Exploratory data analysis and prediction of passenger satisfaction for a US airline company

François Le GAC

Kaggle competition 2020

Abstract

Airline companies are investing considerable amounts of money in order to improve the quality of all their services provided during passengers journey. Nowadays, with the dramatic rise of social networks any passenger can share his flight experience in the blink of an eye. Therefore, narrowing the gap between passenger expectations and delivery hase become a key challenge for airline companies.

But what drives passenger overall satisfaction? In other words, which factors are highly correlated with passenger satisfaction? This paper try to respond to these questions through an exploratory data analysis. At the same time, we draw personas of a loyal customer and a satisfied customer. We conclude this paper with the prediction of customer satisfaction, achieving 93% test accuracy with state of the art machine learning algorithm called Random Forest.

Table of contents

I. Exploratory data analysis

I.1 Dataset description and objectives

In April 2017, the video of a passenger dragged off its flight shocked the internet. The customer, a 69 years old man, had the noise bleed and several other bruises. The next day, the stocks of the american airline company, United Airlines, plummeted by $1 billion dollars. This event resulted from overbooking, but it clearly shows how a bad customer experience can dramatically impact the market value of a company.

Thus, airline companies are constantly trying to improve their quality standards in order to meet growing customer expectations. In 2019, Emirates, the largest airline company from the United Arabe Emirates, led the international airline satisfaction. In order to rank first in this competition, companies continuously send surveys and try to understand what are the most important factors of customer satisfaction.

We will be working on a kaggle dataset entitled *Airline Passenger Satisfaction*. A US airline company has gathered more than 125k passenger reviews such as satisfaction scores on inflight service, on online service etc.. Unfortunately, the name of the company is unknow, the dataset has probably been anonymized.

The aim of this competition is to understand what are the main factors of passenger satisfaction. While going through an exploratory data analysis to answer this question, we will provide interesting insights from segmentation. Different personas will emerge: what are the main characteristics of a typical satisfied customer? How does a dissatisfied customer look like? Finally, we will try to forecast customer satisfaction using state of the art machine learning techniques.

The kaggle competition provides two files: a train set and a test set. Each file contains twenty-three independent variables and one dependent variable, namely the satisfaction variable. The train set consist of about 100k rows and the test set 25k rows. For the data analysis we will combined these two datasets into one big dataset in order to apply data transformations only once. We will take a closer look at this dataset.

I.2 Variables description

In this competition we will have to deal with twenty-three independent variables and one dependent variable, namely the satisfaction. Some variables are categorical such as the customer gender (Male, Female) and some are numerical like the flight distance.

The table below try to sum up the different variables, their type and their description. We will go trough some of them to get a deep understanding of what could have a high influence on the passenger satisfaction.

Type	**Variable**	**Description**
Customer info	Gender	Gender of the passengers (Female, Male)
	Customer Type	The customer type (Loyal customer, disloyal customer)
	Age	The actual age of the passengers
Flight info	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
	Departure Delay in Minutes	Minutes delayed when departure
	Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	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 convenient
	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
	Checkin service	Satisfaction level of Check-in service
	Inflight service	Satisfaction level of inflight service
	Cleanliness	Satisfaction level of Cleanliness
Dependent var	Satisfaction	Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

Fig.1 - Description of the variables

First we have some information about the customer: the gender, his age and the type of customer (Loyal or Disloyal). The latter will be of a great interest for our segmentation. Then we have several variables related to the flight: the type of travel (Business or Personal travel), the class (Business, Eco, Eco+), the flight distance and potential delay for the departure as well as for the arrival.

We also have fourteen variables corresponding to the satisfaction scores for products or services provided by the company. Indeed, passengers were asked to grade their satisfaction between 0 (Very bad) and 5 (Great) regarding these following topics: the inflight wifi service, the food and drinks, the seat comfort, the cleanliness etc… All of these variables are quite straightforward so we will not spend too much time on it.

But maybe we can try to guess which variable will have the highest impact on passenger satisfaction. This is purely personal but it can help us have some intuition for the rest of the

analysis. The variables that, according to me, seem to have the highest impact on passenger overall satisfaction are the departure / arrival delays, the on-board service, baggage handling and the type of travel (Business or Personal).

I.3 Missing values

The dataset contains missing values. Missing values can dramatically decrease the performance of machine learning models. Therefore, understanding and transforming or even getting rid of missing values is critical.

Only one variable has missing values. The arrival delay in minutes contains 393 missing values. We don't exactly know why these values are missing, but instead of getting rid of the corresponding rows, let's try to fill the NaNs with the help of an other variable the 'departure delay in minutes'.

First, we notice that you have 80% chances to arrive on time when you have left the airport on time. Moreover, below are the histograms of the departure and arrival delays.

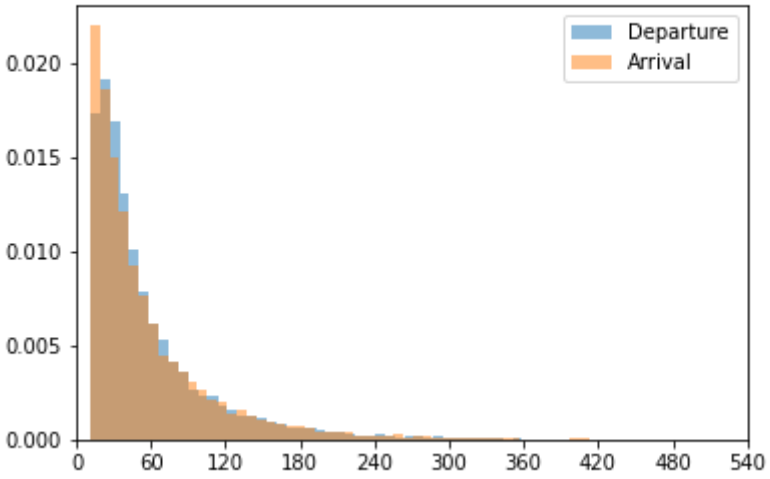

Fig.2 - Histograms of departure / arrival delays (starting from 10 minutes late)

We see that the two histograms are almost layered. We can make the hypothesis that the departure delay is approximately equal to the arrival delay. And we fill the missing values accordingly.

I.4 The dependent variable

The aim of this competition is to predict the dependent variable: the satisfaction. Let's quickly take a look at this categorical variable that has two categories: satisfied vs neutral or dissatisfied.

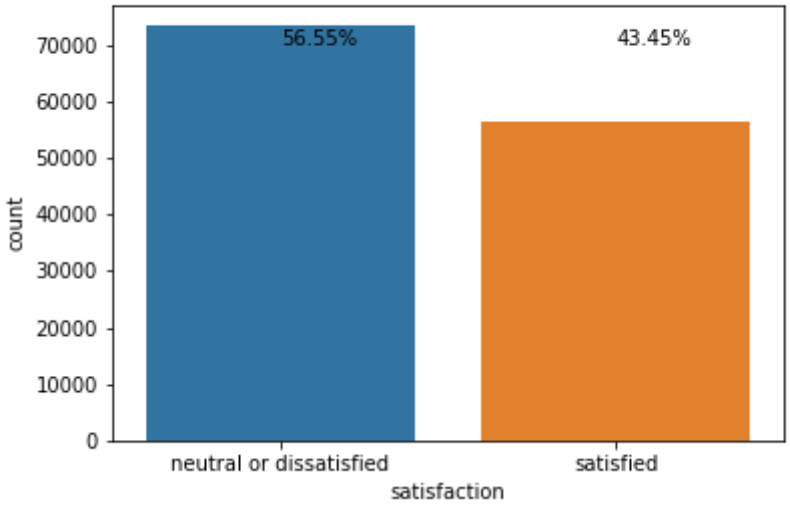

Fig.3 - Barplot of the dependent variable

We can see that about 57% of the passengers are neutral or dissatisfied and consequently 43% of the passengers are satisfied. The dataset is balanced, which confirms that the choice of accuracy as the evaluation metric is a good choice.

I.5 Univariate data analysis

In this part, we will look at some variables of the dataset individually and try to get the main characteristics of our individuals.

Our dataset is equally composed of males and females (~ 50%/50%). We also almost have the same number of passengers travelling in Business as in Eco (~ 45%/45%) and consequently it is interesting to note that the Eco Plus class is not really represented (~ 10%).

On the contrary, there are much more loyal customers (~ 80%) than disloyal customers (~ 20%). Moreover, most the passenger fly for business purposes (~ 70%) while only a few fly for personal purposes (~ 30%).

The mean age is about 40 years old and it seems that we have a wide range of ages among the customers. Below is a histogram representing the distribution of the variable:

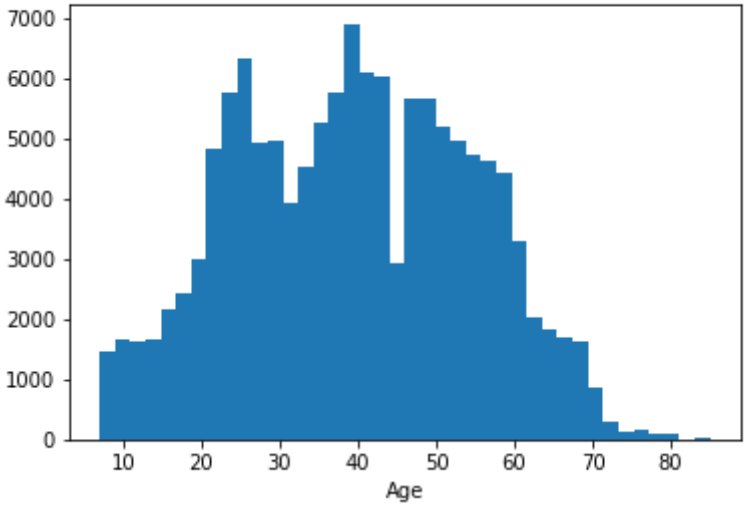

Fig.4 - Histogram of age variable

It is not very convenient to work directly with this numerical variable. We will transform it to a categorical variable with 3 age ranges: (0- 30 years old], (30 - 50 years old] and [more than 50 years old] (representing respectively about 45%, 30% and 25% of the passengers).

Most of our variables are related to satisfaction. Passenger where asked to give a score between 0 and 5 to grade their satisfaction regarding products and services provided by the company. Let's try to find some interesting insights:

Variable name	**Mean sat.**
Inflight service	3.64
Baggage handling	3.63
Seat comfort	3.44
On-board service	3.38
Inflight entertainment	3.35
Leg room service	3.35
Checkin service	3.30
Cleanliness	3.28
Online boarding	3.25
Food and drink	3.20
Departure / Arrival time convenient	3.05
Gate location	2.97
Ease of online booking	2.75
Inflight wifi service	2.72

Fig.5 - Mean satisfaction for each satisfaction variable

The three services where the airline company performs the most are inflight services, baggage handling and seat comfort with the highest mean satisfaction scores, respectively 3.64, 3.63 and 3.44. On the other hand, the company performs quite badly for the gate location, the ease of online booking and the inflight wifi service with the lowest mean satisfaction scores, respectively 2.97, 2.75 and 2.72.

Before providing the company with recommendations on which service it should improve and thereby spend money on, it is important to know which factor has high impact on the overall satisfaction. Indeed, there is no need for the airline company to improve a service if it hang very little in the balance.

Looking at the mean is a common practice but we also need to have an idea of the dispersion. For example, let's imagine that the mean satisfaction score for the inflight wifi service is 2.5. Let's consider two extremes cases in order to illustrate the critical idea of dispersion. The first one, is that all the customers have given the grade 2.5. The second one is that half of the customers have given the grade 0 and the other half 5, which also gives a mean value of 2.5. But as we can see here, the two cases are completely different and this is where dispersion comes in.

Back to our case study, we would like to know if only a few part of the passengers are driving up / down the score or if there is a 'common middle grade'.

Fig.6 - Boxplots of our three best satisfaction variables

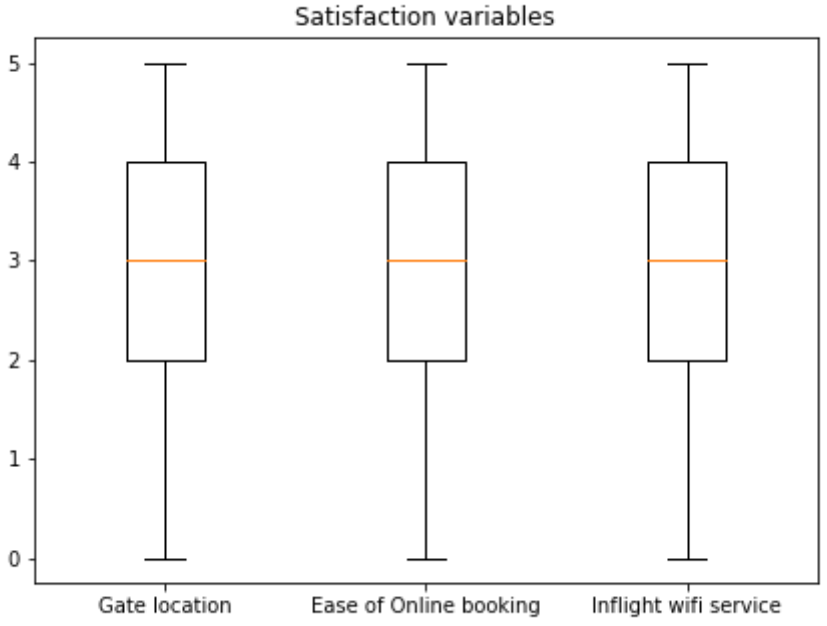

Fig.7 - Boxplots of our three worst satisfaction variables

Boxplots represented above (Fig.6 and Fig.7) provide us with valuable information about customer satisfaction. Without surprise, inflight service, baggage handling and seat comfort have the highest median ~ 4. It means that half of the passengers have given a grade higher than 4 (and therefore the other half have given a grade below 4). But the boxplot provide us with some additional information regarding the dispersion: 75% of the scores for the inflight service and the baggage handling are between 5 and 3 while it's between 5 and 2 for the seat comfort which has more dispersion. Also, it might be interesting to note that the lowest grade for baggage handling is one whereas the lowest value for the inflight service and seat comfort is 0.

Finally, let's quickly look at the variables related to the delay. Below are the two histograms representing the variables:

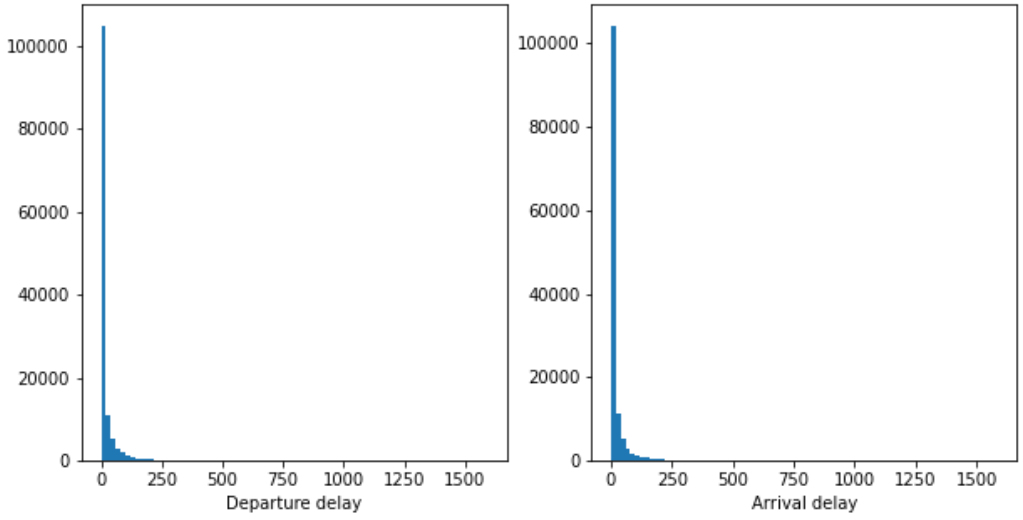

Fig.8 - Full histograms of departure and arrival delays

Again we see that leaving late the airport generally means delay at the arrival. Moreover, we see that the majority of the flights don't have delay.

Now let's take a look a the flight distance. This is a numerical variable and below is a histogram showing the distribution of the variable:

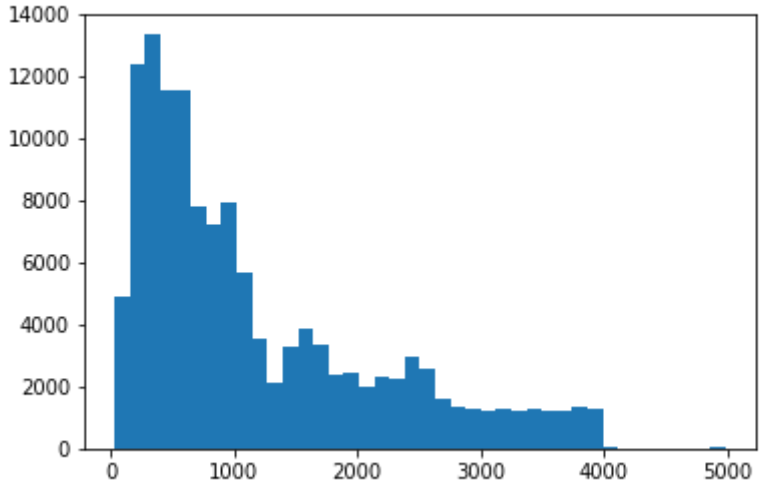

Fig.9 - Histogram of flight distances

Unfortunately we don't know if the distances are in km or in miles. Anyway, we notice that most of the flight distances are below 1,000 km. The maximum distance is about 5,000 km, the minimum distance is about 30 km and the average flight distance is about 1,000 km. Again, for further analysis, we transform this variable into a categorical variable with three categories. The first one corresponds to the short-hauls (0-1,000 km), the second one the medium-hauls (1,000-3,000 km) and third one the long-hauls (more than 3,000 km). Short-hauls account for about 60%, medium-hauls 35% and long-hauls 5%.

I.6 Feature engineering

We remember that one of the aims of this competition is to predict the satisfaction using machine learning algorithms. The independent variables will fuel the machine learning machine during the training step until we get the highest accuracy (% of correct answers) as possible. The winner of the competition is the one who can achieve the highest accuracy. The creation of variables, also called feature engineering is a key element for those who want to achieve the best accuracy results.

We create the first variable called 'total delay' by additionning the departure and arrival delay. We also transform this new numerical variable into a categorical variable with three categories: flights without delay (~45%), flight less than 10 minutes lates (~35%) and flight more than 10 minutes lates (~20%).

Our dataset contains fourteen satisfaction variables regarding each service provided by the company (the Food and drinks, the wifi service etc…). The idea here is to create an indicator that summarizes most of the information of the satisfaction variables. We create the mean satisfaction indicator which is the mean of all the satisfaction variables. Below is the distribution of the mean satisfaction indicator:

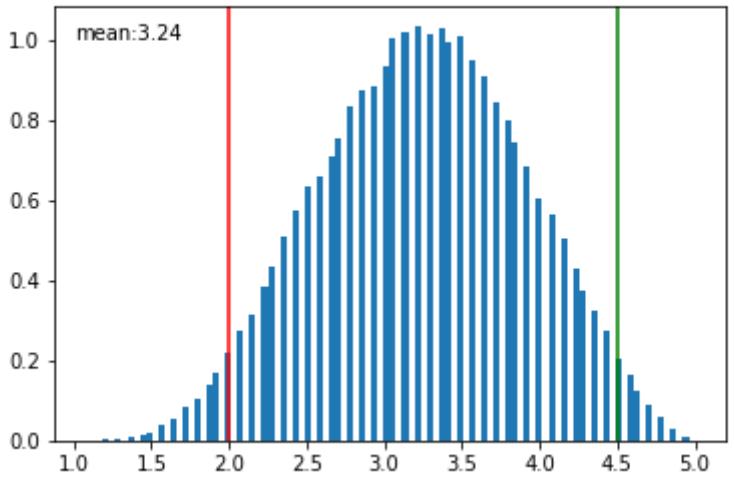

Fig.10 - Histogram of mean satisfaction

The distribution is concentrated around the mean which is 3.24. It is higher than 2.5 so we can say that in overall the company performs pretty well. Thanks to this chart we will be able to identify the very dissatisfied customers with a mean satisfaction score below 2 and the very satisfied customers with a satisfaction score above 4.5.

I.7 Drawing personas

Identifying different customer profiles becomes really important when the management decides the strategy. Segmentation is also crucial for the marketing department in order to better target new customers or to designed new loyalty programs. Moreover, in depth knowledge of your customers allows you to know which communication channel is more likely to be effective. In other words, this also helps the financial department as you know precisely which communication channel you should spend money on.

In this section, we will try to identify the characteristics of the loyal customer (vs a disloyal customer). Then we will present how is a satisfied customer vs a neutral or dissatisfied customer.

I.7.1 Loyal vs disloyal passenger

So first, let's try to draw the profile of the loyal customer. This might be a great help for the marketing department. Moreover, we have seen in the previous section that most of the customers are loyal customers (~80% i.e. 105,000 inds.) so it should help us for the rest of the competition.

	Disloyal Customer			**Loyal Customer**	
Type of Travel	**Size**	**Percent**		**Size**	**Percent**
Business travel	23579	99	Business travel	66114	62
Personal Travel	201	1	Personal Travel	39986	38

Fig.11 - Type of travel vs customer type

Thanks to the above table, we quickly notice that while almost all the disloyal customers are flying for business purposes (99%), 62% of our loyal customers are flying for business purpose and 38% for personal purpose. In other words, this means that almost half of passengers that regularly fly with the company do it for personal purposes.

Let's look at the class variable. The table below is quite surprising:

	Disloyal Customer			Loyal Customer	
Class	**Size**	**Percent**		**Size**	**Percent**
Business	9231	39	Business	52929	50
Eco	13634	57	Eco	44675	42
Eco Plus	915	4	Eco Plus	8496	8

Fig.12 - Class vs customer type

As almost all the disloyal customers are flying for business purposes, we would have expected a higher percentage for the business class. In fact, only 39% of the disloyal customers are in business class whereas 50% of the loyal customers are in business class.

Let's look at the flying distances:

	Disloyal Customer			Loyal Customer	
Flight distance	**Size**	**Percent**		**Size**	**Percent**
Long-haul	24	0	Long-haul	10289	10
Medium-haul	5445	23	Medium-haul	39019	37
Short-haul	18311	77	Short-haul	56792	54

Fig.13 - Flight distance vs customer type

We quickly see that loyal customers tend to fly for longer distances than disloyal customers. Indeed, 10% of the loyal customers take long-hauls while almost no disloyal passenger take long-hauls. Moreover, 77% of disloyal customers take short-hauls while 54% of Loyal customers take short-hauls.

To sum up, loyal customers are equally flying for business purposes as for personal purposes which is different from disloyal customers which are almost all flying for business purposes. Surprisingly, most of the loyal passengers are in business class while most of the disloyal customers are in eco class. Finally, loyal customers tend to travel longer distances than disloyal customers.

I.7.2 Satisfied vs dissatisfied passengers

Now we would like to identify the main characteristics of a satisfied passenger. This should help us to get valuable insights on what drives overall satisfaction.

We start with the type of travel:

	Neutral or dissatisfied			**Satisfied**	
Type of Travel	**Size**	**Percent**		**Size**	**Percent**
Business travel	37337	51	Business travel	52356	93
Personal Travel	36115	49	Personal Travel	4072	7

Fig.14 -Type of travel vs satisfaction

One can notice that almost all the satisfied customers are traveling for business purposes. On the other it's almost 50:50 for the dissatisfied customers. Does that mean that the business class is more appreciated and that it is a determinant of overall satisfaction?

To answer this question we look a the table below:

	Neutral or dissatisfied			Satisfied	
Customer type	**Size**	**Percent**		**Size**	**Percent**
Business	18994	26	Business	43166	76
Eco	47366	64	Eco	10943	19
Eco Plus	7092	10	Eco Plus	2319	4

Fig.15 - Customer type vs satisfaction

The table seems to partly confirm our intuition: most of the satisfied passengers are in business class (76%) whereas only a few of dissatisfied customers are in business class (26%). The majority of dissatisfied customers are in Eco class. This might confirm the fact that people in business class are generally satisfied and that the class is an important factor in the overall satisfaction.

Now let's compare the age distributions:

	Neutral or dissatisfied			Satisfied	
Age_group	**Size**	**Percent**		**Size**	**Percent**
(0-30]	28306	39	**(0-30]**	12574	22
(30-50]	27137	37	**(30-50]**	28261	50
(> 50)	18009	25	**(> 50)**	15593	28

Fig.16 - Age vs satisfaction

We see that the satisfied passengers tend to be older than dissatisfied passengers. Indeed, most of the satisfied passengers are 30-50 years old while most of dissatisfied passengers are under 30 years old. Furthermore, the satisfied group have a higher percentage of over 50 years old than the dissatisfied group (28% vs 25%).

Let’s look at the the flight distance:

	Neutral or dissatisfied			Satisfied	
Flight Distance_group	**Size**	**Percent**		**Size**	**Percent**
Long-haul	2336	3	**Long-haul**	7977	14
Medium-haul	20846	28	**Medium-haul**	23618	42
Short-haul	50270	68	**Short-haul**	24833	44

Fig.17 - Flight distance vs satisfaction

Satisfied customer tend to do longer flight than dissatisfied passengers. Indeed we have 14% long-haul for the satisfied customers versus 3% for the dissatisfied. We can also see this trend for medium-haul: 42% for the satisfied versus 3% for the dissatisfied.

Finally, let’s investigate delay’s differences between these two groups:

	Neutral or dissatisfied			Satisfied	
Total delay	**Size**	**Percent**		**Size**	**Percent**
less than 10min late	13303	18	**less than 10min late**	11177	20
more than 10min late	28716	39	**more than 10min late**	17186	30
on time	31433	43	**on time**	28065	50

Fig.18 - Total delay vs satisfaction

This table is really interesting... because it is highly surprising! Satisfied customers are facing less delay compared to the dissatisfied customers (50% versus 43%) as expected. But what strikes us most is the fact that the distributions are almost the same. In other words, it looks like both group are experiencing the same flights in terms of delay. And we should not forget that because it shows that the flight delays aren’t really important in the overall satisfaction.

To summarize, satisfied customers are almost all traveling for business purposes. They fly in business class, and are generally older than dissatisfied customers. Most of the satisfied passengers are 30-50 years old. They tend to fly longer distance with a high percentage of long and medium-hauls. Finally, satisfied customers are usually experiencing less delays, but the difference with the delays dissatisfied customers are experiencing isn't really significant.

I.8 Bivariate data analysis

Now that we have described our personas, we would like to find some other valuable insights by looking at pairs of variables.

We start by comparing the arrival delay to the departure delay. Unsurprisingly, both variables are positively correlated. The more you have departure delay, the more you have arrival delay.

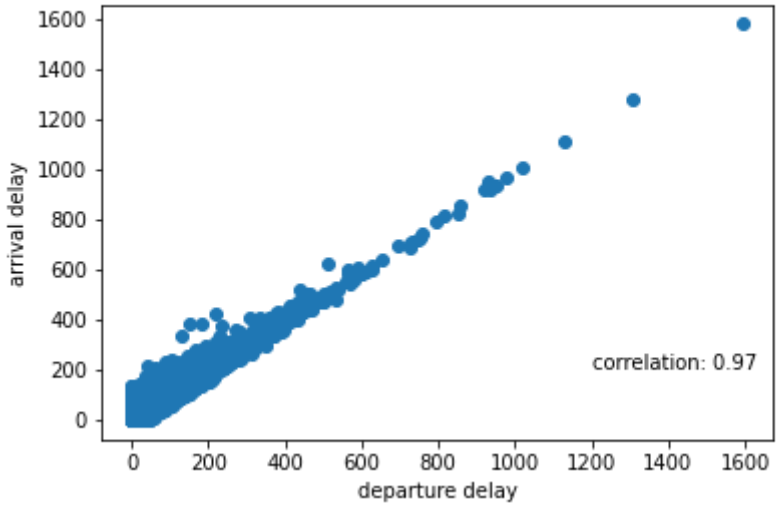

Fig.19 - Scatterplot of departure and arrival delays

Another interesting insight is the link between total delay and the flight distance. We were expecting longer flight to encounter more delays but the graph below show that there is almost no correlation between these two variables:

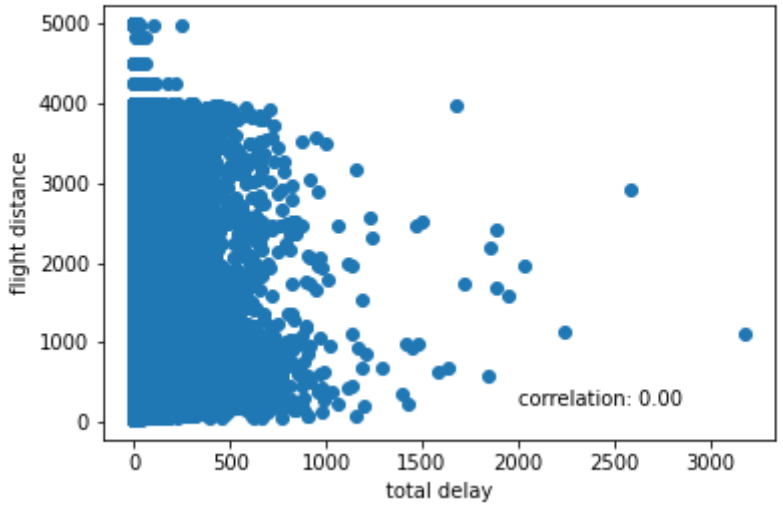

Fig.20 - Scatterplot total delay vs flight distance

II. Predicting passengers satisfaction

We mentioned earlier that the objective of the competition is to predict the passenger satisfaction. In other words, we will use the independent variables to predict the dependent variable, the satisfaction variable. To that end, the competition provided us with a train data set and a test dataset. We will use the train dataset to train our machine learning model and a test dataset in order to evaluate our model.

The metric to evaluate our model is the accuracy, which corresponds to the percentage of good responses. Note that we can make two types of errors:

- predict a customer as satisfied while he is not
- predict a customer as dissatisfied while he is in fact satisfied

To prevent this two types of errors in the same time, we will use the accuracy. Note that both the train and the test dataset contain about half satisfied passengers and half neutral or dissatisfied passengers.

Before starting to create our machine learning models we need to do some data preprocessing, namely create dummy variables, standardize our data and do some cross-validation.

II.1 Preprocessing

II.1.1 Dummy variables creation

What is a dummy variable and why do we need to go through this step ? Well, we said earlier that we will feed our machine learning model with the independent variables in order to output the predicted variable, namely the satisfaction. But you can't feed anything you want to the model. Most of the machine learning models don't like to have categorical variables as inputs. For example, we cannot provide the machine learning model with the variable gender (two categories: 'Male', 'Female'). At least not in that form. We need to transform it into a dummy variable before feeding the model.

Then what is a dummy variable? A dummy variable is a variable that takes only the value 0 or 1 to indicate the absence or presence of a category. Let's take two examples in order to illustrate this concept:

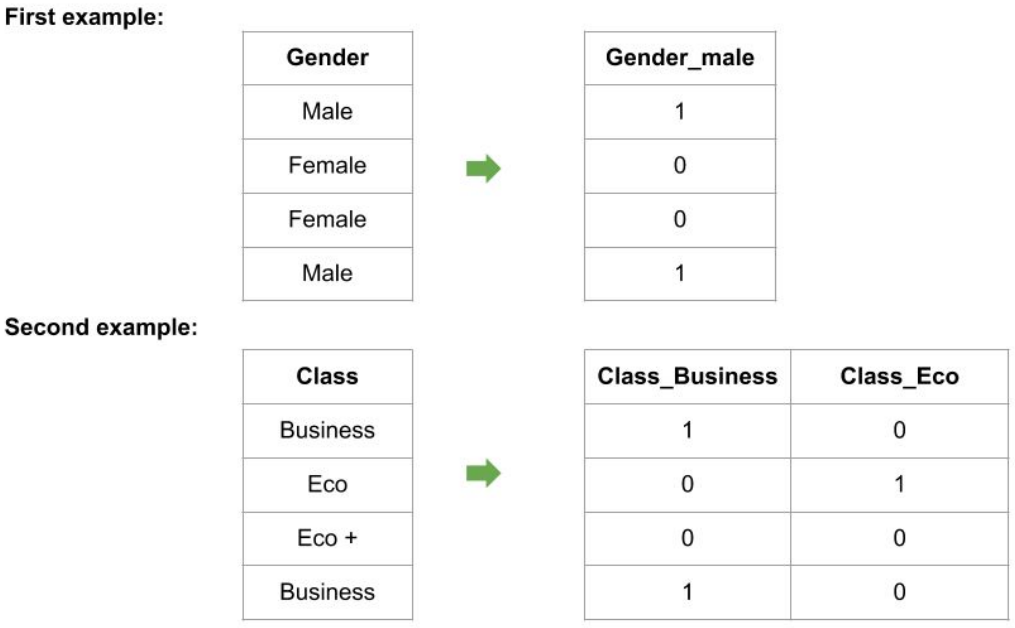

Fig.21 - Dummy variable transformation

We first take the gender variable which has two categories: Male and Female. We select one category, for example Male, and we create a dummy variable replacing 'Male' by 1 and 0 for the other category ('Female'). We don't need to create two dummy variables for the variable gender because one dummy variable contains all the information. Indeed, we deduce that all the other values 0 are Female.

The second example is the variable Class, a categorical variable with three classes: Business, Eco and Eco+. We now need to create 2 categorical variables as a third one would be deductible by the first two. Back to our case study, we transform all our categorical variables into dummy variables.

II.1.2 Standardization

Another common preprocessing step is standardization of the data. Standardization is a technique that allows you to get almost all your values between -2 and +2. For every variable, we retrieve the mean and divide by the standard deviation. Why do we need to go through this step? Let's take an example: a variable with range value between 100,000 and 200,000 will largely outweigh a variable with values between 0 and 1 when feeding the model. Therefore, it is critical to have all your values within the same range if you want your machine learning model to performs well.

II.1.3 Cross validation

The last step of preprocessing is cross validation. This step is quite simple: remember that we have combined our test set and train set into one big dataset for data analysis. Cross validation consists in separating back our big dataset into a train set and a test set. Cross

validation is crucial as it would be too optimistic to train our model and evaluate it on the same data.

II.2 Machine Learning model creation

II.2.1 Logistic regression as the baseline

The first machine learning model that we create is the logistic regression. It is a quite simple algorithm and it will be our baseline algorithm in terms of performance. Logistic regression is the brother of linear regression but for binary classification. Binary classification is when we try to predict a variable with two classes.

We want to get the highest accuracy as possible. In other words, we want to have the percentage of good responses as high as possible. In order to improve the accuracy of our logistic regression model, we will try to find the best hyperparameters. Hyperparameters are parameters specific to a model that can be tuned to improve the performances. We will not describe all these hyperparameters as it goes beyond the limits of this survey, but still let's give an example.

One of the hyperparameters is a coefficient call 'C' in python. This parameter is a coefficient of inverse regularization. Regularization is very useful when it comes to fight overfitting, which is very common problem when the model learns too much the training data and can't generalize with new data. Smaller values of C specify stronger regularization.

After tuning the hyperparameters, our best logistic regression model reach about 87% of test accuracy. It means that 87% of the passengers from the test dataset have been correctly predicted.

II.2.2 Random forest

Random Forest is an other machine learning algorithm famous for winning many kaggle competitions.

Again, we do some hyperparameter tuning. We tune several parameters such as the number of trees, the depth of tree etc… Our best random forest classifier reach a test accuracy of 93% which is better than the logistic regression! Moreover, there is something really interesting with the Random Forest: feature importance. Basically, the algorithm can quickly provide you with the importance of each variable for the prediction of the satisfaction variable. The graph below illustrates this idea:

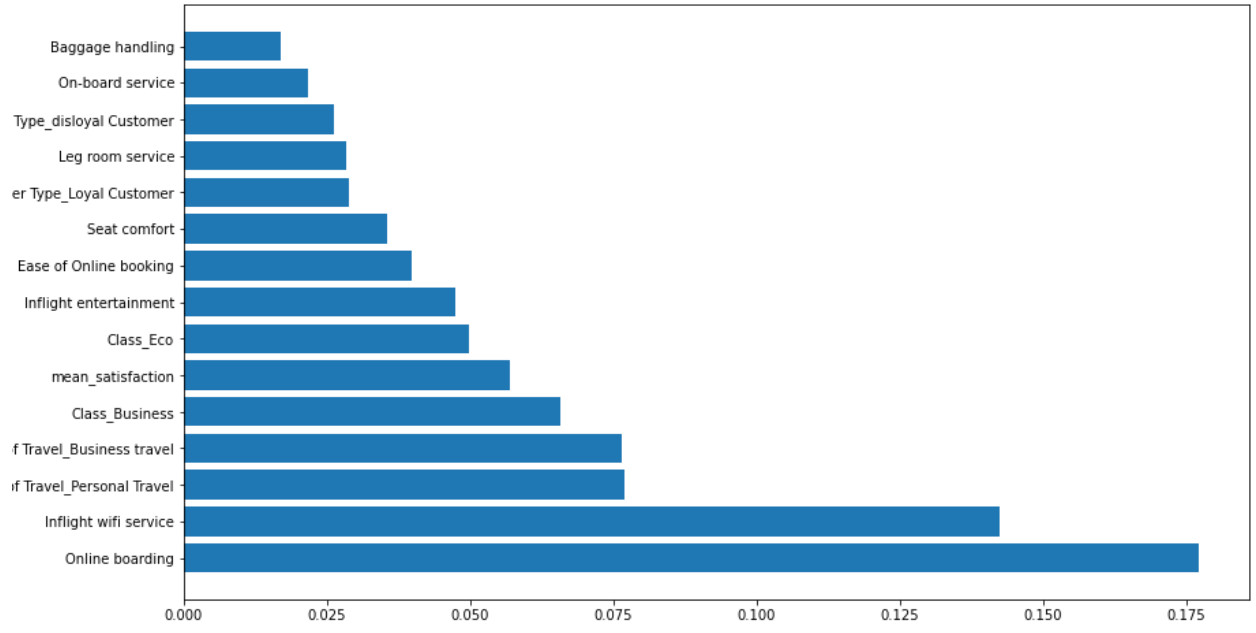

Fig.22 - 15 most important features

And this is highly interesting! We see that the biggest determinant for passenger overall satisfaction is the online boarding experience. Then comes the inflight wifi service, and the

type of travel (business or personal). As a reminder, we tried to guess at the very beginning of this paper, which variables would be the most important. Our first guess was the departure / arrival delays, the on-board service, baggage handling and the type of travel (Business or Personal). So we were right only for one variable, the type of travel! We were not expecting the online boarding experience or the inflight wifi service to respectively rank first and second.

II.3 Recommendations

So what can we recommend to the airline company? Let's start with the most important service: the online boarding experience. Generally a day before the flight, customers receive an email where they can online check-in, choose their seat and receive their boarding pass. This allow them to save time and to check that all details are validated before the flight. Surprisingly this service comes first in the overall satisfaction factor.

Is the company performing well regarding this specific service? Well not exactly: the mean satisfaction for online boarding is 3.25 which is largely lower than the mean satisfaction for the inflight service for example (3.65). Therefore, we recommend the airline company to improve the online onboarding. The company can either decide to create its own app for more flexibility or pay for the services of a digital web agency. In both cases, we recommend the company to invest in this crucial service.

The second most important determinant is the inflight wifi service. Nowadays, inflight wifi service not only gives you access to films or games, it also enable the company to display interactive information about the flight or about the security onboard. For a very long time, it

has been very difficult to bring wifi in the aircraft. Luckily, current technological advances have reduced the gap between passenger expectations and airlines offer.

It turns out that the company is performing really badly regarding this particular service! There is absolutely no doubt that the company must improve drastically the wifi onboard. It probably represent the company's area where there is the biggest room for development. We recommend the airline company to work with expert suppliers in the connectivity space.

The third and last recommendation is to maintain the quality of inflight entertainment. The company is doing well with this service (3.35 of mean satisfaction) and it ranks among the most important features.

II.4 Conclusion

After an exploratory data analysis we have succeeded to draw the profiles of loyal customers as well as satisfied customers. Understanding the former should help the marketing department of the airline company to design loyalty programs. Moreover, understanding what are the main characteristics of a satisfied customer gave us meaningful insights on what drives overall satisfaction.

Regarding the competition, logistic regression has shown good results in predicting if a customer is satisfied or not with a percentage of good responses of 87% on the test set. But the winner is the Random Forest with a test accuracy of 93%. Not only the Random Forest gave us very good results but it has also provided us with valuable and surprising insights: online-boarding, inflight wifi-service and the type of flight are three variables that influence the most customer satisfaction. Based on these discoveries, we delivered the airline company three main recommendations to improve their quality service. The quest of satisfying passengers will be very challenging but also very exciting!