INFO6145 Data Science and Machine Learning

Predicting Passenger Transport Status on the Spaceship Titanic

Wassim Mecheri

28/11/2024

Introduction

- **Objective**: Predict whether passengers aboard the "Spaceship Titanic" were transported to an alternate dimension.
- **Key Challenges**:
  - Handling complex data with both numerical and categorical features.
  - Optimizing models for classification performance.

Dataset Overview

- **Data Source**: Kaggle's Spaceship Titanic competition.
- **Features**:
  - Passenger details: Age, CryoSleep, VIP, PassengerGroup, etc.
  - Service usage: RoomService, FoodCourt, ShoppingMall, Spa, VRDeck.
  - Target: Transported (Yes/No).
- **Missing Data**: The dataset had missing values for almost all features.

Data Preprocessing

- **Feature Engineering**:
  - Created new features by extracting details (PassengerId, Cabin).
- **Handling Missing Data**:
  - Careful imputation (VIP, CryoSleep, Services, HomePlanet, Deck).
  - Mean imputation (Services, Age).
  - Mode imputation (Side, Destination).
- **Feature types**:
  - Ensured every feature was either a float, int, or boolean.
- **Feature Encoding**:
  - One-hot encoding for categorical variables (e.g., CryoSleep, PassengerGroup).

Feature Relations Graphs

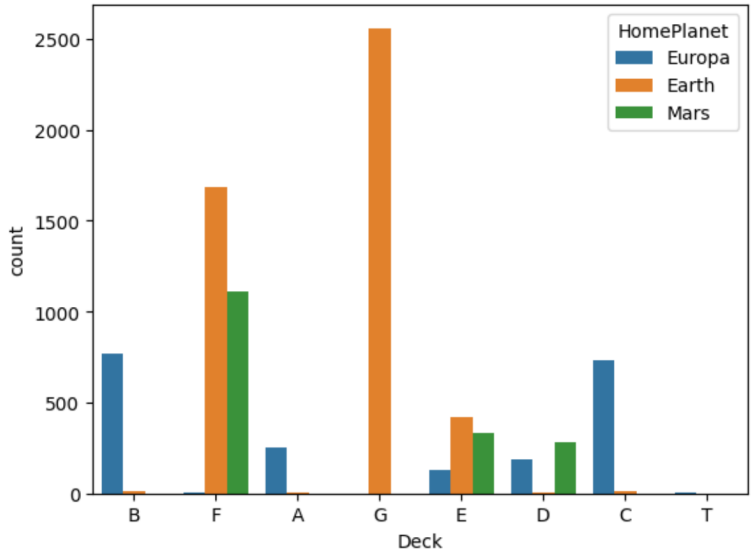

Figure: Relation Deck HomePlanet

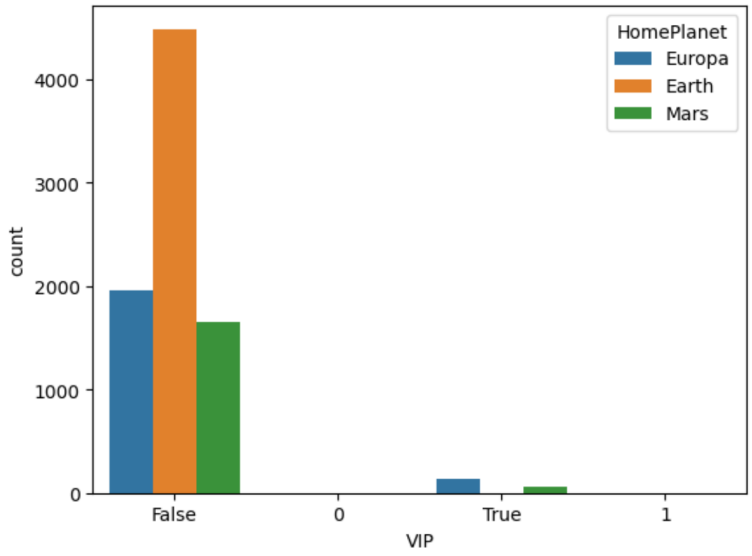

Figure: Relation VIP HomePlanet

Model Selection

- **Models Evaluated**:
  - Logistic Regression
  - MLP Classifier
  - Gradient Boosting Classifier
  - Random Forest Classifier
- **Feature Scaling**:
  - StandardScaler
  - MinMaxScaler
- **Hyperparameter Tuning**: Grid search for optimal hyperparameters (MLP Classifier, Gradient Boosting).
- **Evaluation Metrics**: Accuracy, Precision, Recall, F1 Score, Confusion Matrix.

Model Performance Metrics

Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.7832	0.7950	0.8345	0.7591
MLP Classifier	0.8016	0.8056	0.8162	0.7953
Gradient Boosting	0.8022	0.8076	0.8242	0.7917
Random Forest	0.8079	0.8051	0.7877	0.8234

Table: Model performance using StandardScaler

Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.7579	0.7493	0.7180	0.7833
MLP Classifier	0.7984	0.8033	0.8131	0.7945
Gradient Boosting	0.7990	0.8048	0.8194	0.7886
Random Forest	0.8046	0.8035	0.7862	0.8176

Table: Model performance using MinMaxScaler

Model Performance Confusion Matrix

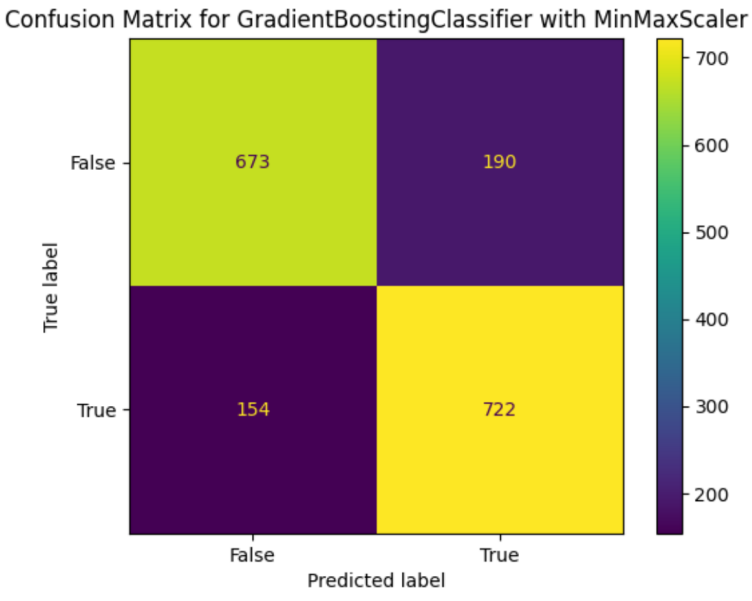

Figure: Gradient Boosting

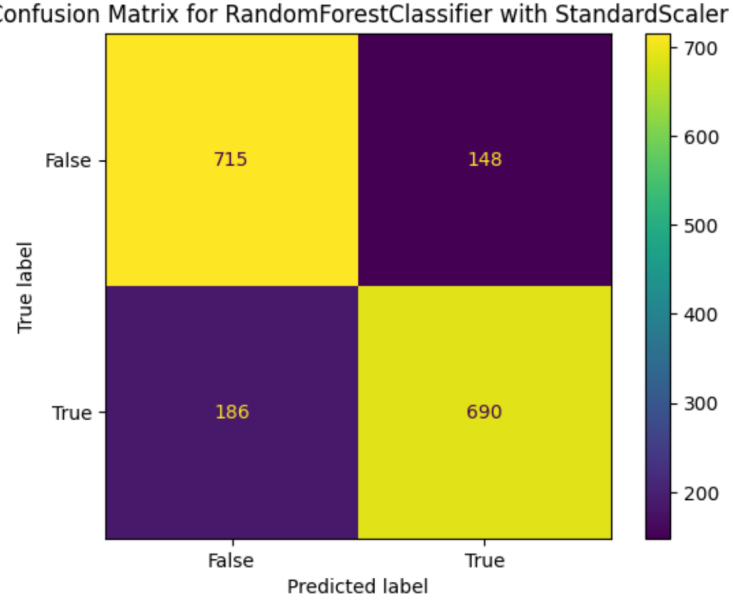

Figure: Random Forest

Model Comparison & Final Results

- **Best Performing Model**: Random Forest with 80.79% accuracy and 82.34% precision.
- **Best Overall Recall**: Gradient Boosting with 82.42%.
- **MLP Classifier**: A bit disappointing, lacked hyperparameter tuning.

Feature Importance Insights

- **Top Features**:
  - RoomService, FoodCourt, Spa, VRDeck, and ShoppingMall.
  - CryoSleep, PassengerGroup.
- **Differences**:
  - Age ranked higher by Random Forest.
  - Deck G ranked higher by Gradient Boosting.

Feature Importance Graphs

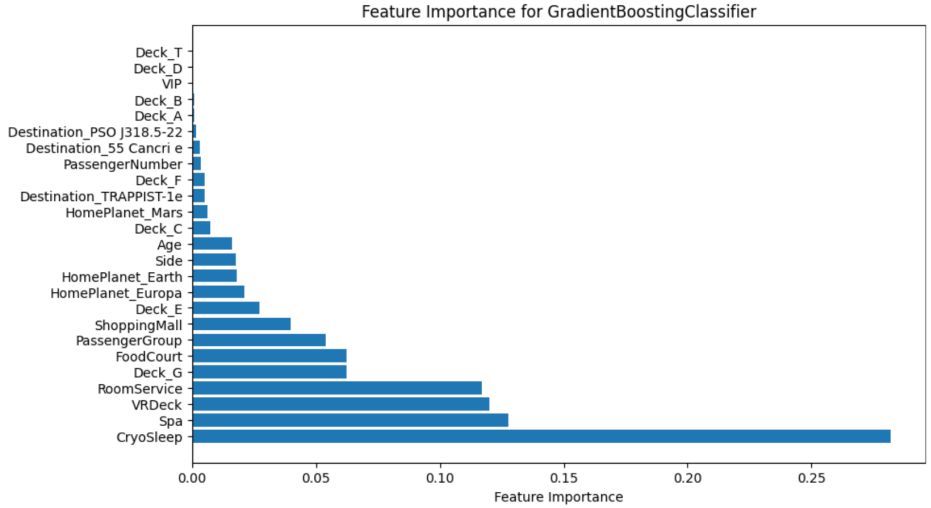

Figure: Gradient Boosting

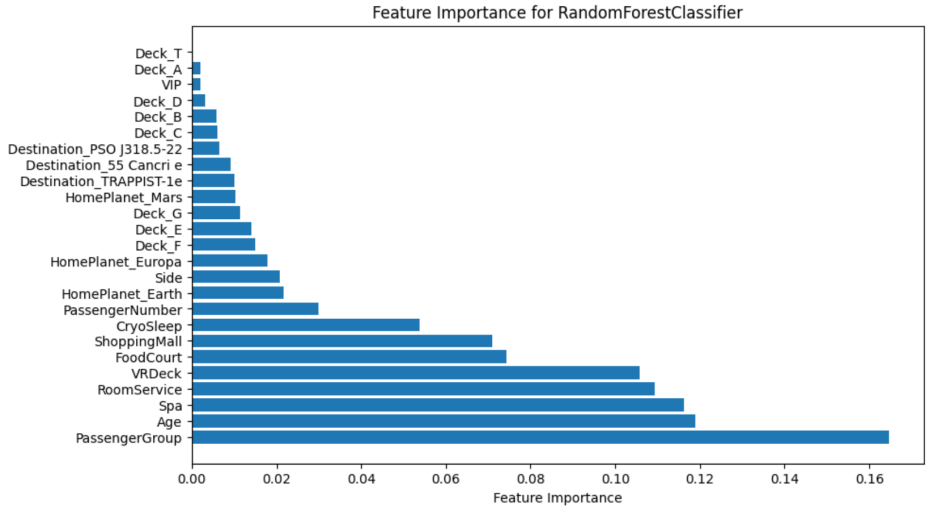

Figure: Random Forest

Conclusion

- **Summary**:
  - Random Forest outperformed other models, achieving higher accuracy and precision.
  - Gradient Boosting was very close, thanks to hyperparameter tuning.
  - MLP Classifier results were disappointing.
- **Future Work**:
  - Better preprocessing.
  - Fine-tuning model hyperparameters.
  - Bigger dataset.
  - Exploring other ensemble methods.
  - Aim for higher accuracy in binary classification (the best was 96%).

References

- Kaggle, *Spaceship Titanic Competition*. https://www.kaggle.com/competitions/spaceship-titanic/overview.
- GeeksforGeeks, *Spaceship Titanic Guide*. https://www.geeksforgeeks.org/spaceship-titanic-project-using-machine-learning-python/.
- Kaggle, *Titanic - Machine Learning from Disaster*. https://www.kaggle.com/c/titanic/overview.
- Kaggle, *First Kaggle Attempt - Feedback Welcome!*. https://www.kaggle.com/code/liamstuart3141/first-kaggle-attempt-feedback-welcome.
- Kaggle, *Spaceship Titanic: Logistic Regression Model Train*. https://www.kaggle.com/code/aslemimolu/spaceship-titanic-logistic-regression-model-train.
- Kaggle, *Spaceship Titanic Using Random Forest and Xgboost*. https://www.kaggle.com/code/hardikgarg03/spaceship-titanic-using-random-forest-and-xgboost.