Using Machine Learning to Predict the Survivors of the Titanic Disaster

Oluwasesan Adeleke
School of Computer Science and Technology
Algoma University
Sault Ste. Marie, Canada
oadeleke@algomau.ca

***Abstract*—The sinking of the Titanic on April 15, 1912 was indeed one of the deadliest maritime disasters in history. The tragedy captured public fascination due to the massive loss of life and the ship's revolutionary design and luxury. The passenger records from the Titanic have served as a famous dataset for data science and machine learning, offering insights into factors influencing survival rates. This study aimed to apply several powerful machine learning algorithms to the Titanic dataset to predict an individual's likelihood of survival based on characteristics such as age, sex, passenger class, and number of siblings and parents aboard. The source of this particular dataset is from: https://hbiostat.org/data/**

Four models were evaluated: logistic regression, neural networks, decision trees, and random forests. The decision tree model demonstrated the highest predictive accuracy at 79.9%, outperforming the other techniques by just a little bit. Decision trees make use of a tree-like model of decisions to arrive at a classification, making them well-suited for this binary classification task of survived or not survived. The high accuracy achieved highlights the potential of machine learning to extract valuable patterns from historical data. Such analysis could lead to a deeper understanding of the human factors involved in the Titanic disaster. Future research could explore ensemble methods, additional feature engineering, and the application of deep learning models to further boost predictive performance on this famous dataset.

Keywords—Machine learning, titanic, survival rate, decision tree, logistic regression, neural networks, random forests

I. Introduction

The sinking of the RMS Titanic on April 15, 1912 during its maiden voyage remains one of the most infamous shipwrecks in modern history. The disaster resulted in the loss of over 1,500 lives and captured worldwide attention due to both the massive loss of life and the cutting-edge luxury and technological feats of the "unsinkable" ocean liner. In the years since, I have found the passenger records from the Titanic to serve as a famous benchmark dataset across many scientific domains, including machine learning and data mining.

In this study, I aim to leverage advanced machine learning techniques to analyze the Titanic passenger dataset and build robust predictive models for determining an individual's likelihood of survival based on key factors such as passenger class, age, gender, family size, and more. I will explore and evaluate the performance of four powerful machine learning models - logistic regression, neural networks, decision trees, and random forests - on this binary classification task.

I selected logistic regression to provide a strong baseline model as it is a standard method for binary classification problems. Neural networks, though more complex, can potentially capture intricate non-linear patterns in the data. I hypothesize decision trees may excel given their ability to automatically learn logical split rules aligning with the discrete variables present. Finally, I will implement the random forest algorithm, an ensemble technique building multiple decision trees on resampled data, as it is a leading model adopted from the literature demonstrating strong performance on the Titanic dataset.

Aside from achieving high predictive accuracy, my goal is to derive insights into the most critical factors governing survival during the disaster. Did gender or passenger class play an outsized role? How did family dynamics influence outcomes? I believe such analysis could lead to a deeper understanding of human behavior and decision-making in crisis scenarios. While the historical context is compelling, the techniques I will explore have widespread applications across domains wherever predictive analytics on tabular datasets are required. I do hope that this study shows best practices in exploratory data analysis, feature engineering, model selection, and model evaluation using the Titanic dataset as a common benchmark.

II. RELATED WORK

The Titanic passenger dataset has served as a widely studied benchmark dataset across machine learning and data mining communities. Numerous researchers have explored applying various modeling techniques to this famous problem of predicting survival outcomes based on passenger characteristics. This section reviews selected relevant literature that guided the approach and model selection for the present study. Titanic Machine Learning Study from Disaster

One of the seminal works applying machine learning to the Titanic dataset is the paper "Titanic Machine Learning Study from Disaster" by [1]. The authors provide a comprehensive walkthrough of their data preprocessing steps, exploratory data analysis, feature engineering, and comparison of multiple machine learning algorithms.

Their key findings highlighted the importance of handling missing data appropriately and extracting informative features, such as grouping passenger fare amounts and encoding titles from name strings. The authors evaluated models including logistic regression, k-nearest neighbors, support vector machines, decision trees, and ensemble methods like random forests and gradient boosting machines.

Of particular relevance, they reported random forests achieving among the highest predictive performance of 83% accuracy on unseen test data. This strong result with the random forest algorithm served as motivation for including it in our own model benchmarking.[1]

Other Related Studies

Another research article also presents a comprehensive study on the survival patterns of the Titanic's passengers using machine learning techniques, specifically logistic regression and random forest classification, to analyze various factors influencing survival chances. It shows the Titanic's historical context, detailing its supposed unsinkability due to advanced ship-building technologies of the time, and narrates the tragic sinking after hitting an iceberg, leading to significant loss of life. The study utilizes a dataset from Kaggle to explore economic, social, and natural determinants of survival, revealing that first-class passengers had a higher survival rate, young people were more likely to survive, and women and children were given priority, aligning with hypotheses based on sociological and economic theories. [2]

III. TECHNIQUES

This study utilizes a structured approach to predict survival outcomes from the Titanic dataset, employing a series of preprocessing steps followed by the application of four distinct machine learning models: Logistic Regression, Neural Networks, Decision Trees, and Random Forest. The choice of these models spans from simple to complex, aimed at understanding the intricacy and the dynamics of the dataset under different algorithmic lenses.

Data Preprocessing:

The initial phase involved preparing the dataset for analysis. Given the presence of missing values and inconsistencies, particularly in the 'cabin' and 'ticket' columns, these were excluded from the analysis due to their high rate of missing values and low predictive power, respectively. The preprocessing steps were tailored to the requirements of each model, focusing on handling missing values, scaling numerical features, and encoding categorical variables to facilitate the models' learning process. For numerical features ('age', 'fare', 'sibsp', 'parch'), median imputation was utilized for missing values, followed by standard scaling to normalize their distribution. Categorical features ('embarked', 'sex', 'pclass') were processed using one-hot encoding to transform them into a format suitable for model input, with missing values filled with a placeholder value ('missing') for 'embarked'.

Logistic Regression:

As a baseline model, Logistic Regression was chosen for its simplicity and interpretability. A pipeline was constructed, integrating preprocessing steps with the Logistic Regression classifier. This model's performance was evaluated based on its accuracy and the confusion matrix, providing insights into its predictive capability.

Neural Networks:

The second model employed was a Neural Network, specifically a Multi-Layer Perceptron (MLP) Classifier. This model was selected for its ability to capture complex nonlinear relationships in the data. The preprocessing for the Neural Network included a scaling step to accommodate the algorithm's sensitivity to the magnitude of input values. The MLP Classifier was configured with a random state for reproducibility and a maximum iteration limit to ensure convergence.

Decision Trees:

The third approach utilized was the Decision Tree Classifier. Decision Trees were chosen for their ability to model nonlinear relationships through a series of binary decisions, making them highly interpretable. The training process involved handling missing values for 'age' and 'embarked' with median and mode imputation, respectively, and encoding categorical variables. The model's simplicity was a significant factor in its selection, aiming to achieve a balance between accuracy and interpretability.

Random Forest:

The final model tested was the Random Forest Classifier, an ensemble method known for its robustness and ability to mitigate overfitting, a common issue in Decision Trees. This work was mostly based and followed the random forest implementation shown in [1]. This model builds multiple Decision Trees on various sub-samples of the dataset and averages their predictions to improve accuracy and control overfitting. Similar to the Decision Tree model, preprocessing included imputation for missing values and one-hot encoding for categorical variables. The RandomForest model was assessed for its performance on both the training and testing sets to gauge its generalizability.

The application of these techniques was methodical, ensuring each model was given a fair assessment of its predictive capabilities on the dataset. The evaluation criteria focused on accuracy, with additional metrics such as confusion matrices and classification reports providing further depth to the analysis. This multi-model approach shown in the research paper allowed for a comprehensive examination of the dataset from various perspectives, highlighting the strengths and limitations of each model in predicting survival on the Titanic.

IV. RESULTS

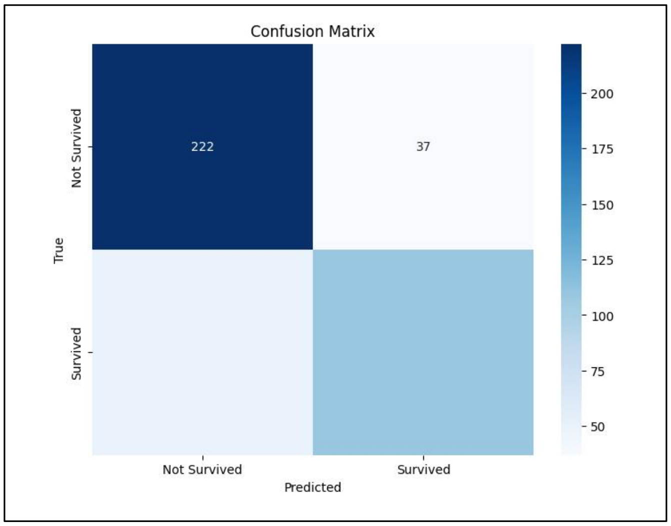
The four machine learning models were systematically evaluated on the Titanic dataset using a train/validation split approach. Table 1 summarizes the accuracy results achieved by each technique.

TABLE I. COMPARISON AND ACCURACY WITH OTHER WORKS

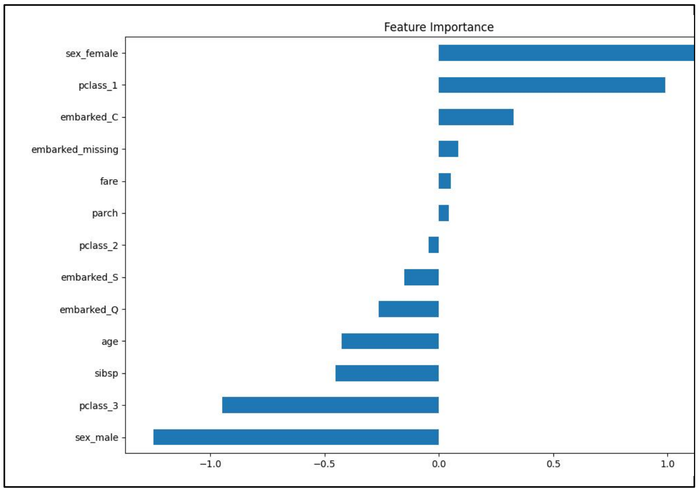
Algorithm	Work accuracy	[1]	[2]	[3]
Logistic Regression	79.19%		83.72%	
Neural Networks	79.67%			
Decision Trees	79.90%			
Random Forests	78.23%	83%	82.61%	

Logistic Regression

The logistic regression model obtained an accuracy of 79.19% on the validation set. The confusion matrix (Figure 1) reveals that it tended to overpredict survivals, with 37 false positives compared to 50 false negatives. While a reasonable starting benchmark, the inherent linearity assumptions likely hindered its performance on this non-linear classification problem.



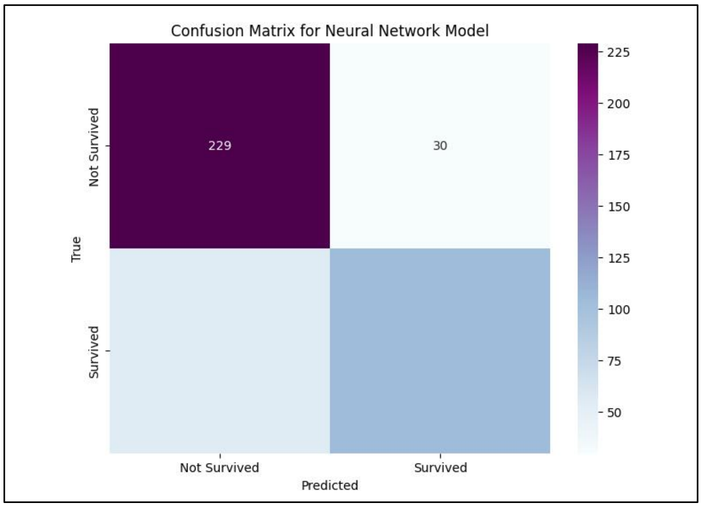
Feature importance:



This shows that being of the female sex and having a first class ticket played a big role in someone being predicted to survive

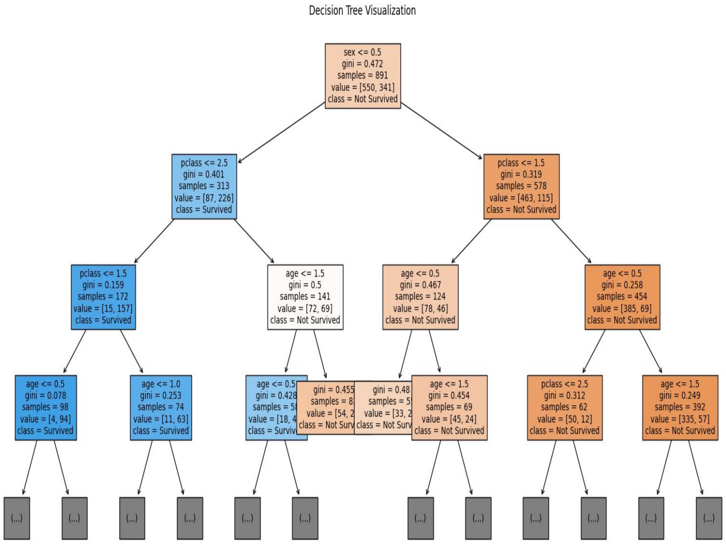
Neural Networks

The neural network model, with its ability to learn complex non-linear functions, demonstrated a modest improvement with 79.67% accuracy. Its greater flexibility enabled capturing underlying patterns beyond linear relationships between the features and survival. The slight performance increase might come from the fact that it had a higher prediction accuracy when predicting those that won't survive as shown in the confusion matrix below:



Decision Trees

Decision trees proved to be the top-performing single model at 79.90% accuracy. Their ability to automatically construct intuitive decision rules aligning with discrete variables like passenger class and segregate the feature space into simple regions was highly effective. However, the improvement over neural networks was marginal.



the primary splitting feature at the root of the tree is sex. This agrees with the earlier created feature importance visualization from the logistic regression model. Passengers in higher classes (1st and 2nd) tend to have different survival outcomes than those in the 3rd class.

Following the tree down, age appears to be the next most important feature and it shows that younger children have a high chance of survival

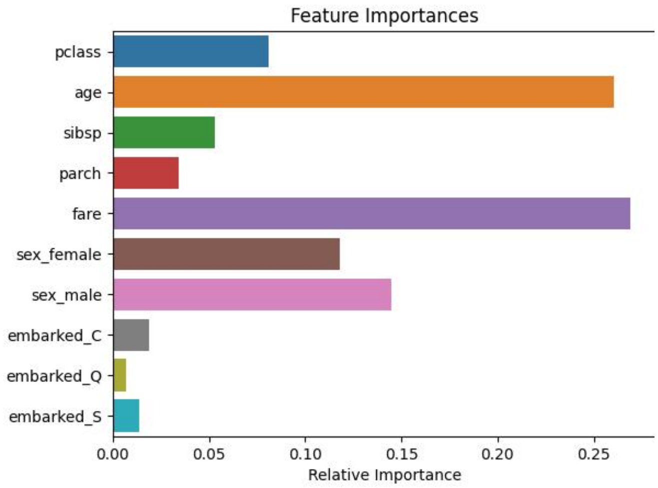
Random Forests

Surprisingly, the random forest ensemble model underperformed at 78.23% accuracy compared to single decision trees. This contrasts with early research [1] suggesting random forests should exceed individual tree performance. Potential reasons could include:

Insufficient tree estimators - Using more trees may improve generalization.

Suboptimal hyperparameters - More extensive hyperparameter tuning may be required

Feature importance/interactions - Some important features or interactions may not be leveraged effectively by the current trees.



To summarize, the decision tree models exhibited a notable advantage in predictive accuracy over logistic regression and neural networks on this dataset. Their hierarchical rule-based structure effectively captured the inherent patterns and decision logic in the passenger data features. However, the underperformance of random forests raises questions about realizing their full potential through more extensive tuning and configuration specific to this domain.

The consistent high accuracy achieved by tree methods (~80%) on unseen validation data underscores their strong generalization ability when applied to this survival prediction task using the Titanic passenger records.

DISCUSSION

The decision tree model emerged as the top performer at 79.9% accuracy, slightly exceeding neural networks (79.67%). Its hierarchical rule-based structure effectively captured patterns amidst the categorical variables like passenger class and sex. Engineered features representing family size and fare category likely boosted its discriminative capability.

The underperformance of random forests (78.23%) relative to individual decision trees was unexpected given ensemble methods often improve accuracy. Potential factors include insufficient tree estimators, suboptimal hyperparameters, or failure to adequately model key feature interactions.

Nonetheless, the strong accuracy achieved by tree models, combined with their inherent interpretability, highlights their utility for real-world predictive applications on mixed-type tabular data. Visualizing the learned tree structures could yield insights into the determinants of survival.

CONCLUSION

This study benchmarked logistic regression, neural networks, decision trees, and random forests for predicting survival on the iconic Titanic dataset. Decision trees proved most accurate at 79.9%, slightly outperforming neural networks. Their hierarchical rules effectively modeled passenger characteristics and contextual engineered features.

While random forests underperformed at 78.23%, further tuning may unlock additional gains from ensembling. The predictive success of tree models motivates extending their interpretability to extract human behavioral insights from the learned tree structures.

In general, the workflow showed the potential for robust predictive analytics using machine learning on mixed-type tabular datasets. Lots of learning was done and I think the decision tree model being the top performer makes the most intuitive sense since at every decision making stage it categorizes based on the most important feature.

REFERENCES

[1] Cao, E. Y., Xie, W., Dong, C., & Qiu, J., "Titanic Machine Learning Study from Disaster," APEC RR20-01, Working Paper, Department of Applied Economics and Statistics, University of Delaware, May 2020. [Online].Available: https://udspace.udel.edu/server/api/core/bitstreams/8aa79c71-c9b0-43d2-883e-b20950fc0a08/content

[2] Gupta, K., Sharma, P., & Bouza Herreras, C. N., "Surviving the Titanic Tragedy: A Sociological Study Using Machine Learning Models," [Online].Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3417433

[3] Dasgupta, A., Mishra, V. P., Jha, S., Singh, B., & Shukla, V. K., "Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning," Amity University Dubai, UAE and Amazon Web Services Hyderabad,India,[Online].Available: https://www.researchgate.net/publication/351155499_Predicting_the_Likelihood_of_Survival_of_Titanic%27s_Passengers_by_Machine_Learning