IT320

Business Intelligence Project Report

Bank Marketing Analysis

Nouran Ghaliounji

Summary

1. The Business Problem

"A term deposit is when you lock away an amount of money for an agreed length of time (the 'term') – that means you can't access the money until the term is up. In return, you'll get a guaranteed rate of interest for the term you select, so you'll know exactly what the return on your money will be."

A Portuguese banking institution wants to access the efficiency of its marketing campaign to know how successful it was in converting customers to subscribe to a term deposit or not.

Throughout, this project I will use the dataset put at my disposal to analyze the efficiency, and the weaknesses of the marketing campaign to make recommendations to the marketing team.

2. The Dataset

The dataset I used is from Kaggle, and was originally posted in the UCI Machine learning Repository and contains data from the year 2012 of a Portuguese Banking institution.
The link to the data: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

It contains the following columns:

- age: (numeric)
- job: type of job (categorical: 'admin.', 'blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- marital: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- education: (categorical: primary, secondary, tertiary and unknown)
- default: has credit in default? (categorical: 'no','yes','unknown')housing: has housing loan? (categorical: 'no','yes','unknown')
- loan: has personal loan? (categorical: 'no','yes','unknown')
- balance: Balance of the individual.
- contact: contact communication type (categorical: 'cellular','telephone')
- month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- Day: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- 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.
- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 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)

- previous: number of contacts performed before this campaign and for this client (numeric)
- poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
- LABEL deposit: has a client subscribed a term deposit (yes or no)

The target variable that we want to access and focus our analysis on is the variable "Deposit" which is a categorical variable represented by yes if the client has subscribed to a term deposit and no if he has not.

3. Data Analysis

Before starting with the data integration and the ETL process, I thought it was an important step for me to analyze the data superficially to know more about the distribution of the numerical variables, as well as the levels of the categorical variables to better handle the data later on. The tool I used at this stage is a Jupyter notebook which is attached along with the files I used for the project.

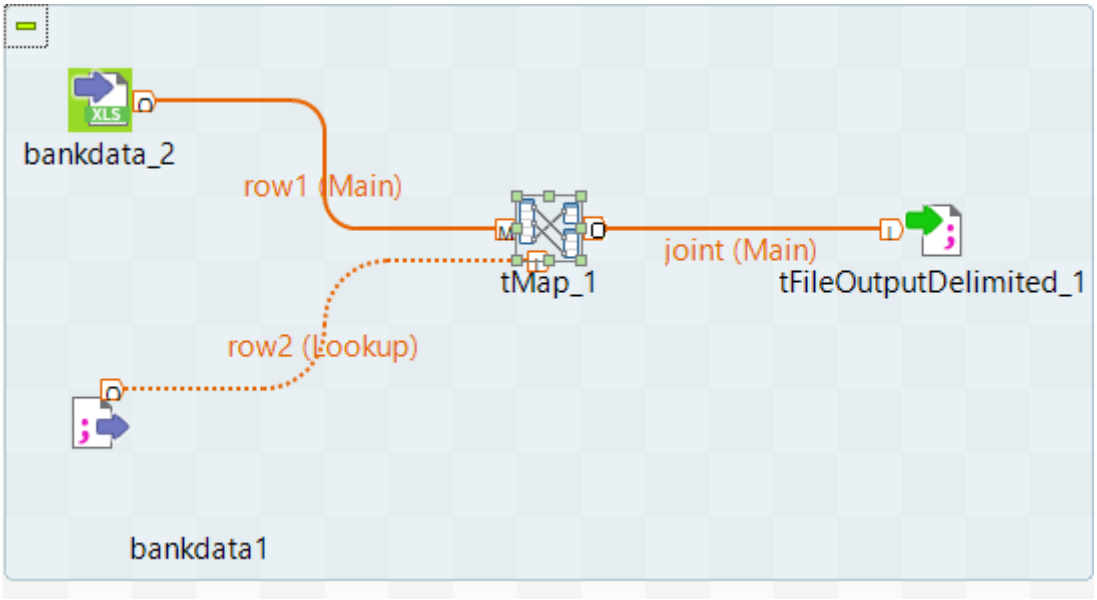
I performed a simple analysis to see the shape of the data, see the columns and their categories, and get the statistical distributions of the numerical columns.

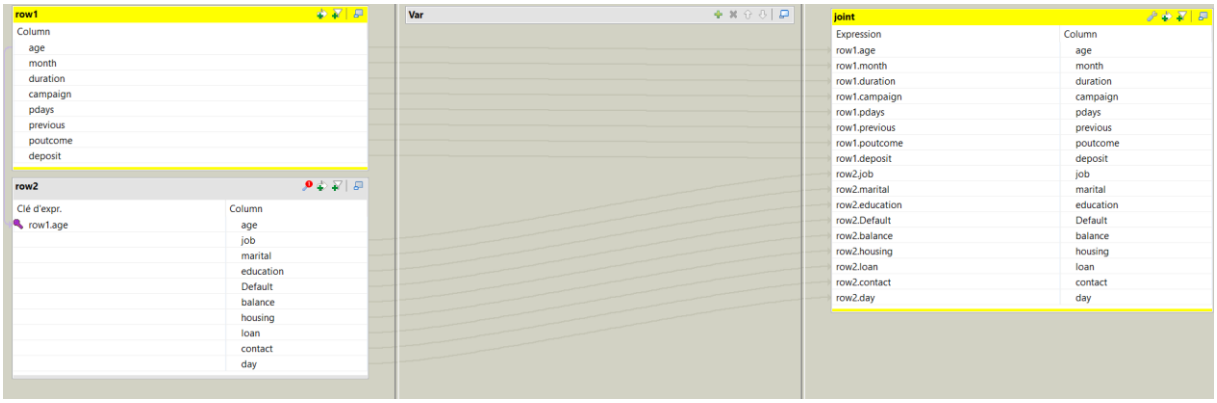
I also used Jupyer notebook later on steps when working with Talend to do simple calculations for the median, check the count of NA values, and display the box plots to check for outliers.

4. ETL with Talend

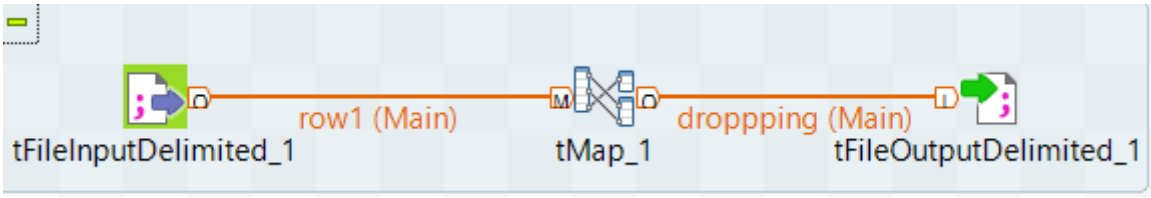
I extracted the data into 2 files: a CSV file and an excel file that both have common the age column.
The extracted files can be found in the project files under the name "Bankdata2.xlsx" and "Bank data11.csv".

In Talend, I used a first job called "Integrating_Files""to bring the 2 files together into a single csv file by using the tmap component. I configured the job as such:

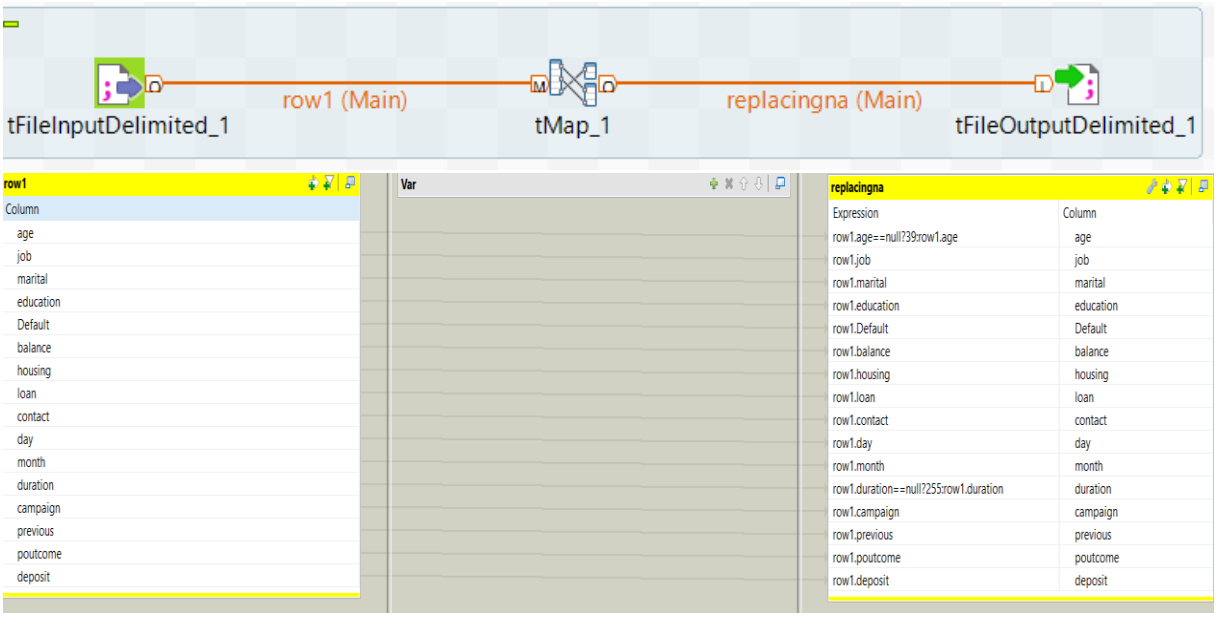




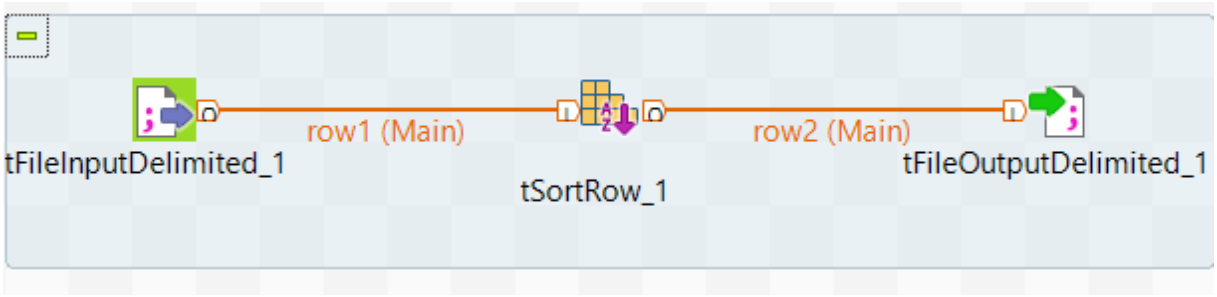
After bringing the 2 files together, I created another job called "Dropping_col" where I dropped a column that I deemed unnecessary for the analysis, and I configured the job as such:



Later on, I created a 3rd job to deal with missing values in the age and duration columns. I had 6 missing values in each of these columns. After checking for outliers in both columns, I decided to impute the missing values with the median for each column instead of the mean to get better-quality of data. For that, I used a tmap component since treplace only works with string values and I insert a java expression to detect the missing values and replace them by the calculated median.

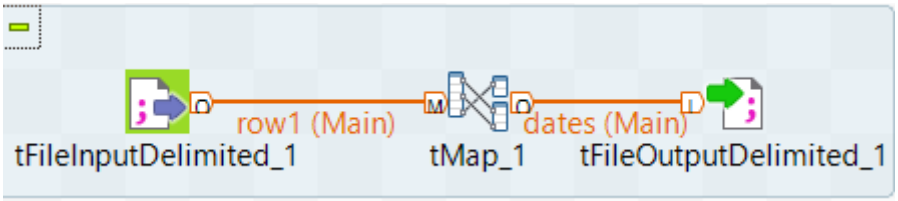


The 4th job called "sorting_duration" I created is to sort by descending order the numerical column duration since I suspect a positive relationship between duration and probability for a client to subscribe to a term deposit. After doing the sorting, my initial assumption seems to be confirmed and it is something that I will check later on when working with PowerBI



Colonne du schéma	tri num ou alpha ?	Ordre asc ou desc ?
duration	num	asc

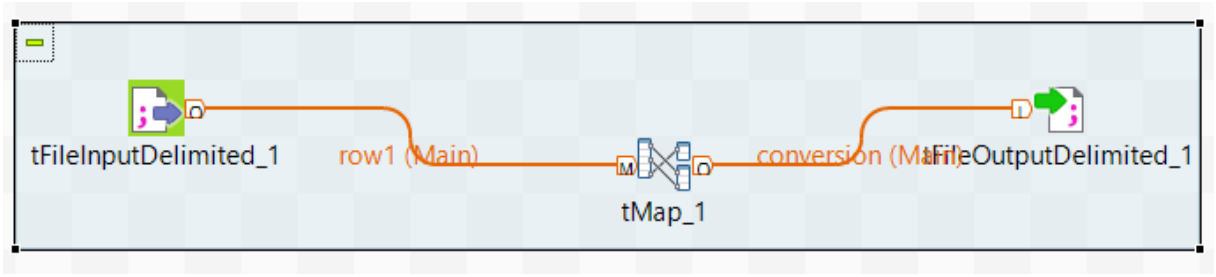
The 5th Job I created called "creating_dates" is to convert the month column into strings that can be recognized by Java as a date. Their initial pattern is "jan", "feb" and I need to convert them into "Jan", and "Feb" to be accepted in MMM format. For that, I also used a tmap in which I inserted the following java expression: INSERT expression
Asides from the months, I also converted the numerical day's column into a string to put in the appropriate format for the java function to convert it as a date

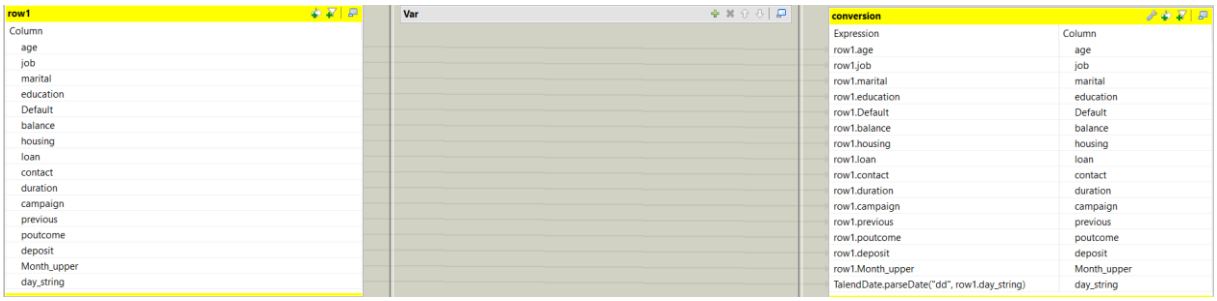


StringHandling.LEFT(StringHandling.UPCASE(row1.month),1) + StringHandling.RIGHT(row1.month,(StringHandling.LEN(row1.month) -1))

Integer.toString(row1.day)

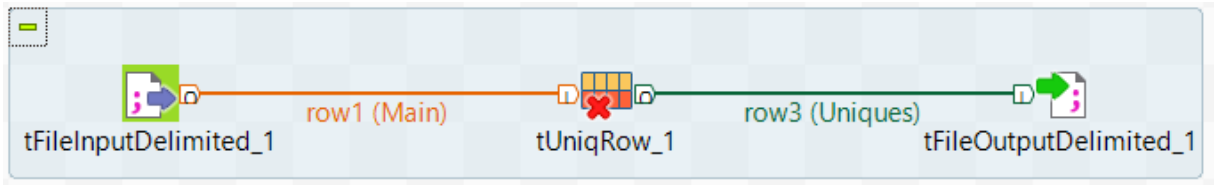
For the 6th Job called “convert_date”I created I used the tmap component to modify the type of the day column from the string into the date





TalendDate.parseDate("dd", row1.day_string)

Finally, the 7th job that is “remove_duplicates” removes the duplicate rows contained in the data by using the tUniqRow_1 component to preserve the quality of the data and avoid redundant information.



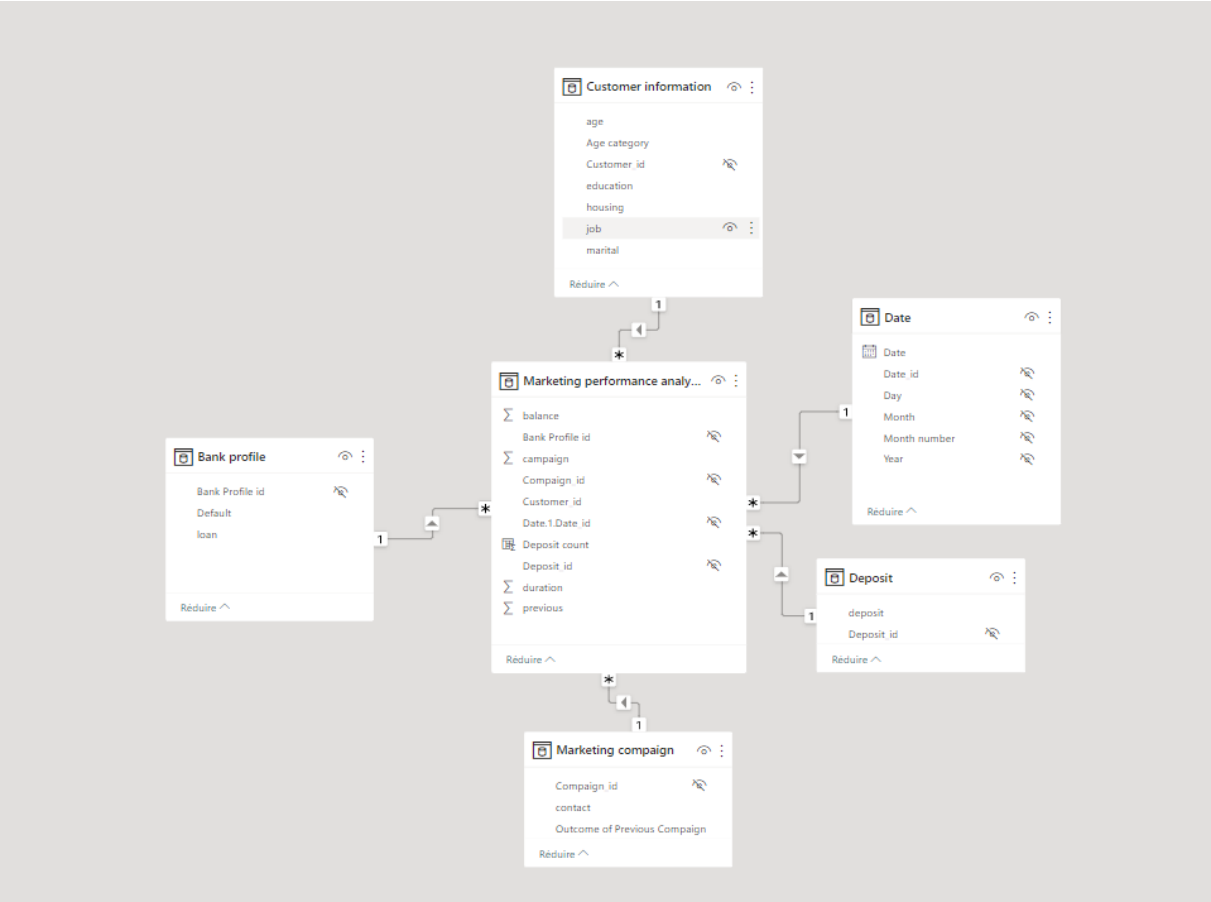
5. Data Modeling

After completing all the previously described jobs with Talend, I took the final output CSV file called “data_with_no_duplicates.csv” and loaded it into PowerBI desktop as a flat file.

Before starting with the visualization, I took the time to clean the data further as some tasks are more simply done by PowerBI using the PowerQuery tool, and then built my model.
The tasks I did in PowerBI to clean the data and put it in an appropriate user-friendly format can be described as such:

a. Renamed the columns into a user-friendly name
b. Created an age category column that ranges the age per category and will be useful later on for visuals
c. Created dimension tables, and fact tables, and identified the measures
d. Created a unique key for each dimension to be referenced in the fact table to obtain a 1 to * relationship which makes the process of parsing data smoother

I obtained the following Star Schema:



6. PowerBI report

The powerBI report I created consists of 6 pages whose ultimate goal is to access the efficiency of the marketing campaign and to make recommendations based on patterns I found in the data.

The report is divided as such:

- **Customer demographics:**
  - This first page is composed of 5 visuals that each represent the number of clients according to the customer profile categories (age, education, housing, jobs marital status)
  - The goal is to know who are the clients and to maybe further identify categories that marketing should focus on more
- **Which Customers are more likely to make a deposit**
  - This category is divided into 2 pages of the report
  - The first one shows the distribution of deposits according to client's jobs. The key takeaway is that retired people are those with the highest ratio of deposits=yes to no. For now, we can say that maybe more marketing efforts should be directed toward that category.
  - The distribution of customers that made deposits by marital status seems to show that maybe single people are more likely to make a deposit
  - The distribution by age clearly shows that old people (+60yr) and young people (18yr-25yr) are the ones that are more likely to deposit not. They should be more targeted by marketing campaigns.
  - The distribution by housing seems to tell that whenever a customer does not own a house, he is slightly more likely to make a term deposit
- **Does an Individual's bank profile have an impact on he/she making a deposit**
  - The key takeaway here is that the larger the balance a customer has in their account, the more likely they are to make a term deposit
  - Loan status and default do not seem to play an important role in determining if a client will make a deposit
  - Marketing should focus on clients that have a large balance
- **The efficiency of the Marketing Campaign**
  - The grouped bar chart "Distribution of clients that made a deposit by contact method" tell us that telephone might be a more efficient way to get customers to make a term deposit
  - Also, the outcome of the previous marketing compaign is important in knowing if a client will make a deposit or not. In the case where the last oucome cas a success there is a probability of 91% that they will make a term deposit. Furthemore, if the last outcome was a failure there is almost an equal chance that a client will make a term deposit.
  - Duration also seems to be a really good predictor of whether or not a client will subscribe to a deposit
  - According to the curve that shows the mean duration by month. July had the highest mean duration and it is also one of the months with the highest amount of term deposits made
- **Accessing the efficiency of the Marketing Campaign through time**

- Focusing on the number of deposits made by month. We notice an interesting trend in the month of may. It is the month with the highest number of deposits made and also the month in which customers did not make a deposit.
- This can be explained by the fact that the month of may was the one in which a lot of contacts were made and the one with the highest sum of compaigns launched.
- From that we can say that there are diminishing returns related to the number of times a customer is reached

7. Recommendation for the campaign

After analyzing the data and creating the Dashboards, I come up with the following recommendation:

- More marketing efforts should be deployed towards Young (between 18 and 25), and old people (+60) as they were the ones with the highest ratio of term deposit subscriptions. Retired people are 1.31 times more likely to subscribe to a term deposit.
- There is a clear correlation between duration and deposit subscription, meaning that the more time a sales agent spends with a customer on the phone, the more likely they are to subscribe to a term deposit. The marketing team needs to find a way to spend more time with people on the phone as this is an important variable in predicting whether a customer will subscribe to a term deposit or not
- Direct contacts by telephone rather than cellular seem to be working more efficiently since the ratio of term deposit=yes is higher for a customer that was contacted by telephone. I would recommend that the marketing team explores this path more, and try to increase the number of contacts made by phone.
- A successful outcome in a previous campaign leads to a very high rate of subscription. Therefore marketing should keep contacting the clients that have had successful outcomes of previous marketing campaigns. The clients that had previous failing campaigns should also be targeted
- A very important element to take into consideration is also the diminishing returns related to marketing investments. We noticed that the month of May is the month during which the most contacts with clients have been made. It is also the month during which most clients chose not to subscribe to a term deposit. Therefore my recommendation is to determine a contact threshold to not face so violently the diminishing returns related to the number of times a client is reached
- Something that had been noticed is that the number of times clients are reached is highly inconsistent over time. A lot of contacts are made during the month of May, June, July, and August which is the summer season, but the other months seem to be neglected. My recommendation is to keep consistent marketing efforts throughout the year
- The customers with the highest median balance are the ones that are more likely to subscribe to a term deposit therefore more marketing efforts should be put at the disposal of that particular category

8. Improvements to be made and conclusions

Although the project was insightful and allowed me to come up with several recommendations, further improvements could be made to make the process automated in the future.

- Data could be stored in an RDBMS like Oracle or MySql
- Talend could directly connect to the database and extract data periodically by scheduling jobs
- More Talend jobs could be added (many of the data cleaning tasks I performed on PowerBI could be periodically performed with Talend automatically)
- Talend could be used to build the star schema model and store it in a data warehouse server
- PowerBI could automatically load the data from the data warehouse and update the reports periodically

Thank you

Nouran Ghaliounji