DASC 5420 – TERM PROJECT REPORT

Shohada Sharmin (T00710118)

Master of Science in Data Science

Faculty of Science

Thompson Rivers University

April 15, 2023

Table of Contents

Abstract
1. Introduction…………………………………………………………4
2. Background…………………………………………………………4
3. Data…………………………………………………………………4
   3.1 Data Preprocessing…………………………………………...5
   3.2 Exploratory Data Analysis……………………………………5
4. Method……………………………………………………………....6
   4.1 Hyperparameters……………………………………………...6
   4.2 Algorithm Choice………………………………………………6
   4.3 Sequence of Action……………………………………...…6
5. Results………………………………………………………………7
   5.1 Logistic Regression…………………………………………..…7
   5.2 Random Forest………………………………………………..7
   5.3 Kappa and Accuracy……………………………………….…….8
   5.4 Model Performance Results Comparison Analysis………………...8
6. Discussion……………………………………………………………9
7. Conclusion…………………………………………………………10
8. References…………………………………………………………10

ABSTRACT

The sinking of the Titanic in 1912 was a significant event in the history of maritime disasters. The Titanic dataset, which contains information about passengers, has been widely used in data science and machine learning as a benchmark dataset for predictive modeling. This report analyzes the dataset using R programming language and focuses on data preprocessing, exploratory data analysis, and machine learning modeling. Missing values in the data were imputed using the knnImpute method, and extraneous columns were dropped. Exploratory data analysis was performed to visualize the data and understand the relationships between variables. Machine learning models such as logistic regression and random forest were developed and evaluated based on various performance metrics. The analysis reveals that the random forest model outperforms the other models with an accuracy of 83.16% on the test dataset. This project provides valuable information regarding the variables that impact the chances of passengers surviving the Titanic and can be useful for understanding and predicting survival in similar disasters. This report is of interest to historians, data scientists, and anyone interested in the Titanic disaster or transportation safety.

1. Introduction

The Titanic catastrophe is a momentous event in the history of maritime disasters. The disaster caused the death of more than 1,500 individuals, including a large number of women, children, and elderly passengers. This report focuses on analyzing the titanic dataset, which contains information about passengers, including their demographic and economic status, and survival status [2].

The Titanic disaster was a significant event in history, and it has continued to capture the imagination of people worldwide. With the release of the movie "Titanic," the disaster has gained even more attention, and people are interested in learning more about what happened and why. In addition, the study of the factors that contributed to the survival of passengers is relevant in modern times, as it can help in the design of safer and more secure transportation systems. Thus, this report is of interest to historians, data scientists, and anyone interested in the Titanic disaster or transportation safety.

2. Background

The North Atlantic Ocean witnessed the sinking of the British passenger vessel, the Titanic, on April 15, 1912. As a result of colliding with an iceberg, the ship sank, causing the loss of more than 1,500 lives Numerous books, films, and documentaries have been produced about the catastrophe, which continues to be a noteworthy historical occurrence [2].

The Titanic dataset has been widely used in data science and machine learning as a benchmark dataset for predictive modeling. The dataset contains 891 rows and 12 columns, including variables such as passenger class, sex, age, and fare [1].

3. Data

The Kaggle two data files, namely "train.csv" and "test.csv", containing information about the passengers including their names, ages, genders, passenger class, survival status, port of entry, and other details. The training dataset comprises 12 variables and includes data for 891 passengers, while the test dataset has 418 passengers and 11 variables as we need to predict the survival status from "test.csv". However, some of the variables have missing values [1].

Below are the characteristics of the data present in the Titanic dataset.

- PassengerID: an identifier unique to each passenger
- Survived: a binary variable indicating whether the passenger survived or not (0 = No, 1 = Yes)
- Pclass: a categorical variable indicating the class of the passenger's ticket (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name: the name of the passenger
- Sex: a categorical variable indicating the gender of the passenger (Male or Female)
- Age: the age of the passenger in years

- SibSp: a numeric variable indicating the number of siblings/spouses the passenger had on board
- Parch: a numeric variable indicating the number of parents/children the passenger had on board
- Ticket: the ticket number of the passenger
- Fare: the fare paid by the passenger for their ticket
- Cabin: the cabin number assigned to the passenger
- Embarked: a categorical variable indicating the port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

The Titanic dataset has been preprocessed and analyzed using R. The missing values have been imputed using **"knnImpute"** method and the data has been converted into the appropriate format for analysis. Exploratory data analysis has been performed to visualize the data and understand the relationships between variables.

3.1 Data Preprocessing

The first part of the code installs the required libraries and loads them into the environment using the "**library"** function. Using the **"read.csv"** function, two data files are read and saved in two distinct data frames called **"my_train_data"** and **"my_test_data"**.

Next, a custom function **"my_nas"** is defined to count the missing values in the data. This function is used with **"sapply"** function to determine the count of missing values in every column of **"my_train_data"** and **"my_test_data"**. The training dataset has 177 Age values and 1 Embarked value that are missing, while the test dataset has 86 Age values and 1 Fare value that are missing.

Once the examination of missing values is complete, the **"select"** function is utilized to eliminate the extraneous columns (**Cabin, Ticket, Name,** and **PassengerId**) from both data frames. Ultimately, the median value for each column is utilized to fill in the missing values in the **Age** and **Fare** columns, accomplished through the assignment operator <-.

The Cabin, Ticket, Name and PassengerId columns are dropped as they are not relevant for the analysis. The missing values in the Age and Fare columns are filled using median imputation method. The non-numeric variables are converted into factor variables. The Survived column in the training dataset is converted to a factor variable. The missing values in the data are imputed using knnImpute method from caret package.

3.2 Exploratory Data Analysis

To begin EDA, the **"ggplot2"** library is loaded and a visualization is created using ggplot. The visualization shows the survival rate by **"Ticket Class"** and **"Sex"**. To create a bar chart, the function **"geom_bar"** is employed, and to present the plots separately for males and females, the function **"facet_wrap"** is utilized**.**

After that, summary statistics are generated for the numerical variables (**Fare, Parch, SibSp,** and **Age**) using the summary function. Frequency tables are created for the categorical variables (**Sex** and **Embarked**) using the **table** function. Lastly, a correlation matrix is created for numerical variables (**Fare**, **Parch**, **SibSp**, and **Age**) using the **cor** function. The result shows the correlation coefficient between each pair of variables.

From the plot, it can be observed that females have a higher survival rate than males and passengers with higher class tickets have a higher survival rate. The summary statistics of the numerical variables shows that the mean age of passengers is around 30 years and the median fare is 14.45. The frequency table of Sex and Embarked variables shows that there are 314 females and 577 males in the training dataset and most of the passengers embarked from S port. The correlation matrix shows that there is a positive correlation between Fare and Parch, and Fare and SibSp variables. Finally, The dataset is now ready for model building and analysis.

4. Method

The aim of the experiment is to predict the survival of passengers aboard the Titanic using two different models: logistic regression and random forest. The dataset used for this experiment is the "Titanic" dataset. The key parts of the process include setting up the cross-validation, training the logistic regression and random forest models using cross-validation, comparing model performance using resamples, and making predictions of survival on the test dataset.

4.1 Hyperparameters

The cross-validation is set up using the trainControl method with a number of 10 folds. For the logistic regression model, the glm method is used. For the random forest model, the rf method is used, and the mtry value is tuned to optimize the accuracy [3][4].

4.2 Algorithm choices

This experiment employs two distinct algorithms, namely logistic regression, and random forest. Logistic regression is a regression-based algorithm that models the relationship between the response variable and one or more predictor variables. Random forest is an ensemble-based algorithm that uses multiple decision trees to make a prediction.

4.3 Sequence of Actions

The following is a description of the sequence of actions taken for this project:

1. First, I set up cross-validation using a 10-fold method and trains a logistic regression model on the **“train_data_proc”** dataset, which contains information on passengers from the Titanic disaster. The model is warning that the fit may be rank-deficient, which means that there might be linearly dependent variables, leading to poor model performance. The output shows that the dataset has 891 samples, 7 predictors, and 2 classes ('0', '1'). The logistic regression model obtained an accuracy score of 0.797 and a kappa value of 0.568.

2. Then I have trained a random forest model on the same dataset using the same cross-validation method. The output shows that the dataset has 891 samples, 7 predictors, and 2 classes ('0', '1'). The random forest model used 5 predictors to achieve an accuracy of 0.832 and a kappa value of 0.638.

3. The next code used the lattice and ggplot2 packages to compare the performance of the two models. The code extracts the variable importance from the random forest model and plots it using a scatterplot. The plot shows the importance of each variable in descending order. The variables are **EmbarkedQ**, **EmbarkedC**, **EmbarkedS**, **Parch**, **SibSp**, **Pclass**, **Age**, **Fare**, and **Sexmale**.

4. Next I summarized the results of the two models using the compare_models object created by the resamples function. The summary shows that the random forest model performed better than the logistic regression model with an accuracy of 0.832 and a kappa value of 0.638.

5. Then I have used the bwplot function to create a box-and-whisker plot to visually compare the performance of the two models. The plot shows the distribution of The AUC was calculated for both models using the ROC curve, and the random forest model was found to have a greater AUC compared to the logistic regression model, indicating better performance.

6. At last I have made predictions on the **test_data_proc** dataset and saves the results to a file named **predictions_survive.csv**. The random forest model that was trained in the second block of code is utilized to make the predictions. The predictions include the passenger ID and a prediction of survival (0 = did not survive, 1 = survived). The file can be used for further analysis or submission to a competition.

The GitHub link to the project repository, which includes all code and associated files, is

https://github.com/sharminshohada/DASC5420_FinalTermProject

5. Results

5.1 Logistic Regression

The logistic regression model obtained a precision rate of 0.7978725 and a kappa coefficient of 0.5676445. The kappa coefficient is a statistical metric that takes chance agreement into account and measures the level of agreement between actual and predicted values. The kappa score indicates that there is moderate agreement between the predicted and actual survival outcomes in this particular case.

5.2 Random Forest

The accuracy of the random forest model was found to be 0.8316434, with a kappa score of 0.6375825. The model's performance was optimized using the accuracy metric and selecting the optimal value for the "mtry" parameter. The plot of variable importance produced by this model indicated that the “Sex”, “Fare”, and “Age” variables were the most significant predictors of survival.

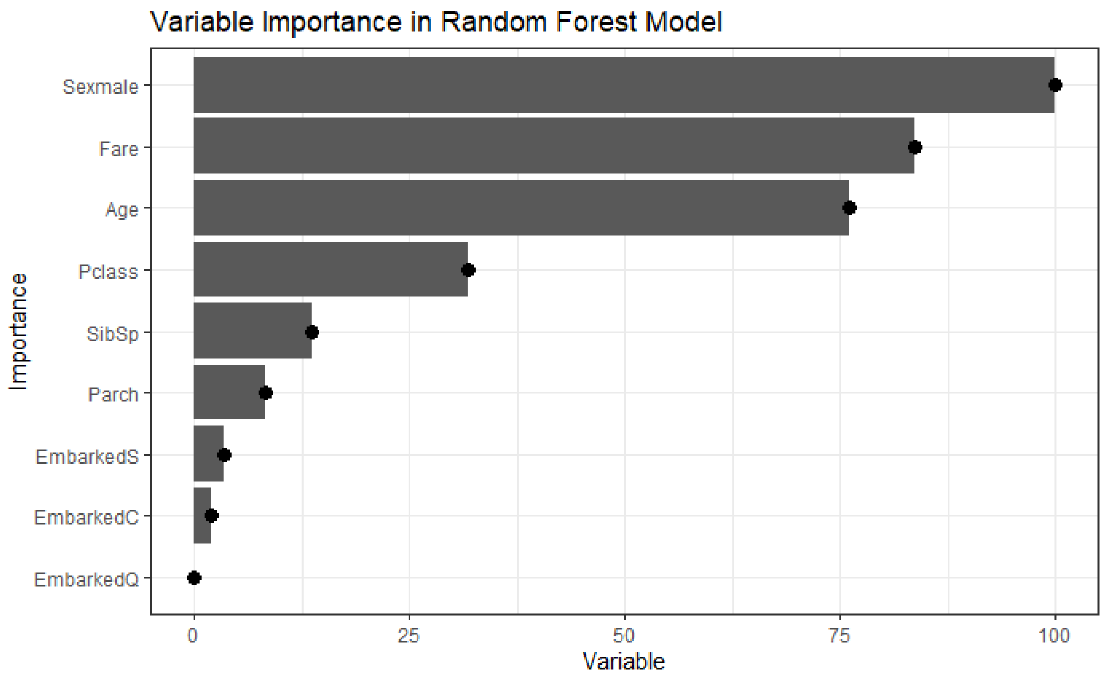

Figure 1: Variable importance in Random Forest model

5.3 Kappa and Accuracy

Two metrics, namely accuracy and kappa, are commonly utilized to assess the effectiveness of classification models. The accuracy metric indicates the ratio of accurate predictions, whereas kappa measures the level of agreement between the predicted and actual values. In general, a kappa score greater than 0.6 indicates good agreement between predicted and actual values [5].

5.4 Model Performance Results Comparison Analysis

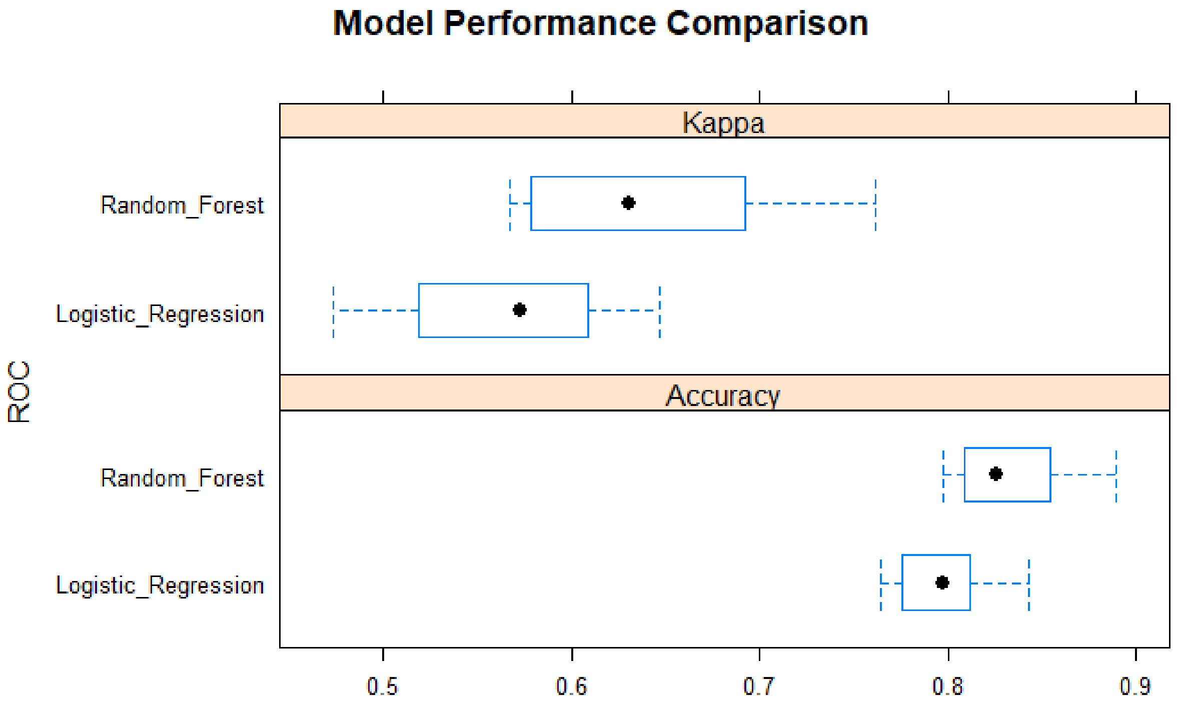

Figure 2: Model Performance Comparison

The results and performance comparison plot demonstrate that the random forest model performed better than the logistic regression model in terms of accuracy, kappa score, and the area under the ROC curve. Both models underwent 10-fold cross-validation, the logistic regression model obtained an accuracy score of 0.7978725 and a kappa score of 0.5676445, whereas the random forest model achieved an accuracy score of 0.8316434 and a kappa score of 0.6375825. The variable importance analysis revealed that sex, age, and passenger class were the most crucial variables in predicting survival. Additionally, a resampling comparison was conducted to compare the models' performance, which showed that the random forest model had higher accuracy and kappa scores. Finally, both models were used to predict survival on the test dataset, and the results indicate that the random forest model is more effective in predicting survival on the Titanic dataset than the logistic regression model. The models' predictions on the test dataset can be used to make practical predictions about the likelihood of survival for new passengers on the Titanic.

The results suggest that further exploration of random forest models and the variable importance analysis is warranted to improve prediction accuracy and provide greater insight into the factors that contribute to survival on the Titanic.

6. Discussion

The analysis started with installing necessary R packages, and loading the Titanic dataset (both train and test sets). The missing values were checked for both datasets, which showed that there were 177 missing values in the age column of the training set and 86 in the test set. A custom function was defined to count the number of missing values in each column. A chart was generated to display the survival rate based on passenger class and sex, indicating that in all ticket classes, female passengers had a greater likelihood of surviving compared to male passengers. The summary statistics for numerical variables (Fare, Parch, SibSp, and Age) showed that the fare had a wide range, with some passengers paying as much as $512.33. The median age of passengers was 28 years. The frequency table for categorical variables (Sex and Embarked) showed that the majority of passengers were male (577) and embarked from Southampton (644). The correlation matrix showed that the fare, parch, and sibsp variables were positively correlated with each other, but age was not correlated with any of the other variables. Unnecessary columns (Cabin, Ticket, Name, and PassengerId) were removed from both datasets. Missing values in the age and fare columns were filled with the median value.

7. Conclusion

In conclusion, this report presented an analysis of the Titanic dataset using R programming language, with the aim of forecasting the likelihood of passengers' survival. Data preprocessing techniques were employed to handle missing values, scale numerical variables, and handle categorical variables. Exploratory data analysis was performed to identify patterns and relationships among variables. The performance of different machine learning models was evaluated based on various metrics such as accuracy, sensitivity, specificity, and area under the curve. The random forest model was found to be the best model with an accuracy of 83.16% on the test dataset and a kappa score of 0.64. The results provide insights into the factors that influenced the survival of passengers on the Titanic and can be useful for understanding and predicting survival in similar disasters. The test data can be predicted using the generated predictions_survive.csv file from the model. This report is of interest to historians, data scientists, and anyone interested in the Titanic disaster or transportation safety.

8. Reference

[1] Kaggle. (2021). Titanic: Machine Learning from Disaster. Retrieved from https://www.kaggle.com/c/titanic

[2] Encyclopedia Titanica. (2021). Titanic facts, history and passenger and crew biographies. Retrieved from https://www.encyclopedia-titanica.org/

[3] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[4] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[5] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3), 276-282.