Data Mining

Assignment 1: Classification

1 Ma INF 2022-2023

Saartje Herman

April 16, 2023

1 Files

All the files can be found in my github repository.

The solution of this project can be found in the jupyter notebook file **DataMining_Assignment1.ipynb**. The selection of customers to send a promotion can be found in the file **selection.txt**.

2 Preprocessing steps

2.1 LabelEncoder

In order to train a classifier with the given data (*existing-customers.xlsx*), the data needed to be encoded. I used the LabelEncoder to encode the target labels with value between 0 and $n_{classes} - 1$. This causes the features to be represented as integers instead of categories.

2.2 Missing Values

To deal with the missing values in the given data, I used Multivariate imputer that estimates each feature from all the others (A strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion). I chose this imputer because it gave better accurancies of the classifiers.

3 Testing Classifiers

To predict the 'class' label of the data *potential-customers.xlsx*, I tried out different classifiers. The classifiers I tried are: Decision Tree, Categorical NB, K-nearest Neighbors, Random Forest, Bagging combined with Decision Tree,

AdaBoosting combined with Decision Tree, Gradient Boosting and HistGradient Boosting. After training and testing these classifiers, HistGradient Boosting seemed to have the best accurancy (approximately $87,7\%$) so I used this classfier to predict the 'class' label of the data *potential-customers.xlsx*.

4 Predicting potential customers

To predict the 'class' label of the potential customers, the same preprocessing steps (LabelEncoder and dealing with missing values) needed to be applied. After doing these step, the HistGradient Boosting is used to predict the label. Result : 3198 people are classified with a high income and 13083 people are classified with a low income. I also applied the method *predict_proba* (probability estimates) to make a list of id's of people to send the promotion to.

5 Revenue Calculations

To calculate the expected revenue, I divided the calculation into 2 main scenarios: the best cases and the worst cases.

5.1 Best Cases

If we chose the people labeled with low income with the lowest probability estimates, there is still a chance that they actually have a high income. In the best case, the 5% of the people labeled low income, have all a high income.

The first approach assumes that we send a promotion to 10% of the people labeled as high income and that all of them accept the promotion. It also assumes that we send a promotion to 5% of the people labeled as low income but that they actually have a high income and that also all of them accept the offer.

The second approach assumes that we send a promotion to 10% of the people labeled as high income and that all of them accept the promotion. It also assumes that we send a promotion to 5% of the people labeled as low income and that also all of them accept the offer.

5.2 Worst Cases

The first approach assumes that we send a promotion to 10% of the people labeled as high income and that only 10% of them will accept the promotion. It also assumes that we send a promotion to 5% of the people labeled as low income but that they actually have a high income and that also only 5% of them accept the offer.

The second approach assumes that we send a promotion to 10% of the people labeled as high income and that only 10% of them will accept the promotion.

It also assumes that we send a promotion to 5% of the people labeled as low income and that also only 5% of them accept the offer.

5.3 Total revenue

The total expected revenue will be the average of these 4 scenarios together namely, approximately 246737.59 euro.

6 ID's of people to send promotion

In this last section, we select the people to who we want to send the promotion. Firstly we take all the people where the label of high income has the highest probability of correctness. Secondly, take all the people where the label of low income has the lowest probability of correctness. These id's are saved in the file **selection.txt**.