Titanic Disaster Prediction Based on Machine Learning Algorithms

Wenqing Liang*

Department of Statistics, University of Toronto, Toronto, Canada

*Corresponding author: wenqing.liang@mail.utoronto.ca

Abstract. The luxury British steamship named Titanic unfortunately sank in 1912 after striking an iceberg, resulting in a severe loss in life and property. This research explains and analyzes what kind of characteristics of passengers were more correlated to the survival rate in the Titanic disaster. The research also demonstrates how to predict the survival rate in the Titanic disaster by using two interpretable machine learning algorithms named Decision Tree and Random Forest and finds which model is better for the survival rate prediction. The feature importance of each model is visualized, and it illustrates that a passenger's age, sex and ticket class are the three most significant causes correlated to the survival rate. The ticket class implied passenger's socioeconomic class, which means physical space of the cabins on the Titanic played an important role in surviving from the Titanic disaster. After data cleaning and model building, the result of accuracy score proves that Decision Tree algorithm performs better than Random Forest.

Keywords: Titanic disaster; survival rate prediction; Random Forest; Decision Tree.

1. Introduction

Both natural disasters and man-made disasters can lead to severe damage to people's life and property, which have been big threats to our society for a long time. Disasters are challenging to be forecasted and controlled due to their abruptness and volatility. One feasible method to prevent potential severe loss is to build a model based on the data collected from past disasters to predict future survival rate in similar disasters. Machine learning models can be applied to predict patients' survival rate, which can help doctors to choose the best individualized therapy with highest survival rate according to the patient's situation since they have been used in other fields successfully [1, 2]. Thus, it can be believed that machine learning model will play an important role in improving survivals in various fields in the future.

As one of the most famous tragedies, the luxury British steamship named Titanic regrettably sank in 1912 after striking an iceberg, resulting in more than 1500 deaths [3]. The disaster shocked and saddened the whole world. Singh et al. centered on R and python and they proved that Logistic Regression model was the best algorithm for Titanic classification problem compared to Naïve Bayes, Decision tree and Random Forest. They also pointed out that both the result of Logistic regression and Random Forest showed that passengers' ticket class, sex, age, number of parents or children travelled with them are the features that are tightly correlated to the survival rate [4]. Dasgupta et al. suggested that females were 4 times more likely to survive than males and families that had less than 3 members had higher survival rate by applying the technique of Logistic Regression [5].

This research paper aims to correctly predict the survival rate of travelers on the Titanic based on a set of demographic data containing passengers' different characteristics and basic personal information. The predictive model is built by applying machine learning algorithm in Python. This research focuses on Random Forest algorithm and Decision Tree algorithm. After analyzing the Titanic dataset, the result could show what kind of characteristics of the passengers are more correlated to survival rate and which model has higher accuracy in predicting survival rate of Titanic Disaster.

2. Dataset Description and Preprocessing

The dataset used for the paper is provided by the Kaggle website [6]. The dataset is composed of 891 rows and 12 columns which contains 12 features. The information of each passenger's id, survival, ticket class, name, sex, age, number of siblings or spouse traveled with this passenger, number of parents or children traveled with this passenger, ticket number, passenger fare, cabin number, port of embarkation is included in the dataset. Among them, Passenger, Pclass, SibSp, parch are numerical variables, Survived and Sex are binary variables, Name and Embarked are categorical variables. Since the correlation between one's ticket number, passenger fare, cabin number and the survival rate were relatively week, the research put emphasis on the rest 9 features: Passengerid, Survived, Pclass, Name, Sex, Age, SibSp, Parch as well as Embarked. The percentage of null value for 'Age' and 'Embarked' are calculated to be 19.8% and 0.22% respectively. In the dataset, 644 people embarked in Southampton, 188 people embarked in Cherbourg and 77 people embarked in Queenstown. Considering that most of the travelers embarked in Southampton, it can be logically assumed that the passengers with no record in the feature Embarked also embarked in Southampton. Hence, the missing values in Embarked are replaced with 'S' presenting Southampton.

Since a passenger's age was an important factor in surviving from the Titanic disaster, filing the missing value in Age with 0 may cause inaccurate results. The dataset is cleaned by two different methods so that the results can be compared to find which data cleaning method has higher accuracy. One method to deal with the null entries in the feature Age is to fill the missing values with the average age of all the passengers. The average age of all the passengers is calculated to be 23.79929292929293. Another method is to classify the title of the passengers and use the average age of people with the certain title to fill the null entries in Age. All types of passengers' titles and their numbers are listed in Table 1. In order to simplify the process of distinguishing different titles in the dataset, titles used by a few passengers like Capt, Col, Don, Dr, Jonkheer, Lady, Major, Rev are replaced with Rare. Countess, Lady, Sir are the titles of the nobility, so they are regarded as Royal. Special titles Mlle, Ms and Mme are substituted with more common titles like Miss and Mrs. The new feature Title need to be switched to numerical values to prepare for future analysis and model choosing. Number 1, 2, 3, 4, 5, 6 represent Mr, Miss, Mrs, Master, Royal, Rare. The next step is to fill the missing values in Age based on the related numerical value in the feature Title.

Table 1. Three Scheme comparing

Sex Title	female	male
Capt	0	1
Col	0	2
Countess	1	0
Don	0	1
Dr	1	6
Jonkheer	0	1
Lady	1	0
Major	0	2
Master	0	40
Miss	182	0
Mlle	2	0
Mme	1	0
Mr	0	517
Mrs	125	0
Ms	1	0
Rev	0	6
Sir	0	1

A passenger's title contained many information, which can not only indicate his / her marital status but also his / her approximate age group. It is assumed that passengers with same titles were in the same age group. Therefore, the missing values in field Age are filled by the average age of the title group the certain passenger belonged to. For instance, if the passenger with no record in age had number 2 in field Title, which means that the passenger had the title Miss, the missing value is replaced with the calculated average age of all passengers with title Miss. After representing the feature Sex and Embarked by dummy variables, cleaning data is finalized and the dataset has 10 features: Survived, Pclass, SibSp, Parch, Age, Sex_female, Sex_male, Embarked_C, Embarked_Q, Embarked_S. the example of the cleaned dataset is provided in Table 2.

Table 2. Cleaned Dataset

	Survived	Pclass	SibSp	Parch	Age	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
0	0	3	1	0	22.00000	0	1	0	0	1
4	0	3	0	0	35.00000	0	1	0	0	1
5	0	3	0	0	32.36809	0	1	0	1	0
6	0	1	0	0	54.00000	0	1	0	0	1
12	0	3	0	0	20.00000	0	1	0	0	1
…	…	…	…	…	…	…	…	…	…	…
766	0	1	0	0	46.00000	0	1	1	0	0
796	1	1	0	0	49.00000	1	0	0	0	1
822	0	1	0	0	38.00000	0	1	0	0	1
848	0	2	0	1	28.00000	0	1	0	0	1
886	0	2	0	0	27.00000	0	1	0	0	1

3. Model Implementation

The research aims to compare two interpretable machine learning algorithms named Decision Tree and Random Forest to find the best Titanic survival rate prediction model. The two datasets gained from different data cleaning methods mentioned before are fit to the two algorithms separately. Each cleaned dataset is divided into two parts. 80% of the dataset is named as training dataset which is used to fit different models and 20% of the dataset is named as testing dataset which is used to test the accuracy of the prediction model. The feature Survived is regarded as the response variable and other features are explanatory variables.

3.1 Random Forest

Random Forest algorithm is a classification algorithm that constructs a multitude of decision trees when training and its output is the class chose by most trees. Each tree in the ensemble is grown in accordance with a random parameter [7, 8]. Random Forest can produce predictions with high accuracy and handle a very large number of input variables without overfitting. Thus, it is regarded as one of the most accurate general-purpose learning techniques.

The Random Forest Classifier is built by applying Sklearn library. For the parameters, n_estimators are set to 1000 which means there are 1000 trees in the forest, random_state is set to 32 to make the train-test splits always deterministic so that the results can be reproduced and verbose is set to False to control the verbosity when fitting and predicting. After building the Random Forest Classifier, fit function is used to train the model on the training dataset and prediction is performed on the testing dataset. In order to provide convictive evidence to compare the Random Forest Model with the Decision Tree Model, Metrics in the Sklearn library are used to calculate the accuracy score of the model.

3.2 Decision Tree

Decision tree is a flow-chart-like tree structure. Each internal node of the decision tree is denoted by rectangles to represent a test on a feature, each branch represents the consequence of the test and

the leaf nodes are denoted by ovals to show a class label [9, 10]. Decision tree is commonly used because its concept is easy to be understood and it can be combined with other decision techniques. Decision tree algorithm has high accuracy when the problem is naturally divided by lines parallel to the different axis, but sometimes it has the problem of overfitting and sensitivity.

The Decision Tree Classifier is built by applying Sklearn library. For the parameters, the criterion of the model is set as entropy, the splitter is set as best to choose the best split and random_state is set as 32 to control the randomness of the estimator. After the model building step, fit function is used to train the Decision Tree Classifier on the training dataset and prediction is performed on the testing dataset. Metrics in the Sklearn library are used to calculate the accuracy score of Decision Tree.

For each model, feature importance is visualized to illustrate which feature was the most important factor influencing the survival rate of Titanic disaster.

4. Result and Discussion

4.1 The Feature Importance of Models

For the dataset cleaned by filling the missing values in the feature Age with the average age of the certain title group the passenger belonged to, the feature importance of the Random Forest Classifier and Decision Tree classifier are illustrated in Fig 1 and Fig 2, respectively. If the dataset is cleaned by replacing null entries in the feature Age with the average age of all passengers, the plots of feature importance under the two models are the same and the result is shown in Fig 3. The histograms all illustrate that Age is the most important feature and Sex, Pclass appear to be the second and the third important feature regardless of different data cleaning methods and models. The result proves that a passenger's survival rate was tightly related to the age, sex and ticket class. The ticket class could imply the passenger's socioeconomic class and it led to the difference in the physical space of the cabins on the Titanic, which means that upper class people were much closer to the boast deck and the lifeboats. This may be the main reason why a passenger's ticket class played an important role in the survival rate, and it can be inferred that passengers with higher ticket class may have higher survival rate than others.

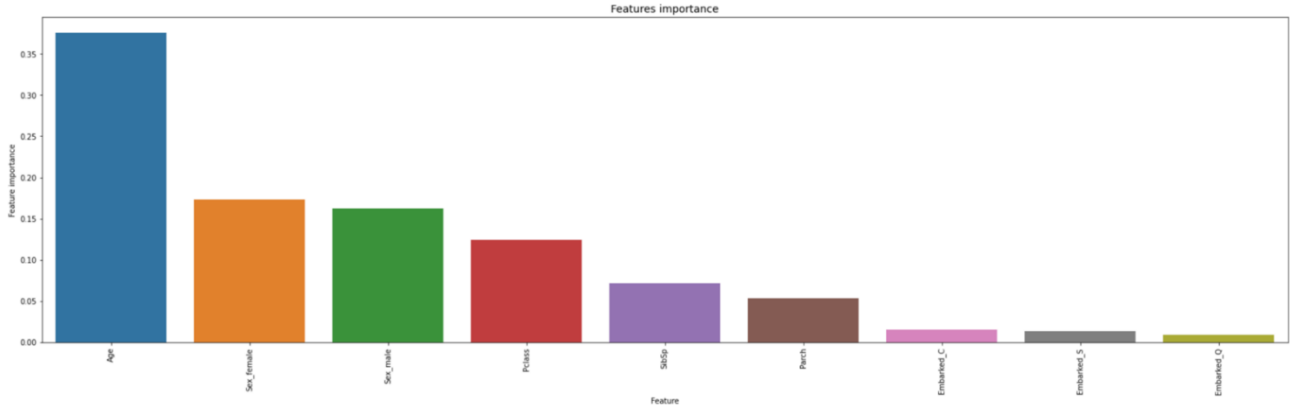

Figure 1. Feature Importance of Random Forest (If fill the null entries in Age with the average age based on each title)

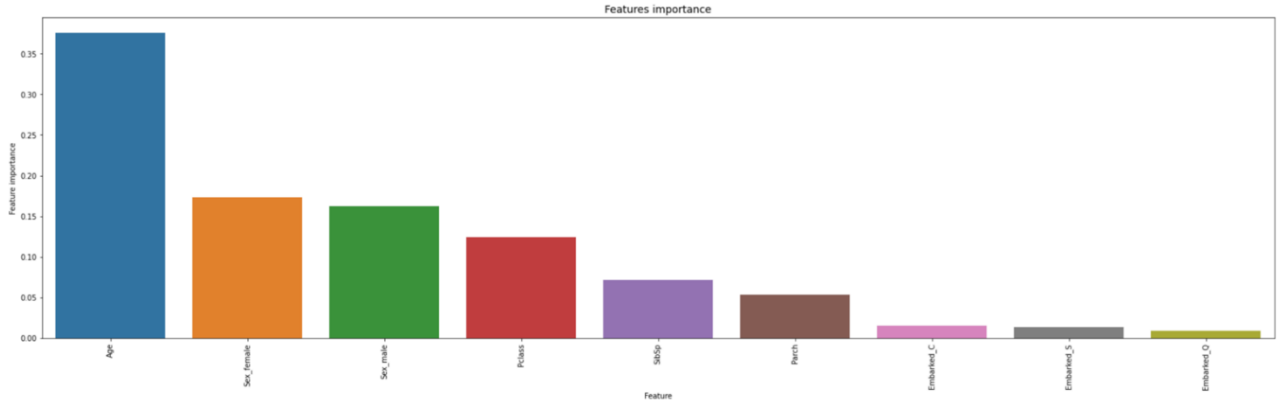

Figure 2. Feature Importance of Decision Tree (If fill the null entries in Age with the average age based on each title)

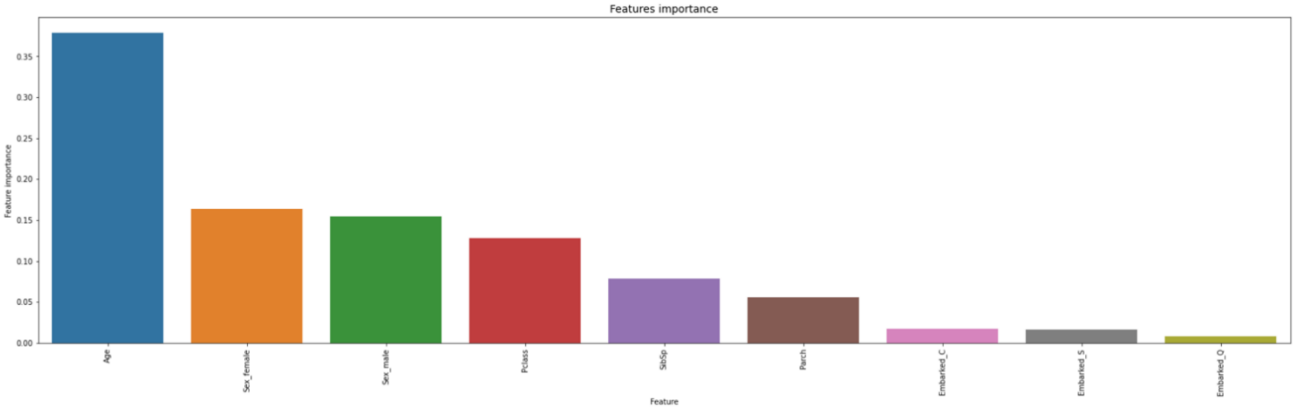

Figure 3. Feature Importance of Random Forest and Decision Tree (If fill the null entries in Age with the average age of all passengers)

4.2 The Performance of Models

The accuracy scores of different models and data cleaning methods is listed in Table 3. If the dataset is cleaned by filling the missing values in the feature Age with the average age of the certain title group the passenger belonged to, the accuracy score of the Random Forest Classifier is 0.759 and that of the Decision Tree Classifier is 0.761. If the dataset is cleaned by replacing null entries in the feature Age with the average age of all passengers, the accuracy score of the Random Forest Classifier and the Decision Tree Classifier are both 0.783.

Table 3. The performance of various schemes

Model	Dataset 1	Dataset 2
Random Forest	0.759	0.783
Decision Tree	0.761	0.783

5. Conclusion

In the research, the tree-based methods were employed to predict the survival rate of the Titanic Disaster. Decision Tree proved to be better than Random Forest due to the high accuracy score. Both Random Forest and Decision Tree suggested that a passenger's age, sex and ticket class are the three most significant causes correlated to the survival rate. Future work might include more advanced machine learning algorithms like neural networks to find an even better model to predict the survival rate of Titanic Disaster. Further research can be conducted to apply the prediction model to other disaster data to prevent severe loss in potential disasters in the future.

References

[1] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1, 2001, 89 - 109.

[2] Dixon, Matthew F., Igor Halperin, and Paul Bilokon. "Machine learning in Finance." Vol. 1406. New York, NY, USA: Springer International Publishing, 2020.

[3] History.com Editors, "Titanic": https://www.history.com/topics/early-20th-century-us/titanic, 2009.

[4] A. Singh, S. Saraswat and N. Faujdar, "Analyzing Titanic disaster using machine learning algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 406 - 411.

[5] A. Dasgupta, V. P. Mishra, S. Jha, B. Singh and V. K. Shukla, "Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning," 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2021, pp. 52 - 57.

[6] Kaggle, "Titanic-Machine Learning from Disaster," https://www.kaggle.com/competitions/titanic/data, 2012.

[7] Biau G. "Analysis of a random forests model." The Journal of Machine Learning Research, 2012, 13 (1): 1063 - 1095.

[8] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25. 2, 2016, 197 - 227.

[9] Priyam A, Abhijeeta G R, Rathee A, et al. "Comparative analysis of decision tree classification algorithms." International Journal of current engineering and technology, 2013, 3 (2): 334 - 337.

[10] Myles, Anthony J., et al. "An introduction to decision tree modeling." Journal of Chemometrics: A Journal of the Chemometrics Society 18.6, 2004, 275 - 285.