Predicting Passenger Transport in the Spaceship Titanic

Using Machine learning to solve a sci-fi Mystery

CSC 44700 Final Project

Zi Xuan Li & Kenny Zhu

Problem & Motivation

Problem Statement

Binary Classification: Predict whether a passenger was transported (True/False) based on passenger records.

Why It Matters

- Rescue Priority: Identifying high-risk groups for targeted search efforts.
- Design Insights: Reinforcing vulnerable areas of the spacecraft.
- Economic Impact: Reducing false negatives (missed rescues) saves more lives, while minimizing false positives helps to avoid wasting resources.

Dataset

Spaceship Titanic

The dataset contains personal passenger records with features relating to their disappearances such as PassengerId, HomePlanet, CryoSleep, Cabin, Age, etc.

Data Provided

- Training data: personal records for about ~8700 of the passengers.
- Testing data: personal records for about ~4300 of the passengers.

Approach

1) Load the datasets into workspace.
2) Analyze the dataset to look for outliers, missing data, negative values.
3) Explore and understand the importance of each feature.
4) Create new features using existing features.
5) Handle missing values by looking for patterns between features.
6) Drop unwanted features.
7) Select the best model.

Key Findings

Top Takeaways

- CatBoost achieved an accuracy of 80.7% accuracy (top 5% finish).
- Passengers in cabin deck B & C were more likely to be transported (saved).
- Solo passengers were less likely to be transported.

	Classifier	Train Accuracy	Valid Accuracy	Overfit Gap	CV Mean	CV Std	Time (mins)
5	CatBoost	0.836785	0.807361	0.029424	0.815071	0.007517	18.72
4	LGBM	0.854328	0.797585	0.056744	0.812914	0.009180	0.70
2	SVC	0.825999	0.791834	0.034165	0.802561	0.005583	16.77
3	RandomForest	0.877912	0.790684	0.087228	0.808455	0.007572	2.43
0	LogisticRegression	0.780414	0.770558	0.009856	0.779697	0.006725	0.12
1	KNN	0.827006	0.745831	0.081175	0.776100	0.008045	0.13
6	NaiveBayes	0.737705	0.719954	0.017751	0.737849	0.012690	0.01

Figure 1: Model comparison showing CategoricalBoost as the top performer with 80.7 validation accuracy.

Data Overview

Dataset Stats

- Passengers:
  - 8,693 passengers.
- Features:
  - 14 raw columns → 23 columns after feature engineering.
- Notable Features:
  - Age, CryoSleep, HomePlanet, Cabin, Destination, Name.
- Balanced label:
  - (Transported: 50.2% True, 49.8% False) No need for undersampling/oversampling.

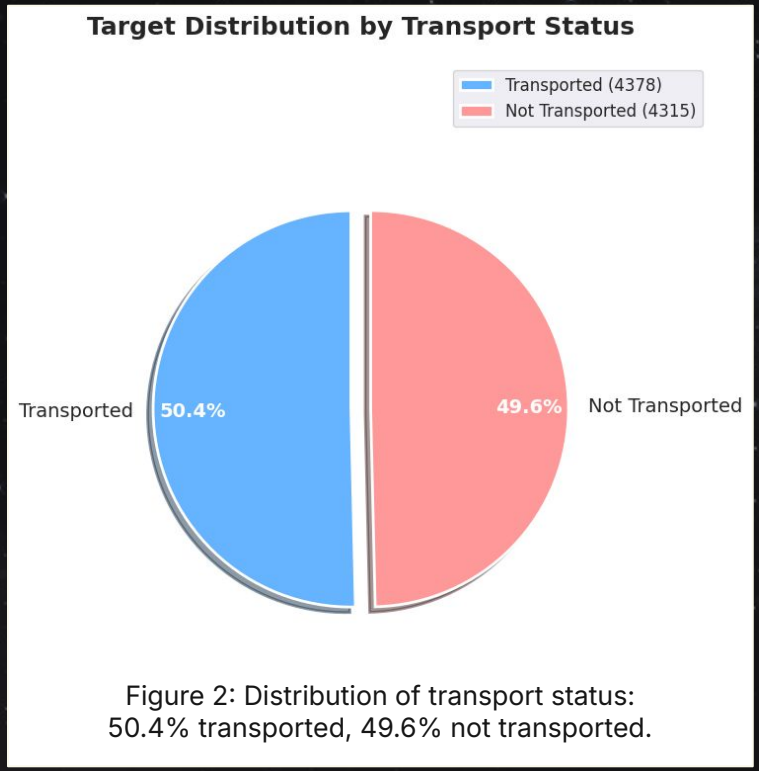

Figure 2: Distribution of transport status: 50.4% transported, 49.6% not transported.

Data Overview (cont.)

Dataset Challenges

- Nearly every single feature had 2% of its data missing.

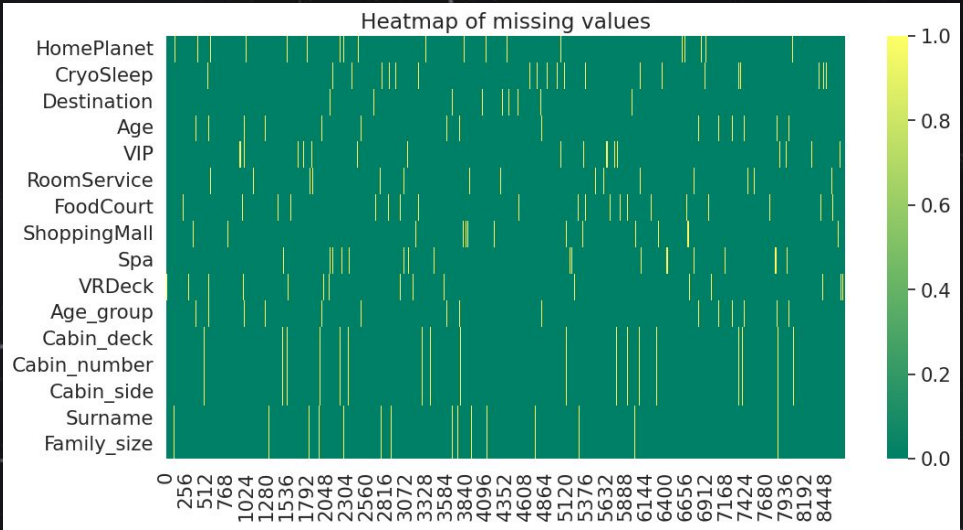

Figure 3: Heatmap of missing values.

	Number_missing	Percentage_missing
HomePlanet	288	2.22
CryoSleep	310	2.39
Destination	274	2.11
Age	270	2.08
VIP	296	2.28
RoomService	263	2.03
FoodCourt	289	2.23
ShoppingMall	306	2.36
Spa	284	2.19
VRDeck	268	2.07
Age_group	270	2.08
Cabin_deck	299	2.31
Cabin_number	299	2.31
Cabin_side	299	2.31
Surname	294	2.27
Family_size	294	2.27

Figure 4: Summary of missing values.

Data Exploration

CryoSleep Matters:

Passengers undergoing Cryosleep were more likely to be transported.

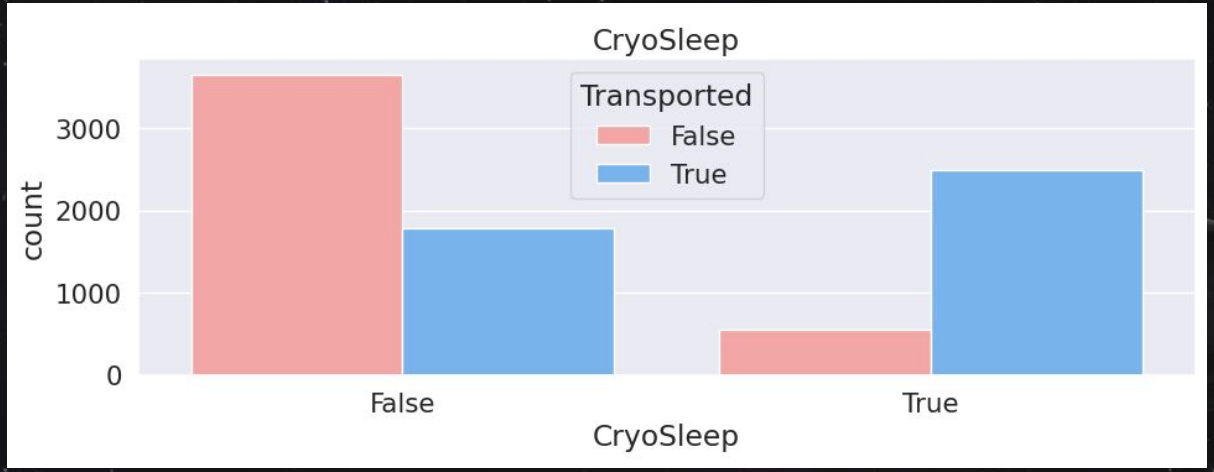

Figure 5: Analysis of transport status of passengers by cryosleep.

Data Exploration (Cont.)

Solo Travellers at Risk:

Solo passengers were less likely to be transported than passengers with a group.

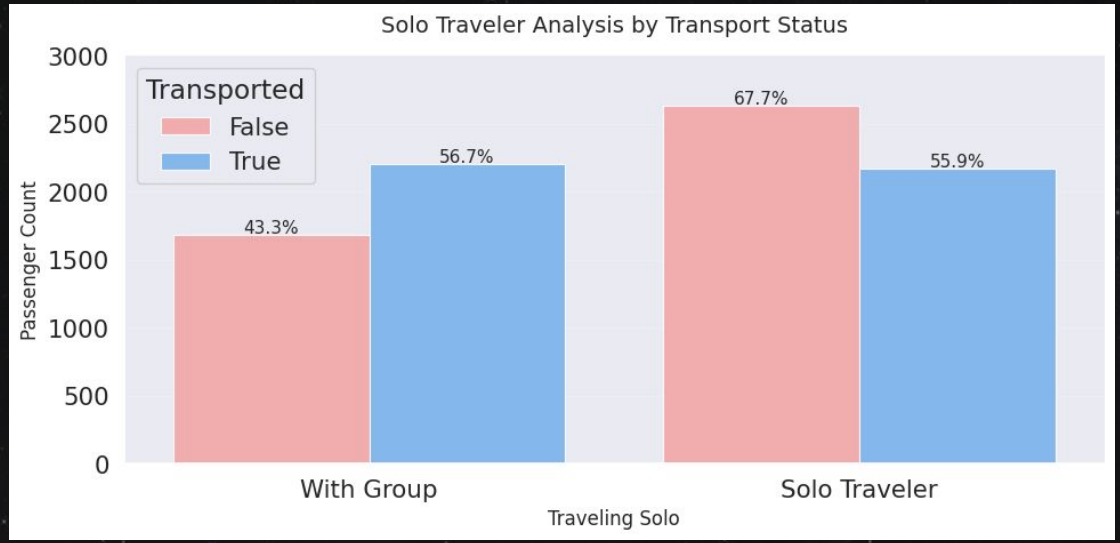

Figure 6: Analysis of transport status of passengers by solo traveler analysis.

Data Exploration (Cont.)

Cabin B & C Hotspot:

Deck B and C passengers were more likely to be transported than any other cabin deck.

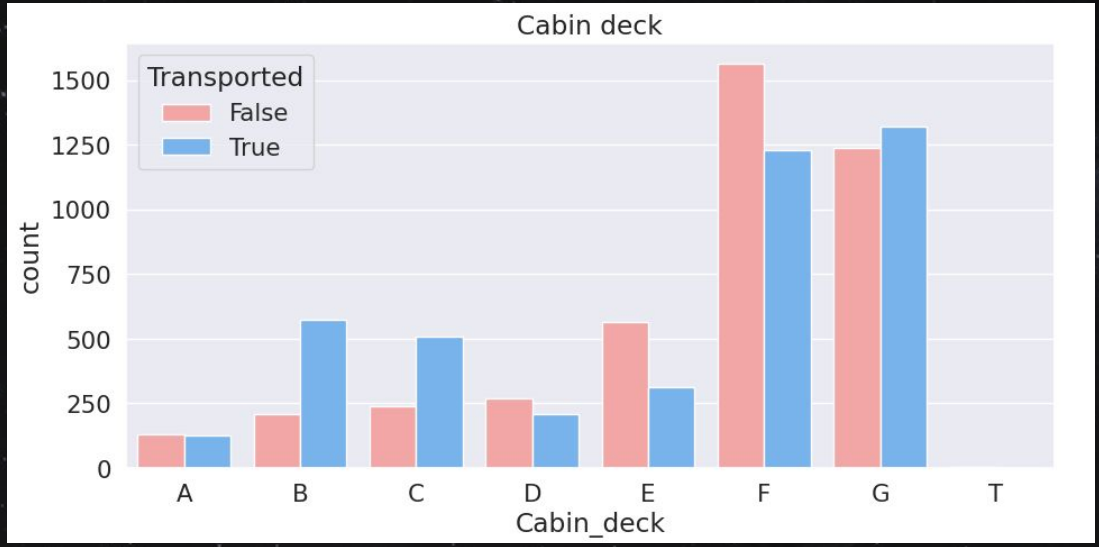

Figure 7: Distribution of transport status of passengers by cabin deck.

Data Exploration (Cont.)

Age Distribution:

Passengers under the age of 18 were more likely to be transported than passengers over the age of 18.

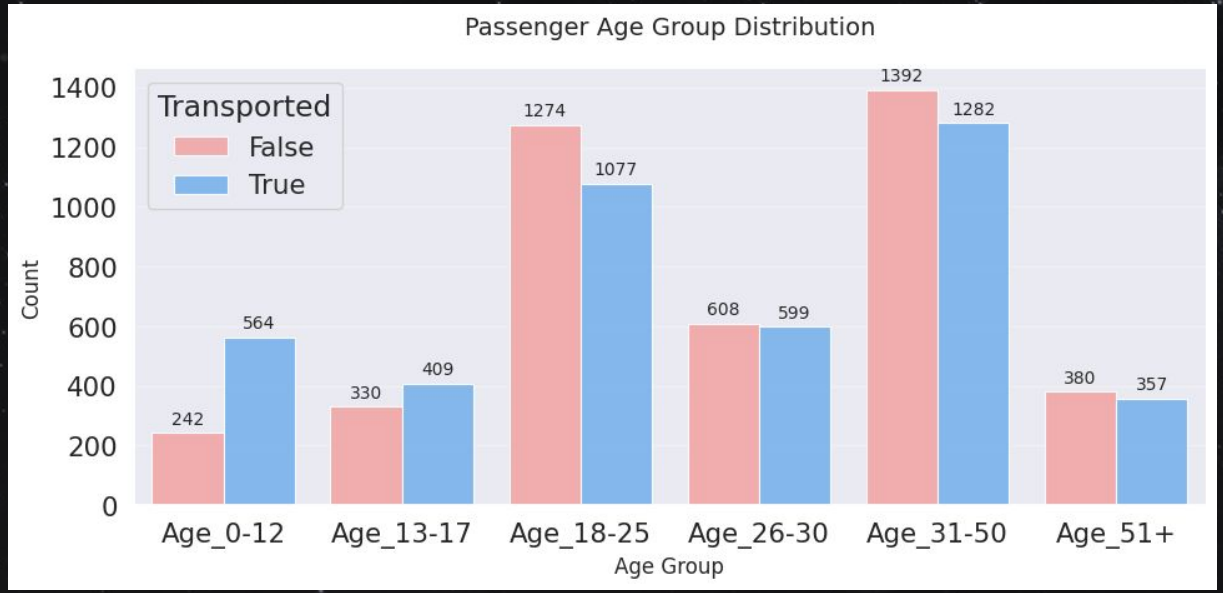

Figure 8: Distribution of transport status of passengers by passenger age.

Feature Engineering

Key Features Created

- Cabin Location Intelligence:
  - Raw Cabin Column (e.g., F,1,S) contained 3 hidden features (Cabin deck, cabin number, cabin side)

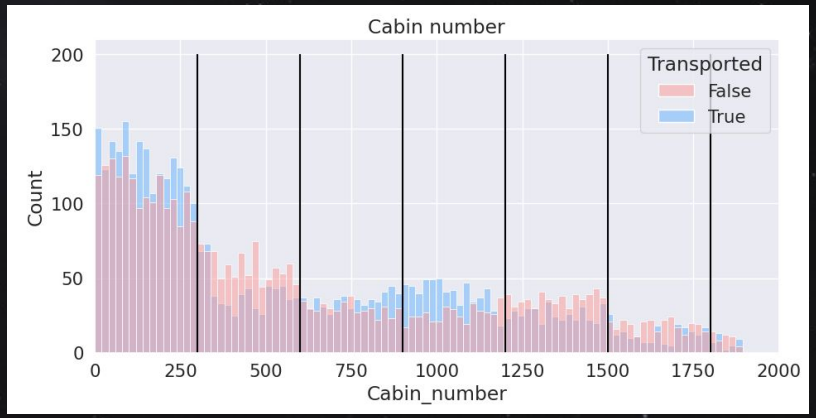

Figure 9: Distribution of transport status of passengers by cabin number.

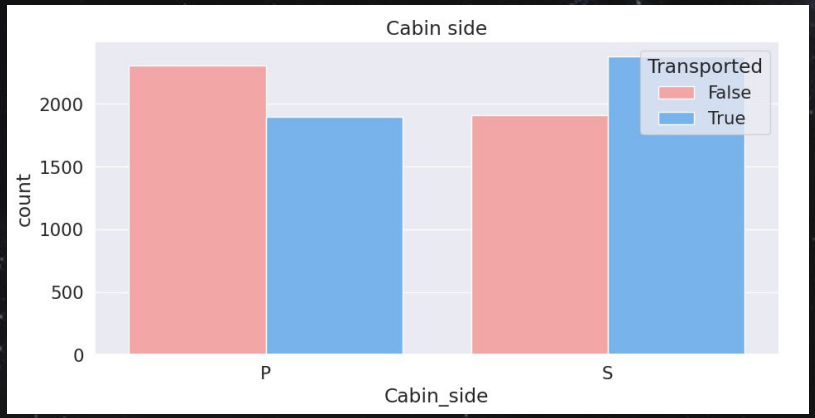

Figure 10: Analysis of transport status of passengers by cabin side.

Feature Engineering (cont.)

Key Features Created

- Group and Family Features:
  - Extracted Group_size from PassengerId (gggg_pp format)
  - Derived Family_size from shared surnames

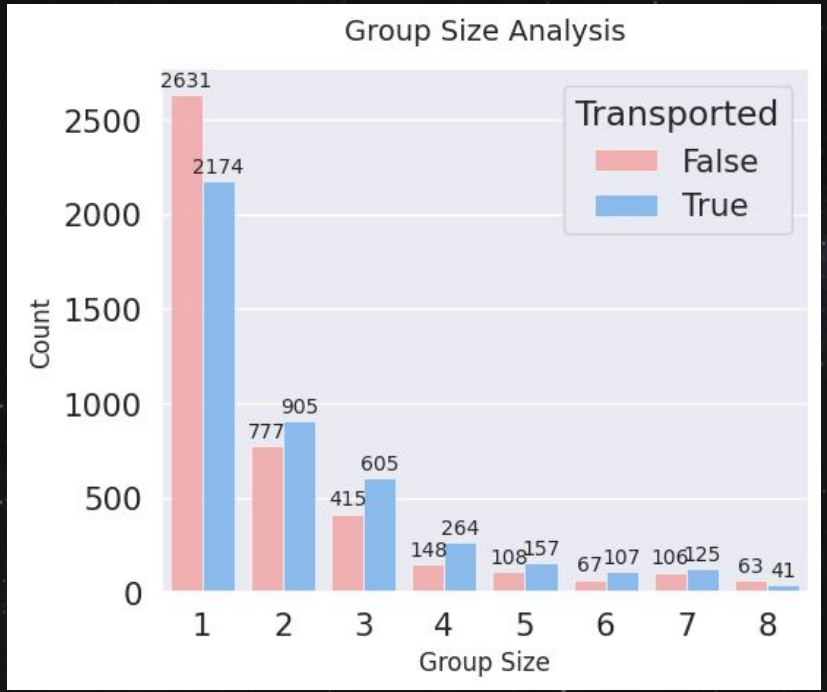

Figure 11: Distribution of transport status of passengers by group size.

Feature Engineering (cont.)

Key Features Created

- Spending Features:
  - Created a new expenditure feature by totalling all luxury amenities spent per passenger in the dataset.
  - Identified passengers with no expenditure.

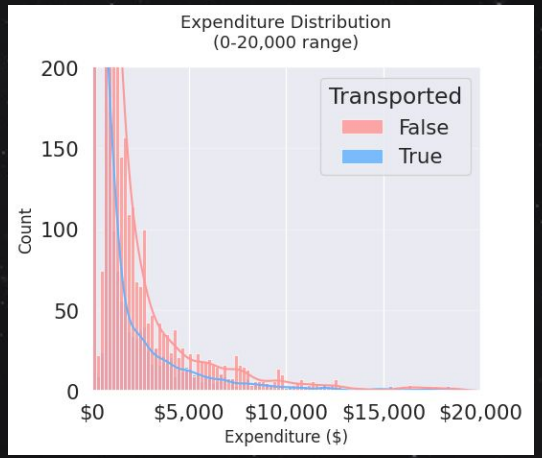

Figure 12: Distribution of transport status of passengers by expenditure.

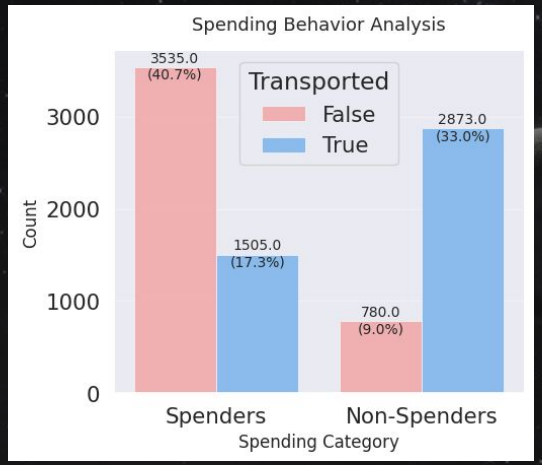

Figure 13: Analysis of transport status of passengers by spending behavior.

Feature Engineering (cont.)

Key Features Created

- Log-Transform for Spending Features:
  - Expenditure features were exponentially distributed and risked models being dominated by outliers.

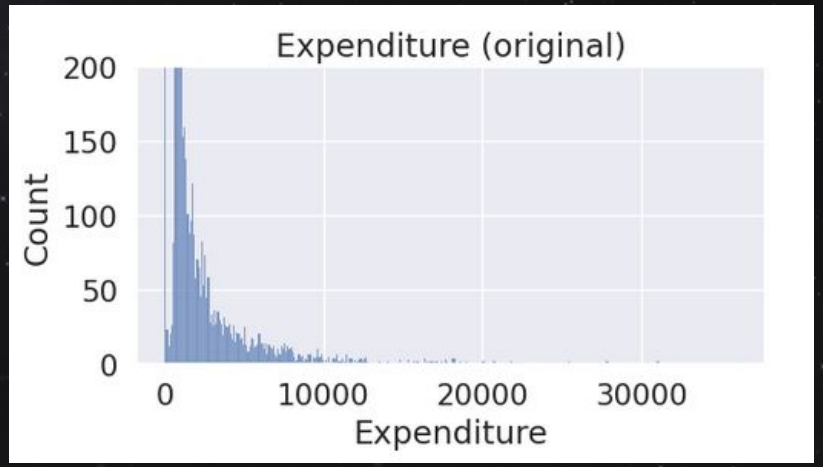

Figure 14: Distribution of transport status of passengers by expenditure before log-transforming.

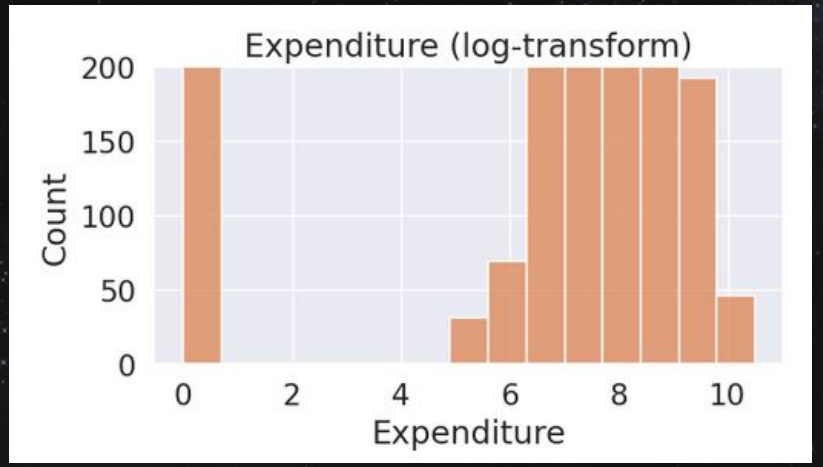

Figure 15: Distribution of transport status of passengers by expenditure after log-transformation.

Missing values

Missing Data

25% of passengers had at least 1 missing value.

Cleaning Strategies

- Group-Based Imputation:
  - Filled missing HomePlanet values based on group (because passengers in the same group would share the same home planet)
- CryoSleep Logic:
  - A person undergoing CryoSleep cannot spend anything, so if expenditure = $0, then the passenger was likely in CryoSleep
- Deck Inference:
  - Used HomePlanet to predict missing Cabin deck (Europa passengers were usually on Deck B)

Missing values

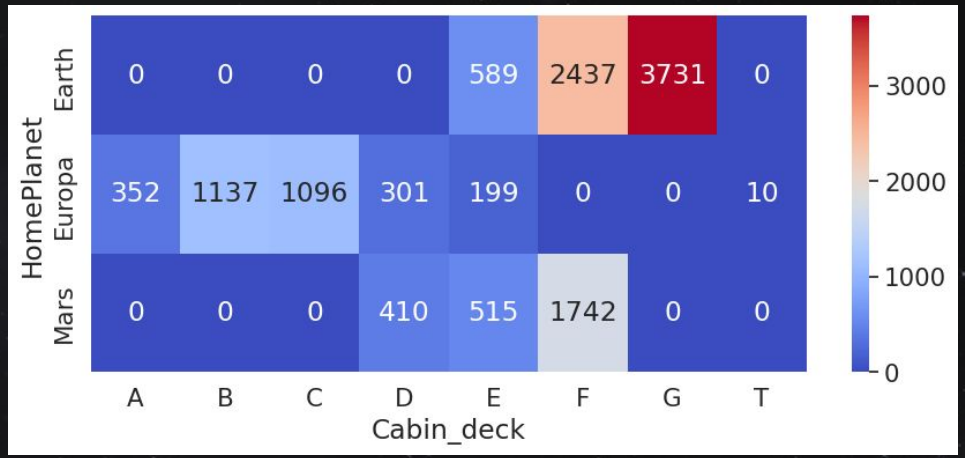

Figure 16: Heatmap of joint distribution of cabin deck and home planet.

HomePlanet	Earth	Europa	Mars
Group			
1	0.0	1.0	0.0
2	1.0	0.0	0.0
3	0.0	2.0	0.0
4	1.0	0.0	0.0
5	1.0	0.0	0.0

Figure 17: Table of joint distribution of home planet and unique groups.

Model Comparison

Models Tested

Logistic Regression, KNN, SVM, Random Forest, LGBM, CatBoost, Naive Bayes

Results

CatBoost: 80.7% accuracy, 0.894 ROC-AUC score.

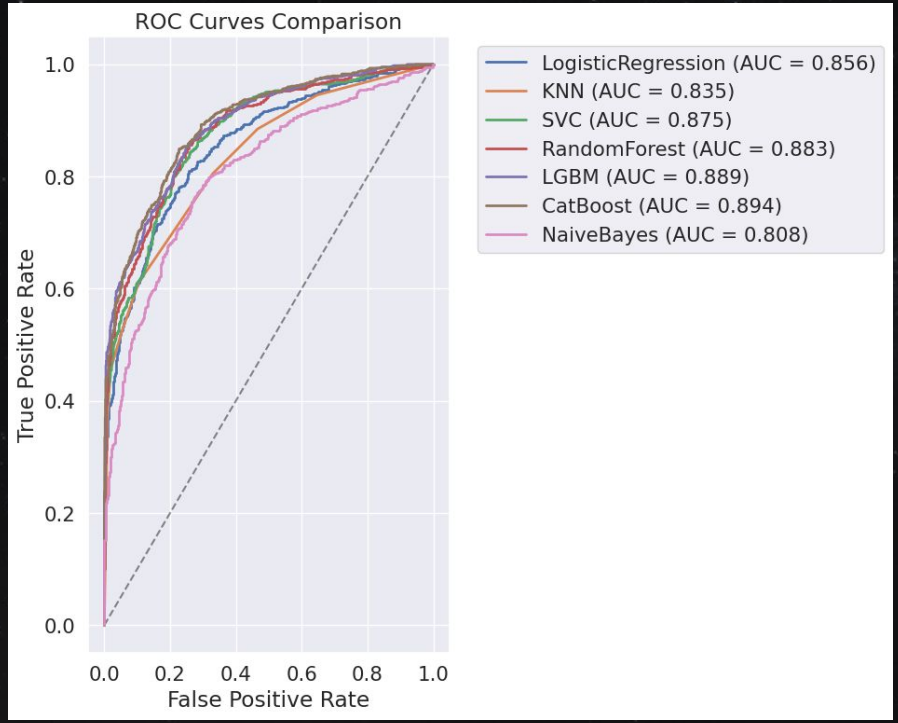

Figure 18: Graph of ROC-AUC score of each tested model.

Model Comparison (cont.)

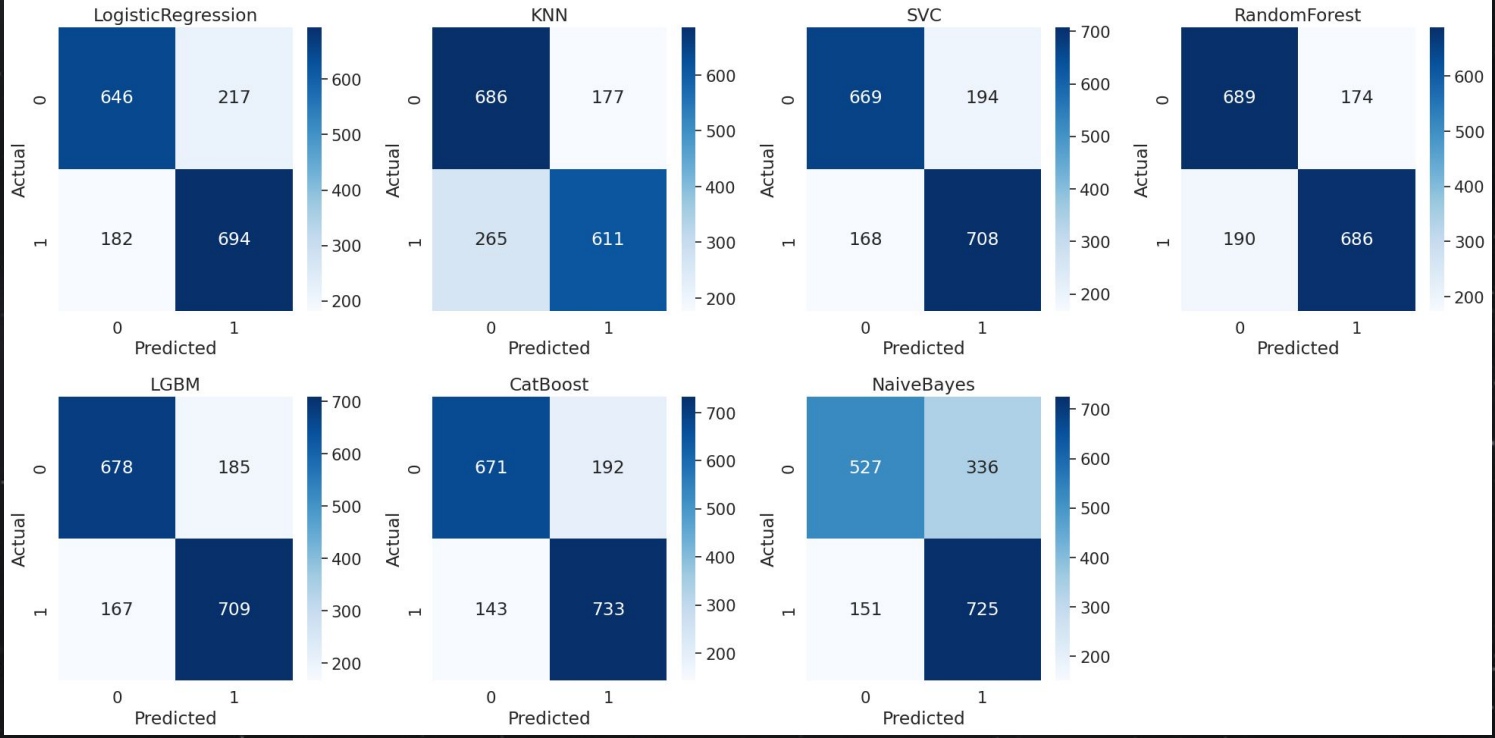

Figure 19: Confusion Matrices of each tested model.

Model Comparison (cont.)

```
=== Detailed Classification Reports ===

LOGISTICREGRESSION CLASSIFICATION REPORT:
              precision  recall  f1-score   support
0                 0.780   0.749     0.764   863.000
1                 0.762   0.792     0.777   876.000
accuracy          0.771   0.771     0.771     0.771
macro avg         0.771   0.770     0.770  1739.000
weighted avg      0.771   0.771     0.770  1739.000

KNN CLASSIFICATION REPORT:
              precision  recall  f1-score   support
0                 0.721   0.795     0.756   863.000
1                 0.775   0.697     0.734   876.000
accuracy          0.746   0.746     0.746     0.746
macro avg         0.748   0.746     0.745  1739.000
weighted avg      0.749   0.746     0.745  1739.000

SVC CLASSIFICATION REPORT:
              precision  recall  f1-score   support
0                 0.799   0.775     0.787   863.000
1                 0.785   0.808     0.796   876.000
accuracy          0.792   0.792     0.792     0.792
macro avg         0.792   0.792     0.792  1739.000
weighted avg      0.792   0.792     0.792  1739.000
```

```
RANDOMFOREST CLASSIFICATION REPORT:
              precision  recall  f1-score   support
0                 0.784   0.798     0.791   863.000
1                 0.798   0.783     0.790   876.000
accuracy          0.791   0.791     0.791     0.791
macro avg         0.791   0.791     0.791  1739.000
weighted avg      0.791   0.791     0.791  1739.000

LGBM CLASSIFICATION REPORT:
              precision  recall  f1-score   support
0                 0.802   0.786     0.794   863.000
1                 0.793   0.809     0.801   876.000
accuracy          0.798   0.798     0.798     0.798
macro avg         0.798   0.797     0.798  1739.000
weighted avg      0.798   0.798     0.798  1739.000

CATBOOST CLASSIFICATION REPORT:
              precision  recall  f1-score   support
0                 0.824   0.778     0.800   863.000
1                 0.792   0.837     0.814   876.000
accuracy          0.807   0.807     0.807     0.807
macro avg         0.808   0.807     0.807  1739.000
weighted avg      0.808   0.807     0.807  1739.000

NAIVEBAYES CLASSIFICATION REPORT:
              precision  recall  f1-score  support
0                 0.777   0.611     0.684   863.00
1                 0.683   0.828     0.749   876.00
accuracy          0.720   0.720     0.720     0.72
macro avg         0.730   0.719     0.716  1739.00
weighted avg      0.730   0.720     0.717  1739.00
```

Figure 20: Classification report of each tested model.

Best Model Insights

Model Selection Rationale

- Classification Report:
  - Highest performing model with nearly the best metrics across all models tested.
- Resistant to Overfitting:
  - Built in regularization (vs. Random Forest tendency to overfit data)
  - Only a 2.94% gap between train (83.68%) and validation (80.74%)

Confusion Matrix Insights

- Key Metrics:
  - Precision: When the model predicts transported, its correct 81% of the time.
  - Recall: Captured 84% of all actual transported passengers

Model Implications

Actionable Insights:

- Prioritize rescues for families on Deck B & C

Broader Impact:

- Similar models could predict earthquake survival or pandemic triage.
- Draws parallel to the Titanic lifeboat allocation problem.

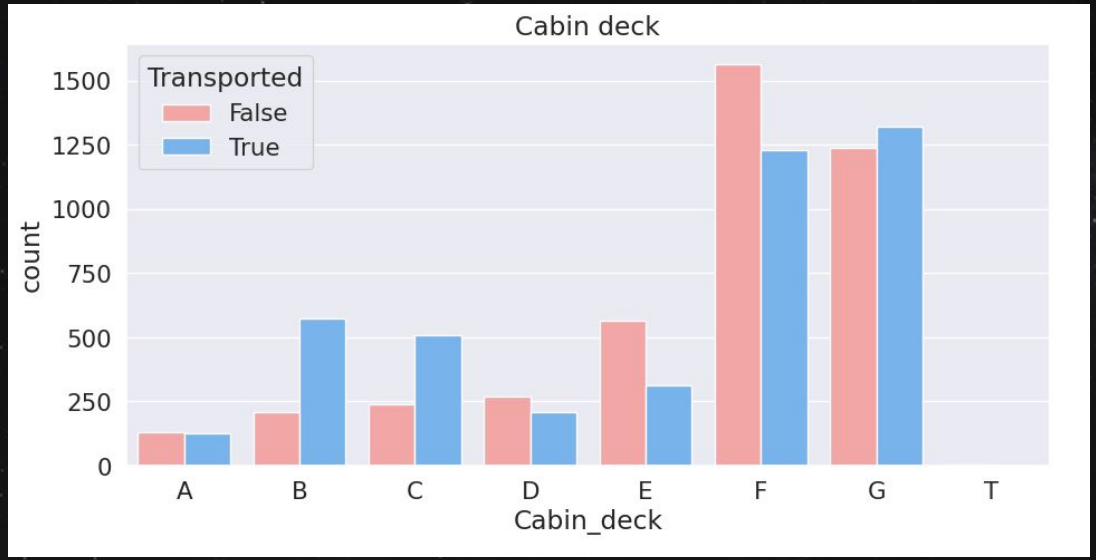

Figure 7: Distribution of transport status of passengers by cabin deck (same as previously show).

Limitations & Future Work

Limitations

- Synthetic Data Constraints:
  - The Kaggle dataset is synthetically created for a competition environment and is purely fictional.
- Feature Gaps:
  - There are no timestamps of the anomaly event or precise cabin coordinates which did not allow us to perform hotspot analysis.

Future Work

- Ethical Considerations:
  - The model currently employs the VIP feature, would this model inadvertently discriminate if VIP was an important feature?

Question & Answer