Capstone Project Report

Bank Direct Marketing

Sewalkar, Himanshu

4/29/18

Udacity Machine Learning Nanodegree

Contents

1 Definition 3

1.1 Background 3

1.2 Problem Statement 3

1.3 Datasets and Inputs 4

1.4 Solution Statement 4

1.5 Benchmark Model 4

1.6 Evaluation Metrics 4

1.7 Project Design 7

2 Analysis 8

2.1 Data Exploration 8

2.1.1 Shape of the data 8

2.1.2 Statistics for Continuous Variables 11

2.1.3 Statistics for Categorical Features 11

2.1.4 Additional Statistical Observations 12

2.2 Exploratory Visualizations 12

2.2.1 Numerical Data Distribution 12

2.2.2 Detailed Exploration 14

3 Methodology 24

3.1 Data Preprocessing 24

3.1.1 Step 1 24

3.1.2 Step 2 24

3.1.3 Step 3 24

3.1.4 Step 4 25

3.1.5 Step 5 25

3.2 Data Correlation 26

3.3 Implementation 27

3.3.1 Developing a Logistics Regression – The Benchmark Model 27

3.3.2 Developing a Naïve Bayes Model – The Solution Model 29

3.4 Improvement 31

3.4.1 Developing a Naïve Bayes Model – The Solution Model (Part II) 31

3.5 Implementation - Extracting Feature Importance 34

3.6 Programming Tips 34

4 Conclusion 35
4.1 Result Summary 35
4.2 Key Remarks 35
4.3 Feature Importance 36
4.4 Recommendations to the Bank 36
4.5 Reflection on my learning 36
4.6 Future Work/Improvement 37
5 References 37

Welcome to my capstone project of the Machine Learning Engineer Nanodegree!

1 Definition

1.1 Background

Direct marketing is a form of advertising where organizations communicate directly to customers through a variety of media including cell phone text messaging, email, websites, online adverts, database marketing, fliers, catalog distribution, promotional letters and targeted television, newspaper and magazine advertisements as well as outdoor advertising. The common form of direct marketing is telemarketing, in which marketers contact customers by phone. The primary benefit to businesses is increased lead generation, which helps businesses increase sales volume and customer base. The most successful telemarketing service providers focus on generating more "qualified" leads that have a higher probability of getting converted into actual sales.[1]

If we consider banking sector, then we can notice that post 2008 economic crisis there is greater pressure on banks to increase profits while reducing costs. Thus, optimization of direct marketing effort under telemarketing has become a key concern. Direct marketing within banking sector is a method where customers are directly informed about banking products, such as credit cards, new savings account types etc. which are analyzed and selected per customer's characteristics and contacted over a communication channel like phone calls.

In direct marketing, the return rate and the effectiveness of the campaigns can be measured using the responses of customers, and improvements on the campaigns can also be made.

1.2 Problem Statement

A Portuguese bank had experienced a revenue decline. It was found the root cause is that bank’s customers are not depositing as frequently as before.

A term deposits benefits a bank in the following way:

- It can invest the deposits in those financial products which lead to higher gain; to make a profit which is higher than the cost associated with deposits.
- Secondly, more the term deposit customers higher is the probability that a bank can upsell or cross sell other products to further increase revenues.

Hence, the Portuguese bank would like to identify existing clients that have higher chance to subscribe for a term deposit and thus focus marketing effort on such clients.

[1] https://en.wikipedia.org/wiki/Direct_marketing

1.3 Datasets and Inputs

The dataset we strive to use in the capstone project is from the UC Irvine Machine Learning Repository and holds information related to a direct marketing campaign of the previously described Portuguese bank.[2]

It was obtained by downloading bank-additional-full.csv (contained in bank-additional.zip) from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

1.4 Solution Statement

The output variable (y) is having binary response which means either the customer will subscribe for a term deposit or not. Therefore, I will use a classification approach to predict which customers are more likely to subscribe for term deposits. The methods I will attempt to try are (i) logistic regression, (ii) naïve bayes.

My solution will aim for the following two outcomes:

- Determining which variables are important
- Implementing the classifier

to predict if a customer will subscribe for a term deposit.

1.5 Benchmark Model

Logistic regression is mainly used in cases where the output is Boolean. So is the case under consideration, the desired target is either a customer subscribes for a term deposit or not. Here, we consider features and outcome Y which can take two values {0,1}. Logistic regression measures the relationship between an output variable Y (categorical) and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. So, logistics regression will be my benchmark model.

1.6 Evaluation Metrics

Naive Bayes classifiers have worked very well in many practical situations, famously known spam filtering. Naive Bayes also assumes that the features are conditionally independent. Real data sets are never flawlessly independent but they can be close. In short Naive Bayes has a higher bias but lower variance compared to logistic regression. Therefore, it will be interesting to use Naïve Bayes as solution model and compare that against the benchmark model based on logistics regression.

[2] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Both Naive Bayes and Logistic regression are linear classifiers, Logistic Regression makes a prediction for the probability using a direct functional form whereas Naive Bayes figures out how the data was generated given the results.[3]

I will the below evaluation metrics for the comparison of the models.

	Naïve Bayes	*Logistics Regression*
Avg. accuracy (test)	A1	A2
10-fold CV avg. accuracy	B1	B2
ROC/AUC Score	C1	C2
f1-score (avg/total)	D1	D2
wall time	E1	E2

Accuracy: Accuracy is the percentage of data points that have been correctly classified out of the total points. This is the most common evaluation metric for classification problems.

We will also use 10-fold cross-validation method to determine mean accuracy of our benchmark and solution models. With *10-fold cross-validation* we aren't just creating multiple test samples repeatedly, but are dividing the complete dataset we have into 10 disjoint parts of the same size.

A study of cross-validation for accuracy estimation and model selection is available by Ron Kohavi here

http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf

ROC/AUC Score: Area under ROC Curve (or AUC for short) is a performance metric for binary classification problems. The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.

F1 score: The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

The formula for the F1 score is: `F1 = 2 * (precision * recall) / (precision + recall)`[4]

In binary classification, precision is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total number of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.[5]

[3] https://medium.com/@sangha_deb/naive-bayes-vs-logistic-regression-a319b07a5d4c

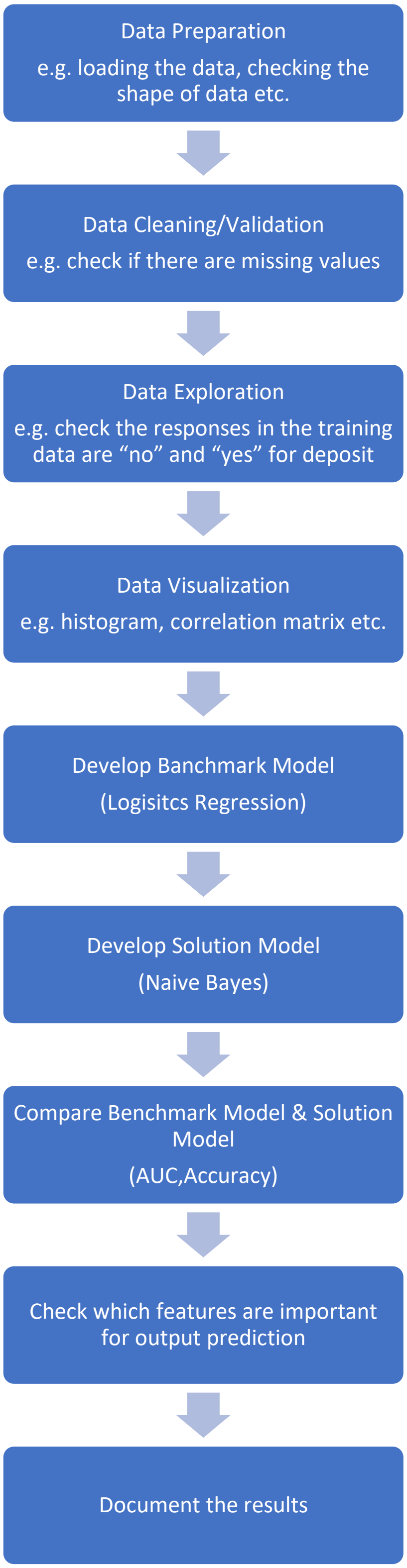
[4] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score

[5] https://en.wikipedia.org/wiki/Precision_and_recall

Wall time: Wall time is the actual time lapsed in completing a job. It is same like timing a job with a stopwatch.[6]

[6] http://www.techbeamers.com/python-time-functions-usage-examples/

1.7 Project Design


Data Preparation
e.g. loading the data, checking the shape of data etc.
Data Cleaning/Validation
e.g. check if there are missing values
Data Exploration
e.g. check the responses in the training data are “no” and “yes” for deposit
Data Visualization
e.g. histogram, correlation matrix etc.
Develop Banchmark Model
(Logisitcs Regression)
Develop Solution Model
(Naive Bayes)
Compare Benchmark Model & Solution Model
(AUC,Accuracy)
Check which features are important for output prediction
Document the results

2 Analysis

2.1 Data Exploration

2.1.1 Shape of the data

The table contains 41,188 rows (i.e., tuples describing telemarketing interaction) and 21 columns (i.e., features of each interaction).

Input Variables

Customer Data

#	Column Name	Description	Type	Values
1	Age	Customer's age	Numerical	
2	Job	Customer's Profession	Categorical	• `admin` • `blue-collar` • `entrepreneur` • `housemaid` • `management` • `retired` • `self-employed` • `services` • `student` • `technician` • `unemployed` • `unknown`
3	Marital	Marital Status	Categorical	• `divorced` • `married` • `single` • `unknown` Note: `divorced` means divorced or widowed
4	Education	Customer's Education Level	Categorical	• `basic.4y` • `basic.6y` • `basic.9y` • `high.school` • `illiterate` • `professional.course` • `university.degree` • `unknown`

5	Default	Indicator if the customer has credit in default	Categorical	• no • yes • unknown
6	Housing	Indicator if the customer has a housing loan	Categorical	• no • yes • unknown
7	Loan	Indicator if the customer has a personal loan	Categorical	• no • yes • unknown

Data related with the last contact of the current campaign

#	**Column Name**	**Description**	**Type**	**Values**
8	Contact	Contact communication type	Categorical	• cellular • telephone
9	Month	Month that last contact was made	Categorical	• jan • feb • ⋮ • dec
10	Day_of_week	Day that last contact was made	Categorical	• monday • tuesday • Wednesday • Thursday • Friday
11	Duration	Duration of last contact in seconds	Numerical	Note: This attribute highly affects the output target (e.g., if duration=0 then y=no). Yet, the duration is not known before a call is performed. Also, after the end of the call, y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other Attributes

#	Column Name	Description	Type	Values
12	Campaign	Number of contacts performed during this campaign for this client (including last contact)	Numerical	
13	Pdays	Number of days since the client was last contacted in a previous campaign	Numerical	**Note:** 999 means client was not previously contacted
14	Previous	Number of contacts performed before this campaign for this client	Numerical	
15	poutcome	Outcome of the previous marketing campaign	Categorical	• failure • nonexistent • success
16	emp.var.rate	Employment variation rate (quarterly indicator)	Numerical	
17	cons.price.idx	Consumer price index - monthly indicator	Numerical	
18	cons.conf.idx	Consumer confidence index - monthly indicator	Numerical	
19	euribor3m	Euribor 3-month rate (daily indicator)	Numerical	
20	nremployed	Number of employees (quarterly indicator)	Numerical	

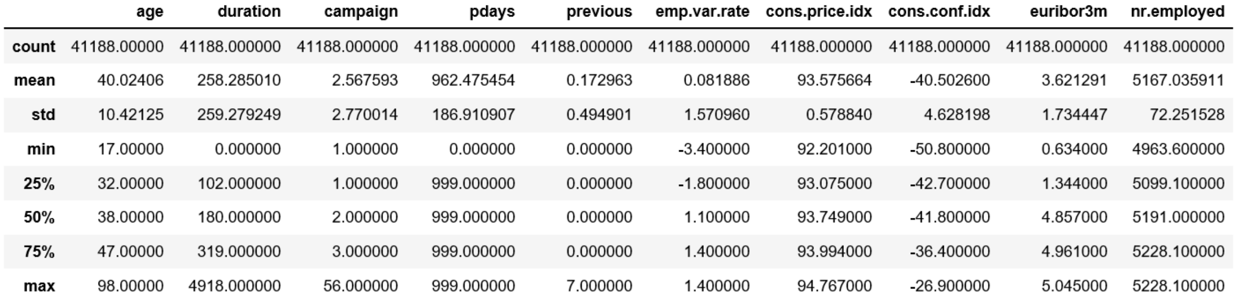
Output Variable (desired target)

#	Column Name	Description	Type	Values
1	Y	Indicator if the client has subscribed for a term deposit	Binary	• yes • no

Observations:

- There are 10 continuous variables 5 of type integer and 5 of type float
- There are 10 variables of type "object" and these are categorical variables
- No missing values

2.1.2 Statistics for Continuous Variables



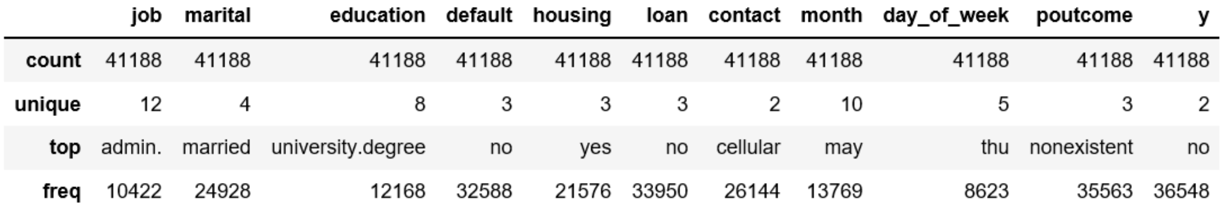
	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-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

Figure 1

Observations:

- The feature 'pdays' has many '999' (missing) values.
- 'previous' = 0 for most of the data
- Remaining numerical features seem OK
- The age of the customers ranges from 17 years to 98 years with mean around 40 years old

2.1.3 Statistics for Categorical Features



	job	marital	education	default	housing	loan	contact	month	day_of_week	poutcome	y
count	41188	41188	41188	41188	41188	41188	41188	41188	41188	41188	41188
unique	12	4	8	3	3	3	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

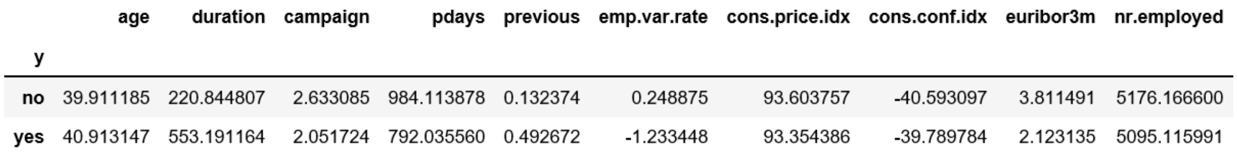
Figure 2

Observations:

- The top job type is admin with a total count of 10422
- 60.52% of the customers are married
- 29.54% of the customers are possessing a university degree
- 79% of the customers have no default
- 52% of the customers have a housing loan
- 82% of the customers do not have a personal loan
- Most contacts to the customer were made in the month of May
- 21% of contacts to the customer were made in on a Thursday
- Outcome of the previous marketing campaign does not exist for majority of the customers i.e. 86%

- Key factor here to note is that majority of the customers i.e. 88% of them have not actually subscribed to a term deposit. Thus, we can see that only 4640 customers have subscribed

2.1.4 Additional Statistical Observations



y	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
no	39.911185	220.844807	2.633085	984.113878	0.132374	0.248875	93.603757	-40.593097	3.811491	5176.166600
yes	40.913147	553.191164	2.051724	792.035560	0.492672	-1.233448	93.354386	-39.789784	2.123135	5095.115991

Figure 3

Observations:

- The average age of customers who bought the term deposit is slightly higher than that of the customers who didn't.
- Campaigns (number of contacts or calls made during the current campaign) are higher for customers who did not opt for the term deposit.
- Number of days since the client was last contacted in a previous campaign which is pdays is higher for customers who did not opt for the term deposit.

2.2 Exploratory Visualizations

2.2.1 Numerical Data Distribution

2.2.1.1 Age, Campaign & Consumer Confidence Index

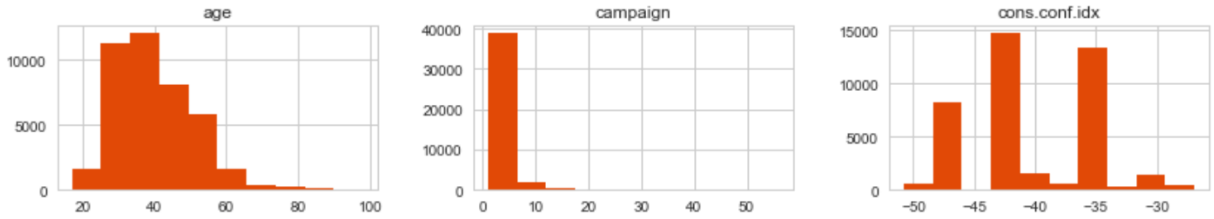


Figure 4

2.2.1.2 Consumer Price Index, Duration & Employment Variation Rate

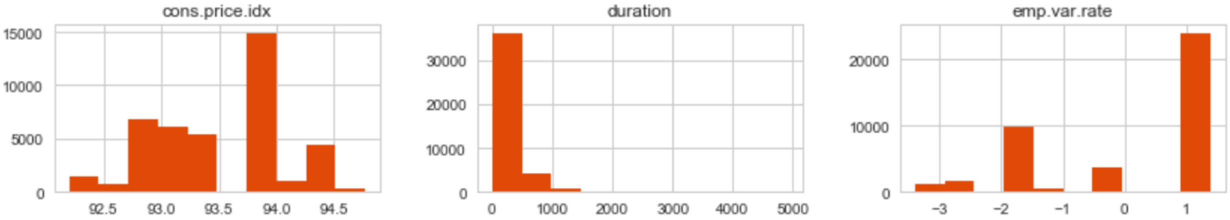


Figure 5

2.2.1.3 Euribor 3-month rate, No. of employees, Number of days since the client was last contacted in a previous campaign, Number of contacts performed before this campaign for this client

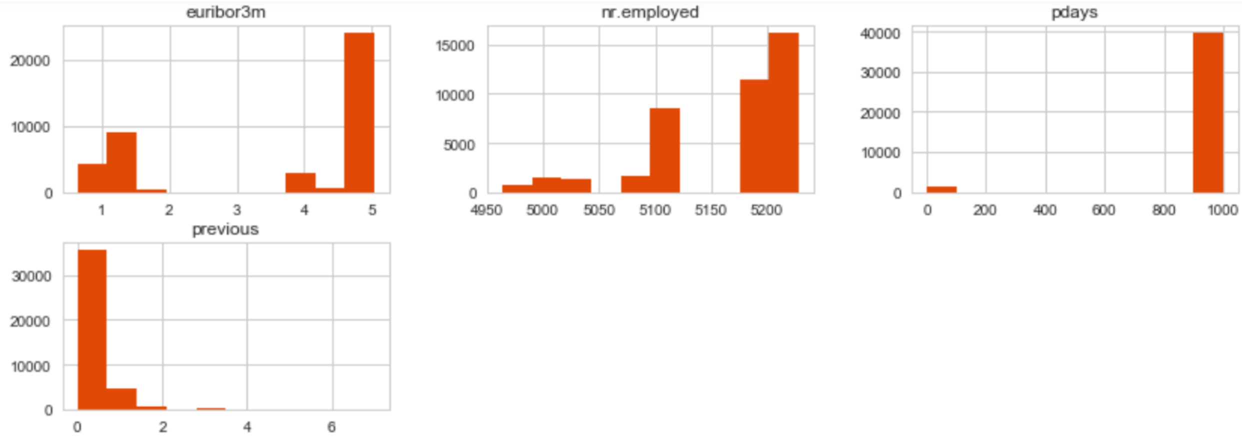


Figure 6

2.2.2 Detailed Exploration

2.2.2.1 Customer's Profession

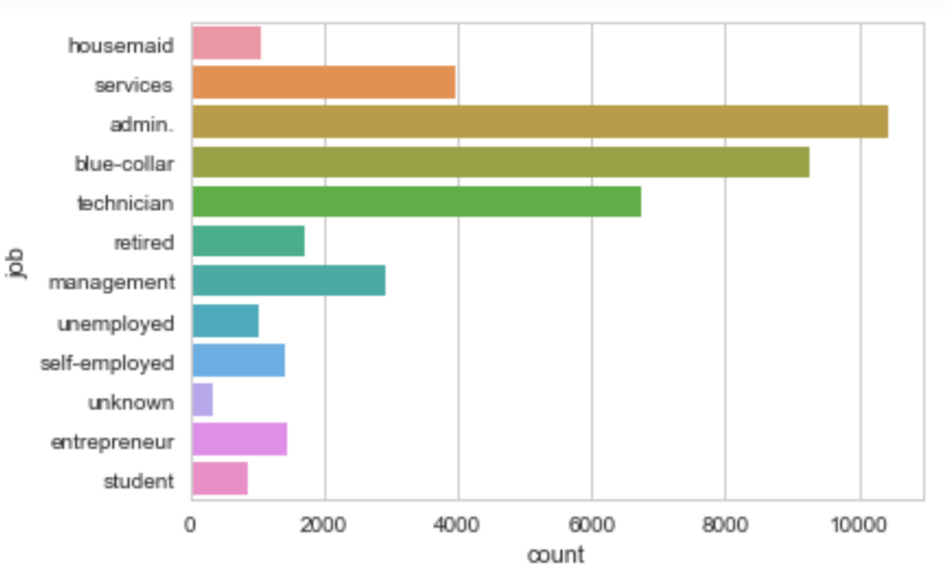
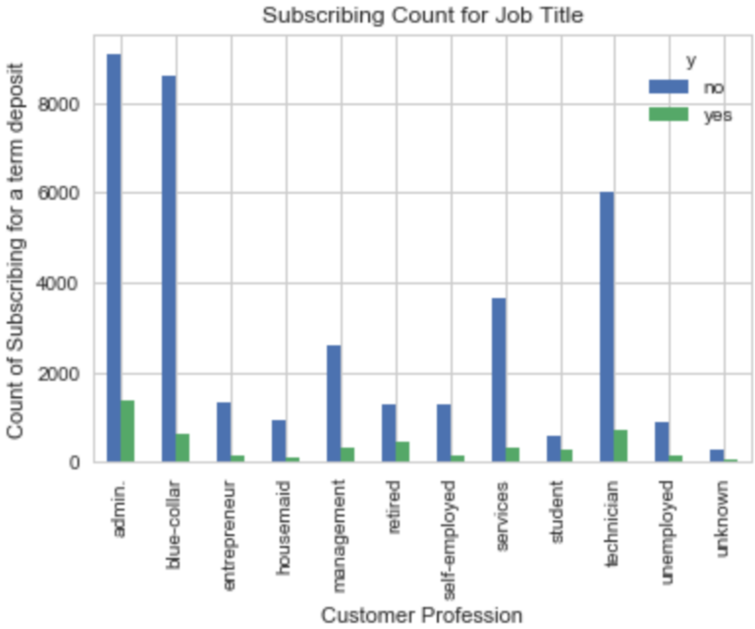


Figure 7

The top job type is admin with a total count of 10422.



The frequency of subscription of the term deposit varies based on the customer's profession. Hence, the job title can be a good predictor of the outcome variable.

2.2.2.2 Marital Status

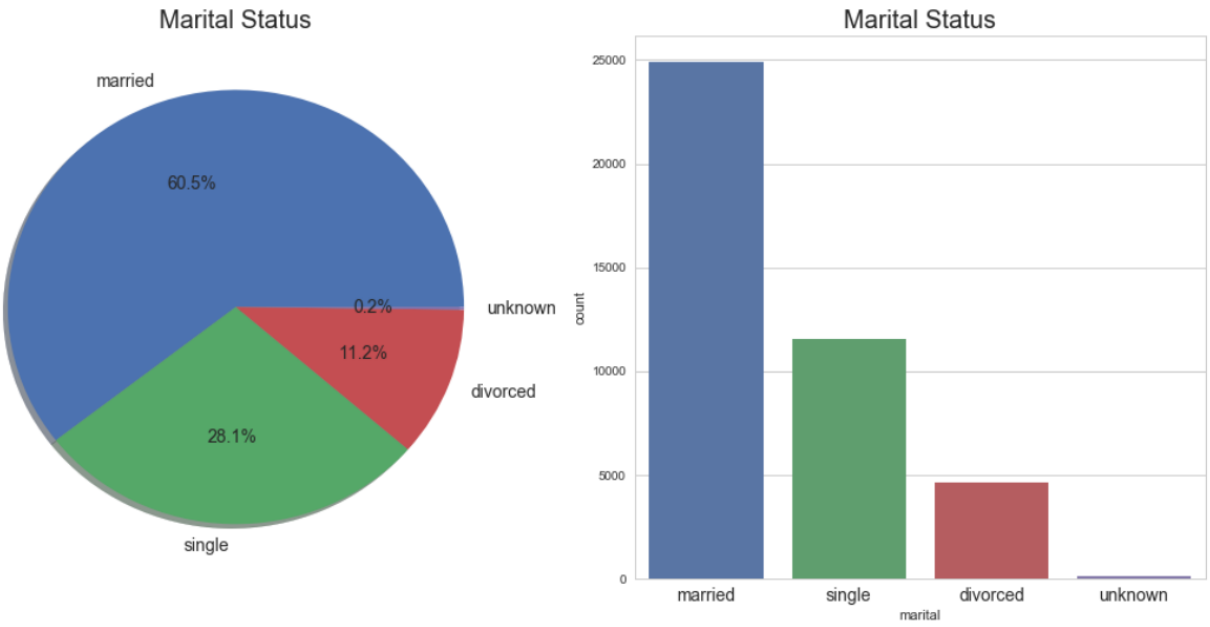


Figure 8

60.52% of the customers are married

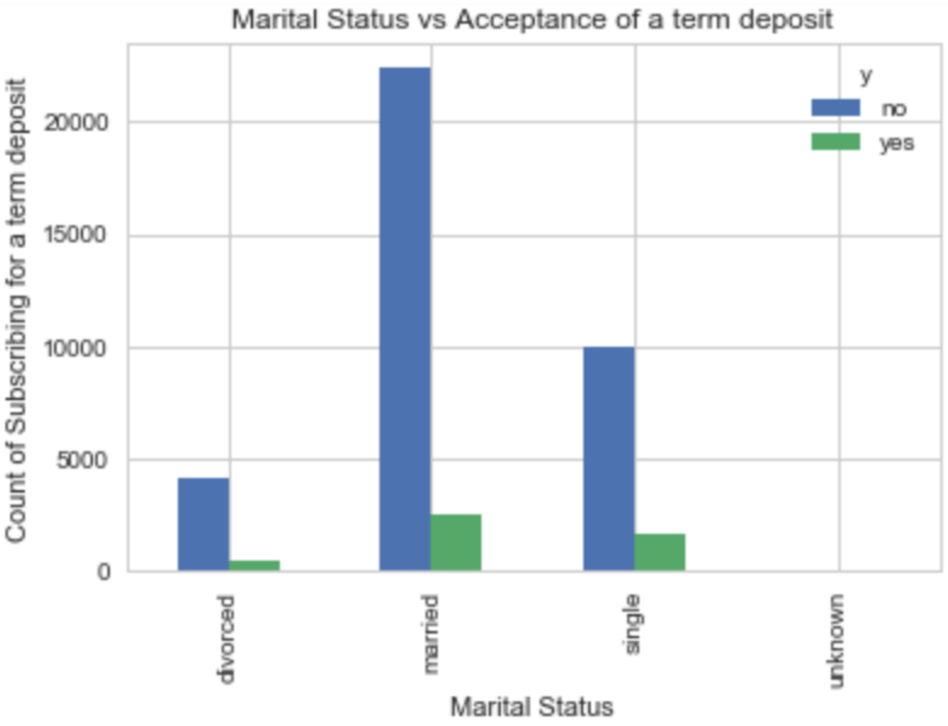


Figure 9

It seems that customers who are divorced show less tendency to open a term deposit.

2.2.2.3 Education

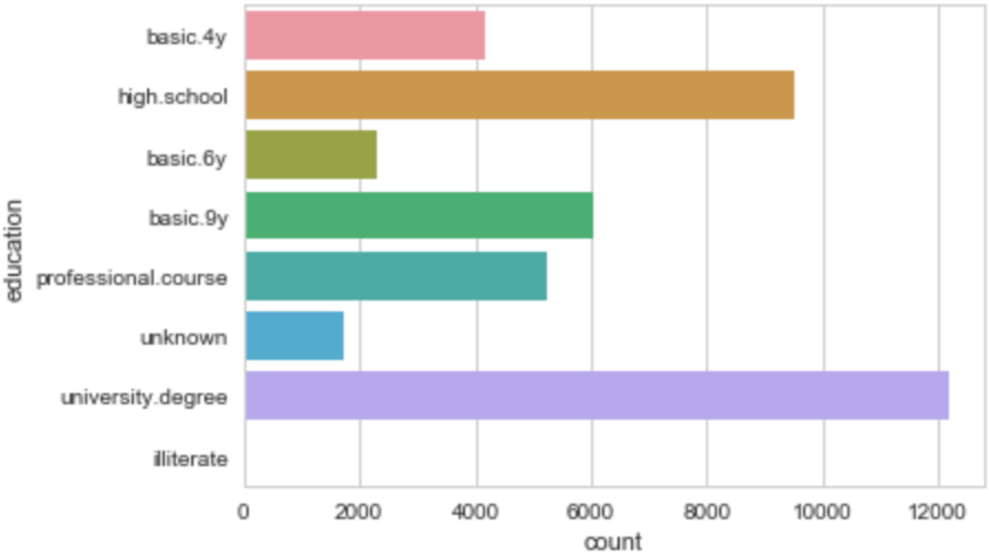


Figure 10

29.54% of the customers are possessing a university degree.

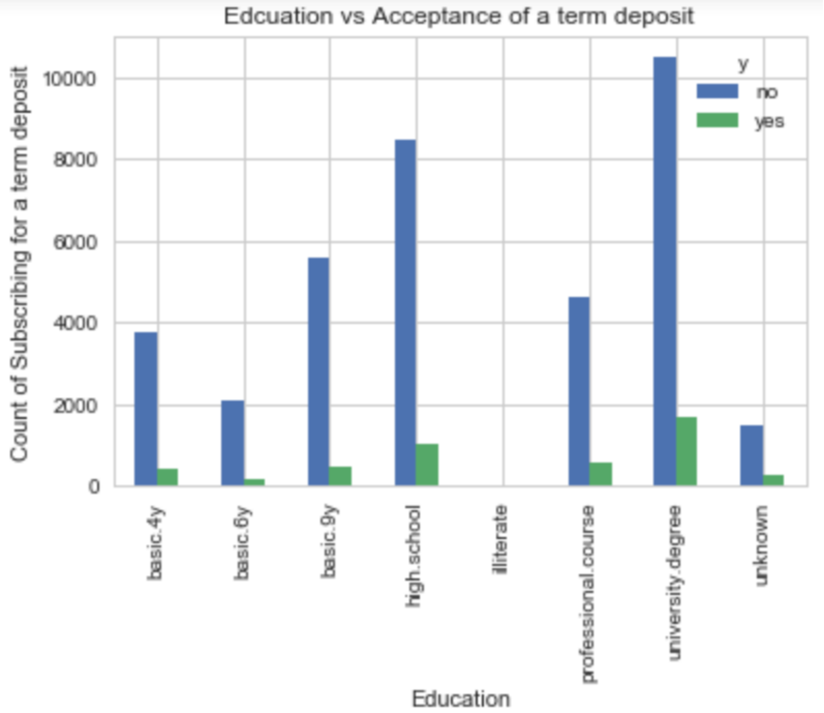


Figure 11

Education appears to be also a very good predictor of the outcome variable.

2.2.2.4 Default

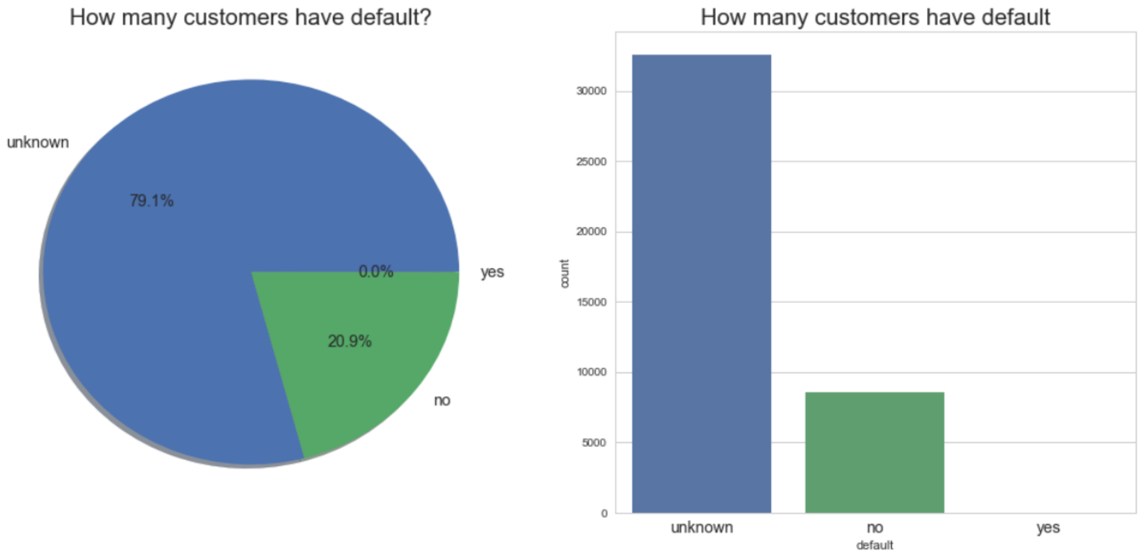


Figure 12

Concretely, we know that 20.9% customers do not have default. The rest is unknown.

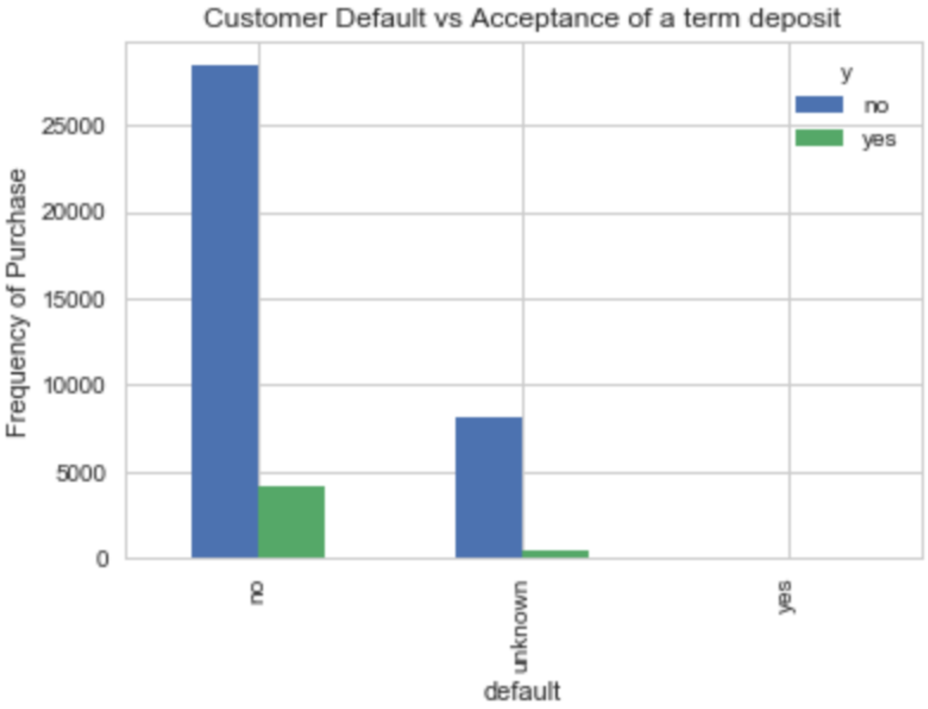


Figure 13

Default is the failure to pay interest or principal on a loan or security when due. This will also be a good feature to consider in prediction.

2.2.2.5 *Housing Loan*

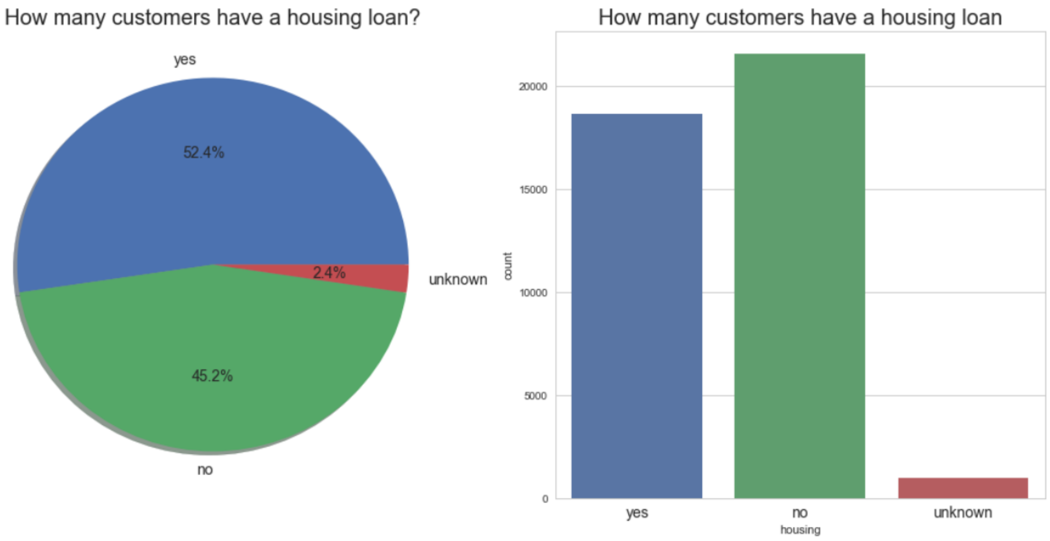


Figure 14

52% of the customers have a housing loan.

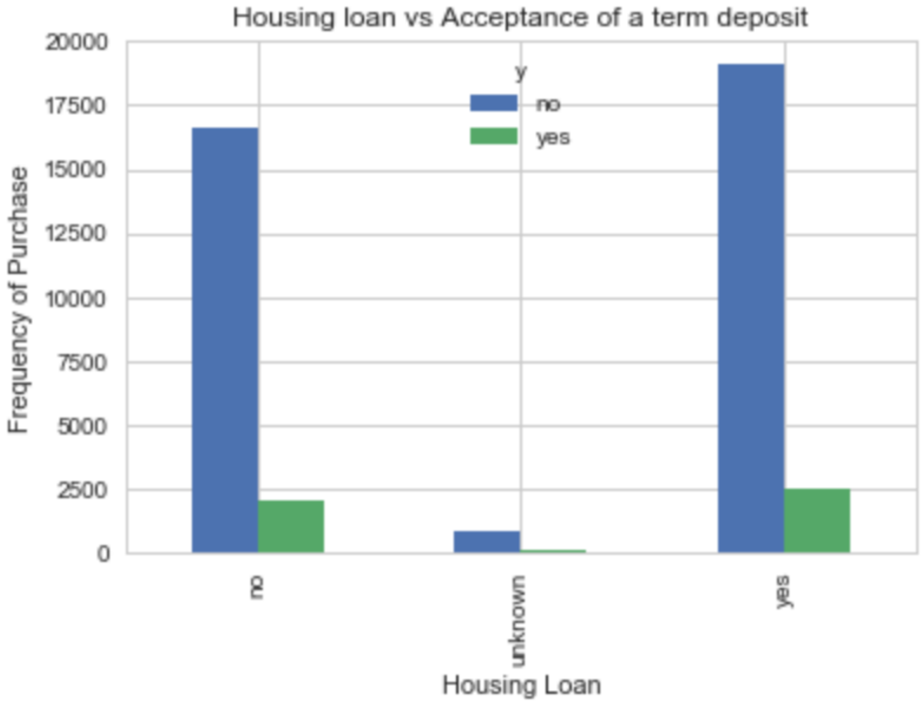


Figure 15

2.2.2.6 Personal Loan

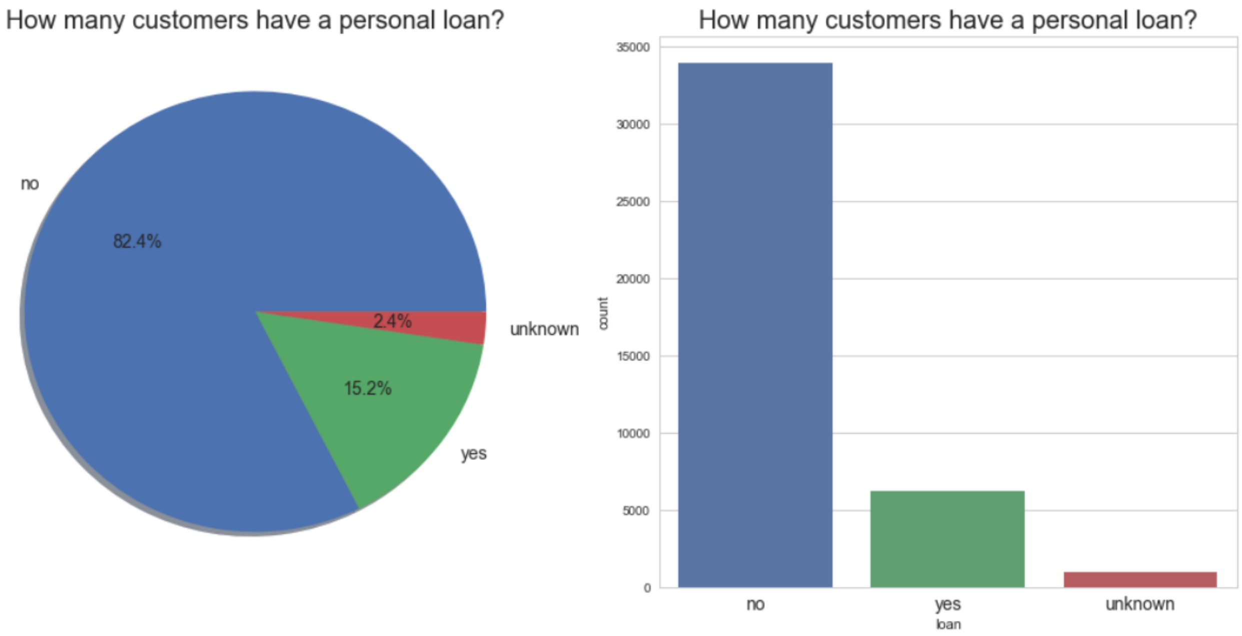


Figure 16

82% of the customers do not have a personal loan.

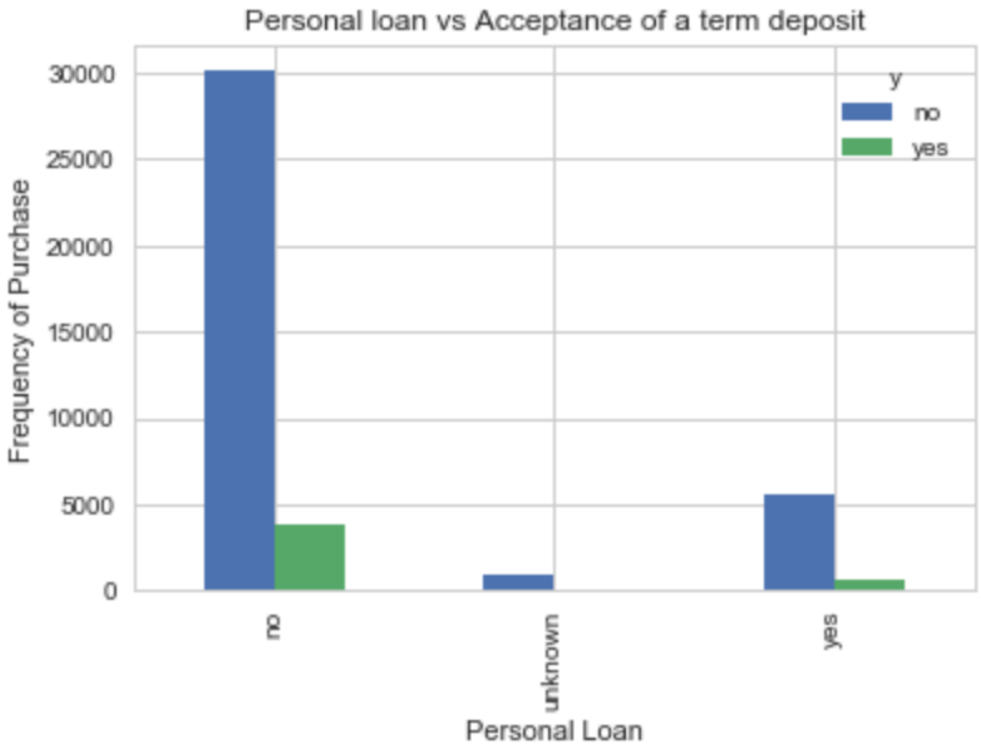


Figure 17

2.2.2.7 Month

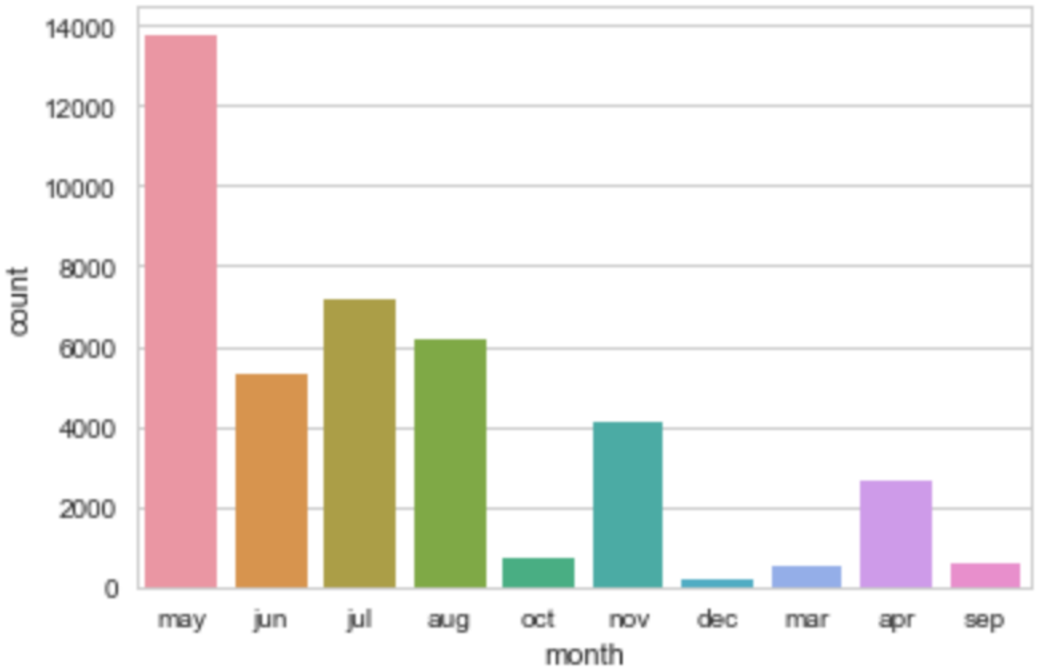


Figure 18

Most contacts to the customer were made in the month of May.

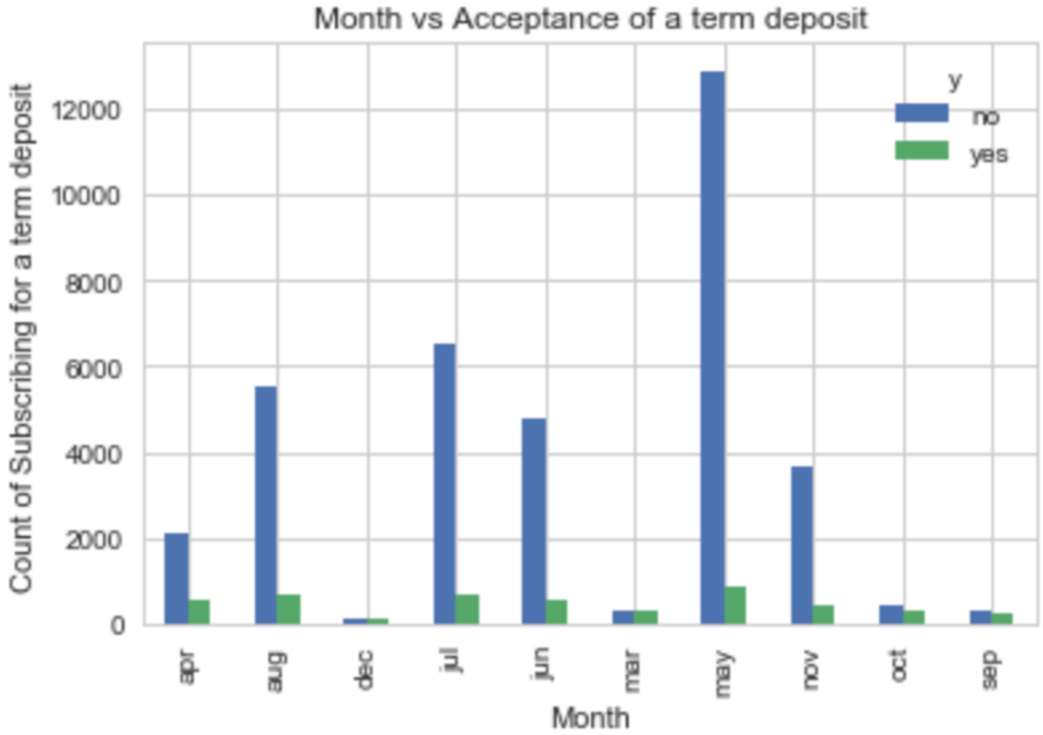


Figure 19

2.2.2.8 Day of the week

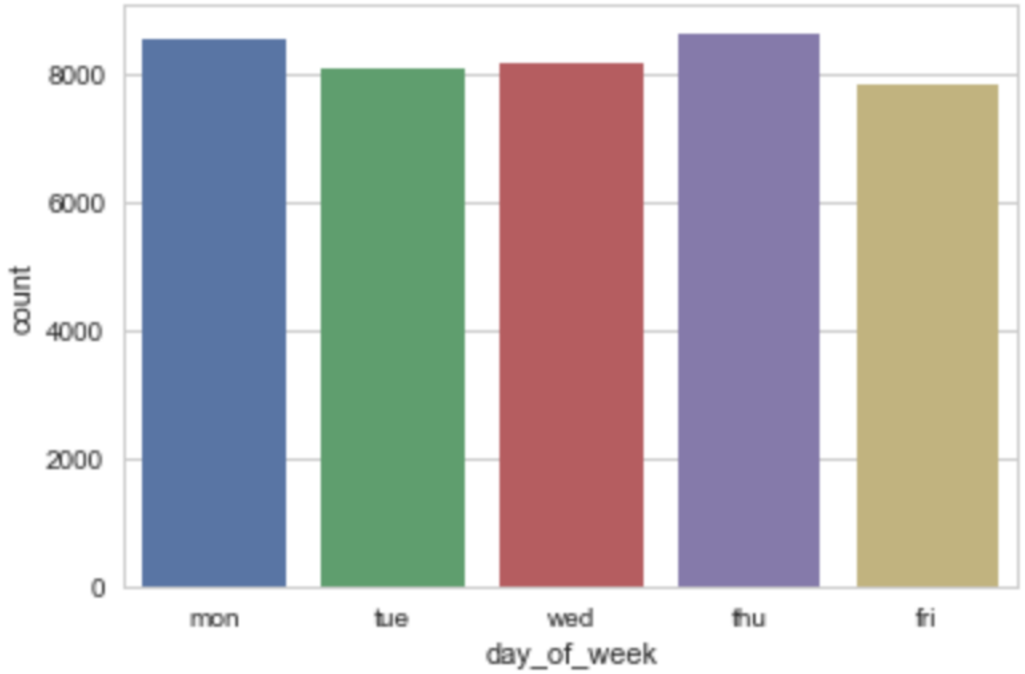


Figure 20

Distribution of the days that last customer contact was made.

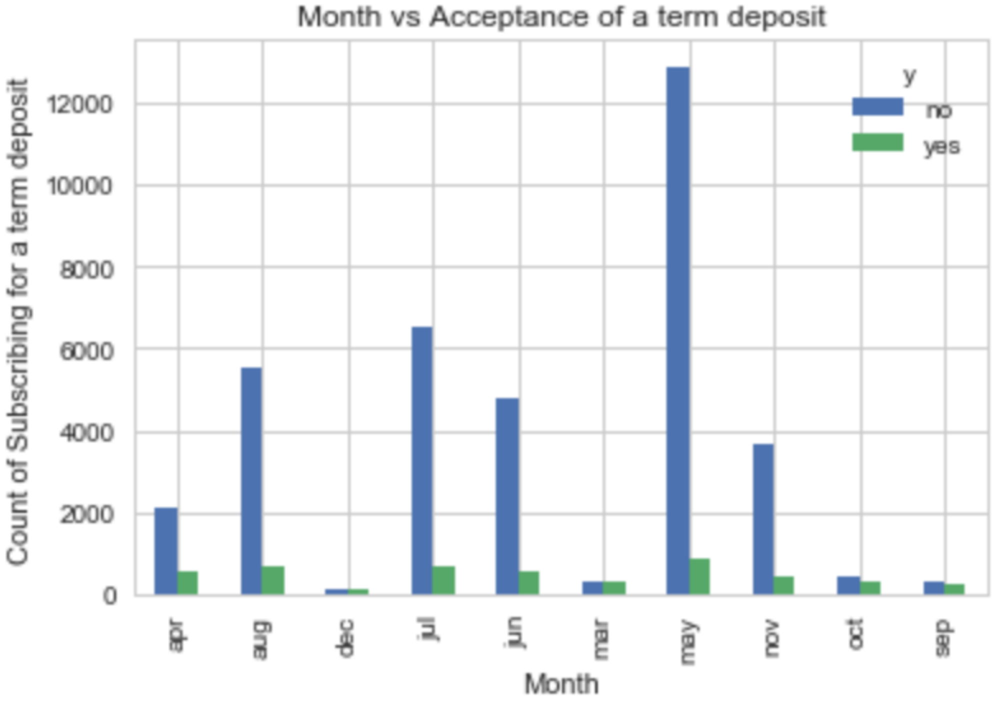


Figure 21

2.2.2.9 Outcome

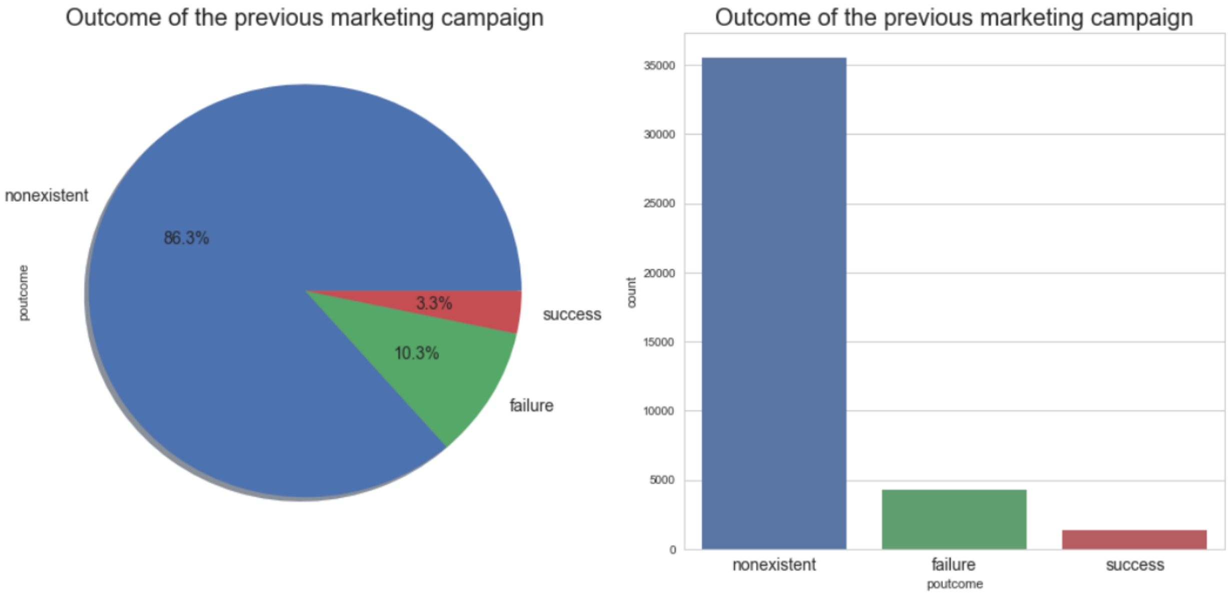


Figure 21

Outcome of the previous marketing campaign does not exist for majority of the customers i.e. 86%.

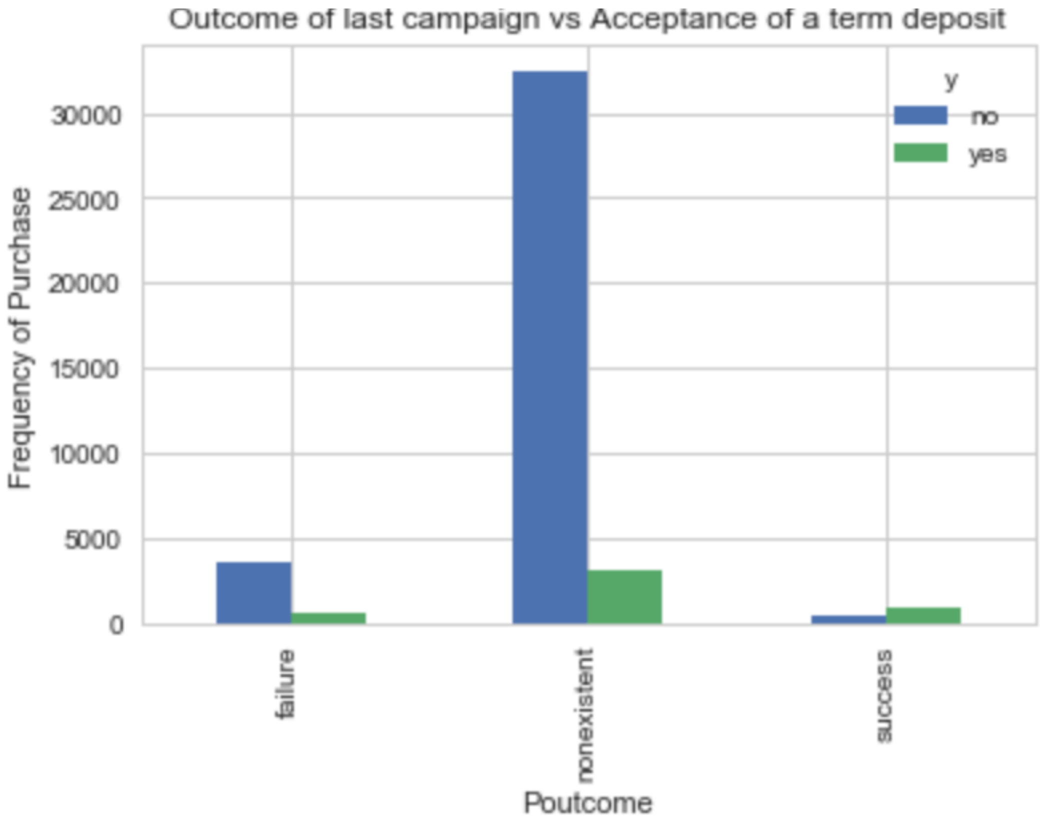


Figure 22

2.2.2.10 Output Variable

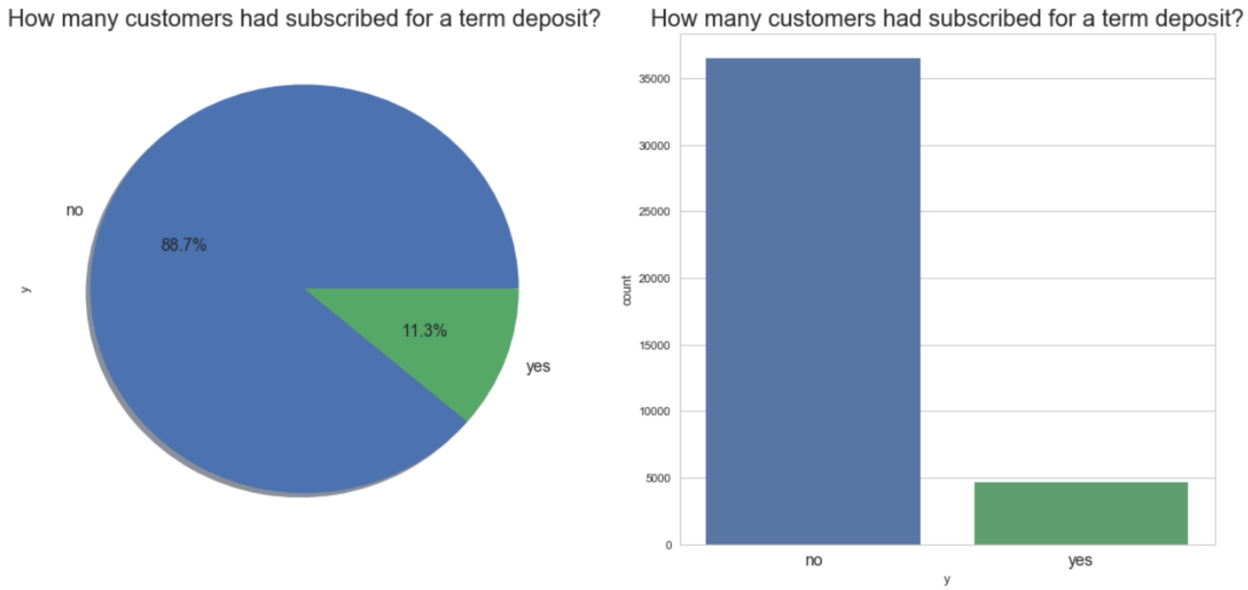


Figure 23

Majority of the customers have not actually subscribed to a term deposit.

2.2.2.11 Key Observations

My expectation is that following features are relatively more relevant for predicting if a customer will subscribe for a term deposit or not.

- Job - Customer's profession is also a good indicator for example students and unemployed are less likely to subscribe.
- Education - Better educated customer could be aware of importance of saving money and might be likely to open a term deposit.
- Marital Status - Married couples tend to have more dependencies and liabilities. It means they would also on focus term deposits which are less risky than equity market as instruments of investments.
- Default - I would assume that a customer in default will have less money at his/her disposal to open a term deposit.
- Housing & Personal Loan - From a bank's perspective these are good features to consider so that bundled offers can be curated for customers. This increases the probability that a customer might open a term deposit.
- Previous - From a bank's perspective outcome of the last campaign - is also a good feature to consider.

On the other hand, all the remaining features are not so relevant in my opinion. Notably attribute "duration" highly affects the output target (e.g., if duration=0 then y=no). Yet, the duration is not known before a call is performed. Also, after the end of the call, y is obviously known. Therefore, I will drop all these features from consideration.

3 Methodology

3.1 Data Preprocessing

3.1.1 Step 1

The education variable has got values such as basic.4y, basic.6y, basic.9y. It will be prudent to treat those as "basic". Thus, the education variable will have the following values.

- basic
- high.school
- illiterate
- professional.course
- university.degree
- unknown

3.1.2 Step 2

According to the aforementioned data exploration and my expectation of the key input variables, I will be dropping the below features from consideration.

- Age
- Contact
- Month
- Day of week
- Duration
- Campaign
- Pdays
- Previous
- emp.var.rate
- cons.price.idx
- cons.conf.idx
- euribor3m
- nremployed

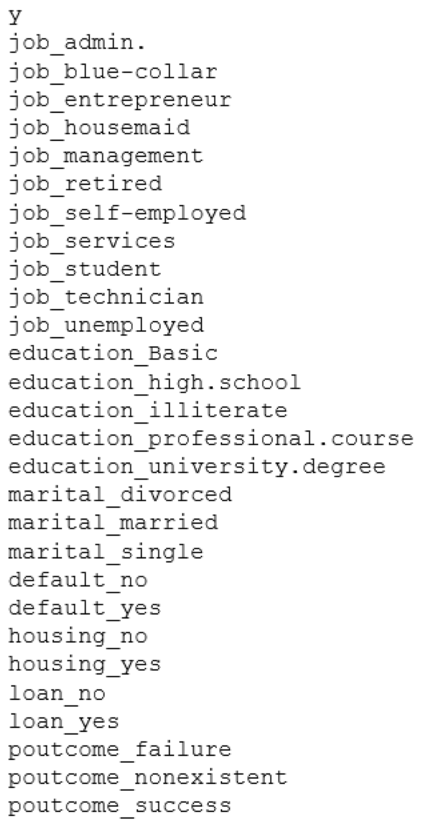
3.1.3 Step 3

In statistics, a dummy variable (also known as an indicator variable, binary variable) is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into mutually exclusive categories (such as smoker/non-smoker, etc.).[7]

In this step, I will create dummy variables for the features under consideration ('job','education', 'marital', 'default', 'housing', 'loan', 'poutcome')

3.1.4 Step 4

It will be good to drop columns with unknown values. So, I will drop the columns such as job_unknown. So, the final list of features will be.



```
y
job_admin.
job_blue-collar
job_entrepreneur
job_housemaid
job_management
job_retired
job_self-employed
job_services
job_student
job_technician
job_unemployed
education_Basic
education_high.school
education_illiterate
education_professional.course
education_university.degree
marital_divorced
marital_married
marital_single
default_no
default_yes
housing_no
housing_yes
loan_no
loan_yes
poutcome_failure
poutcome_nonexistent
poutcome_success
```

Figure 24

3.1.5 Step 5

In this step, we should convert the output variable 'y' values to 0 or 1 and datatype.

[7] https://en.wikipedia.org/wiki/Dummy_variable_(statistics)

3.2 Data Correlation

After plotting the panda's correlation, I did not observe any correlation between different variables. Below is an example visualization. As we see there are no correlations, I will not consider performing Principal component analysis (PCA).

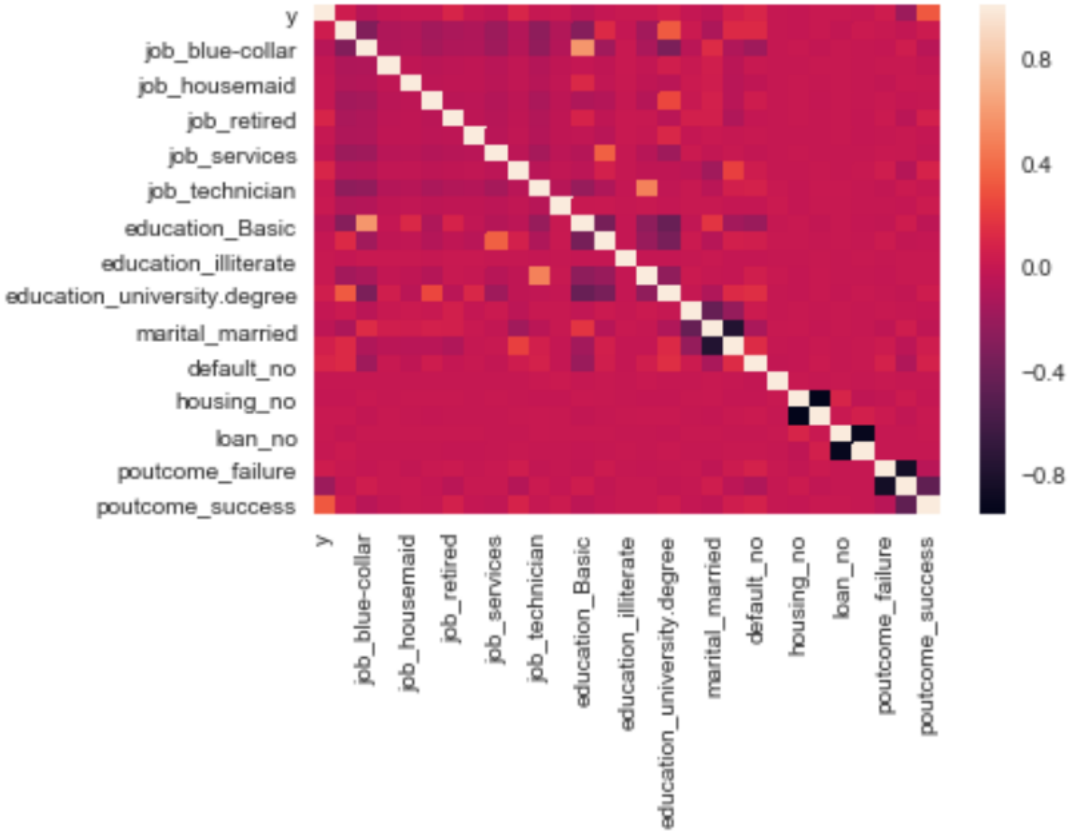


Figure 25

3.3 Implementation

3.3.1 Developing a Logistics Regression – The Benchmark Model

Logistic regression is mainly used in cases where the output is Boolean. So is the case under consideration, the desired target is either a customer subscribes for a term deposit or not. Here, we consider features and outcome Y which can take two values {0,1}. Logistic regression measures the relationship between an output variable Y (categorical) and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. So, logistics regression is my benchmark model.

3.3.1.1 Implementation Steps

Step 1: Split the training and test data in the ratio of 0.7 to 0.3 where 0.3 is test size. Thus, we obtain the following:

- Training set has 28831 samples
- Testing set has 12357 samples

Step 2: Fit the model to training data

Step 3: Predict the test results

Step 4: Plot confusion matrix

Step 5: Find the Accuracy of logistic regression classifier on test set

Step 6: Find the 10-fold cross validation average accuracy

Step 7: Compute classification report

Step 8: Compute ROC/AUC Score

3.3.1.2 Results of the benchmark model

3.3.1.2.1 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known.

		Condition positive	**Condition negative**
Predicted condition	Predicted condition positive	10818	150
	Predicted condition negative	1117	272

We can observe that there are 10818 + 272 correct predictions and 1117+150 incorrect predictions.

3.3.1.2.2 Accuracy of logistic regression classifier on test set

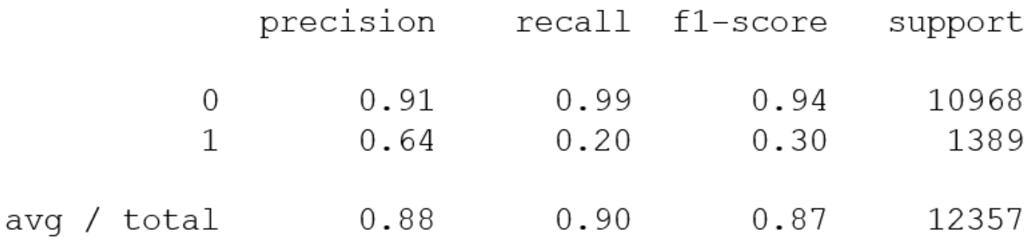
Accuracy of logistic regression classifier on test set is 0.90.

3.3.1.2.3 10-fold cross validation average accuracy

10-fold cross validation average accuracy is 0.898.

The average accuracy stays almost same as that of Logistic Regression model accuracy; hence, we can say that our model generalizes well.

3.3.1.2.4 Classification Report



```
             precision    recall  f1-score   support

          0       0.91      0.99      0.94     10968
          1       0.64      0.20      0.30      1389

avg / total       0.88      0.90      0.87     12357
```

Figure 26

3.3.1.2.5 ROC/AUC Score

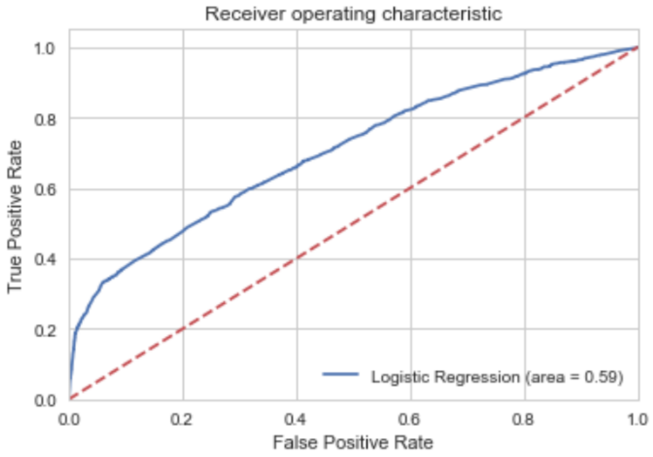


Figure 27

We have got the score as 0.775 which seems to be fair.[8]

3.3.2 Developing a Naïve Bayes Model – The Solution Model

Naive Bayes classifiers have worked very well in many practical situations, famously known spam filtering. Naive Bayes also assumes that the features are conditionally independent.

The Naive Bayes algorithm is called "naive" because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features. For instance, if you are trying to identify a fruit based on its color, shape, and taste, then an orange colored, spherical, and tangy fruit would most likely be an orange. Even if these features depend on each other or on the presence of the other features, all of these properties individually contribute to the probability that this fruit is an orange and that is why it is known as "naive." As for the “Bayes” part, it refers to the statistician and philosopher, Thomas Bayes and the theorem named after him, Bayes' theorem, which is the base for Naive Bayes Algorithm.[9]

3.3.2.1 Implementation Steps

Steps followed are same the ones given in earlier section 3.3.1.1. However, the classifier used is of type **GaussianNB**.

[8] http://gim.unmc.edu/dxtests/roc3.htm

[9] https://www.hackerearth.com/blog/machine-learning/introduction-naive-bayes-algorithm-codes-python-r/

3.3.2.2 Results of the solution model with GaussianNB

3.3.2.2.1 Confusion Matrix

		Condition positive	**Condition negative**
Predicted condition	Predicted condition positive	4366	6628
	Predicted condition negative	273	1090

We can observe that there are 4366+1090 correct predictions and 6628+273 incorrect predictions.

3.3.2.2.2 Accuracy of the classifier on test set

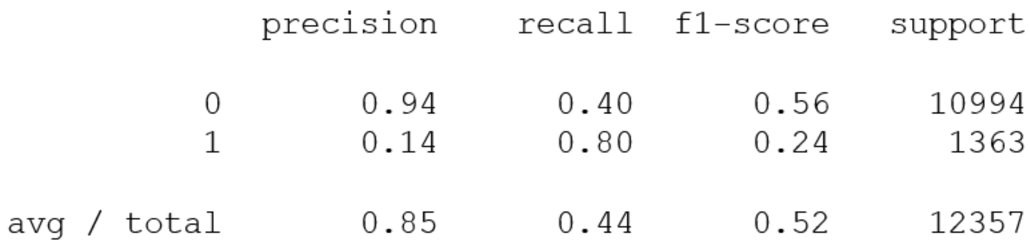
Accuracy of logistic regression classifier on test set is 0.441.

3.3.2.2.3 10-fold cross validation average accuracy

10-fold cross validation average accuracy is 0.444.

The average accuracy stays almost same.

3.3.2.2.4 Classification Report



```
             precision    recall  f1-score   support

          0       0.94      0.40      0.56     10994
          1       0.14      0.80      0.24      1363

avg / total       0.85      0.44      0.52     12357
```

Figure 28

3.3.2.2.5 ROC/AUC Score

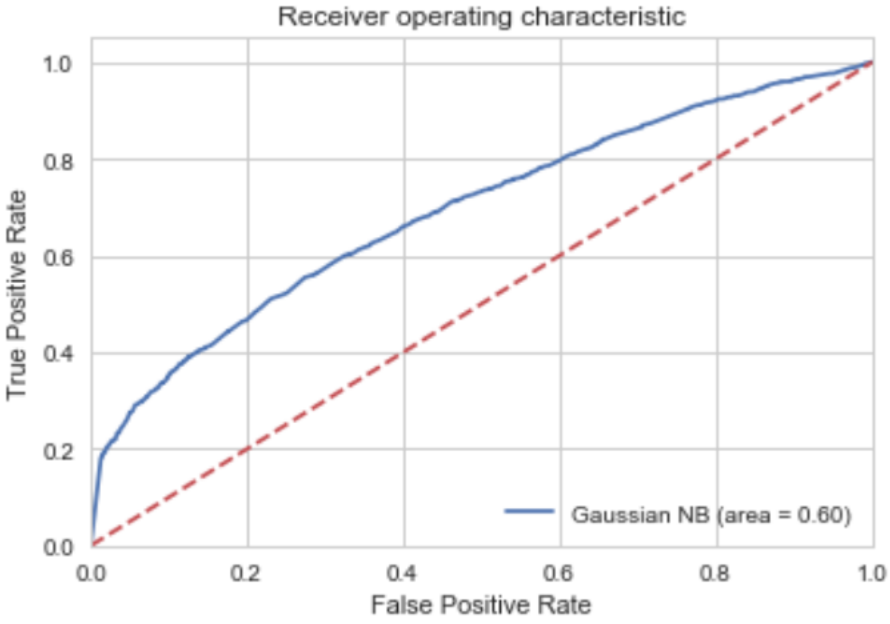


Figure 29

We have got the score as 0.541 which seems to be poor.[10]

3.4 Improvement

3.4.1 Developing a Naïve Bayes Model – The Solution Model (Part II)

It seems that the accuracy of the model with GaussianNB is not up to the mark. BernoulliNB is designed for binary/boolean features. As we have samples represented as binary-valued features it will be interesting to model naive bayes classifier using BernoulliNB. Let me try this.

3.4.1.1 Implementation Steps

Steps followed are same the ones given in earlier section 3.3.1.1. However, the classifier used is of type **BernoulliNB**.

[10] http://gim.unmc.edu/dxtests/roc3.htm

3.4.1.2 *Results of the solution model with BernoulliNB*

3.4.1.2.1 Confusion Matrix

		Condition positive	**Condition negative**
Predicted condition	Predicted condition positive	10865	263
	Predicted condition negative	1112	297

We can observe that there are 1086+297 correct predictions and 1112+263 incorrect predictions.

3.4.1.2.2 Accuracy of the classifier on test set

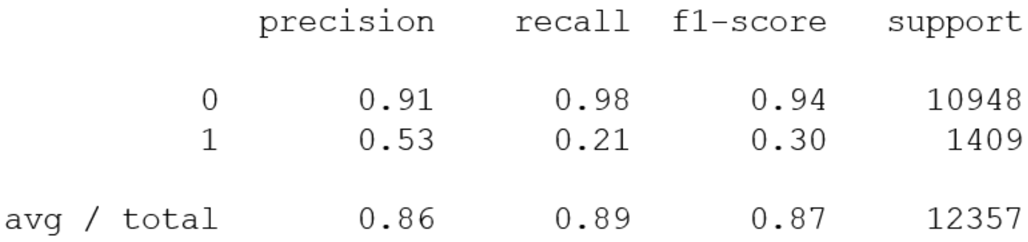
Accuracy of logistic regression classifier on test set is 0.888.

3.4.1.2.3 10-fold cross validation average accuracy

10-fold cross validation average accuracy is 0889.

The average accuracy stays almost same using 10-fold cross validation; hence, we can say that our model generalizes well.

3.4.1.2.4 Classification Report



```
             precision    recall  f1-score   support

          0       0.91      0.98      0.94     10948
          1       0.53      0.21      0.30      1409

avg / total       0.86      0.89      0.87     12357
```

Figure 30

3.4.1.2.5 ROC/AUC Score

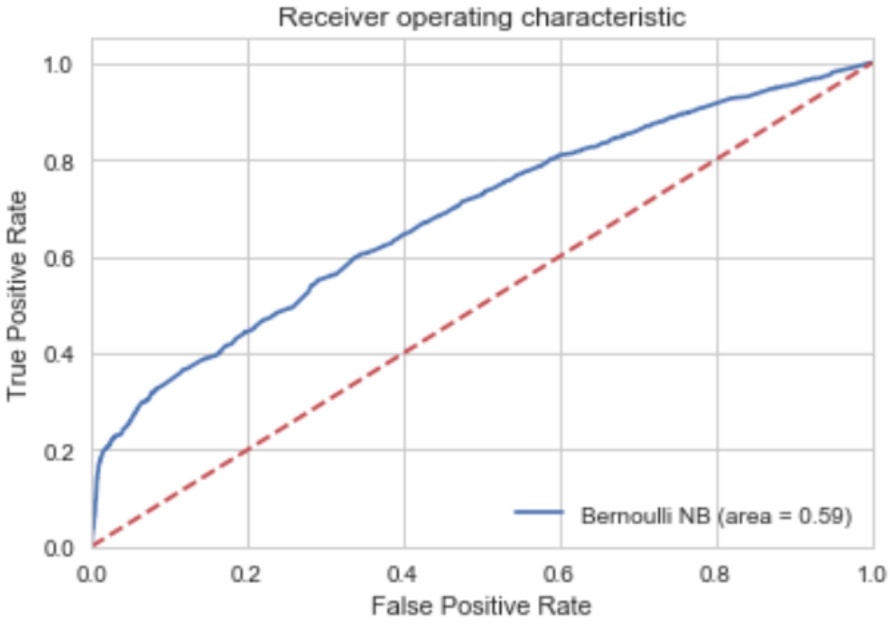


Figure 31

We have got the score as 0.718 which seems to be fair.[11]

[11] http://gim.unmc.edu/dxtests/roc3.htm

3.5 Implementation - Extracting Feature Importance

I chose a scikit-learn supervised learning algorithm RandomForestClassifier that has a feature*importance* attribute available for it.[12] This attribute is a function that ranks the importance of each feature when making predictions based on the chosen algorithm.

Steps:

- Import supplementary visualization code *visuals.py*
- Import a supervised learning model that has *'feature_importances_'*
- Train the supervised model on the training set using *.fit(X_train, y_train)*
- Extract the feature importance using. *feature_importances_*
- Plot

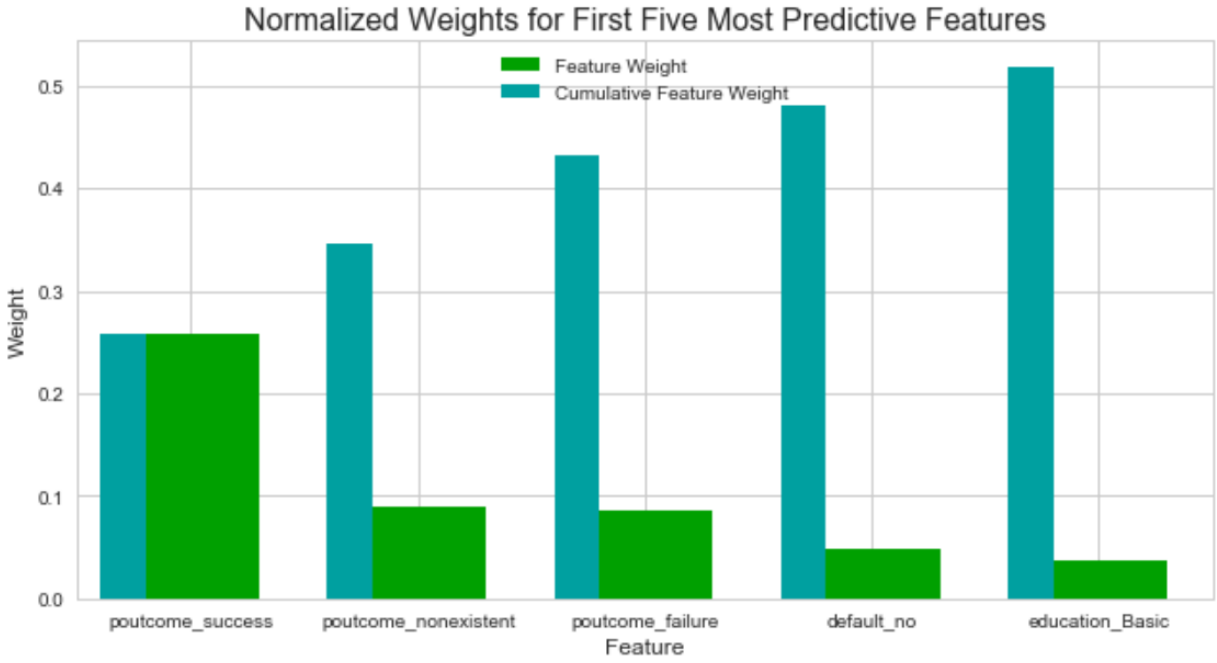


Figure 32

3.6 Programming Tips

- It is important to understand does the data match the column label?
- In this project, one key that should not be missed is to create dummy variables. As previously explained, Dummy variables are used as devices to sort data into mutually exclusive categories.
- With regards to variable values for "Education" it was also worthwhile to consolidate basic.4y, basic.6y,basic.9y as basic only.

[12] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

- We had several values of variables as unknown and unknown values do not help much in analysis. So after the dummy variables were created it was also crucial to drop unknown variables.
- Output variable had values such as yes or no. Hence, the output variable 'y' values were converted to 0 or 1 and also the datatype.
- As previously explained, I did not use PCA in this case due to non-existing correlations.

Having a neat & clean data is key for further exploration and visualization.

4 Conclusion

4.1 Result Summary

Comparison Parameter	*Logistics Regression*	*Naive Bayes (Gaussian)*	*Naive Bayes(Bernoulli)*
Avg. accuracy	0.90	0.441	0.888
10-fold CV avg. accuracy	0.898	0.444	0.889
ROC/AUC Score	0.775	0.542	0.71
f1-score (avg/total)	0.87	0.52	0.87
wall time	145 ms	13 ms	7.98 ms

4.2 Key Remarks

Logistics Regression was the benchmark model of this project while Naive Bayes was the solution model. As we see above comparison, Naive Bayes using Bernoulli came very closer to the benchmark model when we consider accuracy/auc score and f1-score. However, when it came to wall time Naive Bayes using Bernoulli was the fastest. This is in line with the theory where I learned the most appealing strength of Naive Bayes model is that it is relatively faster.

I believe the final solution model using BernoulliNB good enough to apply to the given problem statement which is to predict which customers will be subscribe to a term deposit or not.

Another advantage that can be gained using Naïve Bayes is we would require less training data.

4.3 Feature Importance

Although the focus of this project was on Logistics Regression and Naive Bayes for classification, I took the liberty to try supervised learning algorithm Random Forest Classifier that has a feature importance attribute available. The result of this suggests that below are important features (ordered per weight)

- Outcome of the previous marketing campaign (success/nonexistent/failure)
- If the customer has credit in default?
- Customer's education.

4.4 Recommendations to the Bank

I would like to make the following recommendations to the bank:

- Ensure that customers are not annoyed with frequent calls. Therefore, it is important to keep the number of contacts per customer to an optimum level. This should be supplemented with the bank employee doing his/her homework/preparation about the prospective customer very well (before making a customer contact).
- Prioritize to first tap customers who are married and single.
- With regards to customer's profession it is good to first start with retired people followed by customers in the profession such as admin, services, blue-collar.
- Make sure that not much time is elapsed after a first contact with a customer is made.
- Contacting first the customers who are not defaulters is a common banking sense that should prevail.
- It will be great that bank's marketing department thinks about having some interesting offers for existing customers with home loan or personal loan e.g. offering 0.25% higher rate of return than usual.
- Bank's direct marketing department then should contact those customers where the previous campaign was successful.

In my opinion, this is a good prioritization and optimization of direct marketing efforts to increase probability of customers opening a term deposit.

4.5 Reflection on my learning

The first challenge was to understand the initial problem of the Portuguese bank and the concept of direct marketing. The next was finding relevant public datasets. Once the relevant public dataset was found, the data was downloaded and preprocessed. A benchmark model was created for the classifier using logistics regression. The classifier was trained using the data, splitting dataset in the ratio of 0.7:03. A 10-fold cross validation was also performed on the data.

Subsequently, a solution model was created for the classifier. The classifier was trained using the data, splitting dataset in the ratio of 0.7:0.3. A 10-fold cross validation was also performed on the data.

I also took the liberty to use feature importance attribute of the random forest classifier.

It was also important for me to make the qualitative recommendations for the bank to optimize and prioritize the direct marketing efforts.

4.6 Future Work/Improvement

I tried the most commonly used classifiers such as logistics regression and naïve bayes where each one had some pros and cons. Notably, Naive Bayes can learn individual features importance but can't determine the relationship among features.

Outside the scope of this project, I intend to also try using neural network for predictions. An Artificial Neural Network with an appropriate network structure can handle the correlation/dependence between input variables.[13]

Notably, neural network is growing in popularity. Google recently announced it used neural networks to cut their data center cooling costs by a whopping 40%.[14]

5 References

https://medium.com/@sangha_deb/naive-bayes-vs-logistic-regression-a319b07a5d4c

http://scikit-learn.org/stable/modules/naive_bayes.html

https://en.wikipedia.org/wiki/Direct_marketing

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

https://pdfs.semanticscholar.org/a911/cbe221347b400d1376330591973bb561ff3a.pdf

https://pdfs.semanticscholar.org/cab8/6052882d126d43f72108c6cb41b295cc8a9e.pdf

https://en.wikipedia.org/wiki/Dummy_variable_(statistics)

http://gim.unmc.edu/dxtests/roc3.htm

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://www.quora.com/When-should-I-use-Naive-Bayes-classifier-over-neural-networks

https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

[13] https://www.quora.com/When-should-I-use-Naive-Bayes-classifier-over-neural-networks

[14] https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/