

4YFN-MWC DATATHON

BARCELONA 2022



1 de febrero de 2022

María mellado palacios

**INDEX**

1. Main Goals…………………………………………2
2. Background………………………………………...2
3. The dataset…………………………………………2
4. Exploratory Data Analysis (EDA)…………………3
5. Supervised Classification………………………….6
6. Hyperparameter tuning and cross validation……….7
7. Implementation…………………………………….8
8. Conclusions………………………………………...9

**Main Goals**

The main goal of this project is to create a model that can predict the segment to which the client belongs to. To achieve this goal, it is necessary to analyse sales and client activity, to evaluate the impact of the promotions.

**Background**

Nuwefruit is a start-up that is willing to revolutionize society habits by fomenting daily consumption of fruit. For this reason, the company is specialized in door-to-door fruit sales. This is possible thanks to their optimization algorithm, that allows the company to have low logistic costs and hence low fruit prices.

**The dataset**

For this challenge, two datasets are used, 'CLIENT TABLE', that contains client data and 'ORDERS TABLE', containing information about orders. The datasets contain the following features:

'CLIENT TABLE':

* **CLIENT ID**: unique client identifier.
* **CLIENT\_SEGMENT**: client segment
* **AVG CONSO**: annual average client consumption calculated for last months of 2020.
* **AVG BASKET SIZE**: average size of the client’s basket calculated for last months of 2020.
* **RECEIVED\_COMMUNICATION**: Client received product promotions (1) or the client did not received product promotion (0).

'ORDERS TABLE':

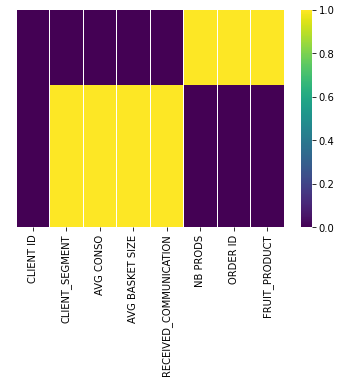
* **CLIENT ID**: unique client identifier.
* **NB PRODS**: Number of 'prods' of the variety of fruits in the order (1 prod = 10 fruit pieces)
* **ORDER ID**: unique order identifier.
* **FRUIT\_PRODUCT**

**Exploratory Data Analysis (EDA)**

In first place, the two datasets were put together in one DataFrame for simplicity. After the creation of the DataFrame, features type and missing values were identified. As it is depicted in the following table, the DataFrame contains three object features, one integer and four float features.

|  |  |
| --- | --- |
|  | **Feature Type** |
| **CLIENT ID** | Integer |
| **CLIENT SEGMENT** | Float |
| **AVG CONSO** | Object |
| **AVG BASKET SIZE** | Object |
| **RECEIVED COMMUNICATION** | Float |
| **NB PRODS** | Float |
| **ORDER ID** | Float |
| **FRUIT PRODUCT** | Object |

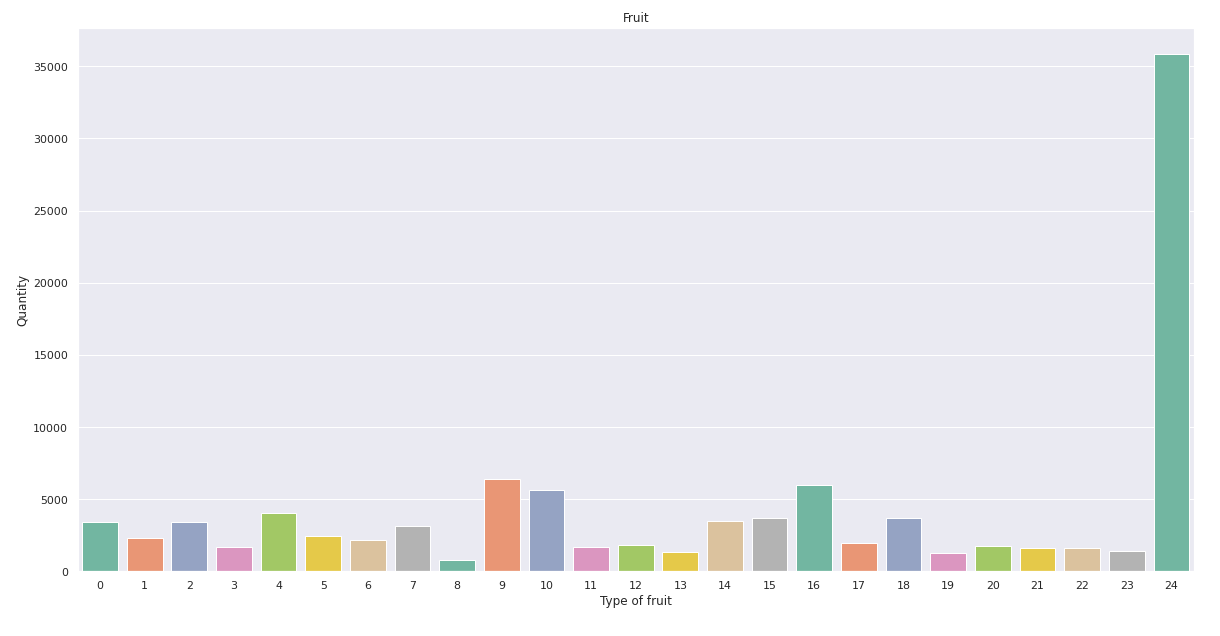
In the following heatmap columns with missing values are depicted. Except the column Client ID, there are missing values in all the columns.



|  |  |
| --- | --- |
|  | **Number of Nan** |
| **CLIENT ID** | 0 |
| **CLIENT SEGMENT** | 66912 |
| **AVG CONSO** | 66912 |
| **AVG BASKET SIZE** | 66912 |
| **RECEIVED COMMUNICATION** | 66912 |
| **NB PRODS** | 35884 |
| **ORDER ID** | 35884 |
| **FRUIT PRODUCT** | 35884 |

The are four columns that have more missing values than real value so mean, or mode imputation would not make sense in this case. Therefore, the missing value imputation is done in this case by applying KKN imputer. For this purpose, 5 neighbours were used and nan euclidean metric.

After missing values imputation, there rest of the features were studied. In the following countplot diagram type of fruits are shown. From this column, 24 types of fruits were identified and the percentage of each one of them were calculated.

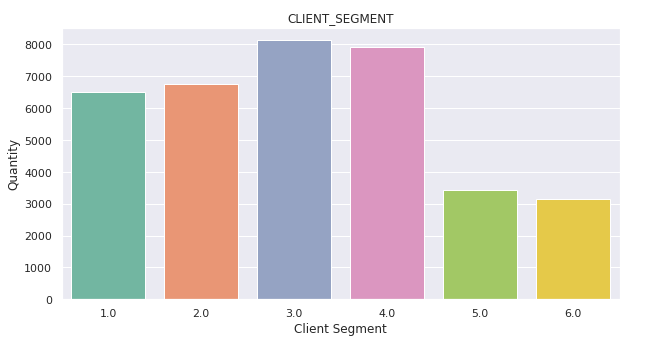


|  |  |
| --- | --- |
|  | **Percentage (%)** |
| **Apple** | 0,03 |
| **Orange** | 0,06 |
| **Kiwi** | 0,06 |
| **Pear** | 0,04 |
| **Cherry** | 0,04 |
| **Watermelon** | 0,01 |
| **Strawberry** | 0,02 |
| **Nectarine** | 0,03 |
| **Grape** | 0,03 |
| **mango** | 0,02 |
| **Blueberry** | 0,02 |
| **Pomegranate** | 0,02 |
| **Nuwe Fruit** | 0,04 |
| **Devil Fruit** | 0,02 |
| **Plum** | 0,02 |
| **Papaya** | 0,02 |
| **Jackfruit** | 0,01 |
| **Pineapple** | 0,01 |
| **Lemon** | 0,06 |
| **Lime** | 0,02 |
| **apricot** | 0,02 |
| **Coconut** | 0,02 |
| **melon** | 0,01 |
| **Banana** | 0,03 |

For facilitating future modelling, some features were label encoded:

* Fruit Product
* Avg basket Size
* Avg Conso
* Received Communication
* Client Segment

Countplot representation of the target feature (Client Segment) is depicted in the following figure. It is seen that there is no unbalanced data, and the mode is client segment 3. As there are six classes under this feature, it is a multiclass classification problem.



**Supervised Classification**

To perform supervised classification, variables X and y were created. Three features were excluded from X variable (client\_id, order\_id, and client\_segment) and the target feature was excluded from y variable (Client Segment). To evaluate the performance of the classification algorithms, precision, recall and f1-score was evaluated for each label. These three parameters can be defined as:

Where:

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

Random Forest Classifier was applied to the data, obtaining an overall score of 85,99%. I the following table, the classification report of the model results is depicted. From this table, it can be concluded that labels 0 to 4 are well classified but labels five and six have low precision and f1-score (for label 5) and very have precision and f1-score (label 6).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **f1-score** |
| **0** | 0,55 | 0,67 | 0,61 |
| **1** | 0,65 | 0,69 | 0,67 |
| **2** | 0,75 | 0,64 | 0,69 |
| **3** | 0,62 | 0,58 | 0,6 |
| **4** | 0,42 | 0,65 | 0,51 |
| **5** | 0,21 | 0,08 | 0,11 |
| **6** | 1,00 | 1,00 | 1,00 |

**Hyperparameter tuning and cross validation**

The performance of Random Forest classifier was optimized by hyperparameter tunning and cross validation. For this purpose, the parameters that were introduced in the Grid Search were:

Max depth: 70,80,90,100

N estimators: 900,1000,1100

For cross validation 3 k folds were defined. Overall, the fitting was performed with 3 folds for each of the twelve candidates, totalling 36 fits. From this fit, the best parameters were estimated as:

Max depth:80

N estimators: 900

After Grid search CV, Random Forest classifier was performed again, this time including the best parameters for fitting:

Class weigh: balanced

Max depth:80

N estimators: 900

In this case, a score of 90.75 % was obtained, being a much higher score than the one obtained before. In the following table, the classification report is shown. One can observe from this table that precision, recall and f1-score of labels 0-4 are in the range of 70-80 % in most of the cases. Nevertheless, for labels five and six very low and very high values are obtained, respectively.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **f1-score** |
| **0** | 0,85 | 0,87 | 0,86 |
| **1** | 0,81 | 0,85 | 0,83 |
| **2** | 0,78 | 0,73 | 0,75 |
| **3** | 0,73 | 0,67 | 0,7 |
| **4** | 0,63 | 0,94 | 0,75 |
| **5** | 0,17 | 0,11 | 0,13 |
| **6** | 1,00 | 1,00 | 1,00 |

**Implementation**

The last dataset (test\_x) was provided for predicting client segment. This dataset contains 19 rows with no missing values and the following features:

* Client ID
* AVG Conso
* Avg basket Size
* Received Communication

As the Dataset does not contain features “FRUIT PRODUCT” and NB\_PRODS”, two columns with fictitious data were created. In the case of fruit product these values were in the range from 0 to 24 and in the case of Nb Prods from -820 to 198.

Attending to object features there were two object variable that were label encoded: AVG Conso and Avg Basket Size. The feature “Received Communication” was transformed into object type.

After this quick EDA, the variable X was created excluding feature “Client ID” and the prediction was calculated with model created before. The resulting prediction was saved in a csv file.

**Conclusions**

It is already well known that machine Learning is an excellent tool for marketing and client segmentation. Client segmentation can help companies save time and money being k-means the unsupervised machine learning algorithms more often used for this purpose. In the present work, a classification model was created, that would successfully perform client segmentation with an overall accuracy higher than 90%.

Due to lack of time, there are several tasks that were not completed and that could improve tremendously the overall performance of this project. Some of these tasks included extended grid search cv. In this project, grid search was performed on Random Forest. However, if more parameters were included in this fit, we could have obtained better results. Besides, performing cross validation with higher k folds could have led as well to better results. Moreover, labels five and six from target feature were not correctly classified in this project. For larger precision values of these two labels a pipeline can be created where two labels are excluded from the dataframe and treated separately.

Client segment models in this project were preformed by using just two classification techniques: Random Forest and Ada Boost. Nevertheless, there are other algorithms that can be used for this purpose and that were not explored this time.

On the other side, not all classification predictive models support multi-class classification. Some algorithms such as the Perceptron, Logistic Regression, and Support Vector Machines do not natively support classification tasks with more than two classes. Instead, other methods can be used to split a multi-class classification problem into multiple binary classification datasets and train a binary classification model each. Two of these methods include: One vs One (OvO) and One vs All (OvA).