AIRLINE CUSTOMERS SATISFACTION: AN EXPLANATORY ANALYSIS

Alice Anna Maria Brunazzi, Alessandro Della Beffa, Daniele Lepre

SUMMARY

This project presents an explanatory analysis of airline passenger satisfaction, using a dataset containing information about various aspects of the passengers' travel experience.

Predictive models and different machine learning techniques were explored to develop a KNIME workflow that allows the user to better understand the variables influencing customer satisfaction and to identify key factors that can enhance the travel experience.

1 INTRODUCTION

The aviation industry nowadays tends to increasingly emphasize customer experience as a central element for the airlines' success. In a competitive market sector, every company needs to view passenger satisfaction as a critical indicator of the quality of service offered, as it tends to be a key factor in customer loyalty and brand reputation.

This report focuses on analyzing passenger satisfaction using a comprehensive data set containing detailed information on passengers' in-flight experiences.

The dataset comprises variables that reflect various aspects of the passenger journey. The main objective of this project is to develop predictive models capable of anticipating passenger satisfaction levels. These models will be used to formulate recommendations aimed at improving the overall customer experience.

2 Dataset & Preprocessing

2.1 Dataset: Airlines customer satisfaction

The dataset uses the data collected by an American company with a fictitious name, 'Invistico', used to protect the privacy of the passengers. This dataset complies with customer details and feedback on various flight aspects.

Every customer was asked to rate a particular flight experience with a rating included between 1 and 5, and this kind of record is also associated with some of the customers' specifics, such as the age, gender, the type of flight class purchased and the loyalty to the company.

The dataset is divided into:

a. Flight related variables:

- *Seat Comfort*: satisfaction level of seat comfort
- *Departure and Arrival Time Convenience:* satisfaction level of Departure/Arrival time convenience
- *Gate Location*: satisfaction level of Gate location
- *Baggage Handling*: satisfaction level of baggage handling
- *Check-in Service*: satisfaction level of Check-in service
- *Online Boarding*: satisfaction level of online boarding
- *Ease of Online Booking*: satisfaction level of online booking

b. Services related variables:

- *Food and Drink*: satisfaction level of Food and drink
- *Inflight Wi-Fi Service*: satisfaction level of the inflight Wi-Fi service
- *Inflight Entertainment*: satisfaction level of inflight entertainment
- *Inflight Service*: satisfaction level of inflight service
- *On-board Service*: satisfaction level of on-board service
- *Leg Room*: satisfaction level of leg room service
- *Cleanliness*: satisfaction level of cleanliness

c. Flight timing related variables:

- *Departure Delay*: minutes delayed when departure.
- *Arrival Delay*: minutes delayed when arrival.
- *Flight Distance*: The flight distance of the journey (expressed in kilometers)

d. Passenger's characteristics:

- *Gender*: gender of the passengers (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)

The whole point of this study is to show the impact of different inflight experiences on the perceived quality of the airline service, expressed through the satisfaction column, a binary variable that takes two values: satisfied or neutral-dissatisfied.

e. Satisfaction:

- *Satisfaction*: Airline satisfaction level (Satisfied, neutral or dissatisfied).

2.2 Preprocessing

2.2.1 Prefatory note

Let us first specify that, for the purpose of processing, ordinal attributes on a Likert scale have been assumed numeric to allow a more in-depth statistical analysis.

2.2.2 Data cleaning

We begin by noting that, on first observation, some ordinal columns assume integer values in [0, 5], and are worth 0 with the meaning of "not applicable": in such cases it is good to replace 0 with a missing value ("?").

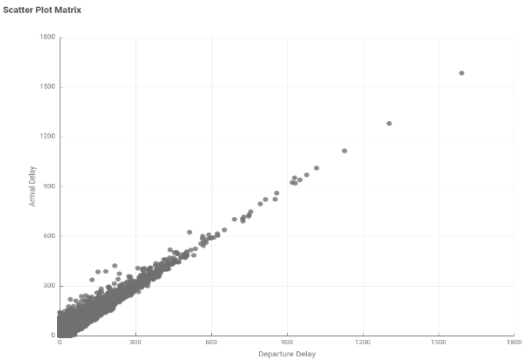
2.2.3 Data exploration and evaluation of outliers

An overview of the descriptive statistics shows that the missing values are mostly in these columns and, to a much lesser extent, in *Arrival Delay*; *Arrival Delay*, like *Departure Delay*, also has an extremely asymmetric distribution, concentrated near 0 with a long tail to the right.

In this situation, the evaluation of the outliers is unclear: the a priori elimination of the outer values of the box plots appears to be a drastic decision, as well as arbitrary; in fact, the scatter matrix of *Flight Distance*, *Arrival Delay*, and *Departure Delay* shows some continuity in the distributions and may justify the more cautious choice of not removing the extreme values.

The similarity between *Arrival Delay* and *Departure Delay*, already suspected and in any case to be expected, is confirmed by the very high correlation index (0.97) and a linear regression that is substantially flattened on the diagonal (as a side note, it should be borne in mind that even among some ordinal variables there are non-negligible correlation indices, which are explored in more detail later).

Figure 1: Arrival Delay and Departure Delay correlation



2.2.4 Numerical feature analysis

Having made these considerations, the ultimate removal of the *Arrival Delay* column is a solution that lends itself to the dual purpose of solving the problem of missing values and avoiding duplication of the same information—an issue, the latter, which can have an undesirable, even severe impact on classification models.

It is interesting, however, to keep track of the difference between *Arrival Delay* and *Departure Delay* in a new column *In-flight Delay*, which represents the minutes of delay accumulated during the trip, and which, unlike *Arrival Delay*, is independent of *Departure Delay*; for records with missing *Arrival Delay*, *In-flight Delay* was set equal to 0. By entirely similar reasoning, one can choose—and this is the case—to create an *Arrival Delay per Distance* variable calculated as the ratio of delay minutes to distance traveled.

2.2.5 Missing values

There remains the issue of missing values in the ordinal columns; the rows involved are about 8 percent of the dataset and have a significantly different distribution from the remaining 92 percent with respect to, for example, *Satisfaction* (a simple chi-square test and the corresponding cross tabulation is sufficient to verify this).

Indiscriminate removal of these rows, therefore, is inadvisable not only because it might result in a loss

of information, but also, and more importantly, it may distort the source data with the risk of affecting model learning.

A more neutral, preferable solution is to replace the missing values with the conditional mean with respect to *Satisfaction* (more accurate than mode and median, in this case), calculated separately between training set and test set.

2.2.6 Training and test sets

The training set and test set were obtained with a random partition of the dataset into 75% and 25% respectively.

This option was preferred over other more refined and wasteful options for two reasons: first, the size involved (about 130,000 rows) is such that absolute randomness is still guaranteed; second, iterative procedures such as cross validation do not allow for a smooth and thorough comparison between different models.

That being said, for sheer scrupulousness, the chosen models were also trained on seven sets iterated in a cross validation, to which no in-depth analysis was devoted and whose results are mentioned only briefly, for comparison purposes, in the conclusion section.

2.2.7 Principal component analysis

At this level, the dataset no longer contains missing values; it is worthwhile, then, to learn more about the ordinal attributes by exploiting their numerical qualities. A Principal Component Analysis, limited to such columns and applied possibly to the training set alone, serves us primarily to identify any relationships between features.

The study of PCA—an essential summary of which follows—is necessarily done on the basis of the correlation matrix, not the covariance matrix. Eight principal components suffice to explain 86% of the total variance of the original fourteen attributes. The first component is correlated with many variables but particularly with *In-flight Entertainment*, *Cleanliness*, *Seat Comfort* and *Food and Drink*; similarly, the second is correlated with the variables *Ease of Online Booking*, *Departure and Arrival Time Convenience*, *Gate Location,* and *In-flight Wi-Fi Service*.

We thus recognize a first group of features closely related to in-flight services and their quality, and a second more inherent to logistical conveniences. After that, beginning with the fourth, the components become less representative, either because they evidently refer to a single attribute, or because they do not clearly refer to any.

The last consideration, together with the difficulties involved in interpreting the components, invites retention of the starting columns.

Ultimately, for the training of the models, it was decided to use all the starting variables, except *Arrival Delay*, with the addition of *In-flight Delay* and *Arrival Delay for Distance*, for a total of twenty-four columns.

3 Classification Models

The classification models used are:

i. *multilayer perceptron;*
ii. *k-nearest neighbors;*
iii. *random forest;*
iv. *logistic regression*.

3.1 Multilayer perceptron

The multi-layer perceptron was set up with a single hidden layer to keep computational complexity and computation time contained.

The model was trained several times with a varying number *n* of neurons (6, 8, 10, 12) in order to identify the most accurate parameter. [*table 1*]. In terms of accuracy, there is a discrete gain from 6 to 8, substantial similarity between 8 and 10, and another improvement from 10 to 12. Other parameters (such as accuracy and specificity) are fluctuating,

confirming the impression that, above the threshold of 10, there is overfitting.

Table 1: Number of neurons

	***n* = 8**	***n* = 10**	***n* = 12**
Recall	0.886	0.896	0.903
Precision	0.908	0.902	0.923
Sensitivity	0.886	0.896	0.903
Specificity	0.932	0.926	0.942
F Measure	0.897	0.899	0.913
Accuracy	0.912	0.913	0.925

The area under the ROC curve is 0.977.

3.2 k-nearest neighbors (KNN)

Even if this classificator only use numeric columns and Euclidean distances (meaning that the columns of non-numeric type are not implemented in the classification model), the *k*-neighbors learner was still applied to the training set, with the number of 'nearest neighbors' spacing from 7 to 11 [*table 2*].

Considering the size of the dataset, this kind of classificator should be unfitted to the analysis (average computational time spacing from 10 to 20 minutes) but is interesting to see the results of the application of a lazy learner and compare them with the results of the eager learners applied on the same dataset. The area under the ROC curve is 0.971.

Table 2: Number of neighbors

	***n* = 9**	***n* = 7**	***n* = 11**
Recall	0.939	0.937	0.94
Precision	0.911	0.912	0.911
Sensitivity	0.939	0.937	0.94
Specificity	0.879	0.881	0.88
F Measure	0.925	0.924	0.925
Accuracy	0.913	0.913	0.914

3.3 Logistic regression

The results for the logistic regression, using the Iteratively reweighted least squares solver with termination conditions $\varepsilon = 10^{-5}$ and maximum number of epochs equal to 100, are presented in the *table 3.*

Table 3: Logistic regression's results

Recall	0.877
Precision	0.874
Sensitivity	0.877
Specificity	0.903
F Measure	0.875
Accuracy	0.892

The aera under the ROC curve is 0.957.

3.4 Random Forest

Random forest, with the splitting Information Gain Ratio criterion, gives these results computed in *table 4.*

Table 4: Random Forest results

Recall	**0.946**
Precision	0.972
Sensitivity	0.946
Specificity	0.979
F Measure	0.959
Accuracy	0.965

The area under the ROC curve is 0.995.

3.5 Feature importance

One aspect that, due to the nature of the data, is of great interest is the evaluation of the importance of features in the obtained models, particularly the random forest and logistic regression (a separate discussion deserves the neural network, which, due

to normalization and binarization requirements, was not comparable). For this purpose, the *Global Feature Importance* node of KNIME offers two types of methods: comparison of the obtained model with a surrogate chosen from three available ones, and the feature permutation method. It was finally chosen to use the second technique for both classifiers—both for generality of application and for the greater intuitiveness of the algorithm—and to measure performance by the area under the ROC curve (AUC).

With the default setting of permuting each feature only once—a choice motivated partly by execution time and partly by the size of the sets—, analysis of the two models took up to a quarter of an hour overall on an average computer. First, we note that the four most important attributes are the same for both classifiers (*Customer Type, In-flight Wi-Fi Service, Online Boarding and Type of Travel*) albeit in slightly different order.

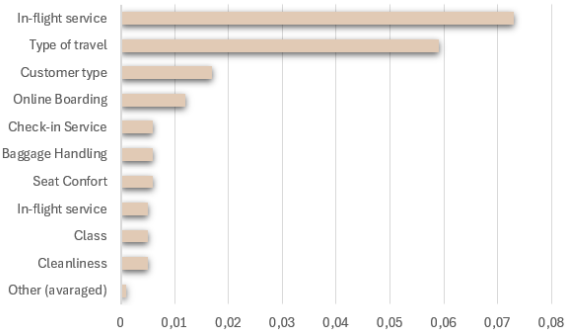
Of these, the second and third concern the client's opinions and therefore, from the airline's perspective, represent aspects on which it is possible, and perhaps advisable, to take proactive measures.

In the case of the random forest [*figure 2*], *In-flight Wi-Fi Service* and *Type of Travel* (first and second features in importance) are the true discriminating attributes; in the case of the logistic regression, it appears that *Type of Travel* is the most influential feature, the others ranking lower at comparable levels.

From the individual chi-square tests and their cross tabulations, moreover, it is inferred that those who travel for personal reasons ("Personal" *Type of Travel*) are 90 percent dissatisfied, as are, for 77 percent, new travelers to the company ("First-time" *Customer Type*).

The neural network model gives results in line with the previous ones, ranking (in order) *Type of Travel, In-flight Wi-Fi Service*, and, with some distance, *Customer Type*.

Figure 2 Global feature importance- Random Forest



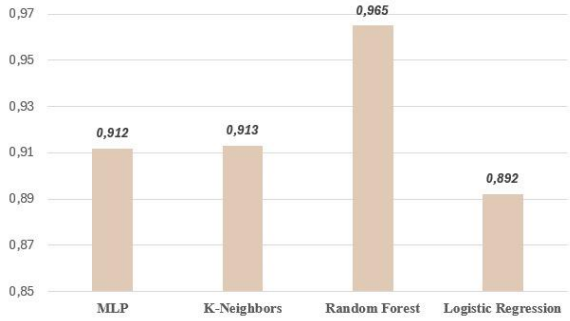
4 CONCLUSIONS

The results computed by the four model [*Figure 3*] trained on a random-sampling splitting training set are approximately identical to the result produced by the same models cross validated in K iterations, so is possible to affirm that:

i. The dataset splitting is balanced, meaning that every class is fairly represented in the models.

ii. The chosen models are robust and exhibit good generalization capability, achieving consistent performance across different data splits, regardless of the training technique used.

Figure 3: Accuracy obtained.



It is also important to notice that:

i. The *k*-neighbors classifier requires too much computational time if compared to other classifiers. The accuracy provided is equal or lower that the one computed by other, faster, models.

ii. The logistic regression classifier produces lower levels of precision and accuracy if applied to this specific dataset.

iii. The multilayer perceptron model requires high computational capacity due to its high demanding nature.

Considering a given degree of scalability and interpretability, we concluded that the best classification model, in terms of its capability to ensure the maximum level of accuracy, speed and robustness, is the Random Forest predictor, as it is the fastest and the more accurate.

From a strictly strategical/analytic point of view, is safe to assert that the 'Invistico' company should invest more into the quality of e-services it provides, *e.g.,* the quality of the Wi-Fi services or its online-booking platform, considering the impact that this kind of variables have on the overall satisfaction of the clients.

Furthermore, considering the influence of variables like *Customer Type, In-flight Wi-Fi Service, Online Boarding, and Type of Travel*, the company should implement targeted strategies to enhance these aspects. This may involve investing in cutting-edge technology for improved in-flight connectivity, optimizing online booking processes, and tailoring services based on the purpose of the passenger's travel.

It would also be useful for the company to better understand the underlying mechanisms that take into account the *type of travel / satisfaction* correlation, to create a better atmosphere inside and outside the gate.

It is essential for the 'Invistico' company to continually gather customer feedback and monitor evolving trends in passenger preferences. This proactive approach will enable the company to adapt its services to the never-ending changing in customer preferences and to maintain a competitive edge in the aviation industry.

In our opinion, considering the nature of the more significant satisfaction- impacting variables, this feedback mechanism could be implemented even more, adding to the customer satisfaction survey another variables, such as: *mobile app/website experience, aircraft cleanliness, post flight support, loyalty program perception, security and safety perception, special events and promotions quality.*