Data Analytics Software Project 2 Report

Bank's Marketing Campaign Analysis with IBM Watson

Wakjira Ashenafi

Date: 08.02.2018

Introduction

In this project we are given a bank's dataset with 42212 records and 16 features and 1 target variable. The first eight features are about the client's information and then we have four features that are related with the last contact made by the bank with its client during their current marketing campaign and the other four features are some how related with the previous marketing campaign and contacts made and outcomes of their previous campaign. The detail of the data information can be seen on Appendix 1.

Given this data the goal of the project is to investigate the elements that leads in a successful marketing campaign for the bank. To do this analysis a cloud-based service by IBM Watson Analytics is used.

Loading the data to IBM Watson

The given data is in CSV format but after downloading it warns me the there might be a loss of data if I save it in CSV format and recommends that I save it in excel format. So, I imported the data with data query editor. After saving the data with excel format then we go to Watson Analytics, New data-> Local file and browse the data that we saved in our local computer and import it to Watson Analytics cloud based service. After importing the data, we get a new data set tile in data tab in Watson Analytics.

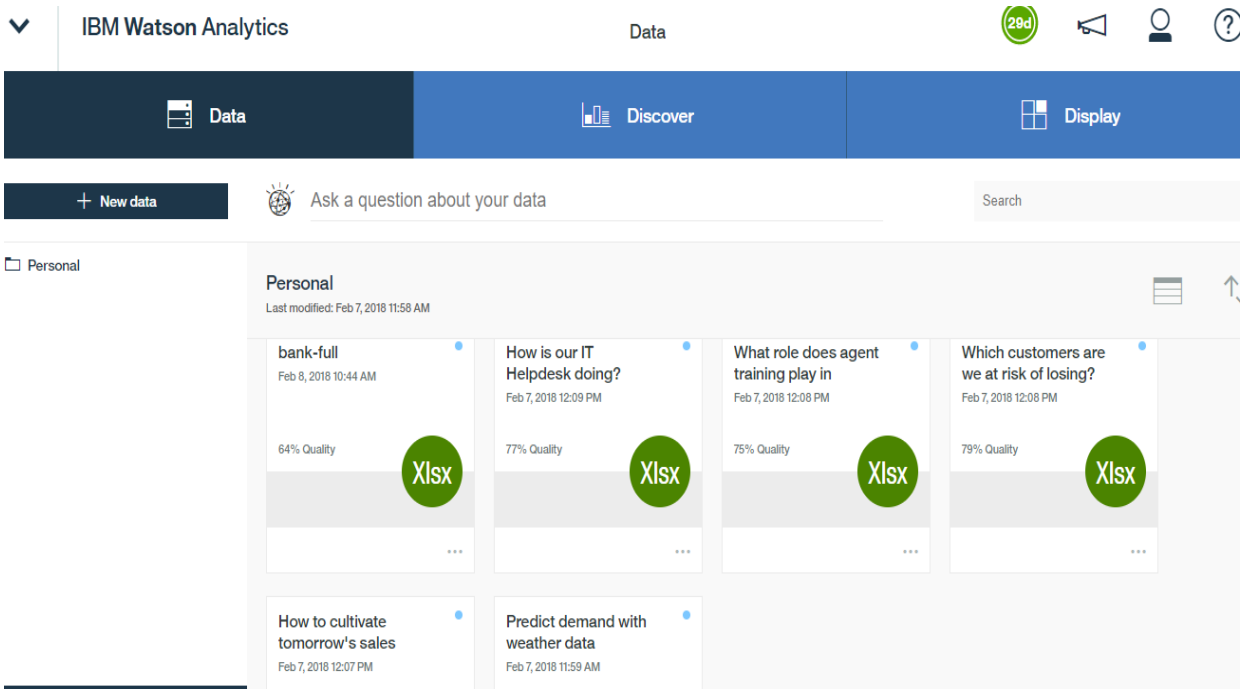

Figure 1. Watson Analytics Interface with the data tabs

As we can see on Figure 1, Our data is imported to Watson Analytics with the same name that we saved in our local machine and the date stamp and data quality. The data quality for our data is 64 %. The data quality score how suitable the data asset is for making prediction, and it is determined missing values, constant values, skewness, influential categories, outliers, imbalance.

Descriptive analysis of the data

Let us first go to our data and check the data quality of each features, check if there is missing values in any of the variables. To do that we go to refine and Data metrices.

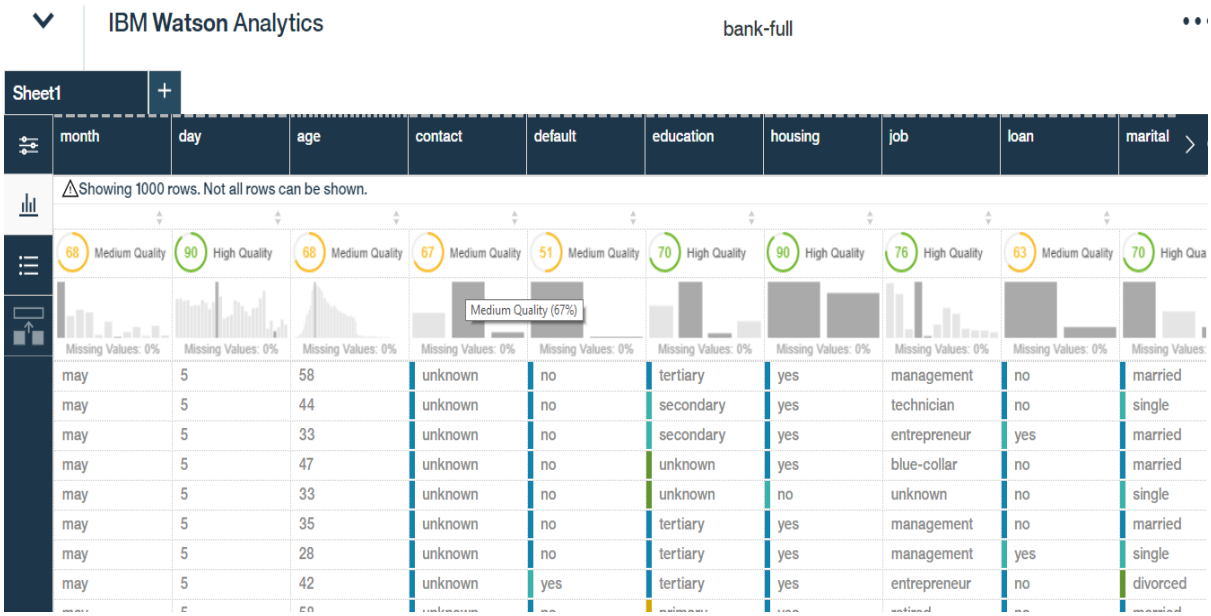

Figure 2. Checking the variables quality and missing values

One thing that Watson analytics is good at is the ability to perform analysis using natural language questions. Depending the type of question, you ask it suggests options to choose from.

Next let us see the frequency distribution.

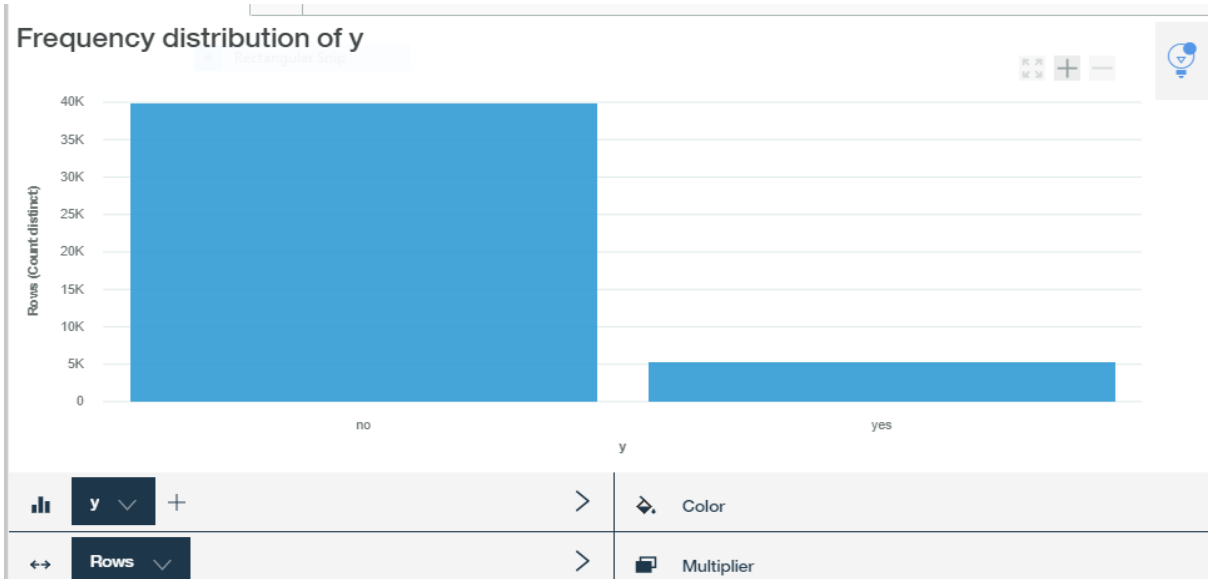

Figure 3. Frequency distribution of y.

What drives the success of the current campaign

The current campaign outcome is y as it is described in the dataset description shown in appendix 1. To find out what drives the success of the current campaign we ask Watson analytics the question "What drives y" and it will suggest with options to select. What is the predictive model for y seems a good choice for this.

Watson gives us a lot of choice to display the result like a spiral, decision rule, tree and more. Here I selected the decision rule. We can select the target category as desired. It also shows the predictive strength of the model which is 89%.

In Figure 3 the target category is set to "Yes" and we can see that the top drivers for this set up are duration = 223.52 – 367.48 and poutcome = success with 76% accuracy. And the second good drivers are duration >367.48, contact = cellular, month = Apr, Feb and housing = No.

From those results we can conclude that the previous campaign and last contact duration with the customers at least for more than 223.52 leads the clients to subscribe a term deposit.

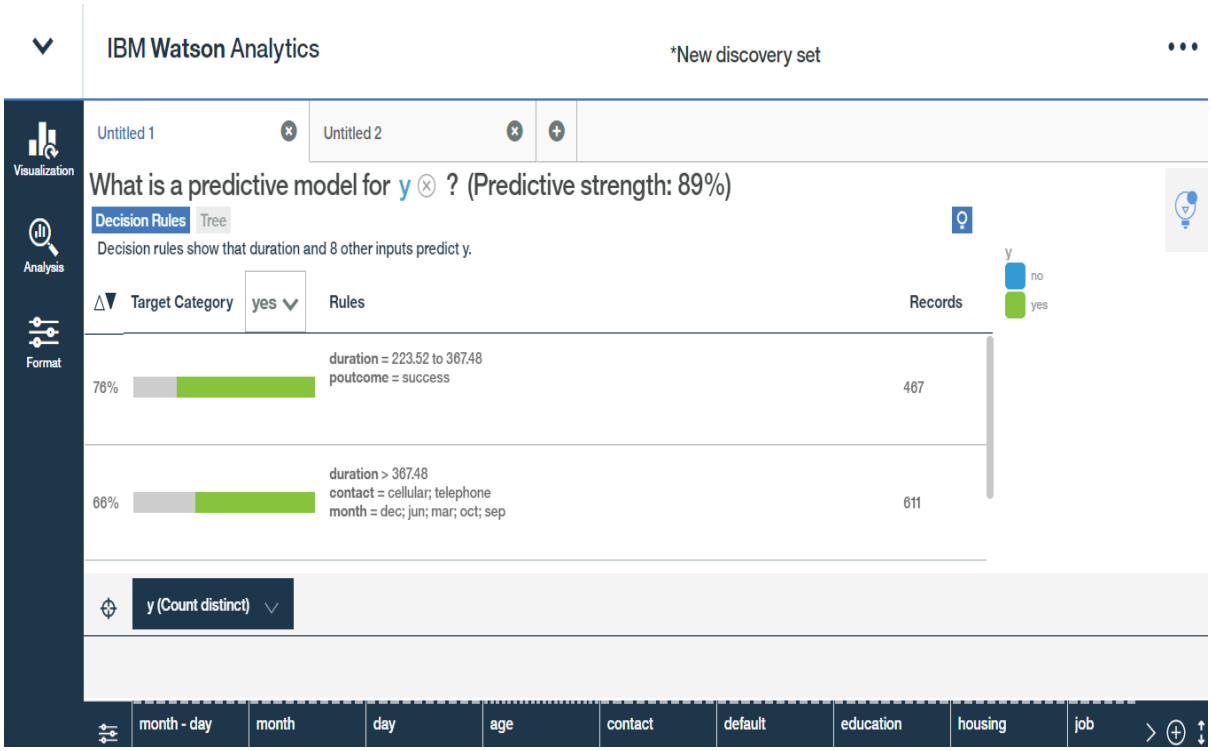

Figure 4. Decision rule for the predictive model of y.

How to improve campaign strategies to attract more customers.

To suggest an improved strategy to attract more customers, we need to see what drives y to failure in addition to what drives y to success. So, the company keeps doing what they are doing fine, and they will also make on their weakness at the same time.

To asses what makes y to failure let us change the exact same figure with target category changed to “No” and check why people don’t subscribed to term deposit.

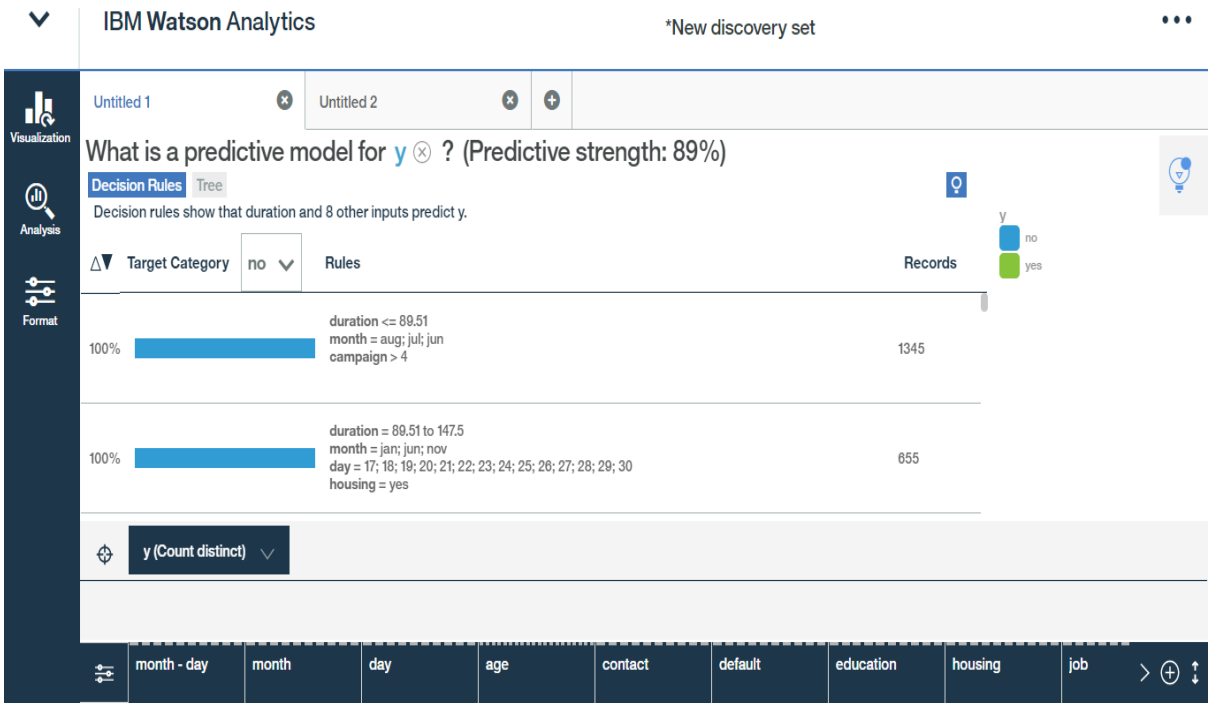

Figure 5. Decision rule showing which features leads to fail to subscribe to the term deposit.

AS we can see from figure for if the duration of the last contact with the customer during this campaign is less than 89.55, if the last contact is made during the months of Aug, June, Jul and a single client is contacted more than 4 times then it leads to 100 % failure to subscribing that client. The other combination of factor that leads to failure to subscribe a client is the one in the second position which is also 100 % lead to failure, duration- between 89.51, month- Jan, Nov, Jun, day between 17-30 and housing = No.

There are a lot of combination factors with 100 % that leads to failure to subscribe the client to the term position. We need to check all and make generalization.

From both successes and failure results one can make suggestion how the company can attract more customers. My suggestion for the company based on the outcomes are:

1. The company should track the entities that previous outcome result which leads to success.
2. Call duration with the customer matters on the success. More talking time at least greater than 223 seconds per customer leads to successes on subscription of a client.
3. Contact medias telephone and cellular are more productive than other medias. As can be seen from the figure bellow the unknown media dominates the telephone. So the company need to figure out what is the unknown media and switch its media to either cellular and telephone since the unknown media is not effective.

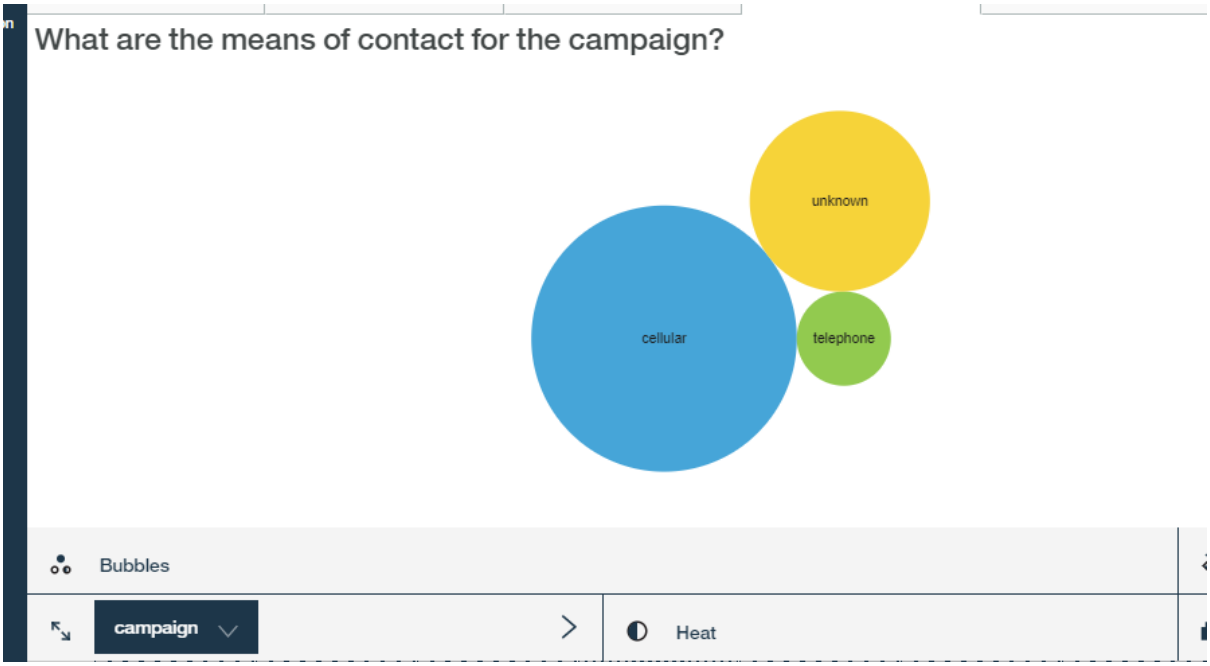

Figure 6. Proportion of campaigning medias

4. August, June, July and some other summer season months campaign seems to fail.
But the truth is most of the campaigns are done during this season. If the campaign is done during this season failure is also expected. May seems the most successful month and most of the successful campaign is already done with that month.

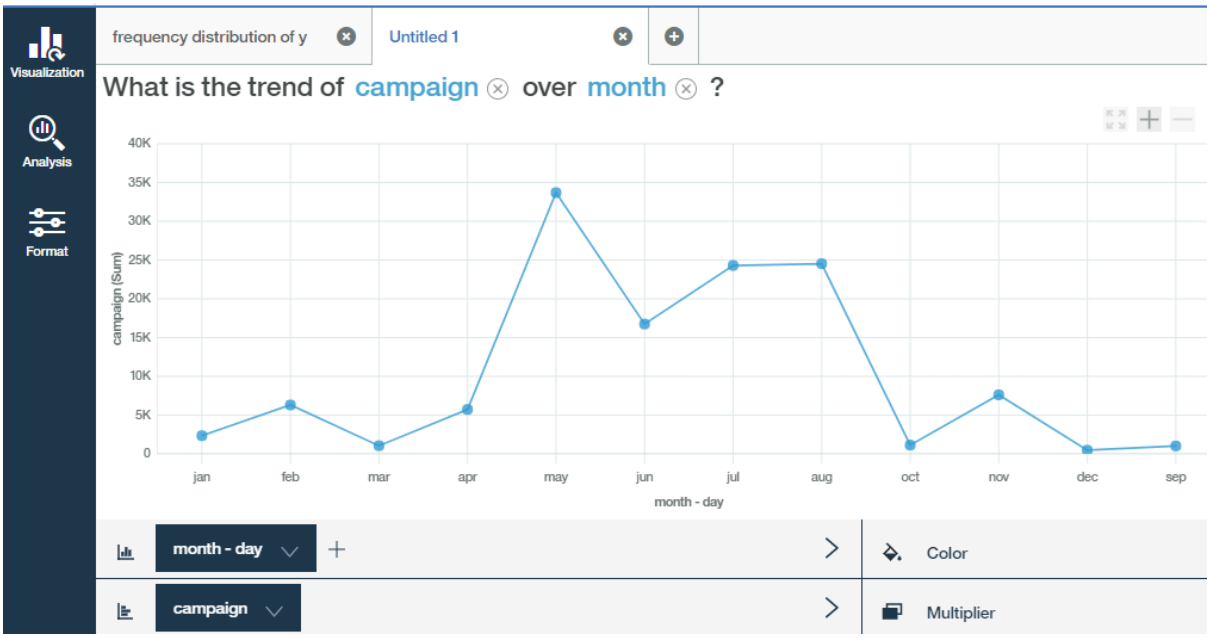

Figure 7. Number of campaigns conducted each month

5. Campaign > 4 tend to fail. I think attempting to convince a customer by contacting repeatedly in this case more than 4 seems not working. It is better for the company to focus on other customers than trying to contact the same customers more than 4 times.

6. Contacting the customers at beginning of the month (1-14 & 31) is more likely be successful than (15-30). This needs further analysis to find out. For me the successful campaign during the month may seems dominating the day value and generalization is made on the month May.

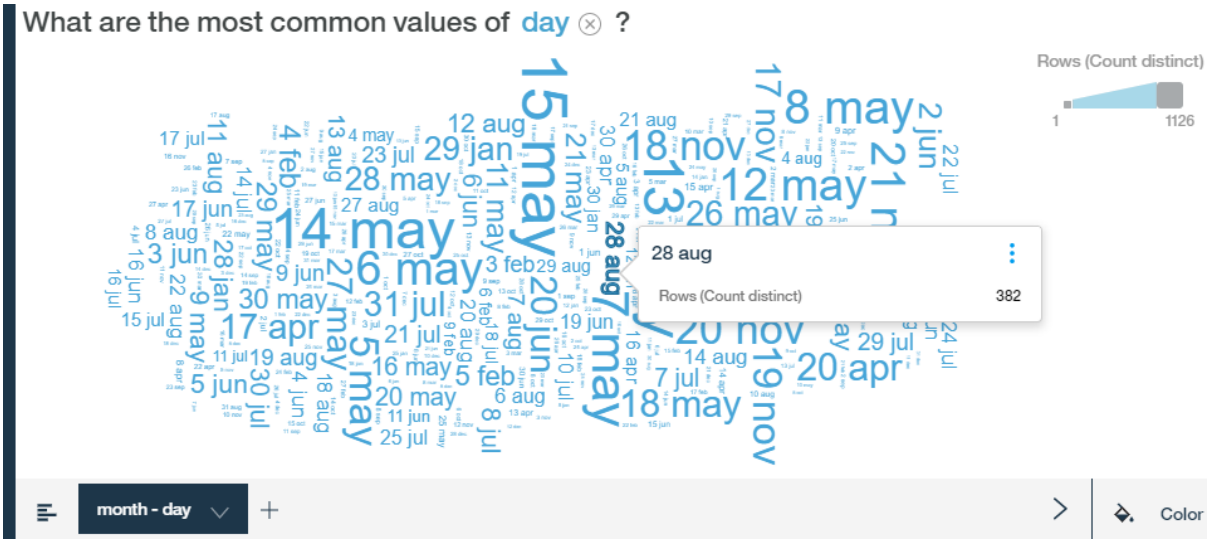

Figure 8. How the number of dates are dominated by the pick campaign Month MAY

7. Those people with housing lone seems that they don't need other commitments. The company need a suitable strategy for this group and focus on those with out housing lone more with the current strategy.

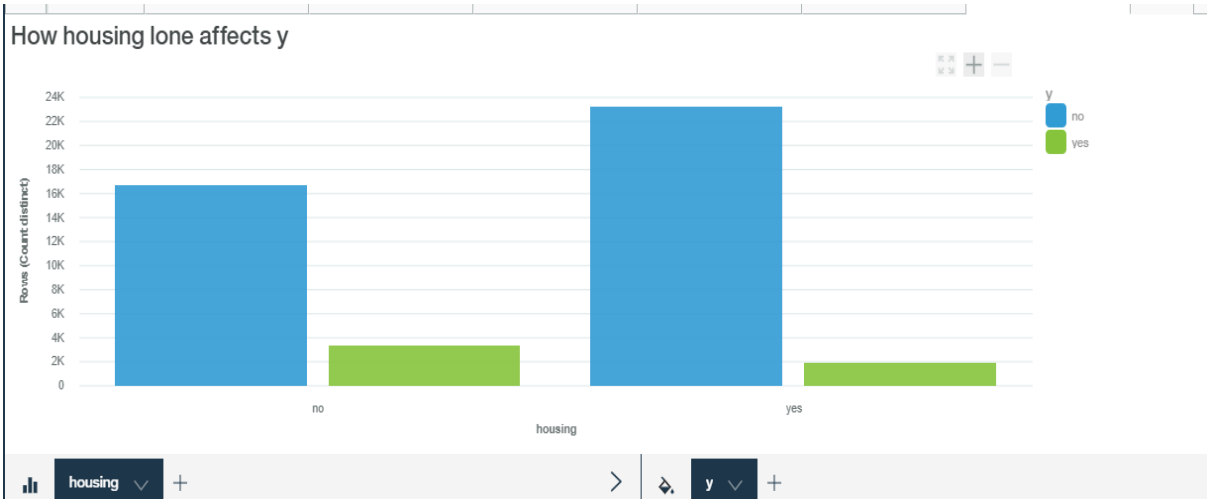

Figure 9. How housing loan affects y. Those with housing loan don't tend to subscribe a term deposit deal during this campaign.

Dashboard

Finally, we can collect all the plots on the display tab and make a dashboard. After downloading it with PDF format I merged the dashboard report at the end of this report as Appendix 2.

Conclusion

In this task I able to import the given data to Watson analytics. The given data is a customer's information of a bank data. In addition to client's basic information the data contains marketing campaign data. More information on the data set is given on appendix 1. From this given data analysis is done on the factors that leads to a success of the campaign. The success of the campaign is when the client is subscribed to a term deposit. According to Watson analysis the top drivers for this campaign are duration and poutcome with 76% accuracy. The overall predictive strength of the model used 89%. Based on the factors that leads to success and failure, I suggested campaign strategies that might attract some more customers. Some relevant pictures are included on the report to give some insight to my justifications. And finally those pictures are collected on a dashboard and show at appendix 2.

Appendix 1

Input variables:
 # bank client data:
1 - age (numeric)
2 - job: type of job (categorical:
"admin.","unknown","unemployed","management","housemaid","entrepreneur","stu
dent", "blue-collar", "self-employed", "retired", "technician", "services")
3 - marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown", "secondary", "primary", "tertiary")
5 - default: has credit in default? (binary: "yes", "no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes", "no")
8 - loan: has personal loan? (binary: "yes", "no")
 # related with the last contact of the current campaign:
9 - contact: contact communication type (categorical: "unknown", "telephone", "cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "Jan", "Feb", "mar", ..., "Nov", "Dec")
12 - duration: last contact duration, in seconds (numeric)
 # other attributes:
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means clients was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")
Output variable (desired target):
17 -y- has the client subscribed a term deposit? (binary: "yes", "no")

Appendix 2