

The Prediction and Feature Importance Investigation in Titanic Survival Prediction

Chutong Huang^a

Beijing Luhe High School International Academy, Beijing, China

Keywords: Artificial Intelligence, Random Forest, Feature Importance.


Abstract: Predicting the survival of Titanic passengers is one of the topics scientists are focusing on. This paper explores the use of the Random Forest (RF) algorithm on a Titanic dataset and analyses the key features that influence the predictions. The RF algorithm is applied to a processed dataset. Feature importance scores are returned for each feature to demonstrate how much it is related to the survival prediction, and the scores are then analyzed in their historical context. Age, fare and sex were found to be the three most significant features in predicting survival. Age is significant as its correlation to survivability, the children are determined to live while the elderly are unable to survive. Fare is a crucial attribute since it is correlated with passenger class, meaning that those paying more are given better information and location to survive. Sex is important because women and children are given priority to survival, while men don't have that chance. The application of Random Forests shows how well Artificial Intelligence (AI) algorithms can predict problems and spot significant patterns in complex data sets. And the analysis could have useful implications for improving predictive models in other areas where attributes are crucial.

1 INTRODUCTION

On April 15, 1912, the RMS Titanic sideswiped an iceberg on its first voyage, causing over 1,500 of the 2,240 passengers and staff members on board to perish in the disaster. This notorious disaster soon became a warning and appears in many films, articles, and novels, warning people the danger of nature (Titanic History, 2024). Enhancing passenger survival in such disasters is an issue that needs to be taken seriously, and Artificial Intelligence, as an emerging technology with strong feature extraction and prediction capabilities, can be considered in conjunction with this task.

AI has evolved rapidly, with significant progress in machine learning, deep learning, and data analytics. It utilizes sophisticated algorithms like logistic regression, random forests, and neural networks to analyze complex data patterns, which is beneficial in forecasting outcomes and improving choices. These advancements have enabled more accurate predictions and enhanced decision-making across various fields, such as medical science. Choi et al. constructed a Recurrent Neural Network (RNN)-

based model that predicts future events of patients (Choi, 2016). Wang et al. utilize a RNN model for the prediction of the future statues of Alzheimer's Disease for patients (Wang, 2018). Among the many prediction tasks, one important direction is the prediction of classification problems like survival of a person. Hsieh et al. illustrated a model of Fuzzy Hyper-Rectangular Composite Neural Network that predicts the survival through the first 24 hours physical data of patients (Hsieh, 2014). Pradeep et al. applied machine learning algorithms include Naive Bayes, classification trees, and Support Vector Machine (SVM) to the information of lung cancer patients and predict their survivability rate (Pradeep, 2018). Kakde et al. determined the impact of each feature to the survival rate and compared algorithms between SVM, decision tree, random forest, and logistic regression (LR). They found that SVM as well as logistic regression perform nearly the best, and there was high influence of age on survival while other features like passenger class, age, fare and "sibsp"(siblings and spouses) all have influences as well (Kakde, 2018). Nair et al. analyzed the correlation between factors of passenger samples and

^a <https://orcid.org/0009-0003-2175-9681>

the survivability of the passengers. They suggested that LR perform the best with the lowest false discovery rate and the highest accuracy. In their model of logistic regression, it tells them that the top5 correlated features are "Pclass"(Passenger class), sibsp, age, children, and sex (Singh, 2017). The effectiveness of AI methods has been demonstrated on many domain tasks, so this paper intends to consider the use of AI algorithms in predicting survival of people in Titanic disaster, but unlike previous studies on the subject, this paper's focus is more on the influence of passengers' features.

In this paper, based on the Kaggle dataset, the random forest algorithm was used for prediction and the feature importance of each feature was compared, as well as the correlation between features and survival was analyzed.

2 METHOD

2.1 Data Preparation

In this study, a dataset from Kaggle is used (Kaggle, 2017). It consists of 1309 samples of passengers on the Titanic, each of these passengers are marked by 8 features such as sex, age and fare. In terms of data preprocessing, it is divided into two parts. First, missing values are handled by imputation using statistical measures, and then categorical variables are converted to numerical formats using label encoding.

2.2 Machine Learning-based Prediction

2.2.1 Introduction of Machine Learning Workflow

Machine learning generally involves several key steps to create an effective predictive model. First, data collection gathers relevant information from various sources to ensure a comprehensive dataset. Next, data preprocessing cleans and transforms this data by fitting missing values, transforming categorical variables, and then scaling features to prepare it for analysis. Feature selection follows, identifying and retaining the most important variables that contribute to the model's performance. Model selection involves choosing the appropriate algorithm

based on the features included in the dataset and the practical problem. It studies from the data to identify patterns during training and make predictions. Performance evaluation assesses the model's accuracy and effectiveness using separate test data. Finally, model tuning adjusts the settings or parameters of the algorithm to enhance its performance, ensuring the model is well-suited to the specific problem and data.

2.2.2 Random Forest

Random Forest (RF) is an accurate and effective supervised machine learning method which combines various decision trees to form a "forest." It is applicable to situations involving both regression and classification. An additional kind of method for data classification is a decision tree. It resembles a flowchart that clearly illustrates the process of progressing choice. It starts at one beginning tree and produces two or more branches in which each tree branch giving a distinct set of possible outcomes. The RF model can achieve high prediction accuracy by combining multiple decision trees. The principle behind this is that several unrelated models of decision tree produce a noticeable improvement when used together. In detail, each 'tree' in the 'forest' casts a 'vote' for a problem, the forest integrates all the votes and chooses the one with the majority of those votes. RF takes different votes from multiple trees, which helps to increase the robustness and reduce the overfitting problem of the algorithm, which is its greatest merit (Careerfoundry, 2023).

In this study, Random Forest is used in Python, scikit-learn (sklearn) provides a random forest classifier library. After applying Random Forest to the dataset, the 1309 survival predictions are fitted into a confusion matrix to calculate the accuracy of the prediction. A confusion matrix is a 2×2 matrix with "True Positive" (TP) and "True Negative" (TN) in one row and "Predict Positive" (PP) and "Predict Negative" (PN) in another row. The accuracy is given by $(TN+TP) / (TN+TP+PN+PP)$. In addition, other evaluations such as recall, precision and F1 score can be calculated from the 4 values above. As for the feature importance, it is returned by the function "feature_importances_" in which list the feature importance of each feature, the greater the value, the more important it is.

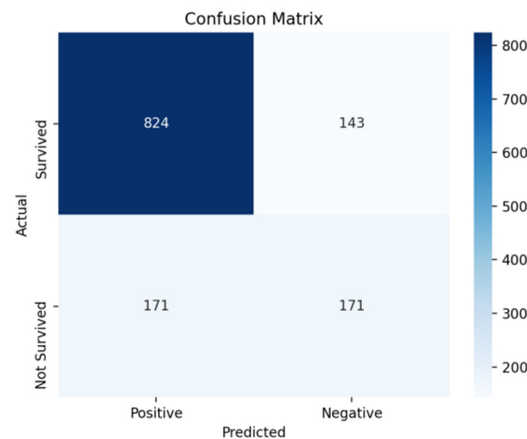


Figure 1: The confusion matrix of the prediction (Photo/Picture credit: Original).

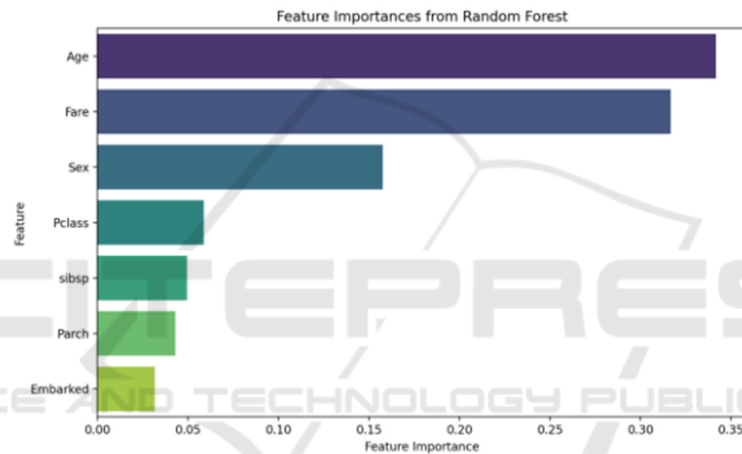


Figure 2: The feature importances of each feature (Photo/Picture credit: Original).

3 EXPERIMENTAL RESULTS AND DISCUSSION

According to Figure 1, the prediction accuracy is 76.0%, which is a good performance. Then moving to Figure 2, the importance of the features shows that the top 2 features are much more important than others, while the third is also way ahead of the rest. Therefore, age, fare, and sex are considered to be dominant features, and the remaining features have no significant effect on the prediction.

3.1 Feature Analysis

In the case of the Titanic disaster, "age" was an essential factor because it was directly related to the chances of survival. Due to the good customs of the

community, children were often given priority for lifeboats, as were women. On the other hand, the elderly usually don't have the same priority as children, and their old age does not allow them to look for survival opportunities like the young, making 'age' a strong predictor. The strong effect of age can be seen in statistical data. Kakde et al. suggest that the 0-10 age group has the survival rate at 53%, which is the greatest, while the 60-70 group has the lowest survival rate at only 23%.

"Fare" is the second most important feature in the prediction. As it is highly correlated with "Pclass", they should be analyzed together. In both Kakde's Kakde (Kakde, 2018) and Bruno's (Frey, 2011) conclusion, passengers traveling in higher class had a better chance of saving themselves compared to passengers traveling in lower class. In fact, there is no direct correlation between "fare" or "Pclass" and survival rate, so the explanation is only speculative:

People who can afford a higher fare live in a higher class, have a better geographical location, have better access to relational information and therefore manage to survive faster than others on board.

"Sex" is significant because women are given priority to survival, as well as children, reflecting the 'women and children first' policy. Kakde et al. also found that people with the title Mrs in their name column had a survival rate of 79%, while people signed Mr had a survival rate of about 16% (Kakde, 2018). There's about a fourfold difference in survival rate between male and female, indicating the importance of "gender" and also that people follow the "women and children first" rule.

In general, features such as "sibsp", "parch" and "embarked" have less weight because they are less relative to survival than age, sex, fare and Pclass. However, these features may be weighted differently in other algorithms.

3.2 Limitations and Future Prospects

In this study, only one algorithm, Random Forest, has been used, which may have resulted in unimportant attributes becoming apparent due to the specificity of the method. As a solution, several machine learning algorithms such as LR, Gradient Boosting, RNN, and SVM should be used and then compared and critically analyzed. Furthermore, a new dataset with a different combination of features could be tried and compared with the previous one, then make a deep dive into how the features affect the predictions.

4 CONCLUSION

In this study, it was found that age, fare and sex were the three most influential features in predicting the survival of Titanic passengers by applying Random Forest to a Kaggle dataset.

The results of this analysis highlight the importance of specific passenger attributes in predicting survival. Recognizing the importance of age, fare and gender can contribute to more accurate forecasts in similar datasets and improve historical analysis. This may have practical implications for improving prediction models in other fields where the importance of characteristics is essential. Furthermore, the use of Random Forests demonstrates the effectiveness of AI algorithms in predicting problems and identifying important patterns in complicated datasets.

However, there are a number of shortcomings in this analysis. Model robustness may be affected by

potential overfitting problems and by focusing on feature importance without considering feature interactions. Future studies could explore other methods for comparative analysis, such as logistic regression or gradient boosting machines. In addition, using a variety of combinations of features, as well as looking at feature interactions, may improve model accuracy and provide a deeper understanding of survival predictions.

REFERENCES

- Careerfoundry, what is Random Forest? 2023. Retrieved from <https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/>
- Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. 2016. Doctor AI: Predicting clinical events via recurrent neural networks. *JMLR Workshop and Conference Proceedings*, 56, 301-318. Epub 2016 Dec 10. PMID: 28286600; PMCID: PMC5341604.
- Frey, B. S., Savage, D. A., & Torgler, B. 2011. Behavior under extreme conditions: The Titanic disaster. *Journal of Economic Perspectives*, 25(1), 209–22. <https://doi.org/10.1257/jep.25.1.209>
- History. Titanic History, 2024. Retrieved from <https://www.history.com/topics/early-20th-century-us/titanic>
- Hsieh, Y.-Z., Su, M.-C., Wang, C.-H., & Wang, P.-C. 2014. Prediction of survival of ICU patients using computational intelligence. *Computers in Biology and Medicine*, 47, 13-19.
- Kakde, Y., & Agrawal, S. 2018. Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. *International Journal of Computer Applications*, 179, 32-38. <https://doi.org/10.5120/ijca2018917094>
- Kaggle, 2017. Titanic Dataset. Retrieved from <https://www.kaggle.com/datasets/heptapod/titanic>
- Pradeep, K. R., & Naveen, N. C. 2018. Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5 and naive Bayes algorithms for healthcare analytics. *Procedia Computer Science*, 132, 412-420.
- Singh, A., Saraswat, S., & Faujdar, N. 2017. Analyzing Titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 406-411). Greater Noida, India: IEEE. <https://doi.org/10.1109/CCAA.2017.8229835>
- Wang, T., Qiu, R. G., & Yu, M. 2018. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. *Scientific Reports*, 8, 9161.