**Spring**

2018

Group 1

Yingjun Pan

BIA 656

Final Project Report

-Titanic: Machine Learning from Disaster

Content

Abstract 3

Introduction 3

Data and Methods 3

Understanding Variables 4

Variables Description 4

Exploring Data 4

Missing Value 10

Features Engineering 11

Applying Methods 15

Random Forest in randomForest Package 15

Results 15

Conclusion 16

Reference 16

# Abstract

The project is to find out what kinds of person are more likely to survive in the Titanic shipwreck. There are 1309 observations and I split 1047 observations, 80% of data as train data and the rest are test data. After exploring variables and filling missing data, I select ticket class, sex, age, numbers of siblings and spouses aboard, numbers of parents and children aboard, fare and embarked as variables. By applying features engineering, there are 4 more variables added in, title transited from name, family size added number of family member, ticket count and age group. I use random forest, decision tree, logistic regression and SVM to predict.

# Introduction

As we all know that the Titanic sank after colliding with an iceberg and killing 1502 out of 2224 passengers and crew. There were about 1300 passengers and the rest people are crew. The major reasons that shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. There was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others like women, children and the upper-class. The purpose of this project is to analysis of what sorts of people were likely to survive.

# Data and Methods

Dataset comes from [www.kaggle.com](http://www.kaggle.com) and there are 1309 observations in the dataset.

## Understanding Variables

### Variables Description

The dataset has 10 variables that include passenger survival or not, ticket class, sex, age, Number of siblings and spouses aboard, Number of parents and children aboard, ticket number, passenger fare, cabin number and port of embarkation. The number o in survival variable represents the passenger did not survival while number 1 means the passenger did survival. There are 3 ticket classes that marked as 1, 2, and 3. There are 3 ports of embarkation, C represented for Cherbourg, Q represented for Queenstown and S represented for Southampton. I make a variables table below to easy understand variables’ names.

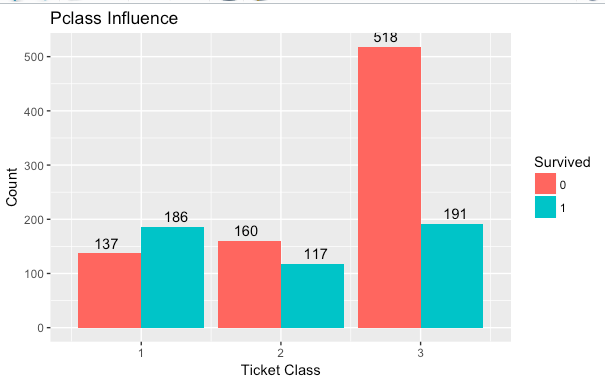
|  |  |
| --- | --- |
| **Variable** | **Description** |
| Survival | Survival or not |
| Pclass | Ticket class |
| Sex | Sex |
| Age | Age in years |
| Sibsp | Numbers of siblings and spouses aboard |
| Parch | Numbers of parents and children aboard |
| Ticket | Ticket number |
| Fare | Passenger fare |
| Cabin | Cabin number |
| Embarked | Port of embarkation |

### Exploring Data

#### Ticket Class – Pclass

Passengers in first class have the highest survival rate while passengers in the third class have the lowest survival rate and passengers in the third class is the majority. I list survival rate in below table.

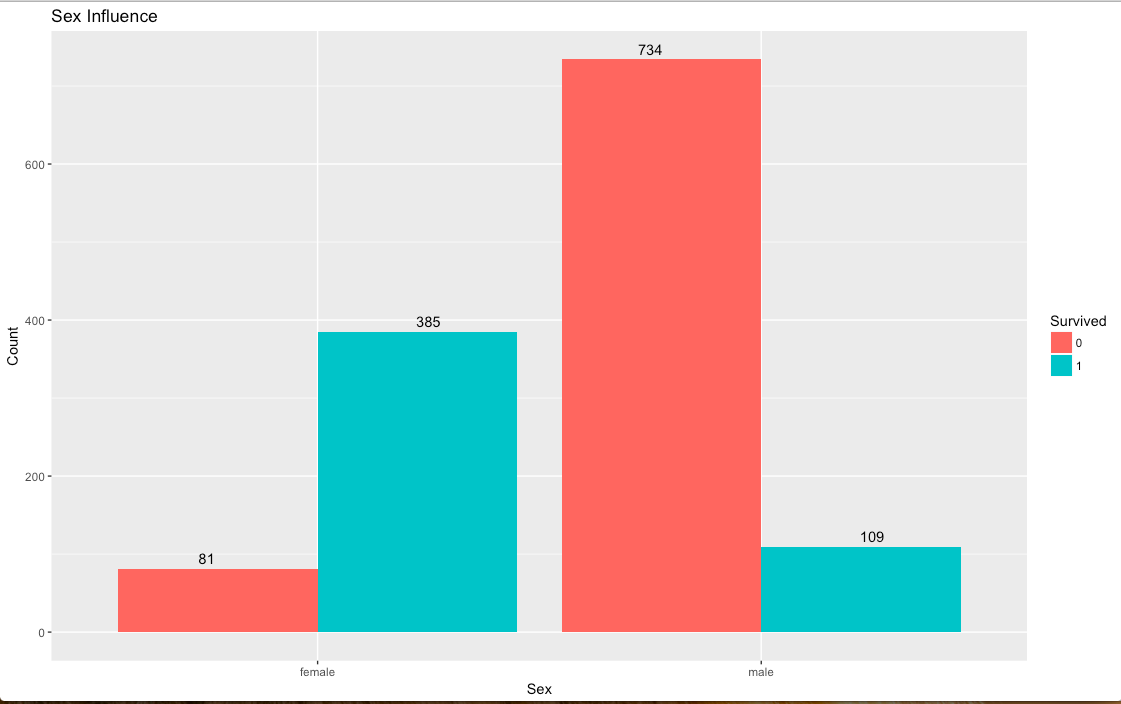
|  |  |
| --- | --- |
| **Ticket Class** | **Survival rate** |
| First class | 57.58% |
| Second class | 42.24% |
| Third class | 26.94% |



#### Sex

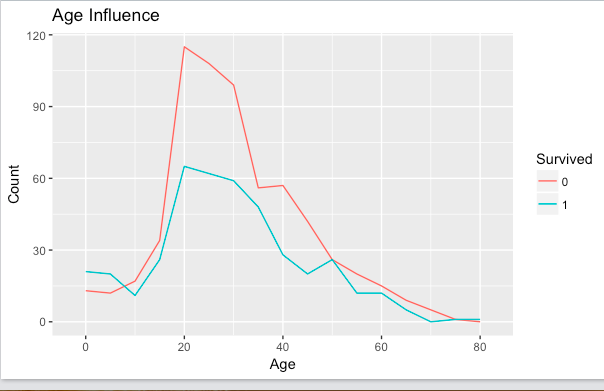
The numbers of male passengers is twice that of female passengers. The survival rate for female is much higher than that of male. It is because most people follow the Lady First principal. I list survival rate for both male and female.

|  |  |
| --- | --- |
| Sex | Survival rate |
| Male | 12.94% |
| Female | 82.62% |



#### Age

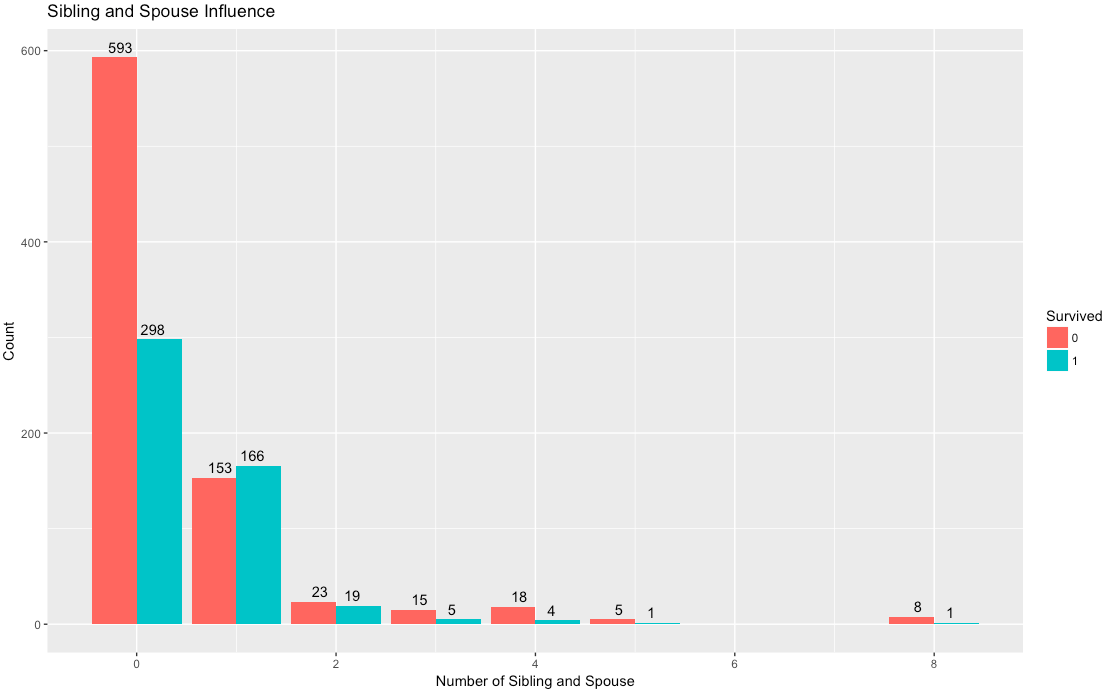
16 to 40 years old are the majority passengers.



#### Numbers of siblings and spouses aboard – Sibsp

Most people aboard with 1 sibling or spouse and few passengers with several siblings and spouse. Passenger with 2 siblings or spouse has the highest survival rate. The survival rate for passenger with 4 or more than 4 siblings or spouse aboard is lower than 20%. The survival rate for passengers with different numbers of siblings and spouse are listed below.

|  |  |
| --- | --- |
| Numbers of siblings and spouses | Survival Rate |
| 1 | 33.45% |
| 2 | 52.04% |
| 3 | 45.24% |
| 4 | 18.18% |
| 5 | 16.67% |
| 6 | N/A |
| 7 | N/A |
| 8 | 11.11% |

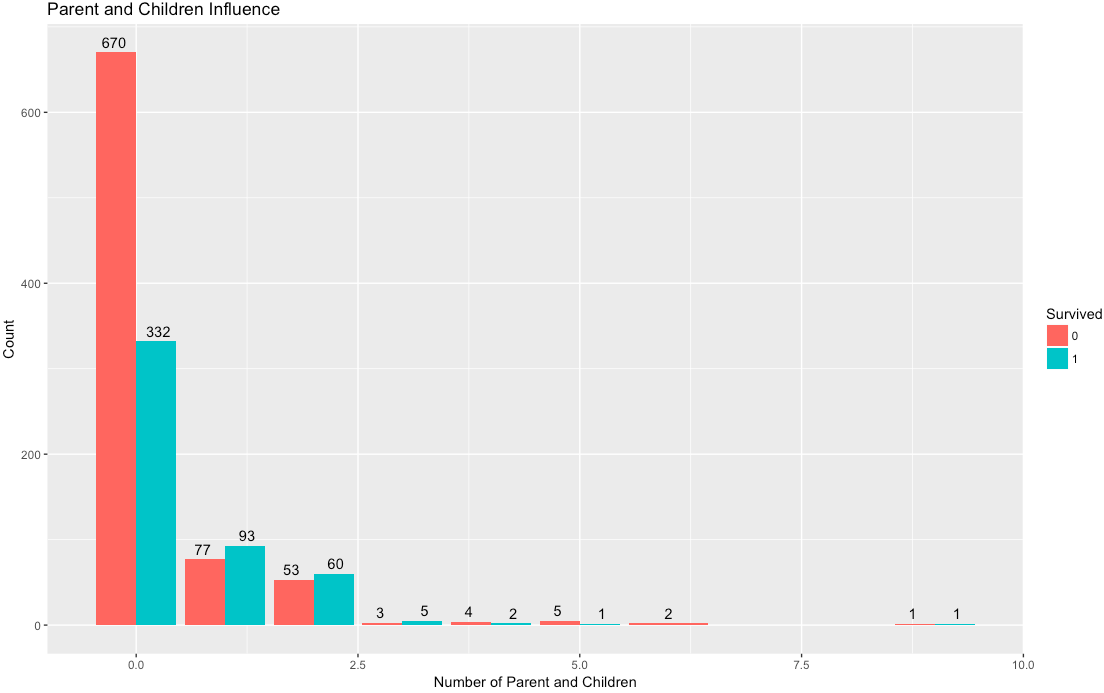


#### Numbers of parents and children aboard

Most people aboard are with 0 parent or child. The total numbers of passenger with more than 2 parents or children are too small to be useful, so I will ignore them for now.

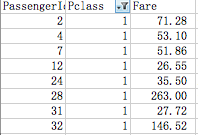
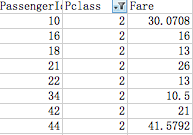
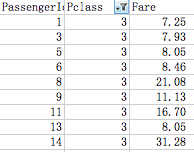
The survival rate for passengers with different numbers of parents and children aboard are listed below.

|  |  |
| --- | --- |
| Numbers of parents and children | Survival Rate |
| 0 | 36.81% |
| 1 | 54.71% |
| 2 | 53.10% |

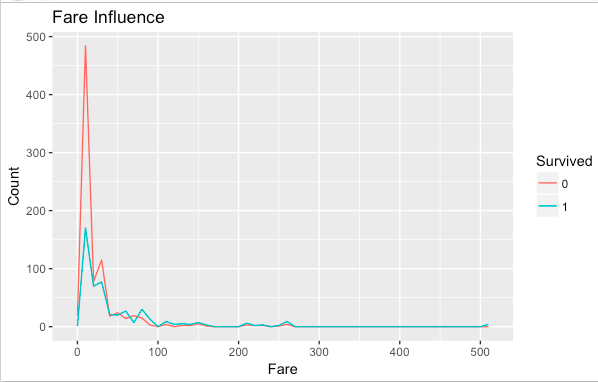


#### Passenger Fare – Fare

I assumed that fare could have a relationship with ticket class since the higher class always requires a higher fare. However, when looking into data, there is no remarkable relationship between ticket class and passenger fare. The fare between 20 and 30 occurred in all of 3 ticket classes. The example of fare for each ticket class is shown below.

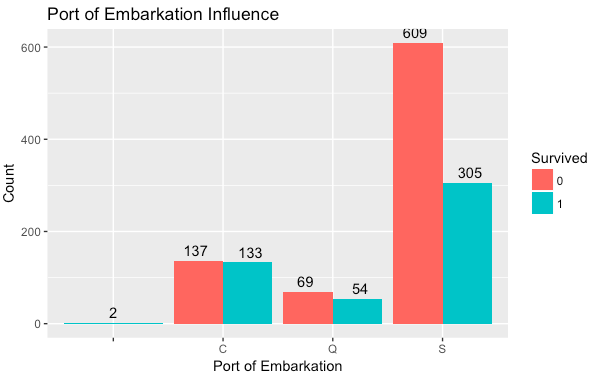
The majority fare is between 5 and 50.



#### Port of Embarkation – Embarked

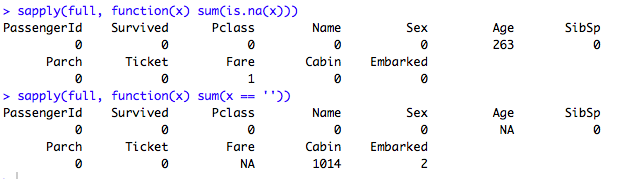
Most passengers were aboard on port Southampton while the survival rate is the lowest at that port. The rest two ports have almost the same survival rate. The survival rates for different ports are list below.

|  |  |
| --- | --- |
| Port of Embarkation | Survival Rate |
| Cherbourg | 49.26% |
| Queenstown | 43.90% |
| Southampton | 33.37% |

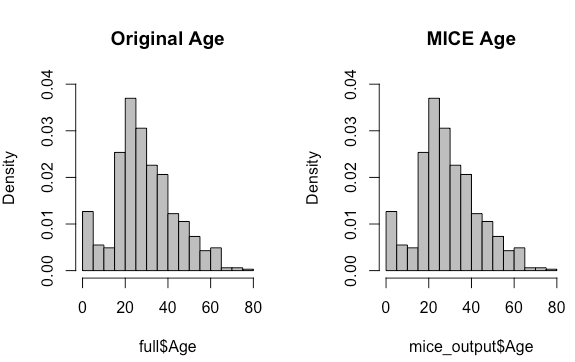


### Missing Value

There are 4 variables, Age, Fare, Cabin and embarked, have missing value, Cabin has too much (1014) missing value to fill accurately so I did not use Cabin in the rest of this project.



The 2 missing value in Embarked are filled by most common value ‘S’. The 1 missing value in fare is filled by similar situation that is the mean fare for passengers who are in third ticket class and in ‘S’ port of embarkation. For the 263(20% of all data) missing value in Age, I use mice that is a prediction method to fill. The shape of frequency distribution for miced age has almost the same shape with original age so this fill method is much better than using mean or mode. The frequency distributions for miced age and original age are shown below.



### Features Engineering

Feature engineering is the process of using domain knowledge of the data to create features that make machine-learning algorithms work. Feature engineering is fundamental to the application of machine learning.

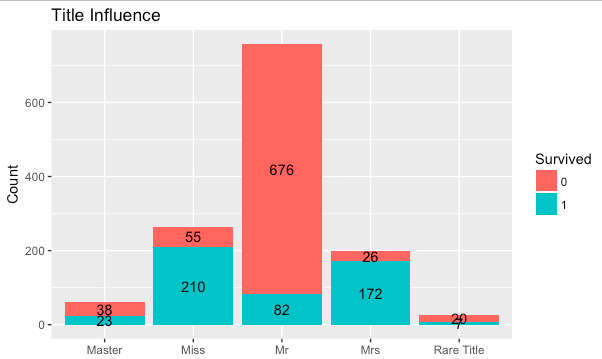
#### Title

The title variable comes from name variable. The second term in a name is title and I pick them up and classify them into groups. There is a sample of data in name variable below. Titles of 'Dona', 'the Countess', 'Capt', 'Col', 'Don', 'Dr', 'Major', 'Rev' and 'Jonkheer' are classified as rare\_title. Titles of ‘Lady’, ‘Mlle’, ‘Ms’ and ‘Miss’ are classified as Miss. Titles of ‘Mme’ and ‘Mrs’ are classified as Mrs.



The majority is Mr. and it is has the lowest survival rate. Survival rates for various titles are listed below.

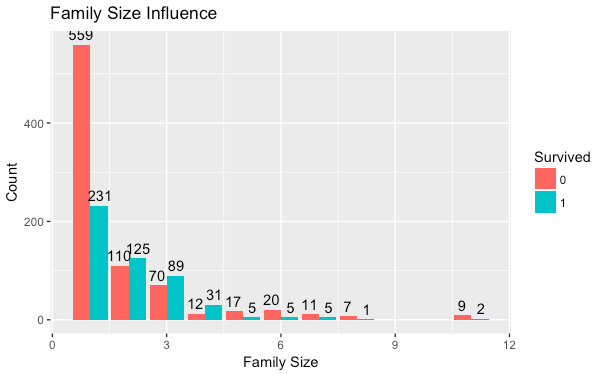
|  |  |
| --- | --- |
| Title | Survival Rate |
| Master | 37.70% |
| Miss | 79.25% |
| Mr. | 10.82% |
| Mrs. | 86.87% |
| Rare Title | 25.93% |



#### Family Size

The calculation of family size is to add the numbers of siblings and spouses and numbers of parents and children and self. Most passengers are travel by with 1, 2 or 3 family members so I list survival rates for them. The numbers of passenger in the rest group are too small to be useful.

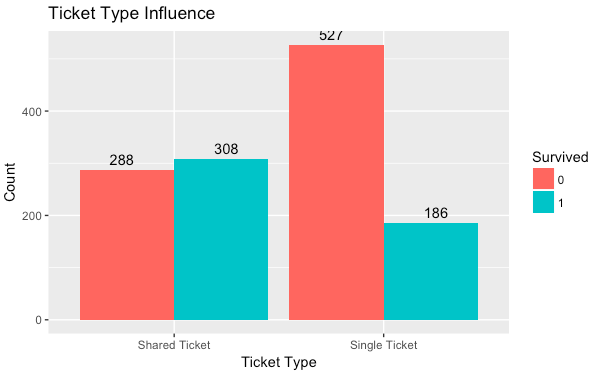
|  |  |
| --- | --- |
| Family Size | Survival Rate |
| 1 | 29.24% |
| 2 | 53.19% |
| 3 | 55.97% |



#### Ticket Count

The new variable of ticket count is from ticket variable. If passengers have a same ticket number then they are more likely know each other so they were not aboard alone, while the others are aboard alone. The numbers of passengers who were aboard alone or not are almost same while the group that were aboard with acquaintance has much higher survival rate than the other group people. The survival rates for two groups are listed blow.

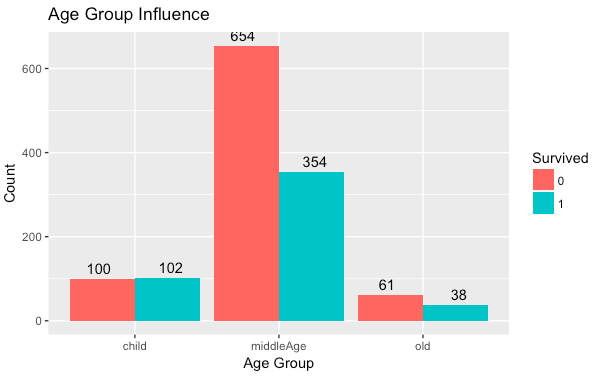
|  |  |
| --- | --- |
| Ticket Type | Survival Rate |
| Shared Ticket | 51.68% |
| Single Ticket | 26.09% |



#### Age Group – Age box

The new variable age group comes from age. I label passengers who are younger than 18 years old as child, passengers who are between 18 and 50 as middle Age and passengers who are older than 50 are old. The middle age is majority, while child has the highest survival rate. Survival rates for each group are list below.

|  |  |
| --- | --- |
| Age Group | Survival Rate |
| Child | 50.49% |
| Middle Age | 35.12% |
| Old | 38.38% |



## Applying Methods

The dataset is split in to 80% training data that contains 1047 observations and 20% testing data that contains 262 observations.

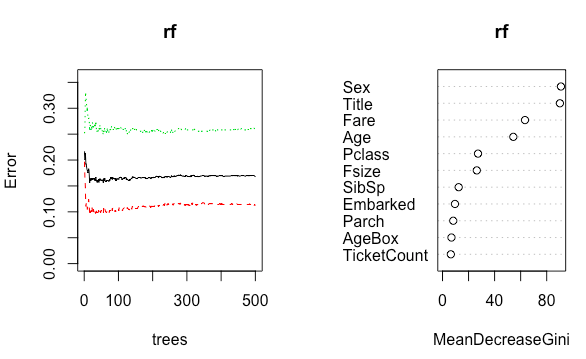
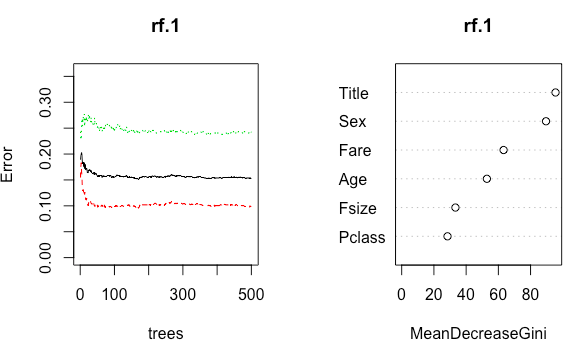
The variable of Survival is to be predicted and variables of Pclass, Title, Sex, Age, SibSp, Parch, Fsize, TicketCount, Fare, Embarked, AgeBox are independent variable.

I applyed random forest in both randomForest package and party package, and applyed decision tree in both rpart and part packages. I applied logistic regression and Support Vector Machine as well.

I will talk more about random forest in randomForest package because I refined this one by using less independent variable. The rest methods I used the same prediction function mentioned above.

### Random Forest in randomForest Package

After applying the above prediction function and plot mean decrease Gini, I found that TicketCount, AgeBox, Parch, Embarked and Sibsp did limited contribution in this prediction so I tried a new prediction function without these variable. The right picture is for refined result.

We can see from plots named tree the Error decrease a little bit.

# Results

Accuracy compared in the blow table.

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| Random Forest | 0.9083969 |
| Refine Random Forest | 0.9160305 |
| Cforest | 0.9465649 |
| Decision Tree (rpart) | 0.9389313 |
| Ctree | 0.9389313 |
| Logistics Regression | 0.9427481 |
| SVM | 0.8587786 |

Method rank based on accuracy:

Cforest > Logistics Regression > Decision Tree(rpart = ctree) > Refine Random Forest > Random Forest > SVM

# Conclusion

To predict this dataset require ability to deal with missing data and find potential variables by using features engineering. If just using given variables, the accuracy is difficult to reach 90%.

In this project, I spent a lot time on finding a better independent variables to refine the prediction, while did not spent much time on try different parameters to improve the accuracy. I believe when using the suitable parameter in the SVM, the accuracy will have a great degree of improvement.

# 

# Reference

1. Kaggle <https://www.kaggle.com/c/titanic>
2. Feature Engineering <https://en.wikipedia.org/wiki/Feature_engineering>
3. Imputing missing data with R; MICE package

https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

1. An Introduction to Statistical Learning with Applications in R