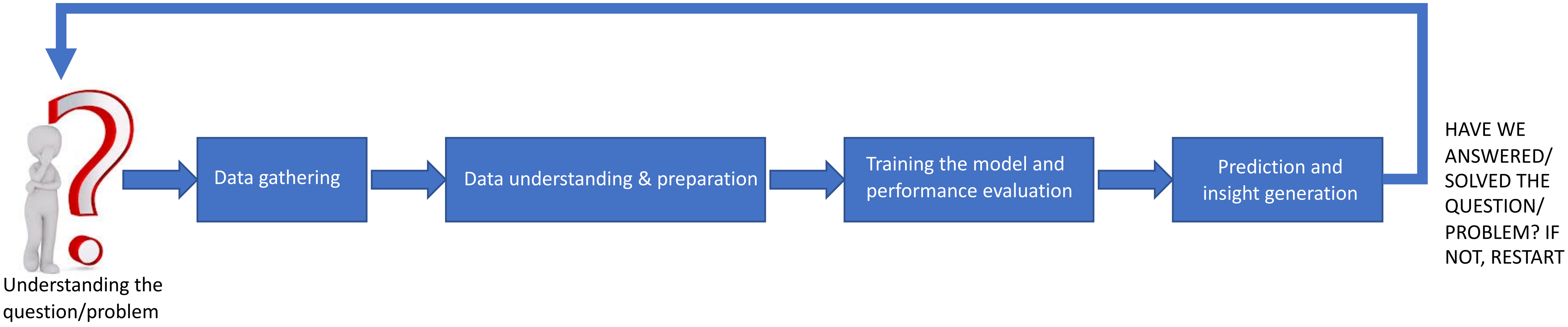
# PROJECT 2: Application of classification in **Marketing and Sales**

***THE GOAL:***

Identify a set of non-customers to sell a Internet-of-Things product of our startup **by means of a machine Learning classifier.**

***METHOD:***

During this project we will follow the common end-to-end Machine Learning process: from understanding the problem, data gathering and cleaning, exploratory data analysis, feature engineering and finally, training and prediction.



Understanding the question/problem

Data gathering

Data understanding & preparation

Training the model and performance evaluation

Prediction and insight generation

HAVE WE ANSWERED/ SOLVED THE QUESTION/ PROBLEM? IF NOT, RESTART

# PROJECT 2: Application of classification in **Marketing and Sales**

Understanding the question/problem

Data gathering

***UNDERSTANDING THE PROBLEM:***

We work as a head of data science and AI in a new Internet of Things (IoT) company. Our company designs, builds and implements wireless IoT products.

Our marketing colleagues are planning to launch a new commercial campaign for capturing new customers. We have to decide which companies are the target to be visited by our sales managers. As the cost to send a sales manager to visit a potential customer is quite high, we have to select from the total market base, those companies that are more likely to buy any of our products and become a new customer.

We will manage key evaluation aspects of a classification as **recall, precision, false positive, etc...** to decide which are the best potential customers.

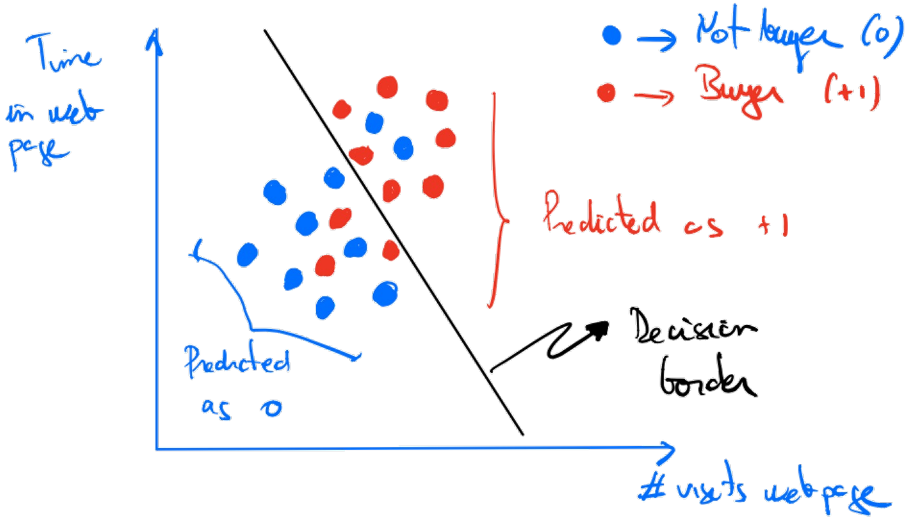
Let's do a recap….

# PROJECT 2: Application of classification in **Marketing and Sales**

Data gathering

Understanding the question/problem

## *UNDERSTANDING THE PROBLEM:*



|  |  | Prediction |  |  |
|---|---|---|---|---|
|  |  | 0 | 1 |  |
| Actual | 0 | 8 | 2 | =10 |
|  | 1 | 3 | 7 | =10 |
|  |  | =11 | =9 |  |

True negative

False positive

False negative

True positive

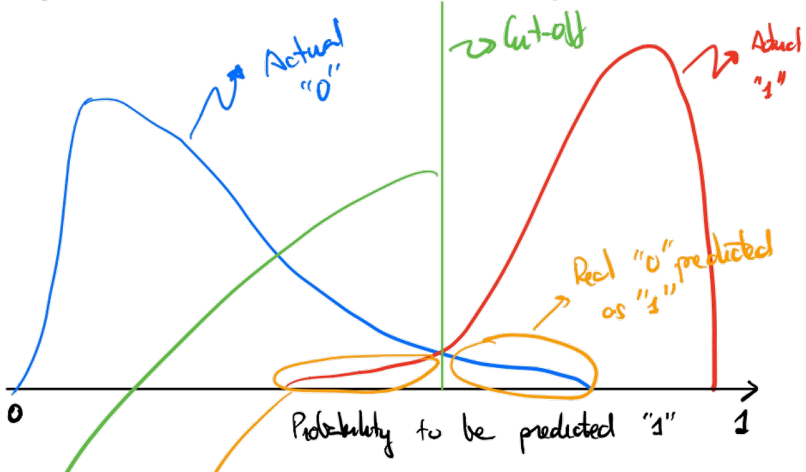
Recall of "1" = 7/10

Recall of "0" = 8/10

Precision of "1" = 7/9

Precision of "0" = 8/11

If we draw the histograms for both classes:



Due to real "1" that overlaps with real "0"

The cut-off is the probability threshold to consider the prediction as "1" or "0"

# PROJECT 2: Application of classification in **Marketing and Sales**

Understanding the question/problem

Data gathering

***DATA GATHERING:*** The data extracted from our Data WareHouse is:

- *Sector*: It's an integer that identifies the sector of the company's activity
- *Revenue*: The annual incomes of the company
- *City: Name of the city where the company is located*
- CNT_EMPLOYEE: Number of employees of the Company
- CNT_CB_DENSITY: Number of companies close
- CNT_CB_MOB_DENSITY: Number of companies with mobile services
- CNT_CB_FN_DENSITY: Number of companies with fixed services
- Legal_Form_Code: It's an integer that identifies the legal type of the company: big, small or medium company
- Mobile potential: It's an estimation of the total annual expense that a company can do in telco services, including IoT
- Customer_Flag: It is a flag that is 1 for current customers and 0 for non-customers

# PROJECT 2: Application of classification in **Marketing and Sales**

Data understanding & preparation

***DATA UNDERSTANDING AND PREPARATION***

Once we know the problem to solve, the next stage is to have a clear understanding of the data we have extracted and to prepare it before modelling. In particular, we will:

- List and verify the type of each variable (object, float, int…). Identify variables with nulls. Measure the memory usage
- Eliminate rows with nulls in order to have a dataset 100% fulfilled
- Exploratory Data Analysis to understand main statistics (mean, standard deviation, min&max values and 25%-50%-75% quartiles) and distribution of the most relevant variables. In this case, we will do the EDA using visualization techniques as **box plots comparing customers and noncustomers.**
- **Remove outliers**
- Transform non-numerical variables using any of he most common transformations **as integer** or **coding as dummies**
- **Create the label or Target variable: customer=1 and noncustomer=0**
- **Create the training and test datasets: 67% for training and 33% for test**
- **Analyze the balancing of both classes in both datasets**

Once this part of the Project is done, we should achieve a deep knowledge about the data. Besides, the dataset will have been processed to be ready to apply the modelling stage.

# PROJECT 2: Application of classification in **Marketing and Sales**

Training the model and performance evaluation

***TRAINING THE MODEL AND PERFORMANCE EVALUATION***

Now we are ready to enter in the training stage of the machine learning models. The common way to procedure is starting with baseline models (i.e. SVM, Decision Trees, etc....) and later, try to improve it adjusting hyperparameters of the models or creating more complex models architectures as ensembles:

- **Baseline of models: Training and evaluation:** In this section we are training a SVM and Decision Tree algorithms. In particular:
  - we will use the X_train and y_train datasets
  - Later on we will evaluate the performance (i.e. **accuracy**, **confusion matrix**, **recall** and **precision**) of each model with the testdataset, i.e. X_test and y_test. To evaluate and compare algorithms, we will apply **KFold cross-validation**

- **Improve the model:** In this section we will test several strategies to improve the performance as adjusting the balance of classes, doing the fine tuning of hyperparameters of the models (e.g. type of Kernel in SVM or minimum samples per leaf in Decision Trees) or using other advance techniques as ensembling. In our exercise, we will practice with **forcing the balancing of both classes and building ensembles of models as bagging, boosting and stacking (or voting).**

# PROJECT 2: Application of classification in **Marketing and Sales**

Insights Generation

***PREDICTION AND INSIGHTS GENERATION***

Once we have the model selected, we are ready to do the prediction. To do it, in this section you test several prediction according to several values of **the cut-off of our model** (i.e. the probability threshold to consider the prediction as 1).

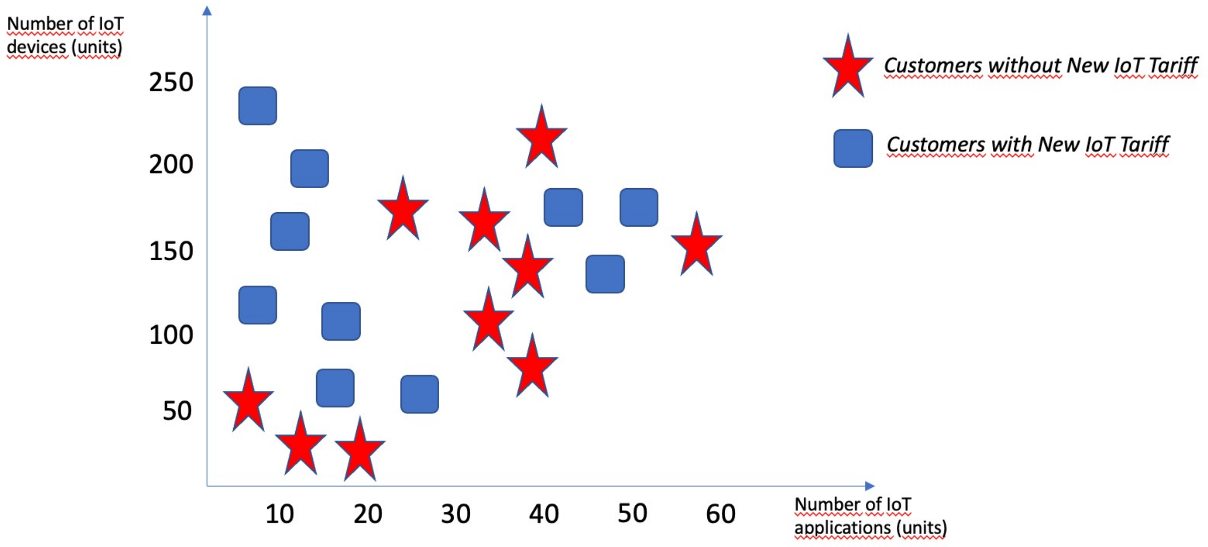
Furthermore, the most important features for the classifier we have selected gives us some ideas about what parameters distinguish a customers from a non-customers.

# PROJECT 2: Application of classification in **Marketing and Sales**

***THINKING "OUT-OF-THE BOX"***

Consider a new campaign focused on accelerating the sales of a **New IoT tariff** to our customers and the scatter plot of the figure. Answer the following questions:

- Which is the target? Which are **target=0** samples? And **target=1**?
- Will the training dataset be balanced or unbalanced?
- Describe in terms of Number of IoT devices and Number of IoT applications traffic the pattern of target 1 customers
- Draw a plane to separate both classes
- According to the previous plane, which are the customers to be phoned to sell the New IoT tariff?
- Could you estimate the precision and recall of the classification?


Number of IoT devices (units)
Number of IoT applications (units)
Customers without New IoT Tariff
Customers with New IoT Tariff