Advanced Data Analysis

Titanic Dataset Model

IJ Joubert

Dr. MP Venter

2019

Plagiarism declaration

I have read and understand the Stellenbosch University Policy on Plagiarism and the definitions of plagiarism and self-plagiarism contained in the Policy [Plagiarism: The use of the ideas or material of others without acknowledgement, or the re-use of one's own previously evaluated or published material without acknowledgement or indication thereof (self-plagiarism or text-recycling)].

I also understand that direct translations are plagiarism, unless accompanied by an appropriate acknowledgement of the source. I also know that verbatim copy that has not been explicitly indicated as such, is plagiarism.

I know that plagiarism is a punishable offence and may be referred to the University's Central Disciplinary Committee (CDC) who has the authority to expel me for such an offence.

I know that plagiarism is harmful for the academic environment and that it has a negative impact on any profession.

Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully (acknowledged); further, all verbatim copies have been expressly indicated as such (e.g. through quotation marks) and the sources are cited fully.

I declare that, except where a source has been cited, the work contained in this assignment is my own work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

I declare that have not allowed, and will not allow, anyone to use my work (in paper, graphics, electronic, verbal or any other format) with the intention of passing it off as his/her own work.

I know that a mark of zero may be awarded to assignments with plagiarism and also that no opportunity be given to submit an improved assignment.

Signature:

Name: Student no:

Date:

Table of contents

[Plagiarism declaration i](#_Toc4499205)

[Table of contents ii](#_Toc4499206)

[List of figures iii](#_Toc4499207)

[List of tables iv](#_Toc4499208)

[1 Introduction 1](#_Toc4499209)

[1.1 Description 1](#_Toc4499210)

[1.2 Background 1](#_Toc4499211)

[1.3 Evaluation 1](#_Toc4499212)

[2 Methodology 2](#_Toc4499213)

[2.1 Data 2](#_Toc4499214)

[2.2 Data Preparation 2](#_Toc4499215)

[2.2.1 Data Overview 2](#_Toc4499216)

[2.2.2 Data Cleaning 5](#_Toc4499217)

[2.2.3 Feature Engineering 6](#_Toc4499218)

[2.3 Predictions 9](#_Toc4499219)

[3 Conclusions 11](#_Toc4499220)

[4 References 12](#_Toc4499221)

List of figures

[Figure 1: Data scatter matrix 3](#_Toc4498213)

[Figure 2: Survived data scatter matrix 4](#_Toc4498214)

[Figure 3: Title survival rates 7](file:///C:\Users\Zack\Google%20Drive\Studies\Meesters\Advanced%20Data%20Analysis\ADA-PG\Kaggle\Titanic\Report.docx#_Toc4498215)

[Figure 4: Family survival rate 8](file:///C:\Users\Zack\Google%20Drive\Studies\Meesters\Advanced%20Data%20Analysis\ADA-PG\Kaggle\Titanic\Report.docx#_Toc4498216)

List of tables

[Table 1: Data categories 2](#_Toc4499224)

[Table 2: Grouped data 5](#_Toc4499225)

# Introduction

## Description

The titanic challenge on Kaggle concerns the classification of the RMS Titanic passenger survival. A dataset with actual recorded data is provided and it is the task of the participant to classify the test dataset into surviving- and not passengers.

## Background

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. [1]

## Evaluation

The scoring for this competition is based on the percentage of accurate predictions. Submissions are made to the Kaggle website where they are evaluated and the score is then shown next to the submission.

# Methodology

## Data

Two datasets are provided. A training set with correct survival predictions and a test set without any survival predictions.

The dataset categories are as follows:

Table 1: Data categories

|  |  |  |
| --- | --- | --- |
| Category | Description | Data type |
| PassengerId | A numerical index of passengers (in no particular order) | Integer |
| Survived | A binary category specifying if the individual lived or died (0 = died, 1 = lived) | Float |
| Pclass | The ticket class of the passenger (1 = first class, 2 = second class, 3 = third class) | Integer |
| Name | The surname, title and name of the passenger (also includes some maiden names) | String |
| Sex | The sex of the passenger (male or female) | String |
| Age | The age of the passenger | Float |
| SibSp | The number of siblings or spouses aboard the Titanic | Integer |
| Parch | The number of children or parents aboard the Titanic | Integer |
| Ticket | The passenger’s ticket number | Sring |
| Fare | The fare paid for said ticket | Float |
| Cabin | The cabin allocated to the passenger | String |
| Embarked | At which port the passenger embarked (C = Cherbourg, Q = Queenstown, S = Southampton) | String |

## Data Preparation

### Data Overview

Firstly, a quick data overview to make a preliminary verdict on which data is most important to an accurate prediction.

The following is a scatter matrix plotting all of the data categories:

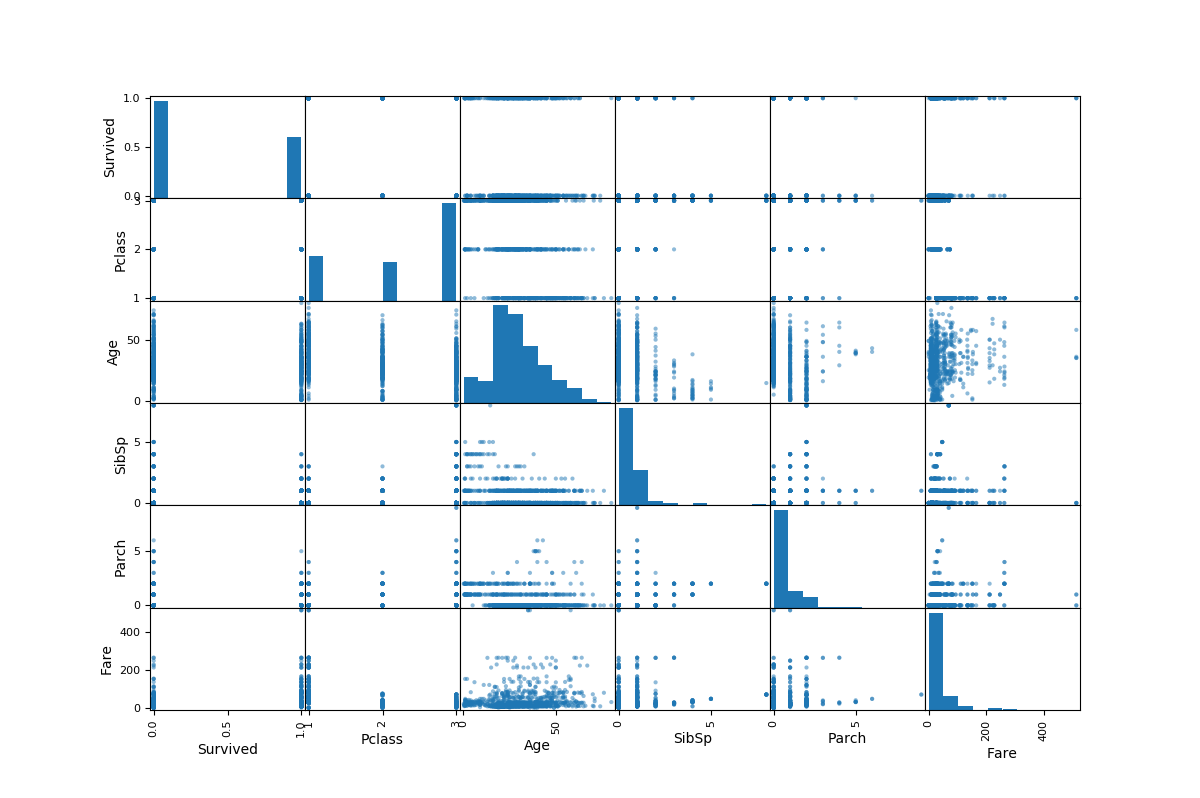


Figure 1: Data scatter matrix

Here we can see why the Titanic classification problem is not a trivial one. At first glance there are no obvious correlations between survival and any of the categories.

Let us rather plot a scatter matrix with Survival statistics:



Figure 2: Survived data scatter matrix

Now there are some more apparent correlations. We can see that Pclass is an important category to classify survival.

### Data Cleaning

As with most datasets, there are missing values in the Titanic dataset.

<class 'pandas.core.frame.DataFrame'>

Int64Index: 1309 entries, 1 to 1309

Data columns (total 11 columns):

Survived 891 non-null float64

Pclass 1309 non-null int64

Name 1309 non-null object

Sex 1309 non-null object

Age 1046 non-null float64

SibSp 1309 non-null int64

Parch 1309 non-null int64

Ticket 1309 non-null object

Fare 1291 non-null float64

Cabin 295 non-null object

Embarked 1307 non-null object

dtypes: float64(3), int64(3), object(5)

memory usage: 122.7+ KB

Here we see that Age is missing 263 values, Survived is missing the test data’s values, which will be predicted, Fare is missing 18 values, Cabin is missing 1014 values and Embarked is missing 2 values.

#### Age

Age is not our most important feature, as seen from the scatter plot, therefore we will simply compute a grouped mean.

The data is grouped by Sex, Pclass and Title as follows:

Table 2: Grouped data

| Sex | Pclass | Title | Age |
| --- | --- | --- | --- |
| female | 1 | Miss | 30 |
|  |  | Mrs | 45 |
|  |  | Rare | 43.5 |
|  | 2 | Miss | 20 |
|  |  | Mrs | 30.5 |
|  | 3 | Miss | 18 |
|  |  | Mrs | 31 |
| male | 1 | Master | 6 |
|  |  | Mr | 41.5 |
|  |  | Rare | 49.5 |
|  | 2 | Master | 2 |
|  |  | Mr | 30 |
|  |  | Rare | 41.5 |
|  | 3 | Master | 6 |
|  |  | Mr | 26 |

When a passenger’s age is missing, their age will be computed based firstly on their sex, then their class, and finally their title. The age values displayed in the table above represents the median of the age values for the preceding categories

#### Fare

This same approach is followed for the missing Fare values. Due to the fact that there are only 18 missing values this was deemed adequate.

#### Cabin

The cabin feature supplies some cabin levels and numbers i.e. C25 is cabin level 3 and cabin number 25. However, there are mostly null values for this category. Therefore this category was dropped

#### Embarked

The embarked feature was also deemed less important and the two missing values will simply filled by the most common value i.e. ‘S’

### Feature Engineering

In order to increase the complexity and accuracy of the model some features were engineered.

#### Title

Title, previously mentioned, is each passenger’s title i.e. Mr, Mrs, etc.

The common titles include Mr, Mrs and Miss. The Master title refers to boys. There are several other miscellaneous titles including Captain, Colonel and Don but these were simply converted to ‘Rare’ titles.

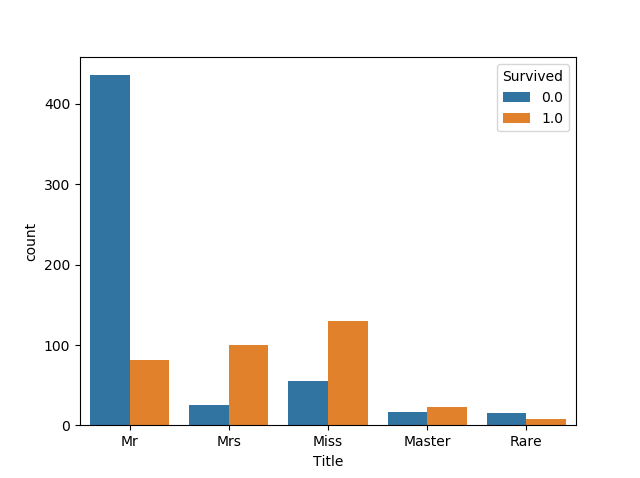


Figure 3: Title survival rates

Here we can already start making assumptions i.e. most of the passengers with title Mr died also, female title passengers tend to survive. This corresponds to the RMS Titanic policy of women-and-children first.

#### SexGroup

This feature is created to help with additional feature engineering. It is a further classification of the ‘Sex’ category and splits each passenger into man, woman or boy.

#### Surname

The surname feature is also created to help with additional feature engineering. It extracts each passenger’s surname from their ‘Name’ category.

#### FamilySize

Some passengers were traveling together with family. This feature simply computes the size of each family.

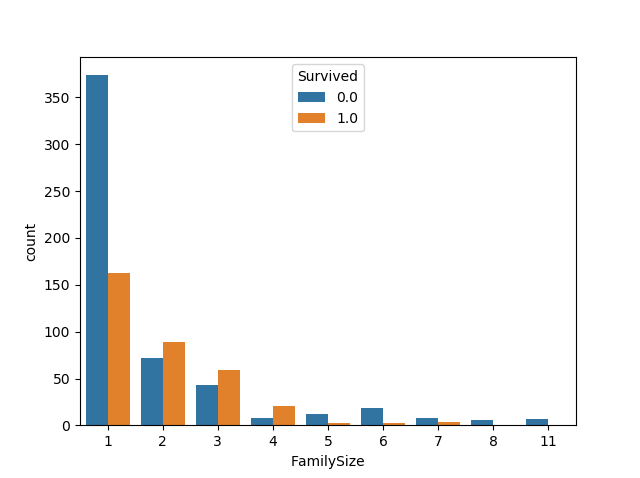


Figure 4: Family survival rate

Here we see that single passengers (FamilySize = 1) were less likely to survive compared to passengers with 2 to 4 family members. However, families larger than 5 were more likely to perish.

#### Groups

As previously mentioned, some passengers were traveling in groups. This group data can be used to improve the accuracy of predictions. In order to improve predictions, men were excluded from these groups as there is little correlation between any of the data and the male survival rate. These predictions will be discussed later

##### GroupID

A very important feature, GroupID, creates groups of passengers traveling together by combining each passenger’s Surname, Class, Ticket number (without the last two digits as some group members had consecutive tickets), Fare and Embarked. This created groups like Palsson-3-3499-21.075-S or Johnson-3-3477-11.1333-S.

##### GroupSurvivalRate

The group survival rate is calculated for the training dataset and then carried over to the applicable groups in the test dataset.

A group’s survival rate is calculated by taking the number of passengers that survived in that group and dividing it by the total number of passengers in the group. This calculation is only done for the passengers in the training dataset thus if a group has one passenger in the training dataset and one passenger in the test dataset, the total group number is two but the group survival rate is calculated only with the training dataset’s passenger’s survival as illustrated:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Survived | Name | GroupID | GroupSurvived | GroupSurvivalRate |
| 1 | Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson) | Asplund-3-3470-31.3875-S | 3 | 0.75 |
| 0 | Asplund, Master. Clarence Gustaf Hugo | Asplund-3-3470-31.3875-S | 3 | 0.75 |
| 1 | Asplund, Miss. Lillian Gertrud | Asplund-3-3470-31.3875-S | 3 | 0.75 |
| 1 | Asplund, Master. Edvin Rojj Felix | Asplund-3-3470-31.3875-S | 3 | 0.75 |
|  | Asplund, Master. Filip Oscar | Asplund-3-3470-31.3875-S |  | 0.75 |
|  | Asplund, Master. Carl Edgar | Asplund-3-3470-31.3875-S |  | 0.75 |

Figure 5: Passenger group survival rate example

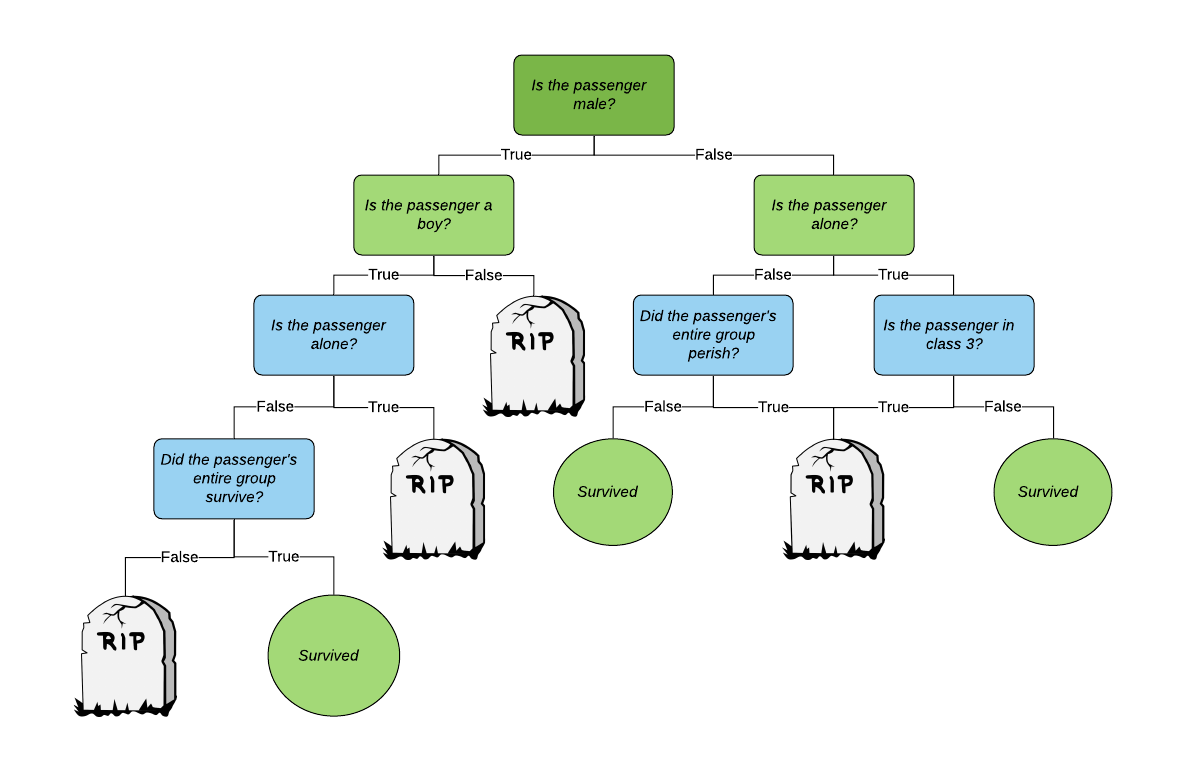
#### Miscellaneous

Due to the high mortality rate of men aboard the RMS Titanic, a fairly accurate assumption to make is that all men perished. This assumption is applicable to this model.

## Predictions

A decision tree classifier was implemented due to the fairly straightforward nature of the passenger survival prediction, taking into account the above-mentioned assumption.

The decision tree is depicted below:



# Conclusions

With this approach, an accuracy of 83.253% was reached.

Therefore, the assumption that all men perish is a fairly accurate one. Further classification of male survivors can be implemented however, the data does not provide a strong correlation or pattern between the data and the surviving male passengers.

The accurate prediction of male survivors requires a more sophisticated model which was deemed unnecessary as the above obtained score places one in the top 4%.

# References

|  |  |
| --- | --- |
| [1] | K. Inc, “Titanic: Machine Learning from Disaster,” [Online]. |