Contents

Executive Summary

This project was taken for building up a deeper understanding of:

- Figuring Out The Variables From The Components Of Problem Statement
- Using Probit And Logit Techniques
- Checking Model Accuracy, Showing The Same On Roc Curve

The problem statement which is being addressed is obtaining to find the best strategies in order to improve for the next marketing campaign ,so that the financial institution have a greater effectiveness for future marketing campaigns.
In order to achieve this we have analyzed the last marketing campaign, the bank performed. Also, identifying the patterns have helped us in finding the conclusions in order to develop future strategies.

Background

As we know marketing is process by which companies create value for customers and build strong customer relationships in order to capture value from customers in return and Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. In order to build a sound marketing strategy for the bank, following factors need to be considered-

Segment of the Population: To which segment of the population is the marketing campaign going to address and why? This aspect of the marketing campaign is extremely important since it will tell to which part of the population should most likely receive the message of the marketing campaign.

As we'll move on further we will discover the main factors and find out customer segments, using data for customers, who subscribed to term deposit and we'll find out the more potential customers as we develop more targeted marketing campaigns.

Problem Statement

The objective of the project is to study the dataset, extract information about next marketing campaign of a financial institution. We need to analyse it in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

Methodology

In order to optimize marketing campaigns with the help of the dataset, we will have to take the following steps:

1. Import data from dataset and perform initial high-level analysis: look at the number of rows, look at the missing values, look at dataset columns and their values respective to the campaign outcome.

2. Clean the data: remove irrelevant columns, deal with missing and incorrect values, turn categorical columns into dummy variables.
3. Use statistical techniques to predict the marketing campaign outcome and to find out factors, which affect the success of the campaign.

Data Source

https://www.kaggle.com/janiobachmann/bank-marketing-dataset/

```
In [3]: df.head()
```

Out[3]:

	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

```
In [4]: df.describe()
```

Out[4]:

	age	balance	day	duration	campaign	pdays	previous
count	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000
mean	41.231948	1528.538524	15.658036	371.993818	2.508421	51.330407	0.832557
std	11.913369	3225.413326	8.420740	347.128386	2.722077	108.758282	2.292007
min	18.000000	-6847.000000	1.000000	2.000000	1.000000	-1.000000	0.000000
25%	32.000000	122.000000	8.000000	138.000000	1.000000	-1.000000	0.000000
50%	39.000000	550.000000	15.000000	255.000000	2.000000	-1.000000	0.000000
75%	49.000000	1708.000000	22.000000	496.000000	3.000000	20.750000	1.000000
max	95.000000	81204.000000	31.000000	3881.000000	63.000000	854.000000	58.000000

Introduction

This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives information about a marketing campaign of a financial institution, which we have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

In general, this marketing data can be used for 2 different business goals:

1. Prediction of the results of the marketing campaign for each customer and clarification of factors which affect the campaign results. This helps to find out the ways how to make marketing campaigns more efficient.
2. Finding out customer segments, using data for customers, who subscribed to term deposit. This helps to identify the profile of a customer, who is more likely to acquire the product and develop more targeted marketing campaigns

Variables

Dependent variable (Y):

- **Deposit** - Has The Client Subscribed A Term Deposit? Binary: 'Yes','No'

Independent variables (X):

- **Age:** Numeric, Shows Age Of Customer Contacted
- **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'; Note: 'Divorced' Means Divorced Or Widowed
- **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 Month ,Numeric: 1,2,3,....29,30
- **Duration:** Last Contact Duration, In Seconds, Numeric.
- **Campaign:** Number Of Contacts Performed During This Campaign And For This Client ,Numeric, Includes Last Contact

Descriptive Statistics

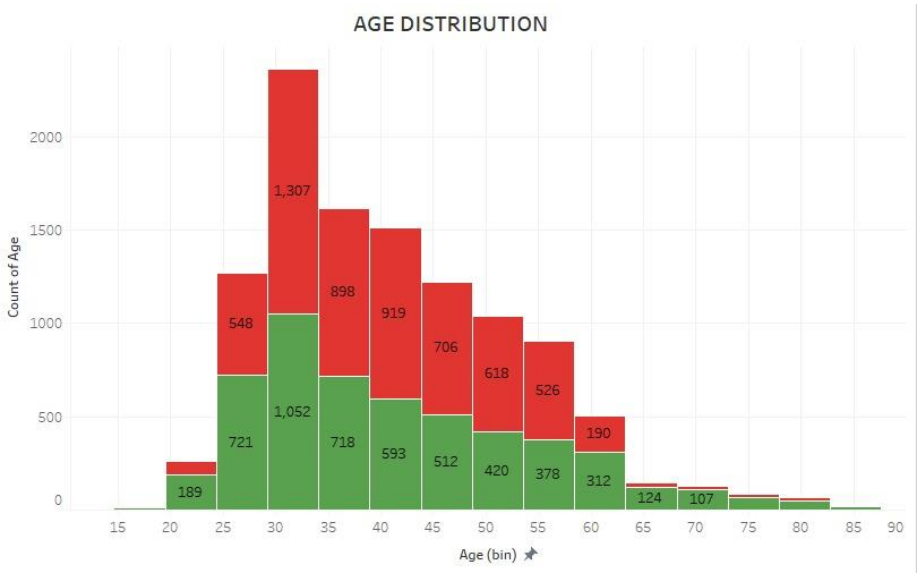
Descriptive statistics provide simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries form the basis of the initial description of the Medical insurance premium prediction data here.

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables, kurtosis ,quartilesand skewness.

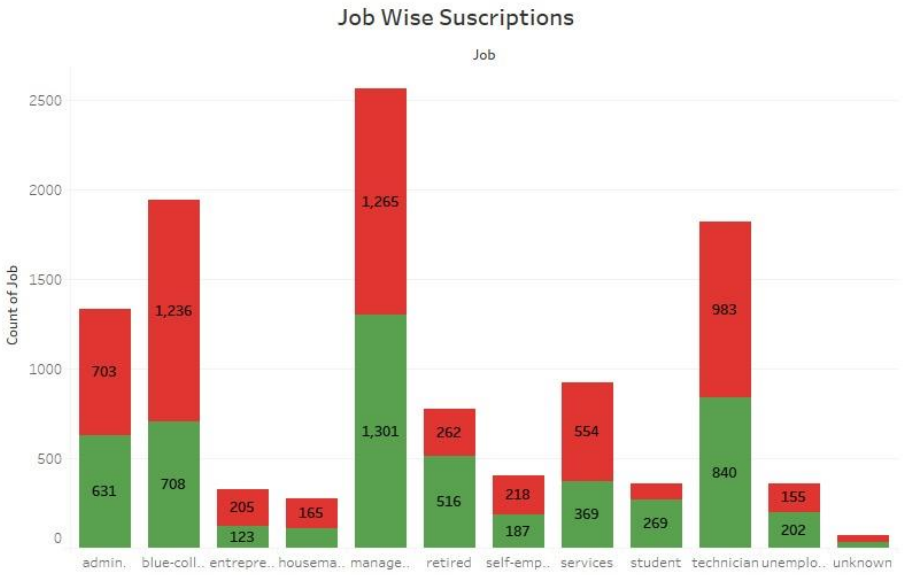
	age	balance	day	duration	campaign	pdays	previous
count	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000
mean	41.231948	1528.538524	15.658036	371.993818	2.508421	51.330407	0.832557
std	11.913369	3225.413326	8.420740	347.128386	2.722077	108.758282	2.292007
min	18.000000	-6847.000000	1.000000	2.000000	1.000000	-1.000000	0.000000
25%	32.000000	122.000000	8.000000	138.000000	1.000000	-1.000000	0.000000
50%	39.000000	550.000000	15.000000	255.000000	2.000000	-1.000000	0.000000
75%	49.000000	1708.000000	22.000000	496.000000	3.000000	20.750000	1.000000
max	95.000000	81204.000000	31.000000	3881.000000	63.000000	854.000000	58.000000

Distributions of independent variables

Distribution across Age

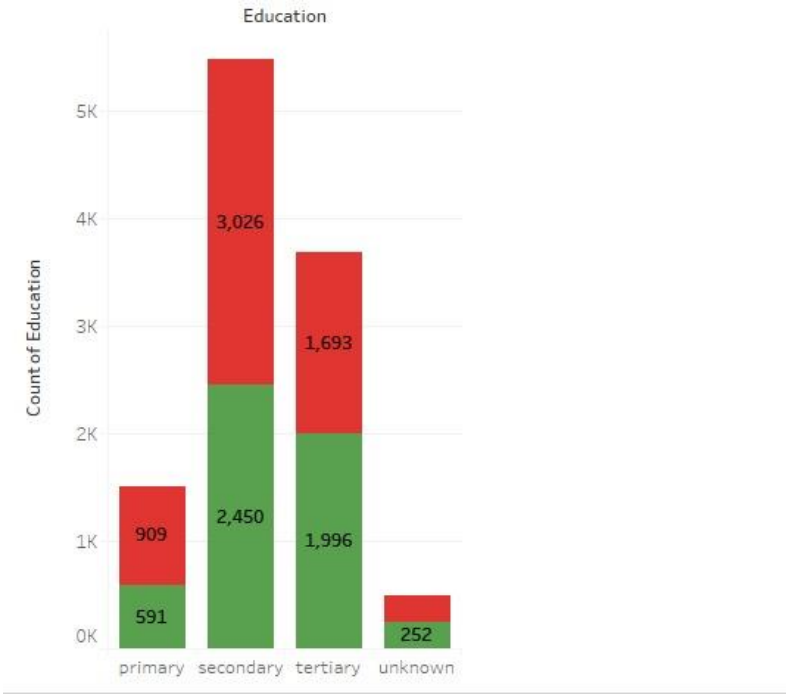


Distribution across Jobs

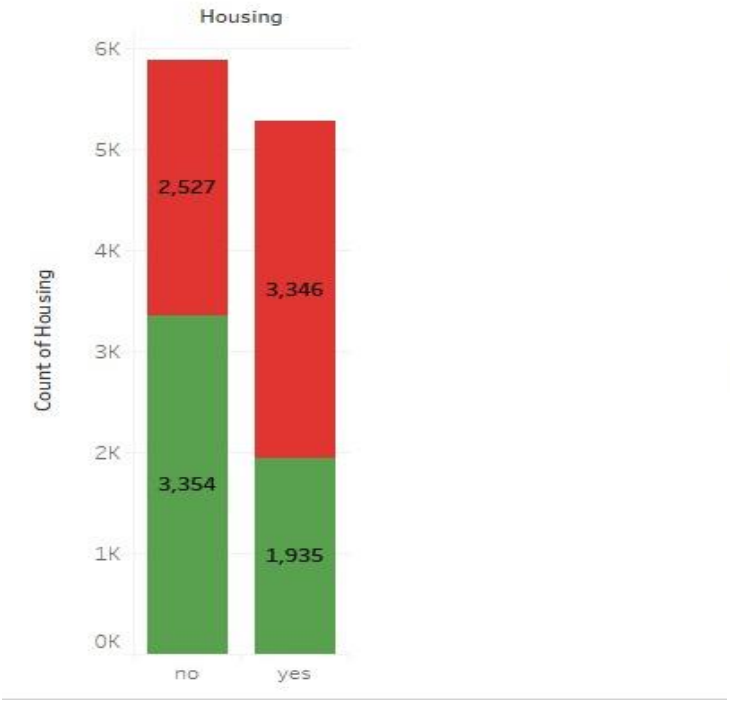


Distribution across Marital Status

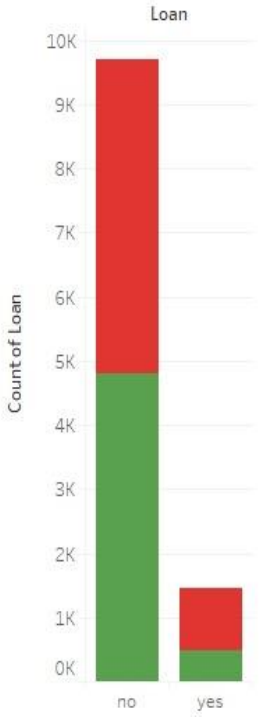
Distribution across Education



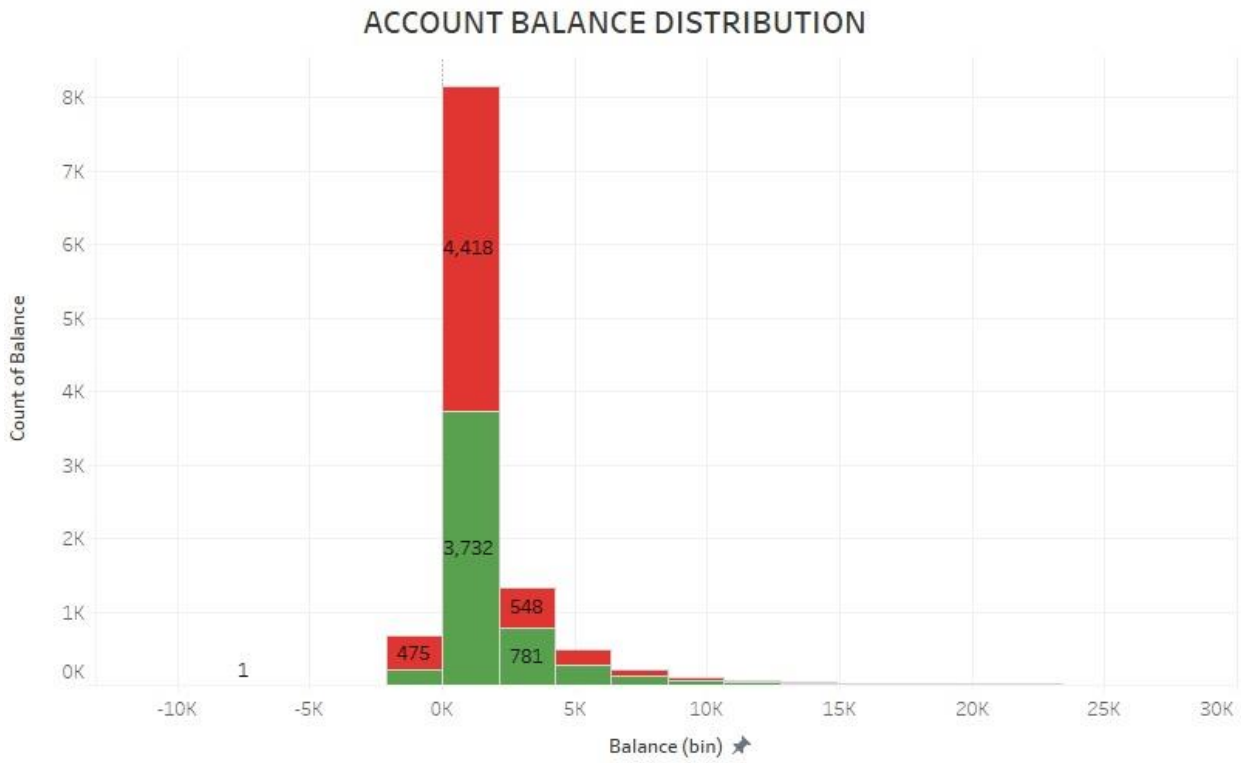
Distribution across Housing Loan



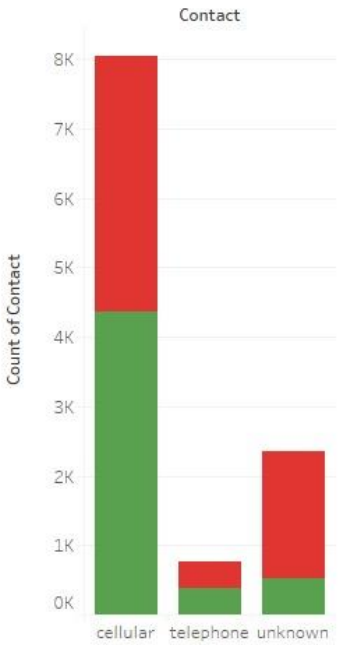
Distribution across Personal Loan



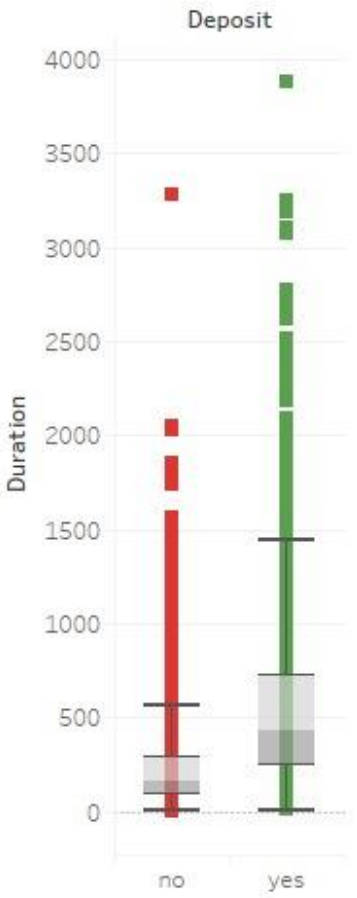
Distribution across Account Balance



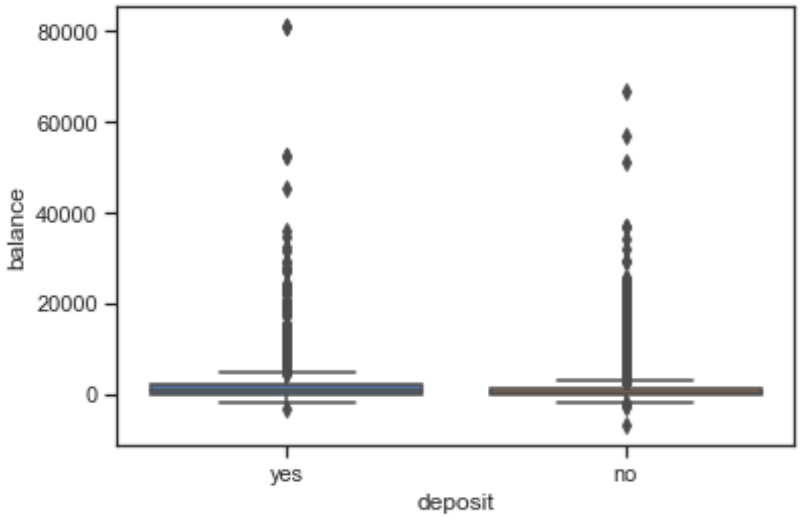
Distribution across Medium of Contact



Box plot for duration attribute



Box plot for balance attribute



Dummy variable trap

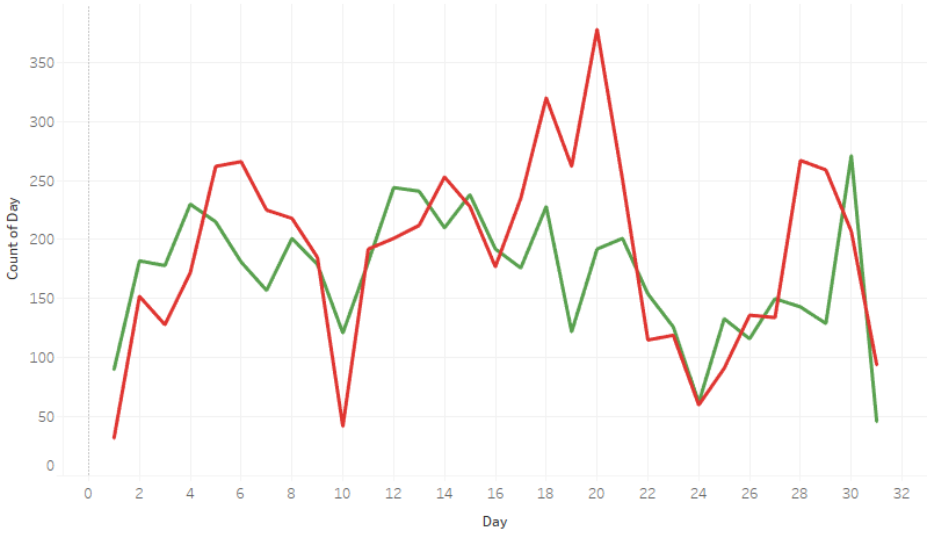
Dummy variables are required to represent a categorical variables the number of dummy variables depends on the number of values that particular categorical variable can take. If we have to represent a categorical variable that can take N different values, we need to define N - 1 dummy variables.

Here in this dataset we have attributes like Job, Month, Day of a month, Education and Marital status which are categorical in nature, In order to include this attribute in the our model we need to encode this to a numerical values. While converting this categorical attribute to numerical attribute we need to take care of dummy variable trap issue.

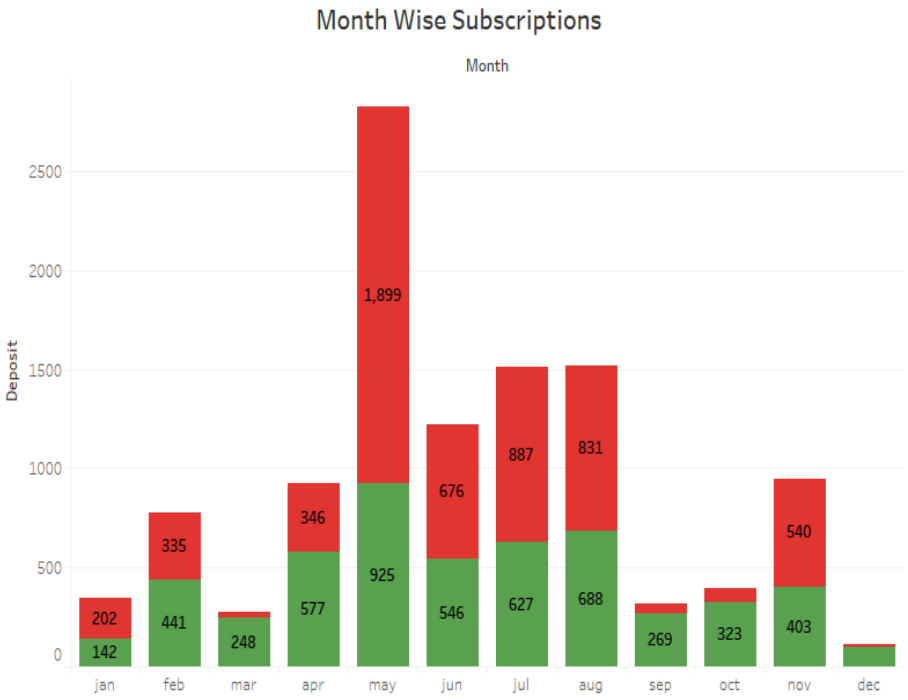
When dummy variables are defined, we need to be careful or else we might end up defining too many variables. If a particular categorical variable takes on N values, it is highly possible that we may define N dummy variables. If we define N dummy variables then we will end up in this trap called Dummy variable trap. This could lead us to linear dependence between these variables so you only need N - 1 dummy variables.

A Nth dummy variable is redundant as it carries no new information. And it creates a severe multicollinearity problem for the Regression analysis. In this dataset we have removed one dimension from each categorical to overcome this problem.

Additional Variables:



From the above plot it is visible that during the first half of the month subscription behavior is different from second half of the month, so a new 0-1 labelled attribute “First half of the month” was created.



The above plot shows the month wise campaign attempts and conversions , It is observed from the above plot from Jan – April and from Sep – Dec, the number of attempts are low when compared to May – Aug , so 2 new 0-1 labelled attribute “season 1” and “season 2” were created. All these variables were encoded and the final number of variables turned out to be 29.

Correlation Matrix and Multicollinearity

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables is linearly related.

In statistical modeling, correlation matrices represent the relationships between variables. Also we need to be aware that correlation does not imply causation. In our problem of predicting deposit subscription we are making use of the correlation matrix to identify the independent variable which has Close relationship with the dependent variable moreover this correlation matrix also help us to identify the multicollinearity associated with the independent variables themselves.

Correlation between dependent and independent variables

Housing	-0.203888
Campaign	-0.128081
Loan	-0.110580
Job_blue-collar	-0.100840
Day	-0.056326
Job_services	-0.044531
Default	-0.040680
Job_entrepreneur	-0.034443
Job_housemaid	-0.024155
Job_technician	-0.011557
Job_self-employed	-0.004707
Job_admin	-0.000610
Job_umemployed	0.033487
Age	0.034901
Job_management	0.036301
Balance	0.081129
Job_student	0.099953
Job_retired	0.103827
Previous	0.139867
Pdays	0.151593
Duration	0.451919
Deposit	1.000000

Most Correlated Independent Variables from Correlation Matrix:

Housing	-0.203888
Campaign	-0.128081
Loan	-0.110580
Job_blue-collar	-0.100840
Default	-0.040680
Job_student	0.099953
Job_retired	0.103827
Previous	0.139867
Pdays	0.151593
Duration	0.451919
Deposit	1.000000

Multicollinearity:

One way to measure multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated.

	VIF Factor	features
0	23.5	Age
1	16.4	Education
2	1.0	Default
3	1.3	Balance
4	2.4	Housing
5	1.2	Loan
6	2.9	Duration
7	2.0	Campaign
8	1.8	Pdays
9	1.6	Previous
10	3.0	Deposit
11	13.7	Job_management
12	7.9	Job_blue-collar
13	8.5	Job_technician
14	6.2	Job_admin
15	4.4	Job_services
16	5.6	Job_retired
17	2.8	Job_self-employed
18	2.1	Job_student
19	2.4	Job_umemployed
20	2.4	Job_entrepreneur
21	2.1	Job_housemaid
22	3.8	Married
23	1.6	divorced
24	5.7	Contact_cellular
25	1.5	Contact_telephone
26	2.3	Season_1
27	5.8	Season_2
28	2.1	First half of month

Interpreting VIF:

A VIF of 1 means that there is no correlation among the *jth* predictor and the remaining predictor variables, and hence the variance is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

From the above chart it can be seen that there are high VIF values for education and some of the jobs. This is obvious that education will be positively correlated with the job. Hence to avoid the issue of multicollinearity we haven't included jobs and education together in any of the models.

Omitted Variable Bias

Omitted variable bias occurs when two conditions are true-

I. When the omitted variable is correlated with the included independent variable and
II. When the omitted variable is a determinant of the dependent variable.

The omitted variables for the given problem statement are as follows-

- Other Similar deposits
- Other Investments
- Taxable income
- Spending Habit

Summary of Models

We fitted regression model between a dependent variable and independent variables. We formulated 4 different models to understand the relationship of deposit subscription with other factors like Housing loan, Age, Account balance , Job, Education ,Marital status etc. The list of models formulated in this study are listed below:

Including duration Variable:

Two models have been trained with the duration variable and the other two without duration variable.

Model 1:

Using Probit:

Deposit = -0.6171 (Housing)+ 0.0025 (Duration) +0.6500 (Contact_cellular) - 0.0834 (Campaign) -0.9548

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9548	0.040	-24.061	0.000	-1.033	-0.877
Housing	-0.6171	0.028	-22.194	0.000	-0.672	-0.563
Duration	0.0025	5.31e-05	47.242	0.000	0.002	0.003
Campaign	-0.0834	0.007	-12.504	0.000	-0.097	-0.070
Contact_cellular	0.6500	0.032	20.098	0.000	0.587	0.713

From probit modelling technique, the confidence intervals obtained after t test for all the independent variables does not contain zero we reject the null hypothesis (Reduced model explains the variance better). And hence we conclude that there is a relationship between the independent and dependent variables. And from the summary we can see that all the p values are very much less than our significance level, Hence we conclude that all the variables included are statistically significant.

Using Logit:

Deposit = -1.0864*(Housing)+ 0.0048*(Duration) +1.1093* (Contact_cellular) -0.1509*(Campaign) -1.7073(const)

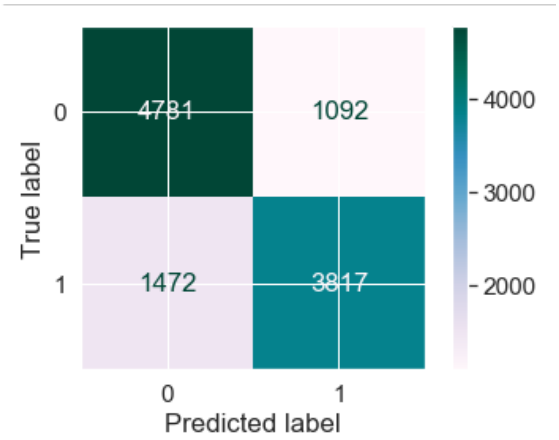
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7073	0.071	-23.941	0.000	-1.847	-1.568
Housing	-1.0864	0.048	-22.433	0.000	-1.181	-0.991
Duration	0.0048	0.000	42.699	0.000	0.005	0.005
Campaign	-0.1509	0.012	-12.162	0.000	-0.175	-0.127
Contact_cellular	1.1093	0.057	19.437	0.000	0.997	1.221

From Logit modeling technique, the confidence intervals obtained after t test for all the independent variables does not contain zero we reject the null hypothesis (Reduced model explains the variance better). And hence we conclude that there is a relationship between the independent and dependent variables. And from the summary we can see that all the p values are very much less than our significance level, Hence we conclude that all the variables included are statistically significant.

Performance Metrics:

Confusion Matrix:

In predictive analytics, a table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than mere proportion of correct classifications (accuracy).

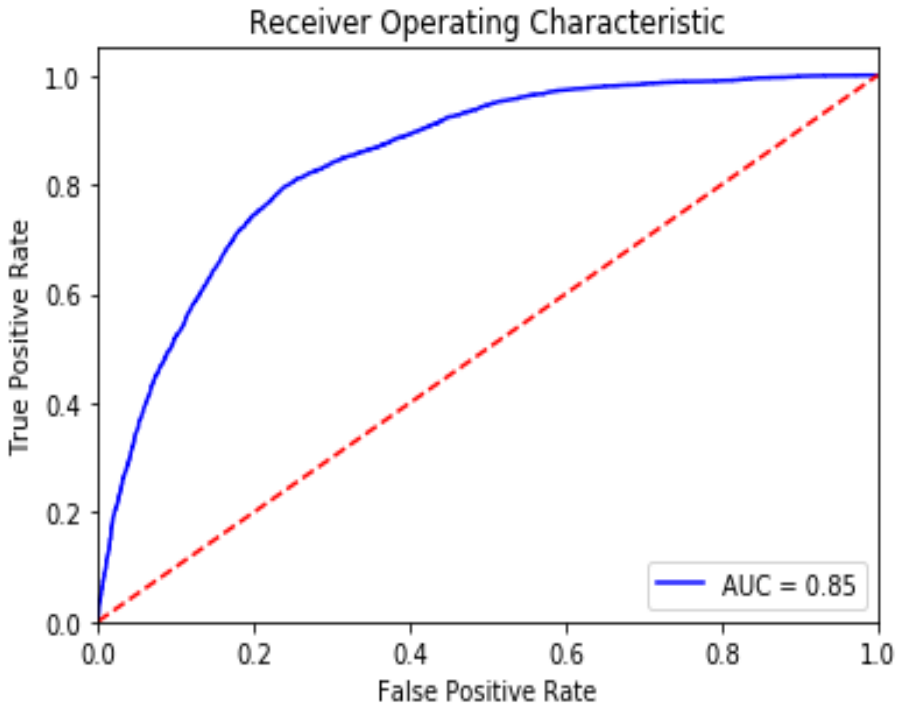


ROC Curve:

There are four possible outcomes for a binary dependent variable problem. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP); however if the actual value is n then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are n, and false negative (FN) is when the prediction outcome is n while the actual value is p.

To draw an ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

An ROC space is defined by FPR and TPR as x and y axes, respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.



Excluding duration Variable:

In our previous model, one of the independent variable is duration but the problem with this variable is duration is not known before a call is performed moreover after the end of the call duration in seconds is obviously known. Thus, this input should only be included only for explaining the variance and should be ignored if the intention is to have a realistic predictive model

Model 2:

Using Probit:

Deposit = 0.0102 + 0.0743 (Education) – 0.0500 (Campaign) – 0.4121 (Housing) + 0.4342 (Contact Cellular) + 0.2011 (First Half of the month) – 0.1785 (Married) – 0.3775(Season_2)

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0102	0.056	0.182	0.856	-0.100	0.120
Education	0.0743	0.018	4.110	0.000	0.039	0.110
Campaign	-0.0500	0.006	-9.061	0.000	-0.061	-0.039
Housing	-0.4121	0.025	-16.427	0.000	-0.461	-0.363
Contact_cellular	0.4342	0.030	14.707	0.000	0.376	0.492
First half of month	0.2011	0.025	7.996	0.000	0.152	0.250
Married	-0.1785	0.025	-7.114	0.000	-0.228	-0.129
Season_2	-0.3775	0.027	-13.879	0.000	-0.431	-0.324

From probit modelling technique, the confidence intervals obtained after t test for all the independent variables does not contain zero we reject the null hypothesis (Reduced model explains the variance better). And hence we conclude that there is a relationship between the independent and dependent variables. And from the summary we can see that all the p values are very much less than our significance level, Hence we conclude that all the variables included are statistically significant.

Using Logit:

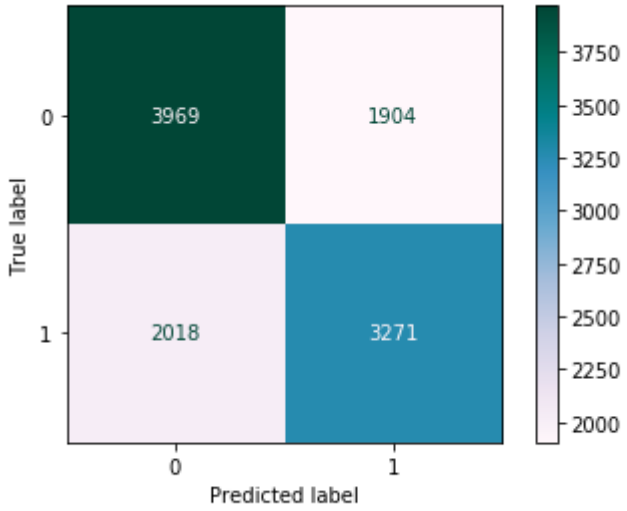
Deposit = 0.0192 + 0.1225 (Education) – 0.0884 (Campaign) – 0.6782 (Housing) + 0.7124 (Contact Cellular) + 0.3323(First Half of the month) – 0.2909 (Married) – 0.6091 (Season_2)

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0192	0.092	0.210	0.834	-0.160	0.199
Education	0.1225	0.030	4.147	0.000	0.065	0.180
Campaign	-0.0884	0.010	-9.058	0.000	-0.108	-0.069
Housing	-0.6782	0.041	-16.570	0.000	-0.758	-0.598
Contact_cellular	0.7124	0.049	14.640	0.000	0.617	0.808
First half of month	0.3323	0.041	8.053	0.000	0.251	0.413
Married	-0.2909	0.041	-7.084	0.000	-0.371	-0.210
Season_2	-0.6091	0.044	-13.753	0.000	-0.696	-0.522

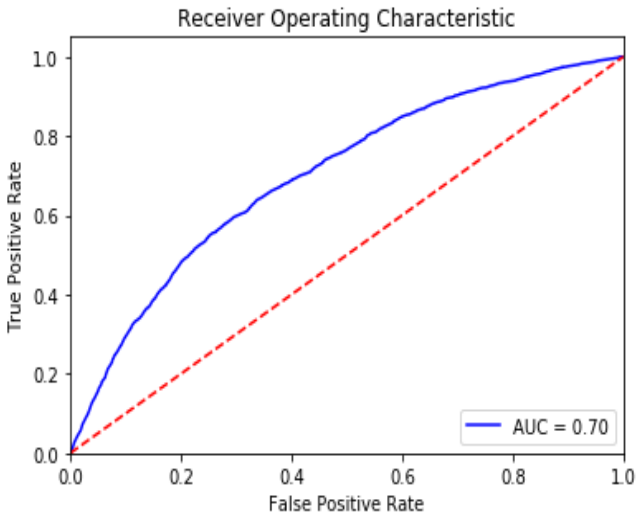
From Logit modelling technique, the confidence intervals obtained after t test for all the independent variables does not contain zero we reject the null hypothesis (Reduced model explains the variance better). And hence we conclude that there is a relationship between the independent and dependent variables. And from the summary we can see that all the p values are very much less than our significance level, Hence we conclude that all the variables included are statistically significant.

Performance Metrics:

Confusion Matrix:



ROC Curve:



	MODEL 1		MODEL 2	
	PROBIT	LOGIT	PROBIT	LOGIT
Intercept	-0.9548	-1.7073	0.0102	0.0192
Duration	0.0025	0.0048		
Education			0.0743	0.1225
Campaign	-0.0834	-0.1509	- 0.0500	-0.0884
Housing Loan	-0.6171	-1.0864	- 0.4121	-0.6782
Contact Cellular	0.6500	1.1093	0.4342	0.7124
First Half of the Month			0.2011	0.3323
Married			- 0.1785	-0.2909
Season 2			- 0.3775	-0.6091

Conclusion

The Recommendations for the next Marketing campaign are-

- **EDUCATION**- Customer with higher level of education should be contacted as they are more likely to subscribe to the term deposit.
- **CAMPAIGN**- Customer should not be contacted a large number of times, as "campaign" increases customer is less likely to subscribe to the term deposit.
- **HOUSING**- Customer with housing loan is less likely to subscribe to the term deposit.
- **CELLULAR CONTACT**- Customer contacted through cell phone is more likely to subscribe to the term deposit.
- **FIRST HALF OF THE MONTH**- Customers contacted in first half of the month are more likely to subscribe to the term deposit.
- **MARRIED** - Unmarried customers should be contacted as married customers are less likely to subscribe to the term deposit.
- **SEASON 2**- Customers contacted in second half of the year are less likely to subscribe to the term deposit.

Appendix

PYTHON CODE

IMPORTING LIBRARIES

```
importnumpyasnp
importpandasaspd
importmatplotlib.pyplotasplt
importseabornassns
```

IMPORTING THE DATASET

```
df = pd.read_csv("bank.csv")
Original = df.copy()
df.head()
```

IDENTIFYING MISSING VALUES

```
dataset.isnull().sum()
```

CONVERTING CATEGORICAL ATTRIBUTES TO NUMERICAL

```
fromsklearn.preprocessingimportLabelEncoder

le = LabelEncoder()
le.fit(df.Default.drop_duplicates())
df.Default = le.transform(df.Default)
# Housing Loan - yes or now
le.fit(df.Housing.drop_duplicates())
df.Housing = le.transform(df.Housing)
# Personal Loan - yes or now
le.fit(df.Loan.drop_duplicates())
df.Loan = le.transform(df.Loan)
# Deposit - yes or now
le.fit(df.Deposit.drop_duplicates())
df.Deposit = le.transform(df.Deposit)


df['Job_management'] = df.Job.map({'management':1,'blue-collar':0,'technician':0,'admin.':0,'ser
vices':0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':0
,'unknown':0 })
df['Job_blue-collar'] = df.Job.map({'management':0,'blue-collar':1,'technician':0,'admin.':0,'se
rvices':0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':
0,'unknown':0 })
df['Job_technician'] = df.Job.map({'management':0,'blue-collar':0,'technician':1,'admin.':0,'ser
vices':0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':0
,'unknown':0 })
```

```
df['Job_admin'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':1,'services'
:0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':0,'unkn
own':0 })
df['Job_services'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'servic
es':1,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':0,'u
nknown':0 })
df['Job_retired'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'service
s':0,'retired':1,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':0,'un
known':0 })
df['Job_self-employed'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'s
ervices':0,'retired':0,'self-employed':1,'student':0,'unemployed':0,'entrepreneur':0,'housemaid'
:0,'unknown':0 })
df['Job_student'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'service
s':0,'retired':0,'self-employed':0,'student':1,'unemployed':0,'entrepreneur':0,'housemaid':0,'un
known':0 })
df['Job_umemployed'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'serv
ices':0,'retired':0,'self-employed':0,'student':0,'unemployed':1,'entrepreneur':0,'housemaid':0,
'unknown':0 })
df['Job_entrepreneur'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'se
rvices':0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':1,'housemaid':
0,'unknown':0 })
df['Job_housemaid'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'servi
ces':0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':1,'
unknown':0 })
df['Job_unknown'] =df.Job.map({'management':0,'blue-collar':0,'technician':0,'admin.':0,'service
s':0,'retired':0,'self-employed':0,'student':0,'unemployed':0,'entrepreneur':0,'housemaid':0,'un
known':1 })

df['Married'] = df["Marital Status"].map({'married':1,'single':0,'divorced':0})
df['Single'] = df["Marital Status"].map({'married':0,'single':1,'divorced':0})
df['divorced'] = df["Marital Status"].map({'married':0,'single':0,'divorced':1})

df['Contact_cellular'] = df["Contact"].map({'cellular':1,'unknown':0,'telephone':0})
df['Contact_unknown'] = df["Contact"].map({'cellular':0,'unknown':1,'telephone':0})
df['Contact_telephone'] = df["Contact"].map({'cellular':0,'unknown':0,'telephone':1})

df.head()

df['Season_1'] = df["Month"].map({'jan':1,'feb':1,'mar':1,'apr':1,'may':0,'jun':0,'jul':0,'aug':
0,'sep':0 ,'oct':0,'nov':0,'dec':0})
df['Season_2'] = df["Month"].map({'jan':0,'feb':0,'mar':0,'apr':0,'may':1,'jun':1,'jul':1,'aug':
1,'sep':0 ,'oct':0,'nov':0,'dec':0})
df['Season_3'] = df["Month"].map({'jan':0,'feb':0,'mar':0,'apr':0,'may':0,'jun':0,'jul':0,'aug':
0,'sep':1 ,'oct':1,'nov':1,'dec':1})
```

AVOIDING DUMMY VARIABLE TRAP

```
df.drop('Job_unknown',axis=1,inplace = True)

df.drop('Job',axis=1,inplace = True)

df.head()

df.drop('Contact',axis=1,inplace = True)

df.drop('Contact_unknown',axis=1,inplace = True)

df.head()


df.drop('Month',axis=1,inplace = True)

df.drop('Season_3',axis=1,inplace = True)

df.head()
```

CHECKING CORRELATION BETWEEN VARIABLES

```
df.corr()['Deposit'].sort_values()
```

```
#With Duration

columns = ['Housing','Duration','Campaign','Contact_cellular','Deposit']
df = df.reindex(columns=columns)
df.head()
```

Modelling:

```
importstatsmodels.api as sm

X = df[['Housing','Duration','Campaign','Contact_cellular']]
Y = df["Deposit"]
X = sm.add_constant(X)

probit = sm.Probit(endog=Y, exog = X)

result = probit.fit()

result.summary()
```

```
X = df[['Housing','Duration','Campaign','Contact_cellular']]
Y = df["Deposit"]
X = sm.add_constant(X)

logit = sm.Logit(endog=Y, exog = X)

result = logit.fit()

result.summary()

#UsingSklearn for plotting roc
```

```
fromsklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X, Y)

y_pred = classifier.predict(X)
fromsklearn.metrics import plot_confusion_matrix
plot_confusion_matrix(classifier, X, Y , cmap='PuBuGn')  # doctest: +SKIP
plt.show()


fromsklearn.metrics import accuracy_score
accuracy_score(Y, classifier.predict(X))

#ROC curve code
importsklearn.metrics as metrics

probs = classifier.predict_proba(X)
preds = probs[:,1]

fpr, tpr, threshold = metrics.roc_curve(Y, preds)
roc_auc = metrics.auc(fpr, tpr)


importmatplotlib.pyplot as plt
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
plt.legend(loc = 'lower right')

plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0.0, 1.05])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
importstatsmodels.api as sm

X = df[['Education','Campaign','Housing','Contact_cellular','First half of month','Married'
,'Season_2']]
Y = df["Deposit"]
X = sm.add_constant(X)

probit = sm.Probit(endog=Y, exog = X)

result = probit.fit()

result.summary()

importstatsmodels.api as sm

X = df[['Education','Campaign','Housing','Contact_cellular','First half of month','Married'
,'Season_2']]
Y = df["Deposit"]
X = sm.add_constant(X)

logit = sm.Logit(endog=Y, exog = X)
```

```
result = logit.fit()

result.summary()

fromsklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X, Y)

y_pred = classifier.predict(X)

fromsklearn.metrics import plot_confusion_matrix
plot_confusion_matrix(classifier, X, Y , cmap='PuBuGn')  # doctest: +SKIP
plt.show()
```

References:

https://stattrek.com/multiple-regression/dummy-variables.aspx

https://en.wikipedia.org/wiki/Correlation_and_dependence#Correlation_matrices

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

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

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

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