Kaggle Group Challenge – Spaceship titanic

# Brief:

The problem statement is to determine which passengers were successfully transported to another dimension.

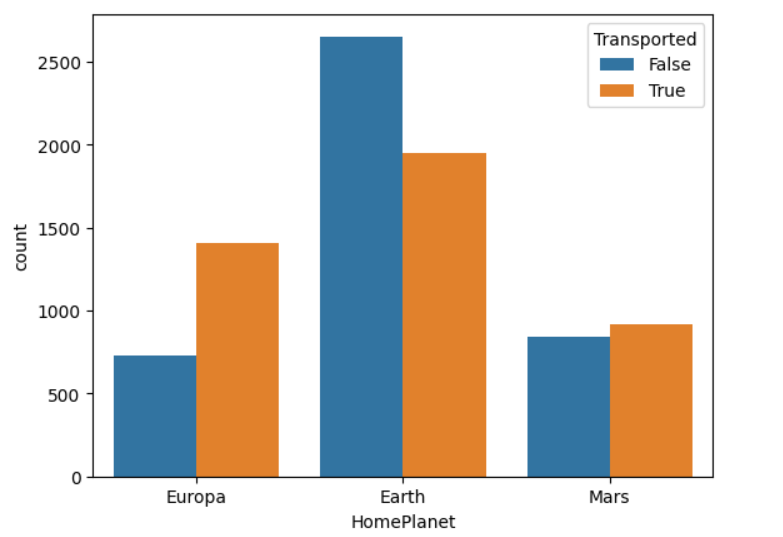
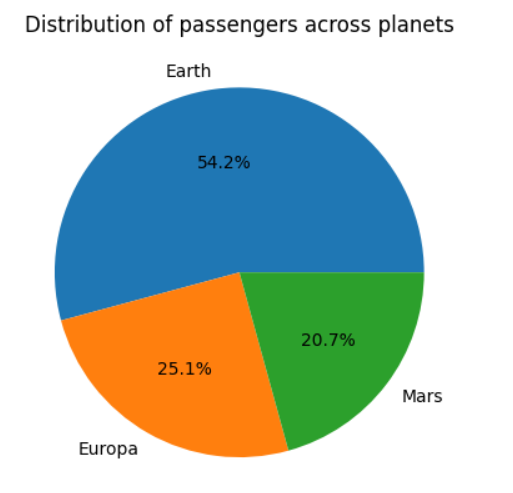
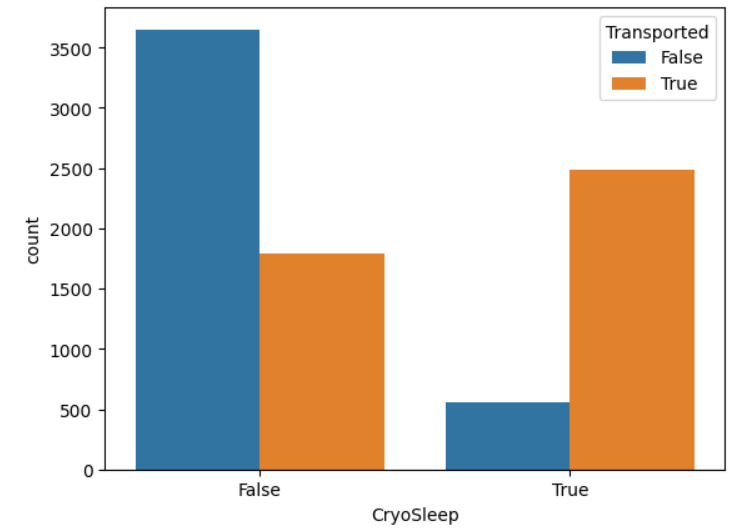
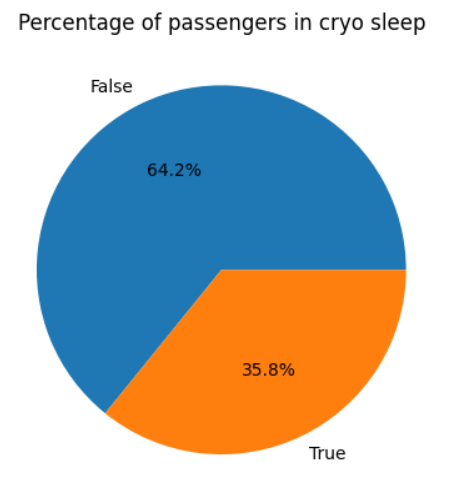
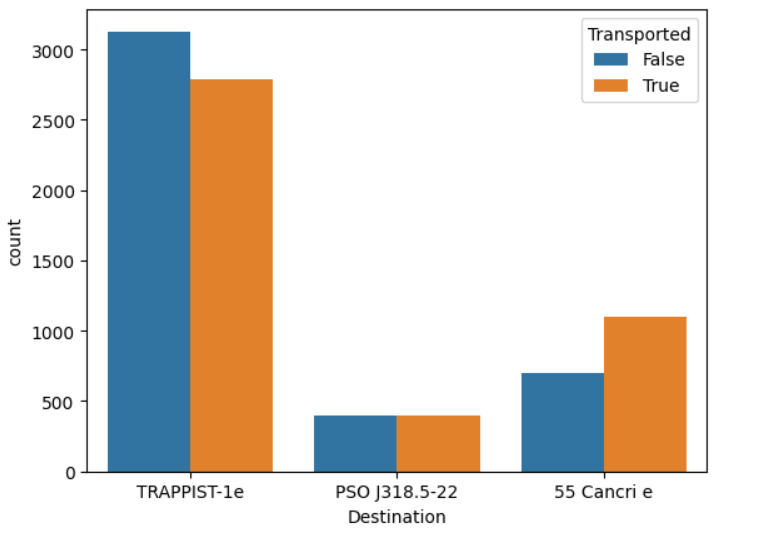
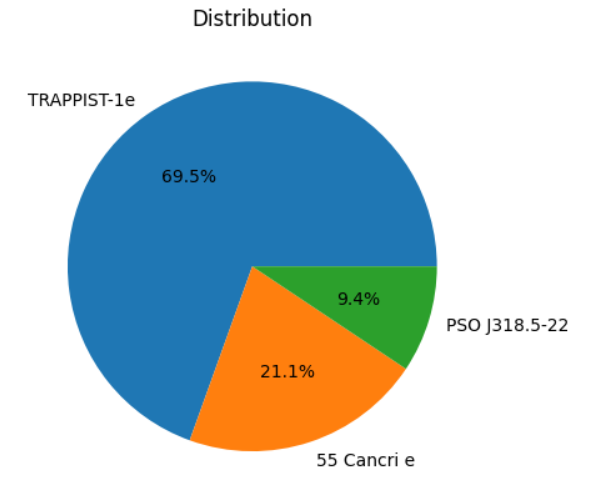
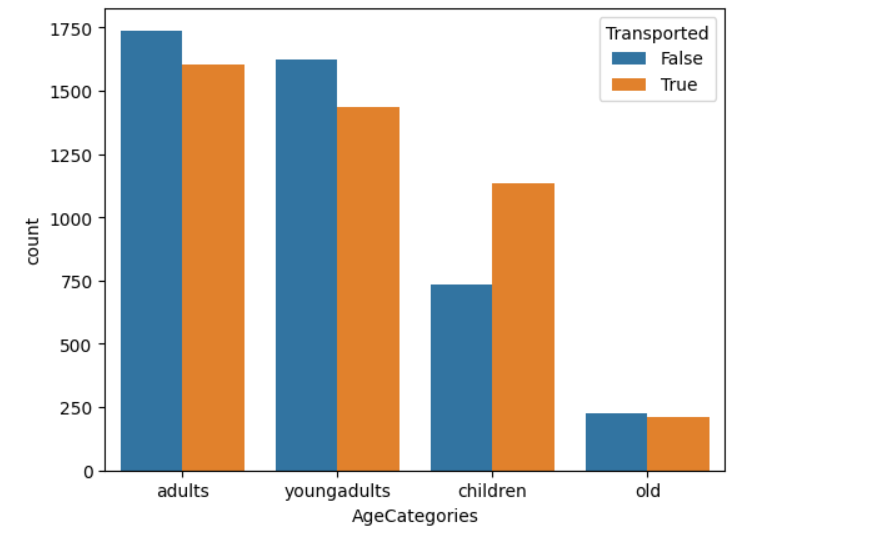
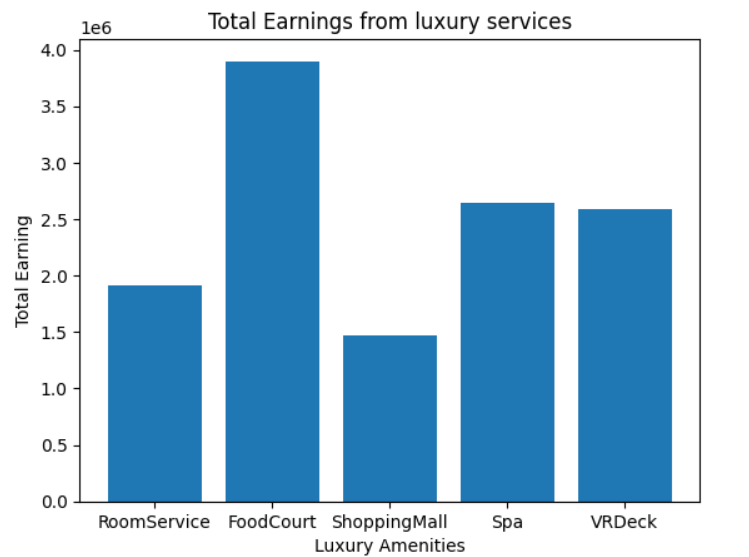
# Data:

The data contains the following columns. The transported column is the target and the rest are features.

1. PassengerId - A unique Id for each passenger. Each Id takes the form gggg\_pp where gggg indicates a group the passenger is travelling with, and pp is their number within the group. People in a group are often family members, but not always.
2. HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
3. CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
4. Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for Port or S for Starboard.
5. Destination - The planet the passenger will be debarking to.
6. Age - The age of the passenger.
7. VIP - Whether the passenger has paid for special VIP service during the voyage.
8. RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.
9. Name - The first and last names of the passenger.
10. Transported - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.

# Exploratory Data Analysis:

The data given has some interesting patterns, few of which are discussed below:

1. How many distinct planets are there and how many people boarded from each?  
   Maximum passengers are from planet earth; the survival rate is more for passenger’s form Europa.
2. Survival of passengers in cryosleep.  
     
   We see that around 36% of the passengers are in cryosleep and majority of them have survived when compared to the other class.
3. How are passengers of different destinations affected?  
     
   More passengers are travelling to TRAPPIST-1e, and the survival rate is more in 55 Cancri e.
4. Relation between age and survival.  
     
   The transported ratio is more for children among all other categories.
5. Max earning luxury amenity  
     
   The amenity the passengers spend most money on is food followed by VRDeck and Spa.

# Preprocessing – Feature Engineering

The data contains numerical and categorical columns, which have some missing values. As part of preprocessing, meaningful columns have been derived from composite columns like cabin.

The numerical columns having null values are imputed with median.

The categorical columns having null values are imputed using the mode of the values of a column.

The categorical columns having ordinal values are encoded using ordinal transformation like cabinDeck.

Other categorical columns are one-hot-encoded for training purposes.

# Modelling

This is a binary classification use case, where transported column is the target and takes values 1 or 0.

The XGBoost classifier is used for the final predictions for this use case.

Detail into the model:

The XGBoost model is Extreme gradient boosting algorithm, which is an ensemble of decision trees and gradient boosting.

It is more efficient when compared to other bagging algorithms as it iteratively learns to adapt from new data and can process faster.

Other methods tried:

1. A logistic regression model used; the accuracy was not satisfy-able.
2. RandomForestClassifier is created but when compared to XGBoost the performance was less.

The training data is divided into train and validation datasets.

The training is done by importing the xgboost class on the training data with default parameters offered by the library.

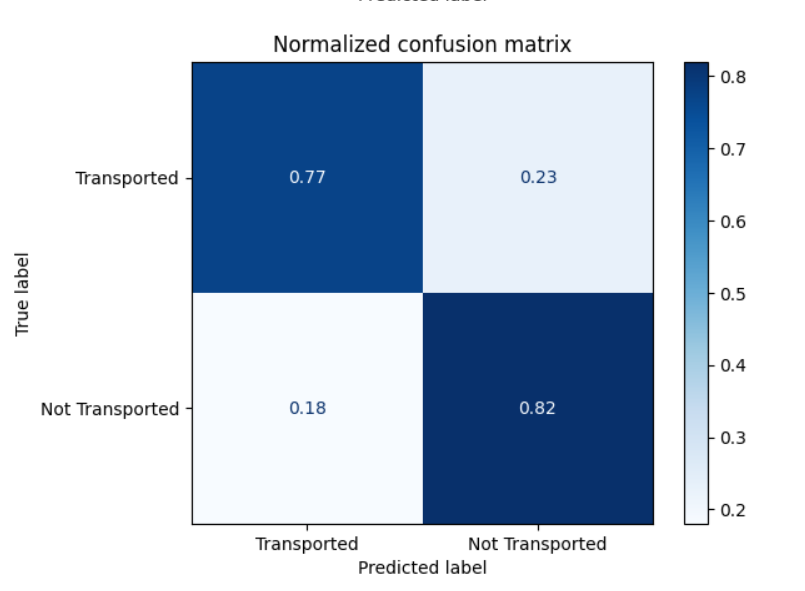
Its objective is set to binary classification as per the use case requirement.

The evaluation metrics are set as Area under the curve of precision and recall, which helps evaluation for imbalanced datasets, with a lot of positive samples.

The model is trained with early stopping In order to avoid overfitting. The model stops training if there is no improvement in the evaluation metric.

# Results analysis

The model is successfully trained and predicted on the test dataset. We are considering confusion matrix for our analysis which would help us better understand the behavior of the binary classification using precision and recall.



We see that our base model correctly predicted 77% of the transported passengers as being transported while incorrectly predicted 23% of the transported passengers as not being transported. Conversely, it predicted correctly predicted 82% of the passengers that were not transported as not being transported and incorrectly predicted that 18% of the passengers that weren't transported as being transported.

This performance of the model can be further increased by fine tuning the hyper parameters of the model to fit our current data.

# Important Links:

CODE LINK: <https://colab.research.google.com/drive/1NEZzNyt3bO3bmOfjacfO-MyO_r_gxgxy?usp=sharing>

REPO LINK: <https://github.com/tejachalmala/Kaggle-Titanic-Spaceship-Survival.git>