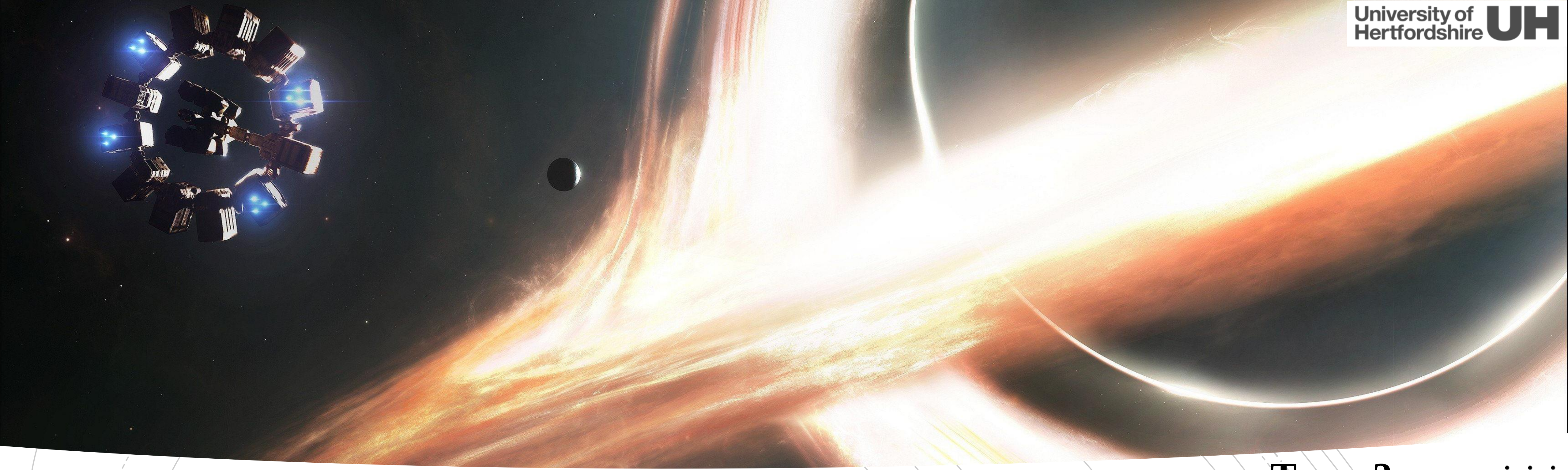


University of Hertfordshire UH

KAGGLE CHALLENGE

Team 2

Bineeth Mathew (22004878)

Anns Tomy (22033815)

Dhanya Davis (22014216)

Meenakshi Rajesh (22012440)

Gobu Chettiakulam Babu (22019388)

Aaron Modiyil Joseph (22018497)

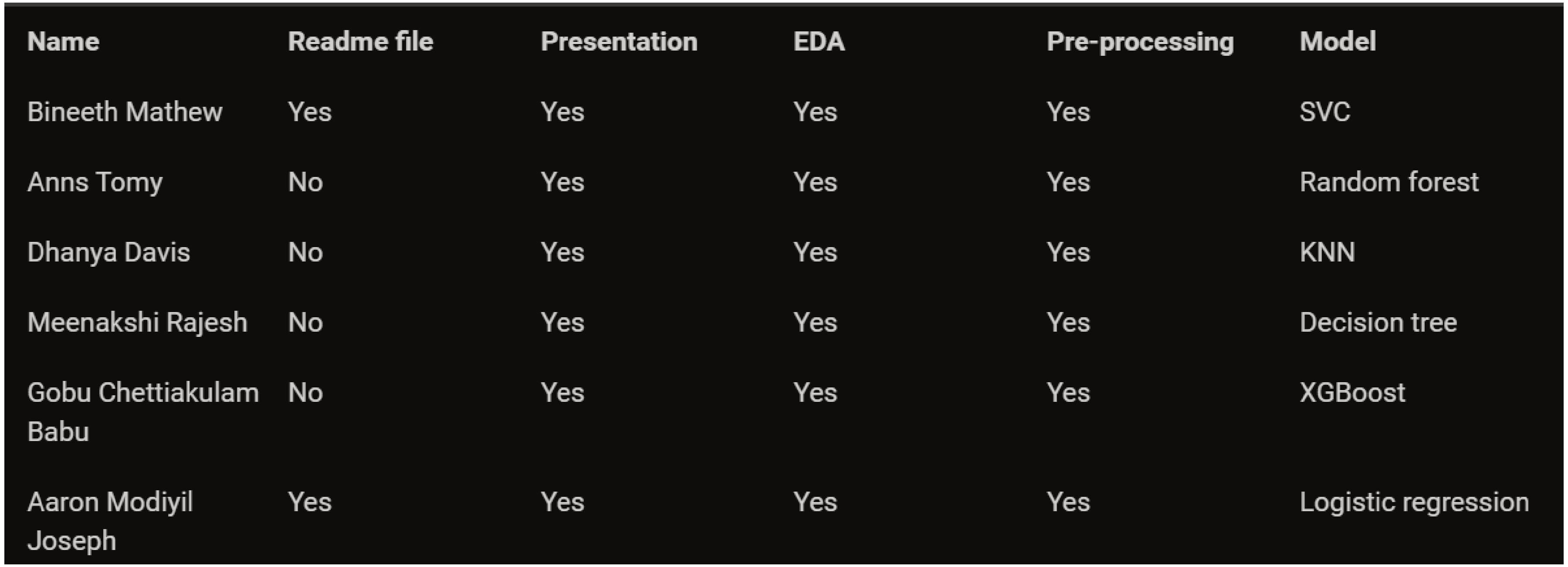
TEAM 2
KAGGLE GROUP PROJECT

GitHub link Colab notebook link

INTRODUCTION

- *The problem*: the Spaceship titanic Kaggle challenge tasks us with predicting which passengers from the interstellar ship were transported to an alternate dimension after its collision with a spacetime anomaly, using data recovered from damaged computer system.
- *Individual Model Building and Training*: some of the initial coding was collaborative. Then each of the six team members took responsibility for building and training a specific model of their choice. This approach allowed us to explore different algorithms and techniques for the same problem and later compare their performance.
- *Team Collaboration*: Team meetings and discussions, in-person and online, were frequently organized for delegating work as collaboration was key. Coding was carried out on a Colab notebook shared by all members. A Github organization, and a repository was created for the same.

WORK DELEGATION OVERVIEW



Name	Readme file	Presentation	EDA	Pre-processing	Model
Bineeth Mathew	Yes	Yes	Yes	Yes	SVC
Anns Tomy	No	Yes	Yes	Yes	Random forest
Dhanya Davis	No	Yes	Yes	Yes	KNN
Meenakshi Rajesh	No	Yes	Yes	Yes	Decision tree
Gobu Chettiakulam Babu	No	Yes	Yes	Yes	XGBoost
Aaron Modiyil Joseph	Yes	Yes	Yes	Yes	Logistic regression

- Four in-person group meetings in one of the LRC group study rooms and two online zoom meetings were held for discussion
- A WhatsApp group chat with all the team members was created on June 6th where discussions were actively taking place.
- In addition to the shared Colab notebook, the PowerPoint file for the presentation was also shared and worked on collaboratively

DATASET OVERVIEW

- Data was gathered from personal records recovered from the damaged computer system of the Spaceship Titanic, an interstellar passenger liner.
- Provided two set of files train.csv and test.csv.
- train.csv contains personal records for about two-thirds (~8700) of the passengers.
- test.csv contains personal records for the remaining one-third (~4300) of the passengers.
- There are 2 nominal features, 5 categorical features and 6 numerical features and 1 Boolean target variable.
- The dataset is sourced from Kaggle.

TEAM 2
KAGGLE GROUP PROJECT

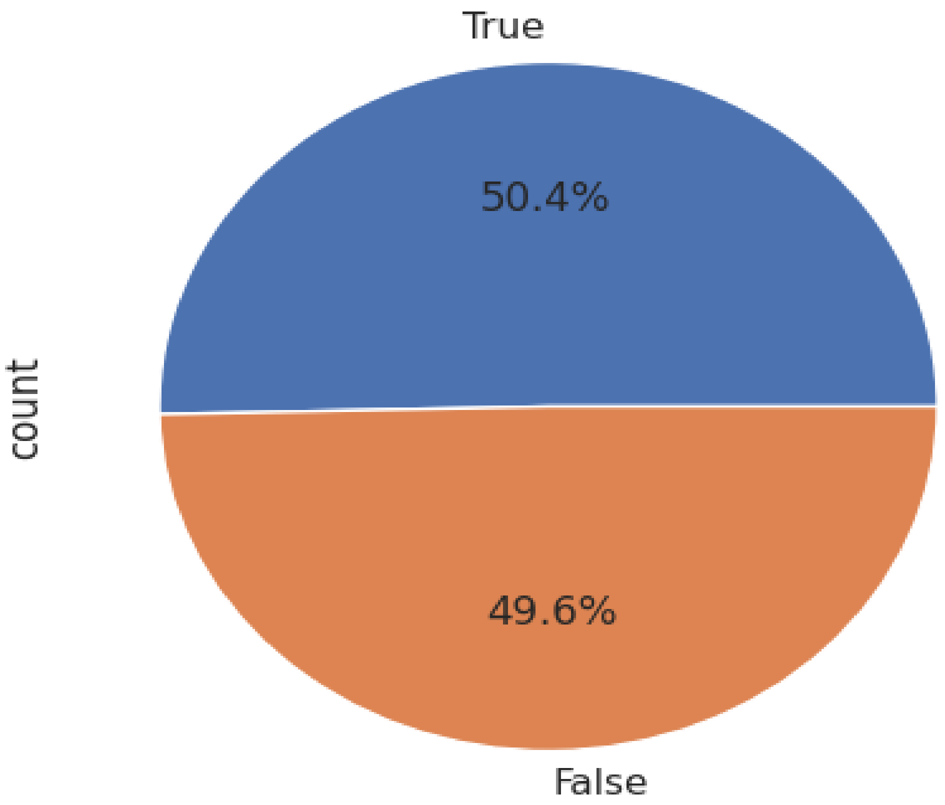
University of Hertfordshire UH

DOMAIN KNOWLEDGE

Column	Description
PassengerId	A unique identifier for each passenger, formatted as gggg_pp where gggg is the group number and pp is their number within the group.
HomePlanet	The planet the passenger departed from, typically their planet of permanent residence.
CryoSleep	Indicates whether the passenger elected to be in suspended animation during the voyage.
Cabin	The cabin number where the passenger is staying, formatted as deck/num/side.
Destination	The planet the passenger will disembark to.
Age	The age of the passenger.
VIP	Indicates whether the passenger has paid for 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.

EXPLORATORY DATA ANALYSIS

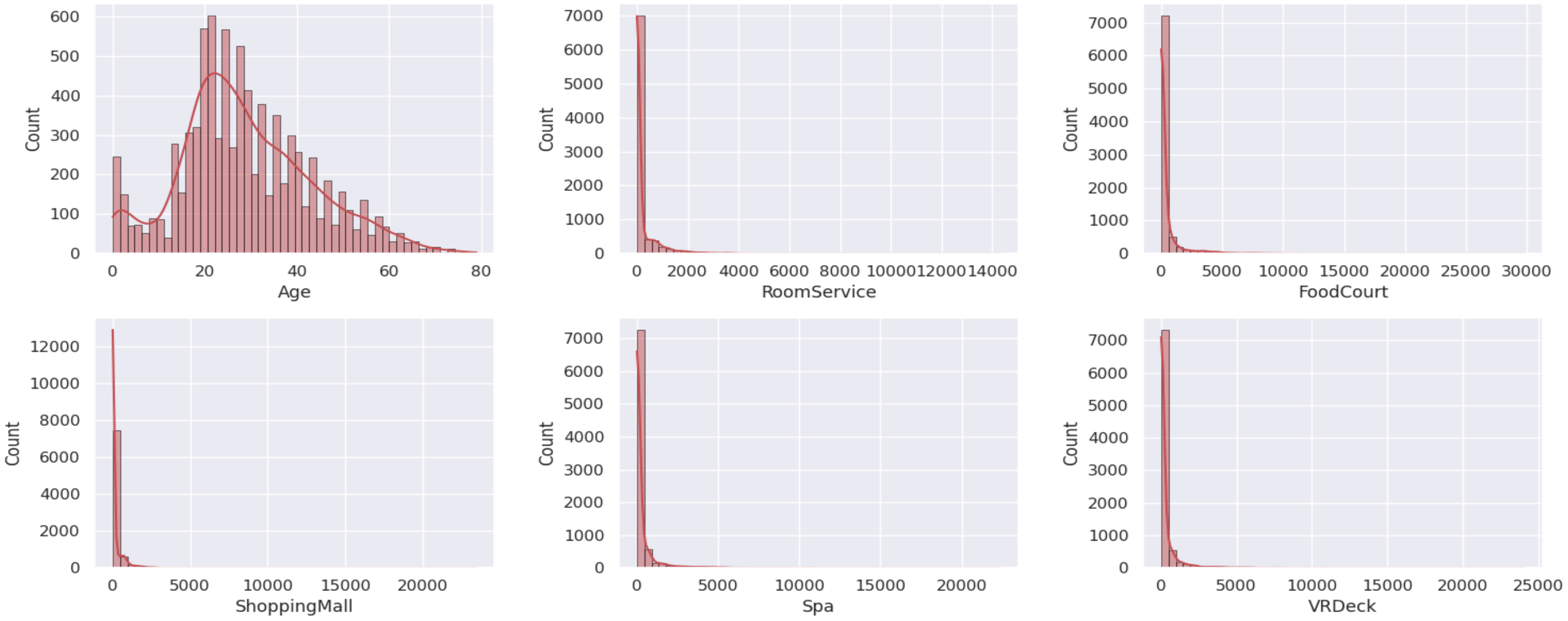
- df.info and df.describe reveal that the dataset has null values and that the data types are not optimal.
- The distribution of the target variable, "Transported," is nearly balanced. This benefits training machine learning models from a balanced distribution, since it eliminates bias towards a certain class.



pie chart - Transported Feature

EXPLORATORY DATA ANALYSIS

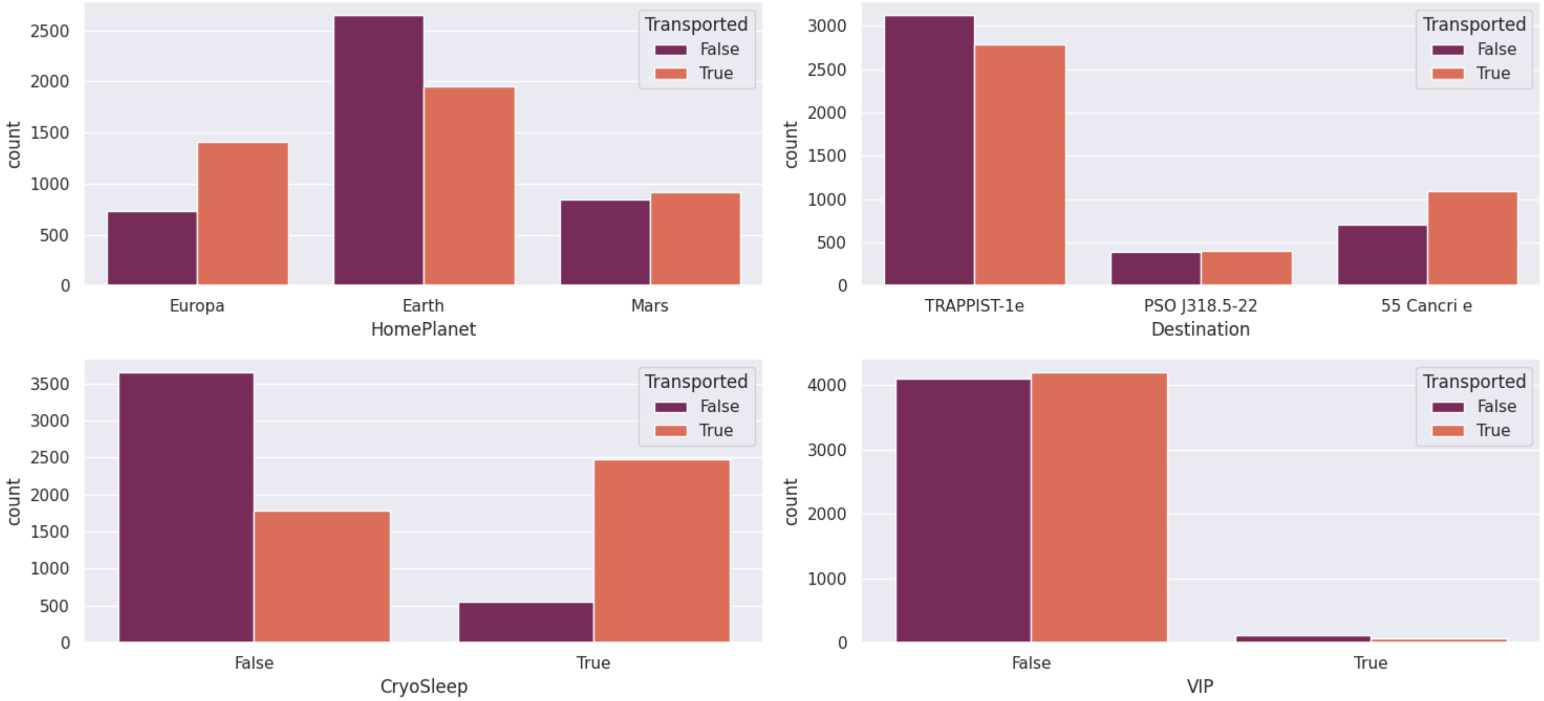
- The “Age" feature nearly resembles a normal distribution, but all other numerical features show skewness. The box plots and df.describe corroborate this observation.



Distribution of each numerical feature

EXPLORATORY DATA ANALYSIS

- The "VIP" feature shows nearly equal distribution of each cases in which people were transported and cases in which they weren't.



Bar plot of each categorical feature

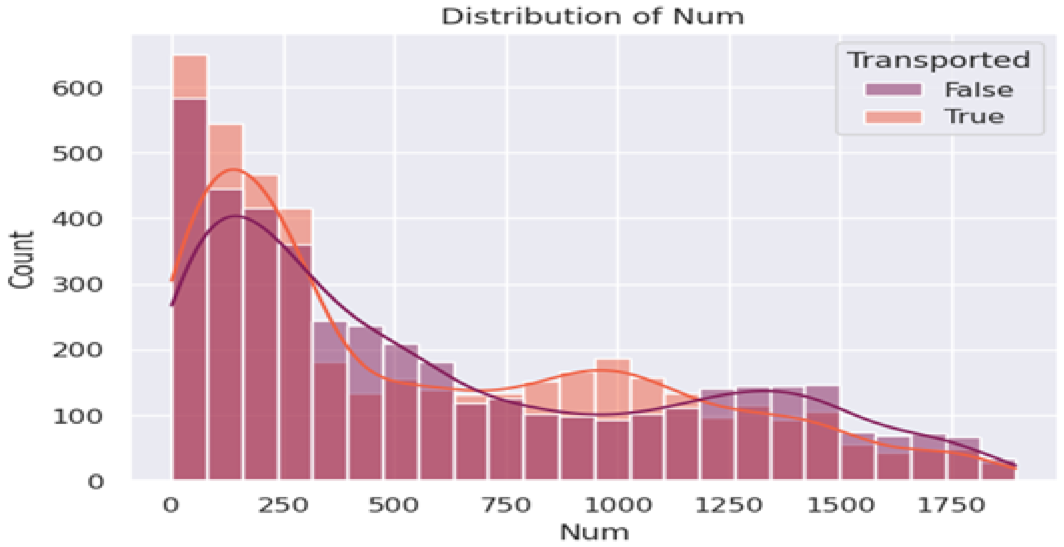
- This evenly distributed data indicates that the "VIP" feature has little effect on the target variable. We dropped the "VIP" feature from additional examination as a result.

TEAM 2
KAGGLE GROUP PROJECT

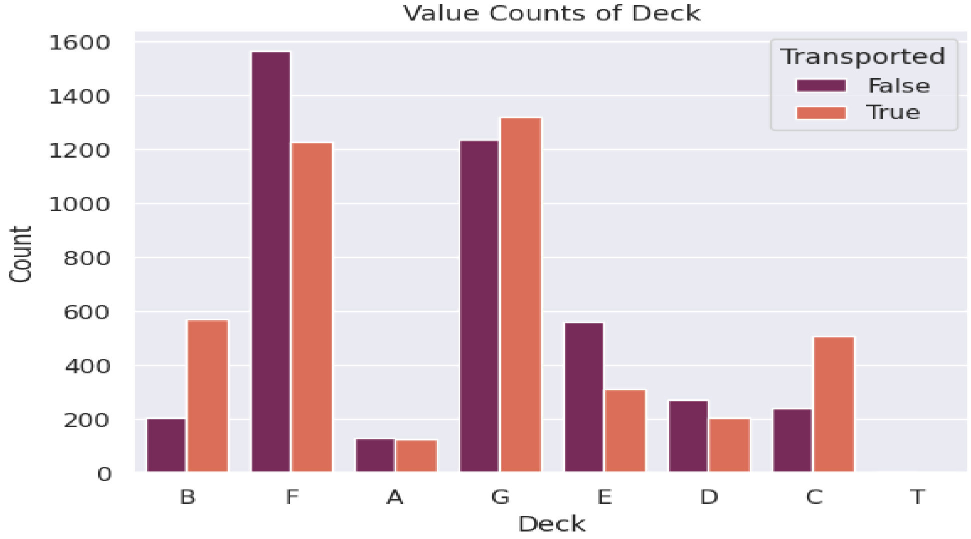
University of Hertfordshire UH

EXPLORATORY DATA ANALYSIS

- The "Num" feature's distribution plot shows that it has a substantial impact on the target variable.
- The "Deck" feature's value counts plot shows how the target variable is influenced differently by the many categories under this feature.



Bar Plot of Num Distribution



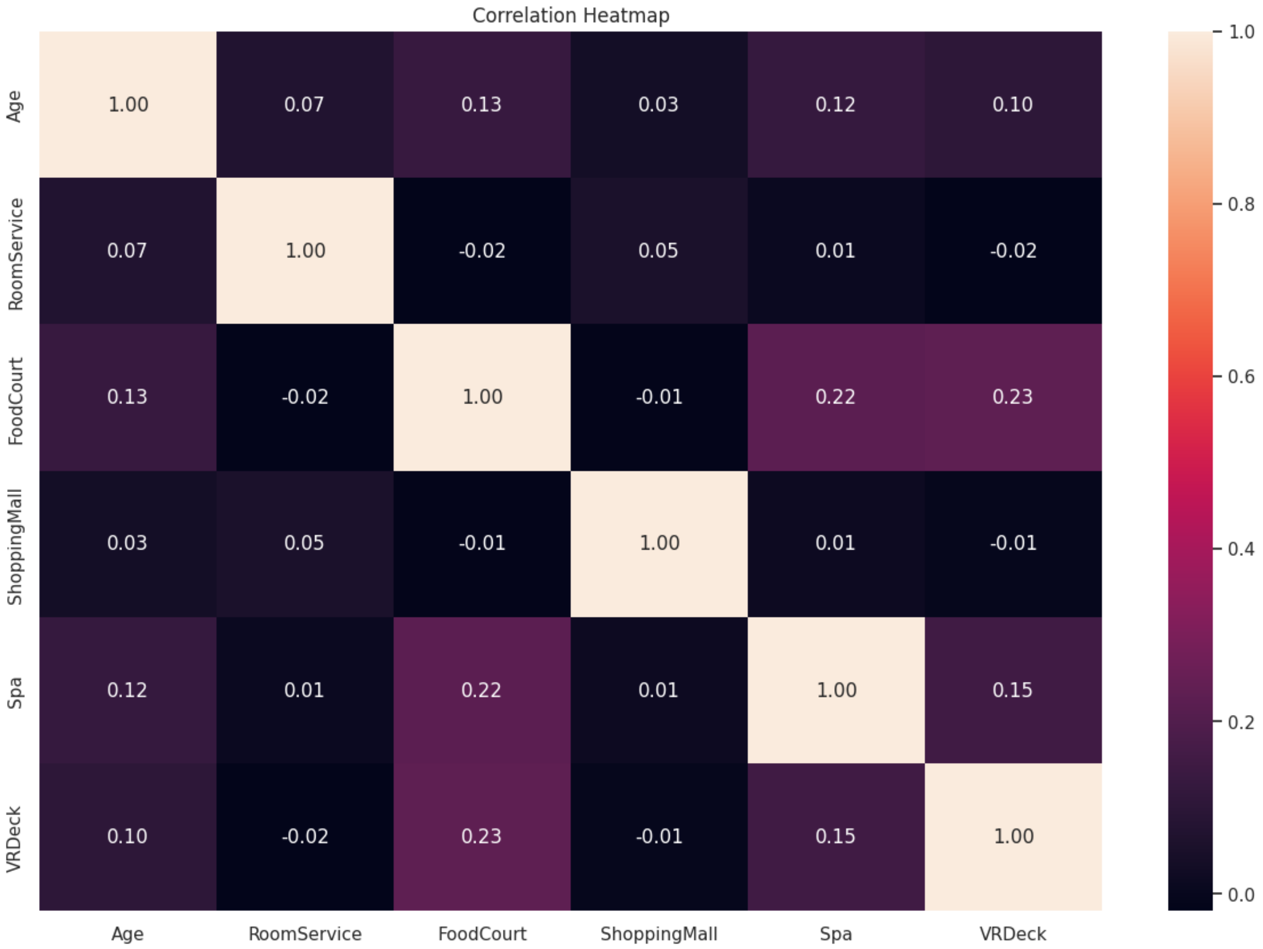
Bar Plot of Value counts of Deck feature

TEAM 2
KAGGLE GROUP PROJECT

University of Hertfordshire UH

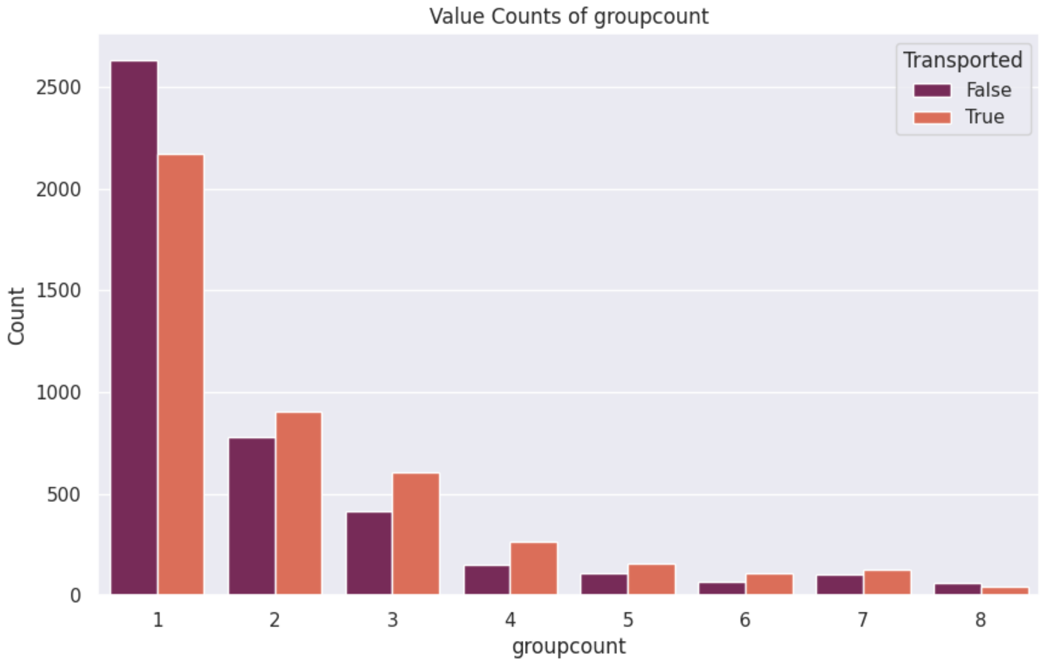
EXPLORATORY DATA ANALYSIS

- The correlation heatmap shows that the numerical features do not significantly correlate with one another. Because there aren't many strong correlations between the numerical characteristics, it's likely that the features are independent of one another.
- This makes machine learning models more effective by lowering multicollinearity and guaranteeing that each feature adds distinct information to the target variable's prediction.



EXPLORATORY DATA ANALYSIS

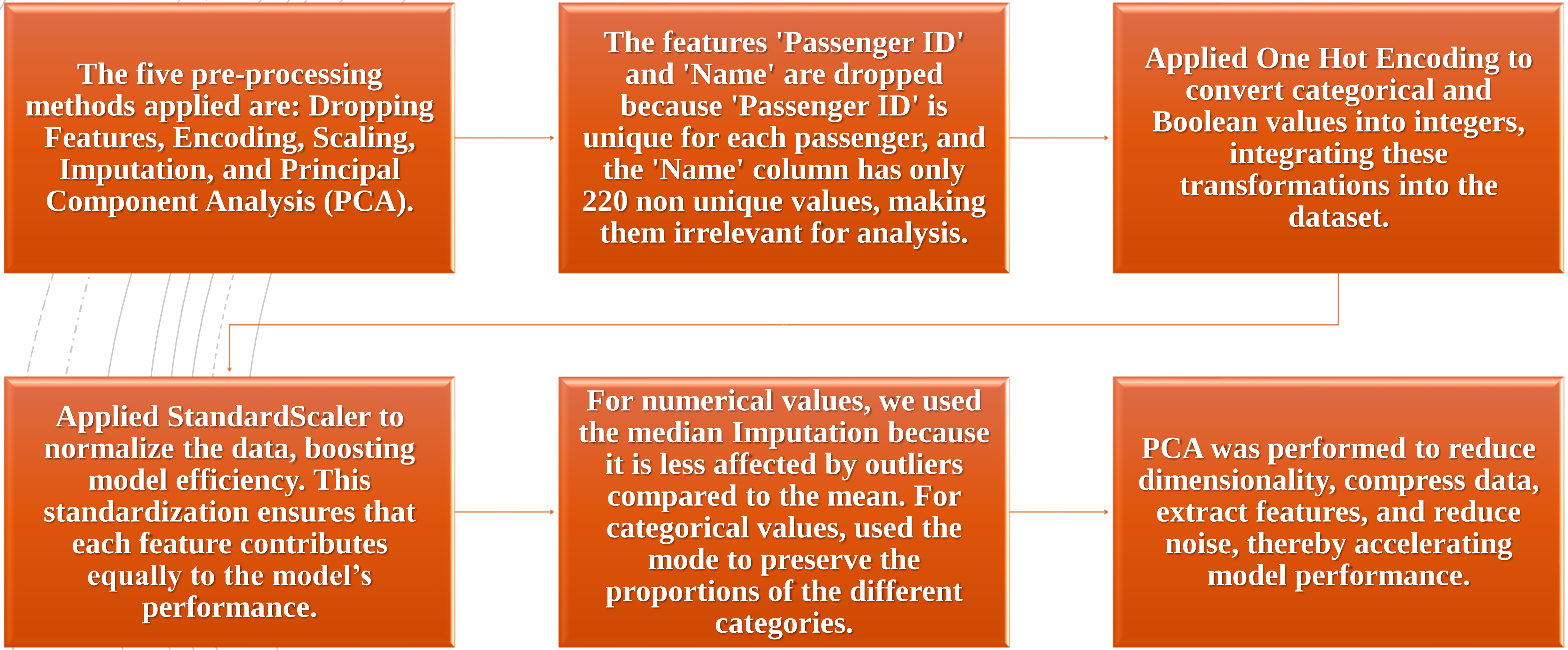
- The ‘Groupcount’ by 'Transported' count graphic makes it evident that Travelers in groups of two to seven are more likely to be transported.
- This implies that the quantity of passengers in a group affects the results of transportation.
- Consequently, two new features, 'Group' and 'Groupcount', have been created to capture this relationship in the dataset.
- Based on the graph indicating lower transportation likelihood for group count 1, a new feature 'SoloTraveler' was also created by checking if the 'Groupcount’ was 1, capturing passengers traveling alone in the dataset.



TEAM 2
KAGGLE GROUP PROJECT

University of Hertfordshire UH

PRE-PROCESSING



The five pre-processing methods applied are: Dropping Features, Encoding, Scaling, Imputation, and Principal Component Analysis (PCA).

The features 'Passenger ID' and 'Name' are dropped because 'Passenger ID' is unique for each passenger, and the 'Name' column has only 220 non unique values, making them irrelevant for analysis.

Applied One Hot Encoding to convert categorical and Boolean values into integers, integrating these transformations into the dataset.

Applied StandardScaler to normalize the data, boosting model efficiency. This standardization ensures that each feature contributes equally to the model’s performance.

For numerical values, we used the median Imputation because it is less affected by outliers compared to the mean. For categorical values, used the mode to preserve the proportions of the different categories.

PCA was performed to reduce dimensionality, compress data, extract features, and reduce noise, thereby accelerating model performance.

TEAM 2
KAGGLE GROUP PROJECT

University of Hertfordshire UH

MODEL

Model	**Test Accuracy**
XGBoost Classifier	0.79448
Support Vector Classifier (SVC)	**0.80149**
Logistic Regression	0.79097
Decision Tree Classifier	0.78442
Random Forest Classifier	0.79611
K-Nearest Neighbors (KNN)	0.7652

From six popular classification machine learning models chosen, SVC had the highest accuracy.

Support Vector Classification (SVC) is a supervised machine learning algorithm used for classification.

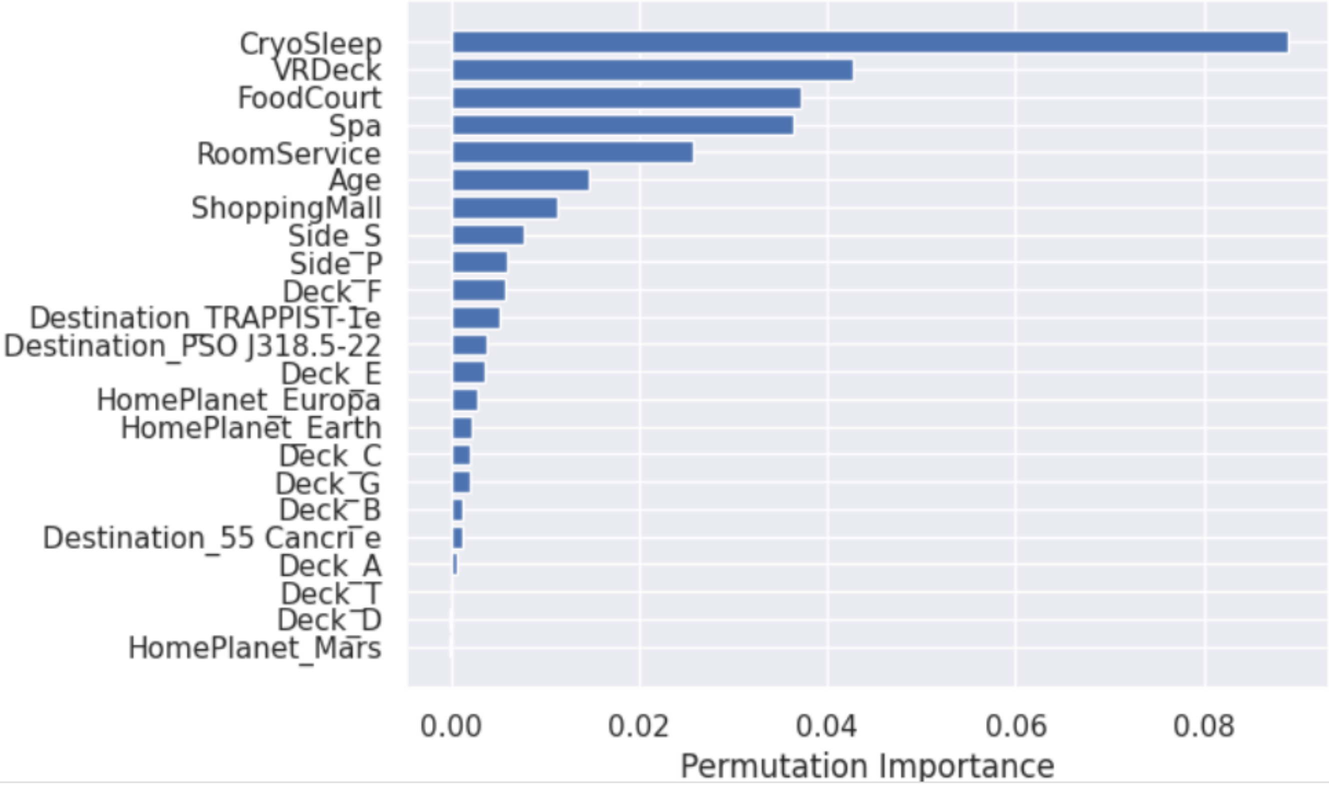
The main goal is to maximize the margin, the distance between the hyperplane and the nearest data points (support vectors) in an N-dimensional feature space.

Model tuning: To determine the ideal hyperparameters (such as C, kernel type, and gamma), cross-validation was done.

Hyper tuning and different preprocessing methods were applied for each Kaggle submissions.

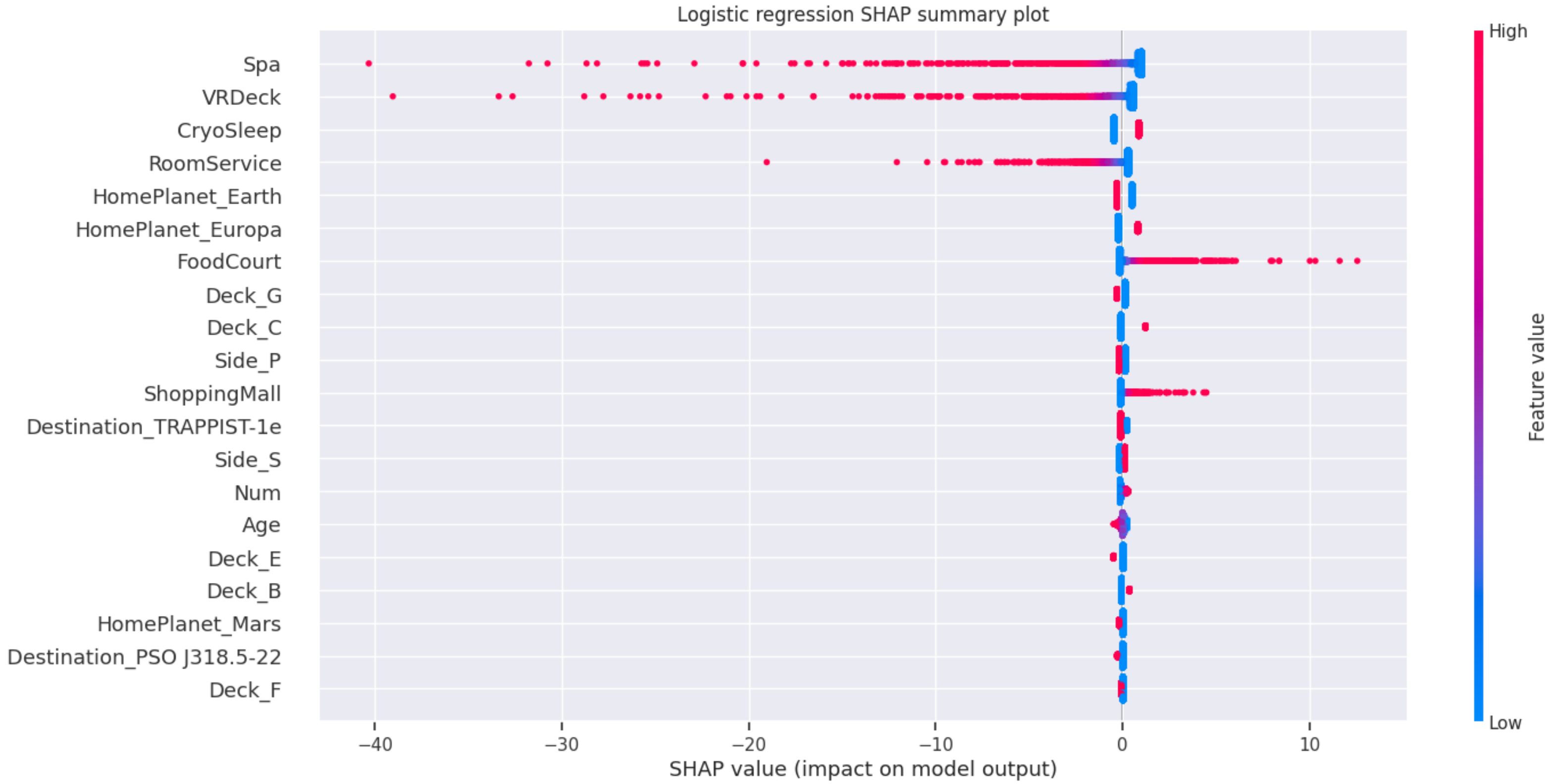
XAI

- CryoSleep is the most important, weighing nearly twice as much as the following item, VRDeck. Food Court and Spa share similar values. The Age feature has reasonably near values, whereas the other features have primarily small values, which are steadily declining. Furthermore, other features seems to have zero importance.



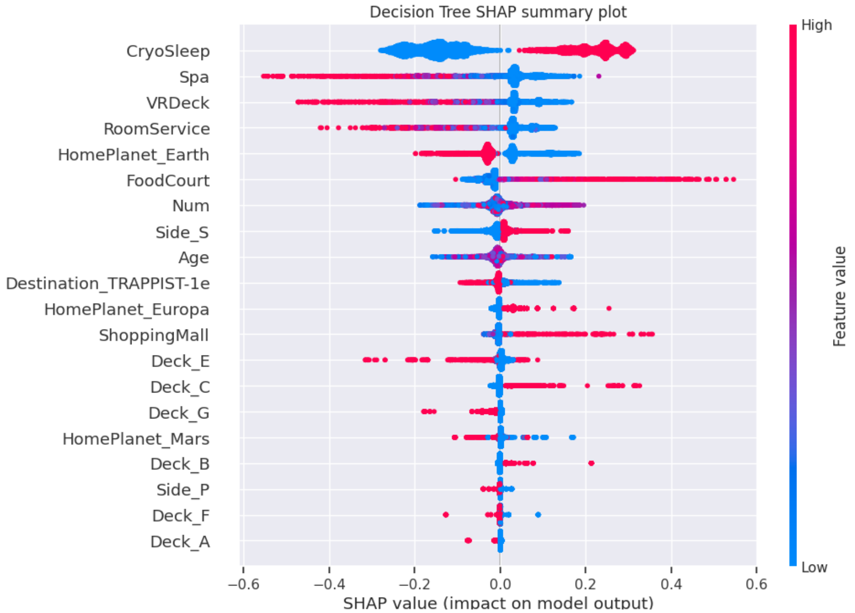
Permutation importance plot for SVC

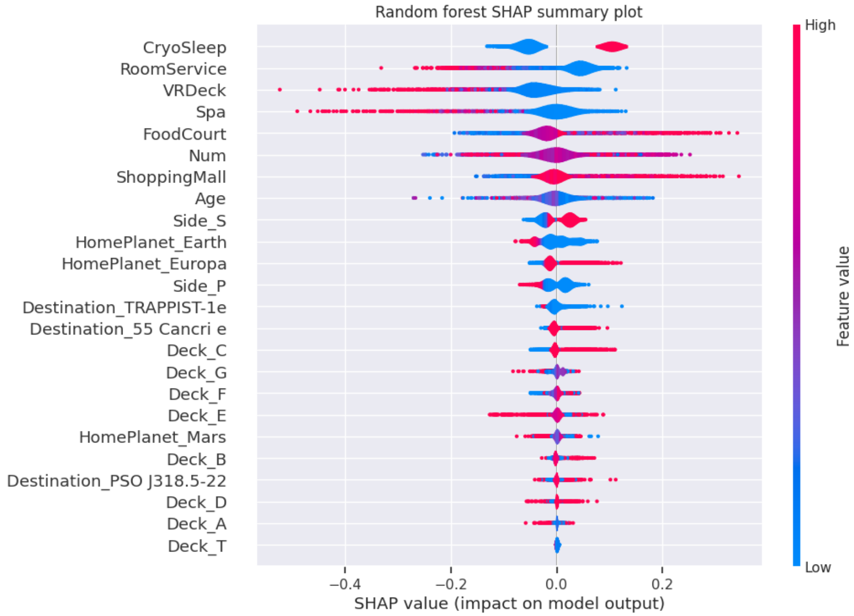
XAI

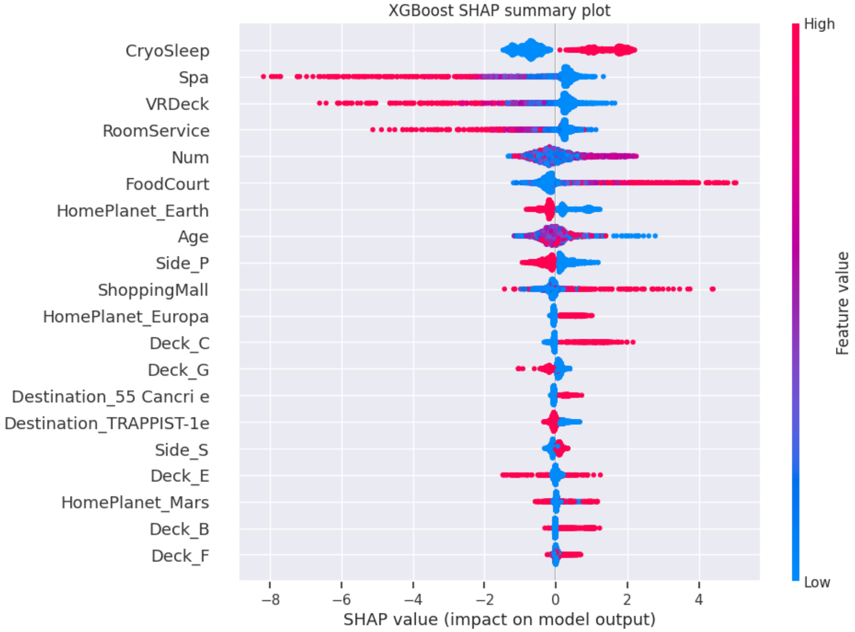


- Spa, VRDeck seems to be the two most important features for the Logistic regression model, followed by RoomService and FoodCourt. Every other feature except ShoppingMall has a relatively smaller importance.
- It can also be noted that there’s a very clear separation between the higher and lower points for CryoSleep feature.

XAI COMPARISON OF DECISION TREE, RANDOM FOREST AND XGBOOST MODELS





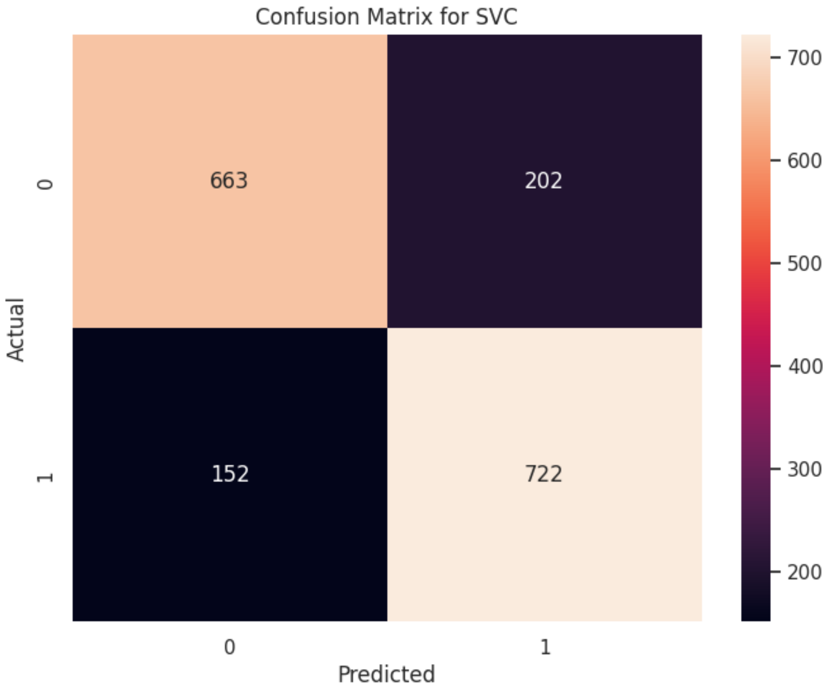


- A comparison of the Shapley summary plots of these 3 methods, all of which are **based on decision trees,** reveals that Spa, VRDeck, RoomService and FoodCourt are the predominant features. However there does not seem to be any recognizable patterns in the behaviour of the remaining features, but most features have some relative importance .

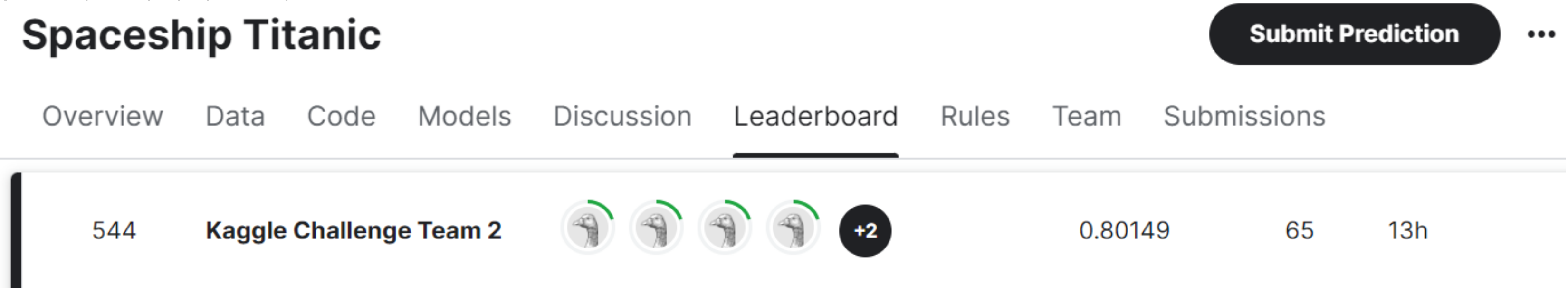
TEAM 2
KAGGLE GROUP PROJECT

University of Hertfordshire UH

RESULTS



- XAI analysis reveals that throughout all models the six features- CryoSleep, Spa, VRDeck, RoomService , FoodCourt and ShoppingMall have a higher importance, with varying levels of significant in each case.
- EDA revealed that 5 of those 6 numerical features(except Cryosleep) had a relatively skewed distributions which we speculate to be the reason behind this. Furthermore, a clear gap can be observed between the high and low feature importance points in CryoSleep, which is probably due to the binary nature of that feature.
- Highest Kaggle submission score was 0.80149 at position 544(at the time of submission). The total number of submission were 65 and all models showed a significant increase in their score from their respective initial submissions.
- SVC had the highest score which went from the initial score of .7814 to .8001 upon hyperparameter tuning and additional pre-processing.
- Our SVC model appears to accurately predict when passengers are transported, as it has fewer False Negatives. This suggests that our model could be well-suited for this task.


Spaceship Titanic

544 Kaggle Challenge Team 2 0.80149 65 13h

CONCLUSIONS

All chosen models improved their initial submission scores upon hyperparameter tuning and additional pre-processing. All chosen models performed our prediction tasks in a considerably well manner.

Tree based approaches showed some similar behaviour, specifically in the feature importance for those models, while the rest of the approaches were fairly distinct.

Six out of the fourteen features seem to have a higher prediction importance in all models, as revealed by the XAI findings. Five of these are the amounts spent on luxury amenities and the remaining one being cryosleep.

SVC outperformed all other models with a higher score on Kaggle, which was improved upon from its initial score of 78.14% to 80.15%. It accurately predicts transported passengers with fewer False Negatives, suggesting it is well-suited for this task.