LUISS

Department of Impresa and Management
Master's degree in Marketing Analytics & Metrics
Chair of Customer Intelligence & Big Data

AI & Machine Learning for Marketing: A Data-Driven Approach to Predict Customer Satisfaction in Air Travel Industry

Prof.
Giuseppe Francesco Italiano
SUPERVISOR

Prof.
Marco Querini
CO-SUPERVISOR

Stefano Faiola
735951
CANDIDATE

Academic Year 2021/2022

INDEX

ABSTRACT 4

INTRODUCTION 4

CUSTOMER SATISFACTION IN GENERAL 6

Airlines Customer Satisfaction 7

DECISION WEIGHTS AND MACHINE LEARNING 8

U.S. AIRLINE PASSENGER SATISFACTION DATASET 10

Customer Ratings 11

DATA VISUALISATION 13

Kernel Density Plot 13

Scatter Plot 16

Bar Chart 18

Correlation 21

DATA PRE-PROCESSING 24

Duplicate Values 24

Missing Values 24

Outliers 25

Labels encoding 26

MACHINE LEARNING MODEL IMPLEMENTATION 27

Logistic Regression 28

Tree-Based Models 30

Classification Tree 30

Random Forest 33

Gradient Boosting 36

Kernel Support Vector Machine 38

Feature Permutation Importance 39

Shapely Additive Explanation (SHAP) 41

RESULTS 43

Academic Implications 44

Managerial Implication 45

APPENDIX 47

Appendix A. Kernel Density Plot 47

Appendix B. Scatter Plot 49

Appendix D. Correlation Heatmap 53

Appendix E. Feature Permutation Importance 55

Appendix F. Shapely Additive Explanation Plot (SHAP) 57

References 59

ABSTRACT

Airlines faced a severe downturn due to COVID-19 air travel disruption, the business is currently recovering but with radical changes. A low price has always been the dominant factor in the air travel market, making it more difficult to reach satisfied consumers. Starting from a dataset of about 130,000 responses; with a marketing perspective, the research uses several Machine Learning models to identify which variables have the greatest influence on consumer satisfaction during their relationship with airlines, from booking to using the service; in this way it is possible to allocate budget to specific areas while sticking to a low-cost strategy without discriminating against consumer satisfaction. The mix of machine learning models has allowed the identification of the best among them, as well as the intersection of the different results, thus obtaining greater interpretability and robustness of the results. the results say that there is a clear gap in satisfaction between the various types of consumers where those who travel for business reasons tend to be more satisfied than those who travel for personal reasons; also, the classes despite the difference in price have a strong influence on satisfaction where customers in business class are more satisfied than those who travel in eco. Finally, digitalisation drives consumer satisfaction; airlines that can simplify and make online boarding efficient and provide good in-flight Wi-Fi would have a powerful positive influence on customer satisfaction.

INTRODUCTION

The air travel business faced a violent shutdown due to COVID-19; the worldwide revenue from passenger air traffic in 2019 was $ 581 billion, dropping to $ 189 billion in 2020. In 2020, the post-COVID loss of airline companies was $ -35 billion in North America and $ -34.5 in Europe. In 2021, the air travel business started to recover slowly, where North American airline losses equalled $ -5.5 billion, and in Europe, airlines losses equalled $ -20.9 billion. Now, the worldwide revenue forecast is $ 378 billion - still not at the levels of 2019, but it is slowly increasing (IATA, 2021).

Certainly, the air travel market is in the eye of the storm. Customer satisfaction could not be viewed as the first major problem, however after every crisis, when the market starts to grow again, and passengers have started to regain confidence in taking airplanes, there is an ample opportunity to re-establish a good relationship with consumers starting with a high level of customer satisfaction. Many studies highlight the importance of customer satisfaction to build a long last relationship and loyalty between companies and customers, which are then translated into a higher

return on investment and profitability (Anderson, Fornell, & Lehmann, 1994; Anderson, Fornell, & Rust, 1997; Fornell, A National Customer Satisfaction Barometer: The Swedish Experience, 1992; Fornell, Johnson, Anderson, Cha, & Everitt Bryant, 1996).

Although there is much research on how to increase customer satisfaction, there is a lack of research on how to increase customer satisfaction in a specific business environment where its drivers are not specifically communicated; indeed, it is vital to offer high service quality, but it is not possible to maximise the quality of every product or service within the company without productivity being undermined.

The air travel industry is characterised by a lower level of consumer satisfaction than other businesses because, in addition to falling under the sale of services and not products, the level of standardisation is very high. In order to achieve a higher level of satisfaction, companies tend to customise rather than standardise, this is not possible and if possible only within certain limits in air travel. Due to the high level of standardisation, individuating and leveraging the most important drivers of customer satisfaction is even more fundamental in the air travel business to increase the number of satisfied customers.

This research aims to provide the most critical drivers of customer satisfaction in the air travel business by using machine learning models in such a way as to offer a machine-side perspective to help professionals gain insights and increase the companies' know-how. Understanding that machine learning is not a substitute for marketers' findings and insights is fundamental. However, machine learning is an additional tool that must be used with the most common marketing research strategies like surveys and in-depth interviews. The most important variable of this research has to be summed up with professional marketing knowledge and used as a complementary tool rather than a substitute.

The research does not offer a complete mathematics and statistics perspective; instead, all findings are interpreted from a marketing perspective, reducing the complexity of the machine learning model and offering a more comprehensive business view related to all the findings. Very often, research like this is done by computer science students who are not close to marketing and business strategy, providing valuable findings with a lack of interpretation. The main answer when tackling machine learning business problems like this is: what is the best model for predicting customer satisfaction that enable me to understand the decision-making process? Secondary, once I know how model decisions are made, are those findings aligned with the company's business strategy?

CUSTOMER SATISFACTION IN GENERAL

To better understand why machine learning can be helpful for the purpose of customer satisfaction prediction and identification of its most essential variables, it is crucial to have a general definition of Customer Satisfaction and how it is influenced in general. The most cited and accepted definition from the extant literature of satisfaction is the following: *"satisfaction is a person's feeling of pleasure or disappointment that results from comparing a product's perceived performance to expectations"* (Kotler & Keller, 2012).

Perceived quality, perceived value and expectations are key drivers of customer satisfaction (Fornell, Johnson, Anderson, Cha, & Everitt Bryant, 1996), highlighting how this KPI can be different among customers. Indeed, the same level of satisfaction can have different influences and be rated with different magnitudes. Although prior research shows the primary drivers of customer satisfaction, a more specific approach is needed. Knowing only that the product having a perceived quality influences customer satisfaction does not explain which attributes in a specific industry are the most important to improve satisfaction. Customers continuously compare products and services and, based on their cumulative past experiences, are able to have intrinsic expectations and evaluations of a product or service (Fornell, 1992); for this purpose, companies' investment should focus on customer satisfaction because there is a relationship creation which drives economic return. It is not the mere single event of being satisfied with a product but the loyalty creation that leads to the relevance of Customer Satisfaction. The more satisfied a customer is, the higher the probability they will become a loyal one. The Statista 2021 report about Airline passenger experiences shows that the companies with the highest rating on the American Customer Satisfaction Index (Southwest Airlines ACSI: 79 and Delta Air Lines ACSI: 79) are the companies with the largest number of loyal customers. In fact, those customers account for almost 50% of the total market (Southwest Airlines, 26% and Delta Air Lines, 22% of the total market) (ACSI, 2021; ValuePenguin, 2021).

Loyalty is one of the most important benefits a firm can have, and leverage it as an asset for profitability due to recurrent purchase, price elasticity reduction, failure cost reduction, higher retention rate and positive word of mouth (Anderson, Fornell, & Lehmann, 1994; Anderson, Fornell, & Rust, 1997; Fornell, 1992).

Today's satisfied or dissatisfied customers will be influenced by future decisions. For this reason, satisfaction, like loyalty, has to be viewed from a long-term perspective (Anderson, Fornell, & Lehmann, 1994). The long-term perspective of satisfaction and loyalty has to be viewed the other

way around when a customer is dissatisfied. It is very difficult to turn negative opinions into positive ones when the company's reputation is damaged. Like loyalty is reached after an extended period, conversion of a dissatisfied customer can never occur or may require greater effort than converting a neutral customer; it is fundamental to understand what customers want and like about a specific product or service. After a long period of work in the same business, managers tend to know the company's customer base, and the manager's assumption is very often correct. However, whether qualitative or quantitative, marketing tools cannot be neglected to have the up-to-date opinion of consumers and increase the company's decision quality. A quantitative tool like machine learning cannot be viewed as a manager's decision substitute; indeed, it is a valuable complementary tool for a manager's decisions.

Airlines Customer Satisfaction

As stated in the American Customer Satisfaction Index (ACSI) from the research of Fornell, Johnson, Anderson, Cha, & Everitt Bryant (1996), satisfaction tends to be lower in the service industry compared with goods. For this reason, it is important to have a different perspective of customer satisfaction in the case of the air travel business rather than making assumptions based on general metrics. Consumers are exposed longer to multiple stimuli in the service industry and even more in the Air Travel industry. In fact, it is sufficient to think about the whole process of taking an airplane from the airport of departure to the airport of arrival. There are many different steps in which the airline is involved. As if that were not enough, with digitalisation and the increase in online ticketing and check-in, the contact between company and consumer takes place even before the day of departure.

Due to the high number of stimuli, inconveniences can occur at every level and time, leading to a high probability of dissatisfied customers or a low Customer Satisfaction Score translated into negative word of mouth and low customer loyalty. Since the relationship between customers and airlines starts even before the day of departure and since the majority of people use the online check-in service, for this reason, customers who buy an air travel service usually follow an online and offline path. There is evidence that depending on the channel, there can be different influences on customer satisfaction; satisfied customers are more likely to become loyal when purchasing from an online channel, and expectations of the product/service quality are higher in the offline channel (Hult, Sharma, Morgeson III, & Zhang, 2018). However, while research on retailing confirms the differences between online and offline customer satisfaction, focused research should be done on the air travel business. Indeed, the amount of complexity in the air travel service industry

is very high, and customer relationships are difficult to manage. As if that were not enough, the price war makes it even more challenging to fulfil this relationship.

Companies seek to improve service or product qualities to leverage the positive relationship between customer satisfaction and quality of service. However, focusing on maximising quality does not necessarily lead to an increase in customer satisfaction. It is not possible to maximise all existing dimensions, and managers have to decide where to allocate resources; this concept is even more important in the case of airlines since today's competitive market is based on a low pricing strategy which drastically reduces the service quality and harms customer satisfaction (Boetsch, Bieger, & Wittmer, 2011). For this reason, the error rate is minimised, which also leads airlines to make low-risk decisions focusing more on productivity (overbooking strategy) rather than being customer-centric. Of course, this is not the case for all airline companies.

In general, customisable products or services lead to higher customer satisfaction and a lower level of productivity (Anderson, Fornell, & Rust, 1997). This is not the case in the air travel business, where the service tends to be standardised to maintain low cost and offer a competitive price. This practice leads to a very similar service offered among airlines, and the number of dimensions in which it is possible to be different is reduced to a minimum level. For this reason, it is fundamental for airlines to understand these dimensions and where to allocate resources to avoid a waste of money and increase the return on investment.

DECISION WEIGHTS AND MACHINE LEARNING

The process of searching for the most influential variable to which lead to a decision, in this case, whether a customer is satisfied or not, is almost the same process that human beings go through for any decision or judgement. A human choice process can be represented as the weighted additive decision rule. Mathematically speaking, humans take some weighted attributes in order to obtain a final score of a specific option:

$$\text{Score (A)} = w_1 * A_1 + \text{w}_2 * \text{A}_2 + w_n * A_n$$

Weights and attributes vary among people because they are based on past experiences. Humans constantly use a large amount of information but do not pick up a pen and write the equation on paper. They just evaluate attributes related to a specific option using past experiences (Soman, 2015). The key concept is to find those weights of a specific decision to understand which attributes are facilitators of the final choice. This concept is exactly what also happens when marketers make decisions and is not only related to consumers' decisions. Professionals weigh all attributes

available unconsciously in order to make a business decision. The higher the seniority, the higher the number of attributes and accuracy of weights. For this reason, high experience may lead to a bad decision because judgment may not be aligned with a fast-changing environment. This is also why companies always say that newly hired people should bring fresh ideas.

Instead of waiting an entire life to learn attributes and weights, processing power can be used to train an algorithm to make a specific decision, in this case, whether a customer is satisfied or not. Learning relationships between attributes and the output variable allows the statistical model to adjust weights to maximise the accuracy of the prediction or reduce the error. In this specific case, there is no interest in predicting customer satisfaction because to do so, one would need some scoring from customers of all the available attributes, which would mean that one could directly ask whether customers are satisfied or not; for this reason, the objective of this research is to pick some classification models and inspect them to understand why and how the decision of the model was made. Practically speaking, the main scope is to look for the weights related to the dataset's attributes to understand which is the most important. It would be like asking a doctor why a specific diagnosis was made. This process answers the question: how much can I rely on a specific variable to understand if customers are satisfied?

There are many different models to choose from. In general, the final scope of machine learning is to find the best one to predict the output variable; however, there are trade-offs in deciding which is the best model to use, and the final decision should always be aligned with the problem that has to be solved. This research aims to have a model to interpret and predict the output variable with an acceptable error. If not, the model would be useless because if it cannot predict the output variable, the interpretability of the model would not be meaningful since understanding how a bad decision is made does not help see which variable is the most important. There are models interpretable by design like Logistic Regression and Classification Tree, less but still interpretable models like Gradient Boosting and Random Forest, and non-interpretable models like Kernel – SVM due to the Radial Basis Function used. This research uses all those models to compare them across accuracy and variable importance. Two interpretability methods like Shapely Additive explanation (SHAP) (Lundberg & Suu-In, 2017) and Feature Permutation Importance (Pedregosa, et al., 2011) are also used for the purpose of local and global interpretability, respectively.

U.S. AIRLINE PASSENGER SATISFACTION DATASET

The U.S. Airline Passenger Satisfaction Dataset (The Dataset) describes customer-related satisfactory experience levels for all U.S. airlines in 2015. The Dataset is publicly available on Kaggle through the following link: https://www.kaggle.com/datasets/johnddddd/customer-satisfaction

The Dataset is composed of 129,880 observations and 24 variables*: id:* Unique Code; *Gender*: sex of the customer (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, Business Travel); *Class*: Travel class in the plane of the passengers (Business, Eco, Eco Plus); *Flight distance*: The flight distance of this journey; *Inflight wifi service*: Satisfaction level of the inflight WiFi service (0:Not Applicable;1-5); *Departure/Arrival time convenient:* Satisfaction level of departure/Arrival time convenience; *Ease of Online booking*: Satisfaction level of online booking; *Gate location:* Satisfaction level of gate location; *Food and drink:* Satisfaction level of food and drink; *Online boarding:* Satisfaction level of online boarding; *Seat comfort*: Satisfaction level of seat comfort; *Inflight entertainment:* Satisfaction level of inflight entertainment; *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 for departure; Arrival Delay in Minutes: Minutes delayed of arrival. For each variable, descriptive statistics and type are reported in Table 1.

Each customer rated a specific service and lastly answered whether, in general, they were satisfied or not. The rating system used in the questionnaire to evaluate a specific variable during the survey ranged from 1 to 5, where values equal to 0 mean that the rating was not applicable. The reason why a rating cannot be applied to a specific flight about a specific variable (e.g., WiFi service) means that the service was not present on that flight; observations with ratings equal to 0 will be retained because machine learning is informative in knowing about the presence of a service. It is possible to consider a rating of 0 as the lowest rating because the service is missing. However, the impact of ratings equal to 0 is minimised due to the reduced presence of the individuals who provided this rating because most airline flights in the dataset have all the services rated by consumers.

Variable	Mean	Median	Min	Max	Range	Type
Gender						Binary
Customer_Type						Binary
Age	39.43	40	7	85	78	Continuous
Type_Travel						Binary
Class						Ordinal
Flight_Distance	1190.32	844	31	4983	4952	Continuous
Inflight_Wifi_Service	2.73	3	0	5	1 to 5	Interval
Departure_Arrival_Time_Convenient	3.06	3	0	5	1 to 5	Interval
Ease_Online_Booking	2.76	3	0	5	1 to 5	Interval
Gate_Location	2.98	3	0	5	1 to 5	Interval
Food_Drink	3.20	3	0	5	1 to 5	Interval
Online_Boarding	3.25	3	0	5	1 to 5	Interval
Seat_Comfort	3.44	4	0	5	1 to 5	Interval
Inflight_Entertainment	3.36	4	0	5	1 to 5	Interval
Onboard_Service	3.38	4	0	5	1 to 5	Interval
Leg_Room_Service	3.35	4	0	5	1 to 5	Interval
Baggage_Handling	3.63	4	1	5	1 to 5	Interval
Checkin_Service	3.31	3	0	5	1 to 5	Interval
Inflight_Service	3.64	4	0	5	1 to 5	Interval
Cleanliness	3.29	3	0	5	1 to 5	Interval
Departure_Delay_Minutes	14.71	0	0	1592	1592	Continuous
Arrival_Delay_Minutes	15.16	0	0	1584	1584	Continuous
Satisfaction						Binary

Table 1 U.S. Airline Passenger Satisfaction (2015)

Customer Ratings

Before starting to interpret the data, it is mandatory to understand the type of variables dealt with, how information was collected, and what can influence each variable of interest. Customers ratings in statistical terms are interval variables where the magnitudes are always equal from score to score; however, the most important digression about ratings is that they are based on customers' perceptions which are not defined by a specific rule - the numerical magnitude is the same, but the real magnitude cannot be defined. Perception is different among customers; this brings a higher

degree of uncertainty to the analysis because the method for measuring the dependent variable and the scores is not standardised.

From a marketing perspective, ratings are very complex variables because they are considered perceptive instead of objective. For a different customer, we can have different perceptions with the same rating or the same perception with different ratings. This difference is driven by the fact that each person has a different background and experiences, influencing the ratings assigned for each variable. For example, a person that used to travel very frequently could be considered an expert, and can compare a higher amount of information related to air travel. Ratings of expert consumers may be more accurate and specific and usually lower compared with ratings assigned by people who do not travel by air very frequently. Moreover, people differ in how they evaluate things; many can be restrictive and others more expansive. Some people do not provide the maximum score very frequently because, for them, it is considered a symbol of perfection.

However, degrees of uncertainty, like in every statistical study, can be reduced by increasing the sample size; the larger the number of observations, the more probable the study would be significant. Moreover, extant literature confirms that objective and subjective measures, even though they differ in type, are strongly correlated (John, 1999). Indeed, it is not possible to have an objective evaluation of customer satisfaction and relying on subjective evaluation is mandatory. From a pure marketing perspective, performances related to consumer behaviour can only be perceptive. For this reason, instead of being called a variable, customer satisfaction should be called a construct because it is conceptualised at the theoretical (abstract) plane (Bhattacherjee, 2012).

The aims of this marketing research are mainly to listen to customers' perceptions; having a subjective evaluation of more than 120,000 customers instead of being viewed as a penalty should be viewed as a strong value creator for marketing research and the company’s manager, looking for patterns in subjective data means to analyse a partial but important part of customer behaviour that should be followed to satisfy their utilitarian and hedonic needs. As marketers, we constantly deal with the construct (to cite some) like perceived risk of a purchase, fear of missing out, loyalty and customer satisfaction; machine learning can be used as an additional tool to better understand those constructs within each business environment.

DATA VISUALISATION

Data visualisation is the practice of translating information into a visual context to make it easier to capture the value in the dataset and build valuable insights. Data visualisation can be used as a pre-analysis tool to understand the dataset distribution and find the variables most correlated with each other and with customer satisfaction. It can also be used as a post-model analysis tool. It is possible to retrieve built graphs to understand the dataset and better interpret the most important variable. For example, knowing that class is an important variable is not very helpful. We need to know what is the level of influence and when a customer is satisfied or not. Moreover, using visualisation, it is also possible to interpret the business of airlines in the dataset, which is strictly related to customer behaviour and the output in satisfaction. Continuous variables distribution will be visualised using the kernel density plot, categorical or nominal variables will be visualised using a bar chart, and a scatter plot will be used to see the position of each individual in a two-dimensional plane. All the visualisation will be provided for satisfied and dissatisfied customers; lastly, the correlation graph will represent the correlation matrix where the highest values are represented in red and the lowest values in blue, where blank cells represent no presence of correlation between two variables. All the data visualisation is provided using the library Ggplot2 in the R environment (R Core Team, 2021; Wickham, 2016).

Kernel Density Plot

Instead of simply plotting a histogram, where it is mandatory to find optimal bins of the dataset of a specific variable, it is possible to plot a kernel density estimation that creates a certain kernel estimate of the data instead of creating bins of the data, in practical terms, many different bins are computed. All of them are summed up in order to have an estimated distribution of the data. It is possible to increase or decrease the smoothness of the kernel density estimation by increasing the value of the bandwidth adjustment for the kernel density plot provided in *figure 1*. A medium value of 4 of the bandwidth adjustment was used to have a less detailed curve but capture changes in data distribution.

Variable age is normally distributed; the mean value is 39.32. When plotting the variable grouping by customer satisfaction, it is possible to see that older people's density is higher in case of satisfaction and also that older people are more likely to be satisfied with regard to younger people. In fact, the mean value of age for neutral or dissatisfied is 37.65, while for satisfied is 41.74. Even though there is a difference, this does not mean it is significant. By simply looking at the graph, it is not possible to statistically state if the difference is significant. However, the mean

values of 37.65 and 41.74 are not very different. A possible assumption related to the variable age could not be related to the fact that the age of customers influences the level of satisfaction, but there is a possibility that the type of flight, destination and services are more likely to fit with older people; however, that information is not available, and it is not possible to further investigate about a possible influence of the age and make specific inferences.

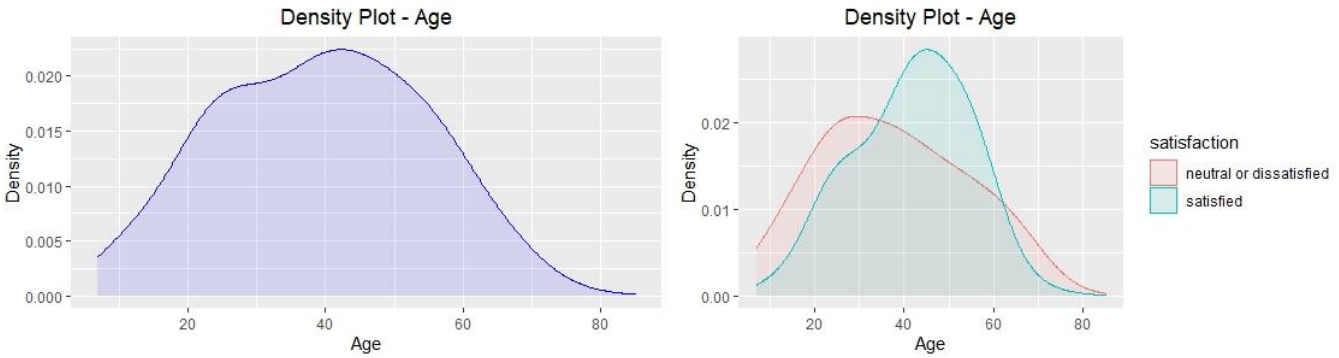


Figure 1 U.S. Airline Passenger Satisfaction (2015)

Flight Distance is a very particular variable, and findings related to it in *figure 2* are fascinating; the observations have a high frequency when a flight is of a short distance. At 1,500 miles, we can individuate an elbow point; after it, the higher the distance, the lower the number of observations. This result is in line with the logical and normal behaviour of trips, where we can state that, in general, the number of short-distance flights is higher because people tend to move more frequently to the nearest point. From a 2017 Statista survey, 44% of respondents said that they have never travelled on a long-haul flight over the past two years, compared with a 19% of respondents that have never travelled on a short-haul flight over the past two years (Statista, 2017).

Computing the density plot distinguishing by customer satisfaction, one can see that flights of high distance are characterised by a higher number of satisfied customers. The density decrease differs significantly from the density plot of neutral or dissatisfied customers. In the density plot of satisfied customers, the elbow point present is missing in the non-grouped by satisfaction graph; this is different in the case of the neutral or dissatisfied customer, where the elbow point is even more visible than the non-grouped by satisfaction density plot.

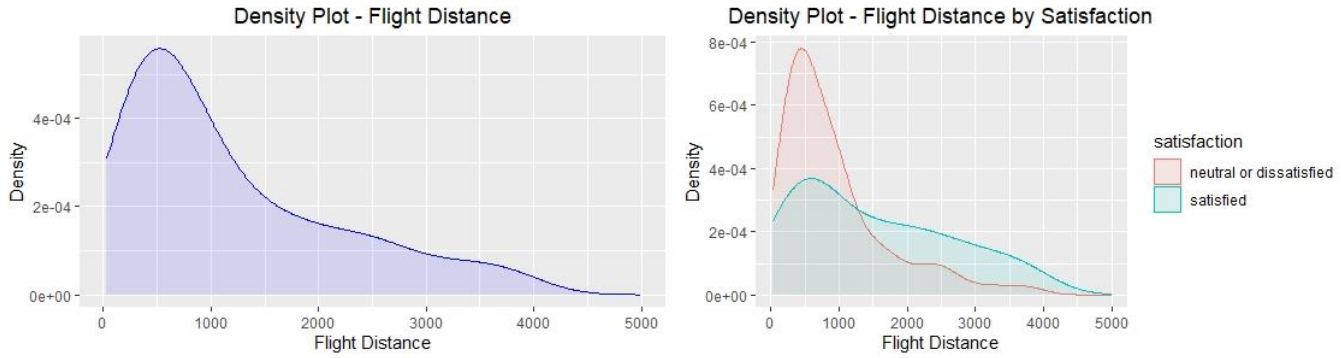


Figure 2 U.S. Airline Passenger Satisfaction (2015)

Arrival and Departure Delay are almost the same variables. Those variables contain a very high number of observations where there is no delay; in order to compute the density estimation and make it visualisable, values equal to 0 minutes of delay are changed to 0.1 to compute the $\log_1$ of the two variables and reduce the magnitude to make the visualisation possible. By reducing the magnitude, it is possible to plot the kernel density estimate for the delays at departure and arrival, shown in *Figures 3* and *4* respectively; the greatest number of flights have a delay of 0 (-2.5 in the graph). The number of flights decrease when a minimal delay is reached (for example, from 1 to 3 minutes). The number of trips starts to increase again when there is a more consistent delay in minutes which ranges from approximately 10 minutes to 140. This behaviour could be justified by the fact that when there is a delay, it is for a real reason which can often be the same. For example, if we had data about the reason for the delay, we would be able to find that those specific events are more frequent than others and require the same number of minutes. The current situation of departure delays is one of the worst in history due to COVID-19 and the pilot shortage, so most of the flights are delayed for the same reasons (Kelleher, 2022). Distinguishing the density plot by customer satisfaction, it is possible to see that arrival delay is a little bit more influential than departure delay; in both cases, when the delay is of many minutes, customers are more likely to be dissatisfied. This assumption is represented by the distribution and it is also logical and accepted by customers. However, a higher number of dissatisfied customers is expected in case of a delay, which is not the case reported in the density plots in *figure 3* and *figure 4*. The density plot is almost the same when distinguished by customer satisfaction; a possible explanation is that very frequently, a delay is characterised by causes of force majeure; consumers recognise this (e.g., bad weather or delay imposed by the airport) and do not blame the airline for the delay, so satisfaction ratings are not likely to be affected. Kernel density plot are also presented in *appendix A* for a more complete and easier data visualization.

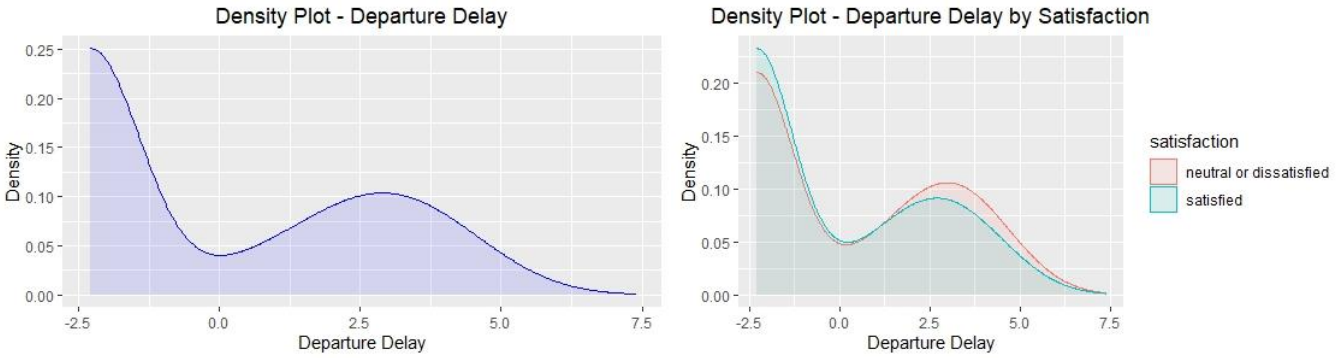


Figure 3 U.S. Airline Passenger Satisfaction (2015)

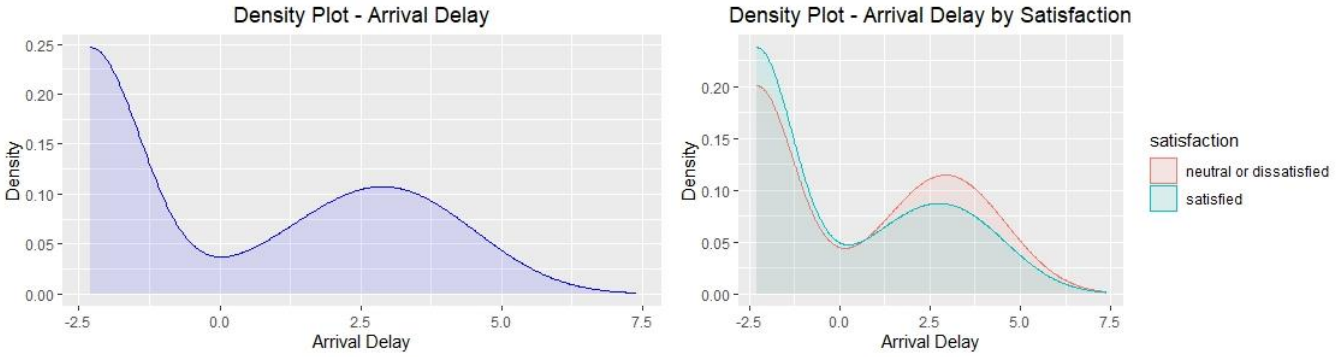


Figure 4 U.S. Airline Passenger Satisfaction (2015)

Scatter Plot

Through scatter plots, it is possible to have a more specific view of each customer, distinguishing between a satisfied and non-satisfied one. Even if age is not correlated with satisfaction (comments on correlation are in the next section), a very interesting pattern in *Figure 5* is obtained by plotting flight distance and age using the scatter method. Satisfaction by age is well distributed, however flights of long distances are composed of a higher number of satisfied customers. In the scatter plot, it is possible to individuate a cluster of individuals in the middle age range from 20 to 60 when the flight distance is longer than 1,500 miles. The rule that correlation does not imply causality, in this case, is mandatory. It is possible to assume that longer flight distances are better equipped in terms of service and the quality could be higher, but from a marketing perspective, the distance of flight is just the objective that the travel service is trying to reach, moving people from one side to another, accomplishing a utilitarian need. What makes the difference in terms of satisfaction is how the service is performed (also because airlines must reach the objective), satisfying the hedonic needs of customers. It is possible to think about utilitarian needs as the goal that should be reached rationally while hedonic needs are the objective that considers the emotions of customers (Solomon, 2018).

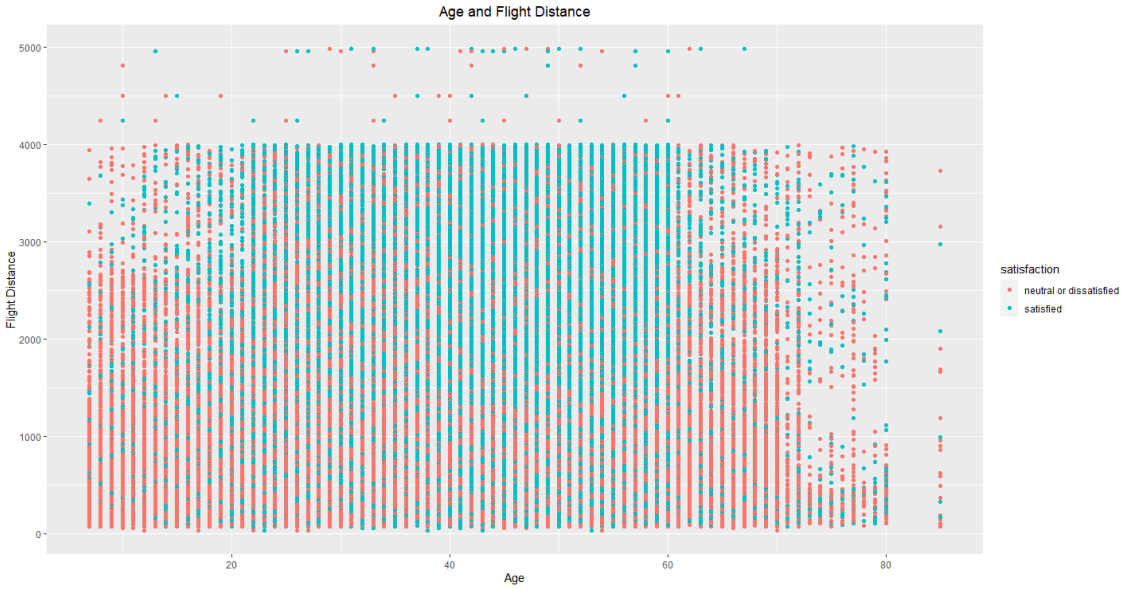


Figure 5 U.S. Airline Passenger Satisfaction (2015)

Another visualisation using the scatter method implies the use of the variable "flight distance" and as the second variable "departure delay in minutes", as seen in *figure 5*. The longer the distance, the more satisfied customers seem to be; what is interesting in *figure 6* is referred to as the second dimension. It would be expected that the higher the delay in minutes, the higher the number of dissatisfied customers. This assumption is unconfirmed; there are still satisfied customers with a delay higher than 1,000 minutes. As explained before, delays could not be associated with the airline's fault, but it is still interesting that customers with such delays are still satisfied. Looking for a pattern like this is a key activity in order to have meaningful assumptions before machine learning is used. It is possible that the same pattern analysed can be used from the model or that the model can find different patterns.

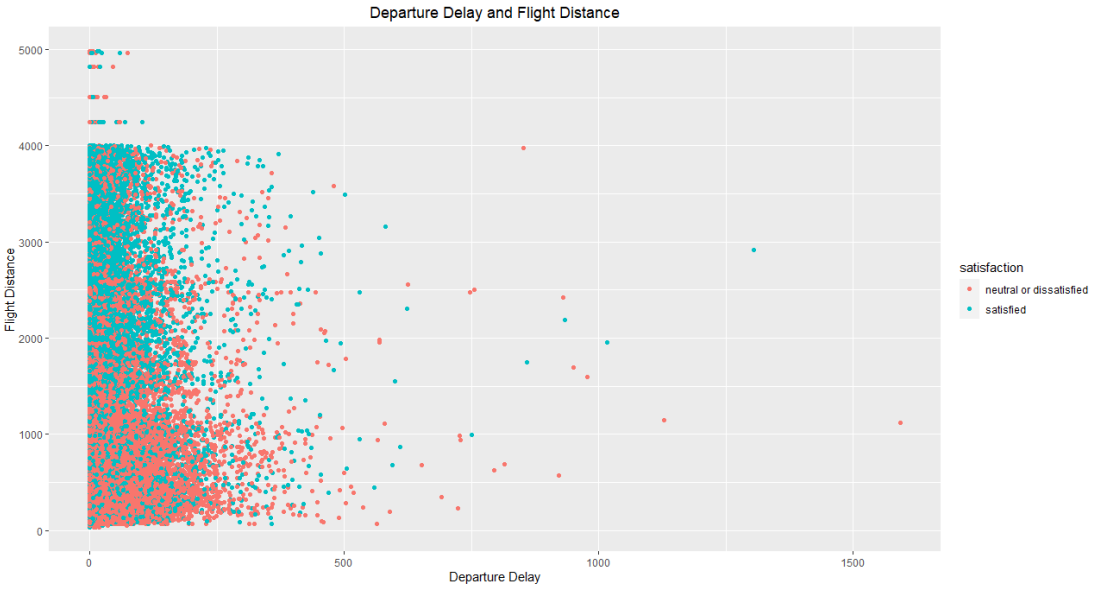


Figure 6 U.S. Airline Passenger Satisfaction (2015)

Bar Chart

Bar charts are one of the simplest graphs but still very effective for this research. Indeed, prior data visualisation findings must be used once the machine learning model is trained to not only have the most important variable but also to provide information about the U.S. market of that specific variable. The most significant influence of each variable can be seen when there are large changes in customer satisfaction for each value. Pay attention that this method completely laid off pattern visualisation because each graph explains only a specific variable and not the relationship of that variable with all the others.

By visualising the bar chart in *figure 7*, the following variable seems to have a low influence on customer satisfaction: Gender; Customer Type; Departure Arrival Time Convenient; Ease of Online Booking; Gate Location; Food and Drink; Check-in Service; Inflight Service. There is no relevant difference between the distribution of neutral or dissatisfied and satisfied customers.

Although Customer Type is not an influential variable, it is considered a consequent satisfaction variable. As anticipated in the previous section, loyalty can be informative to the dataset's quality. In this case, 80% of customers are loyal; loyal customers are more likely to answer and rate services for the company they like. This explains why the percentage is very high; the research aims at having a high-quality dataset in order to provide which variables are the most important for satisfied customers. The fact that prior research confirms the assumption that loyal customers complain and that almost 50% of loyal customers are not satisfied is a positive sign of the quality of the dataset, however, a larger proportion of disloyal customers are also not satisfied with the service.

The bar graphs in *figure 7* also show the variable object of this research, customer satisfaction. it can be seen from the bar graph that the satisfaction variable is more or less evenly distributed, specifically there are fewer satisfied consumers (42%) and more neutral or unsatisfied consumers (58%). However, as mentioned in the section on customer satisfaction, we know that this phenomenon is less in service companies and even less in airline companies due to the high degree of standardisation; the bar graph can therefore serve as further confirmation of the existing literature

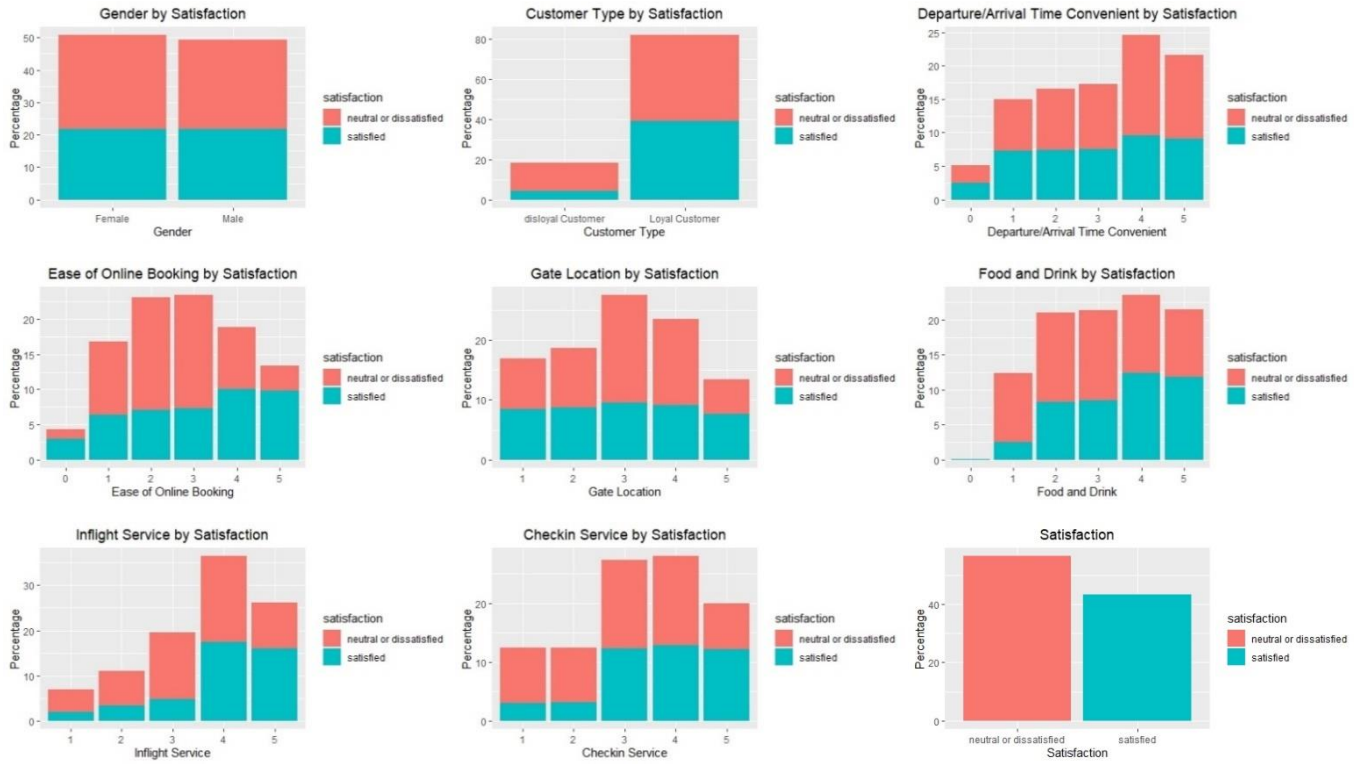


Figure 7 U.S. Airline Passenger Satisfaction (2015)

The following variables, by using bar chart visualisation in showed in *figure 8*, seem to influence customer satisfaction: Type of Travel; Class; Inflight Wi-Fi Service; Online Boarding; Seat Comfort; Inflight entertainment; Onboard Service; Leg Room Service; Baggage Handling; Cleanliness.

The presence of satisfied and unsatisfied consumers seems to be influenced by the change in the mentioned "dimensions."

Type of Travel is a binary variable which reports whether the trip was a Business or Personal trip. This explains the reasoning behind each travel which can be very informative in understanding the customer base. The number of business travel passengers is much higher with regards to personal travel. This could mean that U.S. airlines' dataset business is tailored to focus on a specific customer. Around 70% of customers took a business trip versus 30% of customers who travelled for personal purposes. As demonstrated in the *Type of Travel bar chart* below, more than half of customers who travelled for business purposes were satisfied. Instead, customers who travel for personal reasons are more likely to be unsatisfied. These findings drive even more our assumption that airline companies are tailor-made for business trips because they can satisfy customer needs when travelling for business instead of for personal purposes. 53% of U.S. residents took a business trip once a year in 2017, which fell to 36% in 2022 due to the implementation of remote working (Statista Survey, 2017; Statista, 2022). Considering that the survey that has generated the dataset

is only among people who have travelled, it is evident that the percentage of people that took a business trip is much higher.

Class is a categorical variable with three different values. It is very similar to customer type, but instead of identifying the type of travel (business or personal), it identifies the class used to travel. For this reason, not only are the information of flights provided, but also the hypothetical magnitude in price can be computed assuming that business class is much more expensive than eco and eco plus. A considerable number of observations were made during a trip using the business (48%) and economy (45%) classes. What is interesting in this case is to see the same pattern that we have identified in the customer type dimension. Customers who travel using business class are more likely to be satisfied than those who travel using economy class. However, in this case, the variable class does not explain the reasoning behind a specific trip but the related quality-price. The difference in more satisfied customers in business class may highlight one more time the previous assumption about the fact that the companies could be tailor-made for a business trip. However, in this case, the number of customers in economy class is almost equal to the number of customers in business class and what difference is just the proportion of satisfied customers. Business classes are obviously more comfortable, but the price is higher. For this reason, it is not justified the fact that customers in the economy class are more likely to be dissatisfied, they pay a lower price, and they should expect a lower quality service; for this reason, their evaluation of satisfaction should be balanced. This could highlight the lack of service quality in eco class which seems to be not aligned with consumers' expectations and paid price. Indeed, these assumptions offer the proposition for future research in which it should be understood why customers in the eco class are less satisfied if the price justifies the service.

Inflight Wi-Fi Service seems to have a strong impact on dissatisfied customers. Even with a rating of 3, which is in the middle, the majority of individuals are dissatisfied. At a rating of 5, almost all customers are satisfied. This finding is very interesting because having a rating of 5 on this variable could really have a strong impact on airline companies, however as said before, the influence of other variables is not considered in the bar chart, which means that a 100% focus on inflight Wi-Fi service may be useless if other important variables can drastically reduce the influence of the variable of interest. Another assumption is that the presence of customers on a business trip drives the fact that the presence and the quality of Wi-Fi service are important for work reasons.

Online Boarding shows a very significant gap between low and high values; indeed, this variable could really make the difference in having a large number of satisfied customers having a

rating of 4 or 5 drastically increase the proportion of satisfied customers. However, these ratings are provided by only 50% of the total individuals in the dataset; 70% of customers are satisfied in the case of a rating of 4 and 87% of customers are satisfied in the case of a rating of 5.

Seat comfort, inflight entertainment, leg room service, baggage handling and cleanliness report a similar bar chart in which it is possible to see a greater influence but without evident changes in satisfied customers by each rating. due to the reduced size of the bar graphs in the text, the original size of each bar graph can be found in *appendix C*

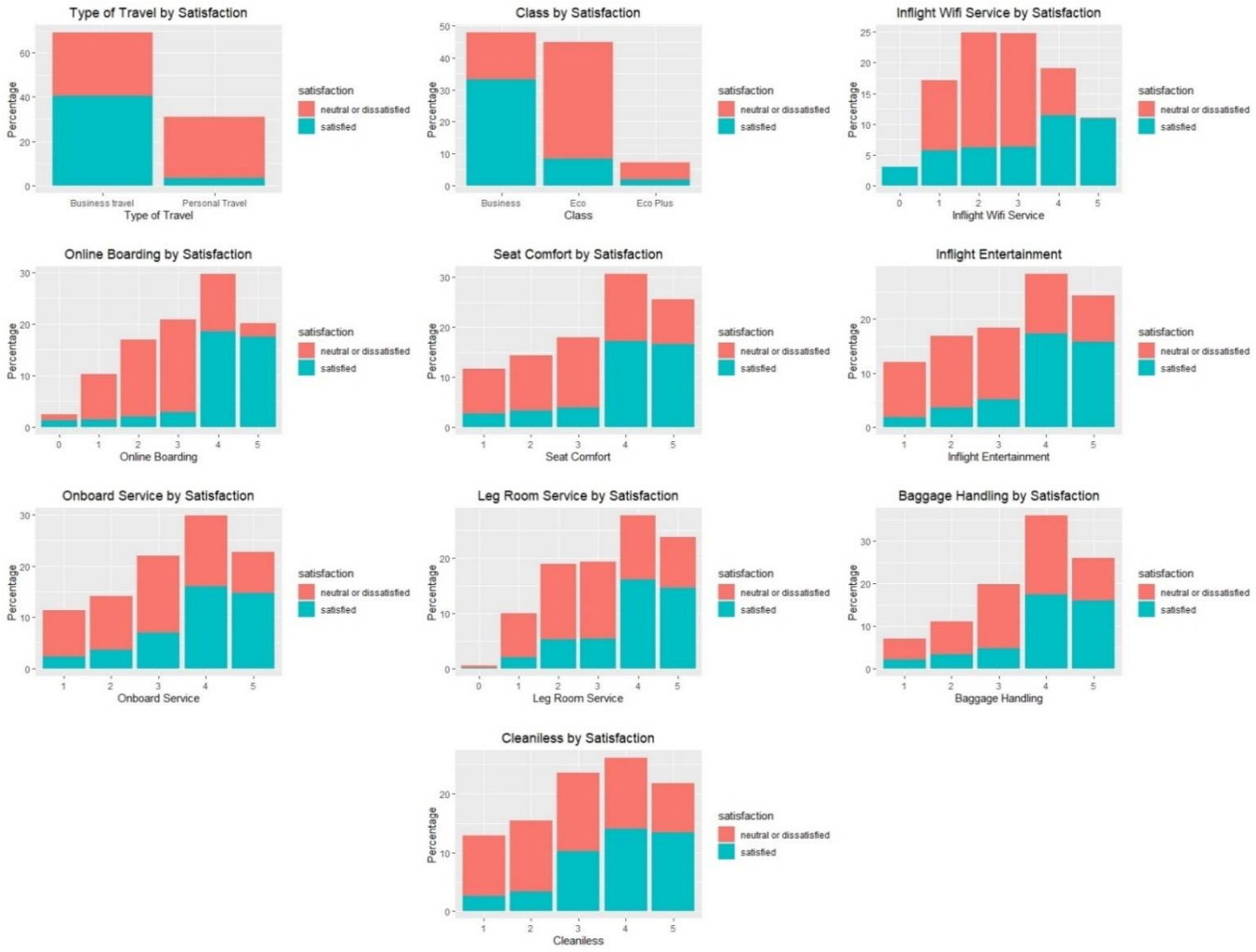


Figure 8 U.S. Airline Passenger Satisfaction (2015)

Correlation

In pre-analysis, the linear correlation is beneficial to identify which variable is mostly related to the outcome variable satisfaction or if there are multicollinearity problems in the dataset. Even though correlation does not imply causality, it is usually used by marketing professionals to evaluate the campaign to test consumers' reactions to different stimuli. It is important to state that computing correlation could seem very similar to the scope of this research. However, compared to the machine learning model in which other variables influence the effect of a variable, correlation

is just a linear relationship between two variables. It does not take into account the effect of other variables. For this reason, correlation is very good at discovering a linear relationship. However, it may not consider some crucial patterns in the dataset. The linear relationship between two variables of the U.S. Airline Passenger Satisfaction Dataset is computed using the Pearson correlation coefficient (1),

$$r_{sv} = \frac{\sum_{i=1}^{n}(x_{i_s} - \mu_{x_s})(x_{i_v} - \mu_{x_v})}{\sqrt{\sum_{i=1}^{n}(x_{i_s} - \mu_{x_s})^2 \sum_{i=1}^{n}(x_{i_v} - \mu_{x_v})^2}} \quad (1)$$

There is not a variable which distinctively influences customer satisfaction. Most of the variables (excluding arrival/departure delay in minutes, gate location, departure arrival time convenience, age and gender) seem to have a middle impact on customer satisfaction. The most correlated variables with customer satisfaction are online boarding, class, type of travel and inflight entertainment; the highest linear relationship with customer satisfaction is with the variable online boarding (r = 0.5). As stated before, with an r score of 0.97, arrival and departure delays are almost identical. Even if there is no influence on customer satisfaction, arrival delay will be removed from the training set to avoid multicollinearity problems. In *figure 9*, the full correlation matrix is reported.

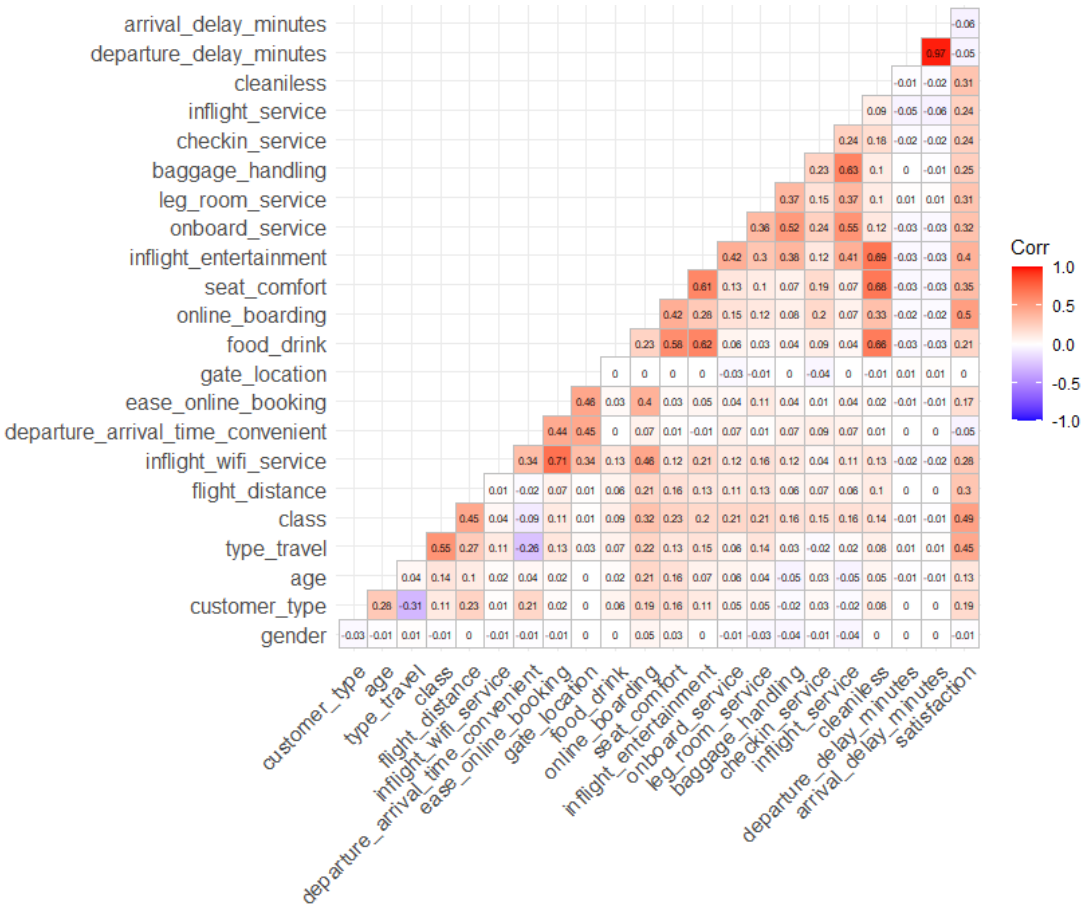


Figure 9 U.S. Airline Passenger Satisfaction (2015)

It may be useful to compute different correlation matrices minimising the number of variables included. In *figure 10* the correlation matrix of only the variables for which customers have rated the service is reported. In this case, it is helpful to understand how variables correlate with customer satisfaction and how all the rating variables correlate with each other. Indeed, it is possible to make some assumptions which can be used in future research.

Inflight WiFi service is highly correlated with ease of online booking (r = 0.71). Those two variables have the online channel in common, so this correlation may be explainable by the fact that a company which scores high in the online channel is more inclined to understand the value of the internet for their customers, so online services are in the top list for investments. Moreover, internet connection is beneficial for business people during travel which may also need ease of online booking to save time. A high proportion of business people may have rated the same variable highly. Cleanliness is highly correlated with inflight entertainment (r = 0.69), seat comfort (r = 0.68) and food drink service (0.66); all these three variables are also correlated with each other. A possible explanation is that customers may perceive more comfort if the environment is clean. This would explain the linear relationship between seat comfort and cleanliness. The same assumption can be made for food and drink service because an environment has to be clean to have lunch or dinner. There is no reasonable assumption for the relationship between inflight entertainment and cleanliness. To simplify the visualisation, several correlation graphs of the same variables have been computed by reducing the number of them so as to highlight the highest scores; it is possible to view the different correlations with their specifications in the *appendix D*

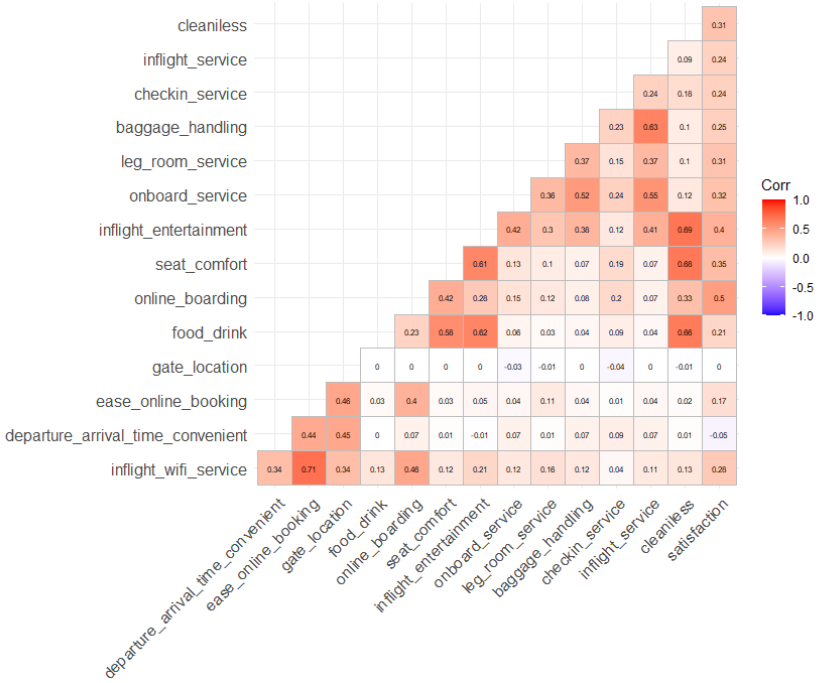


Figure 10 U.S. Airline Passenger Satisfaction (2015)

DATA PRE-PROCESSING

Duplicate Values

There are no duplicate observations by variable “ID”, which is an identifier of customer by trip. For this reason, a customer could have travelled more than once and rated more than one single trip. Also, whether people belong to the same family is unknown. From a marketing perspective, the fact that a customer has left a rating multiple times does not influence the scope of the research because the satisfaction rating is referred to a single trip. There may be an influence on satisfaction if a customer is loyal or not. Evidence shows that higher loyalty leads to a higher willingness to complain (Namkung, Jang, & Choi, 2011; Grégoire & Fisher, 2008). For this reason, it is possible to state that ratings are an informative evaluation of an air trip, knowing that if a customer is loyal, ratings are even more accurate compared with a non-loyal customer. Distribution of loyal customers can be used to express the quality of the dataset, the higher the number of loyal customers, the more accurate the answer to the survey.

Missing Values

Only the “arrival delay in minutes” variable in this dataset contains a missing value. More specifically, there are 393 missing values; knowing that “departure delay in minutes” does not contain missing values allows us to replace missing values with an unusual method. If an airplane is late at the moment of departure, there is a very high chance that it will be late at the moment of arrival. For this reason, replacing those missing values with the mean or median of the variable with missing values is not meaningful. Instead, we can use the same value provided in the departure delay in minutes variable.

Sixty-one thousand two hundred five times, departure delay and arrival delay are exactly the same value; for the rest of the data, those values are not the same. However, if we consider that it is improbable to have the exact same minutes, we can assume that 14 is perceived as 15 minutes of delay by a customer, and since those two variables are strongly correlated ($r = 0.97$), replacing missing values of arrival delay with departure delay should be the right strategy instead of using the mean or median of the variable in question. “Arrival delay in minutes” and “Departure delay in minutes” are almost the same variable. One has to be removed to train the model, but both can be used to provide some interesting data visualisation.

Outliers

The main method used to individuate outliers was both the box plot reported in *figure 11* and *figure 12* and the computation of mild and extreme outliers.

Variables that can contain outliers are Age, Flight Distance, Departure Delay and Arrival Delay. The other variables are nominal or with an upper limit, so they cannot contain outliers. Variable Age does not contain outliers.

Flight Distance contains outliers; more specifically, there are 2,855 mild outliers (values higher than interquartile* 1.5); however, there are no extreme outliers (values higher than interquartile * 3). Many people prefer to remove outliers because they can influence the analysis results. Instead, these findings suggest that outliers can be informative depending on the variable's type. Hence, it was decided not to remove them. The flight distance is a real value, and as tested, there are no extreme outliers; one can assume that there is no error in this variable (e.g., one mln km). Also, the fact that a low number of flights is of very high distance is in line with the airplane travel business. Airlines tend to operate where they belong, for example, ITA airways has a lot of flights from Rome to Milan but a smaller number of flights from Rome to New York. Also, long-distance flights are less frequent due to their cost both for the company and the final customer. 2,855 mild outliers are kept and used in the analysis.

Using the previously reported scatter plot (better visualization) and the boxplot below, we can see the exact number of observations with flight distances higher than 4,000 those observations could be considered outliers. However, considering the longest flight ever recorded which is from Singapore to New York, of 15,349 km/9,537 miles (not present in the dataset); 4,000 miles are not a very large covered distance, emphasising the feasibility of this distance for air travel and thus informative for the analysis.

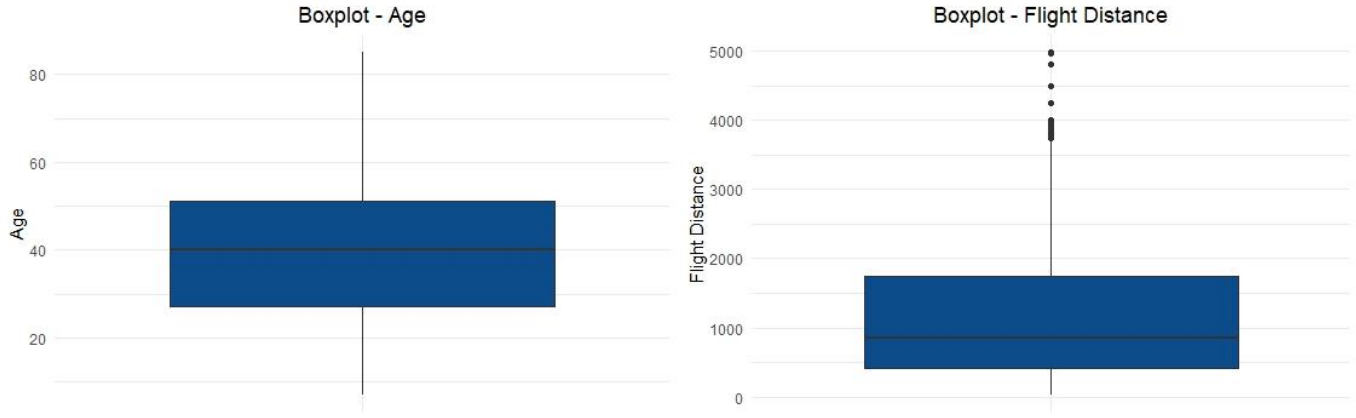


Figure 11 U.S. Airline Passenger Satisfaction (2015)

Arrival Delay and Departure Delay are different types of variables, still continuous, but they contain a lot of 0 values which means that the airplane was not late. Identifying outliers using the boxplot method is a little harder in this case because almost all values greater than one are considered outliers. Hence, these variables need to be transformed. Transformation performed for Arrival Delay and Departure Delay variable consists of removing all values equal to 0 and then computing the log of the remaining values to change the magnitude and reduce the influence due to the high frequency of low values. There are 20 mild outliers for the variable Departure Delay using the log of related values, and 14 outliers for the variable Arrival Delay. The maximum delay for both variables is approximately 26 hours (1,592 minutes for Departure Delay and 1,584 minutes for Arrival Delay). It is likely that a flight scheduled for a specific time can be delayed by many hours; there can be several reasons (weather, technical problems, accidents) that are not specified in the dataset.

The presence and numerosity of outliers in the air company survey are known; observations considered from the dataset will not be removed and included in the machine learning model. Using the scatter plot (*figure 5 and figure 6*) in the previous section, it is possible to see that satisfied and dissatisfied customers are not influenced by outliers.

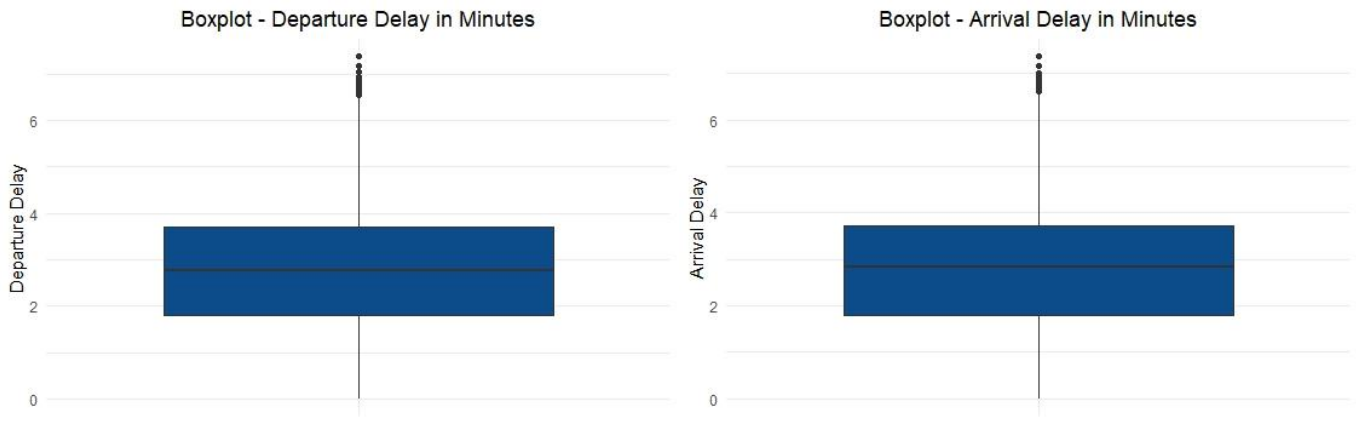


Figure 12 U.S. Airline Passenger Satisfaction (2015)

Labels encoding

To process the dataset, it is essential to transform non-numerical variables to feed the model with meaningful numbers instead of text values. Gender, Type of Travel and Satisfaction are nominal variables. Hence, the numbering just identifies a specific attribute and does not include any magnitude or distances; hence, dummy variables were created for Gender, Type of Travel and Satisfaction. In the case of Gender, the dummy variables created specify when the gender is male; in the case of Type of Travel, the dummy variable reports the presence of business travel, and lastly,

the variable which is the main purpose of this research reports the presence of a satisfied customer; a value of 1 is reported when the attribute of interest is present 0 when it is not.

Class could also be viewed as a nominal variable because each class is just described by a specific value, hence could be possible to create two dummy variables to identify the presence of Business Class and Eco Class; however, marketing reasoning takes place in this specific case because it is possible to identify class as an ordinal variable, where Eco Plus is the lowest level, and Business class is the highest. Indeed, the price and service quality of those three classes (Eco Plus, Eco and Business) follows the ordering principle. It is reasonable to treat this variable as ordinal in which there is a magnitude between each value. For this reason, 0 is equal to Eco Plus, 1 is equal to Eco and 2 is equal to Business Class.

MACHINE LEARNING MODEL IMPLEMENTATION

Once the dataset is cleaned, it is possible to start building machine learning models. The training and test set splits used for all the models is 70% of the training set and 30% of the test set. Due to the purpose of the research, the hold-out set was not used because it is mainly used in the case of maximising the generalisability of models' prediction and comparability between them. In doing this, it is possible not to reduce the number of individuals for the training set too much.

Satisfaction is an antecedent of customer loyalty (Anderson, Fornell, & Lehmann, 1994; Anderson, Fornell, & Rust, 1997; Fornell, 1992). Hence customer type cannot be used as a customer satisfaction predictor and thus removed from the dataset. Another removed variable is arrival delay in minutes to avoid multicollinearity problems due to its high correlation of 0.99 with departure delay in minutes.

The main metric used to evaluate models' performance is the AUC which allows for better evaluation and compares models among each threshold. Indeed it is possible not only to see the AUC score but also the ROC Curve for each model built. Other reported metrics are Accuracy, Recall, Precision and F1 Score, mainly used to detect anomalies but can also be useful for comparing models. For these reasons, all those metrics will be reported in each model section.

The script of each model with the following libraries used: (Pedregosa, et al., 2011) (McKinney, 2010) (Lundberg & Suu-In, 2017) (Harris, et al., 2020), including label encoding, data cleaning, model implementation and interpretability can be downloaded at the following git hub repository: https://github.com/StefanoFaiola/MLCustomerSatisfaction

Logistic Regression

Logistic regression is one of the simplest and most interpretable models for classification. Despite its name, it is used for classification rather than regression. Using the *Logistic Function* (2), it is possible to model the probabilities to describe the possible outcome:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{2}$$

In the standard *logistic function (3)* where L = 1, k = 1, $x_0 = 0$ the output ranges from 0 to 1, this function is also called sigmoid. The exponent of *e* is the formula used in the case of linear regression, where β is the coefficient for each attribute. The higher the coefficients, the greater the influence of the attributes.

$$\mathrm{P(y\ =\ 1)} = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k)}} \tag{3}$$

For this reason, it is possible to still use coefficients to fully interpret a logistic regression model and understand which variable mostly influences customer satisfaction. The logistic function is just a transformation to obtain the probability of a possible outcome occurring. The outcome will be positive if the probability p is higher than the set threshold. It is possible to set many different thresholds to change the model's behaviour. For example, to maximise the recall and capture as many satisfied customers, it is possible to set a lower threshold below 0.5 so that the number of predicted satisfied customers will be higher. On the other hand, more dissatisfied customers will be predicted as satisfied.

The logistic regression model is good at predicting customer satisfaction (AUC = 0.917); hence its coefficients can be used to understand the most influential variables in air travel to have a satisfied customer. Metrics used to evaluate the model are reported in *Table 2*. The threshold used is 0.5, and through the ROC curve in *Figure 13* it is possible to see how the model behaves at different thresholds.

Metrics	Score
Accuracy	0.859
Precision	0.848
Recall	0.823
F1	0.835
AUC	0.917

Table 2 Logistic Regression Scores - U.S. Airline Passenger Satisfaction (2015)

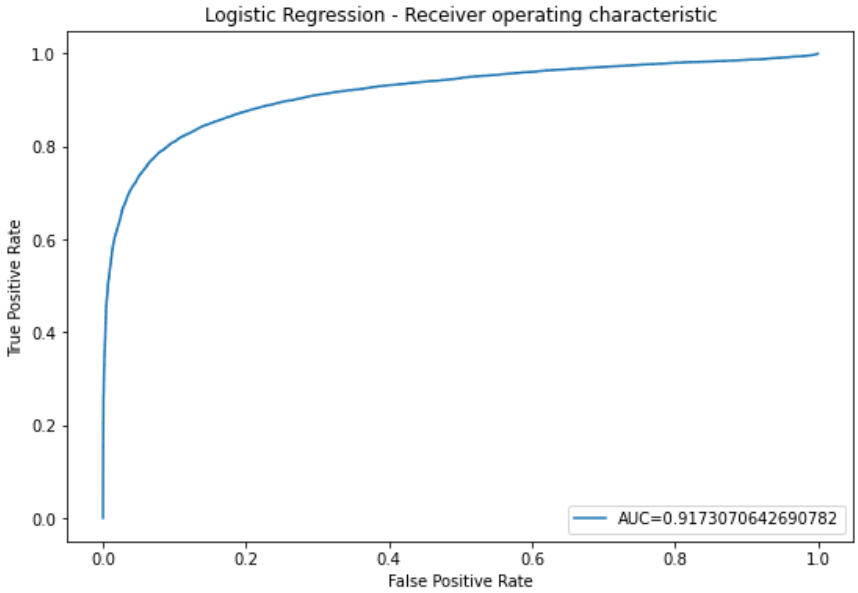


Figure 13 Logistic Regression ROC Curve - U.S. Airline Passenger Satisfaction (2015)

Regression coefficients are strictly related to the variable used in the model. For this reason, to compare them, it is mandatory to previously standardise the input variables to have standardised regression coefficients comparable to each other.

The coefficients of each variable in the logistic regression are shown in *figure 14*, from the highest to the lowest. Type of travel, online boarding and inflight Wi-Fi service are the three most influential variables that positively influence customer satisfaction.

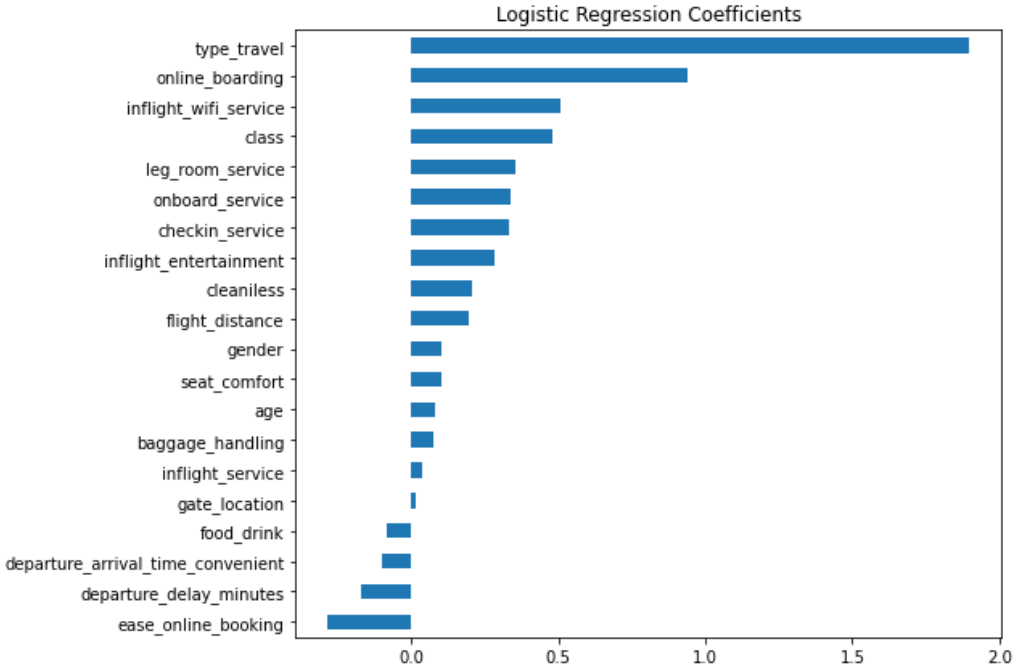


Figure 14 Logistic Regression Coefficients - U.S. Airline Passenger Satisfaction (2015)

Tree-Based Models

Tree-based models can vary from being very interpretable and poorly performing to poorly interpretable and very performing. it is possible to remain in the same context by varying the type of model, which will, however, always be based on the creation of decision trees. The study reports the three most common decision tree-based models, Classification Tree, Random Forest and Gradient Boosting; each of the tree models has a different degree of interpretability and performance.

Classification Tree

Classification Trees are based on asking a question at each node about the data. For this reason, it is possible to interpret the model also by looking at the tree with the possibility to inspect each question asked to determine the prediction at each node level. The process of building a tree begins with the entire dataset. In fact, the first problem is to decide what question should be asked first. This is where building the tree comes in, and the quality of a question needs to be measured to understand how beneficial it is for the tree as a whole.

Tree-based models must choose the question that best splits the data based on the outcome variable. An ideal split will report all the yes in one column and all the no in another. It is the so-called pure node. Ending up with a mixture of people means that the question is not quite so good, and there is a need to minimise the mixture of people asking the right question. The splitting is not random because it is guided by the idea of generating subgroups which are more and more

homogeneous with respect to the response variable. For this reason, there are specific rules depending on whether the problem is of classification or regression.

Since the response variable is binary, in building the model, the method used was the Gini index decrease in impurity. Gini Index (4) ranges from 0 to 1 and is computed for each node to compute the decrease in impurity (5), which stays for heterogeneity. It is a generic concept, something which is impure and not homogeneous. Impurity is employed for evaluating the goodness of the available splits. This means that you can rank the splits. You can choose the best one that separates your data into groups as homogeneous as possible with respect to the response variable.

$$i(t) = 1 - \sum_{j=1}^{j} f_J^2\,(t) \tag{4}$$

$$\Delta i(t,s) = i(t) - [i(t_l)p_l + i(t_r)p_r] \tag{5}$$

For the purpose of just predicting the outcome variable, it is mandatory to set some hyperparameters to avoid overfitting and obtain a generalisable model. A final hyperparameter is obtained through multiple iterations comparing the decrease in accuracy for each max depth of the tree so that accuracy is maximised with the most generalisable model; the hyperparameter is max depth = 5. Metrics used to evaluate the model with the hyperparameter applied are reported in Table 3, plus the ROC curve is reported in *figure 15*. The model performs better with regard to the logistic regression.

Metrics	**Score**
Accuracy	0.903
Precision	0.885
Recall	0.892
F1	0.889
AUC	0.965

Table 3 Classification Tree Scores - U.S. Airline Passenger Satisfaction (2015)

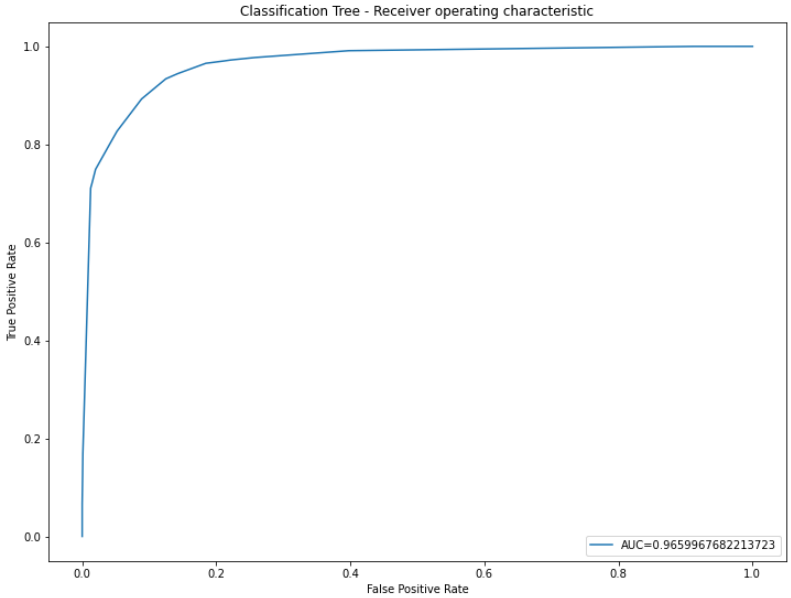


Figure 15 Classification Tree ROC Curve - U.S. Airline Passenger Satisfaction (2015)

In providing features importance using the Gini score, the whole tree was used to have a value for each variable; there are no principal differences in Gini scores using the whole tree or the pruned one. What changes is the magnitude of values which are higher in the case of the max depth tree because the Gini score is summed up each time a variable is chosen for a specific split. The three most important variables, as shown in *figure 16*, are still online boarding, inflight WiFi service and type of travel with a different order compared with logistic regression; online boarding jumped to first place using the classification tree model.

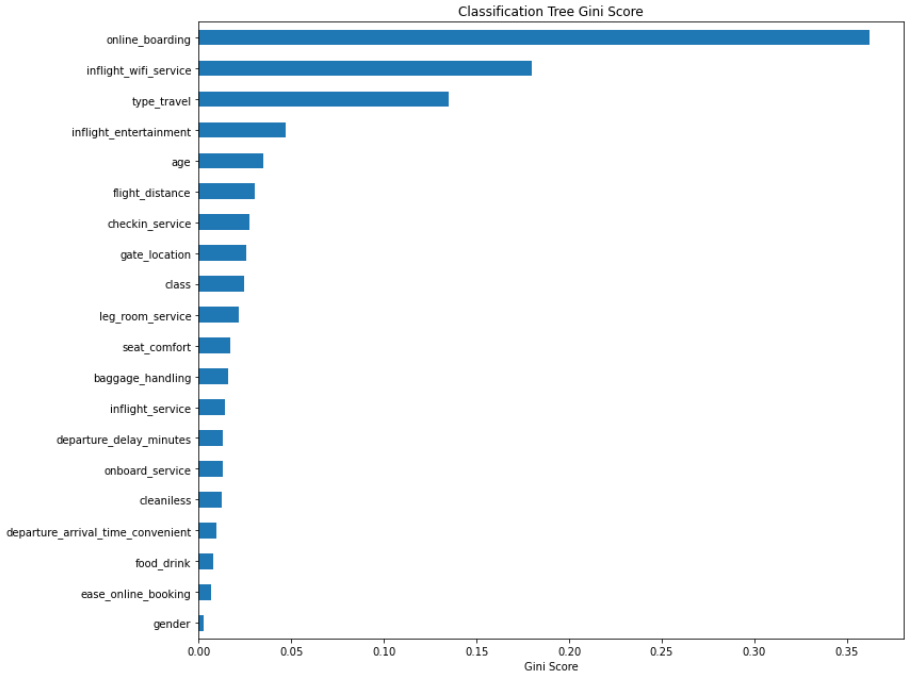


Figure 16 Classification Tree Gini Scores - U.S. Airline Passenger Satisfaction (2015)

The peculiarity of a single tree-based model is that it is possible to inspect how each decision is made at each different split. Indeed it is possible not only to rank the most important variable using the Gini score or the position in the tree graph, but it is also possible to see at which threshold a specific variable was set in order to make the best question possible to the dataset and minimise the impurity of each split. A max depth of 3 Classification Tree of the training set is reported in *figure 17*. The question is reported in the first line of each leaf, and observations are divided based on the answer (true or false) of that question. Value reports the numbers of split samples: on the right, dissatisfied customers; on the left, satisfied customers. In the case of the first node, there are 90,916 samples (70% of 129,880), 51,421 dissatisfied and 39,495 satisfied. If the answer to the question online boarding score <= 3.5 is false, 45,156 consumers are sent to another node, and the number of satisfied customers is 72% of the node. In this way, using this question, the model would be able to predict this level of confidence.

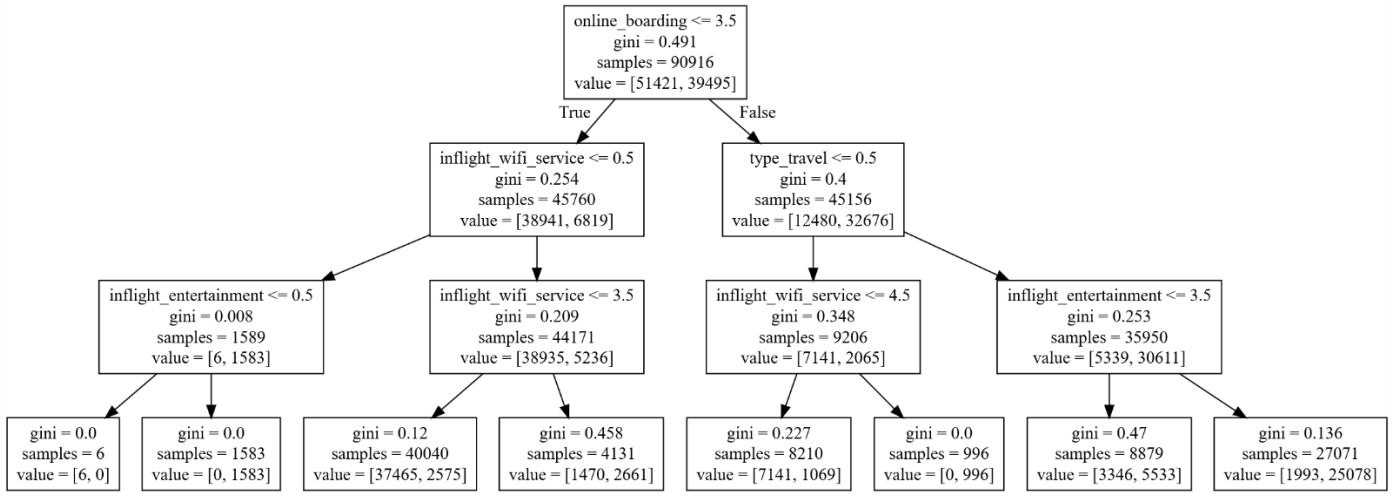


Figure 17 Classification Tree Splits - U.S. Airline Passenger Satisfaction (2015)

Random Forest

A single tree, in general, is a weak learner and is not very performant alone. Moreover, it is highly susceptible to overfitting; indeed, it provides the building blocks for ensemble methods like Random Forest and Gradient Boosting. Random Forest is one of the most common and effective ways of improving the performance of a decision tree. In general terms, a random forest is a group of slightly different trees. After each of those trees is generated, they will vote for the most popular class, and the majority voting will be the assigned decision for the proposed observation. Breiman (2001) defines random forests as a *"classifier consisting of a collection of tree-structured classifiers {h (x, k), k = 1, . . .} where the {k} are independent identically distributed random vectors, and each tree casts a unit vote for the most popular class at input x."*

As normally expected, the random forest model performed better than the single decision tree. Scores and roc curve are reported respectively in *Table 4* and *figure 18*; there are not very large differences in terms of performances compared with the classification tree. In general, even a small percentage increase in accuracy is preferred in machine learning, however for the scope of this research, interpretability should not decrease and a trade-off between model performance and interpretability should be found. In this specific case, an accuracy increases of 2.3% is a good improvement, especially if a large number of observations are under analysis. For example, with 200,000 decisions, the model would be able to correctly predict 4,600 observations more compared with a single decision tree. Indeed, there is a high similarity in the most important feature between the classification tree and the random forest. However, the possibility of understanding how the trees were generated has been completely lost.

Metrics	**Score**
Accuracy	0.923
Precision	0.915
Recall	0.906
F1	0.911
AUC	0.971

Table 4 Random Forest Scores - U.S. Airline Passenger Satisfaction (2015)

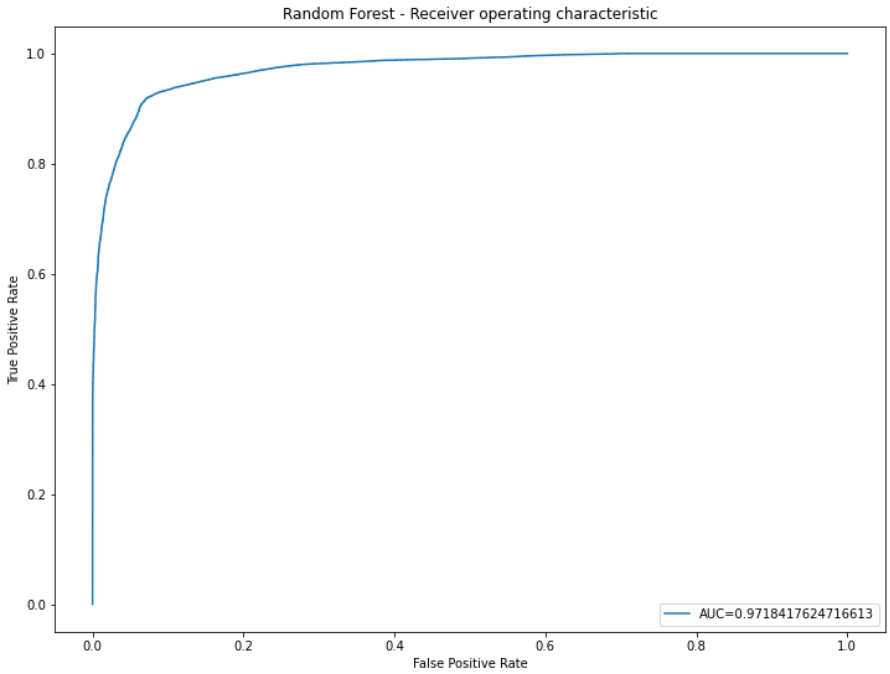


Figure 18 Random Forest ROC Curve - U.S. Airline Passenger Satisfaction (2015)

In building the model to predict customer satisfaction, the number of generated trees and the max depth of each one, are two set hyperparameters, respectively equal to 200 and 5. For this reason, it is impossible to analyse how decisions were made differently from a single tree because having a high number of trees drastically reduces the model interpretability. However, it is still possible to compute the impurity-based feature importance (Gini index decrease in impurity), like in the case of a single tree. The main difference, in this case, is that feature importance (*figure 19*) is computed as the mean accumulation of the impurity decrease within each tree (Pedregosa, et al., 2011). The first two most important features are online boarding and inflight WiFi service; in the case of a single classification tree in the previous section, using random forest, there are four main features that influence customer satisfaction due to variable class being in third place. It was in the fourth position in the logistic regression model and the ninth position in a single classification tree. Using random forest, model accuracy increases, but due to the computed averages of Gini decrease in the impurity of each variable, the difference in scores of each variable is smaller compared with a single classification tree and the most important variables, despite being evident in the graph, appear to have less influence on the overall pattern. The curve generated by the scores has a less steep descent, and it is possible to find an elbow point between the fourth and the fifth variables. However, the elbow is not highlighted as in the single classification tree model.

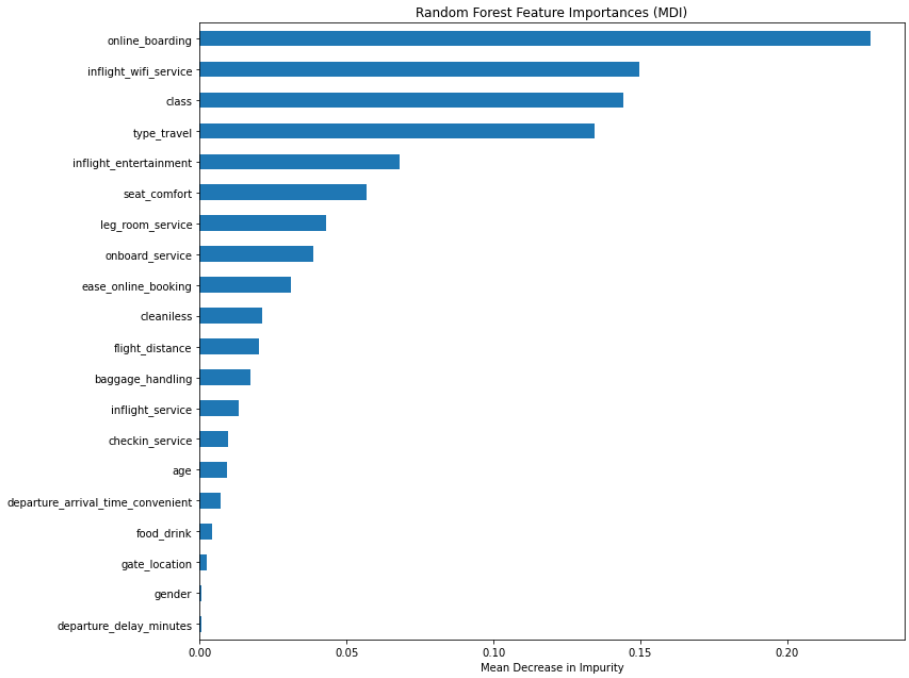


Figure 19 Random Forest Mean Decrease in Impurity (MDI) - U.S. Airline Passenger Satisfaction (2015)

Gradient Boosting

Like Random Forest, gradient boosting is an ensemble model which involves more than one tree that has to be generated. The peculiarity of this model and the main difference with the random forest is that this *"algorithm builds an additive model in a forward stage-wise fashion"* (Pedregosa, et al., 2011); in fact, despite random forests allowing parallelism (generate many trees at the same time), gradient boosting requires a generated tree in order to generate the next one. While random forest aims to come up with a good prediction by simply training more and more models and averaging the results, gradient boosting identifies whether the model is performing poorly and then aims to train the next tree to fix that issue. This is done by looking at the error term (how wrong the model is at predicting certain points). For example, in regression, one can look at the squared error. The model aims to fit this error by adding a weighting term, ensuring that the error term is minimised over time as more trees are built, specifically minimising that error.

In gradient boosting for classification, one needs to measure how correct each tree is on each row. Instead of using a classifier, a regressor needs to be used. This is similar to logistic regression. Each tree will natively output the probability of each individual being of a particular group (satisfied or not satisfied). This gives us the residual, which allows us to prioritise certain individuals above others when training a subsequent tree.

In building the model, some hyperparameters, the number of trees and the learning rate and loss must be applied. For this specific model, the number of trees is equal to 50, the learning rate is equal to 0.3, and the loss function is the same used in logistic regression, also called log loss on the library Scikit-Learn. The gradient boosting classifier is the most performant tree-based model of this research, as shown in *Table 5* and *figure 20*; accuracy is 2.5%, and AUC is 1.8% higher than random forest.

Metrics	**Score**
Accuracy	0.948
Precision	0.953
Recall	0.906
F1	0.940
AUC	0.989

Table 5 Gradient Boosting Scores - U.S. Airline Passenger Satisfaction (2015)

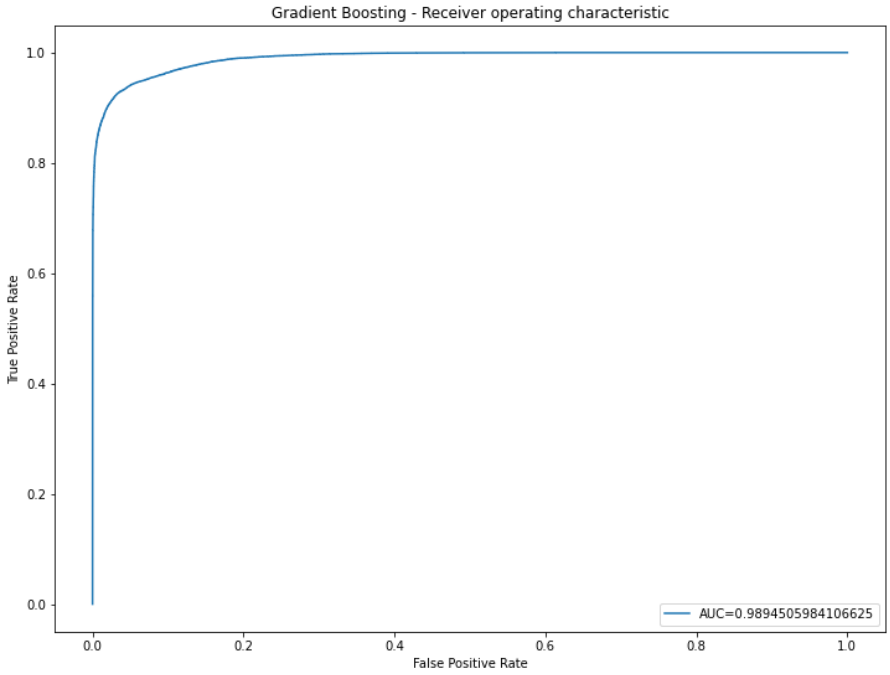


Figure 20 Gradient Boosting ROC Curve - U.S. Airline Passenger Satisfaction (2015)

Even though the number of generated trees is reduced compared with random forests (from 200 to 50), it is not possible to interpret them like in the case of a single tree. However, despite the reduction in interpretability using a gradient boosting classifier, it is still possible to compute the feature importance (*figure 21*), which is computed as the (normalised) total reduction of the criterion brought by that feature. It is also known as the Gini importance. The most critical feature determined by the gradient boosting classifier is online boarding, inflight Wi-Fi service and type of travel. It is possible to see an elbow point between the type of travel and class. Moreover, the same most important feature is individuated in the single classification tree model leading the two outputs to be almost the same. Due to boosting error, the gradient boosting classifier is very good at predicting with a high level of accuracy, almost using only three of the twenty input features.

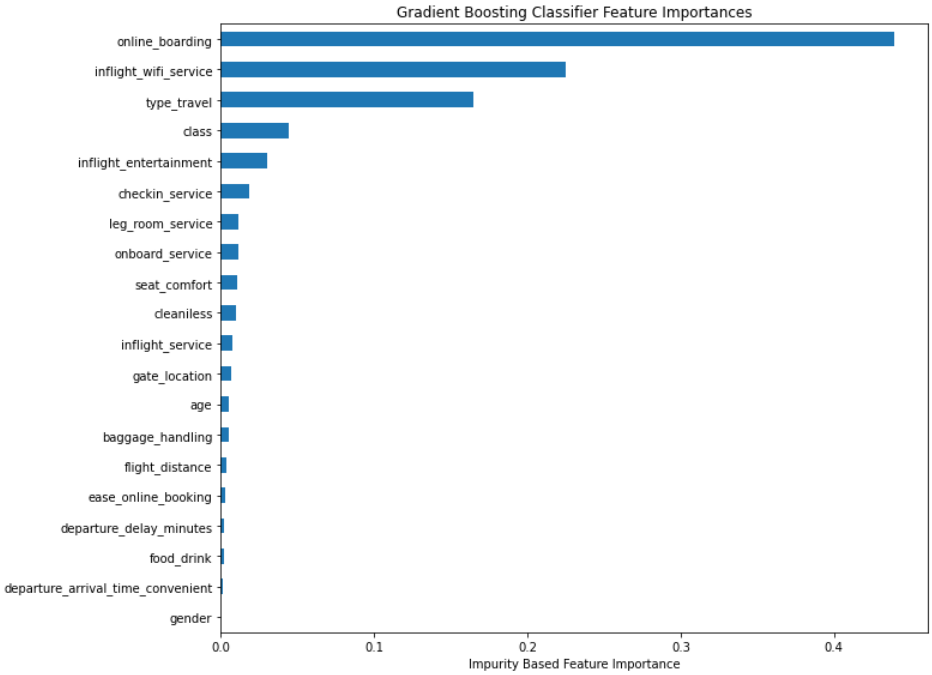


Figure 20 Gradient Boosting Feature Importance - U.S. Airline Passenger Satisfaction (2015)

Kernel Support Vector Machine

The main purpose of a Support Vector Machine (SVM) is to separate observations using a simple line. In the case of a dataset with more than two dimensions, the algorithm separates observations to find the optimal hyperplane. Usually, the classification algorithm works by distinguishing groups of observation, trying to find the most common characteristics, and maximising the differences between classes; SVM uses a different approach. Instead of looking for the differentiator of one class from another, it looks for the most similar samples between two classes. Those two samples are called support vectors, where the algorithm's name came from (Cortes & Vapnik, 1995).

In building the model, the kernel radial basis function was applied to map observation to a higher dimension and increase the goodness of the model. Although Kernel Support Vector Machine (KSVM) is one of the best performing models (*Table 5*) of this research, it is also the most computationally expensive and less interpretable model. Practically speaking, this model is very powerful, but it is not interpretable by design. However, it was used to have a high performant model that could be used with feature permutation importance to compare importance between different models, from the most to the less interpretable ones.

To provide the ROC Curve *figure 21*, it is mandatory to have probability estimates; however, SVM does not provide probability estimates by design. These are calculated using an expensive five-fold cross-validation (Pedregosa, et al., 2011). The decision made by the model could not be

100% aligned with the estimated probabilities. For example, a predicted probability of 0.5 could be a dissatisfied customer instead of a satisfied one, so the AUC score is reliable but cannot be fully compared with other built models in this research.

Metrics	Score
Accuracy	0.948
Precision	0.952
Recall	0.926
F1	0.939
AUC	0.985

Table 6 Kernel Support Vector Machine - U.S. Airline Passenger Satisfaction (2015)

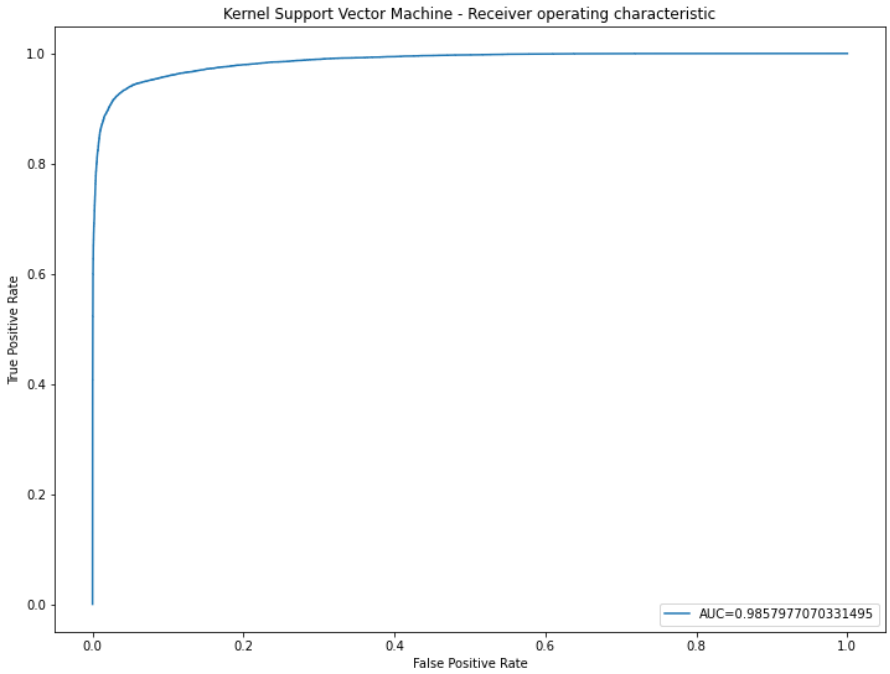
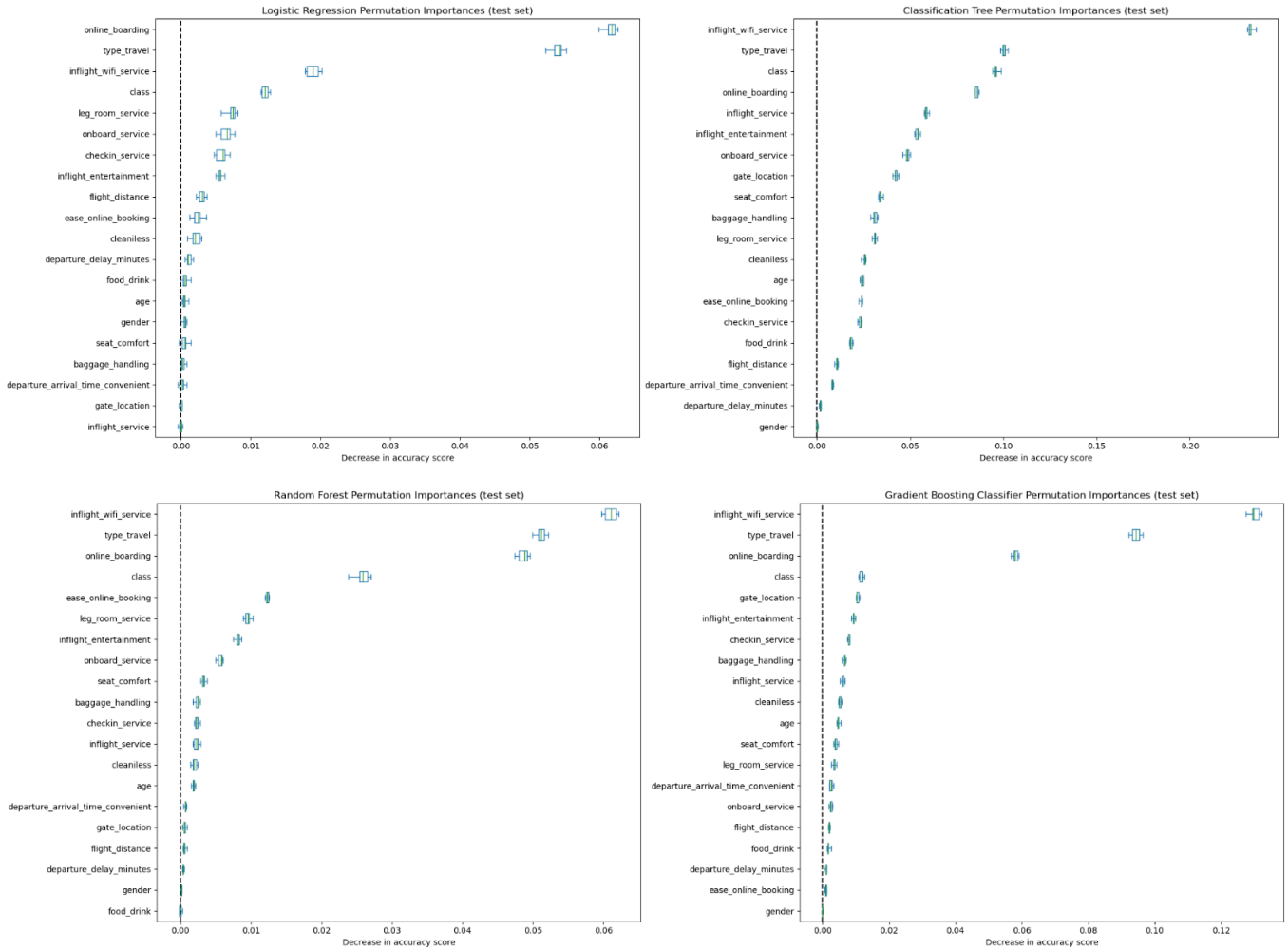


Figure 21 Kernel Support Vector Machine ROC Curve - U.S. Airline Passenger Satisfaction (2015)

Feature Permutation Importance

Feature permutation importance is in the spectrum of Global Interpretability techniques, which explains the model's behaviour across the full range of inputs. It can be used for every model, especially for black box models. Feature importance using permutation is scored using the decrease in accuracy. Specifically, the technique is based on taking one variable at a time and rescoring the model after the chosen variable is shuffled. This means data information of the chosen variable does not make sense anymore, and the accuracy should decrease. If the shuffled variable was a critical variable, the model's performance should drop significantly because the model can no longer rely on the chosen variable. The process is repeated for all the variables in the dataset. Once

it is finished, it is possible to plot the decrease in accuracy of each variable, sorting them from the most to the less important one. The main reason feature permutation importance is a global interpretability technique is that when one variable is shuffled, the model performances are still influenced by all the other variables. This concept is vital to understanding how interpretability by design differs completely from other interpretability techniques. A coefficient defines the model's output when related to a specific variable, which is different in computing the decrease in accuracy of the entire model. It is important to understand that global interpretability helps understand model behaviour but does not explain how it works. For this reason, comparing different results is mandatory to increase the power of global interpretability using feature permutation importance which offers a different perspective based on the model performances. All the results of feature permutation importance of the five used models are plotted in *figure 22*. A more detailed version of the following plot is also reported in Appendix E.



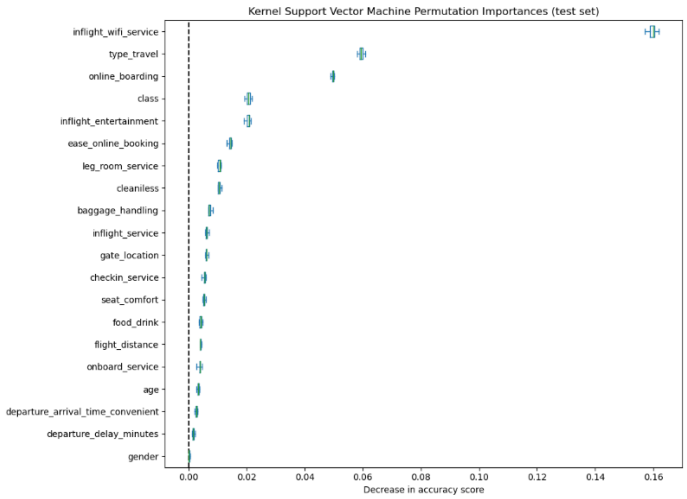


Figure 22 Feature Permutation Importance - U.S. Airline Passenger Satisfaction (2015)

The pattern of feature permutation importance of the five used models is the same provided in the coefficients of each model using interpretability by design. Indeed, there are still large gaps between the three and four most important variables to all the other variables. Inflight Wi-Fi service is at the top of the most important features and in the third position only for the logistic regression feature importance. The type of travel is always in the second position considering the five models permutation importance. Online boarding is in the third position for Random Forest, Gradient Boosting and Kernel Support Vector Machine permutation importance and in the first position for Logistic regression permutation importance. In classification tree permutation importance, online boarding is in the fourth place because class is in third place only for this model. Indeed, it is possible to provide many more comments about the provided partial dependence plots; however, for the sake of simplicity, only the first three variables are commented on. In the results section, a more general view of the variable of influence will be offered to compare all the acquired results.

Shapely Additive Explanation (SHAP)

Shapely additive explanation is introduced by Lundberg and Suu-In (2017). It provides both global and local interpretations and explanations related to interactions between variables. SHAP helps to analyse all possible combinations of a given feature and gives an idea of the individual importance of each feature. However, it also shows how different features affect the model's scoring. It is possible to see in *figure 23* four different SHAP beeswarm plots for the Logistic regression, classification tree, random forest and gradient boosting model; in the case of Kernel SVM, due to lack of processing power, it was not possible to run kernel explainer and compute SHAP values due to the complexity of the model.

SHAP plot reports display data from the most influential to the least. With SHAP's positive values, the prediction increases. In the case of classification, a positive prediction means that a customer is satisfied. For each feature, the beeswarm plot (*figure 23*) also reports its distribution using a minimised scatter plot which can be used to see each individual's influence on SHAP values; a red observation has a high value and a blue one a low value of the variable of interest which is reported on the left. Using the SHAP beeswarm plot, it is possible to individuate when high values influence a positive prediction or vice versa. In the case of this research, the majority of the features report a satisfied customer when the value of a specific feature is high. The reason is that the majority of features are customer ratings; when a rating of a specific service is high, it means that the customer has provided a good grade, which consequently influences a positive outcome in satisfaction; in the case of binary variables, 1 is considered high value and 0 a low value. For Type of travel, it is possible to see only a very strong red and blue. In the case of class, it is possible to see the three colours related to eco plus, eco and business class, where the business class takes the highest value (red dot). Indeed, the reason behind ratings and SHAP results demonstrate the accuracy of the beeswarm plot, which follows the logic behind rating and customer satisfaction (the higher the rating, the higher the customer satisfaction).

Tree based algorithms provide a different SHAP beeswarm plot compared to logistic regression. In fact, for the classification tree, it is possible to see that low values of a specific variable not only lead toward an unsatisfied customer but also towards a satisfied one. This is because trees are based on a specific path of questions, and the output is not computed like in logistic regression in which the coefficient is multiplied by the value of a specific feature; for example, if the coefficient of a hypothetical feature is 2 the value related to that feature will always be multiplied by 2 leading inconsistent output. In the case of trees, an observation with a low value of a specific feature can fall into leaves in which the majority of individuals are satisfied, and the prediction for that feature will be satisfied even if a specific value is low.

Using all SHAP plots, the most critical features are type of travel, online boarding, inflight WiFi service and class. These results are the same feature permutation importance, which is almost aligned with the coefficients of interpretable models by design. Using the SHAP plot, the research provides a more robust analysis of the most influential variables, offering also local interpretability.

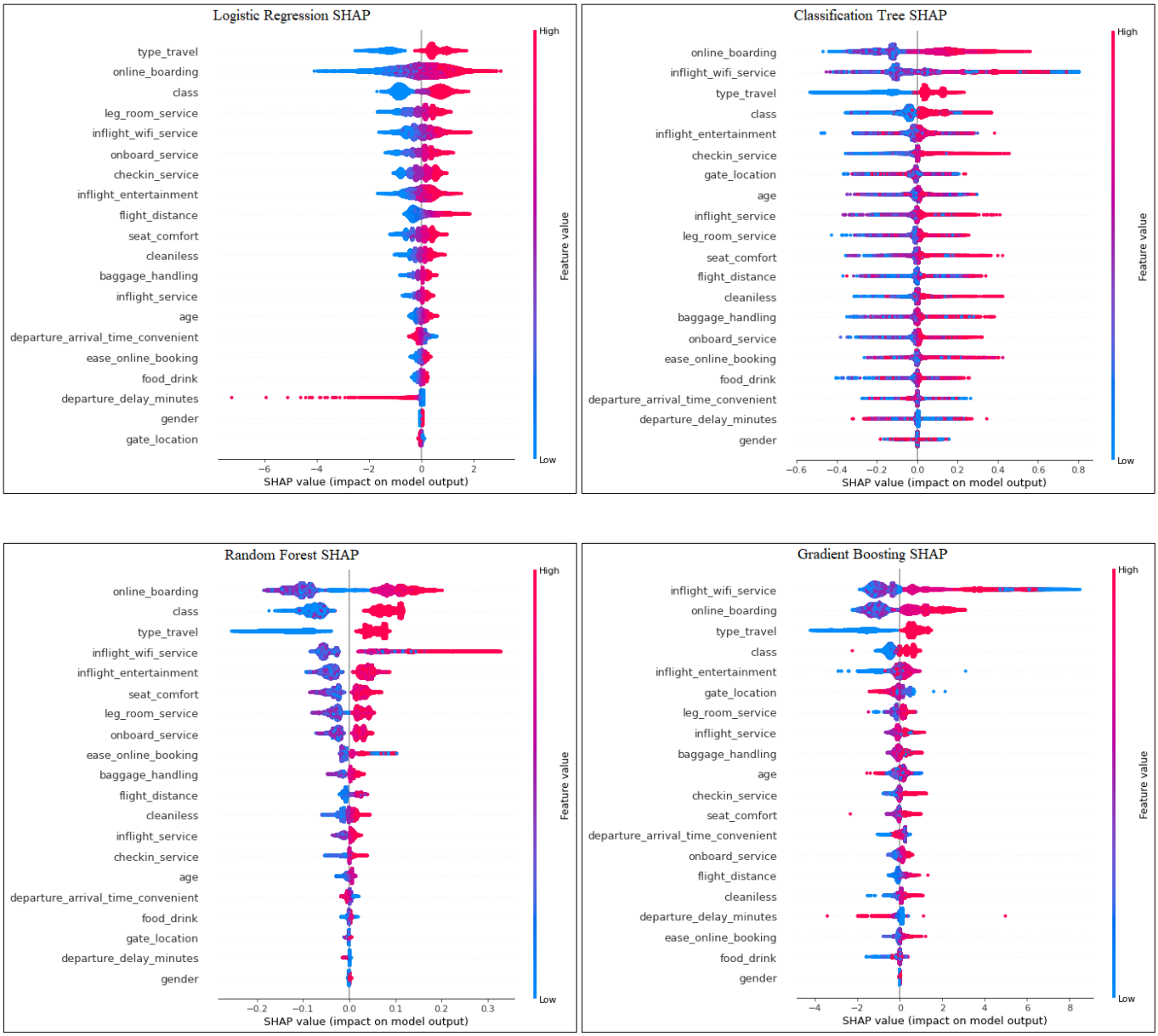


Figure 23 Shapely Additive Explanation Beeswarm Plot - U.S. Airline Passenger Satisfaction (2015)

RESULTS

The results of the most important variables are provided by both interpretability by design and model inspection, like feature permutation and SHAP. To identify the four most important variables, a comparison between different models and inspections was made by computing each method's average of the respective order. In general, results from model coefficients are more accurate with regard to model inspection because they are directly used from the model. Model inspection is a valuable tool to confirm findings from interpretability by design, but it is preferred to not only rely on the results of feature permutation and SHAP if the model analysed is not a black box model; it is possible to fully rely on feature permutation only in the case of Kernel Support Vector Machine because interpretability by design is not possible in this case.

Interpretability by the design of logistic regression, classification tree, random forest and gradient boosting says that the four most important variables influencing customer satisfaction from the first to the fourth are Online Boarding, Inflight Wi-Fi Service, Type of Travel and Class. In the case of feature permutation, the four most important variables from the first to the fourth are Inflight Wi-Fi Service, Type of Travel, Online Boarding and Class. In comparison, for SHAP, the four most important variables from the first to the fourth are Online Boarding, Type of Travel, Inflight Wi-Fi and Class. Both interpretability by design and SHAP provide the same fifth variable, Inflight Entertainment. All three different methods of interpretation are aligned; indeed, it is possible to state that Online Boarding, Inflight Wi-Fi, Type of Travel and Class are the four most important variables without a specific order according to all the methods used. The best model to compromise interpretability and accuracy is a gradient boosting classifier which provides from the most important to the least important Online Boarding, Inflight Wi-Fi, Type of Travel and Class. The best two variables using the output of interpretability by design and model inspection are Online Boarding and Inflight Wi-Fi Service.

Academic Implications

The existing literature is exhaustive regarding customer satisfaction in general and for specific businesses (Anderson, Fornell, & Rust, 1997; Anderson, Fornell, & Lehmann, 1994; ACSI, 2021; Fornell, 1992; Hult, Sharma, Morgeson III, & Zhang, 2018; Kotler & Keller, 2012); however, there is lack of research about the most valuable attributes in each industry in order to improve customer satisfaction. This research aims at providing the most critical dimension of the influence of customer satisfaction in the air travel industry. Indeed, there is a lack of explanatory power for each of the most important variables found. Qualitative research should be applied to the main findings of this research to explain how and why variables of influence are crucial for a satisfied customer.

There is an evident gap between business travel and personal travel. The second type of customers are more likely to be dissatisfied even if the number of people that travel for personal reason is smaller compared to business travel. Future research should understand if the major reasoning behind this is provided by the fact that the company pays for a business trip, and a personal trip is paid for by the customer, increasing the likelihood of being dissatisfied.

Customer satisfaction shows significant changes regarding the travel class, where consumers in business class are more likely to be satisfied than those in eco class, considering that the price paid is relative to the service offered. Future research could provide information about the

misalignment of customer satisfaction in business class and eco class, assuming that the price paid should lower the expectation of an eco class consumer. The importance of understanding customers' behaviour in a different class is even more important considering the disruption of the airline business caused by COVID-19 and the Russian invasion of Ukraine. The CEO of Ryanair said: "There's no doubt that at the lower end of the marketplace, our really cheap promotional fares, the one euro fares, the 99 cent fares, even the 9.99 fares, I think you will not see those fares for the next number of years." (CNBC 2022)

Most individuals in the dataset, and consequently in the United States, travelled for business reasons. These customers are more likely to be satisfied. Future research could be performed from the airlines' side, trying to understand and explain if and why airlines are tailored for business trips. Moreover, a business trip may lead to an increase in the importance of inflight WiFi service providing reasoning for both variables type of travel and inflight WiFi service.

Although cleanliness is not considered an important variable in influencing customer satisfaction, due to COVID-19 major changes have taken place worldwide. Those changes may converge the greatest importance from all the most important variables found to the variable cleanliness. It is important for future research to understand if these changes have taken place and to monitor their impact on customer satisfaction.

Surprisingly, departure delay has no impact on customer satisfaction. Future research should understand why a travel-related variable does not influence customers. It could be inferred that departure delay can be justified by causes of force majeure, which are recognisable by customers.

Managerial Implication

Air travel is a highly competitive business where low-cost companies have influenced the market with low prices compromising customer satisfaction (Boetsch, Bieger, & Wittmer, 2011). For this reason, it is important to allocate resources with the greatest accuracy to maintain low prices and be competitive in the market, mainly due to the increased number of satisfied customers.

Airlines should not neglect the impact of digitalisation and its influence of it on a multi-channel strategy; indeed, online boarding and inflight Wi-Fi services are considered two of the most important variables in influencing customer satisfaction, from NielsenIQ research (12 October 2021) for the "Osservatorio multicanalità". It is evidenced that an increasing number of people use a digital approach to buying and consuming services. Managers should focus on having a satisfied customer even before the travel service takes place. Indeed, it is mandatory to provide a very simple

and user-friendly online boarding experience, compatible with all digital devices, to have half of the job done in having a satisfied customer.

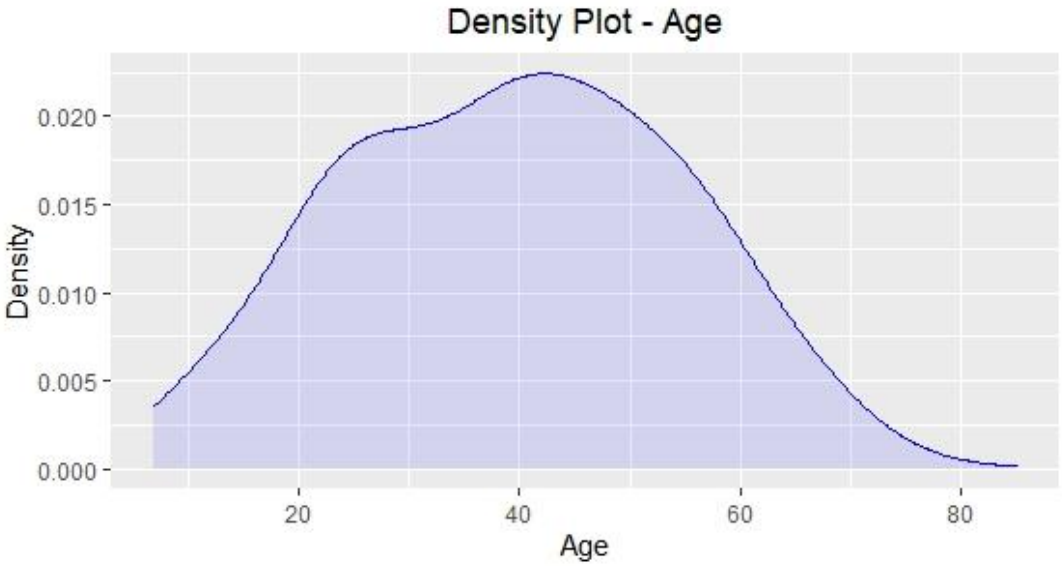
The type of travel is not a variable that airline companies can control. There is no evident explanation as to why customers who take a personal trip are more likely to be dissatisfied. However, knowing there is a strong influence on the type of travel, airline companies could try to understand if their business strategy mainly focuses on business travel rather than personal travel.

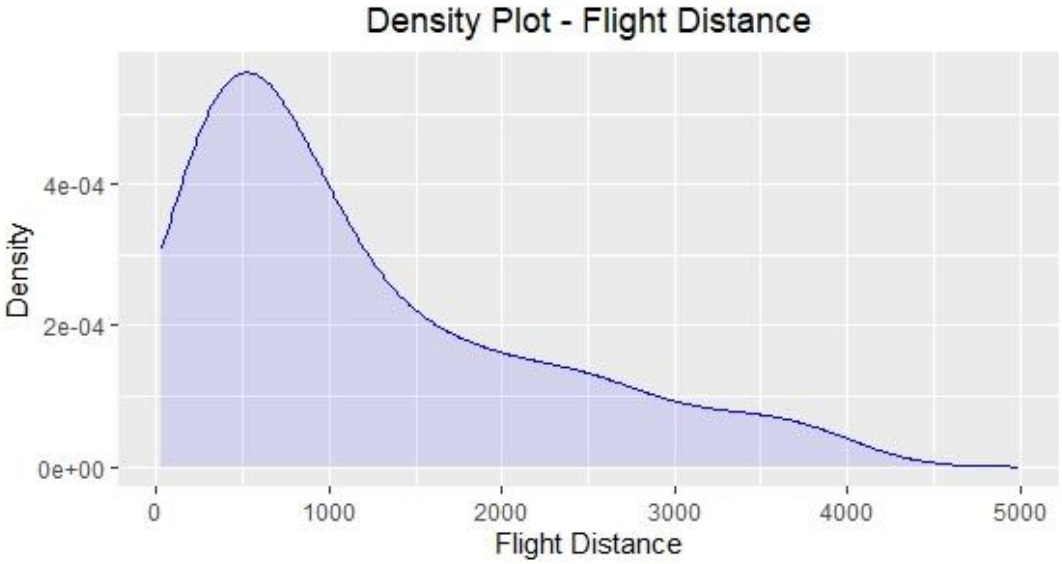
Different class shows different grades of satisfaction. The service in business class is better, and people are more likely to be satisfied. However, the price paid for eco classes should justify the service offered by the airline companies, which customers should understand. Nevertheless, this research shows that this is not the case. Customer satisfaction maximisation may reduce companies' productivity (Anderson, Fornell, & Rust, 1997). For this reason, it is not mandatory to maximise the service quality in eco classes which may lead the companies to become unproductive just to maximise customer satisfaction. A threshold should be found. The main advice is to maximise digital services satisfaction to see if online boarding can reduce the effect of dissatisfied customers from eco classes using the good influence of the first meeting point (online boarding).

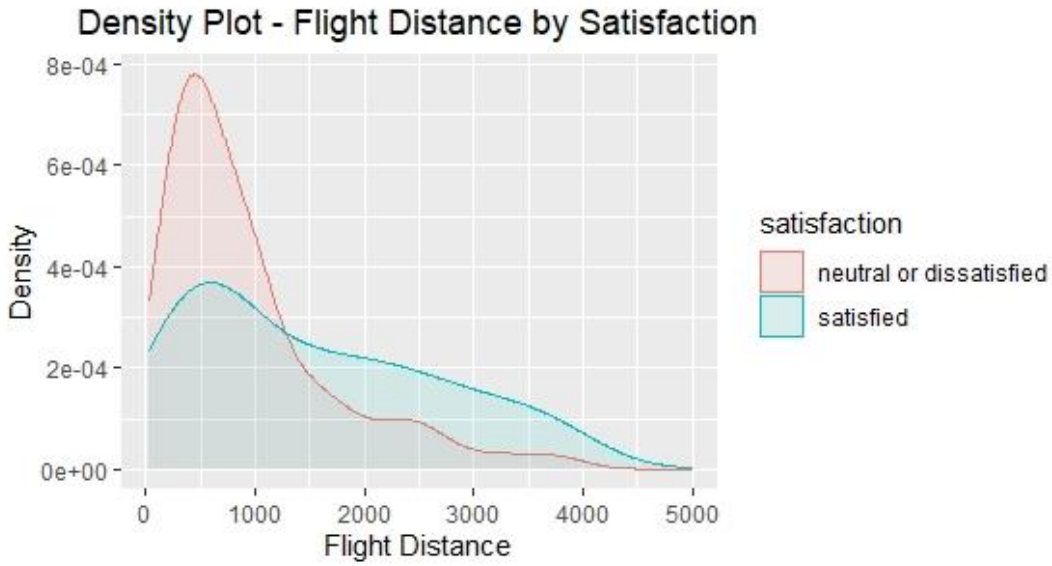
Companies able to leverage customer satisfaction to obtain optimal values on the most important variables could be more competitive in the market trying to fight the low pricing rule. Moreover, a highly digital airline company could use the company’s website to prevent loyal customers from using air travel aggregators, which leads customers to a price comparison mentality.

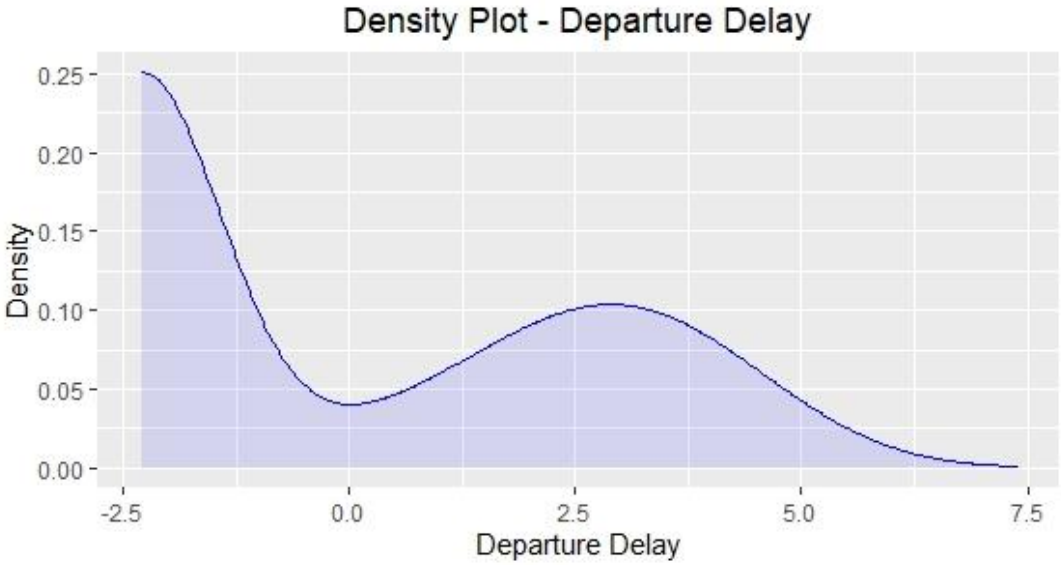
APPENDIX

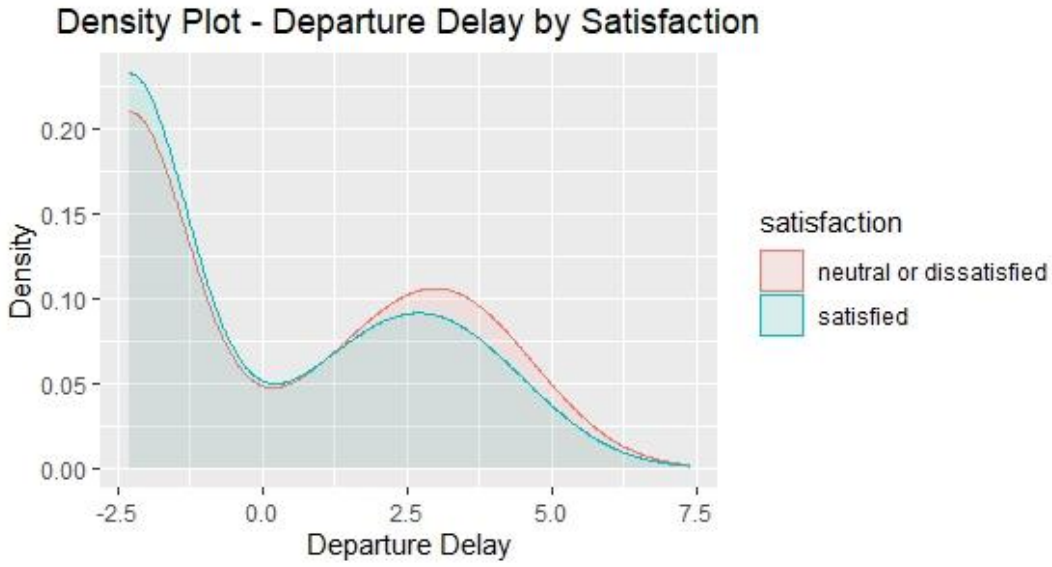
Appendix A. Kernel Density Plot

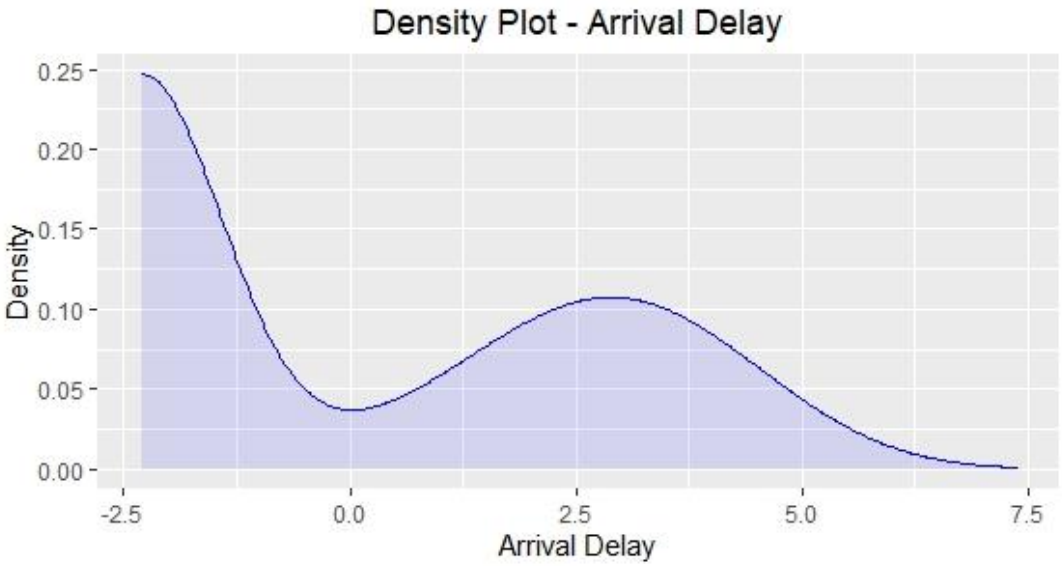


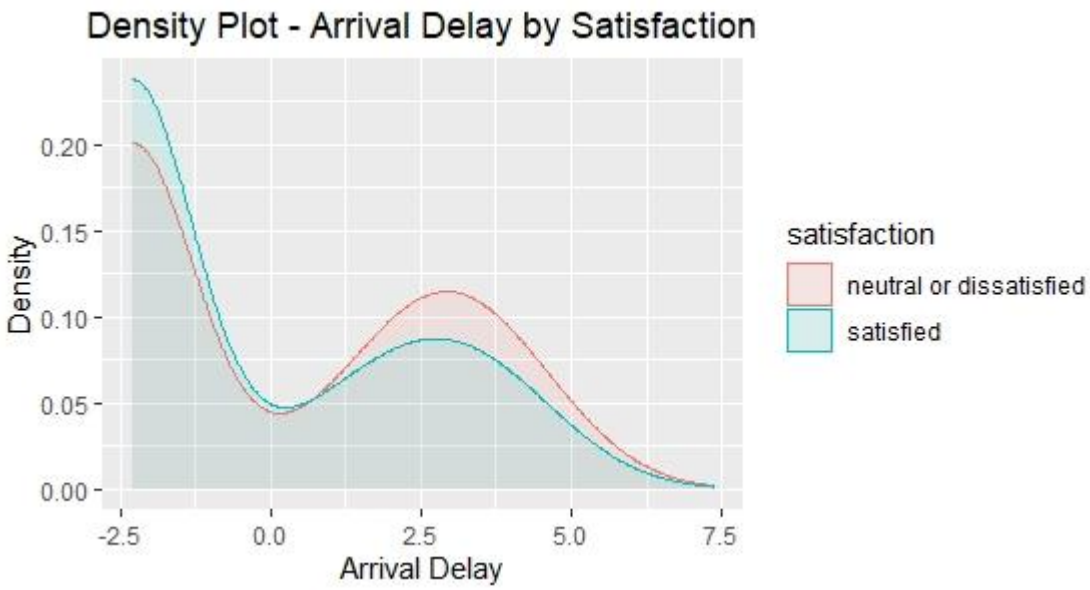



Density Plot - Flight Distance by Satisfaction
8e-04
6e-04
4e-04
2e-04
0e+00
Density
0
1000
2000
3000
4000
5000
Flight Distance
satisfaction
neutral or dissatisfied
satisfied

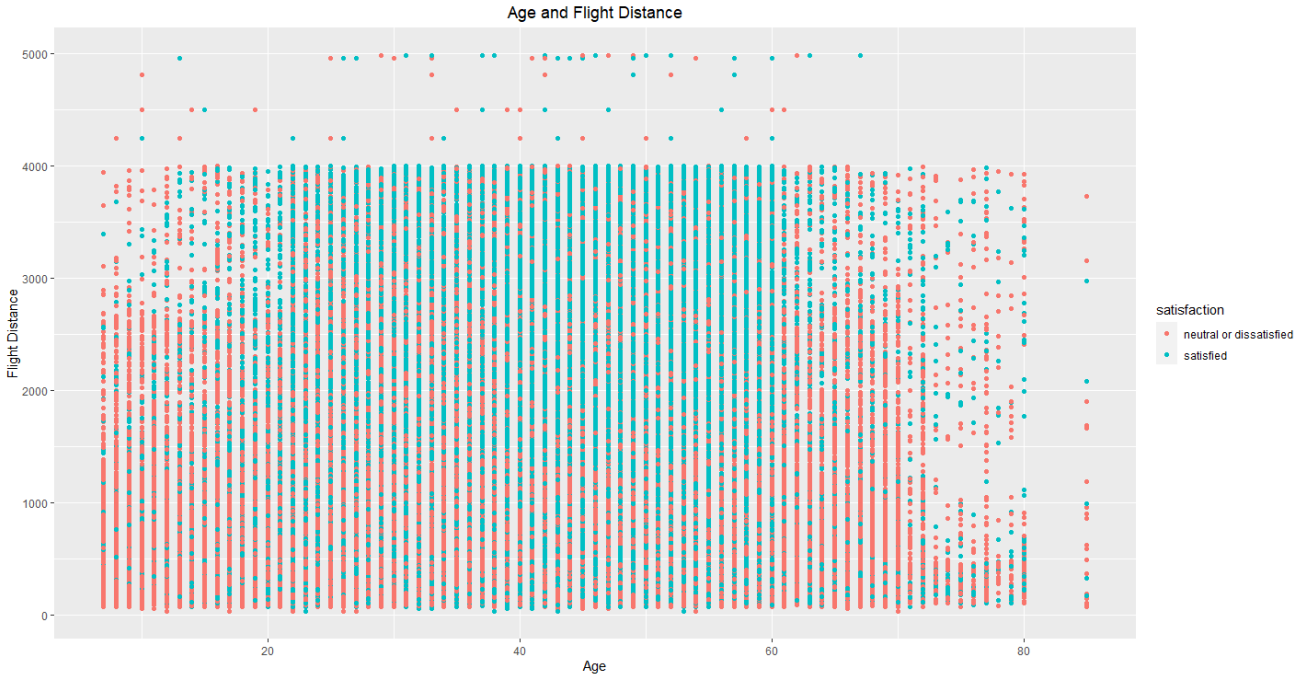

Density Plot - Departure Delay
0.25
0.20
0.15
0.10
0.05
0.00
Density
-2.5
0.0
2.5
5.0
7.5
Departure Delay

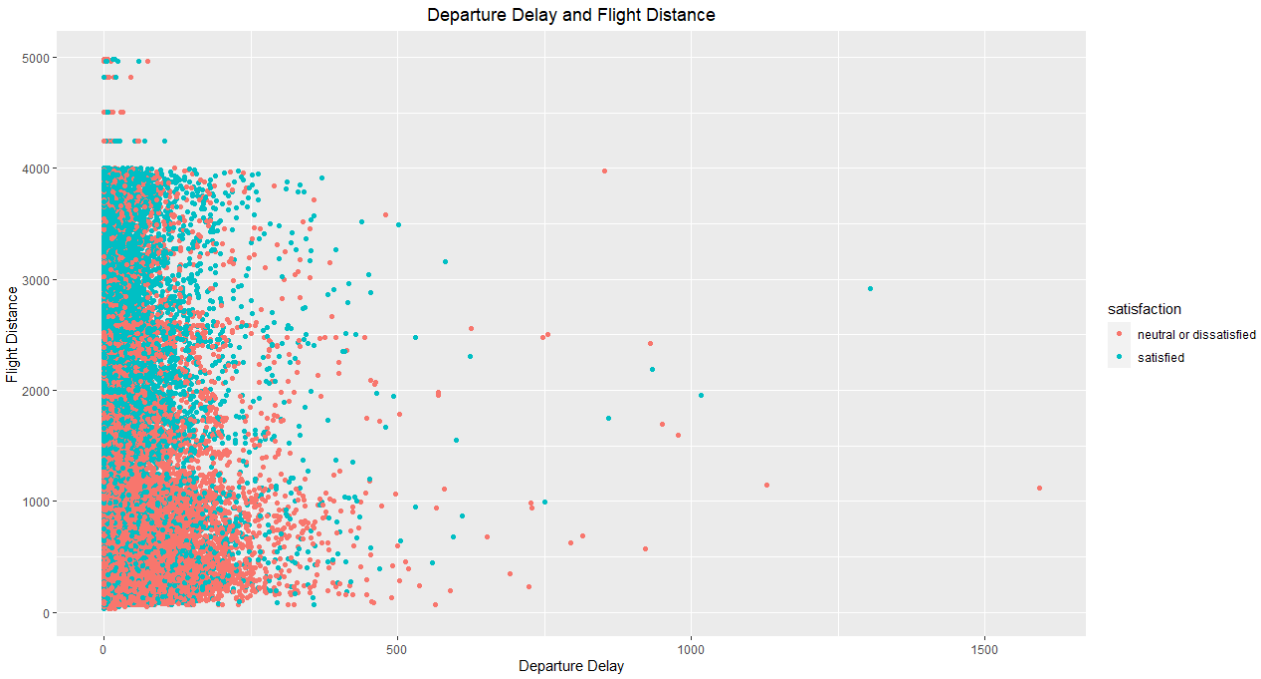

Density Plot - Departure Delay by Satisfaction
0.20
0.15
0.10
0.05
0.00
Density
-2.5
0.0
2.5
5.0
7.5
Departure Delay
satisfaction
neutral or dissatisfied
satisfied



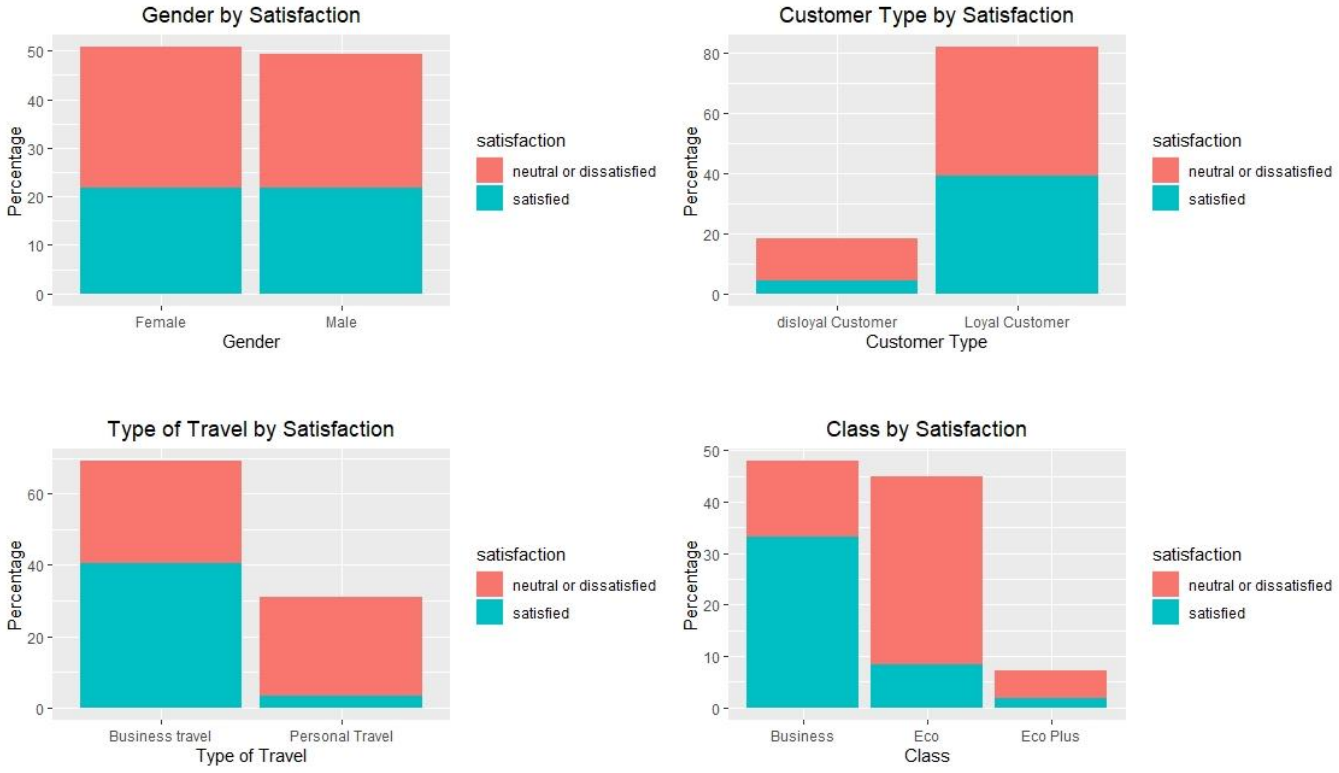


Appendix B. Scatter Plot

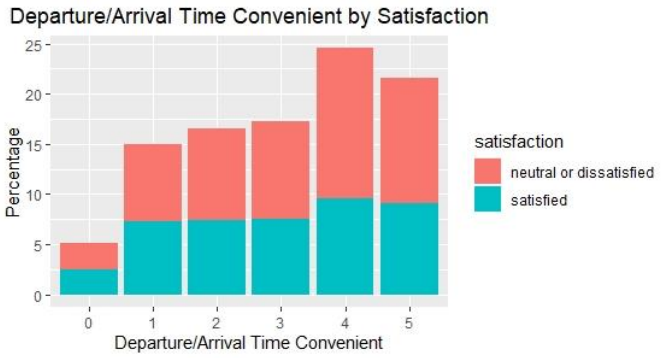




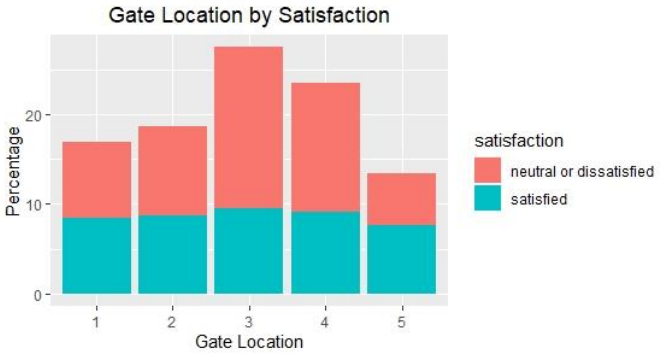
Appendix C. Bar Chart

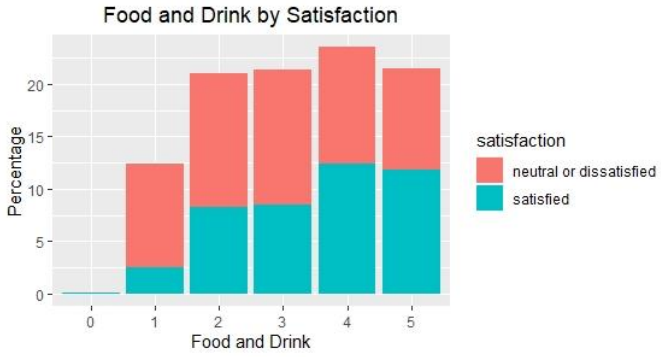



Inflight Wifi Service by Satisfaction
Percentage
Inflight Wifi Service
satisfaction
neutral or dissatisfied
satisfied


Departure/Arrival Time Convenient by Satisfaction
Percentage
Departure/Arrival Time Convenient
satisfaction
neutral or dissatisfied
satisfied

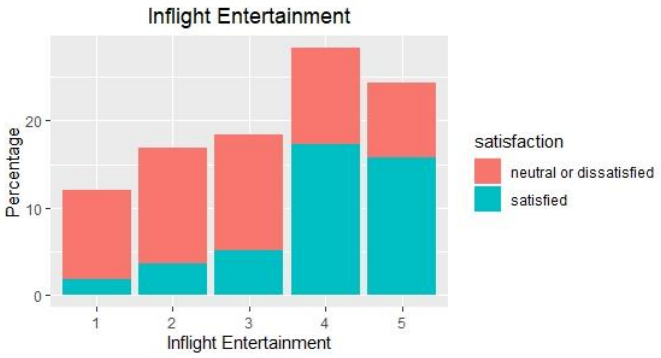

Ease of Online Booking by Satisfaction
Percentage
Ease of Online Booking
satisfaction
neutral or dissatisfied
satisfied

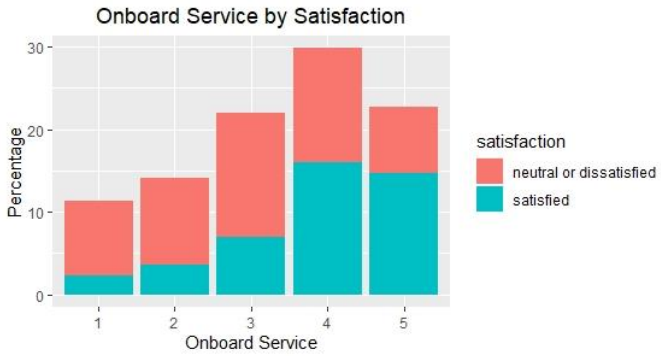

Gate Location by Satisfaction
Percentage
Gate Location
satisfaction
neutral or dissatisfied
satisfied

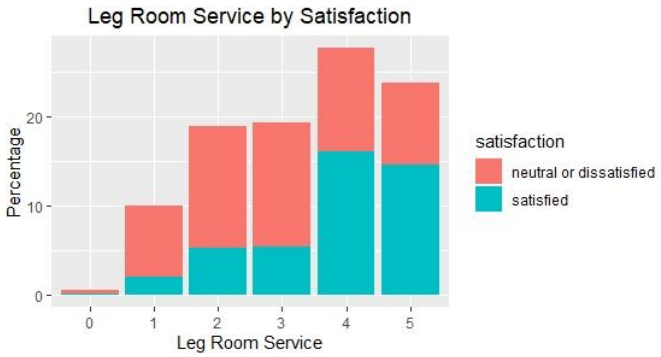

Food and Drink by Satisfaction
Percentage
Food and Drink
satisfaction
neutral or dissatisfied
satisfied

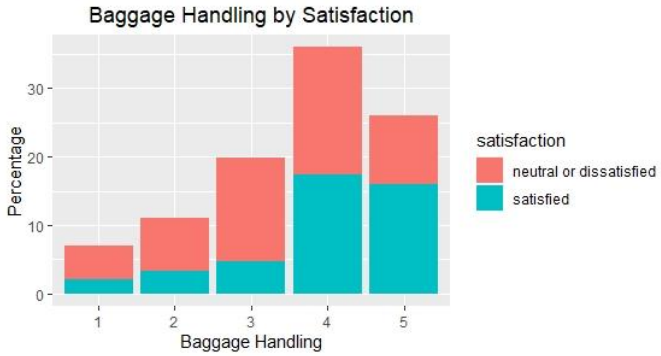

Online Boarding by Satisfaction
Percentage
Online Boarding
satisfaction
neutral or dissatisfied
satisfied

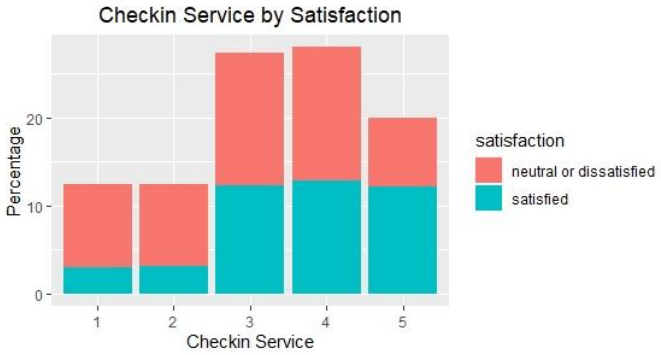

Seat Comfort by Satisfaction
Percentage
Seat Comfort
satisfaction
neutral or dissatisfied
satisfied

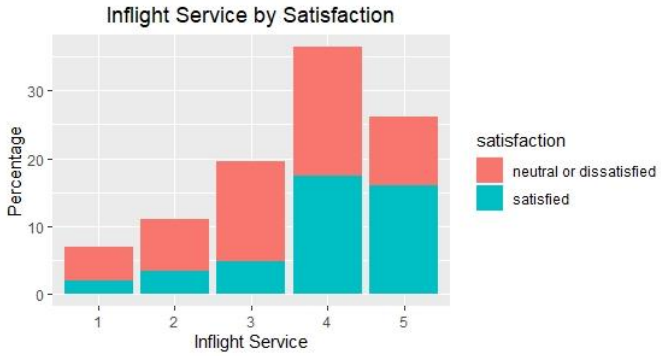

Inflight Entertainment
Percentage
Inflight Entertainment
satisfaction
neutral or dissatisfied
satisfied

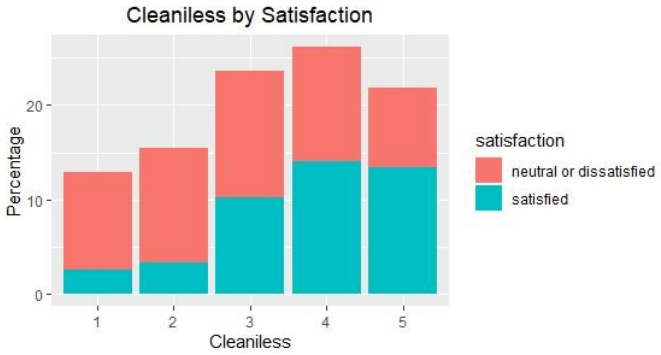

Onboard Service by Satisfaction
Percentage
Onboard Service
satisfaction
neutral or dissatisfied
satisfied

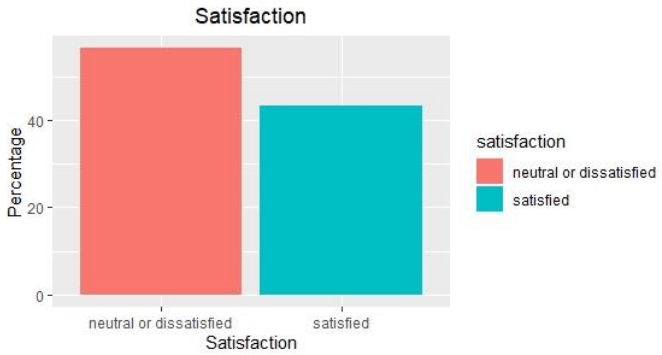

Leg Room Service by Satisfaction
Percentage
Leg Room Service
satisfaction
neutral or dissatisfied
satisfied


Baggage Handling by Satisfaction
Percentage
Baggage Handling
satisfaction
neutral or dissatisfied
satisfied


Checkin Service by Satisfaction
Percentage
Checkin Service
satisfaction
neutral or dissatisfied
satisfied

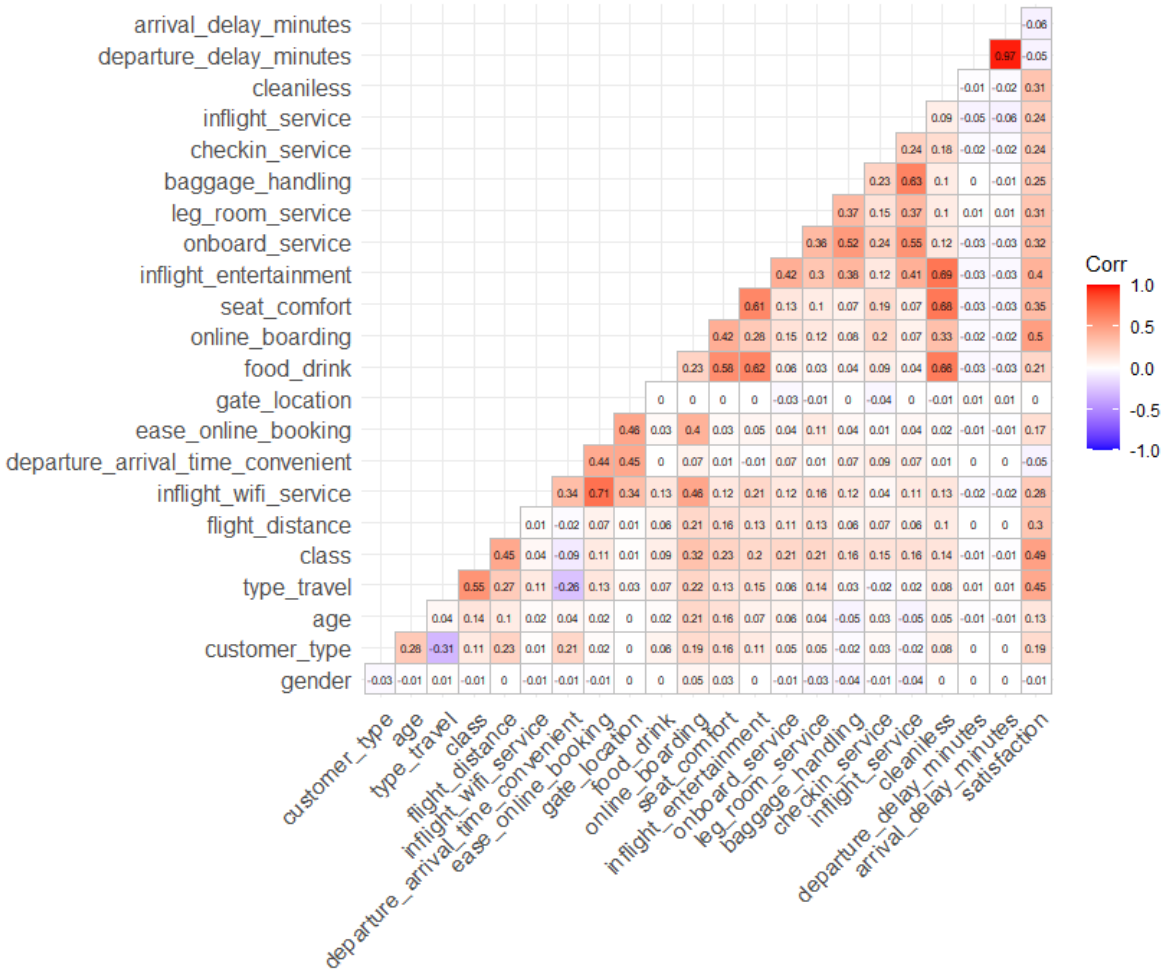

Inflight Service by Satisfaction
Percentage
Inflight Service
satisfaction
neutral or dissatisfied
satisfied


Cleaniless by Satisfaction
Percentage
Cleaniless
satisfaction
neutral or dissatisfied
satisfied

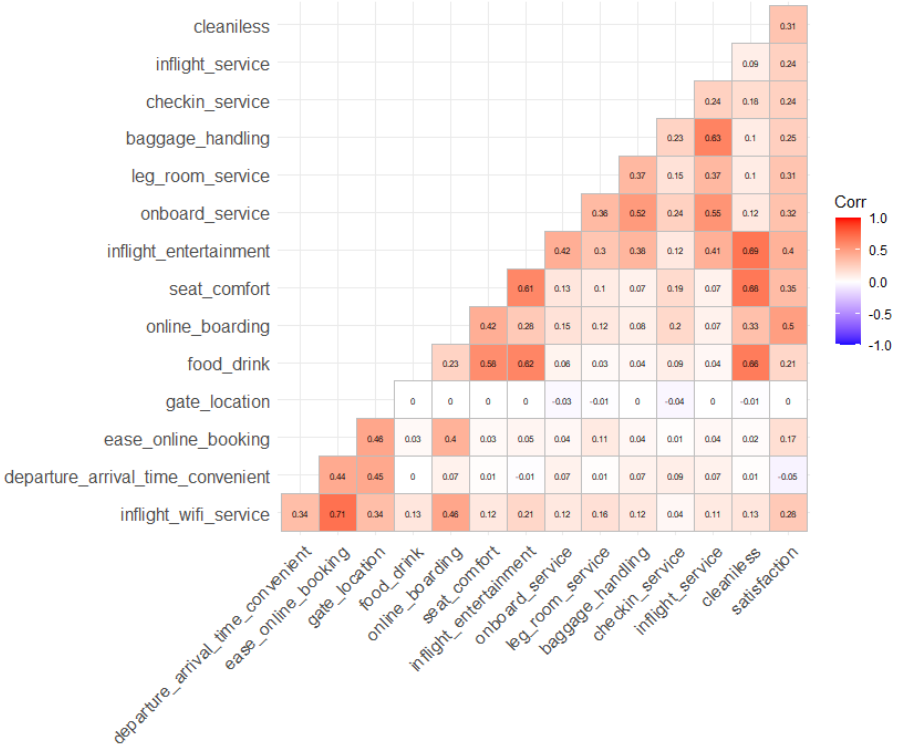

Satisfaction
Percentage
neutral or dissatisfied
satisfied
Satisfaction
satisfaction
neutral or dissatisfied
satisfied

Appendix D. Correlation Heatmap

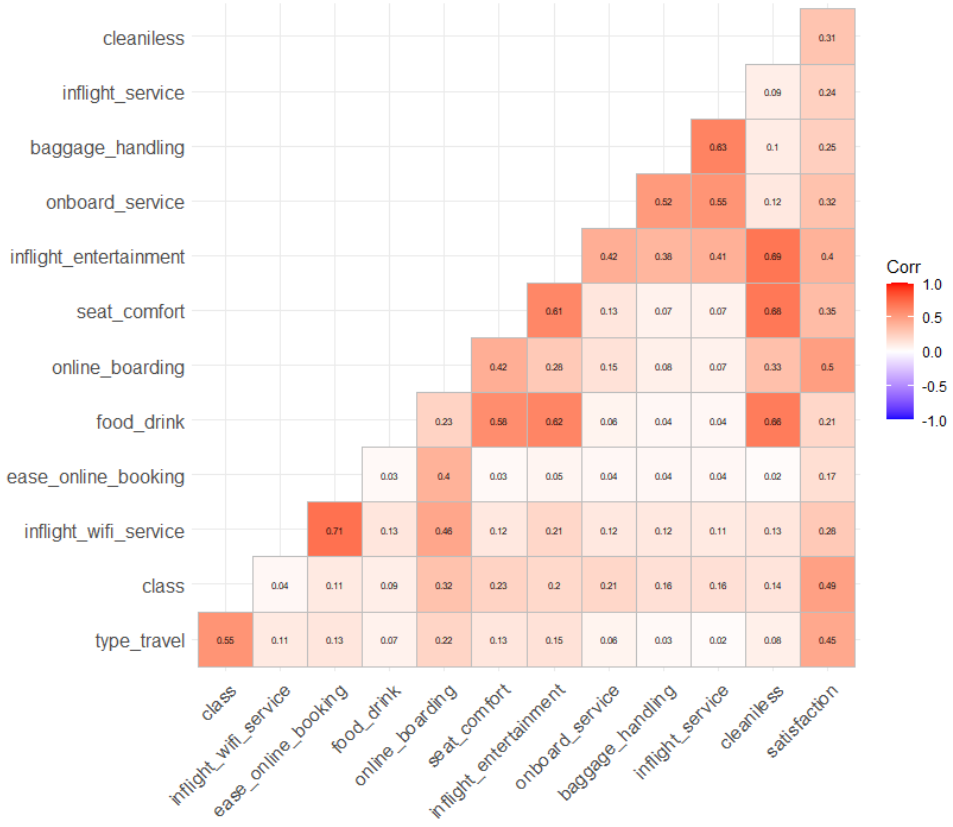
Correlation heatmap with all the variables



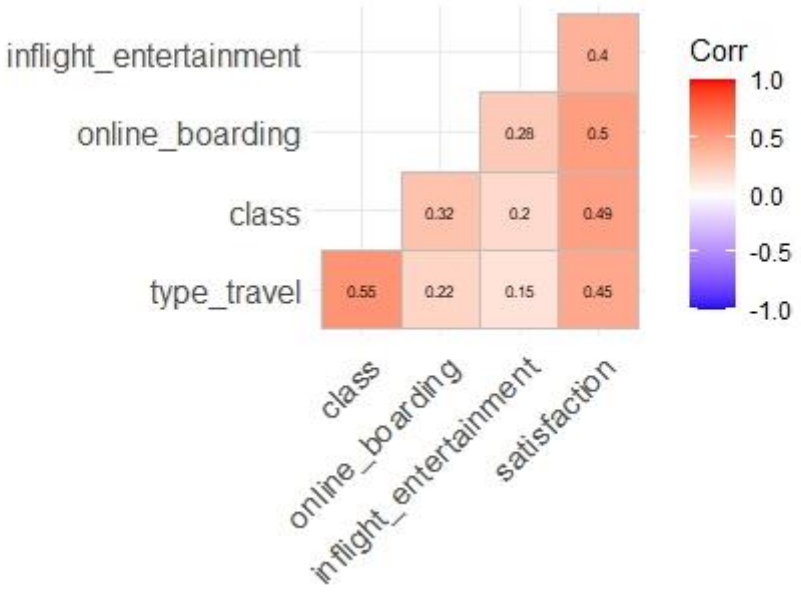
Correlation heatmap with only rating variables



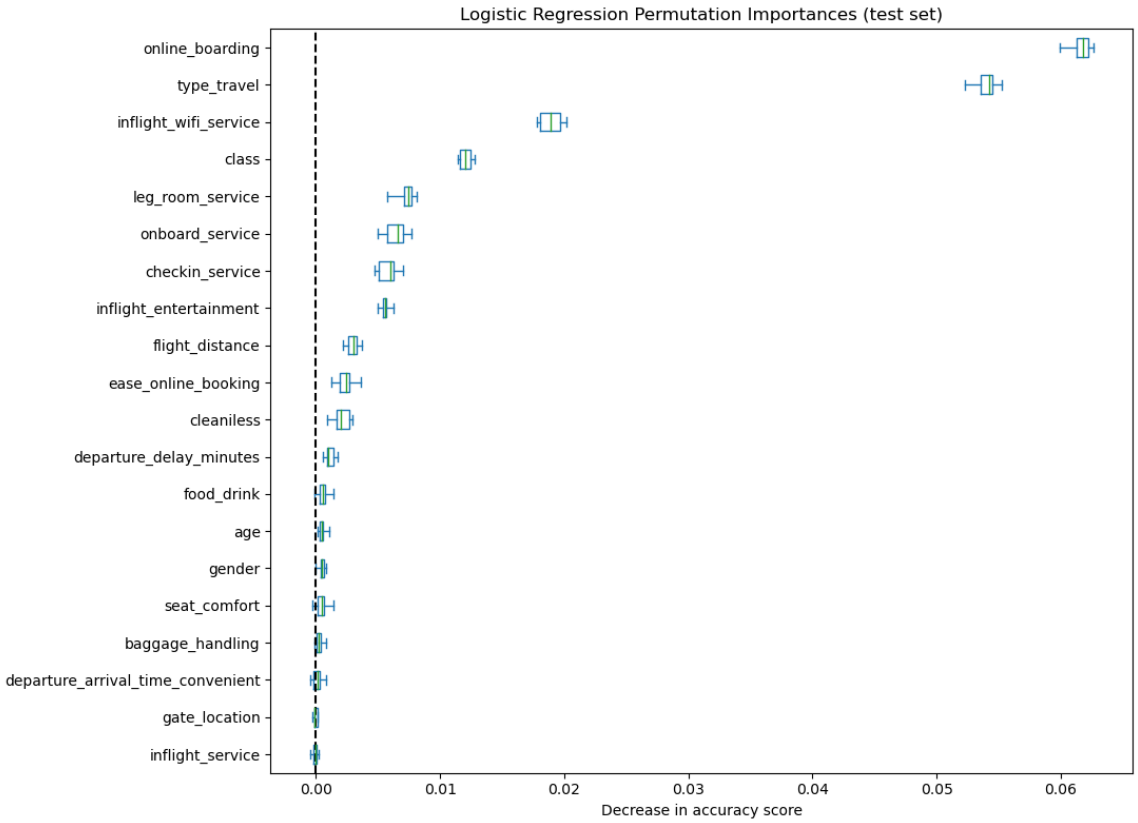
Correlation heatmap with the most correlated variables

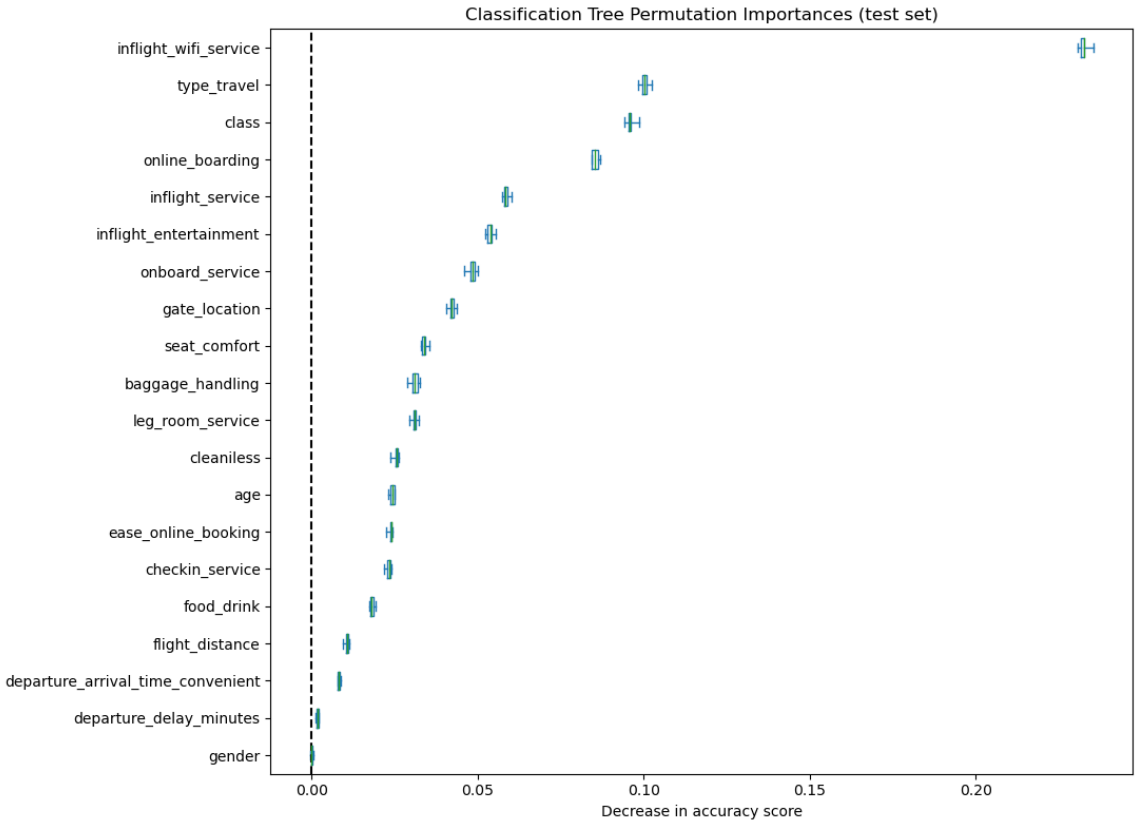


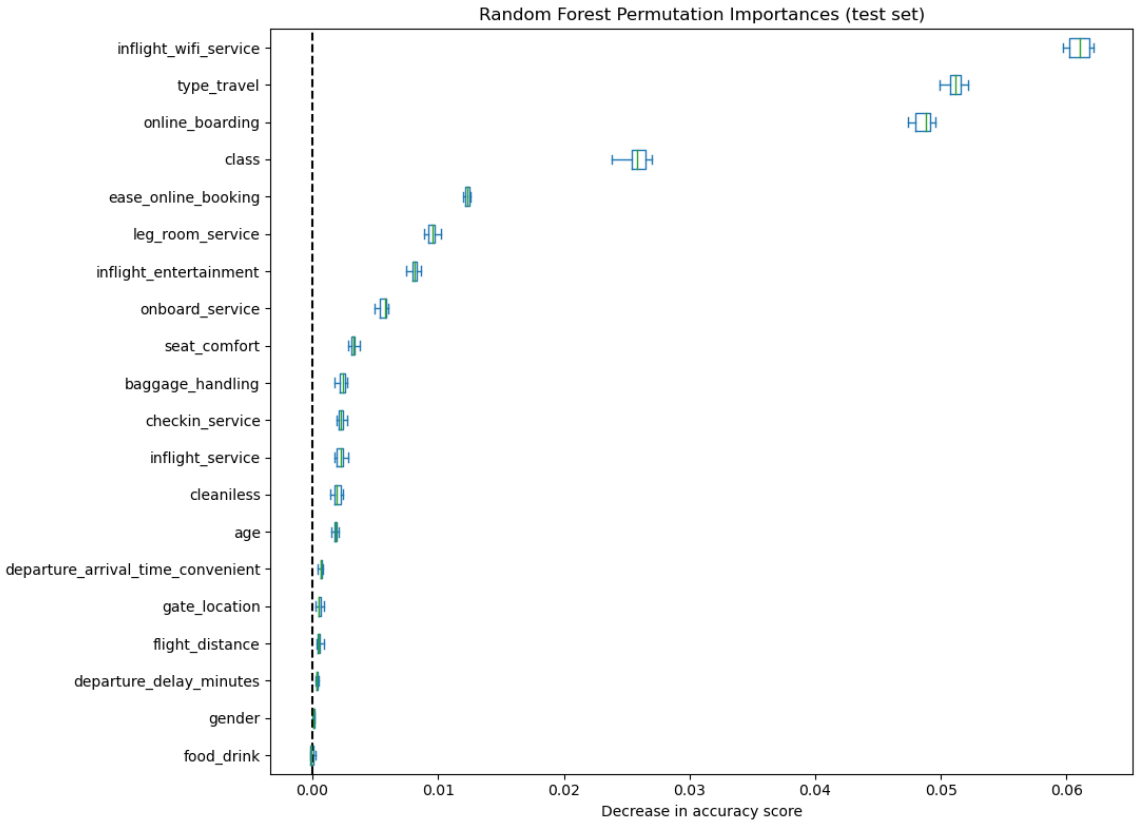
Correlation heatmap with the most correlated variables with customer satisfaction

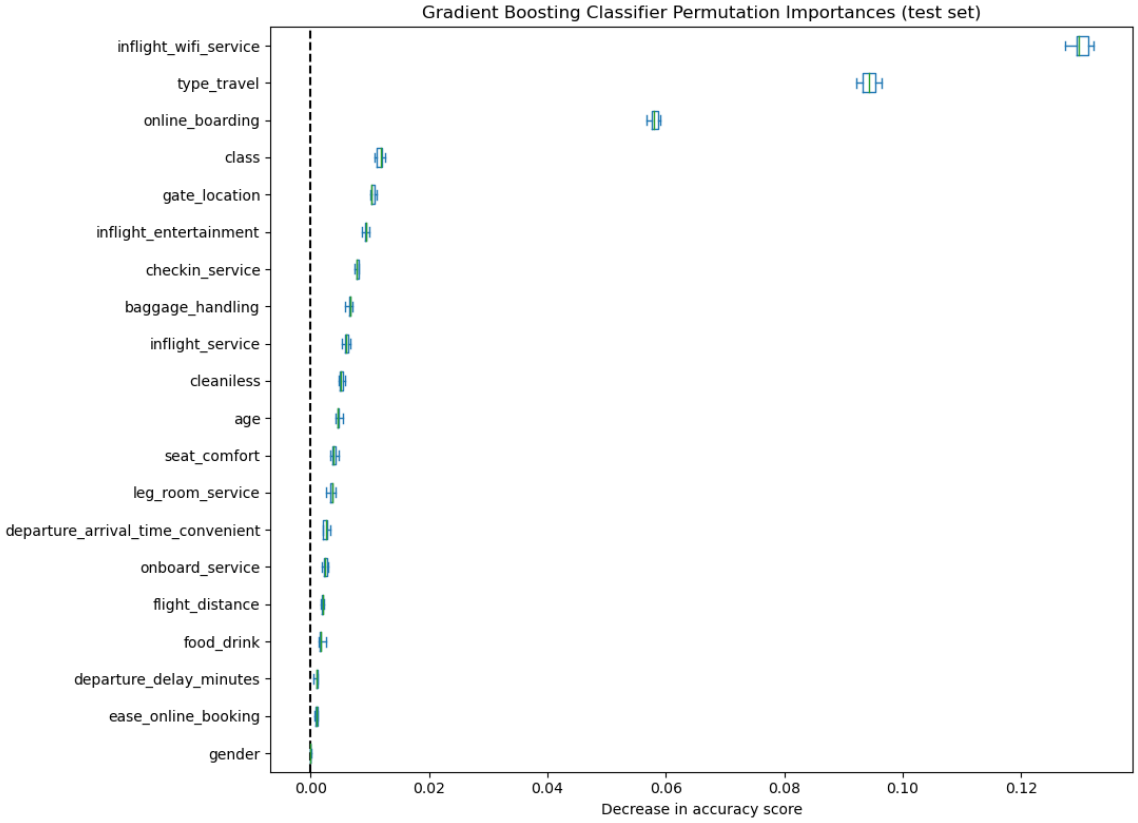


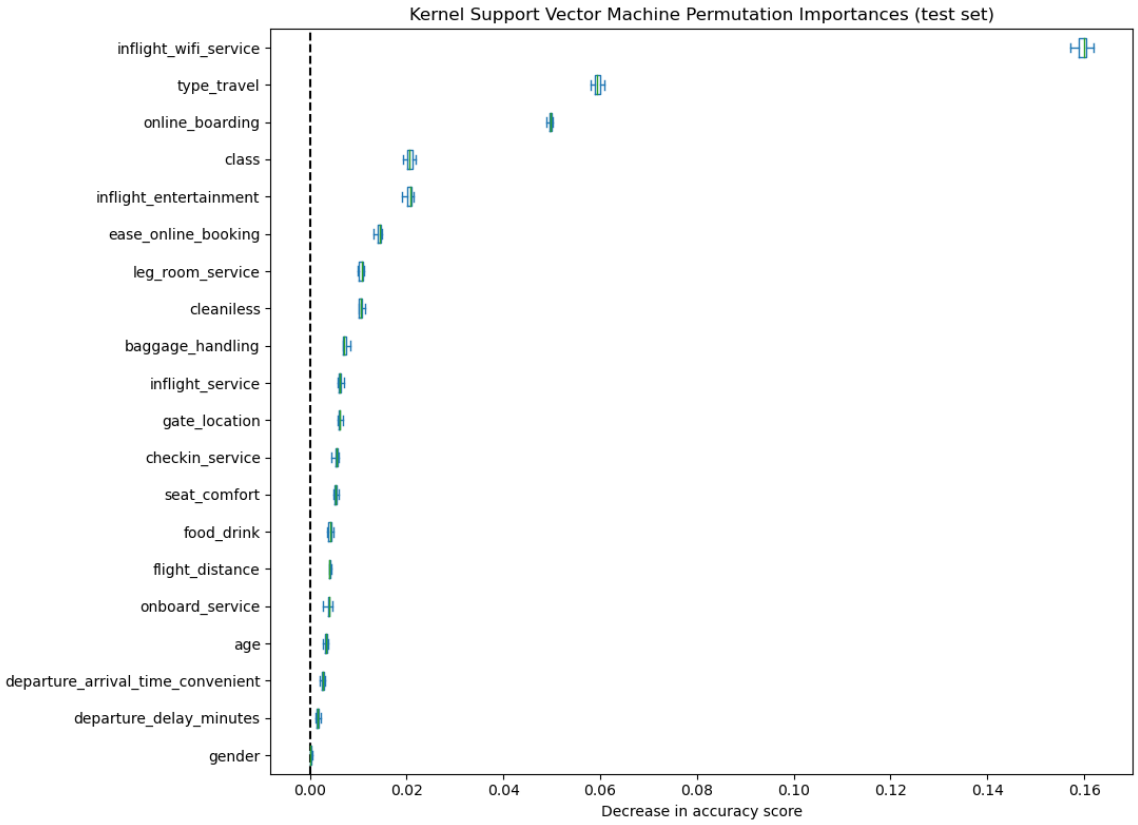
Appendix E. Feature Permutation Importance



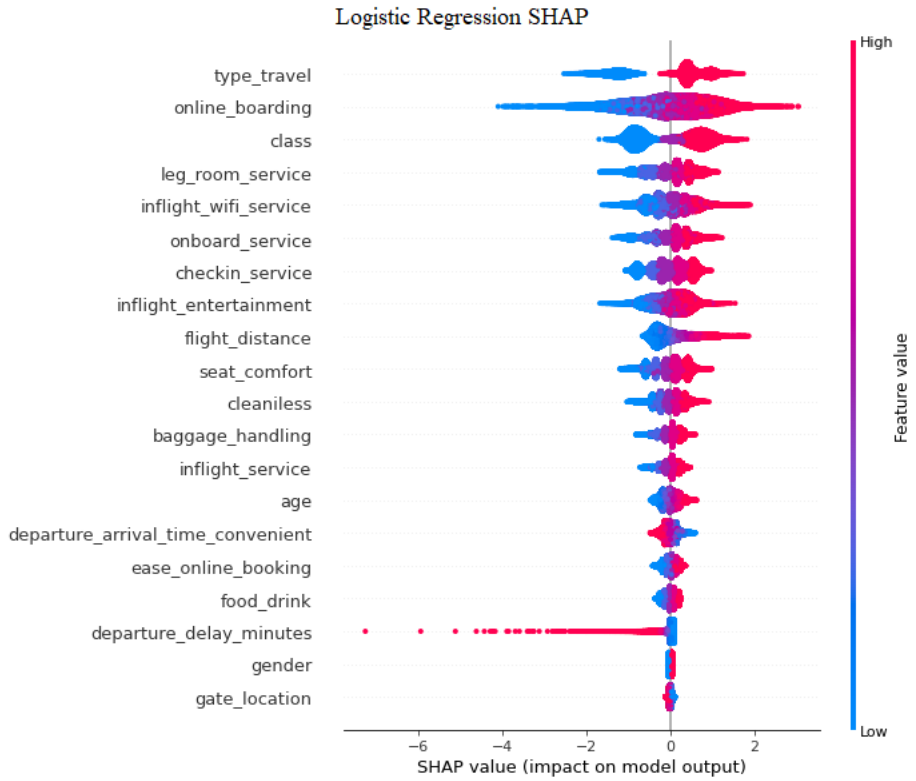


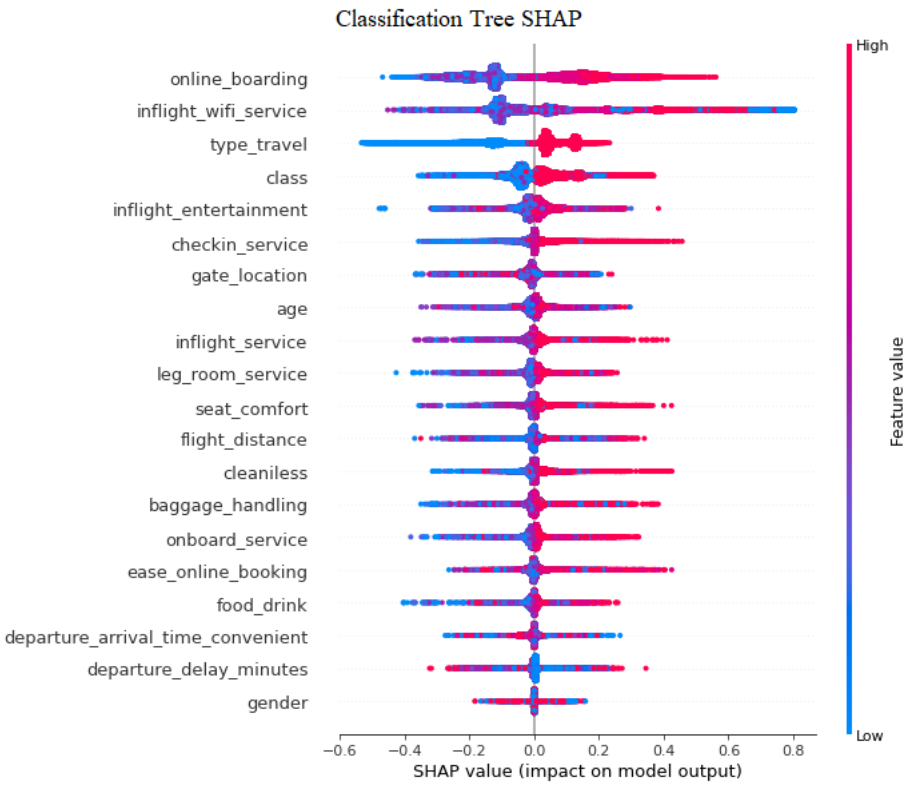

Random Forest Permutation Importances (test set)
inflight_wifi_service
type_travel
online_boarding
class
ease_online_booking
leg_room_service
inflight_entertainment
onboard_service
seat_comfort
baggage_handling
checkin_service
inflight_service
cleaniless
age
departure_arrival_time_convenient
gate_location
flight_distance
departure_delay_minutes
gender
food_drink
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Decrease in accuracy score

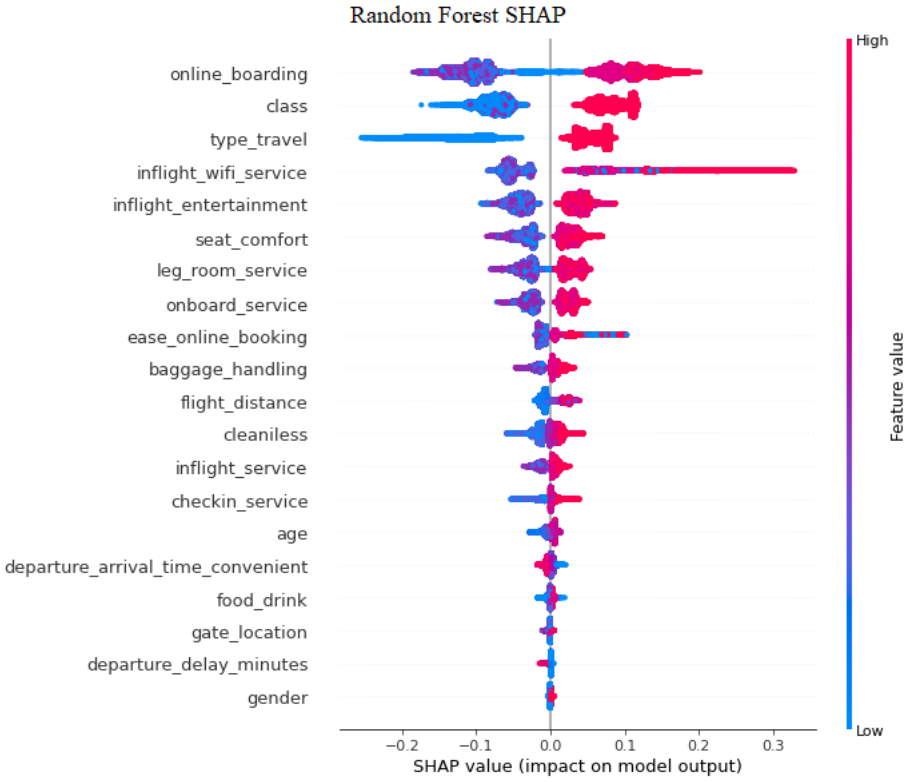

Gradient Boosting Classifier Permutation Importances (test set)
inflight_wifi_service
type_travel
online_boarding
class
gate_location
inflight_entertainment
checkin_service
baggage_handling
inflight_service
cleaniless
age
seat_comfort
leg_room_service
departure_arrival_time_convenient
onboard_service
flight_distance
food_drink
departure_delay_minutes
ease_online_booking
gender
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Decrease in accuracy score

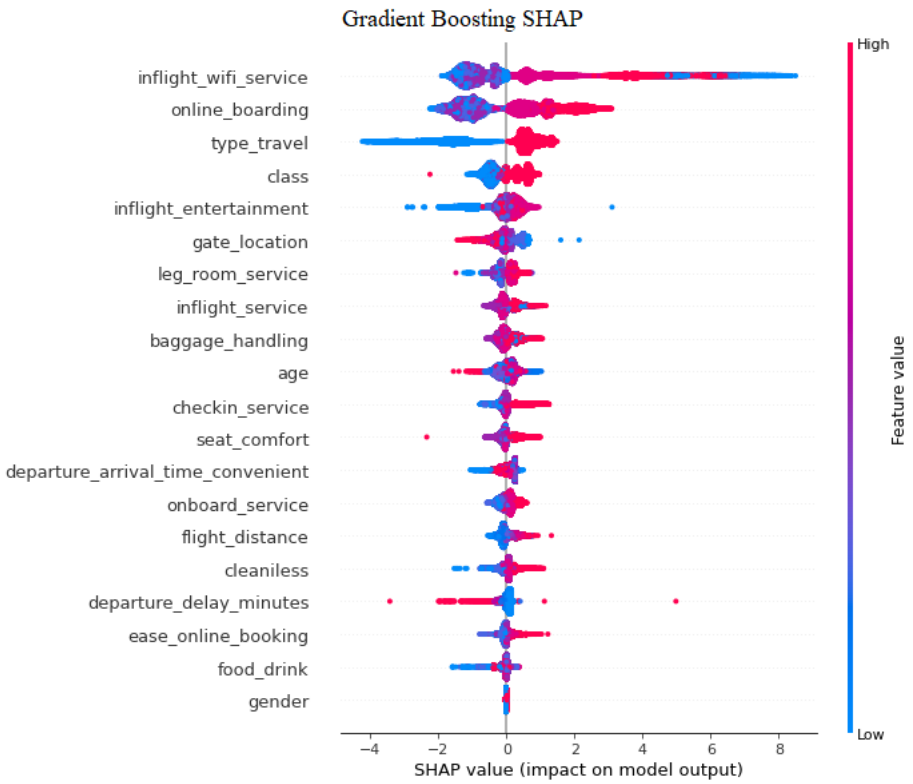


Appendix F. Shapely Additive Explanation Plot (SHAP)




Classification Tree SHAP
online_boarding
inflight_wifi_service
type_travel
class
inflight_entertainment
checkin_service
gate_location
age
inflight_service
leg_room_service
seat_comfort
flight_distance
cleaniless
baggage_handling
onboard_service
ease_online_booking
food_drink
departure_arrival_time_convenient
departure_delay_minutes
gender
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8
SHAP value (impact on model output)
High
Feature value
Low


Random Forest SHAP
online_boarding
class
type_travel
inflight_wifi_service
inflight_entertainment
seat_comfort
leg_room_service
onboard_service
ease_online_booking
baggage_handling
flight_distance
cleaniless
inflight_service
checkin_service
age
departure_arrival_time_convenient
food_drink
gate_location
departure_delay_minutes
gender
−0.2 −0.1 0.0 0.1 0.2 0.3
SHAP value (impact on model output)
High
Feature value
Low



References

ACSI. (2021, May 4). *American customer satisfaction index scores for airlines in the United States from 1995 to 2021*. Retrieved from Statista: https://www.statista.com/statistics/194941/customer-satisfaction-with-us-airlines-since-1995/

Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. *Jorunal of Marketing*, 53-66.

Anderson, E. W., Fornell, C., & Rust, R. t. (1997). Customer Satisfaction, Productivity, and Profitability:. *Marketing Science*, 129-145.

Bhattacherjee, A. (2012). Concepts, Constructs, and Variables. In A. Bhattacherjee, *Social Science Research: Principles, Methods, and Practices* (pp. 10-12). extbooks Collection. 3.

Boetsch, T., Bieger, T., & Wittmer, A. (2011). A Customer-Value Framework for Analyzing Airline Services. *Transportation Journal*, 251-270.

Breiman, L. (2001). Random Forests. (R. E. Schapire, Ed.) *Machine Learning*, 5-32.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 273.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society*, 215-242.

Fornell, C. (1992). A National Customer Satisfaction Barometer: The Swedish Experience. *Journal of Marketing*, 6-21.

Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Everitt Bryant, B. (1996, October). The American Customer Satisfaction Index: Nature, Purpose, and Findings. *Journal of Marketing*, 7-18.

Grégoire, Y., & Fisher, R. (2008). Customer betrayal and retaliation: when your best customers become your worst enemies. *Journal of Academy of Marketing Science*, 247-261.

Hannah, W.-G. (2022, August 11). *No more $10 flights: Budget airline Ryanair says ticket prices will have to rise*. Retrieved from CNBC LLC: https://www.cnbc.com/2022/08/11/no-more-10-flights-budget-ryanair-says-fares-will-have-to-rise.html

Harris, C., Millman, j., Van der Walt, S., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. (2020). Array programming with NumPy. *Nature*, 357-362. Retrieved from https://doi.org/10.1038/s41586-020-2649-2

Hult, G. M., Sharma, N. P., Morgeson III, F. V., & Zhang, Y. (2018). Antecedents and Consequences of Customer Satisfaction: Do They DifferAcross Online and Offline Purchases? *Journal of Retailing*, 10-23.

IATA. (2021, 10 5). *Worldwide revenue with passengers in air traffic from 2005 to 2022 (in billion U.S. dollars) [Graph]*. Retrieved from Statista: https://www.statista.com/statistics/263042/worldwide-revenue-with-passengers-in-air-traffic/

Jhon, D. (2018). *Passenger Satisfaction - US Airline Passenger Satisfaction*. Retrieved from Kaggle: https://www.kaggle.com/datasets/johnddddd/customer-satisfaction

John, D. (1999). The Relationship between Subjective and Objective Company Performance Measures in Market Orientation Research: Further Empirical Evidence. *Marketing Bulletin*, 65-75.

Kelleher, S. R. (2022, July 19). *Why 20,000 Delayed Flights A Day Are Not Going Away Anytime Soon*. Retrieved from Forbes.com: https://www.forbes.com/sites/suzannerowankelleher/2022/07/19/why-delayed-flights-not-going-away-soon/?sh=4d7cc6425727

Kotler, P., & Keller, K. L. (2012). Total Customer Satisfaction. In P. Kotler, & K. L. Keller, *Marketing Management* (p. 128). Upper Saddle River: Prentice Hall.

Lundberg, S. M., & Suu-In, L. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, *Advances in Neural Information Processing Systems 30* (pp. 4765-4774). Long Beach: Curran Associates, Inc.

McKinney, W. (2010). Data structures for statistical computing in python. *9th Python in Science Conference* (pp. 56-61). mckinney proc scipy.

Namkung, Y., Jang, S., & Choi, S. K. (2011). Customer complaints in restaurants: Do they differ by service stages and loyalty levels? *International Journal of Hospitality Management*, 495-502.

Nielsen Consumer LLC. (2021, October 12). *La multicanalità è ormai un fenomeno di massa. Ma come quantificarne la domanda potenziale?* Retrieved from NielsenIQ: https://nielseniq.com/global/it/insights/analysis/2021/la-multicanalita-e-ormai-un-fenomeno-di-massa-ma-come-quantificarne-la-domanda-potenziale/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

R Core Team. (2021). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Retrieved from https://www.R-project.org/

Solomon, M. R. (2018). The Motivation Process: Why Ask Why? In M. R. Solomon, *Consumer Behavior - Buying, Having and Being* (pp. 172-173). Harlow: Pearson Education Limited.

Soman, D. (2015). *The Last Mile: Creating Social and Economic Value from Behavioral Insights.* Toronto: University of Toronto Press.

Statista. (2017). *Passenger experience at U.S. airports.* Statista. Retrieved from https://www.statista.com/statistics/701325/us-air-travel-frequency-of-flying-short-haul/

Statista. (2022, June 01). *Air Transport - United States*. Retrieved from Statista: https://www.statista.com/outlook/io/transportation-storage/air-transport/united-states

Statista. (2022, April 20). *Travel frequency for business purposes in the U.S. in 2022 [Graph]*. Retrieved from Statista: https://www.statista.com/forecasts/997126/travel-frequency-for-business-purposes-in-the-us

Statista Survey. (2017, May 2). *How often do you travel by air for business trips?[Graph]*. Retrieved from Statista: https://www.statista.com/statistics/701067/us-air-travel-frequency-of-flying-business-trips/

ValuePenguin. (2021, September 30). *Which airlines are you most loyal to?* Retrieved from Statista: https://www.statista.com/statistics/1284295/airlines-loyalty-united-states/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org

SUMMARY

The air travel business faced a violent shutdown due to COVID-19; the worldwide revenue from passenger air traffic in 2019 was $ 581 billion, dropping to $ 189 billion in 2020. In 2020, the post-COVID loss of airline companies was $ -35 billion in North America and $ -34.5 in Europe. In 2021, the air travel business started to recover slowly, where North American airline losses equalled $ -5.5 billion, and in Europe, airlines losses equalled $ -20.9 billion. Now, the worldwide revenue forecast is $ 378 billion - still not at the levels of 2019, but it is slowly increasing (IATA, 2021).

Certainly, the air travel market is in the eye of the storm. Customer satisfaction could not be viewed as the first major problem, however after every crisis, when the market starts to grow again, and passengers have started to regain confidence in taking airplanes, there is an ample opportunity to re-establish a good relationship with consumers starting with a high level of customer satisfaction.

To better understand why machine learning can be helpful for the purpose of customer satisfaction prediction and identification of its most essential variables, it is crucial to have a general definition of Customer Satisfaction and how it is influenced in general. The most cited and accepted definition from the extant literature of satisfaction is the following: *"satisfaction is a person's feeling of pleasure or disappointment that results from comparing a product's perceived performance to expectations"* (Kotler & Keller, 2012).

Perceived quality, perceived value and expectations are key drivers of customer satisfaction (Fornell, Johnson, Anderson, Cha, & Everitt Bryant, 1996), highlighting how this KPI can be different among customers. Indeed, the same level of satisfaction can have different influences and

be rated with different magnitudes. Although prior research shows the primary drivers of customer satisfaction, a more specific approach is needed. Knowing only that the product having a perceived quality influences customer satisfaction does not explain which attributes in a specific industry are the most important to improve satisfaction.

The more satisfied a customer is, the higher the probability they will become a loyal one. The Statista 2021 report about Airline passenger experiences shows that the companies with the highest rating on the American Customer Satisfaction Index (Southwest Airlines ACSI: 79 and Delta Air Lines ACSI: 79) are the companies with the largest number of loyal customers. In fact, those customers account for almost 50% of the total market (Southwest Airlines, 26% and Delta Air Lines, 22% of the total market) (ACSI, 2021; ValuePenguin, 2021). Loyalty is one of the most important benefits a firm can have, and leverage it as an asset for profitability due to recurrent purchase, price elasticity reduction, failure cost reduction, higher retention rate and positive word of mouth (Anderson, Fornell, & Lehmann, 1994; Anderson, Fornell, & Rust, 1997; Fornell, 1992). Today's satisfied or dissatisfied customers will be influenced by future decisions. For this reason, satisfaction, like loyalty, has to be viewed from a long-term perspective (Anderson, Fornell, & Lehmann, 1994). The long-term perspective of satisfaction and loyalty has to be viewed the other way around when a customer is dissatisfied. It is very difficult to turn negative opinions into positive ones when the company's reputation is damaged. Like loyalty is reached after an extended period, conversion of a dissatisfied customer can never occur or may require greater effort than converting a neutral customer; it is fundamental to understand what customers want and like about a specific product or service.

As stated in the American Customer Satisfaction Index (ACSI) from the research of Fornell, Johnson, Anderson, Cha, & Everitt Bryant (1996), satisfaction tends to be lower in the service industry compared with goods. For this reason, it is important to have a different perspective of customer satisfaction in the case of the air travel business rather than making assumptions based on general metrics. Consumers are exposed longer to multiple stimuli in the service industry and even more in the Air Travel industry. In fact, it is sufficient to think about the whole process of taking an airplane from the airport of departure to the airport of arrival. There are many different steps in which the airline is involved. As if that were not enough, with digitalisation and the increase in online ticketing and check-in, the contact between company and consumer takes place even before the day of departure. Due to the high number of stimuli, inconveniences can occur at every level and time, leading to a high probability of dissatisfied customers or a low Customer Satisfaction Score translated into negative word of mouth and low customer loyalty. Companies

seek to improve service or product qualities to leverage the positive relationship between customer satisfaction and quality of service. However, focusing on maximising quality does not necessarily lead to an increase in customer satisfaction. It is not possible to maximise all existing dimensions, and managers have to decide where to allocate resources; this concept is even more important in the case of airlines since today's competitive market is based on a low pricing strategy which drastically reduces the service quality and harms customer satisfaction (Boetsch, Bieger, & Wittmer, 2011). For this reason, the error rate is minimised, which also leads airlines to make low-risk decisions focusing more on productivity (overbooking strategy) rather than being customer-centric. Of course, this is not the case for all airline companies. In general, customisable products or services lead to higher customer satisfaction and a lower level of productivity (Anderson, Fornell, & Rust, 1997). This is not the case in the air travel business, where the service tends to be standardised to maintain low cost and offer a competitive price. This practice leads to a very similar service offered among airlines, and the number of dimensions in which it is possible to be different is reduced to a minimum level. For this reason, it is fundamental for airlines to understand these dimensions and where to allocate resources to avoid a waste of money and increase the return on investment.

The process of searching for the most influential variable to which lead to a decision, in this case, whether a customer is satisfied or not, is almost the same process that human beings go through for any decision or judgement. A human choice process can be represented as the weighted additive decision rule. Mathematically speaking, humans take some weighted attributes in order to obtain a final score of a specific option. The key concept is to find those weights of a specific decision to understand which attributes are facilitators of the final choice.

Instead of waiting an entire life to learn attributes and weights, processing power can be used to train an algorithm to make a specific decision, in this case, whether a customer is satisfied or not. Learning relationships between attributes and the output variable allows the statistical model to adjust weights to maximise the accuracy of the prediction or reduce the error. In this specific case, there is no interest in predicting customer satisfaction because to do so, one would need some scoring from customers of all the available attributes, which would mean that one could directly ask whether customers are satisfied or not; for this reason, the objective of this research is to pick some classification models and inspect them to understand why and how the decision of the model was made. This process answers the question: how much can I rely on a specific variable to understand if customers are satisfied?

There are many different models to choose from. In general, the final scope of machine learning is to find the best one to predict the output variable; however, there are trade-offs in deciding which is the best model to use, and the final decision should always be aligned with the problem that has to be solved. This research aims to have a model to interpret and predict the output variable with an acceptable error. If not, the model would be useless because if it cannot predict the output variable, the interpretability of the model would not be meaningful since understanding how a bad decision is made does not help see which variable is the most important. There are models interpretable by design like Logistic Regression and Classification Tree, less but still interpretable models like Gradient Boosting and Random Forest, and non-interpretable models like Kernel – SVM due to the Radial Basis Function used. This research uses all those models to compare them across accuracy and variable importance. Two interpretability methods like Shapely Additive explanation (SHAP) (Lundberg & Suu-In, 2017) and Feature Permutation Importance (Pedregosa, et al., 2011) are also used for the purpose of local and global interpretability, respectively.

The U.S. Airline Passenger Satisfaction Dataset (The Dataset) describes customer-related satisfactory experience levels for all U.S. airlines in 2015. The Dataset is composed of 129,880 observations and 24 variables.

Before starting to interpret the data, it is mandatory to understand the type of variables dealt with, how information was collected, and what can influence each variable of interest. Customers ratings in statistical terms are interval variables where the magnitudes are always equal from score to score; however, the most important digression about ratings is that they are based on customers' perceptions which are not defined by a specific rule - the numerical magnitude is the same, but the real magnitude cannot be defined. Perception is different among customers; this brings a higher degree of uncertainty to the analysis because the method for measuring the dependent variable and the scores is not standardised.

From a marketing perspective, ratings are very complex variables because they are considered perceptive instead of objective. For a different customer, we can have different perceptions with the same rating or the same perception with different ratings. This difference is driven by the fact that each person has a different background and experiences, influencing the ratings assigned for each variable. For example, a person that used to travel very frequently could be considered an expert, and can compare a higher amount of information related to air travel. Ratings of expert consumers may be more accurate and specific and usually lower compared with ratings assigned by people who do not travel by air very frequently. Moreover, people differ in how they evaluate

things; many can be restrictive and others more expansive. Some people do not provide the maximum score very frequently because, for them, it is considered a symbol of perfection. However, degrees of uncertainty, like in every statistical study, can be reduced by increasing the sample size; the larger the number of observations, the more probable the study would be significant. Moreover, extant literature confirms that objective and subjective measures, even though they differ in type, are strongly correlated (John, 1999).

To better understand the dataset it is mandatory to perform a pre-analysis like data visualization. The main insight using kernel density plot is that distinguishing it by customer satisfaction, it is possible to see that arrival delay is a little bit more influential than departure delay; in both cases, when the delay is of many minutes, customers are more likely to be dissatisfied. This assumption is represented by the distribution and it is also logical and accepted by customers. However, a higher number of dissatisfied customers is expected in case of a delay, which is not the case reported in the density plots in *figure 3* and *figure 4*. The density plot is almost the same when distinguished by customer satisfaction; a possible explanation is that very frequently, a delay is characterised by causes of force majeure; consumers recognise this (e.g., bad weather or delay imposed by the airport) and do not blame the airline for the delay, so satisfaction ratings are not likely to be affected. Kernel density plot are also presented in *appendix A* for a more complete and easier data visualization.

Also, a very interesting pattern in *Figure 5* is obtained by plotting flight distance and age using the scatter method. Satisfaction by age is well distributed, however flights of long distances are composed of a higher number of satisfied customers. In the scatter plot, it is possible to individuate a cluster of individuals in the middle age range from 20 to 60 when the flight distance is longer than 1,500 miles. Another visualisation using the scatter method implies the use of the variable “flight distance” and as the second variable “departure delay in minutes”, as seen in *figure 5*. The longer the distance, the more satisfied customers seem to be; what is interesting in *figure 6* is referred to as the second dimension. It would be expected that the higher the delay in minutes, the higher the number of dissatisfied customers. This assumption is unconfirmed; there are still satisfied customers with a delay higher than 1,000 minutes. Looking for a pattern like this is a key activity in order to have meaningful assumptions before machine learning is used. It is possible that the same pattern analysed can be used from the model or that the model can find different patterns.

The following variables, by using bar chart visualisation in showed in *figure 8*, seem to influence customer satisfaction: Type of Travel; Class; Inflight WiFi Service; Online Boarding;

Seat Comfort; Inflight entertainment; Onboard Service; Leg Room Service; Baggage Handling; Cleanliness.

The presence of satisfied and unsatisfied consumers seems to be influenced by the change in the mentioned "dimensions." Type of Travel is a binary variable which reports whether the trip was a Business or Personal trip. This explains the reasoning behind each travel which can be very informative in understanding the customer base. The number of business travel passengers is much higher with regards to personal travel. This could mean that U.S. airlines' dataset business is tailored to focus on a specific customer. Around 70% of customers took a business trip versus 30% of customers who travelled for personal purposes. As demonstrated in the *Type of Travel bar chart* below, more than half of customers who travelled for business purposes were satisfied. Instead, customers who travel for personal reasons are more likely to be unsatisfied. These findings drive even more our assumption that airline companies are tailor-made for business trips because they can satisfy customer needs when travelling for business instead of for personal purposes. 53% of U.S. residents took a business trip once a year in 2017, which fell to 36% in 2022 due to the implementation of remote working (Statista Survey, 2017; Statista, 2022). Considering that the survey that has generated the dataset is only among people who have travelled, it is evident that the percentage of people that took a business trip is much higher.

Class is a categorical variable with three different values. It is very similar to customer type, but instead of identifying the type of travel (business or personal), it identifies the class used to travel. For this reason, not only are the information of flights provided, but also the hypothetical magnitude in price can be computed assuming that business class is much more expensive than eco and eco plus. A considerable number of observations were made during a trip using the business (48%) and economy (45%) classes. What is interesting in this case is to see the same pattern that we have identified in the customer type dimension. Customers who travel using business class are more likely to be satisfied than those who travel using economy class. However, in this case, the variable class does not explain the reasoning behind a specific trip but the related quality-price. The difference in more satisfied customers in business class may highlight one more time the previous assumption about the fact that the companies could be tailor-made for a business trip. However, in this case, the number of customers in economy class is almost equal to the number of customers in business class and what difference is just the proportion of satisfied customers. Business classes are obviously more comfortable, but the price is higher. For this reason, it is not justified the fact that customers in the economy class are more likely to be dissatisfied, they pay a lower price, and they should expect a lower quality service; for this reason, their evaluation of satisfaction should

be balanced. This could highlight the lack of service quality in eco class which seems to be not aligned with consumers' expectations and paid price. Indeed, these assumptions offer the proposition for future research in which it should be understood why customers in the eco class are less satisfied if the price justifies the service.

Inflight WiFi Service seems to have a strong impact on dissatisfied customers. Even with a rating of 3, which is in the middle, the majority of individuals are dissatisfied. At a rating of 5, almost all customers are satisfied. This finding is very interesting because having a rating of 5 on this variable could really have a strong impact on airline companies, however as said before, the influence of other variables is not considered in the bar chart, which means that a 100% focus on inflight Wi-Fi service may be useless if other important variables can drastically reduce the influence of the variable of interest. Another assumption is that the presence of customers on a business trip drives the fact that the presence and the quality of Wi-Fi service are important for work reasons.

Online Boarding shows a very significant gap between low and high values; indeed, this variable could really make the difference in having a large number of satisfied customers having a rating of 4 or 5 drastically increase the proportion of satisfied customers. However, these ratings are provided by only 50% of the total individuals in the dataset; 70% of customers are satisfied in the case of a rating of 4 and 87% of customers are satisfied in the case of a rating of 5.

Seat comfort, inflight entertainment, leg room service, baggage handling and cleanliness report a similar bar chart in which it is possible to see a greater influence but without evident changes in satisfied customers by each rating. due to the reduced size of the bar graphs in the text, the original size of each bar graph can be found in *appendix C*

The last pre-analysis is performed using Pearson correlation, there is not a variable which distinctively influences customer satisfaction. Most of the variables (excluding arrival/departure delay in minutes, gate location, departure arrival time convenience, age and gender) seem to have a middle impact on customer satisfaction. The most correlated variables with customer satisfaction are online boarding, class, type of travel and inflight entertainment; the highest linear relationship with customer satisfaction is with the variable online boarding ($r = 0.5$).

Once the dataset is cleaned, it is possible to start building machine learning models. The training and test set splits used for all the models is 70% of the training set and 30% of the test set. Due to the purpose of the research, the hold-out set was not used because it is mainly used in the case of maximising the generalisability of models' prediction and comparability between them. In doing this, it is possible not to reduce the number of individuals for the training set too much.

Satisfaction is an antecedent of customer loyalty (Anderson, Fornell, & Lehmann, 1994; Anderson, Fornell, & Rust, 1997; Fornell, 1992). Hence customer type cannot be used as a customer satisfaction predictor and thus removed from the dataset. Another removed variable is arrival delay in minutes to avoid multicollinearity problems due to its high correlation of 0.99 with departure delay in minutes.

The main metric used to evaluate models' performance is the AUC which allows for better evaluation and compares models among each threshold. Indeed it is possible not only to see the AUC score but also the ROC Curve for each model built. Other reported metrics are Accuracy, Recall, Precision and F1 Score, mainly used to detect anomalies but can also be useful for comparing models.

The logistic regression model is good at predicting customer satisfaction (AUC = 0.917); hence its coefficients can be used to understand the most influential variables in air travel to have a satisfied customer. Metrics used to evaluate the model are reported in *Table 2*. The threshold used is 0.5, and through the ROC curve in *Figure 13* it is possible to see how the model behaves at different thresholds.

The coefficients of each variable in the logistic regression are shown in *figure 14*, from the highest to the lowest. Type of travel, online boarding and inflight Wi-Fi service are the three most influential variables that positively influence customer satisfaction.

Using Classification Tree, The three most important variables, as shown in *figure 16*, are still online boarding, inflight WiFi service and type of travel with a different order compared with logistic regression; online boarding jumped to first place using the classification tree model. The peculiarity of a single tree-based model is that it is possible to inspect how each decision is made at each different split. Indeed it is possible not only to rank the most important variable using the Gini score or the position in the tree graph, but it is also possible to see at which threshold a specific variable was set in order to make the best question possible to the dataset and minimise the impurity of each split.

As normally expected, the random forest model performed better than the single decision tree. Using random forest, model accuracy increases, but due to the computed averages of Gini decrease in the impurity of each variable, the difference in scores of each variable is smaller compared with a single classification tree and the most important variables, despite being evident in the graph, appear to have less influence on the overall pattern.

Like Random Forest, gradient boosting is an ensemble model which involves more than one tree that has to be generated. Even though the number of generated trees is reduced compared with

random forests (from 200 to 50), it is not possible to interpret them like in the case of a single tree. However, despite the reduction in interpretability using a gradient boosting classifier, it is still possible to compute the feature importance (*figure 21*), which is computed as the (normalised) total reduction of the criterion brought by that feature. It is also known as the Gini importance. The most critical feature determined by the gradient boosting classifier is online boarding, inflight Wi-Fi service and type of travel.

Kernel Support Vector Machine has very good performance, however due to the radial basis function applied it is not possible to interpret it like in the case of other models. For this reason as stated before two interpretability model like Feature Permutation Importance and Shapely Additive Explanation were applied to the built machine learning models.

Feature permutation importance is in the spectrum of Global Interpretability techniques, which explains the model's behaviour across the full range of inputs. It can be used for every model, especially for black box models.

Feature importance using permutation is scored using the decrease in accuracy. Specifically, the technique is based on taking one variable at a time and rescoring the model after the chosen variable is shuffled. This means data information of the chosen variable does not make sense anymore, and the accuracy should decrease. If the shuffled variable was a critical variable, the model's performance should drop significantly because the model can no longer rely on the chosen variable. The process is repeated for all the variables in the dataset. Once it is finished, it is possible to plot the decrease in accuracy of each variable, sorting them from the most to the less important one. The main reason feature permutation importance is a global interpretability technique is that when one variable is shuffled, the model performances are still influenced by all the other variables. This concept is vital to understanding how interpretability by design differs completely from other interpretability techniques. A coefficient defines the model's output when related to a specific variable, which is different in computing the decrease in accuracy of the entire model. It is important to understand that global interpretability helps understand model behaviour but does not explain how it works. For this reason, comparing different results is mandatory to increase the power of global interpretability using feature permutation importance which offers a different perspective based on the model performances. All the results of feature permutation importance of the five used models are plotted in *figure 22*. A more detailed version of the following plot is also reported in Appendix E.

The pattern of feature permutation importance of the five used models is the same provided in the coefficients of each model using interpretability by design. Indeed, there are still large gaps

between the three and four most important variables to all the other variables. Inflight Wi-Fi service is at the top of the most important features and in the third position only for the logistic regression feature importance. The type of travel is always in the second position considering the five models permutation importance. Online boarding is in the third position for Random Forest, Gradient Boosting and Kernel Support Vector Machine permutation importance and in the first position for Logistic regression permutation importance. In classification tree permutation importance, online boarding is in the fourth place because class is in third place only for this model. Indeed, it is possible to provide many more comments about the provided partial dependence plots; however, for the sake of simplicity, only the first three variables are commented on. In the results section, a more general view of the variable of influence will be offered to compare all the acquired results.
Shapely additive explanation is introduced by Lundberg and Suu-In (2017). It provides both global and local interpretations and explanations related to interactions between variables. SHAP helps to analyse all possible combinations of a given feature and gives an idea of the individual importance of each feature. However, it also shows how different features affect the model's scoring. It is possible to see in *figure 23* four different SHAP beeswarm plots for the Logistic regression, classification tree, random forest and gradient boosting model; in the case of Kernel SVM, due to lack of processing power, it was not possible to run kernel explainer and compute SHAP values due to the complexity of the model.

SHAP plot reports display data from the most influential to the least. With SHAP's positive values, the prediction increases. In the case of classification, a positive prediction means that a customer is satisfied. For each feature, the beeswarm plot (*figure 23*) also reports its distribution using a minimised scatter plot which can be used to see each individual's influence on SHAP values; a red observation has a high value and a blue one a low value of the variable of interest which is reported on the left. Using the SHAP beeswarm plot, it is possible to individuate when high values influence a positive prediction or vice versa. In the case of this research, the majority of the features report a satisfied customer when the value of a specific feature is high. The reason is that the majority of features are customer ratings; when a rating of a specific service is high, it means that the customer has provided a good grade, which consequently influences a positive outcome in satisfaction; in the case of binary variables, 1 is considered high value and 0 a low value. For Type of travel, it is possible to see only a very strong red and blue. In the case of class, it is possible to see the three colours related to eco plus, eco and business class, where the business class takes the highest value (red dot). Indeed, the reason behind ratings and SHAP results

demonstrate the accuracy of the beeswarm plot, which follows the logic behind rating and customer satisfaction (the higher the rating, the higher the customer satisfaction).

Tree based algorithms provide a different SHAP beeswarm plot compared to logistic regression. In fact, for the classification tree, it is possible to see that low values of a specific variable not only lead toward an unsatisfied customer but also towards a satisfied one. This is because trees are based on a specific path of questions, and the output is not computed like in logistic regression in which the coefficient is multiplied by the value of a specific feature; for example, if the coefficient of a hypothetical feature is 2 the value related to that feature will always be multiplied by 2 leading inconsistent output. In the case of trees, an observation with a low value of a specific feature can fall into leaves in which the majority of individuals are satisfied, and the prediction for that feature will be satisfied even if a specific value is low.

Using all SHAP plots, the most critical features are type of travel, online boarding, inflight WiFi service and class. These results are the same feature permutation importance, which is almost aligned with the coefficients of interpretable models by design. Using the SHAP plot, the research provides a more robust analysis of the most influential variables, offering also local interpretability. The results of the most important variables are provided by both interpretability by design and model inspection, like feature permutation and SHAP. To identify the four most important variables, a comparison between different models and inspections was made by computing each method's average of the respective order. In general, results from model coefficients are more accurate with regard to model inspection because they are directly used from the model. Model inspection is a valuable tool to confirm findings from interpretability by design, but it is preferred to not only rely on the results of feature permutation and SHAP if the model analysed is not a black box model; it is possible to fully rely on feature permutation only in the case of Kernel Support Vector Machine because interpretability by design is not possible in this case.

Interpretability by the design of logistic regression, classification tree, random forest and gradient boosting says that the four most important variables influencing customer satisfaction from the first to the fourth are Online Boarding, Inflight Wi-Fi Service, Type of Travel and Class. In the case of feature permutation, the four most important variables from the first to the fourth are Inflight Wi-Fi Service, Type of Travel, Online Boarding and Class. In comparison, for SHAP, the four most important variables from the first to the fourth are Online Boarding, Type of Travel, Inflight Wi-Fi and Class. Both interpretability by design and SHAP provide the same fifth variable, Inflight Entertainment. All three different methods of interpretation are aligned; indeed, it is possible to state that Online Boarding, Inflight Wi-Fi, Type of Travel and Class are the four most important

variables without a specific order according to all the methods used. The best model to compromise interpretability and accuracy is a gradient boosting classifier which provides from the most important to the least important Online Boarding, Inflight Wi-Fi, Type of Travel and Class. The best two variables using the output of interpretability by design and model inspection are Online Boarding and Inflight Wi-Fi Service.

The existing literature is exhaustive regarding customer satisfaction in general and for specific businesses (Anderson, Fornell, & Rust, 1997; Anderson, Fornell, & Lehmann, 1994; ACSI, 2021; Fornell, 1992; Hult, Sharma, Morgeson III, & Zhang, 2018; Kotler & Keller, 2012); however, there is lack of research about the most valuable attributes in each industry in order to improve customer satisfaction. This research aims at providing the most critical dimension of the influence of customer satisfaction in the air travel industry. Indeed, there is a lack of explanatory power for each of the most important variables found. Qualitative research should be applied to the main findings of this research to explain how and why variables of influence are crucial for a satisfied customer.

There is an evident gap between business travel and personal travel. The second type of customers are more likely to be dissatisfied even if the number of people that travel for personal reason is smaller compared to business travel. Future research should understand if the major reasoning behind this is provided by the fact that the company pays for a business trip, and a personal trip is paid for by the customer, increasing the likelihood of being dissatisfied.

Customer satisfaction shows significant changes regarding the travel class, where consumers in business class are more likely to be satisfied than those in eco class, considering that the price paid is relative to the service offered. Future research could provide information about the misalignment of customer satisfaction in business class and eco class, assuming that the price paid should lower the expectation of an eco class consumer. The importance of understanding customers' behaviour in a different class is even more important considering the disruption of the airline business caused by COVID-19 and the Russian invasion of Ukraine. The CEO of Ryanair said: “There’s no doubt that at the lower end of the marketplace, our really cheap promotional fares, the one euro fares, the 99 cent fares, even the 9.99 fares, I think you will not see those fares for the next number of years.” (CNBC 2022)

Most individuals in the dataset, and consequently in the United States, travelled for business reasons. These customers are more likely to be satisfied. Future research could be performed from the airlines' side, trying to understand and explain if and why airlines are tailored for business trips. Moreover, a business trip may lead to an increase in the importance of inflight WiFi service providing reasoning for both variables type of travel and inflight WiFi service.

Although cleanliness is not considered an important variable in influencing customer satisfaction, due to COVID-19 major changes have taken place worldwide. Those changes may converge the greatest importance from all the most important variables found to the variable cleanliness. It is important for future research to understand if these changes have taken place and to monitor their impact on customer satisfaction.

Surprisingly, departure delay has no impact on customer satisfaction. Future research should understand why a travel-related variable does not influence customers. It could be inferred that departure delay can be justified by causes of force majeure, which are recognisable by customers. In a Managerial point of view, air travel is a highly competitive business where low-cost companies have influenced the market with low prices compromising customer satisfaction (Boetsch, Bieger, & Wittmer, 2011). For this reason, it is important to allocate resources with the greatest accuracy to maintain low prices and be competitive in the market, mainly due to the increased number of satisfied customers.

Airlines should not neglect the impact of digitalisation and its influence of it on a multi-channel strategy; indeed, online boarding and inflight Wi-Fi services are considered two of the most important variables in influencing customer satisfaction, from NielsenIQ research (12 October 2021) for the "Osservatorio multicanalità". It is evidenced that an increasing number of people use a digital approach to buying and consuming services. Managers should focus on having a satisfied customer even before the travel service takes place. Indeed, it is mandatory to provide a very simple and user-friendly online boarding experience, compatible with all digital devices, to have half of the job done in having a satisfied customer.

The type of travel is not a variable that airline companies can control. There is no evident explanation as to why customers who take a personal trip are more likely to be dissatisfied. However, knowing there is a strong influence on the type of travel, airline companies could try to understand if their business strategy mainly focuses on business travel rather than personal travel.

Different class shows different grades of satisfaction. The service in business class is better, and people are more likely to be satisfied. However, the price paid for eco classes should justify the service offered by the airline companies, which customers should understand. Nevertheless, this research shows that this is not the case. Customer satisfaction maximisation may reduce companies' productivity (Anderson, Fornell, & Rust, 1997). For this reason, it is not mandatory to maximise the service quality in eco classes which may lead the companies to become unproductive just to maximise customer satisfaction. A threshold should be found. The main advice is to

maximise digital services satisfaction to see if online boarding can reduce the effect of dissatisfied customers from eco classes using the good influence of the first meeting point (online boarding).

Companies able to leverage customer satisfaction to obtain optimal values on the most important variables could be more competitive in the market trying to fight the low pricing rule. Moreover, a highly digital airline company could use the company's website to prevent loyal customers from using air travel aggregators, which leads customers to a price comparison mentality.