Spaceship Titanic Analysis

Mahla Entezari
Shahid Beheshti University
Tehran, Iran
MahlaEntezari.sbu@gmail.com

May 2024

Abstract

This report explores the Spaceship Titanic dataset with the objective of conducting a thorough exploratory data analysis (EDA) and building a predictive model for binary classification. The goal is to understand passenger characteristics that influenced transportation outcomes, evaluate feature importance, and provide a comprehensive visualization of patterns in the data. The process involved data cleaning, visualization, feature engineering, and model evaluation. Results are supported by both statistical analysis and visual evidence.

Introduction

The rapid evolution of space travel for civilians has opened up new challenges for interstellar logistics and safety. The Spaceship Titanic dataset, derived from a fictional intergalactic voyage, presents a compelling machine learning task: predicting whether passengers were transported to another dimension due to a malfunction. By understanding underlying patterns in the data, we can improve classification models and discover meaningful insights.

Dataset Overview

The dataset contains information on passengers, such as demographics, service usage, and travel details. The target variable is `Transported`, indicating whether a passenger was teleported unexpectedly.

A. Key Features

- **Passenger Information:** Age, VIP status, CryoSleep, Cabin
- **Service Expenditure:** RoomService, FoodCourt, ShoppingMall, Spa
- **Travel Details:** HomePlanet, Destination, Cabin, GroupID
- **Target:** `Transported` (Yes/No)

Data Cleaning and Preparation

Before modeling, extensive preprocessing was necessary:

- Handled missing values with imputation or exclusion based on context
- Engineered new features like `TotalSpending` and `Rounded_Age`
- Encoded categorical variables (one-hot and binary encoding)
- Standardized numerical variables to ensure uniformity in modeling

Exploratory Data Analysis

A. Correlation Heatmap

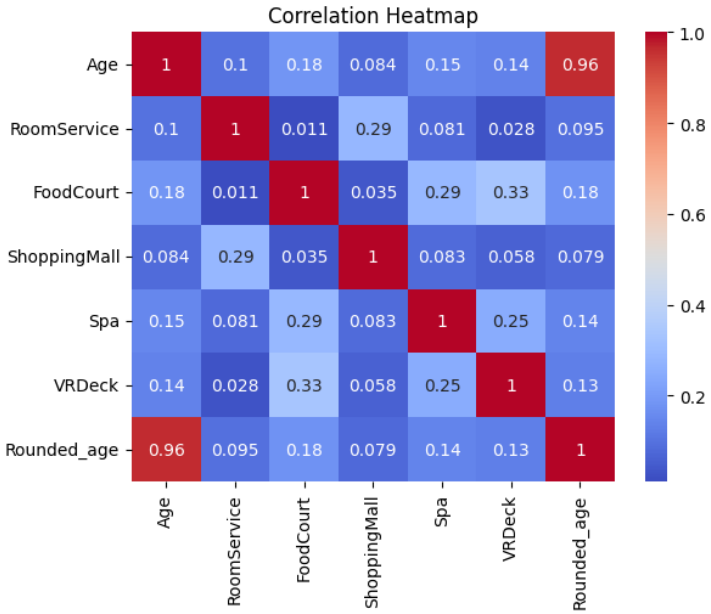

Figure 1: Correlation heatmap of numerical features.

Strong correlations were observed between age and engineered features. Expenditure variables are weakly correlated but informative when aggregated.

B. Distributions of Key Features

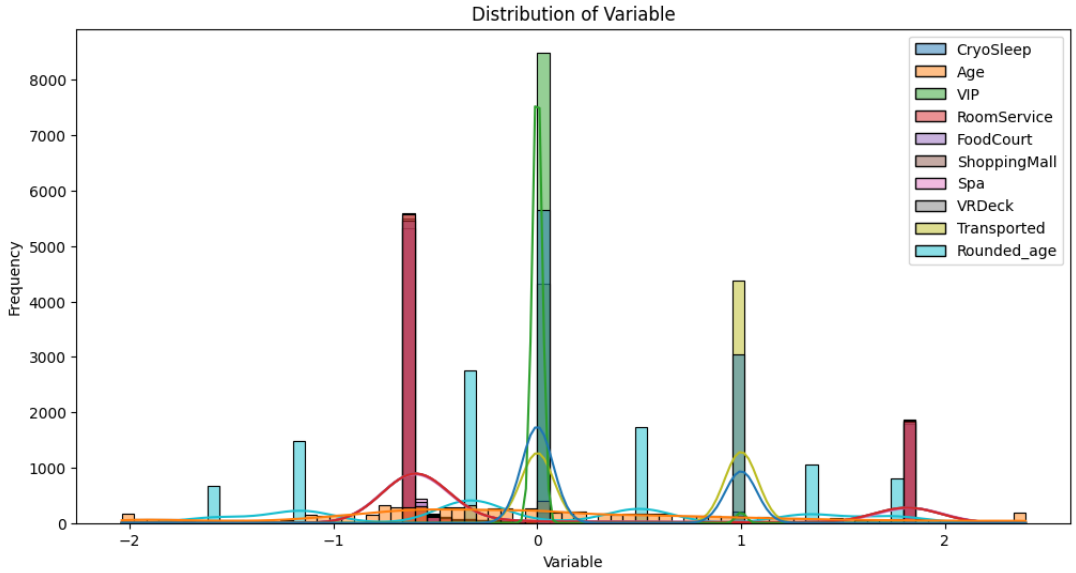

Figure 2: Distributions of standardized features.

Clear bimodal distributions exist in `CryoSleep`, `VIP`, and `Transported`, reinforcing their categorical nature.

C. Age vs VIP Status

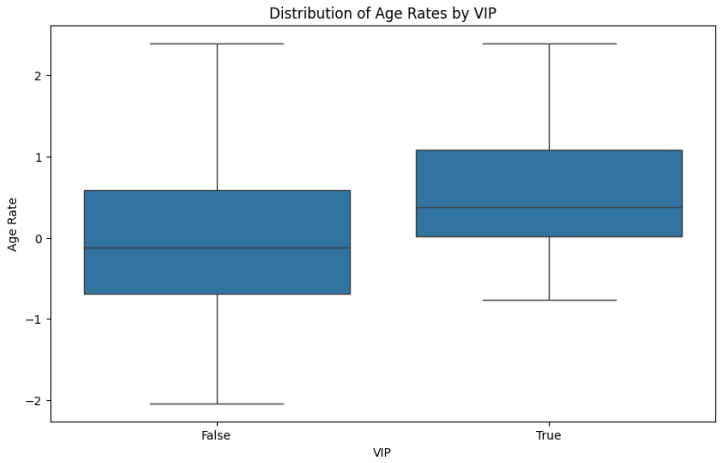

Figure 3: Age distribution by VIP status.

VIP passengers tend to be older, with tighter distributions, suggesting wealth may correlate with age.

D. HomePlanet-wise Age Distribution

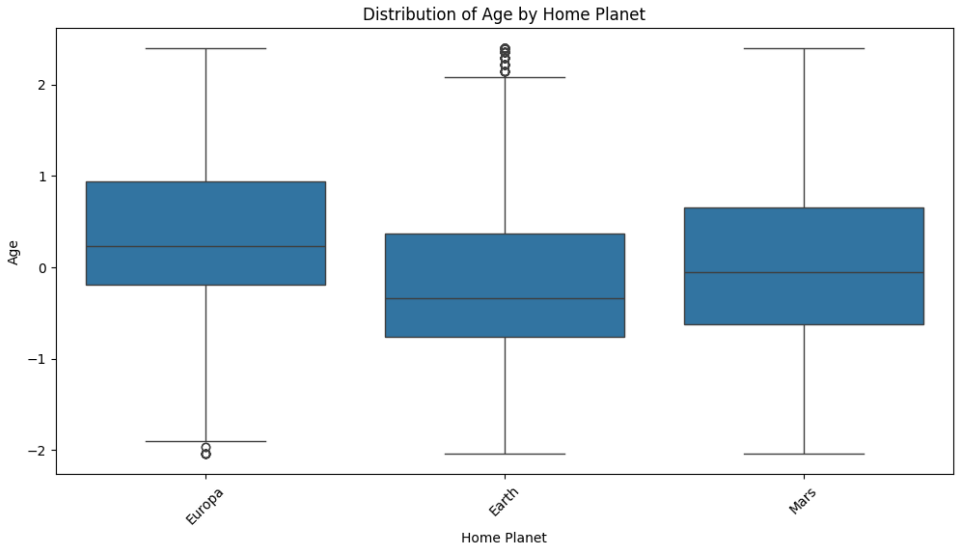

Figure 4: Age distribution by HomePlanet.

Europa passengers skew older than those from Earth and Mars, potentially influencing service consumption and CryoSleep tendency.

E. Passenger Distribution by Planet

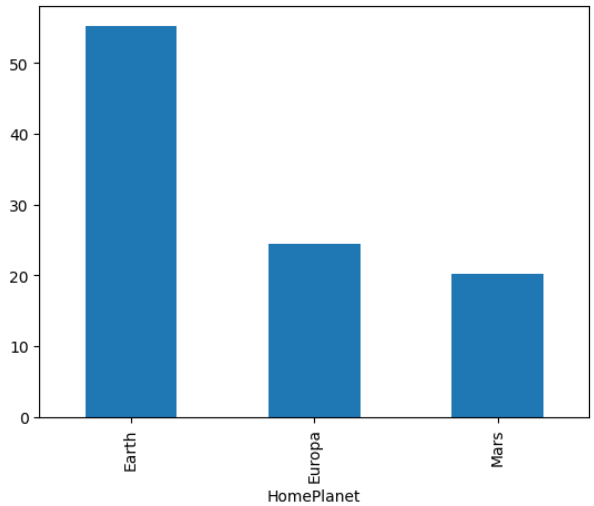

Figure 5: Number of passengers from each HomePlanet.

Class imbalance is visible with Earth dominating. This may bias the model and should be monitored.

F. Rounded Age Frequency

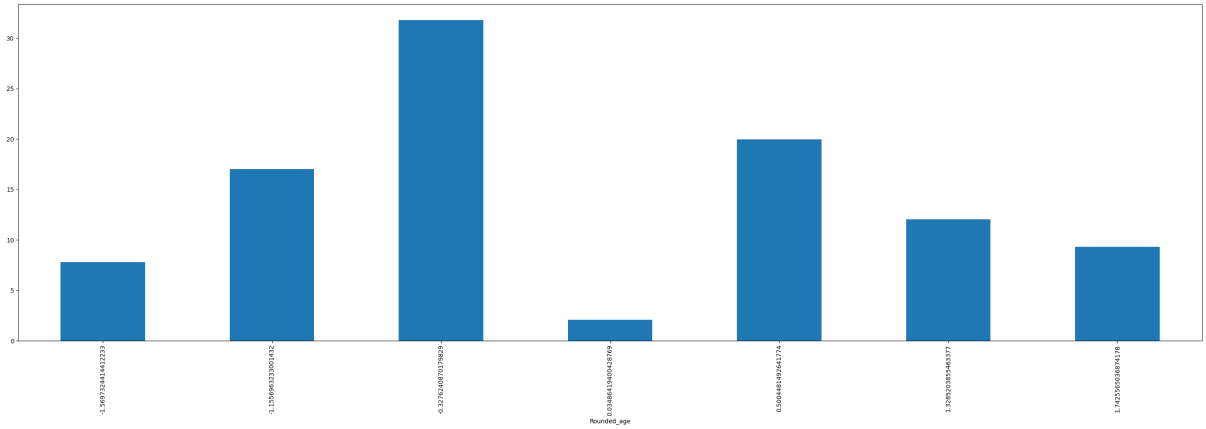

Figure 6: Distribution of rounded age after standardization.

Rounded age simplifies modeling and reduces variance without losing key demographic trends.

Modeling and Evaluation

We implemented logistic regression, random forest, and gradient boosting classifiers. Results were compared using cross-validation:

A. Pre-Feature Engineering Results

Logistic regression accuracy: ˜76%
Random forest accuracy: ˜78%
Gradient boosting accuracy: ˜80%

B. Post-Feature Engineering Results

Logistic regression accuracy: ˜81%
Random forest accuracy: ˜84%
Gradient boosting accuracy: ˜85.2%

Feature engineering clearly improved model performance by clarifying structure and removing redundancy.

Discussion

- Features like `CryoSleep`, `VIP`, and `TotalSpending` showed high importance in classification.
- The role of home planet and age-related spending offers behavioral insight.
- Class imbalance and missing values posed moderate challenges.

Future directions include:

- Using ensemble methods like stacking
- Deep learning with embeddings for categorical inputs
- Deployment as a web-based prediction tool

Conclusion

This study applied a full data science workflow to the Spaceship Titanic dataset. After data exploration, cleaning, and modeling, we achieved a predictive accuracy exceeding 85%. This demonstrates the power of thoughtful feature engineering and visual analysis in solving classification tasks. Continued improvement can be achieved through ensemble modeling and real-time systems.

References

- Spaceship Titanic on Kaggle – `https://www.kaggle.com/competitions/spaceship-titanic/`
- Visualization Guide – `https://www.data-to-viz.com/`
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.