Team member's details: **Individual**

Name: **Caleb Nyakeyo Anthony**

Email: **calebnyakeyo2018@gmail.com**

Country: **Kenya**

College/Company: **University of Embu**

Specialization : **Data Science.**

Problem description

ABC Bank wants to sell it's 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).

This is a binary classification problem. Our two classes are “yes” denoting that the customer subscribed to a term deposit, and “no” denoting that the customer did not subscribe.

Data understanding

We are given the data of direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (target variable y).

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Characteristics:	Multivariate	**Number of Instances:**	45211	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	17	**Date Donated**	2012-02-14
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	1600274

There were four variants of the datasets out of which we chose “ bank-additional-full.csv” which consists of 41188 data points with 21 independent variables out of which 10 are numeric features and 10 are categorical features.

What type of data you have got for analysis

As mentioned above, the dataset consists of direct marketing campaigns data of a banking institution. The dataset was picked from **UCI Machine Learning Repository** which is an amazing source for publicly available datasets. There were four variants of the datasets out of which we chose " bank-additional-full.csv" which consists of 41188 data points with 21 independent variables out of which 10 are numeric features and 10 are categorical features. The list of features available to us are given below:

Input variables:
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 - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

This is a binary classification problem. Our two classes are "yes" denoting that the customer subscribed to a term deposit, and "no" denoting that the customer did not subscribe.

What are the problems in the data (number of NA values, outliers , skewed etc)

What are Outliers? They are data records that differ dramatically from all others, they distinguish themselves in one or more characteristics. In other words, an outlier is a value that escapes normality and can (and probably will) cause anomalies in the results obtained through algorithms and analytical systems. There, they always need some degrees of attention.

1. Skewness

The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

When data skewed, the tail region may behave as an outlier for the statistical model, and outliers unsympathetically affect the model's performance especially regression-based models. Some statistical models are hardy to outliers like Tree-based models, but it will limit the possibility to try other models. So there is a necessity to transform the skewed data to close enough to a Normal distribution.

2. Number of missing values.

n real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

a. Set up a filter in your testing tool

Even though this has a little cost, filtering out outliers is worth it. You often discover significant effects that are simply "hidden" by outliers.

According to Himanshu Sharma at OptimizeSmart, if you're tracking revenue as a goal in your A/B testing tool, you should set up a code that filters out abnormally large orders from test results.

b. Remove or change outliers during post-test analysis

One way to account for this is simply to remove outliers, or trim your data set to exclude as many as you'd like.

This is really easy to do in Excel—a simple TRIMMEAN function will do the trick.

c. Change the value of outliers

Much of the debate on how to deal with outliers in data comes down to the following question: Should you keep outliers, remove them, or change them to another variable?

Essentially, instead of removing outliers from the data, you change their values to something more representative of your data set. It's a small but important distinction: When you trim data, the extreme values are discarded.

d. Consider the underlying distribution

Traditional methods to calculate confidence intervals assume that the data follows a normal distribution, but as with certain metrics like average revenue per visitor, that usually isn't the way reality works.

e. Consider the value of mild outliers

As exemplified by revenue per visitor, the underlying distribution is often non-normal. It's common for a few big buyers to skew the data set toward the extremes. When this is the case, outlier detection falls prey to predictable inaccuracies—it detects outliers far more often.

There's a chance that, in your data analysis, you shouldn't throw away outliers. Rather, you should segment them and analyze them more deeply. Which demographic, behavioral, or firmographic traits correlate with their purchasing behavior? And how can you run an experiment to tease out some causality there?

This is a question that runs deeper than simple A/B testing and is core to your customer acquisition, targeting, and segmentation efforts. I don't want to go too deep here, but for various marketing reasons, analyzing your highest value cohorts can bring profound insights.

Dealing with Missing Values.

(i). Deleting Rows

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

(ii).*Replacing With Mean/Median/Mode*

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with

the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.