What makes people subscribe for a term deposit?

1 Introduction

In this project, I have analyzed the Bank Marketing Data Set, obtained from the UCI Machine Learning Repository. This dataset is related to direct marketing campaigns of Portuguese banking institution. The marketing campaigns were based on the phone calls made to the clients. There are four datasets available in the repository and I chose to work on bank-additional-full.csv. Dataset can be obtained from <https://archive.ics.uci.edu/ml/machine-learning-databases/00222/>.

This dataset consists of the details related to education, age, marital status, number of times clients were contacted during or before the campaign, duration of the last call made to the client, day of the week and month on which client was contacted and output variable represents whether the client has subscribed for the term deposit.

There are 20 attributes provided in the dataset related to the phone calls. I was interested in finding out how are the subscription rates are related to different variables? Are they dependent on any attributes like age, job, and duration of the calls, months or present or previous contact? What should be average duration of the phone call made to client to get higher subscriptions? Does it vary over the months or days of the week? Jobs of which category have high success in subscriptions?

After visualizing the data through different plots using Matplotlib and Pyplot, I have found some interesting results related to duration of call. Subscription rates are correlated to duration of the call, the age of the person, month, day. Based on the analysis, I found that people belonging to the age group of 30-40 have higher subscriptions when duration of the call is around 460 sec and most of them have a university degree. Also, to obtain higher percentage of subscriptions, age groups to be considered are below 30 and above 60 and the average duration of the call should be 400 sec. Higher percentage of subscriptions are observed if the clients were contacted less than 5 times. Students and retired clients are more attracted to this campaign.

2 Data Quality

Data contains 41188 observations with 20 features, 10 are categorical features

and 10 are numeric features. The dataset that I have obtained has no null values.

Profiling the dataset for completeness, uniqueness, validity:

1. To check for completeness, I have calculated the percentage of missing values or null values for each column of the dataset. The result data does have any null or missing or n/a values.
2. To check for uniqueness of the data, I have used dataframe.duplicated() method which returns percentage of rows that are not duplicated and then we can drop the duplicates. Using this property, I have obtained percentage of unique values in the dataset.
3. To check for the validity of the data, I have used the dataframe.describe(include=all) command to compare the attribute values like max and min properties with limits given to the variables in the dataset repository. Include all property of describe commands also shows the properties related to categorical columns like unique values count, frequent value, top value etc.,

Data cleaning:

Some attributes like job, education, loan, poutcome, marital status have values unknown which means those values does not have data or it is not categorized. To analyze the data effectively, I have removed the unknown in the poutcome column, as the previous outcome can be either Yes or No. All other attributes are not completely categorized. So, the columns can have the values that are ‘unknown’, but they are replaced by ‘other’ for ease of understanding.

There are outliers for age and duration columns. So, to analyze the data, removed the outliers for the two columns.

Chart

Description automatically generatedChart, bar chart, box and whisker chart

Description automatically generated

After removing the outliers, area between quantile 0.25 and 0.75 doesn’t change too much.

Chart, box and whisker chart

Description automatically generatedChart, bar chart

Description automatically generated

Data Preprocessing:

Before analyzing the dataset, data needs some preprocessing. Some of the attributes like age, duration, education, month, day of week needs to be grouped into different categories. To do that, I have used pd.categorical() method for object data types and pd.cut() method for numeric data types. The pd.categorical() uses categories and also the order of specification of categories acts the default sort order whereas pd.cut() uses bins of different numbers which can be used to create a left inclusive or right inclusive ranges.

After categorizing the different attribute values, data related to month, days of week are stored in sorted order. So that, it would be easy to analyze months or days in visualizations.

Also, data needs to be normalized accordingly, to get better insights.

Now the data is all set to analyze and perform some visualizations to find interesting facts about the data.

Assessing the quality of the data:

The data has met completeness, uniqueness, and validity properties, but there are outliers which might affect the results or insights of the visualizations. Some attributes also contain unknown values which can be considered as missing values. Based on the above analysis, data is not perfectly clean without any preprocessing and cleaning. In my view, quality of the data is medium. Because the values in the columns are clean and does not require any preprocessing. But the outliers need to be taken care of, to obtain better insights. Also, data required for plotting with months or days needs a sorted order of months or days. Data also needs to be normalized based on the number of subscriptions. This would ease the process of analyzing the patterns.

3 Data exploration and Analysis

Insight 1:

Of all visualizations, the relation between age, education, duration, and number of subscriptions gave different insight of which age group of people has higher number of subscriptions and what education level does most of those age group people possess.

To get the insight, I have started by plotting the education and number of subscriptions for 10 people contacted. Subscriptions rate for every 10 people contacted can be obtained by dividing (the number of successful subscriptions in that group) with (total number calls made to that group) multiplied by 10. This will produce better insights of how many people have successfully subscribed for a term deposit if 10 people are contacted in each education category. I have chosen bar plot to visualize the results as most of the columns are categorical.

The resulting visualization shows that illiterate people are more attracted towards this bank marketing strategies. Followed by people who pursued university degree. They might be in the stage of earning a bit higher amount of salary than other educational category people and wants to save some money for future investments.

Chart

Description automatically generated

Chart, timeline

Description automatically generated

Also, by plotting the different age groups and the normalized subscription rates, we can observe that, to obtain at least 10% of subscriptions, campaigns must consider contacting the people of age groups below 30 and above 60.

To understand the relation education, age, and subscriptions, I have plotted these three columns in a bar plot. The below image explains that most people that were subscribed are between 30-40 age group and most of them have education level of university degree.

Chart

Description automatically generated

To find out average duration of call that needs to be made to clients in order to achieve higher subscriptions, we need to plot age groups, duration columns and their respective subscription rates. Results of which can be seen in the below image.

Chart, bar chart

Description automatically generated

From the results of the two graphs, we can conclude that average duration of call that has to be made for age group 30-40 would be 460 sec nearly and less than 400 sec for groups below 30 and above 60 groups.

Insight 2:

The relation between job, duration and number of subscriptions generated different results when the subscriptions are normalized. Without normalizing the number of subscriptions, the results shows that people working as admin has subscribed for term deposit when the average duration of the call made was 402 sec.

Chart, bar chart

Description automatically generated

But when visualization was made using normalized subscriptions, the plot shows different results. It shows that students and retired people have higher rates of subscriptions compared to other job categories. Retired people might receive PFA which can be invested in the form of term deposit for futuristic purposes. The average duration of the call for these categories of people are in the range of 370 sec.

Chart, bar chart

Description automatically generated

Normalization brings out different insights that are not expected.

Insight 3:

To understand the relation between the number of contacts times during the present campaign and the subscription rate, I have plotted the campaign column with the number of subscriptions.

The result shows that contacts times should not exceed 5, to obtain higher number of successful subscriptions.

Chart, histogram

Description automatically generated

4 Conclusion

Subscriptions could be improved by targeting few groups of clients and contacting them on a few specific days. Students and Retired Clients are mostly likely to subscribe. Talking to Clients less than 400 sec improves securing a subscription. We shouldn’t overwhelm clients by contacting them repeatedly. Ideally, we should contact less then 5 times. Targeting people with age less than 30 or more than 60 and contacting them less than 5 times and talking to them less than 400 sec could improve chances of securing a subscription.