High-Level Design (HDL)

Spaceship Titanic Analysis Prediction

Table of content

SPACESHIP TITANIC ANALYSIS PREDICTION

1. INTRODUCTION:

Welcome to the year 2912, where your data science skills are needed to solve a cosmic mystery. We've received a transmission from four lightyears away and things aren't looking good.

The Spaceship Titanic was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.

While rounding Alpha Centauri en route to its first destination—the torrid 55 Cancri E—the unwary Spaceship Titanic collided with a spacetime anomaly hidden within a dust cloud. Sadly, it met a similar fate as its namesake from 1000 years before. Though the ship stayed intact, almost half of the passengers were transported to an alternate dimension!

To help rescue crews and retrieve the lost passengers, you are challenged to predict which passengers were transported by the anomaly using records recovered from the spaceship's damaged computer system.

Help save them and change history!

2. PROBLEM STATEMENT:

In this article, we will try to solve one such problem which is a slightly modified version of Titanic which is the Spaceship Titanic. The problem statement of this project is like a spaceship having people from different planets on a voyage but due to some reasons, some people have been transported to another dimension. Our task is to predict who will get transported and who will remain on the spaceship.

3. TOOLS USED

Python libraries make it easy for us to handle the data and perform typical and complex tasks with a single line of code.

Pandas – This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.

Numpy – Numpy arrays are very fast and can perform large computations in a very short time.

Matplotlib/Seaborn – This library is used to draw visualizations.

Sklearn – This module contains multiple libraries are having pre-implemented functions to perform tasks from data preprocessing to model development and evaluation.

XGBoost – This contains the eXtreme Gradient Boosting machine learning algorithm which is one of the algorithms which helps us to achieve high accuracy on predictions.

4. PROBLEM STATEMENT:

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms, the best fit for the above case

Operation performed:

1. Data Collection:

Data is taken from kaggle which is a Spaceship Titanic

2. Data Cleaning:

This dataset consists of some missing values and outliers. The categorical missing value are filled with most frequent value i,e. mode while the numerical missing value are filled with mean.

3. Data visualization:

Visualization are performed using the libraries like Matplotlib and Seaborn by plotting Histogram and Countplots, etc.

4. Feature Engineering:

We use get_dummies and label encoder to convert categorical feature to numerical data.

5. Model Training:

Comparing various classification model by training them on the modified dataset.

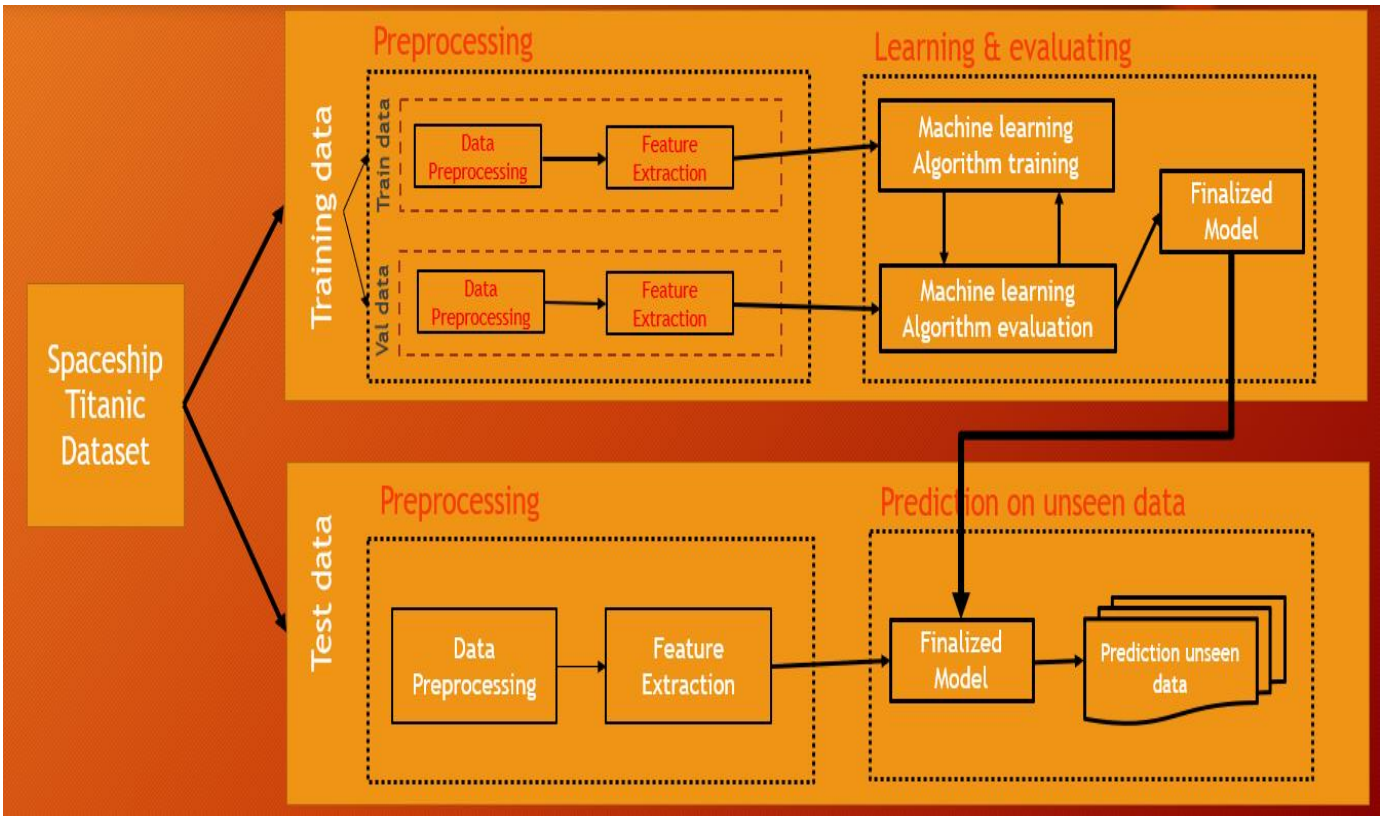
6. Hyperparameter Tuning:

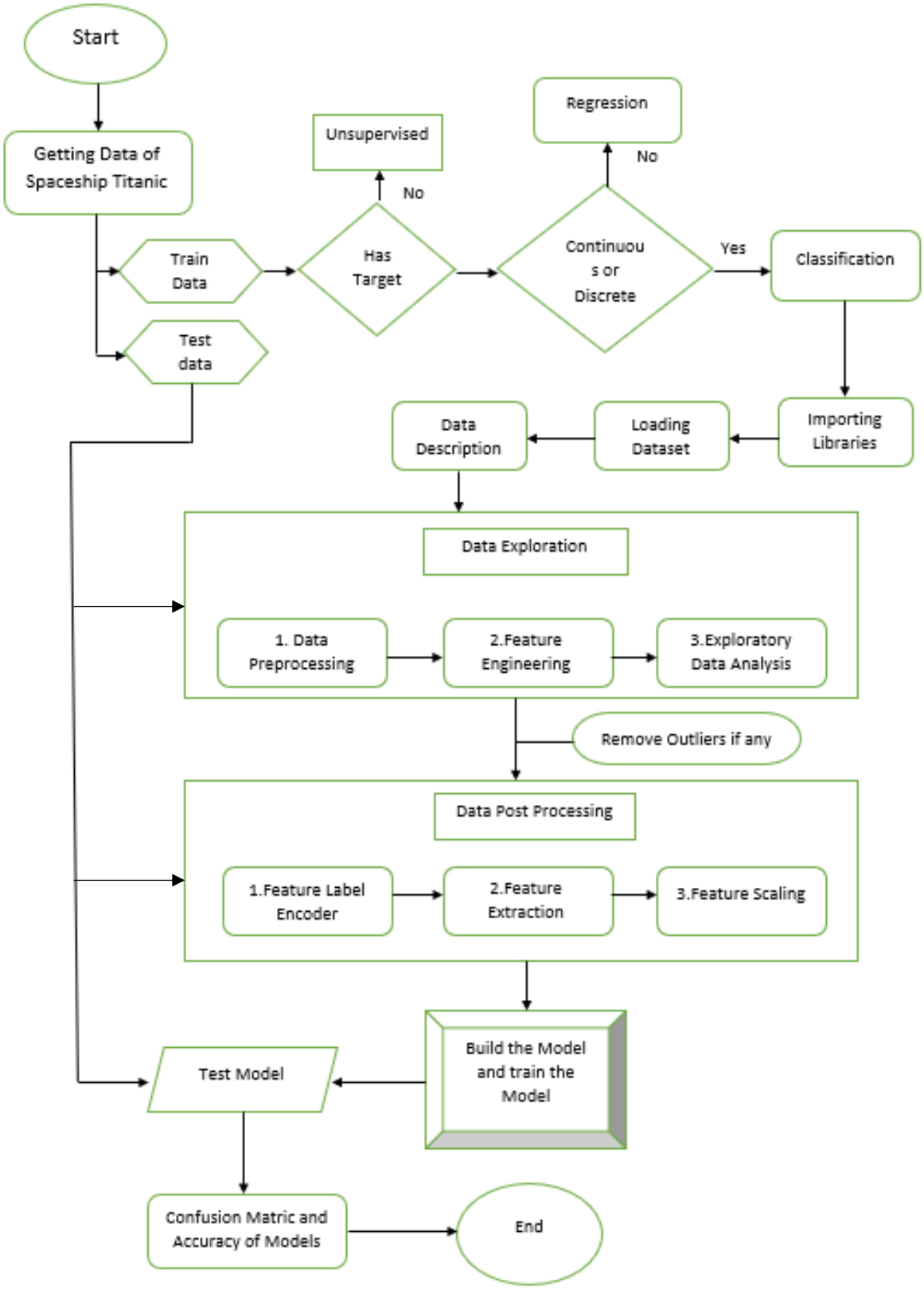
To gain more accuracy on the data we perform hyperparameter tuning. This process take much time but can give us better results.

7. Model Evaluation and Selection:

We choose Our model based on confusion matrix , f1 score and Classification report.

5. WORKFLOW DIAGRAM:




Start
Getting Data of Spaceship Titanic
Train Data
Test data
Has Target
Unsupervised
No
Continuous or Discrete
Regression
No
Yes
Classification
Importing Libraries
Loading Dataset
Data Description
Data Exploration
1. Data Preprocessing
2.Feature Engineering
3.Exploratory Data Analysis
Remove Outliers if any
Data Post Processing
1.Feature Label Encoder
2.Feature Extraction
3.Feature Scaling
Build the Model and train the Model
Test Model
Confusion Matric and Accuracy of Models
End

6. DATASET DESCRIPTION:

Dataset

The dataset is taken from Kaggle Competition Link:-

Click Here: https://www.kaggle.com/competitions/spaceship-titanic/overview

We need to predict whether the passenger is transported or not for the test data as follow.

Features:

- 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.
- HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
- 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.
- 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.
- Destination - The planet the passenger will be debarking to.
- Age - The age of the passenger.
- VIP - Whether the passenger has paid for special VIP service during the voyage.
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.
- Name - The first and last names of the passenger.
- Transported - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.

7. METRICS OF CHOICE/KPI's:

When selecting a metric for a binary classification problem, it is important to consider the specific goals and requirements of the project, as well as the characteristics of the data. Here are some general guidelines for selecting a metric:

I. Accuracy: This is a simple and commonly used metric that measures the proportion of correct predictions made by a model. However, accuracy can be misleading when the data is imbalanced, where one class has many more observations than the other. In such cases, precision, recall, or F1 score might be more informative.

II. Precision: This measures the proportion of true positive predictions among all positive predictions made by a model. It is useful when the goal is to minimize the number of false positives.

III. Recall: This measures the proportion of true positive predictions made by a model among all actual positive observations. It is useful when the goal is to minimize the number of false negatives.

IV. F1 Score: This is the harmonic mean of precision and recall, and it balances both measures. It is useful when the goal is to have a balance of precision and recall.

V. AUC-ROC: This is a measure of how well a binary classifier can distinguish between two classes. It is useful when the goal is to maximize the ability to correctly classify positive observations.

8. PREDICTION RESULTS FOR UNSEEN DATAPOINTS

Passenger Id	Transported
0013_01	TRUE
0018_01	FALSE
0019_01	TRUE
0021_01	TRUE
0023_01	TRUE
0027_01	FALSE
0029_01	TRUE
0032_01	TRUE
0032_02	TRUE

9. CONCLUSION:

We started with the data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important. During this process we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we computed missing values, converted features into numeric ones, grouped values into categories and created a few new features. Afterwards we started training 8 different machine learning models, and for each model we did hyperparameter tuning and checked with the confusion matric and with

classification report which is consist of precision, recall and f-score picked one of them that is Gradient Boosting. But in this model is over fitted so to improve or to avoid this we can use any one of the methods which is mention in the future scope of improvements.

10. FUTURE SCOPE OF IMPROVEMENTS:

As we can see the prediction of the data set of the finalized model, the model is overfitted that is the model is trained well but failed while predicting for the unseen data, so to avoid or to improve the prediction in future days we can use,

- Cross-validation
- Training with more data
- Removing features
- Early stopping the training
- Regularization
- Ensembling

11. REFERENCE:

https://betterprogramming.pub/titanic-survival-prediction-using-machine-learning-4c5ff1e3fa16

https://www.geeksforgeeks.org/

https://www.javatpoint.com/