**Project Titanic**

**Introduction**

In this project, we'll be using Titanic dataset which consists of various features

to predict on test data that whether a person survived or not during titanic sank incident.

We'll be applying Machine Learning to get this done.

In Machine Learning we'll be using :-

1.) data analysis , data visualization , graph plot for feature selection

2.) cross\_val\_score prediction , accuracy\_score prediction , grid\_search for hyperparameter tuning

3.) Various algorithms to find the best model

The dataset contains data for 887 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including :-

* **Survived**(whether they Survived),
* **Age**(age),
* **Pclass** (passenger-class),
* **Sex** (sex),
* **Fare**( fare they paid ) ,
* **Sibsp**( Number of Siblings/Spouses Aboard),
* **Parch**(Number of Parents/Children Aboard) ,
* **Ticket**(Ticket Number) ,
* **Cabin**(Whether they have a cabin),
* **Name**(Name of the paseengers),
* **Passenger ID**(Id’s of the passengers),
* **Embarked**(Port of embarkation)

**Starting with the project**

1. Importing the libraries and dataset

All the libraries needed for the project will be imported along with the dataset.

1. Feature Selection

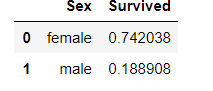
The main part of data analysis and data wrangling is feature selection.

From all the features provided in the dataset , we’ll have to find the one which are mainly dependent on the output(Survival) and will increase it’s accuracy.

For that we’ll check the relation of each feature with the output feature through graphs and grouping them and finding mean values of survival for each category.

1. **Sex**

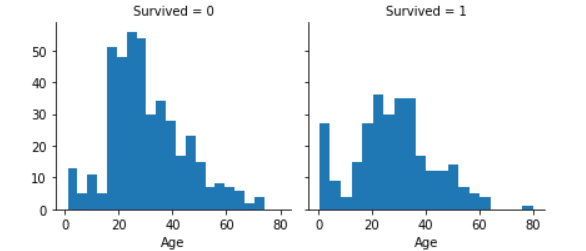
We’ll group by the category of sex and found it’s mean values with the survival.



It is clearly seen that the female have survived more than that of male.

It means that sex is an important feature for our prediction model.

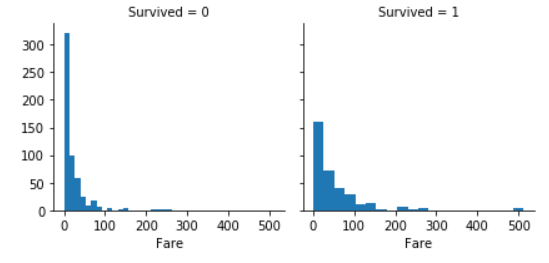
1. **Age**



From this graph we see that:-

* Infants (Age <=4) had high survival rate.
* Oldest passengers have survived.
* Large number of 15-25 year olds did not survive.
* Most passengers are in 15-35 age range.
* So, Age is an important feature for prediction

1. **Fare**

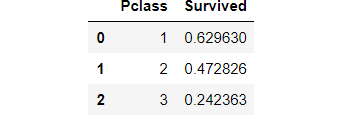


Here we see that, the people with high fares haves more survival rate than the one with less fares.

So, Fare will be an important feature.

1. **Pclass**

We’ll group by and see the mean of survived in Pclass, as it has 3 categories.

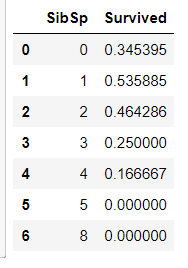


So, here we see that passengers with Pclass= 1 have survived most.

Threrefore, Pclass has to be included in feature selection

1. **SibSp**

We’ll group by and see the mean of survived in SibSp, as it has 3 categories.

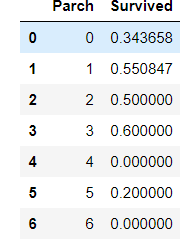


So, here we see that passengers with Sibling = 1 and Sibling = 2 have survived most.

Threrefore, SibSp has to be included in feature selection

1. **Parch**

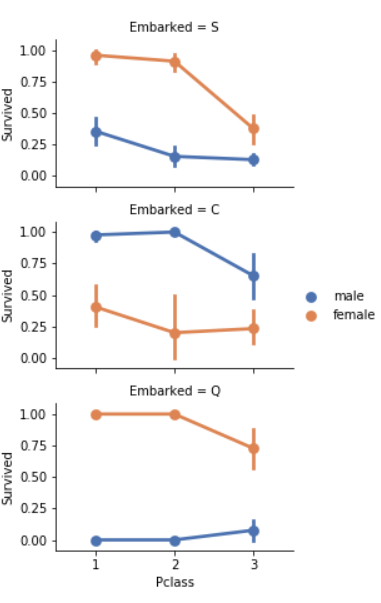
Again , we’ll be seeing the mean of survival by grouping the categories of Parch



So, here we see that passengers with Parch = 1 and Parch = 3 have survived most.

Threrefore, Parch has to be included in feature selection.

1. Embarked

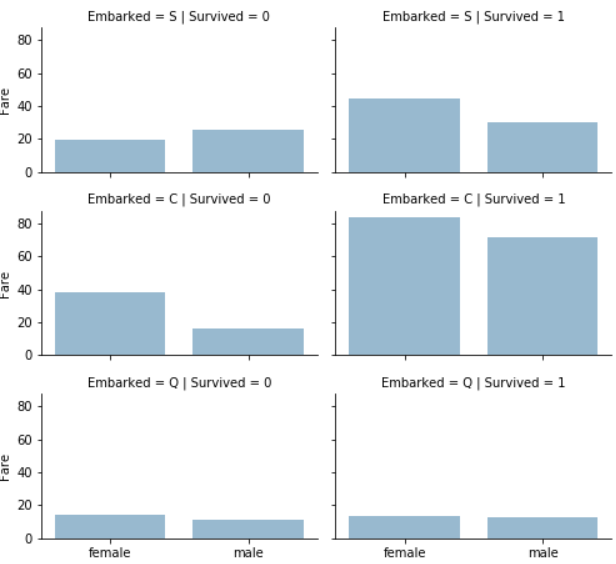


It is relation of Embarked with sex and survived.

Here we see that , Female passengers had much better survival rate than males , as we have seen above.

Exception in Embarked=C where males had higher survival rate. This could be a correlation between Pclass or Fare and Embarked and in turn Pclass and Survived, not necessarily direct correlation between Embarked and Survived.

So , we’ll see the relation of embarked with fare.



Now, here we see that embarked has a relation with fare , as people with ticket of high fare have high chance of survival

So Embarked has to be included in the feature.

1. Finalising the dataset

So, we’ll make our final dataset with the features selected and dropping the one’s which are not useful.

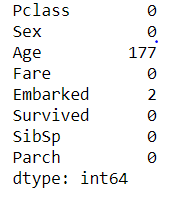
Final features are:-

* **Survived**(whether they Survived),
* **Age**(age),
* **Pclass** (passenger-class),
* **Sex** (sex),
* **Fare**( fare they paid ) ,
* **Sibsp**( Number of Siblings/Spouses Aboard),
* **Parch**(Number of Parents/Children Aboard) ,
* **Embarked**(Port of embarkation)

While features **Name , Cabin , Passenger Id , TicketNo** are not important and will be dropped from the final dataset.

1. Missing Values Imputation

The next step after finalising our dataset is to see whether any feature has NULL values or not and if there , then we have to decide what action should we take.



We see that age , embarked has empty places.

**Embarked has 2 empty places and Age has 177 empty places.**

**So , as Age has 177 values it is preferred to fill those empty values with the mean of the age value and Embarked as ‘S’.**

1. Encoding

After our dataset is finally ready , we’ll have to encode the numeric values present in the feature.

For eg:- If a feature consist of 4 uniques values like 1,2,3,4 then like continuous data, then we would assume that 4, is more similar to 3, than it is to 1 or 2, which are of other types. Thus, it would be more likely to cluster the category with 4s and 3s together than the 4s and 1s together. In contrast, if we treat these numbers like categorical data, then we treat each one a separate category that is no more or less similar to any of the other categories. Thus, the likelihood of clustering category with 4s with 3s is the same as clustering 4s with 1s, and that approach is more reasonable.

We’ll encode Pclass , Embarked , Parch and Sex features.

1. Input/Output Feature Selection , Scaling and Splitting Dataset

* Next , we’ll make ‘Survived’ as the output feature and rest as input features.
* After that we’ll scale values in ‘Fare’ , ‘Age’ column.

We'll scale the values to remove prioritisation of one feature over another based on its higher value than another.

Moreover, SVC algorithm needs scaled values for best prediction accuracy.

* And at last , we’ll split the dataset into train and test datasets.

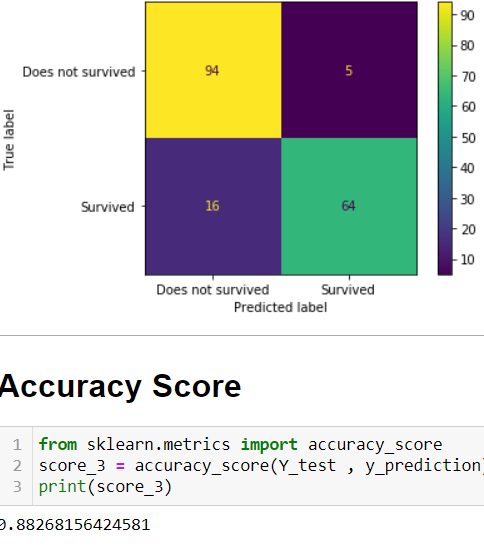
1. Algorithm Selection

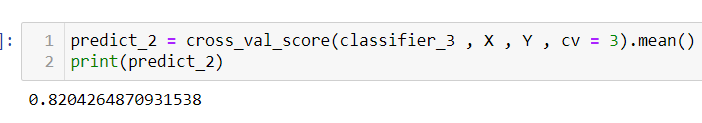
We’ll use different algorithms on the dataset and see which algorithm gives the best result.

We have used :-

* Random Forest Classifier
* SVM
* Logistic Regression

**Results**





**As here we see , the best model accuracy is achieved by SVC model which is 88%.**