**Predicting the Survival of Titanic Project Using Machine Learning.**



In this blog, we will be going through the whole process of creating the Machine Learning model on the Titanic disaster data set. For those who don’t know about Titanic disaster lets me quickly explain you, The RMS **Titanic**, a luxury steamship, sank in the early hours of April 15, 1912, off the coast of Newfoundland in the North Atlantic after sideswiping an iceberg during its maiden voyage. Now let us quickly go through our problem statement.

**Problem Definition:**

The Titanic Problem is based on the sinking of the ‘Unsinkable’ ship Titanic in early 1912. It gives you information about multiple people like their ages, sexes, sibling counts, embarkment points, and whether or not they survived the disaster. Based on these features, we have to predict if an arbitrary passenger on Titanic would survive the sinking or not.

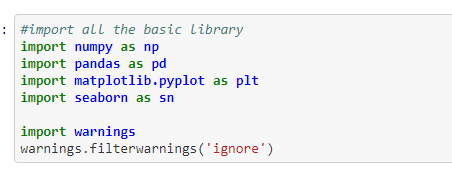
**Which passenger this disaster Survived?**



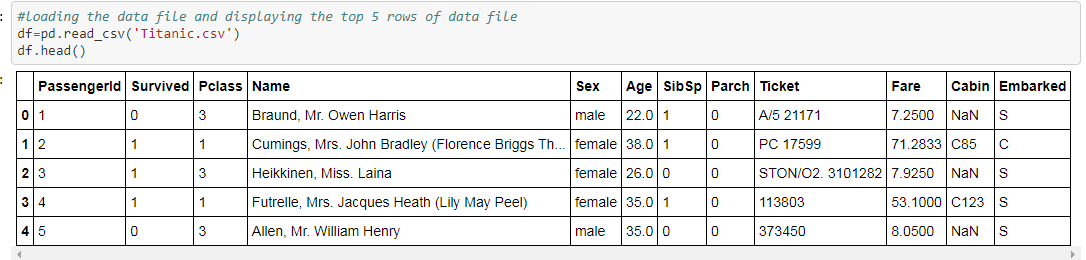
**Data Analysis:**

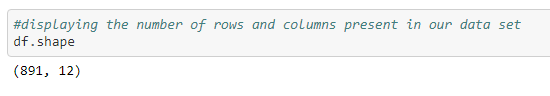


Data helps us to get the insight of the disaster. Data analysis helps us to discover the useful information from which we might be able to classify easily. Apart from this Data analysis help us to get rid of unwanted data and also the outlier in the data. Now let us quickly go through our titanic data set, before that lets us quickly import the entire basic library.



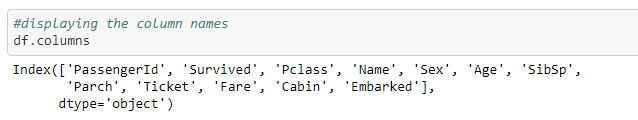
Let’s import our data set

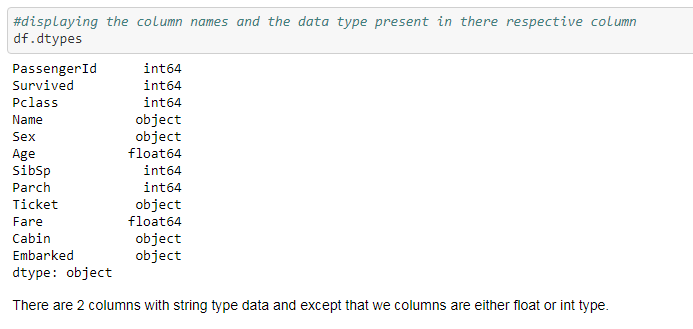




* We have 12 columns and 891 rows, small data set we must be bit careful while training the data.
* Here column contains the Attributes/Variable and rows contains the record of the each passengers.
* Our Dependent variable is “Survived” and looking at this we can say this problem is Binary Classification or a Logistic regression.

Now it’s time to go through each column of our data set. By studying the column we will be able to get a lot of information and understanding about our data set.

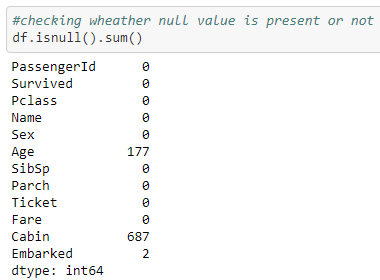


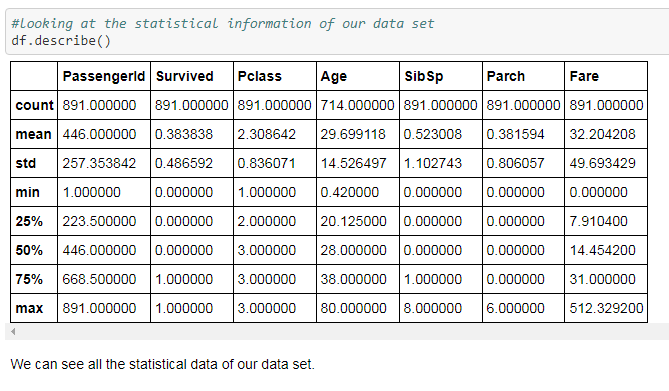


From above picture we can see that

|  |  |
| --- | --- |
| **Name** | **Description** |
| PassengerId | Numerical Value |
| Survived | Outcome of survival (1 = Yes, 0 = No) |
| Pclass | Social-economic class (1 = Upper class, 2 = Middle class, 3 = Lower class) |
| Name | Name of passenger |
| Sex | Sex of the passenger(male, female) |
| Age | Numerical continuous Value (Age of the passenger) |
| SibSp | Number of siblings and spouses of the passenger aboard |
| Parch | Number of parents and children of the passenger aboard |
| Ticket | Ticket number |
| Fare | Amount paid by the passenger |
| Cabin | Cabin number of the passenger |
| Embarked | Passenger boarding point (C = Cherbourg; Q = Queenstown; S = Southampton) |

* We can see that our data set contains integer, float and object type of data set. We must a keep in mind that all the objective type of data set has to changed into integer type
* Below you can see that there are some missing values in three columns they are Age, Cabin and Embarked. We need to get rig of this missing value either by filling them or by dropping this column

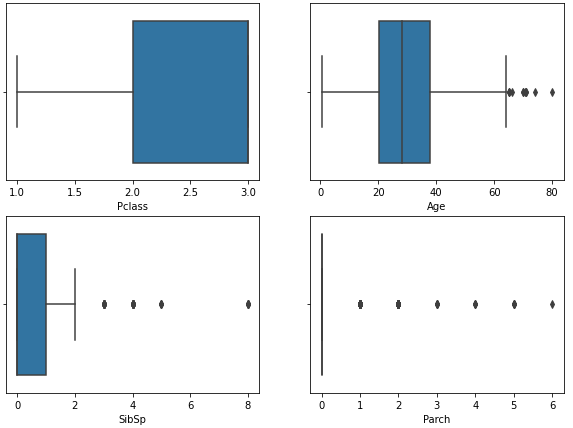


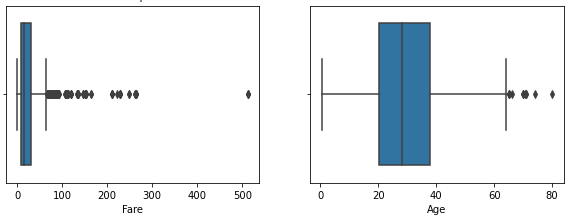


* Above we can see that the Age of the passenger ranges from 0.42 to 80, from this we can tell that some ones age can’t be 0.42 this are all the outlier we must also treat that before building the model.
* Fare ranges from 0.00 to 512.326200

Now lets us carry out data analysis by graphical method.

Let’s plot the box plot to see the **outlier.**

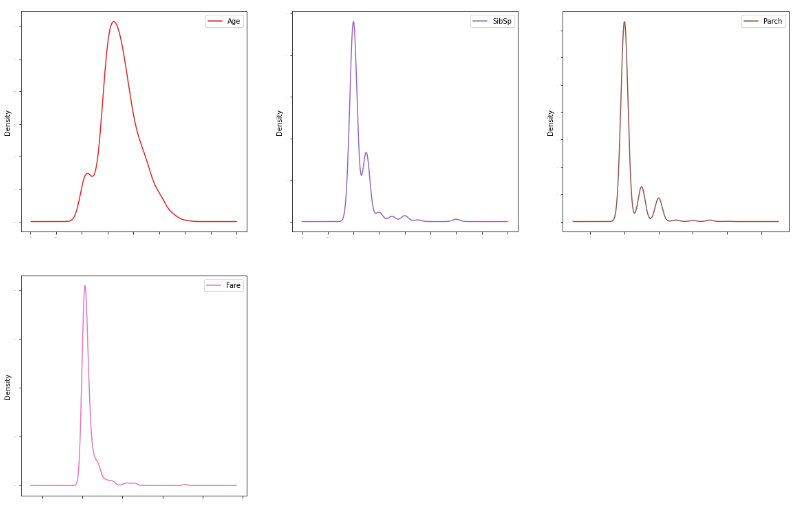
****

****

Outlier is present in all the numerical column and if we pass on this outlier into our model there is high change that our model would be having high variance.

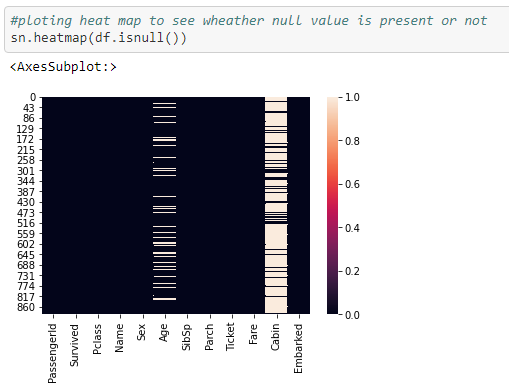
Most of the Machine learning Model assume that our model has no **skewness** and Tail data will act as the outlier this will affect the performance of the data. So we must minimize the skewness as much as possible.



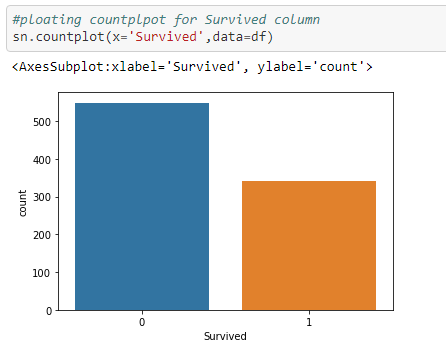


We can see the skewness in the column Age, SibSp, Parch, Fare in the later part we will we treating and strive hard to make it as minimum as possible.

From the function isnull() we can see that we have any **null/missing value** present in our datat set. Lets plot the heatmap to display the null value, from the below heat map we can easily see that cabin and Age has got lot of missing value

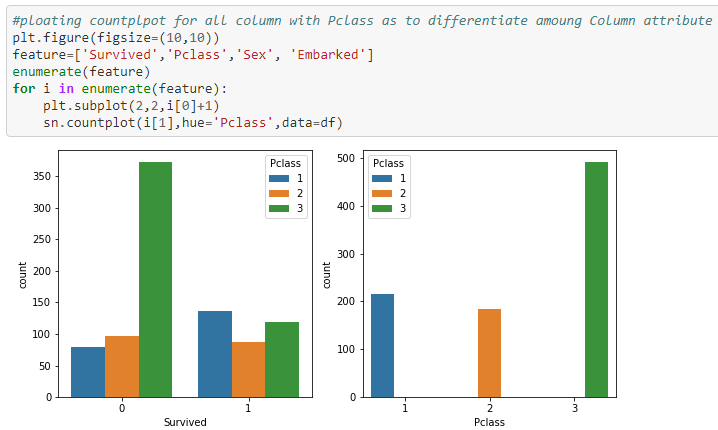


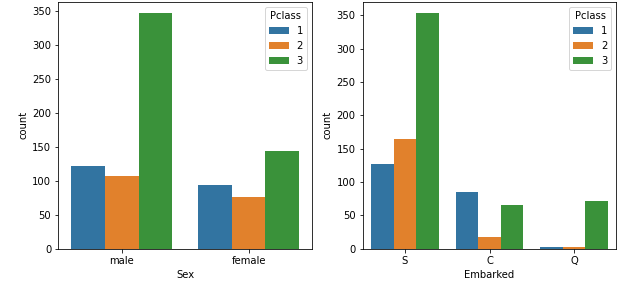
Firstly if we have classification problem we need to check whether we have **class imbalance** or not. Because if we have class imbalance in our model, model will become biased towards majority class which has got more records.



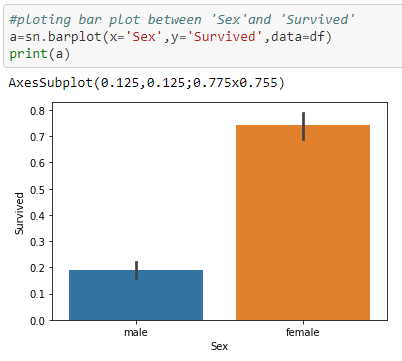
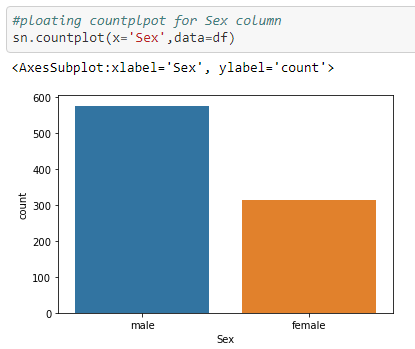
We have class imbalance we need to equalize it either by over sampling or under sampling.

Let’s plot some other graphs to get the insight of our data

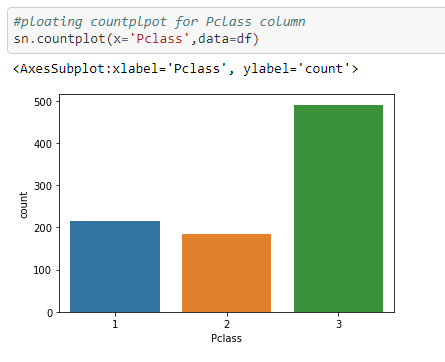




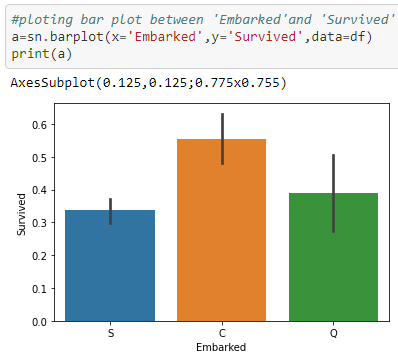
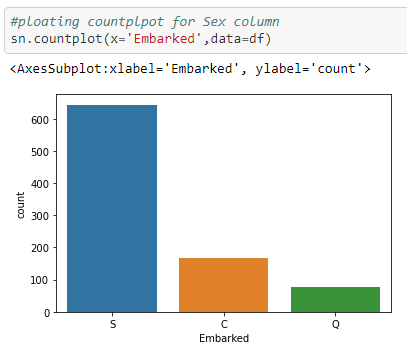
* People from Pclass 3 has lost their lives largely after that Pclass 2 followed by Pclass 1
* People from Pclass 3 in more in count followed by Pclass 1 and Pclass 2



* Male are more in count compared to female.
* Survival Rate of female is more compared to male that you can confirm by right side graph

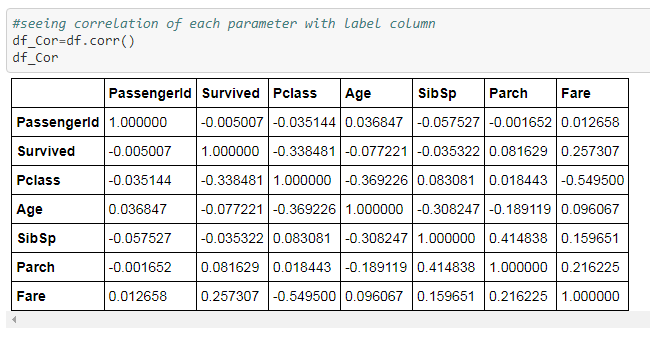


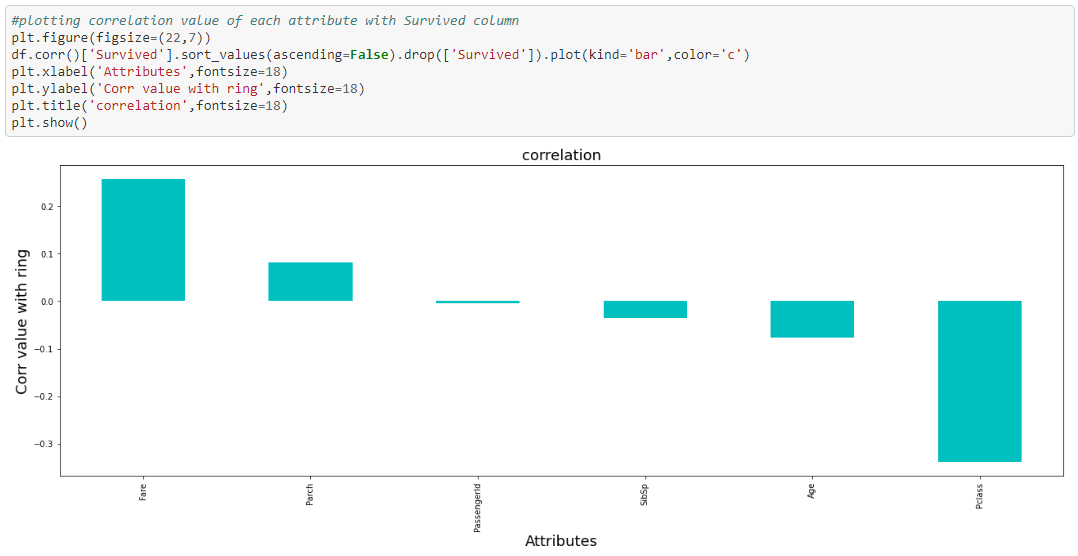
* From the above diagram we can see that people from Pclass 3 is more followed by Pclass 1 and Pclass 2
* But When we look at graph on right side we can see that Survival from Pclass 1 is more followed by Pclass 2 and Pclass 3



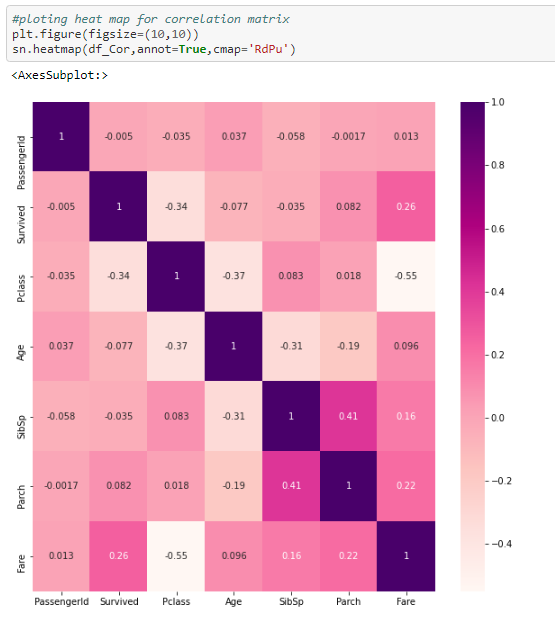
* We can see that the people starting the voyage from point P is more followed by point C and point Q.
* But survival rate of point C is more followed by point Q and Point S

We should also see **correlation** of our dependent variable with independent variable and also the correlation with dependent variables with other independent variable.



* From above graph we can see that Fare price has good positive correlation with survival rate that means that the probability of survival rate increases with increase with the Fare price.
* And Pclass high negative correlation with Survival rate that means if Pclass increases the survival probability decreases. That means people for Pclass 3 survival rate is less compared with other 2. We must keep in mind that the People from Pclass 3 belongs to Lower class.
* Parch has low positive correlation with the dependent variable and SibSp and Age has low negative correlation with the dependent variable.

Now let’s plot heat map to our data set

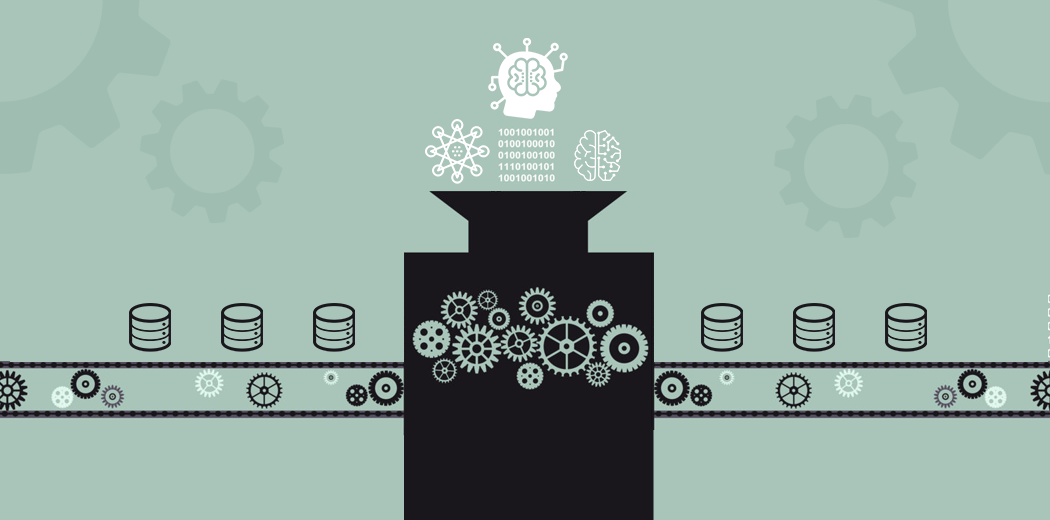


* From the heat map we will be able to tackle the multi co linearity in our data set if we see any multi co linearity we must get rid of it as it can effect the efficiency of our data set. If we find the multi co linearity between two columns we must drop either of the columns.
* We can see that none of the column has great co linearity with other columns so we can move forward

**EDA remarks** :

* We have missing values in 3 column with missing values. In column Age we have 177 missing values, in column Cabin 687 missing values and in Embarked we have 2 missing values.
* Column PassengerId and Name has unique value for all the rows. So the information from this columns is nill.
* Ticket has got 681 unique values and the information from this is not so satisfying.
* We have 5 columns with object type which has to be converted into numerical form.
* In our data set we saw that the outlier is present in some columns which has to be treated before the model building.
* Skewness is present in our data set and this skewed data can act as the outlier and this will decrease our efficiency of our model so we need to treat that.
* Scale of each column is different and id we have all the column in same range our model will give equal importance to all the columns.
* Class imbalance is present in our data set and it’s likely that our model will become biased towards majority class. So we need to balance our class in our dependent variable.

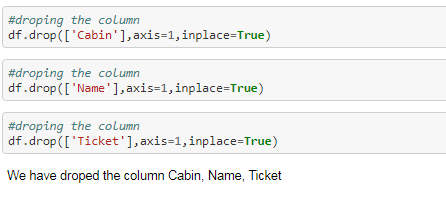
**Pre-processing Pipeline**

****

From EDA we have got to know that our data is semi-structured so we should make it structured before building the model.

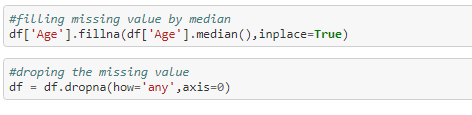
We have come across some less useful variable, class imbalance, outlier presence, skewness in our data, huge scale difference between each variables. Lets Treat one by one.

**Dropping the variables/columns:**



I have dropped the column Cabin, Name, Ticket as we are not getting more information from this variables.

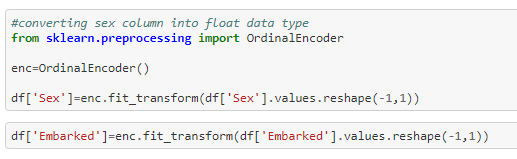
**Treating the missing values:**

****

I am filling the missing value in Age by median as we have some skewness in that columns and apart from that we had missing value in cabin so I have dropped that columns and we had 2 missing value in Embarked so I am dropping that 2 rows.

**Encoding:**

Encoding is nothing but converting the object type into numerical value after dropping the columns we are left with 2 columns with object type. For converting them I am using ordinal encoding method. And the code for that can be seen below.



## Outlier removal:

## Outlier may affect the efficiency of the model so I will see whether should I remove the outlier or not because removing the outlier would cause the data loss so we must be very careful has our data set is very small. I will be using the 2 method 1st Z-score method and 2nd IQR method.

## 

## From the Z-score method we can see that we are losing 7.9% method of data so we must most loos so much of data.

## Let’s try IQR method

## 

## From above we can see that we are losing almost half of our data. So, I will not remove the outlier I will be keeping all the entry of our data.

## Skewness removal:

## 

## Above we can see how much skewness is present. This skewed data act as outlier which in turn decreases the efficiency of our model so I have used the power transform and log transfer to remove the skewness and I have applied this method only on the independent variable.

## Scaling the data:

## 

## Scaling the data is very much important if we don’t scale down the data our model will give more importance to the variable with larger scale and less importance to variable with lower scale. But most of the time it won’t be true in most of the real world scenario. So, I have scaled down the data using the Standard scalar.

## Also I have splitted our model into x and y, x contains all the independent variable and y contains the dependent variable Survived.

**Building Machine Learning Models:**

## The Best Resources I Used to Teach Myself Machine Learning

Let’s import the entire library that is needed to build our model

## 

## And the model which I will be using are Random Forest Classifier Model, AdaBoostClassifier Model, DecisionTreeClassifier Model, KNeighborsClassifier Model.

## Now lets select the Best Random state

## 

## Here we can see that we have got best random state 81 with 80% accuracy using the DecisionTreeClassifier Model. I have splitted the data into train and test. Our test data size is 33% and our train data size is 67%.

## After that I have called the DecisionTreeClassifier Model and fitted my train data into it and predicted the test data. And for evaluation I have used the accuracy score which will calculate the difference between the actual value and the predicted value, We have got 80% correctness.

## Train, Test, Split:

## 

## I have splitted my data into Train and Test. Our train data contains 75% and test data contains 25%.

## Now let’s fit it into our model

## Random Forest Classifier Model:

## 

## 

## AdaBoostClassifier Model

## 

## DecisionTreeClassifier Model

## 

## KNeighborsClassifier Model

## 

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Accuracy Score | Cross validation score | Diff between acc and cross val |
| Random Forest Classifier Model | 84% | 80% | 4% |
| AdaBoostClassifier Model | 85% | 72% | 13% |
| DecisionTreeClassifier Model | 78% | 76% | 2% |
| KNeighborsClassifier Model | 77% | 79% | 2% |

## From the above 4 model we can see that the our model Random Forest Classifier Model has accuracy of 84% and deviation of 4%.So I have selected that as our best model and now let’s carry out the Hyper parameter tuning on that.

## 

## We have carried out the Hyper parameter tuning on the Random Forest and we gave got the best parameter. So let’s push this best parameter into Random Forest and build the model.

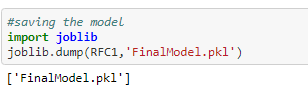
## 

## We can see that there is no such improvement in our model. So, let’s save base Random forest model as the best model. And now let’s draw roc auc curve

## 

Our Final model is **RFC1** with accuracy of 84.75% with Auc\_Score of 81.97%

**Saving the model:**



**Concluding Remarks:**

* We were able to passenger survival using Machine Learning.
* We were able to identify what all factor influenced the survival of the passenger.
* We were able to get the probability of survival rate.
* We were able to see which all class, sex people survived the most

**Link to the solution of this problem:**

[**https://github.com/rs6044922/DataTrained/blob/main/TITANIC\_Project/Titanic..ipynb**](https://github.com/rs6044922/DataTrained/blob/main/TITANIC_Project/Titanic..ipynb)

**Link to my LinkedIn profile:**

<https://www.linkedin.com/in/ravikumar-s-a874081b8>