Airline Customer Satisfaction

Project Report

DSCI 5240 Data Mining and Machine Learning for Business

University of North Texas

Team 8: Priyadharshini Ayyampalayam Ramajeyam, Sufia Nooreen, Chandrika Jilla, Santoshi Nikitha Modala, Kavya Damera

Table of Contents

Introduction

In this world of cut-throat competition and ever-changing customer demands, understanding the underlying factors that determine passenger satisfaction is imperative for airline companies to stay dominant in the market. This project studies the experiences of various airline passengers by utilizing a dataset of an unidentified airline company that includes a wide range of variables pertaining to their travel. By employing prediction models, this report aims to provide valuable insights into the factors that influence passenger satisfaction.

This study begins with a thorough examination of the dataset through exploratory data analysis, identifying distributions, trends, and relationships between variables. This exploratory study not only offers important background information, but also reveals hidden trends and possible anomalies that need more research. Then, in order to guarantee the integrity of the dataset, data preparation tasks are carried out. These tasks include imputing missing values, identifying and handling outliers, and transforming the variables using different types of transformations wherever necessary.

To explore the complexities of passenger satisfaction with airline services, we use a combination of logistic regression models in our thorough analysis, including stepwise logistic regression models with and without polynomial terms. we also plan to incorporate sophisticated methods like *random forests* and *decision trees* in the final report of our study. Our findings could improve customer experiences and point out areas for improvement, which would have a significant impact on the airline sector.

Research Objective

The objective of this study is to provide an analysis of customer satisfaction as it relates to air travel. The study seeks to accomplish the following major goals with the use of a dataset that includes different factors associated with travel experiences: Analyzing the Distribution and Features of Data, and understanding the distribution, patterns, and relationships between the variables in the dataset is the goal of the study. This involves identifying skewness, analyzing the distribution of ordinal and interval variables, and displaying data patterns.

We further make use of box plots to identify outliers in the dataset. Such outliers will be handled during the data preprocessing phase of our study. In order to prepare the dataset for modeling, the research performs data preparation, which involves addressing outliers, handling missing values, and making any necessary changes like variable transformations. Furthermore, we will be generating a correlation heatmap to understand the relationship between variables. This process can reveal which factors might affect customer happiness. The study creates prediction models, such as a logistic regression model with a 70/30 data split, and splitwise selection with and without polynomial terms. In the future, the goal is to perform more prediction models like decision trees and random forests to generate and validate better models for our dataset and present them in our final report.

From this report, we hope to gain important insights into the variables influencing customer happiness in the context of airline travel. Airlines can use these insights to increase overall customer happiness, identify areas for improvement, and improve services. A deeper understanding of passenger experiences in the airline sector is also aided by the study, which helps stakeholders make decisions that will improve customer relations and service quality.

Exploratory Data Analysis

Dataset Description

Variable	Variable Type	Description
Satisfaction	Nominal	Target variable, it represents the satisfaction levels of customers whether 'Satisfied' or 'Dissatisfied'
Gender	Nominal	Represents whether the customer is 'Male' or 'Female'
Customer Type	Nominal	Represents whether the customer is 'Loyal Customer' or a 'Disloyal Customer'
Age	Interval	Represents the age of every customer from 7 to 85
Type of Travel	Nominal	Represents travel type: 'Business' or 'Personal'
Class	Nominal	Represents travel class: 'Business', 'Eco Plus' or 'Eco'
Flight Distance	Interval	Distance covered by the flight
Seat Comfort	Ordinal	Rating from 0 to 5
Departure/Arrival time convenient	Ordinal	Rating from 0 to 5
Food and drink	Ordinal	Rating from 0 to 5
Gate location	Ordinal	Rating from 0 to 5
Inflight wifi service	Ordinal	Rating from 0 to 5
Inflight entertainment	Ordinal	Rating from 0 to 5
Online support	Ordinal	Rating from 0 to 5
Ease of Online booking	Ordinal	Rating from 0 to 5
On-board service	Ordinal	Rating from 0 to 5
Leg room service	Ordinal	Rating from 0 to 5
Baggage handling	Ordinal	Rating from 0 to 5
Checkin service	Ordinal	Rating from 0 to 5
Cleanliness	Ordinal	Rating from 0 to 5
Online Boarding	Ordinal	Rating from 0 to 5
Departure Delay in minutes	Interval	Represents delay in departure time in minutes
Arrival Delay in minutes	Interval	Represents delay in arrival time in minutes

Table 1.

Descriptive Statistics

The dataset is sourced from an airline company and encompasses customer satisfaction data. It includes approximately 129,880 records and 23 variables. Among these are 5 nominal, 14 ordinal, and 4 interval input variables. The variable 'satisfaction' is identified as the target variable, representing two levels of satisfaction, i.e. 'satisfied' and 'dissatisfied'.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	39.42796	15.11936	129880	0	7	40	85	-0.00361	-0.71914
Arrival_Delay_in_Minutes	INPUT	15.09113	38.46565	129487	393	0	0	1584	6.670125	95.11711
Baggage_handling	INPUT	3.695673	1.156483	129880	0	1	4	5	-0.74304	-0.23754
Checkin_service	INPUT	3.340807	1.260582	129880	0	0	3	5	-0.39244	-0.79351
Cleanliness	INPUT	3.705759	1.151774	129880	0	0	4	5	-0.756	-0.20889
Departure_Arrival_time_convenien	INPUT	2.990645	1.527224	129880	0	0	3	5	-0.25228	-1.08937
Departure_Delay_in_Minutes	INPUT	14.70774	37.9694	110398	19482	0	0	1592	6.660143	97.20699
Ease_of_Online_booking	INPUT	3.472105	1.30556	129880	0	0	4	5	-0.49172	-0.91065
Flight_Distance	INPUT	1981.409	1027.116	129880	0	50	1925	6951	0.466748	0.364306
Food_and_drink	INPUT	2.851994	1.443729	129880	0	0	3	5	-0.11681	-0.98673
Gate_location	INPUT	2.990422	1.30597	129880	0	0	3	5	-0.05306	-1.08982
Inflight_entertainment	INPUT	3.383477	1.346059	129880	0	0	4	5	-0.60483	-0.53279
Inflight_wifi_service	INPUT	3.24913	1.318818	129880	0	0	3	5	-0.19112	-1.12145
Leg_room_service	INPUT	3.483895	1.293146	110398	19482	0	4	5	-0.49486	-0.84246
On_board_service	INPUT	3.465075	1.270836	129880	0	0	4	5	-0.50527	-0.78502
Online_boarding	INPUT	3.352587	1.298715	129880	0	0	4	5	-0.3665	-0.93805
Online_support	INPUT	3.519703	1.306511	129880	0	0	4	5	-0.57536	-0.81057
Seat_comfort	INPUT	2.837126	1.393061	110398	19482	0	3	5	-0.09062	-0.94417

Fig. 1

Fig. 1 reveals that the data has missing values for certain variables like Arrival Delay in Minutes, Departure Delay in minutes, Leg room service, and seat comfort. The missing values can be handled during the data-cleaning process. The figure also reveals information on the skewness degrees for each variable. For instance, Arrival Delay in Minutes has the highest skewness of 6.67 indicating a positive skew. This suggests that the data is concentrated towards the left side, implying that the majority of instances experienced shorter wait times, while only a few instances had longer wait times. In contrast, age exhibits a lower skewness of 0.00361 and is also negatively skewed. Additionally, the flight distance and departure delay in minutes have positive skewness while all other variables have negative skewness.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Class	INPUT	3	0	Business	47.86	Eco	44.89
TRAIN	Customer_Type	INPUT	2	0	Loyal Customer	81.69	disloyal Customer	18.31
TRAIN	Gender	INPUT	2	0	Female	50.74	Male	49.26
TRAIN	Type_of_Travel	INPUT	2	0	Business travel	69.06	Personal Travel	30.94
TRAIN	satisfaction	TARGET	2	0	satisfied	54.73	dissatisfied	45.27

Fig. 2

The range of the dataset is defined by the difference between its maximum and minimum values and encompasses the range of values present within the dataset. The mean and standard deviation are calculated on both interval and ordinal variables within the dataset to identify the central tendency and the distribution of data. Furthermore, the mode shown in Fig. 2 is determined for nominal variables to identify the value that occurs most frequently. In our analysis, Flight Distance has a higher standard deviation of 1027.116, so the transformation is required to reduce the variance in the variable. Moreover, the distribution of target variables—including whether or not they are equally represented—is revealed by the StatExplore results. According to the statistics we utilized, 45.27% of the consumers were not satisfied and 54.73% of the customers were satisfied.

Visualizing Data Distributions

Histograms are specifically employed to assess the distribution of interval variables in our data. They are used to identify whether data follows a normal distribution or not.

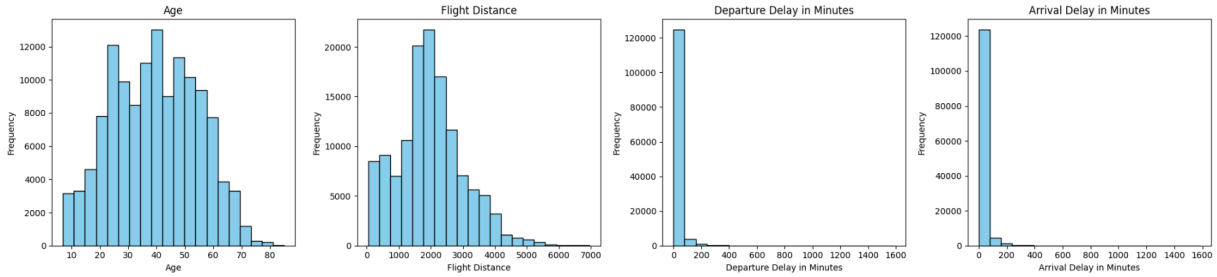

Fig. 3

Age has a relatively normal distribution, indicating a well-balanced range of values. Flight Distance, on the other hand, has a right-skewed distribution, indicating a concentration of shorter distances with a tail extending toward longer ones. Departure Delays in Minutes and Arrival Delays in Minutes both deviate from normal distribution patterns. The majority values in both of these variables are concentrated within a narrow range, exhibiting a rightward skew. It is worth noting that the distribution of these two variables is quite similar.

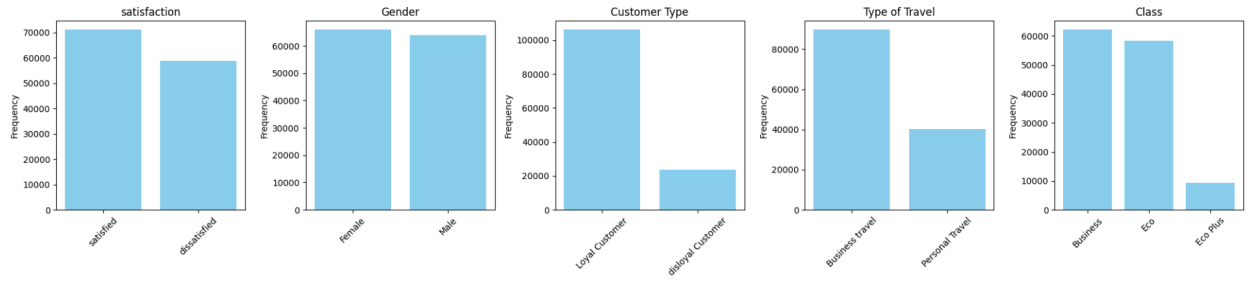

Fig. 4

Satisfaction, Gender, Customer Type, Type of travel, and Class fall under the Nominal class of variables. Bar charts were employed to showcase data distribution across distinct categories of each variable (see Fig. 4).

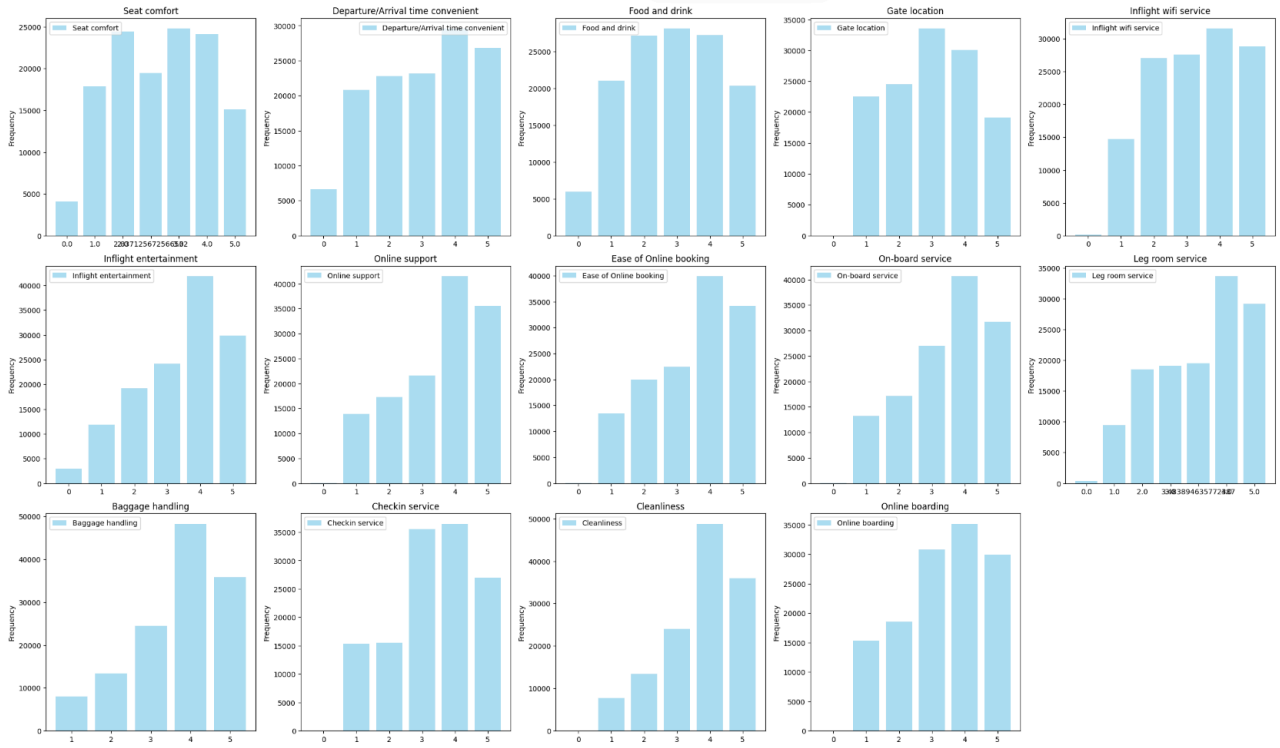

Fig. 5

Bar charts, similar to a histogram, were employed to visualize the ordinal variables. Each bar represents the frequency of instances within a specific rank, making it easy to compare different levels. All ordinal variables in the dataset have ratings ranging from 0 to 5 (See Fig. 5).

Visualizing Outliers

Box Plots are an effective tool for detecting outliers in a dataset. Outliers can be found in variables such as flight distance, departure delay in minutes, arrival delay in minutes, on-board service, and leg room service. The concept of standard deviation for outliers, as seen in interval data, may not apply to ordinal variables such as on-board service, leg room service, or check-in service.

Categories with frequencies significantly outside the norm are flagged as outliers in these cases, as shown in Fig. 6. Meanwhile, numerous data points have been identified as outliers in

interval variables such as flight distance, departure delay in minutes, and arrival delay in minutes, warranting data cleaning consideration.

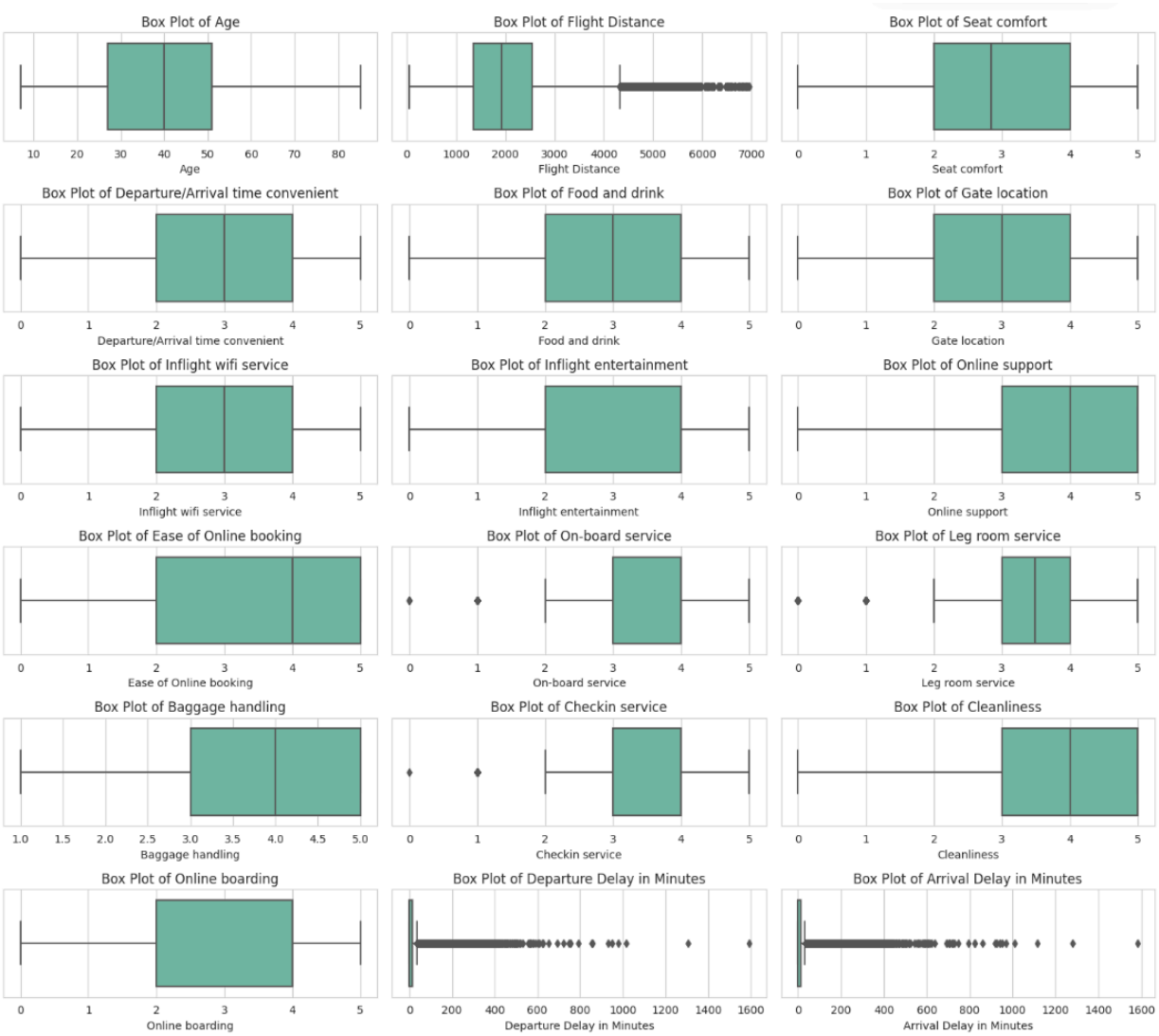

Fig. 6

Data Preprocessing:

We begin the data preprocessing steps by imputation of missing values. In this case, we have employed the median to address the skewed and outlier-prone data and imputed four different variables. This ensures that our dataset remains complete and credible for further analyses. We employed the Interquartile Range (IQR) approach to detect outliers, notably in our ordinal data, which provides a strong indication of data points that deviate from the central tendency. We proceeded to execute appropriate transformations such as logarithmic, square root, and Box-Cox transformations. This was done to reduce the impact of influencing outliers for better performance of the models.

Correlation Analysis:

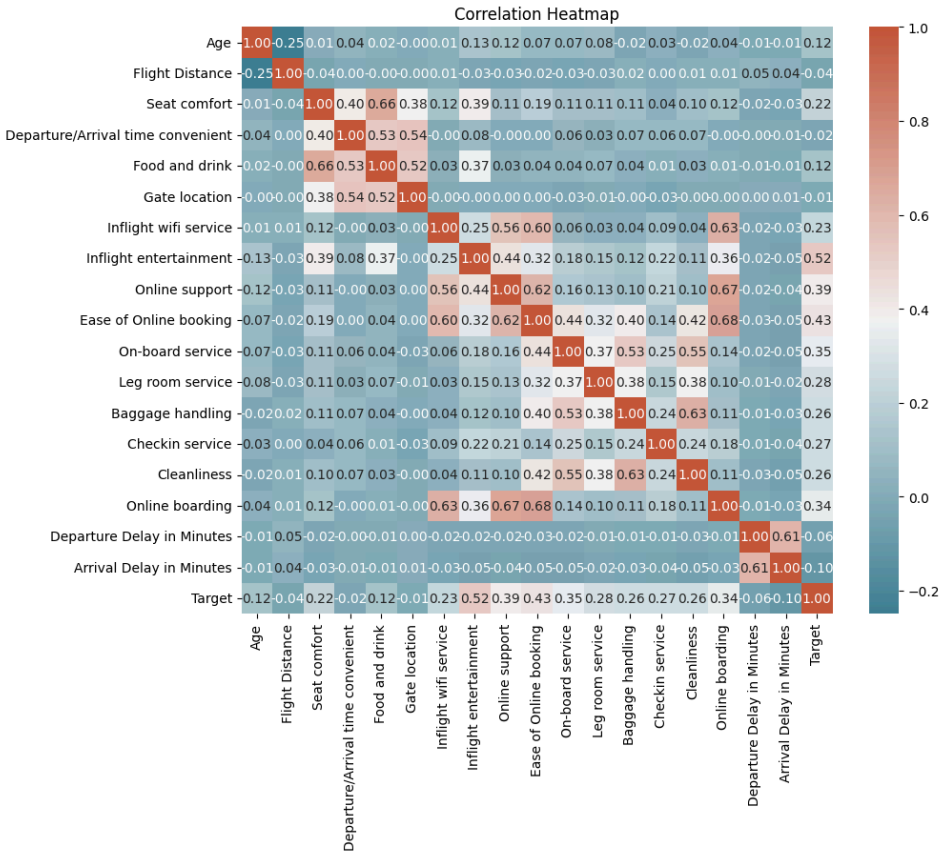

Fig. 7

Prediction Models

Model 1: Logistic Regression

Using every variable that was available after the data cleaning, we performed a logistic regression with a 70/30 data split in the first model to determine how much each variable affected customer satisfaction. Without any variable selection, this functioned as the baseline model. We evaluate the importance of individual predictor variables in our logistic regression model. Understanding the impact of each variable on the target variable requires an understanding of the variable's significance. Using p-values to quantify variable significance helps us determine whether the predictor variable has a statistically significant impact on the target variable. As demonstrated in Fig. 8, this model demonstrates a strong overall statistical significance, with a p-value of less than 0.0001.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
125219.278	70079.112	55140.1661	22	<.0001

Fig. 8

However, the model includes a variable that is not statistically significant and may affect the performance of the model. Hence, it is suggested to remove this variable and re-run the model. Specifically, we observed that the p-value of the variable 'Departure_delay_in_minutes' is found to be 0.49. This p-value holds significance implications regarding the variables's statistical significance. Following the removal of this variable, we re-ran the model. As shown in Fig. 9, the overall performance of the model was significant with all parameters demonstrating statistical significance. Therefore, the model can be considered as a valid option.

Analysis of Maximum Likelihood Estimates

Parameter		satisfaction	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		satisfied	1	-7.1829	0.0760	8938.60	<.0001		0.001
Age		satisfied	1	-0.00846	0.000683	153.60	<.0001	-0.0705	0.992
Arrival_Delay_in_Minutes		satisfied	1	-0.3107	0.0155	403.13	<.0001	-0.1329	0.733
Baggage_handling		satisfied	1	0.1210	0.0110	120.92	<.0001	0.0774	1.129
Checkin_service		satisfied	1	0.2946	0.00826	1271.48	<.0001	0.2051	1.343
Class	Business	satisfied	1	0.4866	0.0180	731.30	<.0001		1.627
Class	Eco	satisfied	1	-0.2177	0.0164	175.20	<.0001		0.804
Cleanliness		satisfied	1	0.0880	0.0115	58.83	<.0001	0.0560	1.092
Customer_Type	Loyal Customer	satisfied	1	0.9973	0.0150	4439.59	<.0001		2.711
Departure_Arrival_time_convenien		satisfied	1	-0.1880	0.00809	540.16	<.0001	-0.1583	0.829
Departure_Delay_in_Minutes		satisfied	1	-0.0139	0.0204	0.46	0.4973	-0.00448	0.986
Ease_of_Online_booking		satisfied	1	0.2554	0.0138	343.82	<.0001	0.1840	1.291
Flight_Distance		satisfied	1	-0.00013	0.000010	152.54	<.0001	-0.0713	1.000
Food_and_drink		satisfied	1	-0.1581	0.0106	221.22	<.0001	-0.1259	0.854
Gate_location		satisfied	1	0.1052	0.00912	132.94	<.0001	0.0759	1.111
Gender	Female	satisfied	1	0.4879	0.00982	2466.53	<.0001		1.629
Inflight_entertainment		satisfied	1	0.7052	0.00984	5138.54	<.0001	0.5232	2.024
Inflight_wifi_service		satisfied	1	-0.0723	0.0105	47.18	<.0001	-0.0527	0.930
Leg_room_service		satisfied	1	0.2035	0.00871	546.31	<.0001	0.1353	1.226
On_board_service		satisfied	1	0.3129	0.00983	1014.26	<.0001	0.2196	1.367
Online_boarding		satisfied	1	0.1666	0.0118	199.11	<.0001	0.1194	1.181
Online_support		satisfied	1	0.0847	0.0108	61.72	<.0001	0.0611	1.088
Seat_comfort		satisfied	1	0.2297	0.0108	449.53	<.0001	0.1627	1.258
Type_of_Travel	Business travel	satisfied	1	0.3988	0.0139	818.33	<.0001		1.490

Fig. 9

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
satisfaction	satisfaction	_DFM_	Model Degrees...	23	.
satisfaction	satisfaction	_DFT_	Total Degrees ...	90915	.
satisfaction	satisfaction	_DIV_	Divisor for ASE	181830	77930
satisfaction	satisfaction	_ERR_	Error Function	70079.11	30136.86
satisfaction	satisfaction	_FPE_	Final Prediction...	0.119463	.
satisfaction	satisfaction	_MAX_	Maximum Abso...	0.998669	0.998044
satisfaction	satisfaction	_MSE_	Mean Square E...	0.119433	0.119651
satisfaction	satisfaction	_NOBS_	Sum of Freque...	90915	38965
satisfaction	satisfaction	_NW_	Number of Esti...	23	.
satisfaction	satisfaction	_RASE_	Root Average S...	0.345546	0.345906
satisfaction	satisfaction	_RFPE_	Root Final Pred...	0.345634	.
satisfaction	satisfaction	_RMSE_	Root Mean Squ...	0.34559	0.345906
satisfaction	satisfaction	_SBC_	Schwarz's Baye...	70341.72	.
satisfaction	satisfaction	_SSE_	Sum of Square...	21710.93	9324.416
satisfaction	satisfaction	_SUMW_	Sum of Case W...	181830	77930
satisfaction	satisfaction	_MISC_	Misclassificatio...	0.165759	0.16733

Fig. 10

The validation set accuracy is 1-0.16733 = 83%, whereas the training set accuracy is 1-0.165759 = 82%. This 82% accuracy suggests that our model accurately predicted customer satisfaction in the majority of cases. In the given case, the accuracy of the validation dataset and the training dataset shows a very slight difference, which does not strongly indicate the overfitting of the model.

Model 2: Logistic Regression with Stepwise Selection

Model 2 is presented in this phase of the project, which employs logistic regression utilizing stepwise variable selection. This model aims to improve on the Model 1 forecast of airline customer satisfaction. Stepwise variable selection is an approach for selecting the most relevant variables for our logistic regression model in a systematic way. It is an iterative procedure in which predictor variables are added or removed based on their statistical significance. As shown in Fig. 11, the overall performance of the model is statistically significant as the p-value is less than 0.0001 and the variables are also significant.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
125219.278	70079.112	55140.1661	22	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		satisfaction	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		satisfied	1	-7.1840	0.0760	8945.10	<.0001		0.001
Age		satisfied	1	-0.00846	0.000683	153.68	<.0001	-0.0705	0.992
Arrival_Delay_in_Minutes		satisfied	1	-0.3171	0.0123	661.27	<.0001	-0.1356	0.728
Baggage_handling		satisfied	1	0.1210	0.0110	120.81	<.0001	0.0774	1.129
Checkin_service		satisfied	1	0.2946	0.00826	1271.38	<.0001	0.2051	1.343
Class	Business	satisfied	1	0.4865	0.0180	731.19	<.0001		1.627
Class	Eco	satisfied	1	-0.2176	0.0164	175.04	<.0001		0.804
Cleanliness		satisfied	1	0.0880	0.0115	58.86	<.0001	0.0560	1.092
Customer_Type	Loyal Customer	satisfied	1	0.9974	0.0150	4441.23	<.0001		2.711
Departure_Arrival_time_convenien		satisfied	1	-0.1880	0.00809	540.23	<.0001	-0.1583	0.829
Ease_of_Online_booking		satisfied	1	0.2554	0.0138	344.03	<.0001	0.1840	1.291
Flight_Distance		satisfied	1	-0.00013	0.000010	153.22	<.0001	-0.0714	1.000
Food_and_drink		satisfied	1	-0.1581	0.0106	221.10	<.0001	-0.1258	0.854
Gate_location		satisfied	1	0.1052	0.00912	132.92	<.0001	0.0759	1.111
Gender	Female	satisfied	1	0.4879	0.00982	2466.48	<.0001		1.629
Inflight_entertainment		satisfied	1	0.7052	0.00984	5138.64	<.0001	0.5232	2.024
Inflight_wifi_service		satisfied	1	-0.0723	0.0105	47.21	<.0001	-0.0527	0.930
Leg_room_service		satisfied	1	0.2035	0.00871	546.13	<.0001	0.1353	1.226
On_board_service		satisfied	1	0.3129	0.00983	1014.20	<.0001	0.2196	1.367
Online_boarding		satisfied	1	0.1666	0.0118	199.04	<.0001	0.1194	1.181
Online_support		satisfied	1	0.0847	0.0108	61.70	<.0001	0.0610	1.088
Seat_comfort		satisfied	1	0.2296	0.0108	449.41	<.0001	0.1627	1.258
Type_of_Travel	Business travel	satisfied	1	0.3989	0.0139	819.15	<.0001		1.490

Fig. 11

The model has chosen to retain all variables with p-values less than 0.0001, indicating their statistically significant contribution to predicting customer satisfaction. Model 2 exhibits a validation accuracy of 83.26 % and a training accuracy of 83.42% and we conclude that the model is not overfitting as there is a minimal difference between the two sets of data.

Model 3: Logistic Regression with Stepwise Selection and Polynomial Degree 2

In model 3, we implement stepwise variable selection to identify significant predictor variables. To account for potential non-linear relationships between variables, polynomial terms are introduced. These allow the model to capture curvilinear effects, potentially improving the representation of the data. Running this regression model resulted in a p-value less than 0.0001, as seen in Fig. 12, indicating that the model is statistically significant. Further analyzing the output results, we noticed that all variables were statistically significant as well. The model chooses variables that have a p-value less than 0.05. With a validation set accuracy of 90% and a training set accuracy of 89%, it can be concluded that the model is not overfitting.

However, this model yielded 65 different combination variables, each of them statistically significant. While under normal circumstances, we would consider such a model to be valid for making predictions, however in this case the sheer volume of variables renders the model very complex.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood		
Intercept Only	Intercept & Covariates	Ratio Chi-Square	DF	Pr > ChiSq
125219.278	44378.785	80840.4928	66	<.0001

Fig. 12

Conclusion

In this project report, we studied the dataset of an anonymous airlines company encompassing data on passenger satisfaction. The application of prediction models based on logistic regression, after a thorough process of data preparation and exploratory data analysis, has provided some valuable insights into the factors that influence passenger satisfaction. Airlines, analysts, and other stakeholders can use these results to improve the passenger experience, decide on service enhancements with greater knowledge, and forge closer bonds with their customers.

In this report, the team has executed our initial algorithm consisting of three models–the first model being a basic logistic regression, the second model employing logistic regression using stepwise selection, and the third model implementing logistic regression with stepwise selection and polynomial degree terms of 2. Initial analysis concludes that all the three models are valid without exhibiting any overfitting concerns. Furthermore, the third model involves a high amount of variables and their combinations for making predictions as compared to the other models. However, we are not selecting either of these models as we do not consider them the optimal choice yet. The project has advanced significantly, as we understood our data and run some initial prediction models yielding us some results to work with. Moving forward, we intend to incorporate prediction models such as Decision Trees, Random Forests, and additional algorithms to identify the most suitable model for our dataset.

References

Bogicevic, V., Yang, W., Bujisic, M., & Bilgihan, A. (2017). Visual Data Mining: Analysis of airline service quality attributes. *Journal of Quality Assurance in Hospitality & Tourism*, *18*(4), 509–530. https://doi.org/10.1080/1528008x.2017.1314799