

Capstone project data science

Titanic Survival



March 20, 2019

ItGuts

Delhi

Contents

[Titanic Survival Project 2](#_Toc4007501)

[Steps:- 2](#_Toc4007502)

[1. Business Problem Understanding 2](#_Toc4007503)

[2. Data Collection 2](#_Toc4007504)

[3. First Prediction Model 2](#_Toc4007505)

[4. Data Preparation/Cleaning 2](#_Toc4007506)

[5. Data Exploration 2](#_Toc4007507)

[6. Feature Engineering 2](#_Toc4007508)

[7. Model Building by applying multiple machine learning algorithms 2](#_Toc4007509)

[8. Model Evaluation 2](#_Toc4007510)

[9. Model Tuning 2](#_Toc4007511)

[10. Final Conclusion and deployment 2](#_Toc4007512)

# Titanic Survival Project

## Steps:-

### Business Problem Understanding

### Data Collection

### First Prediction Model

### Data Preparation/Cleaning

### Data Exploration

### Feature Engineering

### Model Building by applying multiple machine learning algorithms

### Model Evaluation

### Model Tuning

### Final Conclusion and deployment

#### Business Problem Understanding:-

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.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

#### Data Collection

<https://www.kaggle.com/c/titanic/data>

MyAccount- <https://www.kaggle.com/itguts>



#### First Prediction Model

Without applying any machine learning tools and steps, we can apply our common sense and predict accordingly. It is called as Lazy Prediction.

1. Let’s assume all the passengers of the test data will die
2. When we see the count of survived based on gender I find that the female survival rate is higher. So lazily, I predict that all the female passengers in the test dataset will survive.
3. After each prediction, you can submit the file to kaggle and check score.

All codes in R can be found here:- 

#### Data Preparation

We should consider below factors:-

1. Missing value imputation:-

If missing value of a parameter/filed/column is less than 20%, we should remove the parameter

1. Outlier detection:-

It is important to know the outliers, as at the time of imputing missing values, we can select either mean or median. If outliers is high then median is the best option.

1. Imbalanced Data Treatment
2. Data Type Conversion
3. Data Manipulation

Code 

#### Data Exploration

It is also known as EDA. Need to analysis each parameter that how they are contributing the Survival rate.

1. Univariate Analysis
2. Bivariate Analysis
3. Multivariate analysis

Code - 

#### Feature Engineering

If you think that by creating new feature/field will help in analysing the fact that would be great for your model.

1. Child:- whether the passenger is child or not(age<20)
2. Title:- Extracting title from the name of the passenger and see what titles of the passengers survived most
3. Family Size:- adding self+ sibling+parch will give family size and check the survival probability of a passenger who has family size > 1

Code- 

#### Model Building

We can apply many suitable models after identifying the type of machine learning. This assignment comes under classification supervised machine learning, so we can use logistics, knn, naïve bayes, svm, decision tree or random forest. After that, we can the accuracy and robustness of each model.

We are using Logistic regression, Decision tree and Random forest.

##### Logistic Regression



##### Decision Tree



##### Random Forest



##### SVM(Support Vector Machine)

