Project Report

On

ANALYSIS AND PREDICTION OF AIRLINE PASSENGER SATISFACTION

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

ASHITHA E

(Register No. SM21AS004)

(2021-2023)

Under the Supervision of

ARUNIMA P S

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

APRIL 2022

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **ANALYSIS AND PREDICTION OF AIRLINE PASSENGER SATISFACTION** is a bonafide record of the work done by Ms. **ASHITHA E** under my guidance as partial fulfillment of the award of the degree of **Master of Science in APPLIED STATISTICS AND DATA ANALYTICS** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:
Place: Ernakulam

Arunima P S
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

ST. TERESA'S COLLEGE DEPT. OF MATHEMATICS ERNAKULAM

Smt. Betty Joseph
Associate Professor & HOD,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1: Dr. Joseph Justin Pothilla 18/5/23

2: SARI THOMAS

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **ARUNIMA P S**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam.

Date: 18/05/23

ASHITHA E

SM21AS004

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry this work. Their continuous invaluable knowledgably guidance throughout the course of this study helped me to complete the work up to this stage

I am very grateful to my project guide (Arunima P S) for the immense help during the period of work

In addition, very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to faculty, teaching and non-teaching staff of the department and Colleagues.

I also very thankful to HoD for their valuable suggestions, critical examination of work during the progress.

ASHITHA E
SM21AS004:

Ernakulam.
Date: 18/05/23

ABSTRACT

With the continuous improvement of people's living standards, airline industries are growing, and people have put forward higher requirements for aviation service quality. This paper analyses the airline Passenger satisfaction data to determine the factors that highly influencing passenger satisfaction, to fit a regression and machine learning models, and to compare them and predict with the best model. The Data for the investigation has been collected from Kaggle.

The findings of the study show that there are five most important factors affecting the satisfaction of airline passengers. For fitting a regression and machine learning models, Logistic regression, K Nearest Neighbor classification, support vector machine classification, Decision tree classification, and Random forest classification are taken into consideration and it is found that the random forest has the highest accuracy about 95.44%, so the best model to predict customer satisfaction for the data is random forest, and prediction using random forest is done. The result based on this study will help airlines better serve the passengers by understanding the factors influencing passengers satisfaction.

ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM

Certificate of Plagiarism Check for Thesis

Author Name	Ashitha E
Course of Study	MSc. Applied Statistics & Data Analytics
Name of Guide	Ms. Arunima P S
Department	Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	Analysis and prediction of Airline passengers satisfaction
Similarity	3%
Paper ID	724318
Submission Date	2023-04-19 10:04:12

Signature of Student

Signature of Guide

Checked By
College Librarian

* This report has been generated by DrillBit Anti-Plagiarism Software

Contents

CERTIFICATE ii
DECLARATION iii
ACKNOWLEDGEMENTS iv
ABSTRACT v
CONTENT v

1 Introduction 2
1.1 Airline Passenger Satisfaction 2
1.2 Importance of Airline Passenger Satisfaction Analysis . . 3
1.3 Objectives 4

2 Literature Review 5

3 Methodology 8
3.1 Data 8
3.2 Exploratory Data Analysis 8
3.3 Factor Analysis 8
3.4 Model Building 9
3.4.1 Logistic Regression 9
3.4.2 k-Nearest Neighbor 9
3.4.3 Support Vector Machine. 9
3.4.4 Decision Tree 10
3.4.5 Random forest model 10

4 Dataset And Exploratory Data Analysis 11
4.1 Priliminaries 11
4.1.1 Data Sample 12

4.1.2 Pre-processing 12
4.1.3 Descriptive Statistics 12
4.2 Exploratory Data Analysis 13
4.2.1 Univariate Visualisations 13
4.2.2 Multivariate Visualisatons 16

5 Result And Discussions 18
5.1 Factor Analysis 18
5.2 Logistic Regression 20
5.3 Model Fitting 21
5.4 Comparing The Models Predicting With The Best Model 22
5.4.1 Comparing The Models 22

6 Conclusion 24

7 Reference 25

Chapter 1

Introduction

Now a days, transportation services have become one of the basic needs of the community for both daily activities and travel needs. For a long distance journey, aeroplanes have always been one of the primary choices for people to travel because of their convenience and safety. The global aviation industry is gaining more and more weight, which in turn significantly adds to national economies, since it plays an important role in moving individuals and products locally or internationally. Air transport can reach places that cannot be reached by other modes of transport such as land and sea; in addition, it is able to move faster and have a straight, practically bare-free path, which is also selected for its efficiency and effectiveness of time.

1.1 Airline Passenger Satisfaction

In any buisiness field customer satisfaction is increasingly recognized as a determinant of business performance and a strategic tool for gaining a competitive advantage. Richard N Cardozo first put forward the viewpoint of customer satisfaction in 1965, and was introduced into the field of marketing for the first time [1] and the theory of customer satisfaction also made continuous development possible.

Satisfaction is not only considered as a customer goal to be derived as a result of degrading services, but is also a company's goal, since it is a major way of getting higher customer retention rates and ways of generating profits. If the services are provided in accordance with

customer's expectations, then the satisfaction level of the customer will be high and increase the level of customer loyalty, and conversely, if the service quality is not up to the expectations of the customer, service quality will be considered as bad, which in turn causes the decrease of customer loyalty. Customer loyalty is considered as commitment or customer principle to always choose the service/product continuous and consistent in the future.

Research has revealed that the aspects of the overall service offered by airlines to passengers are crucial determinant for passenger's decisions regarding the choice of airlines. According to Adam & John, airline managers are expected to comprehend the value placed by passengers set on different aspects of the service quality mix, as these choices will directly affect the airline's strategic positioning.

1.2 Importance of Airline Passenger Satisfaction Analysis

It is safe to say that customer service is one of the biggest factors for service-oriented industries to increase their revenues; therefore, it comes as no surprise that customer satisfaction is given a high priority. Today the global aviation industries are highly competitive, which causes airline companies to aim for top positions with a heavy significance on customer satisfaction. Customer satisfaction can add value to the airline company, helps to increase the company's revenue, increase customer loyalty of the company, etc. so the study on customer satisfaction has a vital role in aviation industry since it can improve service quality and to know the pulse of customers/passengers like what they need in their travelling mean time which increases the satisfaction level. This definitely improves the airline's business.

1.3 Objectives

- **To find the factors that highly contribute to the variable satisfaction**
- **To fit a regression and ML models to the given data**
- **By comparing the models and predicting with most accurate model.**

Chapter 2

Literature Review

- **Adarsh Anand and Gunjan Bansal in the paper 'Predicting Customer's Satisfaction Using Logistic Regression' they provide a predictive model to identify customer's satisfaction. For this study, used mobile phones as a product, and due to dichotomous nature of the dependent variable Logistic Regression has been applied for the validation**

- **In the paper 'Applying Factor Analysis in Assessing Customer Satisfaction with Service Quality: The Case of Indian Banks' by Mustafa Raza Rabbani and Zehra Zulfikar by using the Experience Survey of 568 respondents from SBI and ICICI Bank in the National Capital Region of Delhi ,from the analysis, out of 13 factors given, the four major factors have been identified, namely, "Work culture and timeliness", "Basic amenities and accessibility factor", "Procedural complexities" and "Location and cost factor which are highly influencing the respondents towards satisfaction and help the customers choose the bank.**

- **In the paper 'A machine learning approach to analyze customer satisfaction from airline tweets' by Sachin Kumar and Mikhail Zymbler, used machine learning models such as Support Vector Machine (SVM), an Artificial Neural Network and a convolutional neural network (CNN) is trained on the data and its performance were compared with the best model among SVM and ANN models and**

the results show that CNN outperformed all other models in terms of accuracy and performance.

- **In the paper 'A Logistic Regression Model of Customer Satisfaction for Online Food Delivery Services' by Dr. Kali Charan Modak and Dr. Kumkum Sinha, concluded that out of the eight independent variables four variables are negatively related to satisfaction of customers of online food delivery services. Better discount is the independent variable that makes most customers satisfied compared to others Data on 100 consumers was collected from Indore region. They used SPSS.**

- **In the paper Customer 'Analysis Using Machine Learning Algorithms: A Case Study Using Banking Consumer Dataset' by R.Siva Subramanian, D.Prabha,B.Maheswari , and J.Aswini, used NB, KNN, LR, c45, Random Forest The empirical outcomes indicate C4.5 model ends with 89.8786% higher accuracy compared to other approaches.**

- **Moulay Smail Bouzakraui, Abdelalim Sadiq, Abdessamad Youssfi Alaoui in the paper 'Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques', they proposed a new method for facial emotion detection to recognise customer's satisfaction using machine learning techniques. After using SVM, KNN, Random Forest, Adaboost, and Decision Tree they got an 98.66% as accuracy for the most performance SVM classifier**

- **Jiong Mu, Lijia Xu, Xuliang Duan, and Haibo Pu in the paper 'Study on Customer Loyalty Prediction Based on RF Algorithm' has used the random forest algorithm to improve the accuracy of the prediction of customer loyalty in order to guarantee higher customer loyalty for enterprises**

- **In the paper 'The Application of Neural Network and Logistics Regression Models in Predicting Customer Satisfaction in a Student-operated Restaurant' by Aisyah Larasati, Camille DeYong, and Lisa Slevitch, built a neural network model and Logistic Regression**

model to predict overall customer satisfaction and the result shows that a neural network model has a better performance to predict overall customer satisfaction than a logistic regression model.

- **Abubakari S. Gwelo's aim in the paper 'Multinomial modelling of customer satisfaction in the education sector' was to analyse the influence of quality services offered by higher education institutions on student satisfaction they used a multinomial logistic regression model and found that improving services on responsiveness, reliability, tangibility, assurance, and empathy is bound to lead to maximization of satisfaction which would result in increased student's enrolment**

- **Peter Josephat Kirigiti & Abbas Ismail,in the paper 'A Logistic Regression Model of Customer Satisfaction of Airline', a sample size of 272 passengers was selected through a questionnaire. Findings showed that out of the five independent variables only one variable is negatively related to the satisfaction of airline passengers (customer services).Airline safety, schedule integrity, and on board-services are positively related with passengers satisfaction.**

Chapter 3

Methodology

3.1 Data

The dataset was collected through Kaggle. It has 129880 observations and 25 attributes.

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach using descriptive and graphical tools to better understand data. It is used mainly to increase insight into a data set, detect outliers and anomalies, and test underlying assumptions. It is a robust first step before the application of other statistical methods.

3.3 Factor Analysis

Factor analysis is a data reduction method that enables to investigate concepts that cannot easily be measured directly. Factor analysis is a statistical method used to describe variability amoung observed, correlated variables with a lower number of unobserved variables called factors. The observed variables are modelled as a linear combination of latent factors.

3.4 Model Building

3.4.1 Logistic Regression

Logistic regression, sometimes called the logistic model or logit model, analyses the relationship between multiple independent variables and a categorical dependent variable, and by fitting data to a logistic curve, it estimates the probability of the occurrence of an event. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical.

$$f(z) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

$$X = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{3.2}$$

3.4.2 k-Nearest Neighbor

kNN is one of the simplest but most powerful classification algorithms. It assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity, which is when new data appears then it can be easily classified into a well suited category by using kNN. kNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

3.4.3 Support Vector Machine.

SVM is a model, that can do linear classification as well as regression. SVM is based on the concept of a surface, called a hyperplane, which draws a boundary between data instances plotted on the multidimensional feature surface. The SVM algorithm builds an N-dimensional hyperplane model that assigns future instances to one of the two possible output classes.

3.4.4 Decision Tree

Decision tree learning is one of the most widely adopted algorithms for classification. As the name indicates, it builds a model in the form of a tree structure. Its grouping precision is focused on different strategies, and it is exceptionally productive. It is used for multi-dimensional analysis with multiple classes, and its goal is to create a model that predicts the value of the output variable based on the input variables in the feature vector.

3.4.5 Random forest model

Random Forest is an ensemble classifier that is a combining classifier that uses and combines many decision tree classifiers. Ensembling is usually done using the concept of bagging with different feature sets. The reason for using a large number of trees in a random forest is to train the trees enough such that contribution from each feature comes in a number of models. After the random forest is generated by combining the trees, majority vote is applied to combine the output of the different trees. The result from the ensemble model is usually better than that from the individual decision tree models.

Chapter 4

Dataset And Exploratory Data Analysis

4.1 Priliminaries

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 the flight of the passengers(Personal Travel, Buisiness Travel)
Class.	Travel class in the plane of the passengers(Buisiness, Eco, Eco Plus)
Flight distance.	The flight distance of this journey
Inflight wifi service.	Satisfaction level of the inflight service (0:Not Applicable;1-5)
Departure/Arrival time convenient.	Satisfaction level of Departure/Arrival time convenient
Ease of online booking.	Satisfaction level of online booking
Gate location.	Satisfaction level 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
On-board service.	Satisfaction level of On-board service
Leg room 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 departure
Arrival delay in Minutes.	Minutes delay in Arrival
Satisfaction.	Airline satisfaction level(satisfaction, neutral or dissatisfaction)

4.1.1 Data Sample

	id	Gender	Customer	Age	Type of Tr	Class	Flight Dist	Inflight	Depart	Ease of O	Gate loca	Food and	Online boa	Seat comf	Inflight ent	On-board	Leg room	Baggage h	Checkin se	Inflight s	Cleanlir	Departu	Arrival Del	satisfaction
0	19556	Female	Loyal Cust	52	Business tr	Eco	160	5	4	3	4	3	4	3	5	5	5	5	2	5	5	50	44	satisfied
1	90035	Female	Loyal Cust	36	Business tr	Business	2863	1	1	3	1	5	4	5	4	4	4	4	3	4	5	0	0	satisfied
2	12360	Male	disloyal Cu	20	Business tr	Eco	192	2	0	2	4	2	2	2	2	4	1	3	2	2	2	0	0	neutral or dissatisfied
3	77959	Male	Loyal Cust	44	Business tr	Business	3377	0	0	0	2	3	4	4	1	1	1	1	3	1	4	0	6	satisfied
4	36875	Female	Loyal Cust	49	Business tr	Eco	1182	2	3	4	3	4	1	2	2	2	2	2	4	2	4	0	20	satisfied
5	39177	Male	Loyal Cust	16	Business tr	Eco	311	3	3	3	3	5	5	3	5	4	3	1	1	2	5	0	0	satisfied
6	79433	Female	Loyal Cust	77	Business tr	Business	3987	5	5	5	5	3	5	5	5	5	5	5	4	5	3	0	0	satisfied
7	97286	Female	Loyal Cust	43	Business tr	Business	2556	2	2	2	2	4	4	5	4	4	4	4	5	4	3	77	65	satisfied
8	27508	Male	Loyal Cust	47	Business tr	Eco	556	5	2	2	2	5	5	5	5	2	2	5	3	3	5	1	0	satisfied
9	62482	Female	Loyal Cust	46	Business tr	Business	1744	2	2	2	2	3	4	4	4	4	4	4	5	4	4	28	14	satisfied
10	47583	Female	Loyal Cust	47	Business tr	Eco	1235	4	1	1	1	5	1	5	3	3	4	3	1	3	4	29	19	satisfied

4.1.2 Pre-processing

There are 25 columns, out of which two columns, that is unnamed and id are not needed for this study. So the two columns are deleted before beginning the data cleaning. In the data set the majority does not have any null values except one, Arrival Delay in Minutes, which has 310 missing values which was replaced by its mode.

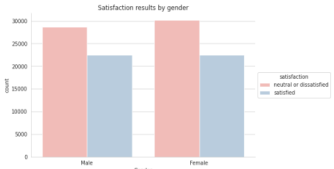
4.1.3 Descriptive Statistics

Descriptive statistics describe, show, and summarise the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements.

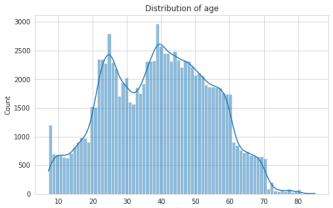
	count	mean	std	min	25%	50%	75%	max
Inflight wifi service	129487.0	2.728544	1.329235	0.0	2.0	3.0	4.0	5.0
Departure/Arrival time convenient	129487.0	3.057349	1.526787	0.0	2.0	3.0	4.0	5.0
Ease of Online booking	129487.0	2.756786	1.401662	0.0	2.0	3.0	4.0	5.0
Gate location	129487.0	2.976909	1.278506	0.0	2.0	3.0	4.0	5.0
Food and drink	129487.0	3.204685	1.329905	0.0	2.0	3.0	4.0	5.0
Online boarding	129487.0	3.252720	1.350651	0.0	2.0	3.0	4.0	5.0
Seat comfort	129487.0	3.441589	1.319168	0.0	2.0	4.0	5.0	5.0
Inflight entertainment	129487.0	3.358067	1.334149	0.0	2.0	4.0	4.0	5.0
On-board service	129487.0	3.383204	1.287032	0.0	2.0	4.0	4.0	5.0
Leg room service	129487.0	3.351078	1.316132	0.0	2.0	4.0	4.0	5.0
Baggage handling	129487.0	3.631886	1.180082	1.0	3.0	4.0	5.0	5.0
Checkin service	129487.0	3.306239	1.266146	0.0	3.0	3.0	4.0	5.0
Inflight service	129487.0	3.642373	1.176614	0.0	3.0	4.0	5.0	5.0
Cleanliness	129487.0	3.286222	1.313624	0.0	2.0	3.0	4.0	5.0
Departure Delay in Minutes	129487.0	14.643385	37.932867	0.0	0.0	0.0	12.0	1592.0
Arrival Delay in Minutes	129487.0	15.091129	38.465650	0.0	0.0	0.0	13.0	1584.0

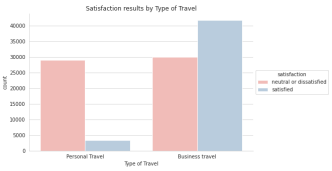
4.2 Exploratory Data Analysis

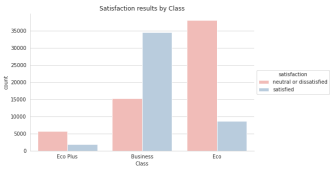
4.2.1 Univariate Visualisations

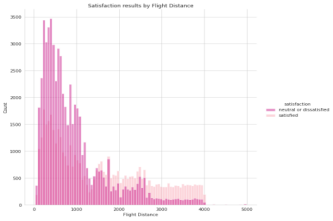
[i]


[ii]


[iii]


[iv]


[v]


[vi]


- **Figure (i) shows the satisfaction level of the attribute gender. The gender split in the data set is relatively even, with 49.25 % of the passengers being male. It appears that the satisfaction level for both genders were relatively similar, with over 55 % of individuals from both genders reporting that they were not satisfied with the service.**

- **Figure (ii) indicates the customer type, that is, whether a passenger is loyal or disloyal, with a large skew in favour of loyal customers. There is a high number of loyal customers who are dissatisfied with the services. Over 76 % of disloyal customers stated that they were dissatisfied with the service.**

- **Figure (iii) shows the attribute age. Examining the age variable, some aspects appear prominently. Firstly, it appears that the majority of those satisfied with the quality of the service were above average age, between 40 and 60. However, after approximately 60 years, satisfaction suddenly drops. Likewise, many of the younger passengers found issues with the service.**

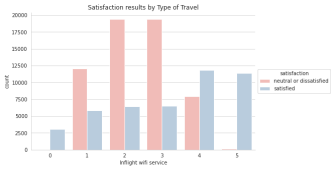
- **Figure (iv) shows the type of travel. It appears that it may specialise in business travellers as the large share of the passengers**

utilised the airline for business travel. Furthermore, such travellers were more likely to find the service satisfactory.However, those who travelled for personal reasons found the experience extremely poor, approximately 90%. This is concerning, and the airline should examine the issues relating to their service and rectifying any issues with the utmost urgency, if they wish to retain customers.

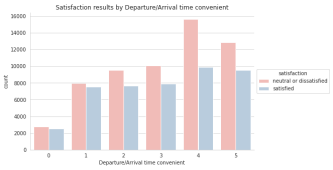
- **From figure (v), classes also highlight the dominance of business travel as the majority of those who travelled via business class were satisfied with the service they received. In contrast, those who were in Eco-class was particularly dissatisfied with the service, with 81% voicing dissatisfaction. Given that 45% of the passengers travelled Via Eco, the airline should focus on improving their services in this area.This is particularly the case, as even the business travellers who used Eco were disappointed with the service. The poor quality of Eco is also appears to be the major factor in the low ratings amongst those who travelled for personal reasons and they were more likely to use eco class.**

- **From figure (vi), it appears that the airline holds specific strengths in long distance travel. Further investigation reveals that the rate of dissatisfaction differs if the individual has travelled more or less than 1500. For those who travelled less than 1500 km, dissatisfaction was as high as 66%. However, if the passenger travelled more than 1500 km, they were only 34% likely to be dissatisfied with the service.**

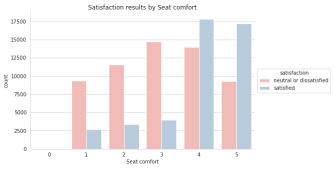
[vii]



[viii]



[ix]



- **Figure (vii) shows in-flight WiFi service.For this variable; it seems**

that in general, if the passenger rates the service between 1 and 3, they were more likely to consider that the overall service was poor. In contrast, as travel type had a major impact on satisfaction, the data can be viewed with respect to this factor.In doing so, it highlights that for business passengers, they followed the overall trend previously identified. Between 1 and 3 ratings for WIFI, the overall rate of dissatisfaction rises. However, at a score of 4, the rating drops significantly.Personal travellers also recorded a similar pattern. On the other hand, as they were overall more likely to be dissatisfied, for scores between 1 and 3 for in-flight WIFI, all travellers were likely to rate the overall service dis-satisfactory. That said, the biggest level of satisfaction was identified amongst the personal travellers who identified that the WIFI was 5 stars.

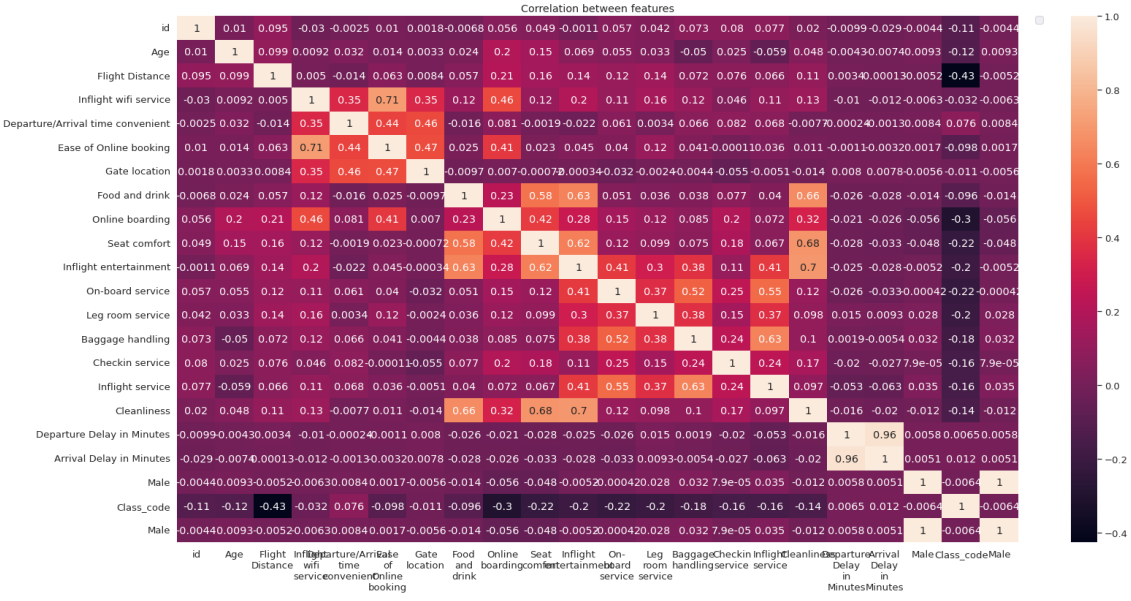
- **For Departure/Arrival time convenience, it appears that the rate of convenience had a little impact on the overall satisfaction.For those who rated the timing as extremely convenient (5 stars), their dissatisfaction with the service was only 30%.**
- **Similar to WiFi, those who were satisfied with the booking service were more likely to rate the overall service satisfactory**
- **On the other hand, for personal travellers, if the booking service was poor, they were extremely likely to find the overall service poor**
- **It appears that gate location may have a relatively limited impact on the overall satisfaction. General level of correlation is low and regardless of travel type, there was no major fluctuation in the overall satisfaction is dependent on the convenience of the gate location.**
- **The impact on overall satisfaction based on food and drink appears to be somewhat weak. However, for business travellers, if they found the food to be particularly poor, they were more likely to rate the overall satisfaction as poor. As such, there may be benefits in ensuring that the food is not rated as 1 by those who access it.**

- **In general, it appears that the online boarding services are impacting the customer's overall level of satisfaction. This is, however, felt more by business travellers. If they experienced excellent online boarding services, they were more likely to rate the service highly. It seems that for these travellers, they value convenience highly, and therefore ensuring such services will garner more business travellers.**

- **Seat comfort on an overall level shows that those who rated the seats extremely comfortable, also rated the overall experience with the airline highly.This is particularly prominent with business travellers, for whom if the seats were found to be particularly conformable, the rate of dissatisfaction fell to 28% or lower.**

- **The trends in seat comfort are also reflected in in-flight entertainment and onboard service. The better the experience, the higher the satisfaction rating. However, business travellers were more sensitive to this than personal travellers. Personal travellers maintained a similar level of dissatisfaction regardless of their rating of the in-flight entertainment/ onboard services.**

4.2.2 Multivariate Visualisatons



- **There is a high positive correlation between arrival delay and departure delay. If a flight departs late, it is likely to arrive late as well. such as seat comfort, food and drink, and in-flight entertainment tend to have a positive correlation.**
- **Variables such as WIFI service, online booking, and gate location tend to have a positive correlation.**
- **Cleanliness is correlated with gate location, food and drink, seat comfort and in-flight entertainment.**

Chapter 5

Result And Discussions

5.1 Factor Analysis

1. **3.80389837**
2. **2.3721328**
3. **2.16956004,**
4. **1.96182794**
5. **1.06300589**
6. **0.94961228**
7. **0.69607891**
8. **0.53723377**
9. **0.51374997**
10. **0.46784182**
11. **0.36634408**
12. **0.32884042**
13. **0.29332142**
14. **0.25443665**
15. **0.18743275**
16. **0.03468288**

Here, there are five eigen values that are greater than one. So here there is a need to choose five factors (or unobserved factors). In factor analysis, a factor with an eigenvalue of 1 accounts for as much variance as a single variable, and the logic is that only factors that explain at least the same amount of variance as a single variable is worth keeping. From this figure, it can be said that,

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Online boarding	0.289549	0.122385	-0.009535	0.108246	0.754005
Inflight wifi service	0.095122	0.134786	-0.009001	0.614102	0.465372
Ease of Online booking	-0.032297	0.031080	-0.002353	0.772955	0.448606
Seat comfort	0.756388	0.079526	-0.013844	-0.026458	0.209397
Checkin service	0.113416	0.287751	-0.013051	-0.027097	0.133295
Cleanliness	0.854195	0.084949	0.000647	-0.001291	0.097845
Leg room service	0.057830	0.486148	0.023440	0.043128	0.092634
On-board service	0.085271	0.701342	-0.019281	0.010336	0.047134
Food and drink	0.770830	0.004101	-0.018019	0.030665	0.034680
Inflight entertainment	0.767526	0.466055	-0.007833	0.040945	0.023256
Departure Delay in Minutes	-0.015680	-0.014231	0.968664	0.000091	-0.006186
Departure/Arrival time convenient	-0.009576	0.055463	-0.000297	0.589526	-0.006491
Arrival Delay in Minutes	-0.017345	-0.019420	0.995885	-0.000800	-0.008277
Baggage handling	0.036425	0.764506	0.006939	0.046204	-0.035087
Inflight service	0.035749	0.799371	-0.044377	0.046369	-0.058022
Gate location	0.012585	-0.046715	0.004773	0.682653	-0.111332

1. **Factor 1 has high factor loadings for cleanliness, Food and Drink, Inflight Entertainment, Seat comfort**
2. **Factor 2 has high factor loadings for Inflight services, Baggage Handling, Onboard Services, Leg room**
3. **Factor 3 has high factor loadings for Arrival and Departure Delay**
4. **Factor 4 has high factor loadings for Online Booking, Gate Location, Inflight Wifi, and Departure/Arrival Time Convenience**
5. **Factor 5 has high factor loadings for Online Boarding.**

The Cumulative variance shows the amount of variance explained by n+(n-1) factors. In this case, the result shows that the five factors together are able to explain 61
Findings: From the result, it is obvious that data reduction was done by using Factor Analysis. The 5 factors can be named as

1. **Comfort (cleanliness, food and drink, inflight entertainment, seat comfort)**
2. **Service (inflight services, baggage handling, onboard services, leg room)**
3. **Arrival and Departure delay**
4. **Covenience (online booking, gate location, inflight wifi, departure/arrival time covenience)**
5. **Online Boarding.**

These are the factors that are expected to have a high influence on airline passenger satisfaction.

5.2 Logistic Regression

The aim is to fit a regression model for the given data. Here logistic regression model is used since the dependent variable (satisfacton) is categorical.
For this data, satisfied is set to 1 and dissatisfied to 0.
The coefficient values are -5.8305,.0406,-2.0360,-2.7204,-0.7365,-0.8520,-0.0083,-1.711e-05,0.3947,-0.1248,-0.1429,0.0293,-0.0276,0.6124,0.0668,0.0639,0.3016,0.0089.

- **To interpret the overall significance of the model, it is taken as loglikelihood of our model Vs log likelihood of null model.**
- **Since a good would mean by the likelihood of the data generated by the model is high (close to 1) therefore log(likelihood) should be closer to 0.**

- **In this case, log likelihood of our model -34000 and of null moel is -71000 with a p value of 0.00 therefore we have a significant model.**
- **Unlike a linear regression, there is no straight forward way to calculate r-square in logistic regression and therefore psuedo-r-square is used. A value closer to 1 would mean that our independent variables 'explain'most of the variation in the dependent variable. A value closer to 0 would mean that the independent variables help explain nothing about the dependent variable.**
- **In this case, the value is 0.46, and therefore we don't have much explanatory power in our dependent variables and would need more (external) information about our passengers to determine what makes them satisfied.**

Findings:

- **The Accuracy percentage of the model is 87.6%**
- **Most variables turn out to be statistically significant as their p-values are close to 0**
- **Only flight distance seems to be a statistically insignificant variables at a 95% confidence interval**

5.3 Model Fitting

To find the best model to fit the data, here used the machine learning algorithms such as K-Nearest Neighbor Classification, Support Vector Machine Classification, Decision Tree Classification, and Random Forest Classification.

ROC AUC SCORE:-

The ROC AUC score tells us how efficient the model is. The higher the AUC, the better the models' performance at distinguishing between positive and negative classes. An AUC score of 1 means the classifier can perfectly distinguish between the entire positive and the negative class points.

K-Nearest Neighbor Classification

The ROC AUC score of model is 0.919

Support Vector Machine Classification **The ROC AUC score of model is 0.942**

Decision Tree Classification

The ROC AUC score of model is 0,935

Random Forest Classification

The ROC AUC score of model is 0.954

5.4 Comparing The Models Predicting With The Best Model

The models are compared according on the basis of their accuracy and the prediction is done with the most accurate one.

5.4.1 Comparing The Models

The models are compared and their accuracies are as in following table. The Random Forest Classification has the highest accuracy than

Model Name	Accuracy	F1 Score	ROC AUC Score
Logistic Regression	0.876	0.858	0.874
K Nearest Neighbours Classification	0.919	0.905	0.919
Support Vector Machine Classification	0.943	0.934	0.942
Decision Tree Classification	0.937	0.928	0.935
Random Forest Classification	0.954	0.948	0.954

the other classifications. So the same is used for prediction of airline passenger satisfaction.

For predicting with random forest the data set was divided into two set such as Training and Test data set.

The Test data is predicted by Random Forest Model.

The predicted values are given as below.

	actual value	predicted value
0	0	0.00
1	0	0.53
2	1	1.00
3	1	1.00
4	0	0.00
...	...	...
7763	0	0.00
7764	0	0.00
7765	1	1.00
7766	1	0.32
7767	1	1.00

Chapter 6

Conclusion

The purpose of this project was to analyse and predict airline passenger satisfaction, and all the objectives of this project have been achieved. The first aim was to find the factors that influence passenger satisfaction highly, and it was found that comfort, service, arrival and departure delay, convenience, online boarding were those factors. A Logistic Regression model was fitted. Amid different Machine Learning Models Random Forest Model performs the best, and prediction was done using the same.

This survey helps the airlines understand the most important factors that have a direct impact on passenger satisfaction. Airline companies should strive for value creation and appropriate allocation of resources in international air travel by creating more realistic consumer's expectations about the promises that airlines make, as this may increase the level of passenger satisfaction, for example, through meeting consumer's desired service levels,dealing effectively with dissatisfied customers, and confronting customer complaints positively. This study can be extended by considering more factors that could have affected the passenger satisfaction.

Chapter 7

Reference

1. Saikat Dutt,Subramanian Chandramouli, Amit Kumar Das,(2018), Machine Learning

2. Adarsh Anand, Gunjan Bansal(2016). Predicting Customer's Satisfaction Using Logistic Regression,DOI:10.33889/IJMEMS.2016.1.2-009.

3. Mustafa Raza Rabbani, Zehra Zulfikar (2013).Applying Factor Analysis in Assessing Customer Satisfaction with Service Quality: The Case of Indian Banks[165].

4. Sachin Kumar, Mikhail Zymbler (2019).A machine learning approach to analyze customer satisfaction from airline tweets,DOI:10.1186/s40537-019-0224-1.

5. Dr Kali Charan Modak, Dr. Kumkum Sinha(2019). A Logistic Regression Model of Customer Satisfaction for Online Food Delivery Services[320].

6. R.Siva Subramanian,D.Prabha,B.Maheswari ,J.Aswini(2021).Analysis Using Machine Learning Algorithms: A Case Study Using Banking Consumer Dataset, DOI:10.3233/APC210263.

7. Moulay Smail Bouzakraui, Abdelalim Sadiq, Abdessamad Youssfi Alaoui (2020).Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques,DOI:10.25046/aj050470.

8. **Jiong Mu, Lijia Xu, Xuliang Duan, Haibo Pu(2013).Study on Customer Loyalty Prediction Based on RF Algorithm,DOI:10.4304/jcp.8.8.2134-2138.**

9. **Aisyah Larasati, Camille DeYong, Lisa Slevitch (2012).The Application of Neural Network and Logistics Regression Models in Predicting Customer Satisfaction in a Student-operated Restaurant,DOI:10.1016/j.sbspro.2012.11.097.**

10. **Abubakari S. Gwelo(2019).Multinomial modelling of customer satisfaction in the education sector,doi: 10.22367/jem.2019.35.04.**

11. **Peter Josephat Kirigiti,Abbas Ismail(2012).A Logistic Regression Model of Customer Satisfaction of Airline,DOI:10.5296/ijhrs.v2i4.2868.**