David Pérez Carrasco - 241375
Guillem Escriba Molto - 242123

upf. Universitat Pompeu Fabra Barcelona

Project Report

[EX1]

First of all, we are going to focus just on the extracted variables, since there are more in the original file, but we are going to work just with the extracted ones.

The City variable is the more special one, since it uses the String format, or list of characters, but it is stored as an Object. This is because Pandas recognise a Datatype whose all instances have the same number of bytes, 8 bytes for int64 and float64 for example, but in the case of the String each one has a different size, so each one needs a different number of bytes, and since all the strings of the dataset are not of the same size,Pandas cannot recognize that as a String, and says that's an Object, which will not help us to use that column, since an Object cannot necessarily be a String and hence we could not make operations with them. On the other hand, Revenue and the Mobile Potential are floats and the rest are integers, although all of them except Customer Flag and CNT_EMPLOYEE (which correctly are stored as integers) are stored as floats with no decimals in the data extraction.

The Sector Variable has some Nulls, although in relation to the total number of instances it is practically negligible, and the same happens with the legal_form_code, which has even less Nulls. Moreover, CNT_CB_DENSITY, CNT_CB_MOB_DENSITY and CNT_CB_FN_DENSITY have multiple Nuls, while City, Customer_flag, CNT_EMPLOYEE and Mobile_potential have all no Nulls.

Finally we have to say that the size of the complete dataset is 13335 entries, and there originally are 10 variables taken into account.

[EX3]

As we can see, the customer boxplot is more compact than the noncustomer one, we can notice how the non-customer one has a lot more of outliers and the IQR box is more expanded for customer boxplot, which means more regular data. All of that makes sense since customers would have more similar variables because they all have in common that they are users of the same company, while noncustomers could be from really different profiles, which will cause a less effective boxplot.

We also notice that Customer dataset have higher CNT_EMPLOYEE, since both the quantiles and the medium values are higher in Customer boxplot, although nonCustomer data set has outliers that reach the same maximum value. When it comes to Revenue outliers, non-customer (234) have more than customer (35). The quantiles for Revenue are Q1= 903986 and 1047500, Q2=1750000 and 2200000, Q3= 3501123 and 4195000 for

non-customers and customers respectively, and for the Mobile potential variable we obtain Q1= 1975 and 2090, Q2=2278 and 2401, Q3= 2632.
We can notice that when deleting the outliers the boxplot takes a more uniform shape, where the boxplot represents the most part of the dataset, and not the outliers as happened before. Moreover, now the quantiles have changed, and indeed the new outliers have appeared, but the plot is more representative now.

[EX5]

Once we have calculated the ratios of the City variable, we see that for the Customer dataset there are several cities with more than 1%, while in the Non-Customer dataset all the categories are below 1%, which shows that the customers are more centralized in some cities, probably because the company is located in some particular cities, ant then the non-customers could be anywhere else.

[EX6]

We get a X_train size of 80% and X_test size of 20%, which is perfectly aligned with the test_size value selected (0.20)

[EX7]

Both datasets, y_trained and y_test are imbalanced as we can see in the histogram. The number of noncustomers (Target 0) is several times higher than customers due to this it is possible that the model becomes overfitted and could tend to classify the instances to class 0, noncustomers. For this reason it is necessary to obtain a balanced sample to avoid overfitting.

[EX12]

We can see that Decision Tree has more accuracy than the SVM model and it also has better precision. The recall is very similar in both models but is slightly better in SVM. I would recommend the SVM although it is less accurate, the better recall is more important to classify both classes.

[EX13]

Although it is less accurate than the non ensemble models we can see that has a better recall than svm, this is because if there are more class 0 (noncustomers) than class 1 (customers) and the model tends to classify every instance to class 0, it would have a good

recall of class 0 and a good accuracy due to the limited number of class 1. This model is not better than the previous model due to the distribution of the recall, despite of class 0 has a good recall, class 1 has a low recall value.

[EX14]

The Random Forest model is the best option that we have used until now, it has higher accuracy than previous models (using a balanced dataset) and it has a better recall so it would be the optimal model.

[EX16]

As we can see this model is even better than Random Forest, having higher accuracy, higher precision and higher recall so it makes Gradient Tree Boosting the best option to choose in order to create an optimal model to predict and classify Targets.

[EX17]

In the plot we can see that this histogram is more balanced than the previous one and has higher probabilities. While the first histogram only goes from 0'2 to 0'8, the second one goes from 0 to 1 making possible to predict with high precision the classification of the instance, for this reason the Gradient Tree Boosting is better.

[EX18]

The noncustomers that we would send to the sales manager are those who have been classified as class 1 (customer) without being a customer in fact, this means that those customers have a similar pattern of features as real customers what makes them potential customers. Depending on the cut-off value, the number of targets (noncustomers with customer pattern) vary. For example with a cutoff of 0'5 we would have targets as we can see on the confusion matrix, this is the highest value that we have tried, we could use a smaller cutoff but then the probability of success decreases. A good balance between both features would be 0'65 cutoff with 23 targets.

[EX19]

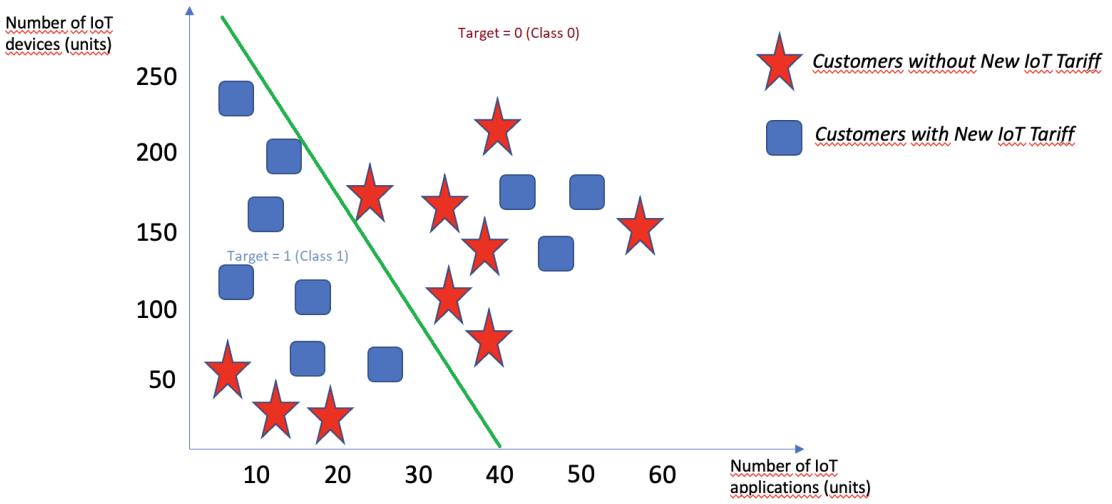
The top 3 features to discriminate between noncustomers and customers are (by order of importance): Mobile_potential, CNT_CB_DENSITY and Legal_Form_Code.

[EX20]

For the new campaign our target=1 samples would be the customers who have our New IoT tariff and the target=0 the rest of our customers. With that classification we will find those customers who have a similar pattern than the customers that have acquired our new tariff, making them potential targets. Adding more data, despite increasing computational cost, could improve the campaign giving to the model more features to work with. However it can also overfit the model so we should be careful processing the new data. Also, the choice of including the new variables would depend on the tariff, if it is so related with these variables we have to include the new data. If the new tariff is very popular among our customers the dataset will be unbalanced because there will be so many customers who already have our new tariff so customers with and without the tariff will be a similar amount.

The customers of our new tariff, target=1, follow the next pattern: high number of IoT devices with low number of IoT applications, so it seems that number of IoT devices is positively correlated with being a new tariff customer and having a high number of IoT apps is negatively correlated with being a new tariff customer. In other words, the higher amount of devices and the lower amount of applications implies higher probability of becoming a new tariff customer and vice versa.

The Scatter plot shows us 20 instances, 10 of Customers with the New Tariff and 10 without it so we can conclude that the dataset is balanced. We can clearly see that the plan which separate both classes would be the following line:



According to the plane, the customers to be phoned are the ones who are in the left-bottom side of the plot, the three customers who are below 20 IoT apps and aren't already customers of the new tariff. As we can see those customers who will be phoned would be classified as target=1 by the model due to its similar pattern to the current customers, but

also the right-top ones who are customers of the new tariff will be classified as target=0 so we can estimate the recall of class 0 (True positive = 7) as (10-3)/10 = 70% and the same for class 1, and precision will be calculated for both classes as True positive/(True positive + False positive (10-3)/(10`+3) = 53’85%.

We hereby declare that, except for the code provided by the course instructors, all of our code, report, and figures were produced by ourselves.