**Business Statistics Project Report**

On

**“Developing a Classification Model Using Random Forest to find the Passenger Status in a Spaceship Accident Scenario”**

Submitted in partial fulfillment for award of

**Master’s in Business Administration**

Degree

In

General Management

**By**

**Team 09**

|  |  |
| --- | --- |
| Shreya K | 22BM63066 |
| Deepak Singh | 22BM63042 |
| Sarath R | 22BM63106 |

Under the guidance of

**Professor Sujoy Bhattacharya**



Vinod Gupta School of Management

Indian Institute of Technology, Kharagpur

Kharagpur, India

**November, 2022**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Section** | **Title** | **Page No.** |
|  |  |  |
| 1. | Introduction | 2 |
| 2. | Problem Statement | 2 |
| 3. | Methodology | 2 |
|  | Approach | 3 |
|  | Formulation | 3 |
|  | Dataset: Explained | 5 |
|  | Model Development | 8 |
| 4. | Results and Discussion | 15 |
| 5. | Conclusion | 15 |
| 6. | Bibliography | 15 |
| 7. | Appendix – Contribution Table | 16 |

1. **Introduction**

In order to understand the complexities of the world, humans have always resorted to field of statistics in order to formulate the real life events as a workable model. Starting from the ages of manual calculations, the field has advanced into a much advanced independent dimension, commonly referred as Data Analytics. This has presented an opportunity to utilize and implement the learnings in the field of Business, where Executives deal with huge amount of data in daily basis in order to arrive at a meaningful decision. The purpose of a Business Analytics is to assist the decision making by backing it up with a quantifiable data and providing a viable solution to the problems faced by the target entity. It utilizes various data analytics tools in order to predict and analyze the available data.

One such tool is Random Forest. Random Forest is a supervised learning algorithm that uses ensemble methods to solve both regression and classification problems. It operates by constructing many decision trees at training time and outputs the mean/mode of prediction of individual trees. We have used this algorithm for our model construction.

1. **Problem Statement**

The problem statement pertains to a fictional scenario in year 2912 where a spaceship TITANIC, transporting passengers from 3 different planets to 3 different destinations meets a space anomaly, resulting in the teleporting of several passengers into alternative dimension. The objective is to assist the rescue team by analyzing the database of damaged ship and identifying whether passenger has been teleported or not. We have been provided with different attributes associated with passengers in order to develop a training model and implement the algorithm in test dataset, returning the results in form of Passenger ID along with the status whether he/she was teleported or not.

1. **Methodology**

The given problem statement pertains to a case of **Classification Model** withrequirement of predicting the status of passengers, that is, essentially classifying if the passenger has been teleported or not by utilising the given dataset.

* **Decision Making Tool: Random Forest**

**Why** Random Forest over Decision Tree?

Random forest adds additional randomness to the model while growing trees. When splitting a node, it searches for the best feature among a random subset of features instead of looking for the most important feature. Thus, it reduces the overfitting problem in decision trees and lessens the variance, improving accuracy.

**Target Variable**

TRUE : Transported

FALSE : Not Transported

Since our **target** variable is **categorical** and our **independent** dataset is **discrete and categorical**,primarily, the problem pertains to a ***Classification Model*** that can be easily formulated as well as solved using Random Forest.

The final ***evaluation*** of our decision-making model is based on the ***Classification Accuracy***, calculated through:

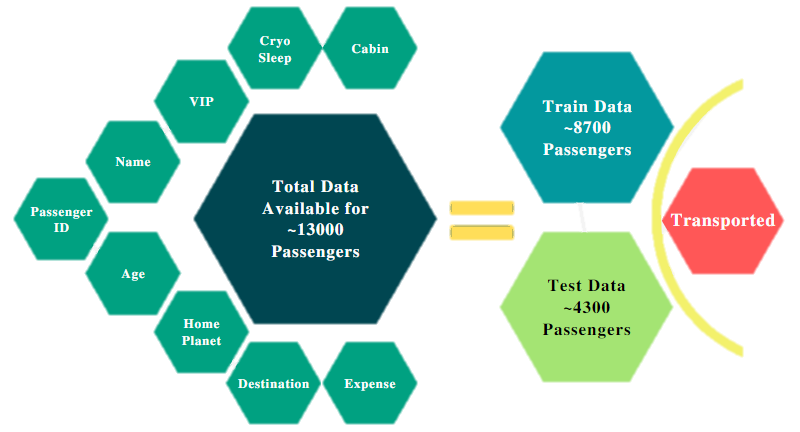
*Classification Accuracy:*

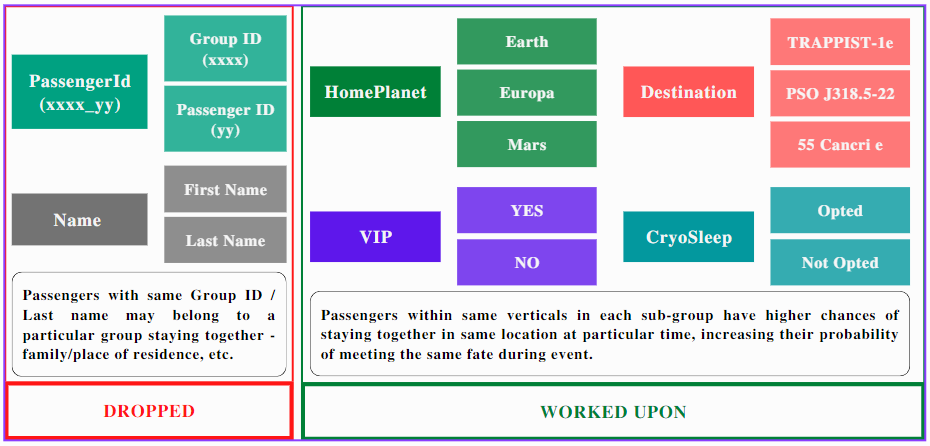
* **Approach**

Our approach can be segregated under two stages:

1. Formulating problem
2. Developing Model and Analysis
3. **Formulation**

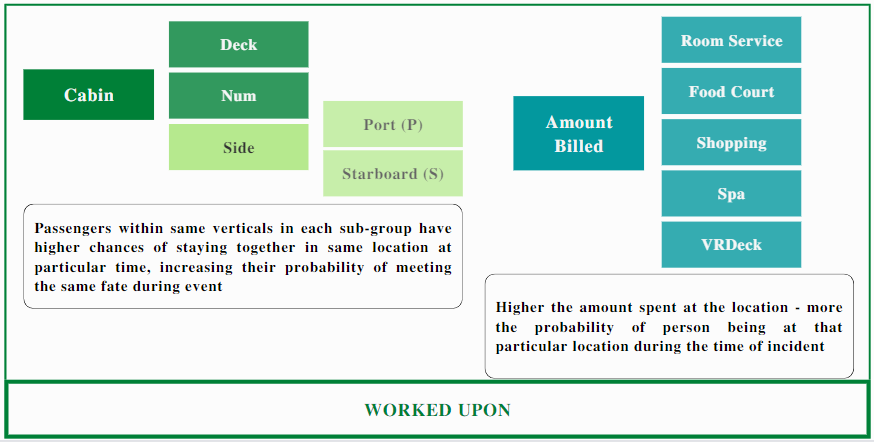
Formulation involved defining our variables as independent and target variables and segregating them further on the basis of their utility in our Test Model. ***Figure 3.1*** depicts various input variables that were provided with the problem statement along with the target variable that was to be determined.

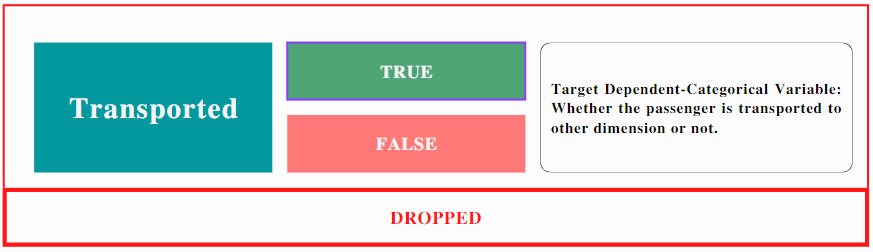




*Fig. 3.2.1 Dataset Available*

*Fig. 3.1 Model Development - Outline*





*Fig. 3.2.2 Dataset Available*

*Fig. 3.3 Target Variable*

* **Dataset Explained:**

The dataset retrieved from damaged ship was primarily segregated into following groups:

* **Passenger ID : xxxx\_yy**

Data Type : **Numerical**

Significance :

Passenger ID depicted the Unique ID assigned to the passengers for the journey. Entire database was with reference to this ID and thus final outcome was also assigned to it. Also, the ID comprised of both Group ID (xxxx) that was assigned to a particular group travelling together along with unique number associated with the passenger (yy).

Since it was unique to the passenger, the dataset could not have contributed in training our model and hence was dropped from our test dataset.

* **Name**

Data Type : **Categorical**

Significance :

Name comprises of First and Last name of an individual. The last name might reflect the person belonging to a particular group, example a family, same residential location, etc.

Considering each name unique to an individual, the dataset has been dropped out of our test model.

* **Age**

Data Type :  **Numerical**

Significance :

Person belonging to different age groups have different capacities to respond to a given emergency situations. Hence the data has been retained in our working model.

* **VIP**

Data Type : **Categorical**

Significance :

Since all the VIP persons might be placed together by the crew, hence they can be grouped together. The data has been retained.

* **Home Planet**

Data Type : **Categorical**

Significance :

The dataset classifies passengers among three categories on the basis of their origin planet: Earth, Europa, and Mars. As persons belonging to same locations often are clustered together, their chances of meeting the same fate are higher during any adverse event. Hence, data has been retained in our model.

* **Destination Planet**

Data Type : **Categorical**

Significance :

The dataset classifies passengers among three categories on the basis of their destinations. Persons can be grouped on the basis of their destinations, which raises their chances of meeting the same fate during any adverse event. Hence, data has been retained in our model.

* **Cryo-sleep**

Data Type : **Categorical**

Significance :

Data classifies passengers on the basis of their choice to opt for Cryo-sleep. All passengers opting for cryo-sleep have high probability of being placed together, and hence ending up in similar situation as others in same group. Thus, the data has been retained in our model.

* **Cabin**

Data Type : **Categorical**

Significance :

Passengers have been segregated on the basis of their cabin’s position in the ship – as Port side or Starboard Side, along with the number of the cabin occupied. Persons on the same side of the ship and closer to the assigned number have higher chances of meeting the same fate. Thus, the data has been retained in our model.

* **Expenses**

Data Type : **Categorical**

Significance :

Expenses at a particular location can be used to determine the probable location of the passenger at the time of the event, under the assumption that more the amount a passenger spends at a particular location, higher the amount of time spent by her at that location.

* **Transported**

Data Type : **Categorical**

Significance :

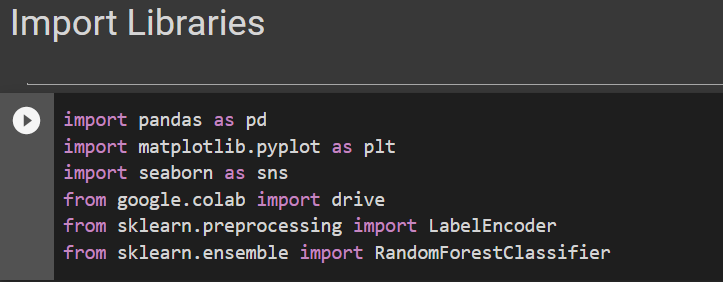
This is our Target Variable, and it signifies whether the passenger was teleported or not. Since this was our target variable, it has been segregated to the variable ‘Y’ signifying that it is our target variable.

1. **Model Development**

Various **stages** involved in the process were:

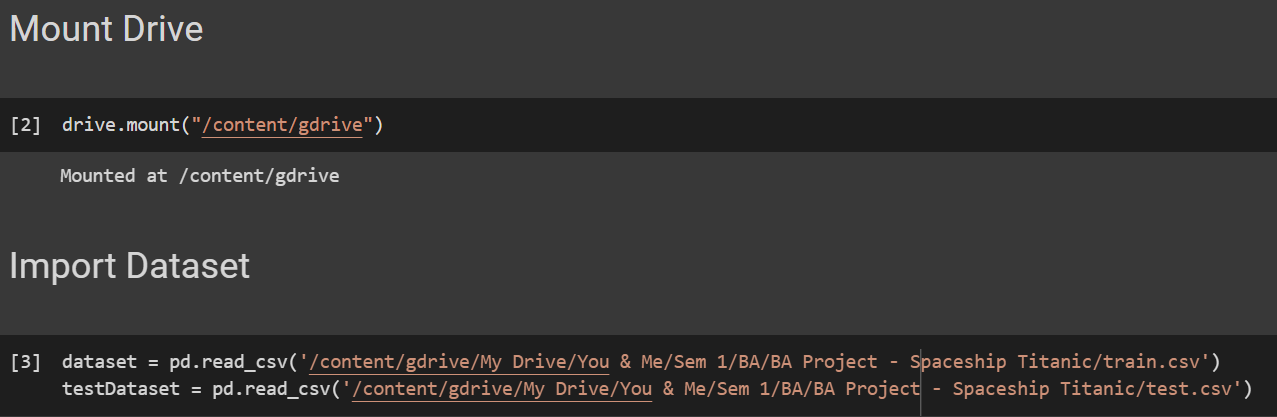
1. Platform Selection : *Google Collab*

Selecting Library : *pandas, sklearn, matplotlib, seaborn*

**

*Fig. 3.4 Importing Libraries*

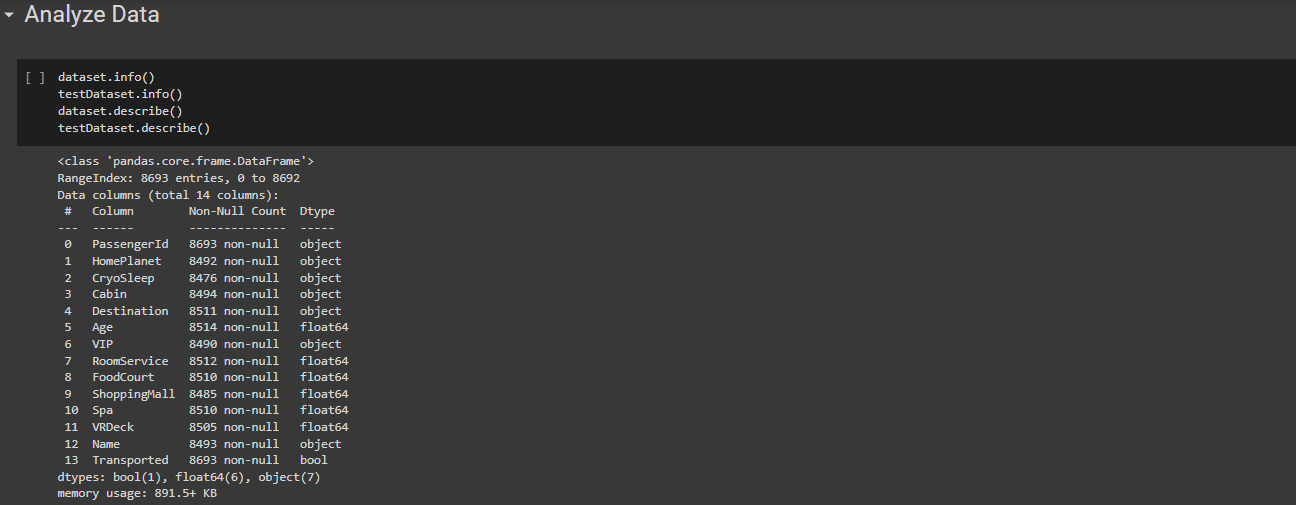
1. **Importing** source dataset: *Train dataset* + *Test Dataset*



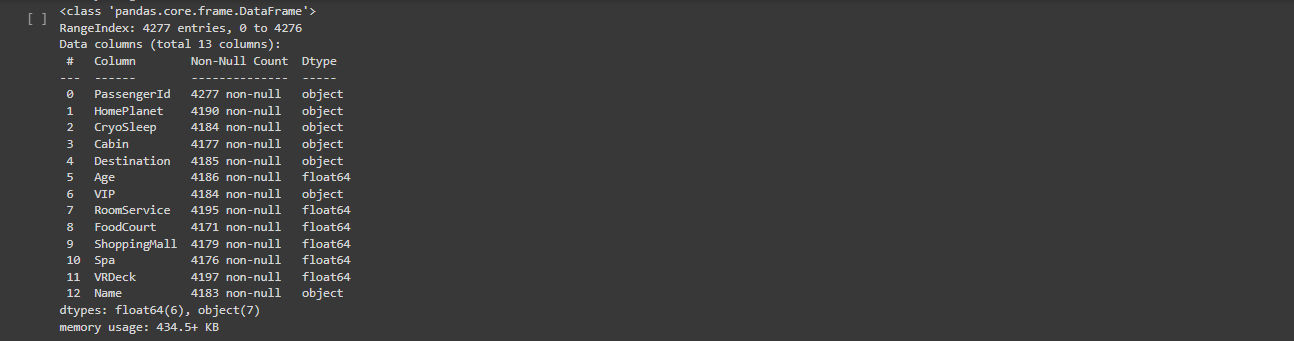
*Fig. 3.5 Importing Dataset from Google Drive*

1. **Analyse** the given datasets for developing a descriptive statistic (to be used in assigning values to vacant columns) and determining the unfilled rows and columns.

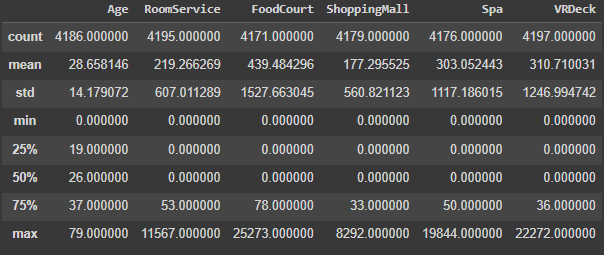
* Info() method gives us the number of non-null entries and the data type of each feature.
* Describe() method gives us the descriptive statistics of the dataset.



*Fig. 3.6 Data Type and Non-null Entries in Training Dataset*



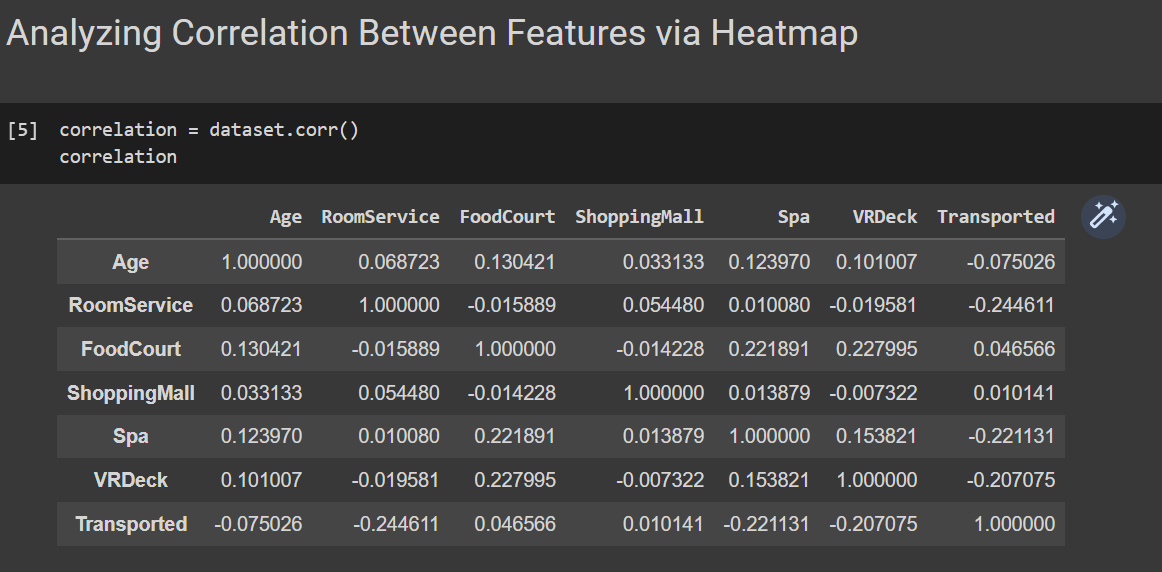
*Fig. 3.7 Data Type and Non-null Entries in Testing Dataset*



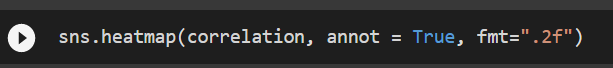
*Fig. 3.8 Descriptive Statistics of Training Dataset*

1. **Understanding the correlation** between different features

In order to understand how closely the features are related to each other and to the target variable, we are computing the correlation among them. Then we are plotting the correlation on a heatmap to get a visualization of the same. We can take decision regarding which features to keep or drop in the subsequent stages of model development.



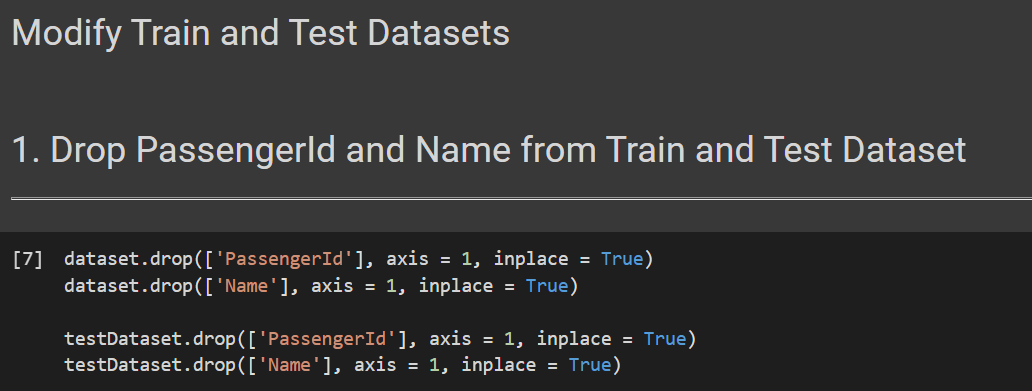
*Fig. 3.9 Correlation Map Between Features*





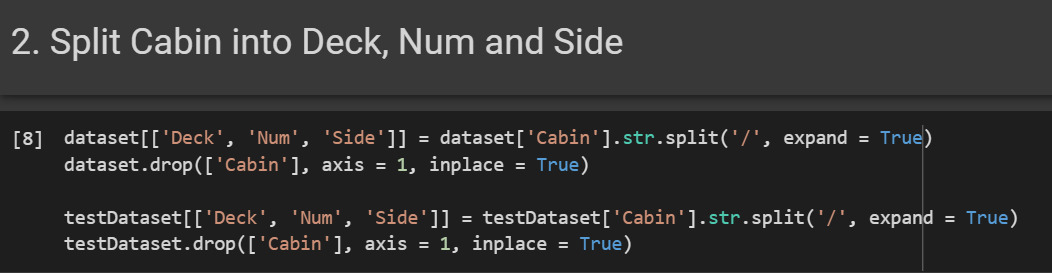
*Fig. 3.10 Heatmap Map Between Features highlighting Correlation*

1. **Modifying** the dataset
   * + *Dropping* Passenger ID and Name (unique to passengers – no pattern developed)



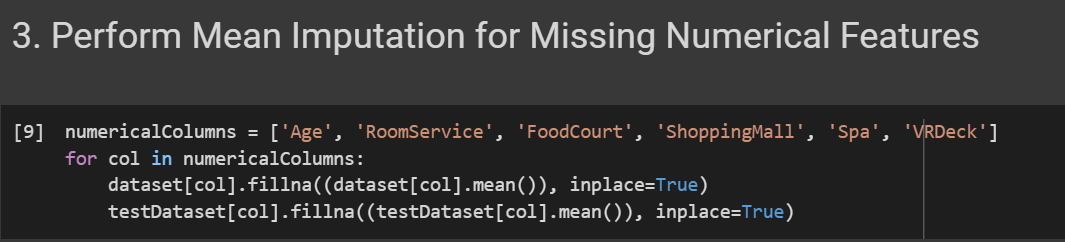
*Fig. 3.11 Dropping Passenger Id and Name from Training and Test Datasets*

* + - *Splitting* Cabin data into respective datasets. We are doing this because, cabin is a combination of Deck, Number and Side each of which represent a position in the spaceship thereby having a contributing factor in determining if the passenger will be transported or not.



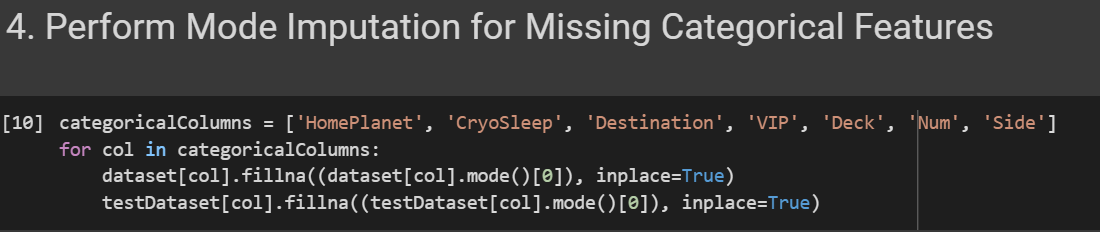
*Fig. 3.12 Splitting Cabin into Further Features*

* + - For missing *numerical data*, we performed *Mean Imputation* on the dataset.



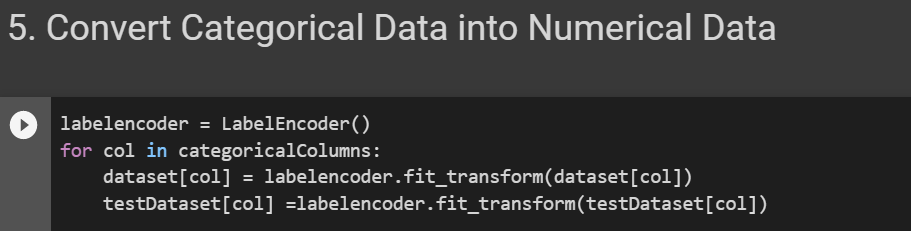
*Fig. 3.13 Mean Imputation*

* + - For missing *categorical* values, we performed the *Mode Imputation* on the dataset

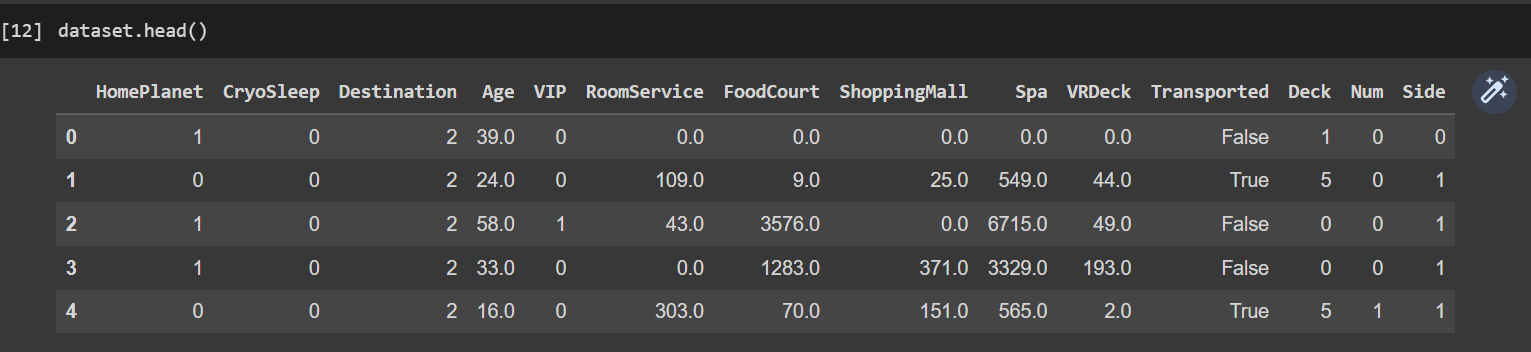


*Fig. 3.14 Mode Imputation*

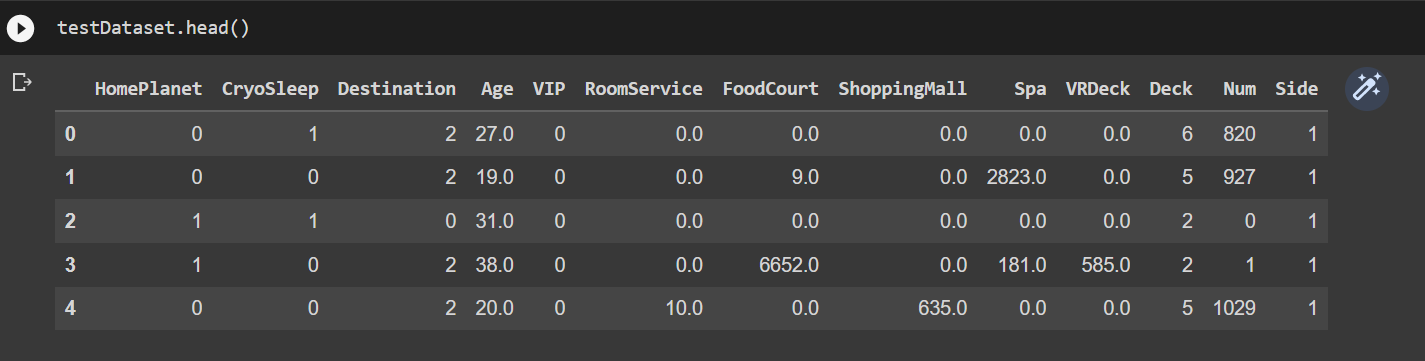
* + - 1. Further the categorical dataset was converted into numerical dataset in order to develop a relation among our independent variable.
    - Affirmative/Negative Responses converted into binary responses.
    - Home and Destination planets were assigned numerical values.



*Fig. 3.15 Conversion of Categorical to Numerical Data*

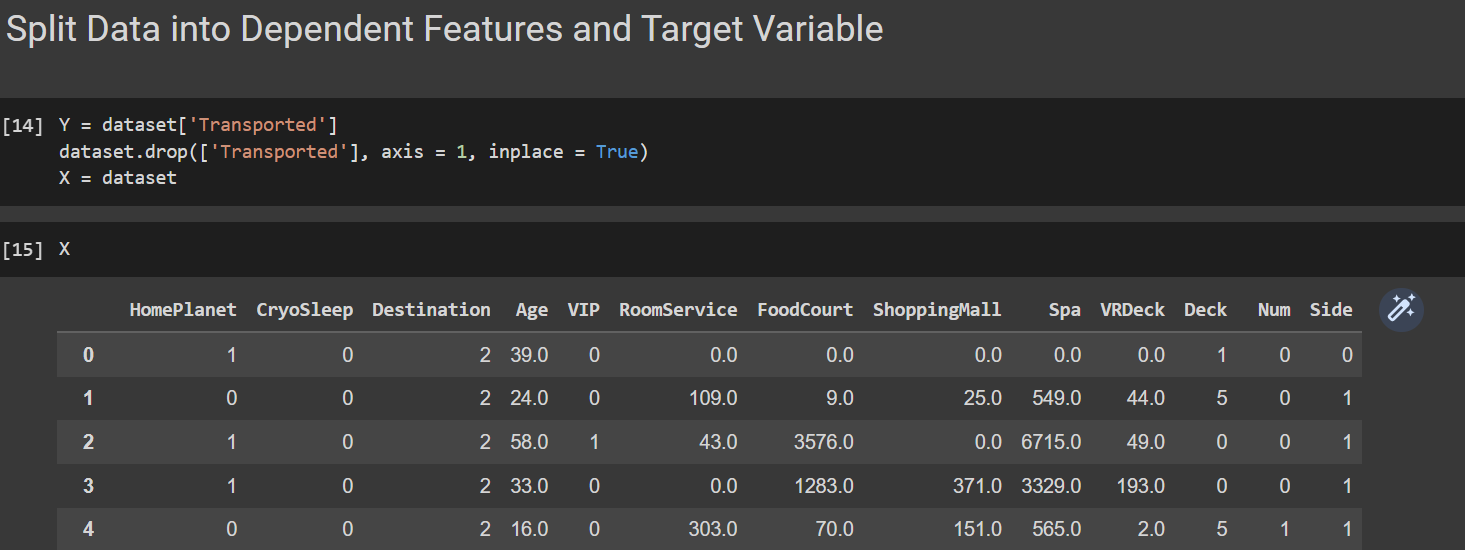


*Fig. 3.16 Converted Training Dataset*

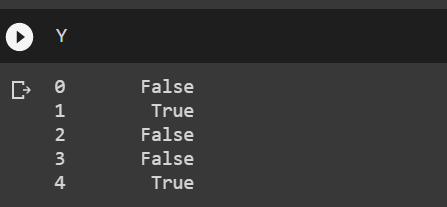


*Fig. 3.17 Converted Testing Dataset*

1. Further dataset was further split into *independent features* and the *target variables*:

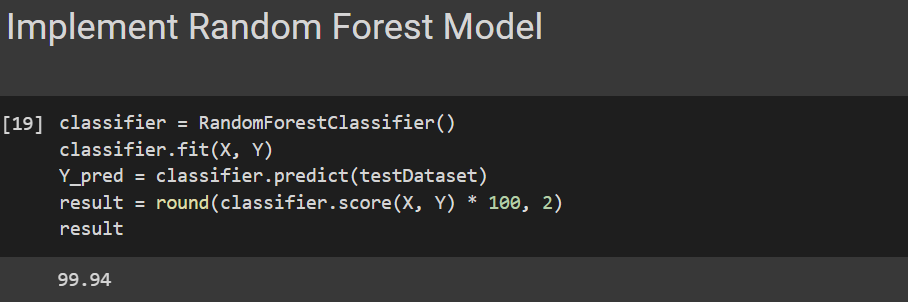


*Fig. 3.18 Independent Features Represented by ‘X’*



*Fig. 3.19 Dependent Feature Represented by ‘Y’*

1. Finally *running* the *Random Forest model*



*Fig. 3.20 Random Forest Implementation*

1. **Results and Discussion**

Our model returned a Classification Accuracy of ***99.94 %***. Therefore, the accuracy of the model is very good. With the use of Random Forest Algorithm over Decision Trees, the potential of over-fitting has decreased significantly though it may occur due to the limited amount of input data available with us. This can result in conveying false information for some passengers, and needs to be corrected by including some more information in the dataset.

1. **Conclusion**

The undertaken study helps us in developing a basic understanding about Business Analytics as a tool for decision-making. It further exposes the limitations of the Random Forest as a tool with limited dataset. The undertaken study can be used in other decision-making scenarios as well, provided there is sufficient data available for model to run successfully, without *over-fitting* or *under-fitting* our model.

The field of Analytics is equipped with several other tools to deal with similar decision-making dilemmas, which can be used along with Random Forest in order to arrive at a better conclusion.

1. **Bibliography**

* [*Spaceship Titanic: Kaggle link*](https://www.kaggle.com/competitions/spaceship-titanic/data)
* *[Coding and model development : Google Collab link](https://colab.research.google.com/drive/1orafbNtjIEmAZoGbOkf_rkZrh-AY2efk?usp=sharing)*
* *Data Set:*

**

1. **Appendix: Contribution Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task / Roll No.** | **22BM63042** | **22BM63066** | **22BM63106** |
| **Problem Identification** | **✓** | **✓** | **✓** |
| **Methodology Development** | **✓** | **✓** | **✓** |
| **Coding** | **✓** | **✓** | **✓** |
| **Analysis** | **✓** | **✓** | **✓** |
| **Report Compilation** | **✓** | **✓** | **✓** |
| **Contribution (%)** | **33** | **34** | **33** |