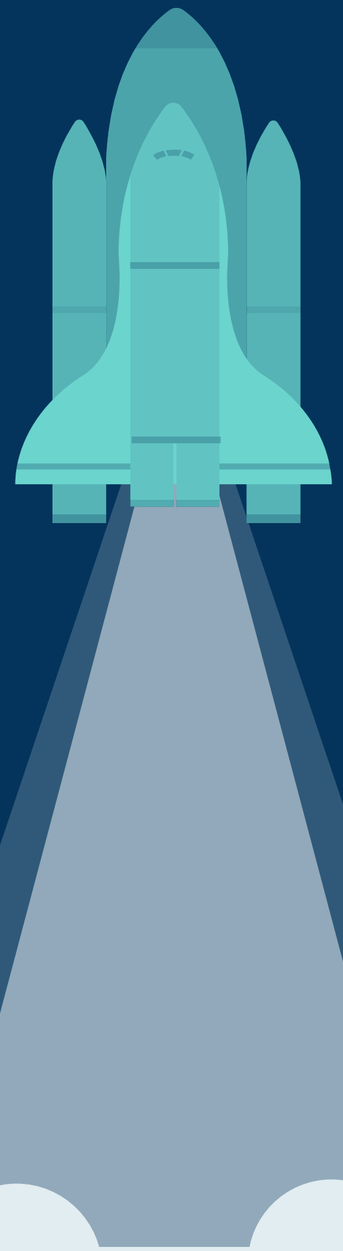
SPACESHIP PROJECT

Michalina Hulak

INTRODUCTION

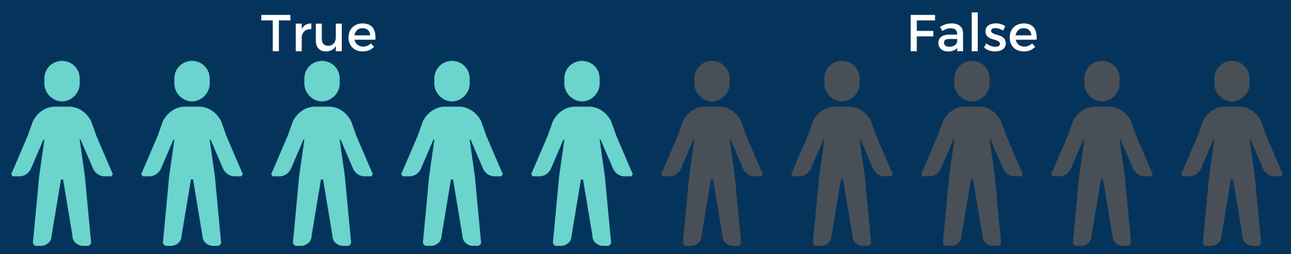
The Spaceship Titanic was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.
While rounding Alpha Centauri en route 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!



The introduction and dataset comes from kaggle.com

GOAL

The aim of the project is to **predict which passengers were transported** by the anomaly using records recovered from the spaceship's damaged computer system.



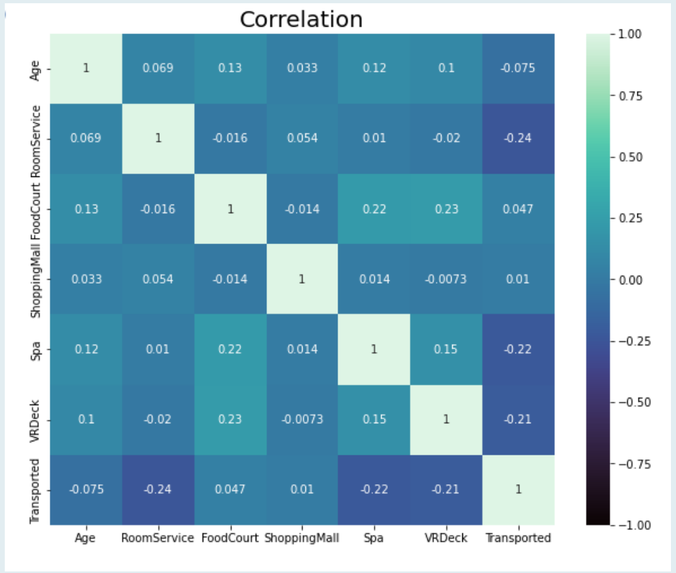
The prediction is based on the data contained in the train.csv file.

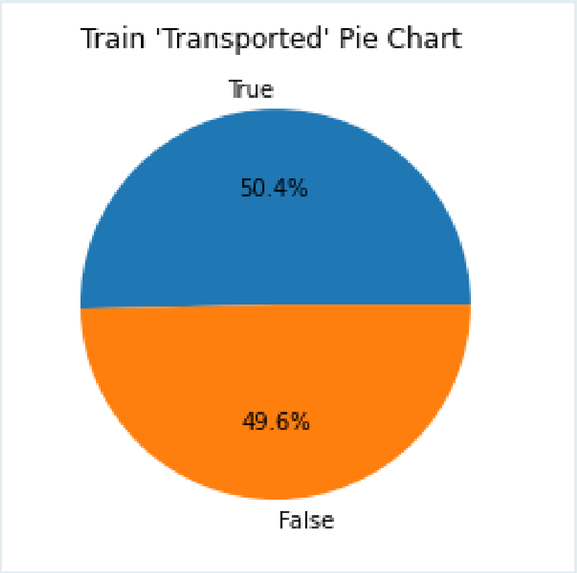
```
[ ]  train_dataset.head(5)
```

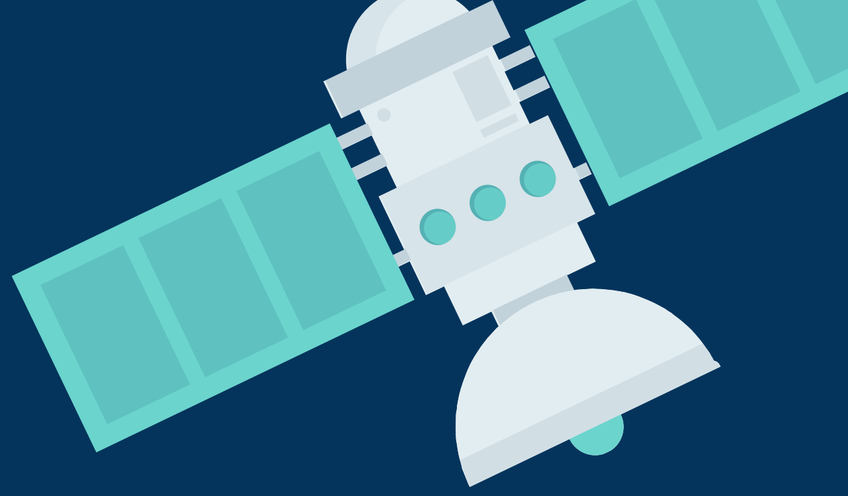
PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True

Research phase of this project

1. Pandas profiler to generate report in html format
2. Visualizations
3. Filling in missing values
4. Replacing categorical values with One-Hot-Encoding and Ordinal Encoding
5. Using 4 models: Random Forest, XGBClassifier, LGBMClassifier, CatBoost, with default parameters
6. Finding the best hyperparameters using Optuna
7. Choosing the best model and prediction for the test_dataset







RESULTS

Best hyperparameters:
'n_estimators': 28,
'max_depth': 17,
'min_samples_split': 8,
'min_samples_leaf': 7,
'max_features': 'sqrt'

Metrics:
Accuracy: 0.78
Precision: 0.78
Recall: 0.78
F-Score: 0.78
Total: 1739
Mislabeled: 377

Best hyperparameters:
'max_depth': 5,
'eta': 0.07231591481728795,
'gamma': 1.058351741651363e-07,
'grow_policy': 'depthwise',
'subsample': 1.0,
'colsample_bytree':
0.30000000000000004,
'min_child_weight': 26,
'n_estimators': 219

Metrics:
Accuracy: 0.79
Precision: 0.79
Recall: 0.79
F-Score: 0.79
Total: 1739
Mislabeled: 363

Best hyperparameters:
'lambda_l1': 3.417199593159779,
'lambda_l2': 0.017340606461167682,
'num_leaves': 36,
'max_depth': 6,
'learning_rate': 0.016682548296829414,
'feature_fraction': 0.5126266483695845,
'bagging_fraction': 0.4866852303282229,
'bagging_freq': 5,
'min_child_samples': 95,
'n_estimators': 618

Metrics:
Accuracy: 0.79
Precision: 0.8
Recall: 0.79
F-Score: 0.79
Total: 1739
Mislabeled: 359

Best hyperparameters:
iterations=531,
learning_rate=0.04311710835109832,
depth=5,
l2_leaf_reg=0.24703700368322665,
bagging_temperature=1.1165404356275512,
random_strength=0.6939220143617256

Metrics:
Accuracy: 0.8
Precision: 0.8
Recall: 0.8
F-Score: 0.8
Total: 1739
Mislabeled: 346

CONCLUSIONS

1. The CatBoost model has the highest metrics and the lowest number of missed predictions.

2. All models seem to perform relatively well on both the training and validation sets, with accuracies ranging from 0.78 to 0.80, and relatively low numbers of mislabeled instances. However, XGBoost and CatBoost have slightly higher accuracy, precision, recall, and F1 scores than the other models, suggesting that they may be better at generalizing to new data.

3. However, we also want to avoid overfitting, which occurs when the model performs well on the training set but poorly on the validation set, indicating that it has learned to memorize the training data rather than generalizing to new data. It seems that CatBoost model will predict the answers best.