Big Data Fundamentals Project

Ranking customer satisfaction for an airline

Project made by:
David Larisa
Gâta Alexandra-Denisa

Contents

1. Introduction

Companies are always interested in increasing customer satisfaction. There are many factors that can influence consumers' perception of the products or services offered. Therefore, following marketing research on customer perception, it is important to be able to analyze which factors are relevant in creating the final opinion.

In this project, we will analyze the case of an airline company that conducted a customer satisfaction survey. Airline services include not only the transportation itself, but also a wide range of other services provided before and during the flight, with the aim of increasing customer satisfaction and differentiating themselves from the competition. It is therefore essential for them to be able to analyze the importance of the services they provide.

The dataset that has been used is the result of a questionnaire on the customers' experience during their last flight with the analyzed airline. The data collected was related to the various services and facilities offered. These will be detailed in the chapter "Data set".

The set of research questions for which we propose to conduct the analysis is:

1. Is there any link between customer data (Gender, Customer Type, Age, Flight distance) and satisfaction?
2. If so, how strong is the link?
3. Is it possible to estimate whether or not the customer will be satisfied, taking into account the customer's opinion of the airline's service?

Analyzing by customer data (Gender, Customer Type, Age, Flight distance) is relevant to avoid subjectivity given by certain customer segments. For example, people traveling long distances might be prone to have a negative opinion regardless of the rest of the services provided. This example is only a possible conjecture, which needs to be verified in order to create an accurate conclusion.

In the case of data regarding the consumer's opinion of the services offered by the airline, after analyzing the data, we want to check whether a prediction can be made as to whether the customer is satisfied or not. This prediction is useful in detecting customers who are prone to be dissatisfied, so that the company can intervene by various means to decrease the degree of dissatisfaction created. These could be offering discount coupons, additional services or other benefits to improve the customer experience.

Preventing consumer dissatisfaction is important to create a positive company image and prevent negative reviews. Thus, paying attention to this analysis can improve the customer experience and as a result, the company will have an increase in performance.

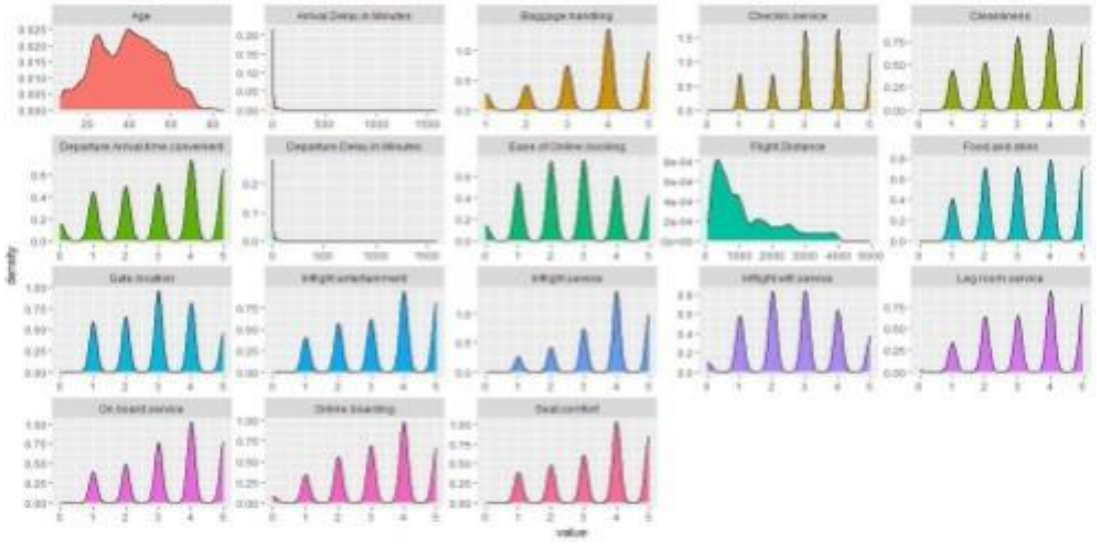
2. data set

The dataset used in this project was taken from https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction. It is a dataset collected from a customer survey of customers who have used the airline's services.

Attributes in the dataset are **Gender:** customer gender (Female, Male), **Customer Type:** customer type (Loyal customer, disloyal customer), **Age:** current age of the customer, **Type of Travel:** purpose of the flight (Personal Travel, Business Travel), **Class:** class flown (Business, Eco, Eco Plus), **Flight distance:** distance of the trip, **Inflight wifi service:** level of satisfaction with in-flight wi-fi service (0:Not applicable; 1-5), **Departure/Arrival time convenient:** level of satisfaction with departure/arrival time, **Ease of Online booking:** level of satisfaction with ease of online booking, **Gate location:** level of satisfaction with gate location, **Food and drink:** level of satisfaction with food and drink, **Online boarding:** level of satisfaction with online booking, **Seat comfort:** level of satisfaction with seat comfort, **Inflight entertainment:** level of satisfaction with in-flight entertainment, **On-board service:** level of satisfaction with on-board services, **Leg room service:** level of satisfaction with leg room, **Baggage handling:** level of satisfaction with baggage handling services, **Check-in service:** level of satisfaction with check-in services, **Inflight service:** level of satisfaction with in-flight services, Cleanliness**:** level of satisfaction with cleanliness, **Departure** Delay **in Minutes:** minutes of delay at take-off, **Arrival Delay in Minutes:** minutes of delay at landing. The target attribute is **satisfaction** which indicates the customer's class: satisfied or "neutral or dissatisfied".

The first step is to remove null data. In addition, attributes that had categories stored in string data (Gender, Customer Type, Type of travel, Class and satisfaction) have been transformed into factors.

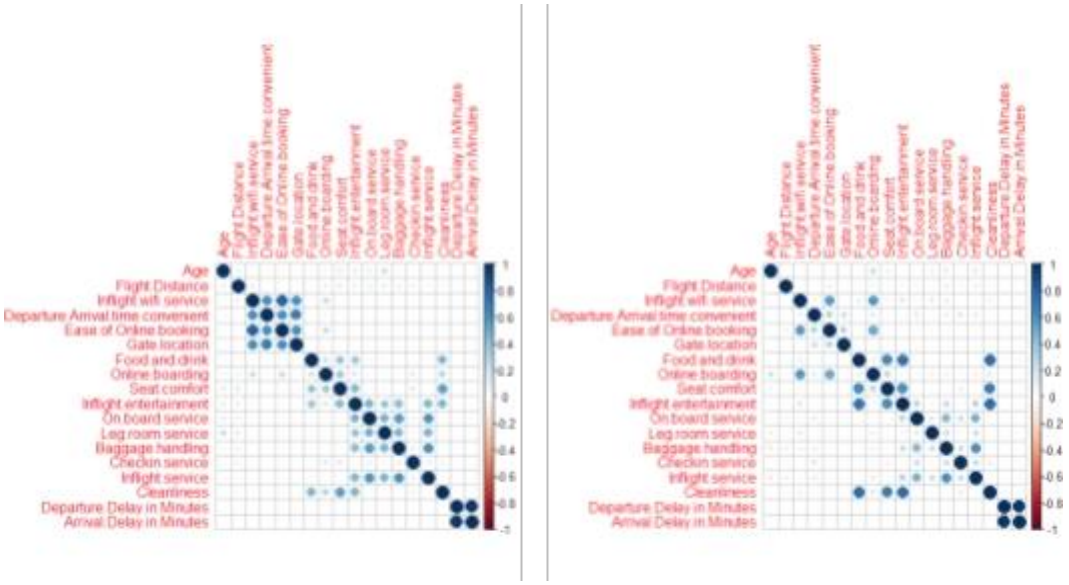
Due to the fact that the attributes were numerous and there was a risk of high correlation between them, which would lead to a decrease in the accuracy of the methods, we performed a cleaning process. The first step was to visualize the numerical data.



It can be seen that most of the data are data extracted from questionnaires with variants from 0 to 5 to rank the level of satisfaction with different services provided. This data will have to be transformed into factors in order to be correctly interpreted in realizing the prediction methods. The attributes to be transformed into factors are: Inflight wifi service, Departure/Arrival time

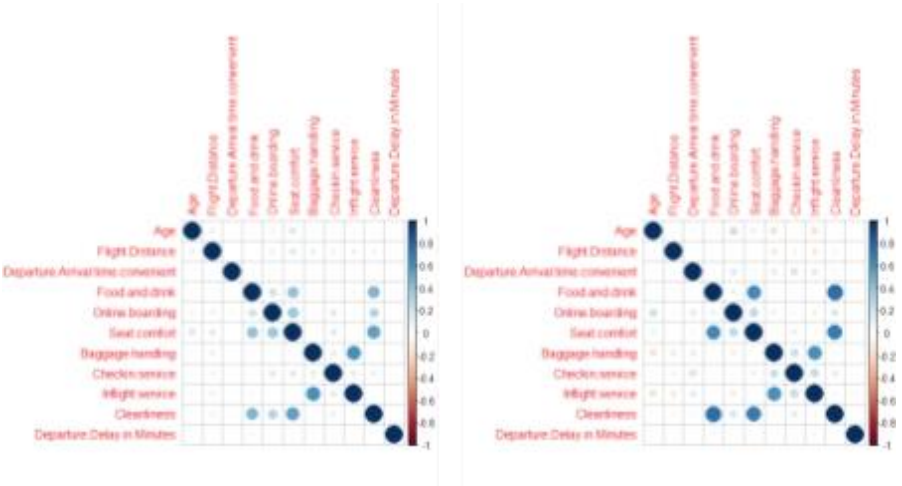
Convenience, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service and Cleanliness.

We also checked the correlation between the numerical attributes. The first representation is for customers in the "satisfied" class, and the second in the "neutral or dissatisfied" class.



Several correlations between attributes can be observed. Some attributes can also be observed that contain values related to the same services. For example, Leg room service is included in the attribute Seat comfort.

Taking into account previous observations, we removed a number of attributes: Arrival Delay in Minutes, Inflight wifi service, Ease of Online booking, Gate location, Inflight entertainment, On-board service and Leg room service. We again realized the correlation representation between the attributes. The first representation is for customers in the "satisfied" class and the second one is for customers in the "neutral or dissatisfied" class.



It can be seen that correlations between attributes have not been completely removed. We decided this because even though they appear to be correlated, the attributes Cleanliness, Food and drink and Seat comfort each refer to a different service. Thus, we decided to keep these values. The results of the cleaning will still be used for each method.

3. Results and discussions

3.1. Overview

The methods chosen are Naive Bayes, Logistic Regression and Decision Trees. These methods have been chosen in order to be able to make a classification prediction and to be able to compare the results obtained from each method. The comparison will be performed in order to be able to choose the method with the best accuracy, thus ensuring the quality of the results.

3.2. Naive Bayes

The first method used was Naive Bayes. Its implementation was performed on the airline's dataset which was cleaned as described in the "Dataset" chapter.

The prediction process starts by splitting the data into training data (70%) andtest data (30%). The test data will be used only once, to check the overfitting situation. The distribution over the two classes is:

Training dates		Test date	
neutral or dissatisfied	satisfied	neutral or dissatisfied	satisfied
41087	31427	17610	13470

Note that there is more data in the "neutral or dissatisfied" class.

To obtain higher accuracy using this dataset, we used the cross-validation method with k=5. The resulting model will have an accuracy of 84.64158% if kernel is not used and 84.93670% if kernel is used. The resulting confusion matrix (using kernel) is as follows:

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                    49.7       8.1
  satisfied                                   7.0      35.2

 Accuracy (average): 0.8494
```

The accuracy is good at 84.94%. There is a preference for class prediction
"The values for false positives (7.0%) and false negatives (8.1%) are low. Thus, the model created is good, but can be improved.

Next we created a new model on the Naive Bayes method, looking for an optimal combination of parameters. The parameters that will be adjusted in the search are: whether or not the kernel will be used, the Laplace factor will be set to 0.5, and the range of kernel fit is from 0 to 5, with step of 1. The most optimal model resulting was the one with kernel fit 4 and Laplace factor equal to 0.5. It resulted in accuracy of 85.25940% and the confusion matrix is as follows:

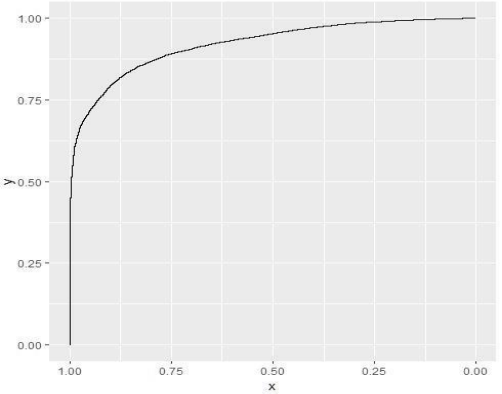
```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                    50.6       8.7
  satisfied                                   6.1      34.7
```

There is an increase in the prediction of the "neutral or dissatisfied" class, with 50.6% true negatives, but a decrease in true positives to 34.7%. The values for false positives decreased to 6.1% and false negatives increased to 8.7%. As it is more important to detect dissatisfied customers, the evolution of the values is favorable.

Next we created the prediction on the test set with the fitted model. The results after comparison with the actual values are as follows:

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                   15743      2711
  satisfied                                  1867     10759

               Accuracy : 0.8527
                 95% CI : (0.8487, 0.8566)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

            Sensitivity : 0.8940
            Specificity : 0.7987

       'Positive' Class : neutral or dissatisfied
```

The final accuracy is 85.27%, and the confidence interval is (0.8487, 0.8566). The P-value is small (< 2.2e-16), resulting in higher accuracy than NIR (No Information Rate) which was recorded at 56.66%. The sensitivity (0.8940) is higher than the specificity (0.7987), showing that the model can better identify customers in the "neutral or dissatisfied" category than the "satisfied" ones.



Finally, we generated the ROC curve, in order to analyze the graphical representation of the rate obtained for true positives (sensitivity), compared to the rate of false positives (specificity). The area of the ROC curve is large, showing the efficiency of the adjusted Naive Bayes method.

3.3. Back Logistics

The second method used was Logistic Regression. Its implementation was carried out on the airline's dataset which was cleaned as described in the "Dataset" chapter.

In realizing the first logistic regression model, all attributes were included. We decided this method in favor of analyzing each attribute individually, because in this way the attributes are numerous and we could analyze what are the relationships between them, avoiding confounding. The model created had the following result:

(Intercept)	-1.929e+01	2.225e+02	-0.087	0.93091
Seat.comfort1	-1.248e+00	2.052e+02	-0.006	0.99515
Seat.comfort2	-1.527e+00	2.052e+02	-0.007	0.99406
Seat.comfort3	-2.441e+00	2.052e+02	-0.012	0.99051
Seat.comfort4	-1.992e+00	2.052e+02	-0.010	0.99226
Seat.comfort5	-1.464e+00	2.052e+02	-0.007	0.99431
	Estimate	Std. Error	z value	Pr(>|z|)
Inflight.service1	1.179e+01	1.035e+02	0.114	0.90934
Inflight.service2	1.179e+01	1.035e+02	0.114	0.90928
Inflight.service3	1.164e+01	1.035e+02	0.112	0.91049
Inflight.service4	1.250e+01	1.035e+02	0.121	0.90390
Inflight.service5	1.303e+01	1.035e+02	0.126	0.89981
Cleanliness1	1.297e+01	5.768e+01	0.225	0.82204
Cleanliness2	1.310e+01	5.768e+01	0.227	0.82027
Cleanliness3	1.359e+01	5.768e+01	0.236	0.81367
Cleanliness4	1.350e+01	5.768e+01	0.234	0.81501
Cleanliness5	1.378e+01	5.768e+01	0.239	0.81118

Following these results, after analyzing the p-value of each attribute, the attributes will be removed: Seat comfort, Inflight service and Cleanliness. Their elimination was because the p-value had a high value, resulting in insignificant influence on the final model of the attributes. Next, another regression model was performed, but with only the remaining attributes. However, high p-values were recorded for Check-in services and Flight Distance:

	Estimate	Std. Error	z value	Pr(>|z|)
Checkin.service1	8.523e+00	7.246e+01	0.118	0.906368
Checkin.service2	8.647e+00	7.246e+01	0.119	0.905013
Checkin.service3	9.109e+00	7.246e+01	0.126	0.899970
Checkin.service4	9.061e+00	7.246e+01	0.125	0.900486
Checkin.service5	9.757e+00	7.246e+01	0.135	0.892895
Flight.Distance	1.351e-05	1.422e-05	0.950	0.342328

Thus, for the next model these attributes were also removed. In addition, in realizing the next model the data were divided into training data (70%) and test data (30%). The results on the training set are:

	Estimate	Std. Error	z value	Pr(>|z|)

```
(Intercept)                              2.8130069 0.3425544    8.212  < 2e-16 ***
GenderMale                               0.0909008 0.0248291    3.661 0.000251 ***
Customer.TypeLoyal Customer              2.3815849 0.0387665   61.434  < 2e-16 ***
Age                                     -0.0051836 0.0008991   -5.765 8.15e-09 ***
Type.of.TravelPersonal Travel           -3.4938383 0.0424709 -82.264  < 2e-16 ***
ClassEco                                -0.7231214 0.0297616 -24.297  < 2e-16 ***
ClassEco Plus                           -0.8993076 0.0505866 -17.778  < 2e-16 ***
```

```
Departure.Arrival.time.convenient4   -0.5278244   0.0581688   -9.074  < 2e-16 ***
Departure.Arrival.time.convenient5   -0.5352999   0.0599510   -8.929  < 2e-16 ***
Food.and.drink1                      -2.4076420   0.3331428   -7.227 4.94e-13 ***
Food.and.drink2                      -2.0030888   0.3325232   -6.024 1.70e-09 ***
Food.and.drink3                      -1.9886686   0.3324025   -5.983 2.19e-09 ***
Food.and.drink4                      -1.6948816   0.3324529   -5.098 3.43e-07 ***
Food.and.drink5                      -1.6763848   0.3328052   -5.037 4.73e-07 ***
Online.boarding1                     -3.4091621   0.0765379  -44.542  < 2e-16 ***
Online.boarding2                     -3.7148040   0.0730487  -50.854  < 2e-16 ***
Online.boarding3                     -3.7534944   0.0710751  -52.810  < 2e-16 ***
Online.boarding4                     -1.4853596   0.0667396  -22.256  < 2e-16 ***
Online.boarding5                      0.5346019   0.0703932    7.595 3.09e-14 ***
Baggage.handling2                    -0.1352036   0.0566528   -2.387 0.017008 *
Baggage.handling3                    -0.1959209   0.0528245   -3.709 0.000208 ***
Baggage.handling4                     1.1646320   0.0495057   23.525  < 2e-16 ***
Baggage.handling5                     2.0696802   0.0530220   39.034  < 2e-16 ***
Departure.Delay.in.Minutes           -0.0044881   0.0003328  -13.484  < 2e-16 ***
```

It can be seen that the p-values are small, resulting that the remaining attributes significantly affect the final prediction. Attributes influencing the decision towards the class "satisfied" are Gender (Male), Customer type (Loyal Customer), Online boarding (5) and Baggage handling (4 and 5), and the others influence the decision towards the "neutral or dissatisfied" class.

By setting the resulting prediction values to be divided into "neutral or dissatisfied" if below 0.5 and "satisfied" if above 0.5, the resulting matrix is:

```
       neutral or dissatisfied satisfied
FALSE                    15865      2098
TRUE                      1745     11372
```

The data show a better prediction of the class "neutral or dissatisfied" than "satisfied". Also the number of false positives and false negatives are low.

The following model was realized using cross-validation with all attributes. The resulting accuracy is 88.97105%. The confusion matrix still shows a better estimation of the class "neutral or dissatisfied" and a small number of values for false positives and false negatives.

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                    51.9       6.3
  satisfied                                   4.7      37.0
```

Following prediction on the test set, the values obtained are as follows:

```
               Accuracy: 0.8879
                 95% CI : (0.8844, 0.8914)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa: 0.7705

 Mcnemar's Test P-Value: < 2.2e-16

            Sensitivity: 0.9198
            Specificity : 0.8463

       'Positive' Class : neutral or dissatisfied
```

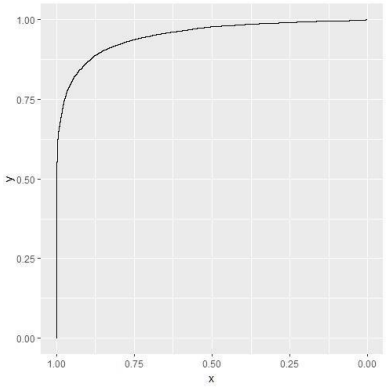
The final accuracy is 88.79%, and the confidence interval is (0.8844, 0.8914). The P-value is small (< 2.2e-16), resulting in a higher accuracy than NIR (No Information Rate) which was recorded at 56.66%. Sensitivity (0.9198) is higher than specificity (0.8463),

showing that the model can better identify "neutral or dissatisfied" customers than "satisfied" ones.

Next, we performed a model in which we excluded attributes that previous analysis assumed not to influence prediction. However, the results obtained were weaker:

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                   15865      2098
  satisfied                                  1745     11372

               Accuracy: 0.8764
                 95% CI : (0.8726, 0.88)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

             Sensitivity: 0.9009
             Specificity : 0.8442

       'Positive' Class : neutral or dissatisfied
```

In conclusion, the model chosen will be the one using cross-validation with all attributes, and its results have been detailed above. The ROC curve has been constructed on this model and is plotted:



3.4. Tree

The last method used was Decision Trees. Implementing it on the airline dataset that has been cleaned as described in the "Dataset" chapter. In addition, for this method we chose to transform the numerical variables from Age and Flight.Distance to intervals, to make it easier when splitting the tree on nodes, also making it easier to interpret.

The prediction process starts by splitting the data into training data (70%) andtest data (30%). The test data will be used only once, to check the overfitting situation. The distribution over the two classes is:

Training dates		Test date	
neutral or dissatisfied	satisfied	neutral or dissatisfied	satisfied
41087	31427	17610	13470

We continued with the creation of the first m1 tree. From this, it emerges that the most important variables are Online boarding, Seat comfort, Type of travel, Class, Cleanliness, Age group and Food and drink.

```
Online.boarding    Seat.comfort  Type.of.Travel           Class
             33              15              13              13
    Cleanliness       Age_Group  Food.and.drink
             10               9               7
```

On the basis of these variables, the tree was created, and then we created the confusion matrix.

```
                          Reference
      Prediction          neutral or                    satisfied
      dissatisfied neutral or dissatisfied                 3061
                          15845                           10409
      satisfied                                1765
```

From here we observe that we had 15845 instances of true negatives and 10409 instances of true positives. The instances of false negatives (3061) and false positives (1765) represent a small value compared to those correctly predicted. Next, we analyzed the resulting prediction statistics:

```
               Accuracy : 0.8447
                 95% CI : (0.8406, 0.8487)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa: 0.6802

   Mcnemar's Test P-Value < 2.2e-16
                        :

            Sensitivity : 0.8998
            Specificity : 0.7728
         Pos Pred Value : 0.8381
         Neg Pred Value : 0.8550
             Prevalence : 0.5666
         Detection Rate : 0.5098
   Detection Prevalence : 0.6083
      Balanced Accuracy : 0.8363

       'Positive' Class : neutral or dissatisfied
```

We have relatively good values, both in accuracy, sensitivity and specificity, but they could be improved. Therefore, we decided to also create an untrimmed tree to analyze if the difference in the results will be a major one. Thus, we have the tree m2, where we set the complexity parameter cp=0, so that our tree has the maximum number of nodes and splits, having a complex tree without pruning. The results obtained by it on the training set is one with good values, but it doesn't bring a major improvement, that's why we decided not to use this tree, because of the very large size it has, and the results were not favorable enough. But we decided to make a new tree m2_prunned , where we set the complexity parameter cp = 0.02. After analyzing the confusion matrix, we found that it has identical values to those obtained for m1. Therefore, the final variant for which we used as optimization parameter error , is the m1 tree.

Since we wanted to obtain even better values, we decided to build trees based on other optimization parameters, namely entropy and gini.

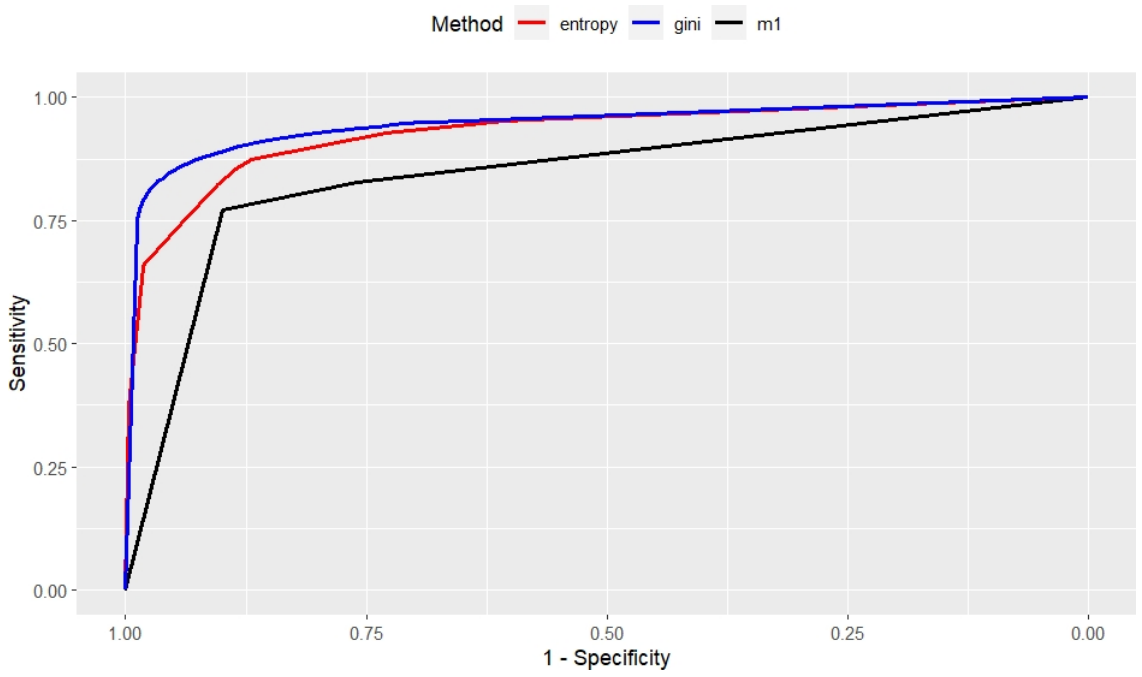
Confusion matrix results for the entropy parameter tree:

```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   15593      1945
  satisfied                                  2017     11525

               Accuracy : 0.8725
                 95% CI : (0.8688, 0.8762)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : <2e-16

                  Kappa: 0.7406

 Mcnemar's Test P-Value : 0.2593

            Sensitivity : 0.8855
            Specificity : 0.8556
         Pos Pred Value : 0.8891
         Neg Pred Value : 0.8511
             Prevalence : 0.5666
         Detection Rate : 0.5017
   Detection Prevalence : 0.5643
      Balanced Accuracy : 0.8705

       'Positive' Class : neutral or dissatisfied
```

Confusion matrix results for the gini parameter tree:

```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   16657      1922
  satisfied                                   953     11548

               Accuracy : 0.9075
                 95% CI : (0.9042, 0.9107)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa: 0.81

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9459
            Specificity : 0.8573
         Pos Pred Value : 0.8965
         Neg Pred Value : 0.9238
             Prevalence : 0.5666
         Detection Rate : 0.5359
   Detection Prevalence : 0.5978
      Balanced Accuracy : 0.9016

       'Positive' Class : neutral or dissatisfied
```

Following these results, we observed that the tree with the highest accuracy, a small p-value and also the best values for specificity and sensitivity, is the tree that uses gini as optimization parameter. We have also chosen to make a representation of the ROC curves, in order to see as clearly as possible which of the variants is the most optimal.



3.5. Optimal method

In order to determine the optimal method for the presented classification problem, we took into account the results obtained after analyzing the dataset by the methods presented above.

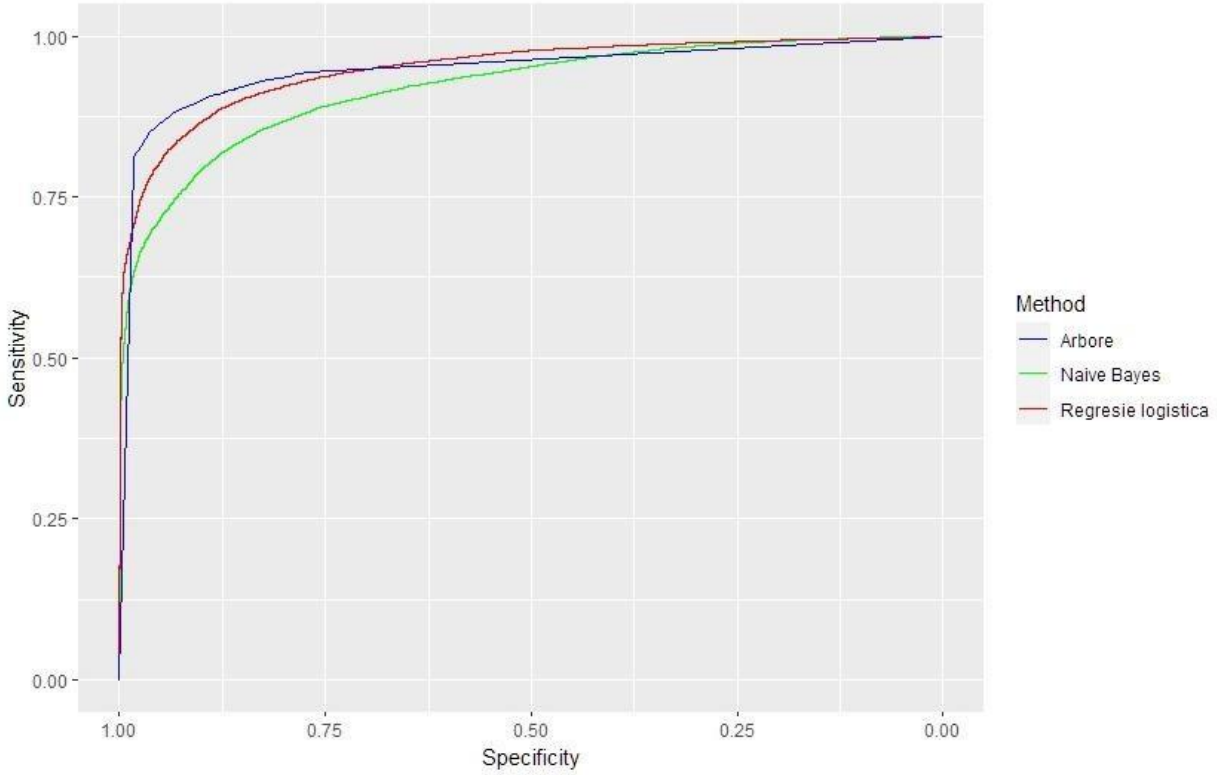
We performed a comparison between Naive Bayes, Logistic Regression and Gini parameter tree methods. In this comparison we used accuracy, sensitivity and specificity values. Thus, we could analyze which method is more efficient in estimating the correct class. In addition, we could observe how accurate the estimates are for each class.

All methods had a higher sensitivity than specificity, thus they provide a better estimate of the "neutral or dissatisfied" class. This result is favorable for the chosen set of research questions.

It can be observed that the best values are recorded for the Gini parameter tree method and the worst values are recorded for the Naive Bayes method.

	Naive-Bayes	**Logistic regression**	**Tree (with gini parameter)**
Accuracy	0,8527	0,8764	0,9075
Sensitivity	0,8940	0,9009	0,9459
Specificity	0,7987	0,8442	0,8573

Next, we compared the ROC curves obtained by each method. Following the graphical representation below, it can be seen again that the best results are obtained using the Gini parameter tree method.



In conclusion, the method chosen as the most optimal method in class estimation on the analyzed dataset is the Gini parameter tree method.

3.6. Limitations of analysis

The first limitations are given by the dataset analyzed. The number of records is limited, and therefore the cross-validation method had to be used. Because of this, some cases may not have been addressed. For example, in the Naive Bayes framework some pairs may not have existed, the remedy being provided by the Laplace factor.

Another limitation of this set is the attributes that were recorded. Some of them were similar and could confuse customers in a questionnaire, resulting in highly correlated attributes.

These limitations may lead to a decrease in the results obtained by the methods presented above.

4. Conclusion

Following the analysis presented above, the Gini parameter tree method will be used to answer the set of research questions.

For the first questions "Is there any relationship between customer data (Gender, Customer Type, Age, Flight distance) and satisfaction?" and "If yes, how strong is the link?", the importance of the attributes was analyzed. The most

Important attributes were in order: Online boarding, Seat comfort, Type of travel, Class, Cleanliness, Age group and Food and drink. Thus, most of the important attributes are related to the services provided. The exception is the attribute Age group which shows an influence of age group on satisfaction. However, it can be said that the most important aspects in determining customer satisfaction are service-related aspects, while personal data has less influence.

The last question "Is it possible to estimate whether or not the customer will be satisfied, taking into account the customer's opinion of the airline's service?" has a positive answer. It will be possible to estimate customer satisfaction on the basis of opinions about the services provided by the airline. The estimate will be accurate and it will be possible to detect particularly dissatisfied customers so that remedial actioncan be attempted.

In conclusion, the analysis yielded positive results in relation to the proposed research questions. Thus, the chosen method can provide a reliable prediction for managing customer satisfaction in airline.