**Capstone Project 1 - Milestone Report**

**1. Problem**

*“Predict if the client will subscribe a term deposit.”*

To drive the sales of the enrollment for a campaign (term deposit), understanding the clients we’re contacting plays a very significant role. The goal of this project is; given a client’s attributes, predict whether they end up subscribing for a term deposit.

**2. Client**

The data is related with direct marketing campaigns of a Portuguese banking institution (name of the firm has been anonymized, for confidentiality reasons). The marketing campaigns were based on phone calls.

**3. Data set**

This dataset is collected from [University of California, Irvine – Machine Learning Repository](mailto:https://archive.ics.uci.edu/ml/datasets/bank+marketing).

|  |  |
| --- | --- |
| **Bank Client Data** |  |
| 1 | age (numeric) |
| 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') |
| **Related with the last contact of the current campaign:** |  |
| 8 | contact: contact communication type (categorical: 'cellular','telephone') |
| 9 | month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec') |
| 10 | day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri') |
| 11 | 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. |
| **Other attributes** |  |
| 12 | campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) |
| 13 | 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) |
| 14 | previous: number of contacts performed before this campaign and for this client (numeric) |
| 15 | poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success') |
| **Social and Economic context attributes** |  |
| 16 | emp.var.rate: employment variation rate |
| 17 | cons.price.idx: consumer price index |
| 18 | cons.conf.idx: consumer confidence index |
| 19 | euribor3m: euribor 3 month rate |
| 20 | nr.employed: number of employees |
| **Output variable (desired target):** |  |
| 21 | y has the client subscribed a term deposit? (binary: 'yes','no') |

**4. Explain your initial findings**

1. ‘age’ – Age of the client is slightly skewed to the right. 70% of the clients contacted for this campaign are between the ages 23 – 48. The ages of the client are binned into three categories ‘young\_adult’, ‘adult’ and ‘senior’.
2. ‘age and marital’ – A new feature is created that combines the age and marital status of the client.
3. 'job', 'education' and 'month' -- These variables are consolidated on the percentage of positive and negative responses.

**5. Other potential data sets I could use**

The data provided could actually be considered very rich in terms of predicting the client’s behavior for a given campaign. However, given additional data pertaining to client’s financial spending such as income disposal, large credit purchases, demographic of the client.