KAGGLE SPACESHIP TITANIC CHALLENGE

TEAM MEMBERS:

SHAHZAIB SAEED (22027918)
HAFIZ MALIK FARHAN (22037101)

AHMAD RAZA (22028088)

MUHAMMAD USAMA BADAR (22023675)

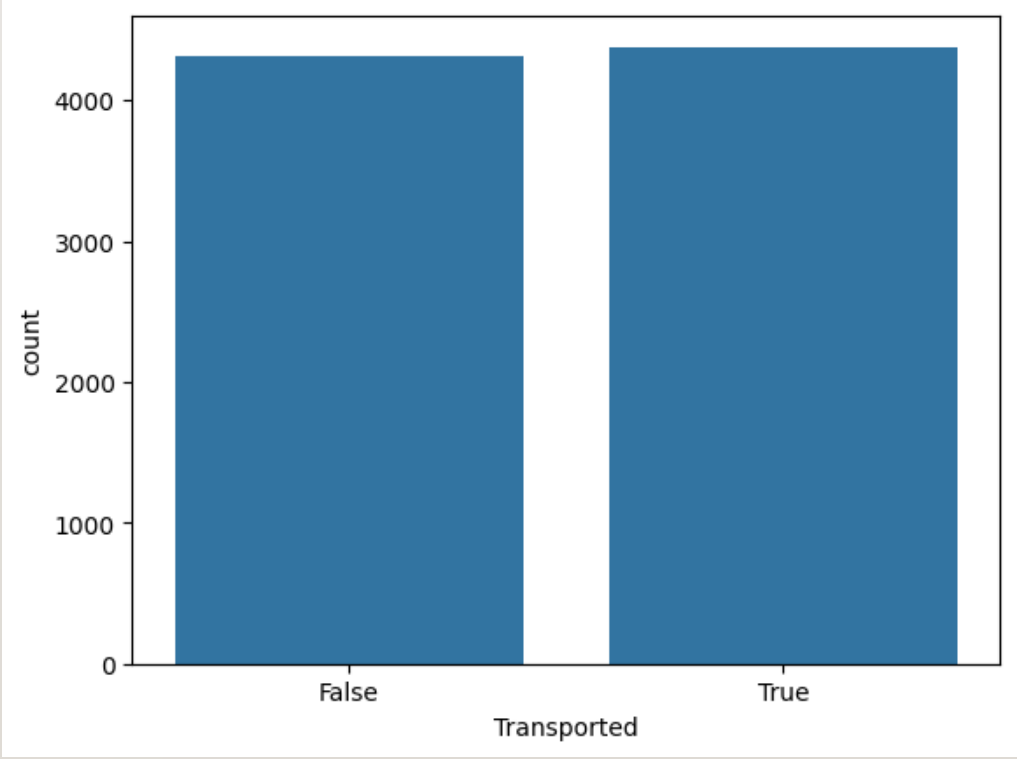
NOMAN AMIN (22017418)

GOOGLE COLLAB LINK:
HTTPS://COLAB.RESEARCH.GOOGLE.COM/DRIVE/1AHRS0MZAUI_TZJL1UOXJKCILEP4VBJCA?USP=SHARING

GITHUB LINK: HTTPS://GITHUB.COM/7PAM2015-0509-2023-GROUP12/KAGGLE_CHALLENGE

INTRODUCTION

- **Objective**:
  To predict whether a passenger was transported to another dimension using the dataset provided by Kaggle.
- **Dataset**:
  The dataset includes features such as demographics, expenses, and travel details of passengers.



EXPLORATORY DATA ANALYSIS (EDA)

- **Missing Values**:
  Visualized missing values to understand the extent of missing data.
- **Distribution of Target Variable**:
  Visualized the distribution of the Transported variable using a count plot.

```
PassengerId       0
HomePlanet      201
CryoSleep       217
Cabin           199
Destination     182
Age             179
VIP             203
RoomService     181
FoodCourt       183
ShoppingMall    208
Spa             183
VRDeck          188
Name            200
Transported       0
```

DATA PRE-PROCESSING

- **Dropped Columns**:
  Removed irrelevant columns: PassengerId, Name, and Cabin.
- **Feature Engineering**:
  1. **Categorical Features**: HomePlanet, CryoSleep, Destination, VIP
  2. **Numerical Features**: Age, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck
- **Imputation**:
  I. Used median strategy for numerical features.
  II. Used most_frequent strategy for categorical features.

ENCODING AND SCALING

- **Label Encoding**:
  Used LabelEncoder for categorical features to convert them to numerical format.
- **Feature Scaling**:
  Used StandardScaler for numerical features to standardize the data.

MODEL TRAINING AND EVALUATION

- **Models Used**:
  - RandomForestClassifier
  - XGBClassifier
  - LogisticRegression
- **Evaluation Strategy**:
  - Split data into training and validation sets (80-20 split).
  - Used GridSearchCV for hyperparameter tuning.
  - Evaluated models based on validation accuracy.

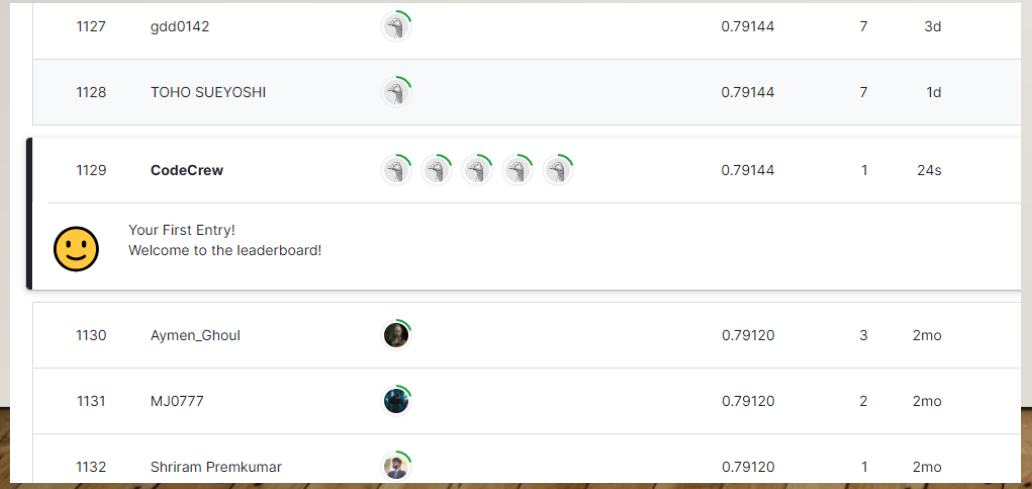
MODEL PERFORMANCE

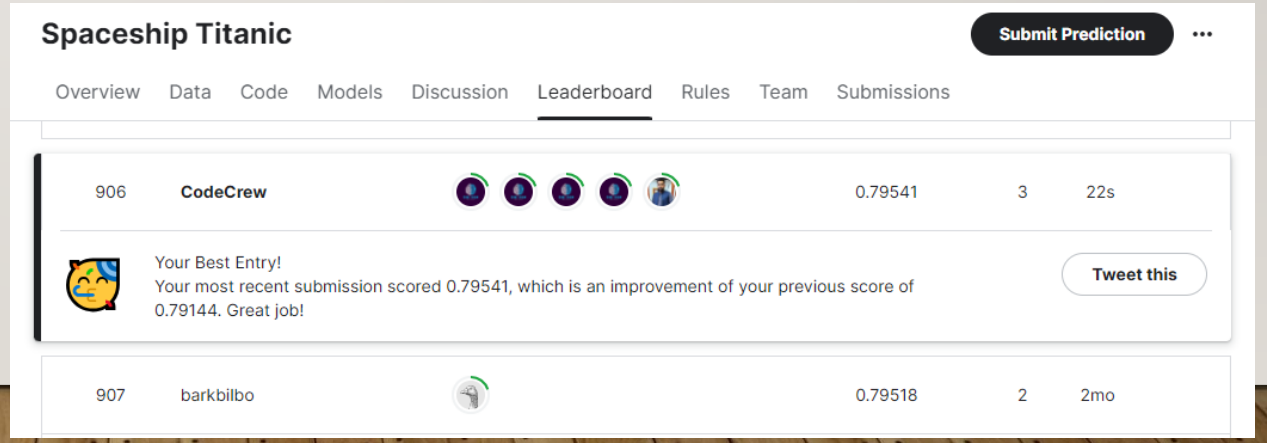
- **Best Model**: XGBoost Classifier
- **Validation Accuracy**: 78.61%
- **Classification Report**:

```
Validation Accuracy: 0.7860839562967222
              precision    recall  f1-score   support

       False       0.82      0.73      0.77       861
        True       0.76      0.84      0.80       878

    accuracy                           0.79      1739
   macro avg       0.79      0.79      0.79      1739
weighted avg       0.79      0.79      0.79      1739
```

KAGGLE SUBMISSION

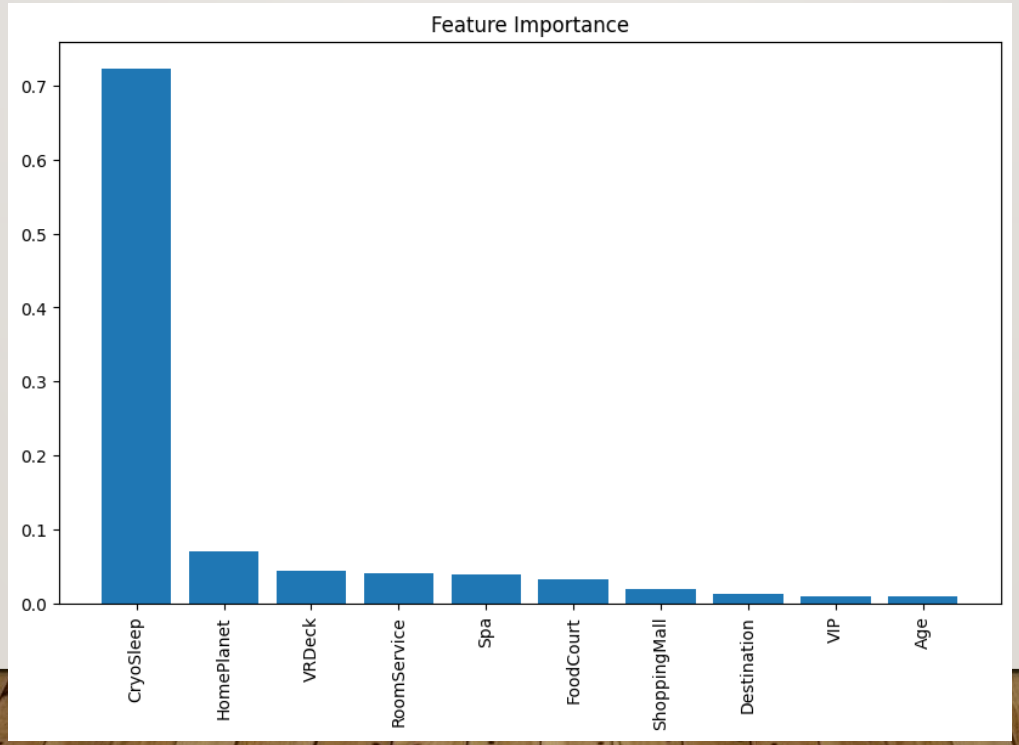
- **Submission Process**:
  - Predicted on the test set.
  - Converted predictions from 1/0 to True/False.
  - Prepared the submission file and uploaded it to Kaggle.
- **Kaggle Score**: First try: 0.79144, Second try: 0.79144, Third try: 0.79541





FEATURE IMPORTANCE

- **Feature Importance**:Visualized feature importance for the best model using a bar chart.



MISCLASSIFICATION ANALYSIS

- **Misclassified Instances**: Analyzed misclassified instances to understand where the model could improve.

Misclassified instances analysis

	Age	RoomService	FoodCourt	ShoppingMall	Spa	\
count	359.000000	354.000000	359.000000	353.000000	357.000000	
mean	27.727019	143.988701	365.181058	198.407932	107.207283	
std	14.223009	399.212227	1123.299304	418.373274	367.308793	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	18.000000	0.000000	0.000000	0.000000	0.000000	
50%	26.000000	0.000000	0.000000	0.000000	0.000000	
75%	38.000000	46.000000	76.500000	33.000000	10.000000	
max	70.000000	2997.000000	10153.000000	2473.000000	4103.000000	

	VRDeck
count	356.000000
mean	155.441011
std	478.200240
min	0.000000
25%	0.000000
50%	0.000000
75%	13.000000
max	5063.000000

MODEL SUCCESSES AND FAILURES

- **Successes**:
  - Achieved a validation accuracy of 78.61%.
  - Kaggle score of 0.79541 indicates good model performance.
- **Failures**:
  - Some misclassifications, indicating potential areas for improvement.

COLLABORATIVE EFFORT

- **Collaboration Tools**:
  - Used GitHub for version control and collaboration.
  - Utilized Google Colab for interactive development and sharing.
- **Team Contributions**:
  - Described individual contributions to data preprocessing, model training, and evaluation.

CONCLUSION

Summary:

- In this project, we aimed to predict whether a passenger on the Spaceship Titanic was transported to another dimension using the dataset provided by Kaggle. We performed extensive exploratory data analysis (EDA) to understand the distribution of the data and the presence of missing values. We then preprocessed the data by imputing missing values, encoding categorical variables, and scaling numerical features. We evaluated several machine learning models, including RandomForestClassifier, XGBClassifier, and LogisticRegression, using GridSearchCV for hyperparameter tuning. The best performing model was the RandomForestClassifier, which achieved a validation accuracy of 78.61%. Our final submission to Kaggle yielded a score of 0.79541, indicating that our model performed well on unseen data.

FUTURE WORK

Future Work:

While our model performed well, there are several areas for improvement that could potentially increase its accuracy and robustness:

Advanced Feature Engineering:

Further exploration of feature interactions, polynomial features, and domain-specific feature creation could provide additional predictive power.

Ensembling Methods:

Combining the predictions of multiple models (e.g., stacking, boosting, bagging) could help improve performance by leveraging the strengths of different algorithms.

THANK YOU