19MAM67 Big Data Analytics CAT 2 - Term Deposit

Successful Marketing Campaign

1934013 Jeevasurya V,
1934018 Kishore Kumar R

Abstract

The following project focuses on analyzing a dataset 'Bank Marketing', which contains 11000+ entries about customers and aims to get valuable insights from it and predict if a customer will accept a deposit offer or not, and if so, the best possible month to approach.

About Dataset

age (numeric)
job type of job (categorical:'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
marital marital status (categorical: 'divorced','married','single','unknown')
education (categorical: primary, secondary, tertiary and unknown)
default has credit in default? (categorical: 'no','yes','unknown')
housing has housing loan? (categorical: 'no','yes','unknown')
loan has personal loan? (categorical: 'no','yes','unknown')
balance Balance of the individual.
contact contact communication type (categorical: 'cellular','telephone')
month last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
day last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
duration last contact duration, in seconds (numeric).
campaign number of contacts performed during this campaign and for this client (numeric, includes last contact)
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)
previous number of contacts performed before this campaign and for this client (numeric)
poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
deposit has the client subscribed a term deposit? (binary: 'yes','no')

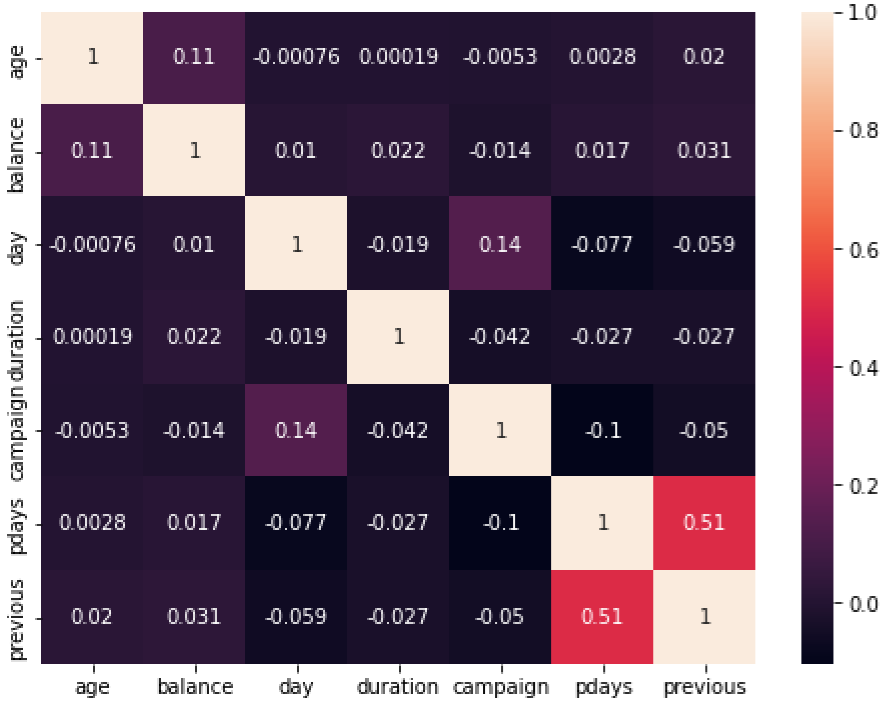
Pre Visualization

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
2	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
3	55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes
4	54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes

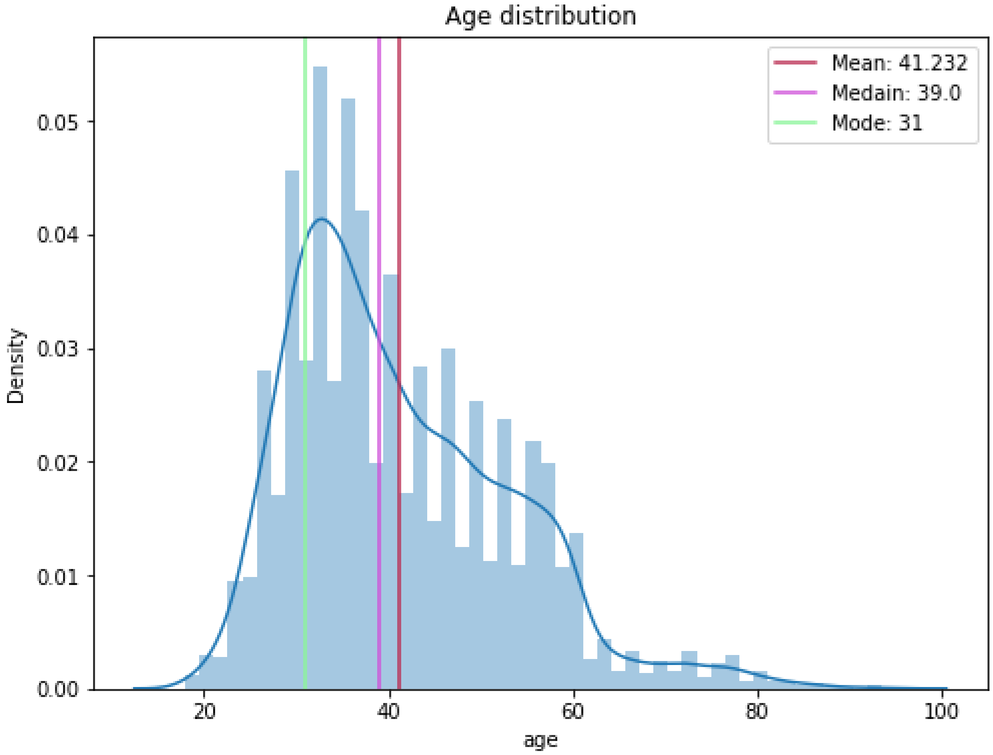
Balance Check:

	deposit	count
0	no	5873
1	yes	5289

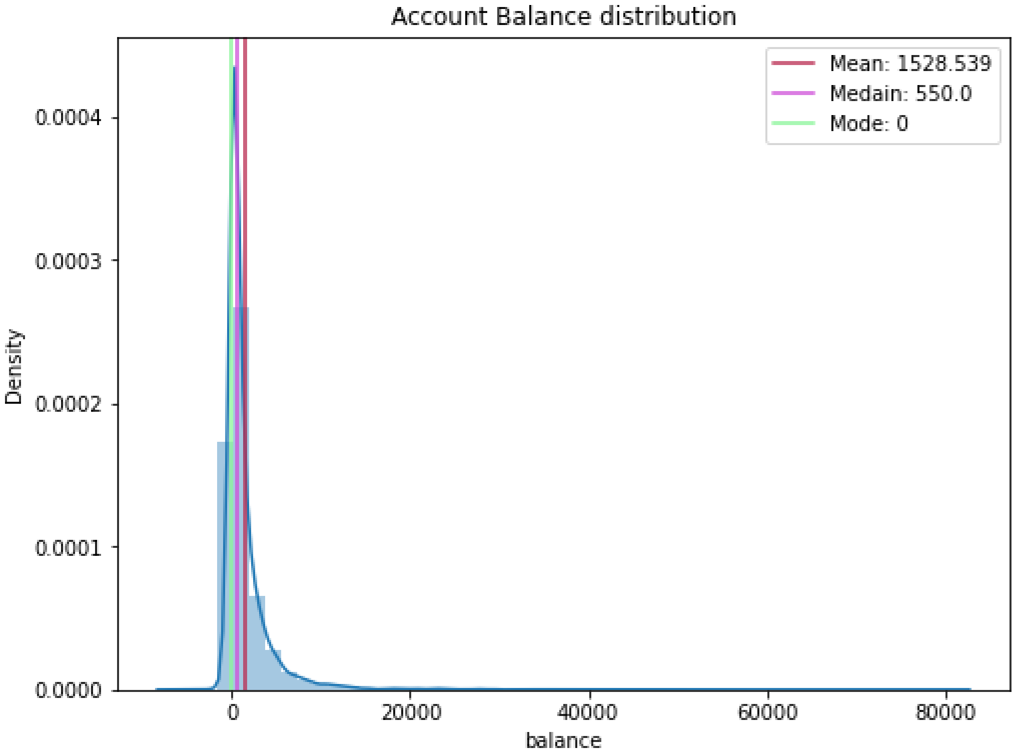
Correlation Heatmap



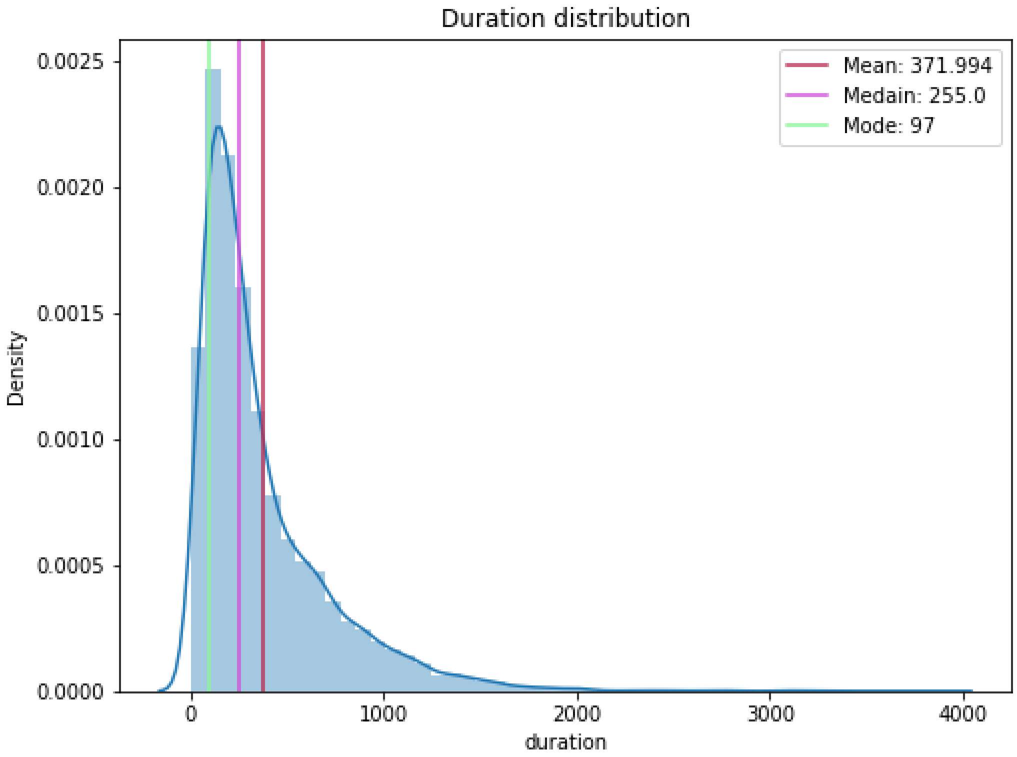
- There is no significant inter-correlation among the variables, other than pdays and previous.
- That is expected, as it is mentioned that pdays = -1 means that the customer has not been contacted yet and the previous will be 0.



- Age is Skewed to the right. The mean and median age is close but the mode is low.
- People the age of 20 and above are contacted. People at the age of 32 are contacted more frequently. People above the age of 60 are not contacted as frequently as other young customers.

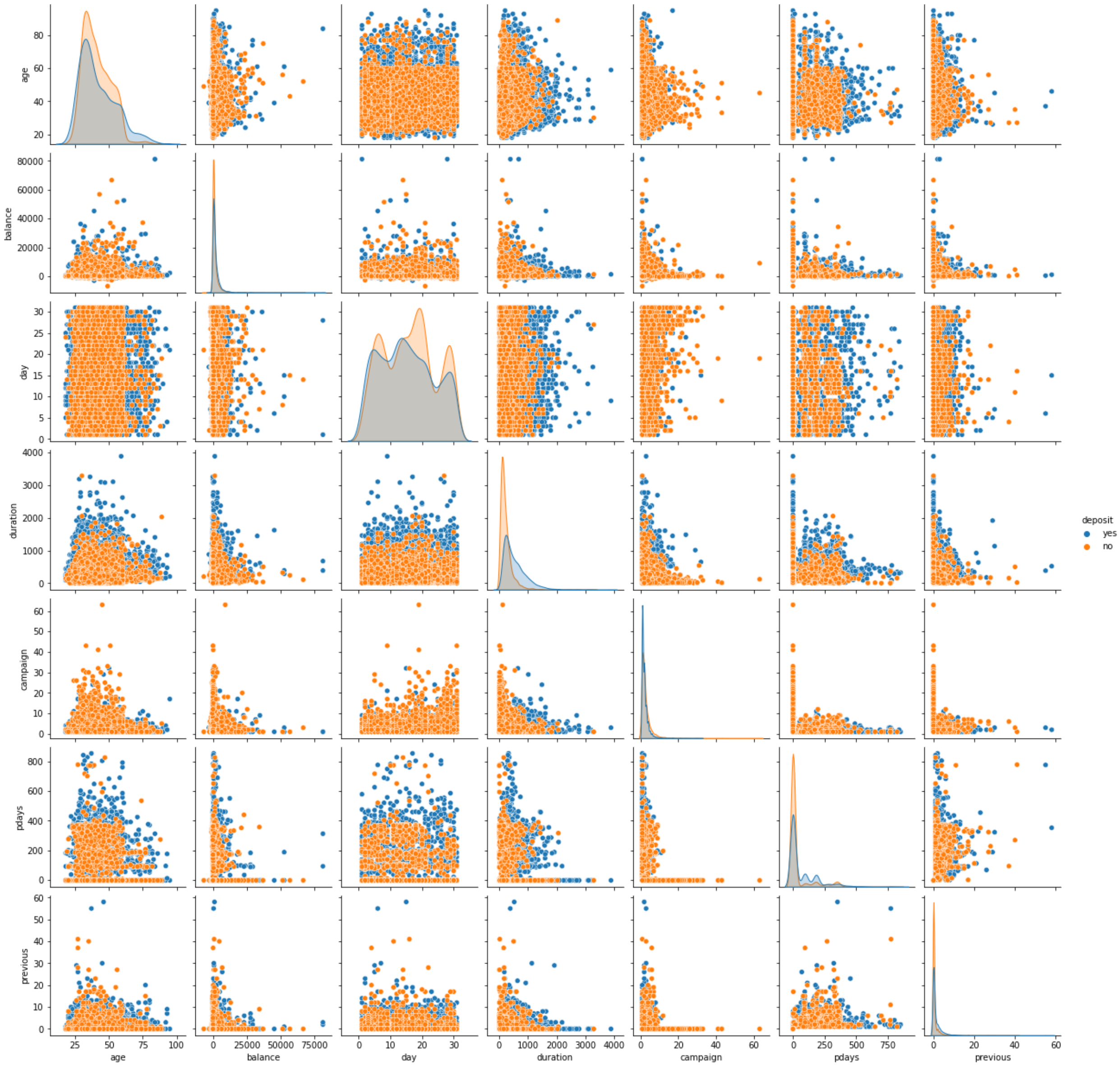


Balance is highly right-skewed. The mean balance is 1528. There are few people with high balance amounts, this is why the right tail goes around 80000 and few people are having a negative balance. Most people have 0 balance.

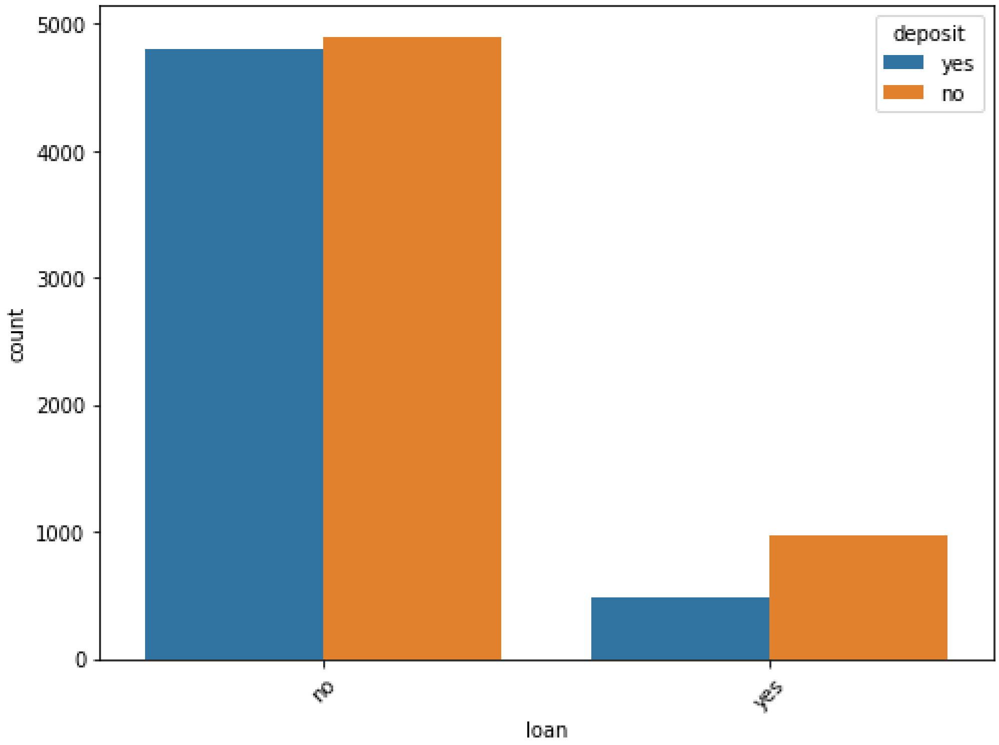


Most call durations are 97 seconds. Suggesting most people end the call within the first 2 minutes of the campaign. The average time a customer is engaged in 6 minutes.

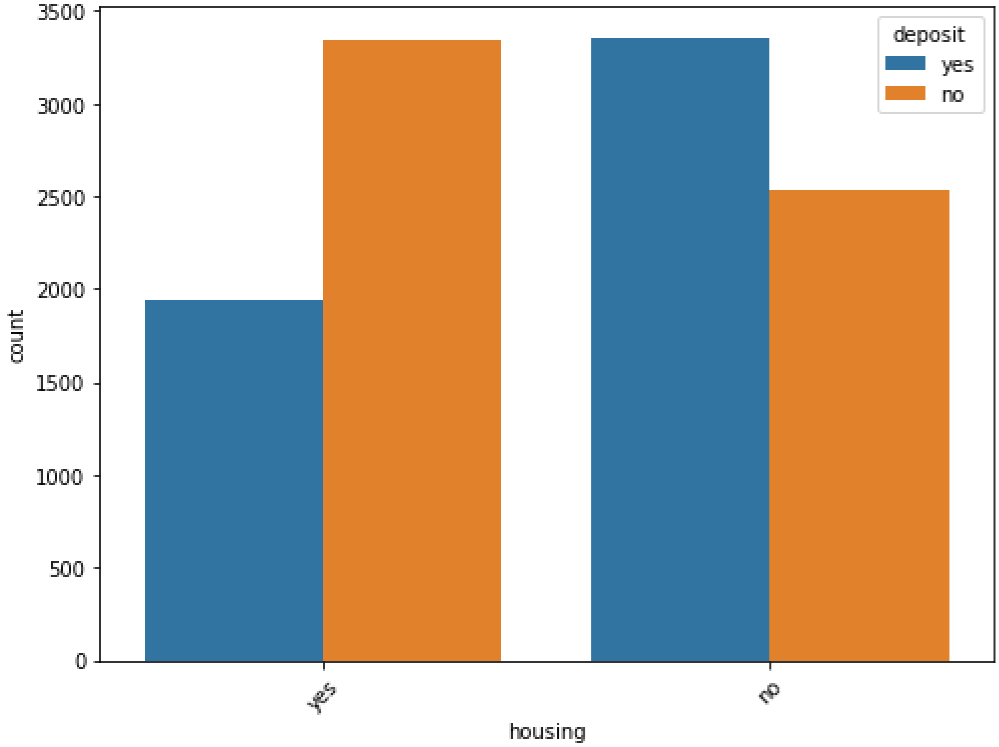
New Customers contacted: 8324
New Customers percentage: 74.574 %



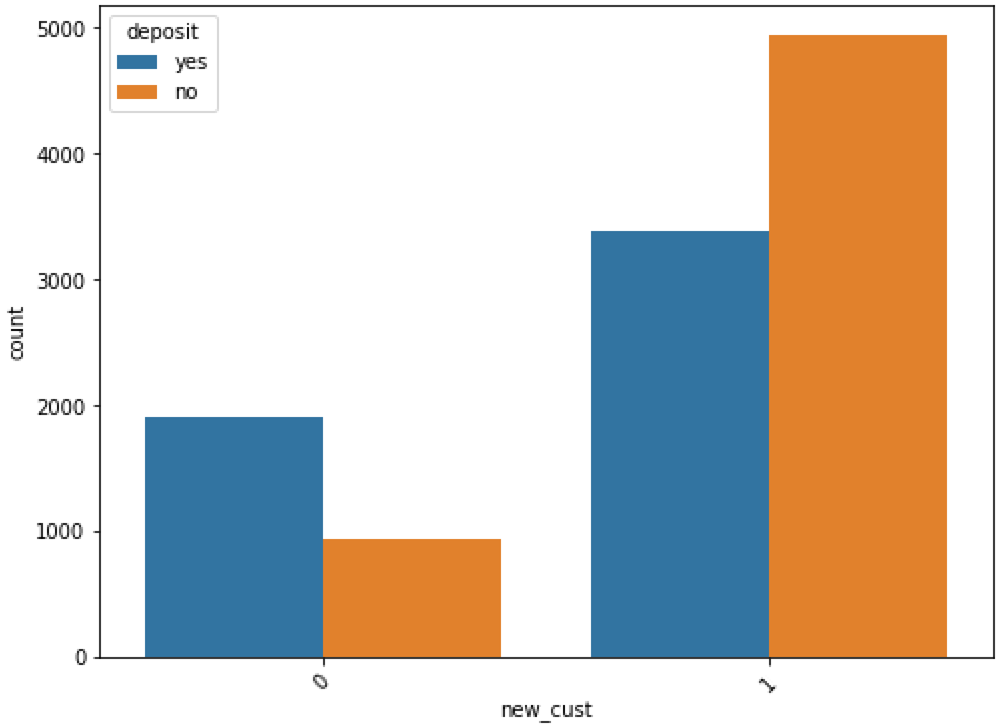
- Most of the plots in the pair plots are overlapping.
- That is both yes and no responses have similar distribution among continuous variables.
- Few Yes classes have higher values in some of the pairs.
- It is difficult to draw inferences from this pair plot



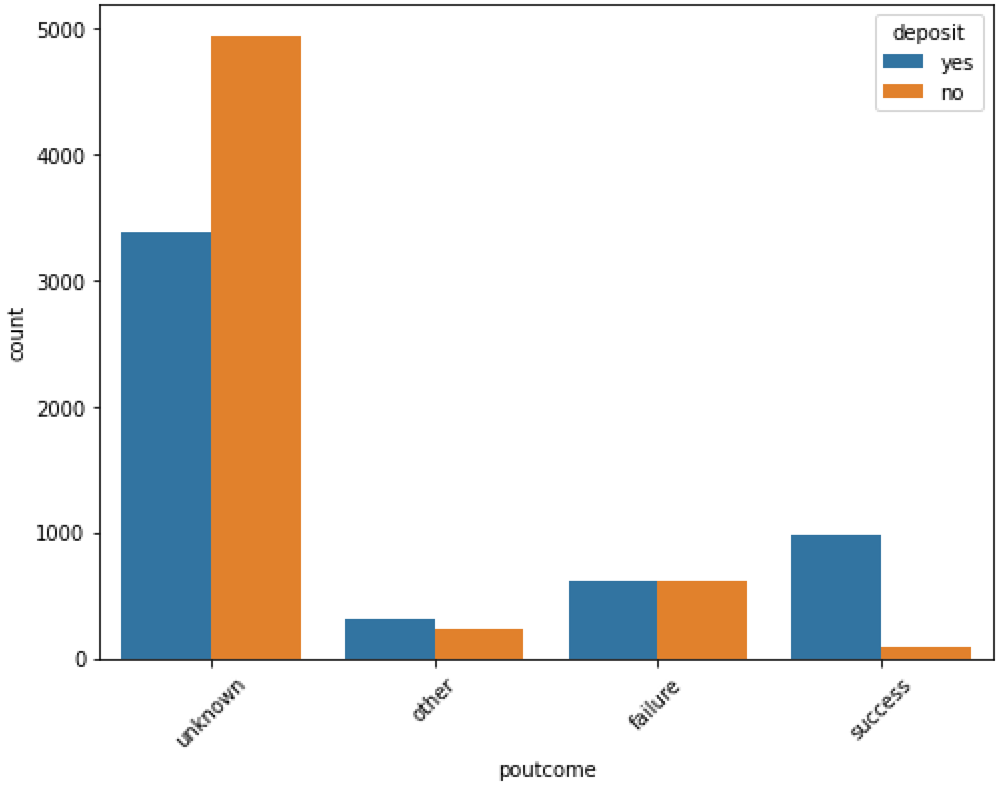
Customers with personal loans tend to reject the deposit more.



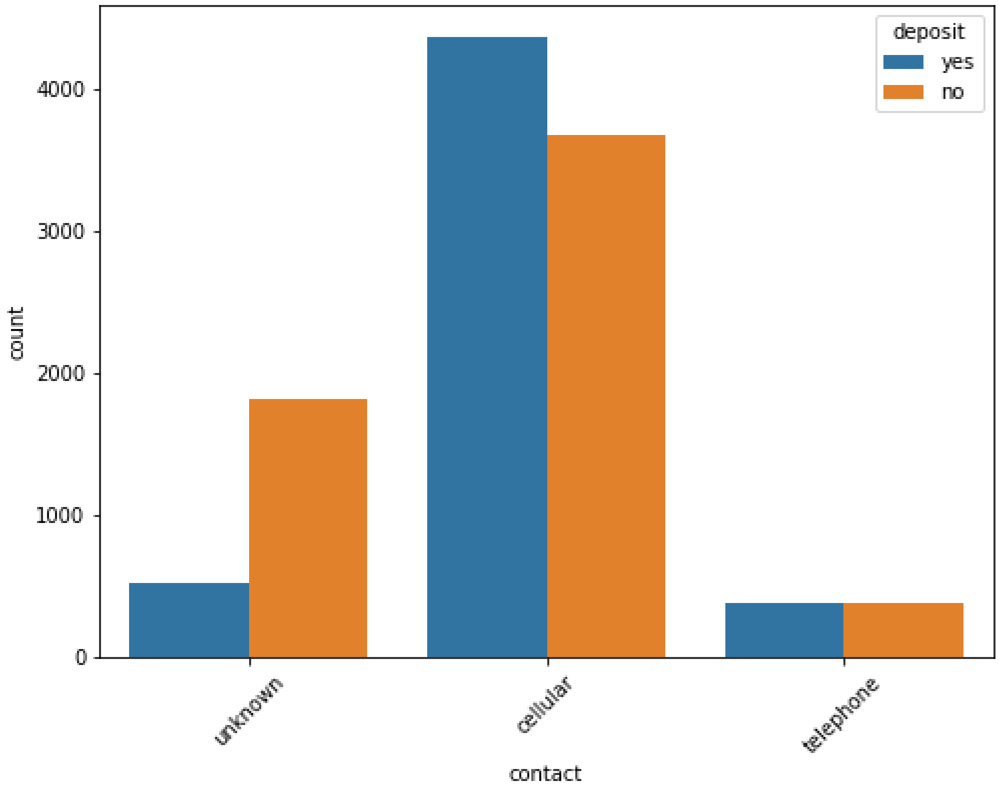
The same is the case with Housing loans. People with housing loan tends to reject the deposit.



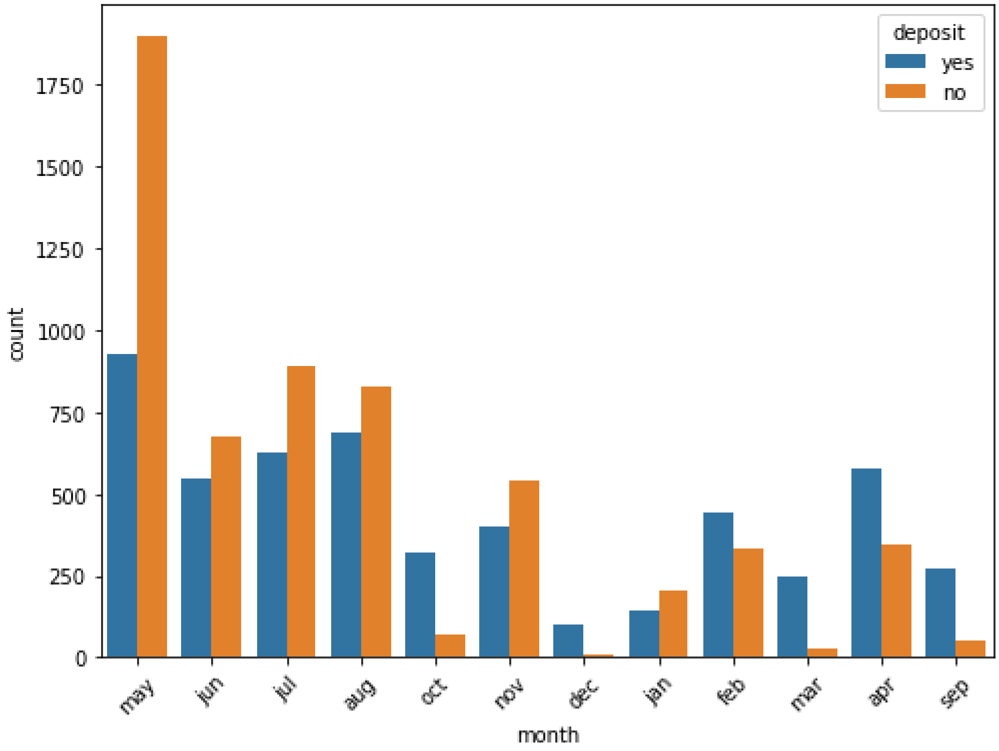
New customers are more rejective. Customers who were previously contacted seem to again accept the deposit. Also, people who deposited during the last contact tend to again deposit for the term.



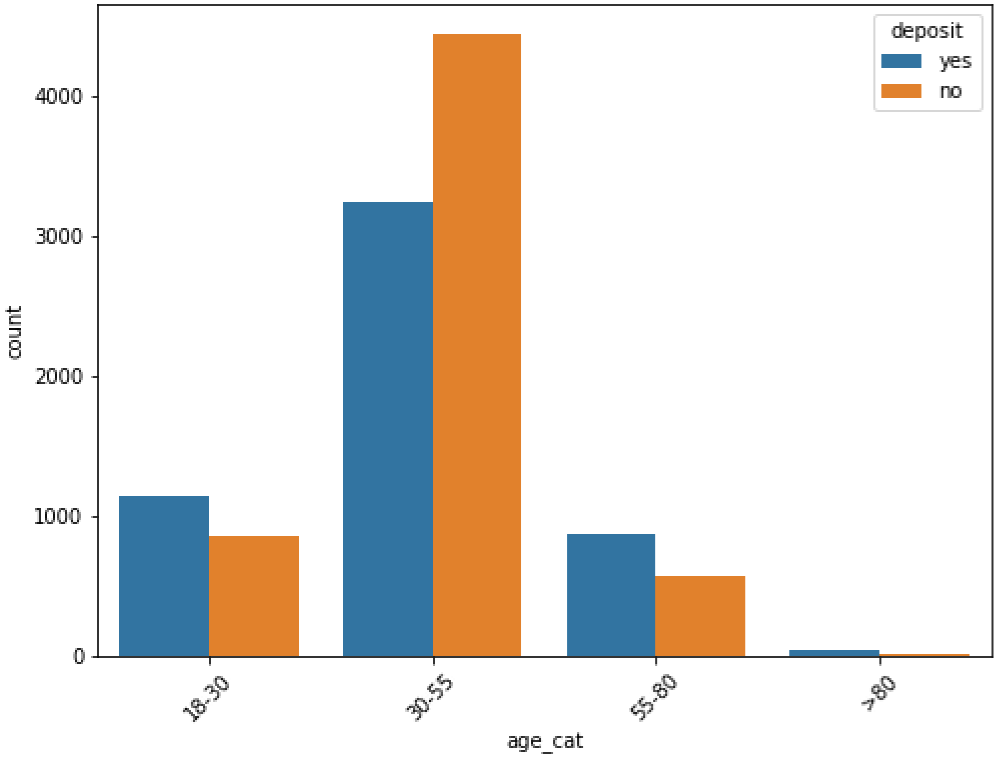
People who said ‘no’ have also made a deposit.



Customers who were contacted via cellular are more accepting.



People accept the term deposit more in the months of February, March, April, September, October, and December



Most people in their middle age are rejecting deposits. Young people and old(retired) people are more accepting of the deposit.

Libraries Used

SparkSession - To build a new session,
StringIndexer - Indexing Categorical Value,
OneHotEncoderEstimator - OneHot Encoding,
VectorAssembler - Combine the features into one vector column,
Pipeline - Pipelining the stages to specify the ML workflow,
LogisticRegression,
RandomForestClassifier,
BinaryClassificationEvaluator - Calculate Receiver Operating Characteristic Curve, and Area Under the Curve

Preprocessing

Indexed all categorical columns using Pyspark module *StringIndexer.*
Converted the indexed categories to one-hot variables.
StringIndexer was applied again to generate label Indices.
Vector assembler is used to combine the features into one vector column.

Model Training

Target: Deposit

Logistic regression

areaUnderROC
Train data: 0.903623753467109
Test data: 0.9051773856966862

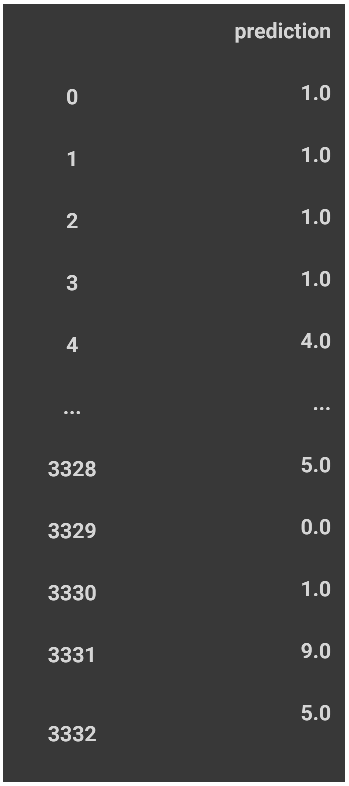
RandomForestClaassifier

areaUnderROC
Train data: 0.9034560677047303
Test data: 0.8982238331825921

- The performance of both the models is almost similar.
- The target audience must be young age and elderly(retired) people.
- Their previous Outcome is positive.
- The Customer must be engaged for at least 6 minutes.

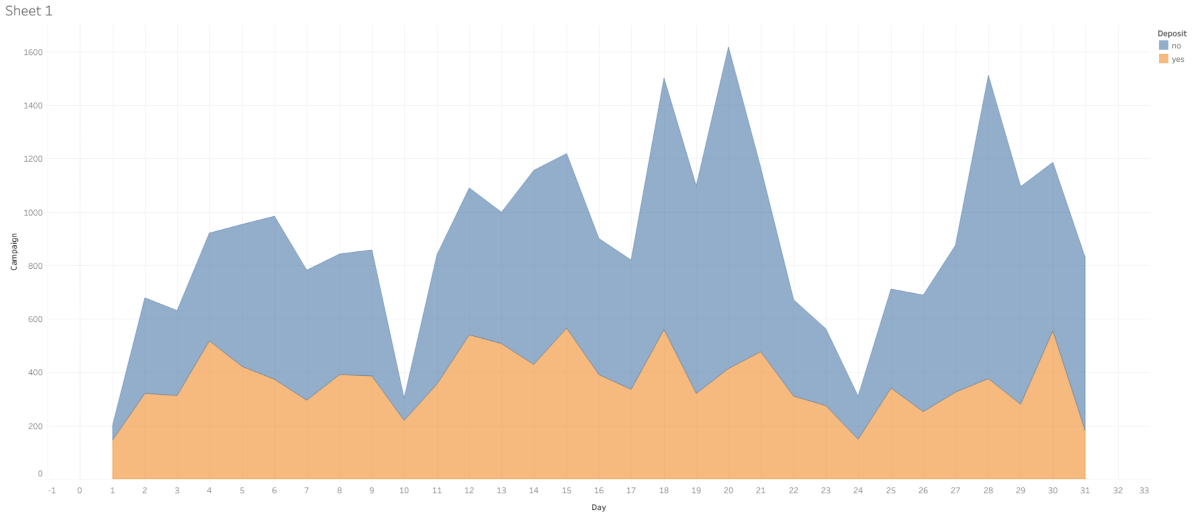
Target: Month

RandomForestClassifier



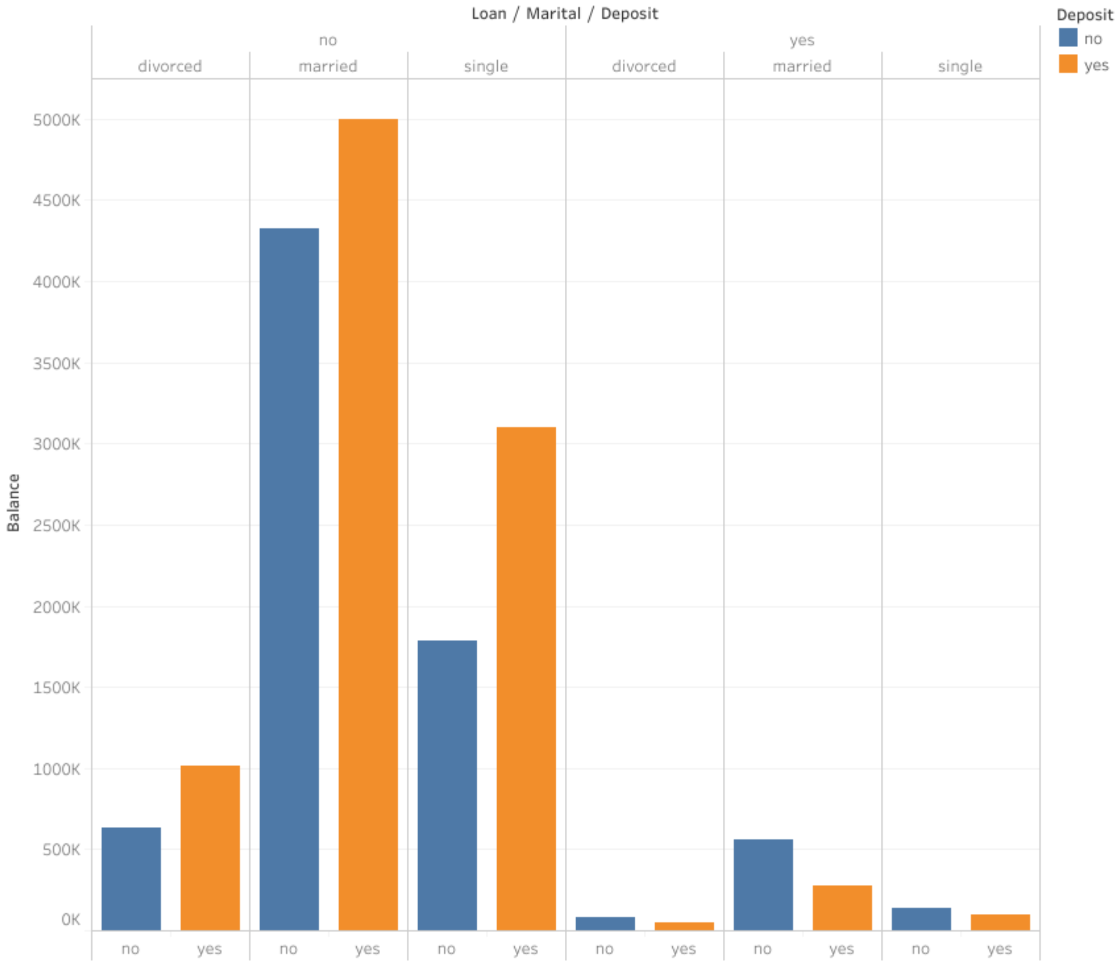
	prediction
0	1.0
1	1.0
2	1.0
3	1.0
4	4.0
...	...
3328	5.0
3329	0.0
3330	1.0
3331	9.0
3332	5.0

Using Big Data Tool



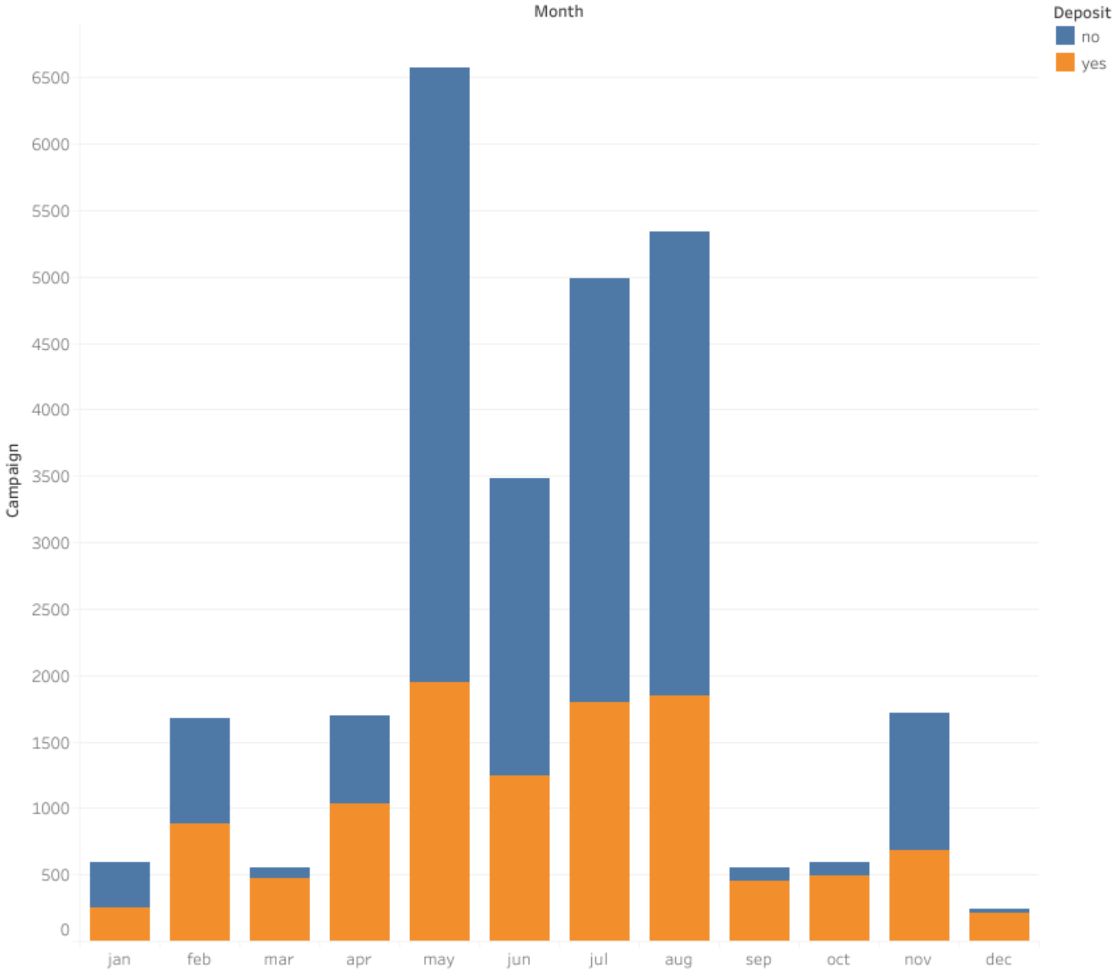
Deposits made by Days in Month.

Sheet 1



Deposits Made Based on Loan and Marital Status.

Sheet 1



Deposits made based on Months.