PROJECT REPORT

BIG DATA ANALYTICS

Fall 2022

Factors affecting the subscription of term deposit after being contacted by marketing campaign

Group :

Mirza Ahsan Baig

Kashan Ahmed Khan

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/6171452303199300/3641816614758093/8796473721586237/latest.html

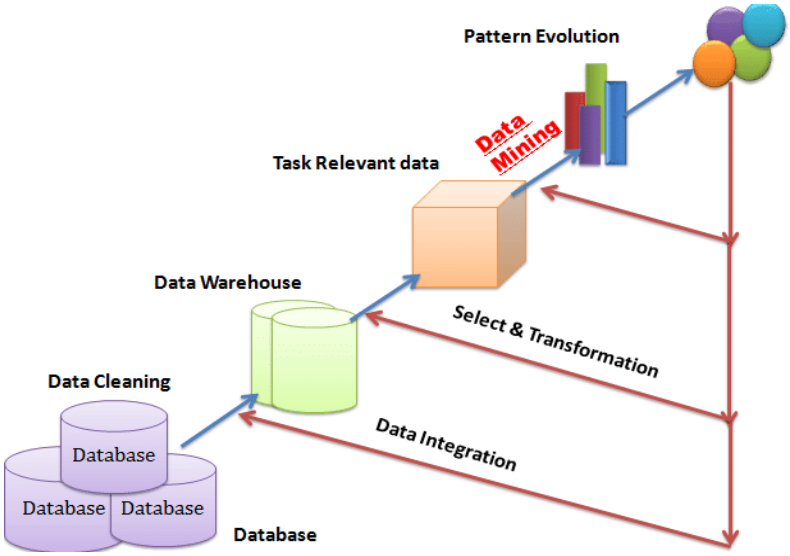


Fig 1 : Steps of knowledge discovery in database

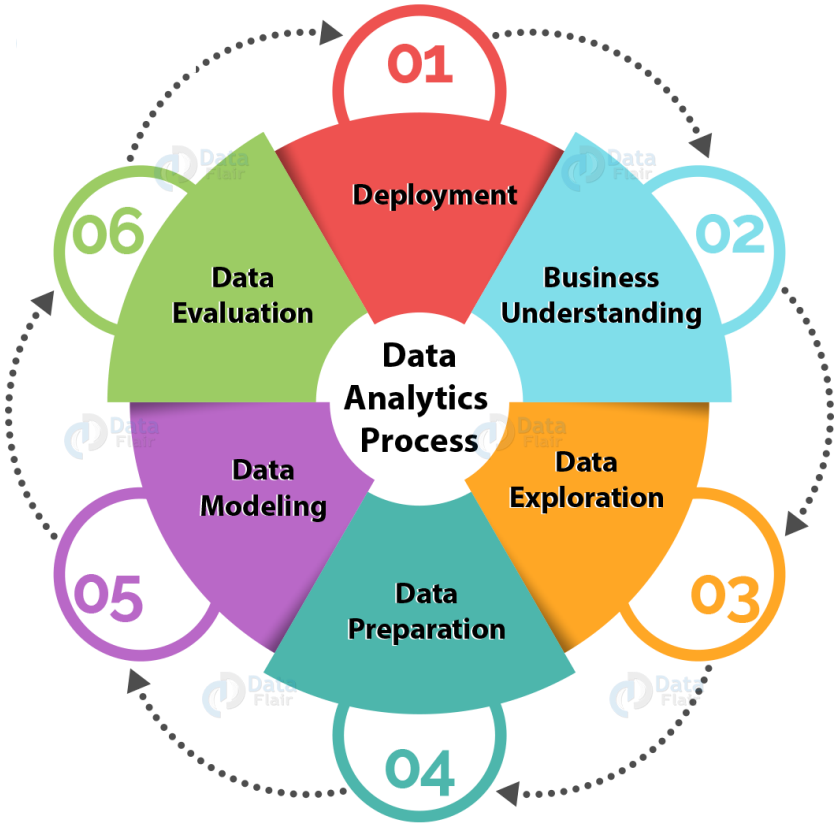


Fig 2 : Data Analytics Process

TOOLS AND FRAME WORK :

This project has been done on "Databricks". It is an American enterprise software company founded by the creators of Apace Spark. Databricks develops a web-based platform for working with Spark, that provide automatic cluster management and Python-style notebooks. It provides a unified, open platform for datasets and empowers data scientists, data engineers and data analyst with a simple collaborative environment to run interactive and scheduled data analysis workloads.

LIBRARIES :

Apache Spark

Apache Spark is an open-source, distributed computing system used for big data processing and analysis. It offers a unified, high-performance, and easy-to-use platform for data engineers, data scientists, and business analysts.

Numpy

NumPy is a powerful open-source library for the Python programming language, providing support for large, multi-dimensional arrays and matrices of numerical data, as well as a large collection of mathematical functions to operate on these arrays. It is a fundamental package for scientific computing with Python.

PANDAS

Pandas is a powerful open-source data manipulation and data analysis library for the Python programming language. It provides easy-to-use data structures and data analysis tools for handling and manipulating numerical tables and time series data.

MATPLOTLIB

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. It allows to easily create static, animated, and interactive visualizations in Python.

SEABORN

Seaborn is a data visualization library for the Python programming language, based on Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics. It also allows to easily visualize complex datasets and relationships between multiple variables, making it a powerful tool for exploratory data analysis.

ANALYTICS FLOW FOR CAMPIGN SUBSCRIPTION DATASET

1. DATA COLLECTION :

The data set of campaign subscription was provided by the course supervisor Dr. Saad Ahmed.

2. DATA PRE-PROCESSING :

The initial data set had 4521 records with 17 columns. The names of columns from left to right are as follows :

age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome and y.

The few first records are presented below :

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays
1	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1
2	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339
3	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330
4	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1
5	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1
6	35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176
7	36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330

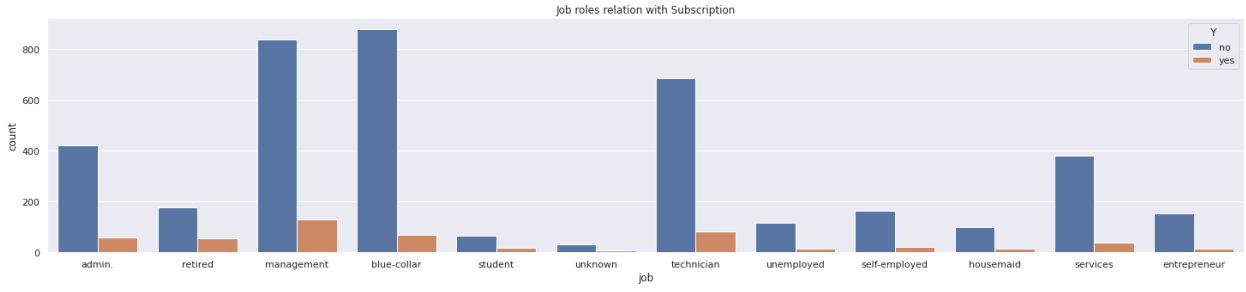
The data set was already cleaned and didn't had any NaN/missing or garbage values.

3. DATA EXPLORATION :

Data exploration is done to deduce useful knowledge from the given data set. For this purpose, we used Spark SQL to the data set to answer the following questions.

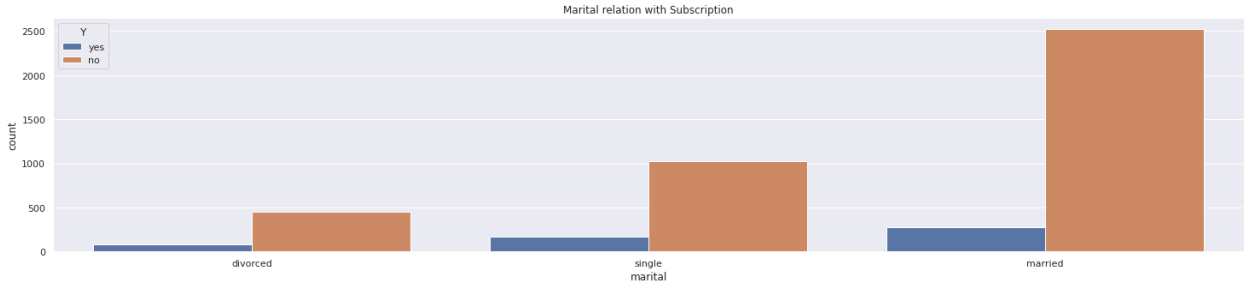
S. No	Research questions answered through dataset exploration
1.	What is the impact of the job nature on the subscription of campaign ?
2.	What is the impact of the marital status on the subscription of campaign ?
3.	What is the impact of the education on the subscription of campaign ?
4.	What is the impact of the housing loan on the subscription of campaign ?
5.	What is the impact of the personal loan on the subscription of campaign ?
6.	What is the impact of age on subscription of campaign ?

1. Impact of the job nature on the subscription of campaign :



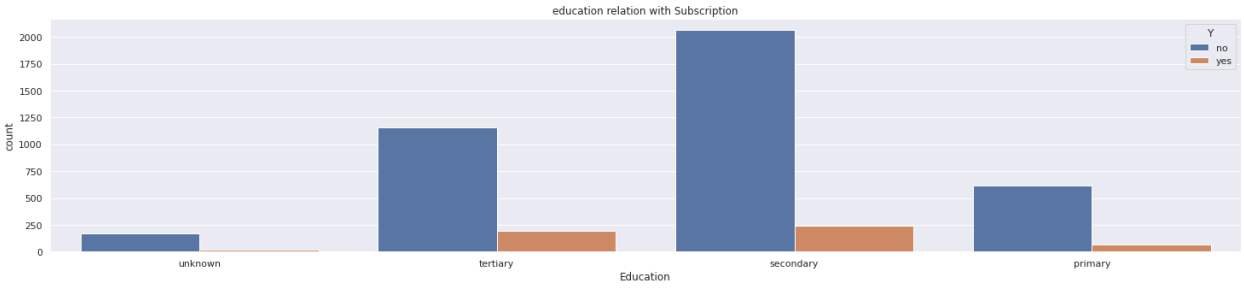
It is observed from the plotted graph that Customers having Jobs in Management, Technicians, Admin are most likely to subscribe.

2. Impact of the marital status on the subscription of campaign:



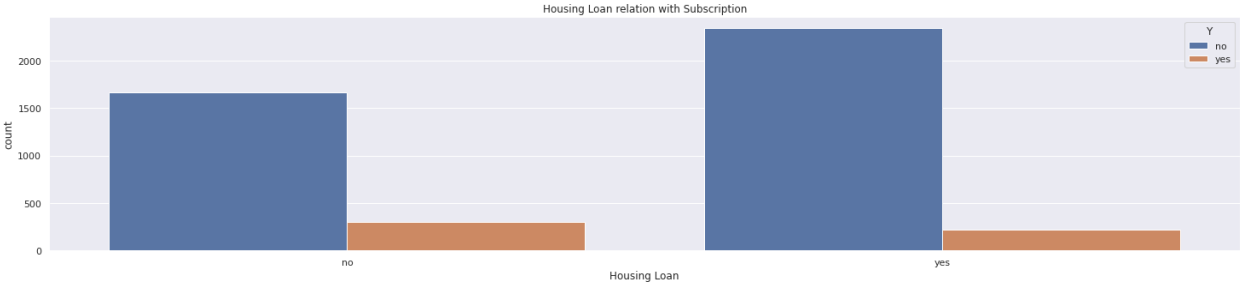
It is observed from the above plotted graph that Customers who are married usually likely to declined subscription.

3. Impact of the education on the subscription of campaign :



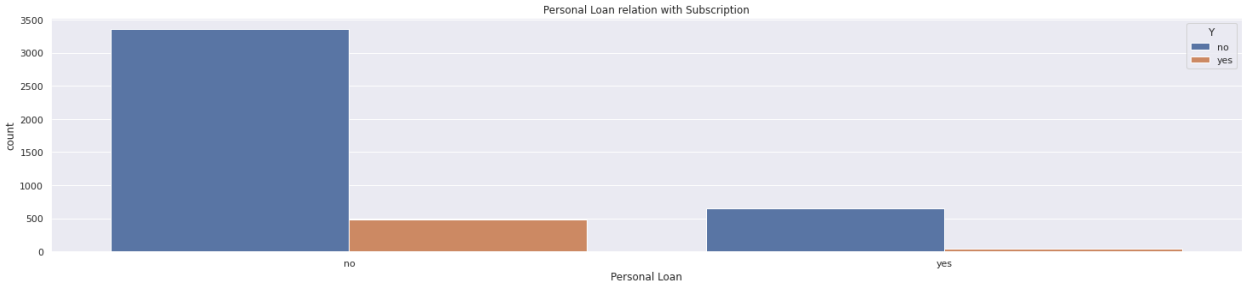
It is observed from the above plotted graph that usually costumers having secondary or tertiary education has higher declining rate.

4. Impact of the housing loan on the subscription of campaign:



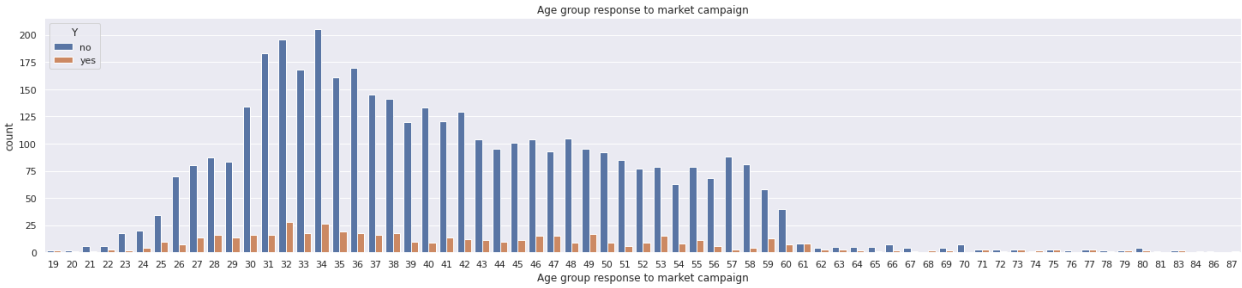
It is observed from the above plotted graph that Customers having housing loan usually declined to subscribe service.

5. Impact of the personal loan on the subscription of campaign:



It is observed from the above plotted graph that Customers don't having Personal loan usually likely to subscribe service.

6. Impact of age on campaign subscription :



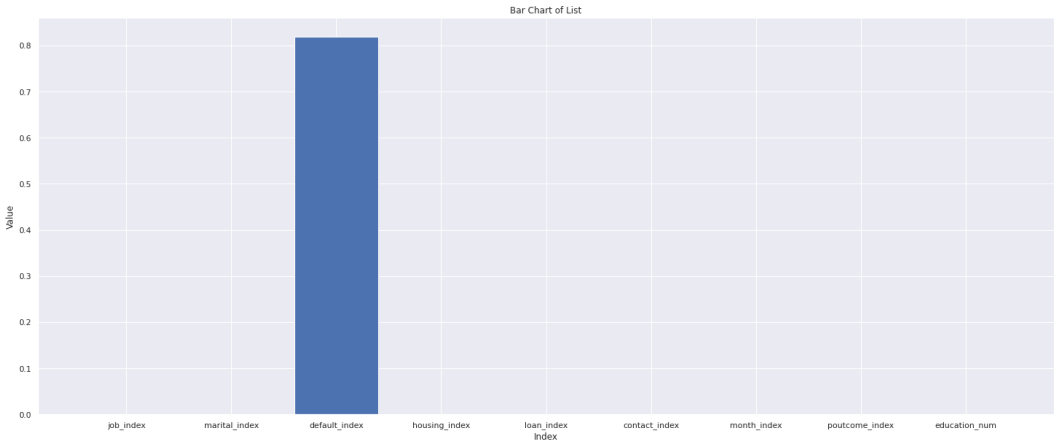
Age group between 31 to 36 are most likely to declined to market campaign.

4. DATA MODELLING :

Before applying machine learning algorithms, we're going to select most important features.

Categorical Feature Selection :

For the selection of categorical features, we first encoded ordinal and nominal features and then we applied Chi-Square test.

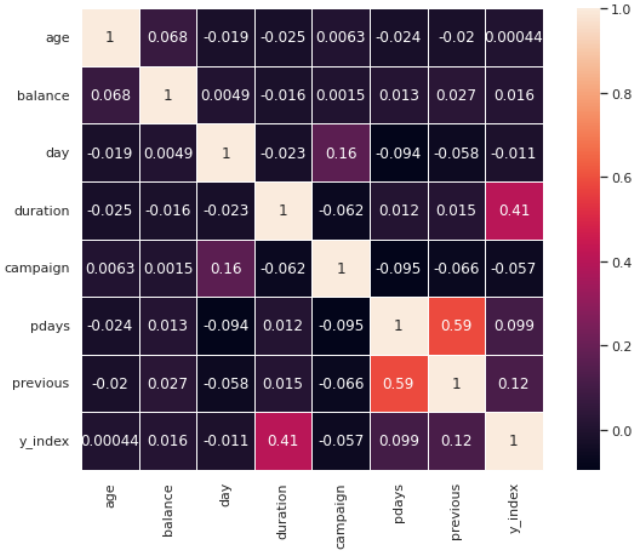


The test showed that default had the high P-value that is why it was eliminated. The final categorical features were :

Job, Marital status, housing, loan, contact, month, poutcome and education.

Numeric Features Selection :

For numeric features selection, we applied pearson's co-relation. Following are the results.



Based on the results, we dropped three features, namely, previous, duration and campaign. The final numeric features were :

age, balance, day, pdays and y.

APPLICATION OF ML ALGORITHMS

A) K-MEANS CLUSTERING :

K-means clustering is a unsupervised machine learning algorithm, which means that it doesn't need labeled input data. But for understanding, we applied K-Means on our dataset. Following results were obtained.

```
only showing top 20 rows

+----------+-----+
|prediction|count|
+----------+-----+
|         1|  257|
|         0| 4186|
+----------+-----+

+-------+-----+
|y_index|count|
+-------+-----+
|    0.0| 3953|
|    1.0|  490|
+-------+-----+
```

B) LOGISTIC REGRESSION :

The dataset was split into 70% training and 30% testing. Then the model was fitted on the data.

```
train, test = LR_df.randomSplit([0.7, 0.3], seed = 2018)
print("Training Dataset Count: " + str(train.count()))
print("Test Dataset Count: " + str(test.count()))
```

```
Training Dataset Count: 3151
Test Dataset Count: 1292
```

Then, we evaluated our model using evaluation metrics. Following were the results,

```
Test Area Under ROC 0.6901957920234664
Accuracy : 0.8955108359133127
Precision : 0.7222222222222222
Recall : 0.09090909090909091
True Neg Rate : 0.9956483899042646
F-Score : 0.16149068322981366
```

C) GRADIENT-BOOSTED TREE CLASSIFIER :

We applied gradient-boosted tree classifier on our dataset. The results are shown below :

```
Test Area Under ROC: 0.743309779863306
Accuracy : 0.8877708978328174
Precision : 0.4642857142857143
Recall : 0.09090909090909091
True Neg Rate : 0.9869451697127938
F-Score : 0.15204678362573099
```

CONCLUSION :

In conclusion, it is observed from the exploratory data analysis of the dataset, the people having jobs in management, administration and are technicians are mostly likely to subscribe the campaign. People who are not married, literate and educated often declines the subscription. People, having any kind of loan, either housing or personal, usually don't like to subscribe the campaign. People, lying the age bracket of 31-36 are most likely to decline the subscription.

We applied three ML algorithms on our dataset, namely, K-Means Clustering, Logistic regression and Gradient-Boosted tree classifier. It is evident from the results that the best performing and suited ML algorithm on or data is Logistic regression with 89% accuracy, whereas, gradient-boosted tree classifier has accuracy 88%.