

Project 4: Spaceship Titanic | Kaggle competition

Daniela Montiel
Esaú Cervantes
Jessica Montoya



Introduction: Kaggle competition

What is a Kaggle competition?

A Kaggle competition is a data science competition hosted on the Kaggle platform, where participants compete to develop the most accurate predictive models or solutions for specific real-world problems. Kaggle competitions typically provide datasets, evaluation metrics, and problem descriptions.

Participants build and train machine learning models using the provided data and submit their predictions or solutions to Kaggle. The submissions are evaluated against a holdout dataset, and participants are ranked on a leaderboard based on the performance of their models.

Kaggle competitions serve as a platform for data scientists, machine learning practitioners, and enthusiasts to showcase their skills, learn from others, and tackle challenging problems in various domains.





Spaceship Titanic

Situation

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, the challenge is to predict which passengers were transported by the anomaly using records recovered from the spaceship's damaged computer system.

Dataset description | Variables

train.csv - Personal records for about two-thirds (~8700) of the passengers, to be used as training data.

- **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.

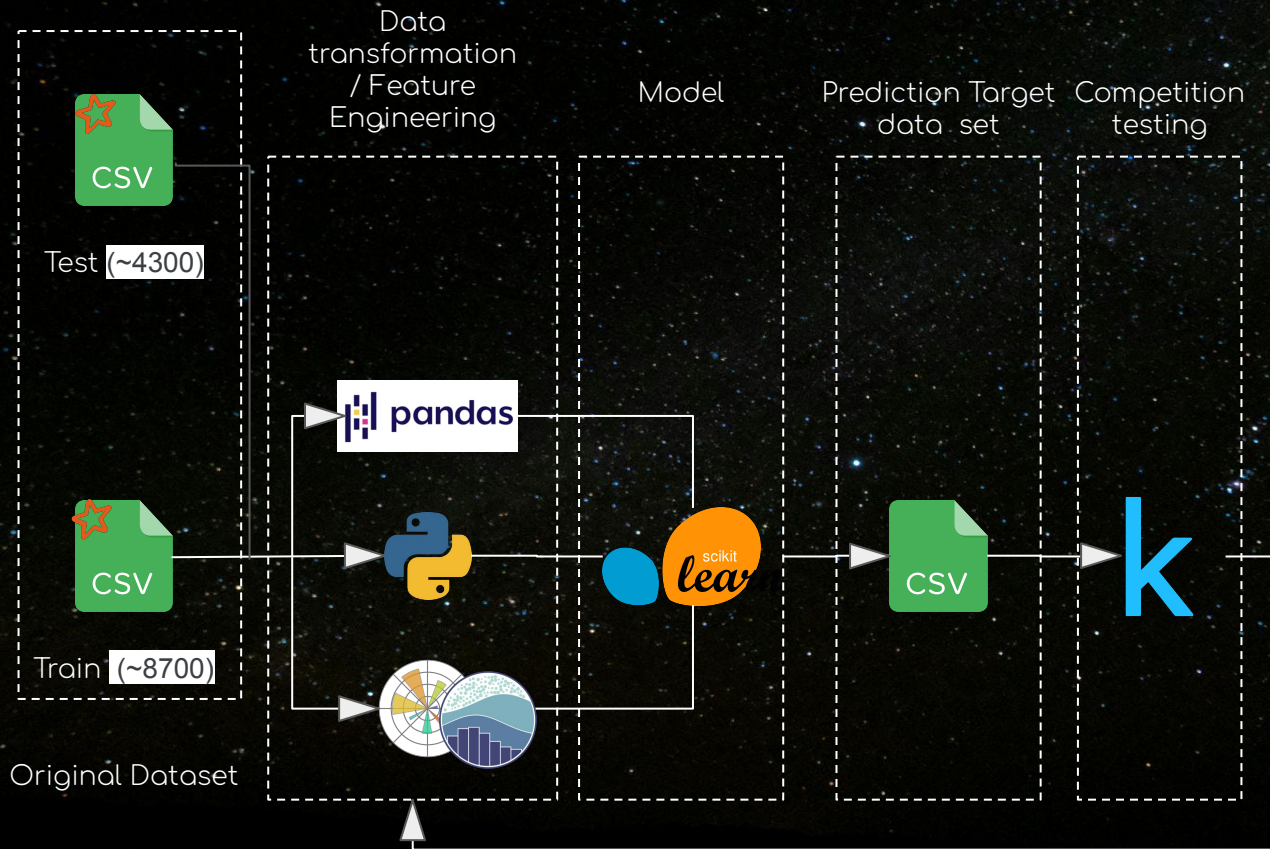
Dataset description | Variables

test.csv - Personal records for the remaining one-third (~4300) of the passengers, to be used as test data. Your task is to predict the value of **Transported** for the passengers in this set.

sample_submission.csv - A submission file in the correct format.

- **PassengerId** - Id for each passenger in the test set.
- **Transported** - The target. For each passenger, predict either **True** or **False**

Data journey



Data journey



Data EDA

Analyzing and exploring data to understand its characteristics, relationship and patterns. Dataset split into training and testing subsets for machine learning tasks.



Feature Engineering

Transforming raw data to enhance model performance by selecting, creating and modifying the features.



Data ML Model: Random Forest

A popular machine learning algorithm for classification and regression tasks, created and optimized by tuning hyperparameters and assessing performance.

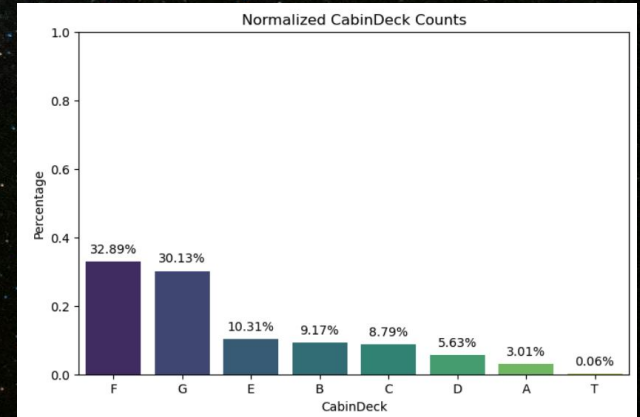
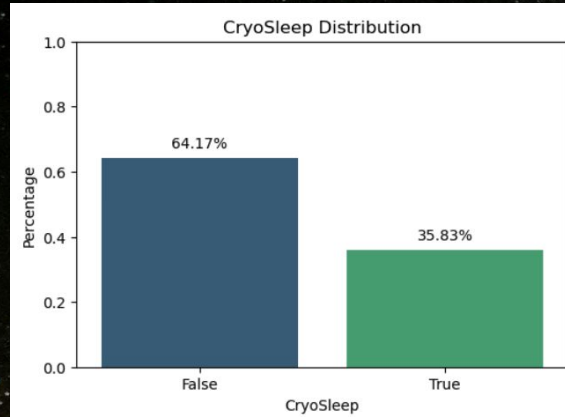
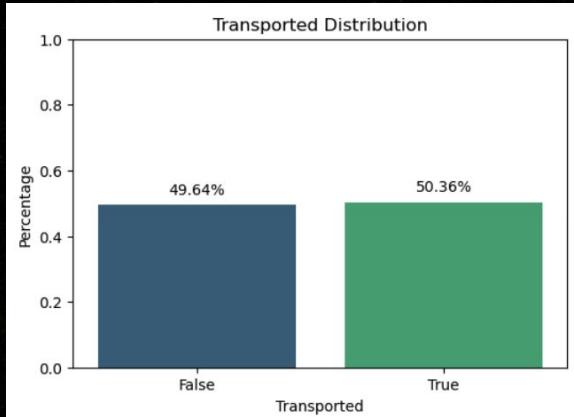


Kaggle competition testing

Competition on Kaggle requires using the provided dataset and submitting predictions for evaluation. Results are showcased on the comprehensive dashboard, enabling tracking for each submission.

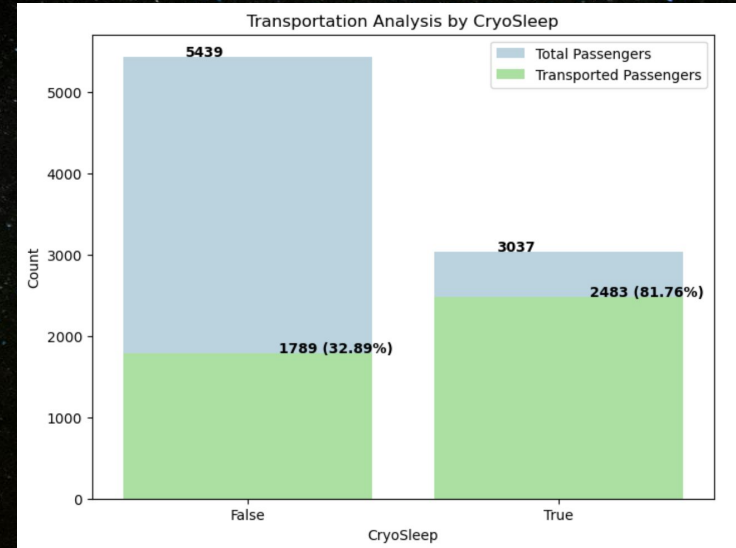
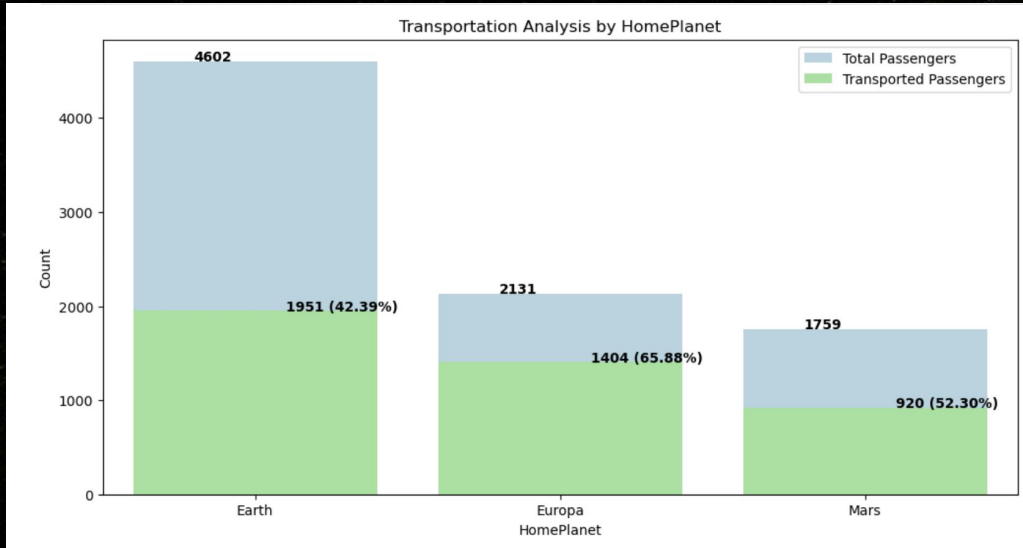
Data journey: EDA

We begin data exploration. The provided dataset 'Train' contains 8693 rows and organizes information into 14 columns: 'PassengerId', 'HomePlanet', 'CryoSleep', 'Cabin', 'Destination', 'Age', 'VIP', 'RoomService', 'FoodCourt', 'ShoppingMall', 'Spa', 'VRDeck', 'Name', 'Transported'. However, the dataset contains null values; to enhance null data filling, we decided to conduct an initial exploration of variables versus the main challenge variable: 'Transported'



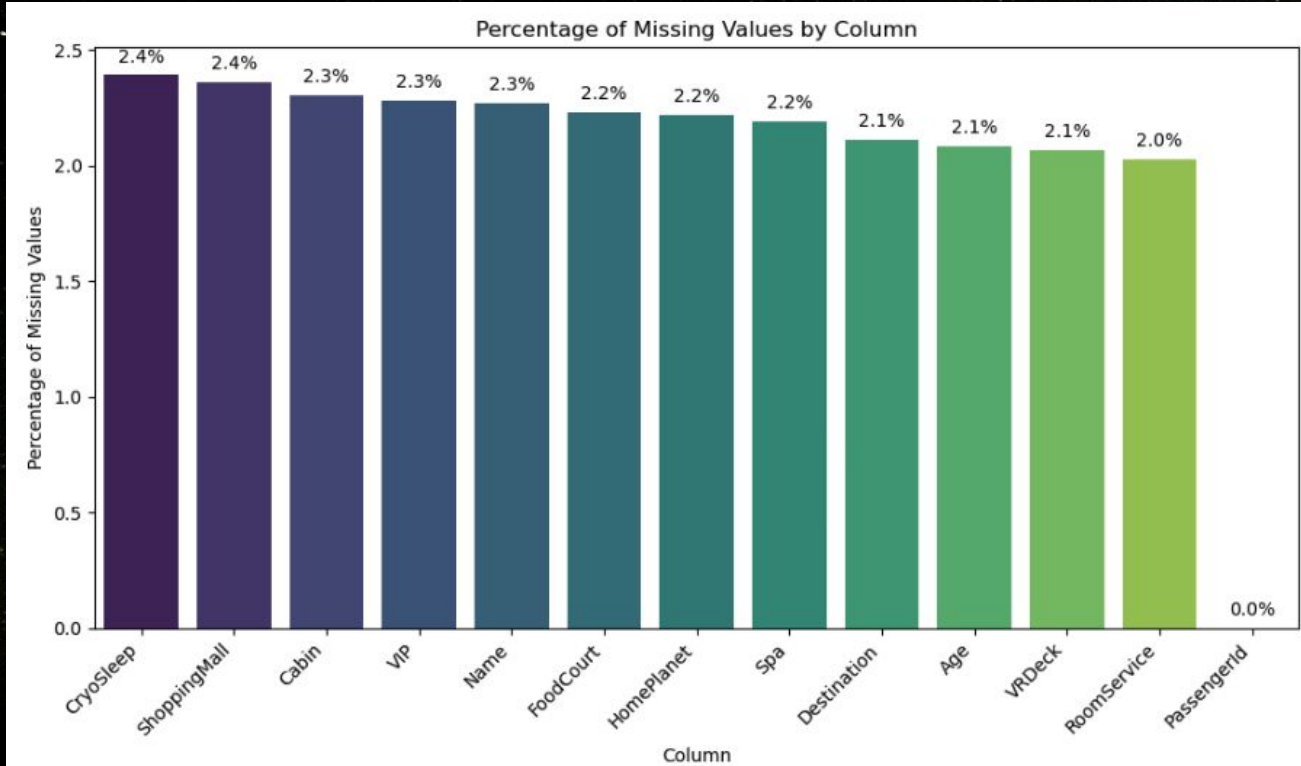
Data journey: EDA

Afterwards, we analyze the data distribution and merged with bivariable analysis.



Data journey: EDA

There are 3,083 rows with null values, about 23.77% of the data has some null value.

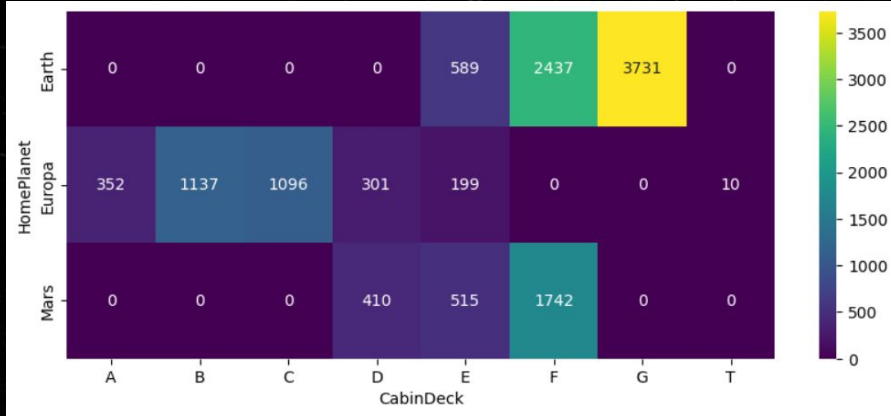


Data journey: Null Values treatment

After reviewing the data, the team adopted the following criteria. Null values would not be dropped from the dataset; instead, they would be filled with the following approaches:

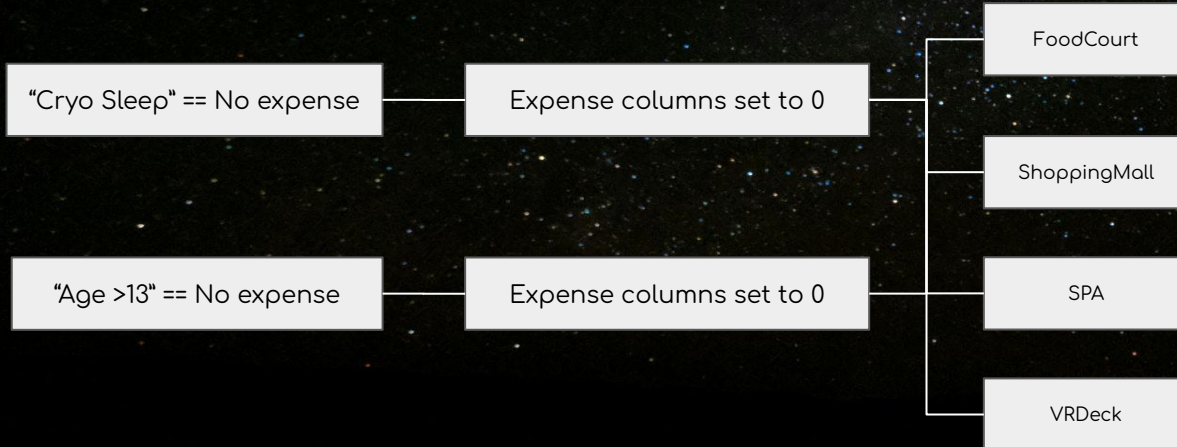
- Infer hidden rules from the data (Insights)
- Use Machine Learning to predict the null value from the known values
- Use the mean or the mode for each feature to input the null value

Null Values treatment: Infer hidden rules from the data (Insights)



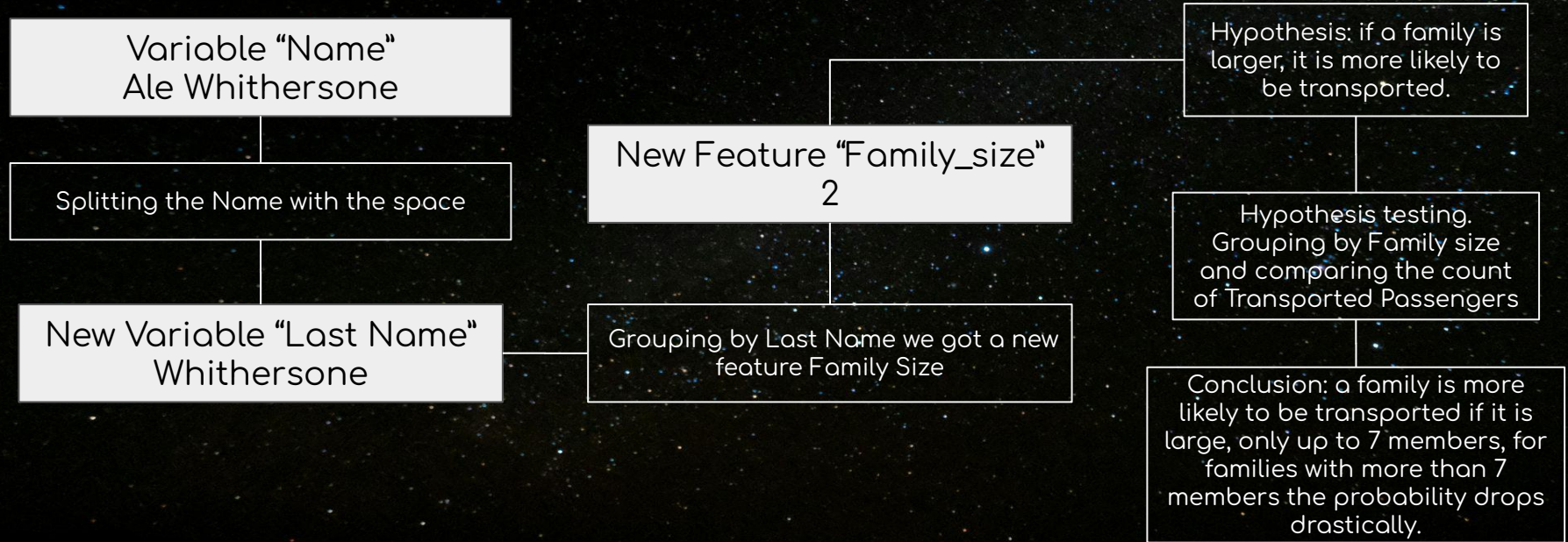
Hidden rules

- All the people in Cabin Decks A, B, C and T came from Europa
- All the People in Cabin deck G came from Earth
- People in Cryo Sleep didn't spend money
- Kids under 13 years also didn't



Data journey: Feature Engineering

Feature engineering is the process of selecting, manipulating and transforming raw data into **new features** suitable for machine learning models.



Data journey: Feature Engineering

Features considered for the models:

- Original Features:
 - Numerical: Age, Room Service, Food Court, Shopping Mall, Spa, VR Deck
 - Categorical : Cryo Sleep, Home Planet, Destination, VIP
- Engineered Features:
 - Numerical: Group size, Family size, Luxury expenses, Regular expenses, Total expenses
 - Categorical: Cabin Deck, Cabin Side, Is Alone

Preprocessing:

- Pandas get dummies for categorical columns
- Standard scaling was used in the numerical columns for the Neural Network Models but it wasn't necessary for the Random Forest models

Data journey: ML Models tested | Attempts summary

1. Decision Tree Model

- Hierarchical structure representing a sequence of decisions.
- Each internal node represents a decision based on a feature.
- Builds tree recursively by selecting best feature splits.
- Interpretable and easy to understand, but may overfit without proper pruning.

.7898

Model discarded due to the limitation of its components.

2. Neural Network Model

- Computational model inspired by the human brain.
- Consists of interconnected nodes organized in layers.
- Learns complex patterns and relationships in data through adjusting weights between nodes.
- Capable of handling various tasks like classification, regression, and clustering.

.7776

Model discarded due to its low accuracy and overfitting

3. Random Forest Model

- Ensemble learning method combining multiple decision trees.
- Each tree is trained on a random subset of data and features.
- Final prediction is the average (regression) or mode (classification) of individual tree predictions.
- Known for high accuracy and robustness.

.80406

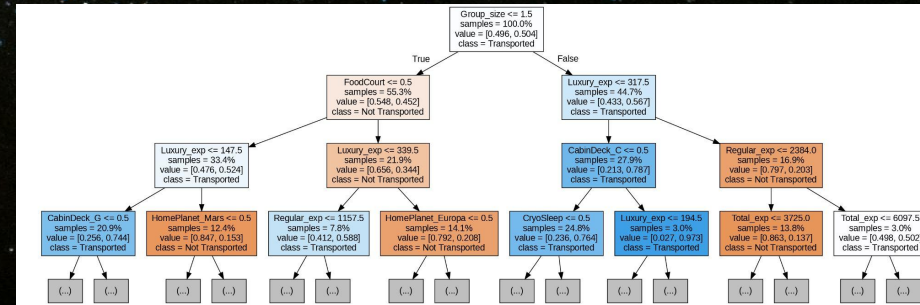
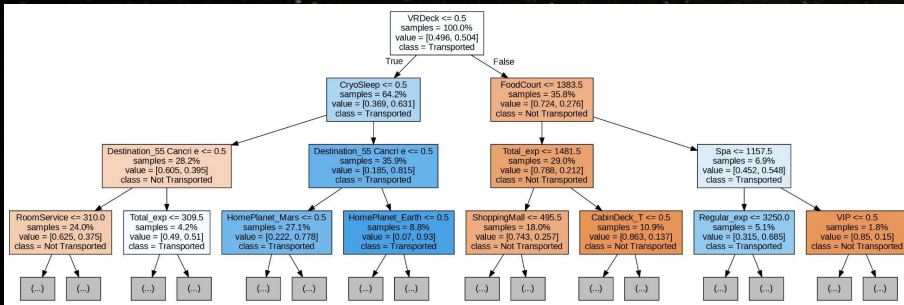
Selected

Data journey: Data ML Model | Random Forest

Random Forest is a machine learning algorithm that operates by constructing multiple decision trees during training and outputs the mode (classification) or the mean prediction (regression) of the individual trees.

Each decision tree in a Random Forest is trained on a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and decorrelate the trees, leading to more diverse and robust predictions.


```
best_model = RandomForestClassifier(criterion='gini', bootstrap=False, max_depth=9, max_features='auto',  
min_samples_split=3, n_estimators=1000, random_state=1)  
best_model.fit(X_encoded, y)
```




Data journey: Kaggle competition

Once the process is completed, the team submits their predictions for the testing data to Kaggle. Submissions are in the form of a CSV file containing the predicted values for each instance in the testing set. Kaggle evaluates the submissions using predefined evaluation metrics specific to the competition.

The model created for this project, according to its best score, ranks in the 458th place out of 2521 participants.

458	3312 Team		0.80406	21	42m
-----	-----------	---	---------	----	-----



Your Best Entry!
Your most recent submission scored 0.80406, which is an improvement of your previous score of 0.80383. Great job!

[Tweet this](#)

Conclusions:

In conclusion, a machine learning model such as this entails three critical steps for its optimization and enhancement.

- Firstly, addressing null values and opting not to drop them by default; comprehending the provided dataset, and, most importantly, understanding the situation and the problem at hand, as clarity greatly benefits data treatment.
- Secondly, creating new features derived from existing variables, establishing a logical sequence for their creation.
- Lastly, refining the parameters set within the model to be tested at each iteration. Understanding the mathematics behind the model, such as `max_depth` and `n_estimators`, is paramount.

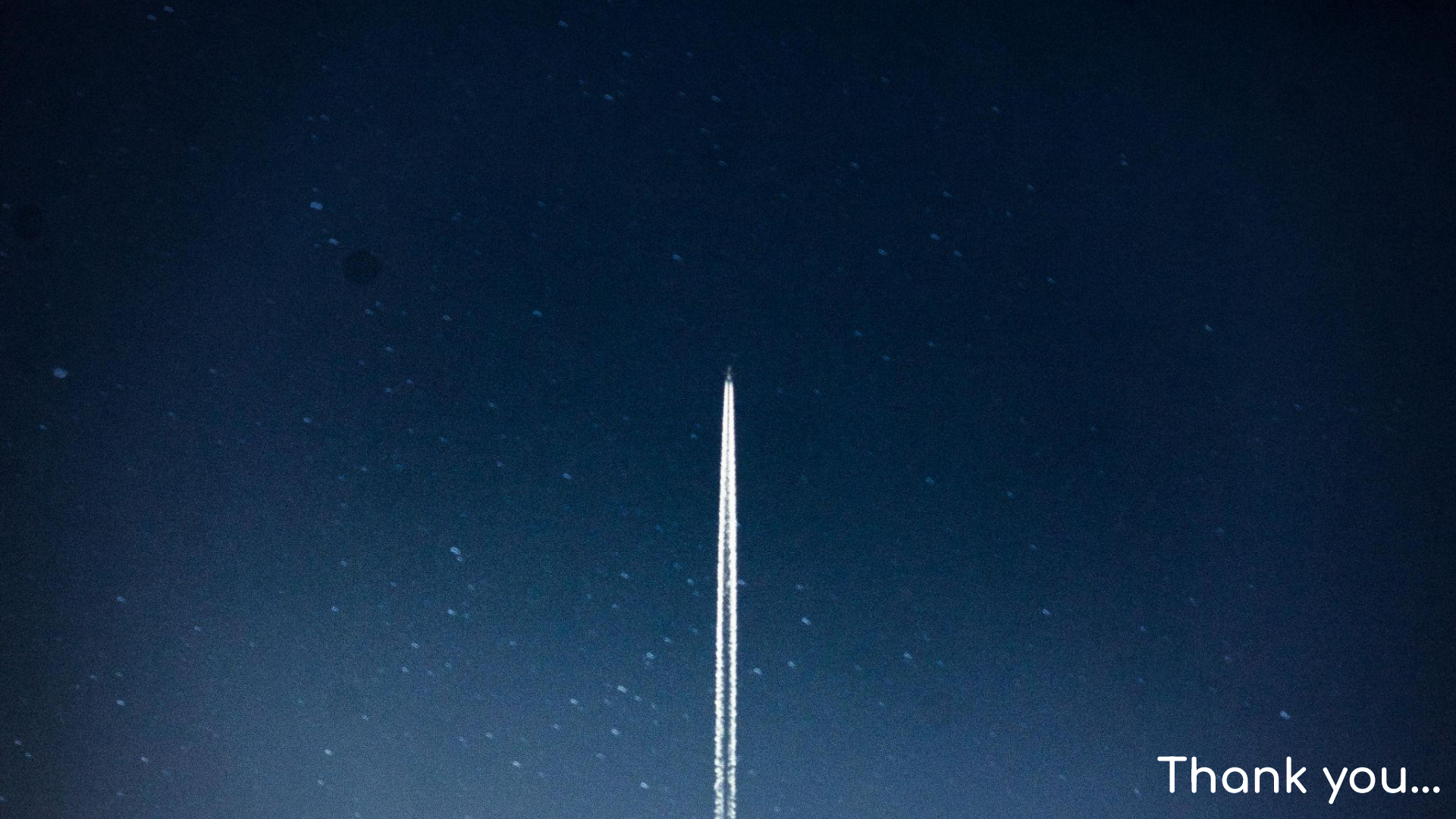
The aforementioned statements stem from other conclusions the team wishes to convey regarding the lessons and insights gained from the bootcamp.

The benefit of employing a logical thought process to digest data management and enhancing data acumen in handling information for addressing various problems is underscored.

Furthermore, we highlight the developed sensitivity in understanding the problem to be solved and firmly justifying the use of one tool over another, rather than solely promoting the use of hot technologies due to market trends.

The skills acquired and the methods learned are applicable and transferrable to other problems and industries within the realm of data analysis.

In summary, this project enabled us to readily identify the role of the data analyst. While code and statistics facilitate the execution of necessary calculations, it is the analyst who must make decisions regarding the significance of variables, establish relationships between them, and determine the appropriate methods for their treatment.



Thank you...