A PROJECT ON

"AIRLINE PASSENGER SATISFACTION PREDICTION"

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
PG DIPLOMA IN BIG DATA ANALYTICS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY
"Sunbeam IT Park", Phase 2 of
Rajiv Gandhi Infotech Park,
Hinjawadi, Pune - 411057,
MH-INDIA

SUBMITTED BY:

Gaurav Rajpurohit (75818)

Sanket Dongare (75767)

UNDER THE GUIDENCE OF:

Mrs. Manisha H
Faculty Member
Sunbeam Institute of Information Technology, PUNE.

CERTIFICATE

This is to certify that the project work under the title 'Airline Passenger Service Satisfaction' is done by **Sanket Dongare** & **Gaurav Rajpurohit** in partial fulfillment of the requirement for the award of Diploma in Big Data Analytics Course.

Mrs. Manisha Hingane
Project Guide

Mrs. Pradnya Dindorkar
Course Co-ordinator

Date:

ACKNOWLEDGEMENT

A project usually falls short of its expectations unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Pradnya Dindorikar (Course Coordinator, SIIT, Pune) and Project Guide Mrs. Manisha Hingane.

We are deeply indebted and grateful to them for their guidance, encouragement, and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Sanket Dongare
DBDA March-2023 Batch,
SIIT Pune

Gaurav Rajpurohit
DBDA March-2023 Batch,
SIIT Pune

.

TABLE OF CONTENTS

1.Introduction

1.1 Introduction And Objectives:

Develop a machine learning model to predict Airline passenger Service Satisfaction based on a dataset of passenger feedback and flight-related information. The goal is to understand the factors that influence passenger satisfaction and provide airlines with insights (with help of Data Visualisation tool Tableu) to enhance their services. The data is provided on daily basis. Here we are owing to Analyze the given dataset.

1.2 Why this problem needs to be Solved?

Data analysis is essential for Airlines to enhance their operational efficiency, improve customer satisfaction, ensure safety, optimize revenue and profits, taking business related decision to cater more customers, and maintain a competitive edge in a dynamic and highly regulated industry. Airlines that leverage data effectively are better positioned to thrive in a complex and competitive market. Airlines that effectively harness data analysis gain a competitive advantage. They can respond quickly to changing market conditions, adapt to customer demands, and make data-driven decisions that improve their overall performance.

1.3 Dataset Information

Airline.csv:

It has 24 columns.
Gender: Gender of the passengers (Female, Male)
Customer Type: The customer type (Loyal customer, Disloyal customer)
Age: The actual age of the passengers
Type of Travel: Purpose of travel of passengers (Personal Travel, Business Travel)
Class: Travel class in the plane of the passengers (Business, Economy)
Flight distance: The flight distance of this journey
Ease of Online booking: Satisfaction level of online booking
Gate location: Satisfaction level of Gate location
Food and drink: Satisfaction level of Food and drink
Online boarding: Satisfaction level of online boarding
Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment
Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient.
On-board service: Satisfaction level of On-board service
Legroom service: Satisfaction level of Leg room service
Baggage handling: Satisfaction level of baggage handling
Check-in service: Satisfaction level of Check-in service
Inflight service: Satisfaction level of inflight service
Cleanliness: Satisfaction level of Cleanliness.
Departure Delay in Minutes: Minutes delayed when departed
Arrival Delay in Minutes: Minutes delayed when Arrived
Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)
Cleanliness: Satisfaction level of Cleanliness.

Train.csv

It has 12 columns. Store:

Age: The actual age of the passengers
Type of Travel: Purpose of travel of passengers (Personal Travel, Business Travel)
Class: Travel class in the plane of the passengers (Business, Economy)
Flight distance: The flight distance of this journey
Ease of Online booking: Satisfaction level of online booking
Online boarding: Satisfaction level of online boarding
Seat comfort: Satisfaction level of Seat comfort
Inflight entertainment: Satisfaction level of inflight entertainment
On-board service: Satisfaction level of On-board service
Legroom service: Satisfaction level of Leg room service
Baggage handling: Satisfaction level of baggage handling

Test.csv: is the same as train.csv except it does not have ‘Satisfaction’ Column.

2. Problem Definition and Algorithm:

2.1 Problem Definition

The objective of this machine learning project is to develop a predictive model that accurately assesses and predicts passenger service satisfaction for an airline. The project aims to understand the factors influencing passenger satisfaction, enable airlines to proactively address service issues, and ultimately enhance the overall passenger experience.

The success of this project will be determined by:

- The model's ability to accurately predict passenger service satisfaction.
- The Identification of key factors influencing passenger satisfaction.
- The practicality and feasibility of using model predictions to enhance the airline's services and improve passenger satisfaction.

2.2 Algorithm Definition

Logistic Regression: Logistic regression is a statistical model used for binary classification tasks. It predicts the probability of an event occurring, typically "1" (positive class), based on one or more predictor variables. It's widely applied in fields like medicine and marketing. The model uses a logistic function to constrain the output between 0 and 1, making it suitable for probability estimation. Parameters are estimated using maximum likelihood estimation. Despite its simplicity, logistic regression can provide valuable insights and is a foundation for more complex models like neural networks. It's interpretable and can handle both continuous and categorical input variables.

Decision Tree: A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It models decisions as a tree-like structure, where each internal node represents a feature or attribute, and each leaf node represents a class label or numerical value. The tree is constructed by recursively splitting the data based on features to minimize impurity or maximize information gain. Decision trees are easy to understand and visualize, making them valuable for interpreting model decisions. However, they can be prone to overfitting, which can be mitigated using techniques like pruning.

Random Forest: Random Forest is a powerful machine learning ensemble method. It constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Each tree is trained on a random subset of data and features. By aggregating the results, Random Forest provides robust and interpretable models for classification and regression tasks. It's known for handling high-dimensional data and maintaining model generalization while offering feature importance insights.

Support Vector: Support Vector Machine (SVM) is a supervised machine learning algorithm for classification. It finds a hyperplane that best separates data into different classes while maximizing the margin between them. SVM is effective in high-dimensional spaces and can handle linear and non-linear data separation using various kernel functions. It aims to achieve good generalization and is resistant to overfitting. SVM is widely used in image classification, text categorization, and bioinformatics due to its versatility and strong theoretical foundations.

3.Experimental Evaluation:

3.1 Methodology:

Loading data into Notebook:

```
df =pd.read_csv('airline_passenger_satisfaction.csv')
print(df)
print(df.head())
```

Preprocessing:

Filling the null values with appropriate values:

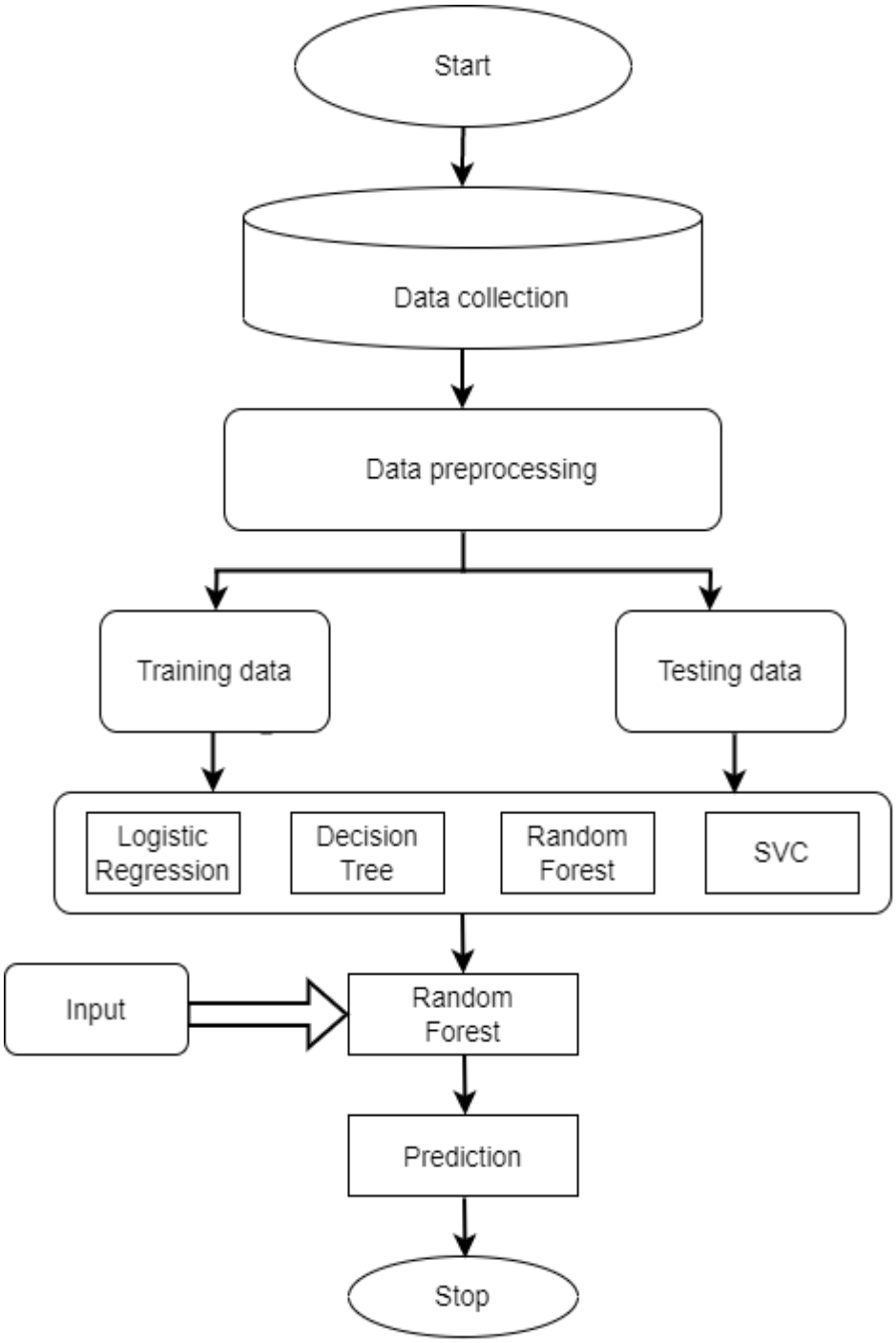
```
df['age']=df['age'].fillna(df['age'].mean())

for val in df.columns:
    if df[val].isna().sum()==6:
        df[val]=df[val].fillna(df[val].mode()[0])
    else:
        df[val]=df[val].fillna(df[val].median())
```

Outlier treatment:

```
sns.boxplot(df['flight_distance'],orient="h")
plt.show()
Q1 = df['flight_distance'].quantile(0.25)
Q2 = df['flight_distance'].quantile(0.5)
Q3 = df['flight_distance'].quantile(0.75)
print(Q1,Q2,Q3)
IQR = Q3 - Q1
Print(IQR)
upper_bound = Q3 + (1.5*IQR)
lower_bound = Q1 - (1.5*IQR)
upper_bound , lower_bound
total_outlier_count = df['flight_distance'] >
upper_bound).sum()
total_outlier_count
index = df['flight_distance'] > upper_bound
df.loc[index, 'flight_distance'] = upper_bound
```

Project Flow Diagram :



3.2 Exploratory Data Analysis

- The Percentage of Customers Class travelling from Airline travel Categorised into Business class and Economy Class can be visualized by the Pie Chart which Shows that 52% comes in Economy Class while 48% comes in Business Class

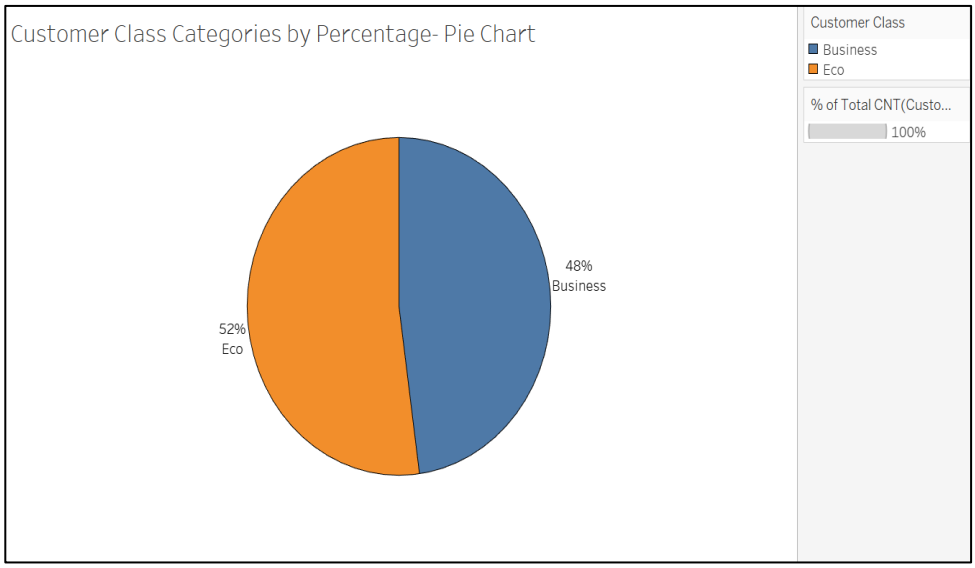


Fig 1: Pie- chart showing Customer Class Category Percentage

- Count of customers from Loyal and Disloyal Category from Customer Type is shown below with the help of Bar Chart.

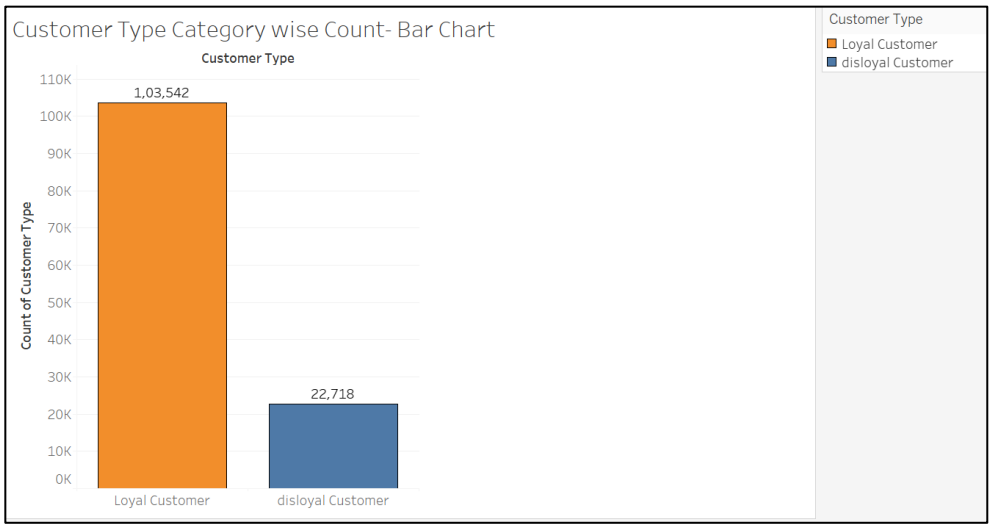


Fig 2: Customer Type- Loyal and Disloyal Customer Count with Bar Chart

- Gender Ratio of all Customers travelling from Airline is plotted below with Pie Chart.

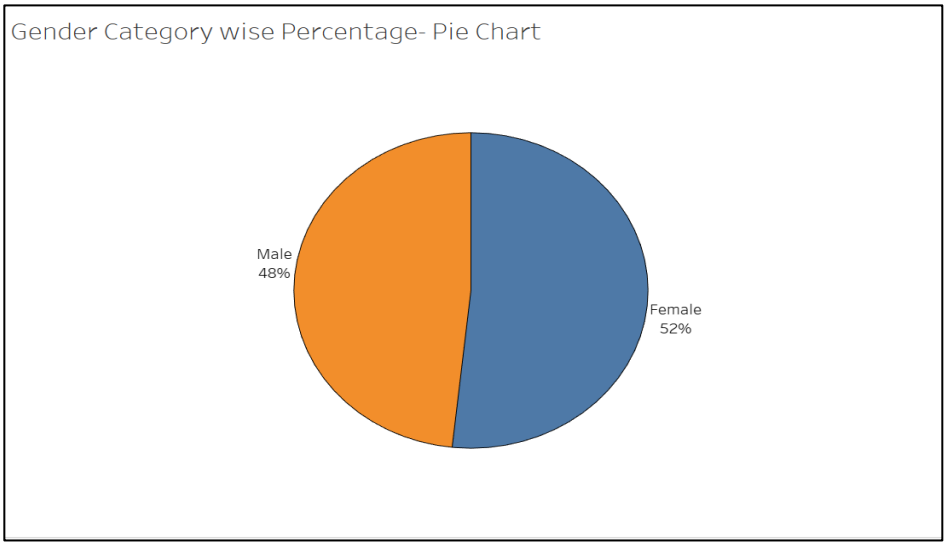


Fig 3: Customer Gender Ratio Percentage Pie Chart

- Based on the All services provided by Airline, People have given Feedback as whether they are Satisfied or Not. This pie Chart Shows the Percentage of People Satisfied with overall Percentage.

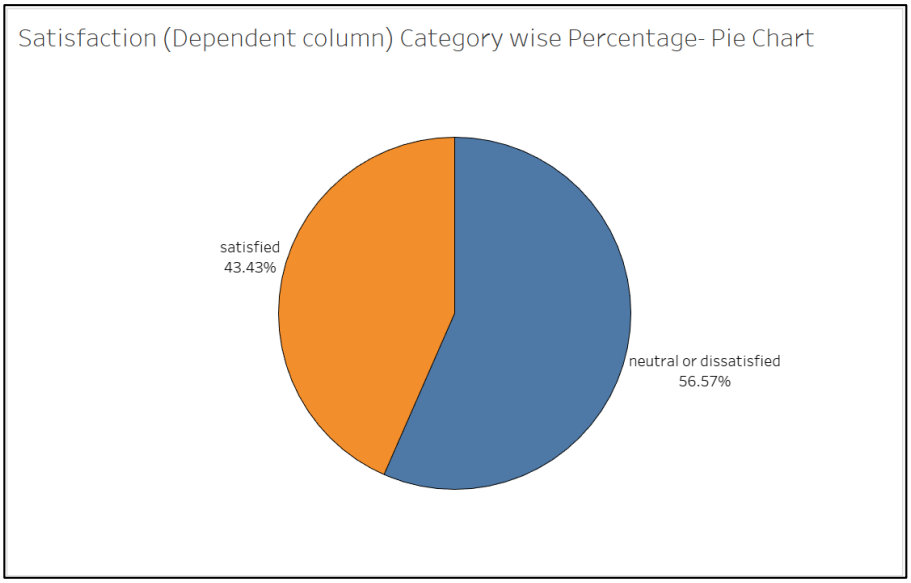


Fig 4: Percentage of people Satisfied or Not Satisfied- Pie Chart

- Age wise Customers count plot on Histogram considering their Satisfaction level. The Majority of Customers belong to age group of 20 - 60 year travelling through Airline.

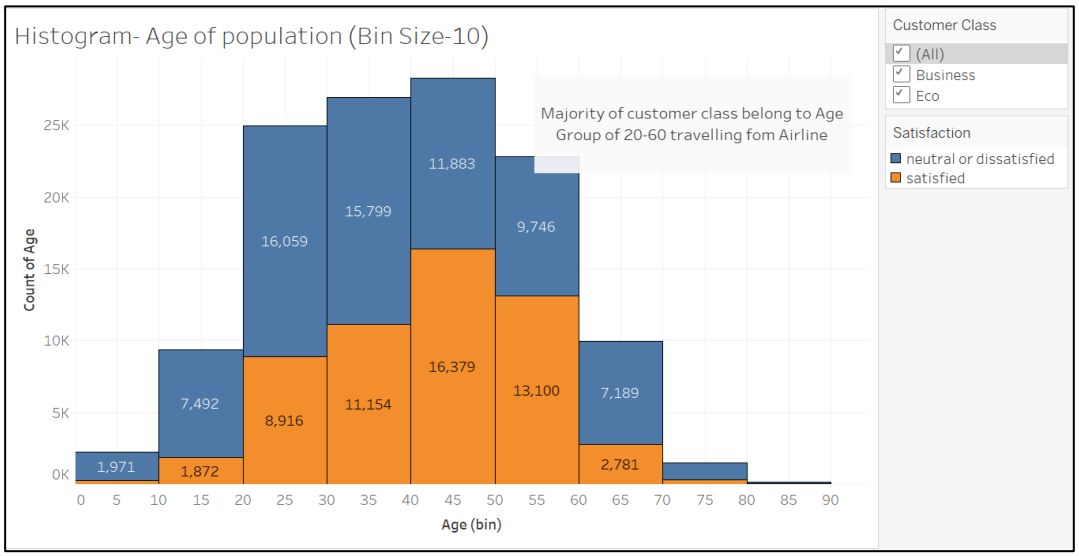


Fig 5: Histogram of Age of Customers- Considering their Satisfaction Count

- The Histogram below Shows count of Customers of different Age groups travelled through Airline. From this we can conclude that the Majority of Customers travelling from Economy class are not Satisfied with the Services Provided.

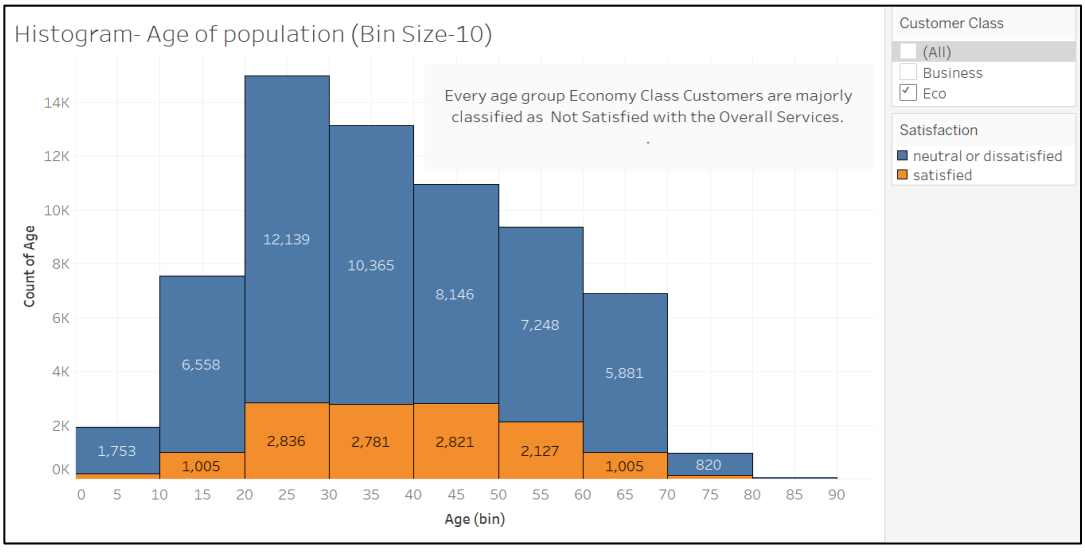


Fig 6: Age wise Economy Class Customers Count of Satisfaction – Histogram

- Baggage Handling is One of the Services provided by Airline. The following Bar chart provides the details of ratings given by Customers of Economy Class to Baggage Handling Service. From chart we can Conclude that, almost 50% of customers rated it either 3 or below out of 5.

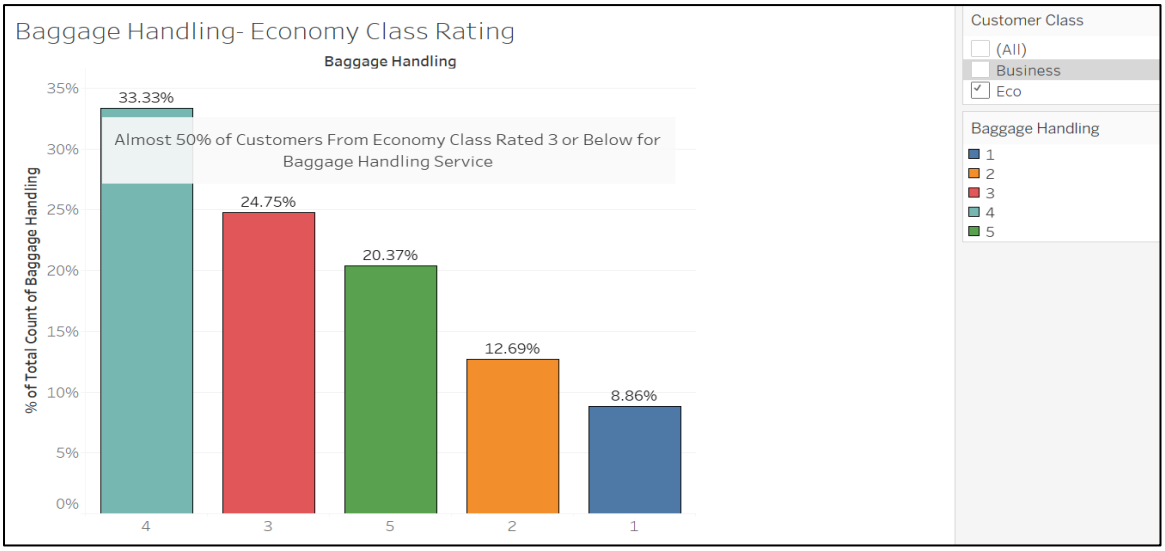


Fig 7: Rating for Baggage Handling by Economy Class Customers- Bar Chart

- On Contrary, 70% customers from Business class rated the Baggage Handling 4 or more out of 5.

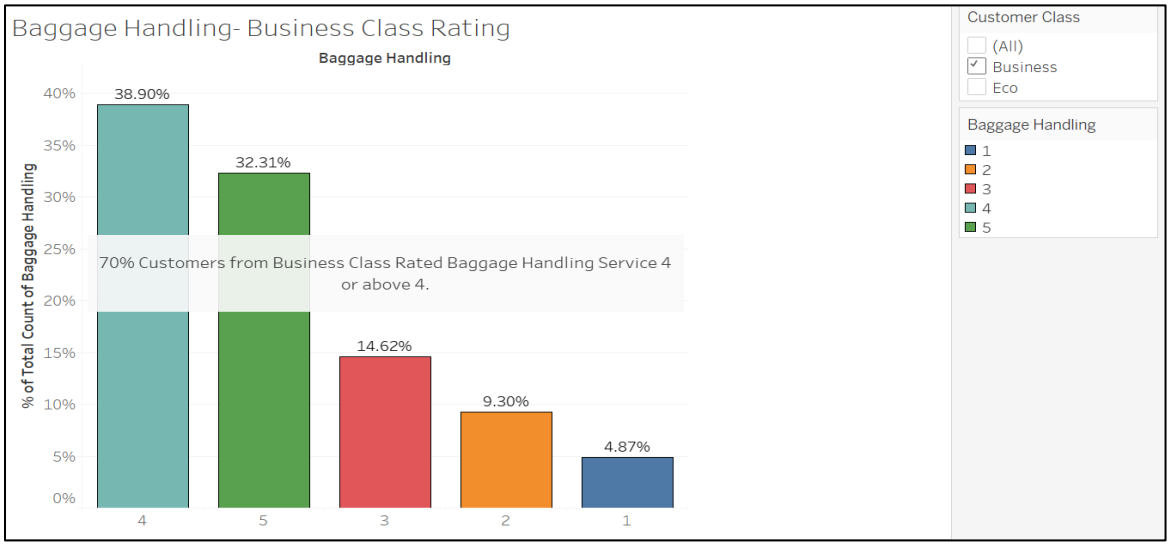


Fig 8: Rating for Baggage Handling by Business Class Customers- Bar Chart

- Cleanliness: The 60% of people who reviewed as Not Satisfied with overall Services have rated the Cleanliness at 3 or below 3 out of 5.

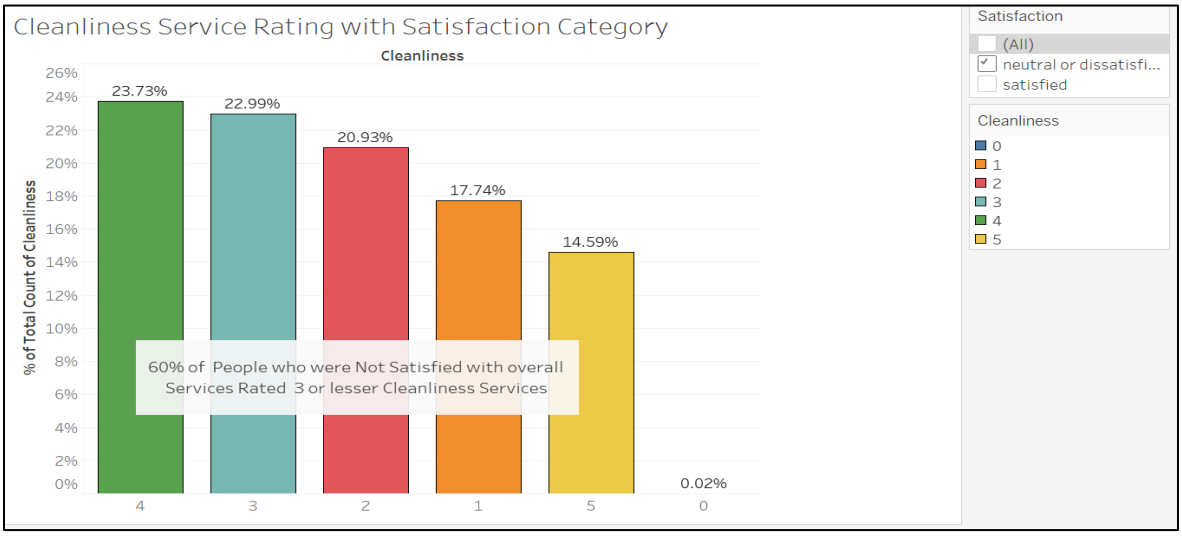


Fig 9: Rating for Cleanliness by Customers who were Not Satisfied with overall Services- Bar Chart

- Leg Room Service: Only 30% people from Business class rated Leg Room Service 4 or 5 out of 5 who were not Satisfied with overall Services.

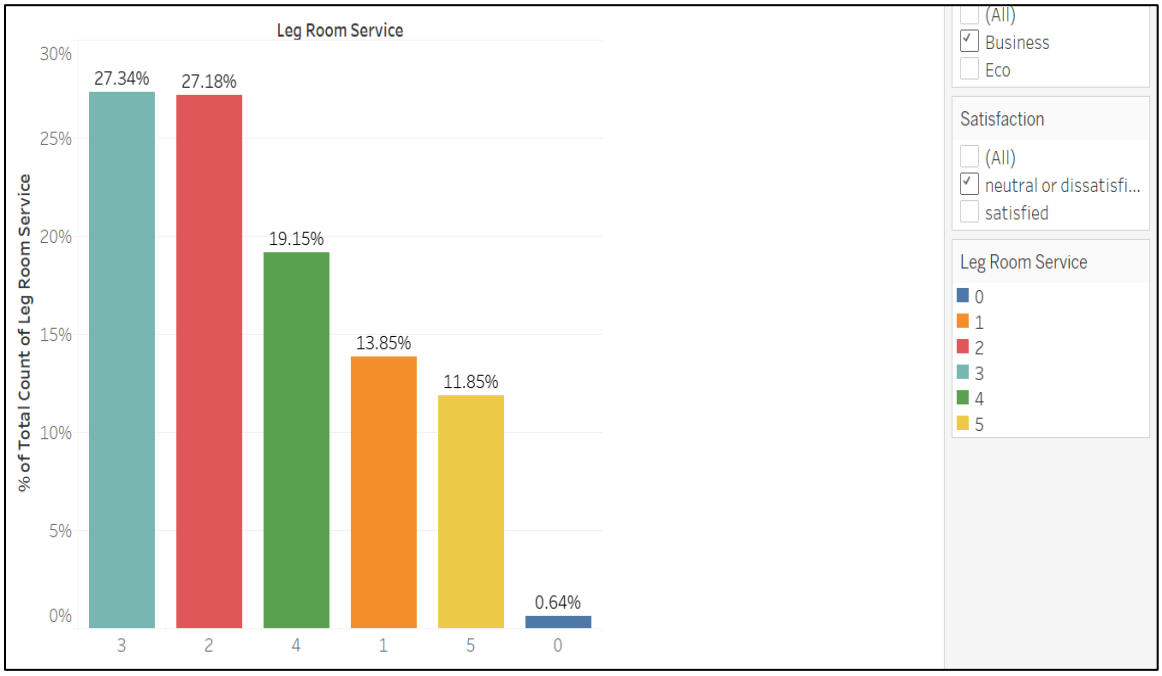


Fig 10: Rating for Leg Room Service by Customers who were Not Satisfied with overall Services- Bar Chart

- Only 40% people from Business class rated Arrival and Departure Time Convenience 4 or 5 out of 5.

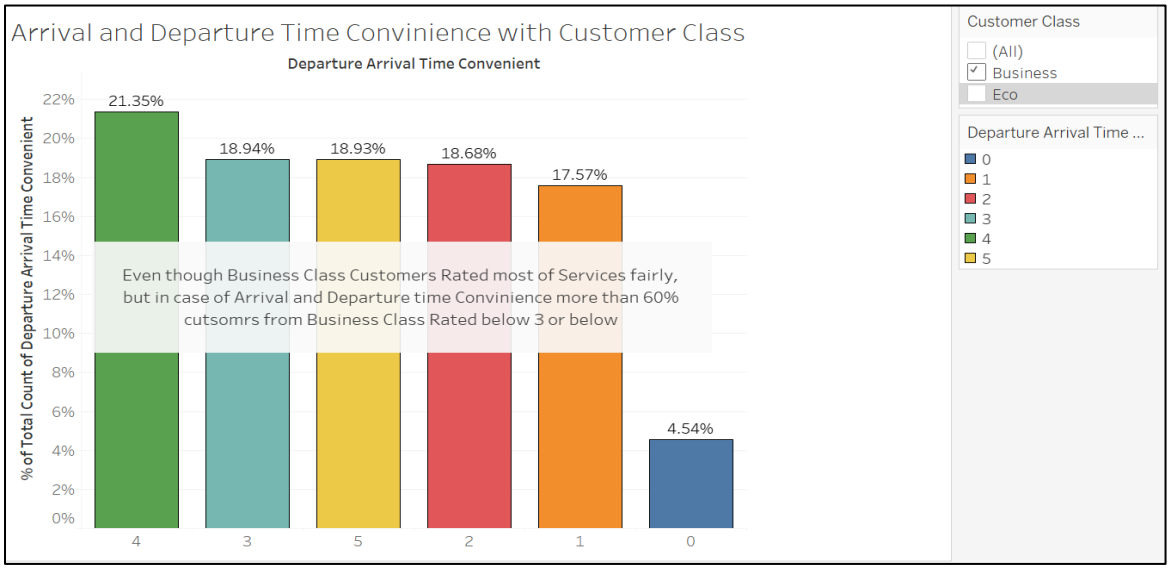


Fig 11: Arrival and Departure Time Convenience Bar Chart for Business Class Customers

- 60% people from Economy class rated Inflight Entertainment as 3 or below out of 5.

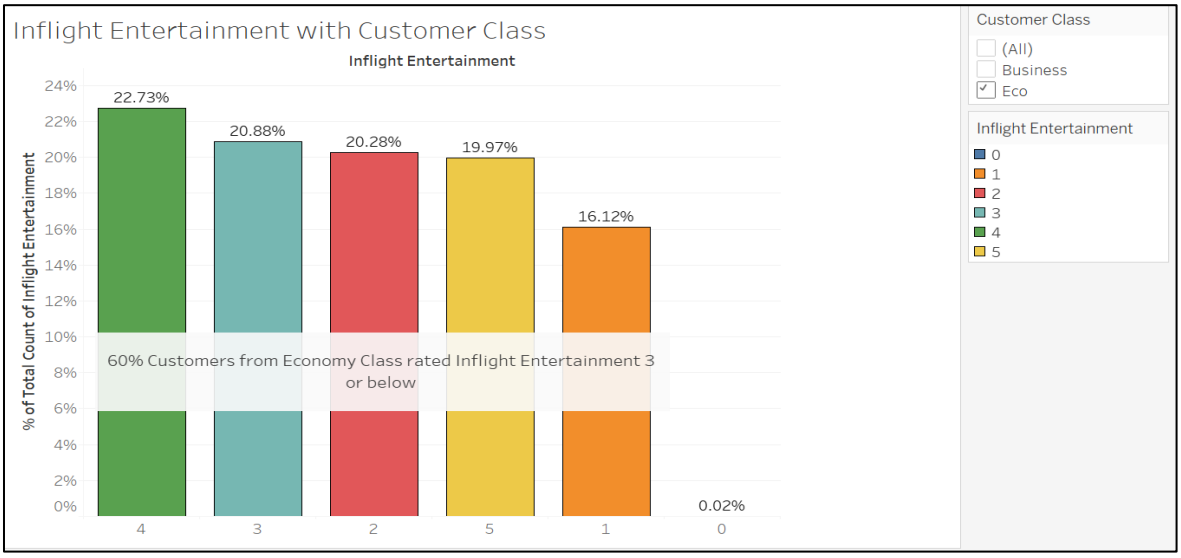


Fig 12: Ratings for Inflight Entertainment by Economy Class Customers- Bar Chart

- The Food and Drink Category was rated average by both Loyal and Disloyal customers.

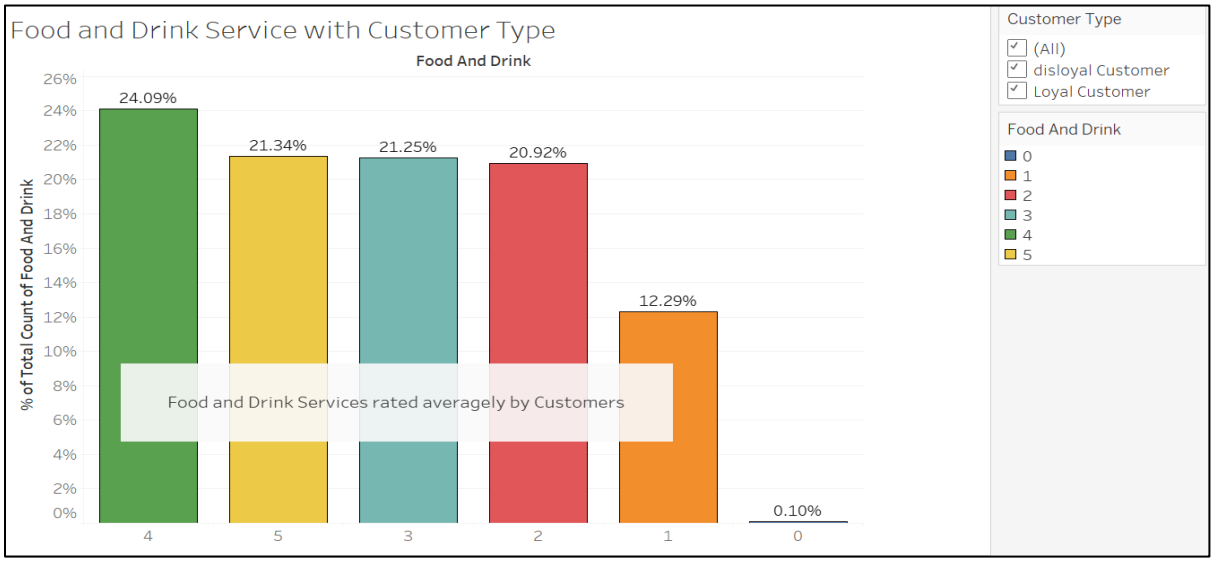


Fig 13: Rating for Food and Drink Service by Customer Type- Bar Chart

4. Results and discussion:

Following are the readings received for Different Algorithms on the Dataset
From the all algorithms, we have selected Random Forest for our model as it is the algorithm with most Accuracy, Precision, Recall and F1 amongst all other Algorithms.

Algo	Accuracy	Precision	Recall	F1
LG	0.86	0.85	0.82	0.83
SVM	0.66	0.66	0.46	0.54
DT	0.90	0.92	0.85	0.88
RF	0.92	0.91	0.90	0.91
GB	0.93	0.92	0.91	0.92
Ada	0.91	0.91	0.89	0.90

5. GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

Sample Templates of GUI-

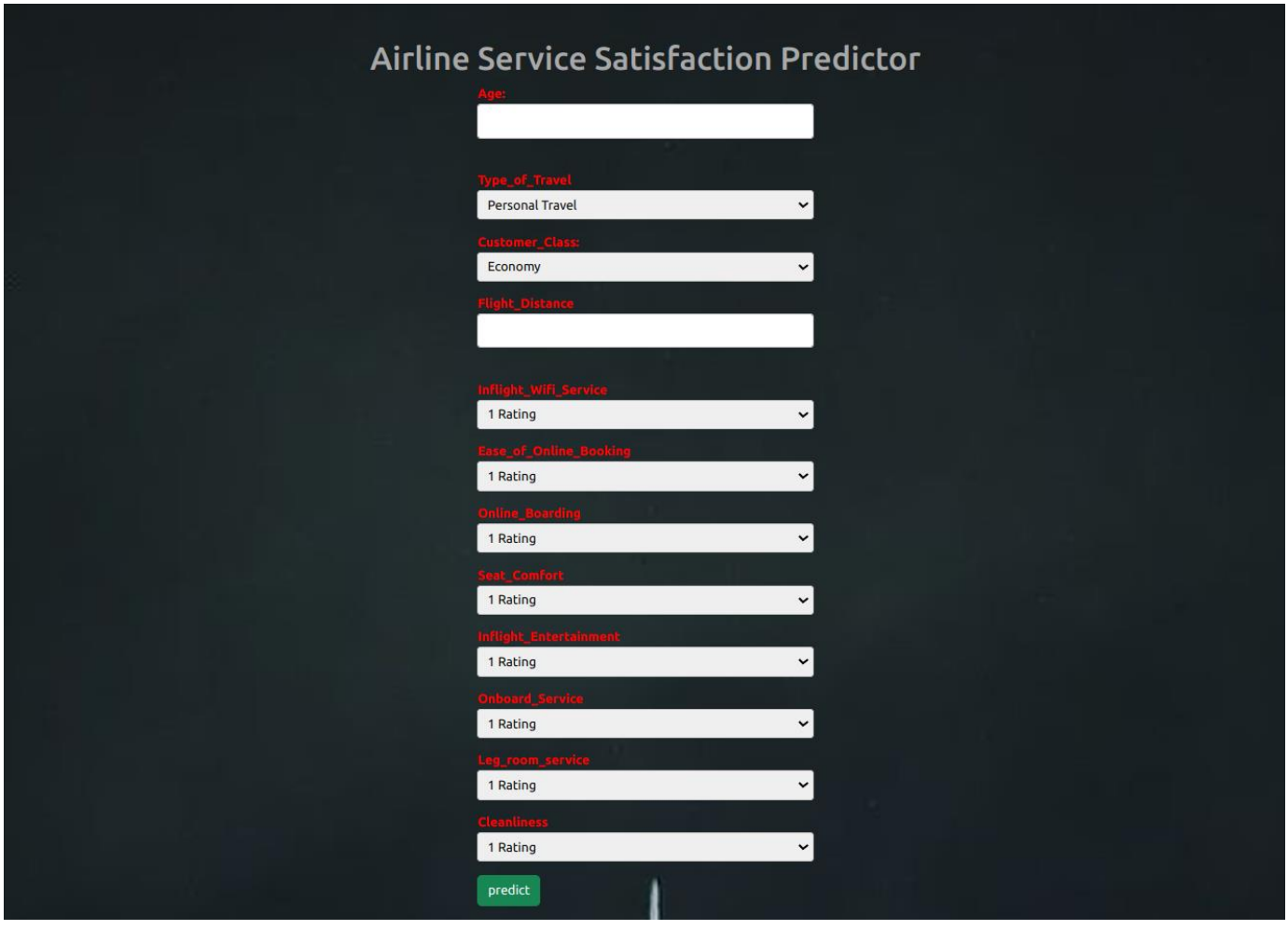


Fig- Input Page Template

Fig- Result Page Template

6. Future work And Conclusion:

6.1 Future Work:

1. **Incorporate More Data Sources**: Expand the dataset by incorporating additional sources of data, such as social media sentiment analysis, weather conditions, and passenger demographics. This can provide a more comprehensive view of factors affecting customer satisfaction.
2. **Deep Learning Models**: Experiment with more advanced machine learning models, including deep learning techniques such as neural networks and recurrent neural networks (RNNs), to capture complex patterns and dependencies in customer satisfaction data.

6.2 Conclusion:

In conclusion, our machine learning project focused on predicting airline customer satisfaction has successfully demonstrated the ability to forecast customer satisfaction levels with a reasonable degree of accuracy. By analyzing a combination of factors such as flight punctuality, in-flight services, seat comfort, and customer feedback, we have developed a predictive model that can help airlines identify areas for improvement and enhance overall passenger experiences. This predictive tool can enable airlines to proactively address customer concerns, leading to increased satisfaction, loyalty, and ultimately, improved business outcomes. However, it's essential to note that customer satisfaction is influenced by various complex factors, and continuous data collection and model refinement are necessary to maintain predictive accuracy over time.

- Sales are also dependent on the department of the store as different departments showed different levels of weekly sales

- Among the trained models for predicting the future sales, Gradient Boosting Machine performs the best.