Direct Bank Marketing

Jyoti Mote

AIT 580

1.Abstract:

The financial sector is one of the sectors which is changed the most by analytics. In this project bank direct marketing data set is analyzed, the data is from UCI Machine Learning Repository. Bank-Full.csv is the data set which contains 17 attributes. Direct marketing is a form of marketing which is done through direct communicating or selling to individual customers. In this dataset direct marketing is used to present the product (bank term deposit) to clients over a phone call. In this paper the data is visualized and analyzed in order to check if the client has subscribed to the term deposit ('yes' or 'no') subscribed and the factors influencing it.

(Bertram, n.d.)

2.Introduction:

In the world of big data, data is the king, if it is used effectively, it can create huge impact on the business. When it comes to the finance market, analytics and big data are important. Firms are actively working on effective customer wooing and retention tactics. To increase revenues, financial and banking companies are learning how to balance Big Data with their services. Banks have strengthened their data patterns and repetitive operations by automating them. Financial data processing, data mining, and business analysis all use data visualization. It is the use of computer-assisted and immersive visual representation to enhance perception and communicate complex concepts underlying details. Charts, diagrams, and design features are used to effectively execute this strategy. In this paper we will see the various visualization techniques used to analyze the customers, the visualizations plots that are used are Stacked bar plot, grouped bar plot, box plot, count bar.

(Roy, 2019)

_(Alaa Abu-Srhan, 2019)

2.1 Research Questions:

The question answered in this paper are as follows:

1. Factors influencing the purchase of a bank product.
2. People with which profession have the maximum loan (personal loan)?
3. To find the average duration time of customers who have subscribed and not subscribed the bank term.
4. Analyze and compare between number of calls performed between current campaign and previous the campaign.

   It is important to study the dataset to know the factors influencing the purchase of the bank product. We can see which profession is taking the highest loan which will help us understand the overall financial needs with respect to the job. We can also analyze if there were any extra efforts put from the bank side to increase more customers during the campaign.

3.Literature review:

- **Bank Direct Marketing Analysis of Data Mining Techniques**

Summary:

Massive amounts of data are held in banks about their clients. This information can be used to build and maintain consistent relationships with customers in order to target them individually for particular goods or banking offers. Typically, selected consumers are approached directly by personal contact, telephone, mobile phone, mail, email, or some other contact to promote a new product/service or make an offer. This type of marketing is known as direct marketing. In reality, many banks and insurance firms use direct marketing as a primary method of engaging with their customers.

This paper is on the bank direct marketing data set written by Hany Elsalamony from Helwan University. Since these data sets are huge, it is difficult for a human researcher to locate interesting data that would assist in decision-making. They are using multilayer perception neural network (MPLNN), Naïve Bayes (TAN), logistic regression (LR), and C5.0 decision tree classification model. The data set has been used to evaluate the performance of the above listed Machine Learning techniques. The results of the experiments indicate that these models are more accurate in predicting the best campaign contact with clients for subscribing deposits. To measure the results, three mathematical methods are used: definition accuracy, sensitivity, and specificity.

Review:

The topic and data set are the same but the moto of in this paper they are checking the performance of the techniques and in my project, I am focusing on the factors that influence people to purchase the bank product.

(Elsalamony, 2013)

- **Visualization and Analysis in Bank Direct Marketing Prediction**

Summary:

The aim of this study was to develop a visualization mechanism for simple classification tasks. For a direct marketing campaign of a Portuguese bank, experiments were performed on an unbalanced data collection. Different oversampling methods, as well as five commonly used classifiers from the literature, were used in the experiments. A bank's direct marketing data collection is subjected to many visualization strategies in this report. To improve the accuracy of the forecast of a client's subscription to a term deposit, certain oversampling methods are employed. The impact of oversampling techniques on multiple classifier output is used to measure visualization reliability. The findings reveal that agglomerative hierarchical clustering outperforms other oversampling methods, with the Naive Bayes classifier delivering the best prediction results. This study is restricted to the use of the most commonly used oversampling methods, and it was only conducted on a small sample of a Portuguese banking institution's direct marketing strategy.

Review:

Their research is relevant to my research questions as they have also done analysis on diferent factors that are influencing customers to buy the bank product (term deposit), I would be using the same method and visualize to see the outcome in a single type of plot.

The difference between their research and my research questions and method is they are focusing on the different oversampling (visualization) methods, their performance and accuracy by calculating G-mean, I would be using a single type of plot and I would not be calculating the G-mean and accuracy of the visualizations and not all my research questions are answered.

(Alaa Abu-Srhan1, 2019)

- **Predicting Demographic and Financial Attributes in a Bank Marketing Dataset**

Summary:

In this research data from a Portuguese bank's telemarketing campaign was mined using data mining techniques. Individual customers are contacted by bank representatives with offers via telemarketing, for example. This telemarketing methods can be enhanced by integrating them with data processing techniques that make consumer knowledge and preferences more predictable. In this thesis, bank telemarketing data from a Portuguese banking institution was examined in order to ascertain the predictability of many customer demographic and financial characteristics, as well as the most important contributing factors in each. Data was preprocessed to ensure accuracy, and data mining models for the attributes were created using Orange's data mining platform, which included logistic regression, support vector machine (SVM), and random forest. Precision, memory, and F1 score were used to evaluate the findings.

Review:

The research paper is relevant to my research questions as they are analyzing the factors influencing the customers to purchase bank product. The duration of phone call for current and previous campaign, the methods used is the first method imputed the attributes and the class using the average/most common method; the second method did not impute the attributes and did not impute the class. I would be using the first method to calculate the same. People with which profession takes maximum loan is not reflected in this paper, Further research can be done on this topic and further methods can be found to do the same in my research questions.

(Ejaz, 2016)

4. Data:

4.1 Dataset:

The Data is linked to a Portuguese bank's direct marketing campaigns. Calls were used in the marketing campaigns. In order to determine if the product (bank term deposit) will be subscribed ('yes') or not ('no'), several contacts with the same client were often essential. The csv file taken is bank-full.csv which has 17 inputs. Attributes used for visualization are: [job], [marital], [education], [loan], [duration], [campaign], [poutcome], [y].

Column Name:	Description	Type
age	Age of the client	Nominal/Ordinal
job	type of job	Nominal: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'
marital	marital status	Nominal: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed

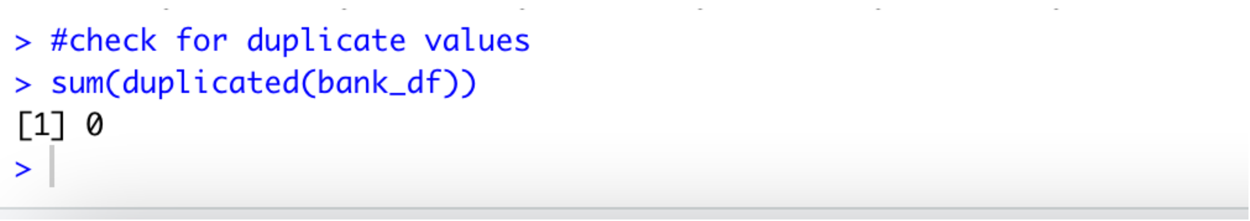
education	Education of a client	Nominal:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'
default	has credit in default?	Nominal: 'no','yes','unknown'
housing	has housing loan?	Nominal: 'no','yes','unknown'
loan	has personal loan?	Nominal: 'no','yes','unknown'
contact:	contact communication type	Nominal: 'cellular','telephone'
month	last contact month of year	Interval/ Ordinal Categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec'
day	Day of contact	Interval/ Ordinal: 'mon','tue','wed','thu','fri'
duration	last contact duration, in seconds	Ordinal
campaign	number of contacts performed during this campaign and for this client	Ordinal
pdays	number of days that passed by after the client was last contacted from a previous campaign	Ordinal
previous	number of contacts performed before this campaign and for this client	Ordinal
poutcome	outcome of the previous marketing campaign	Nominal: 'failure','nonexistent','success'
y	has the client subscribed a term deposit?	Nominal: 'yes','no'

(Bank Marketing Dataset, n.d.)

4.2 Data Cleaning and Validation:

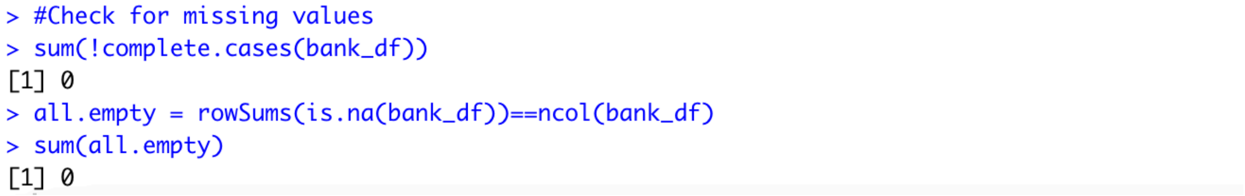
The data cleaning process is done in R, to check if the data is clean the following parameters were considered:

1. Check for duplicate values:
   There are no duplicate values in the dataset.



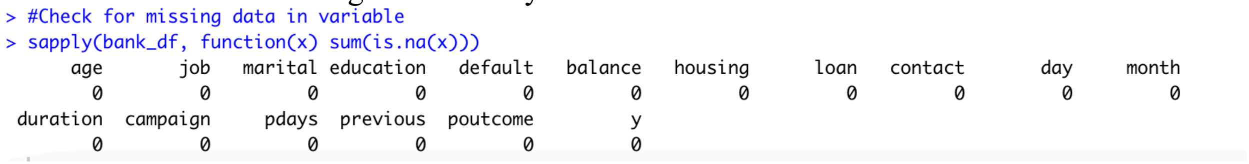
```
> #check for duplicate values
> sum(duplicated(bank_df))
[1] 0
> 
```

2. Check for missing values for all rows:
   There are no missing values in the rows.



```
> #Check for missing values
> sum(!complete.cases(bank_df))
[1] 0
> all.empty = rowSums(is.na(bank_df))==ncol(bank_df)
> sum(all.empty)
[1] 0
```

3. Check for missing values for individual attribute:
   There are no missing values in any attribute.



```
> #Check for missing data in variable
> sapply(bank_df, function(x) sum(is.na(x)))
      age       job   marital education   default   balance   housing      loan   contact       day     month 
        0         0         0         0         0         0         0         0         0         0         0 
 duration  campaign     pdays  previous  poutcome         y 
        0         0         0         0         0         0 
```

5.Visualizations:

To get an idea of how many clients have subscribed and not subscribed in total without considering any factors a simple bar plot is used. The total number of clients that have subscribed to the term is 5289 and the clients who have not subscribed is 39922.

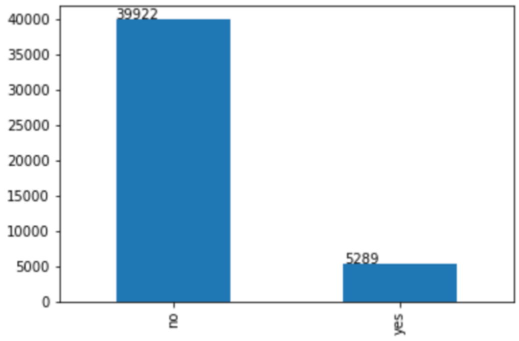


Fig1: Total count of clients who has subscribed and not subscribed to the term

Now let's see the factors influencing the purchase of a bank product. As it is important for any business to analyze the effecting their profit and loss, to check the same there are 3 factors considered:

1. Job
2. Marital Status
3. Education

I have used grouped bar plot to answer the first question as it shows each data category in a frequency distribution. Bar plots are easy to understand, as there are two categories 'Yes' and a 'No' for the term deposit grouped bar plot are the best fit to show this kind of information.

Visualization: Job vs Bank deposit.
Tool used: Jupiter Notebook
Plot used: Grouped Bar Plot

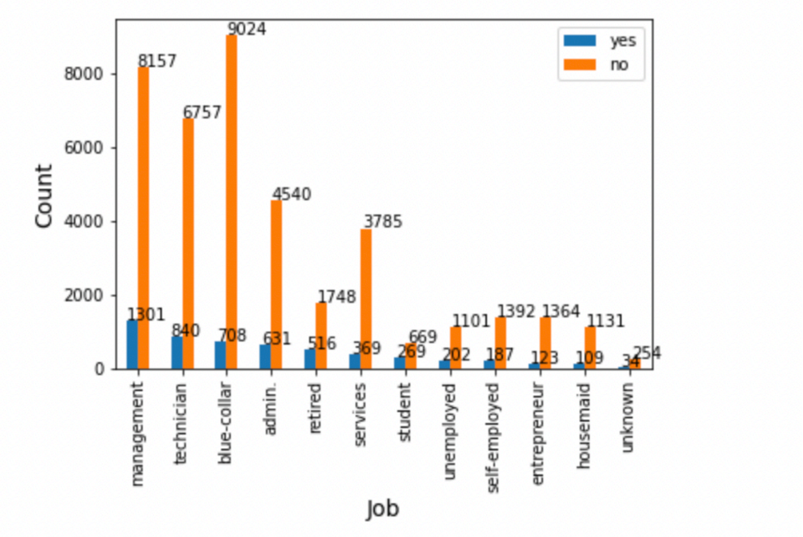


Fig2: Job vs Deposit

Looking at the above visualization we can clearly say that the maximum number of clients who have not subscribed for the deposit are having a blue-collar job. The maximum number of clients who have subscribed are from the management job.

Visualization: Education vs Bank deposit.
Tool used: Jupiter Notebook
Plot used: Grouped Bar Plot

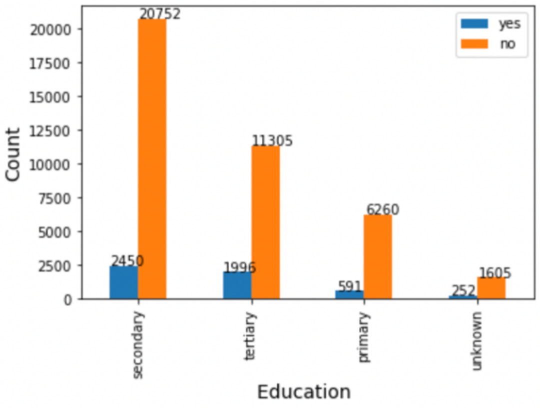


Fig3: Education vs Deposit

The visualization is for education and term deposit here the analysis is done based on education. The above plot depicts that clients who have education till secondary level have majorly not subscribed for the bank term. Looking at this plot we can also interpret that most clients contacted by the bank might be who have completed secondary education as the number of clients who have subscribed are also from secondary.

Visualization: Marital vs Bank deposit.
Tool used: Jupiter Notebook
Plot used: Grouped Bar Plot

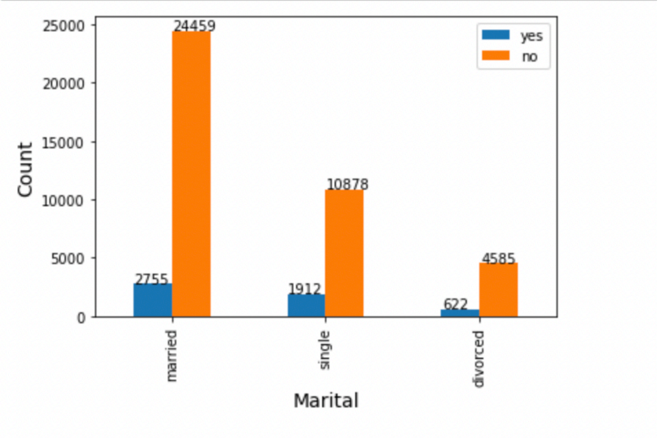


Fig4: Marital vs Deposit

The next factor considered is Marital Status, the Marital status is taken on the x axis and Count of deposit is taken on the y axis, here majority of the clients who have not subscribed to the bank term deposit are who are married.

The second question people with which profession have the maximum loan (personal loan)? Is answered using the below Visualization. I have used a stacked bar plot because when we have to compare the results, it is much easy to do comparison when the results are stacked on top of each other.

Visualization: Job vs Loan.
Tool used: R Studio
Plot used: Stacked Bar Plot

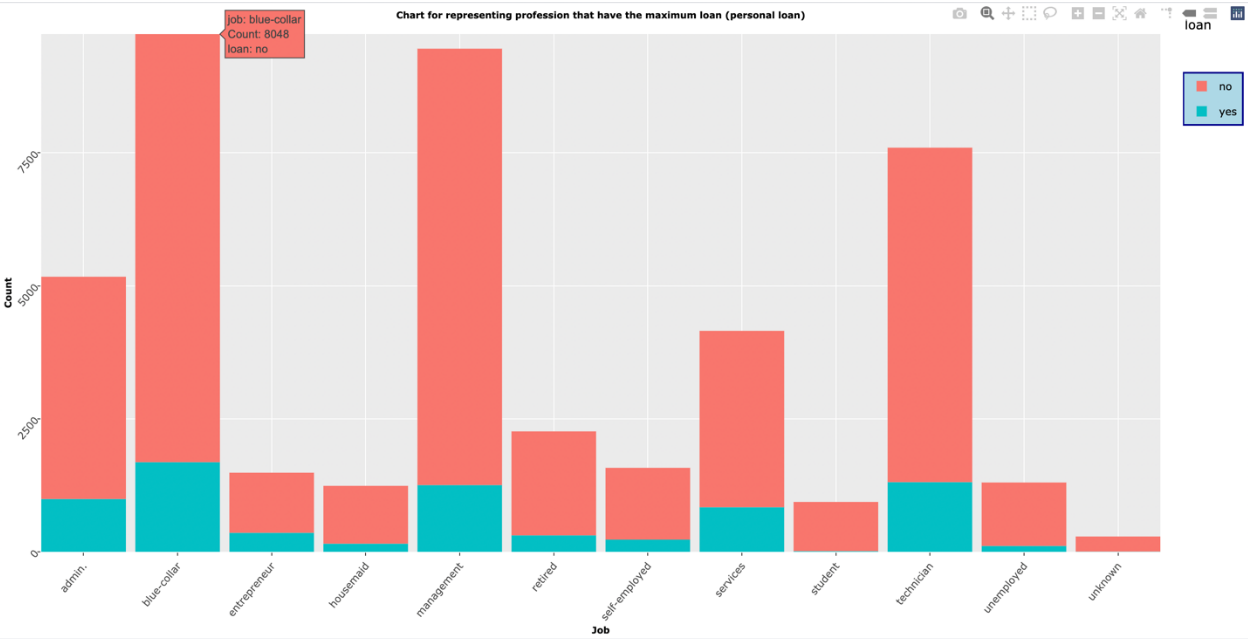


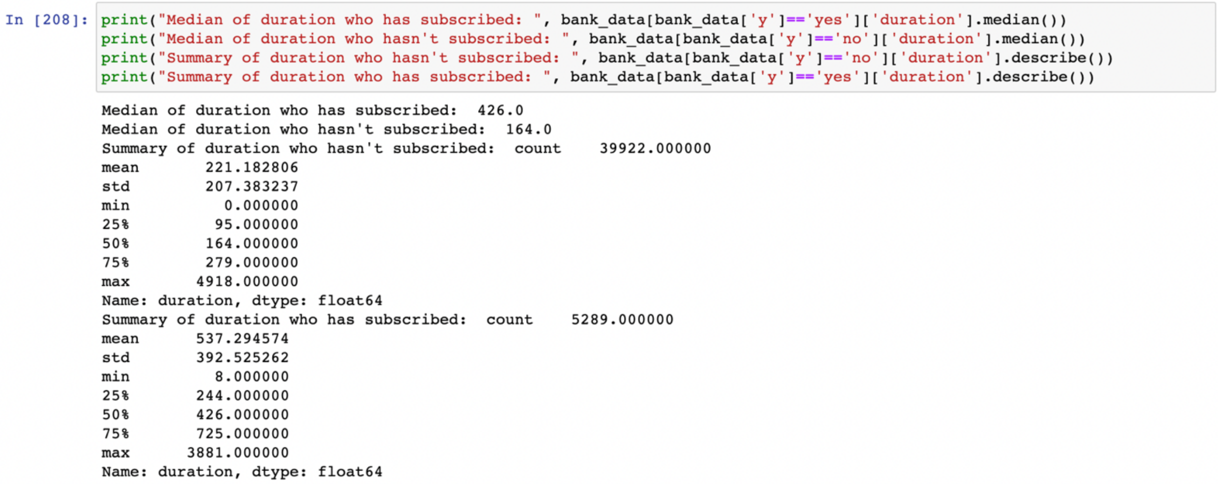
Fig5: Job vs Loan

The profession which has the maximum loan is blue-collar. To represent this a stacked bar plot is used, where the job is taken on the x axis and loan is taken on the y axis. Dynamic data labeling is used to show the data labels. The profession which has least loan is unknown followed by students.

Looking at this graph and comparing it with the Job vs deposit graph that we used to answer the first question we can interpret that blue-collar profession has the maximum loan and blue-collar profession had not subscribed to the bank term we can say that the main problem here can be people with blue-collar profession might not have enough money to subscribe for the deposit.

To answer the third question a box plot is used as to find the average duration time of customers who have subscribed and not subscribed to the bank term and box plot is best to Summarize variation in large datasets visually it also displays the outliers if any.

Visualization: Duration vs Deposit.
Tool used: Jupiter Notebook
Plot used: Box Plot



```
In [208]: print("Median of duration who has subscribed: ", bank_data[bank_data['y']=='yes']['duration'].median())
          print("Median of duration who hasn't subscribed: ", bank_data[bank_data['y']=='no']['duration'].median())
          print("Summary of duration who hasn't subscribed: ", bank_data[bank_data['y']=='no']['duration'].describe())
          print("Summary of duration who has subscribed: ", bank_data[bank_data['y']=='yes']['duration'].describe())
```

```
Median of duration who has subscribed:  426.0
Median of duration who hasn't subscribed:  164.0
Summary of duration who hasn't subscribed:  count    39922.000000
mean       221.182806
std        207.383237
min          0.000000
25%         95.000000
50%        164.000000
75%        279.000000
max       4918.000000
Name: duration, dtype: float64
Summary of duration who has subscribed:  count    5289.000000
mean      537.294574
std       392.525262
min         8.000000
25%       244.000000
50%       426.000000
75%       725.000000
max      3881.000000
Name: duration, dtype: float64
```

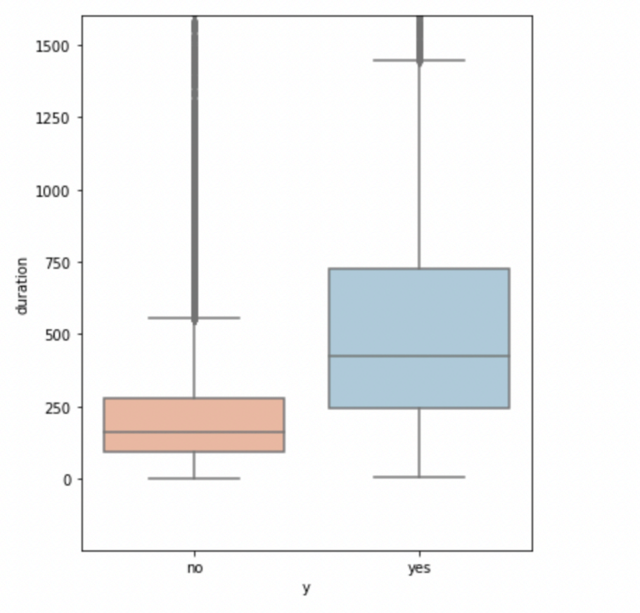


Fig6: Duration vs Deposit

The average of duration who has subscribed is 537.294(in seconds). The average of duration who has not subscribed is 221.182(in seconds). Looking at the values we can say that clients who stay on the call longer are more likely to subscribe.

Now coming to the fourth question we need to analyze and compare between number of calls performed between current campaign and previous the campaign to analyze these two parameters are taken into consideration the column [campaign] and [poutcome]. A count plot is used for this visualization as it shows the number of occurrences based on the category here, we are analyzing the count of clients who have subscribed and not subscribed by the number of contacts done in the campaign and comparing with the results of the previous campaign. Count plot is the best fit for such kind of visualization.

Visualization: Campaign vs Deposit and poutcome vs Deposit
Tool used: Jupiter Notebook
Plot used: Count plot

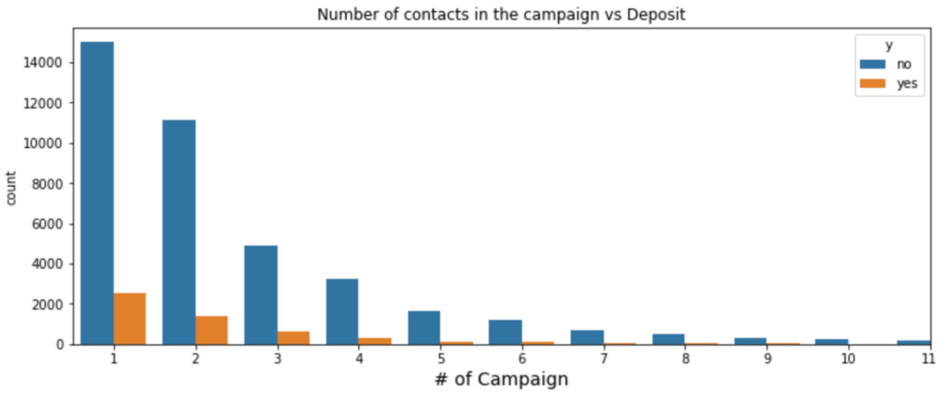


Fig7: Campaign vs Deposit

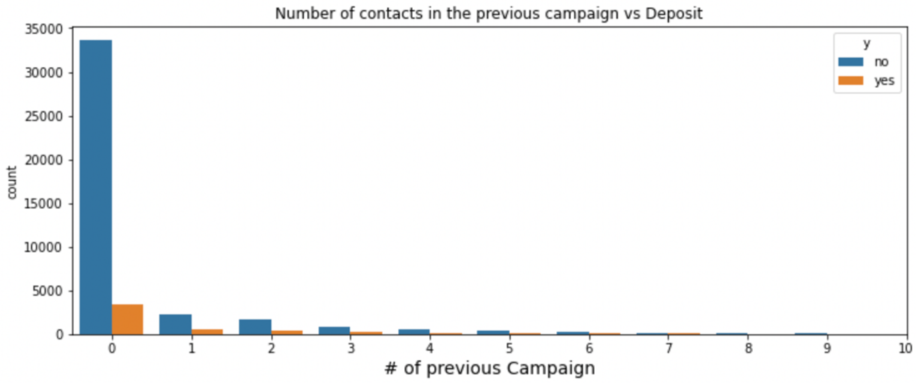


Fig8: Previous vs Deposit

Looking at the above plot we can say that the number of contacts to the client is increasing the subscription to the deposit is decreasing in both the campaign.

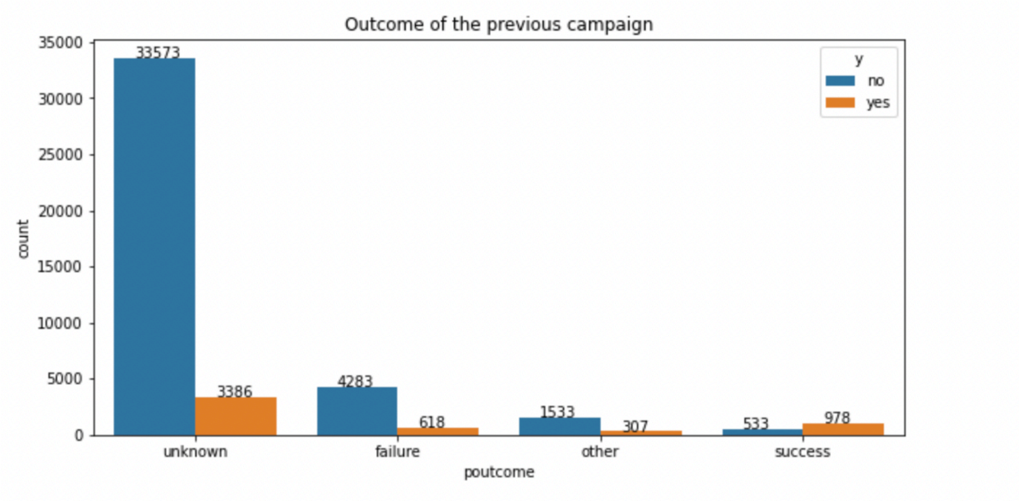


Fig8: poutcome vs Deposit

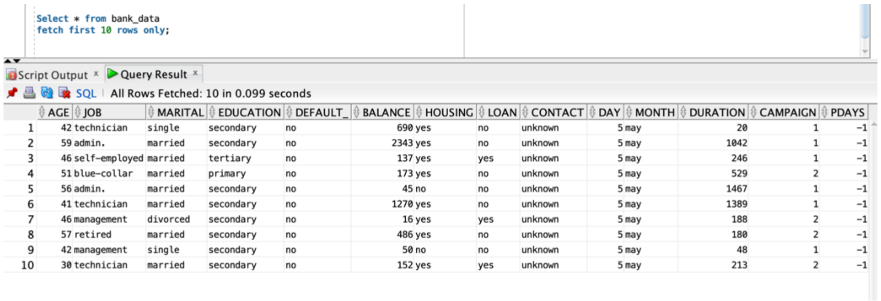
Looking at the previous campaign outcome plot we can say that for success category there are high chances every second customer would subscribe to the deposit. There are chances that if the previous contact is a failure there are chances, they would subscribe to the deposit.

6.SQL:

Create table Command to create a table:

```
CREATE TABLE "JMOTE"."BANK_DATA"

  (          "AGE" NUMBER(38,0),

             "JOB" VARCHAR2(26 BYTE),

             "MARITAL" VARCHAR2(26 BYTE),

             "EDUCATION" VARCHAR2(26 BYTE),

             "DEFAULT_" VARCHAR2(26 BYTE),

             "BALANCE" NUMBER(38,0),

             "HOUSING" VARCHAR2(26 BYTE),
             "LOAN" VARCHAR2(26 BYTE),

             "CONTACT" VARCHAR2(26 BYTE),

             "DAY" NUMBER(38,0),

             "MONTH" VARCHAR2(26 BYTE),

             "DURATION" NUMBER(38,0),

             "CAMPAIGN" NUMBER(38,0),

             "PDAYS" NUMBER(38,0),

             "PREVIOUS" NUMBER(38,0),

             "POUTCOME" VARCHAR2(26 BYTE),

             "Y" VARCHAR2(26 BYTE)

  )
```

Select Statement:



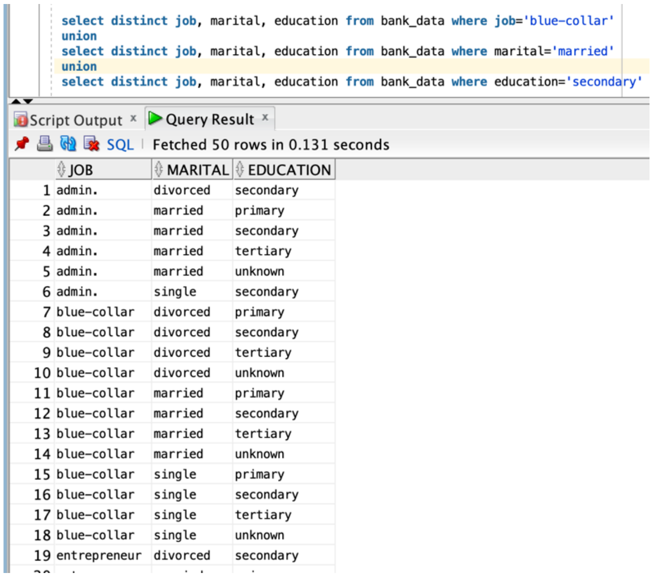
```
Select * from bank_data
fetch first 10 rows only;
```

All Rows Fetched: 10 in 0.099 seconds

	AGE	JOB	MARITAL	EDUCATION	DEFAULT_	BALANCE	HOUSING	LOAN	CONTACT	DAY	MONTH	DURATION	CAMPAIGN	PDAYS
1	42	technician	single	secondary	no	690	yes	no	unknown	5	may	20	1	-1
2	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1
3	46	self-employed	married	tertiary	no	137	yes	yes	unknown	5	may	246	1	-1
4	51	blue-collar	married	primary	no	173	yes	no	unknown	5	may	529	2	-1
5	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1
6	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1
7	46	management	divorced	secondary	no	16	yes	yes	unknown	5	may	188	2	-1
8	57	retired	married	secondary	no	486	yes	no	unknown	5	may	180	2	-1
9	42	management	single	secondary	no	50	no	no	unknown	5	may	48	1	-1
10	30	technician	married	secondary	no	152	yes	yes	unknown	5	may	213	2	-1

Fig 9: Select top 10 records

Statement using union



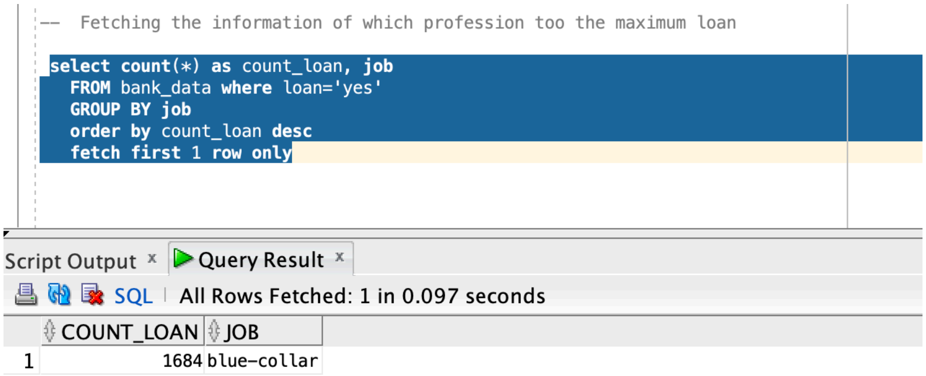
```
select distinct job, marital, education from bank_data where job='blue-collar'
union
select distinct job, marital, education from bank_data where marital='married'
union
select distinct job, marital, education from bank_data where education='secondary'
```

Fetched 50 rows in 0.131 seconds

	JOB	MARITAL	EDUCATION
1	admin.	divorced	secondary
2	admin.	married	primary
3	admin.	married	secondary
4	admin.	married	tertiary
5	admin.	married	unknown
6	admin.	single	secondary
7	blue-collar	divorced	primary
8	blue-collar	divorced	secondary
9	blue-collar	divorced	tertiary
10	blue-collar	divorced	unknown
11	blue-collar	married	primary
12	blue-collar	married	secondary
13	blue-collar	married	tertiary
14	blue-collar	married	unknown
15	blue-collar	single	primary
16	blue-collar	single	secondary
17	blue-collar	single	tertiary
18	blue-collar	single	unknown
19	entrepreneur	divorced	secondary

Fig 10: Fetching the information of factors who have the majorly not subscribed to the term deposit.

Statement to check which profession has taken the maximum loan



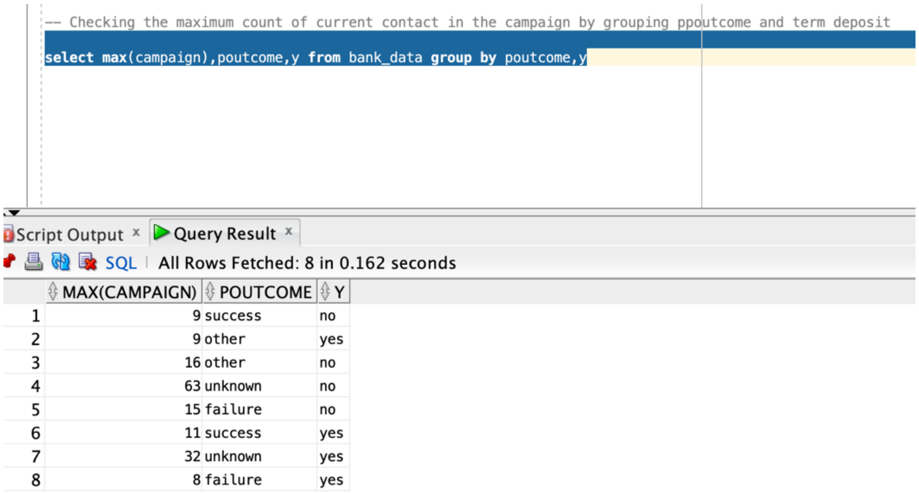
```
--  Fetching the information of which profession too the maximum loan

select count(*) as count_loan, job
  FROM bank_data where loan='yes'
  GROUP BY job
  order by count_loan desc
  fetch first 1 row only
```

All Rows Fetched: 1 in 0.097 seconds

	COUNT_LOAN	JOB
1	1684	blue-collar

Fig 10: Fetching the information of which profession too the maximum loan

Fetching the data of the maximum number of contacts in the current campaign by grouping it to the outcome of previous campaign and term.



```
-- Checking the maximum count of current contact in the campaign by grouping ppoutcome and term deposit

select max(campaign),poutcome,y from bank_data group by poutcome,y
```

All Rows Fetched: 8 in 0.162 seconds

	MAX(CAMPAIGN)	POUTCOME	Y
1	9	success	no
2	9	other	yes
3	16	other	no
4	63	unknown	no
5	15	failure	no
6	11	success	yes
7	32	unknown	yes
8	8	failure	yes

Fig 11: Checking the maximum count of current contact in the campaign by grouping poutcome and term deposit

7. Limitations and Future Scope:

In this paper I have only selected few factors to analyze the subscription of the bank product. Here there is a lot of scope to check with respect to other factors such as age, balance, month, year etc. Other attributes can also be used for analysis out of 17 I have only utilized the following attributes: [job], [marital], [education], [loan], [duration], [campaign], [poutcome], [y]. pdays, previous could also be used for analyzing and visualizing.
There is a vast scope to use this data set for regression modelling and predicting the deposit. I have not explored the predictive modelling in this paper. The classifiers can be applied and tested for the best performance.

8. Lessons Learnt:

The lessons learnt after doing this project are I got to know about the factors that are influencing the subscription of the deposit. Certain parameters are not influential. I also learnt about various plots in python and R. Like stacked bar plot with dynamic labels and box plot.

8. Conclusion:

The aim of this paper is to visualize and analyze the different parameters and factors influencing the purchase of the bank product (Deposit). After considering different factors it is seen that clients having blue-collar job has the highest count of not subscribing to the deposit, and the maximum of clients who have subscribed from are from management job. We can also conclude that clients from blue-collar job have taken the highest loan (personal loan) which can be interpreted as blue-collar job clients do not have enough money, so they have high tendency to not subscribe to the deposit. The average duration of clients who have subscribed is more than that of who did not subscribe. The clients who stay on the call for longer time will mostly subscribe. There is a lot of scope for further studies and more precise conclusions can be made by other performance paraments and measures.

References:

7 Big Data Examples: Applications of Big Data in Real Life. (2016, July 13). Retrieved from Intelli Paat: https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/

Alaa Abu-Srhan, S. B. (2019). *Visualization and Analysis in Bank Direct Marketing Prediction*. Retrieved from Visualization and Analysis in Bank Direct Marketing Prediction: https://thesai.org/Downloads/Volume10No7/Paper_85-Visualization_and_Analysis_in_Bank_Direct_Marketing.pdf

Alaa Abu-Srhan1, S. A.-S. (2019). *Visualization and Analysis in Bank Direct Marketing Prediction*. Retrieved from Visualization and Analysis in Bank Direct Marketing Prediction: https://thesai.org/Downloads/Volume10No7/Paper_85-Visualization_and_Analysis_in_Bank_Direct_Marketing.pdf

Bank Marketing Dataset. (n.d.). Retrieved from UCI Machine learning Repository: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Bertram, M. (n.d.). *How Big Data Impacts The Finance And Banking Industries* . Retrieved from Smart Data Collective: https://www.smartdatacollective.com/how-big-data-impacts-finance-and-banking-industries/#:~:text=Improvement%20in%20risk%20management%20operations,organizations%20with%20better%20risk%20coverage

Ejaz, S. (2016, May). *Predicting Demographic and Financial Attributes in a Bank Marketing Dataset*. Retrieved from Predicting Demographic and Financial Attributes in a Bank