Machine Learning in Business

MIS710 – A1

Case Study (Business Report)

VayuAir Australia

Sachin Bhat, 218676233

DEAKIN
BUSINESS
SCHOOL

CONTENTS

Executive Summary

The project was commissioned by Ms. Lyn Gray, Customer Insights Manager, VayuAir and prepared by business analyst Sachin Bhat. Their objective is to extract insights from the data and predict customer satisfaction while identifying the ways to boost customer engagement and retention. We have identified it as a classification problem and used the supervised machine learning's logistic regression approach through the LogisticRegression functionality from the scikit-learn library to solve this. Factors such as flight distance, online check in quality, airport check-in process, baggage handling quality, cabin-crew service, seat comfort, inflight entertainment, inflight wifi, leg room, and cleanliness positively influence satisfaction. While maintaining and improving the quality of the factors positively influencing customer satisfaction, VayuAir should focus on improving service quality across the factors negatively influencing customer satisfaction through targeted actions. The expected value gains include enhanced customer retention, operational efficiency, great word-of-mouth, an increased customer based, and enhanced revenues.

Business Analysis Core Concept Model

Need: Based on the customer survey data, the primary need is to identify the main factors that are influencing customer satisfaction. VayuAir wants to establish prediction capabilities around customer satisfaction outcomes as this would enable proactive betterment of the services along with targeted improvements in customer experience.

Context: Airline industry has an extremely competitive environment involving changing customer expectations and strict operational limitations. Organisations in this industry need to have a customer-centric approach which is vital in valuing customer feedback and continuous improvement. The technologies are advancing and play a significant role in improving the operational effectiveness and customer experience.

Stakeholder: Stakeholders in this project include VayuAir as the business owner, Ms. Lyn Gray - Customer Insights Manager at VayuAir, service enhancement teams, operational staff, marketing team, and the executive management. The clients include businesses and regular travellers. Other stakeholders include regulatory and airport authorities, service partners managing catering and in-flight entertainment.

Change: VayuAir is new airline company in Australia whose objective is to enhance overall customer satisfaction to improve its retention rates and competitive position in the airline industry. To enhance the customer experience, the change will focus on implementing strategic improvements across different aspects of customer service, including booking, airport arrangements, flight scheduling, and in-flight services.

Value: Value proposition to VayuAir includes enhanced customer retention and loyalty, improved operational efficiency, increased competitive edge and a good word-of-mouth leading to an upsurge in customer base. Also, the customers get a comfortable, well-tailored, and a satisfactory travel experience being at par with their expectations.

Solution: Through data analysis, the high-impact areas have been identified. It would be best to implement immediate enhancements to them. This would mean enhancing in-flight services such as seat comfort, cleanliness, entertainment, Wi-Fi, cabin crew service etc. while streamlining and improving the online check-in process. Creating a predictive model based on customer satisfaction will yield long-term benefits allowing VayuAir to anticipate and cater to areas involving potential dissatisfaction, customise services depending on traveller type, and constantly monitor areas to enhance service quality on the basis of customer feedback and model insights.

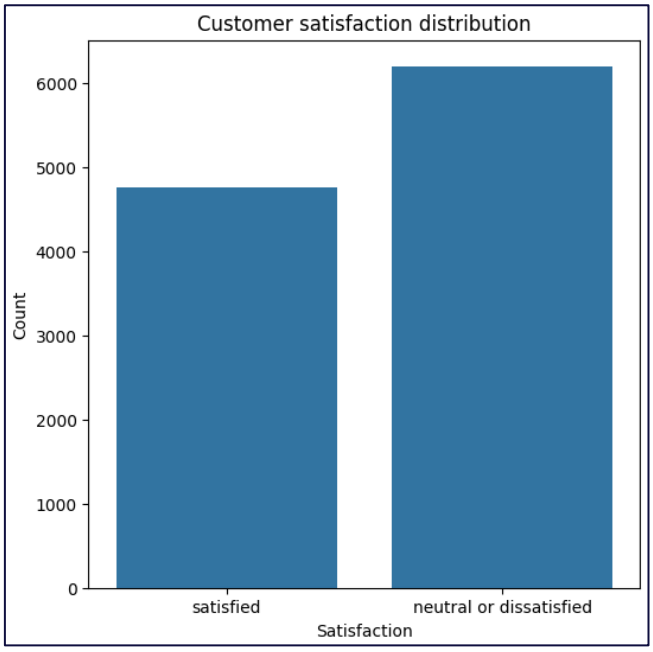
Data details and Univariate Analysis

The data sample we have been provided with is obtained from VayuAir's customer survey dataset and they include a wide range of aspects shedding light on their customers' flight experience. The dataset includes **categorical** factors like satisfaction and gender along with **numerical** factors like age and flight distance.

Initially we had 11006 rows and 25 columns. We found 53 missing values under 'Arrival delay in minutes' and since the number is small, we removed the rows having missing values under this column. Then we got 10953 rows and 25 columns.

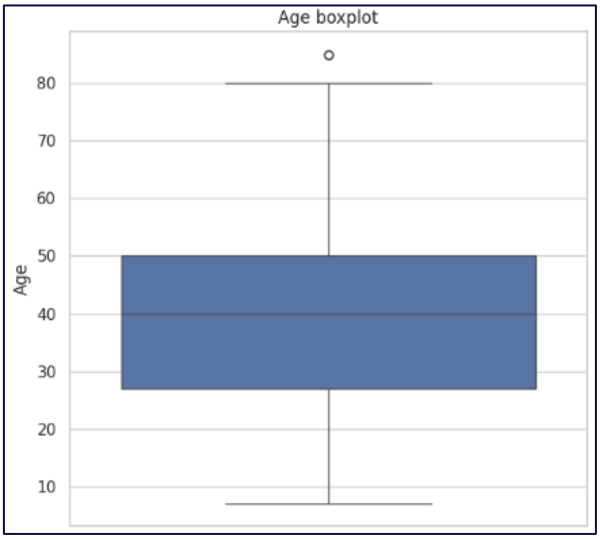
Analysing features individually:

Satisfaction:



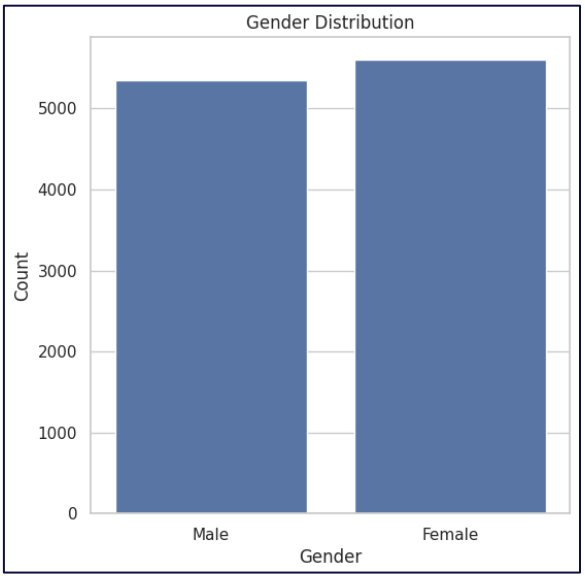
6200 customers are neutral or dissatisfied accounting for 56.60%. 4753 customers are satisfied accounting for 43.49%. Most customers are not satisfied which isn't a good sign. VayuAir has to devise strategies to increase this number.

Age:



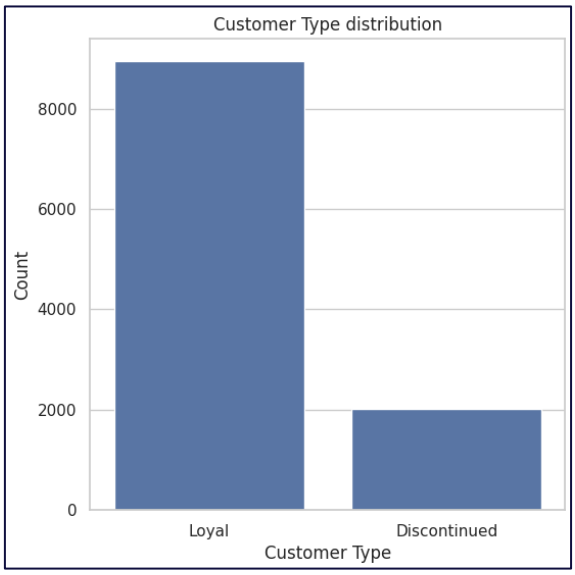
25% of their customers are 27 years old or younger. 50% of them are younger than 40 years and the other half is older. 75% of them are 50 years old or younger. The minimum age is 7 and maximum is 85 also being the outlier as its atypical for this data range while the average age is 39 years.

Gender:



5600 of the customers are female accounting for 51.13%. 5353 of the customers are male accounting for 48.89%. Most customers are female. VayuAir can create tailored services for each gender.

Customer Type:



8940 customers are loyal accounting for 81.62%. 2013 customers have discontinued using VayuAir accounting for 18.38%. VayuAir can produce personalised offers to win them back.

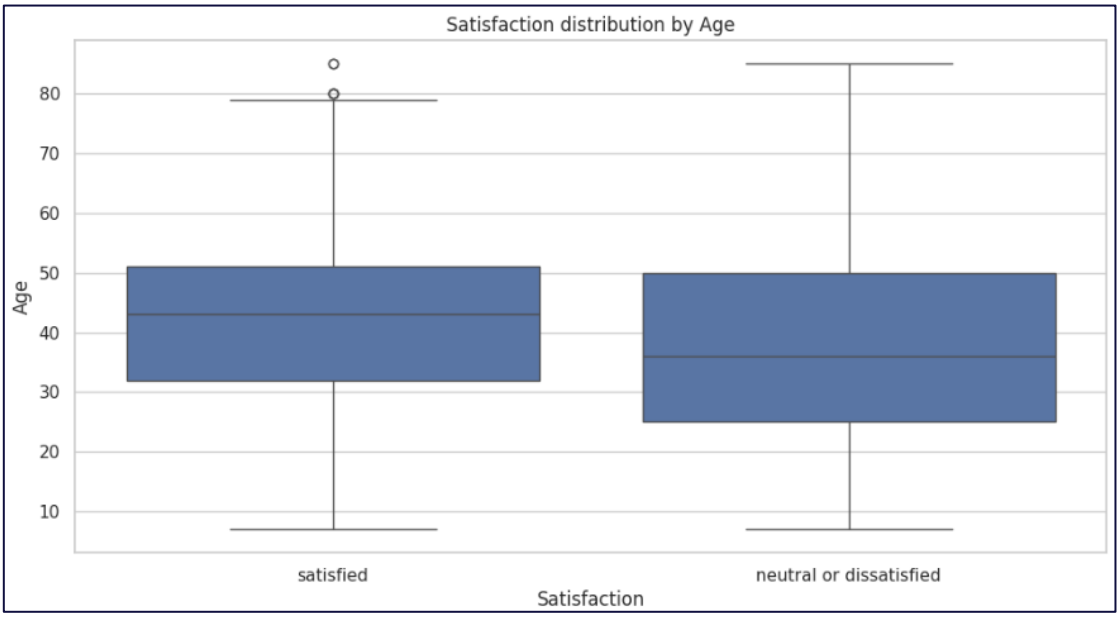
Travel Type:

7555 customers have travelled for business purposes accounting for 68.98%. 3398 customers have travelled for personal purposes accounting for 31.02%. People majorly use VayuAir for business purposes.

Bivariate Analysis

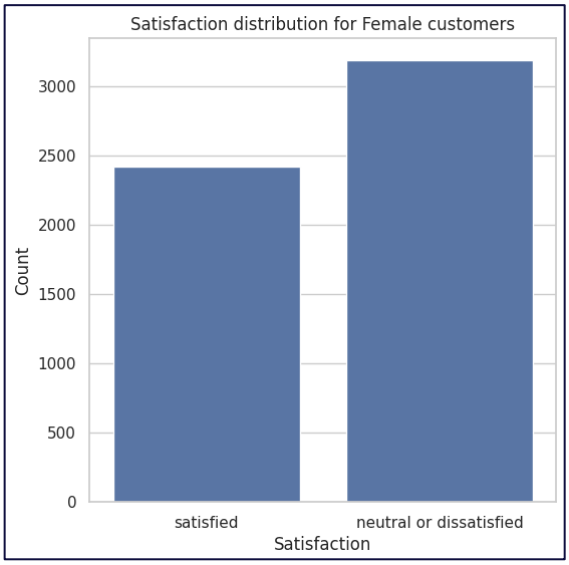
Analysing one feature at a time with respect to satisfaction:

Q1. Satisfaction by Age:

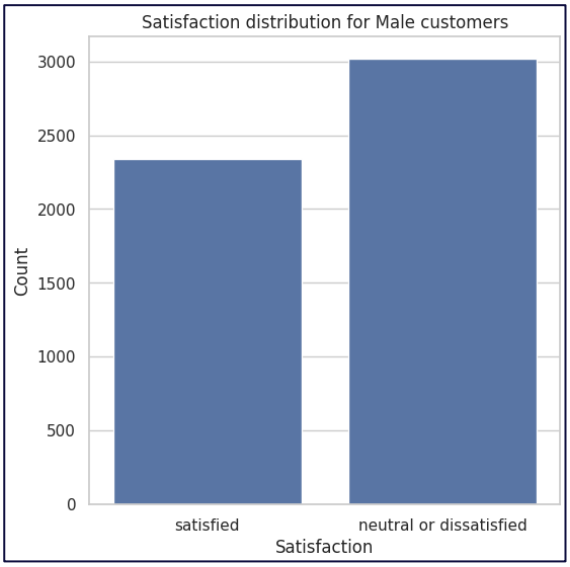


The median age of satisfied customers is around middle 40s which is slightly higher than late 30s of neutral or dissatisfied customers. Satisfied group has a few outliers who are older than the typical ages for their group. The neutral or dissatisfied group tends to have slightly younger customer base as the median range is lower and the age range is broader.

Q1. Satisfaction by Female and Male genders:



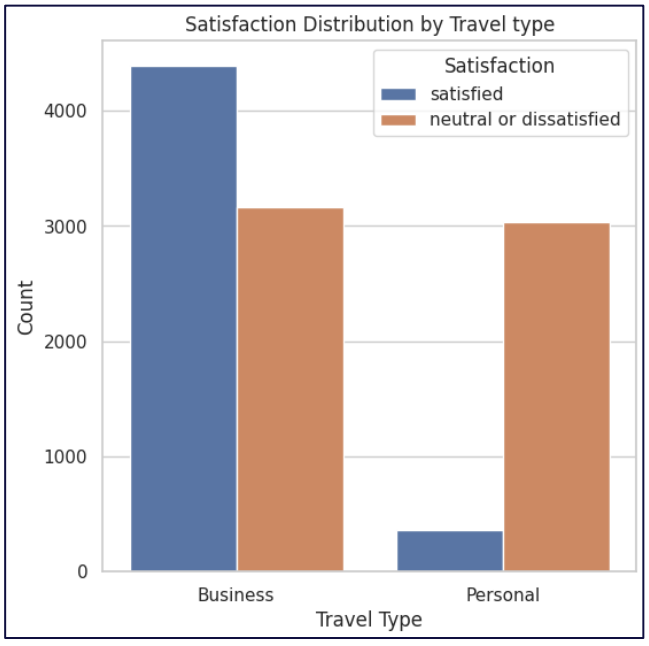
Female



Male

3184 females are neutral or dissatisfied accounting for 56.86%. 2416 females are satisfied accounting for 43.14%. 3016 males are neutral or dissatisfied accounting for 56.34%. 2337 males are satisfied accounting for 43.66%.

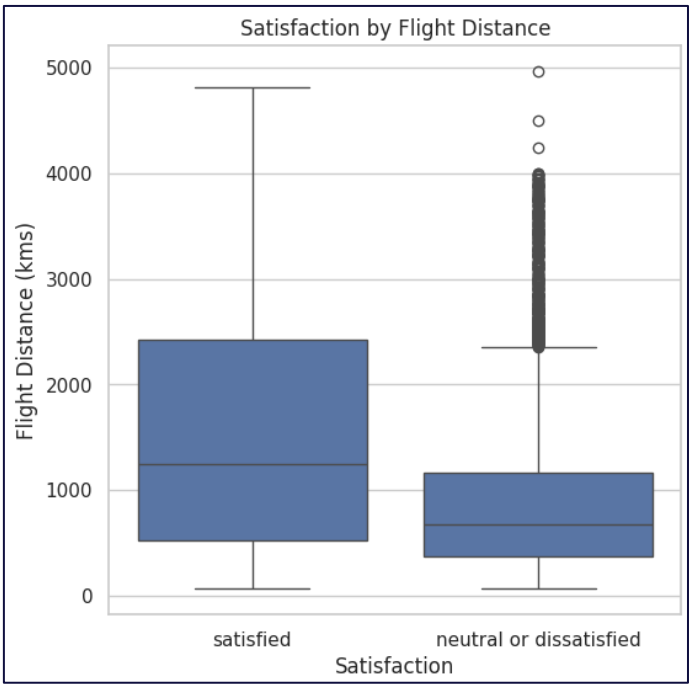
Q2. Satisfaction by Travel Type:



Travelling for business reasons: 4389 customers are satisfied and 3166 are neutral or dissatisfied.

Travelling for personal reasons: 364 customers are satisfied and 3034 are neutral or dissatisfied. Most people travelling for personal reasons are dissatisfied.

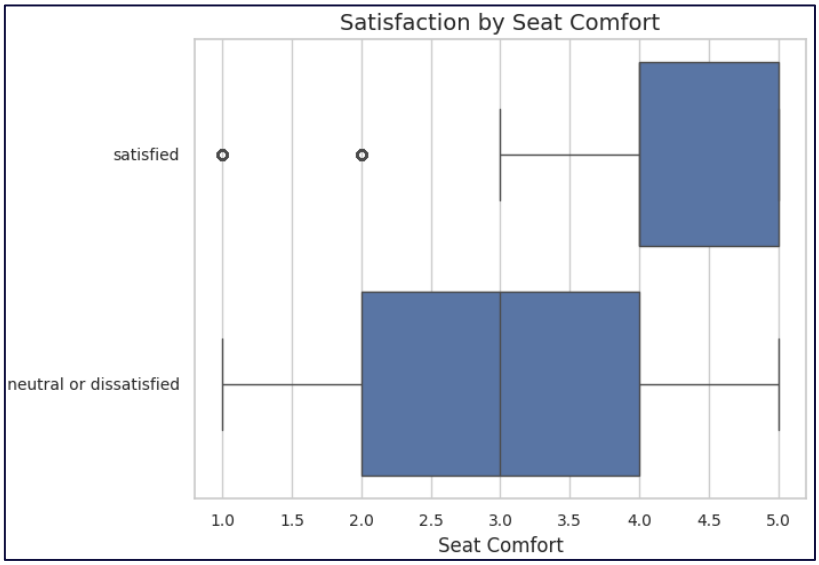
Q3. Satisfaction by Flight Distance:



For satisfied customers, the median flight distance is above 1000 kms suggesting that they are usually flying longer distances. There are varying flight distances among satisfied customers. Customers flying longer flight distances tend to have a better experience possibly because of the factors like services offered on those flights being at par with customers' expectations.

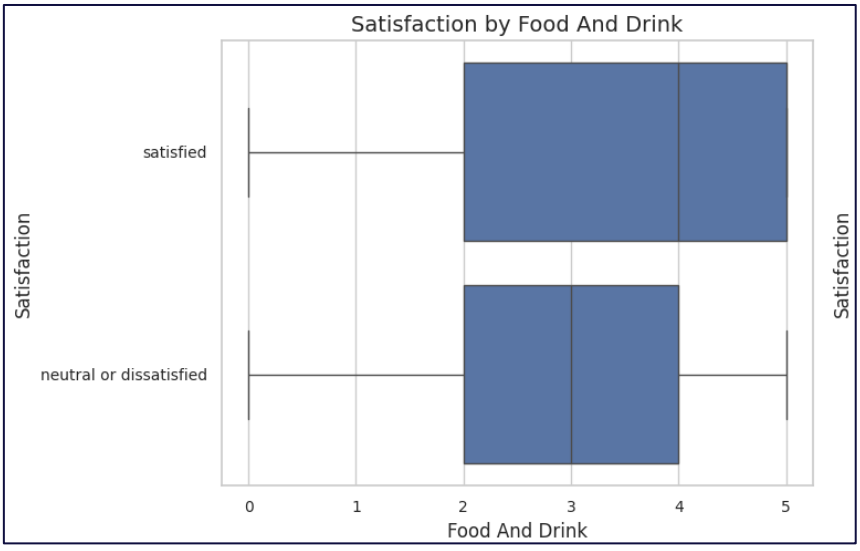
For neutral or dissatisfied customers, the median flight distance is below 1000 kms suggesting that they are usually flying shorter distances. The outliers in this group tell us that there have been a few instances of extremely long flights. Customers flying shorter flight distances are more likely to be neutral or dissatisfied. This should be addressed.

Q4. Satisfaction by Seat comfort:



Most satisfied customers have given high ratings ranting from 4-5 for seat comfort indicating that it is a critical factor influencing satisfaction. Outliers here could mean that other aspects of the flight experience overshadowed seat comfort in terms of overall satisfaction. The neutral or dissatisfied customers have given middle ratings ranting from 2-4. No bars at extreme ends could mean that extremely poor seat comfort is rare and customers who gave highest ratings generally had satisfying experience.

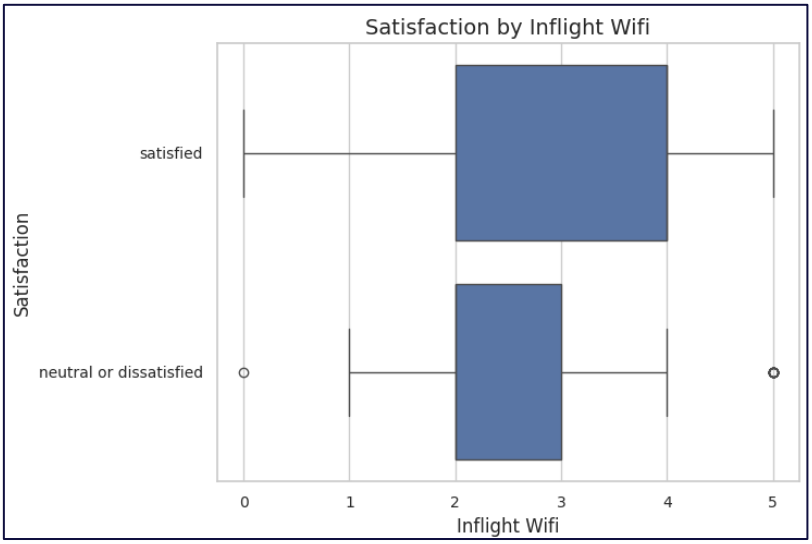
Q4. Satisfaction by Food quality:



Many satisfied customers have given high ratings ranging from 4-5 indicating food and drinks positively affect overall satisfaction. Lack of low ratings might suggest that VayuAir consistently provides above-average food and drinks.

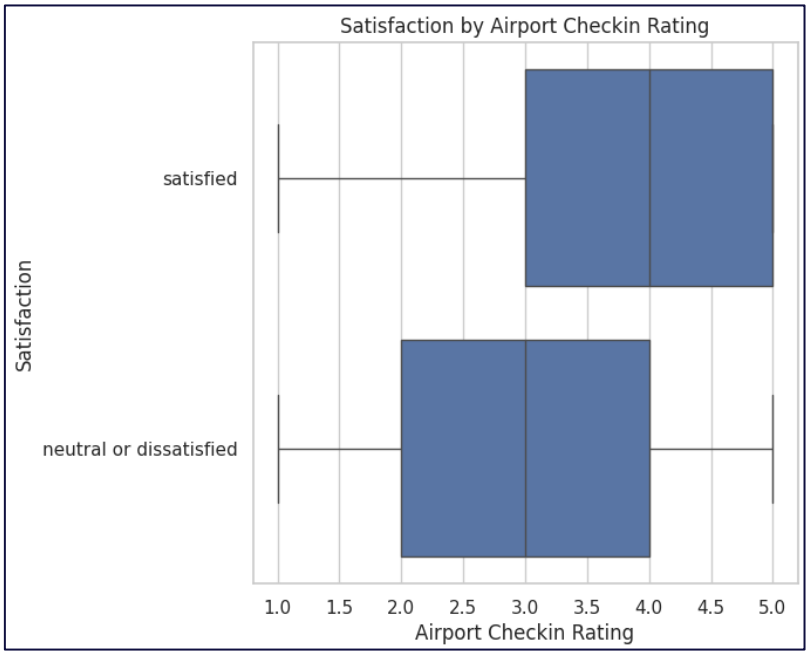
Neutral or dissatisfied customers have given a range of ratings with a higher concentration in the middle-range between 2-4. Lack of extreme customer ratings i.e. 0 and 1 might suggest that few people give extreme ratings, or their experiences don't greatly influence overall satisfaction.

Q4. Satisfaction by Inflight wifi:



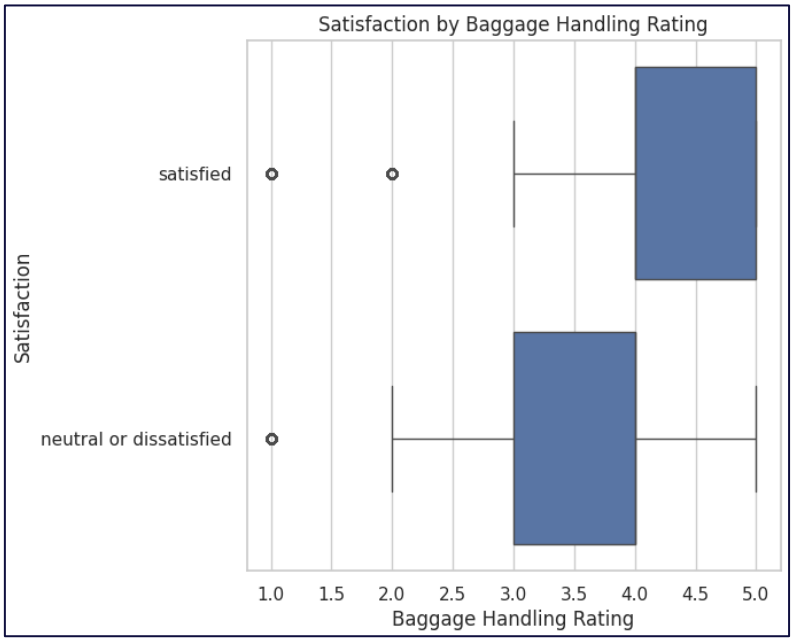
Most satisfied customers have given high ratings to inflight wifi which indicates that good wifi service could be a key contributor to overall satisfaction. The neutral or dissatisfied customers' ratings for inflight wifi are evenly distributed suggesting that it might not be the only factor influencing overall satisfaction. This plot indicates that while wifi is an important factor its influence on overall satisfaction varies among individuals.

Satisfaction by Airport checkin:



Most satisfied customers have given high ratings between 4-5 to airport checkin. This suggests that positive views of the airport checkin are connected to overall customer satisfaction. The neutral or dissatisfied customers have given ratings from low to high ranges which are concentrated around 3. This suggests that customers who rate airport checkin very highly are mostly satisfied customers. High quality airport checkin likely leads to higher overall satisfaction.

Satisfaction by Baggage handling:



Satisfied customers usually rate baggage handling highly suggesting its relevance to overall satisfaction. Neutral or dissatisfied customers have given varied ratings indicating that other factors also influence their overall satisfaction. Both groups have outliers implying different experiences among customers.

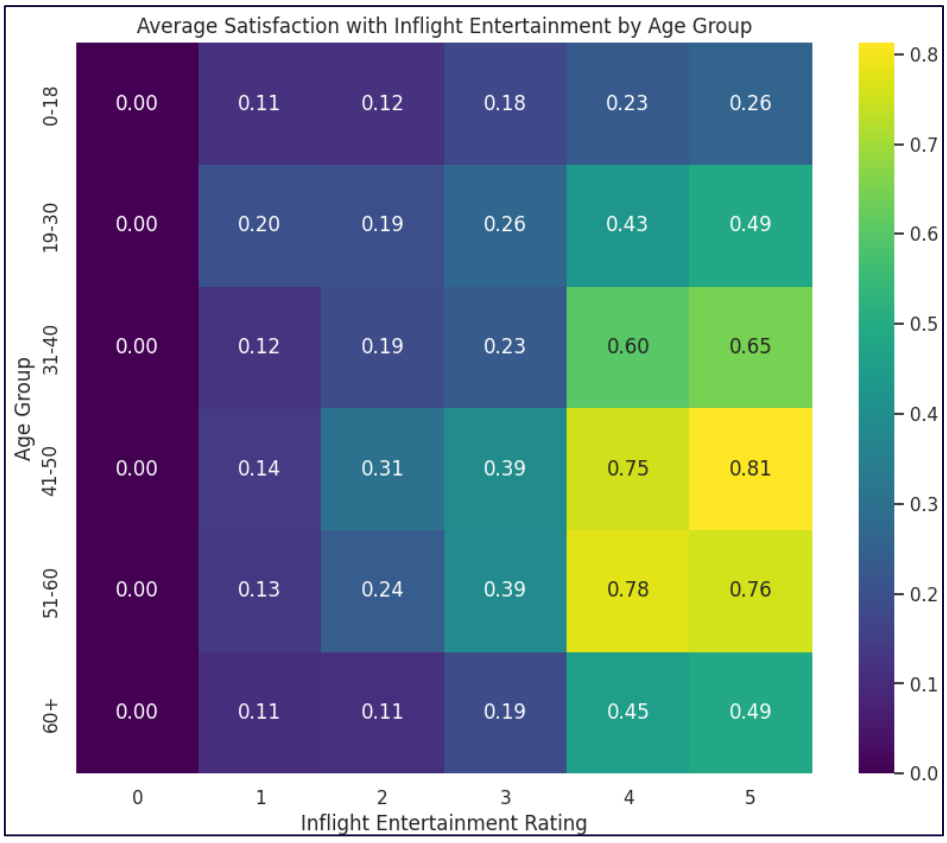
Multivariate Analysis (Additional Insight)

Encoding Variables:

Satisfaction has been encoded using Label encoder with 'satisfied' as 1 and 'neutral or dissatisfied' as 0. Gender, Customer type, Travel type and Ticket type have been one-hot encoded.

Analysing two features at a time with respect to satisfaction:

Satisfaction by age groups and inflight entertainment:



This heatmap helps us to establish that average customer satisfaction based on inflight entertainment ratings tends to increase with age. Customers between 31-40, 41-50 and 51-60 appear to have achieved higher satisfaction levels while enjoying the inflight entertainment. The ratings are lower for people from the 0-18, 19-30, and 60+ age groups which might suggest that they are less interested in the offered inflight entertainment.

Feature selection and Data Split

We have set Satisfaction_encoded as the label. We have removed the variables we believe are unrequired and irrelevant as predictors: Module_id, Response_ID, Age, Satisfaction, Satisfaction_encoded, Departure delay in minutes, Arrival delay in minutes, Customer type_Loyal, Flight schedule suitability, Gender_Male.

The relatable variables influencing Satisfaction_encoded have been taken as: Booking service, Online checkin, Airport checkin, Baggage handling, Boarding service, Food and drink, Inflight wifi, Seat comfort, Inflight entertainment, Cabin crew service, Leg room, Inflight amenities quality, Cleanliness, Ticket type_Economy, Ticket type_Economy Plus, Flight distance, Total_Delay_in_Minutes, Travel type_Personal.

We have split the data into: 80% for training the model and 20% for testing its performance. Random state = 42 to get same train and test sets across different executions. This is crucial for reproducibility, debugging, and comparing the results.

Machine Learning Approach

We have used supervised ML approach of Logistic Regression using LogisticRegression from scikit-learn. As the target is categorical, this is a classification problem. The datapoints are independent, there are no extreme outliers, no severe collinearity among the predictors. There exists a linear relation between each predictor and the logit of the target. The sample size is large enough.

Model Formula: log_odds = -0.5134 + 0.1911*Flight distance + -0.3142*Booking service + 0.9472*Online checkin + 0.3545*Airport checkin + 0.0771*Baggage handling + -0.0797*Boarding service + 0.2880*Cabin crew service + 0.0791*Seat comfort + -0.1119*Food and drink + 0.4552*Inflight entertainment + 0.5555*Inflight wifi + 0.3971*Leg room + -0.0804*Inflight amenities quality + 0.0789*Cleanliness + -0.8048*Travel type_Personal + -0.5408*Ticket type_Economy + -0.2420*Ticket type_Economy Plus + -0.1578*Total_Delay_in_Minutes

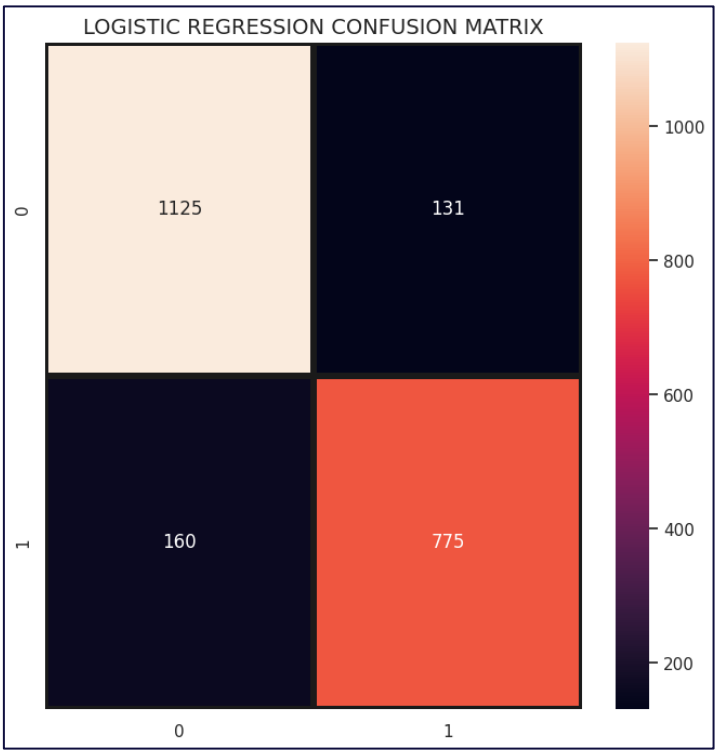
-0.5134 is the intercept. The log odds of a customer being satisfied increases with increase in: flight distance, online check-in quality, airport check-in process, baggage handling quality, cabin-crew service, seat comfort, inflight entertainment, inflight wifi, leg room, and cleanliness. Customer satisfaction decreases with decrease in: booking service quality, boarding service quality, food and drinks quality, inflight amenities quality, personal travel type, economy ticket type, economy plus ticket type, and total delay in minutes.

Model Evaluation, Pros and Cons

	Precision	Recall	f1-score
Class 0: likely neutral or dissatisfied	0.88	0.90	0.89
Class 1: likely satisfied	0.86	0.83	0.84

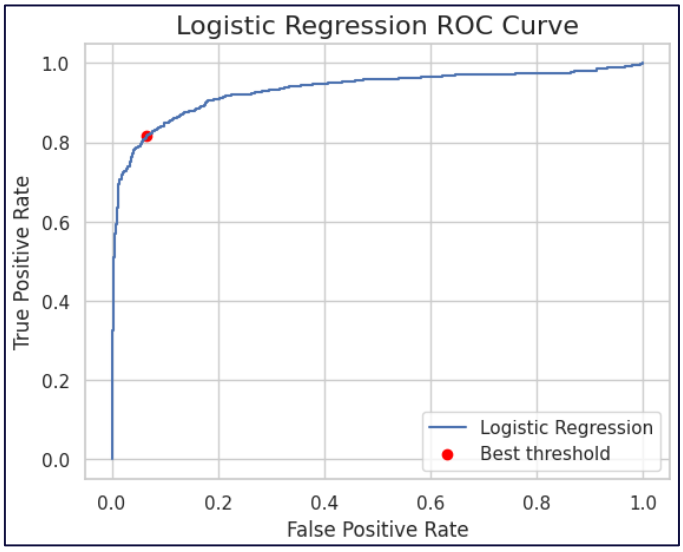
Our Model has low false positive and high true positive rates. It is relatively more balanced at predicting class 0.

Confusion matrix:



Our model correctly predicted ‘class 0’ as ‘class 0’: 1125 times (true negatives) and ‘class 1’ as ‘class 1’: 775 times (true positives). It incorrectly predicted ‘class 1’ as ‘class 0’: 160 times (false negatives) and ‘class 0’ as ‘class 1’: 131 times (false positives).

ROC Curve:



The true positive rate and false positive rate are plotted at various thresholds. The extreme rise towards the top-left depicts a high TPR for a low FPR suggesting a nice classification performance. Accuracy of the test is reassured by looking at this ROC space as the curve follows the left-hand border and then the top border. Best threshold= 0.5781730422749346.

Pros: It is easily explainable due to the coefficients' relationships with target variables. Visual representation (ROC curve) has been done. Less effort is needed for data preparation, no need for normalisation but our model underwent scaling which was necessary. Works for both numerical and categorical predictors.

Cons: Target variable needs to be categorical as it is designed for binary/categorical results. Large datasets are required, overfitting occurs if they are too small. Can get complex when having multi-class targets.

Recommendations and Improvement

VayuAir should use the logistic regression model for predicting customer satisfaction (classification) while focusing on ‘neutral or dissatisfied’ customers across different factors with targeted actions to improve their experience. Booking service quality, boarding service quality, food and drinks quality, inflight amenities quality, personal travel type, economy ticket type, economy plus ticket type, and total delay in minutes negatively influence satisfaction. I recommend VayuAir to improve the services negatively affecting customer satisfaction while maintaining those at which they are good.

After removing Food and drink from the predictors, we achieved improved model performance. The model is able to better identify the true negatives and true positives. The number of false positives and false negatives has decreased.