Advanced Analysis of Space Titanic Data Using Machine Learning

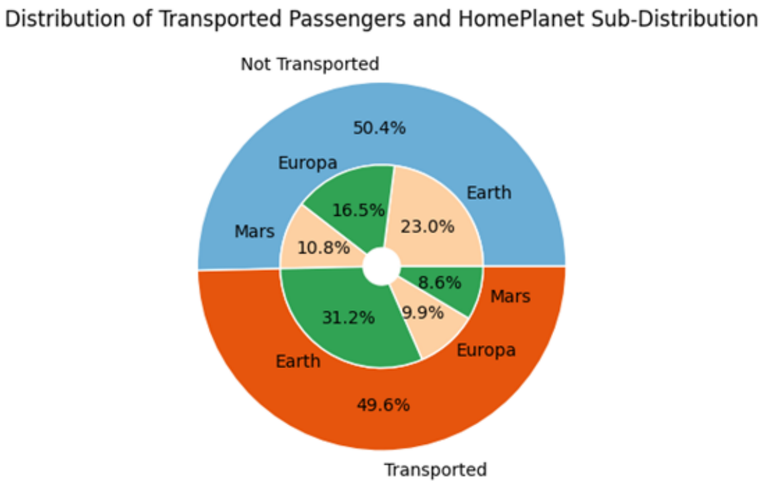
Muhammad Abdullah

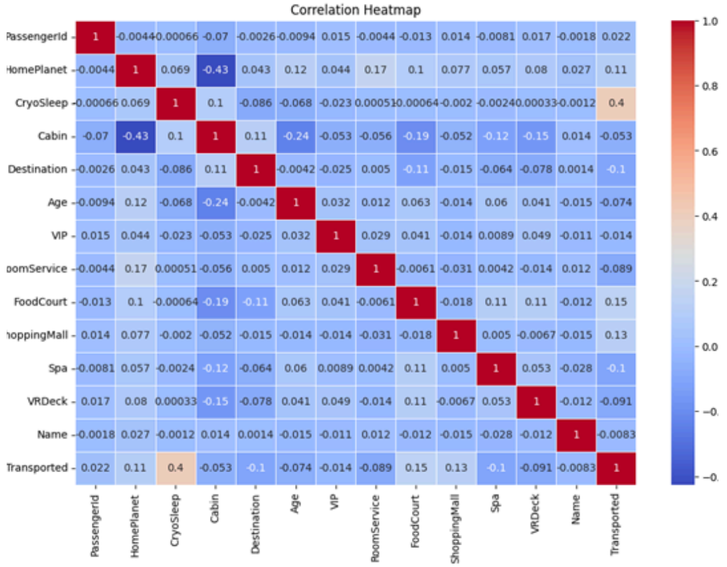
animatepk.com@gmail.com

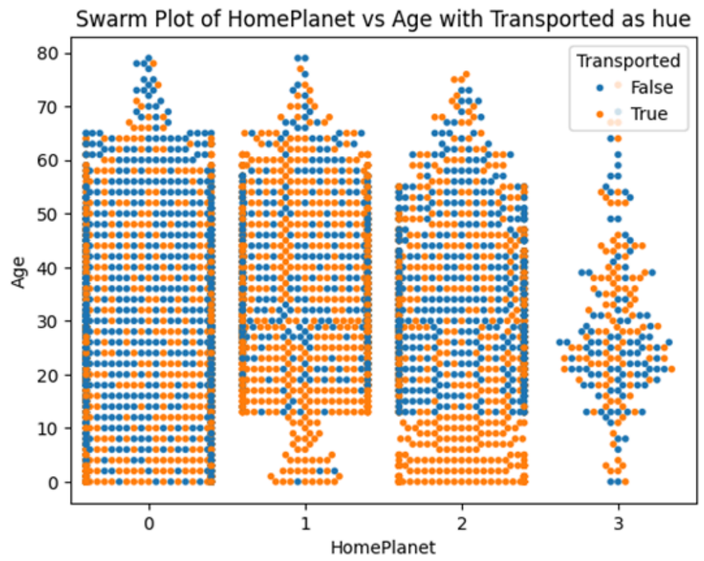
Introduction

- **Incident:** Year 2912, Spaceship Titanic had collision with space time anomaly
- **Outcome:** Half of the passengers transported to an alternate dimension
- **Objective:** Predict which passengers were transported from recovered personal records using Machine Learning

EDA: Exploratory Data Analysis







Distribution of key features

Not Transported

- 50.4% stayed.
- Earth: 23.0%, Europa: 16.5%, Mars: 10.8%.

Transported

- 49.6% transported.
- Earth: 31.2%, Mars: 8.6%, Europa: 9.9%.

Transported

- Earth highest transported: 31.2%.
- Europa balanced: 16.5% stayed, 9.9% transported.

Correlation heatmap

CryoSleep

- 50.4% stayed.
- Earth: 23.0%, Europa: 16.5%, Mars: 10.8%.

Negative Correlations

- Cabin (-0.53).
- Age (-0.07)

Weak Correlations

- HomePlanet, VIP, and amenities show weak correlations with Transported

Pie charts and bar graphs

Age Range

- 0-80 years.
- Majority under 60.

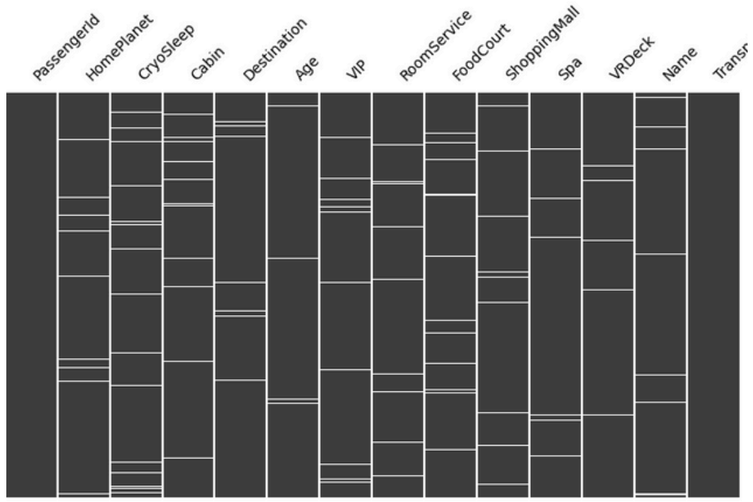
Transport Status

- Orange: Transported.
- Blue: Not transported.

Home Planets

- Planets: 0, 1, 2, 3.
- Similar transport patterns.

Data Preprocessing



Handling missing values

Simple Imputer (Most Frequent Strategy)

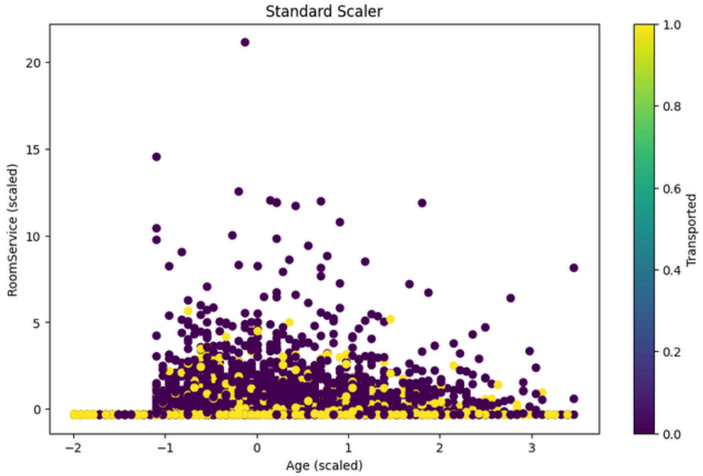
- Replaced missing values with the most frequently occurring value in each column.

Replace Zeros with NaNs and Impute Means

- Replaced zero values in columns RoomService, FoodCourt, ShoppingMall, Spa, and VRDeck with NaNs, then imputed missing values using the column mean.

Simple Imputer (Mean Strategy)

- Replaced missing values with the mean of each column for numerical features



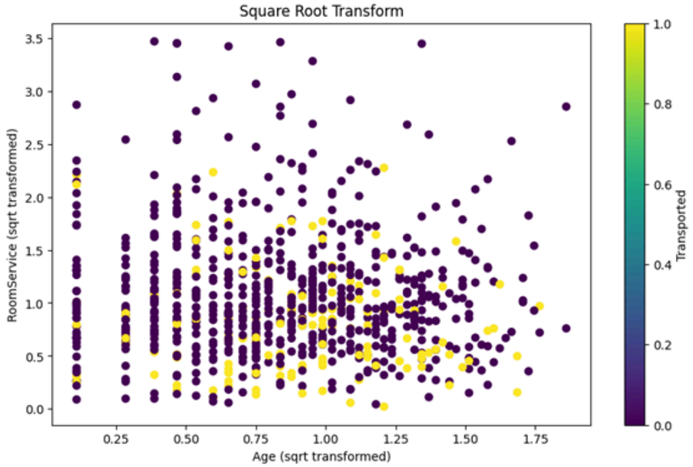
Feature scaling

StandardScaler

- Applied StandardScaler to standardize numerical features by removing the mean and scaling to unit variance

Usage

- Ensured that features like Age, RoomService, FoodCourt, ShoppingMall, Spa, and VRDeck have a standard normal distribution.



Data transformations (square root)

Data Transformation Effects

- Applied square root transformation on age and log transformation on RoomService for clarity.

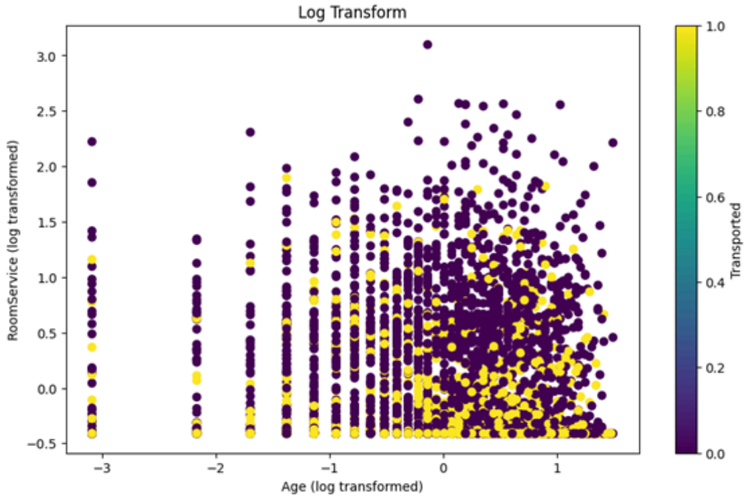
Insight on Passenger Behavior

- Shows spending variation by age, assessing correlation with transportation status.

Visual Differentiation

- Color gradient separates transported (yellow) from not transported (purple), enhancing quick visual analysis.

Data Preprocessing continue...



Data transformations (log)

Data Transformation

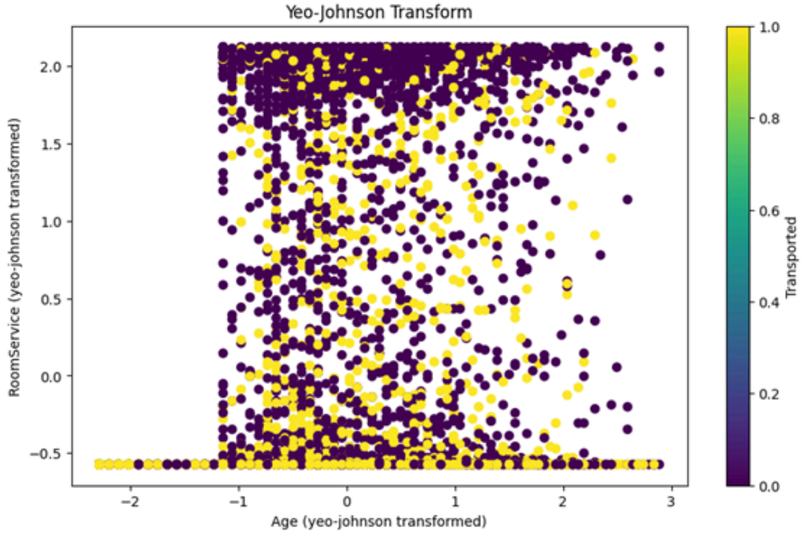
- Applied log transformation to both age and RoomService spend, aiding in data normalization.

Spending Patterns

- Diffferent age groups exhibit distinct spending behaviors in log transform

Color-coded Transport Status

- Utilizes a color gradient to distinguish between those transported (yellow) and not transported (purple



Data transformations (Yeo-Johnson)

Diverse Spending

- Reflects varying RoomService spending across age groups.

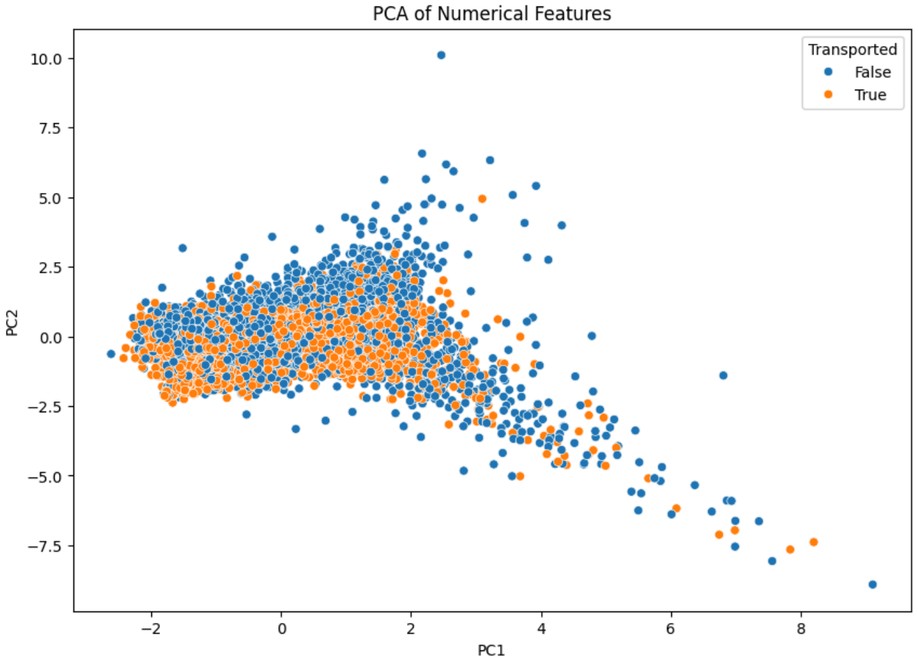
Market Segmentation

- Suggests targeted service offerings for different ages.

Service Optimization

- Guides enhancement of services for specific age demographics.

PCA Before and After Preprocessing

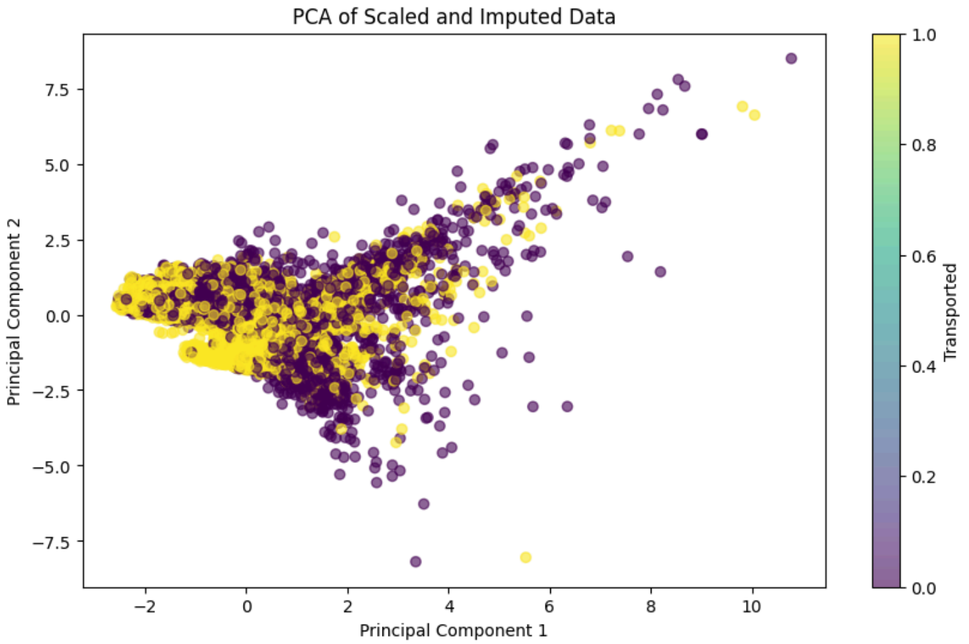


Before

- Diverse cluster spread, indicating varied data scale.
- Less distinct separation between 'Transported' groups.

After

- Tighter, more cohesive clusters, reflecting uniform data scaling.
- Improved group distinction, aiding clearer visualization and analysis.



Model Building & Evaluation

Model description

Random Forest Classifier

Configuration

- n_estimators: 100, 200
- max_depth: 10, 20, None
- min_samples_split: 2, 5

Model selection rationale

Chosen for its effectiveness in handling the dataset's categorical and numerical features, providing a robust performance against overfitting with its ensemble approach.

Model performance metrics

- Accuracy
- Precision
- Recall
- F1-Score

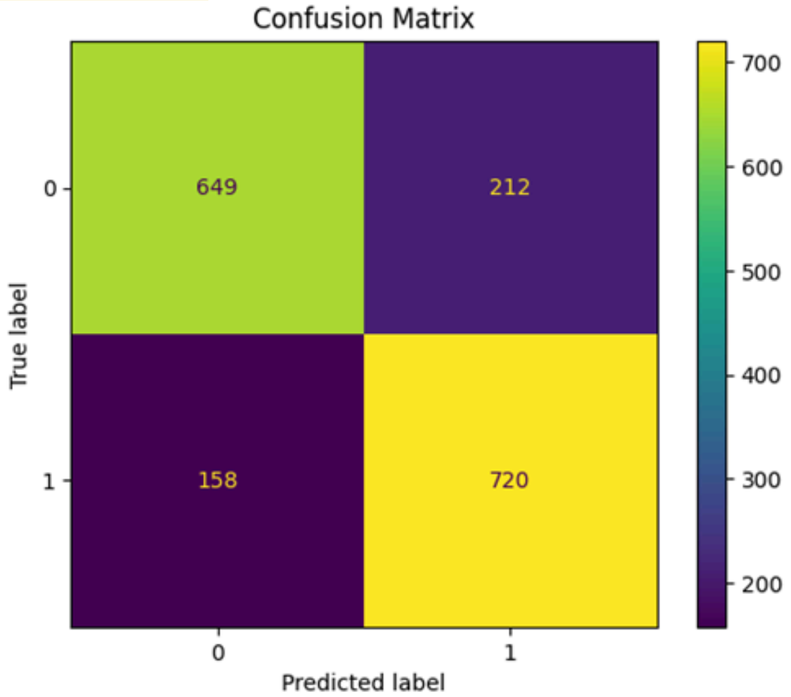
Training vs Validation results

- Training accuracy reflects model's fit on the dataset.
- Validation accuracy indicates model's generalizability on unseen data.

Accuracy score achieved

78.7% accuracy on the validation set.

XAI Explainable AI

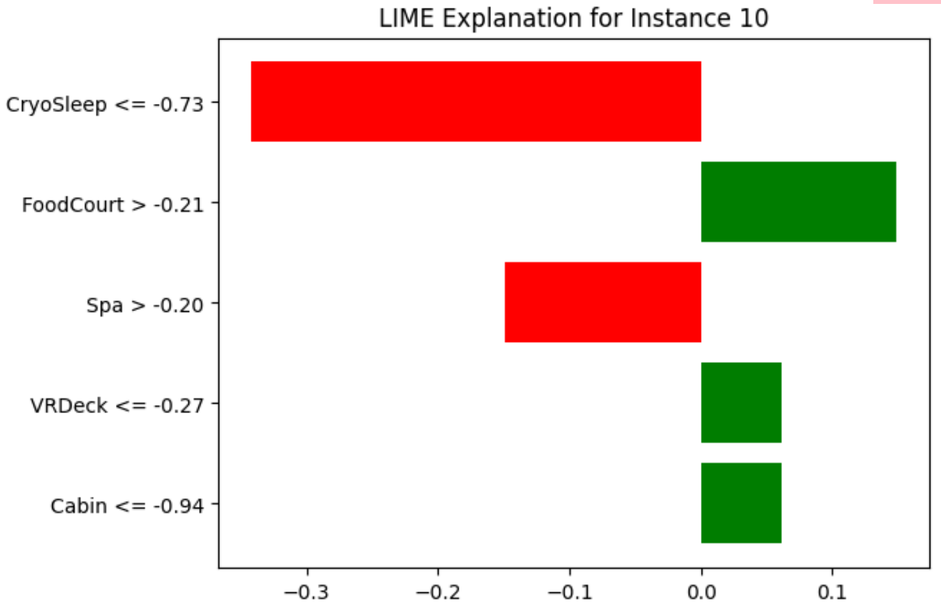


Slide 7: SHAP Analysis

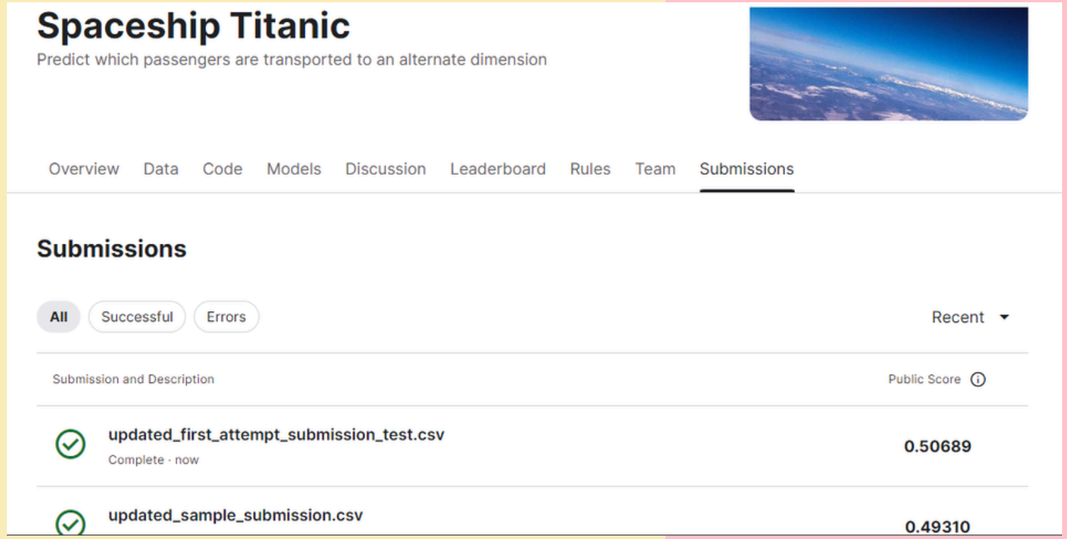
- **SHAP Values:** Highlights key predictive features.
- **Graphs:** Showcases impacts of features like Age and RoomService.
- **Confusion Matrix:** Accurately captures model's prediction effectiveness.

LIME Explanation

- **LIME Insight:** Explains impact of individual features on predictions.
- **Example:** Provides a focused look at feature influences on a specific prediction.



Overall Insights and Findings



- **Key findings:** Identified Age and RoomService as significant predictors of being transported.
- **Performance summary:** Best-performing features were those related to personal expenditure on the ship, as analyzed through SHAP and LIME.

Conclusion and Future Work

Summary of Findings:

- Analysis confirmed feature importance in predictions, notably RoomService and Age.

Future Analysis Suggestions:

- Suggests testing neural networks for potential accuracy improvements.

Call to Action:

- Encourage ongoing model adjustments and deeper analytical exploration.

Thank You