Data Analysis Project- Bank Marketing

Spring 2020

Tal Ifargan and Ronny Bloch

Table of Contents

1. DATASET

The dataset we chose to work with is "Bank Marketing".

This dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. in addition to data about each client (record in the dataset) there is also a class column that represents if the costumer agreed to the offered bank term deposit - i.e. what was the target result.

2. QUESTIONS

The questions that were interesting in our opinion and we tried to answer during the preforming the different tasks:

1. **"Are people that get contacted several times have less chance to subscribe a term deposit?"**

   This question came out to our minds when we were thinking about ourselves in that situation, getting a phone call from our bank about subscribing a term deposit. It seemed like a good question to ask because it also has meaning in terms of the data we have and what conclusions derived from it can improve the next marketing campaign impact and of course it also something interesting to figure out.

2. **"Which people will subscribe a term deposit and which don't? What features will be the best predictors in KNN classification of that target? What insights we can get from training a classifier to predict that target class?"**

   This question was interesting for us to explore because we think that this is the essence of data analysis, getting real life consequences using a data that can improve currently ongoing projects and existing workflows. In the case of this question the answers we might get can supply answers for questions like – "What an average person that subscribes looks like?", "How companies that run marketing campaigns can improve their campaign by calling their potential customers on

specific times? or using specific platform rather than the one that already used? , etc."

3. ANALYSIS & FINDINGS

3.1. Attributes

For the begging we conducted a general data analysis including exploring the different features we have and if they are qualified to help us answering our questions. Here is the list of all the attributes that was part of our dataset:

Table 3-1: Attributes

No.	Attribute	Description
Bank client data:		
1.	age	Age in years (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")
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).
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")

3.2. Preprocessing

The next step was preprocessing the data, choosing which attributes to keep. In order to cope with the goals we have in analyzing the data, we decided to drop some of the records and attributes which was either unnecessary or affecting the results in a way that not allowing exploring and modeling the data to get genuine conclusions.

3.3. Exploring the Data

To perceive our data better we created several visualizations of either a distribution of a variable or a bivariate relationship of two variables.

3.3.1. Age

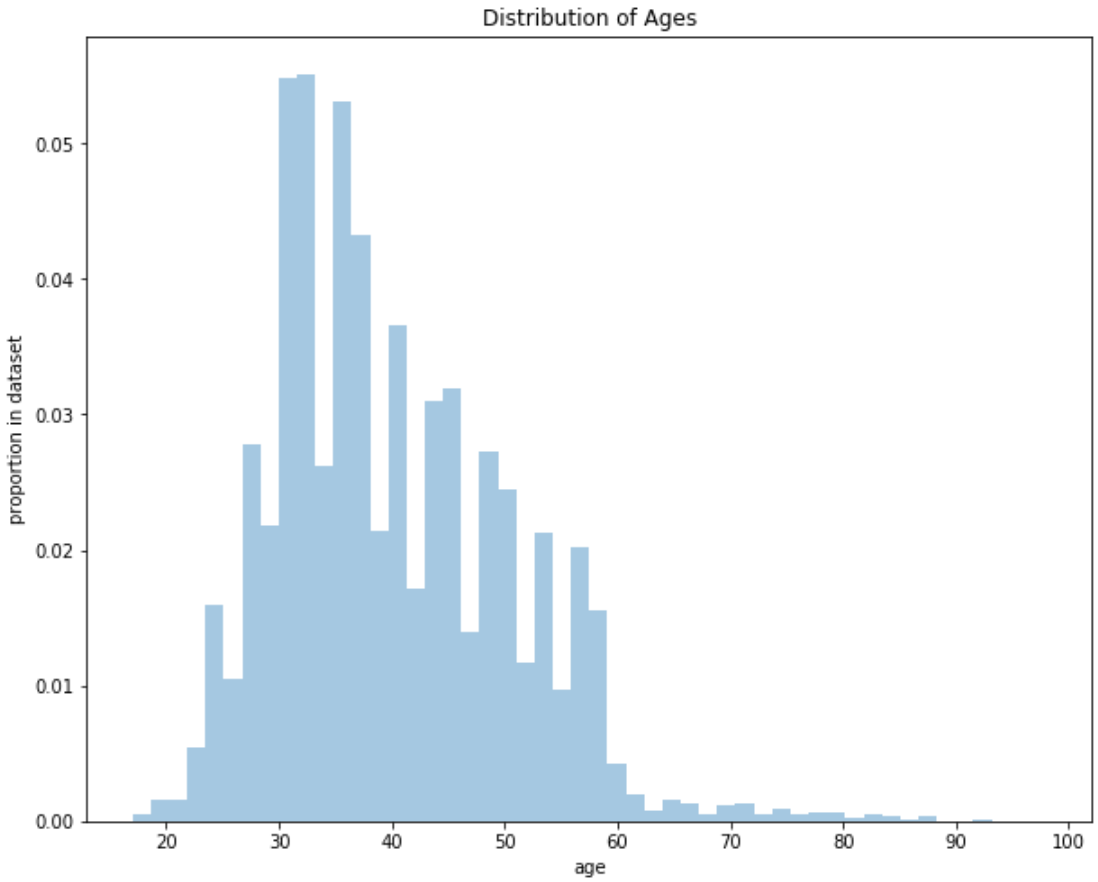


Figure 3-1: Distribution of Ages

3.3.2. Duration

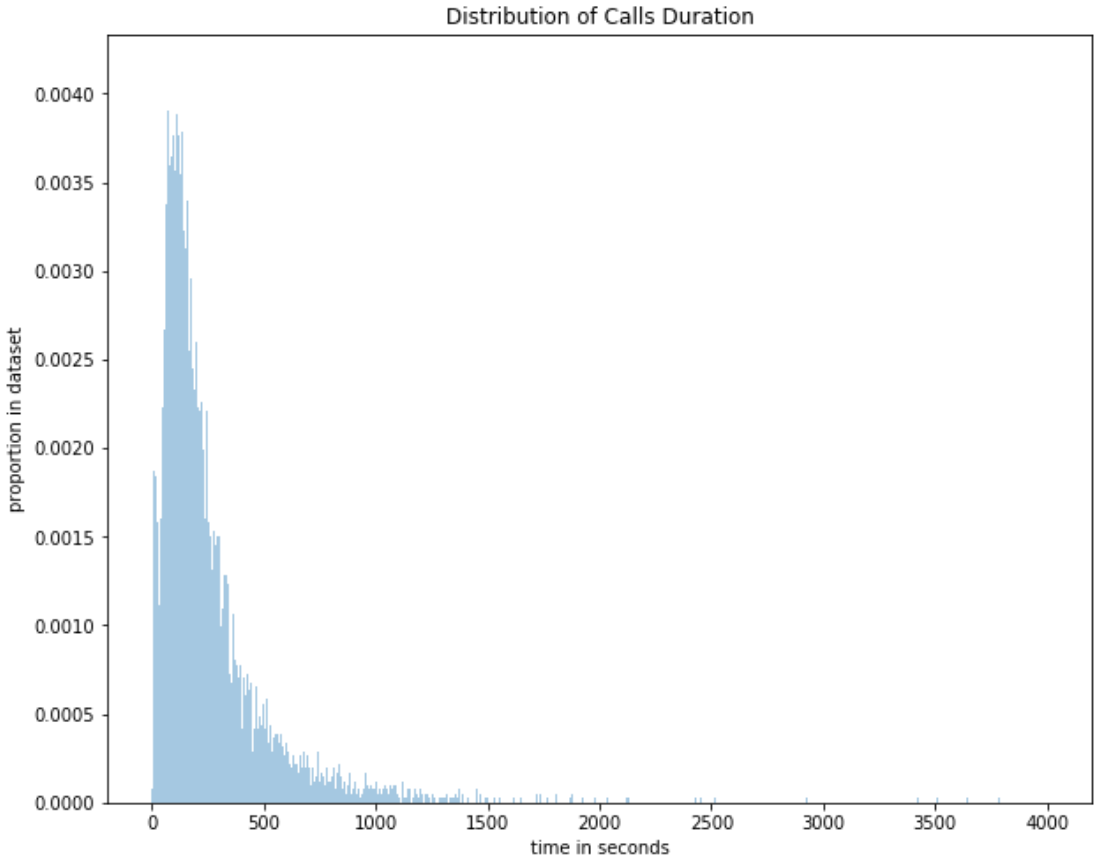


Figure 3-2: Distribution of Calls Duration

As we already seen in previous "time related" distributions, the duration has an exponential distribution.

3.3.3. Job

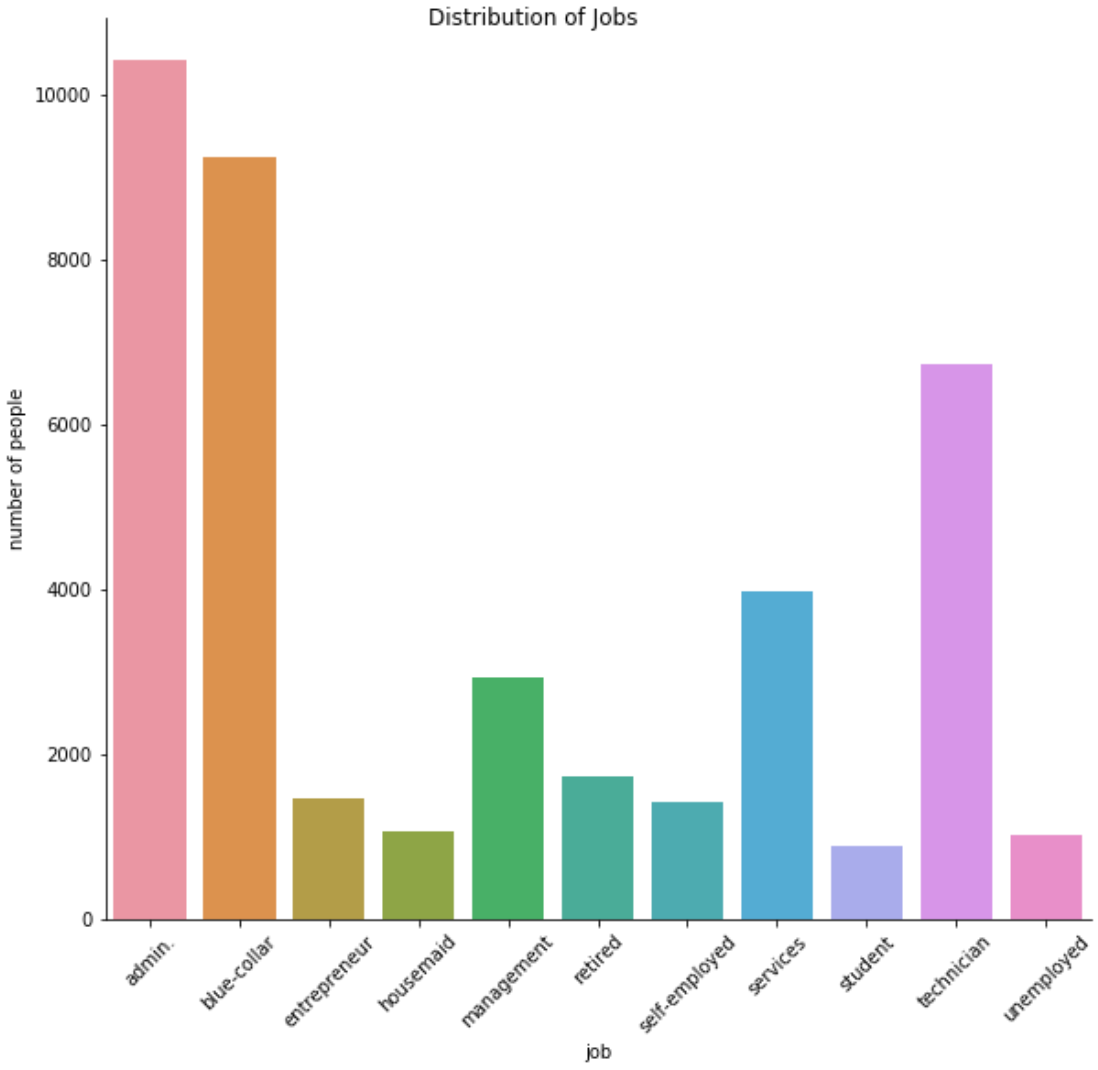


Figure 3-3: Distribution of Jobs

This distribution made us curious about the percentage of unemployment, so we created the following respected visualization.

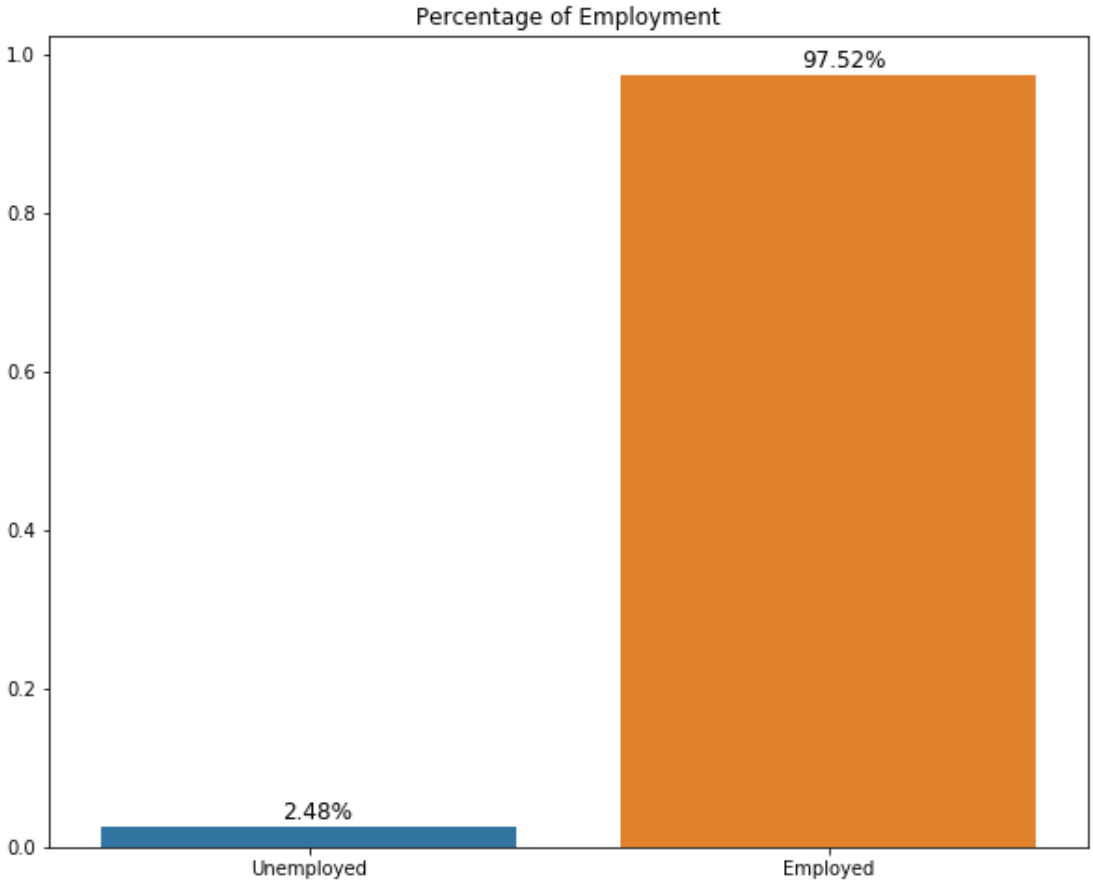


Figure 3-4: Percentage of Employment

3.3.4. Previous

As a reminder for the reader, the previous attribute represents the number of contacts performed before this campaign and for this client.

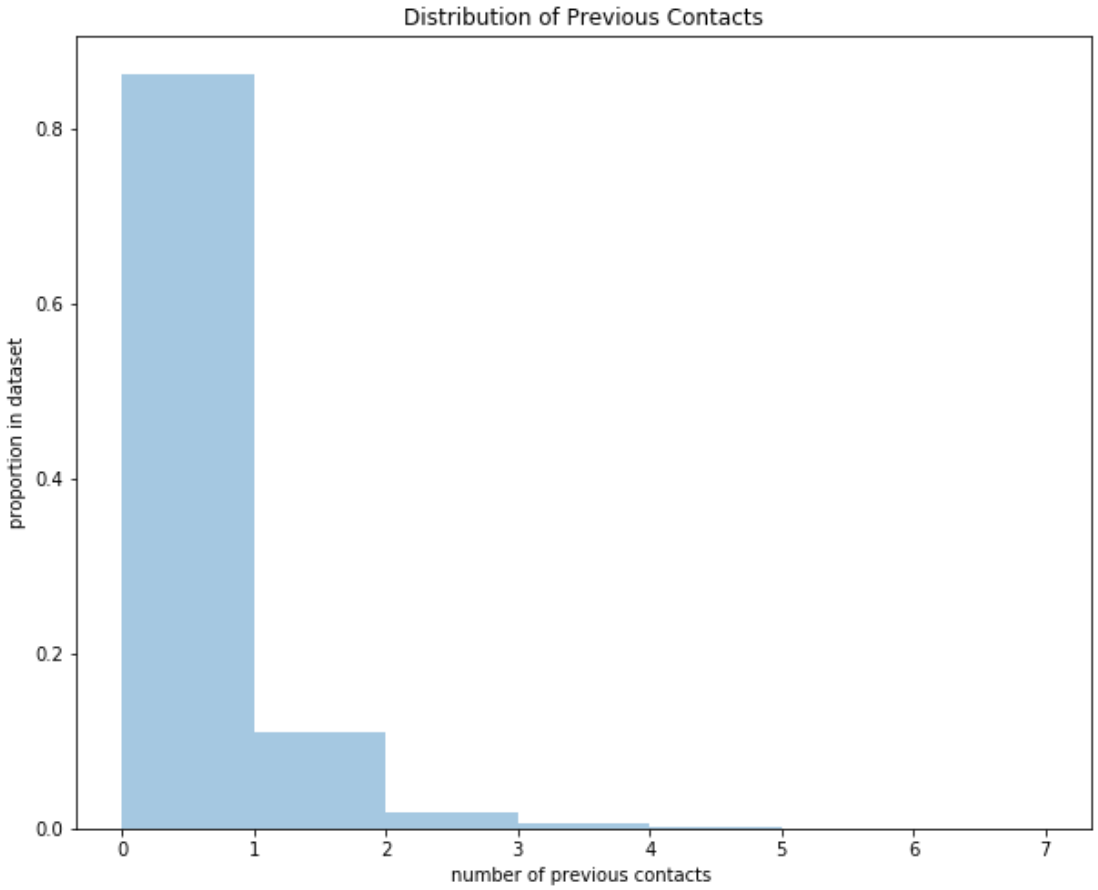


Figure 3-5: Distribution of Previous Contacts

We found the data of previously contacted ('previous' column) interesting and we are going to further examine it in question 3. We created a new binary attribute that separates people that have been contacted in the first time and those who have been contacted before. later we will check if this information affects the decision to subscribe a term deposit.

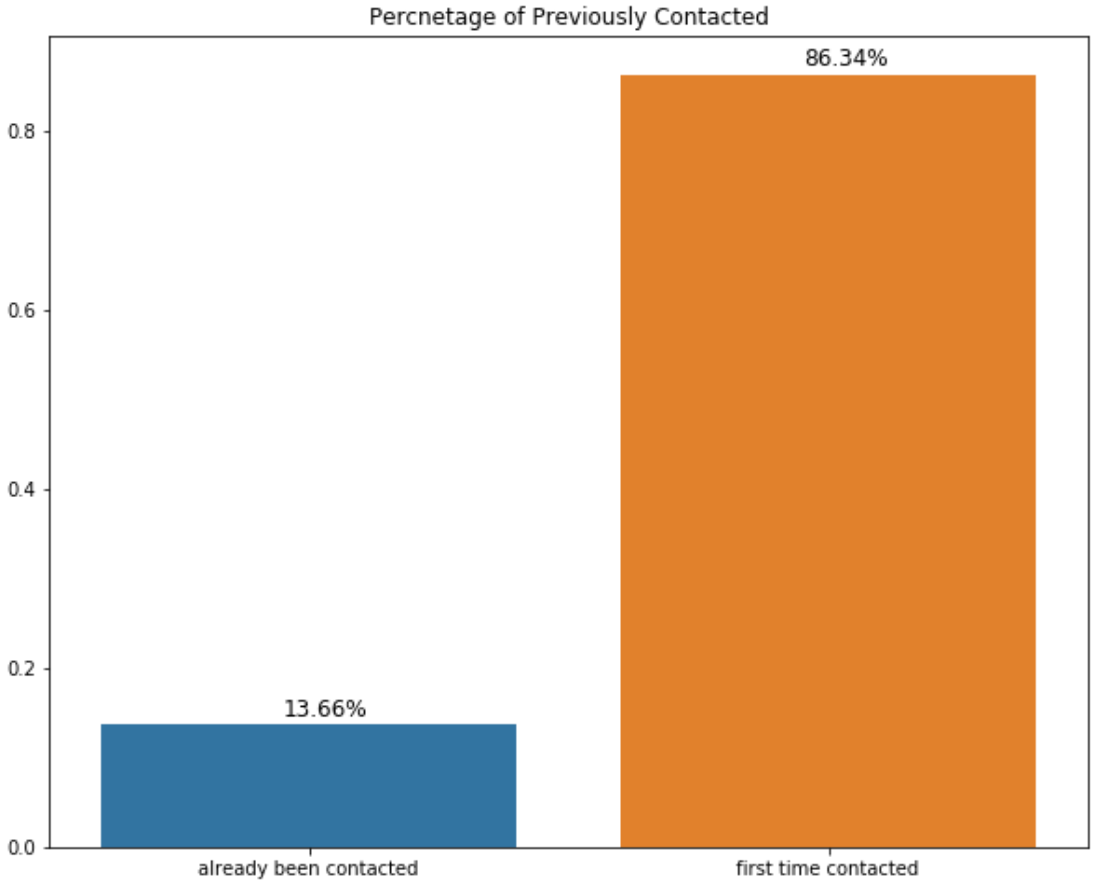


Figure 3-6: Percentage of Previously Contacted

3.3.5. Subscribed

We wanted to change the 'y' column from categorical to numerical so it will be more convenient to analyze and visualize it. In addition we want to give meaningful name to the column rather than y, because eventually the main goal of this dataset is to see how successful was the marketing campaigns (how many people have subscribed a term deposit), and this column directly represent it.

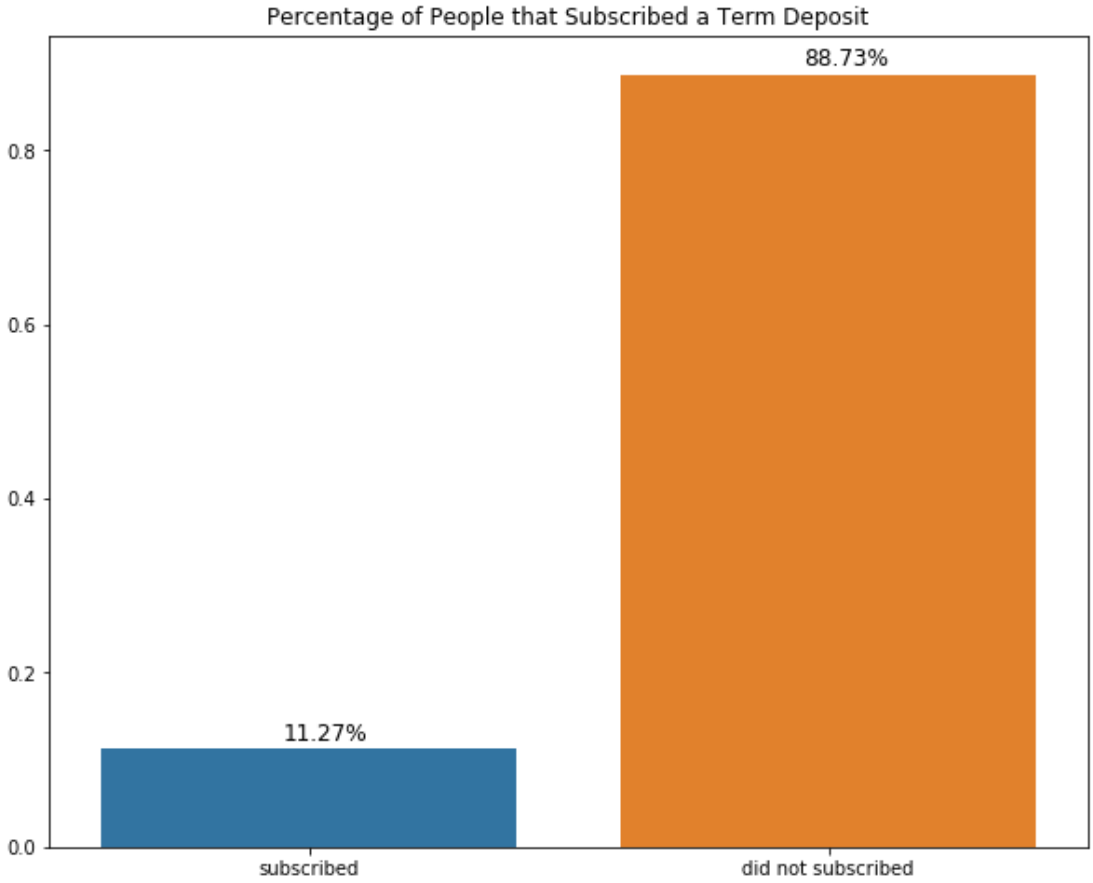


Figure 3-7: Percentage of People that Subscribed a Term Deposit

3.3.6. Campaign

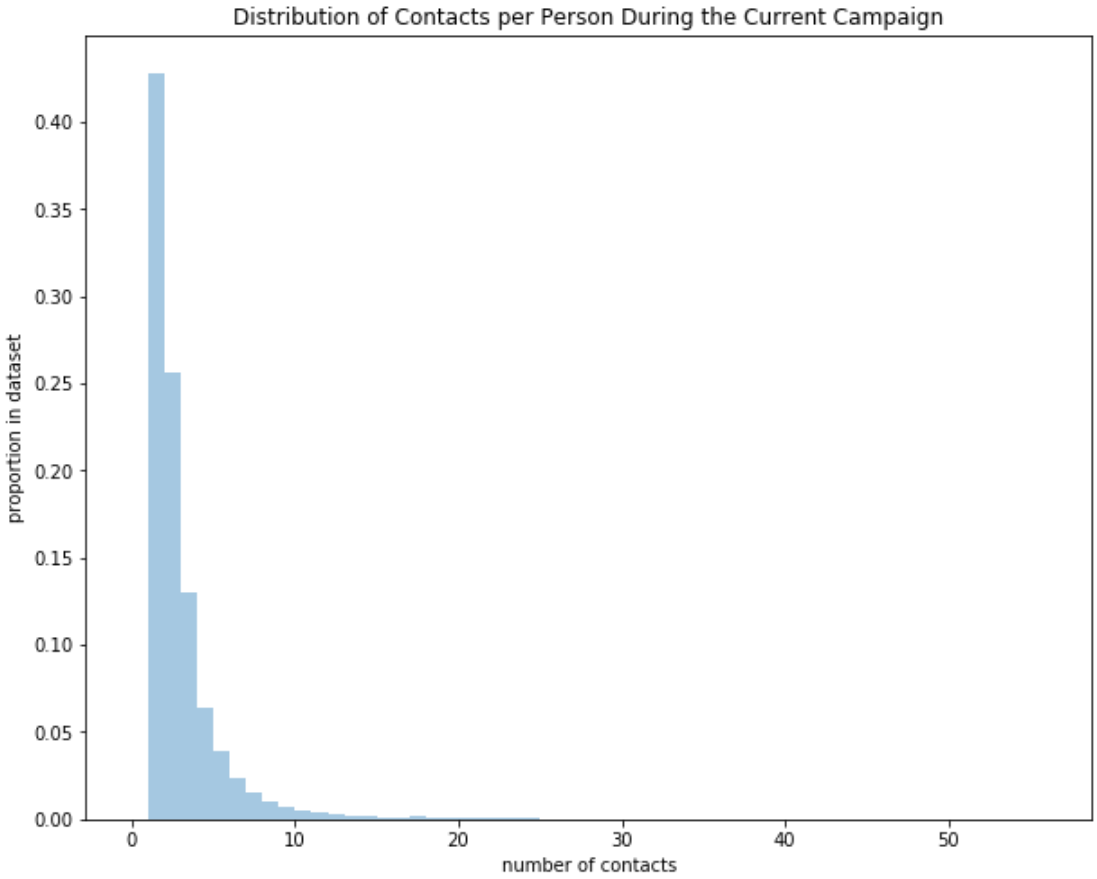


Figure 3-8: Distribution of Contacts per Person During the Current Campaign

We will try to display the graph with y axis on logarithmic scale to see if there are additional insights we can get.

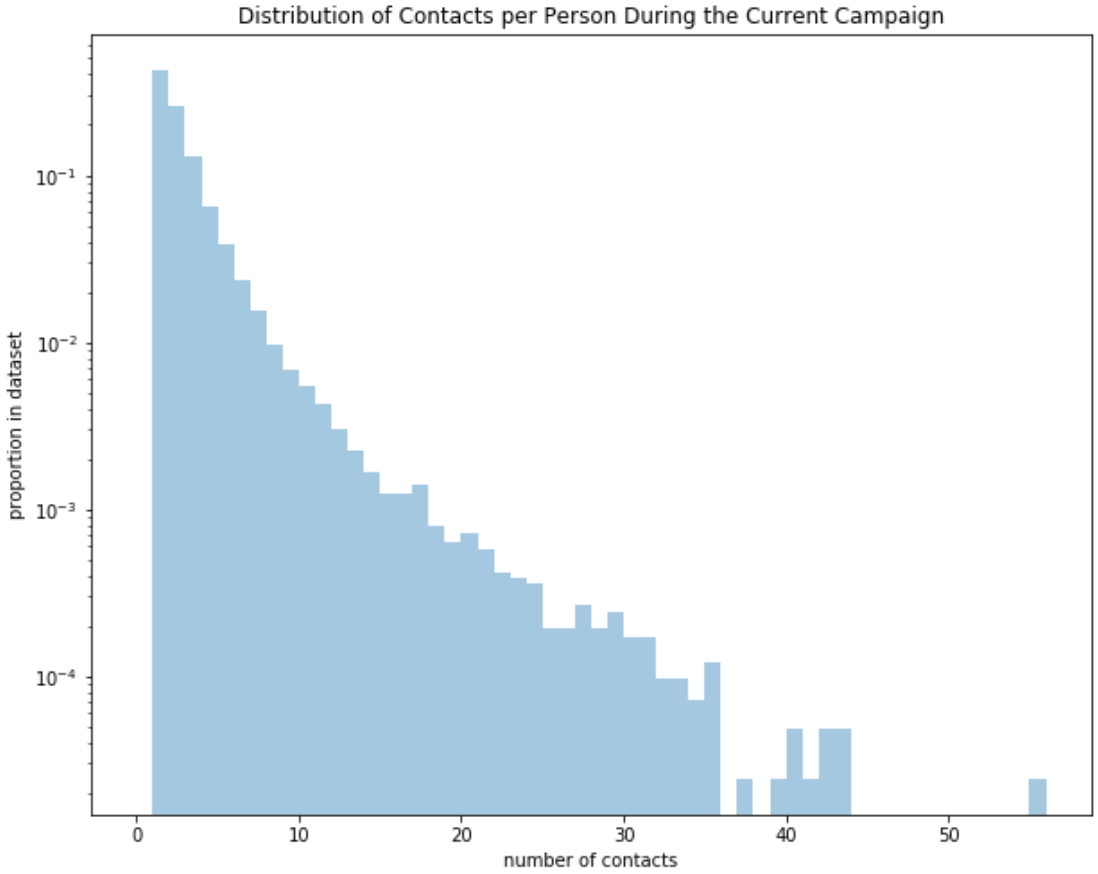


Figure 3-9: Distribution of Contacts per Person During the Current Campaign on Log Scale

we can infer from this distribution that during collecting the data there might have been some measurement bias because its sounds not reasonable that there is a person that have been contacted 56 times during one campaign!

3.3.7. Relationships

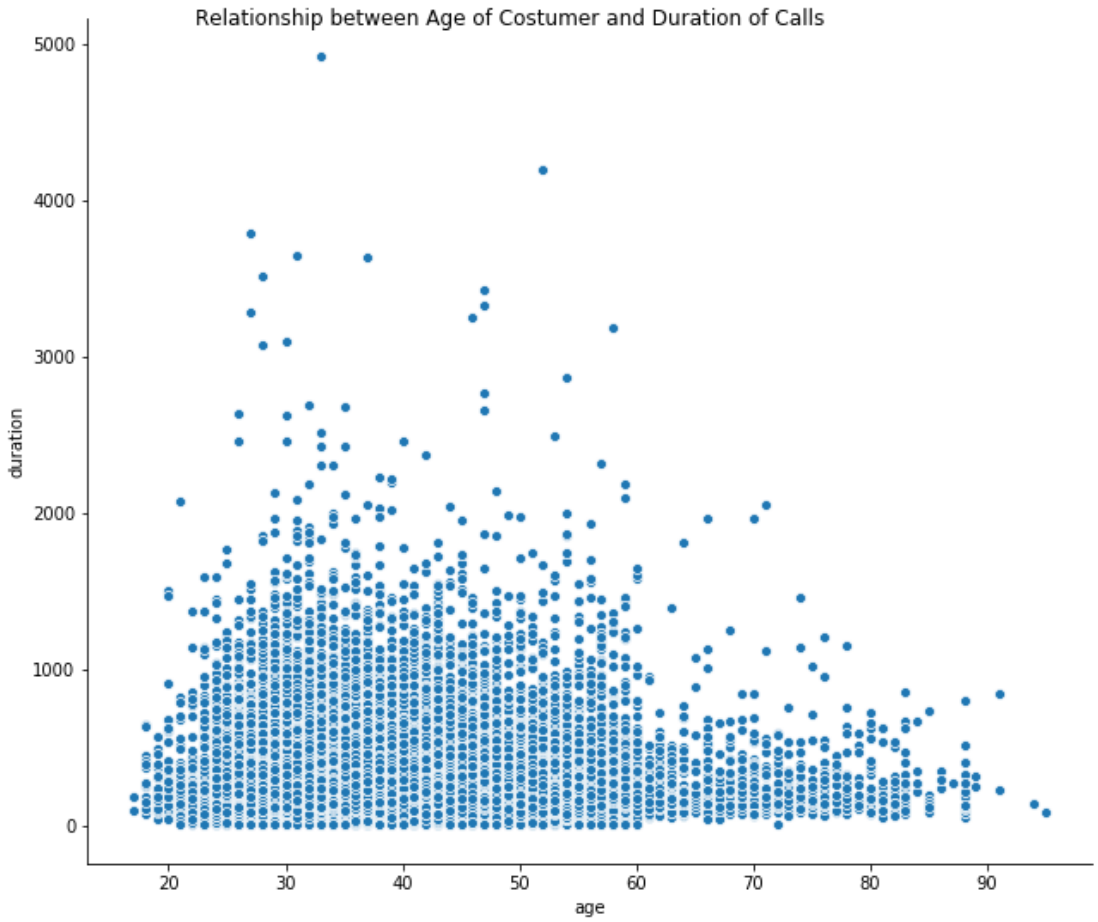


Figure 3-10: Relationship between Age of Costumer and Duration of Calls

We decided to test the correlation between age and duration. at first, we thought that there is might be a positive correlation but from the graph above it seems to be no correlation between them. we know that most of the people ages are between 30-50, that is why the biggest concentration of records is within that range.

The correlation using Pearson method is -0.0008149591515499178, i.e. no correlation.

The other correlation we are exploring is the correlation between calls duration to subscribing, we know that usually correlation visualization done using scatterplot graph but in our case there are many points that go on top of each other and it is not clear what is really going on so we used boxplot and barplot to check for correlation.

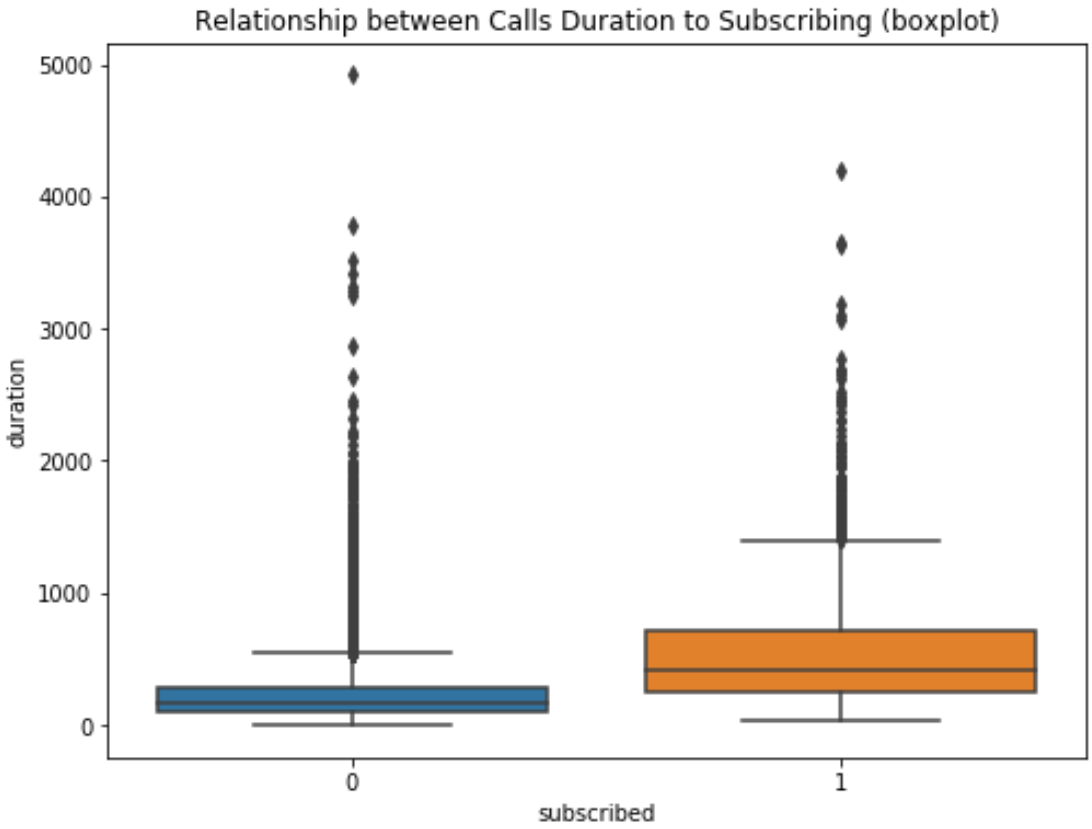


Figure 3-11: Relationship between Calls Duration to Subscribing (boxplot)

We can see from the visualization above that the 25th percentile, median and 75th percentile calls duration is higher among people that subscribed what might state on positive correlation between these parameters.

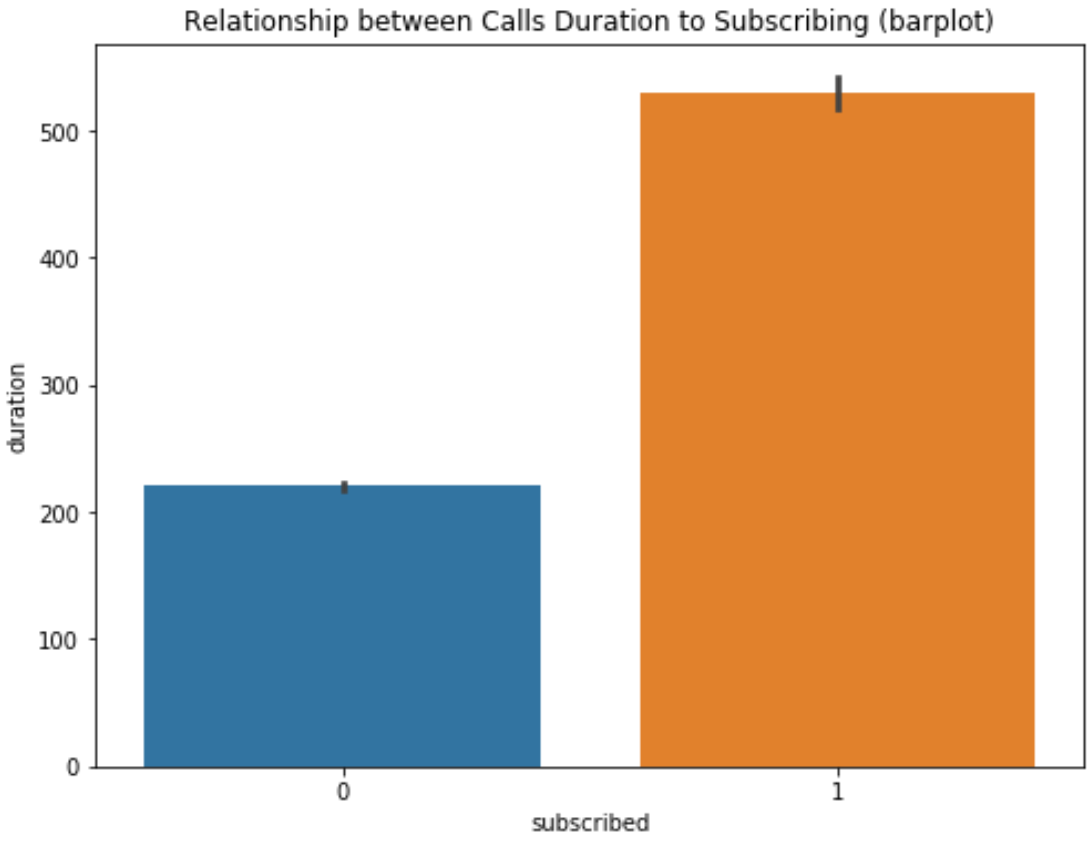


Figure 3-12: Relationship between Calls Duration to Subscribing (barplot)

It is clear from this graph that there is positive correlation between the two variables.

The correlation value using Pearson method is 0.3935099724688079.

3.4. Estimation and Hypothesis Testing

As mentioned in chapter 2 the question we decided to explore is: "are **people that get contacted several times have less chance to subscribe a term deposit?**"

When we thought about what might affect our decision if we were offered to subscribe a term deposit from our bank, we got to the conclusion that if we get the same offer or similar one several times, than the chance that we will agree to that offer is getting smaller.

We don't really have a profound reason for that but maybe something about that the bank contacts you more than one time about that same offer they have does the opposite - creates the feeling that it's not worthy and that the bank can't sell this product to its customers.

So we are going to formally state an hypothesis test it out using bootstrapping from our dataset.

3.4.1. Hypothesis

$\boldsymbol{H_0}$ – the mean number of previous contacts with people that have subscribed a term deposit is **equal** to the mean number of previous contacts with people that did not subscribed a term deposit.

$\boldsymbol{H_1}$ – the mean number of previous contacts with people that have subscribed a term deposit is **not equal** to the mean number of previous contacts with people that did not subscribed a term deposit.

3.4.2. Hypothesis Testing

The hypothesis testing made under the assumptions that we have a big and random dataset with diversity of adult men and women, what makes our dataset good representation of the population. This assumption is crucial to procced using bootstrapping to test our hypothesis.

According to the discipline learned in class, we chose to test the null hypothesis under 95% confidence interval. That means that to eliminate the null hypothesis the value of difference between the two means (which is 0) in our $\boldsymbol{H_0}$ should be found somewhere outside the 95% confidence interval.

3.4.3. Results

The 95% bootstrap confidence interval for difference between previous contacts means [0.3304528569258993, 0.3863377045902207].

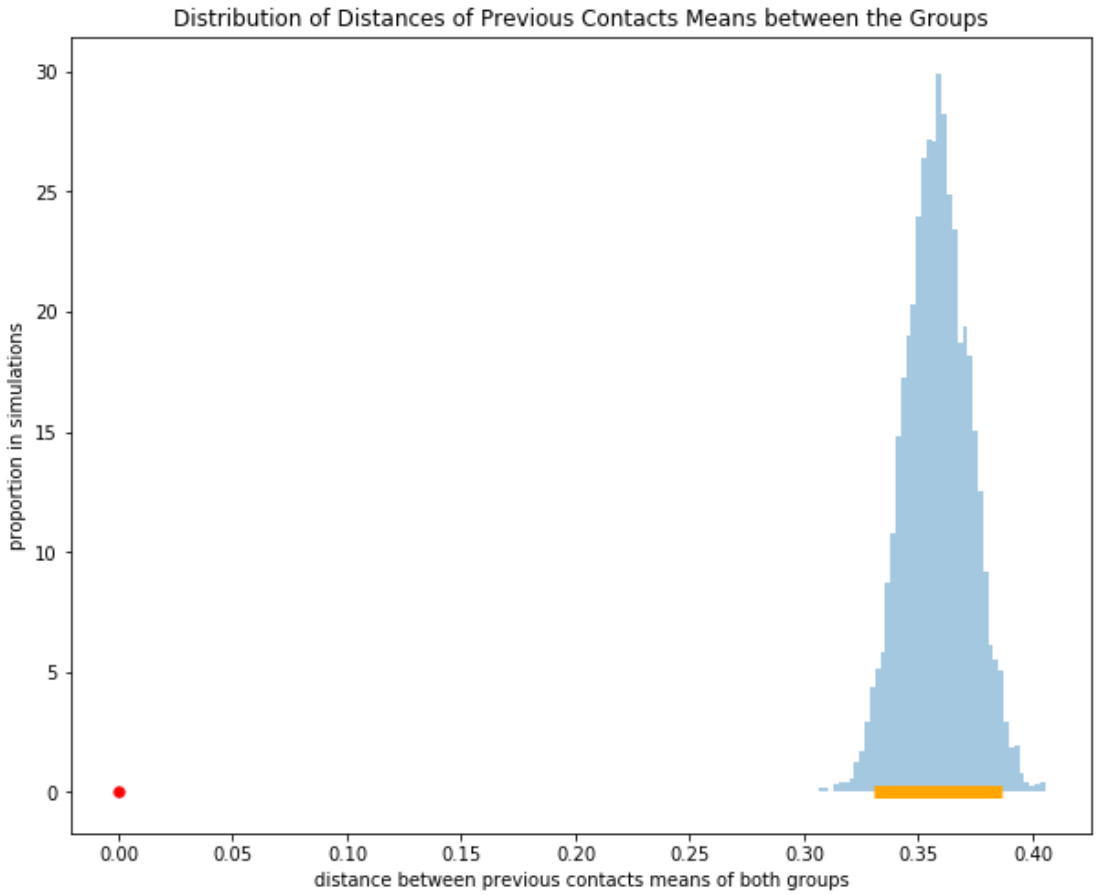


Figure 3-13: Distribution of Distances of Previous Contacts Means between the Groups with Null Hypothesis Marker

We can see that the value that suits the null hypothesis placed well outside the confidence interval which means we can eliminate the null hypothesis with 95% certainty!

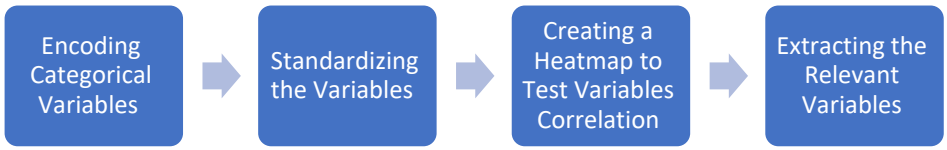
3.5. Prediction

3.5.1. The Question We Explored

As an extension to the questions presented in chapter 2; We wanted to classify people that subscribed a term deposit and those who do not. this information can be very meaningful if we have a data about a specific person and we would like to use the data about him if he is likely to subscribe a term deposit. Other option will be to give advices to the marketing team about best ways and times for interacting with their possible costumers using the information we can get from the data.

3.5.2. Data Processing Workflow

First, before choosing what features we want to use to predict the 'subscribed' target variable we need to check which of the features have a good correlation and what features we need to avoid in our classification. The steps were as follows:



3.5.3. Encoding Categorical Variables

The goal of this stage is to get equal distances between all possible values. We first dropped any rows that had 'unknown' values, after that we turn every categorical variable that had only two possible value to binary and eventually treated every variable that stayed as nominal categorical value and used one hot method to encode them.

3.5.4. Standardizing the Variables

The goal of this stage is to prevent prediction differs due to different units of the variables.

We used $\frac{X_i - \min(X)}{\max(X) - \min(X)}$ AKA min-max normalization to normalize all the values we have to scale of $[0,1]$.

3.5.5. Creating a Heatmap to Test Variables Correlation

The goal of this stage is to visualize the correlation values.

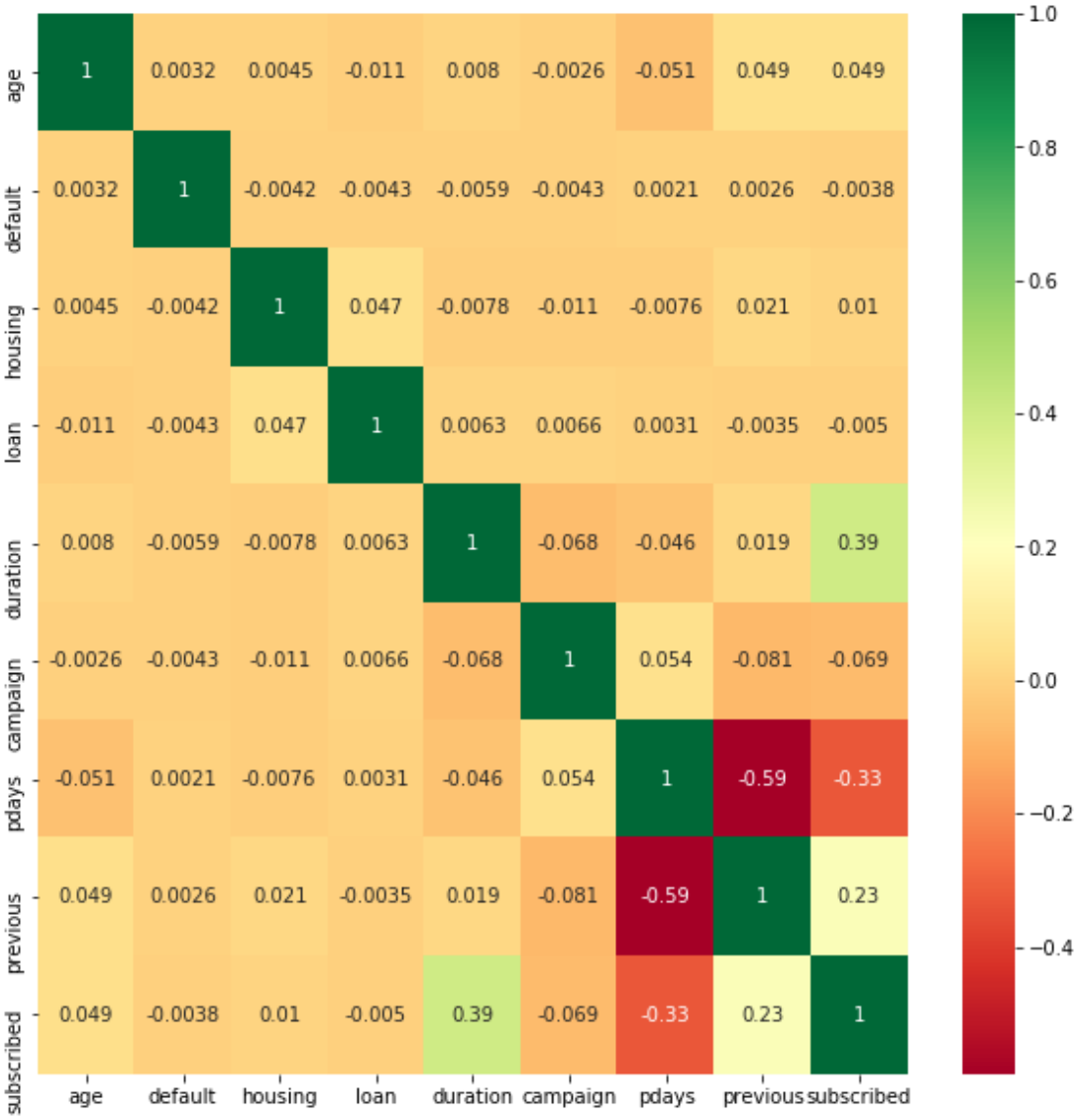


Figure 3-14: Correlation Heatmap Before Encoding

We can see here that the variables that have the highest correlation with the target are 'pdays' and 'duration'.

Note: we also created heatmap for after encoding the variables but because there are 50 variables after encoding, the figure is almost unreadable and

does not add any value to the results of the heatmap before encoding. This heatmap still remains part of our notebook as a reference and explanation for our choices.

3.5.6. Extracting the Relevant Variables

We chose to use the variables with the highest correlation with the target, so eventually we have 'pdays' and 'duration' as variables in the classification and 'subscribed' as the target containing two possible values – 1 for 'subscribed' and 0 for 'not subscribed'.

3.6. Classification Process

We will use cross validation to test range of different K (1 to 100) values to choose which one is the best and then were going to train a KNN classifier to predict the test results and compare it to the true values of the target. The results of the cross-validation appear in the following figure.

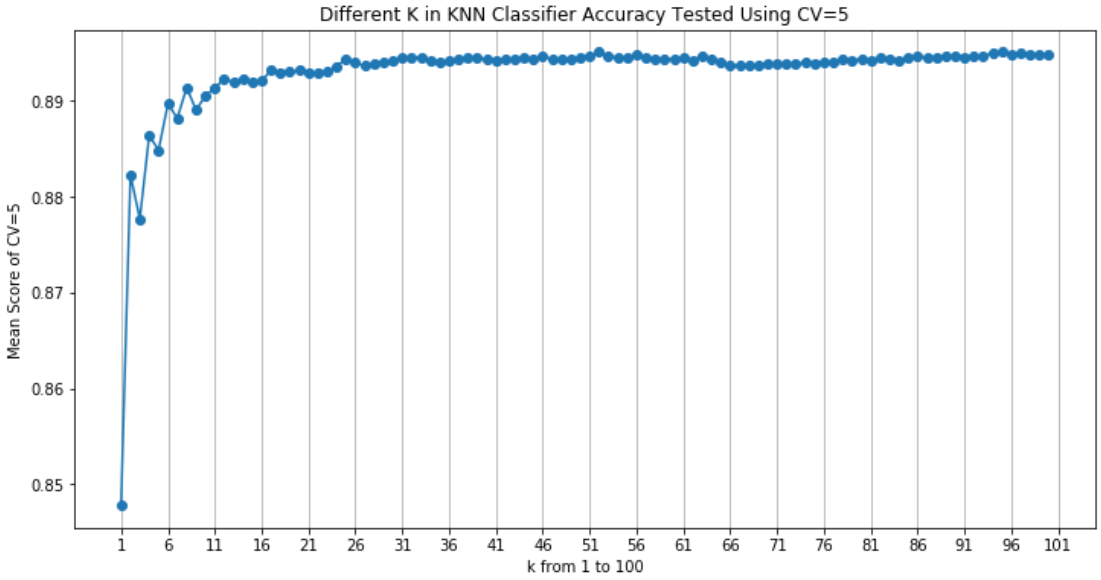


Figure 3-15: Different K in KNN Classifier Accuracy Tested Using CV=5

In this run the best K found using the cross-validation with parameter 5 was 51 with the score of 0.895. Of course, that in each different run the K can vary but we know that the score should be around the same result we got.

Finally, we can train our classifier (with 51 neighbors)!

And the score the classifier got using the test was 0.8948 which is close to the score we got using the cross-validation tests.

It can be observed that even though the correlation between the variables we used as our predictors in our KNN classification to the class was not very high (around |0.35|) and yet the classifier got a pretty good score that we are going to further analyze right away.

3.7. Results Analysis

3.7.1. Confusion Matrix

The confusion matrix is a summary of prediction results on a classification problem. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives insights not only into the errors being made by your classifier but more importantly the types of errors that are being made.

The Confusion Matrix we got in this run was as follows:

Table 3-2: Confusion Matrix

	True Label	**False Label**
True Classification	5179	133
False Classification	508	277

3.7.2. Rates

The interesting rates we are going to analyze according to the run we had will be:

- Sensitivity \ Recall
- Specificity
- Precision
- F1 Score

3.7.2.1. Sensitivity \ Recall

Probability of a positive classification given that observation is positive.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \equiv \frac{True\ Positive}{Positive}$$

The Recall rate is 0.91.

3.7.2.2. Precision

Precision rate is the probability of actual positive out of predicted positive.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \equiv \frac{True\ Positive}{Classifed\ as\ Positive}$$

The precision rate is 0.975.

3.7.2.3. Specificity

Probability of a negative classification given that observation is negative.

$$Specifity = \frac{True\ Negative}{True\ Negative + False\ Positive} \equiv \frac{True\ Negative}{Negative}$$

The Specificity rate is 0.675.

3.7.2.4. F1

F1 provides a single score that balances both the concerns of precision and recall in one number.

To assess the F1 score we will need to consider the precision and the recall calculations, F1 score is the harmonic mean of the precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Our F1 score is 0.941, and as we know as much as the score is close to 1 that means the classification is better.

3.7.3. Summary

In overall we managed to get very good results using the classification parameters we chose. The different rates we got suggests that when it comes to the important conclusion of classifying the individuals that are

going to subscribe to the service (Recall rate) we got a high score in addition to the F1 that as in our eyes gives a great "balanced" score for the classifier.

4. LIMITATIONS

Potential limitations and biases of our data:

- Categorical Data
- Low Correlation Between Variables
- Generic Variables
- Missing Values
- Selection Bias
- Measurement Bias

4.1. Categorical Variables

The dataset we analyzed composed of mainly categorical variables, which in terms of using the predictions algorithms we have learnt are not working as well as continuous or discrete variables (after encoding, of course). In addition, it was harder to find correlations between the categorical variables to the target class.

4.2. Low Correlation Between Variables

Finding high correlation between two variables in data was hard, we had to preprocess the data and the results was still disappointing for us – the highest correlations between the target variable and any other variable in the data according to the Pearson method was 0.39 which is pretty low in comparison to the scores we have seen in other datasets.

4.3. Generic Variables

Our dataset included some variables (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed) which have repeated values for many records, as they are representing an index measurement in the given time of the contact with the costumer. As a result, that information does not contribute for us in the analyzing process because we are not familiar with any tool that can help us process data that represented in that way.

4.4. Missing Values

Our data included over 10,000 records with missing data, what makes it difficult for us to use specific analysis actions on our data.

4.5. Selection Bias

We think that for answering our main question ”which people will subscribe a term deposit” we will need more information from more banks, because considering only our data when using bootstrap method may cause a selection bias, We do not know the financial state of the main crowed of this bank in relation to other banks.

Furthermore, to the fact that a proper randomization in sampling of all possible population is not achieved in that way.

4.6. Measurement Bias

In paragraph 3.3.6 we already mentioned a potential measurement bias we suspect that might happen. We know that this data is real data that have been documented by the bank workers and its very likely that some mistakes have occurred during this manual process.

5. FUTURE DIRECTIONS

As we already said in paragraph 4.5 for achieving our goal in analyzing the data and answering our main questions it would be better to data from more than one bank, which may represent wider parts of the population. It might also allow getting better conclusions about conducting this kind of marketing campaigns and even expanding them outside the boundaries of this specific bank customers.

5.1. Possible Question

During the process of processing and analyzing the data we had came up with idea of creating six tailored plans that can be offered for each costumer based on different data we have regarding him. That implies to the following questions – if there is a connection between a person salary and subscribing a term deposit? and if so, is it worth creating special plans for him?

This kind of questions can lead to a whole new data analysis project. The questions in that project will lead for further investigation of the improvement to the marketing campaign led by tailored fit plans using the costumer's information to the target of the campaign (subscribing a term deposit).

5.2. Additional Data

We think that data which describing the financial status of a costumer would be very helpful trying to answer the questions we raised, because this kind of information could imply better understanding of the customer needs and finical capabilities that can be used to improve the tailored plan to the right costumer.

As additional data, for begging we would be happy to have a column filled in by the suiting plan for each costumer (lest assume that there are 6 different plans that ranked by levels when subscribing), and we would use that to

explore if we can reach better sales. This data will later be used to have better understanding of which levels we should offer to each costumer by his specific data.