MGT 6203 Group Project Final Report - Summer 2023

Team 6: Predicting Customer Satisfaction in the Airline Industry

Members: Joshua Farina (joshuapfarina), Henri Salomon (henri_salomon), Raajitha Middi (Raajitha_Middi), Ryan Chandler (zujin87), Alejandro Martinez (amzeta)

1. Introduction/Background/Motivation

1.1. Objective/ Problem statement

We intend to use airline passenger satisfaction survey data to determine which factors drive customer satisfaction (or dissatisfaction) so that we can provide actionable recommendations on how money should be invested or re-allocated to keep passengers satisfied. Customer satisfaction is a key factor in attracting and retaining business. Identifying factors that have the strongest effect on customer satisfaction will provide key insights for airlines to enhance their services, improve the customer experience and optimize business operations.

1.2. Business Justification/ Impact

The airline industry has undergone significant changes with the emergence of hybrid business models, blurring the line between full-service and low-cost carriers, and facing financial difficulties stemming from mismanagement or external factors like fluctuating oil prices. To thrive in this competitive landscape, enhancing air passenger satisfaction is crucial for companies as it can positively impact both revenues and costs:

- Firstly, airlines can strive for operational efficiency while addressing individualized customer needs to elevate satisfaction levels and minimize unnecessary expenditures, offering better "value for money."
- Secondly, boosting customer satisfaction leads to reduced customer churn and improved loyalty. Estimates suggest that a mere 5% increase in customer retention can boost profits by 25%-95% (Reichheld and Sasser). Analyzing key factors that influence customer satisfaction enables airlines to identify specific areas for improvement that can reduce customer churn and ultimately lead to higher profitability.
- Lastly, enhancing customer experience contributes to building a strong brand image and reputation in marketing. When passengers have positive experiences, they are more inclined to share them through word of mouth and social media, thereby attracting new passengers and further bolstering the airline's success.

1.3. Research questions

Our primary research question (RQ) was: *What are the most important factors in ensuring that an airline passenger is satisfied*? Other supporting research questions were:

1. Does grouping variables by category (demographics, in-flight service quality, timeliness and delays) provide additional insight?
2. Can we predict airline passenger satisfaction based on a limited set of factors?
3. Do characteristics which would imply a higher level of service (business class) result in higher levels of satisfaction?
4. Are there any noteworthy interactions which would affect the probability of a satisfied customer?

1.4. Hypotheses

Our first hypothesis was that certain factors would have a significant impact on passenger satisfaction while others may not. The following factors had been predicted as the most significant: flight distance, inflight wifi service, departure/arrival time convenience, food and drink, seat comfort, inflight entertainment, on-board service, legroom service, inflight service, departure delays in minutes, and arrival delays in minutes. Additionally, we hypothesized that the relevance of these factors will vary depending on the class, type of travel, demographic factors (e.g., age, gender), and flight distance.

1.5. Methodology/ Approach

Our methodology encompassed the following steps. First, we split the data set into training, validation, and testing sets (60/20/20). After performing data cleaning operations, we conducted Exploratory Data Analysis (EDA) to gain insights into the data set's characteristics.

Secondly, we narrowed down the focus of our modeling to: i) Logistic Regression, ii) Decision tree, iii) Random Forest, and iv) Support Vector Machines (SVM). For logistic regression, we systematically tested variables individually, assessed their distributions, and considered transformations if needed. In the case of decision trees, we constructed a hierarchical structure that enables effective prediction and interpretation. Random forests were utilized to evaluate feature importance. SVM was included to explore it's potential applications and to develop deeper insight into our dataset.

Thirdly, throughout the analysis, we interpreted the results of each model, examined the significance of variables and drew conclusions regarding their impact on passenger satisfaction. We compared the performance of different models based on evaluation metrics, such as accuracy and Receiver Operating Characteristic (ROC) curve and computing the Area Under the Curve (AUC). Additionally, we considered insights gained from EDA and statistical tests to support our conclusions.

Finally, upon selecting the most suitable model, we interpreted its predictions and drew conclusions regarding the factors contributing to passenger satisfaction. We considered the overall model performance and the importance of different features in influencing satisfaction levels. By following this methodology (detailed in Appendix 4), we aimed to gain a comprehensive understanding of the data set, select the most appropriate model, interpret its results, and draw meaningful conclusions about the factors affecting air passenger satisfaction.

1.6. Literature review

In our research project, we reviewed several relevant research papers. Pereira et al. (2023) used a text mining tool to analyze written reviews and identified the critical features influencing passenger satisfaction. They found that staff behavior, employee attitude towards dissatisfied customers, booking and cancellation policies, baggage handling, seat quality, boarding, check-in, customer service, and food and drink were important factors. Their study concluded that the COVID-19 pandemic increased the importance of cleanliness and refunds, which is not applicable to our pre-pandemic data set.

Another study by Lucini et al. (2020) used Latent Dirichlet Allocation to analyze customer reviews and found similar factors to be influential, including cabin staff, onboard service, value for money, and seats. They also highlighted the importance of segmentations such as nationality and cabin flown. The practical implications to maximize customer satisfaction by cabin flown are: focus on customer service for first class passengers, comfort for premium economy passengers, and checking luggage and waiting time for economy class travelers).

Lastly, Ouf (2023) conducted an optimized deep learning analysis using the same data set we are using. They found that factors such as online boarding, class, type of travel, and in-flight entertainment were important.

It is worth noting that our data set may have limitations, as critical factors such as cabin crew quality were not considered and there may be biases due to customers not taking the questionnaires seriously. Incorporating approaches like text mining could enhance our understanding of passenger satisfaction by capturing a wider range of factors and reducing noise in the data.

2. Overview of data

2.1. Overview of data sets

Our primary data set is taken from surveys from a US airline and is publicly available on Kaggle (see details in *7. Work cited and reference*). It includes user ratings for multiple aspects of air travel, such as food and drink and seat comfort, as well as customer information, such as age and gender. It contains 22 features (see details in Appendix 1) and is made of 2 different files of 103,904 (train.csv) and 25,976 records (test.csv).

We initially intended to use other sources of data in relation to the business problem, but after in-depth analyses, we concluded that i) merging with our primary data set was impossible, or ii) conducting a parallel analysis of customer satisfaction was not feasible for the identified data sets:

- ACSI data set, containing customer satisfaction scores for 4 industries, incl. airline industry
- Flight route data set, containing flight route information and airport code
- British Airways review data set, containing scraped information from a website

2.2. Data cleaning process

While the primary dataset was of high-quality, we did need to take some steps to pre-process it. To ensure that our approach consistently and uniformly applied, we developed a pre-processing script to be called at the start of each piece of analysis.

HANDLING MISSING VALUES. There were two cases of missing values. Firstly, in the ordinal (Likert scale) variables, some rows had 0 (“Not applicable”) when passengers didn’t rate the respective variables. As they accounted for less than 5% of the total data for each variable, we replaced those values with their corresponding mode values. To eliminate bias, modes were derived using the training data set only. Secondly, the variable *Arrival.Delay.in.Minutes* has 310 missing values. The high correlation between arrival delay and departure delay supported our assumption that the latter significantly influences the former, prompting us to drop *Arrival.Delay.in.Minutes* and consider only *Departure.Delay.in.Minutes* as the preferred feature.

DROPPING UNNECESSARY COLUMNS. We dropped the ID column as it is irrelevant for analysis.

TRANSFORMING VARIABLE TYPES. Specific columns were transformed using appropriate techniques. The *Age* column was converted to a numeric format, while the Likert scale columns into factors, which enabled categorical representation and facilitated subsequent analysis. The *Satisfaction* column was transformed into binary values, with "satisfied" being encoded as 1 and other values as 0.

LOG TRANSFORMATIONS. Log transformations were used on *Flight.Distance* and *Departure.Delay.in.Minutes* to address skewed distributions, normalize the data, and mitigate the impact of extreme values.

OUTLIERS. While the InterQuartile Range (IQR) method identified outliers only in the log transformed *Departure.delay.in.Minutes* and not in *Flight.Distance*, we decided to keep these values because they represented less than 0.05% of the data set and they could represent real situations where a flight would be delayed by 6 or more hours.

TRAINING, VALIDATION, AND TEST DATA SETS. We decided to divide the entire data set into training, validation and test data sets with a ratio of 60:20:20 (while we kept the original test.csv file as the test set, we divided the train.csv file into training and validation sets). Models were trained on the training set. We then used the validation set to make sure that the models generalized well and conducted hyperparameter tuning if necessary. The best model was chosen based on performance on validation data.

2.3. *Insights from Exploratory Data Analysis (EDA)*

CORRELATION ANALYSIS: We performed different correlation analyses. Firstly, we focused on numeric variables, with the below correlation heatmap. This shows that *Arrival.Delay.in.Minutes* and *Departure.Delay.in.Minutes* are highly correlated (refer to 2.2 for actions undertaken as a result).

We generated a heat map for the categorical variables also. Most values are not highly correlated (<0.7) but there are a few higher correlation values for *Cleanliness* to *Food.and.drink, Inflight.entertainment,* and *Seat.comfort,* as well as between *Inflight.wifi.service* and *Ease.of.Online.booking.*

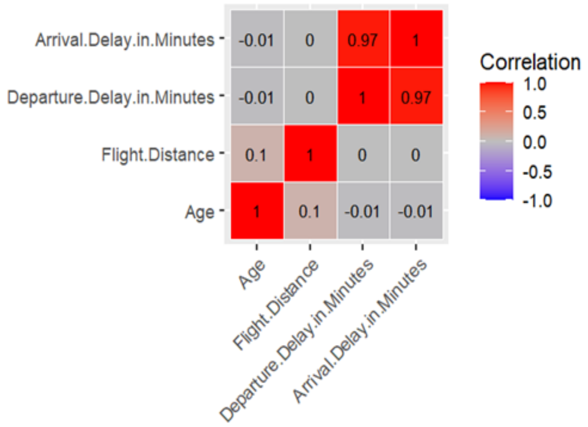


Figure 1: Heatmap for numeric variables

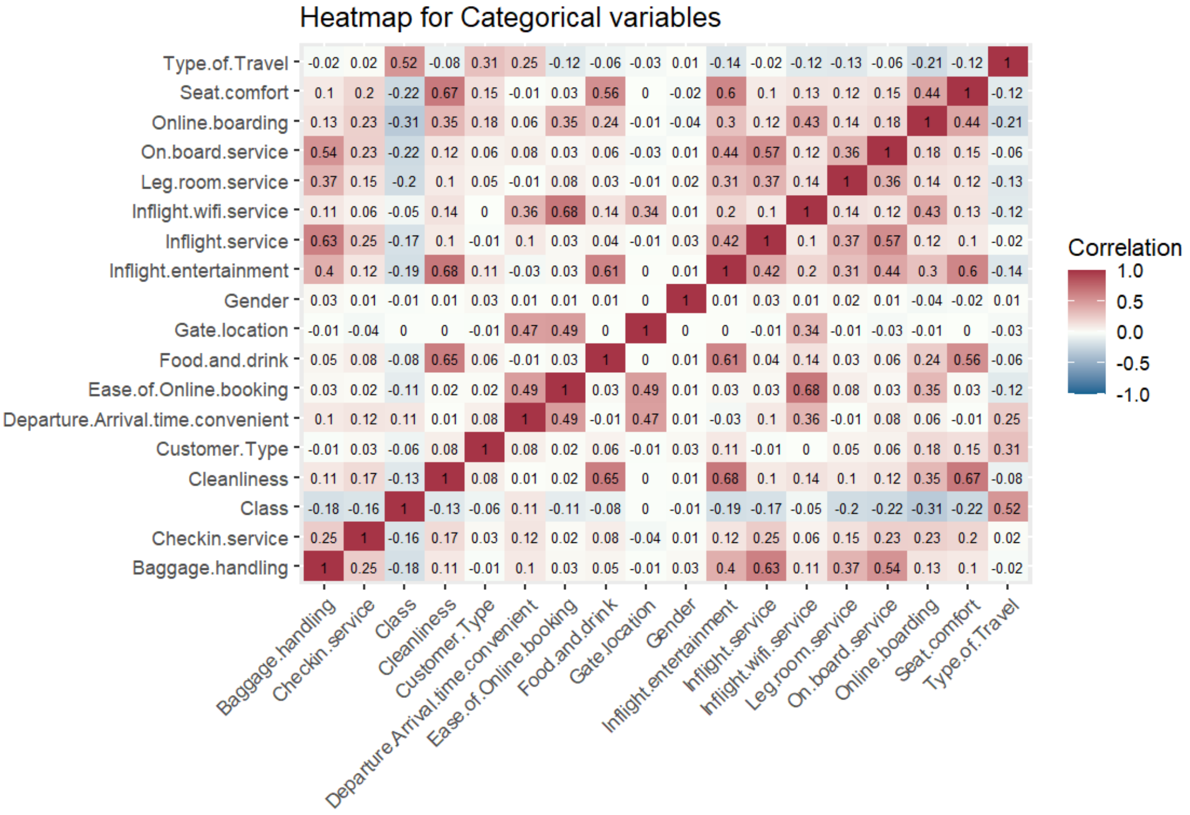


Figure 2: Heatmap for categorical variables

DATA VISUALIZATION: We also looked at the visualizations of categorical variables with respect to the response variable. Few key observations from the plots (Figure 3 and Appendix 3):

- Gender doesn't seem to play a major role as both have almost the same satisfaction levels.
- People in the age group of approximately 20-35 seem more dissatisfied than any other age group; interestingly, people in this age group are more likely to be disloyal.
- Business class passengers have higher satisfaction levels, particularly in *Checkin.service*, *Seat.Comfort* and *Food.and.drink* compared to the Eco plus and Economy class (Appendix 3)

- Surprisingly, even with ratings of 4 and 5, *cleanliness* does not translate to passenger satisfaction.
- Passengers express greater dissatisfaction with services like *Inflight.wifi.service*, *Inflight.entertainment*, *Baggage.handling*, *On.board.service*, *Inflight.service*. Only when the ratings of these services are above 3, there are more passengers who are satisfied, but then this number doesn't seem to be significantly large.

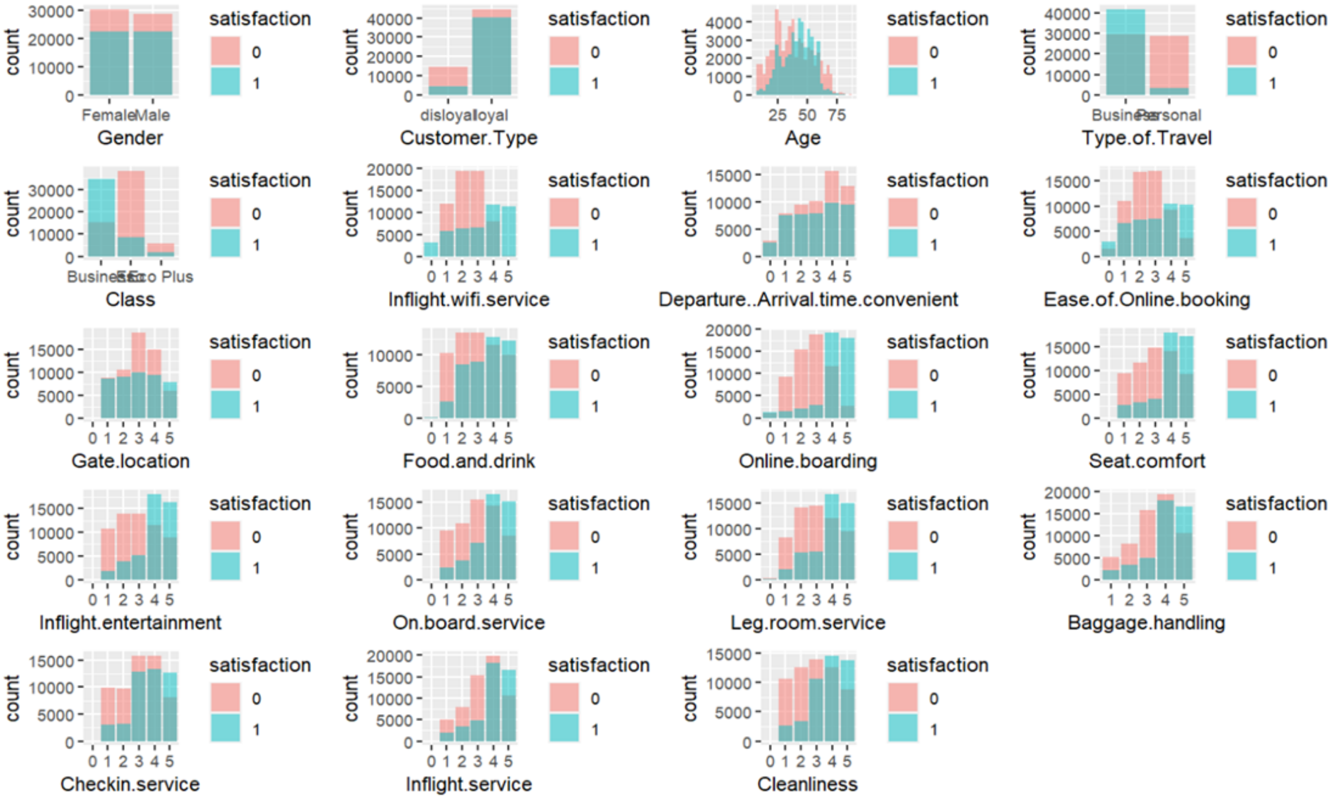


Figure 3: Categorical Variables (except Age) against response variable

NUMERIC VARIABLES AND NORMAL DISTRIBUTION. Among the numeric variables, *Flight.Distance* ranges from 31 to 4983, while *Departure.Delay.in.Minutes* ranges from 0 to 1592. We conducted visualizations using QQ plots, histograms, and box plots after applying various transformations like log, square root, and box-cox. Although none of the resulting transformations yielded a normal distribution, applying a log transform significantly reduced skew and seemed the most appropriate.

3. Overview of Modelling

3.1. Types of model used

Afterward, we evaluated and compared various models using their AUC on a validation set (trained on the training set). Among these models, logistic regression with forward selection achieved the highest AUC on the validation set.

Model	Variables	Accuracy & other metrics (data set)
Logistic regression + Forward selection	Online.boarding4 and 5 (+), Type.of.Travel Personal Travel (-), Inflight.wifi.service4 and 5 (+), Customer.TypeLoyal (+)	● Accuracy: 91.9% (validation) ● Validation AUC: 0.971 ● Test AUC: 0.972
Random Forest	Checkin.service, Inflight.wifi.service, Type.of.Travel, Seat.comfort, Baggage.handling, Online.boarding, Customer.Type	● Accuracy: 91.2% (validation) ● AUC: 0.959 ● Out-Of-Bag Error: 8.94%
Decision Tree (pruned)	Online.boarding, Inflight.wifi.service, Type.of.Travel	● Accuracy: 86.2% (Validation)

		● AUC: 0.886
SVM	Inflight.wifi.service5 (+), Type.of.Travel Personal Travel (-), Customer.TypeLoyal (+), Online.boarding5 (+)	● Accuracy: 92.2% (validation) ● AUC: 0.969

3.2. Model performance, comparison and optimization

3.2.1. Logistic regression

3.2.1.1. *Forward Selection.* To understand the important features to be included in the modeling, we used forward selection to choose the factors (Appendix 5). Except *Flight.Distance*, forward selection chose all the features. We built a logistic regression model to check if there were still any insignificant variables.

3.2.1.2. *Logistic Regression.*

First, we built the model with all factors except flight distance. For Likert scale variables, the base class was chosen to be 1 by the model. Few classes of the categorical variables, such as *Inflight.wifi.service_2*, *On.board.service_2*, turned out to be insignificant (p-value>0.05). This suggests that there is no significant difference between the levels 1 and 2.

Therefore, we removed those factor levels 1 and 2 from the model making both levels as base case. We ran the model again and we repeated the process of removing variables and running the model until we got all the variables to be significant (for a summary of Logistic Regression, refer to Appendix 6.2).

The final model has a validation accuracy of 91.9% and an area under ROC curve of 97.1%. We checked the variance inflation factor (VIF) values for all the variables in the final model (Appendix 6.2).

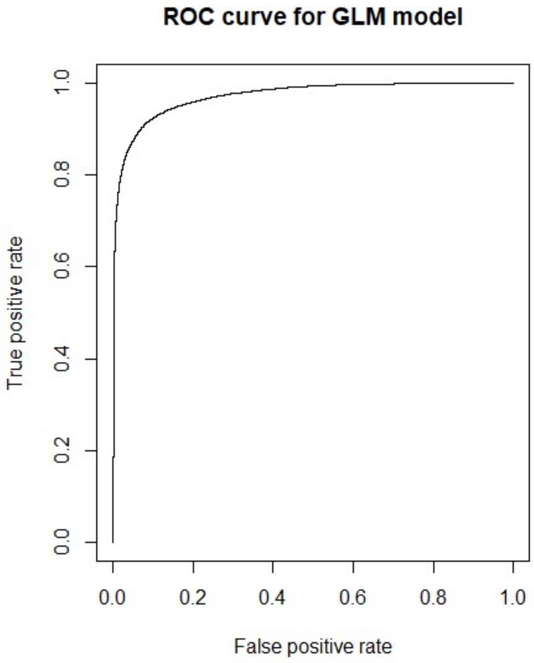


Figure 4: ROC Curve for Logistic Regression

3.2.1.3. *Insights from the model.* The top 5 factors as per the forward selection are *Online.boarding*, *Type.of.Travel*, *Inflight.wifi.service*, *Customer.Type*, *Inflight.service*. All the classes of these variables are significant in the final logit model. The difference between null deviance and the residual deviance is huge. Null deviance is when the model has only intercept and residual deviance is when the model has all factors. The greater the difference, the better our model is, which implies the model is a strong candidate.

3.2.2. Random Forest

3.2.2.1. The primary random forest factors were decided using each parameter's importance. The random forest model's most important factors are *Checkin.service*, *Inflight.wifi.service*, *Type.of.Travel*, *Seat.comfort* with *Baggage.handling*, *Online.boarding*, and *Customer.Type* having very similar importance levels. These factors were chosen as they all have an importance of 60 or higher. With these 7 factors, the accuracy of the model is 91.2%.

3.2.2.2. Optimizing the model included plotting the error convergence and testing how many predictors to try at each split. It was found that the error had converged around 100 trees, so 250 were selected to ensure that any anomalies would be captured. The best number of factors to test at each split was 13 but the out-of-bag error rate decreased by less than 1% as compared to the default setting of 4. Due to the amount of time it took to generate the model, it was decided to use the default setting of 4. By reducing the number of trees down to 250 and leaving the number of factors to try at each split at 4, the processing time was reduced from over 60 seconds down to less than 20 seconds with minimal reduction in model quality.

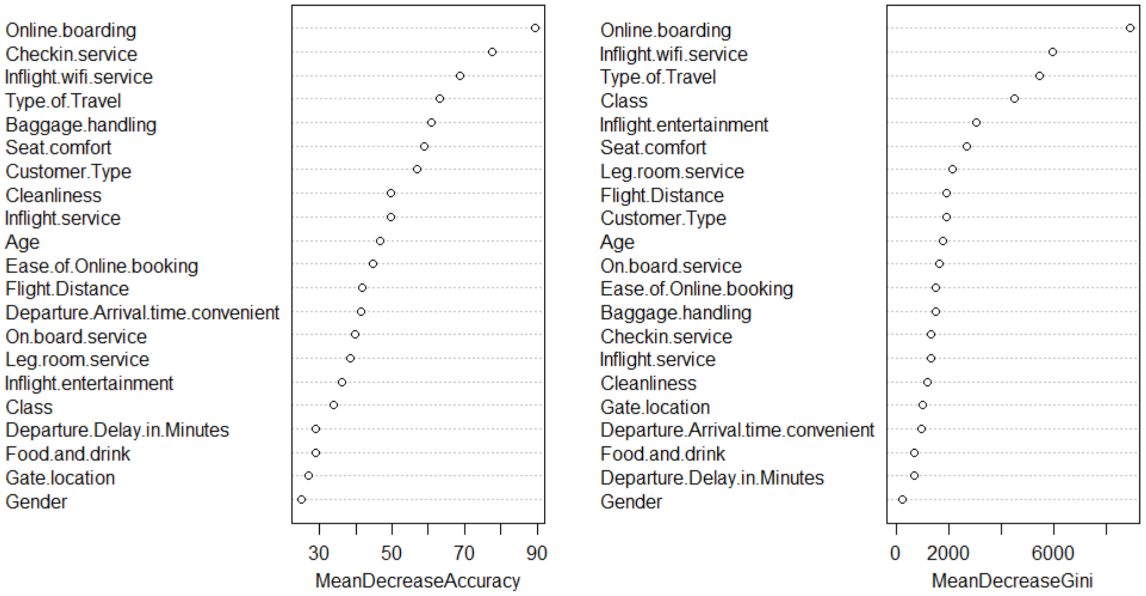


Figure 5: Variable Importance for random forest model

3.2.3. Decision Tree

3.2.3.1. The decision tree model provides a fairly easy to explain model. With an accuracy of 86.2%, it provides good insight and predictability of overall customer satisfaction. Based on the decision tree model, the top factors are *Online.boarding*, *Inflight.wifi.service*, and *Type.of.Travel*. The model shows that a score of 3 or below for *Online.boarding* and *Inflight.wifi.service* generally leads to an unsatisfied customer indicating that a focus on connectivity and easy online boarding will lead to an overall higher number of customers being satisfied.

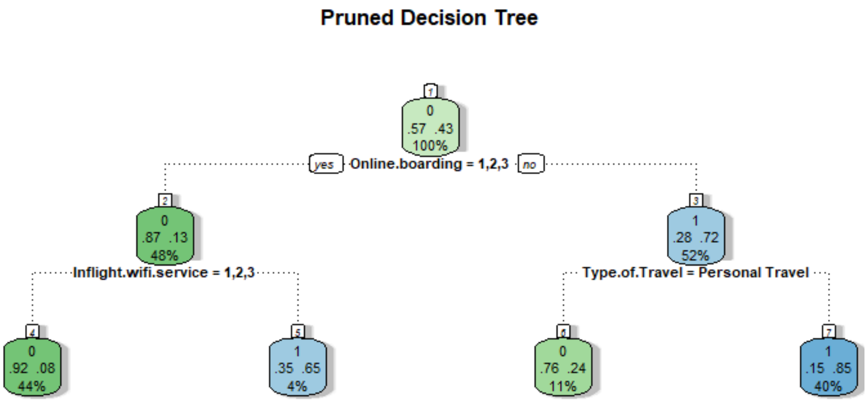


Figure 6: Pruned decision tree

3.2.3.2. Models exploring a single *Class* or *Type.of.Travel* were not as accurate as the "all data" model overall but did have a few insights. The Business Type of travel has an accuracy of 74.3% with a pruned tree of 4 splits as compared to the original model with 7 splits with an accuracy of 76.2%. It seems that Business travelers have a higher expectation for *Inflight.entertainment* than Personal travelers. The expectation of *Inflight.entertainment* is reinforced when grouping Business Class travelers. It should also be noted that there are over twice as many business travelers than personal travelers.

3.2.3.3. The initial decision tree generated required very little pruning. The only pruning required was the output of the "Yes" side of the *Type.of.Travel* split. The reason for this pruning was that the split below *Type.of.Travel* resulted in one of the responses only containing 1% of the overall data. This indicated that the model was potentially overfitting the training data set.

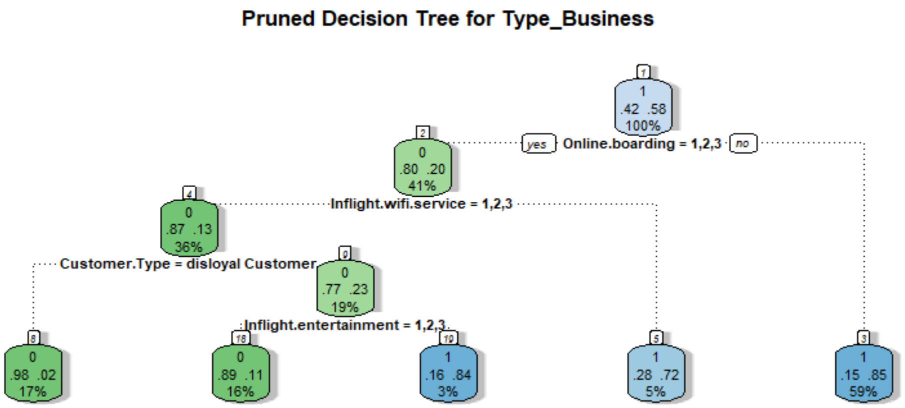


Figure 7: Pruned Decision for Business trips

3.2.4. Support Vector Machines (SVM)

3.2.4.1. We adopted a sub-sampling approach in training our SVM, utilizing a small section (15,000 randomly selected observations) of the overall training set. The reason for this limitation was SVM's known tendency to scale poorly to larger data sets. For the kernel, we selected a linear kernel. for its interpretability relative to other kernels.

3.2.4.2. Models were compared by varying the penalty parameter, C, within a range of 0.001 to 100. The results demonstrated a trend of improving accuracy and AUC up to approximately C=1. Beyond this point, further increases in C yielded only marginal enhancements, while the time to train became prohibitive.

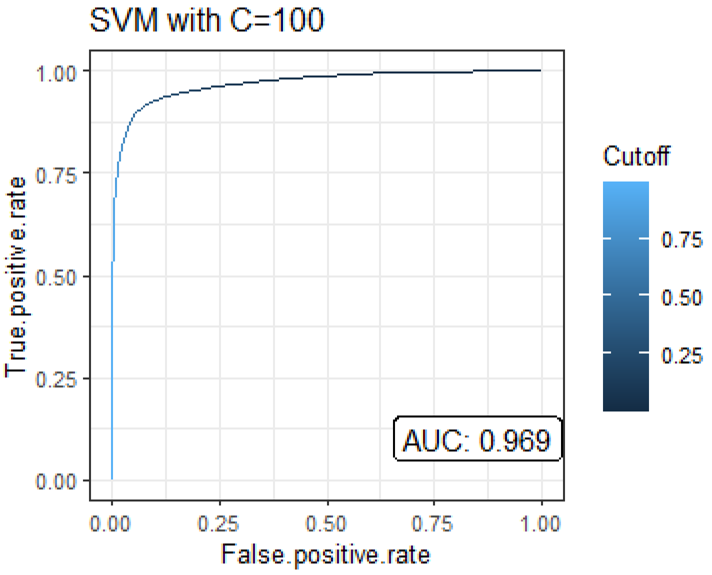


Figure 8: ROC Curve for SVM model

3.2.4.3. The strong results show that our data is linearly separable by the SVM model. Extracting the coefficients shows that Inflight WiFi Service = 5, Type of Travel, and Customer Type to be the most important factors. Remaining factors show a gradual decrease in relevance (Appendix 7) .

4. Discussion & interpretation of results

Q1: Does grouping variables by category (demographics, in-flight service quality, timeliness and delays) provide additional insight?

Grouping customers by *Class* and *Type.of.Travel* indicated that, for the most part, most travelers had similar requirements for satisfaction. *Seat.comfort* seems to be of some concern for Business Class customers (not to be confused with Business Travelers), but it was not a strong indicator based on the random forest model. Overall, the models all indicate that customer facing technology is extremely important to all customers, regardless of travel type or passenger class.

Q2: Can we predict airline passenger satisfaction based on a limited set of factors?

The explored models achieved an overall good quality at predicting customer satisfaction with very few factors. An accuracy of over 85% can be achieved with as few as 7 factors with the Random Forest model. As such, this would enable the airline to focus on key areas to ensure overall customer satisfaction without having to spend money on multiple factors that will not significantly impact customer satisfaction. Subsetting the Types and Classes of customers could also help go after a particular subset of customers that may either be underserved or more profitable.

In addition, after comparing different models using their AUC on a validation set, logistic regression with forward selection achieves the highest AUC, and hence, we concluded that it was the best model to predict air passenger satisfaction despite using most of the variables.

Q3: Do characteristics which would imply a higher level of service (business class) result in higher levels of satisfaction?

Business class passengers have higher satisfaction levels, particularly in *Checkin.service*, *Seat.comfort* and *Food.and.drink* compared to the Eco plus and Economy class. These findings suggest that offering premium services can positively impact customer satisfaction and create a more enjoyable travel experience for those passengers who opt for higher-class options.

Q4: Are there any noteworthy interactions which would affect the probability of a satisfied customer?

Based on their correlations, there appear to be two primary groupings of factors. The technology grouping includes *Inflight.wifi.service*, *Ease.of.Online.booking*, and *Online.boarding*. Another grouping appears to be focused on service within the aircraft which includes *Cleanliness*, *Food.and.drink*, *Inflight.entertainment*, and *Seat.comfort*. Poor *Inflight.wifi.service* scores had a high indication of customer non-satisfaction which is indicated by both the regression model and the Personal Traveler decision tree models.

5. Conclusion and key takeaways

KEY FINDINGS. The most important features that increase airline passenger satisfaction are *Customer.TypeLoyal*, *inflight.wifi.service*, and *Online.boarding*. People that travel for personal reasons are significantly more likely to be dissatisfied than people who travel for business. This would suggest that investing in good complimentary high-speed wifi would be worthwhile since it would lead to higher customer satisfaction and potentially increase the number of loyal customers, which are also more likely to be satisfied. Further investments could be made to improve the online boarding process for customers so that it is easier and faster, which would also lead to increased satisfaction. While airlines have no control over whether people are traveling for personal or business reasons, money may be better spent trying to attract business travelers instead of people traveling for personal reasons.

CHALLENGES. Contrary to our initial approach, we were not able to merge two other previously identified (Flight Route database and ACSI data set) due to the lack of direct mapping of factors driving customer satisfaction. In addition, handling missing values or “not applicable” values and determining the most appropriate imputation techniques posed challenges during the research process. Finally, the SVM and Random Forest models posed difficulties in interpretation, while selecting the most appropriate model for our research question was challenging given the various techniques' unique strengths and limitations.

LIMITATIONS. We identified several limitations to our research project, related to the limited scope of the data set, potential missing factors, the nature of survey data against other types of data (e.g., written reviews) and the reliability of surveys. Our data set's limited scope, focusing on a US airline and pre-Covid, may not fully capture factors relevant to customer satisfaction in other regions or airlines today. Important factors such as queuing time, lounge comfort, cabin crew quality, and other external factors like airport infrastructure and weather were not included in our data set and could play an important role. More broadly, customer surveys assume known factors, potentially overlooking critical variables that impact passenger satisfaction, while novel approaches identified in our literature review analyzed publicly available written reviews using natural language processing (NLP) to figure out which features airline customers care about. Finally, customer tendency to not take questionnaires seriously and fill them randomly might introduce noise and affect the accuracy of responses.

6. Further research

Having already performed forward selection to identify the most important features in the logistic regression model, further research should now focus on addressing the issue of multicollinearity observed in the logistic regression model (see appendix 6.2), particularly with regards to variables such as *Inflight.wifi.service*, *Ease.of.Online booking* and *Inflight.entertainment* which exhibited higher VIF values. Exploring methods such as feature selection, dimensionality reduction, or transformation techniques could help mitigate multicollinearity issues and enhance the model's interpretability - for example, we could investigate and compare the performance of the logistic regression model with ridge regression or lasso regression.

In addition, analyzing airport data could provide valuable insights into air passenger satisfaction as the airport experience is the first and last touchpoint for passengers and impacts overall perception. Various data sets could be used to identify additional factors impacting customer satisfaction, such as airport passenger surveys, “feedback buttons”, feedback forms or online reviews. Acquiring a comprehensive and precise data set can however present challenges, as well as the process of merging it with our primary data set.

7. Work cited and reference

Lucini, F. R., Tonetto, L. M., Fogliatto, F. S., & Anzanello, M. J. (2020). Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. *Journal of Air Transport Management*, 83, 101760. ISSN 0969-6997.

Pereira, F., Costa, J. M., Ramos, R., & Raimundo, A. (2023). The impact of the COVID-19 pandemic on airlines' passenger satisfaction. *Journal of Air Transport Management*, 112, 102441. ISSN 0969-6997

Ouf, S. (2023). An Optimized Deep Learning Approach for Improving Airline Services. *Computers, Materials & Continua*, 75(1), 1213-1233.

Morgeson, F. V., Hult, G. T. M., Sharma, U., & Fornell, C. (2023). The American Customer Satisfaction Index (ACSI): A sample dataset and description. *Data in Brief*, 48, 109123. ISSN 2352-3409

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1), 12–40.

Reichheld, F., and W. E. Sasser Jr. "Zero Defections: Quality Comes to Services." Harvard Business Review 68, no. 5 (September–October 1990): 105–111.

Airline Passenger Satisfaction data set (Kaggle), available on:
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

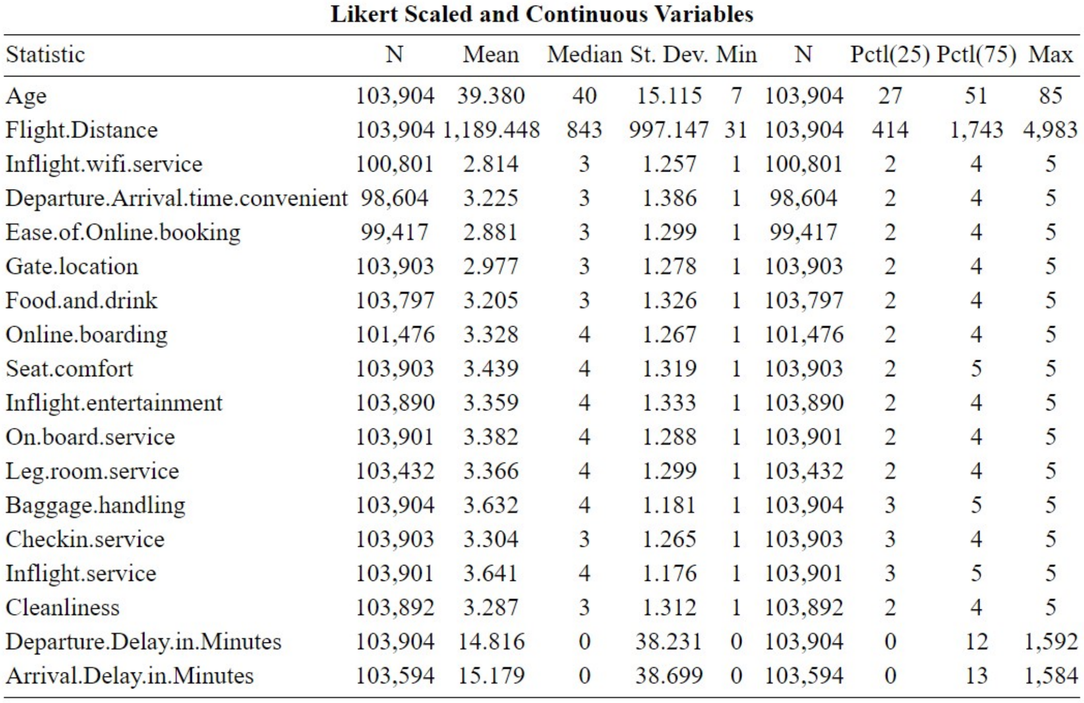
ACSI data set, available on:
https://data.mendeley.com/datasets/64xkbj2ry5/1/files/e4e4b8b0-0d7d-41a9-a2be-c1e586897d7e

Flight Route Database (Kaggle), available on:
https://www.kaggle.com/datasets/open-flights/flight-route-database

British Airways Reviews data set (Kaggle), available on:
https://www.kaggle.com/datasets/lapodini/british-airway-reviews

Hlavac M (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Social Policy Institute, Bratislava, Slovakia. R package version 5.2.3https://CRAN.R-project.org/package=stargazer

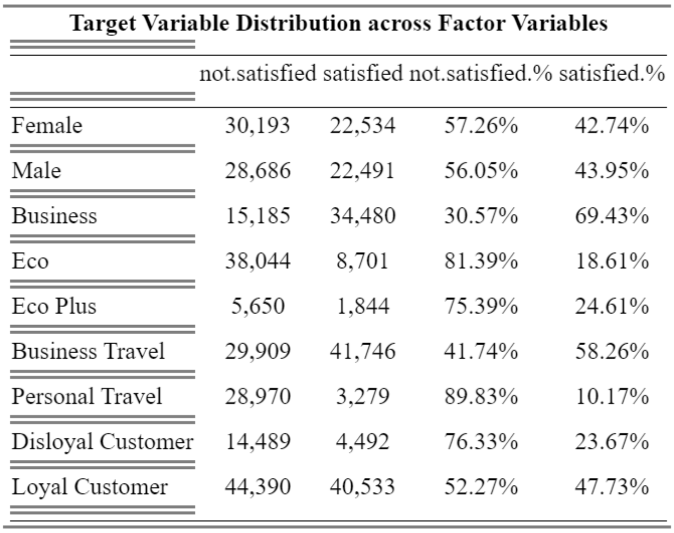
Appendix



Likert Scaled and Continuous Variables

Statistic	N	Mean	Median	St. Dev.	Min	N	Pctl(25)	Pctl(75)	Max
Age	103,904	39.380	40	15.115	7	103,904	27	51	85
Flight.Distance	103,904	1,189.448	843	997.147	31	103,904	414	1,743	4,983
Inflight.wifi.service	100,801	2.814	3	1.257	1	100,801	2	4	5
Departure.Arrival.time.convenient	98,604	3.225	3	1.386	1	98,604	2	4	5
Ease.of.Online.booking	99,417	2.881	3	1.299	1	99,417	2	4	5
Gate.location	103,903	2.977	3	1.278	1	103,903	2	4	5
Food.and.drink	103,797	3.205	3	1.326	1	103,797	2	4	5
Online.boarding	101,476	3.328	4	1.267	1	101,476	2	4	5
Seat.comfort	103,903	3.439	4	1.319	1	103,903	2	5	5
Inflight.entertainment	103,890	3.359	4	1.333	1	103,890	2	4	5
On.board.service	103,901	3.382	4	1.288	1	103,901	2	4	5
Leg.room.service	103,432	3.366	4	1.299	1	103,432	2	4	5
Baggage.handling	103,904	3.632	4	1.181	1	103,904	3	5	5
Checkin.service	103,903	3.304	3	1.265	1	103,903	3	4	5
Inflight.service	103,901	3.641	4	1.176	1	103,901	3	5	5
Cleanliness	103,892	3.287	3	1.312	1	103,892	2	4	5
Departure.Delay.in.Minutes	103,904	14.816	0	38.231	0	103,904	0	12	1,592
Arrival.Delay.in.Minutes	103,594	15.179	0	38.699	0	103,594	0	13	1,584

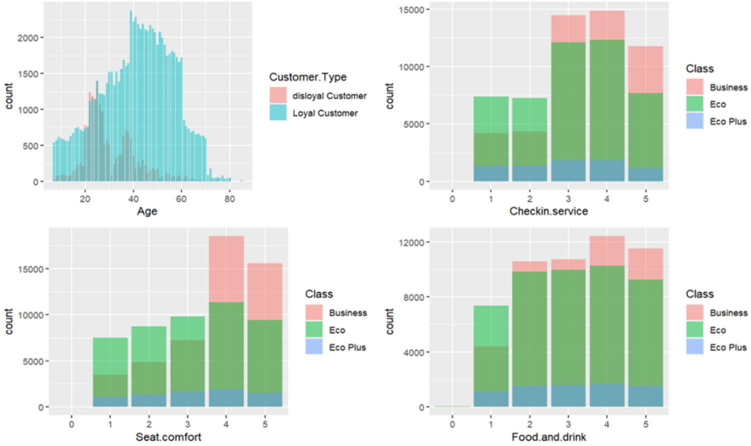
Appendix 1: Predictor Description



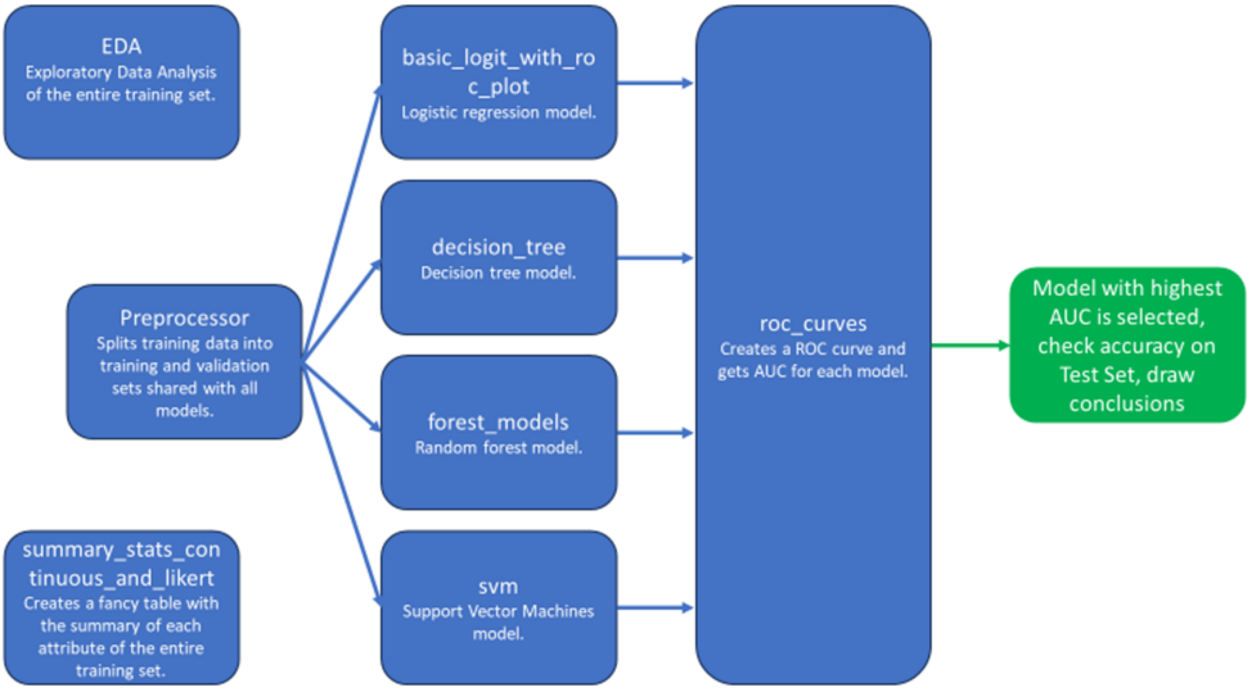
Target Variable Distribution across Factor Variables

	not.satisfied	satisfied	not.satisfied.%	satisfied.%
Female	30,193	22,534	57.26%	42.74%
Male	28,686	22,491	56.05%	43.95%
Business	15,185	34,480	30.57%	69.43%
Eco	38,044	8,701	81.39%	18.61%
Eco Plus	5,650	1,844	75.39%	24.61%
Business Travel	29,909	41,746	41.74%	58.26%
Personal Travel	28,970	3,279	89.83%	10.17%
Disloyal Customer	14,489	4,492	76.33%	23.67%
Loyal Customer	44,390	40,533	52.27%	47.73%

Appendix 2: Analysis of demographic predictors



Appendix 3: Categorical Variables visualizations



EDA
Exploratory Data Analysis of the entire training set.

Preprocessor
Splits training data into training and validation sets shared with all models.

summary_stats_continuous_and_likert
Creates a fancy table with the summary of each attribute of the entire training set.

basic_logit_with_roc_plot
Logistic regression model.

decision_tree
Decision tree model.

forest_models
Random forest model.

svm
Support Vector Machines model.

roc_curves
Creates a ROC curve and gets AUC for each model.

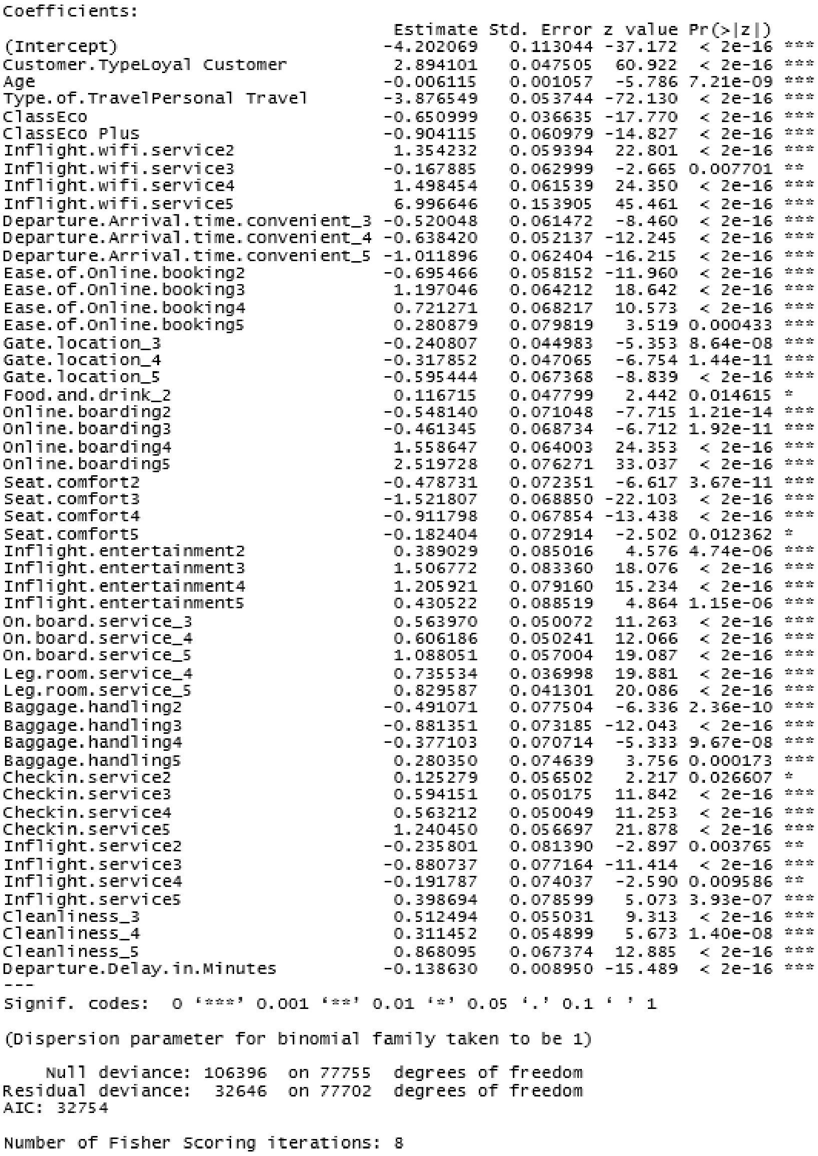
Model with highest AUC is selected, check accuracy on Test Set, draw conclusions

Appendix 4: Methodology and approach followed

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1		NA	NA	103885	142165.13	142167.13
2	+ Online.boarding	-4	43771.50806	103881	98393.62	98403.62
3	+ Type.of.Travel	-1	19766.82322	103880	78626.8	78638.8
4	+ Inflight.wifi.service	-4	10164.54812	103876	68462.25	68482.25
5	+ Customer.Type	-1	6546.868551	103875	61915.38	61937.38
6	+ Inflight.service	-4	7182.501528	103871	54732.88	54762.88
7	+ Leg.room.service	-4	2211.843474	103867	52521.04	52559.04
8	+ Checkin.service	-4	1750.071613	103863	50770.97	50816.97
9	+ Seat.comfort	-4	1186.632167	103859	49584.34	49638.34
10	+ Baggage.handling	-4	980.203787	103855	48604.13	48666.13
11	+ Inflight.entertainment	-4	864.653205	103851	47739.48	47809.48
12	+ Class	-2	874.104126	103849	46865.37	46939.37
13	+ Departure..Arrival.time.convenient	-4	652.479314	103845	46212.9	46294.9

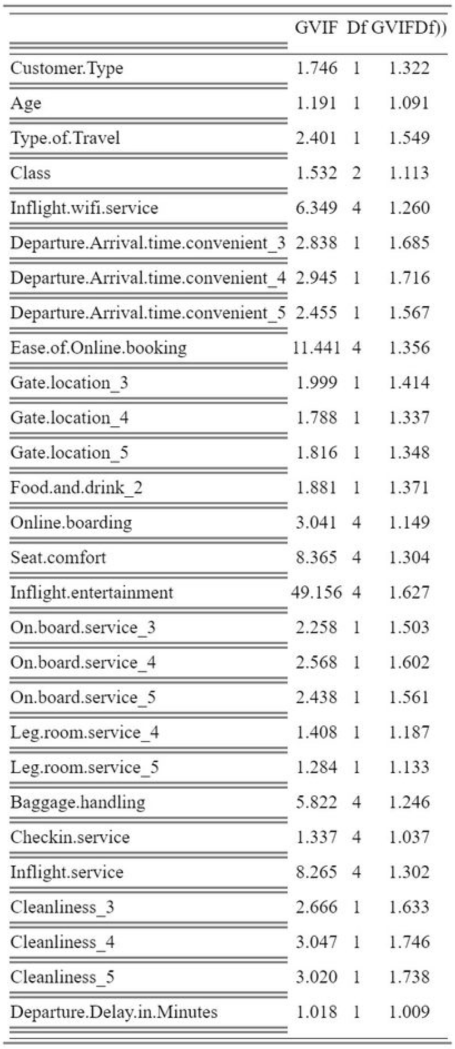
14	+ Ease.of.Online.booking	-4	539.367682	103841	45673.53	45763.53
15	+ On.board.service	-4	409.296861	103837	45264.23	45362.23
16	+ Departure.Delay.in.Minutes	-1	341.223488	103836	44923.01	45023.01
17	+ Cleanliness	-4	219.708532	103832	44703.3	44811.3
18	+ Gate.location	-4	135.231073	103828	44568.07	44684.07
19	+ Age	-1	50.884056	103827	44517.18	44635.18
20	+ Food.and.drink	-4	29.758659	103823	44487.42	44613.42
21	+ Gender	-1	4.628407	103822	44482.8	44610.8

Appendix 5: Forward Selection



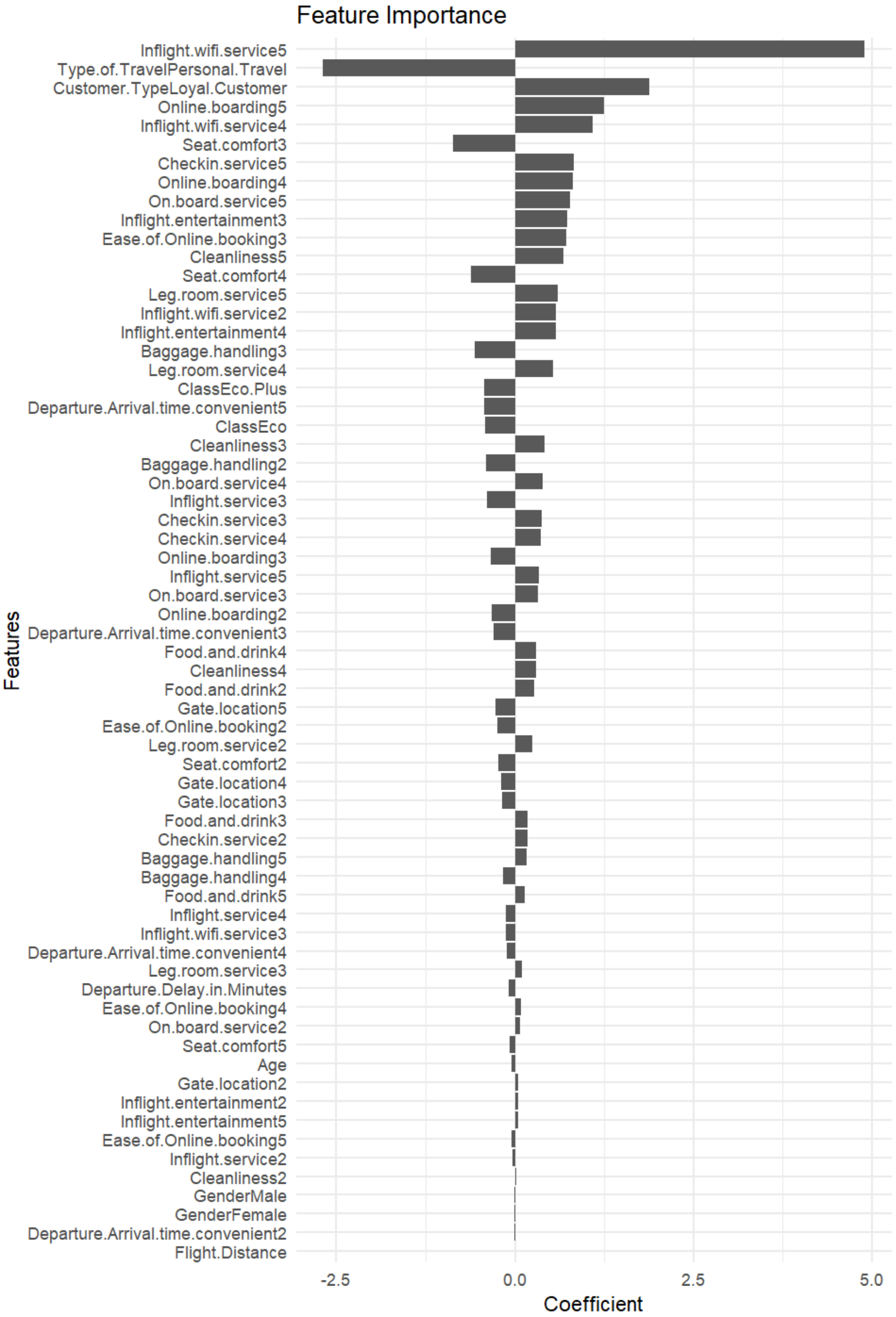
```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -4.202069   0.113044 -37.172  < 2e-16 ***
Customer.TypeLoyal Customer       2.894101   0.047505  60.922  < 2e-16 ***
Age                              -0.006115   0.001057  -5.786 7.21e-09 ***
Type.of.TravelPersonal Travel    -3.876549   0.053744 -72.130  < 2e-16 ***
ClassEco                         -0.650999   0.036635 -17.770  < 2e-16 ***
ClassEco Plus                    -0.904115   0.060979 -14.827  < 2e-16 ***
Inflight.wifi.service2            1.354232   0.059394  22.801  < 2e-16 ***
Inflight.wifi.service3           -0.167885   0.062999  -2.665 0.007701 ** 
Inflight.wifi.service4            1.498454   0.061539  24.350  < 2e-16 ***
Inflight.wifi.service5            6.996646   0.153905  45.461  < 2e-16 ***
Departure.Arrival.time.convenient_3 -0.520048   0.061472  -8.460  < 2e-16 ***
Departure.Arrival.time.convenient_4 -0.638420   0.052137 -12.245  < 2e-16 ***
Departure.Arrival.time.convenient_5 -1.011896   0.062404 -16.215  < 2e-16 ***
Ease.of.Online.booking2          -0.695466   0.058152 -11.960  < 2e-16 ***
Ease.of.Online.booking3           1.197046   0.064212  18.642  < 2e-16 ***
Ease.of.Online.booking4           0.721271   0.068217  10.573  < 2e-16 ***
Ease.of.Online.booking5           0.280879   0.079819   3.519 0.000433 ***
Gate.location_3                  -0.240807   0.044983  -5.353 8.64e-08 ***
Gate.location_4                  -0.317852   0.047065  -6.754 1.44e-11 ***
Gate.location_5                  -0.595444   0.067368  -8.839  < 2e-16 ***
Food.and.drink_2                  0.116715   0.047799   2.442 0.014615 *  
Online.boarding2                 -0.548140   0.071048  -7.715 1.21e-14 ***
Online.boarding3                 -0.461345   0.068734  -6.712 1.92e-11 ***
Online.boarding4                  1.558647   0.064003  24.353  < 2e-16 ***
Online.boarding5                  2.519728   0.076271  33.037  < 2e-16 ***
Seat.comfort2                    -0.478731   0.072351  -6.617 3.67e-11 ***
Seat.comfort3                    -1.521807   0.068850 -22.103  < 2e-16 ***
Seat.comfort4                    -0.911798   0.067854 -13.438  < 2e-16 ***
Seat.comfort5                    -0.182404   0.072914  -2.502 0.012362 *  
Inflight.entertainment2           0.389029   0.085016   4.576 4.74e-06 ***
Inflight.entertainment3           1.506772   0.083360  18.076  < 2e-16 ***
Inflight.entertainment4           1.205921   0.079160  15.234  < 2e-16 ***
Inflight.entertainment5           0.430522   0.088519   4.864 1.15e-06 ***
On.board.service_3                0.563970   0.050072  11.263  < 2e-16 ***
On.board.service_4                0.606186   0.050241  12.066  < 2e-16 ***
On.board.service_5                1.088051   0.057004  19.087  < 2e-16 ***
Leg.room.service_4                0.735534   0.036998  19.881  < 2e-16 ***
Leg.room.service_5                0.829587   0.041301  20.086  < 2e-16 ***
Baggage.handling2                -0.491071   0.077504  -6.336 2.36e-10 ***
Baggage.handling3                -0.881351   0.073185 -12.043  < 2e-16 ***
Baggage.handling4                -0.377103   0.070714  -5.333 9.67e-08 ***
Baggage.handling5                 0.280350   0.074639   3.756 0.000173 ***
Checkin.service2                  0.125279   0.056502   2.217 0.026607 *  
Checkin.service3                  0.594151   0.050175  11.842  < 2e-16 ***
Checkin.service4                  0.563212   0.050049  11.253  < 2e-16 ***
Checkin.service5                  1.240450   0.056697  21.878  < 2e-16 ***
Inflight.service2                -0.235801   0.081390  -2.897 0.003765 ** 
Inflight.service3                -0.880737   0.077164 -11.414  < 2e-16 ***
Inflight.service4                -0.191787   0.074037  -2.590 0.009586 ** 
Inflight.service5                 0.398694   0.078599   5.073 3.93e-07 ***
Cleanliness_3                     0.512494   0.055031   9.313  < 2e-16 ***
Cleanliness_4                     0.311452   0.054899   5.673 1.40e-08 ***
Cleanliness_5                     0.868095   0.067374  12.885  < 2e-16 ***
Departure.Delay.in.Minutes       -0.138630   0.008950 -15.489  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 106396  on 77755  degrees of freedom
Residual deviance:  32646  on 77702  degrees of freedom
AIC: 32754

Number of Fisher Scoring iterations: 8
```

Appendix 6.1: Summary of Logistic Regression



	GVIF	Df	GVIFDf))
Customer.Type	1.746	1	1.322
Age	1.191	1	1.091
Type.of.Travel	2.401	1	1.549
Class	1.532	2	1.113
Inflight.wifi.service	6.349	4	1.260
Departure.Arrival.time.convenient_3	2.838	1	1.685
Departure.Arrival.time.convenient_4	2.945	1	1.716
Departure.Arrival.time.convenient_5	2.455	1	1.567
Ease.of.Online.booking	11.441	4	1.356
Gate.location_3	1.999	1	1.414
Gate.location_4	1.788	1	1.337
Gate.location_5	1.816	1	1.348
Food.and.drink_2	1.881	1	1.371
Online.boarding	3.041	4	1.149
Seat.comfort	8.365	4	1.304
Inflight.entertainment	49.156	4	1.627
On.board.service_3	2.258	1	1.503
On.board.service_4	2.568	1	1.602
On.board.service_5	2.438	1	1.561
Leg.room.service_4	1.408	1	1.187
Leg.room.service_5	1.284	1	1.133
Baggage.handling	5.822	4	1.246
Checkin.service	1.337	4	1.037
Inflight.service	8.265	4	1.302
Cleanliness_3	2.666	1	1.633
Cleanliness_4	3.047	1	1.746
Cleanliness_5	3.020	1	1.738
Departure.Delay.in.Minutes	1.018	1	1.009

Appendix 6.2: VIF values of Logistic Regression



Appendix 7: Feature importance of SVM Classifier