Titanic Project

DataTrained

Submitted by: Prashant Pathak

Batch No.: DS0522

Gmail: rudraksha7505021514@gmail.com

1. Problem Definition:

Titanic British luxury passenger liner that sank on April 15, 1912, en route to New York from Southampton, England, on its maiden voyage.

The largest and most luxurious ship afloat, the *Titanic* had a double-bottomed hull divided into 16 watertight compartments. Because four of these could be flooded without endangering its buoyancy, it was considered unsinkable. Shortly before midnight on April 14, it collided with an iceberg southeast of Cape Race, Newfoundland; five compartments ruptured, and the ship sank. Some 1,500 of its 2,200 passengers died.

After the disaster, new rules were drawn up requiring that the number of places in lifeboats equal the number of passengers (the *Titanic* had only 1,178 lifeboat places for 2,224 passengers) and that all ships maintain a 24-hour radio watch for distress signals (a ship less than 20 mi [32 km] away had not heard the *Titanic*'s distress signal because no one had been on duty). The International Ice Patrol was established to monitor icebergs in shipping lanes. In 1985 Robert Ballard found the wreck of the *Titanic* lying upright in two pieces at a depth of 13,000 ft (4,000 m). American and French scientists explored it using an uncrewed submersible.

So far, we know what Titanic was and what happened to him but now we discuss about in prediction problem, and we have this problem statement -

Problem Statement:

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.

Titanic train dataset file download link:

https://github.com/dsrscientist/dataset1/blob/master/titanic_train.csv

GitHub Solution link in ipynb format: Practice-Projects/Titanic Survived Prediction (1).ipynb at main · prashantpathakji/Practice-Projects (github.com)

2. Data Analysis:

First step is import required library -

- NumPy
- Pandas
- Matplotlib
- Seaborn
- Trian test split
- Logistic Regression
- Accuracy Score
- Decision tree
- RandomForestClassifier

In the next step we'll download titanic dataset from the link and upload dataset in our Jupiter notebook for analysis.

Now we'll use head, tail, and shape keywords for checking some basic information of dataset. Apart from that we use some more keywords for checking duplicate, unique, and Nan values in our dataset.

After this we'll handle the missing values and remove outliers from our dataset. Apart from that we'll visualize the data with the help of graphs. We'll also use describe keyword for knowing basic statistical information of dataset and will check correlation

also using heatmap.

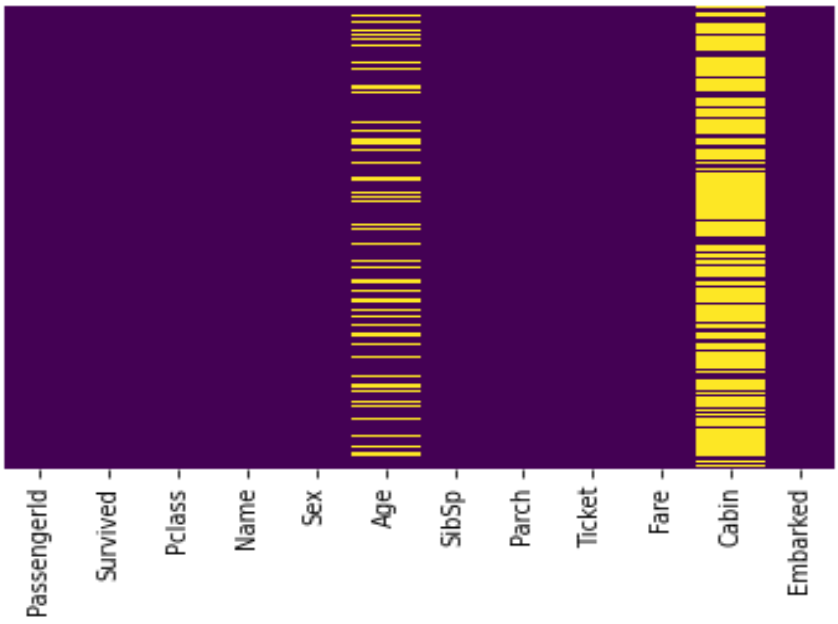
3. EDA Concluding Remark:

We have 891 rows and 12 columns in our dataset, and we don't have any duplicate value in our dataset.

The name of all columns is;

1. PassengerId- It's unique id of passengers. In this column the Starting point is 1 and ending point is 891.
2. Survived – in this column we have 2 categories 0 and 1. [0 = not Survived, 1 = Survived]
3. Pclass- in this column we have 3 classes. [1=first class,2=second class, 3=third class]
4. Name – in this column we have the names of passengers.
5. Sex- in this column we have 2 categories male and female.
6. Age- in this column we have the age of passengers.
7. SibSb- in this column we have information about passenger's siblings.
8. Parch- in this column we have 7 categories.
9. Ticket- in this column we have the ticket number of passengers.
10. Fare
11. Cabin
12. Embarked

Age column has 177 missing values and Cabin column has 687 missing values and Embarked column has only 2 missing values.



Before handling missing values.

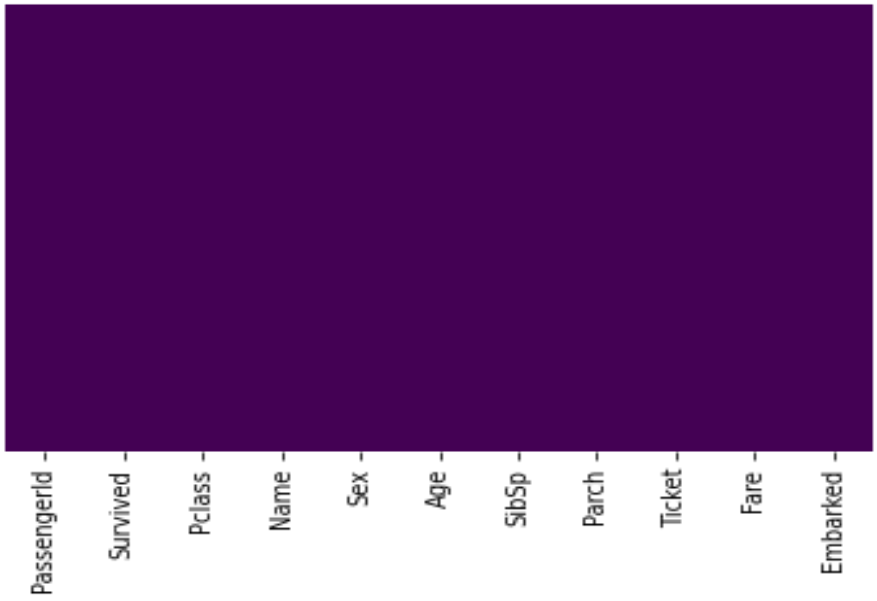
Now we'll handle the missing values:

Replace the missing value of age column by mean of age and

replace the missing value of the "Embarked" column with mode value.

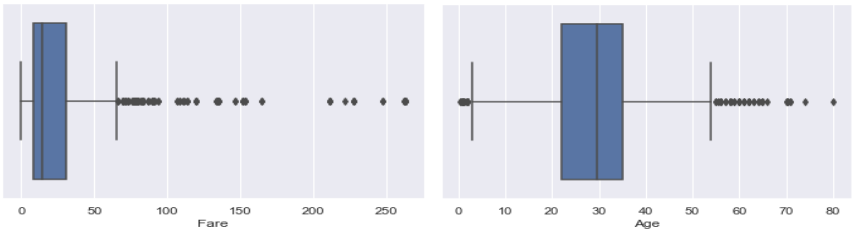
After this,

Drope the "Cabin" column from the Dataset because it has too many missing values.



After handling missing values.

After handling missing values we'll handle outliers:



Data type of every column:

Categorical:

- **Nominal (**variables that have two or more categories, but which do not have an intrinsic order.)
    - **Cabin**
    - **Embarked (**Port of Embarkation)

C(Cherbourg)
Q(Queenstown)
S(Southampton)

- **Dichotomous (**Nominal variable with only two categories)
  - **Sex** Female
    Male
- **Ordinal (**variables that have two or more categories just like nominal variables. Only the categories can also be ordered or ranked.)

  * **Pclass** (A proxy for socio-economic status (SES))

    1(Upper)

    2(Middle)

    3(Lower)

Numeric:

- Discrete
  - Passenger ID (Unique identifying # for each passenger)
  - SibSp
  - Parch
  - Survived (Our outcome or dependent variable)

    0
    1
- Continuous
  - Age
  - Fare

Text Variable:

- **Ticket** (Ticket number for passenger.)
- **Name** (Name of the passenger.)

Some Statistical information about in Dataset:

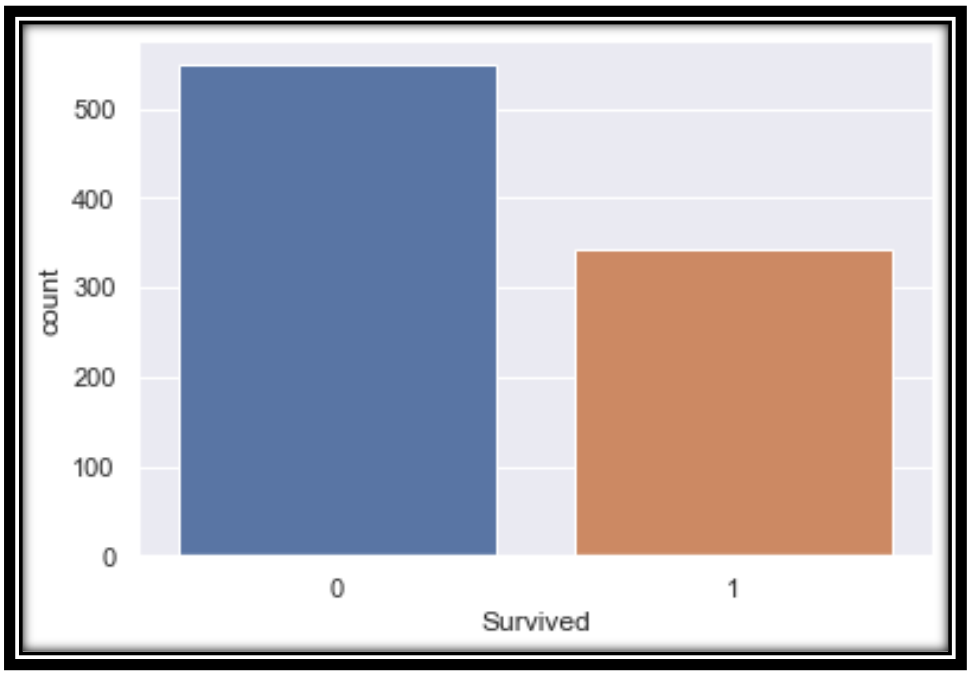
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	31.364716
std	257.353842	0.486592	0.836071	13.002015	1.102743	0.806057	43.257927
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	22.000000	0.000000	0.000000	7.910400

50%	446.000000	0.000000	3.000000	29.699118	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	263.000000

Visualization part:

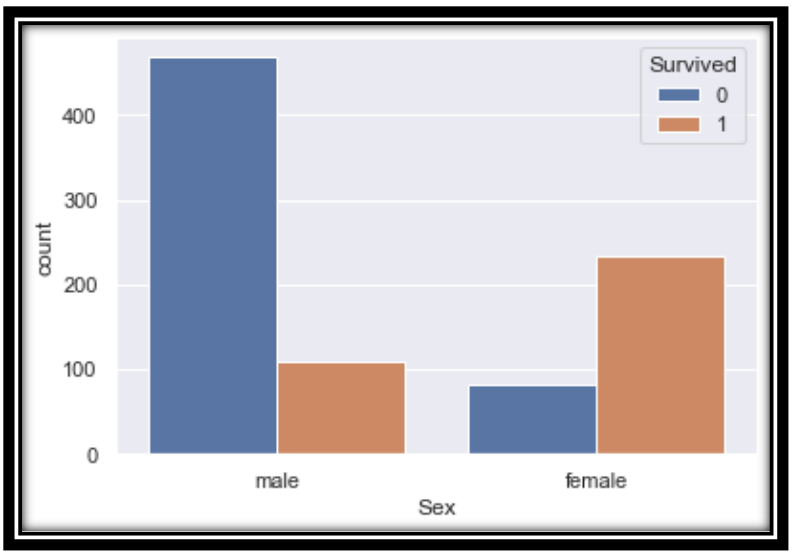
Survived column is our target column.

(i) Number of Survived – 342 (ii) Number of not Survived – 549

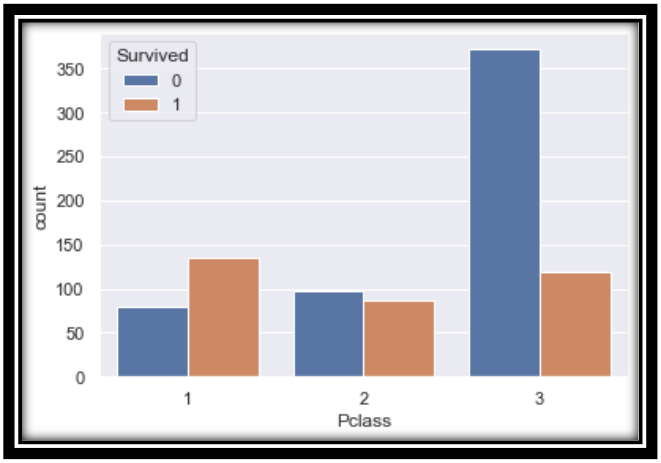


Result:- Survived passengers are less than in compare of not survived passengers.

577 male and 314 female are in our data set.



Result:- Number of Survived females are Greater the number of survived males.



We have 3 pclass in our ship and count of passengers in every single class is-

First class – 216

Second class – 216

Third class - 491

Result: - More passengers died in third class in compare of first class and second class.

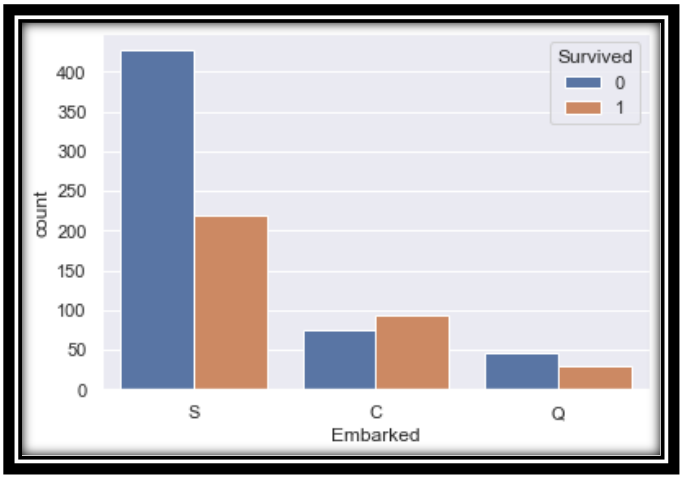
Embarked column analysis:

Count of S = 168

Count of C = 168

Count of Q = 168

Result: - Survived Passenger (S>C>Q)



Correlation using heatmap:

4. Pre-Processing Pipeline:

Separating column into training data and testing data and drop 'PassengerId','Name', and 'Ticket' column because these columns are useless for prediction.

We make 2 separate data frames, their names x and y.

x is our training dataset and y is our testing dataset.

```
x=titanic_df.drop(columns=['PassengerId','Name','Ticket','Survived'])
```

```
y=titanic_df['Survived']
```

After this we use train test split using this code:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20,random_state=1)
```

5. Building Machine Learning Models:

We'll train 3 ML Models **(i) Logistic Regression as LR**
(ii) Decision Tree as DT

(iii) Random forest as RF

6. Concluding Remarks:

After testing our all models, we have some result

```
**************************************************
 ------------------------------------------------
 Final Result of LR:
 Accuracy Score for Traning Data -  0.8033707865168539
 Accuracy Score for Test Data-  0.7988826815642458
 ------------------------------------------------
 ------------------------------------------------
 Final Result of DT:
 Accuracy Score for Traning Data -  0.9873595505617978
 Accuracy Score for Test Data DT-  0.7988826815642458
 ------------------------------------------------
 ------------------------------------------------
 Final Result of RF:
 Accuracy Score for Traning Data -  0.9873595505617978
 Accuracy Score for Test Data-  0.7988826815642458
 ------------------------------------------------
```

Final Result:- Logistic Regression is perfect model for prediction because accuracy score is for traning data and testing data is closer to each other in comparison of another models.

Thankyou

DataTrained
Keep Skilling, Keep Growing