A

Project Report

On

Predictive analysis of the Titanic's mortality risk

**Table of Contents**

[Abstract: 3](#_Toc121514578)

[Introduction: 3](#_Toc121514579)

[Objective: 3](#_Toc121514580)

[About the collection of Dataset: 3](#_Toc121514581)

[Methodology: 4](#_Toc121514582)

[Implementation: 4](#_Toc121514583)

[Import the libraries,datasets and read the dataset 4](#_Toc121514584)

[Data Exploration/Analysis 5](#_Toc121514585)

[Data Preparation and Exploratory Analysis 5](#_Toc121514586)

[Data Cleaning 5](#_Toc121514587)

[Exploratory Analysis and Data Processing 10](#_Toc121514588)

[Age vs Survival 10](#_Toc121514589)

[Sex vs Survival 12](#_Toc121514590)

[Pclass vs Survival 13](#_Toc121514591)

[Modeling 14](#_Toc121514592)

[Feature Engineering 14](#_Toc121514593)

[Model Training 16](#_Toc121514594)

[Model Evaluation 20](#_Toc121514595)

[Prediction 21](#_Toc121514596)

[Conclusion: 21](#_Toc121514597)

[References: 22](#_Toc121514598)

# 

# **Abstract:**

In this project, we will implement predictive analysis of titanic survival rate in R.

In this data analytics project, we will explore the data upon which we will be building our segmentation model. Also, in this data science project, we will see the descriptive analysis of our data and then implement several modeling techniques.

# **Introduction:**

One of history's most well-known shipwrecks is the RMS Titanic's sinking. Out of 2224 passengers and crew, 1502 died when the Titanic sank on April 15, 1912, during her maiden voyage after striking an iceberg. The international society was stunned by this shocking catastrophe, which prompted improved ship safety rules.

The lack of lifeboats for the passengers and crew was one of the factors that contributed to the shipwreck's high death toll. Some groups of people had a higher chance of surviving the sinking than others, such as women, children, and the upper class, even though there was some element of luck involved.

# **Objective:**

Our objective is to figure out what features would influence the survival, we are going to go deep into the data to explore the relationship between each attribute and survival.

# **About the collection of Dataset:**

The largest passenger ship ever built was involved in an iceberg collision on April 15, 1912. Of the 2224 passengers and crew, 1502 died when the Titanic sank. The international society was stunned by this shocking catastrophe, which prompted improved ship safety rules. The lack of lifeboats for the passengers and crew was one of the factors that contributed to the shipwreck's high death toll. Some groups of people had a higher chance of surviving the sinking than others, even though there was some element of luck involved.

The titanic.csv file contains data for 891 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived (S), their age (A), their passenger-class (C), their sex (G) and the fare they paid (X).

The data has been split into two groups:

training set (train.csv)

test set (test.csv)

Machine learning models are constructed using the training set. We offer the result (sometimes referred to as the "ground truth") for each passenger in the training set.Our project  will be based on "features" like the class and gender of the passengers. To develop new features, you can also employ feature engineering.

To evaluate how well our project performs on untested data, utilise the test set. We do not offer the ground truth for every passenger in the test set. It is your responsibility to foresee these results. Using the project we trained to estimate

# **Methodology:**

Construct training data frame

* Logistic Regression
* Random Forest
* Decision Tree
* SVM

# **Implementation:**

**Import the libraries,datasets and read the dataset-**

library('dplyr') *# data manipulation*

library('ggplot2') *# Data Visualization*

library('ggthemes') *# Data Visualization*

options(warn = -1)

*# load train.csv*

train <- read.csv('../input/train.csv', stringsAsFactors = F)

*# load test.csv*

test <- read.csv('../input/test.csv', stringsAsFactors = F)

*# combine them as a whole*

test$Survived <- NA

full <- rbind(train,test)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

filter, lag

The following objects are masked from ‘package:base’:

intersect, setdiff, setequal, union

**Data Exploration/Analysis**-

*# check the data*

str(full)

'data.frame': 1309 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...

$ Sex : chr "male" "female" "female" "female" ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : chr "" "C85" "" "C123" ...

$ Embarked : chr "S" "C" "S" "S" ...

We've got a sense of our variables, their class type, and the first few observations of each. We know we're working with 1309 observations of 12 variables. In which 891 observations are from train data set, and 418 observations are from test data set. When separate the variables by type, we have ordinal variable **PassengerId**, lable variable **Name** and **Ticket**, numeric variables such as **Age**, **SibSp**, **Parch**, **Fare**, and categorical variables like **Survived** ,**Pclass**, **Sex** ,**Cabin**, and **Embarked**.

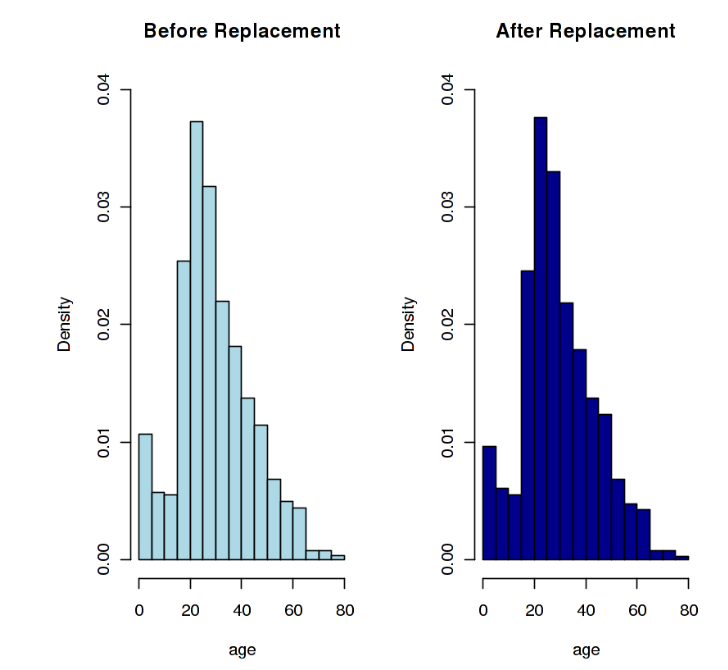
**Data Preparation and Exploratory Analysis-**

**Data Cleaning-**

From the data set, we notice that there are missing values in Age, Cabin ,Fare and Embarked column. We are going to replace missing values in Age with a random sample from existing ages. For Cabin, since cabin number makes little sense to the result, we are going to create a new Cabin column to indicate how many cabins the passenger has.

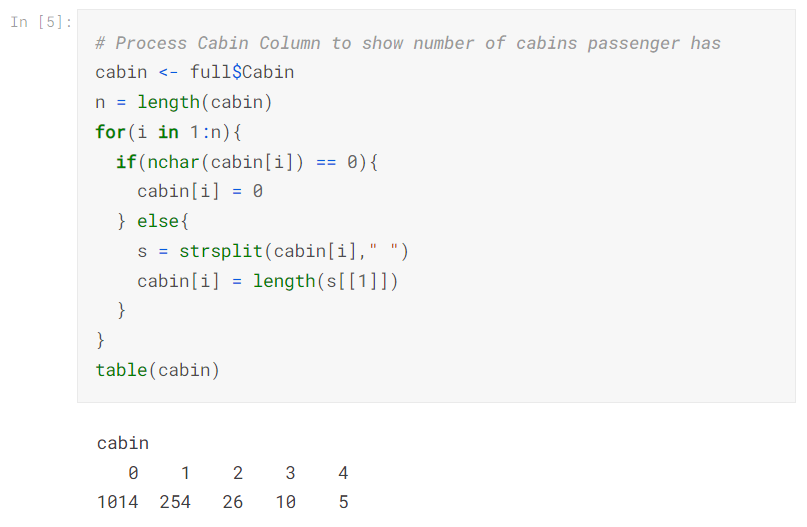
Age

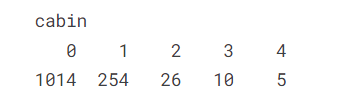




We can see from the histograms above that there is not much significant change of age distribution, which means the replacement is appropriate. Next we are going to process Cabin Column. We are going to create a new Cabin column to indicate how many cabins the passenger has.

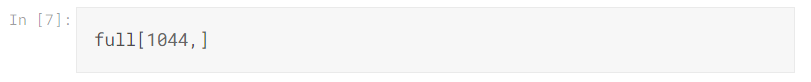
Cabin:



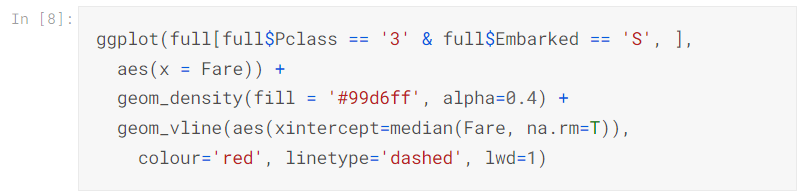


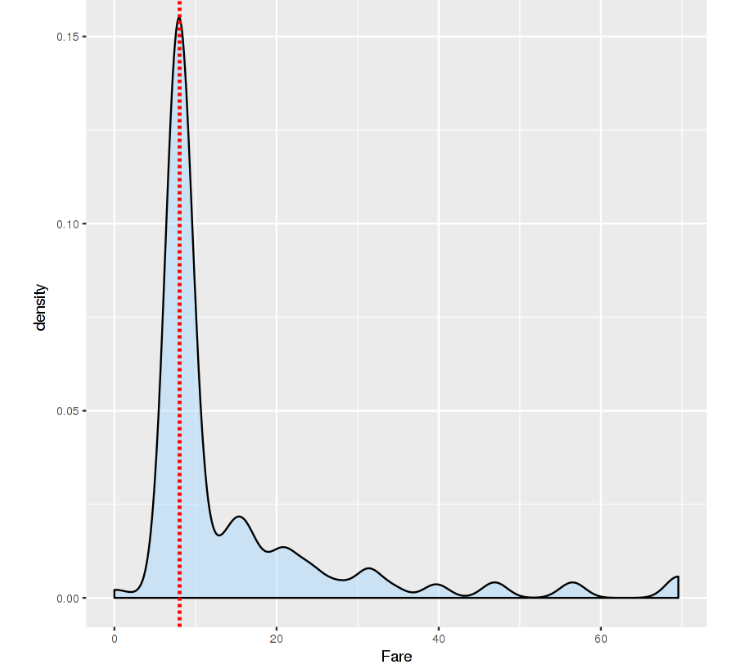
Fare-

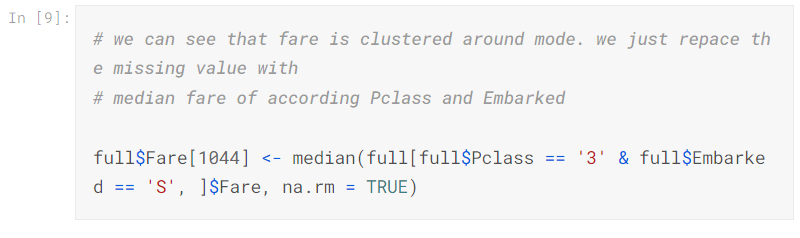




The passenger has feature value Pclass = 3 and Embarked =S. We then check the fare distribution of the same feature value.







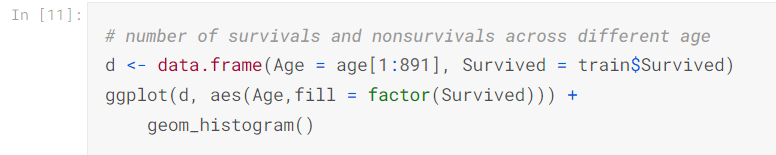
Embarked



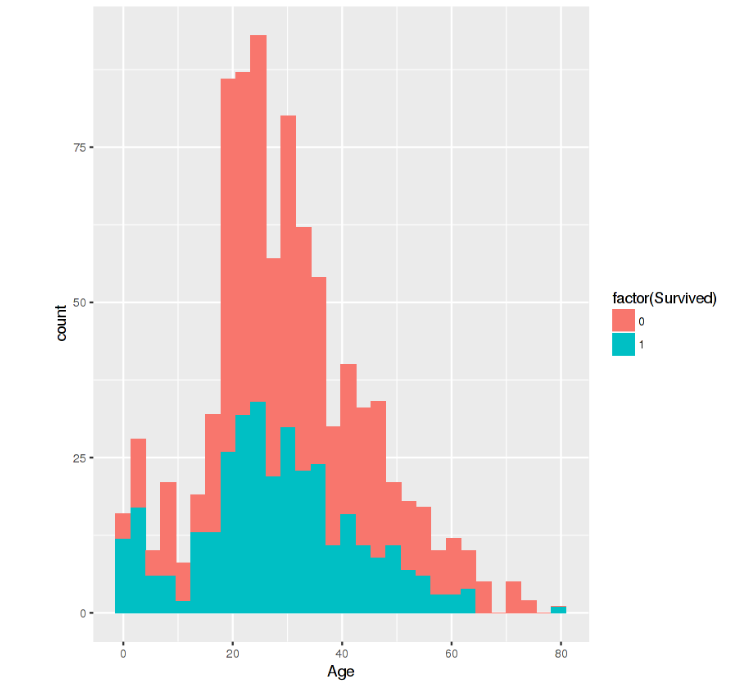
## **Exploratory Analysis and Data Processing**

As our objective is to figure out what features would influence the survival, we are going to go deep into the data to explore the relationship between each attribute and survival.

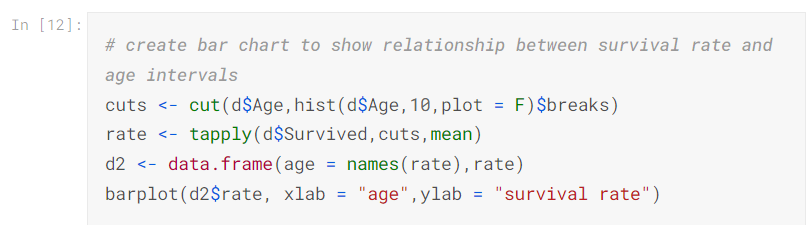
### **Age vs Survival**

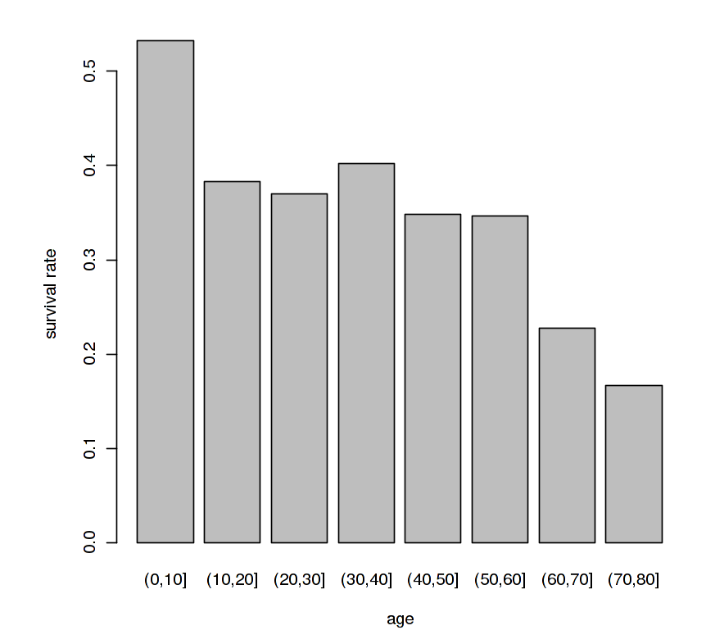


`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



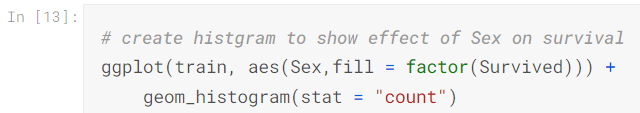
From the histogram, it seems that kids with very young age have a respectively higher survival rate, and elder people have a respectively lower survival rate. To verify it, I create a bar chart to show the relationship between survival rate and age intervals.

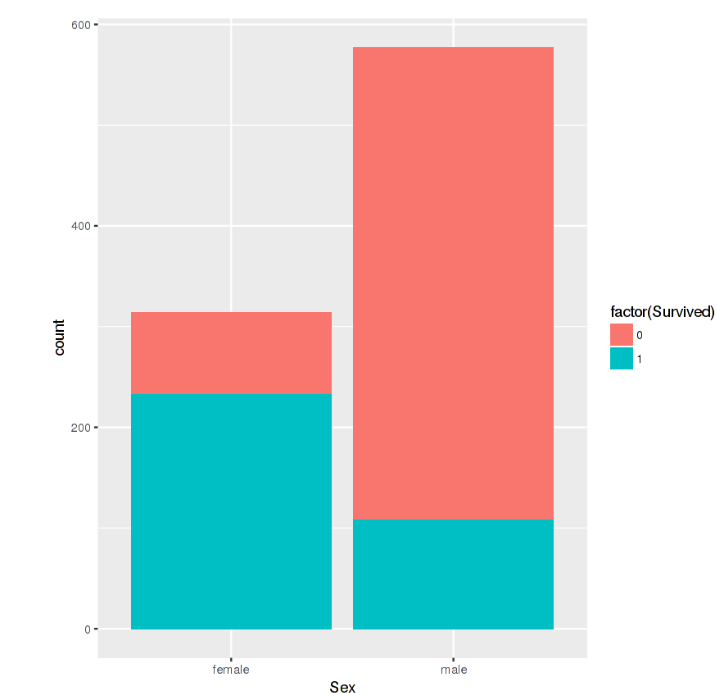




We can see clearly from the bar chart above that survival rate decreases as age increases. Kids below 10 years old have a higher survival rate above 0.5, people who's age is between 10 to 60 have a relatively constant survival rate around 0.4, and elder people above 60 years old has a lower survival rate around 0.2.

### **Sex vs Survival**

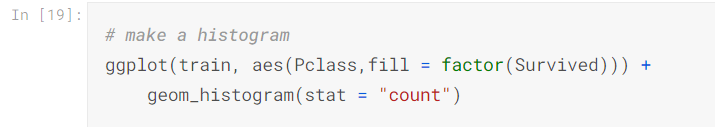


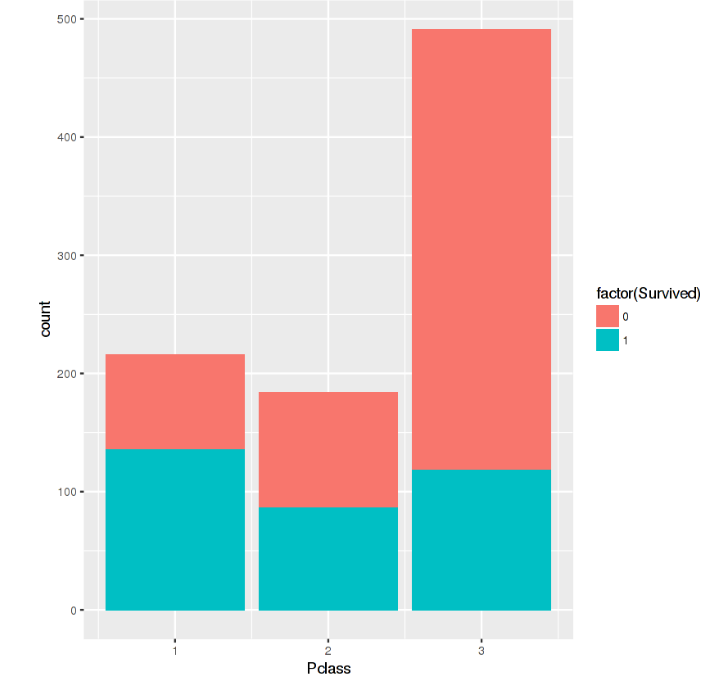


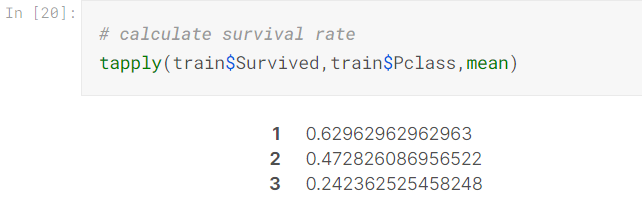
We can see from the histogram above that **female's survival rate is greater than male's**.

The survival rate of female is 0.74, while the survival rate of male is 0.19.

## **Pclass vs Survival**







From the histogram and table, we notice that Pclass = 1 group has the highest survival rate, then is Pclass = 2 group, and Pclass = 3 group has the lowest survival rate within these three groups.

Similarly,we calculate the survival rates with family size,cabin,fare,embarked and the results are as following.

**Family size vs survival**

We can see that the survival rate increases as the family size increases from 0 to 3. When family size becomes greater than 3, survival rate decrease dramatically.

**Cabin vs Survival**

We notice that passenger who has no cabin has a lower survival rate, and passenger who has one or more cabins has higher survival rate.

**Fare vs Survival**

We notice that Passengers who's fare is lower than 50 has a relatively lower survival rate. Passengers who's fare is extremely high (500-550) have very high survival rate.

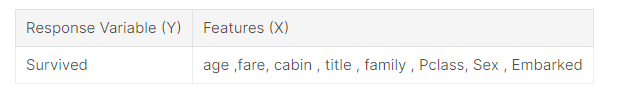
**Embarked vs Survival**

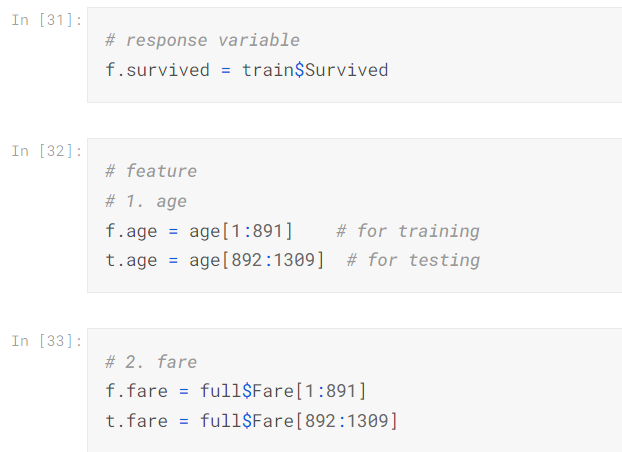
We notice that Embarked C group has a relatively higher survival rate than other 2 groups.

## **Modeling**

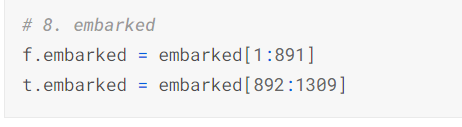
### **Feature Engineering**

In this section, we are going to prepare features used for training and predicting. We first choose our features that have significant effect on survival according to the exploratory process above. Here we choose Survived column as response variable, age (after filling), title, Pclass, Sex, family size, Fare, cabin(cabin count), Embarked these 8 column as features.



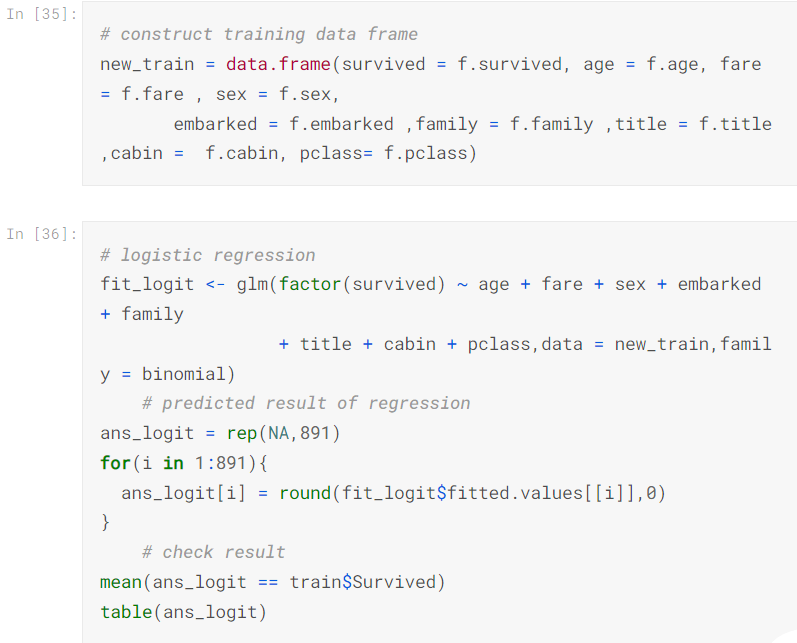


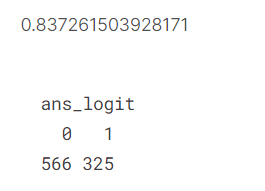




## **Model Training**

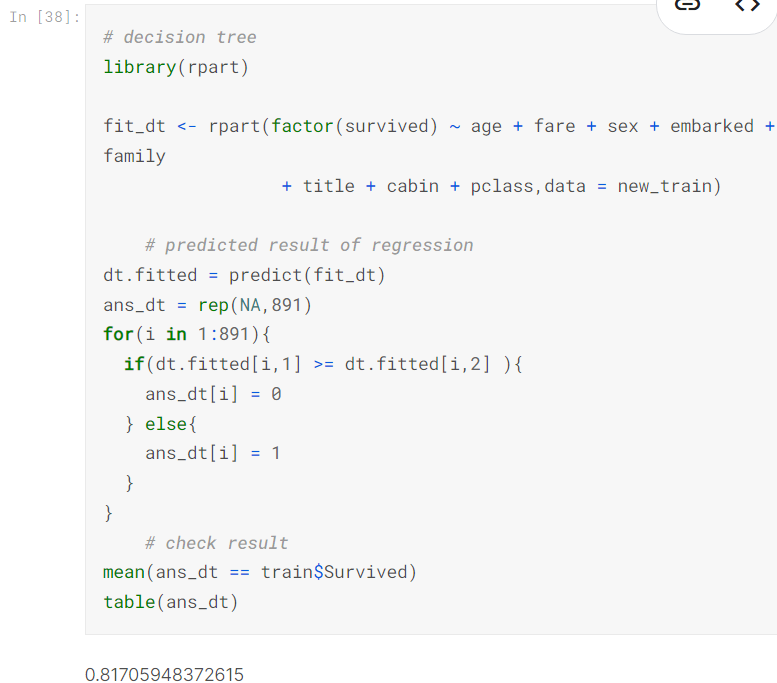
We tried to build basic learners using common machine learning model such as Logistic Regression, Decision Tree, Random Forest, SVM.

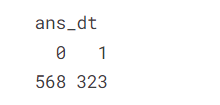


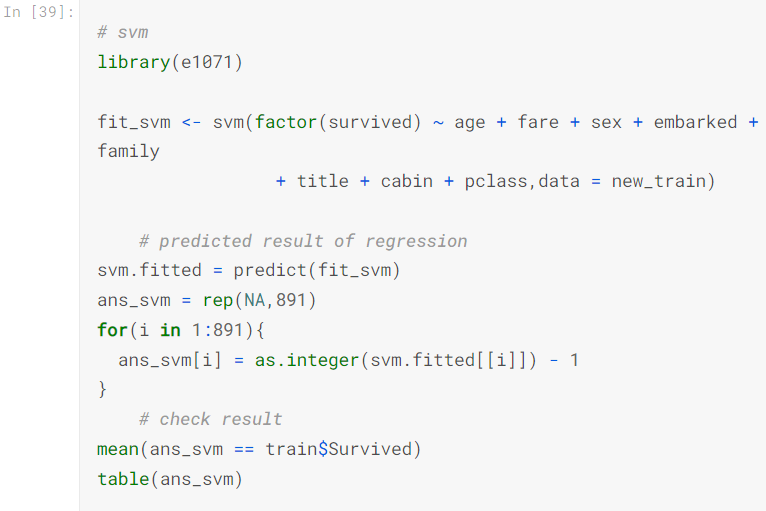


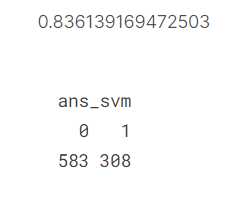












## **Model Evaluation**

We built 4 basic learner in last section. Then we evaluated model accuracy using Confusion Matrix.Of all the above models both logostic regression and Svm work well for training data set.

## **Prediction**

Since we got models that have reasonable predictive power, we can perform them to our test data set to make prediction. Here we choose SVM to perform prediction as an example.



## **Conclusion:**

With the help of clustering and some machine learning algorithms such as SVM, Logistic Regression, Random Forest and Decision Tree, we identify the the accuracy of predictive analysis of the Survival rate of Titanic based on some factors such as cabin no, sex. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.

## **References:**

* <https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8#:~:text=Our%20model%20predicts%2081%25%20of,the%20people%20who%20actually%20survived>.
* <https://www.kaggle.com/code/vincentlugat/titanic-data-analysis-rf-prediction-0-81818>