Kaggle Data Science Challenge: Spaceship Titanic Dataset
Predicting Outcomes of passengers with Machine Learning Models

MSc Data Science with Advanced Research
University of Hertfordshire

Abstract

In this study, we applied Logistic Regression, Random Forest Classifier and XGBoost algorithm for the predictive analysis of the transportation of passengers on a fictional spaceship named Titanic, using the Kaggle Spaceship Titanic dataset. The analysis required data preprocessing, model fitting and prediction on a sample test set. We were successfully able to evaluate the performance of Logistic Regression and Random Forest Classification for binary classification tasks and examine their interpretability using Shapley values. The model achieved 79% accuracy on Kaggle competition. We found XGBoost performed slightly better than the others. However, it is important to note that the quality of data preprocessing plays a significant role in improving the model's performance as was the case in our work.

Keywords: Logistic Regression, Random Forest, XGBoost, Kaggle, Spaceship Titanic, Shapley Values, Model Interpretation

1. Introduction

Binary classification is one of the most prevalent supervised machine learning problems encountered in various domains from fraud detection to disease diagnosis spanning across industries: finance, healthcare, manufacturing and security and more. Over the years, numerous algorithms such as Logistic Regression, Naïve Bayes Classifier, K-Nearest Neighbours, Decision tree, Gradient Boosting and Random Forest Classifiers have been developed to address this challenge. Although each of these methods has their own strengths and weaknesses, there is no one single method for all classification problems and the effectiveness of different methods can vary considerably depending on factors such as the type of data, its quality, and the extent of the dataset being used.

Logistic Regression is a simple and effective method for binary classification problems.[3] Because of its simplicity it is the go-to algorithm for binary classification problems. The goal of the algorithm is to find the decision boundaries between discrete classes in supervised classification problems. The model predicts the probability of a data point to belong in a particular class using a logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as:

$$\text{logistic}(\eta) = \frac{1}{1 + exp(-\eta)}$$

It performs well when the classes are linearly separable and produce interpretable results while being computationally effective. However, this algorithm struggles with datasets containing highly correlated features and sometimes is prone to overfitting.

Random Forest (RF) Classifiers are made up of decision trees utilizing both bagging and feature randomness to create an uncorrelated forest. RF algorithm requires three hyperparameters: node size, the number of trees, and the number of features sampled. Using these RF can predict classes. Random Forests are improvements of decision trees and have less chances of overfitting. However, it is a complex model and requires higher computing power. Also RF does not perform well with noisy, imbalanced dataset.

Gradient Boosting builds a predictive model by combining multiple weak models (typically decision trees) in a sequential manner. Gradient boosting works by training a series of models, where each subsequent model is built to correct the errors made by the previous models. This iterative process continues until the model achieves optimal performance.

Gradient Boosting is known for its ability to handle complex datasets and produce highly accurate predictions in a variety of tasks, including classification and regression. The limitations of this includes: susceptibility to overfitting, sensitivity to noise and outliers and complex interpretability.

The titanic dataset is a widely known example in the field of data analysis and machine learning which has extensively been used to explore and demonstrate various data analysis techniques, feature engineering methods, and predictive modelling algorithms. This dataset is specifically suitable for binary classification analysis. The 'Spaceship Titanic' is a dataset similar to the classic titanic dataset. The challenge here is to predict whether a passenger on the spaceship has been transported to another dimension.

2. Dataset

The Spaceship Titanic is a fictional interstellar passenger liner that collided with a spacetime anomaly while en route to three newly habitable exoplanets. The ship stayed intact, but almost half of the passengers were transported to an alternate dimension. The dataset contains information on the passengers. The transportation status of some of the passengers is unknown. This challenge is to identify which ones were transported based on the available data.

The dataset columns are: PassengerId, HomePlanet, CryoSleep, Cabin, Destination, Age, VIP, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck, Name and Transported containing both numeric and string data. The columns - RoomService, FoodCourt, ShoppingMall, Spa, VRDeck are the amount a passenger was billed in respective facilities. The rest of the column names are self-explanatory.

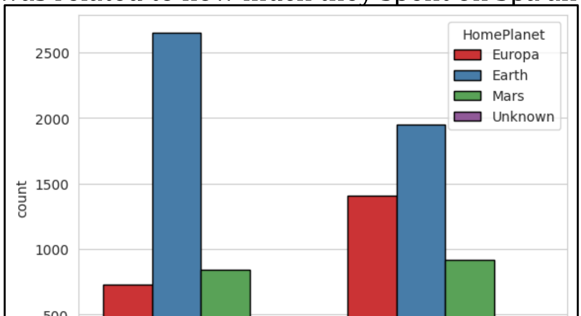
Missing Values						
PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP
0	201	217	199	182	179	203
RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
181	183	208	183	188	200	0

Table 1: Missing data in the dataset.

2.1 Data preprocessing:

Before starting the analysis, it was necessary to address the issue of the missing values. The first step was to replace the empty cells in non-numeric data column to 'Unknown'. Then the missing values on the numeric columns were imputed by their mean values after grouping them by HomePlanet. Following this step was one hot encoding of the string data columns. We also applied fancyimpute but the result was not very different from this method.

To run exploratory data analysis we did tried different plots including histograms, boxplots, distribution and pair plots. The plots showed that the likelihood of a passenger not transported was related to how much they spent on Spa and VRDeck.



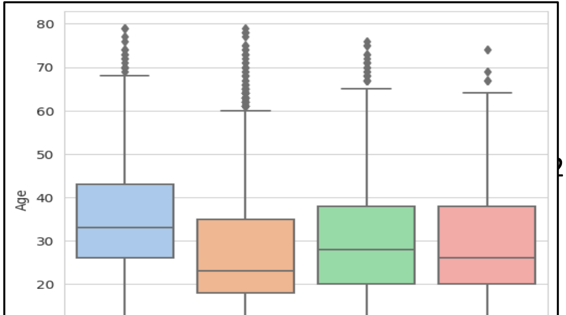


Figure 1: EDA Plots.

3. Models

We imported LogisticRegression and RandomForestClassifier modules from sklearn and XGBClassifier from XGBoost. Both normalised and original dataset were used to investigate model fit. But there were no noticeable differences.

Logistic Regression			
	precision	recall	f1-score
0	0.80	0.77	0.78
1	0.77	0.8	0.79
	Accuracy	0.785	

Random Forest			
	Precision	recall	f1-score
0	0.77	0.80	0.78
1	0.79	0.77	0.78
	Accuracy	0.78	

XGBoost			
	precision	recall	f1-score
0	0.83	0.75	0.79
1	0.77	0.84	0.81
	Accuracy	0.79	

Table 2: Results of the three methods applied.

Different hyperparameters were tested for RF model number of estimator and criterion. The accuracy improvements were too insignificant.

For XGBoost the best parameters were iterated with GridSearchCV.

4. Model interpretation

The Shapley values showed that the amount a passenger had spent on various amenities on the ship had the most impact on the prediction.

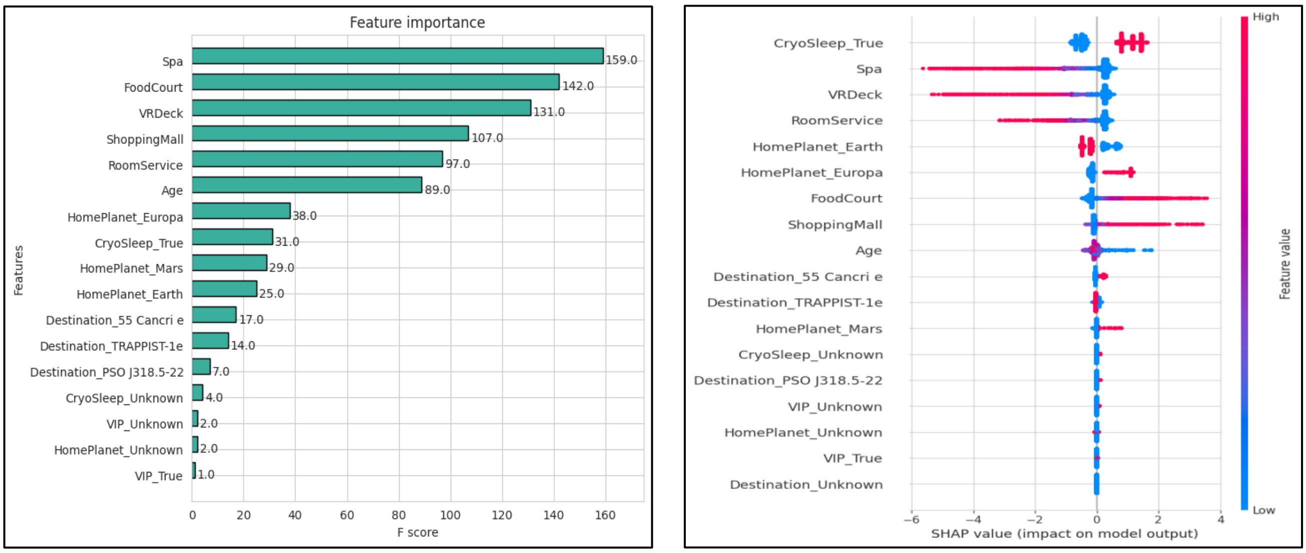


Fig 2: Shapley values.

5. Discussion and Conclusions

Overall, the model accuracy achieved was 80% with XGBoost. The data preprocessing function excluded features like cabin data and also ignored possible missing values in: HomePlanet, Destination and VIP column when missing values were put into ‘Unknown’ category. Some of the features in the dataset can be further explored with better. Future improvement on the dataset could be splitting the passenger ID to identify the group a passenger was travelling with and cabin data can be split to identify more precise location of the passenger on the ship. This could help discover potential underlying pattern which could lead to better prediction.

6. References

Lecture and Tutorial Notes:. (2023). Rafael S. de Souza, Research Methods in Data Science, University of Hertfordshire.

Scapin, D., Cisotto, G., Gindullina, E., & Badia, L. (Year). Shapley Value as an Aid to Biomedical Machine Learning: a Heart Disease Dataset Analysis. Dept. of Information Engineering (DEI), University of Padova, Italy, Dept. of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Italy, Inter-University Consortium for Telecommunications (CNIT), Italy.

Subasi, A. (2020). Practical Machine Learning for Data Analysis Using Python.

scikit-learn contributors. (n.d.). RandomForestClassifier. In scikit-learn: Machine Learning in Python. Retrieved March 31, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Breiman, L. (n.d.). Random Forests. Retrieved March 31, 2023, from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro

XGBoost contributors. (n.d.). XGBoost Model. Retrieved March 31, 2023, from https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Wikipedia