Spaceship Titanic ∗

Transported Passengers prediction

Pranjal Jain  
 Information Technology and Web Science  
 Rensselaer Polytechnic Institute  
 Troy New York USA  
 [pranj.jain95@gmail.com](mailto:pranj.jain95@gmail.com)

Executive Summary

The Spaceship Titanic was an interstellar passenger liner with almost 13,000 passengers on board set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars. The training data set has 14 features including the target variable and 8693 rows and the testing data set has 4277 rows and 13 features. The spaceship titanic training dataset will be used to build predictive model to predict which passengers were transported by the anomaly using records recovered from the spaceship’s damaged computer system. The spaceship titanic dataset comes from a Kaggle competition. The dataset has 6 continuous value columns 'Age', 'RoomService', 'FoodCourt', 'ShoppingMall', 'Spa', 'VRDeck'. Cabin column is a combination of deck/number/side. The columns which are of nominal type are Transported, CryoSleep, cabin, destination, VIP and there are 3 columns with Boolean data type including one ground truth column ‘Transported’, The ‘Transported’ column has 4378 True labels and 4315 False labels. The ‘PassengerId’ column has 8693 unique values in the training dataset and 4277 in testing dataset. ‘HomePlanet’ has three values Earth, Europa, and Mars. The PassangerId column is split into two parts first part has the first 4 characters and the second part has all the characters after underscore. Cabin column is split into three parts, cabin deck, cabin number, and cabin side. The rows which have null or missing values in continuous variables are filled with median values for the training dataset and the testing dataset. The continuous variables in the training data and the testing data are standardized as well. The description of feature columns is as follows.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columns** | **Description** | **Datatype** | **Has missing values** | **Example** |
| PassengerId | A unique Id for each passenger. Each Id takes the form gggg\_pp where gggg indicates a group the passenger is travelling with and pp is their number within the group. People in a group are often family members, but not always. | String | No | 0001\_01 |
| HomePlanet | The planet the passenger departed from, typically their planet of permanent residence. | String | Yes | Earth |
| CryoSleep | Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins. | Boolean | Yes | True/False |
| Cabin | The cabin number where the passenger is staying. Takes the form deck/num/side | String | Yes | F/1/S |
| Destination | The planet the passenger will be debarking to. | String | Yes | TRAPPIST-1e |
| Age | The age of the passenger. | Float | Yes | 39.0 |
| VIP | Whether the passenger has paid for special VIP service during the voyage. | Boolean | Yes | True/False |
| RoomService | Amount the passenger has billed for room service. | Float | Yes | 109.0 |
| FoodCourt | Amount the passenger has billed at the food court. | Float | Yes | 9.0 |
| ShoppingMall | Amount the passenger has billed at the shopping mall. | Float | Yes | 25.0 |
| Spa | Amount the passenger has billed at the spa. | Float | Yes | 549.0 |
| VRDeck | Amount the passenger has billed at the VR deck. | Float | Yes | 44.0 |
| Name | The first and last names of the passenger. | String | Yes | Maham Ofracculy |
| Transported | Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict. | Boolean | No | True/False |

The models used for prediction of the target variable ‘transported’ are Logistic Regression, Random Forrest, and Multi-layer perceptron neural network. The accuracy values for prediction task are calculated by taking the average of 5-fold cross validation. The model accuracy for Logistic Regression on the training data is 0.87 and the average accuracy after performing 5-fold cross validation is 0.79 , The model accuracy for Random Forest on the training data is 0.85 and the average accuracy after performing 5-fold cross validation is 0.79 and, the model accuracy for Multi-layer perceptron on the training data is 0.93. the three models have performed well on the prediction task which can be confirmed by the accuracy values of the models.

Benchmarking of other solutions

Three models from Kaggle’s spaceship titanic submissions are used to benchmark solutions for predicting which passenger was transported to another dimension. The categorical columns namely HomePlanet, CryoSleep, Destination, VIP, CabinDeck, CabinSide are converted to integer values based on class values of each column. Then the data is split into training and validation set. The data is split into 5 parts using stratified cross validation for testing the performance. The performance of the classifier is measured by taking the average of accuracy of each validation set. in preprocessing steps for continuous variables, the missing values are filled with median values and then normalized. For categorical variables the missing values are filled with constant values, then one hot encoding is performed on the categorical variables. The three models used for benchmarking solutions are Logistic regression, Support vector machine and K nearest neighbors. The model accuracy for Logistic Regression on the training data is 0.79, The model accuracy for Support vector machine on the training data is 0.80 and the model accuracy for K nearest neighbors on the training data is 0.78. Support vector machine has the best accuracy on the prediction task.

|  |  |  |  |
| --- | --- | --- | --- |
| **Notebook Name** | **Feature Approach** | **Model Approach** | **Train/Test Performance** |
| Spaceship titanic - EDA + Modeling | 'Age', 'RoomService', 'FoodCourt',  'ShoppingMall',' Spa', 'VRDeck',  'PassengerGroupSize', 'CabinNumber' | Logistic regression | 0.793 |
| Spaceship titanic - EDA + Modeling | 'Age', 'RoomService', 'FoodCourt',  'ShoppingMall',' Spa', 'VRDeck',  'PassengerGroupSize', 'CabinNumber' | SVM classification | 0.805 |
| Spaceship titanic - EDA + Modeling | 'Age', 'RoomService', 'FoodCourt',  'ShoppingMall',' Spa', 'VRDeck',  'PassengerGroupSize', 'CabinNumber' | KNN clustering | 0.78 |

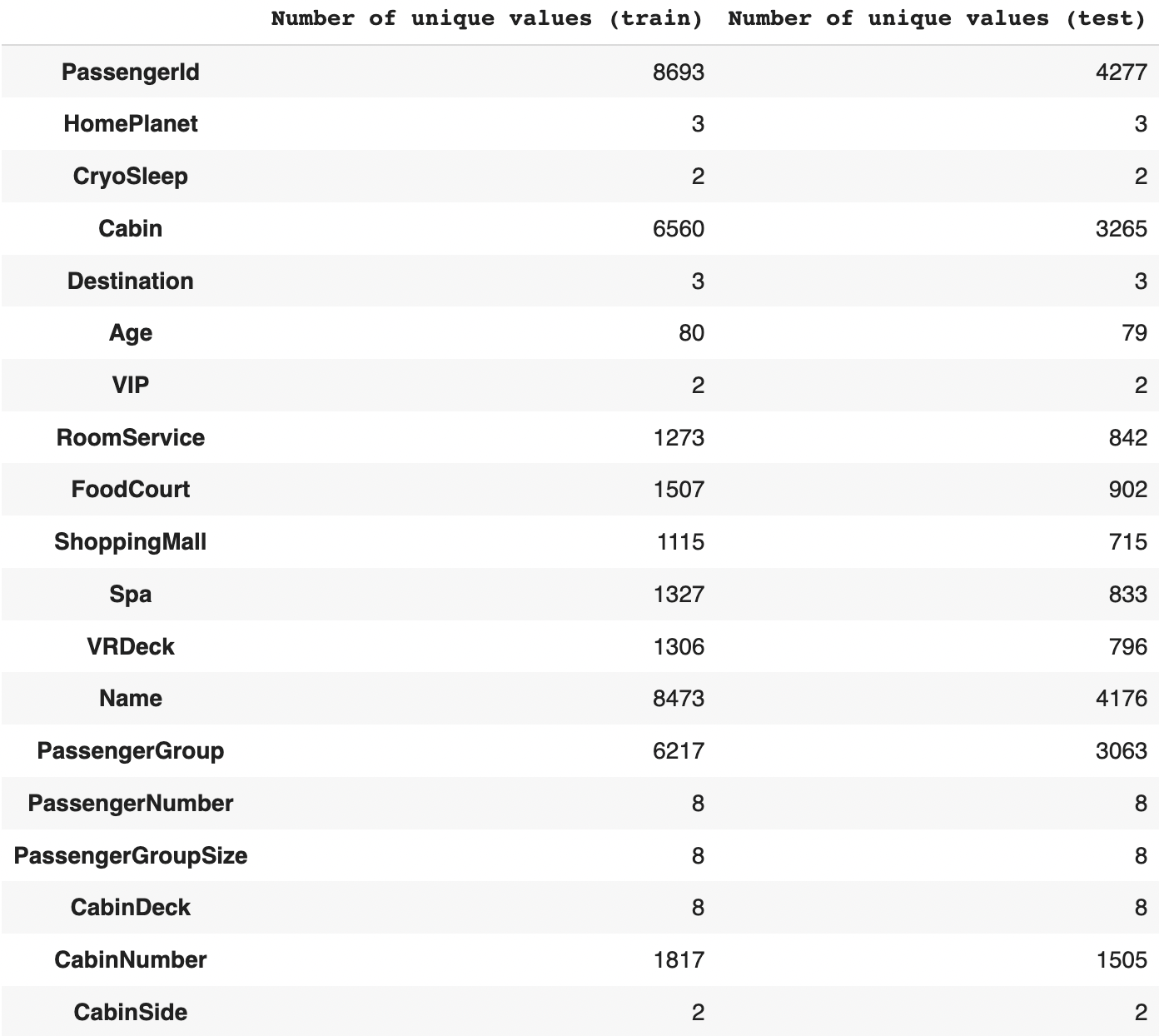
**Data Description and initial processing**

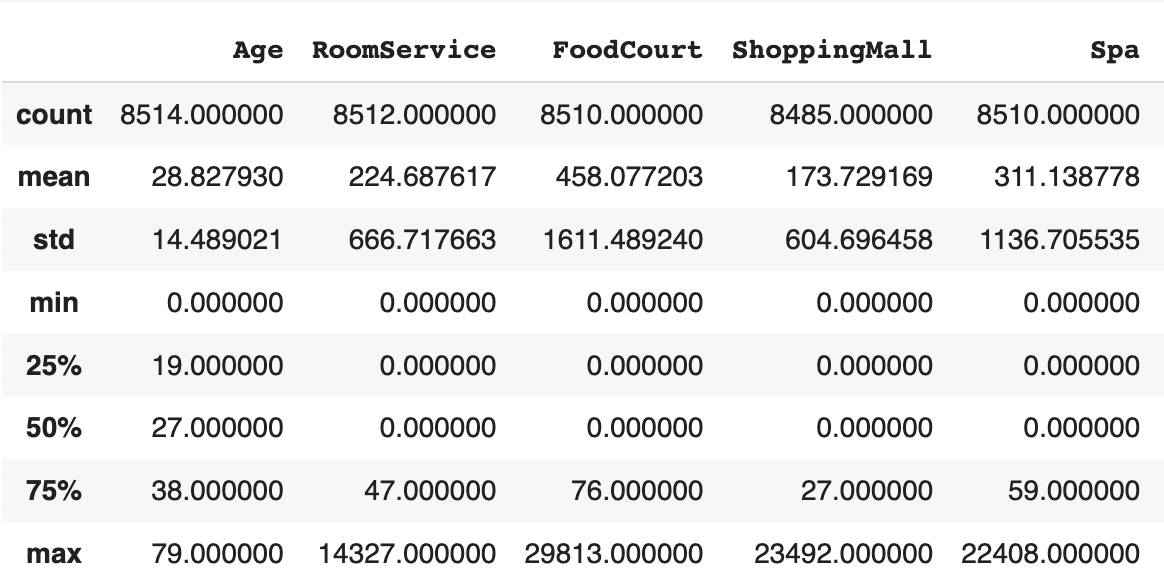
**Preprocessing**

In data preprocessing the passangerId column is split into two columns with underscore as the delimiter. The new columns are named PassangerGroup and PassangerNumber . Next, The Cabin column is split into three columns using ‘/’ as the delimiter, the new columns created in this step are CabinDeck, CabinNumber, and CabinSide. Then one hot encoding is performed on the categorical features and after that the data is standardized and the missing values are filled with the median values from the respective feature. The columns name, passangerId and Cabin are dropped because name has no relevance to our target variable. passangerId and Cabin are used to create other features and can be dropped without losing any information.

**Exploratory Data Analysis**

The cardinality of the data gives the unique values in each of the feature variable. The table below describes the cardinality.



The statistical description of the data gives information about the various statistics of the features. 

Table

Description automatically generated

The correlation between feature variables describes how the features are related to each other and the trend they follow.

Table

Description automatically generated

Table

Description automatically generated

The correlation among all the feature variables is low which means that none of the features can be discounted for building the models.

The plots of counts of different unique values in terms of the target value can give some good insights about the data. The plots are as follows.

Chart, histogram

Description automatically generated

The above plot shows that most of passengers were from age group 18 to 40.

Chart

Description automatically generated

The above plot shows majority of passengers did not order from room service.

A picture containing diagram

Description automatically generated

The above plot shows majority of passengers did not order from food court.

A picture containing chart

Description automatically generated

The above plot show that most of the passengers did not shop at the shopping mall.

Chart, histogram

Description automatically generated with medium confidence

The above plot show that most of the passengers did not use at the spa.

Chart, histogram

Description automatically generated

The above plot show that most of the passengers did not use at the VR Deck.

Diagram

Description automatically generated

The above plot shows that majority of the passengers were traveling alone or in group of 2 or 3.

Chart, histogram

Description automatically generated

The above plot shows that most of the passengers stayed in cabin with cabin numbers higher than 4.

Chart, bar chart

Description automatically generated

The above plot shows that most of the passengers were from earth.

Chart, bar chart

Description automatically generated

The above plot shows that most passengers were not in Cryo sleep which raises the question why a smaller number of passengers used room service, spa and VR Deck.

Chart, bar chart

Description automatically generated

The above plot shows that most of the passengers were traveling to TRAPPIST-1ePOS.

Chart, bar chart

Description automatically generated

The above plot shows that most of the passengers were not VIPs. The number of VIP passengers is very small compared to non VIP passengers.

Chart, bar chart

Description automatically generated

The above plot shows that deck F and G had the largest number of passengers. A large number of passengers were transported to different dimension from deck B,G and C.

Chart, bar chart

Description automatically generated

The above plot shows that there were almost equal number of passengers on both cabin sides but more passengers were transported from side S.

Chart

Description automatically generated

The above plot shows the ratio of passangers by homePlanet and CryoSleep. The largest ratio of passengers who were transported to another dimension are from Europa.

Modeling

For training models, three algorithms are chosen which are logistic regression, random forest, and multi-layer perceptron. The reason behind choosing these algorithms is complexity of the algorithms, where logistic regression is the simplest, MLP is the most complex and random forest has medium complexity. The different complexity levels of these three algorithms enables the understanding of overfitting and underfitting in the models.

Three models are used to do the prediction task and their accuracy values are calculated by performing 5-fold cross validation and taking the average of the accuracy of all folds. Since, there are 28 features in the dataset, and they have low correlation among them no feature can be dropped directly from the data except for name, passengerId and cabin. Name is of no use in the prediction task and passengerId and cabin are split to create more columns. Dimensionality reduction is performed on the data set to reduce number of dimensions used for training the models, Principal component analysis is performed on the training and testing data and first 21 principal components are used for training the models. The first 21 principal components capture more than 99 percent of the variance in the data. Logistic regression, Support vector machine and Multi-layer perceptron algorithms are used to train the models.

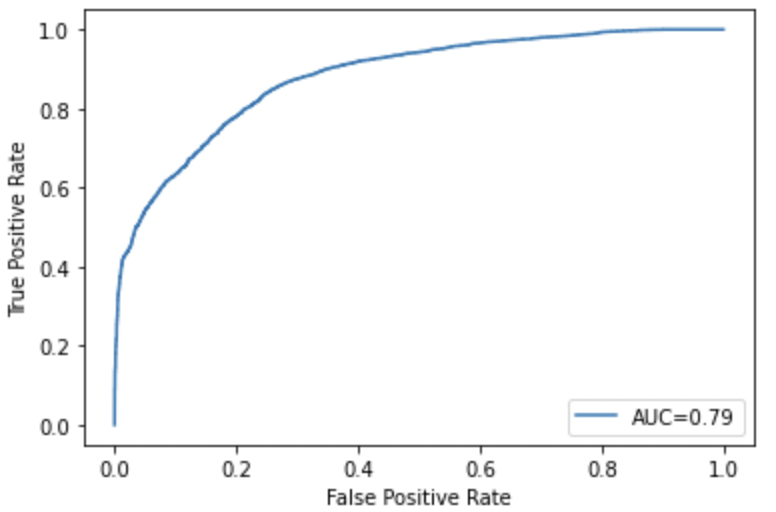
In this prediction task the model tries to predict if the passenger was transported or not. This is a supervised learning classification task because the training data set provides the ground truth labels. The choice of algorithm is based on the complexity of the algorithm. The algorithms are chosen to range from low complexity to high complexity, Logistic regression is the least complex algorithm, random forest has medium complexity and multi-layer perceptron has the highest complexity. Since PCA is used to reduce the dimensionality of the data set it is difficult to say which feature is the most important or the least important for learning the target variable.

Logistic Regression

Logistic regression is a classification algorithm that predicts a binary outcome based on a series of independent variables it one of the most basic methods of binary classification in machine learning. The logistic regression algorithm can help predict the likelihood of events by looking at historical data points. For training the logistic regression model the first 21 principal components are used.

The model gives an accuracy of 0.79 which is at par with the models from benchmarking section. The model performance is satisfactory. The 5-fold cross validation accuracies are 0.77113283 0.78895917 0.79815986 0.787687 0.78078251 which averages to 0.79. The f-score is 0.79, the precession is 0.79 and the recall is 0.81 on the training data.

The below plot describes the ROC curve of the logistic regression model.



**Random forest**

Random forest is a supervised learning algorithm. The “forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

For training the random forest model the first 21 principal components are used. The model gives an accuracy of 0.79 which is at par with the models from benchmarking section. The model performance is satisfactory. The 5-fold cross validation accuracies are 0.7573318 0.78723404 0.78953422 0.79056387 0.80034522 which averages to 0.79. The f-score is 0.84, the precession is 0.87 and the recall is 0.81 on the training data. the max depth of the model is 8, the max depth parameter is decided by trial-and-error method where the model was trained with different values of max depth parameter and the best result was obtained by setting max depth to 8.

The below plot describes the ROC curve of the random forest model.

**Chart

Description automatically generated**

**Multi-layer Perceptron**

Multilayer Perceptron (MLPs) are the building blocks of neural network. They are comprised of one or more layers of neurons. Data is fed to the input layer, there may be one or more hidden layers providing levels of abstraction, and predictions are made on the output layer, also called the visible layer. They are very flexible and can be used generally to learn a mapping from inputs to outputs. This flexibility allows them to be applied to other types of data.

The multi-layer perceptron model has three layers in a fully connected architecture, the input layer has 21 nodes with Relu activation which is dictated by the 21 principal components used meaning each observation has 21 dimensions, the hidden layer has 10 nodes with Relu activation, and the output layer has one node with sigmoid activation because the goal is to perform binary classification. The model is trained for 150 epochs and inputs are feed in batches of 100 points to the model. The network uses binary cross entropy loss. For training the multi-layer perceptron model using the first 21 principal components. The model gives an accuracy of 0.93 which is at much better than the models from benchmarking section. The model performance is satisfactory.

The below plot describes the ROC curve of the MLP model.

Chart

Description automatically generated

The below ROC curve compares the performance of the three models used for the prediction task. Random forest and Multi level perceptron performs better than the logistic regression model. random forest and Multi level perceptron have similar performance but based on accuracy value multi-layer perceptron has the best performance.

**Chart

Description automatically generated**

**Appendix**

The three solutions mentioned in the benchmarking section of this report do not use PCA for dimensionality reduction they use other methods for dimensionally reduction. Using PCA for dimensionality helped in training the models with less data. The models described in the benchmarking section use only 2 components which explains less than 90 percent of the variability in the data. The models described in this report performs better than the benchmark solution because these models use 21 principal components which capture more than 99 percent of the variance.

The models in benchmarking section do not use complex algorithms like multi-layer perceptron for the predication task where the MLP model described in this report surpasses every model from the benchmarking section with an accuracy of 0.92.

The model with the highest accuracy from benchmarking section is support vector machine with an accuracy of 0.805 whereas multi-layer perceptron has an accuracy 0.92 which is 12 percent higher than support vector machine model from benchmarking section.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00