**ML Classification Assignment – Report** **Studentcode code: 968641**

The analysis is aimed at developing a model for the company to predict customer satisfaction considering some explanatory variables, using f1 as metrics to validate it. Before going through the model creation step, a time-consuming preliminary process of inspection and preparation of the provided dataset is required.

**Data analysis and data preparation**

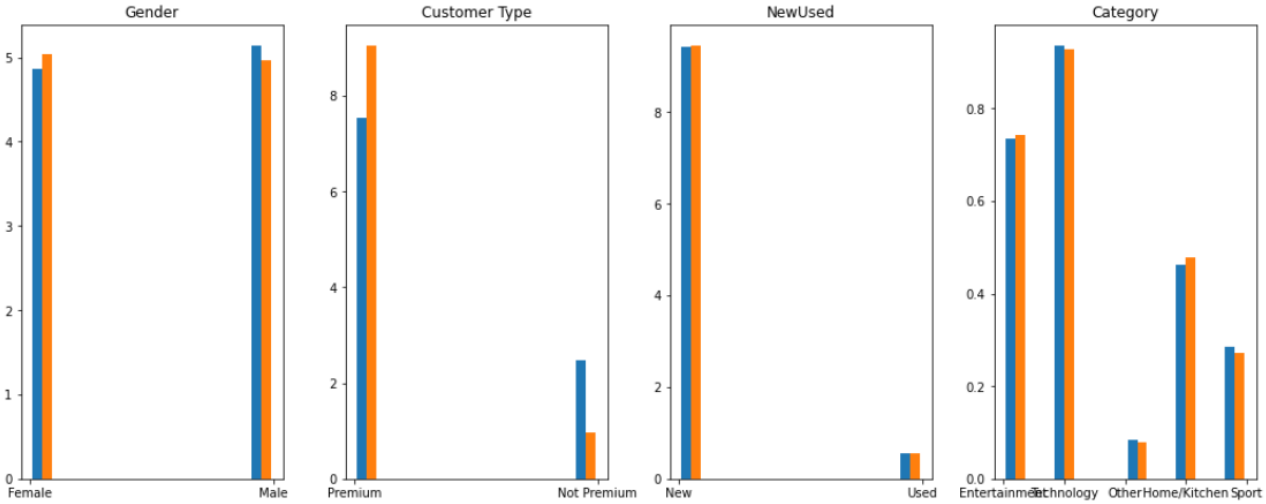
The first step was to analyse in general the whole dataset. It contains 5000 observations and 18 different explanatory variables of which 5 categorical ones ('id', 'Gender', 'Customer Type', 'New Used' and 'Category'). The target variable 'Satisfaction' was converted to a binary numerical variable: 0 = 'Not Satisfied' and 1 = 'Satisfied'.

The only variable containing missing values was the 'Age' variable. If these values are considered separately, they include a higher percentage of non-satisfied customers (85,1%) with respect to the whole dataset (59%). Therefore, NaN values contain important information that can help the company to discriminate between satisfied and non-satisfied customers and for this reason they have to be deeply investigated. Since it was observed that also the percentage of non-satisfied customers whose age is higher than 70 years was pretty high (71,8%), it was decided to substitute NaN values with a high value (120) distant from all the other observations, rather than a very low value.

The next step was to detect duplicates. There were no duplicates in the entire dataset, but it was found an equal identification code associated to two different customers (one is a male and another one is a female with a different age) and for this reason these two observations were deleted, as well as the variable 'id' because it did not add any information.

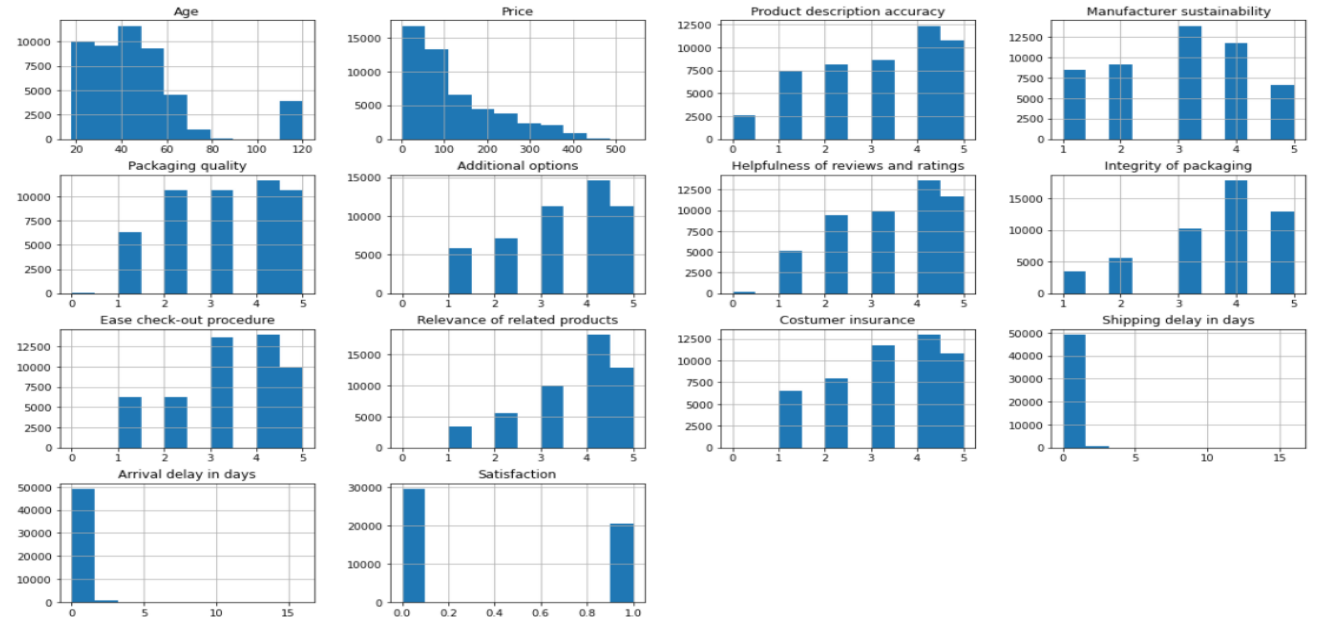
At this point the dataset was split into two sub-datasets: one containing just the categorical variables and the other one containing just the numerical ones and the former dataset was firstly investigated.

To understand which categorical variables were relevant for a discriminant analysis regarding the satisfaction of a customer, four histograms associated to four different categorical variables ('Category', 'Gender', 'New/Used' and 'Customer Type') were plotted. In each histogram the frequency of the satisfied and non-satisfied customers was shown. By looking at them, it was decided not to consider the 'Category' and the 'New/Used' variables since the difference in the frequency of non-satisfied and satisfied customers is negligible.




Then, the remaining categorical variables were converted into numerical ones using dummy variables.

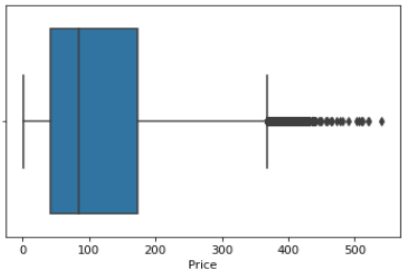
The numerical sub-dataset was firstly analysed by plotting the histograms counting the frequencies for each variable to detect some interesting distributions.



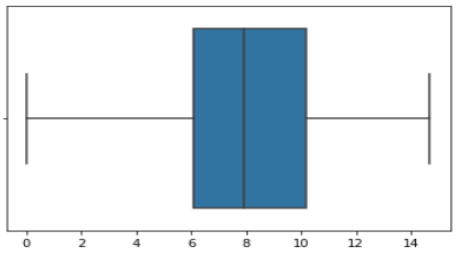

'Age' (notice that 120 is the conventional value discussed above), 'Price', 'Shipping delay in days', and 'Arrival delay in days' presented a left skewed distribution. A box-plot for the last three variables was used to make further considerations. For both the arrival and the shipping delay in days variables, most of the observations were equal to 0. Then, since they seemed to be very similar by looking at the histograms and the box-plots, the correlation coefficient between the two variables was computed (0.91) and it came out that they were strongly positively linearly correlated. Since they both carry more or less the same information, just the 'Arrival delay in days' variable was kept.

A further analysis was conducted considering the 'Arrival delay in days' variable: when it increased, as it was expected, the percentage of satisfied customers slightly decreased. There was a strange situation for what concerned the satisfaction frequencies of the observations associated to 5, 6 and 13 days of delay. Indeed, even though the delay was high, the percentage of satisfied customers increased. This could have introduced some biases in the algorithms developed and, considering the very low number of observations in which the arrival delay was equal or higher than five days, it was decided to remove the related observations.

There was an attempt to transform the price distribution into a normal one through a Box Cox transformation. Even if it was not possible to obtain normality, this transformation allowed to reduce the number of observations falling in the left tail. Before the transformation the 'Price' variable presented a lot of points out of the whiskers of a box-plot, but after the transformation no point felt outside the whiskers.



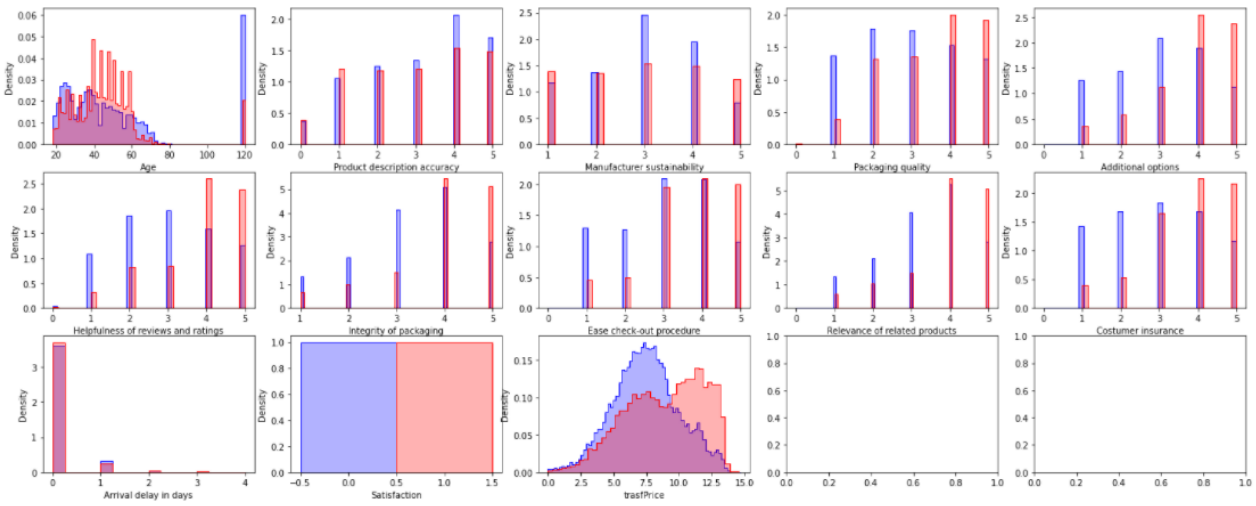

before




after

Then, considering the 'Easy check out procedure' variable, it was noticed that just one observation presented a value equal to 0 and it was deleted.

Following the same line of reasoning used in the categorical dataset, the relevance of each numerical variable for a satisfaction discrimination analysis was studied.



By looking at the histograms only the 'Arrival delay in days' variable seemed to be not so relevant because of the scale, so it was decided to keep it.

The last step of the data preparation process was to standardize all the numerical variables to avoid biases in the algorithms caused by the different magnitude of the values.

**Model training and testing**

At this point data was ready to train the different algorithms. Data was split into training set (75%) and test set (25%). A stratification technique was adopted so that the proportion of the number of satisfied customers and of non-satisfied customers was the same in the training and in the test set.

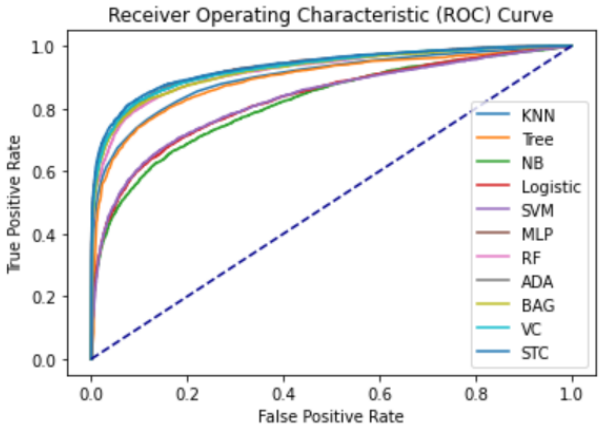
The final goal was to find the best supervised learning algorithm in terms of f1.

To do so a Grid-search was adopted to find the optimal hyperparameters of the different classification models. A threefold cross validation (with k=3 the computational effort is not so high) was used to increase the robustness of the algorithms, in order not to make the results dependent just on the particular split chosen for the dataset. In this way the best combination of hyperparameters for each model refers to the mean of the f1 score metrics.

In some cases it occurred that the hyperparameters found in the grid search led to an overfitting problem in terms of f1 score and for this reason, starting from the grid search results, some hyperparameters were changed to reduce the distance between f1 train and f1 test, since the goal is to predict (as a rule of thumb, it was considered that there was an overfitting problem when the distance was higher or equal to about 4%).

The table below summarizes the metrics obtained using different algorithms, while the graph shows the ROC curves associated to them.

| Model | F1 train | F1 test | AUC |
|---|---|---|---|
| Naive Bayes(NB) | 0.69 | 0.69 | 0.82 |
| Categorical Tree(Tree) | 0.831 | 0.798 | 0.89 |
| Logistic regression(Logistic) | 0.713 | 0.710 | 0.83 |
| Random Forest(RF) | 0,838 | 0,810 | 0.92 |
| Knn(KNN) | 0,808 | 0,788 | 0.9 |
| Mlp(MLP) | 0.837 | 0.830 | 0.92 |
| Adaboosting(ADA) | 0.865 | 0.833 | 0.93 |
| Svm(SVM) | 0.715 | 0.713 | 0.83 |
| Bagging(BAG) | 0.835 | 0.822 | 0.92 |
| Voting classifier(VC) | 0.846 | 0.833 | 0.93 |
| Stacking(STC) | 0.867 | 0.844 | 0.93 |



## Models evaluation and final remarks

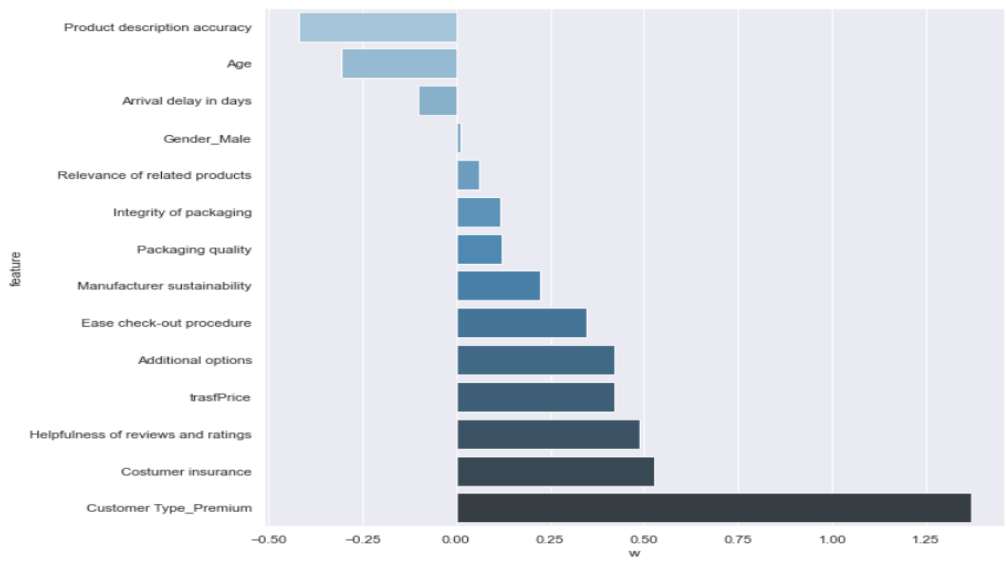
The first consideration to underline is that the hypothesis of independence between the explanatory variables is not always respected (Customer Type and Customer insurance for example) and for this reason the Naïve Bayes classifier is not considered for the final evaluation of the best model. Moreover, even if the performances of the KNN classifier are quite good, the Euclidean distance was strongly affected by the choice of the value to substitute the NaN and this could have led to some biases.

For what concerns the interpretability the Random Forest and the Logistic Regression algorithms showed very interesting insights. 'Gender' and 'Arrival delay in days' variables are those that less influence the customer satisfaction. However, while in the logistic regression model the weight given to the 'Gender' variable is quite low and the one given to the 'Arrival delay in days' is modest, in the application of the Random Forest the opposite situation occurs.

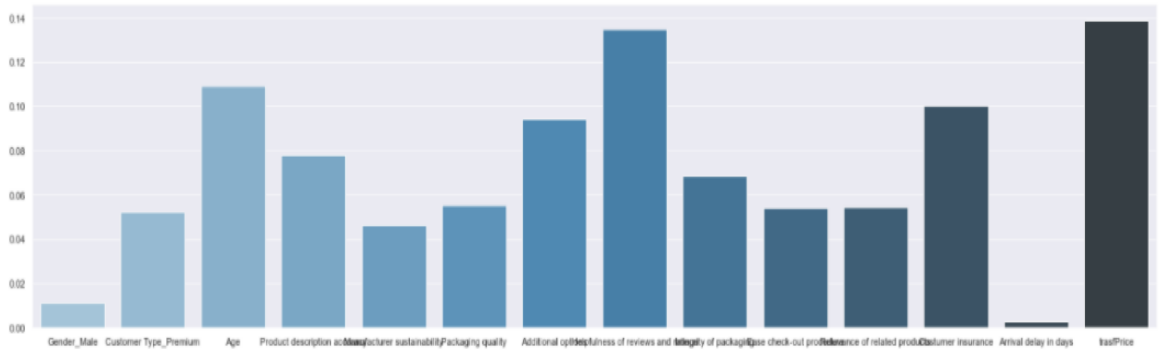
It is important to mention that negative impact of the 'Age' variable on the customer satisfaction shown in the logistic curve graph was accentuated by NaN values that were substituted with 120.

In the end it was noticed that the more the production description accuracy was high the more the customers were non-satisfied, maybe because it could have increased their expectations regarding the product.

*Logistic Regression*



*Random Forest*




For what concerns the Classification Tree algorithm, even though one of its main strengths is the interpretability, in the analysis the excessive ramifications of the best tree model (max depth=18) did not allow to draw interesting conclusions.

The algorithms that performed better in terms of f1 test and AUC were:

- Multi-layer perceptron classifier

  *Ensemble methods*:
- Adaboost classifier
- Random Forest
- Bagging with 100 multi-layer perceptron classifiers
- Voting classifier considering the best models obtained: Random Forest, KNN, Multi-layer perceptron classifier, Adaboost and Bagging, with a 'soft' voting procedure, that returns the class label as argmax of the sum of the predicted probabilities.
- Stacking classifier considering the same algorithms used in the Voting one but without the Bagging classifier. The Stacking classifier consists in stacking the output of each estimator and use a classifier to compute the final prediction, which in this case was the Logistic Regression classifier.

From a computational effort point of view the ensemble methods and the SVM are the most expensive. In particular, it was not possible to exploit all the potentialities of SVM because of time constraints.

In order to choose the algorithm to adopt for the prediction phase a trade-off between the computational effort, the interpretability, the f1 test and the AUC must be considered. The Logistic Regression and the Random Forest should be preferred for their high interpretability. The KNN and the Decision Tree classifier showed good results with a very low computational cost. The Adaboost, the Voting and the Stacking

classifiers were the best in terms of performances; the drawback was represented by the high computational effort.

Taking in mind all these considerations and the goal set, the Stacking classifier is the one selected to predict future observations, because its high performances outweigh the low interpretability and the computational effort.