BANK MARKETING (CAMPAIGN).

Virtual Internship
Data Science
Project Report.

Group Name: LISUM01: Data science Group 1
Members:
1. William Ogweli Okomba, willokomba@gmail.com, Kenya
2. Ece Kurnaz, eceeee.kurnaz@gmail.com, Turkey
3. Collin Mburugu, colinmburugu@gmail.com, Kenya
4. Udbhav Balaji, udbhavbalaji@gmail.com, India

Name: Bank Marketing(Campaign)
Report date: 15/08/2021
Internship Batch: LISUM01
Version:1.0
Data intake by: William Ogweli Okomba
Data intake reviewer: Intern who viewed the report
Data storage location:

INTRODUCTION.

ABC bank (a Portuguese banking institution) has a term deposit product that is desired to be sold to clients. We will focus on customer's past interactions with the bank or other financial institutions to have a better understanding on whether these particular clients will buy this product or not. Developing a model with using machine learning for this aim is reasonable. With performing this project, our aim is to save resources and time for ABC bank.

Business Objective.

The main objective of this project is;

- To create a bank term deposit model to predict whether a customer will accept the product or not based on the historical data in the given dataset. Select one or several suitable learning algorithms and a suitable metric for assessing quality model.
- To be able to identify relationships between products purchased and customer behaviour.
- Come up with insights that help with marketing strategies.

Assessing the Data.

1. Resource Inventory.
   - Datasets

     We were provided by the dataset

     Dataset link: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
   - Software(Python, Jupytor)
   - Personnel - Team members

2. Assumptions.
   - ➢ The available dataset was complete and no data losses.
   - ➢ All the information needed for the study was captured in the dataset.

3. Constraints.
   - ➢ There are no constraints on working on the dataset.

Data Mining Goals.

1. To determine the relationships between product purchase and previous customer behaviour.
2. Come up with insights that help with marketing strategies.
3. To identify features that determine customer chance of buying the product.

Data Mining Success Criteria.

- We'll consider our project successful when we achieve an accuracy of at least 65 %, recall score of 81% or Roc/AUC score of 65% in the model.

2. Data Understanding.

Data Description.

We had 2 dataset, bank_additional_full.csv (bank_df) and it's sample bank_additional.csv (bank)

Bank campaign dataset contains 41188 rows and 21 columns;

1. age (Age of the Customer) - Numerical
2. job (Type of Job) - Categorical - Possible Values - ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. marital (Marital Status) - Categorical - Possible Values - ('divorced','married','single','unknown'); NOTE: 'divorced' includes divorced and widowed
4. education (Education Level) - Categorical - Possible Values - ('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5. default (Has credit in Default) - Categorical - Possible Values - ('no','yes','unknown')
6. housing (Has Housing Loan) - Categorical - Possible Values - ('no','yes','unknown')

Related with the last contact of the current campaign:

7. loan (Has Personal Loan) - Categorical - Possible Values - ('no','yes','unknown')
8. contact (Type of Communication) - Categorical - Possible Values - ('cellular','telephone')
9. month (Month of Last Contact) - Categorical - Possible Values - ('jan', 'feb', 'mar', ..., 'nov', 'dec')

10. day_of_week (Day of Week of Last Contact) - Categorical - Possible Values - ('mon','tue','wed','thu','fri')

11. (Last Contact Duration in seconds) - Numerical; IMPORTANT NOTE = (If duration=0, y="No')

Other attributes:

12. campaign (Number of Contacts performed during this campaign for this client) - Numerical

13. pdays (Number of days passed after client was contacted from a previous campaign ; 999 - Not Previously Contacted)

14. previous (Number of contacts performed before this campaign and for this client) - Numerical

15. poutcome (Outcome of the previous marketing campaign) - Categorical Values - Possible Values - ('failure','nonexistent','success)

Social and economic context attributes41188

16. emp.var.rate: employment variation rate - quarterly indicator(Quarterly Indicator of Employment Variation Rate) - Numerical

17. cons_price_idx (Monthly Indicator of Consumer Price Index) - Numerical

18. cons_conf_idx (Monthly Indicator of Consumer Confidence Index) - Numerical

19. euribor3m (Daily Indicator of Euribor 3 Month Rate) – Numerical (Quarterly Indicator of Number of Employees) - Numerical

20. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21.y (Target Feature - Has the client subscribed to a term deposit) - Binary - Possible Values - ('yes','no'

Verifying Data Quality.

- Completeness

The job, marital status, education, default, housing, and loan variables had missing values.

We imputed with the mode and education variable was imputed by “N/A”.

- Relevance

  All the provided columns and column entries are relevant to our study.

- Uniformity

  The bank_df dataset had 12 duplicates, we dropped them to make the dataset uniform.

- Validity

  All entries are valid and accurate.

3. Data Preparation.

Data preparation procedures involve;

Libraries used ;

- Numpy
- Pandas
- Searborn
- Matplotlib and pyplot
- plotly

I. Loading the Data.

We loaded the csv file into the Jupytor environment using the function read_csv and converted it into a dataframe before working on it.

The datasets has been named; bank_df, and bank.

II. Cleaning the Data.

During data exploration, we did a few adjustments to the dataset before exploratory data analysis.

a. Checking for null values. The job, marital status, education, default, housing, and loan variables have missing values. We imputed with the mode and education variable was imputed by “N/A”.

b. Checking for duplicate values.
The dataset displayed 12 a record of duplicated entries. We dropped them.

c. Checking on outliers.
Most features in the dataset have outliers that we didn’t get rid of. The reason for this is that since we are working with the bank customer’s information which are genuine. However, we removed outlier in age variable using interquartile range.

d. Checking on dataset description.
Dataset description gives us a glimpse on the statistical analysis of the features.

- The job with the most frequent respondents was admin.
- The married people were the most respondents etc.

```
#3.6a concise summary of the numerical datatypes of bank_df dataset
bank_df.describe()
```

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

```
#3.6b summary for non numerical values
bank_df.describe(include=["O"])
```

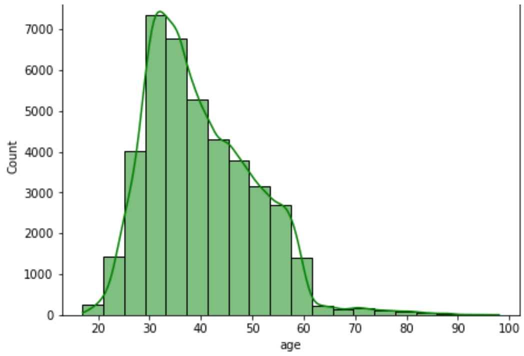
	job	marital	education	default	housing	loan	contact	month	day_of_week	poutcome	y
count	40858	41108	39457	32591	40198	40198	41188	41188	41188	41188	41188
unique	11	3	7	2	2	2	2	10	5	3	2
top	admin.	married	university.degree	no	yes	no	cellular	may	thu	nonexistent	no
freq	10422	24928	12168	32588	21576	33950	26144	13769	8623	35563	36548

EXPLORATORY DATA ANALYSIS.

Univariate Analysis.

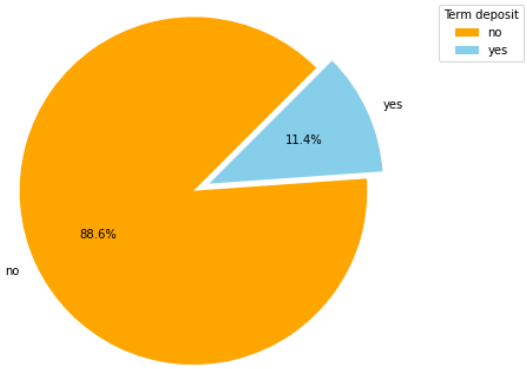
1. Age

The age variable is skewed to the right (positive skewness), this means the mean is greater than the mode



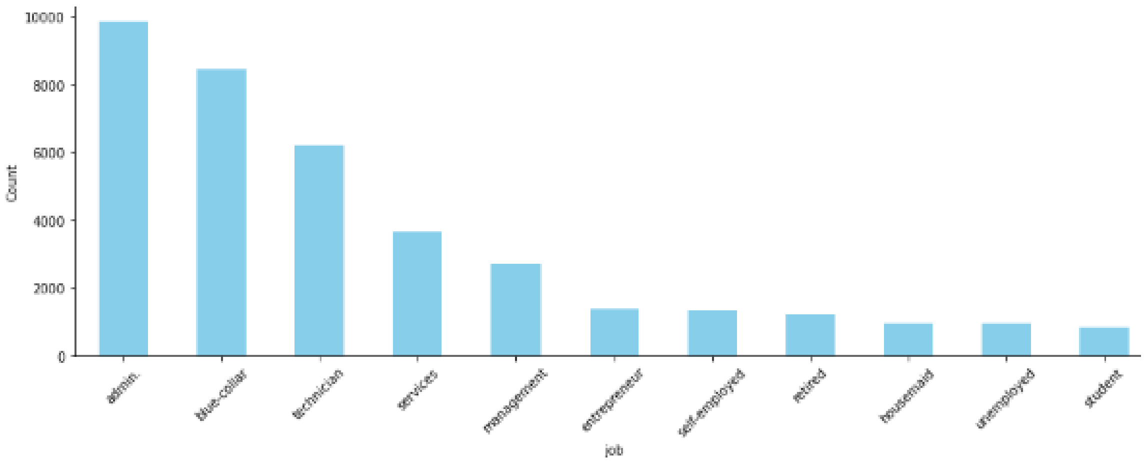
2. Term Deposit

Only 11.4% of respondents have term deposit product. This is our target variable. Class is imbalanced



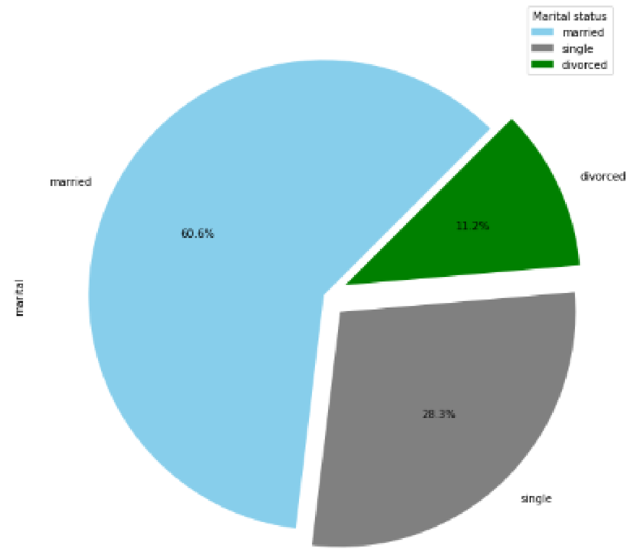
3. Job

Respondents in admin job were the highest followed by Blue-collar job.



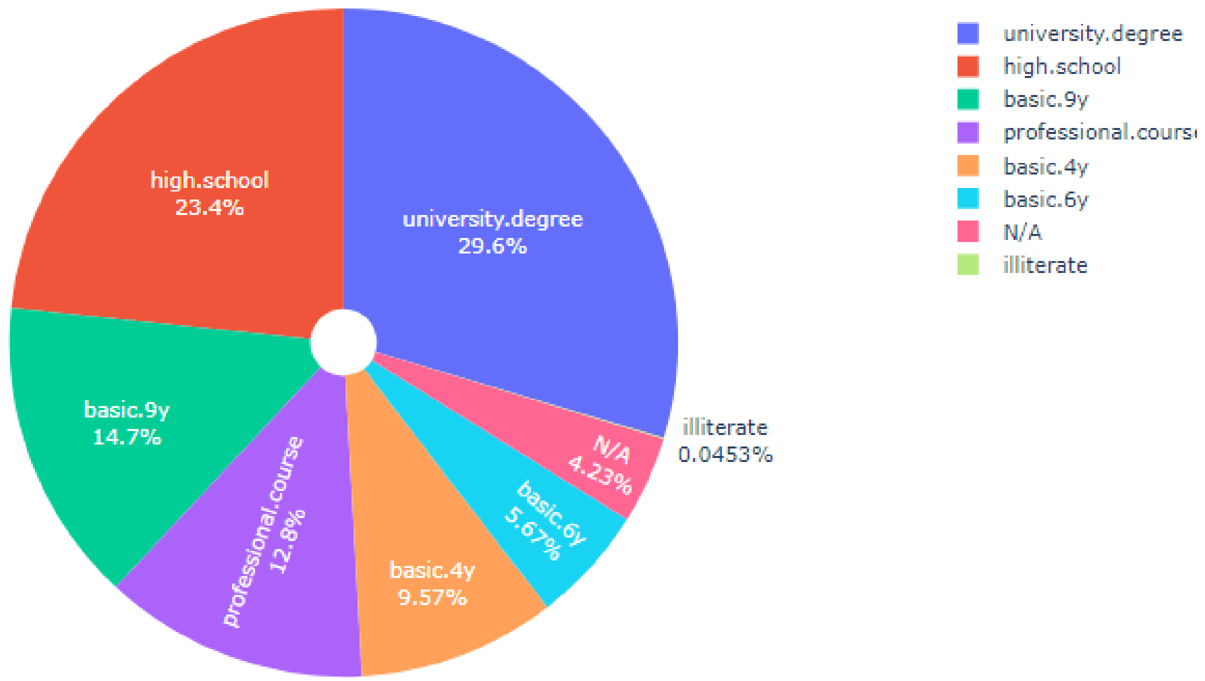
4. Marital status

The married respondents comprised of 60.6% of the total respondents, followed by singles at 28.3%



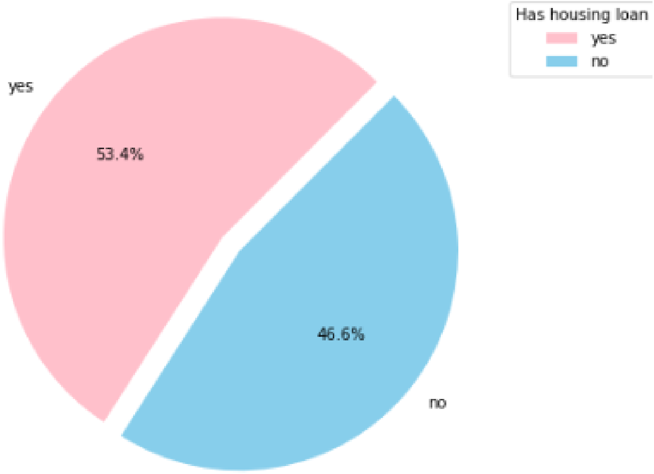
5. Education

Those with a university degree had the highest respondents at 29.6%, followed by high school.



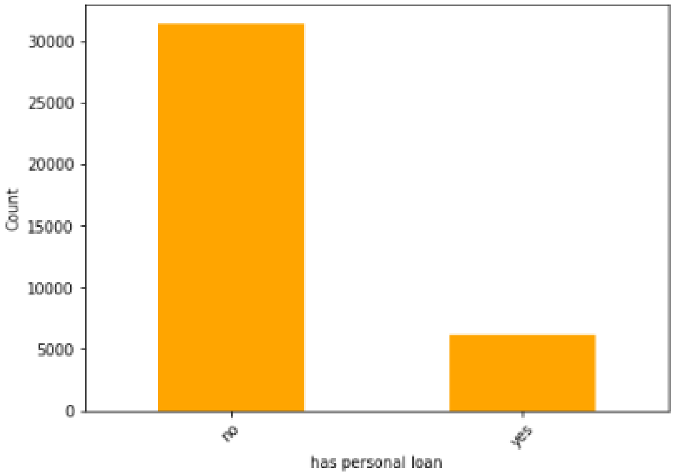
6. Housing loan

53.% of the respondents has housing loan.



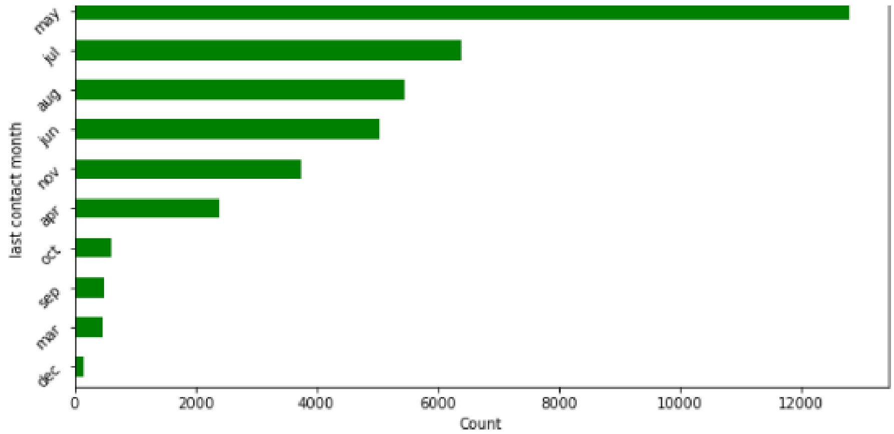
7. Personal loan

Most of the respondents did not have personal loan



8. Last contact month

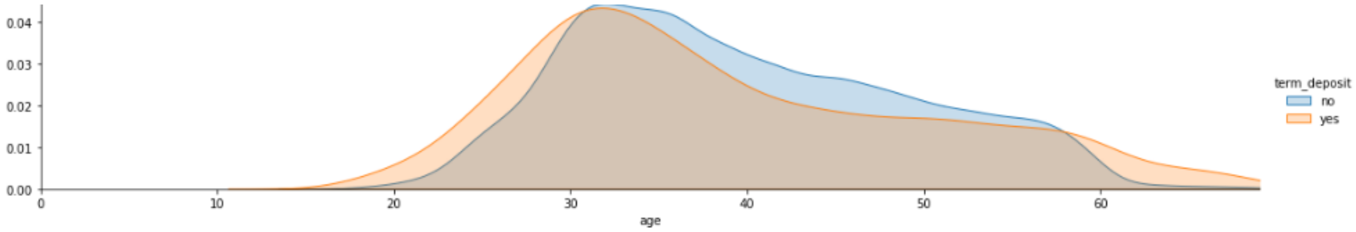
Most of the respondents were contacted in the month of May



BIVARIATE ANALYSIS.

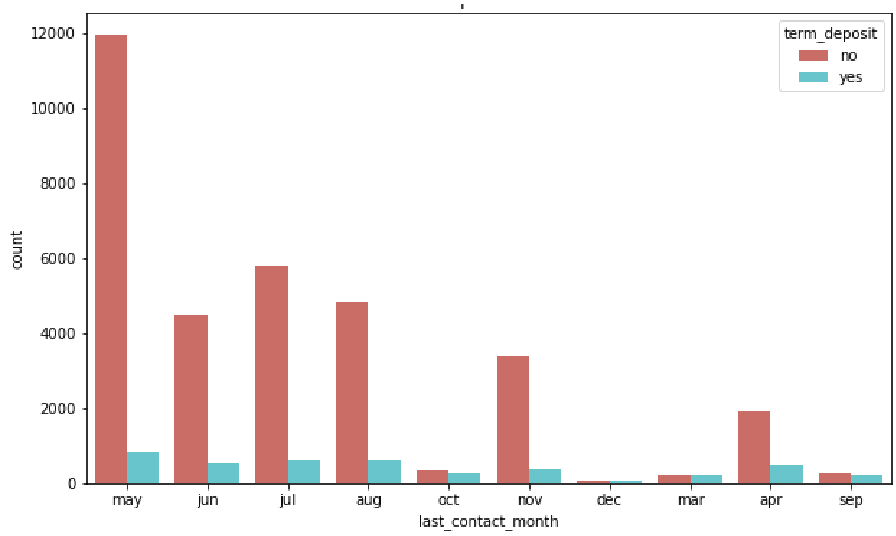
I. Age vs term deposit :

There is a significant difference in the ages of those who have and those who didn't accept the term product. Though those without are the majority.



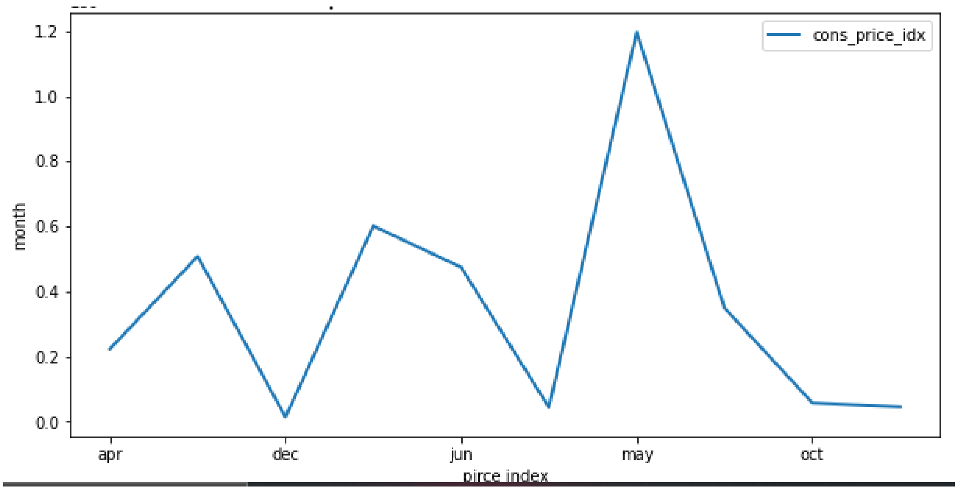
Ii. Month vs term deposit

Customers did not buy the term deposit package on May. December was the lowest in accepting term deposit as well not accepting the product



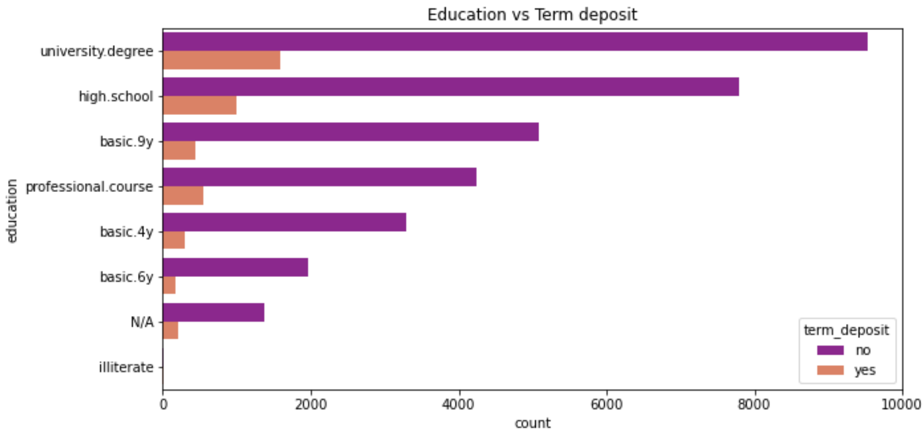
Iii. Consumer price index vs term deposit:

Consumer price index was high in May, this is attributed more customers acquiring the term deposit in the same month



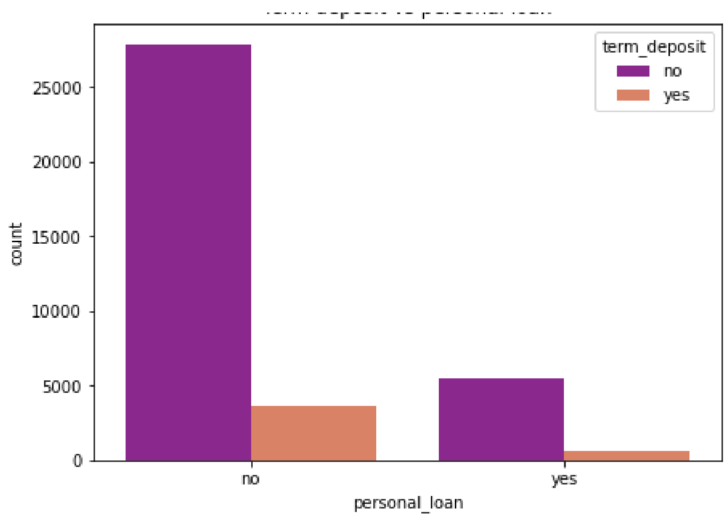
Iv. Education vs term deposit

Those with university degree got the term deposit product the most



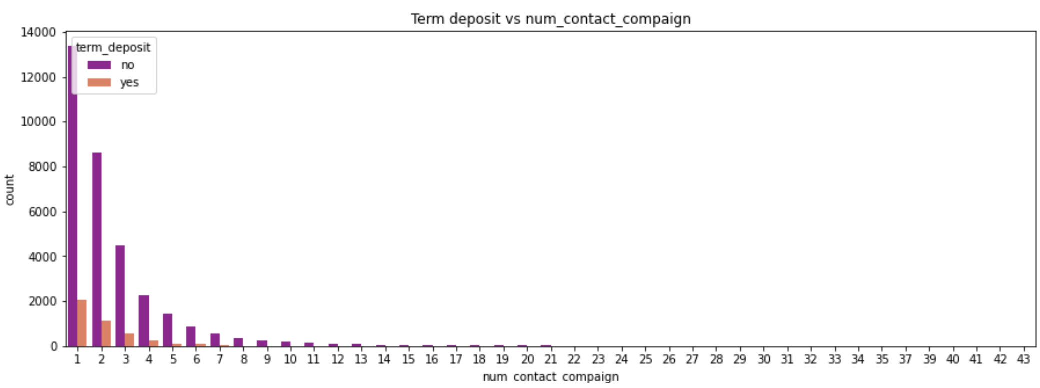
V. personal loan Vs term deposit.

Those without personal loans accepted the term deposit more



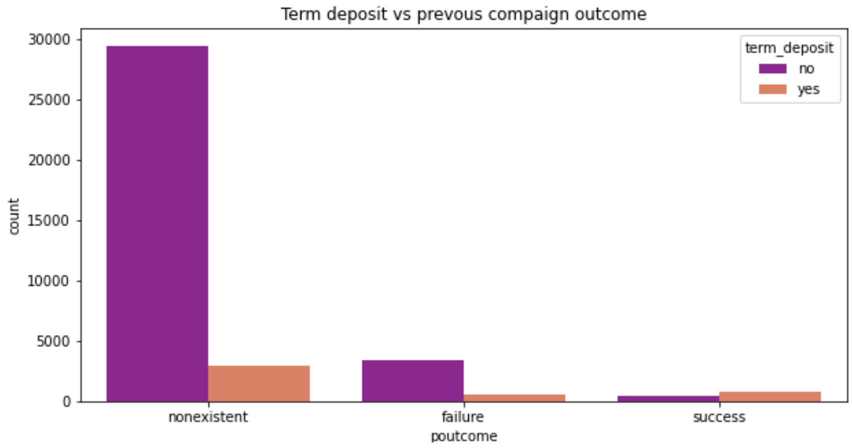
Vi. num_contact_compaign vs term deposit.

1. Most of the customers who acquired the term deposit were contacted once, meaning only one call was enough for the person to decide on whether to have the product or not.
2. The more the calls the customer received the less they were interested in the product



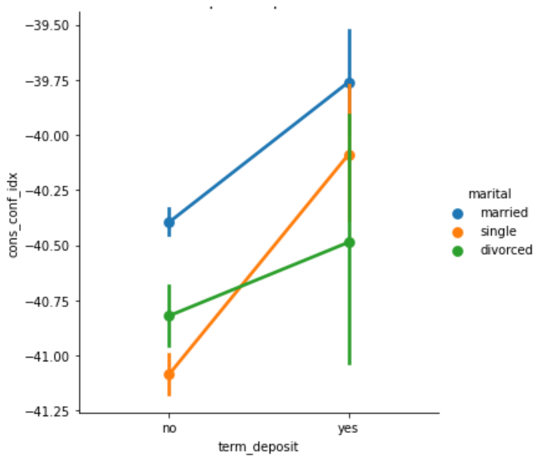
Vii. Previous campaign outcome vs term deposit

The acquisition of the term deposit product did not depend on previous campaign



Viii. Consumer confidence rate

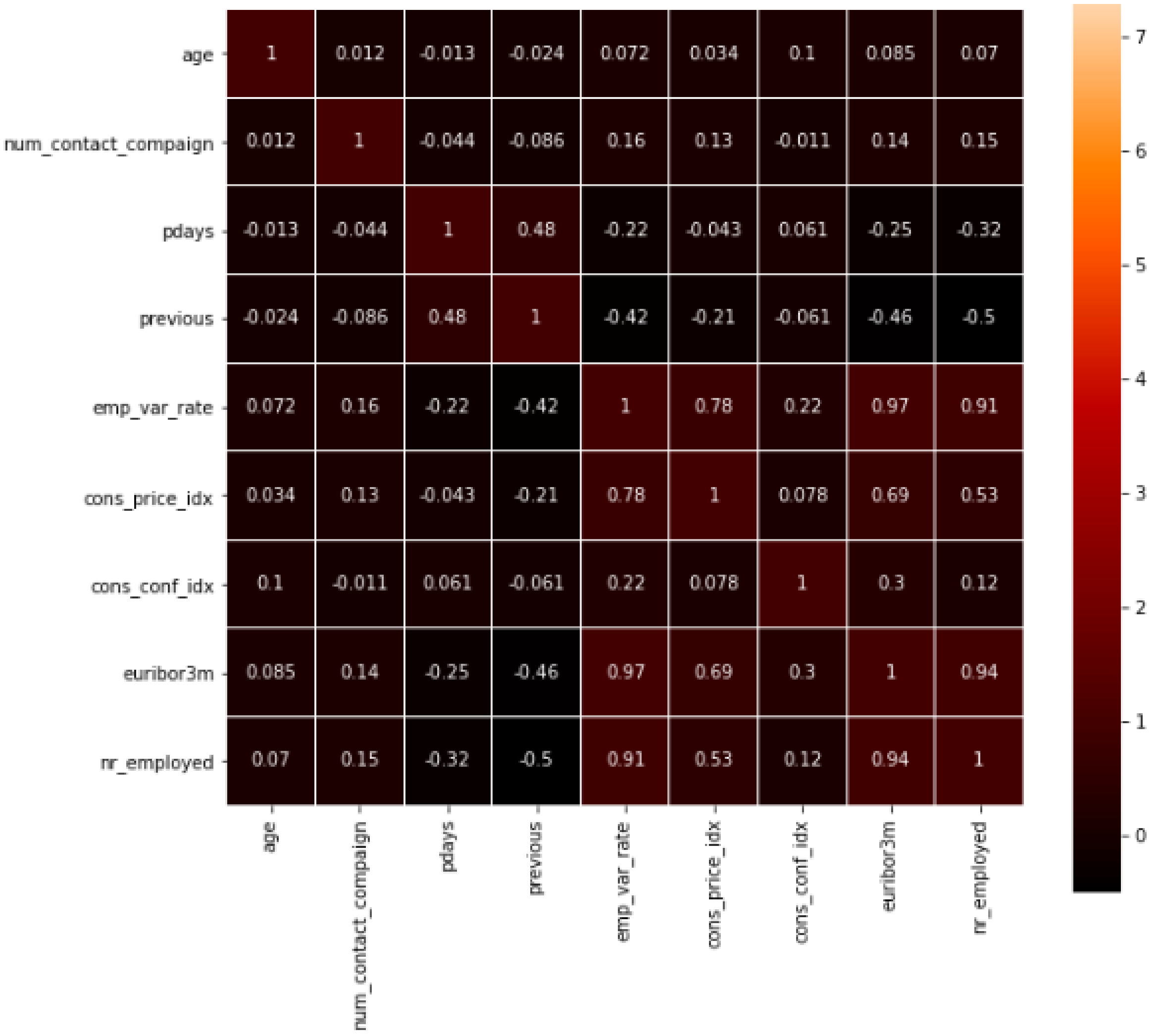
The married is very high regardless of term_depsit, and acquisition of the term high as consumer index increases



CORRELATION.

Library used;

- Corrplot.
- "emp_var_rate" & "euribor3m", and "euribor3m" & "nr_employed" variables are strongly positive correlated at 0.97 and 0.94 respectively.
- "num_contact_compaign" and "cons_conf_idx" variables are strongly negatively correlated at -0.011. This means the more the calls the customer received the less he was interested in the product.



Checking multi-collinearity

Correlation above shows the relationship between variables. The coefficient of 1 across the diagonal shows that a variable is perfectly correlated to itself.

The above will be used to compute the VIF (Variance Inflation Factor) score for each variable, by finding the inverse matrix of the correlations matrix

Multi-collinearity detected as the VIF is above 5 for some variables.

```
#computing VIF(variance inflation factor)
pd.DataFrame(np.linalg.inv(corr.values), index = corr.index, columns=corr.columns)
```

	num_contact_compaign	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed
num_contact_compaign	1.036577	0.017345	-0.333672	-0.117313	-0.060947	0.696548	-0.437227
previous	0.017345	1.345625	-0.276951	0.025912	0.036893	-0.023153	0.926819
emp_var_rate	-0.333672	-0.276951	32.963576	-7.177394	0.984353	-24.149561	-3.502694
cons_price_idx	-0.117313	0.025912	-7.177394	6.226373	2.016902	-7.021506	9.653831
cons_conf_idx	-0.060947	0.036893	0.984353	2.016902	2.599836	-9.189778	6.444607
euribor3m	0.696548	-0.023153	-24.149561	-7.021506	-9.189778	63.822345	-33.719298
nr_employed	-0.437227	0.926819	-3.502694	9.653831	6.444607	-33.719298	30.710216

Now there is No multi-collinearity detected as the VIF is between 1 and 3 and none is heading to 5 or greater than 5

```
#computing VIF(variance inflation factor)
pd.DataFrame(np.linalg.inv(corr.values), index = corr.index, columns=corr.columns)
```

	num_contact_compaign	previous	cons_price_idx	cons_conf_idx	euribor3m
num_contact_compaign	1.033790	0.033131	-0.044158	0.047794	-0.141092
previous	0.033131	1.318621	-0.305930	-0.141924	0.842390
cons_price_idx	-0.044158	-0.305930	2.034539	0.318939	-1.608404
cons_conf_idx	0.047794	-0.141924	0.318939	1.128908	-0.578125
euribor3m	-0.141092	0.842390	-1.608404	-0.578125	2.658193

Pre-processing

Changing object variable to the right categorical variables

Then hot encoding them to numerical datatypes.

IMPLEMENTATION.

1. Logistic regression, this was our baseline model.

   Libraries for implementation;

   Sklearn, LogisticRegression

   Step 1.

   - Normalizing the data.

     The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

   Step 2.

   Train the model.

   Predict and check confusion matrix

   Tune the model

Note: other model followed the same procedure

Outcomes

model	**Recall score**	**ROC/AUC SCORE**	**Accuracy**
Logistic regression	**0.77**	**0.61**	**0.61**
Decision tree	**0.80**	**0.57**	**0.57**
Random Forest	**0.85**	**0.58**	**0.65**
Gradient Boost	**0.86**	**0.68**	**0.67**
ANN			**0.78**

Conclusion

The Recall score for the gradient boost model was higher than other models, even it was the best at predicting true negatives.
Overall, we can still retain to use the gradient boast classifier, as it still outperformed other models even without any optimisation done.
Therefore the best model for this problem is Gradient boost or use artificial neural network.

RECOMMENDATIONS.

- Customers whose education level is at the university level should be targeted mostly. Their chances of buying the product is higher than other customers with a lower education level.
- More customers should be contacted in the months of May, June, July, August but mostly May. One is likely to buy the product during this month.
- Most customers who were contacted once accepted the product compared to those who were contacted more than once. So just a single contact is better than multiple contacts.
- They should also target customers without personal loans with the bank. They are more likely to buy the product.