**Titanic**

****

* [Introduction](https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic#introduction)
* Data Collection
  + Understanding of the dataset
  + will be the findings
  + Dataset Analysis
* Nominal/categorical
* Numerical
  + Null Values
  + Exploratory data analysis
* Count Plot
* Scatter Plot
* Heat map
* DistPlot
* Box Plot
  + Skewness
  + Outliers
  + [Prediction](https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic#prediction)
* LogisticRegression
* RandomForestClassifier
* DecisionTreeClassifier
* KNeighborsClassifier
* SVC
  + Hyper Parameter Tuning
  + ROC Curve
  + Saving the Model
  + [Conclusion](https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic#conclusion)

**Introduction –**The titanic ship was first underwater ship which was sun in North Atlantic Ocean on 1912. There were lots of people died and some survived. This was big accident of that time in the world. Here we are going to calculate how many people died and how many people survived.

**Data Collection -**  The dataset is available on different platform like Github and Kaggle, but I got this dataset from DataTrained.

**Understanding –** The dataset contain all details of the passengers. There were lots of people were died and some of survived. I have to build model to predict that people who were travelling in the ship, from them how many died and how many survived. This survival will predict with the help of passenger details.

**What will be the findings –** I have to find the how many passengers are live and how many died on the behalf of available dataset.

**Dataset Analysis –** The dataset has 12 columns and 891 row, including feature and target. There are some columns which does not has any relation with target. So, I will drop these column. These columns are Passenger Id, Name, Ticket, Cabin these column dose not has any relation with the target.

Now only 9 columns and 891 rows remain. These columns are mentioned below –

* **Survived**
* **Pclass**
* **Sex**
* **Age**
* **SibSp**
* **Parch**
* **Fare**
* **Embarked**

**Type of Data –**

**Nominal/categorical**– **Survived, Pclass, Sex, Sibsp, Parch and Embarked has Categorical data.**

* **Survived -** This feature has categorical values which show who is survived or who died. 1 means the person is survived and 0 means the person is died. This is target column and I have to use this column to predict.
* **Pclass-** Pclass is column which describe the class in the ship.
* **Sex -** sex is the categorical values which is Male and female. This will help to identify the survival person is male or female.
* **Sibsp –** This column has categorical values and show the siblings of the passengers.
* **Parch -** This column has categorical values.
* **Embarked -** This column has categorical values and shows the compartment in the ship.

**Numerical**: Age and Fare has numerical continuous data.

* **Age –** Age column has continuous values, the Age column show the age of the passengers.
* **Fare -**  Fare column has continuous values and fare also decide the category of the Compartment in the ship.

**Null Values -** I checked the null values in the dataset. The Age has 177 null values. Filled this null value with Simple Inputer with Mean of option and filled the null values.

**Categorical to Integer –** Change the categorical value to integer with the help of Label Encoder.

**Exploratory data analysis –** As most of the featureshas categorical values. So, first I plot the count plot ad checked what is the count of the value.

**Count Plot**

Count plot help to identify the count of the values in the column. this method is used to Show the counts of observations in each categorical bin using bars..

* **Survived -** This column has only two vales which is died or live. 0 represent died and 1 represent live. On the behalf of the dataset more people is died in the ship accident.
* **Pclass –** The Pclass has three values. These values represent the class in the ship. These 3 values show the 3 Category of the ship.
* **Sex -** This column has only 2 values which is male and female. 0 is for female and 1 is for male. More male passengers

Was travelling in the ship comparison to the female passengers.

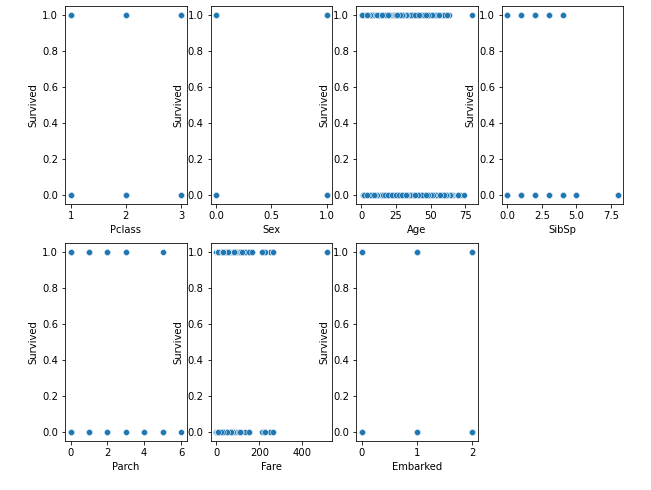
* **Age -** This column has continuous values. So, I cannot count the value.
* **SibSp –** This column represent the sibling and family member of the passengers. This column has 9 values which shows that how many sibling or family member of the passenger has and who sunk or who is swim in the water. Mostly passenger has no sibling or family member.
* **Parch-** This column has categorical value and depend on the SibSp column parch will assign to those who has sibling or family member. Parch also has the approx. same value to SibSp. If anyone has with them got the extra room according to the number of member. We can also say those who has family or sibling got the more chance to survive.
* **Fare-** This column has continuous values. So, I cannot count the value.
* **Embarked-** This column has categorical values and depend on the Fare column. Embarked has three column which has category of room. S class owed more passengers compared to the other passenger

**Scatter Plot**

A scatter plot is **a diagram where each value in the data set is represented by a dot**. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis.

X= All feature and y = Survived

* **Pclass-**The Pclass has directly relation with the target values.
* **Sex –** The Sex feature has relation with survival and I can say more male passenger has more survival rate comparison to the female.
* **Age –** The age column played good contribution to predict the survival. The age after the 50 lost their life in the accident and the age below 50 has mixed more chances to survive.
* **SibSp –** This column has more relation with target. Those passengers who has more sibling or family member had more probability to die.
* **Parch-** Parch column with 1-5 has good relation with target but more passenger died with 5 -7 has more probability to die.
* **Fare-**This column has good relation with Fare.
* **Embarked-** All Embarked class has more good relation with the target. All three category has trend with the target.

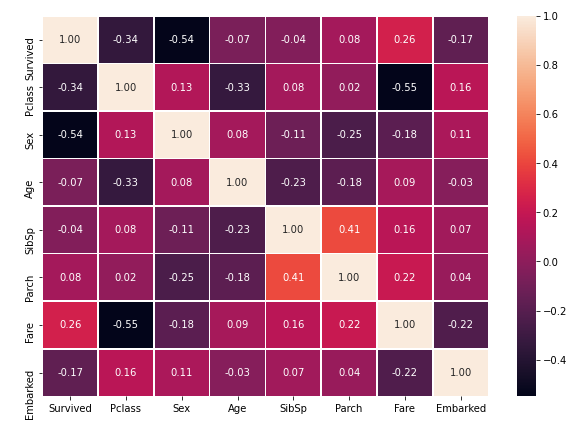
****

**Checking multicollinearity**

Multicollinearity occurs when two or more independent variables (also known as predictor) are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

No feature has multicollinearity with each other only Parch and Sibsp has some relation which is 41%. So, I do not need to worry about that.

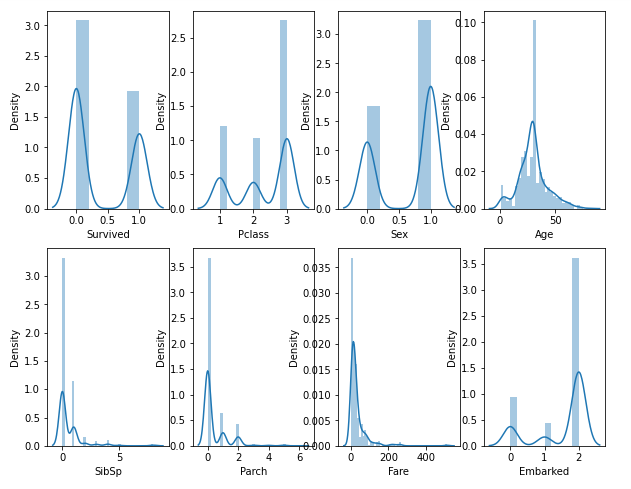
* **Pclass -**Does not has any multicollinearity with any feature.
* **Sex -** Does not has any multicollinearity with any feature**.**
* **Age -** Does not has any multicollinearity with any feature.
* **SibSp-** Has some multicollinearity with Parch which has maximum with all other and percentage is 41% which is acceptable to build model no need to drop any column.
* **Parch -** Has some multicollinearity with SibSp which has maximum with all other and percentage is 41% which is acceptable to build model no need to drop any column.
* **Fare-** Does not has any multicollinearity with any feature**.**
* **Embarked -** Does not has any multicollinearity with any feature.

****

**DistPlot**

A **Distplot**or distribution plot, depicts the variation in the data distribution. Seaborn Distplot represents the overall distribution of continuous data variables.

* **Survived-** I cannot define the Distplot in this column as this feature has categorical value.
* **Pclass -** I cannot define the Distplot in this column as this feature has categorical value.
* **Sex -** I cannot define the Distplot in this column as this feature has categorical value.
* **Age -** This column has good value distribution but has some Skewness, which is right skewed.
* **SibSp-** I cannot define the Distplot in this column as this feature has categorical value.
* **Parch -** I cannot define the Distplot in this column as this feature has categorical value.
* **Fare -** This column has good value distribution but has some Skewness, which is right skewed.
* **Embarked-** I cannot define the Distplot in this column as this feature has categorical value.

****

* **Checking the Statistical distribution of the dataset –** It is very important to check the Statistical values of the dataset. I check the stats of the dataset with Heat map. All column has normal stats like – Mean, 25%, 50%,75% and 100%. Only Fare column has some disturbance, the difference between Max and 75% is more than difference between others 25%, 50%. It might be outliers. I will check this with Boxplot and try to remove this.

**Skewness**

Skewness is **the measure of how much the probability distribution of a random variable deviates from the normal distribution.** The Skewness in the features value will mislead the model accuracy. So, checking the Skewness and remove the Skewness is very important. I check the Skewness and results are mentioned below –

Pclass -0.630548

Sex -0.618921

Age 0.434488

SibSp 3.695352

Parch 2.749117

Fare 4.787317

Embarked -1.264823

Some features has Skewness and we need to remove this Skewness. This I removes the Skewness with the Power Transformer. Skewness is removed only for continuous values and results are mentioned below –

Pclass -0.441438

Sex -0.618921

Age 0.064420

SibSp 0.808608

Parch 1.228795

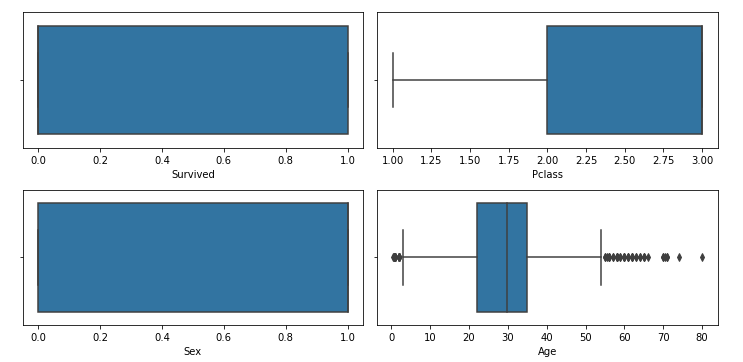
Fare -0.040329

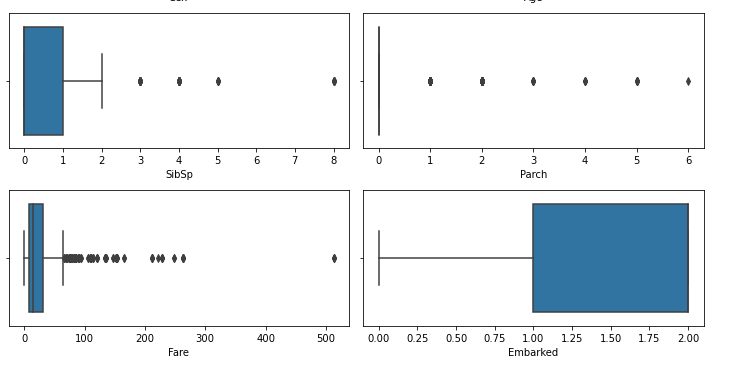
Embarked -1.064527

**Boxplot for Outliers**

It is **a type of chart that depicts a group of numerical data through their quartiles**. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

* **Survived –** This column has no outliers and all set to proceed further.
* **Pclass-** This column has no outliers and all set to proceed further.
* **Sex -** This column has no outliers and all set to proceed further.
* **Age –** This column has so much outliers. I will check and remove this.
* **SibSp-** This column has so much outliers. I will check and remove this.
* **Parch -** This column has some outliers. I will check and remove this.
* **Fare-** This column has so much outliers. I will check and remove this.
* **Embarked-** This column has no outliers and all set to proceed further.

****

****

**Removing the Outliers –** As some of the feature has outlier I checked and removed the outliers with help of Z-score.

**Now the data is clean are ready to build model**

Separate the dataset into feature and target. Now dataset is ready to build the model import the basic libraries and models. I am going to build below mentioned model and perform the accuracy and other model matrix.

**LogisticRegression**

**RandomForestClassifier**

**DecisionTreeClassifier**

**KNeighborsClassifier**

**SVC**

**I performed 5 model prediction Accuracy score, those score are mentioned below –**

**Accuracy score for LogisticRegression model 81.70731707317073**

**Accuracy score for RandomForestClassifier model 80.48780487804879**

**Accuracy score for DecisionTreeClassifier model 76.82926829268293**

**Accuracy score for KNeighborsClassifier model 81.70731707317073**

**Accuracy score for SVC model 81.70731707317073**

From the above, I conclude that I will go forward with KNeighborsClassifier because three model has same accuracy score but we can do better hyper parameter tuning with KNeighborsClassifier. May the Accuracy will increase.

**Hyper Parameter Tuning**

Hyper parameter determining the severity of the penalty. As the value of the penalty increases, the coefficients shrink in value in order to minimize the cost function

Accuracy score for after tuning parameter of KNeighborsClassifier model 84.14634146341463

Hyper parameter tuning increased the score of the model now the accuracy of the **model is 84.14 %. Now checking the classification report, Confusion matrix and F1 Score.**

**Classification report**

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model. Classification report of the KNeighborsClassifier is mentioned below –

**Confusion matrix**

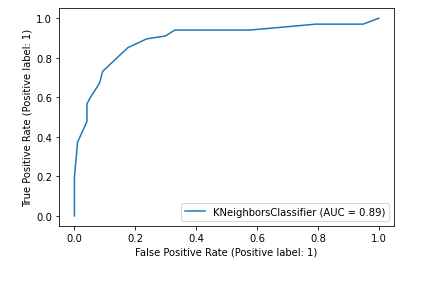
In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Confusion matrix of the KNeighborsClassifier is mentioned below –

**[[88 9]**

**[17 50]]**

**ROC AUC plot**

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. This helped to know the model trend with noise. So, the ROC curve for KNeighborsClassifier is mentioned below –



The ROC curve accuracy for the KNeighborsClassifier is very good which is 84% and this show that the model will perform better with input feature.

**Saved the model with JobLib**

**Conclusion**

In the whole process of the model building I imported the data, cleaned the data and build the model. I found that many more people died in this accident. There are 7 feature and no one is strongly has relation with the target but every feature has some relation with model.

The male passengers lost more lives compared to the female. Those passengers who are older than 50 years, lost their lives compared to the other passengers.

The passenger who has more siblings and family member, lost less live compared to those who don’t has any siblings or family. Passengers who bought the costly ticket got the better cabin and saved their lived more than others.

I can say all the feature has some contribution to predict the model. I cannot say that any specific feature contribute more than other. I build 5 classification model as I have to predict the survival of the passenger which died or live. In this scenario I can only perform classification model. All the model are performed very well but KNeighborsClassifier model performed very well and when I tune the parameter of the KNeighborsClassifier than I got best accuracy which is 84%. The 84% accuracy should be considered as very good model Accuracy.

The model accuracy say that it will predict 84% accurate survival and 16% predication will be false. Overall the model is performing very well.

Thanks   
Shashwat