Week 8 deliverable

Group name: LatinosDS

Name: Sebastián J. Castro, Ignacio Solórzano, Freddy Tapia, Omar Jazouli

Country: Argentina, Mexico, Spain

Specialization: Data Science

Problem Statement: ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

The data provided is a csv file that contains the following information:

Clients Data

The bank provide us with a data of almost 41 K clients which were contacted during a marketing campaign. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

1. age: int
2. job: type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. marital: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. education:(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5. default: has credit in default? (categorical: 'no','yes','unknown')
6. housing: has housing loan? (categorical: 'no','yes','unknown')
7. loan:has personal loan? (categorical: 'no','yes','unknown')

Data related with the last contact

1. contact: contact communication type (categorical: 'cellular','telephone')
2. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
3. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
4. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a

call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Data related with the last campaign

1. campaign:number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
3. previous: number of contacts performed before this campaign and for this client (numeric)
4. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Data related with social and economic context

1. emp.var.rate: employment variation rate - quarterly indicator (numeric)
2. cons.price.idx: consumer price index - monthly indicator (numeric)
3. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
4. euribor3m: euribor 3 month rate - daily indicator (numeric)
5. nr.employed: number of employees - quarterly indicator (numeric)

Result of the current campaign

1. y - has the client subscribed a term deposit? (binary: 'yes','no')

Problems found:

- The dataset contains several features with an “unknown” value for some instances. This value will be treated as a NaN value.
- The target variable y is imbalanced.
- The feature duration is left-skewed.
- There are many outliers in many features.

In order to overcame these problems, we will try different approaches. For NaN values we will analyze them and determine what is the best solution: imputation or deletion. We will experiment with these two possibilities and we will determine what it the best one for this case. For outliers we will try two approaches: IQR and WOE. IQR and WOE are both the best approaches for treating outliers as part of data preparation for machine learning algorithms. Besides, for overcoming the unbalanced target we will used SMOTE technique, which is one the most common solutions for this kind of problem. Best performace have been observed in machine learning algorithm after implementing SMOTE for oversampling.