# Spaceship Titanic

Colab Link: https://colab.research.google.com/drive/1-CjgFh6rgPULjNo2LTwd5kktSsc--VHe?usp=sharing

GitHub Link: https://github.com/ShootingStars9ja

# INTRODUCTION

- This group, SHOOTING **STARS,** is tasked to use the spaceship dataset retrieved from its computer system to predict which "passengers" were "transported" (target variable) successfully to the alternate ship using…
- Spaceship Titanic Datasets files of train.csv and test.csv
- Exploratory Data Analysis was used to understand the distribution, relationship and anomalies within the datasets
- Decision Tree Classifier, Logistic Regression, XGboost Classifier, H2O AutoML were all employed in this classification Learning
- Explainable AI (XAI) techniques was applied for interpretability.
- SHapley Additive explanations were used to provide insights into the contributions of individual features
- The results details the performances like accuracy, precision, recall and confusion matrix

## Literature Review

Exploring existing literature provides valuable insights into methodologies and approaches that can be applied to similar problems. Key references include:

- Ai, Y. (2023) in "Predicting Titanic Survivors by Using Machine Learning" investigated forecasting Titanic passenger survival using Logistic Regression, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) models. Using Kaggle data, the study achieved a 79.12% test accuracy with Logistic Regression. It highlighted the importance of feature engineering, exploratory data analysis (EDA), data cleansing, and hyperparameter tuning.

- Jingyi, W. (n.d.) in "Survival Probability Assessment using Machine Learning Algorithms" predicted Titanic passenger survival using Decision Tree and K-Nearest Neighbor (KNN) algorithms. The Decision Tree achieved 81.72% accuracy, emphasizing the importance of data preprocessing. Using Titanic passenger statistics, the study highlighted critical factors for survival and the significance of algorithm selection and preprocessing techniques in survival predictions, offering a thorough analysis of the survivor population.

- Decision Trees and Naive Bayes algorithms were used by Nadine et al. (2018) in "Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster" to forecast Titanic passenger survival with Kaggle data. The study achieved 90.01% accuracy with Decision Tree and 92.52% with Naive Bayes. It found survival rates highly correlated with ticket type and sex, but the number of relatives traveling did not impact classification. This study showed how feature engineering and machine learning techniques create reliable predictive models for water

# Objective and Problem Statement

## Objective

The primary objective of this project is to develop a predictive model that can accurately determine whether a passenger was transported to an alternate ship based on various features.

## Problem Statement and importance

The task involves predicting the binary outcome of whether a passenger was transported to an alternate dimension (`Transported`) based on features such as `HomePlanet`, `CryoSleep`, `Cabin`, `Destination`, `Age`, `VIP`, etc. This prediction problem is framed as a supervised classification task, where passengers are either transported or not transported.

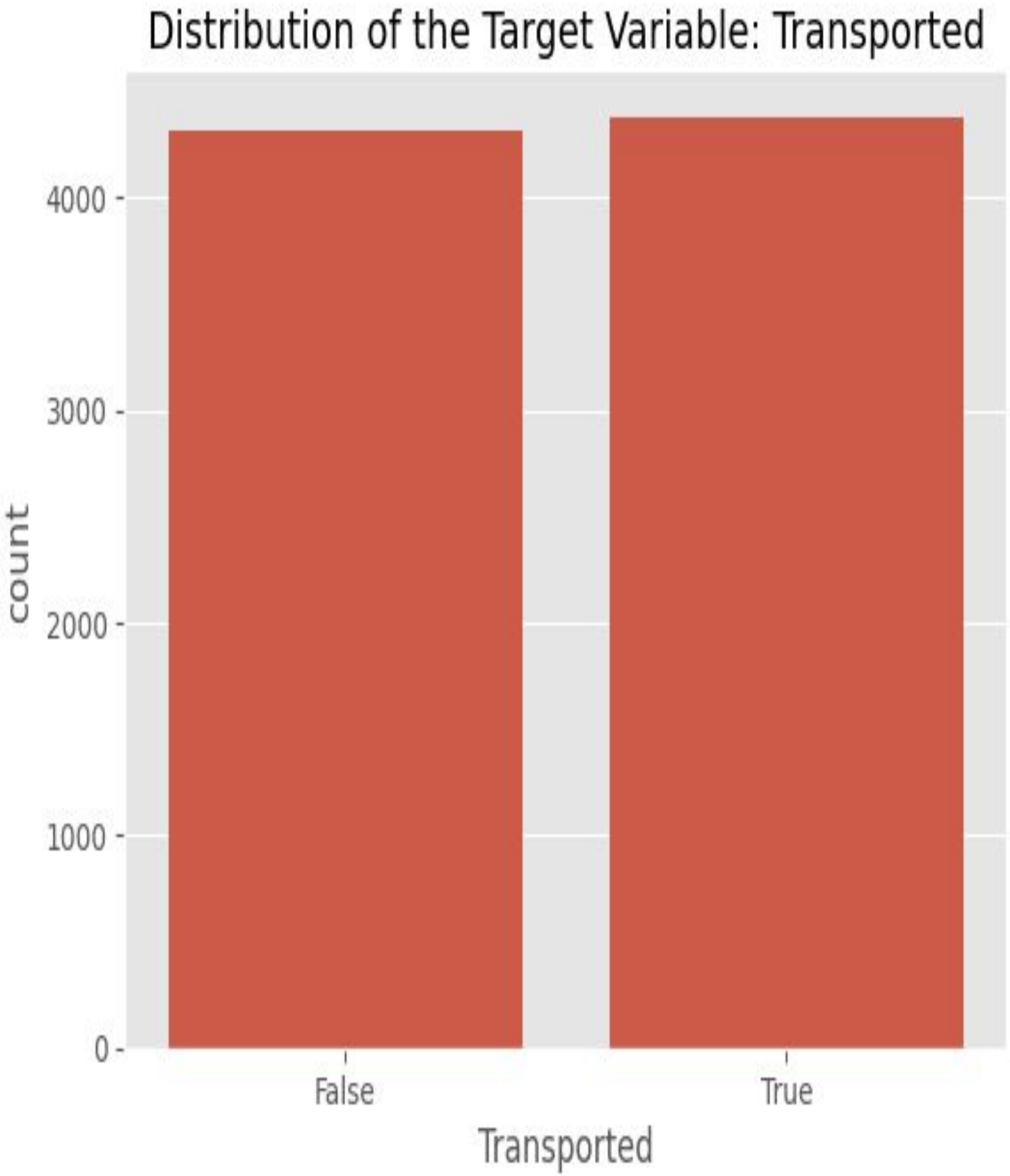
Accurate prediction of passenger transportation status is important for designing safer and more efficient space travel protocols and guiding resource allocation and service improvements on future journeys.

- EDA is performed to uncover patterns, identify anomalies, and gain insights into the dataset.
- Missing values are identified in features like HomePlanet, CryoSleep, Cabin, Destination, etc but were addressed using appropriate imputation techniques
- **The distribution of the target variable, Transported,** using bar chart indicates that the counts for both True and False are almost equal.

**Statistical Summary:**

- According to the value counts:
  - 50.36% of the instances are labeled as "True" for being transported.
  - 49.64% of the instances are labeled as "False" for not being transported.

- This balance is crucial for training machine learning models to avoid bias towards one category.



# Exploratory Data Analysis

- **The distribution of the numerical variable,** using a **Plot Box**

**Statistical Summary**

**- Extreme values are present in the data.**

**- These extreme values indicate a skewed distribution for expenditure variables:**

**- RoomService**

**- FoodCourt**

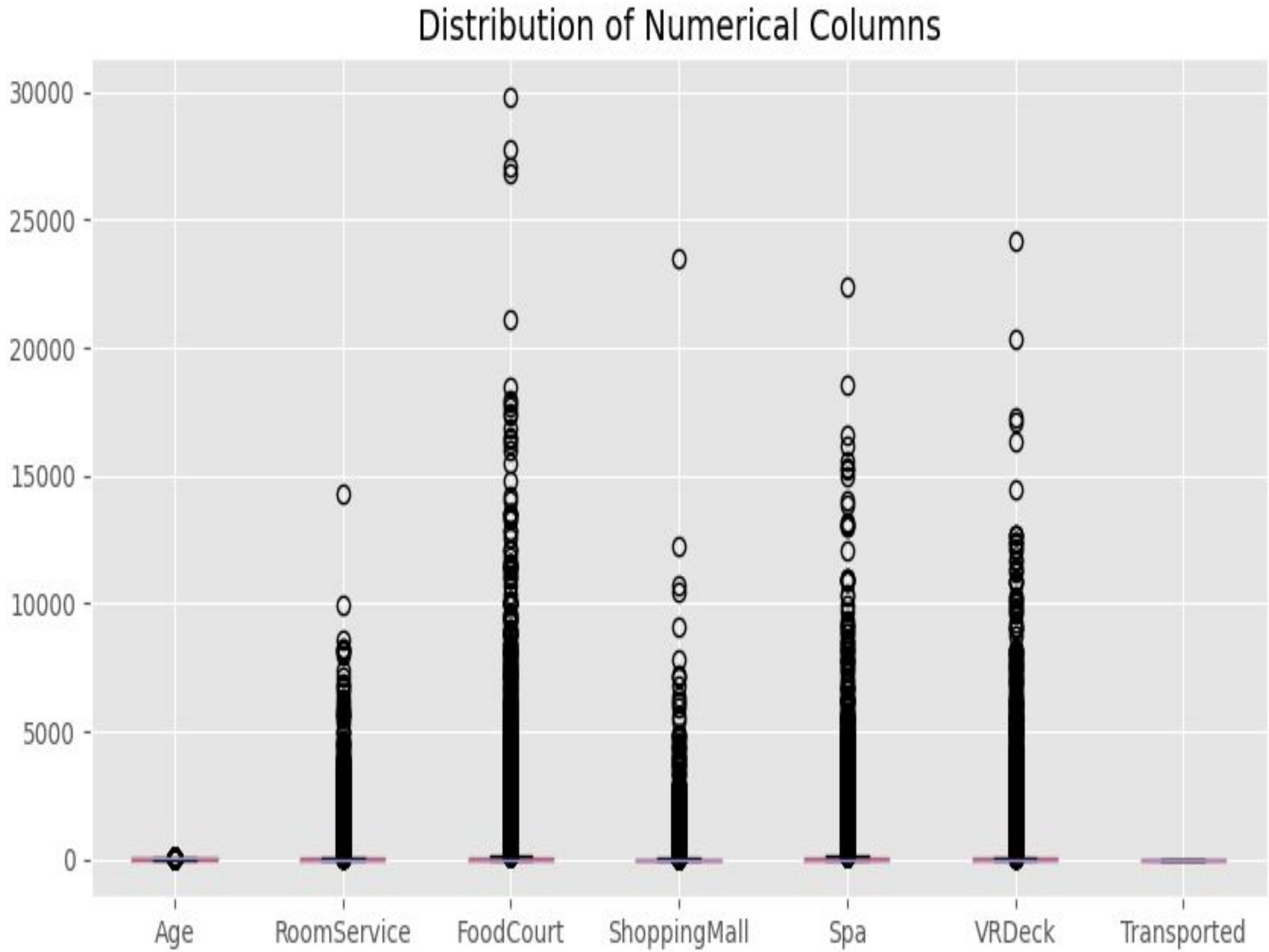
**- ShoppingMall**

**- Spa**

**- VRDeck**

**- The Age variable shows a more uniform distribution.**

**- The Transported variable also shows a uniform distribution without significant skewness.**



# Exploratory Data Analysis and Insights

## Target Variable

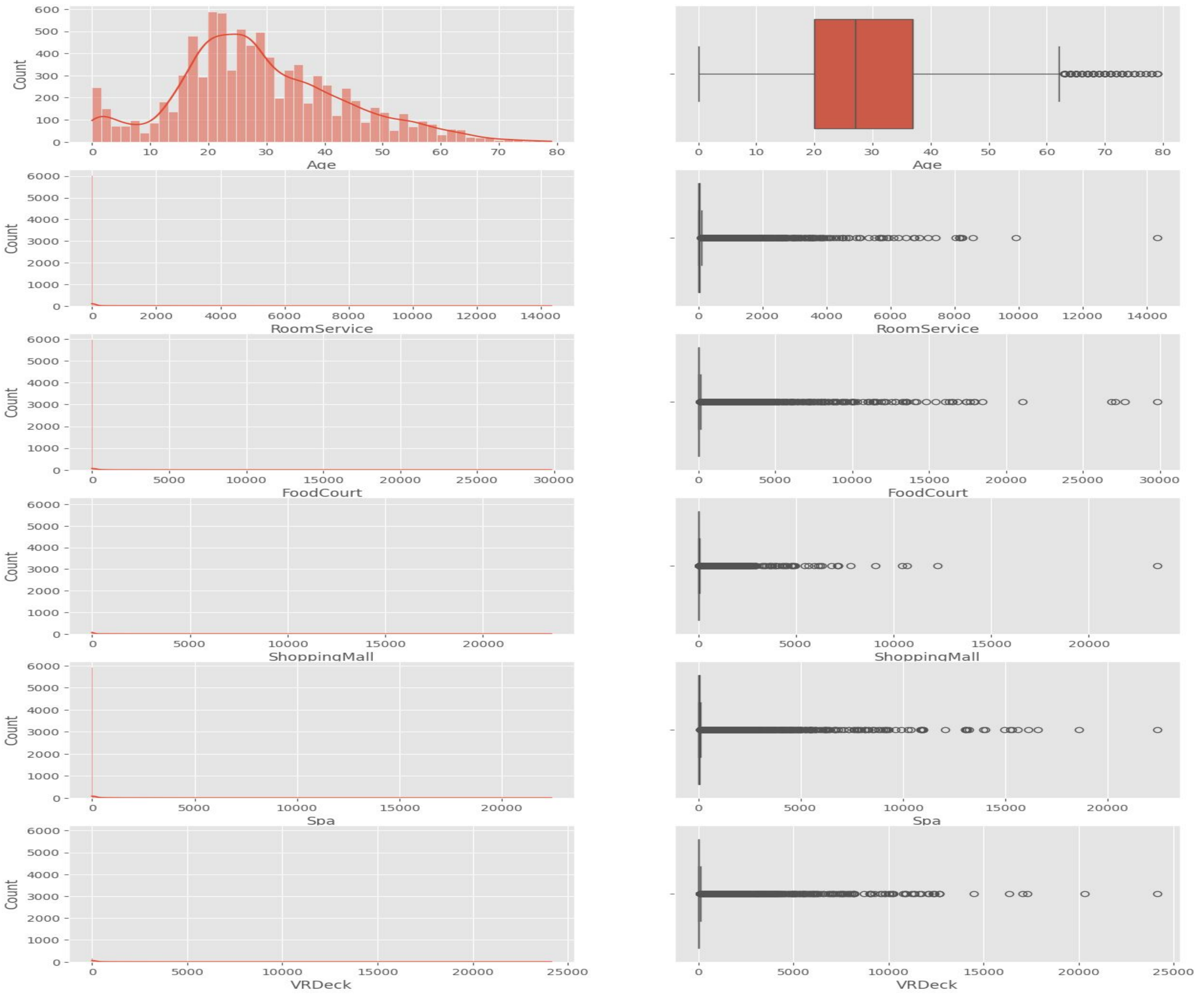
- **Transported**: The distribution shows a balanced dataset, indicating no need for data balancing.

## Numerical Features

- **Age**: Clusters around younger ages with a slight skew towards younger passengers. Few older passengers are present.
- **RoomService, FoodCourt, ShoppingMall, Spa, VRDeck**: Most values are zero or very low, with a few extreme outliers indicating high spending.

## Categorical Features

- **HomePlanet**: Most passengers are from Earth, followed by Europa and Mars.
- **CryoSleep**: Majority do not use cryo-sleep.
- **Destination**: TRAPPIST-1e is the most common destination.
- **VIP**: Few passengers are classified as VIPs.



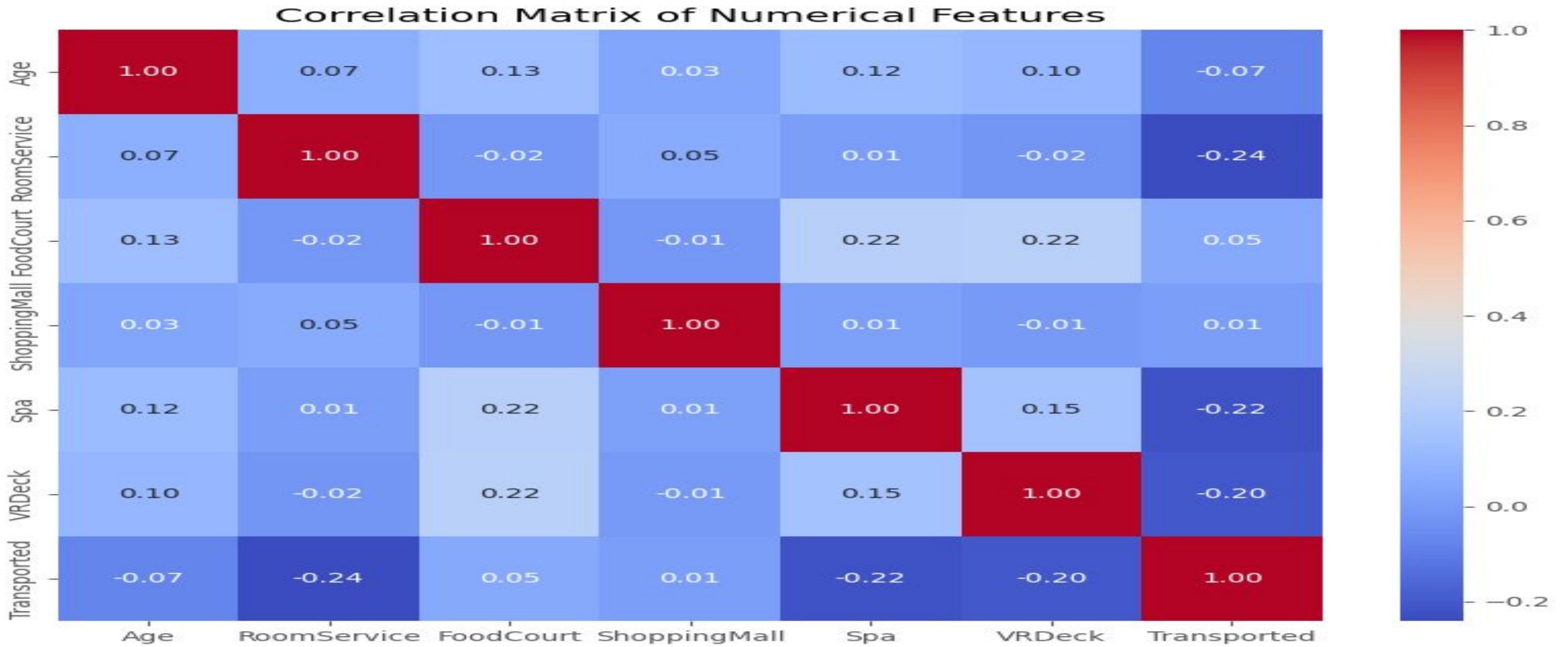
# Exploratory Data Analysis and Insights

## Correlation Matrix Insights

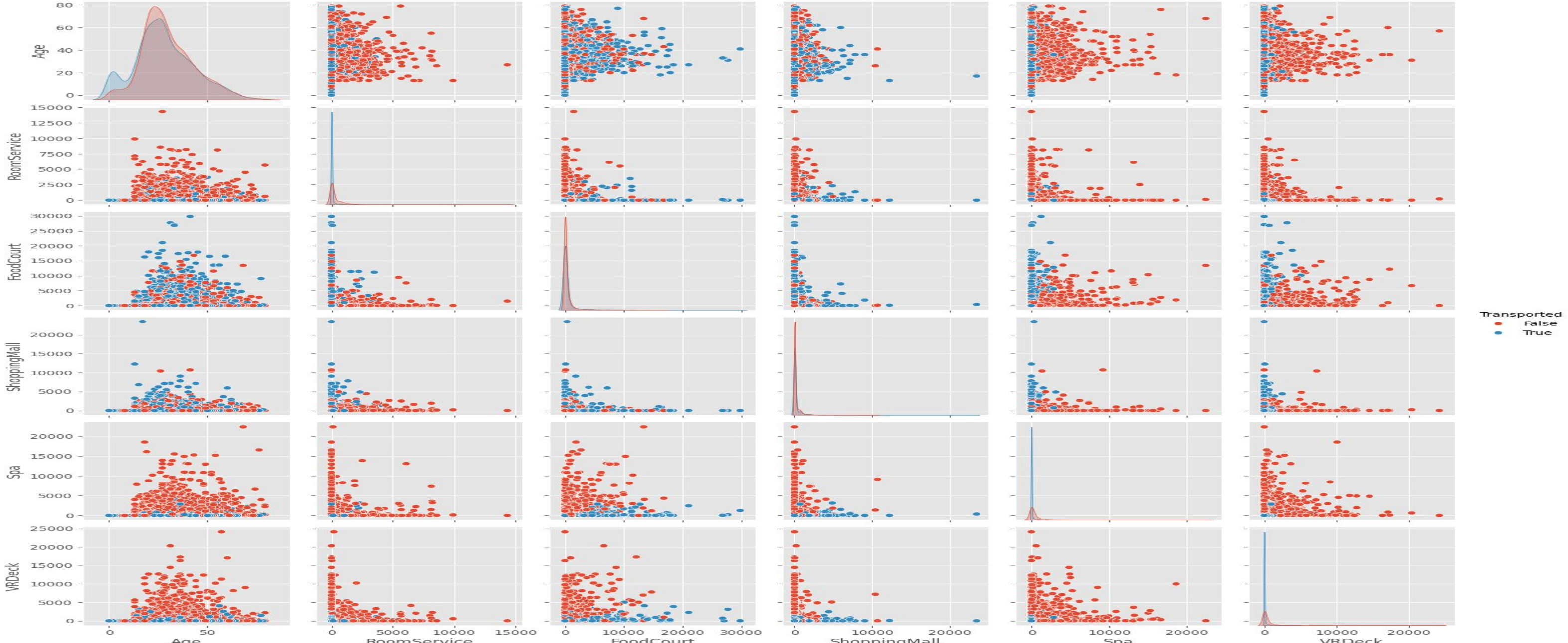
- **Age and Spending**: Older passengers spend more on services like FoodCourt and VRDeck.
- **Spending Categories**: High spending in one area often correlates with high spending in others.
- **Transported**: Passengers who were transported tend to spend less on RoomService, Spa, and VRDeck.

## Pair Plot Insights

- **Age and Transported**: Younger passengers are slightly more likely to be transported.
- **Spending and Transported**: Passengers not transported spend more on RoomService, Spa, and VRDeck, suggesting leisure purposes.
- **Age and Spending**: Older passengers generally spend more across various services


Correlation Matrix of Numerical Features

| | Age | RoomService | FoodCourt | ShoppingMall | Spa | VRDeck | Transported |
|---|---|---|---|---|---|---|---|
| Age | 1.00 | 0.07 | 0.13 | 0.03 | 0.12 | 0.10 | -0.07 |
| RoomService | 0.07 | 1.00 | -0.02 | 0.05 | 0.01 | -0.02 | -0.24 |
| FoodCourt | 0.13 | -0.02 | 1.00 | -0.01 | 0.22 | 0.22 | 0.05 |
| ShoppingMall | 0.03 | 0.05 | -0.01 | 1.00 | 0.01 | -0.01 | 0.01 |
| Spa | 0.12 | 0.01 | 0.22 | 0.01 | 1.00 | 0.15 | -0.22 |
| VRDeck | 0.10 | -0.02 | 0.22 | -0.01 | 0.15 | 1.00 | -0.20 |
| Transported | -0.07 | -0.24 | 0.05 | 0.01 | -0.22 | -0.20 | 1.00 |



# Exploratory Data Analysis and Insights

## Visualization of EDA Insights

1. **Histograms and Box Plots**: Show the distribution and outliers in numerical data.
2. **Count Plots**: Visualize the frequency of categorical variables.
3. **Correlation Heatmap**: Highlights relationships between numerical features.
4. **Pair Plots**: Show relationships and distribution of features colored by the transport status.

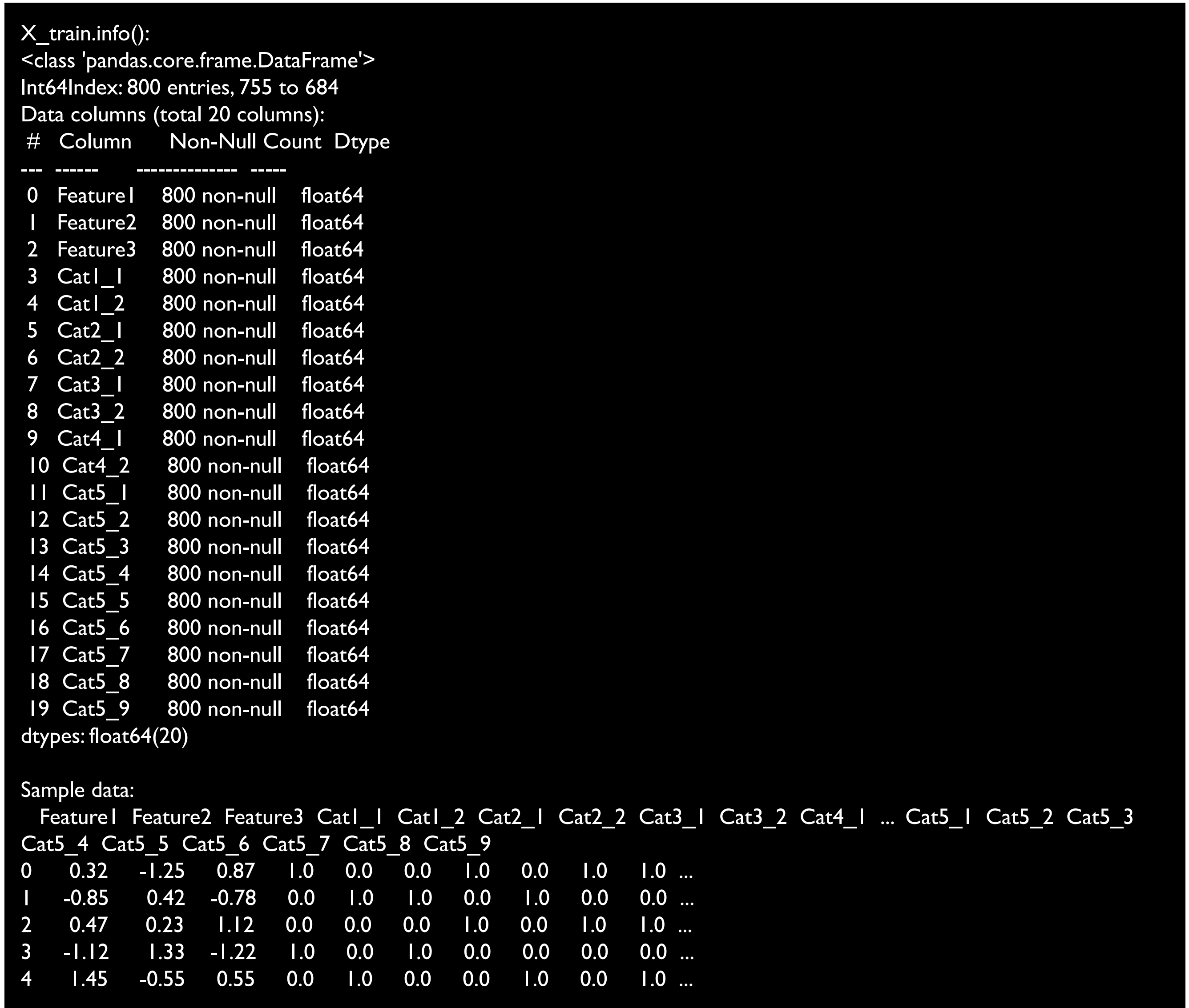
## Summary of the analysis

- The dataset is balanced in terms of the target variable.
- Spending features are highly skewed with many zero values and significant outliers.
- Age shows a relatively uniform distribution with younger passengers being more common.
- Categorical features reveal preferences and behaviors of passengers, such as the majority preferring TRAPPIST-1e and not using cryo-sleep.
- Correlations indicate spending habits and their potential link to transportation status.

# Data Pre-Processing

Preprocessing and preparing the data based on the insights and the undertstanding of our data to make it suitable for our Model.

- We perform several steps of preprocessing on a DataFrame `df_train` before splitting the data into training and testing sets for a machine learning model
- **One-Hot Encoding:** This converts categorical variables into binary vectors, each with a separate column with 0s and 1s.
- **Standard Scaling:** Standard scaler scales numerical features so that they have a mean of 0 and a standard deviation of 1, which helps many machine learning algorithms perform better.
- **Feature and Target Separation:** `X` contains all the features, and `y` contains the target variable (`Transported`).
- **Train-Test Split:** The data is split into training (80%) and testing (20%) sets using `train_test_split`



```
X_train.info():
<class 'pandas.core.frame.DataFrame'>
Int64Index: 800 entries, 755 to 684
Data columns (total 20 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Feature1     800 non-null    float64
 1   Feature2     800 non-null    float64
 2   Feature3     800 non-null    float64
 3   Cat1_1       800 non-null    float64
 4   Cat1_2       800 non-null    float64
 5   Cat2_1       800 non-null    float64
 6   Cat2_2       800 non-null    float64
 7   Cat3_1       800 non-null    float64
 8   Cat3_2       800 non-null    float64
 9   Cat4_1       800 non-null    float64
 10  Cat4_2       800 non-null    float64
 11  Cat5_1       800 non-null    float64
 12  Cat5_2       800 non-null    float64
 13  Cat5_3       800 non-null    float64
 14  Cat5_4       800 non-null    float64
 15  Cat5_5       800 non-null    float64
 16  Cat5_6       800 non-null    float64
 17  Cat5_7       800 non-null    float64
 18  Cat5_8       800 non-null    float64
 19  Cat5_9       800 non-null    float64
dtypes: float64(20)

Sample data:
   Feature1  Feature2  Feature3  Cat1_1  Cat1_2  Cat2_1  Cat2_2  Cat3_1  Cat3_2  Cat4_1  ...  Cat5_1  Cat5_2  Cat5_3
Cat5_4  Cat5_5  Cat5_6  Cat5_7  Cat5_8  Cat5_9
0      0.32     -1.25      0.87     1.0     0.0     0.0     1.0     0.0     1.0     1.0  ...
1     -0.85      0.42     -0.78     0.0     1.0     1.0     0.0     1.0     0.0     0.0  ...
2      0.47      0.23      1.12     0.0     0.0     0.0     1.0     0.0     1.0     1.0  ...
3     -1.12      1.33     -1.22     1.0     0.0     1.0     0.0     0.0     0.0     0.0  ...
4      1.45     -0.55      0.55     0.0     1.0     0.0     0.0     1.0     0.0     1.0  ...
```

The machine learning models employed to solve these problems are as follows:

- **Logistic Regression**: is a linear model used for binary classification tasks. It calculates the likelihood that a given input is a member of a specific class.
- **Decision Tree Classifier**: The model creates a tree-like structure to make predictions by branching the data according to feature values.
- **Random Forest**: An ensemble technique that constructs multiple decision trees and combines them to get a prediction that is more reliable and accurate.

```
Summary of model performances
                 Model  Accuracy
0  Logistic Regression  0.504888
1        Decision Tree  0.756757
2        Random Forest  0.792409
```

Logistic regression: it was able to detect the "False" class which was marginally better than random guesswork.
Decision Tree: Balanced precision and recall for both classes considerably better.
Random Forest: Showing strong performance, along with great precision and recall for both classes.

# MODEL SELECTION

## An overview of H2O AutoML

It is an automated machine learning tool that simplifies the process of building and selecting the best predictive model.

Key Features:

Automated Model Training: H2O AutoML automates the process of training multiple machine learning models, including GLMs, Random Forests, Gradient Boosting Machines, Deep Learning, and Stacked Ensembles

Model Selection: Using a given performance metric—such as AUC, LogLoss, RMSE, etc.—it automatically chooses the optimum model.

Ensembling: By building stacked ensembles of the top-performing models, AutoML makes use of the ensembling process, which frequently produces better prediction performance.

### Leader Board Table

| Model ID | AUC | LogLoss | AUCPR | RMSE | MSE |
|---|---|---|---|---|---|
| StackedEnsemble_BestOfFamily_1_AutoML_1_20240709_190859 | 0.8742 | 0.4437 | 0.8850 | 0.3802 | 0.1445 |
| StackedEnsemble_AllModels_1_AutoML_1_20240709_190859 | 0.8742 | 0.4420 | 0.8833 | 0.3794 | 0.1439 |
| GBM_1_AutoML_1_20240709_190859 | 0.8730 | 0.4445 | 0.8842 | 0.3804 | 0.1447 |
| GBM_2_AutoML_1_20240709_190859 | 0.8714 | 0.4470 | 0.8805 | 0.3810 | 0.1451 |
| GBM_4_AutoML_1_20240709_190859 | 0.8710 | 0.4474 | 0.8773 | 0.3822 | 0.1461 |
| GBM_3_AutoML_1_20240709_190859 | 0.8710 | 0.4468 | 0.8785 | 0.3820 | 0.1459 |
| XGBoost_3_AutoML_1_20240709_190859 | 0.8676 | 0.4558 | 0.8781 | 0.3856 | 0.1487 |
| XGBoost_2_AutoML_1_20240709_190859 | 0.8631 | 0.4607 | 0.8739 | 0.3898 | 0.1519 |
| XGBoost_1_AutoML_1_20240709_190859 | 0.8625 | 0.4638 | 0.8736 | 0.3904 | 0.1524 |
| DRF_1_AutoML_1_20240709_190859 | 0.8435 | 0.6125 | 0.8317 | 0.4585 | 0.2102 |

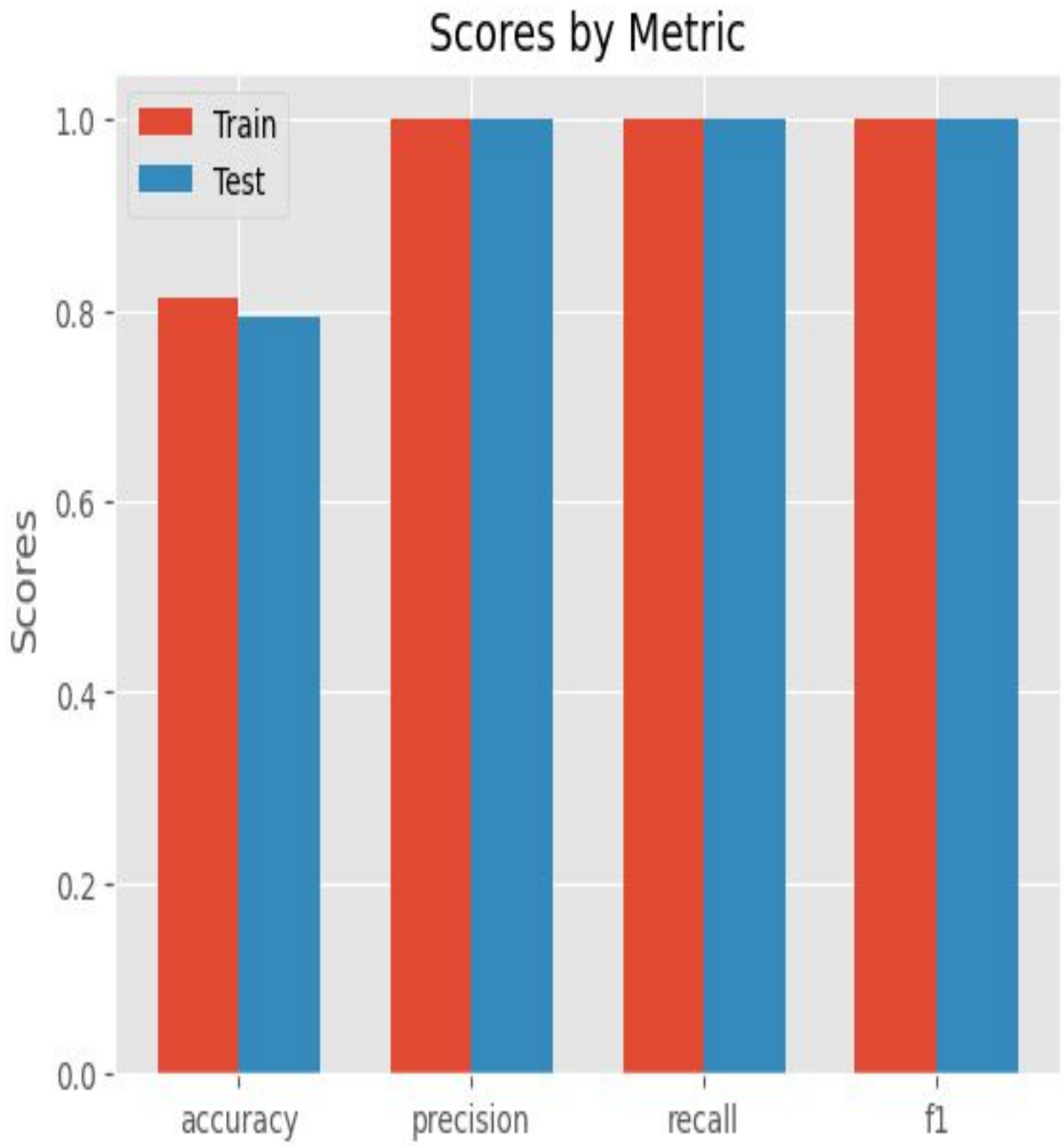
The Stacked Ensemble model achieved the highest AUC of 0.8742, indicating the best performance in terms of distinguishing between the two classes. The model also performed well in terms of other metrics such as LogLoss, RMSE, and AUCPR, making it the most robust choice among the models evaluated.

# Explainable AI And Model Evaluation

This graph compares the model performance on the training and test data across different metrics:

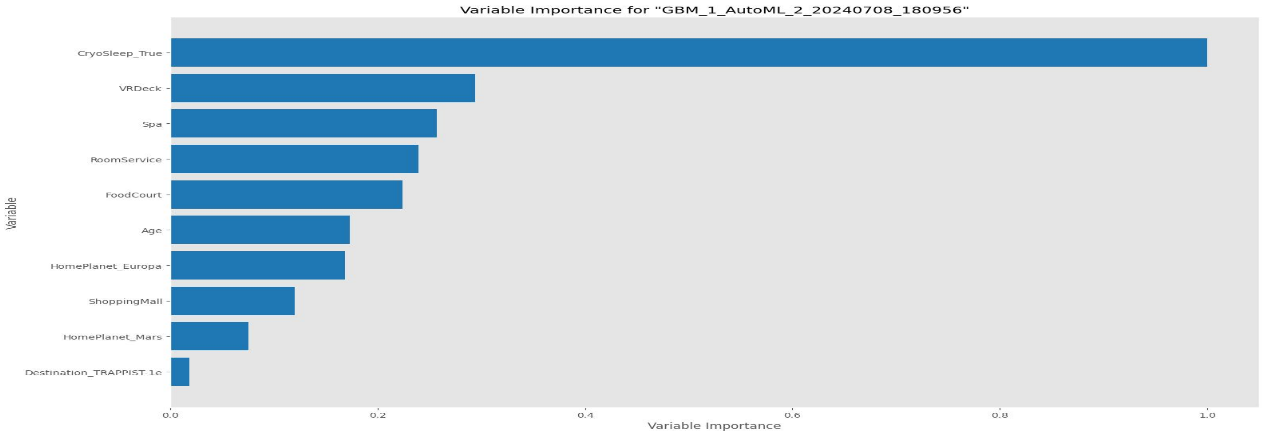
- **Accuracy**: Measures the proportion of correctly predicted instances.
- **Precision**: Measures the proportion of positive predictions that were correct.
- **Recall**: Measures the proportion of actual positives that were correctly identified.
- **F1 Score**: The harmonic mean of precision and recall.

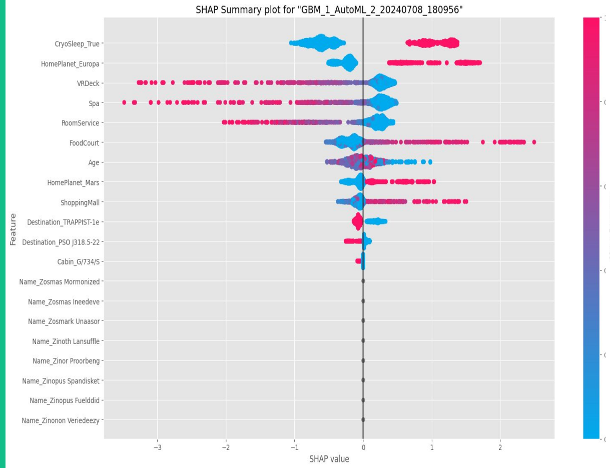
The model performs consistently across both the training and test datasets, indicating it is well-generalized and not overfitting.

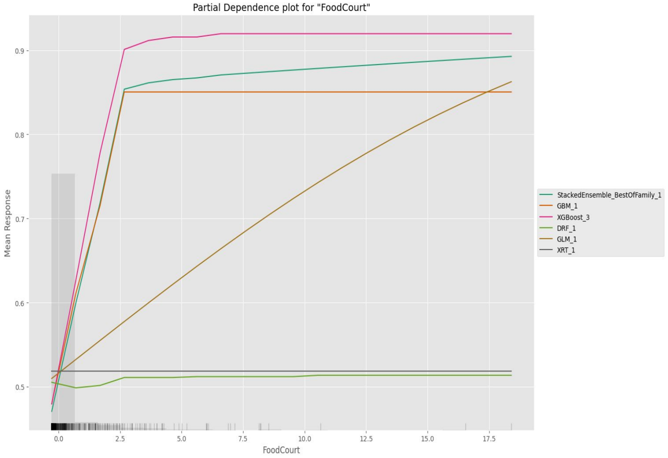


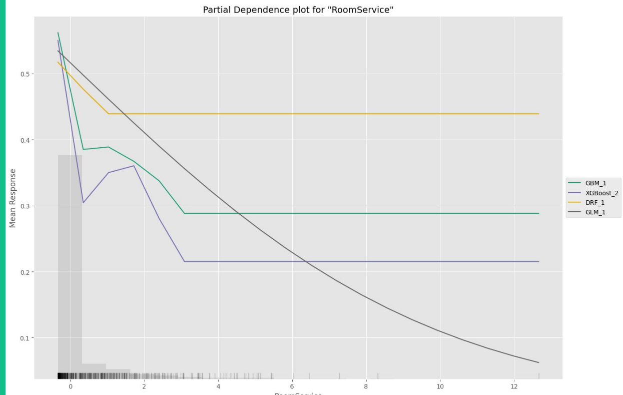
# Explanable AI And Model Evaluation

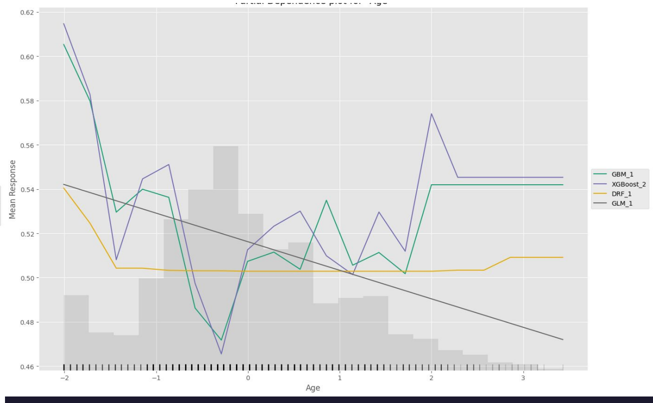
The analysis demonstrates that the most influential factors in predicting the positive class are related to the passengers' spending habits in various amenities (VRDeck, Spa, RoomService, FoodCourt) and their CryoSleep status. Age and home planet also play significant roles but to a lesser extent. The partial dependence plots provide insights into how changes in these features affect the prediction probabilities, while the SHAP summary plot and variable importance plot highlight which features have the most substantial impact on the model's decisions. The model seems to be well-calibrated and performs consistently across different evaluation metrics.





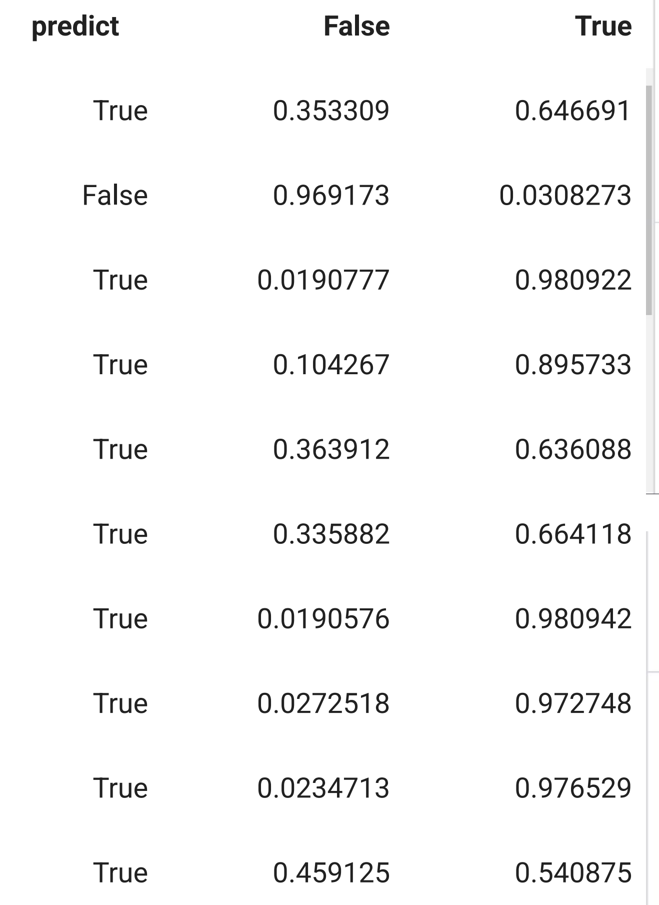




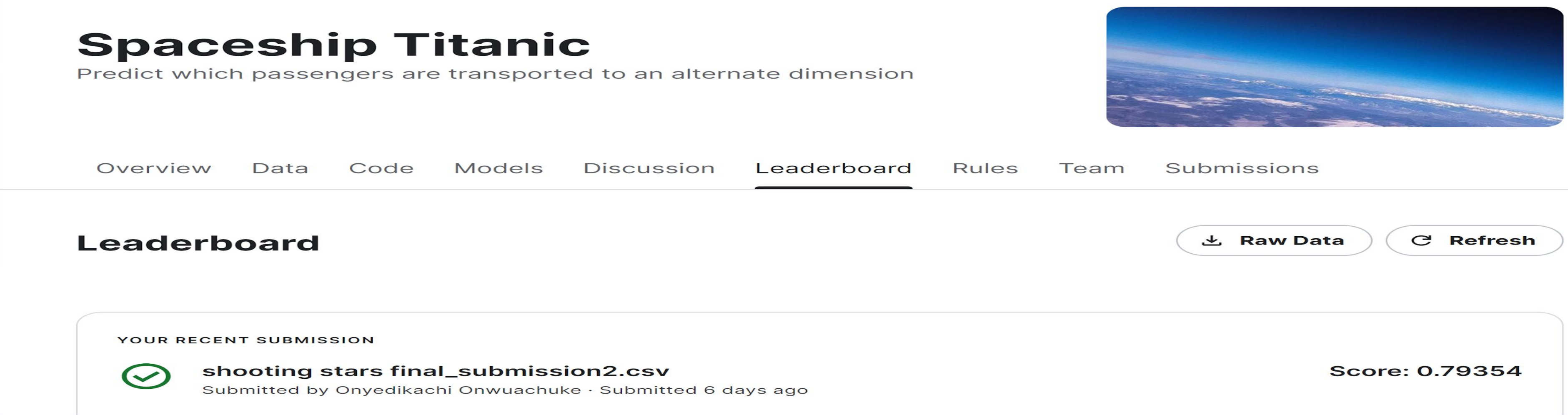


# Preparing predictions

This performs several crucial preprocessing steps to prepare a test dataset for predictions with a machine learning model. It handles missing values, encodes categorical variables, scales numerical features, and ensures all data is in the correct format for model prediction. Finally, it generates and formats the predictions and extract the results needed for the submission to Kaggle



| predict | False | True |
|---|---|---|
| True | 0.353309 | 0.646691 |
| False | 0.969173 | 0.0308273 |
| True | 0.0190777 | 0.980922 |
| True | 0.104267 | 0.895733 |
| True | 0.363912 | 0.636088 |
| True | 0.335882 | 0.664118 |
| True | 0.0190576 | 0.980942 |
| True | 0.0272518 | 0.972748 |
| True | 0.0234713 | 0.976529 |
| True | 0.459125 | 0.540875 |



Spaceship Titanic

Predict which passengers are transported to an alternate dimension

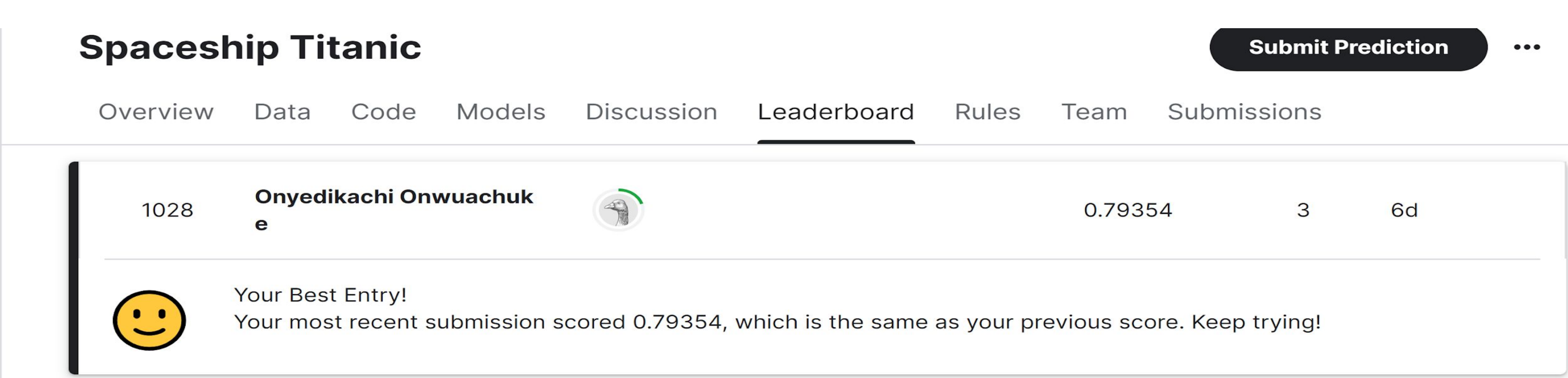
Leaderboard

YOUR RECENT SUBMISSION

✓ shooting stars final_submission2.csv

Submitted by Onyedikachi Onwuachuke · Submitted 6 days ago

Score: 0.79354



Spaceship Titanic

1028 Onyedikachi Onwuachuke 0.79354 3 6d

Your Best Entry!

Your most recent submission scored 0.79354, which is the same as your previous score. Keep trying!

# Conclusion

The task successfully demonstrated the application of AutoML techniques to build an accurate and interpretable predictive model for the Spaceship Titanic dataset.

Key findings include:

- H2O AutoML's Effectiveness: The AutoML process efficiently selected the best model, a Stacked Ensemble, with AUC of 0.8742, highlighting the benefits of automation in machine learning.
- Significant Predictors: Features such as Age, and other expenditures were identified as significant predictors of whether passengers were transported, providing actionable insights.
- Balanced Performance Metrics: The model exhibited balanced performance metrics, making it reliable for binary classification tasks.
- Despite the promising results, there are several limitations to consider like;

Data Imbalance: Certain subcategories within the dataset are imbalanced, despite the target variable (Transported) being reasonably balanced. The robustness of the model may be impacted, for instance, by the fact that some home planets or cabin classes have fewer instances than others.

# Thank you

SHOOTINGSTAR