**Data Pre-processing and Analysis for Travel Package Purchase Prediction**

**By:**

**Riya Shukla , Sahil Goyal , Sameer Khatwani**

**"Unlocking the Hidden Treasures of Company Data: A Journey into Understanding Customer Behavior"**

### Greetings and Welcome to our blog that will take our readers on an exciting and enlightening journey through the obstacles our organization faced, the painstaking data processing that came next,data visualization and the thrilling finish in the shape of an approachable model deployment site. You will have a thorough grasp of how we used data to empower our team with binary customer classification and make well-informed decisions by the conclusion of this read.

**Section 1: Knowing about the Problem**

We'll get right to the core of the issue in our first section. What issue does our business have, and how did we identify it? Additionally, we'll expose you to the dataset that served as the basis for our solution. Understanding a problem is the first step towards solving it.

**Section 2: Data preprocessing**

Data processing is typically a tricky maze that we expertly negotiated. We will present our methodical and intelligent approach in managing undesired features,errors in the dataset and missing values. Readers will discover how to tame and transform even the most complex data into a useful resource.

**Section 3: Data Visualization**

In this section our readers will come across univariate and bivariate analysis of the company’s data, which will enable all our readers to get captivated insights of the data through an impactful visualizations which will result in profound understanding of our data.

**Section 4: Model Deployment Techniques**

The application of models is what makes theory a reality. Our team has created a platform that enables binary classification. Our company is now better able to anticipate client decisions, customize our services, and provide a more customized customer experience thanks to analytics and an intuitive platform.

**Section 1: Knowing about the Problem**

**Introduction**

Our travel agency, Travelea, has unveiled a cutting-edge trip package with the main goal of growing our clientele and creating a viable business plan. We have five different travel packages available right now: Basic, Standard, Deluxe, Super Deluxe, and King. But part of our old marketing approach was contacting prospective clients without doing any preliminary research. This strategy resulted in a meager 18% conversion rate and substantial marketing expenses.

To draw in and keep a wider audience, we are now concentrating on honing our marketing approach, expanding our knowledge of our clientele, and upgrading our products. A key to accomplishing these goals is the launch of the new travel package, and we're dedicated to using data-driven insights and smart marketing strategies to meet our expansion targets and control expenses.

Inorder to improve the customer base of the company we will be working on the past data collected about the customers which has been supplemented from kaggle and then has been imputed with few extra features through web research and information extraction to enhance the quality of our analysis.

**Problem Statement**

Travelea has the significant challenge of achieving significant commercial expansion and sustainability in the face of a stagnant clientele and inadequate marketing tactics. With our current 18% conversion rate and growing marketing costs, we really need to implement a data-driven marketing plan, gain a deeper understanding of our customer base, and reorganize our product line. The introduction of our innovative travel package is a critical first step in accomplishing these objectives, and we're dedicated to use data-driven insights and cost-effective marketing tactics to foster expansion and long-term success.

**Understanding The Dataset**

* We improved the Kaggle dataset by proactively resolving missing data using logical imputation techniques, giving priority to important variables like the customer's insurance and budget (these two characteristics we have imputed into the dataset).
* We used information extraction techniques and web scraping to get relevant data from many internet sources.
* As a result, our dataset has expanded to include 4,888 rows and 22 columns, creating a thorough and complete information source.
* The dataset consists of both Text and Numeric Values.

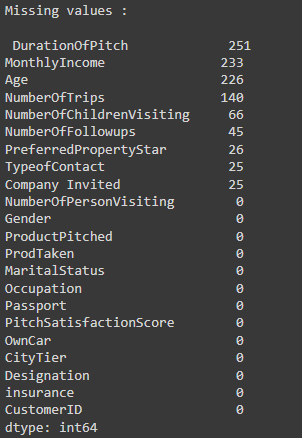
**Dataset Descriptions**

* + customerID: A distinct customer ID
  + productTaken: The status of the product taken, either zero or one.
  + Age: Customer's age
  + TypeofContact: The method of communication with the consumer (self-inquiry or company invited)
  + CityTier: A city's growth, population, amenities, and level of living all affect its tier. The groups are arranged in a i.e.1,2,3
  + DurationOfPitch: The length of a salesperson's presentation to a client
  + Gender: The customer's gender
  + NumberOfPersonVisiting: The total number of people that are scheduled to go with the client
  + Number of Follow-Ups: Following the sales presentation, the salesperson has completed a total of number of follow-ups.
  + ProductPitched: The salesman's product pitch
  + PreferredPropertyStar: The customer's preferred hotel property rating
  + MaritalStatus: The customer's marital status
  + NumberOfTrips: The typical number of trips a client takes annually
  + Passport: Whether or not the consumer possesses a passport (0: No, 1: Yes)
  + PitchSatisfactionScore: The satisfaction score of a sales pitch
  + OwnCar: Indicates if the clients have a car or not (0: No, 1: Yes).
  + NumberOfChildrenVisiting: The total number of kids under five that are there ,arranging to travel with the client
  + Designation: The customer's designation inside the present company
  + MonthlyIncome: The customer's gross monthly income

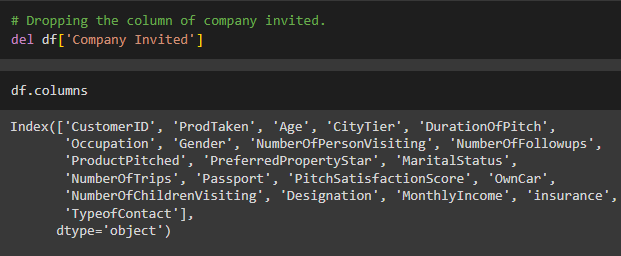
**Section\_2 : Data Preprocessing**

We incorporated two new features in our dataset which were budget and insurance which are important aspects to be taken into consideration for our Binary Classification.

In data preprocessing we started with Analyzing the missing values in the dataset. There were in total 9 features with missing values in them.

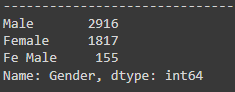
****

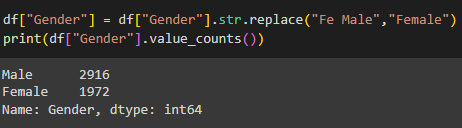
**Here, we can also observe that Company Invited and Type of Contact were the same . So, we chose to drop the column of Company Invited.**

****

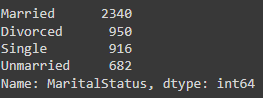
**Dealing with the errors in the dataset:**

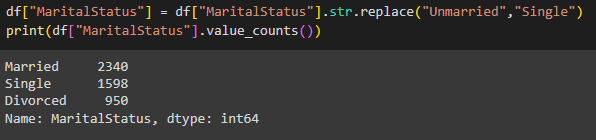
1. **Gender :** Gender Distribution has an error ,as it is considering Fe male as a unique value. So, we replaced the string ‘Fe male’ with ‘Female’.

****

****

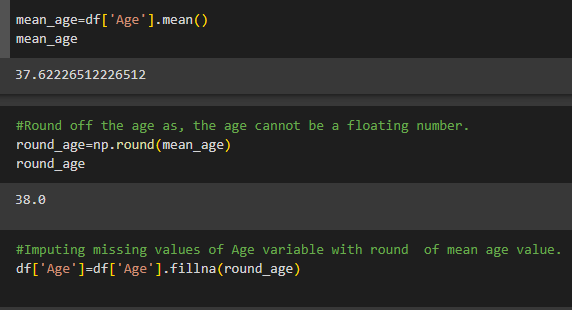
1. **Marital status :** In marital status single and unmarried are occuring which are of the same meaning. So,we replaced ‘Unmarried’ with ‘single’.

****

****

**Dealing with Missing Values:**

1. **Age :** We used mean imputation to fill in the missing values for the feature "age," considering that it followed a normal distribution.

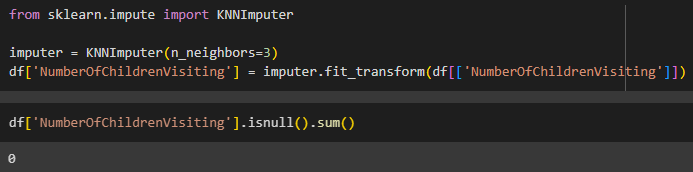


1. **Number of Children Visiting :** Approached to fill the missing valuesby rearranging the dataset according to marital status and then estimating the number of children who visit with them using K-nearest neighbors (KNN) imputation.

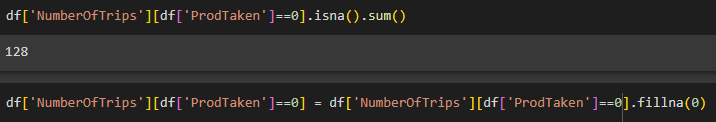
Sorting according to marital status

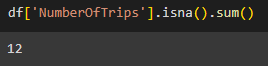


Then ,applying KNN Imputation

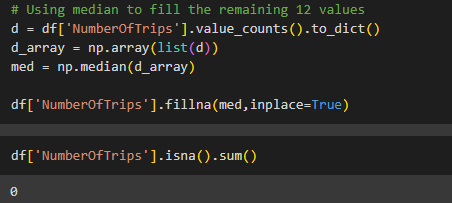


**Number of Trips:** If a customer's "Product Taken" value is 0, it is reasonable to infer that the "Number of Trips" will also be 0. This is because if the customer has not previously acquired any products from the company, it is unlikely that they would participate in trips offered by the same company.

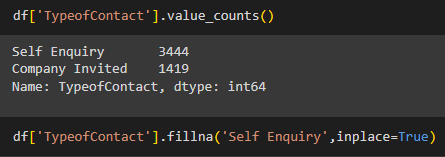




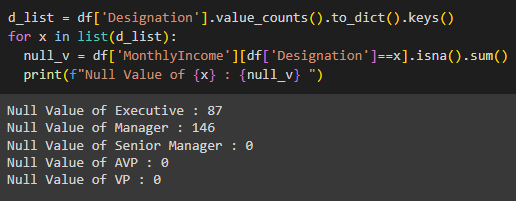
But , still we were left with 12 missing values. So we imputed the missing values through median imputation.



1. **Type of Contact :** Imputed the missing values using the mode, which is the value in the dataset that occurs the most frequently. So, Self Enquiry appeared most of the time so we imputed the value with it.

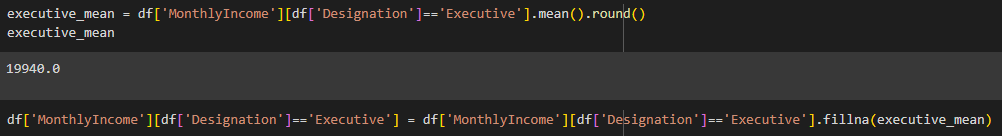


1. **Monthly Income :** Mapping the relationship between Monthly Income and Designation of the Customer and filling the null values with respect to it.

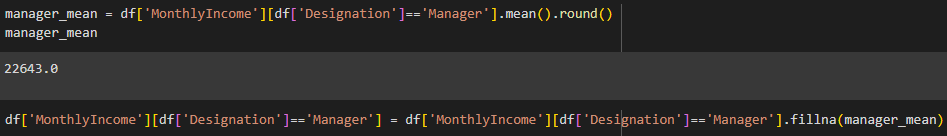


We noticed that the monthly income of only customers with the designations of Executive and Manager was absent. In light of the customer's mean wage and the same designation, the missing figures for monthly income were therefore imputed.

For the Executive’s Income



For the Manager’s Income



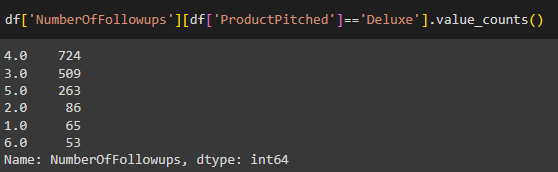
1. **Preferred Property Star :** Filled up the Preferred Property Star's null values in relation to the Mean of the corresponding Designation.After analyzing the counts of property star preferred by the customer according to his post we concluded that customers of all designation preferred 3.0 star property . So, we impute the Null value with 3.0



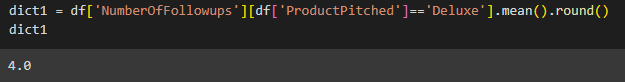
1. **Number Of Follow Ups :** Number of Follow Ups feature is imputed with respect to the Product Pitched . We checked the null values Number Of Follow Ups where the product pitched were Basic,Deluxe,King,Super Deluxe ,and Standard.

We observed that null values were found only where the product pitched were Deluxe and Basic.

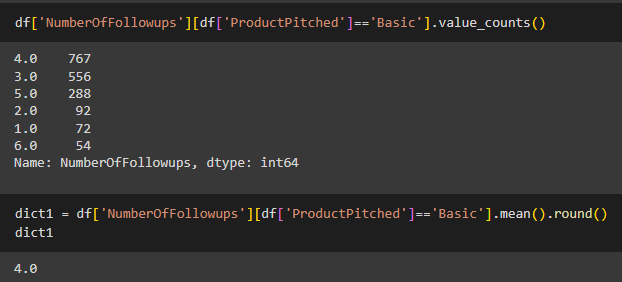
So , we counted the Number of Follow Ups when the product pitched was Deluxe.



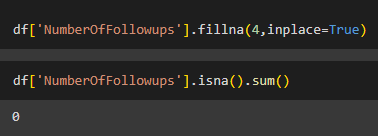
Imputed the missing values with mean . The median was also coming out to be same .



Same results appeared for Basic product as well



Finally, imputed the missing value for both Basic as well as Deluxe with mean .



**8.Duration Of Pitch :** We plotted a distribution graph for the feature and found that it was positively skewed so imputed the missing values by sorting the product pitched and then applying K-NearestNeighbours approach.

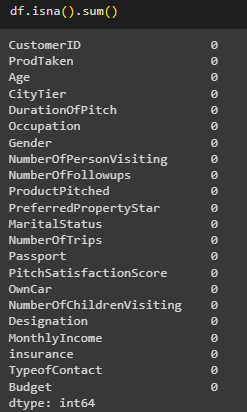
Sorted according to the product pitched



Applied KNN with 3 nearest neighbors.



Finally Checking for the null Values in the dataset



## 

## **Section-3: Data Visualization**

Data visualization is like a magic wand to harness the dataset of travel package purchases. It turns complex data into easy-to-understand visuals, helping businesses make smarter marketing decisions and improve the travel experience. In this blog, we'll see how it works and why it's crucial for adapting to travelers' changing preferences.

### **Univariate Analysis**

### 

#### **Fig 1 : CustomerID**

### 

#### ***Observations***

|  |  |
| --- | --- |
| 1. | CustomerID showcases balanced data as it serves as the primary key in our dataset. |
| 2. | CutomerID is only for identifying the customer record. |

#### **Fig 2 : Age**

### 

#### ***Observations***

|  |  |
| --- | --- |
| 1. | Age variable is almost normally distributed with no outliers. |
| 2. | We can observe that most customers are in the age brackets 30- 45 yrs. |

#### **Fig 3 : Product Taken**

### 

***Observations***

|  |  |
| --- | --- |
| 1. | 81.2% of the people, dont take the product |
| 2. | The Product has been taken by only 18.8% |

#### **Fig 4 : Type Of Contact**

### 

#### 

***Observation***

|  |  |
| --- | --- |
| 1. | Self-Enquiry is the most preferred contact method by 71% of customer |

#### 

#### **Fig 5 : Type Of Contact**

### 

### 

***Observation***

|  |  |
| --- | --- |
| 1. | 65.3% of customers are from Tier 1 cities and Tier3 cities customer's 30.7%. |
| 2. | Minimum Customer come from tier 2 |  |  |

#### 

#### **Fig 6 : Duration Of Pitch**

### 

### 

### 

***Observation***

|  |  |
| --- | --- |
| 1. | The Duration of Pitch is Positively Skewed |
| 2. | There are few outliers after 40 min |
| 3. | We can observe that most of the Data is highly concentrated before 20 min |

#### 

#### **Fig 7 : Occupation**

### 

### 

***Observation:***

1. Almost 90% of the Customer's were either Salaried or Small Business

#### **Fig 8 : Gender**

### 

***Observation***

1. Number of Male customers are higher than the Female customers

|  |  |
| --- | --- |
|  |  |

#### **Fig 9 : No Of Person Visiting**

### 

***Observation***

|  |  |
| --- | --- |
| 1. | Almost 10% customers either go alone, or take 5 people with them |
| 2. | It is more likely that the customer took three people with them |

#### 

#### **Fig 10 : No Of Person Visiting**

#### 

#### 

***Observation***

|  |  |
| --- | --- |
| 1. | 43.2 % customers had 4 followups. |
| 1. | 30% customers have 3 followups |

#### **Fig 11 : Preferred Property Star**

### 

### 

***Observation***

|  |  |
| --- | --- |
| 1. | 65.3% of customers prefer 3 star property. |

#### **Fig 12 : Marital Status**

### 

### 

***Observation***

|  |  |
| --- | --- |
| 1. | 47.9 % of the Customer are Married |

#### **Fig 13 : No. Of Trips**

### 

***Observation***

|  |  |
| --- | --- |
| 1. | NumberofTrips is right-skewed a little and majority of the customers seem to take at least 2 trips per year. |
| 2. | We also see very few outliers in the higher end. |

#### **Fig 14 : Passport**

### 

***Observation***

|  |  |
| --- | --- |
| **1.** | 70.9% of the customers do not possess Passport |

#### **Fig 15 : Pitch Satisfaction Score**

### 

***Observation***

|  |  |
| --- | --- |
| 1. | Only 30.2% of customers rated the sales |
| 2. | Even though 18.7% customers rated at 4. |
| 3. | 19.8% rated a pitch score of 5. |
| 4. | we also see that 19.3% rated the Sales pitch score at 1. |
| 5. | This shows a need for improvement in this area. |

#### **Fig 16 :Own Car**

### 

***Observation***

|  |  |
| --- | --- |
| **1.** | 62 % of the Customers have their own car. |

#### **Fig 17 :Designation**

### 

***Observation***

|  |  |
| --- | --- |
| 1. | Executive (37.7%) and Manager(35.4%) are the highest Designations of the customers in the dataset. |

#### **Fig 18 :Monthly Income**

### 

***Observation***

|  |  |
| --- | --- |
| 1. | MonthlyIncome is positively-skewed |
| 2. | The majority of customers lie between the income bracket 20K and 3oK rupees. |
| 3. | We also observe two outliers in the low end and above the upper whisker. |

#### **Fig 18 :Monthly Income**

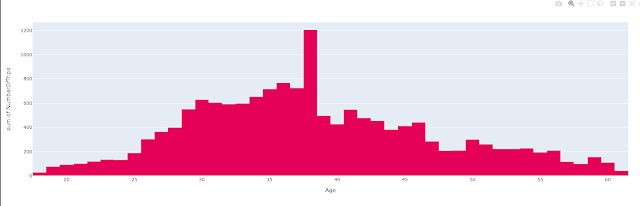
### 

***Observation***

|  |  |
| --- | --- |
| 1. | 57.7% of the customers don't have insurance with them i.e around 2820 customers. |

### **Bivariate Analysis**

#### **Fig 1:Age vs Number of Trips**

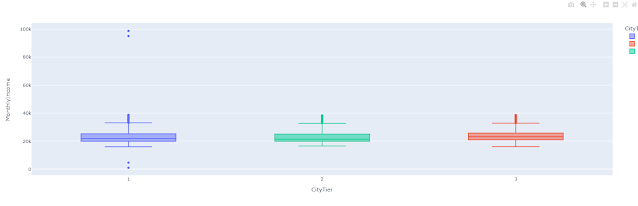
[](https://www.blogger.com/)

***Observation***

#### 1. We can see that it is a normal distribution and the highest is the customers at age 38

#### 

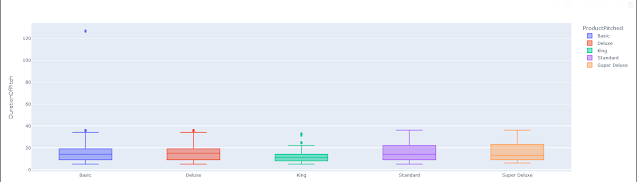
#### **Fig 2 :City Tier vs Monthly Income**

[](https://www.blogger.com/)

***Observations***

1. The Monthly income of customers in all the City Tier range from 15k to 35k. However it has some outliers where the income is below 15k and above 35k

#### **Fig 3: Product Pitched vs Duration Of Pitch**

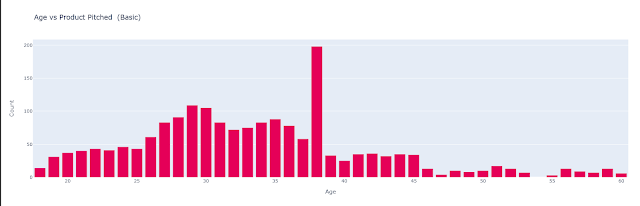
[](https://www.blogger.com/)

***Observations :***

1. Median for all the packages offered ranges with [11,14].
2. Duration of pitch for Basic ,Deluxe ,King witnessed few outliers which are not really the outliers as they are very close to the upper whisker.

#### 

#### **Fig 4: Age vs Product Pitched (Basic)**

[](https://www.blogger.com/)

***Observation***

1. Basic product is primarily aimed at individuals aged between 20 and 40, with 38 being the highest among these.

#### **Fig 5 :Age vs Product Pitched (Deluxe)**

[A graph of a number of red bars

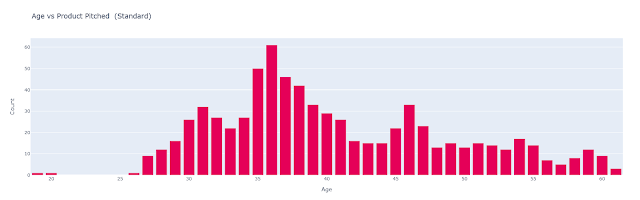
Description automatically generated with medium confidence](https://www.blogger.com/)

***Observation***

1. Deluxe Product is not pitched to customer's below the Age of 25

 2. Deluxe product is primarily aimed at individuals aged between 25 and 60, with 36 being the highest among these.

#### **Fig 6:Age vs Product Pitched (Standard)**

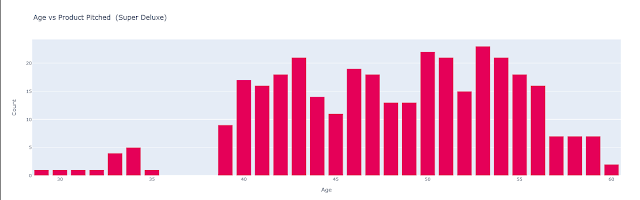
[](https://www.blogger.com/)

***Observations***

1. Standard Product is not pitched to customer's below the Age of 25

2. Standard product is primarily aimed at individuals aged between 25 and 60, with 36 being the highest among these.

**Fig 7: Age vs Product Pitched (Super Deluxe)**

[](https://www.blogger.com/)

***Observations***

1. Super Deluxe Product is not pitched to customer's aged between 35 and 39

 2. Super Deluxe product is primarily aimed at individuals aged between 40 and 60, with 53 being the highest among these.

#### **Fig 8 :Age vs Product Pitched (King)**

[A graph with red bars

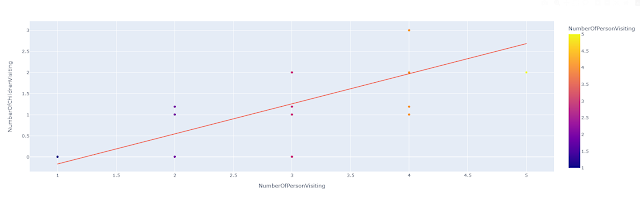
Description automatically generated with medium confidence](https://www.blogger.com/)

***Observations :***

1. King Product is not pitched to customer's aged below 39

2. King product is primarily aimed at individuals aged between 40 and 60, with 50 being the highest among these.

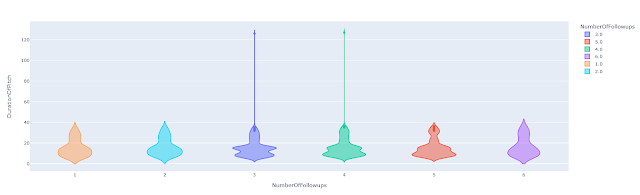
#### **Fig 9 :Number Of Person Visiting vs Number Of Children Visiting**

[](https://www.blogger.com/)

***Observation:***

1. There exists a positive correlation between the number of people visiting vs Number of children.

#### **Fig 10 : Number Of Followups vs Duration Of Pitch**

[[](https://www.blogger.com/)](https://www.blogger.com/)

***Observations :***

1. For all the Number of followups , median of Duration of pitch was ranging from 13 to 14 mins. 2. In Number of Followups 3 and 4 we can observe few outliers.

### **Distribution Plots**

#### **Fig 1 : Age Distribution**

[A pink line on a white background

Description automatically generated](https://www.blogger.com/)

***Observations:***

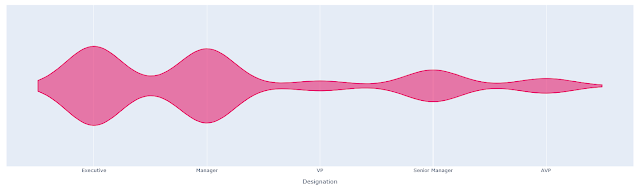
1. 31% of the data (age) is lying below q1.

2. The median age of the customer is 37.

3. No such outliers in the age

4. Maximum probability density is associated with kde:0.998

#### **Fig 2 :Designation Distribution**

[](https://www.blogger.com/)

Observations:

1. For the Executive we observed the maximum probability density of 0.999.

2. In case of manager we saw maximum probability density of 0.94

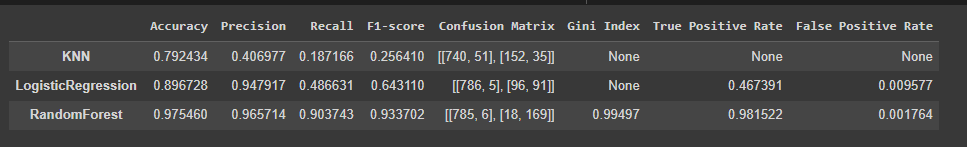
3. For the VP's and AVP's its 0.125 and 0.185 respectively. Means, Very Less customers are designated as VP's.

4. Maximum probability density is assocated with kde:0.403 for the senior manager.

**Section 4: Model Deployment Techniques**

We have trained and tested our dataset using KNN,Random Forest and Logistic regression classifiers . We splitted the data into 80% training and 20% testing .

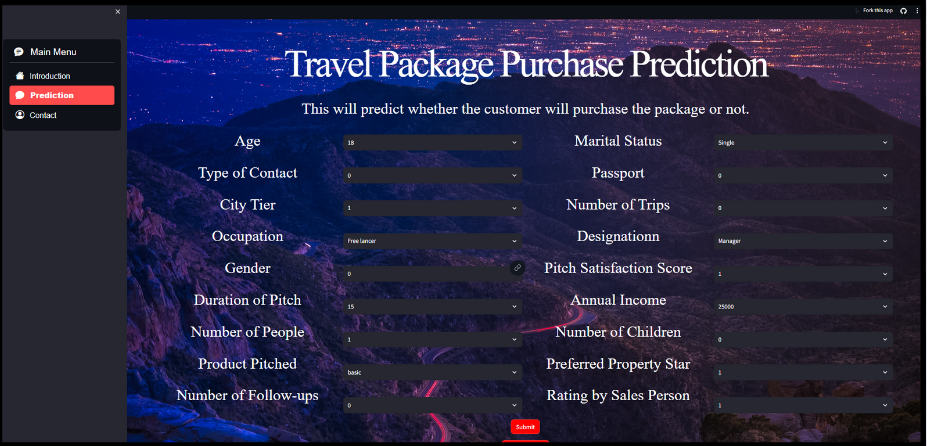
1. **KNN Classification :** for KNN modeling we used 5 nearest neighbors into consideration.
2. **Random Forest :** for performing Random forest classification we used 100 n\_estimators into consideration.
3. **Logistic Regression:** This model gave us the accuracy of 0.896728 with a false positive rate of 0.009577.

These are the respective classification report for the Model Analysis:

Finally,

We deployed the classification model using streamlit where we have also provided the provision for the company to enter the details for the new customer as well .

Deployed website appears to be like ,



To access the website, the link is provided below

<https://travelpackagepurchasepredictor.streamlit.app/>

GitHub link: <https://github.com/riya0785/mini_project_TravelPackagePurchaseprediction>

**Reference**

**[1]** <https://github.com/foos0016/Travel-Package-Purchase-Prediction>

**[2]** <https://www.kaggle.com/code/yogidsba/travelpackageprediction-ensemble-techniques>