Titanic Prediction of Survival of Passengers

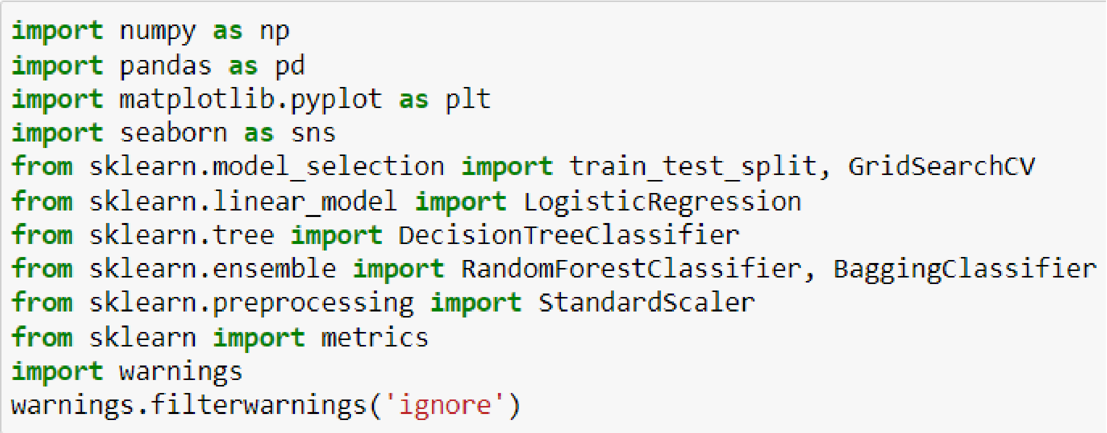
By : Ashutosh Chaudhary

Problem Statement :

To predict that whether passenger survived or not that were sailing on Titanic ship on the basis information present in various other columns in dataset.

Step Taken For Analysis :

1. Importing Libraries For Task To get Completed :

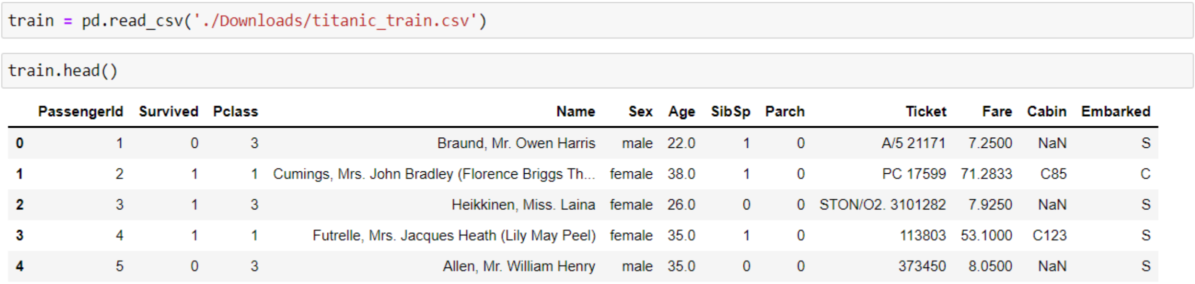


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, BaggingClassifier
from sklearn.preprocessing import StandardScaler
from sklearn import metrics
import warnings
warnings.filterwarnings('ignore')
```

We imported all the above mentioned libraries, so, we can get the task completed & achieved our objective of making the survival prediction.

2. Importing Dataset :

We imported the dataset using pandas library.



```
train = pd.read_csv('./Downloads/titanic_train.csv')
```

```
train.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

After importing dataset & observing it, we concluded that target feature is named Survived in it.

3. Checking For Dimensions, Description & Null Values Presence in Dataset :

By doing this we get to know this about dataset :

- Dataset contains 891 columns & 12 columns.
- From description we get to know about mean, max value, etc of every column present in dataset containing continuous data.
- There is pressence of null values in Age, Cabin and Embarked columns in the dataset.

4. EDA :

We plot the graphs of every column in dataset one by one to gather information from it to make prediction :

PassengerId :

After looking at values present in PassengerId & plotting graphs we concluded that they were just numbers given to people like in index & does not affect target feature in anyway, so, we dropped it.

Pclass :

After looking at both graphs of Pclass we concluded that most passengers are present class 3 & every category of Pclass have both categories of Survived.

Name :

After looking at values present in Name we concluded that we can extract their initials to refine the information by creating another column named Initial & then categories initial into 4 categories (Mr., Mrs., Miss & Master) & plotted the graphs using it to conclude that most passengers present are of Mr. initial & every category of initial are present in both categories of Survived.

Sex :

After looking at both graphs of Sex we concluded that most passengers present on ship are male sex & both categories of Sex are present in both categories of Survived.

Age :

As we know that there null values are present in this column & we filled it on the basis of Initial. After that we plot the graph and concluded that data is positively skewed & is having highest density at around 37. Also we conclude that most of female passengers survived & most male passengers did not survived.

SibSp & Parch :

We combined both of these columns to create 2 another columns which are Family & Alone. Then we plot graphs for both new columns & concluded that most passengers on the ship are present with no family members with them.

Ticket :

After observing the values in column we concluded that it is of no use & neither can we extract any useful information from it. So, we move to next column.

Fare :

After looking at both graphs of Fare we concluded that data is positively skewed with having highest density at around 10 or 15. We also concluded that all range of Fare is present in both categories of Survived. Also most females survived & most male does not survived.

Cabin :

We know that there is presence of null values present in this column & we filled them randomly to not create data bias. After that we refined the information by categorising information on the basis of first letter in every value of column & then we plotted the graphs. We concluded that most passenger lives in Cabin C & every category of Cabin is present in both categories of Survived.

Embarked :

We know there are only 2 null values present in this column , so, we filled them with the mode value of column & then we plot the graphs. After observing at graphs we concluded that most passenger embarked from S & every category of Embarked is present in both categories of Survived.

5. Changing Datatypes :

We changed the data types of every column to int or float data types to create heatmap.

6. Creating Heatmap :

We created heatmap to check the correlation of every column with target column. From we concluded that target column has high positive correlation with Sex, Fare, Embarked & Fare columns & target column also high negative correlation with Family. After that we dropped PassengerId, Name, SibSp, Parch & Ticket columns to reduce multi collinearity.

7. Splitting Data Into Train & Test Set :

We splitted data into train & test set using train_test_split with test size being 25% of total data, so that we can check various models performance on it to find best suited model for the dataset.

8. GridSearchCV :

We used on GridSearchCV on various models to find best suited hyper parameters while working on the training dataset & models we used on this dataset to check are :

- Logistic Regression Model
- Decision Tree Classification Model
- Random Forest Classification Model
- Bagging Classification Model

Out of all the models we checked, it was observed that best training score is given by Random Forest Classification Model & we used it make prediction on testing dataset.

9. Using Random Forest Classification Model On Test Dataset :

From GridSearchCV we get to know that best parameters for Random Forest Classification model are :

Criterion = 'gini'

max_depth = 5

min_samples_leaf = 2

min_samples_split = 3

n_estimators = 120

Using these parameters we create Random Forest Classification Model named (rf) & used it to make prediction on dataset.

10. Scaling the Training & Testing Dataset :

We scaled both training & testing dataset to remove outliers before applying model on it.

11. Applying Model on Dataset :

After scaling dataset we applied the model on it to make prediction & then we create dataframe comparing actual & prediction values.

12.Using various metrices on model :

We used metrices to check performance of Random Forest Classification Model :

- Accuracy score : 0.7982
- Precision score : 0.6210
- Cohen Kappa score : 0.5716
- Confusion matrix : [119, 36
9, 59]

It is showing the model is performing quite well on testing dataset.

13. Saving Model :

After checking performance of model we saved model using pickle library & task is completed.