ATILIM UNIVERSITY

SCHOOL OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

CMPE 468

MACHINE LEARNING FOR ENGINEERS

PROJECT REPORT

Submitted to	:	Assoc. Prof. Dr. Kasım Murat KARAKAYA		
Project Members	:	Mehmet Ege OĞUZMAN	–	160302019
		Ceyda KANBUROĞLU	–	160301025
		Gökçe Lal ARSLAN	–	17243610042
		Nazif Tugay ERBAĞI	–	19243710075

2020-2021

SPRING

TABLE OF CONTENTS

1. Introduction

The new trends of the global world have affected many sectors and these sectors can operate actively with a rapid adaptation process. Rapid changes in the transportation sector change the way people travel and prefer faster means of transportation in order to adapt to fast life. [2] Due to this advantage it provides, providing passenger satisfaction in airline transportation plays an important role as the number of competitors increases day by day. In the airline transportation sector, which can be defined as an abstract product, the product perception of the passengers is the combination of each transaction they make. Thus, satisfaction, defined as a feeling, pleasure, the degree of comfort, the distance between performance and service expectations, is a total assessment that depends on the consumption experience and the total goods and services purchased over time.

Satisfying passengers is more of a way to create a sustainable competitive advantage that allows airlines to retain customers, rather than being an option. If the passenger is not satisfied with the quality of the service provided, they will rethink for the next flights and most likely feel the need to go to another airline. In addition, it is another factor that concerns the economic profit of the companies and the studies on this subject are developing day by day because it is important.

It is important to determine the factors that affect the satisfaction of the passengers in order to provide a competitive advantage for the airline companies. Using the results of the customer satisfaction survey, the most basic features such as seat comfort, check-in service that provide satisfied customers can be predicted. In this project, it is aimed to examine the results of the survey conducted by US Airlines to measure the satisfaction of its passengers and classify them with the help of machine learning algorithms. These algorithms are KNN, Naive Bayes, logistic regression, trees and random forest. All results and reviews are in the section below.

2. Problem Definition and Algorithm

The task of machine learning is to find out how customers choose their seats, based on customer satisfaction surveys. If we can find out what customer choices are based on, we can improve those aspects.

2.1 Task Definition

According to the survey results, customer satisfaction rates vary according to some factors. These are age, flight distance, seat comfort, online boarding. The survey results give us the points that passengers pay attention to in their seat selection.

2.2 Selected ML Algorithm(s)

Based on the results obtained from the survey data, the main purpose should be classification. The subjects we classified were the seat choices of the passengers. With these results, the points that passengers pay attention to in their seat selection are determined. The data set is compatible with controlled classification methods.

Models are called classifiers, guess class tags that are categorical. It limits the machine learning methods that can be used. Therefore, the dataset is compatible with controlled binary classification methods. In this report, the following methods are utilized;

- Random Forest
- Decision Trees
- Logistic Regression
- Naive Bayes
- KNN

- **Logistic Regression:** Logistic regression is a statistical method used to analyze a data set with one or more independent variables that determine a result. It should not be confused with linear regression. The dependent variable is categorical. It is a predictive algorithm and is based on probability. The purpose of logistic regression is to find the most suitable model to describe the relationship between a set of independent variables related to its bidirectional characteristic. It is used to classify categorical or numerical data. It does not assume a linear relationship between the dependent variable and the independent variables, but the linear relationship between the log it of the explanatory variables and the response.

- **Random Forest:** Random Forest is an ensemble method. Which means it consists of individual decision trees. Random Forest makes predictions based on majority of votes from each of the decision trees made. Since we are working on multiple decision trees there is a chance of less over-fitting and it is the biggest advantage of random forest classifiers.

  Another plus for Random Forest classifiers is they are easy to understand and visualize by using tree diagrams.

- **K-Neirest Neigbors:** KNN is a distance-based supervised learning algorithm. To predict the value in a new data point, the algorithm finds the closest data point in the training data set. KNN is one of the simplest machine learning algorithms. KNN, lazy learning, is a non-parametric algorithm. Uses various classes of data to estimate the classification of the new sample point. but there are some disadvantages. on the contrary, as the value of k increased, the predictions became more stable, but we began to witness errors with increasing numbers after a certain point. We did not use this model because KNN could not make the predictions at the time it was supposed to, by significantly slowing down as the data volume increased. But there are faster algorithms that can produce more accurate classification and regression results.

- **Naive Bayes:** Naive models generally provide poorer performance than linear classifiers. It learns by looking at each feature individually, and each feature collects simple statistics per class. Naïve Bayes has a higher speed for lots of training. naive bayer is compatible with high dimensional datasets but slower than linear models in the training process. The assumption that all properties are independent is often not the case in real life. hence making the naive algorithm less accurate than complex algorithms (price of speed).
- **Decision Trees:** The purpose of the algorithm is to create a model that predicts the value of the target variable. Therefore, decision trees use tree representation to solve the following problems: a leaf node corresponds to a class tag and attributes are represented on internal nodes. Unlike other supervised learning algorithms, decision tree algorithms can also be used to solve regression and classification problems.

  Decision tree is a supervised learning technique that can be used for classification and regression problems, but it is a popular one for solving classification problems. It is a tree-structured classifier in which the internal nodes represent the attributes of the data set, the decision rules of the branches, and the leaf node the result. The decision button has two nodes: a node and a leaf node. While decision nodes can be used for any decision and have multiple branches, leaf nodes do not include the output of those decisions and others. Decisions or tests based on "a dataset" were performed. A tree-like structure that looks like a tree and expands more into a branch, starting from the root node, is called a decision tree. We use the CART algorithm (short for "Classification and Regression Tree" algorithm) to construct the tree. The decision tree asks only one and divides it into subtrees in its decision according to the answer (yes / no). The algorithm fits our model.

3. Summarize Data

In this project we obtained our data from Kaggle platform. This data set called Airline satisfaction survey and as the name suggest this dataset contain an airline passenger satisfaction survey. This dataset originally published by John.D but cleaned up by TJ Klein which we are using[1]. The last update on the data at 2020.

This data set has 23 features and there are 129880 samples in our dataset. Data is splitted by %25 and %75 for test and train respectively. List of our features as follows;

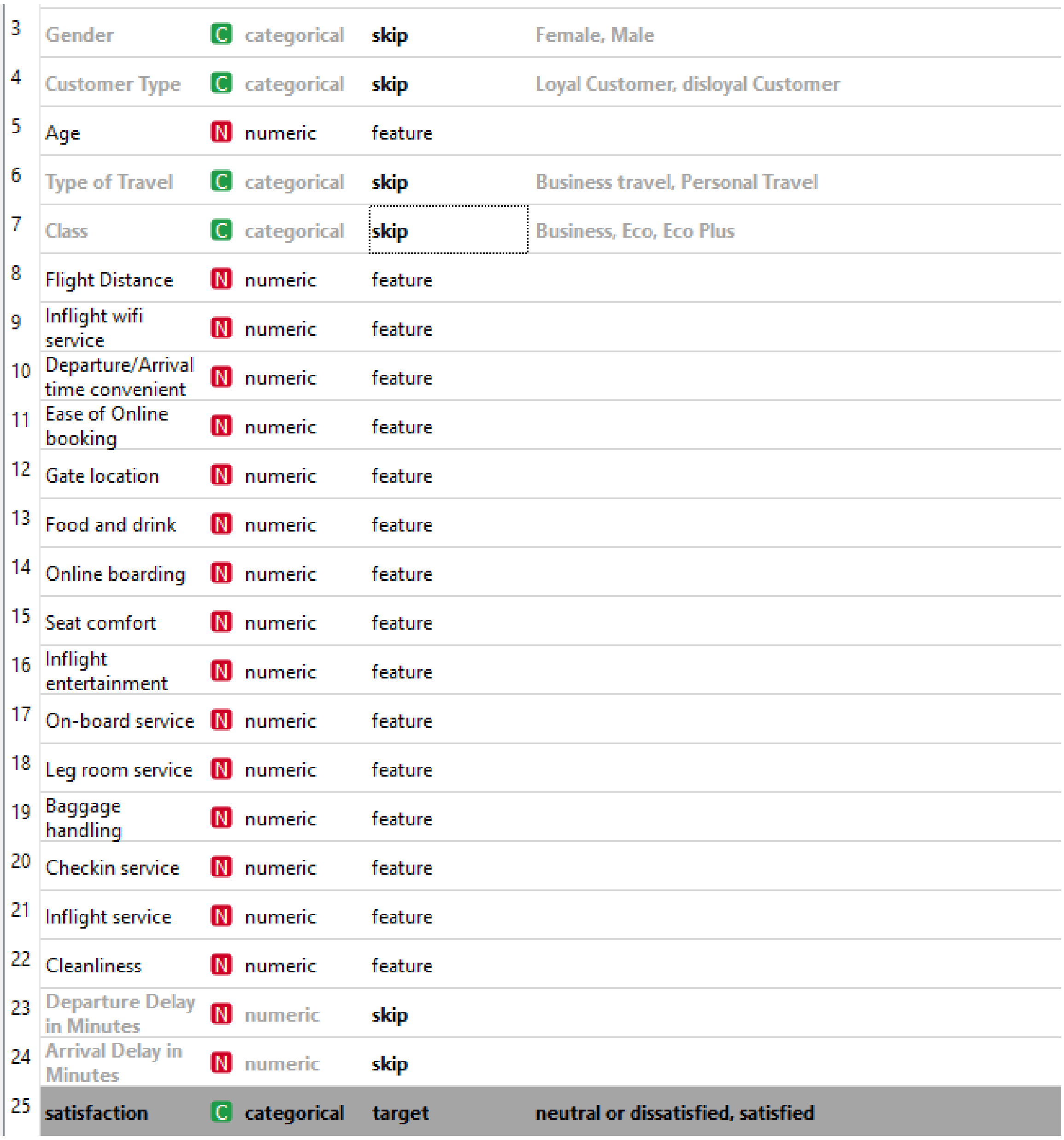
- Gender: Gender of passengers.
- Customer Type: The customer type(Loyal customer or Disloyal customer).
- Age: The actual age of the passengers.
- Type of Travel: Purpose of the flight of the passengers (Personal travel, Business Travel).
- Class: Travel class in the plane of the passengers (Business, Eco, Eco plus).
- Flight distance: The flight distance of this journey.
- In-flight wi-fi service: Satisfaction level of the in-flight wi-fi service.
- 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 drinks: Satisfaction level of Food and drinks.
- Online boarding: Satisfaction level of online boarding.
- 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.
- In-flight service: Satisfaction level of in-flight service.
- Cleanliness: Satisfaction level of Cleanliness.
- Departure delay in minutes: Minutes delayed when departure.
- Arrival delay in minutes: Minutes delayed when arrival.
- Satisfaction: Airline satisfaction level (Satisfied, neutral or dissatisfied).

Also this data set is ranked as 8.8 usable which tells us it is almost as clean as possible. But of course we needed to have some adjustments according to our needs or get rid of some missing values and etc. Our dataset is binary classification which means “1” or “0”. For neutral or dissatisfied it is 0 and for satisfied it is 1. For the numeric values. Besides from age, flight distance, departure delay in minutes and arrival delay in minutes all our features rank between 0 to 5. 0 is worst and 5 is best.

3.1 Descriptive Statistics

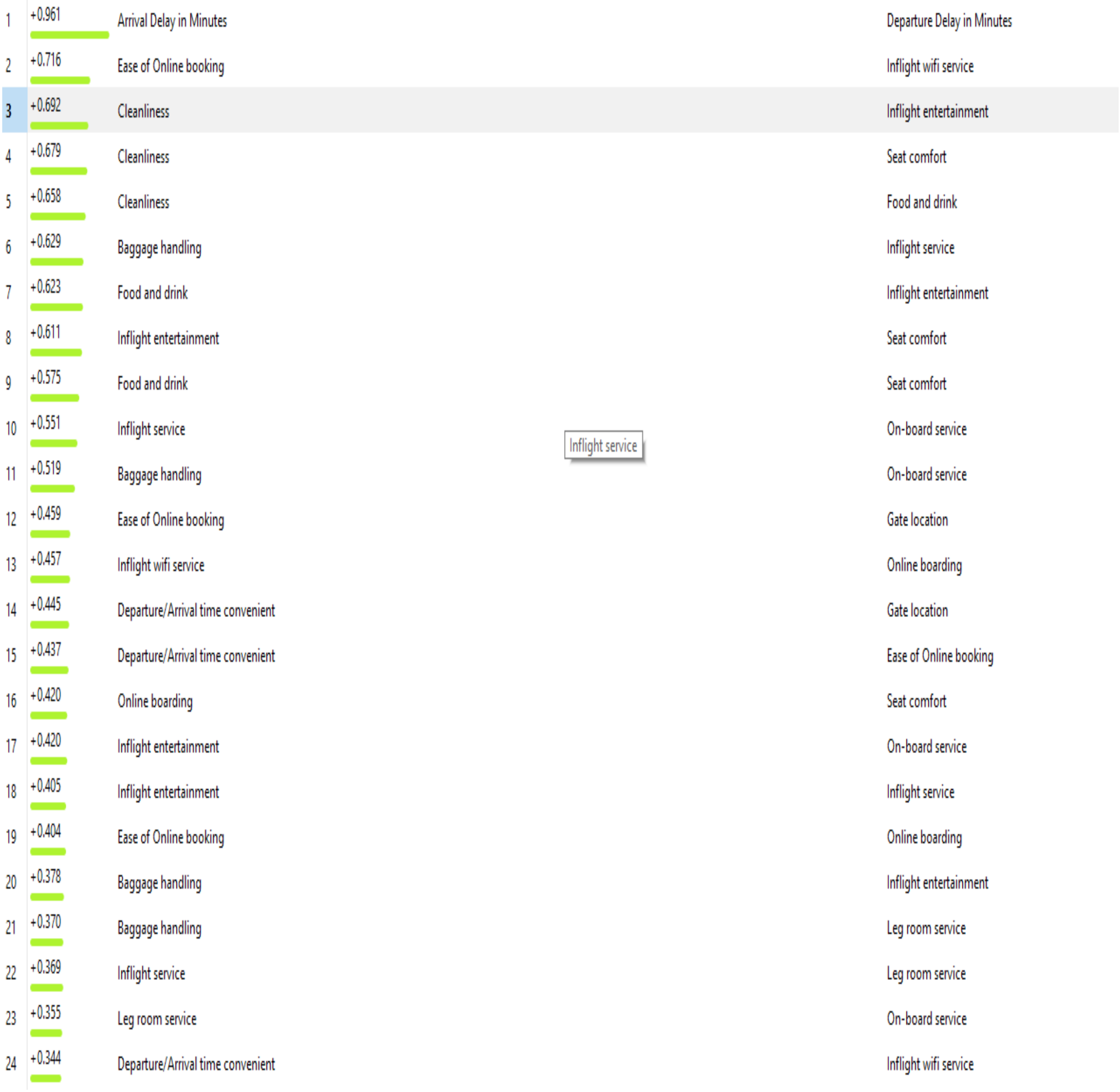
In this data set we have 5 categorical and 18 numerical features as shown in Figure 3.1.



3	Gender	C	categorical	skip	Female, Male
4	Customer Type	C	categorical	skip	Loyal Customer, disloyal Customer
5	Age	N	numeric	feature	
6	Type of Travel	C	categorical	skip	Business travel, Personal Travel
7	Class	C	categorical	skip	Business, Eco, Eco Plus
8	Flight Distance	N	numeric	feature	
9	Inflight wifi service	N	numeric	feature	
10	Departure/Arrival time convenient	N	numeric	feature	
11	Ease of Online booking	N	numeric	feature	
12	Gate location	N	numeric	feature	
13	Food and drink	N	numeric	feature	
14	Online boarding	N	numeric	feature	
15	Seat comfort	N	numeric	feature	
16	Inflight entertainment	N	numeric	feature	
17	On-board service	N	numeric	feature	
18	Leg room service	N	numeric	feature	
19	Baggage handling	N	numeric	feature	
20	Checkin service	N	numeric	feature	
21	Inflight service	N	numeric	feature	
22	Cleanliness	N	numeric	feature	
23	Departure Delay in Minutes	N	numeric	skip	
24	Arrival Delay in Minutes	N	numeric	skip	
25	satisfaction	C	categorical	target	neutral or dissatisfied, satisfied

Figure 3.1 Feature List

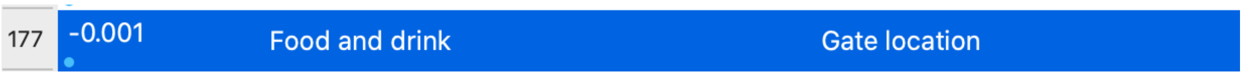
For the statistic part lets have a look at Correlations. Actually in our dataset highest correlation is between Arrival delay in minutes and Departure delay in minutes which is +0.961 it is close to +1 but still can be used but we eliminated one of them and second highest is +0.716 and it is usable. Here is some of correlations in Figure 3.2.



1	+0.961	Arrival Delay in Minutes	Departure Delay in Minutes
2	+0.716	Ease of Online booking	Inflight wifi service
3	+0.692	Cleanliness	Inflight entertainment
4	+0.679	Cleanliness	Seat comfort
5	+0.658	Cleanliness	Food and drink
6	+0.629	Baggage handling	Inflight service
7	+0.623	Food and drink	Inflight entertainment
8	+0.611	Inflight entertainment	Seat comfort
9	+0.575	Food and drink	Seat comfort
10	+0.551	Inflight service	On-board service
11	+0.519	Baggage handling	On-board service
12	+0.459	Ease of Online booking	Gate location
13	+0.457	Inflight wifi service	Online boarding
14	+0.445	Departure/Arrival time convenient	Gate location
15	+0.437	Departure/Arrival time convenient	Ease of Online booking
16	+0.420	Online boarding	Seat comfort
17	+0.420	Inflight entertainment	On-board service
18	+0.405	Inflight entertainment	Inflight service
19	+0.404	Ease of Online booking	Online boarding
20	+0.378	Baggage handling	Inflight entertainment
21	+0.370	Baggage handling	Leg room service
22	+0.369	Inflight service	Leg room service
23	+0.355	Leg room service	On-board service
24	+0.344	Departure/Arrival time convenient	Inflight wifi service

Figure 3.2 Correlation Table

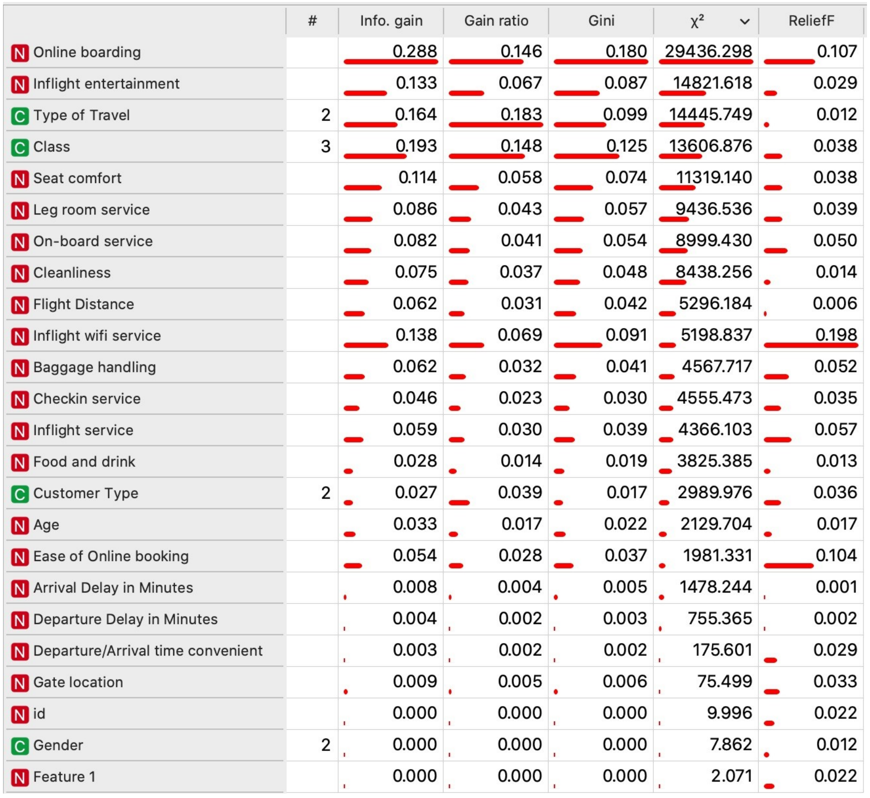
The lowest correlation in our dataset is between Food and drink and Gate location which is -0.001 this tells us they are completely different from each other. Here you can see in Figure 3.3.



177	-0.001	Food and drink	Gate location

Figure 3.3 Lowest Correlation

In this part I would like to show the rankings we have because it is one of the most important descriptive statistic we have. It basically gives us reliable information about our features and wich one is better according to algorithms we choose. Here in Figure 3.4.



	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF
N Online boarding		0.288	0.146	0.180	29436.298	0.107
N Inflight entertainment		0.133	0.067	0.087	14821.618	0.029
C Type of Travel	2	0.164	0.183	0.099	14445.749	0.012
C Class	3	0.193	0.148	0.125	13606.876	0.038
N Seat comfort		0.114	0.058	0.074	11319.140	0.038
N Leg room service		0.086	0.043	0.057	9436.536	0.039
N On-board service		0.082	0.041	0.054	8999.430	0.050
N Cleanliness		0.075	0.037	0.048	8438.256	0.014
N Flight Distance		0.062	0.031	0.042	5296.184	0.006
N Inflight wifi service		0.138	0.069	0.091	5198.837	0.198
N Baggage handling		0.062	0.032	0.041	4567.717	0.052
N Checkin service		0.046	0.023	0.030	4555.473	0.035
N Inflight service		0.059	0.030	0.039	4366.103	0.057
N Food and drink		0.028	0.014	0.019	3825.385	0.013
C Customer Type	2	0.027	0.039	0.017	2989.976	0.036
N Age		0.033	0.017	0.022	2129.704	0.017
N Ease of Online booking		0.054	0.028	0.037	1981.331	0.104
N Arrival Delay in Minutes		0.008	0.004	0.005	1478.244	0.001
N Departure Delay in Minutes		0.004	0.002	0.003	755.365	0.002
N Departure/Arrival time convenient		0.003	0.002	0.002	175.601	0.029
N Gate location		0.009	0.005	0.006	75.499	0.033
N id		0.000	0.000	0.000	9.996	0.022
C Gender	2	0.000	0.000	0.000	7.862	0.012
N Feature 1		0.000	0.000	0.000	2.071	0.022

Figure 3.4 Ranks

As you can see this list is ranked according to algorithms above but specifically ranked according to Chi square (X square). As you can see other then ReliefF bottom part features are useless and Top parts are very useful.

Now I would like to move to feature statistic part in our dataset. We had some missing value in Arrival delay in minutes and it is 310. The easiest way to see the missing values is from feature statistic and there we check there first. Actually there are some methods to overcome this problem but we did not need it because we just eliminated it. Here in Figure 3.5 you can see.

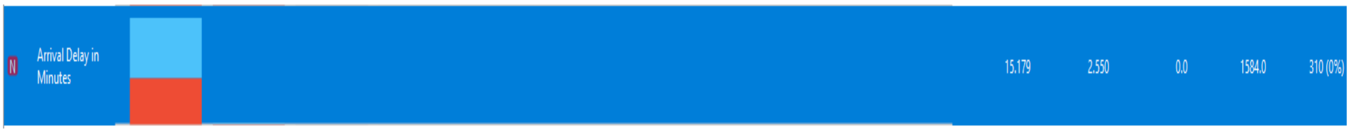


Figure 3.5 Missing Value

Also at the feature statistic we saw that minimum value for Flight distance is shown as minimum 31 mile but it seemed weird to us and we started researching on it. We found that in some situations this might happen and since the number of samples with this kind of value is negligible we did not touch it but here it is shown in Figure 3.6.

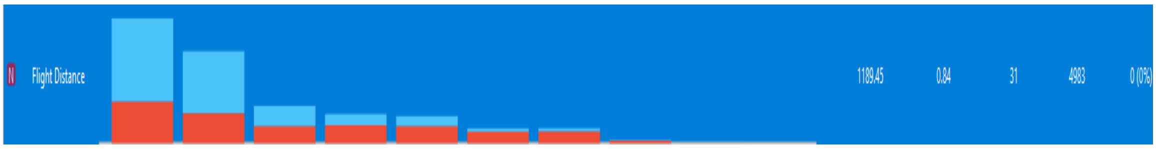


Figure 3.6 Min Value

For the most reliable features we have I would like to point ot some feature statistic graphs too because it can be clearly seen from here too. Like flight distance since the values differ from each other it means that it is giving us information. You can see in Figure 3.7.

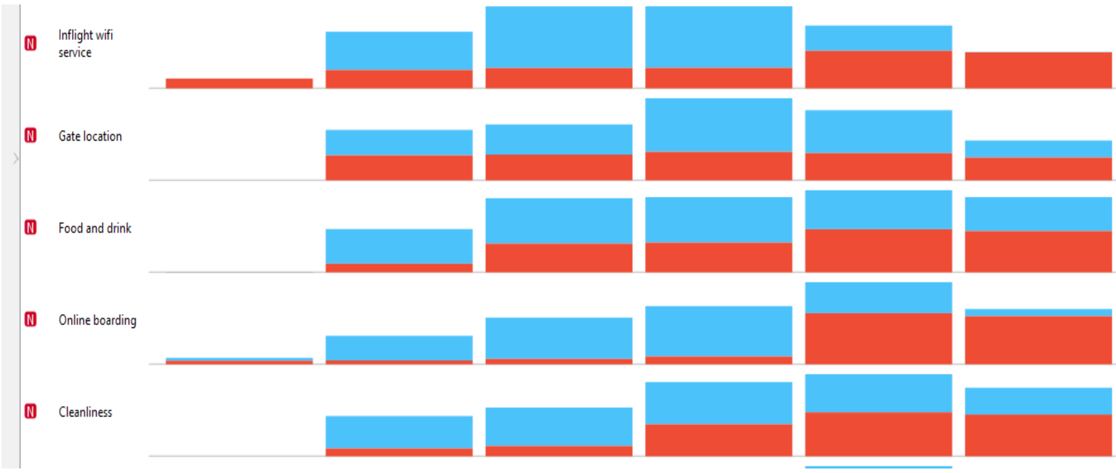


Figure 3.7 Feature Statistic

As you can see above for in flight wifi service although some people gave 4 to this part they are dissatisfied or although they give 0 they are satisfied. Which tells us that this feature is informative.

3.2 Data Visualization

In order to understand your data and see what is better or worst it is important to use visual aids to sort some things out or even to understand what is going on. For example we know that if correlation is low numbers should de distributed as evenly as possible and we know that if correlation is high they are gathered in one spot. There for I would like to start with high correlation scatter plot here in Figure 3.8.

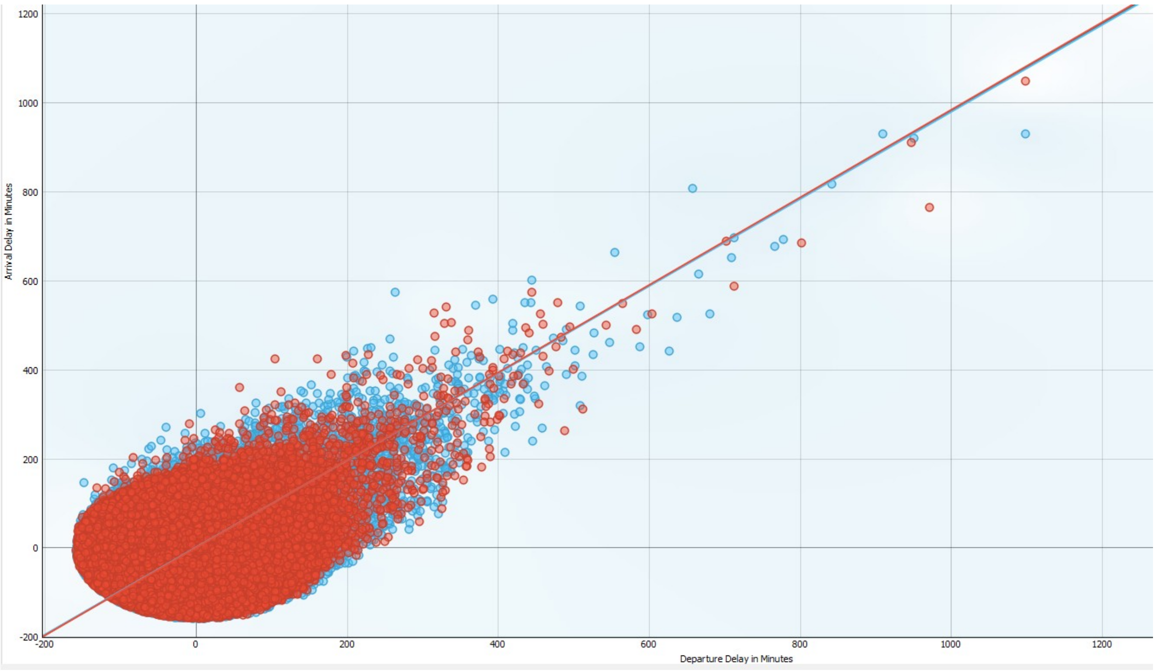
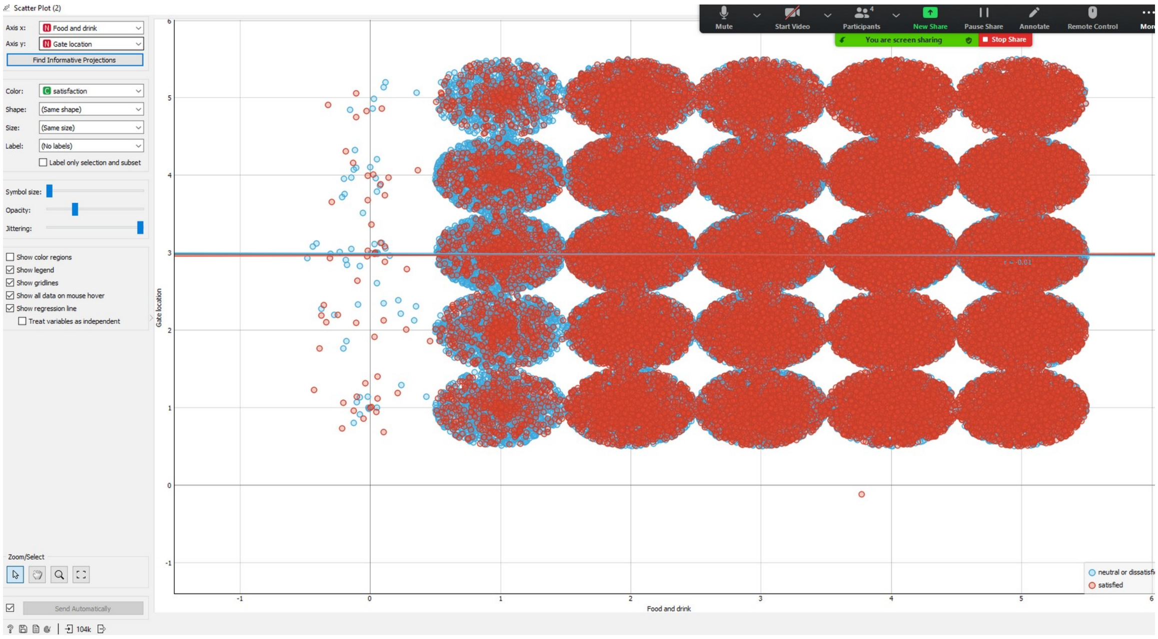


Figure 3.8 Arrival Delay in Minutes & Departure Delay in Minutes

On the Figure 3.8 you are seeing the scatter plot representation of Arrival delay in minutes and Departure delay in minutes. As you can see all the data is gathered on a place and regression line is linear also without even knowing their correlation we can easily say that these features are highly correlated with each other. Now lets look at low correlation example here in Figure 3.9.



Scatter Plot (2)

Axis x: Food and drink

Axis y: Gate location

Color: satisfaction

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

☐ Label only selection and subset

☐ Show color regions

☑ Show legend

☑ Show gridlines

☑ Show all data on mouse hover

☑ Show regression line

☐ Treat variables as independent

☑ Send Automatically

Figure 3.9 Food and Drink & Gate Location

On the Figure 3.9 you are seeing the scatter plot representation of Food and drink and Gate location. As you can see all the data mostly scattered evenly and the regression line is cutting the data in half. With these information in hand we can easily say that they are not correlated with each other. Also it tells us that we can use both of them to get some information according to our survey.

For lastly I would like to talk about Tree view tool. It is very useful when it comes to sort out and understand your data. Especially in our case it helped us highly here in Figure 3.10 you can see it.

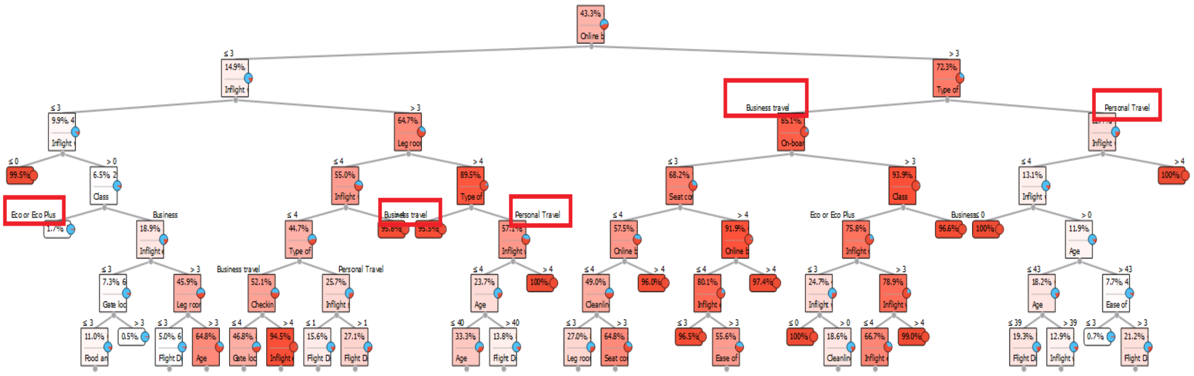


Figure 3.10 Tree View

The ones circled with red are other classes such as Type of travel or etc. and once we saw this representation we decided that we only want to know the satisfaction of the customer regardless of their gender or class or etc. So we skipped those features. Also we can clearly see what makes them more satisfied or dissatisfied from this representation because Dark reds means highly satisfied and whites are highly dissatisfied or neutral.

This analysis can be expanded but explaining all the scatter plots or other visual aids are not really necessary for our case therefore only the necessary parts are explained in this area.

4. Prepare Data

Most machine learning algorithms need data to be formatted in a very particular way, but datasets typically need any amount of planning until they will produce valuable insights. Some datasets have missing, invalid or otherwise difficult values for a processing algorithm. If data is missing, the algorithm can’t use it. If data is invalid, the algorithm produces less accurate or even misleading outcomes. Some datasets are relatively clean, but must be shaped and many datasets lack a useful business context, so enrichment is necessary. Good data preparation produces clean and well-cured data, leading to practical, precise model results.

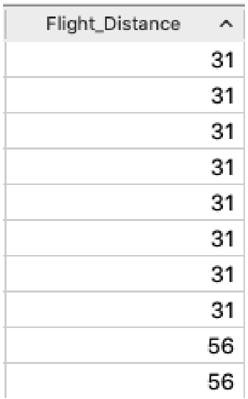
Initially, our raw data set contained 25 features. Since the problem is tried to solve by using supervised machine learning techniques, structures of dataset and features become crucial. So that reason, feature engineering and data pre-processing techniques are applied on the dataset.

First of all, missing values are detected to be removed. After that, feature engineering is applied on dataset. In that part some features are encoded to increase the comprehensibleness and to do some calculations easier with machine learning algorithms. Lastly, feature selection process has been done according to some analysis, such as, correlations, feature ranks and the relevance of features.

4.1 Data Cleaning

A clean dataset is sufficient to produce sensitive results. Even if we try to create a model on a dataset, we can exponentially boost the results by inspecting and cleaning our files. Feeding an unnecessary or misleading model will decrease the accuracy of the model. A smarter dataset gives us a better result than a fantastic one. A clean dataset will also facilitate the potential work with those in your enterprise.

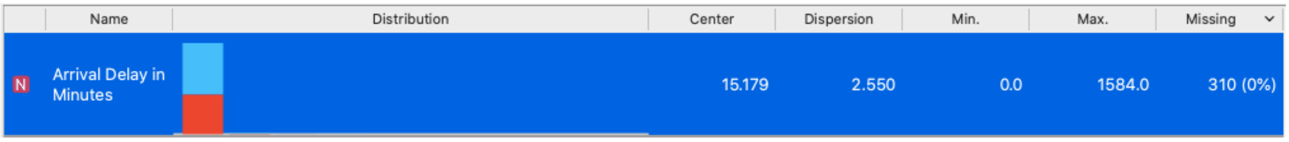
The dataset worked on did not require much data cleaning. There is no incorrect or irrelevant value in the dataset. In fact, “Flight Distance” feature has some values less than 50 miles and it is considered as an incorrect data at first. Then they are transformed to new random values which are more than 100 miles but it did not effect on our results. Also, according to the researches that has been done, there are some flights with less than 50 mile flight distance. Moreover, some values for “Flight Distance” feature are less than 50 miles in the survey results because of the technical problems of the flight. So, the decision made was based on the fact that “Flight Distance” data were not changed.



Flight_Distance ^
31
31
31
31
31
31
31
31
56
56

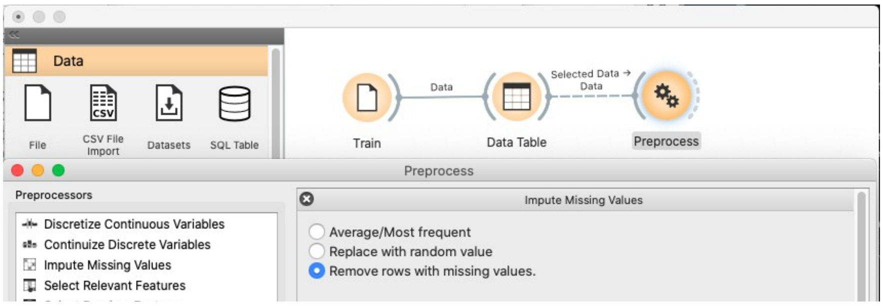
Figure 4.1 – Flight Distance Sample Values

The only feature that has some missing values is “Arrival Delay in Minutes”. It had 310 missing values to be removed. By using Orange 3 pre-processing module, the entire line with the missing value has been deleted.



	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N	Arrival Delay in Minutes		15.179	2.550	0.0	1584.0	310 (0%)

Figure 4.2 – Missing Value Rate Distribution



Preprocess

Preprocessors

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features

Impute Missing Values

○ Average/Most frequent
○ Replace with random value
● Remove rows with missing values.

Figure 4.3 – Impute Missing Value

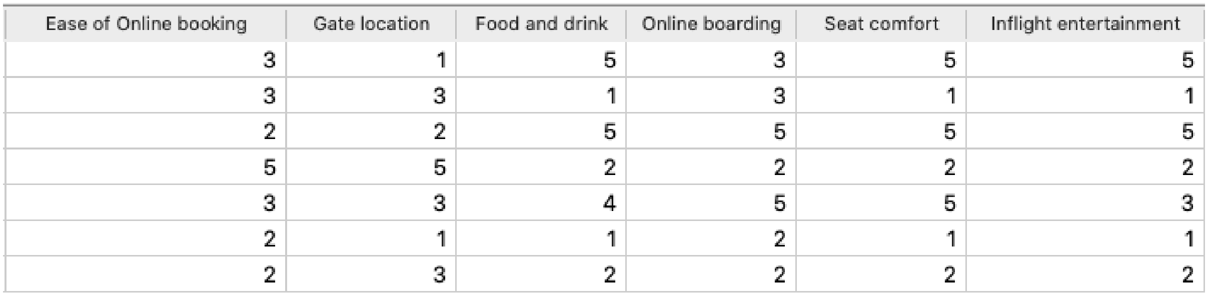
4.2 Feature Engineering

Feature engineering is also one of the most important parts. If properly carried out, the predictive ability of machine learning algorithms is increased by generating features from raw data to promote machine learning.

Basically, all algorithms used to generate outputs are using those input data. This data has elements that are typically ordered columns. Algorithms require sufficient functionality for such special functions.

Our dataset contains some features that present ratings for services of airline companies. For example, passengers rated the inflight entertainment service according to their satisfaction level. They gave a score for these services between zero and five. Zero means not applicable, 1 is very dissatisfied, and 5 is very satisfied. To work on such categorical data better, they have been encoded. Since these features are ordinal, integer encoding or ordinal encoding is applied instead of one-hot encoding. Because it is clear that there is an ordinal relationship between these values. Also, one-hot encoding would increase the number of independent variables and it is the thing that we don't want to. As a result, our categorical and ordinal features are encoded by the integer encoding method.

Also, no binning method was used to for numeric features because we have already knew that these features are not useful for the problem. Such as “Age”, “Arrival Delay in Minute” or “Departure Delay in Minute”.



Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment
3	1	5	3	5	5
3	3	1	3	1	1
2	2	5	5	5	5
5	5	2	2	2	2
3	3	4	5	5	3
2	1	1	2	1	1
2	3	2	2	2	2

Figure 4.4 – Encoded Feature Examples

4.3 Feature Selection

Feature selection helps us to control the complexity of the model and it reduces the chance of overfitting. Also that process increases the model’s interpretability by revealing the most informative factors driving the model’s outcomes. So that reason, we analysed feature pairs, feature ranks with different kind of parameters and correlation scores to select most fruitful features to create the machine learning model with high test accuracy without overfitting problem.

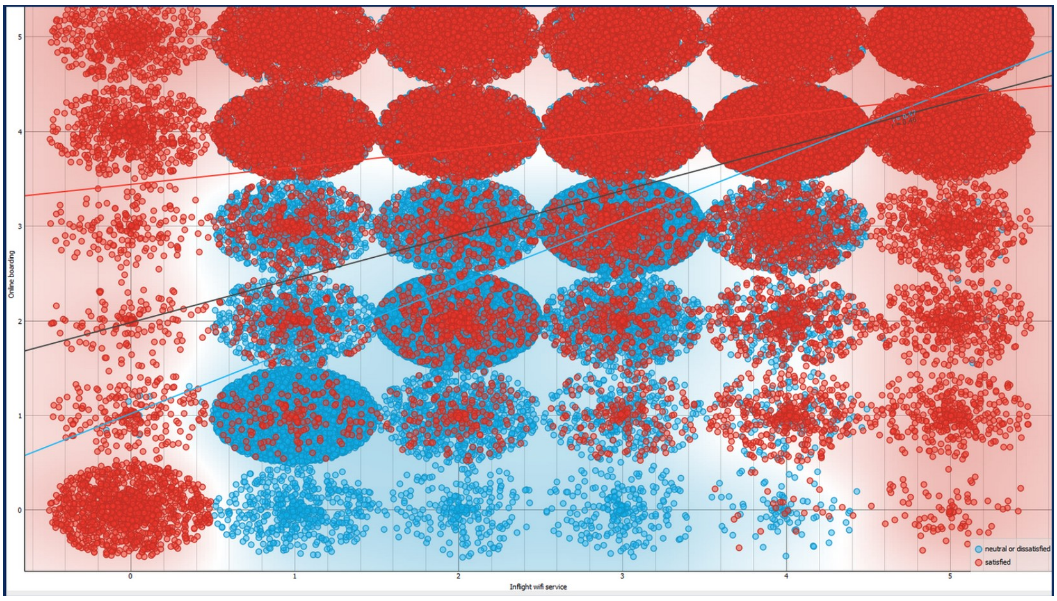
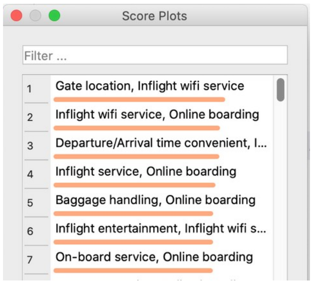


Figure 4.5 – Feature Pair Example

These are the second most related features that is calculated by Orange3 and they are namely; inflight wifi service and online boarding. As you can see, this pair gives us remarkable information. As we expected, from 1 to 3 rating causes dissatisfaction and the classes that are close to five are satisfied in general. When the answers are not applicable or even 0, satisfaction level is still high. So these graphs are giving us some clues but we also have a lot of graph to analyze to solve this problem.



1 Gate location, Inflight wifi service

2 Inflight wifi service, Online boarding

3 Departure/Arrival time convenient, I...

4 Inflight service, Online boarding

5 Baggage handling, Online boarding

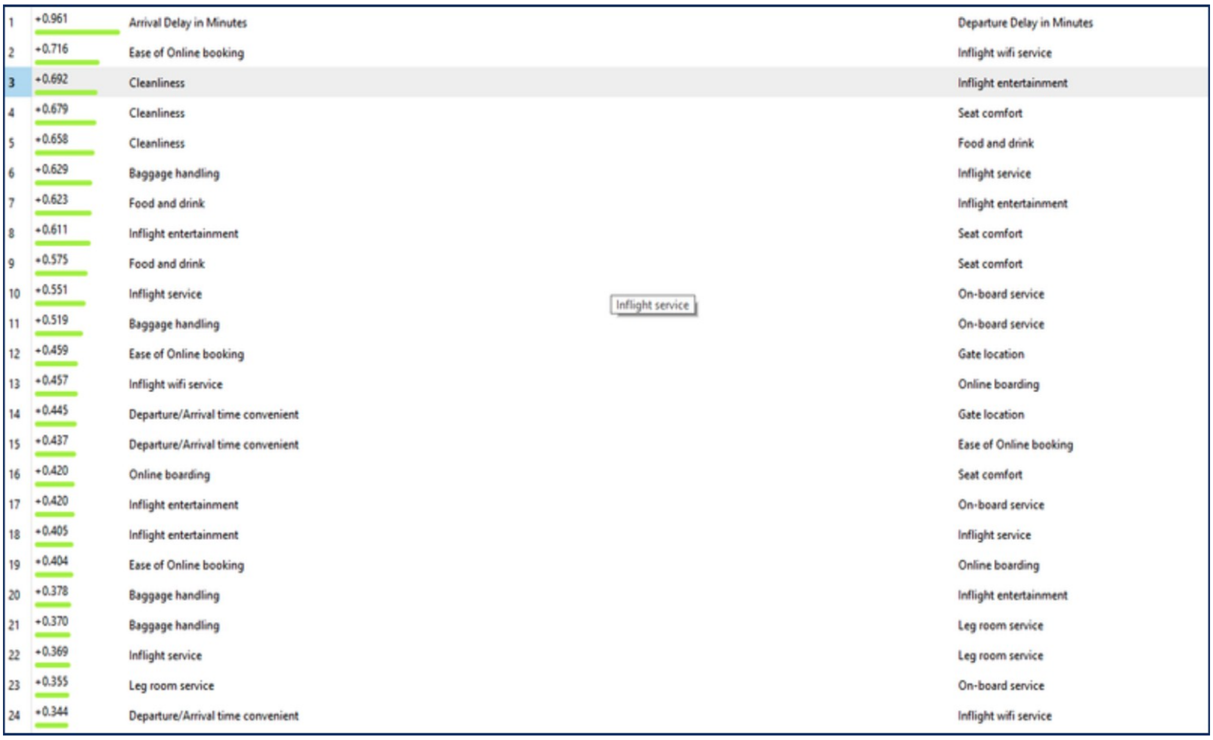
6 Inflight entertainment, Inflight wifi s...

7 On-board service, Online boarding

Figure 4.6 – Feature Pair Ranks

4.3.1 Correlation Detection

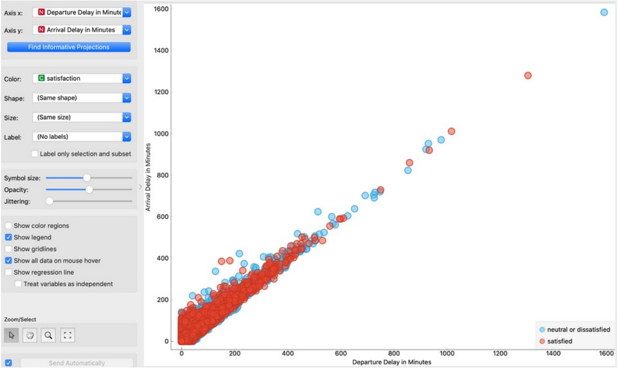
We also checked correlations scores by using pearson method for every possible feature pair. As a result, we get both possitive and negative correlations. We tried to find closest scores to plus 1 or minus one. Since, “Arrival Delay in Minutes” and “Departure Delay in Minutes” get highest possitive correlation score with zero point 9, we decided to get rid of one of them. But the other pairs are okay for us. Hence, we are not discarding any other feature.



1	+0.961	Arrival Delay in Minutes	Departure Delay in Minutes
2	+0.716	Ease of Online booking	Inflight wifi service
3	+0.692	Cleanliness	Inflight entertainment
4	+0.679	Cleanliness	Seat comfort
5	+0.658	Cleanliness	Food and drink
6	+0.629	Baggage handling	Inflight service
7	+0.623	Food and drink	Inflight entertainment
8	+0.611	Inflight entertainment	Seat comfort
9	+0.575	Food and drink	Seat comfort
10	+0.551	Inflight service	On-board service
11	+0.519	Baggage handling	On-board service
12	+0.459	Ease of Online booking	Gate location
13	+0.457	Inflight wifi service	Online boarding
14	+0.445	Departure/Arrival time convenient	Gate location
15	+0.437	Departure/Arrival time convenient	Ease of Online booking
16	+0.420	Online boarding	Seat comfort
17	+0.420	Inflight entertainment	On-board service
18	+0.405	Inflight entertainment	Inflight service
19	+0.404	Ease of Online booking	Online boarding
20	+0.378	Baggage handling	Inflight entertainment
21	+0.370	Baggage handling	Leg room service
22	+0.369	Inflight service	Leg room service
23	+0.355	Leg room service	On-board service
24	+0.344	Departure/Arrival time convenient	Inflight wifi service

Figure 4.7 – Correlation Scores

As you can see, it is clear that there is a correlation problem between these features. Also when we analyze them on scatter plot, we can see that they are linearly distrubeted. Moreover color regions are not distingusihed. So both of them give us same information. Thus we can drop one of them.



Axis x: N Departure Delay in Minute
Axis y: N Arrival Delay in Minutes

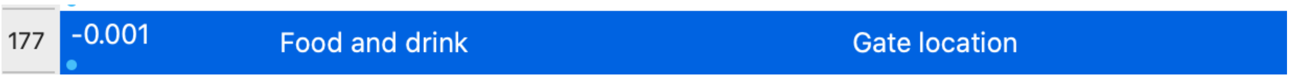
Color: C satisfaction
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
□ Label only selection and subset

□ Show color regions
☑ Show legend
□ Show gridlines
☑ Show all data on mouse hover
□ Show regression line
□ Treat variables as independent

☑ Send Automatically

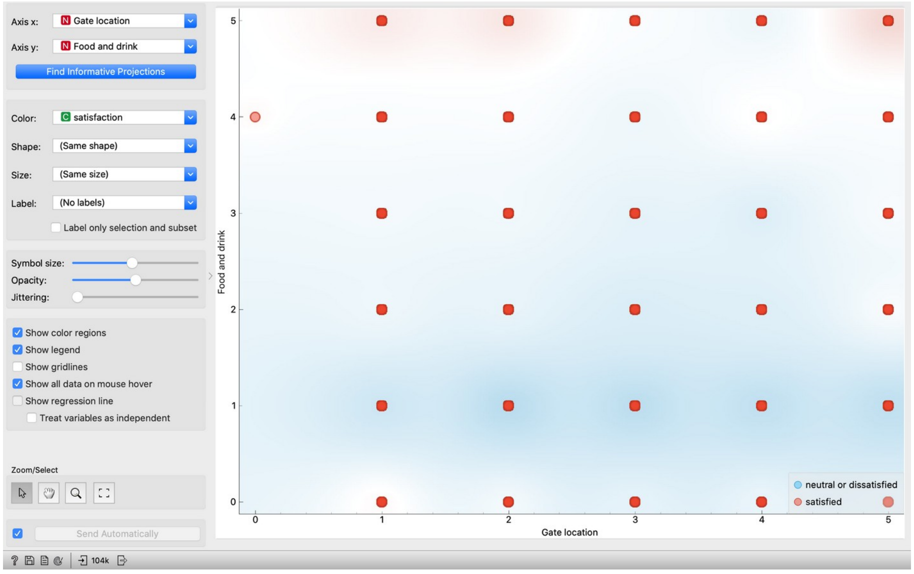
Figure 4.8 – Correlated Pair Scatter Plot

Also we analyzed you one of the non-correlated pair to make comparison easier. It is clear that they give us different kind of information. Altough passangers very satisfied with gate location, the region is still blue and target value is dissatisfied. Because they are dissatisfied with food and drink service. So we should the check other pairs to interpret gate locations service. They do not give us same information.



177 -0.001 Food and drink Gate location

Figure 4.9 – Non-Correlated Pair



Axis x: Gate location

Axis y: Food and drink

Color: satisfaction

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

☑ Show color regions

☑ Show legend

□ Show gridlines

☑ Show all data on mouse hover

□ Show regression line

□ Treat variables as independent

Figure 4.10 – Non-Correlated Pair Scatter Plot

4.3.2 Feature Ranks

Feature ranking is the most critical part in the feature engineering processes. Features are ranked by using different kind of methods that are provided by Orange 3. For that project, these feature ranking methods are used;

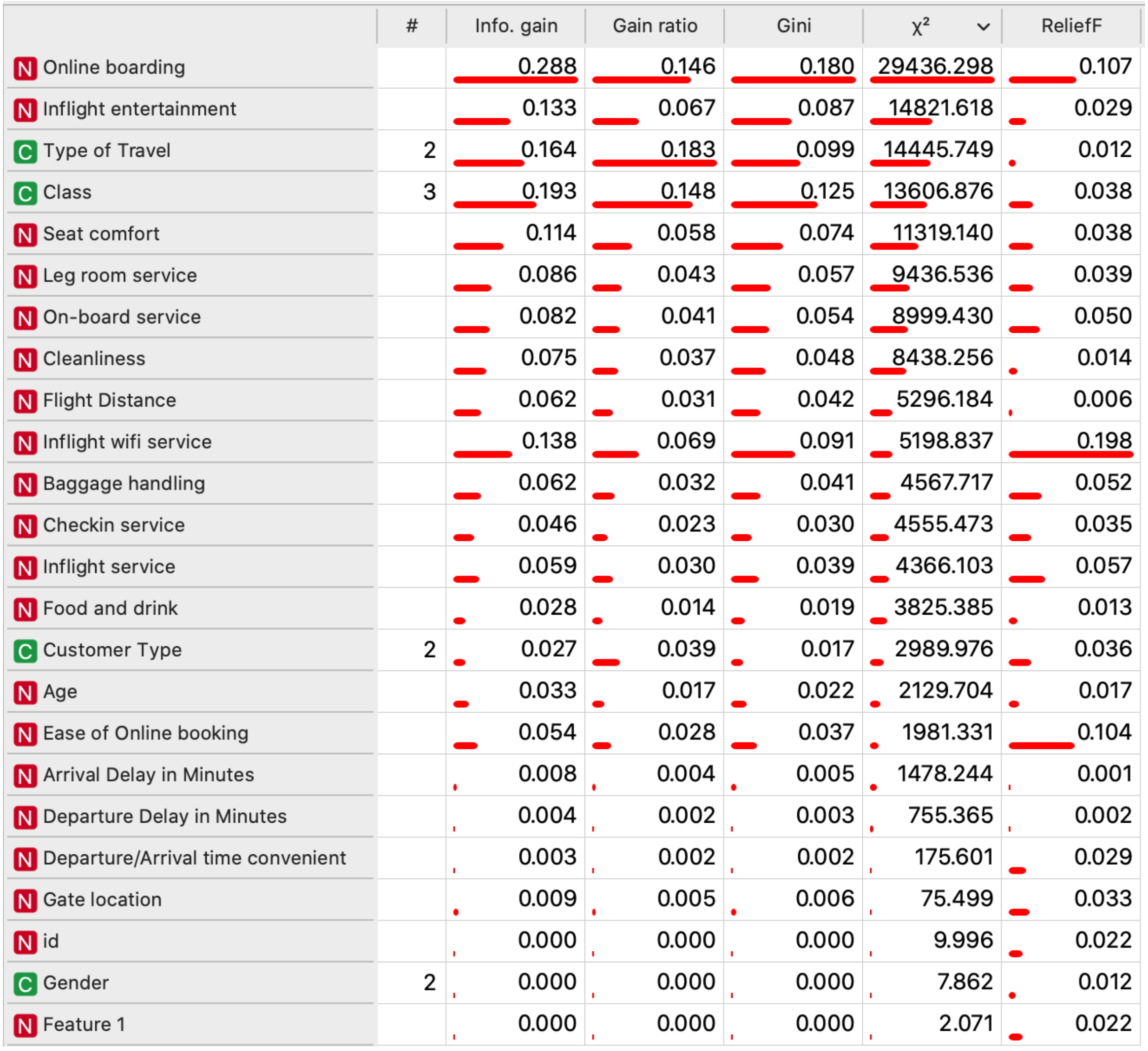
Information Gain: Information gain is the reduction in entropy (degree of uncertainty).

Gain ratio: Gain Ratio is used to normalize the information gain of an attribute against how much entropy that attribute has.

Gini: Gini computes the degree of probability of a specific variable that is wrongly being classified.

Chi^2(X^2): Dependence between the feature and the class as measure by the chi-square statistic.

ReliefF: The ability of an attribute to distinguish between classes on similar data instances.



	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF
N Online boarding		0.288	0.146	0.180	29436.298	0.107
N Inflight entertainment		0.133	0.067	0.087	14821.618	0.029
C Type of Travel	2	0.164	0.183	0.099	14445.749	0.012
C Class	3	0.193	0.148	0.125	13606.876	0.038
N Seat comfort		0.114	0.058	0.074	11319.140	0.038
N Leg room service		0.086	0.043	0.057	9436.536	0.039
N On-board service		0.082	0.041	0.054	8999.430	0.050
N Cleanliness		0.075	0.037	0.048	8438.256	0.014
N Flight Distance		0.062	0.031	0.042	5296.184	0.006
N Inflight wifi service		0.138	0.069	0.091	5198.837	0.198
N Baggage handling		0.062	0.032	0.041	4567.717	0.052
N Checkin service		0.046	0.023	0.030	4555.473	0.035
N Inflight service		0.059	0.030	0.039	4366.103	0.057
N Food and drink		0.028	0.014	0.019	3825.385	0.013
C Customer Type	2	0.027	0.039	0.017	2989.976	0.036
N Age		0.033	0.017	0.022	2129.704	0.017
N Ease of Online booking		0.054	0.028	0.037	1981.331	0.104
N Arrival Delay in Minutes		0.008	0.004	0.005	1478.244	0.001
N Departure Delay in Minutes		0.004	0.002	0.003	755.365	0.002
N Departure/Arrival time convenient		0.003	0.002	0.002	175.601	0.029
N Gate location		0.009	0.005	0.006	75.499	0.033
N id		0.000	0.000	0.000	9.996	0.022
C Gender	2	0.000	0.000	0.000	7.862	0.012
N Feature 1		0.000	0.000	0.000	2.071	0.022

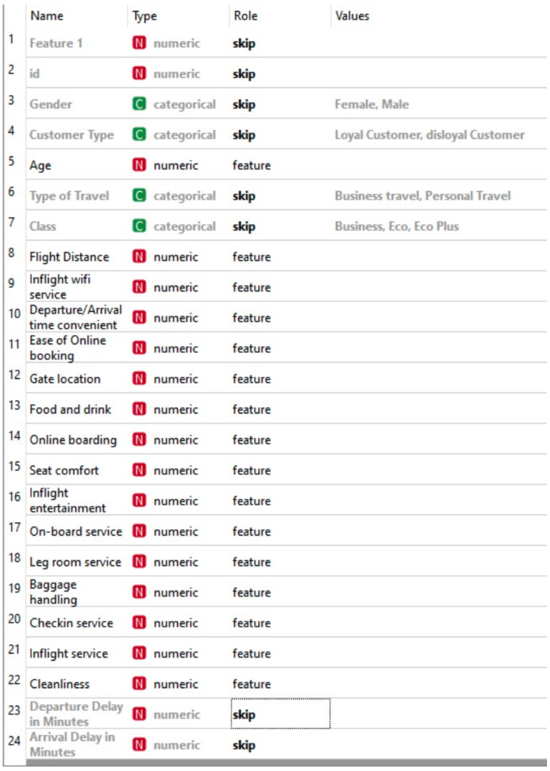
Figure 4.11 – Feature Ranks

4.3.3 Feature Elimination

According to all these analyses, some features should be selected to be used for the model and some of them should be dropped.

According to feature rank scores “Feature 1” and “id” features are totally useless for classification and they are dropped from the dataset. Moreover, since only the satisfaction of the passengers is investigated, categories such as their gender and flight class are also quite meaningless. So that reason, features such as, “Gender”, “Customer Type”, “Type of Travel” and “Class” are dropped.

When we analyze the “Departure Delay in Minutes” and “Arrival Delay in Minutes” features, they are eliminated too because they are correlated with 0.9 correlation score which is very close to 1 so we need to avoid one of them and since “Arrival Delay in Minutes” has missing values we also eliminated that one. For the “Departure Delay in Minutes”, it does not do good when it comes to the rankings so we eliminated it too.



	Name	Type	Role	Values
1	Feature 1	N numeric	skip	
2	id	N numeric	skip	
3	Gender	C categorical	skip	Female, Male
4	Customer Type	C categorical	skip	Loyal Customer, disloyal Customer
5	Age	N numeric	feature	
6	Type of Travel	C categorical	skip	Business travel, Personal Travel
7	Class	C categorical	skip	Business, Eco, Eco Plus
8	Flight Distance	N numeric	feature	
9	Inflight wifi service	N numeric	feature	
10	Departure/Arrival time convenient	N numeric	feature	
11	Ease of Online booking	N numeric	feature	
12	Gate location	N numeric	feature	
13	Food and drink	N numeric	feature	
14	Online boarding	N numeric	feature	
15	Seat comfort	N numeric	feature	
16	Inflight entertainment	N numeric	feature	
17	On-board service	N numeric	feature	
18	Leg room service	N numeric	feature	
19	Baggage handling	N numeric	feature	
20	Checkin service	N numeric	feature	
21	Inflight service	N numeric	feature	
22	Cleanliness	N numeric	feature	
23	Departure Delay in Minutes	N numeric	skip	
24	Arrival Delay in Minutes	N numeric	skip	

Figure 4.12 – Features That Are Dropped

5. Experimental Evaluation

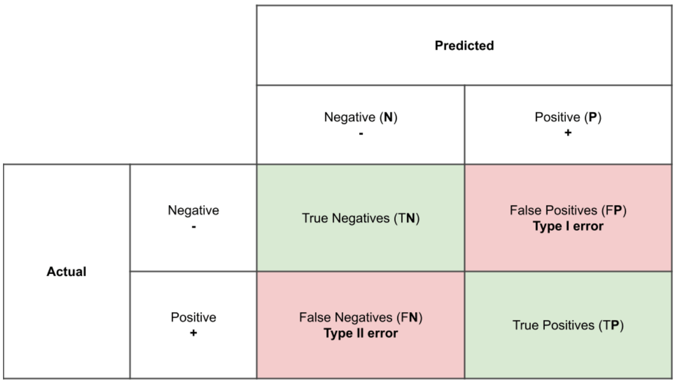
5.1 Methodology

When assessing machine learning models (ML), it is important to choose the right metric. Different metrics are formulated in different implementations to test ML models. You may not get the full picture of the issue you are resolving with certain applications looking for a single measurement, and you might like to use a subset of the metrics addressed in this article to evaluate the models in detail.

As a solution was sought on a classification problem, classification metrics were used to compare and test classification algorithms.

Confusion Matrix: [3] In order to evaluate the performance of the classification models used in machine learning, the error matrix, in which the predictions and actual values of the target attribute are compared, is frequently used. Regardless, classification estimates will have one of four assessments;

- True Positive – TP
- True Negative – TN
- False Positive – FP
- False Negative – FN



		Predicted	
		Negative (**N**) **-**	Positive (**P**) **+**
Actual	Negative **-**	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive **+**	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figure 5.1 – Confusion Matrix

Accuracy: It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}}$$

Figure 5.2 – Accuracy Formula

Precision: It is specified among all examples predicted to belong in a certain class as a fraction of the positive examples.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Figure 5.3 – Precision Formula

Recall: It is the number of correct positive results divided by the number of all samples that should have been identified as positive.

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Figure 5.4 – Recall Formula

F1 Score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

Figure 5.5 – F1 Formula

AUC – ROC: AUC – ROC curve is an efficiency calculation in different threshold settings for classification problems. ROC is a probability curve, and AUC represents a separability degree or metric.

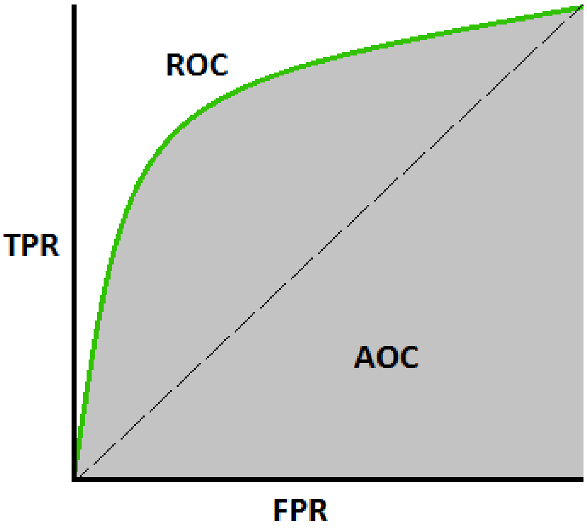
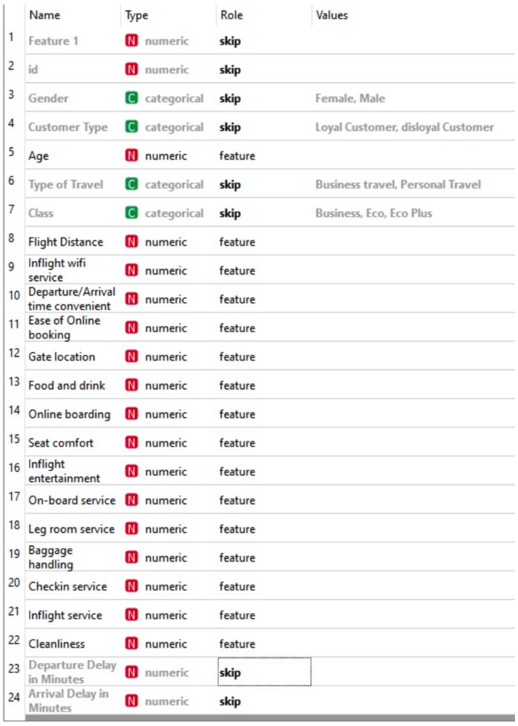


Figure 5.6 – AUC-ROC Curve Explanation

5.2 Latest Version of Dataset

Our data set as a result of data preprocessing and feature selections that we applied in the previous sections is as shown in Figure 5.7.



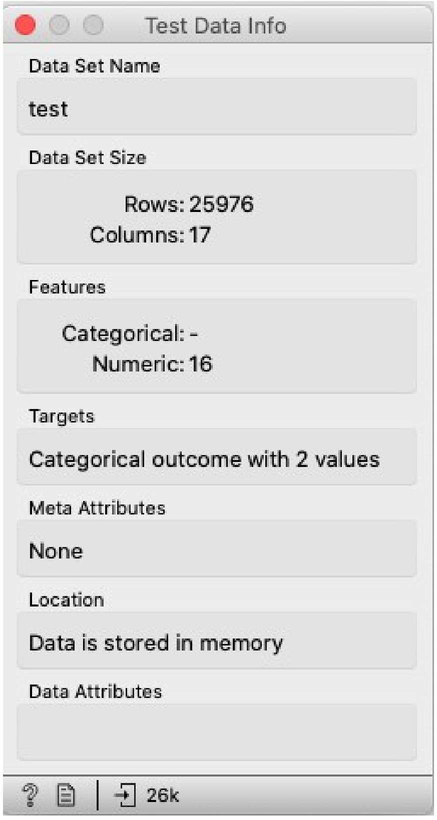
	Name	Type	Role	Values
1	Feature 1	N numeric	skip	
2	id	N numeric	skip	
3	Gender	C categorical	skip	Female, Male
4	Customer Type	C categorical	skip	Loyal Customer, disloyal Customer
5	Age	N numeric	feature	
6	Type of Travel	C categorical	skip	Business travel, Personal Travel
7	Class	C categorical	skip	Business, Eco, Eco Plus
8	Flight Distance	N numeric	feature	
9	Inflight wifi service	N numeric	feature	
10	Departure/Arrival time convenient	N numeric	feature	
11	Ease of Online booking	N numeric	feature	
12	Gate location	N numeric	feature	
13	Food and drink	N numeric	feature	
14	Online boarding	N numeric	feature	
15	Seat comfort	N numeric	feature	
16	Inflight entertainment	N numeric	feature	
17	On-board service	N numeric	feature	
18	Leg room service	N numeric	feature	
19	Baggage handling	N numeric	feature	
20	Checkin service	N numeric	feature	
21	Inflight service	N numeric	feature	
22	Cleanliness	N numeric	feature	
23	Departure Delay in Minutes	N numeric	skip	
24	Arrival Delay in Minutes	N numeric	skip	

Figure 5.7 – Features After Preprocessing

After these operations, our number of features decreased from 25 to 16. After analyzing our data set, some features that we deem unnecessary to solve the problem have been skipped as shown in the figure. In this way, we expect to get more accurate and cleaner results from our machine learning algorithms.

5.3 Latest Dataset Information

Not only features but also some data preprocessing has been applied on the data. Our data set was obtained on Kaggle as two separate files, "test" and "train". Following operations such as deletion of missing data and encoding, the final version of our data set is as follows.



Test Data Info

Data Set Name

test

Data Set Size

Rows: 25976
Columns: 17

Features

Categorical: -
Numeric: 16

Targets

Categorical outcome with 2 values

Meta Attributes

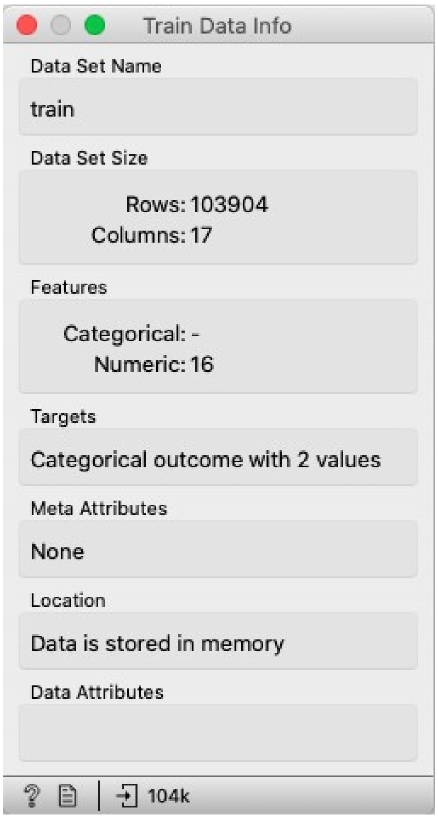
None

Location

Data is stored in memory

Data Attributes

Figure (a) – Test Data Information



Train Data Info

Data Set Name

train

Data Set Size

Rows: 103904
Columns: 17

Features

Categorical: -
Numeric: 16

Targets

Categorical outcome with 2 values

Meta Attributes

None

Location

Data is stored in memory

Data Attributes

Figure (b) – Train Data Information

5.4 Machine Learning Model Selection

Logistic Regression

Logistic Regression is a regression method for classification. It is used to classify categorical or numerical data. [4]

- General formula:

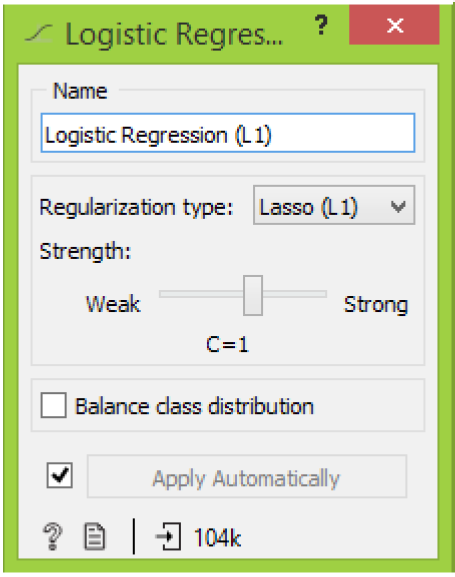
$$\hat{y} = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b$$

- w: coefficients and b: intercept
- Model parameter: C
- Classification error = 1- Classification accuracy

The model has 2 types of regularizations. These are Lasso (L1) and Ridge (L2) regularizations. Also, for Logistic Regression, there is the exchange parameter that determines the strength of the regularization, called C, and the C parameter is very important. The C parameter is also valid for Lasso and Ridge regression.

When moving to the left along the scale, both training and test set accuracy decreases relative to the default parameters. If the capacity is small, regularization will be strong. Also, if the capacity is high, Classification Error is getting higher. Therefore, our capacity number should be low.

First, let's see what score is achieved with default parameter values by applying Logistic Regression. The default arrangement types Ridge (L2) and Lasso (L1) C parameter equals 1. The parameter tab of the model is shown in Figure A and B.



Logistic Regres...

Name

Logistic Regression (L1)

Regularization type: Lasso (L1)

Strength:

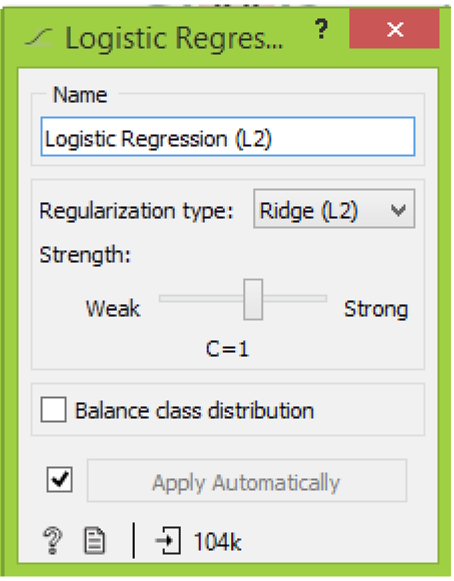
Weak — Strong

C=1

□ Balance class distribution

☑ Apply Automatically

104k



Logistic Regres...

Name

Logistic Regression (L2)

Regularization type: Ridge (L2)

Strength:

Weak — Strong

C=1

□ Balance class distribution

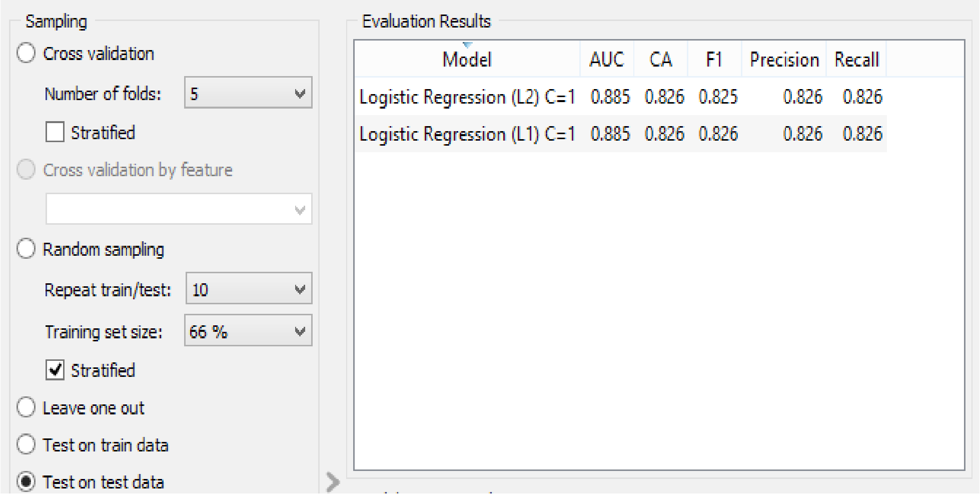
☑ Apply Automatically

104k

Figure (a) – L1 Regularization

Figure (b) – L2 Regularization

Now let's discuss about the model by examining Logistic Regression. First, the same C parameter values are assigned in both applications to examine. These values are the highest score and the lowest score for both settings. C parameter values are 0.001, 1, 1000, respectively.



Sampling

○ Cross validation

Number of folds: 5

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

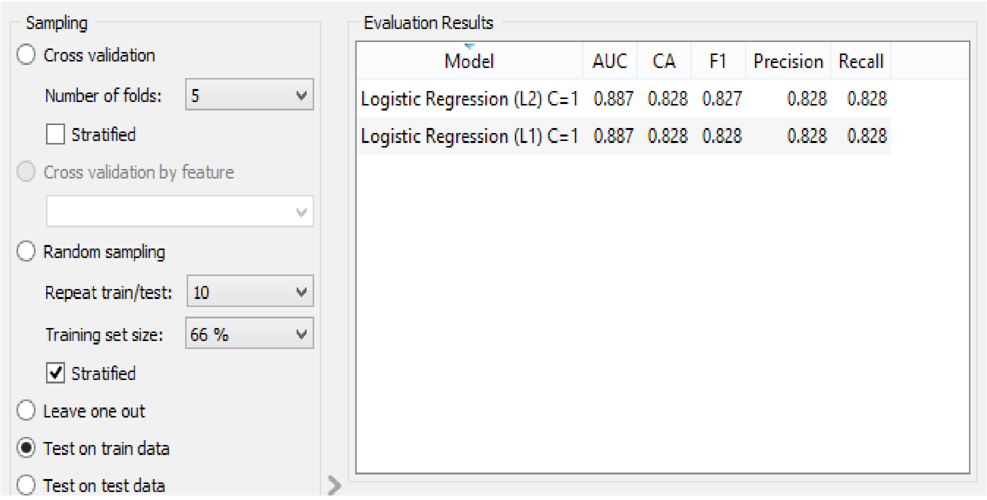
○ Test on train data

● Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression (L2) C=1	0.885	0.826	0.825	0.826	0.826
Logistic Regression (L1) C=1	0.885	0.826	0.826	0.826	0.826

Figure 5.8 – Test Score Results at C = 1



Sampling

○ Cross validation

Number of folds: 5

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

● Test on train data

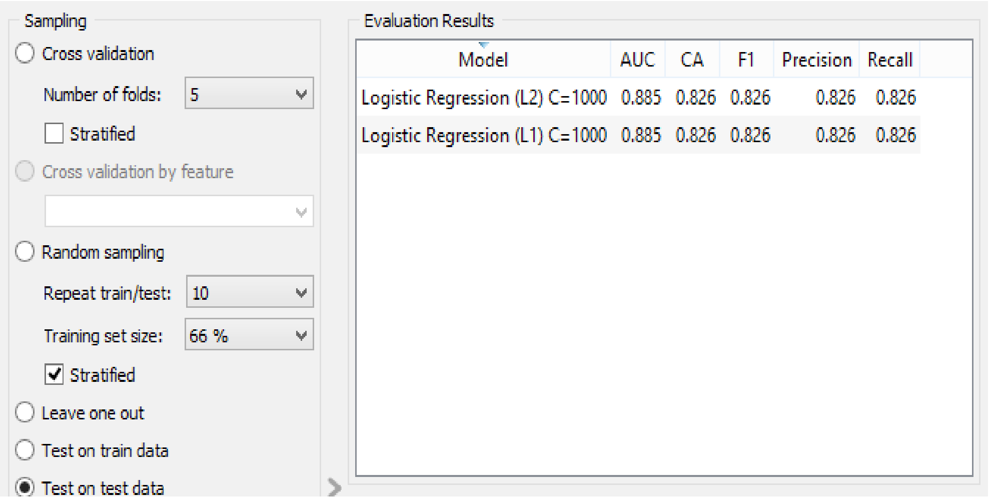
○ Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression (L2) C=1	0.887	0.828	0.827	0.828	0.828
Logistic Regression (L1) C=1	0.887	0.828	0.828	0.828	0.828

Figure 5.9 – Train Score Results at C = 1

As seen in the Figure 5.10, in L2 regularization, the default value of C=1 provides quite good performance, with 82.8% accuracy on the training and 82.6% accuracy on the test set. It is under-fitting because training and test set performance are very close. The same results is seen in L1 regularization.



Sampling

○ Cross validation

Number of folds: 5

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

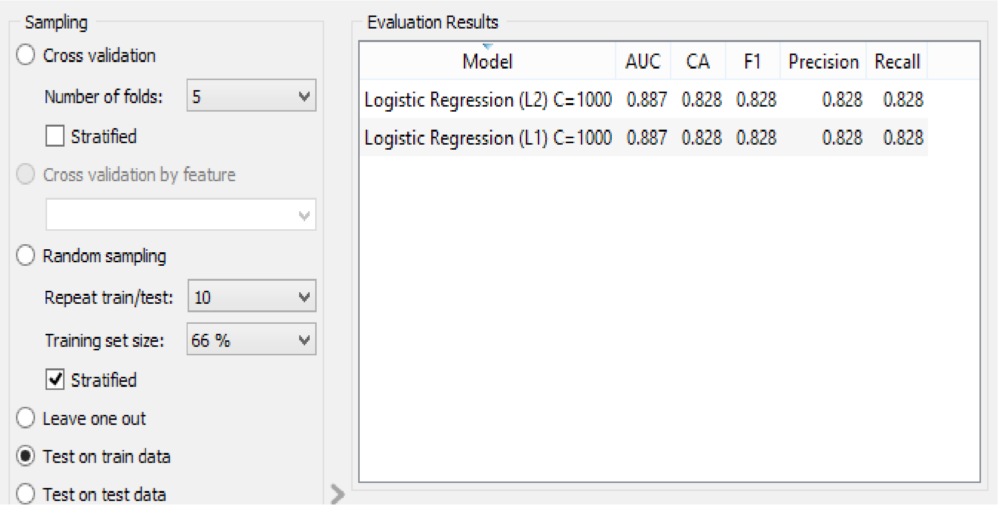
○ Test on train data

● Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression (L2) C=1000	0.885	0.826	0.826	0.826	0.826
Logistic Regression (L1) C=1000	0.885	0.826	0.826	0.826	0.826

Figure 5.10 – Test Score Results at C = 1000



Sampling

○ Cross validation

Number of folds: 5

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

● Test on train data

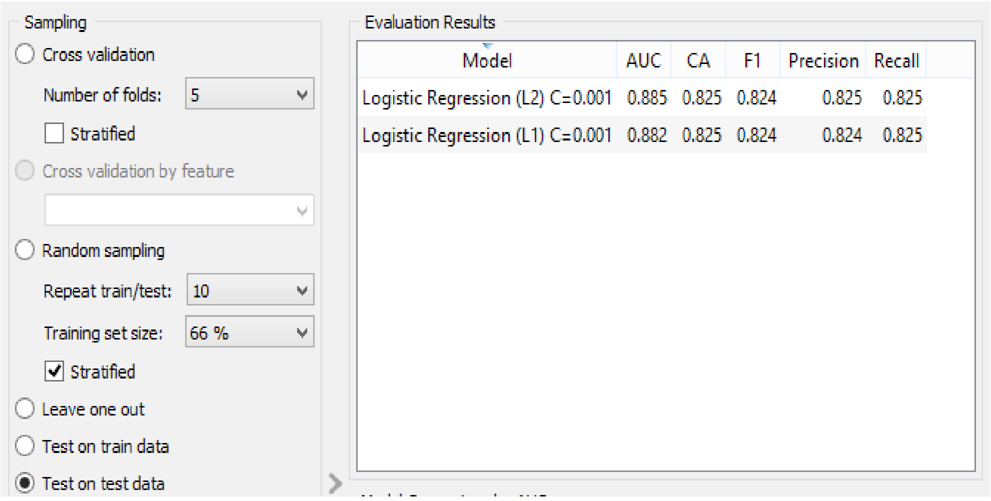
○ Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression (L2) C=1000	0.887	0.828	0.828	0.828	0.828
Logistic Regression (L1) C=1000	0.887	0.828	0.828	0.828	0.828

Figure 5.11 – Train Score Results at C = 1000

The default value of C = 1000 provides the same result as C = 1, with 82.8% accuracy in training and 82.6% accuracy in the test set. The same results is seen again in L1 regularization. Figure is shown in 5.11.



Sampling

○ Cross validation

Number of folds: 5

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

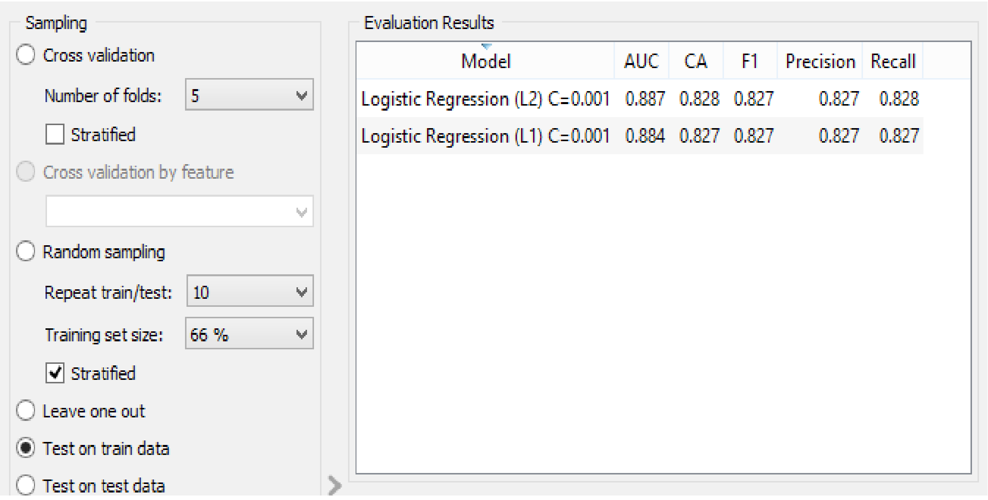
○ Test on train data

● Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression (L2) C=0.001	0.885	0.825	0.824	0.825	0.825
Logistic Regression (L1) C=0.001	0.882	0.825	0.824	0.824	0.825

Figure 5.12 – Test Score Results at C = 0.001



Sampling

○ Cross validation

Number of folds: 5

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

● Test on train data

○ Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression (L2) C=0.001	0.887	0.828	0.827	0.827	0.828
Logistic Regression (L1) C=0.001	0.884	0.827	0.827	0.827	0.827

Figure 5.13 – Train Score Results at C = 0.001

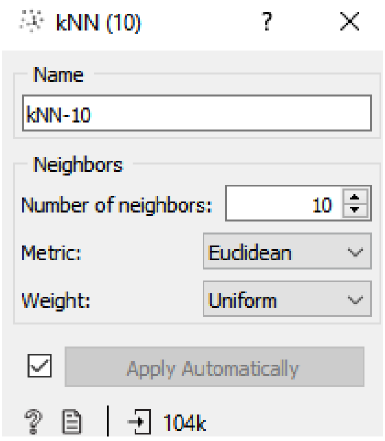
If we set C = 0.001; There is an accuracy of 82.8% in education and an accuracy of 82.5% in the test set. In L1 regularization, there is an accuracy of 82.7% in education and an accuracy of 82.5% in the test set. Figure is shown in 5.13.

Finally, when our model was examined by logistic regression, the results of the classification accuracy are close to each other, even the same, regardless of the C parameters. As a result, in logistic regression, no trials yielded complete results. Therefore, it is not suitable for our model.

K-Nearest Neighbors

KNN (k-nearest Neighbors) is a distance-based supervised learning algorithm. To predict the value in a new data point, the closest data point in the algorithm trend data set is found. For using a single neighbor, the test set accuracy is lower than when k = 1 more neighbors are used, indicating that the use of the closest neighbor leads to a very complex pattern.

To try the KNN algorithm, we first assigned some numbers to the k value, and we assigned our 2, 3, 4, 5, 6 and 10.test set as data -> test data. We dropped our train set as data-> data.



Name

kNN-10

Neighbors

Number of neighbors: 10

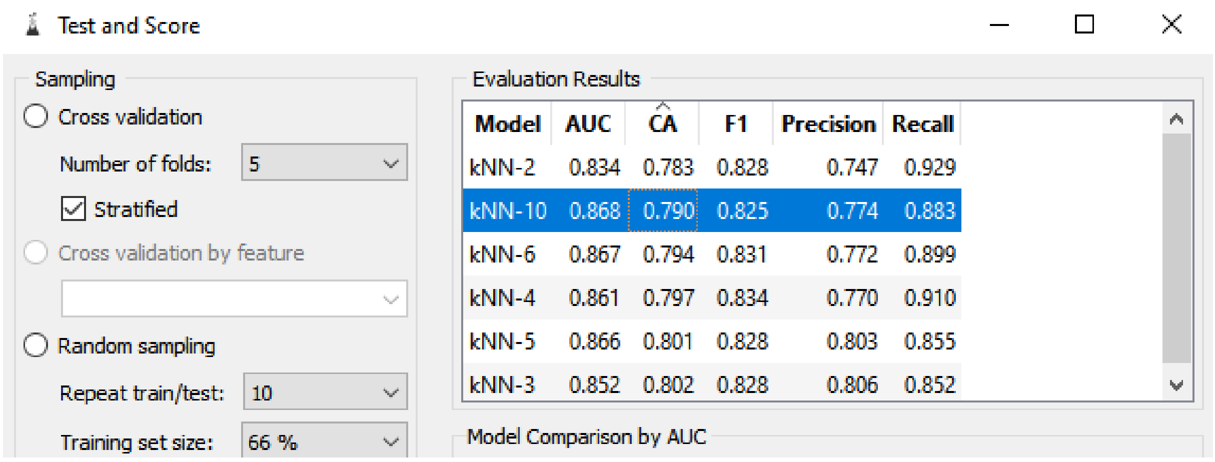
Metric: Euclidean

Weight: Uniform

☑ Apply Automatically

Figure 5.14 – Parameter Selection

The training results the with k values as 2,3,4,5,6 and 10 in the KNN model. Here we found the most suitable k value by trying the value of k one by one. This value was 10 because as the value of k fell to 1, it became overfit. Conversely, as the value of k increased, the predictions became more stable, but until a certain point we began to witness increasing numbers of errors.



Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

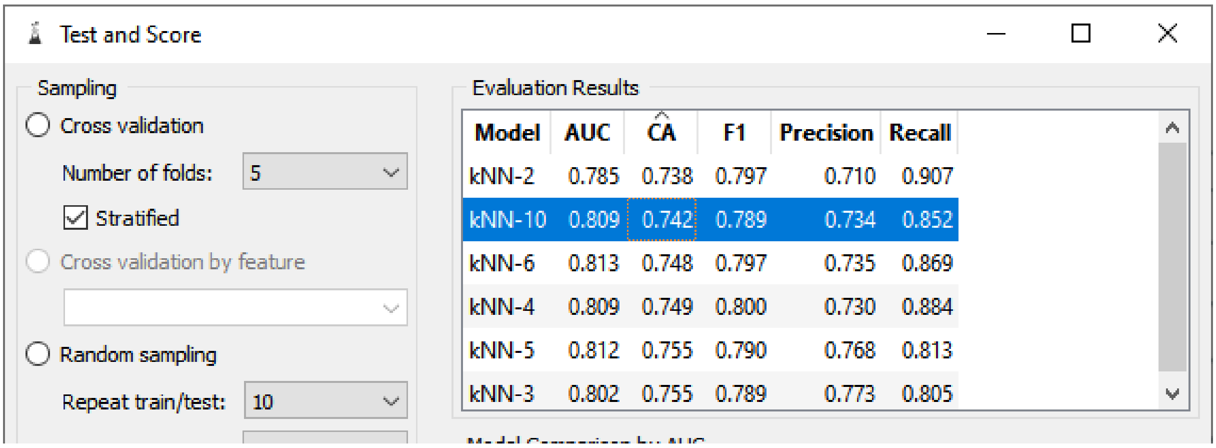
Training set size: 66 %

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN-2	0.834	0.783	0.828	0.747	0.929
kNN-10	0.868	0.790	0.825	0.774	0.883
kNN-6	0.867	0.794	0.831	0.772	0.899
kNN-4	0.861	0.797	0.834	0.770	0.910
kNN-5	0.866	0.801	0.828	0.803	0.855
kNN-3	0.852	0.802	0.828	0.806	0.852

Model Comparison by AUC

Figure 5.15 – KNN Training Results



Test and Score

Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN-2	0.785	0.738	0.797	0.710	0.907
kNN-10	0.809	0.742	0.789	0.734	0.852
kNN-6	0.813	0.748	0.797	0.735	0.869
kNN-4	0.809	0.749	0.800	0.730	0.884
kNN-5	0.812	0.755	0.790	0.768	0.813
kNN-3	0.802	0.755	0.789	0.773	0.805

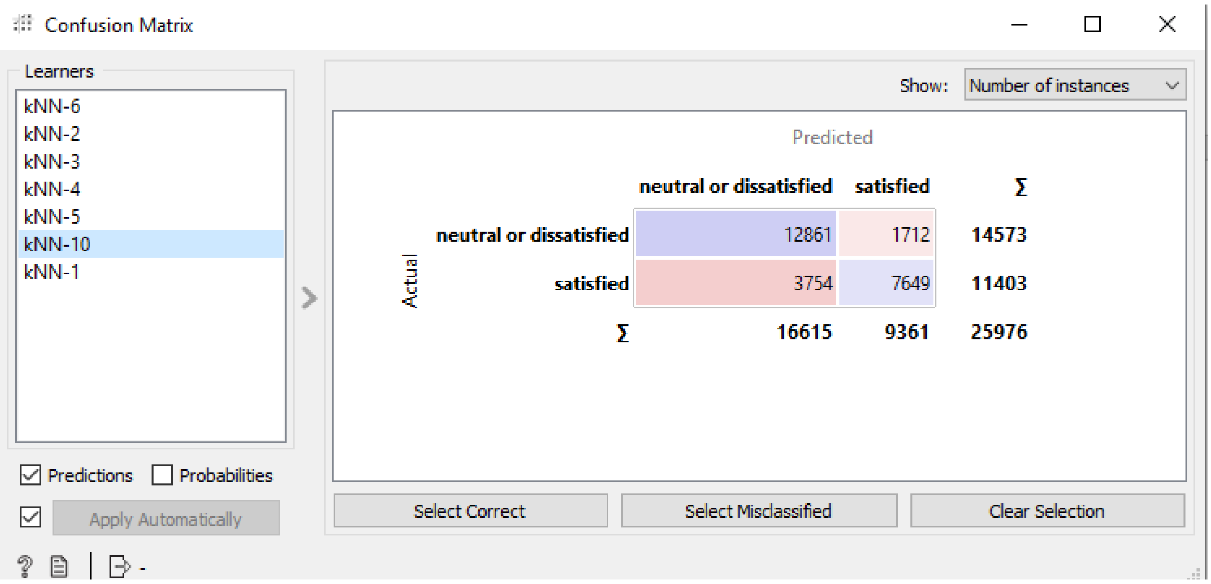
Model Comparison by AUC

Figure 5.16 – KNN Test Results

Our results are as follows for k=10:

- train (0.790)
- testing (0.742)

Here it is also the representation in confusion matrix format in Figure 5.17.



Confusion Matrix

Learners

kNN-6
kNN-2
kNN-3
kNN-4
kNN-5
kNN-10
kNN-1

Show: Number of instances

Predicted

Actual	neutral or dissatisfied	satisfied	Σ
neutral or dissatisfied	12861	1712	14573
satisfied	3754	7649	11403
Σ	16615	9361	25976

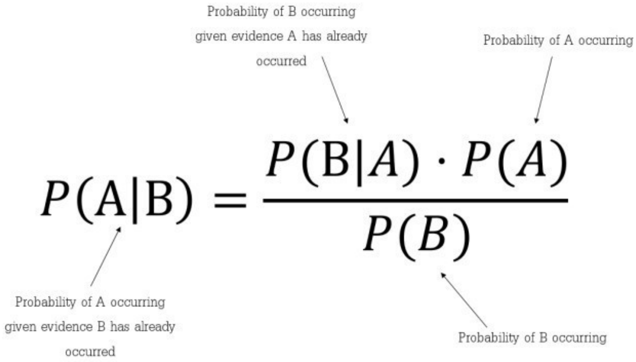
☑ Predictions ☐ Probabilities

☑ Apply Automatically

Figure 5.17 – KNN Confusion Matrix

Naive Bayes

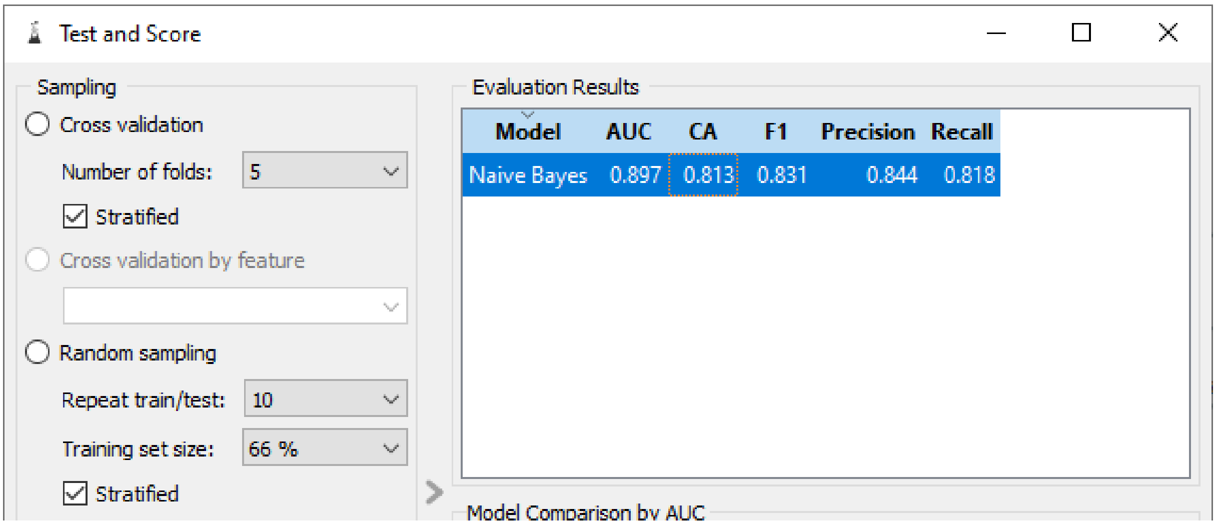
It learns each feature by looking at it separately and collects simple statistics per class from each feature.



$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

Probability of B occurring given evidence A has already occurred
Probability of A occurring
Probability of A occurring given evidence B has already occurred
Probability of B occurring

Figure 5.18 – Conditional Probability Formula



Test and Score

Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

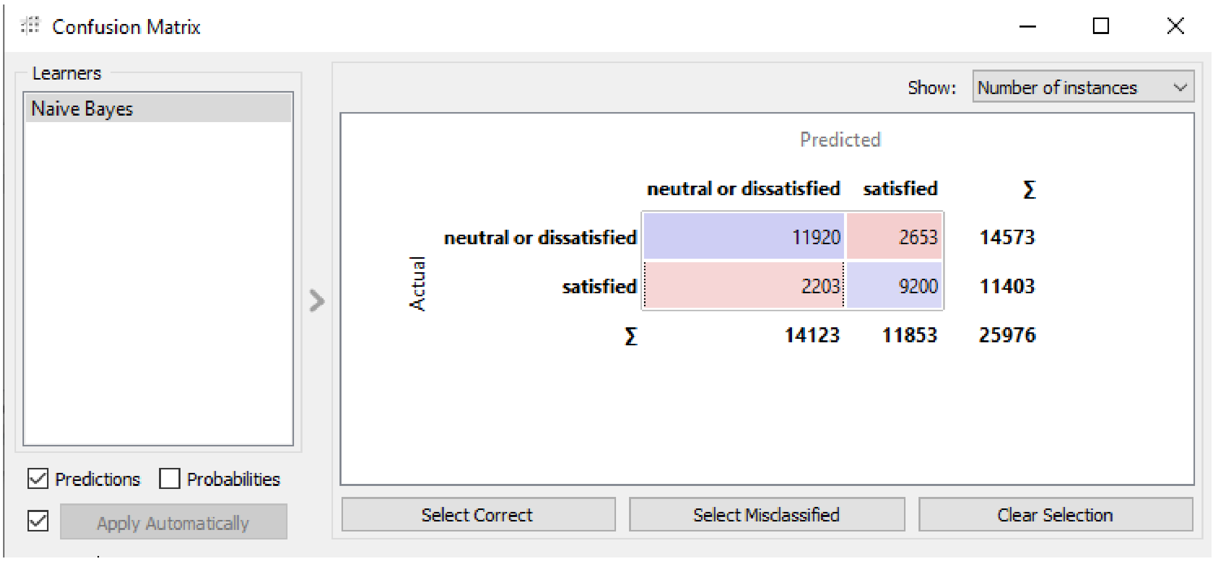
☑ Stratified

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.897	0.813	0.831	0.844	0.818

Model Comparison by AUC

Figure 5.19 – Naive Bayes Test Score



Confusion Matrix

Learners

Naive Bayes

Show: Number of instances

		Predicted		
		neutral or dissatisfied	satisfied	Σ
Actual	neutral or dissatisfied	11920	2653	14573
	satisfied	2203	9200	11403
	Σ	14123	11853	25976

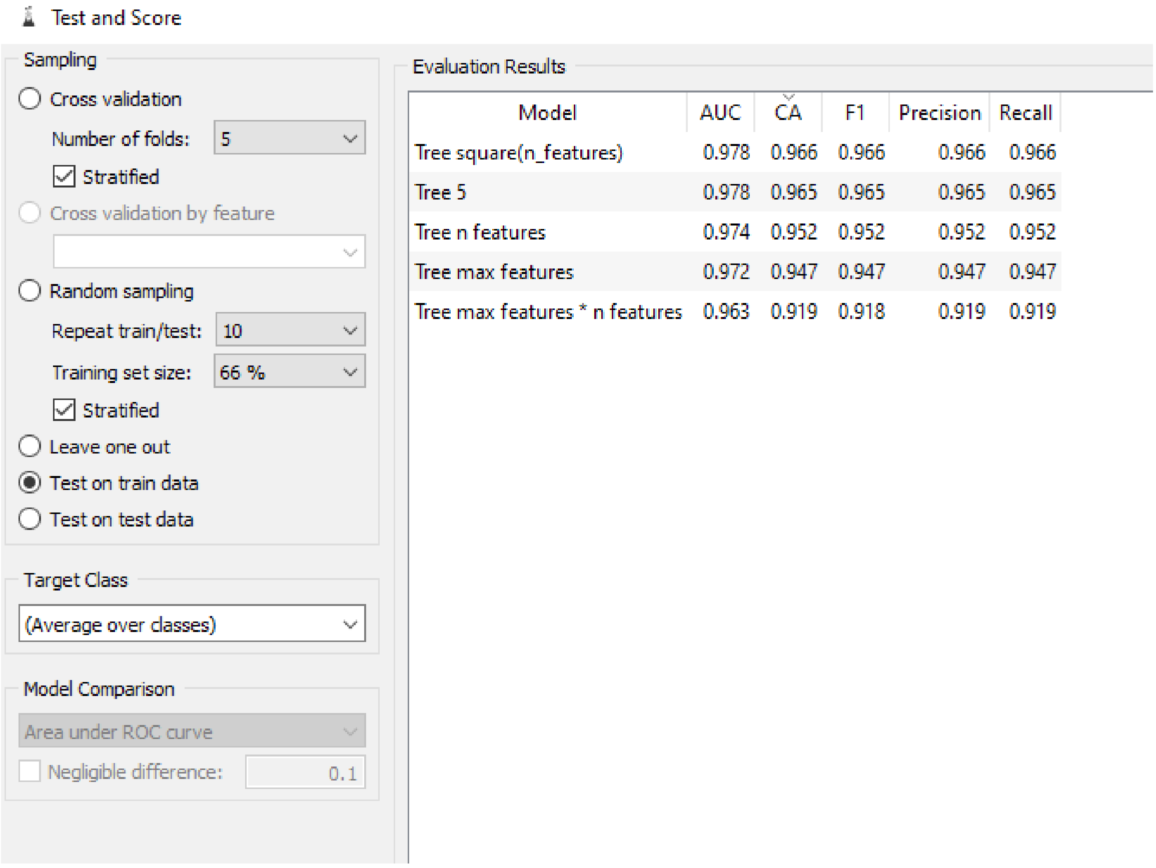
☑ Predictions ☐ Probabilities

☑ Apply Automatically

Figure 5.20 – Naive Bayes Confusion Matrix

Decision Tree

In this model the usage of tree model is explained. Tree modeling is really works well on classification problems because it can be modified accordingly. In our example we choose the number of leaves as 2 because it is a binary classification problem. This number of leaves will never change during the process because it is a set number. That being said we are first going to examine the number of split. For the first attempt we did not limit the max tree dept in order to see the effects of split. On Figure 5.21 and 5.22 we are going to see the test on test data and test on train data.



Test and Score

Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

● Test on train data

○ Test on test data

Target Class

(Average over classes)

Model Comparison

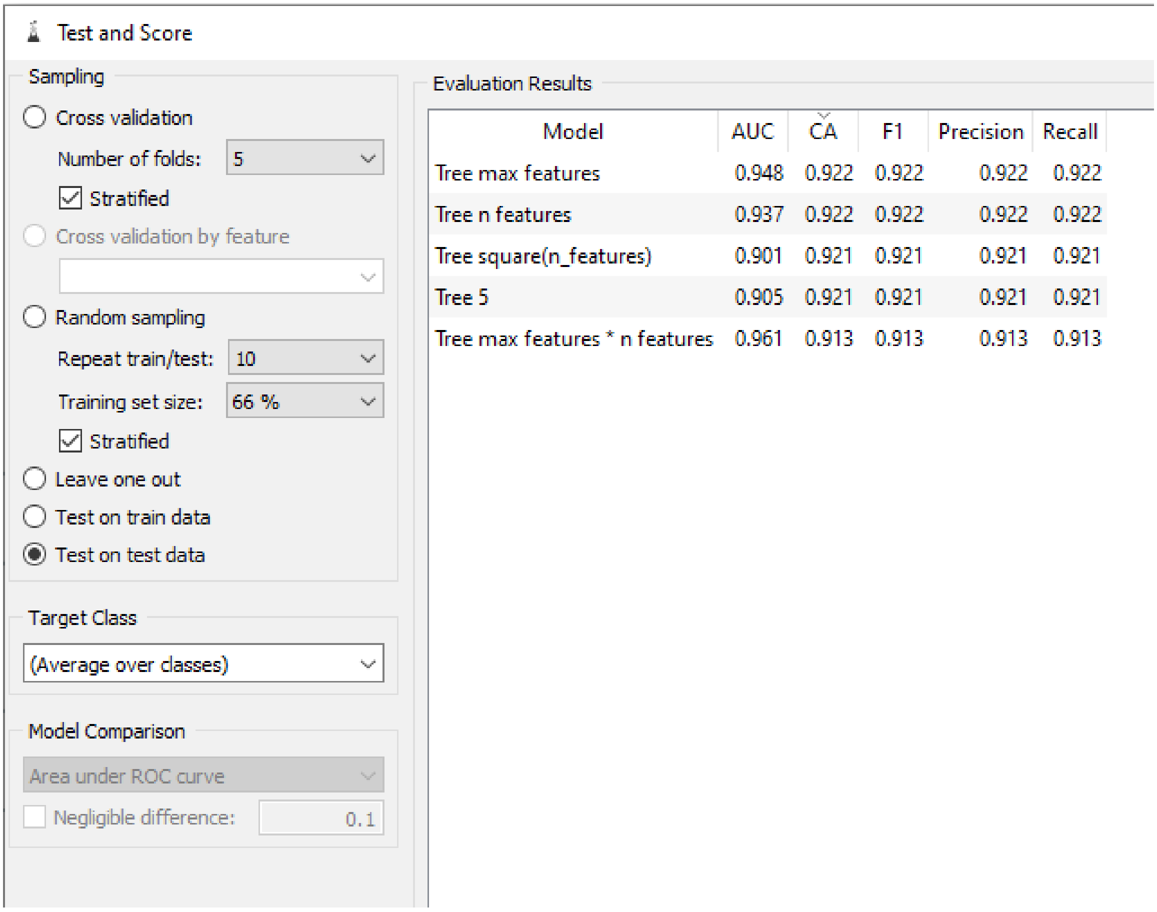
Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree square(n_features)	0.978	0.966	0.966	0.966	0.966
Tree 5	0.978	0.965	0.965	0.965	0.965
Tree n features	0.974	0.952	0.952	0.952	0.952
Tree max features	0.972	0.947	0.947	0.947	0.947
Tree max features * n features	0.963	0.919	0.918	0.919	0.919

Figure 5.21 – Decision Tree Test Training Results



Test and Score

Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

● Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree max features	0.948	0.922	0.922	0.922	0.922
Tree n features	0.937	0.922	0.922	0.922	0.922
Tree square(n_features)	0.901	0.921	0.921	0.921	0.921
Tree 5	0.905	0.921	0.921	0.921	0.921
Tree max features * n features	0.961	0.913	0.913	0.913	0.913

Figure 5.22 – Decision Tree Test Results

We take 5 parameters according to [5] and they are;

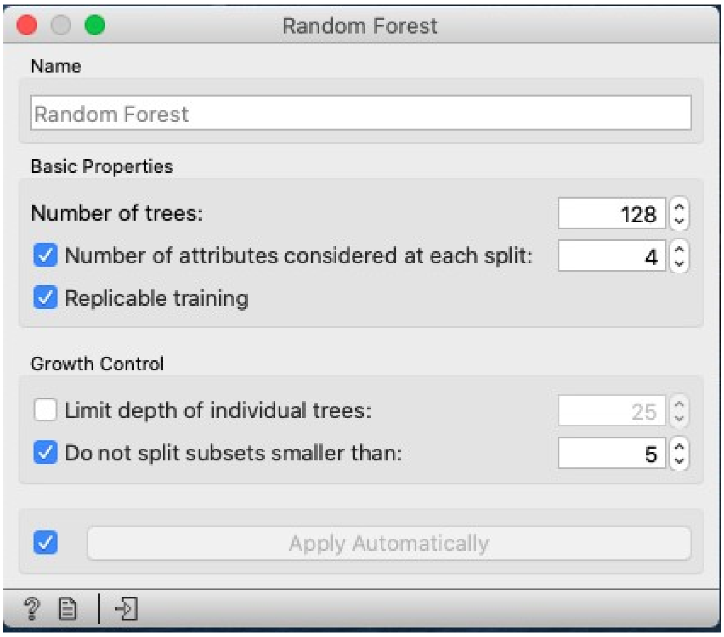
- Square root of selected features (sqrt(16) = 4).
- Square root of max features (sqrt(25) = 5).
- Number of selected features (n features = 16).
- Number of max features (max features =25).
- Number of max features multiply by number of selected features (25*16 = 400).

According to test on train data classification accuracy and precision first one is the best option in hand but according to test on test data second and third option takes the lead but the difference is negligible. Of course these result are taken with no limitation oof max tree dept. Actually not limiting the max tree dept is optimal for most of the case because it may cause over fitting or under fitting if not being careful. For example, we tried tree depth as 100 and 1000 and the result did not change not even a little bit. But below 25 the accuracy started dropping lower and stopped at 0.788 classification accuracy when the depth becomes 1. There for we decided not to limit the tree dept as instructed. Since the best result belongs to first parameter we choose that one for our project.

Random Forest

Random Forest is an ensemble method and it consists of individual decision trees. Random Forest makes predictions based on majority of votes from each of the decision trees made. Since it is considering working on multiple decision trees there is a chance of less over-fitting and it is the biggest advantage of random forest classifiers. Another plus for Random Forest classifiers is they are easy to understand and visualize by using tree diagrams.

Several parameter values are tested and base on suggestions that have been written on forums it is okay to use between 128-256 number of tree for that size of dataset. So, 128 is used as a number of trees parameter.



Random Forest

Name

Random Forest

Basic Properties

Number of trees: 128

☑ Number of attributes considered at each split: 4

☑ Replicable training

Growth Control

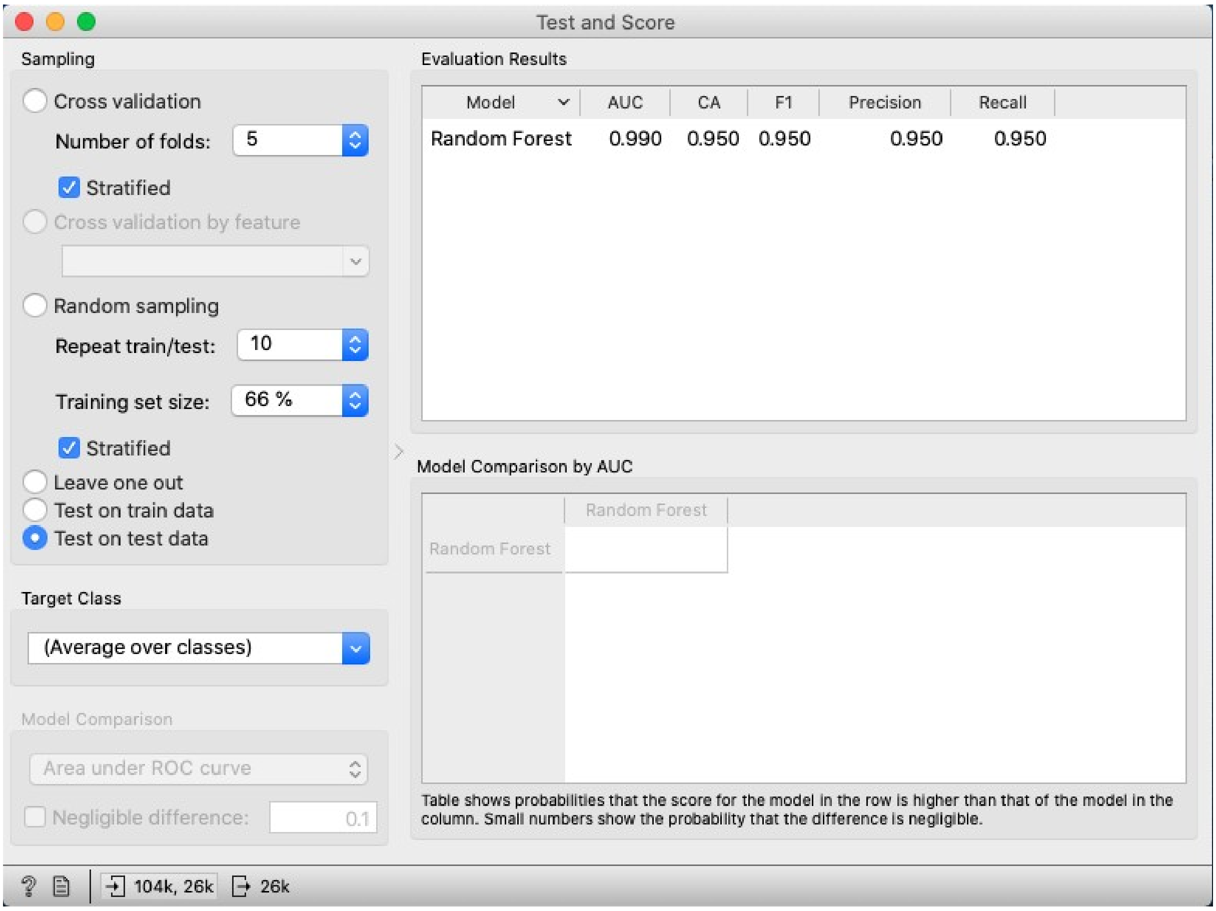
□ Limit depth of individual trees: 25

☑ Do not split subsets smaller than: 5

☑ Apply Automatically

Figure 5.23 – Random Forest Parameters

But more important parameters is the number of attributes considered for each split. It has crucial impact on bootstrap process. So, we choose that 4 as a square root of the number of total features.



Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

● Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.990	0.950	0.950	0.950	0.950

Model Comparison by AUC

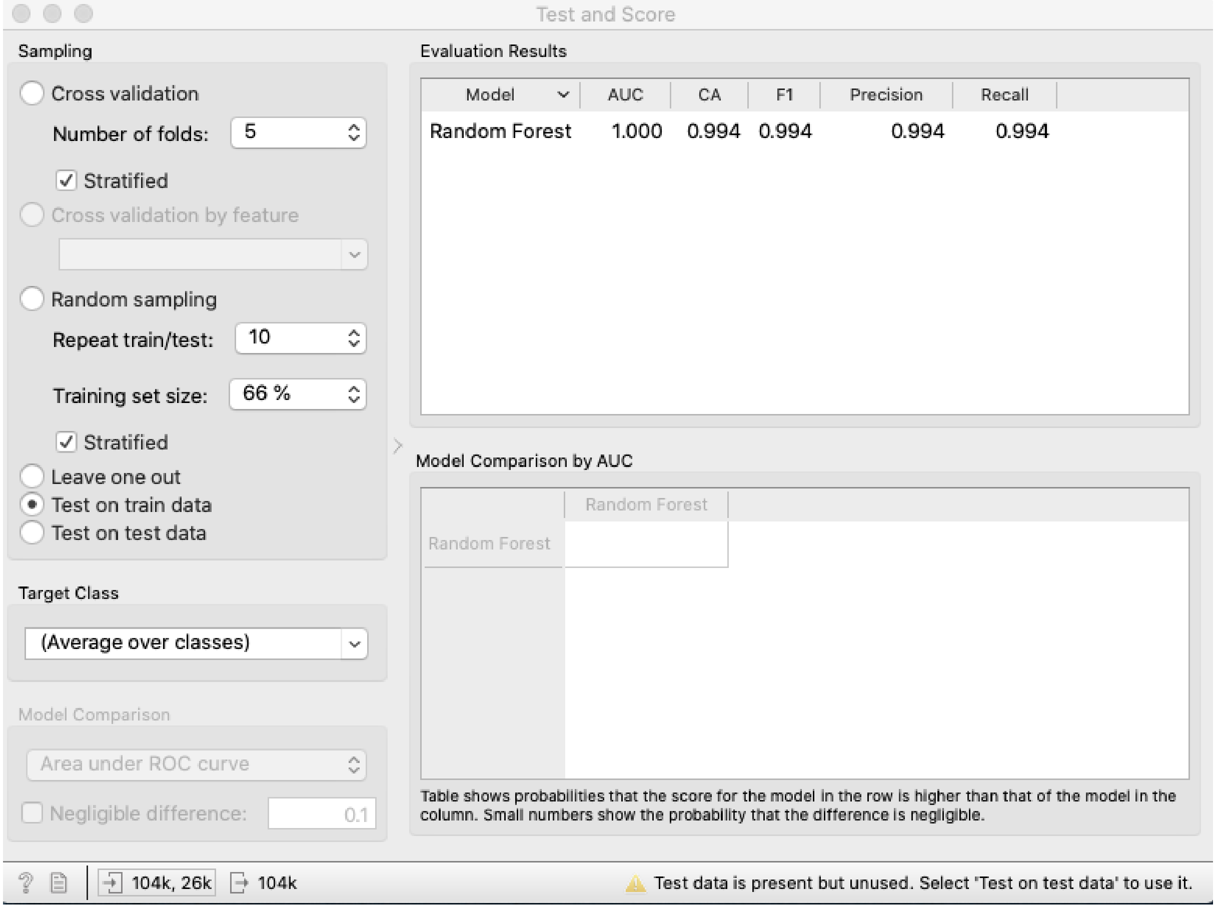
	Random Forest
Random Forest	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

104k, 26k 26k

Figure 5.24 – Random Forest Test Score

As it expected Random Forest has high training accuracy, when it is compared with the other classifiers. With the 0.994 training score and with the 0.950 testing score, it is clearly that the Random Forest Model that is trained is almost perfectly generalized for the binary classification problem.



Test and Score

Sampling

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

● Test on train data

○ Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	1.000	0.994	0.994	0.994	0.994

Model Comparison by AUC

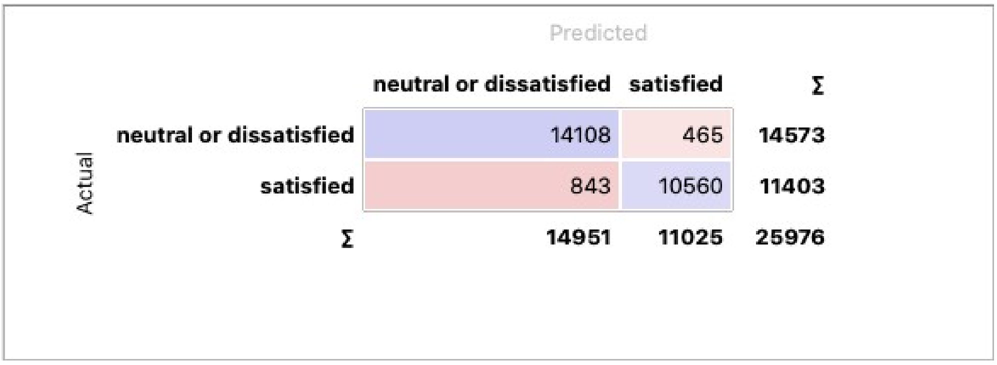
	Random Forest
Random Forest	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

104k, 26k 104k

⚠ Test data is present but unused. Select 'Test on test data' to use it.

Figure 5.25 – Random Forest Training Score



Predicted

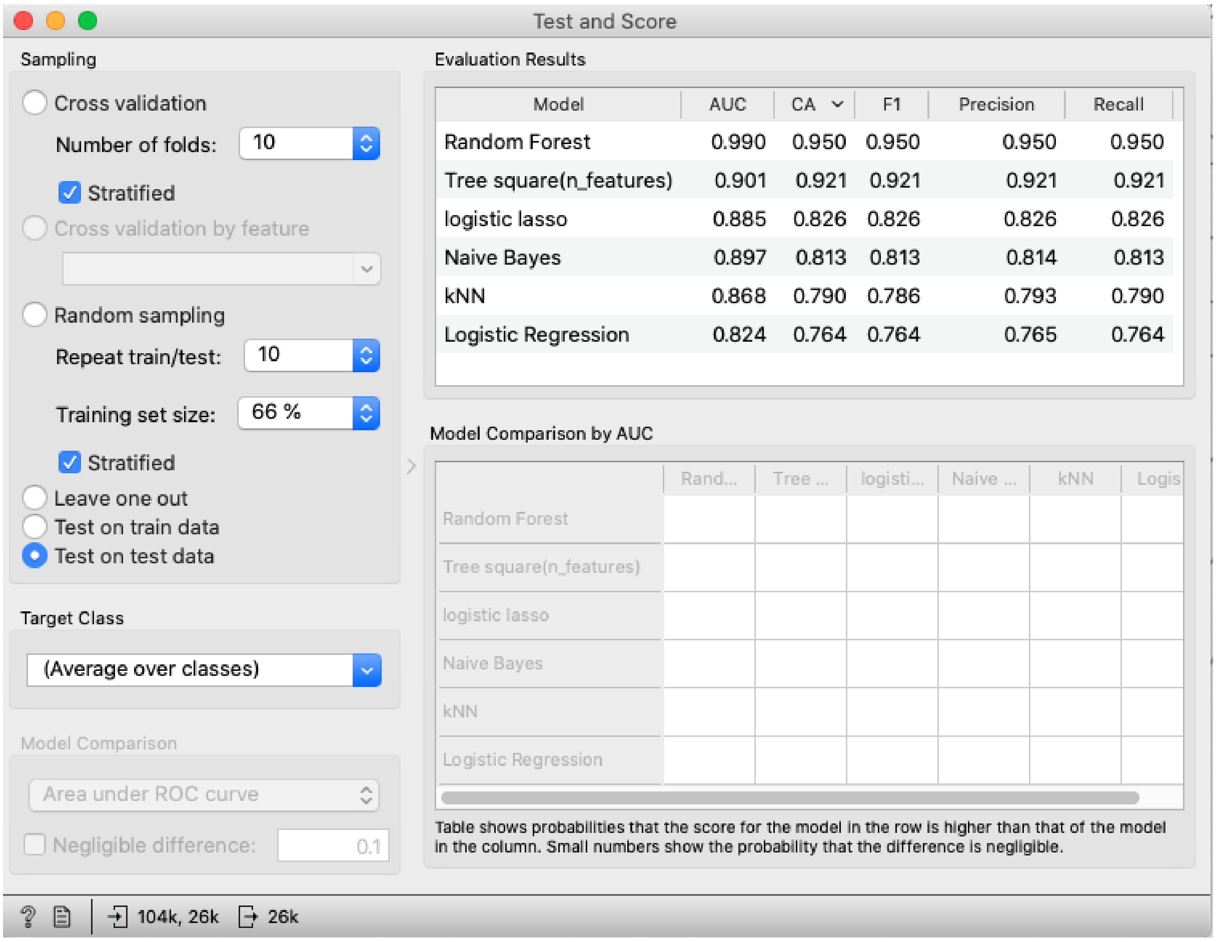
Actual	neutral or dissatisfied	satisfied	Σ
neutral or dissatisfied	14108	465	14573
satisfied	843	10560	11403
Σ	14951	11025	25976

Figure 5.26 – Random Forest Confusion Matrix

According to our results, our confusion matrix is like that. The random forest model has made 1308 mistakes totally and remaining predictions are true. These numbers are showing us the accuracies.

5.5 Results and Discussion

As mentioned earlier, five different classification algorithms were used to solve the project. The AUC-ROC graph and results of these classification algorithms are as follows.



Test and Score

Sampling

○ Cross validation

Number of folds: 10

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

● Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.990	0.950	0.950	0.950	0.950
Tree square(n_features)	0.901	0.921	0.921	0.921	0.921
logistic lasso	0.885	0.826	0.826	0.826	0.826
Naive Bayes	0.897	0.813	0.813	0.814	0.813
kNN	0.868	0.790	0.786	0.793	0.790
Logistic Regression	0.824	0.764	0.764	0.765	0.764

Model Comparison by AUC

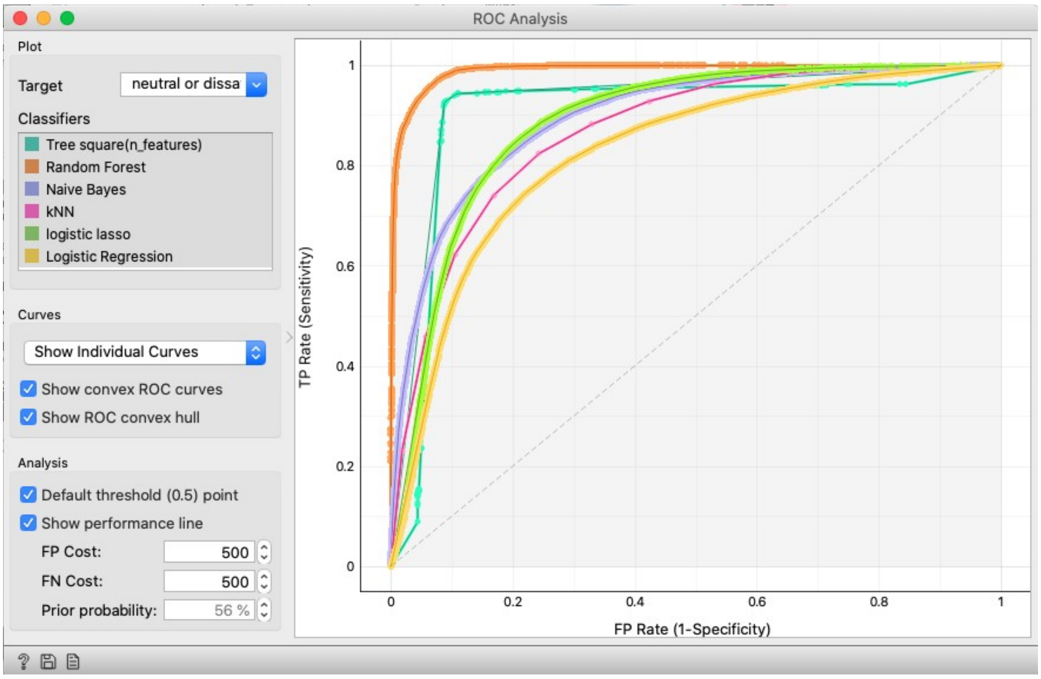
	Rand...	Tree ...	logisti...	Naive ...	kNN	Logis
Random Forest						
Tree square(n_features)						
logistic lasso						
Naive Bayes						
kNN						
Logistic Regression						

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

104k, 26k 26k

Figure 5.27 – Final Test and Score

According to the analysis, the Random Forest model seems to be the most optimal solution to the problem. Since the accuracy rate of both training and test results are high and these values are close to each other, it can be said that our model is able to generalize.



ROC Analysis

Plot

Target: neutral or dissa

Classifiers

- ■ Tree square(n_features)
- ■ Random Forest
- ■ Naive Bayes
- ■ kNN
- ■ logistic lasso
- ■ Logistic Regression

Curves

Show Individual Curves

☑ Show convex ROC curves

☑ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

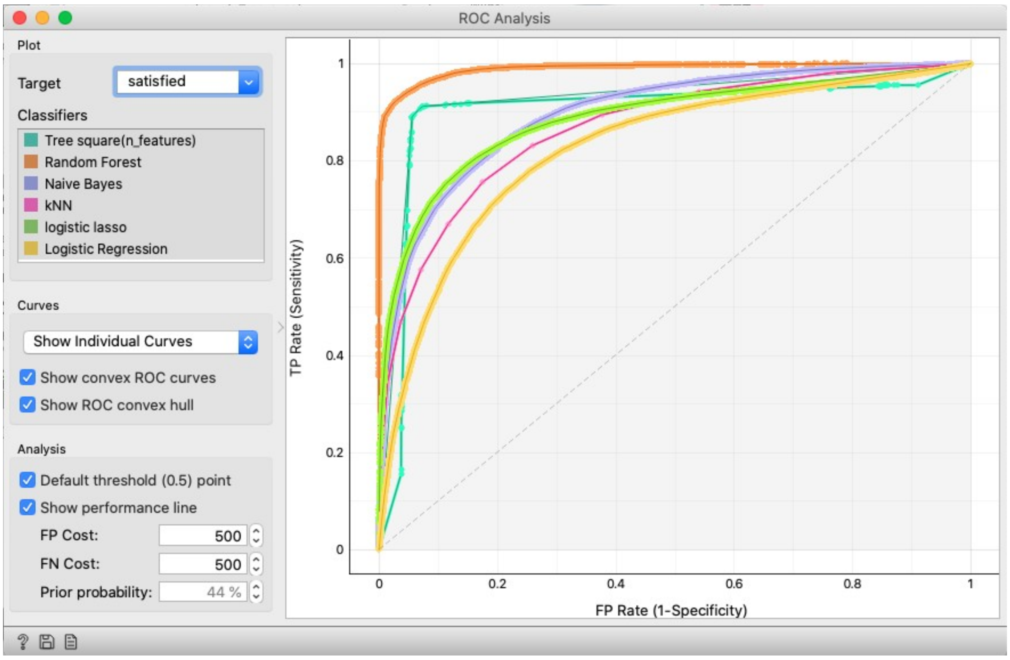
☑ Show performance line

FP Cost: 500

FN Cost: 500

Prior probability: 56 %

Figure 5.28 – ROC Curves For Target “Dissatisfied”



ROC Analysis

Plot

Target: satisfied

Classifiers

- ■ Tree square(n_features)
- ■ Random Forest
- ■ Naive Bayes
- ■ kNN
- ■ logistic lasso
- ■ Logistic Regression

Curves

Show Individual Curves

☑ Show convex ROC curves

☑ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

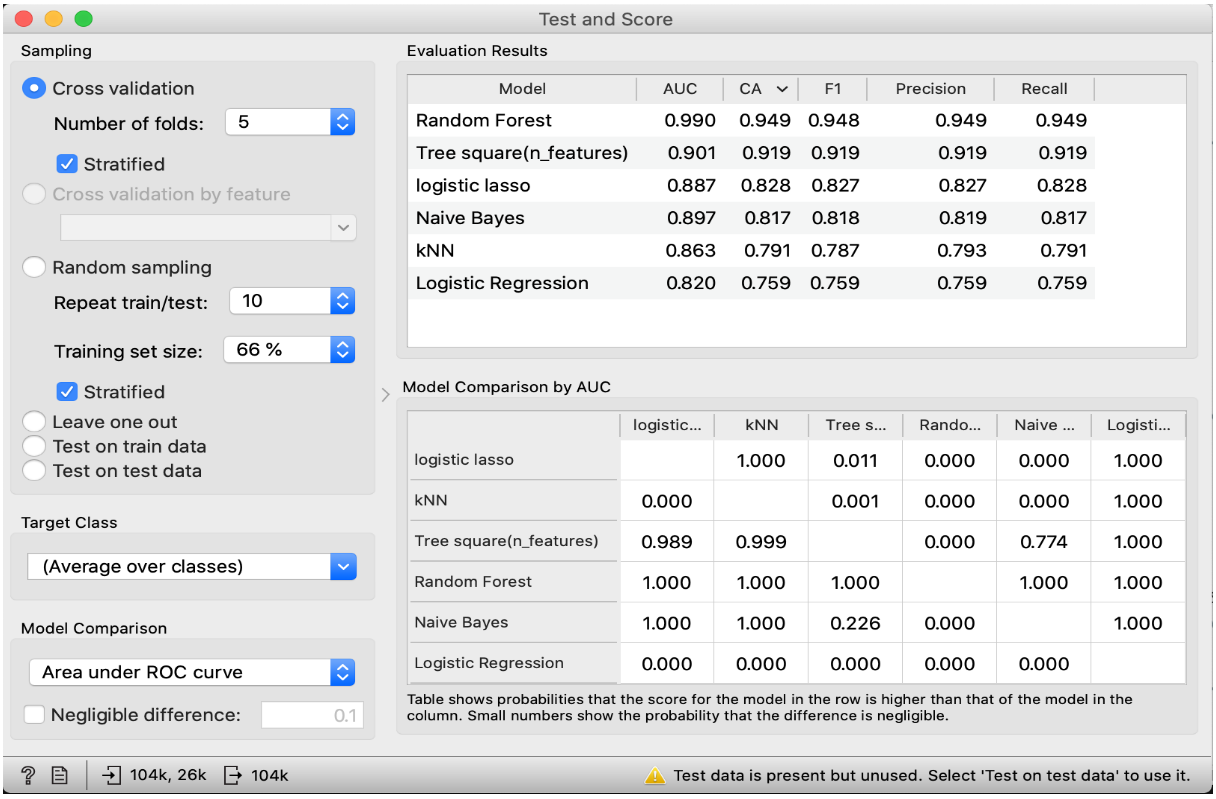
FP Cost: 500

FN Cost: 500

Prior probability: 44 %

Figure 5.29 – ROC Curves For Target “Satisfied”

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample and tested the models with two different k value, 5 and 10. As a result of that the accuracies did not change and that shows that the models are working pretty well.



Test and Score

Sampling

● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.990	0.949	0.948	0.949	0.949
Tree square(n_features)	0.901	0.919	0.919	0.919	0.919
logistic lasso	0.887	0.828	0.827	0.827	0.828
Naive Bayes	0.897	0.817	0.818	0.819	0.817
kNN	0.863	0.791	0.787	0.793	0.791
Logistic Regression	0.820	0.759	0.759	0.759	0.759

Model Comparison by AUC

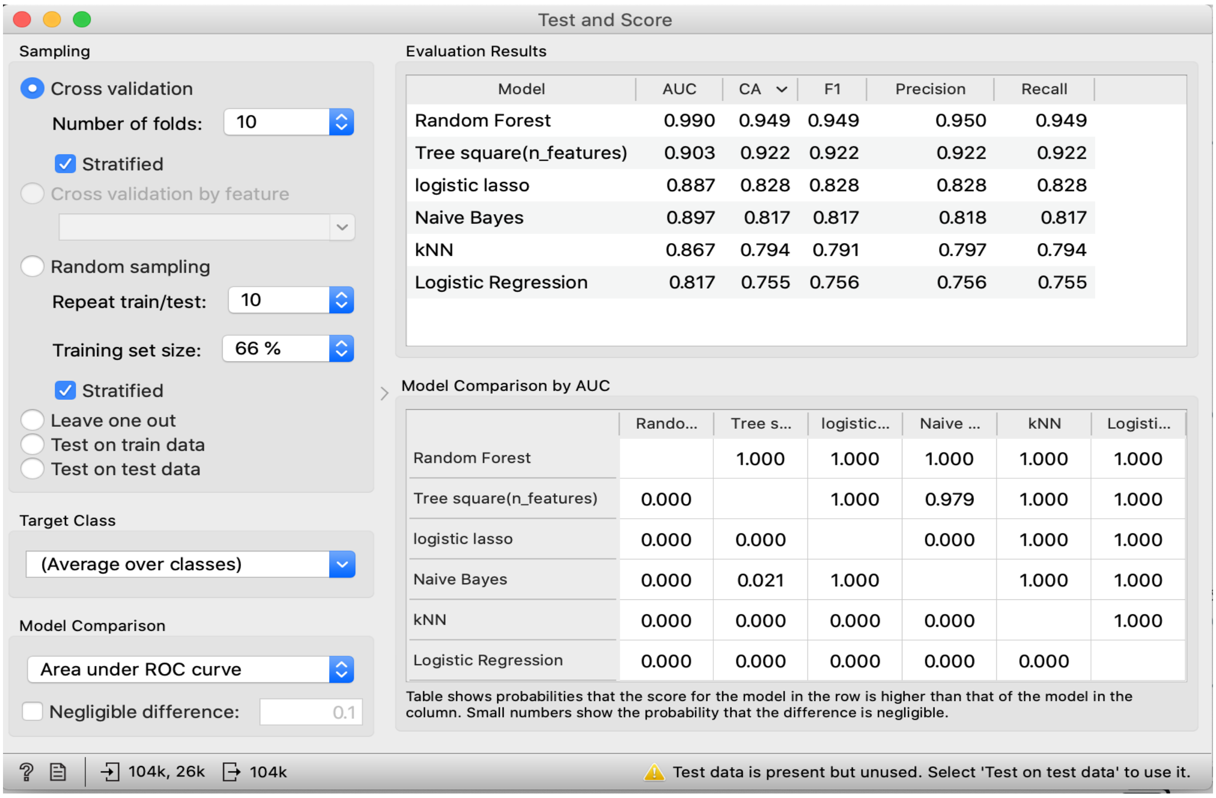
	logistic...	kNN	Tree s...	Rando...	Naive ...	Logisti...
logistic lasso		1.000	0.011	0.000	0.000	1.000
kNN	0.000		0.001	0.000	0.000	1.000
Tree square(n_features)	0.989	0.999		0.000	0.774	1.000
Random Forest	1.000	1.000	1.000		1.000	1.000
Naive Bayes	1.000	1.000	0.226	0.000		1.000
Logistic Regression	0.000	0.000	0.000	0.000	0.000	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

104k, 26k 104k

⚠ Test data is present but unused. Select 'Test on test data' to use it.

Figure 5.30 – Cross Validation Test with Number of Folds "5"



Test and Score

Sampling

● Cross validation

Number of folds: 10

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.990	0.949	0.949	0.950	0.949
Tree square(n_features)	0.903	0.922	0.922	0.922	0.922
logistic lasso	0.887	0.828	0.828	0.828	0.828
Naive Bayes	0.897	0.817	0.817	0.818	0.817
kNN	0.867	0.794	0.791	0.797	0.794
Logistic Regression	0.817	0.755	0.756	0.756	0.755

Model Comparison by AUC

	Rando...	Tree s...	logistic...	Naive ...	kNN	Logisti...
Random Forest		1.000	1.000	1.000	1.000	1.000
Tree square(n_features)	0.000		1.000	0.979	1.000	1.000
logistic lasso	0.000	0.000		0.000	1.000	1.000
Naive Bayes	0.000	0.021	1.000		1.000	1.000
kNN	0.000	0.000	0.000	0.000		1.000
Logistic Regression	0.000	0.000	0.000	0.000	0.000	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

104k, 26k 104k

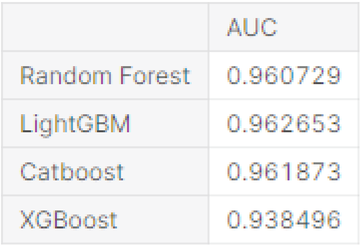
⚠ Test data is present but unused. Select 'Test on test data' to use it.

Figure 5.31 – Cross Validation Test with Number of Folds "10"

6. Related Work

Surveys are one of the most common studies among machine learning practices. In our case we studied on airline passenger satisfaction survey. This subject mostly studied based on the other classes in our dataset. Such as, they did not look for only the satisfaction but they consider the type of travel or etc. and then look for satisfaction. Best way to diminish peoples needs best way to accomplish this is machine learning because it reduces the biased point of view and faster in all kinds of ways. By using the machine learning techniques we aimed to achieve what makes people satisfy in a general manner while choosing an airline. The basics of machine learning is that you need have a big dataset more the better. Here are some relevant works on this dataset.[6],[7].

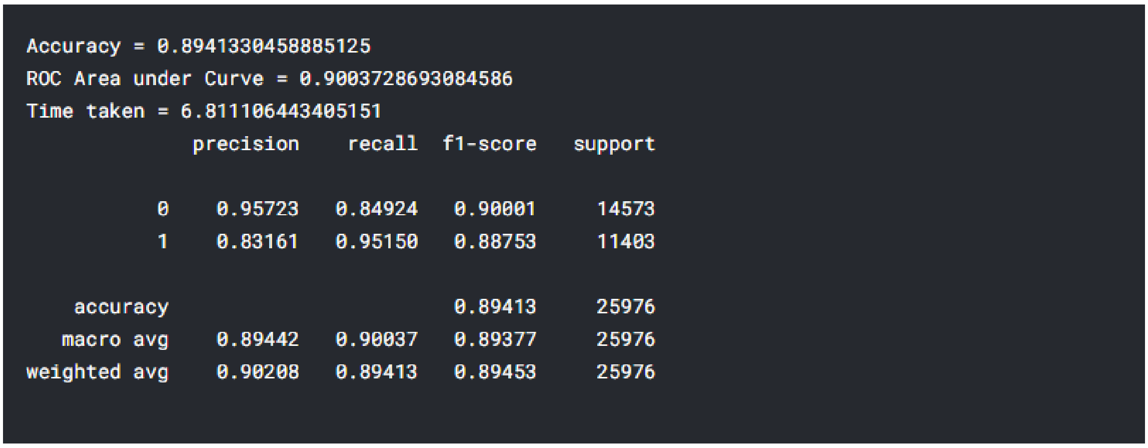
On the first related work[6], the same dataset is being used and although the target is satisfaction they did not eliminate other classes like we do. At first They started with the cleaning process which is the very basic of machine learning and start sorting related and unrelated features from each other. They used encoding for some categorical variables. As we mentioned before according to correlation we did not have to eliminate any feature and this study thinks the same and they keep all the features. However, this study approached model choosing different from us aside from random forest. They compare Random forest, LightGBM, Cat-boost and XGBoost and as matter of fact they all give some great results but they choose random forest as best. Here are the results.



	AUC
Random Forest	0.960729
LightGBM	0.962653
Catboost	0.961873
XGBoost	0.938496

Figure 6.1 – AUC Results

On the second related work [7], again the same dataset is used. This time dataset is studied on thoroughly and cleaned carefully. Before making any assumptions they get rid of the missing variables and then used encoding. After that they get rid of the outliers and then checked the correlation. As the first related work no feature is eliminated in this study too. For the ranking this study used Chi-square. After that they compared 8 different model which are Logistic regression, Naïve Bayes, K-Nearest neighbor, Decision Tree, Neural Network, Random forest, Extreme Gradient Boosting and Adaptive Gradient Boosting. According to their observation random forest and ADA-boost performed equally well but they choose to with random forest because of time it takes is less. Here is the result of random forest.



```
Accuracy = 0.8941330458885125
ROC Area under Curve = 0.9003728693084586
Time taken = 6.811106443405151
              precision    recall  f1-score   support

           0    0.95723   0.84924   0.90001     14573
           1    0.83161   0.95150   0.88753     11403

    accuracy                        0.89413     25976
   macro avg    0.89442   0.90037   0.89377     25976
weighted avg    0.90208   0.89413   0.89453     25976
```

Figure 6.2 – Random Forest Results

7. Conclusion and Future Work

The purpose of this project is to measure passenger satisfaction by examining the results of the survey conducted by US Airlines in order to ensure passenger satisfaction. Classification criteria in machine learning have been used and evaluated to compare and test the classification algorithms. At the beginning of the project, some classification algorithms to be applied to this data set have been selected. KNN, Random Forest, Tree, Logistic Regression and Naive Bayer models have been used in the project. The evaluations are discussed in the Experimental Evaluation section. First of all, the number of features have been examined and unnecessary ones have been removed. In this way, it has been observed that more accurate results are obtained in algorithms. It is also crucial for ROC analysis and evaluating the performance of Confusion Matrix models. After being evaluated in this section, it has been observed that the algorithms used do not give proper results and the Random Forest model is the most suitable model for this data set.

The value of the KNN algorithm used in our model is 0.790, the value of the Naive Bayer is 0.813, the value of the Tree algorithm, it is 0.924, and in the Logistic Regression algorithm, the value is 0.856. On the other hand, Random Forest has a successful result of 0.950. As a result, the ML algorithm chosen in our project is Random Forest. It is an improved version of Decision Trees algorithm and includes many decision trees. It is observed that the decision tree algorithm works well when we have features at completely different scales in Random Forest. In addition, this algorithm is a very powerful methodology for classification problems and is preferred because of its large size.

As a result, we have considered how our model will work in real life. It has usability in other countries and thus it can facilitate the work to be done and the decisions to be made in this area. Also, it may be informative for future studies, in terms of this data set and its features, how the models are compatible with the data set. In addition, the satisfaction data set in US air transportation can be used for a wide range of applications, such as satisfying passengers, increasing the economic profit of companies and providing sustainable competitive advantage.

As a result, we have considered how our model will work in real life. It has usability in other countries and thus it can facilitate the work to be done and the decisions to be made in this area. In addition, the satisfaction data set in US air transportation can be used for a wide range of applications, such as satisfying passengers, increasing the economic profit of companies and providing sustainable competitive advantage.

References

[1] Airline Passenger Satisfaction – *https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction*

[2] Canan Yılmaz Uz (2019) – *Havayolu Hizmet Kalitesinin Yolcu Memnuniyeti Üzerine Etkisi*

[3] Confusion Matrix – *https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce*

[4] Logistic Regression – *https://medium.com/@ekrem.hatipoglu/machine-learning-classification-logistic-regression-part-8-b77d2a61aae1*

[5] Decision Tree – *https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680*

[6] Related Work – *https://www.kaggle.com/teejmahal20/classification-predicting-customer-satisfaction*

[7] Related Work 2 – *https://www.kaggle.com/chandrimad31/flight-passenger-satisfaction-eda-and-prediction*