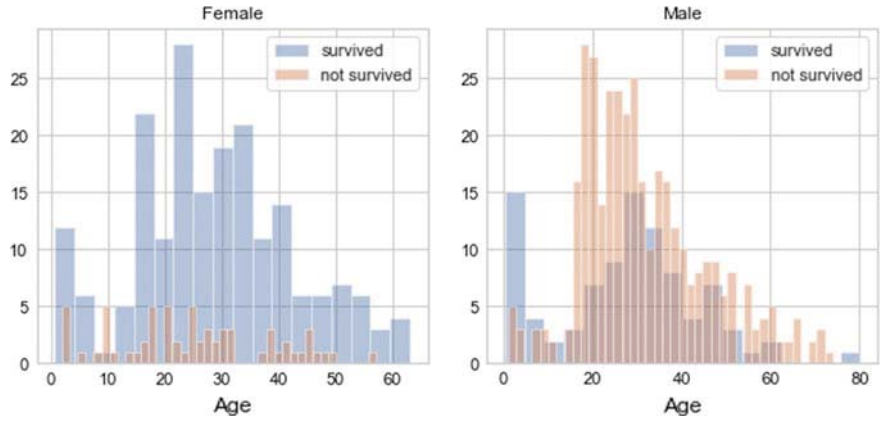

Fig. 2 Survival rate of males and females based on age

Table 2 Snapshot of the dataset after feature selection

P-class	Sex	Age	SibSp	Parch	Nickname
3	1	22	1	0	Mr.
1	0	38	1	0	Mrs.
3	0	26	0	0	Ms.

3.5 Hardware and Software Used

The experiments were carried out on Windows 10 64-bit OS, Intel(R) Core(TM)-i5-7200U@ 2.50GHz Processor, 8 GB RAM.

4 Results

In this paper, the confusion matrix is used as the metric evaluation. Accuracy is a technique to measure how much the model will predict. The model with higher accuracy is a better model. Apart from accuracy to deal improper distribution of data, we have used other parameters like precision, recall, and F1-score. Table 3 shows the results of implementation for different classification models, and Tables 4, 5 and 6 show the results of implementation for different classification models by using K-cross-validation for K equal 10, 5 and 4, respectively. From the results, it can be observed that the data in the dataset is evenly distributed as all the performance measures are giving consisitent results. We have also used k-fold cross-validation to avoid probable fluctuations in the result of classifiers for each random selection of training and testing set.

Table 3 Result for different classifiers for 891 training and 418 testing instances

Algorithm	F1_score	Accuracy	Recall	Precision
Logistic regression	95.45	95.45	94.07	93.46
KNN	86.35	86.36	80.92	81.45
SVM	95.46	95.45	95.39	92.35
Naive Bayes	98.80	98.80	97.36	99.32
DT	80.71	80.62	75.65	72.32
RF	81.49	81.33	78.94	72.28

Table 4 Results for different classifiers with $K = 10$

Algorithm	F1_score	Accuracy	Recall	Precision
Logistic regression	87.53	87.56	82.64	83.23
KNN	87.24	87.27	82.59	82.66
SVM	86.90	86.94	81.62	82.59
Naive Bayes	86.11	86.19	79.59	81.86
DT	87.01	87.06	81.35	82.80
RF	86.38	86.45	80.52	82.21

Table 5 Results for different classifiers with $K = 5$

Algorithm	F1_score	Accuracy	Recall	Precision
Logistic regression	85.66	85.79	78.58	82.76
KNN	84.49	84.57	77.84	80.58
SVM	86.12	86.25	79.42	83.26
Naive Bayes	85.51	85.71	78.48	82.88
DT	81.82	81.97	73.10	77.68
RF	83.16	83.27	75.49	79.31

Table 6 Results for different classifiers with $K = 4$

Algorithm	F1_score	Accuracy	Recall	Precision
Logistic regression	85.38	85.48	78.37	82.84
KNN	83.02	83.12	75.33	79.18
SVM	86.12	86.25	78.97	83.96
Naive Bayes	85.45	85.56	78.23	83.2
DT	80.42	80.59	71.12	76.27
RF	82.59	82.66	76.03	78.02

5 Conclusion and Future Work

In this research, we focused on preprocessing dataset and selection features and extracted main information from the name column; the final features selected as a significant feature are (P-class, sex, age, SibSp, ParCh, nickname). Naive Bayes is the best algorithm for the Titanic classification dataset where the accuracy is 98.80% as compared to other algorithms without K-cross-validation, and with using K-cross-validation to different K values equal (10,5,4), we found the accuracy is better when K equals 10 because the instances for training dataset are more than the instances for K equal 5 or 4.

In future, we will use ranking feature selection to extract some meaningful information from dataset and use deep learning algorithms to predict the survival of passengers.

References

1. Kaggle. https://www.kaggle.com/c/titanic. Last accessed 1 June 2019
2. Lam E, Tang C (2012) Titanic machine learning from disaster. In: Lam Tang
3. Balakumar B, Raviraj P, Sivaranjani K (2019) Prediction of survivors in Titanic dataset: a comparative study using machine learning algorithms. Sajrest Arch 4(4)
4. Farag N, Hassan G (2018) Predicting the survivors of the Titanic Kaggle, machine learning From disaster. In: Proceedings of the 7th international conference on software and information engineering, pp 32–37
5. Kakde Y, Agrawal S (2018) Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. Int J Comput Appl 32–38
6. Singh A, Saraswat S, Faujdar N (2017) Analyzing Titanic disaster using machine learning algorithms. In: International conference on computing, communication and automation (ICCCA). IEEE, pp 406–411
7. Ekinci EO, Acun N (2018) A comparative study on machine learning techniques using Titanic dataset. In: 7th international conference on advanced technologies, pp 411–416
8. Nair P (2017) Analyzing Titanic disaster using machine learning algorithms. Int J Trend Sci Res Develop (IJTSRD) 2(1):410–416