CAPSTONE PROJECT 1 MILSTONE REPORT

Anshul Gupta

Springboard

Problem Statement

The goal of the business problem is to predict if a customer is going to peruse the banking service or not. This is a binary classification problem where the response variable is 'yes' or 'no' where yes is that a customer is ready to avail the banking service and no is that customer will not avail any service from the bank.

Clients

The client will be banks who want to expand their customer base. The prediction model will save time and help to narrow down the potential customers of the bank from the rest.

Dataset

Data associated with direct marketing campaigns of a Portuguese banking institution. The Dataset is available from https://archive.ics.uci.edu/ml/datasets/bank+marketing .

Dataset Description

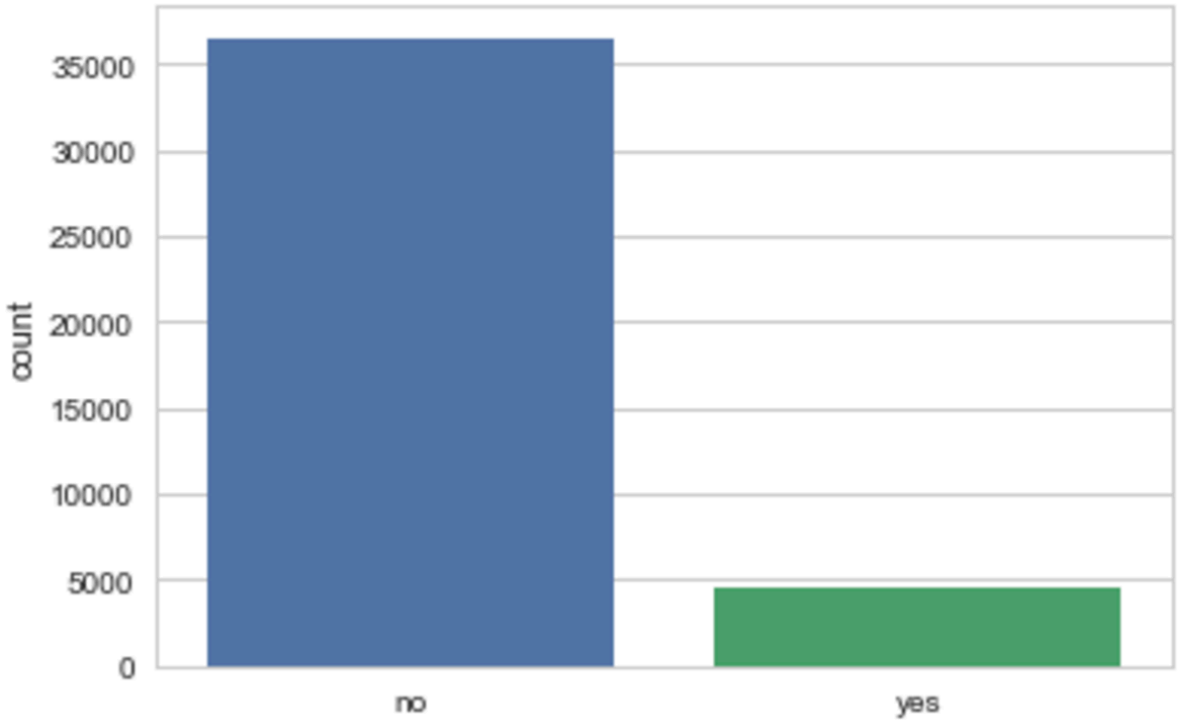
The bank's marketing dataset has about 40K+ instances, 20 features, and a binary response variable. Using this dataset, I have implemented various Machine Learning Binary Classification Models to predict whether a client will subscribe to the product being offered by the bank, or not. The table below summarizes the fields in the dataset, and their types.

Dataset Field	Type
Age	Numeric (Integer)
Job	Categorical
Marital	Categorical
Education	Categorical
Default (Indicate whether the client has credit in default)	Categorical
Housing	Categorical

Loan	Categorical
Contact	Categorical
Month	Categorical
Day of the week	Numeric (Integer)
Duration	Numeric (Integer)
Campaign	Numeric (Integer)
PDays (Number Of Days passed by since the client was last contacted in a previous campaign)	Numeric (Integer)
Previous (Number of contacts performed before this campaign for this client)	Numeric (Integer)
POutcome (Outcome of previous marketing campaign)	Categorical
Emp.var.price (Quarterly Employee variation rate)	Numeric (Decimal)
Cons.price.idx (Monthly consumer price index)	Numeric (Decimal)
Cons.conf.idx (Montly consumer confidence index)	Numeric (Decimal)
Euribor3m	Numeric (Decimal)
Nr.Employed	Numeric(Decimal)
Y	Categorical

The dataset contains 36548 negative labels and 4640 positive labels. From the perspective of the business problem, a positive label is vital as these are associated with customers who have decided to subscribe to offered bank products. Given the marked difference between positive and negative labels in the dataset, this type of classification problem is called “imbalanced”.

The chart below graphically compares the number of positive (“yes”) and negative (“no”) labels.



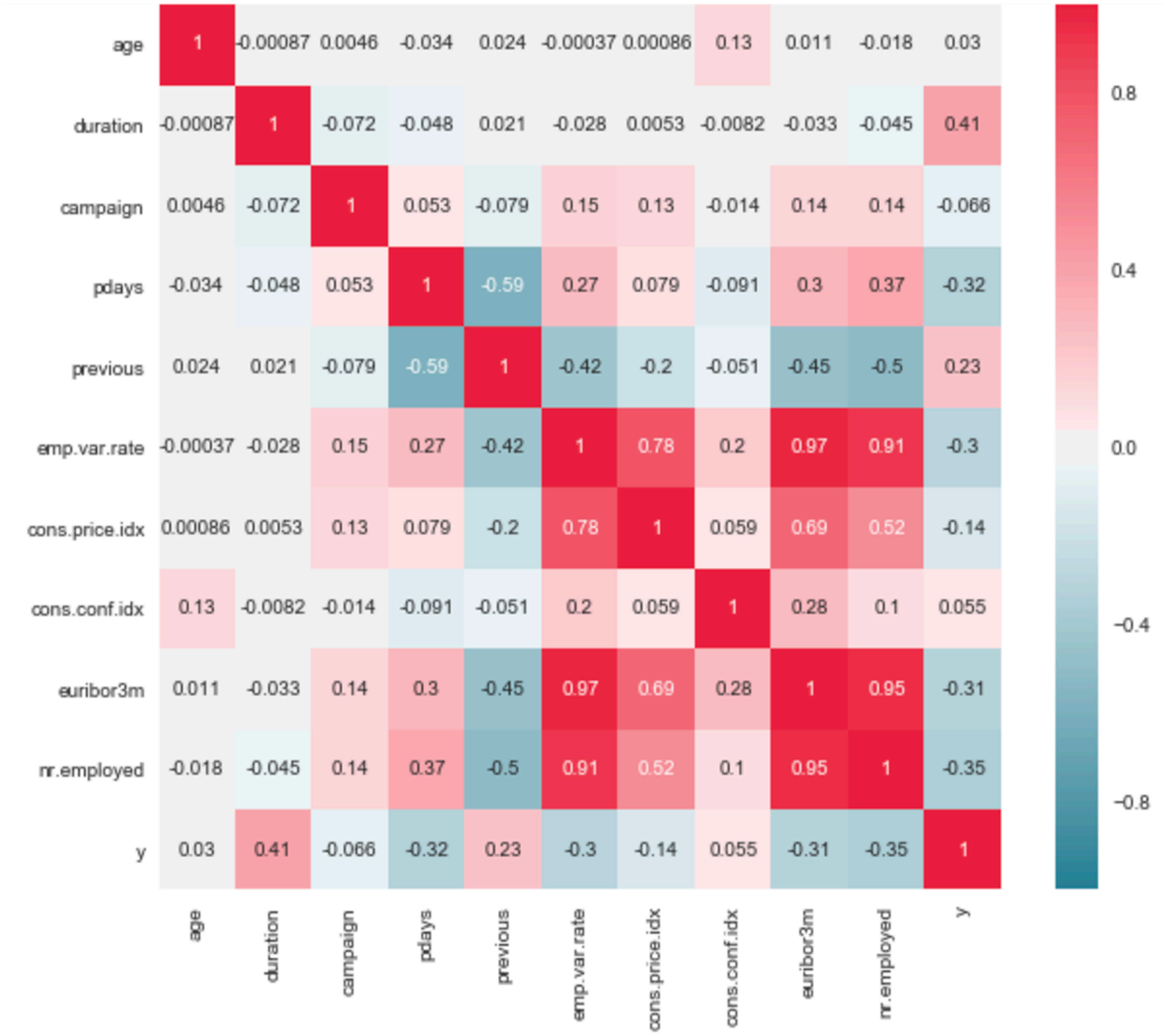
There are about 21 attributes out of which Age, Job, Marital Status, Education, Default, Housing and Loan are the customer features. Contact, Month, Day Of Week and Phone Call Duration are data recorded by telemarketing campaign based phone calls. Employment Variation Rate, Consumer Price Index, and Consumer Confidence index are the social economic factors included in the Dataset . The response variable has categorical values ‘yes’ or ‘no’. This business problem is modeled as a machine learning classification problem. Therefore, the data science problem is to build a binary classifier that can be used to assist the clients in addressing the business problem.

Data Wrangling and Exploration

There are about 10 features which have categorical values, which were “one-hot encoded” by introducing multiple “dummy variables” that take numeric values

(either 0 or 1). For that, Pandas provide a function called pd.get_dummies to one-hot encode the categorical features to fit with various classification and regression problems. After performing this transformation, the number of attributes increased from 21 to 64.
I examined the potential correlation between pairs of attributes to determine any positive or negative correlation with respect to response variable ‘y’. Correlation describes the degree of linear relationship between pairs of variables.

Following correlation chart shows correlation of numerical columns with response variable and among themselves. The diagram illustrates no particularly strong correlation with respect to the response variable.



Descriptive Statistics

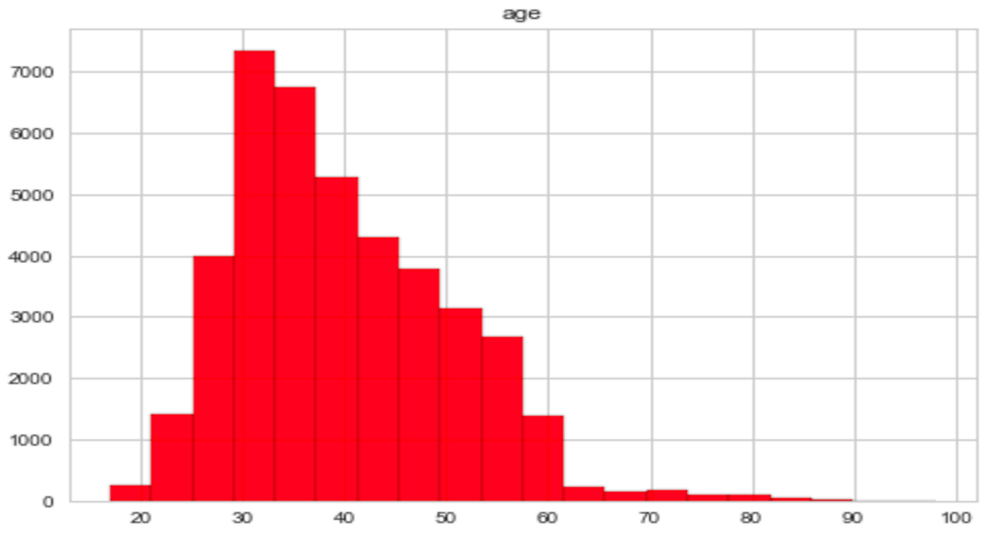
The Data contains no missing values hence there was no need to impute missing values. The following tables shows statistics for numerical columns.

	Age	Duration	Campaign	Pdays	Previous
count	41188	41188	41188	41188	41188
mean	40.02	258.29	2.58	962.48	0.17
std	10.42	259.28	2.77	186.91	0.49

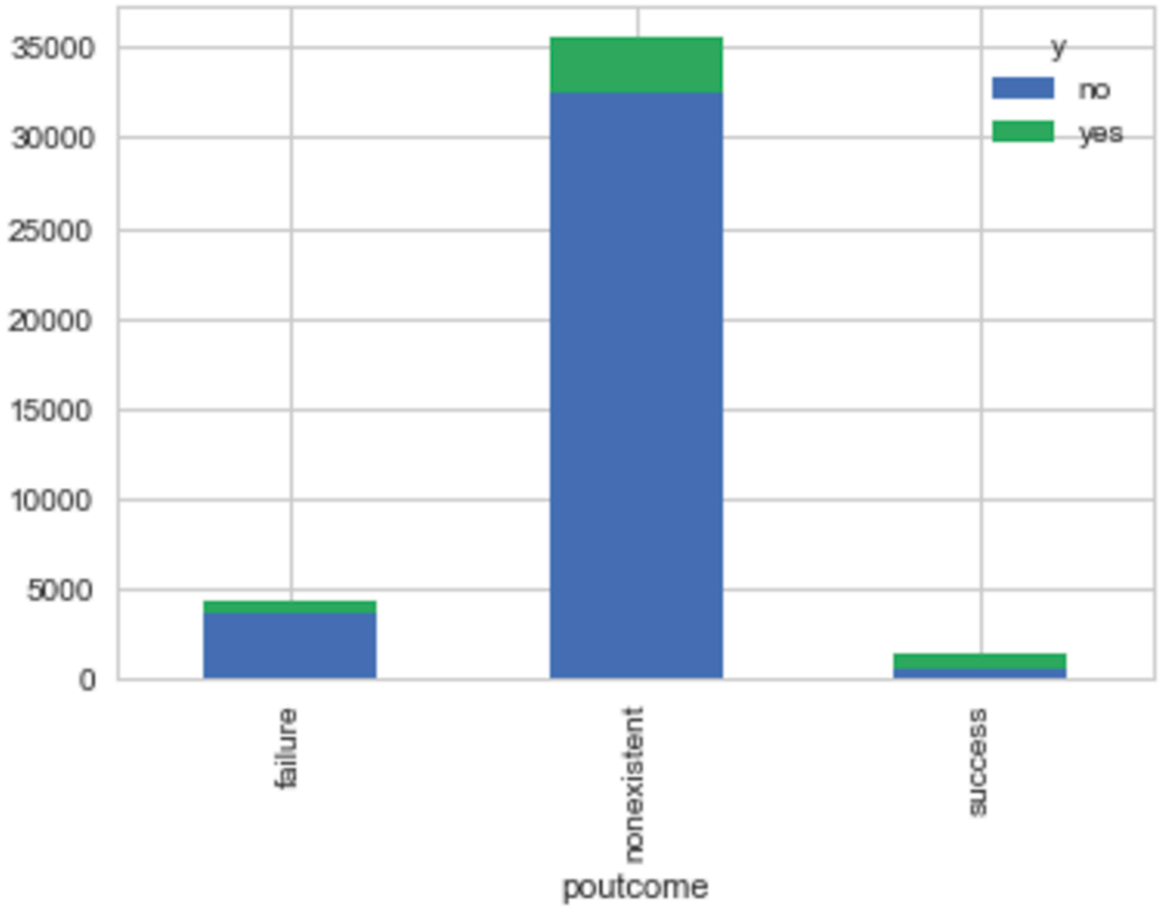
min	17	0	1	0	0
25%	32	102	1	999	0
50%	38	180	2	999	0
75%	47	319	3	999	0
max	98	4918	56	999	7

	Emp.var.rate	**Cons.price.idx**	**Cons.conf.idx**	**Euribor3m**	**Nr.Employed**
count	41188	41188	41188	41188	41188
mean	0.08	93.57	-40.50	3.62	5167.03
std	1.57	0.57	4.63	1.73	72.25
min	-3.4	92.20	-50.8	0.634	4963
25%	-1.8	93.07	-42.70	1.34	5099.1
50%	1.10	93.749	-41.8	4.857	5191
75%	1.4	93.99	-36.4	4.96	5228.1
max	1.4	94.77	-26.9	5.04	5228.1

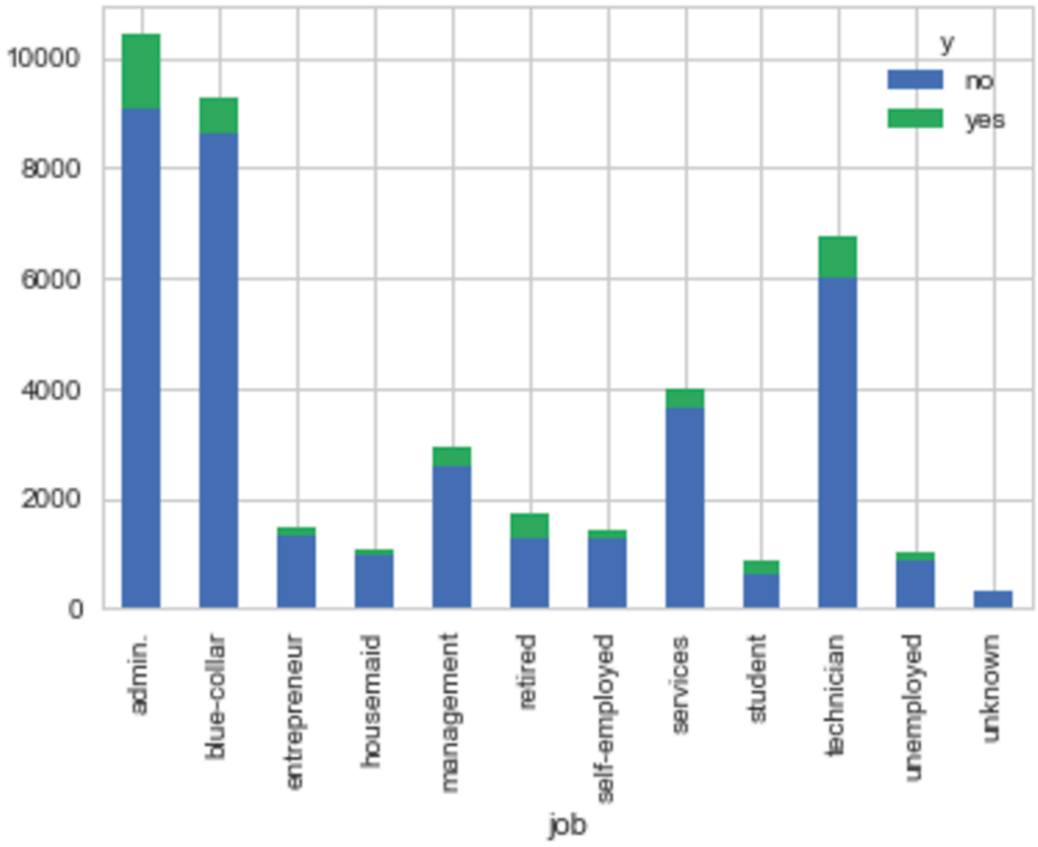
The following histogram shows that age group 30-40 shows maximum number of customers in the Dataset.



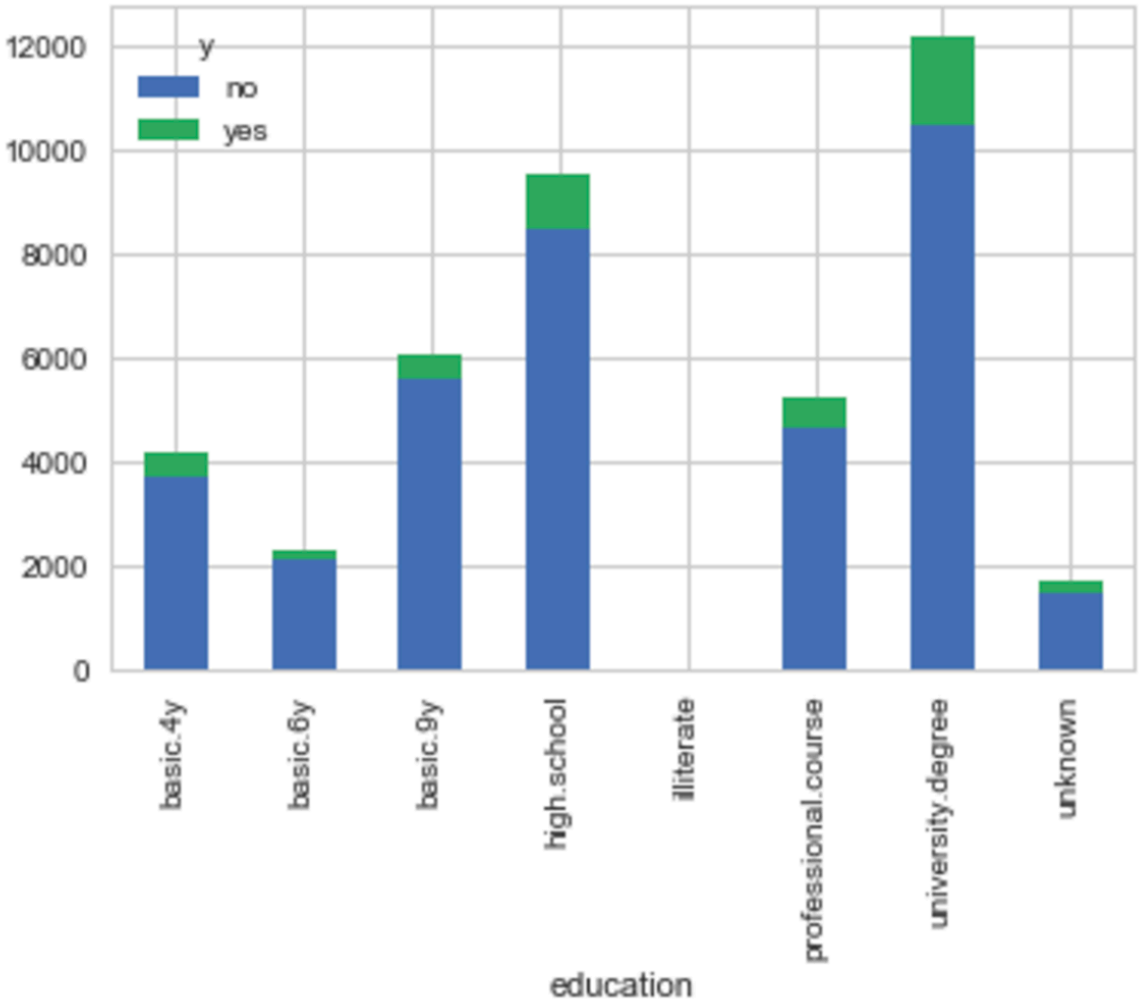
The following plot shows the distribution of values for the variable associated with previous outcome.



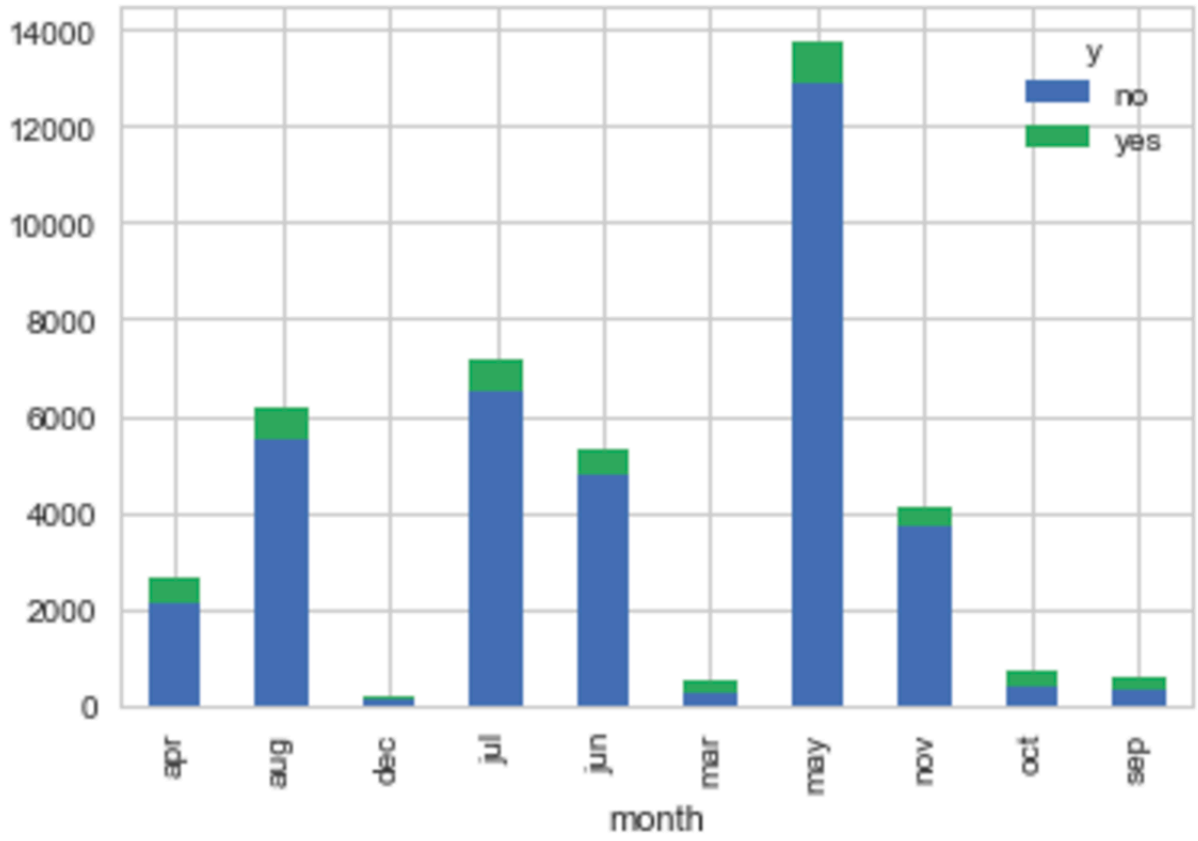
The three top jobs people were contacted about were Admin, Blue-Collar and technician which comprise more than half of the total people contacted by the Telemarketing company.



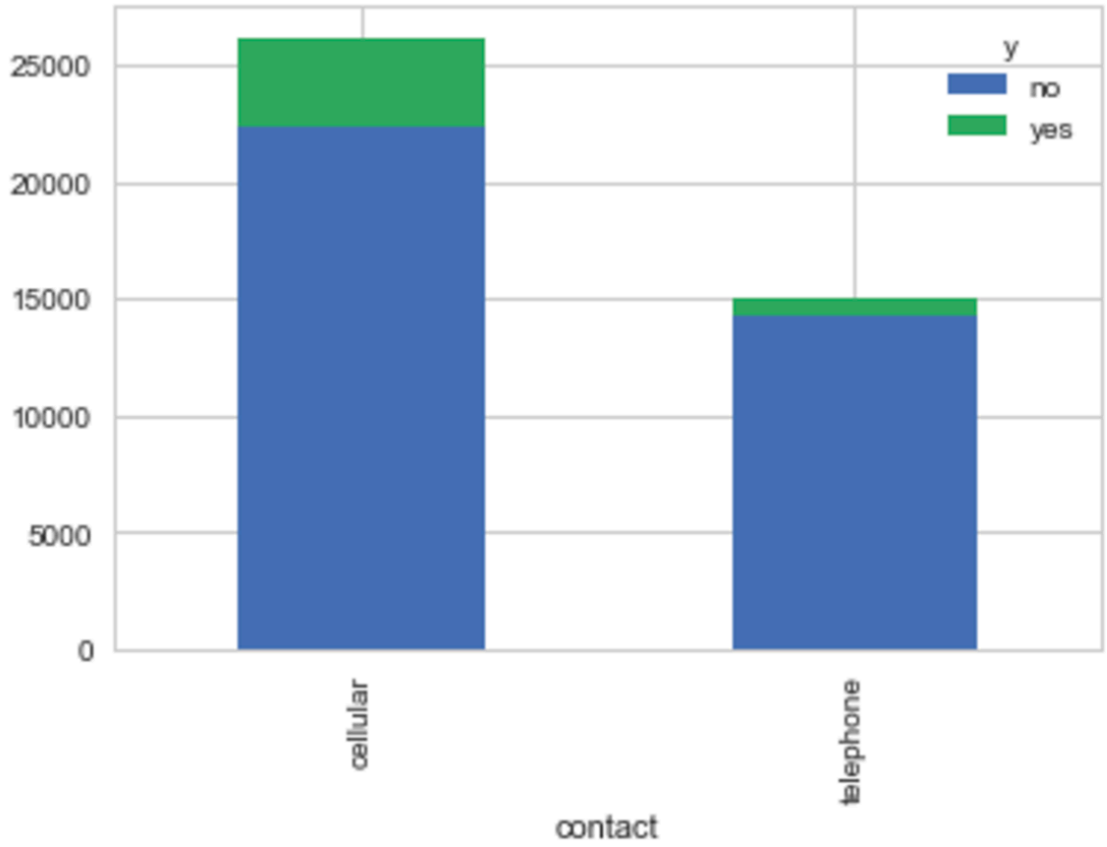
The following plot shows that people with university degrees were contacted the most**.**



The following plot shows that the month of May was when maximum calls were made which is almost double than second highest month that is July.

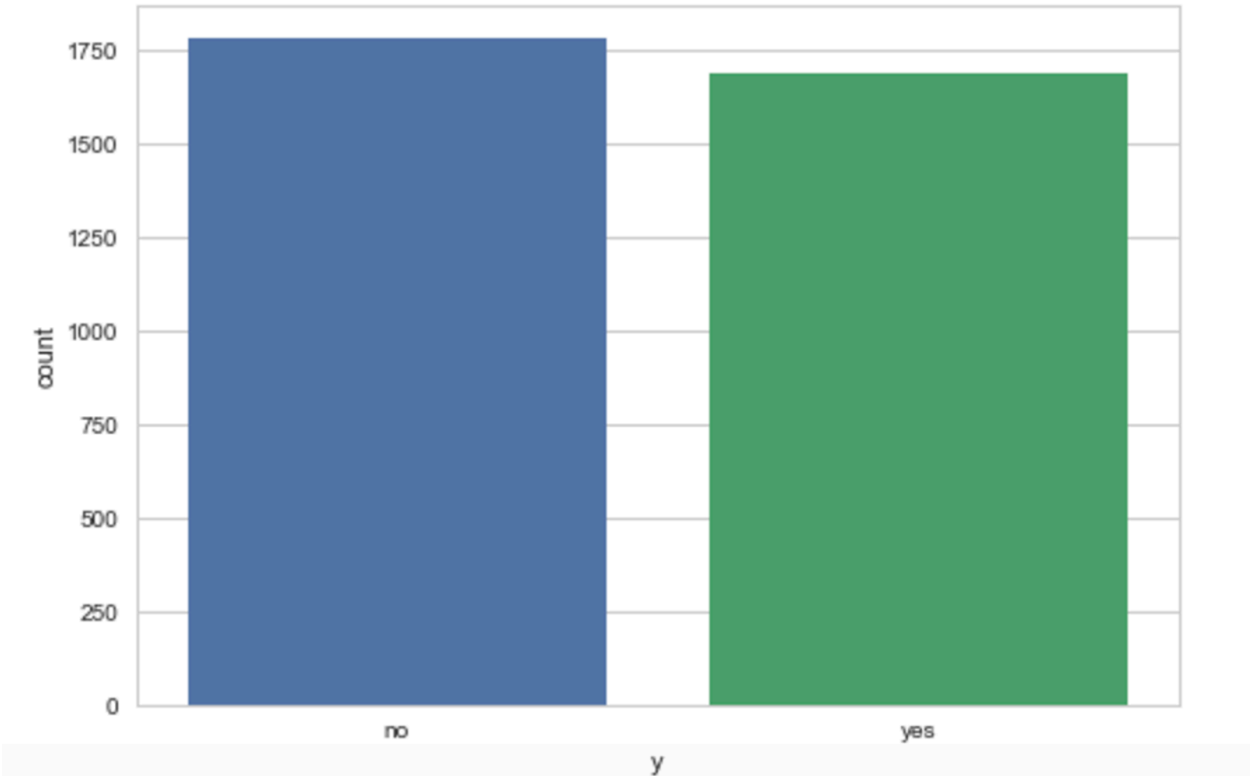


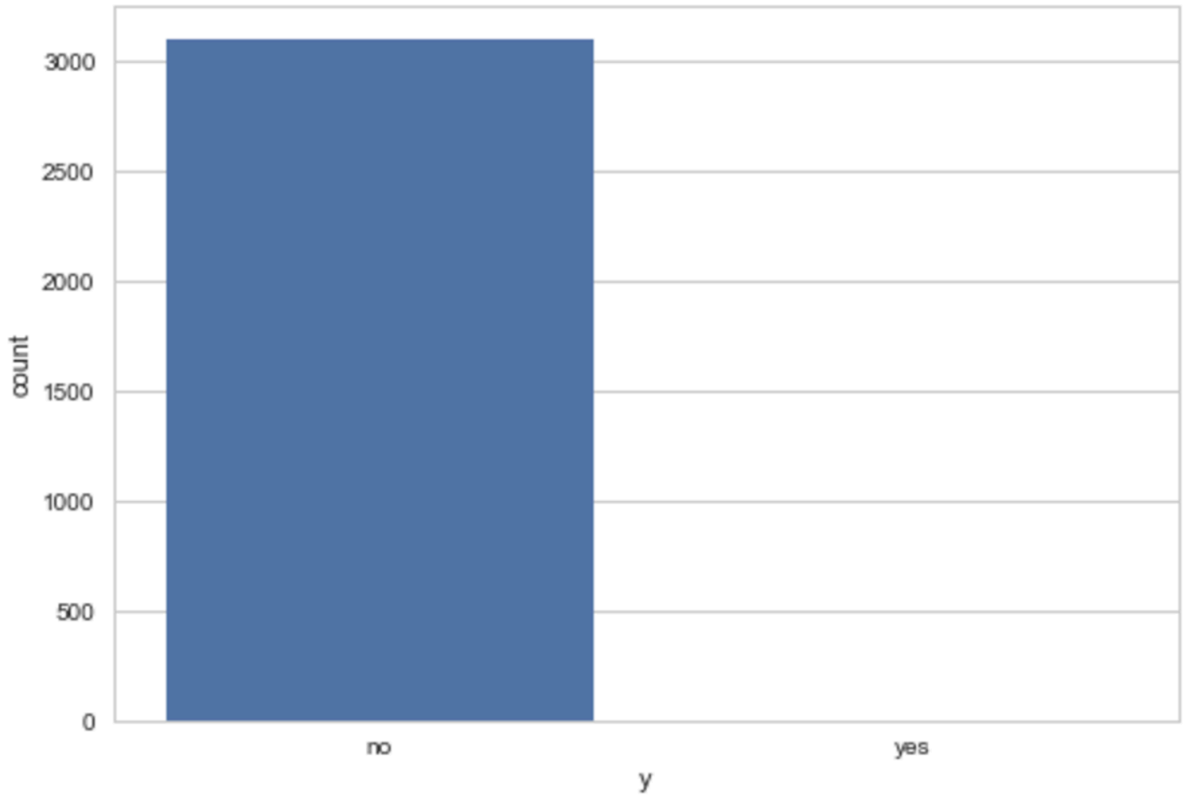
The following chart shows that there were two types of method of contact: one was via a mobile phone and other was a regular phone. The bars show yes or no for mobile and regular phones. It shows that the former was preferred over the latter.



The call duration attribute could highly influence the response variable. Longer the call duration the more chances are for a yes. When the duration of call is more than 600s, the response variable has value ‘no’ with frequency 1780, and has value ‘yes’ with frequency 1684. When the call duration is less than 50 seconds, the response variable is ‘no’ 3101 times, and ‘yes’ only in one case.

So, we can infer that call duration plays an important role in getting a positive response for the bank. But as soon as the call is over we get to know the outcome, and hence this can be used as benchmark that longer call duration is a sign that customer is going to avail a bank product.


1750
1500
1250
1000
750
500
250
0
count
no
yes
y


3000
2500
2000
1500
1000
500
0
count
no
yes
y