

Titanic Survival Prediction Based on Machine Learning Algorithms

Yuming Zhang*

RCF Experimental Scholl, Beijing, 100000, China

* Corresponding Author Email: lulubelle4956@email.phoenix.edu

Abstract. This report is aimed at demonstrating the application of machine learning techniques for predicting the survival of passengers who boarded the Titanic. After analyzing the Titanic dataset, which includes variables Pclass, Sex, Age, SibSp, Parch, Fare, Ticket, and Cabin, two machine learning algorithms, Logistic Regression and Random Forests Model, are used to give survival predictions. Models are compared to find an accuracy difference, and the magnitudes that each factor has on survival are also identified. Data preprocessing is the essential technique that will be used to adjust the data set. Before this process, correlations between variables are analyzed to give directions for feature engineering. And for feature engineering, data conversion and vacancy filling are first implemented. Afterwards, features are selected while new features are gained for model implementation. In understanding the final outputs of the model, new features, like name length, combine to give insight into more implicit survival factors that have been previously ignored.

Keywords: Data preprocessing, logistic regression, random forest model.

1. Introduction

The Titanic event is one of the most famous tragedies in history: the Olympic-class ship Titanic sank on her maiden voyage to America in April 1912 after hitting an iceberg. 1502 passengers out of the total of 2224 passengers were killed, along with crew numbers. By studying such disasters, researchers can understand several survival factors. This paper aims to use machine learning to identify the different influences that each factor has on Titanic survival and test the accuracy of survival prediction.

The factors involved in survival on the Titanic are very complex. However, a comprehensive data set about personnel on the ship, including name, age, and ticket price of passengers, is provided by Kaggle to launch a competition. Both foreign and domestic scholars have found correlations between the above criteria and survival. In some papers, Haque et al. identified specific directions for feature engineering and applied decision tree models to make predictions [1]. Survival status was also analyzed using clustering techniques. This research suggested that passengers traveling with a small family size of 2–4 people survived more compared to death. Sherlock et al. provided essential data classification guidance and further explored the application of the random forest model, concluding that the influences that different features have on survival obey this order from top to bottom: Sex, class, incarcerated, and age group [2]. Singh et al. used a variety of models, including logistic regression, k-nearest neighbors, SVM, naive Bayes, decision trees, and random forests, to make survival predictions. In their work, algorithms with high accuracy could be identified as accuracies of the algorithms based on features fed to them that had been compared in tabular form [3, 4]. In order to predict the survival rate of people, Shekhar et al. proposed an analytical approach on the Titanic ship by applying suitable classifiers with the help of Python [5]. It is also shown that through the spot check algorithm, the survival prediction is as close to 86.29% as possible, which is the highest result as most of the other algorithms accuracy was between 72 and 83%.

Kakde had provided a process flow chart [6]. This paper gave insight into handling missing values, analyzing data relationships, converting variables, designing feature engineering, and combining models. Overall, this paper inspired intuitive analysis of ideas.

Aside from research on the topic of Titanic survival, there are also papers introducing fundamental ideas in data preprocessing, which is the key to determining prediction accuracy. Guyon I et al.

provided an introduction to feature extraction, which could be used to create a subset of new features by combining existing features [7, 8]. In combination with their work, Heaton's analysis of feature engineering provided guidance for data preprocessing [9]. Yates et al. also contributed information related to the confusion matrix to summarize predictive performance [10]. It also introduced cross-validation, allowing parameters with high accuracy to be obtained. In summary, this article will mainly focus on logistic regression and random forest models, combined with cross-validation, to yield Titanic survival predictions.

2. Methodology

2.1. Data Source

The data for this literature is from the Kaggle website, which was released by Alexis Cook in 2022 in the Titanic: Machine Learning from Disaster competition for 891 individuals.

2.2. Variable Selection

The data used in this paper is from a train set provided by the Kaggle website, counting a total of 891 people onboarded. From analyzing the data, we know that a total of 342 of 891 people survived, of whom 233 were female. The original dataset has many nulls for variable Cabin. Due to the large number of data vacancies, this paper chose to eliminate the cabin column. There are 9 variables in the dataset that are relevant to survival: Pclass, Sex, Age, SibSp, Parch, Fare, Ticket, Cabin, and Embarked. The specific descriptions are listed in Table 1.

Table 1. Variable Description

Variable	Definition	Key
Survival	Survival	0 = no, 1 = yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	none
Age	Age in years	none
SibSp	Number of siblings	none
Parch	Number of parents/children onboard the Titanic	none
Fare	Passenger fare	none
Ticket	Ticket	none
Cabin	Cabin	none
Embarked	Port of Embarkation	C = Cherbourg, Q =Queenstown,S =Southampton

For variables that have vacancies, data preprocessing is implemented to modify the data set while creating new features based on the above variables.

2.3. Method Introduction

This paper uses Logistic Regression and Random Forest Model for machine learning to determine the outcome of individuals on the Titanic and compares the prediction with the actual outcome of passengers to produce a survival rate. Various data preprocessing techniques, including feature engineering, are also adopted in this research.

The Logistic Regression model directly gives the probability of each value belonging to 1 or 0. It is a model that can carry out the sorting process by finding the best fit curve from existing data. The Random Forest Model is able to randomly choose data and features in the construction of each decision tree. The use of Random Forest Model investigates the influence that each feature has on the survival of people onboard. In the forest, each decision tree contains different combinations of features.

Along with the use of the above models, Cross Validation is also implemented. The idea behind Cross Validation is to adjust the parameters by dividing the training dataset into several parts. In this prediction, the train set is divided into several parts to serve the roles of gaining more accurate model reflections and providing test data. By analyzing the relationships between variables, features that are crucial in the data set are identified for feature engineering.

3. Results and Discussion

3.1. Correlation Analysis

Table 2 records the correlations between each feature and survival. Table 2 demonstrates that females are more likely to survive; their survival rate is 74.20%, whereas males only have a survival rate of 18.90%. Passengers of higher class are also more likely to survive, with a survival rate of 63.0%.

Based on the 891 specimens provided, the average age is 30 years old, the standard deviation is 13.5 years old, the minimum age is 0.42 years old, and the maximum age is 80 years old. Passengers are divided into children, teenagers, adults, and the elderly according to age. Table 2 records that the survival rate for children seems to be higher, and passengers that have families onboard also have a higher survival chance.

Table 2. Survival Rate

Vairables	Categories	Survival Rate
Sex	Female	0.742
	Male	0.189
Pclass	1	0.630
	2	0.473
	3	0.242
Age group	0 to 12	0.506
	12 to 18	0.467
	18 to 65	0.365
	65 to 100	0.125
Siblings	Yes	0.466
	NO	0.345
Parents and Chilren	Yes	0.488
	No	0.344

The appellation is also related with the survival, shown by Figure 1. Figure 1 further shows that women of middle or higher status have a relatively higher survival rate. It also correlates with the idea that people of higher-status are more likely to survive, indicating nobility is an important determinant.

The relationships between embarked and survival rate showed that C had the highest survival rate and S had the lowest survival rate, with Q in the middle. After identifying each variable's correlations with survival, connections between variables are understood.

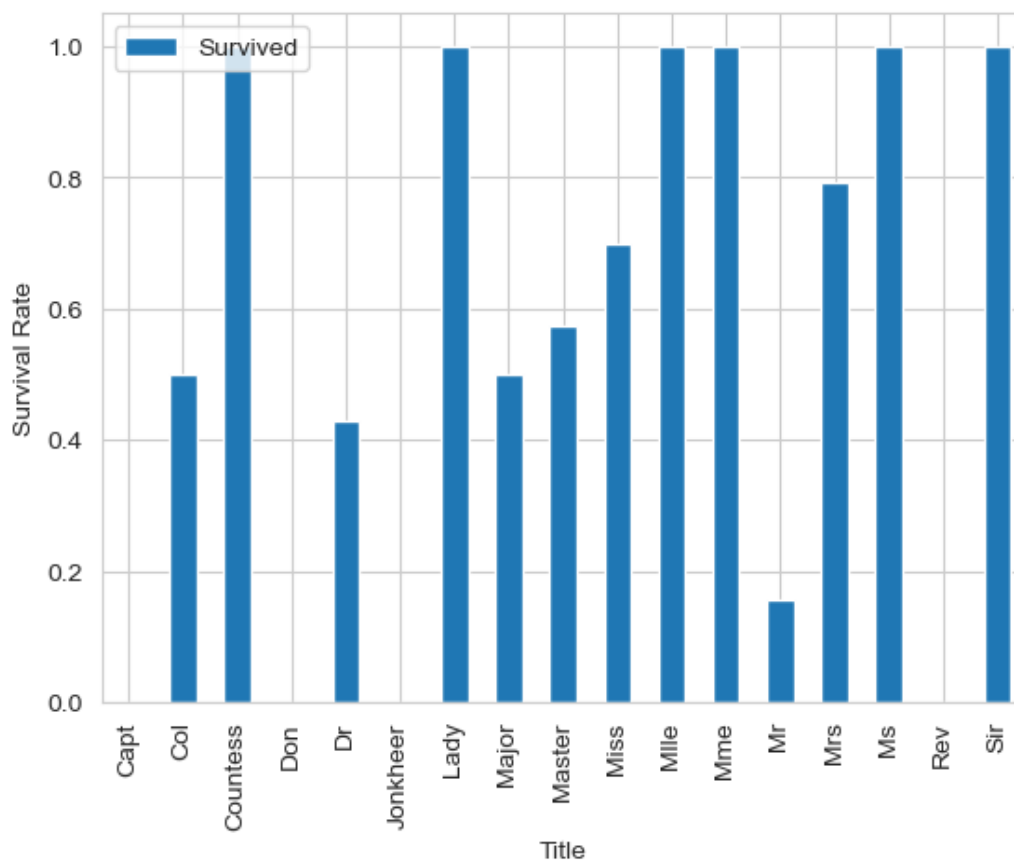


Fig. 1 Survival Rate by Appellation

3.2. Data Preprocessing

Variable conversion is needed before inputting data into the algorithms. There are two kinds of data: qualitative and quantitative. For variables Embarked and Sex, dummy is used as a qualitative conversion. Dummy is also used after extracting the appellation from the Name to sort different appellations. For the Fare part, the previous analysis shows that some of the ticket number data is duplicated. After combining the data of the number and name of relatives and the comparisons of the ticket price and cabin class, it is indicated that there are family tickets or group tickets, so procedures that divide the ticket price of group tickets among individuals in the group are needed. In addition, binning is used to classify the ticket price base into five levels. For Pclass, no more preprocessing is actually required. However, the ticked price may also indicate the location of the cabin in each class. Therefore, the highest and lowest price for each cabin are distinguished. From previous analysis, Parch and Sibsp were discovered to have limited influences, and the new feature Family_Size was constructed by combining these two variables. For the age column, since there are many missing values, mode or mean are not suitable for filling in the blanks. Therefore, machine learning is needed to create new features based on given variables. In this section, train data is data with Age, while test data is the wanted values to fill in the blanks in Age column. After using machine learning (Ensemble Learning), Age vaccancies are diminished. For Ticket, there are both alphabets and numerical series. The alphabets may indicate different locations or classes of the cabin, which is crucial in determining survival. Therefore, the alphabets and numbers are divided. For the last variable, Cabin, there are too many missing values, leaving little room for preprocessing. As a result, this feature was removed. The process can be seen in Figure 2:

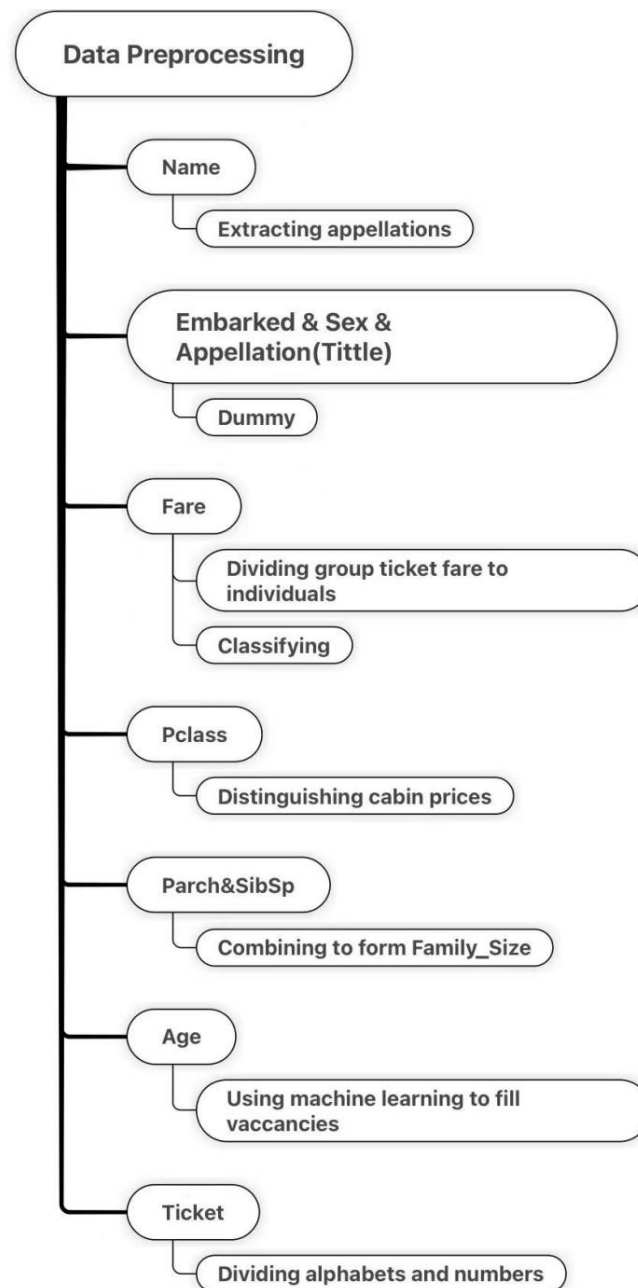


Fig. 2 Flow Chart

After factorizing a variety of features, regularization is implemented for Age, Fare, and Name_Length to prevent overfitting. Too many features will not only affect the speed of model training but may also make the model overfit. Therefore, in this case, where too many features are generated, different models can be adopted to screen all the features and select the first n features.

3.3. Model Implementing

After data preprocessing, data is input into the models, along with Cross Validation. Since the test set provided by Kaggle doesn't include the actual outcomes of individuals, Cross Validation is also used as a way to divide the train set data to compare accuracies.

The accuracy of the model is reflected by comparing the estimated individual survivors to the ground truth of each passenger in train set.

In Logistic Regression model, this sorting model can directly give the probability of each value belonging to 1 or 0, and its accuracy of prediction is 79.57%. Linear Regression can also be considered to conduct Cross Validation, however, it is less accurate, giving only an accuracy of 78.34%.

In Random Forest model, procedures that determine how many trees there are in the forest, what the minimum number of rows is, and what the minimum number of samples there are for a tree branch to end are needed. After adjusting the parameters, the accuracy of the model floats between 85.0% and 87.5%, as shown in Figure 3:

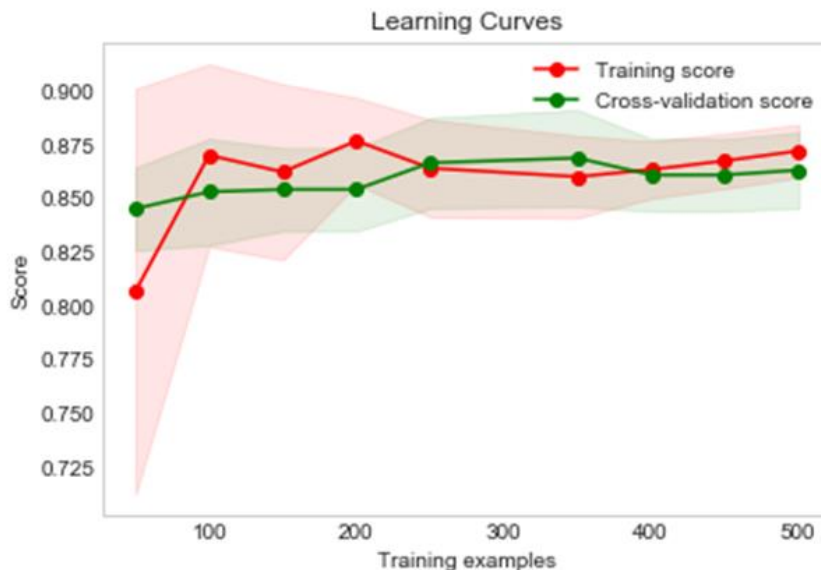


Fig. 3 Accuracy of Random Forest Model

The influences that different features have on survival is shown in Figure 4:

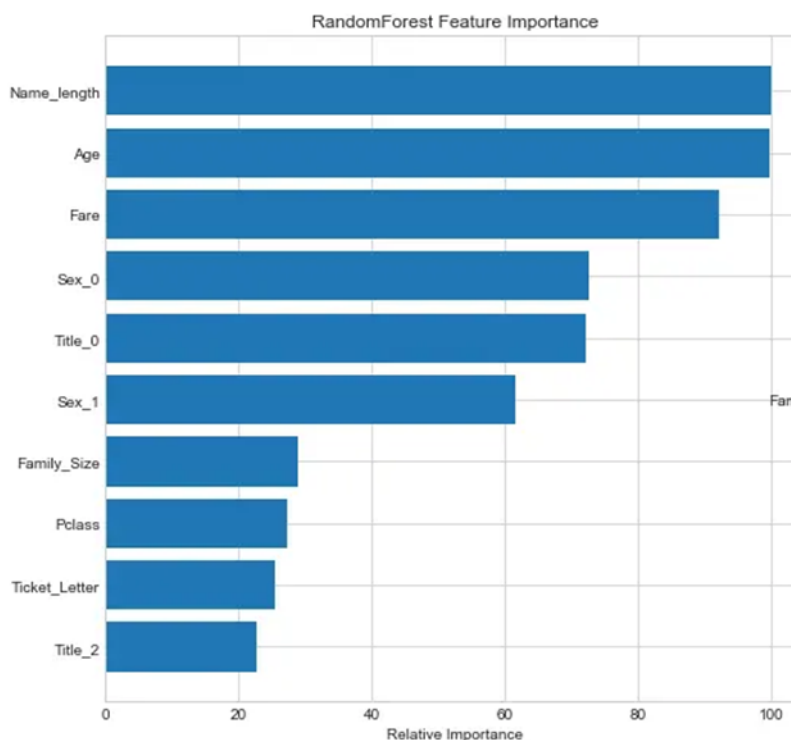


Fig. 4 Influences of different features

In general, Random Forest Model is a more accurate model for survival prediction. Its high accuracy is due to the fact that it can capture complex non-linear relationships between multiple factors, whereas Logistic Regression can only capture the best-fit curve through values. The Random Forest Model is also robust to outliers in the data as it compares with multiple decision trees, reducing the impact of outliers on final predictions. Overall, the Random Forest Model can model nonlinear trends and other intricate patterns that exist in a set of data while being robust to outliers.

4. Conclusion

The study selected 891 people onboard the Titanic as samples, which have nine variables, for survival prediction. By implementing data into both Logistic Regression and the Random Forest Model, the survival rate is checked, and the Logistic Regression yields a more accurate result. The magnitude of the influences of different features on survival is also determined by the Random Forest Model. Name_length and Age seem to have the highest relative influence, which is indeed surprising. However, longer name length often indicates greater nobility, while Fare, the third important influence, is also a sign of position, and Age can also be interpreted to have connections with status. To summarize the result, social status may be the overall most important determinant, followed by sex, perhaps linked with the code of “lady first”.

It cannot be denied that due to limited parameter adjusting, the result may be less accurate and the variables’ correlations may be exaggerated, causing inaccuracies. Nevertheless, the study still has some merits. Firstly, the approach taken in feature engineering is relatively comprehensive and sober that machine learning is used for imputing values into vacancies in Age, and similar features are combined to form new features, reducing computing power. A geographical approach was also implemented to analyze the data. This enables visualization, making the data clearer and more readable. Moreover, reasons behind accuracy differences are also analyzed from the perspective of algorithm theory, and models are compared, enabling a better understanding of machine learning algorithms. This paper also discovers many good data-analysis methods and procedures and combines these advantages together to summarize important ideas. It also interprets the result from a social angle, which is the work that a few papers have carried out.

References

- [1] Haque A, Shivaprasad G, Guruprasad G. Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster. IOP Conference Series: Materials Science and Engineering, 2021, 1065(1): 12042.
- [2] Sherlock J, et al. Classification of Titanic passenger data and chances of surviving the disaster. Cornell University, 2018.
- [3] Singh K, Nagpal R, Sehgal R. Exploratory data analysis and machine learning on Titanic Disaster Dataset. IEEEEXPLORE, 2020.
- [4] Tabbakh A, Rout J K, Rout M. Analysis and prediction of the survival of Titanic passengers using machine learning. In Lecture notes in networks and systems, 2020, 297-304.
- [5] Shekhar S, Arora D, Sharma P. Classifying Titanic Passenger Data and Prediction of Survival from Disaster. In Lecture notes in networks and systems, 2020, 181-187.
- [6] Kakde Y, Agrawal S. Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. International Journal of Computer Applications, 2018, 179 (44): 32-38.
- [7] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning Proc. Int. Conf. on Science and Information (London), 2014, 372-378.
- [8] Guyon I, Elisseeff A. An Introduction to Feature Extraction (Feature Extraction. Studies in Fuzziness and Soft Computing Vol 207 Springer, Berlin, Heidelberg), 2006.
- [9] Heaton J. An empirical analysis of feature engineering for predictive modeling 2016 Proc. SoutheastCon (Norfolk VA), 2016, 1-6.
- [10] Ao Huajian, Yu Kaichao. Design of the Titanic Survival Prediction Model. Information Technology, 2023, 47(5): 6-12.