MBC 638 Final Project

Bank Term Deposit Analysis

- Ankita Vartak
- Dhrumil Parekh
- Nivesh Vaze
- Pranav Seth

Descriptive Analytics Part A

PART A: Descriptive Analysis

- Dataset Description and Source
- Variable Description
- Descriptive and Statistical Analysis of Variables

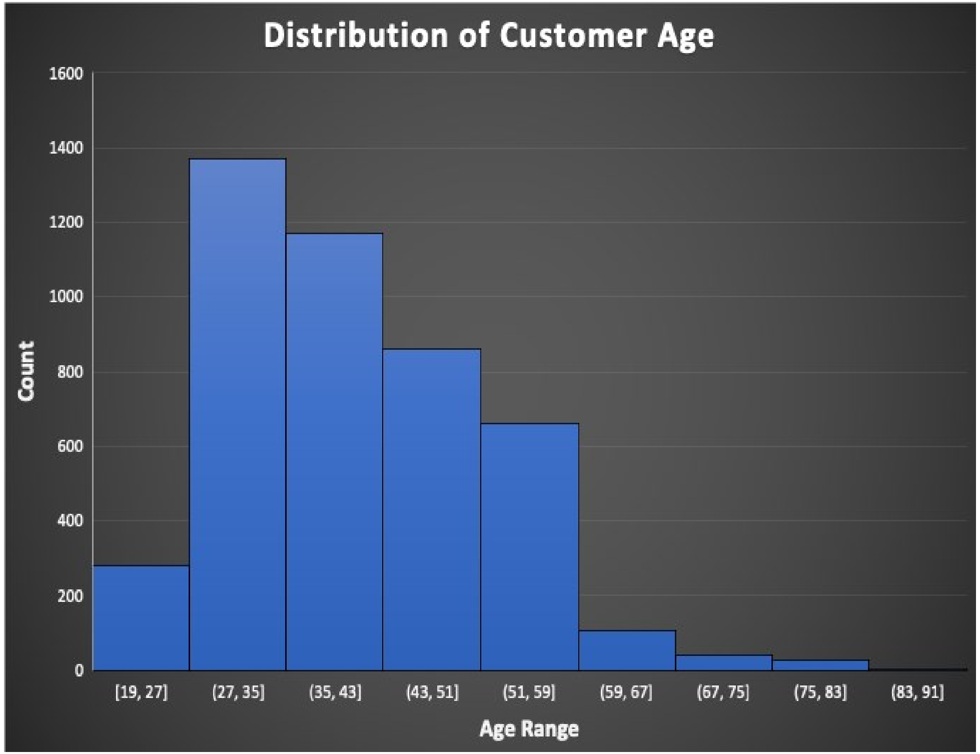
Data Description and Source

- Data is related with direct marketing campaigns of a Portuguese banking institution
- Marketing campaigns were based on phone calls where often more than one contact to the same client was required, in order to access if the product (bank term deposit) would ('yes') or would not ('no') be subscribed
- Classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y)
- Dataset Source: UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems

VARIABLE DESCRIPTION

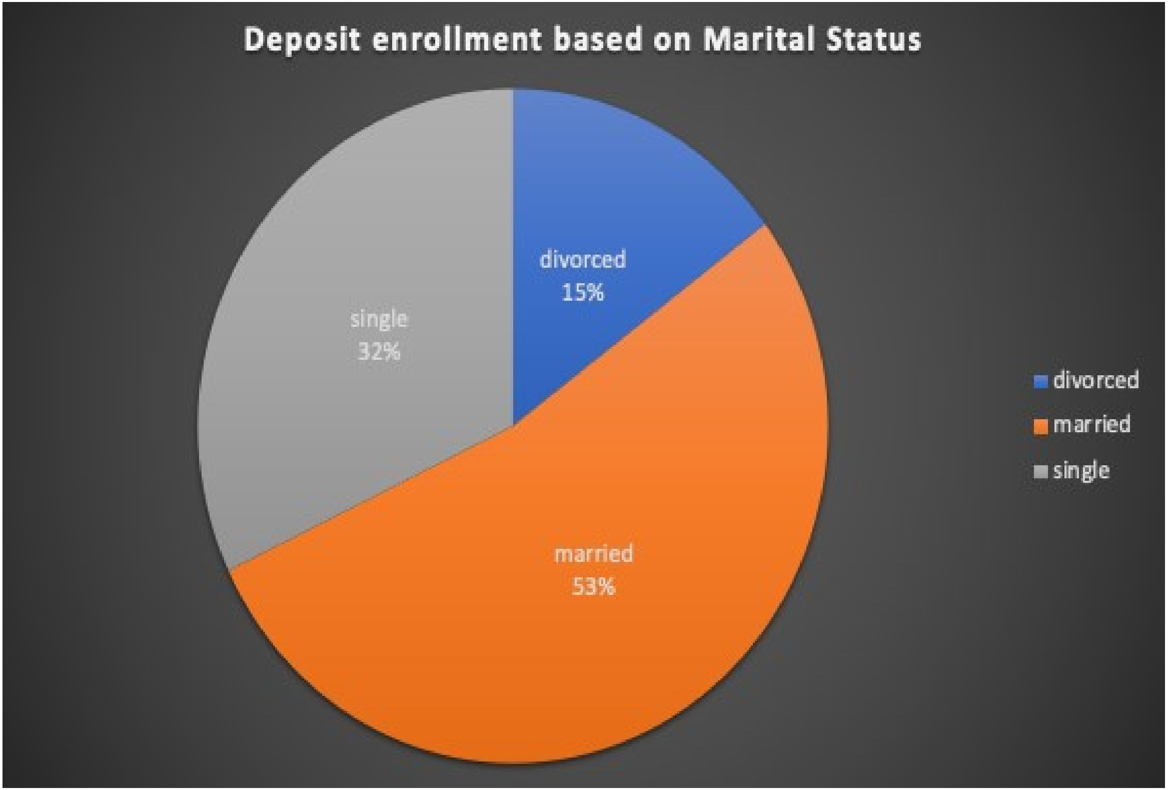
VARIABLES	DESCRIPTION
Age	Age of the customer (in Years)
Job	Type of job (eg. Admin, student, technician, retired)
Marital	Marital status (singly, married, divorced)
Education	Education Categories (primary, secondary, tertiary)
Default	Default in credit (yes, no)
Contact	Contact communication type (cellular, telephone, unknown)
Housing	Has a Housing loan (yes, no)
month	Last contact month of year (eg. Jan, feb, mar)
day_of_week	Last contact day of the week (1,2,3, etc)
duration	Last contact duration (in seconds)

campaign	Number of contacts performed during this campaign and for this client
pdays	Number of days that passed by after the client was last contacted from a previous campaign
previous	Number of contacts performed before this campaign and for this client
poutcome	Outcome of the previous marketing campaign
emp.var.rate	Employment variation rate - quarterly indicator
cons.price.idx	Consumer price index - monthly indicator
cons.conf.idx	Consumer confidence index - monthly indicator
euribor3m	Euribor 3 month rate - daily indicator
nr.employed	Number of employees - quarterly indicator
Y(Term deposit enrollment)	Has the client subscribed a term deposit



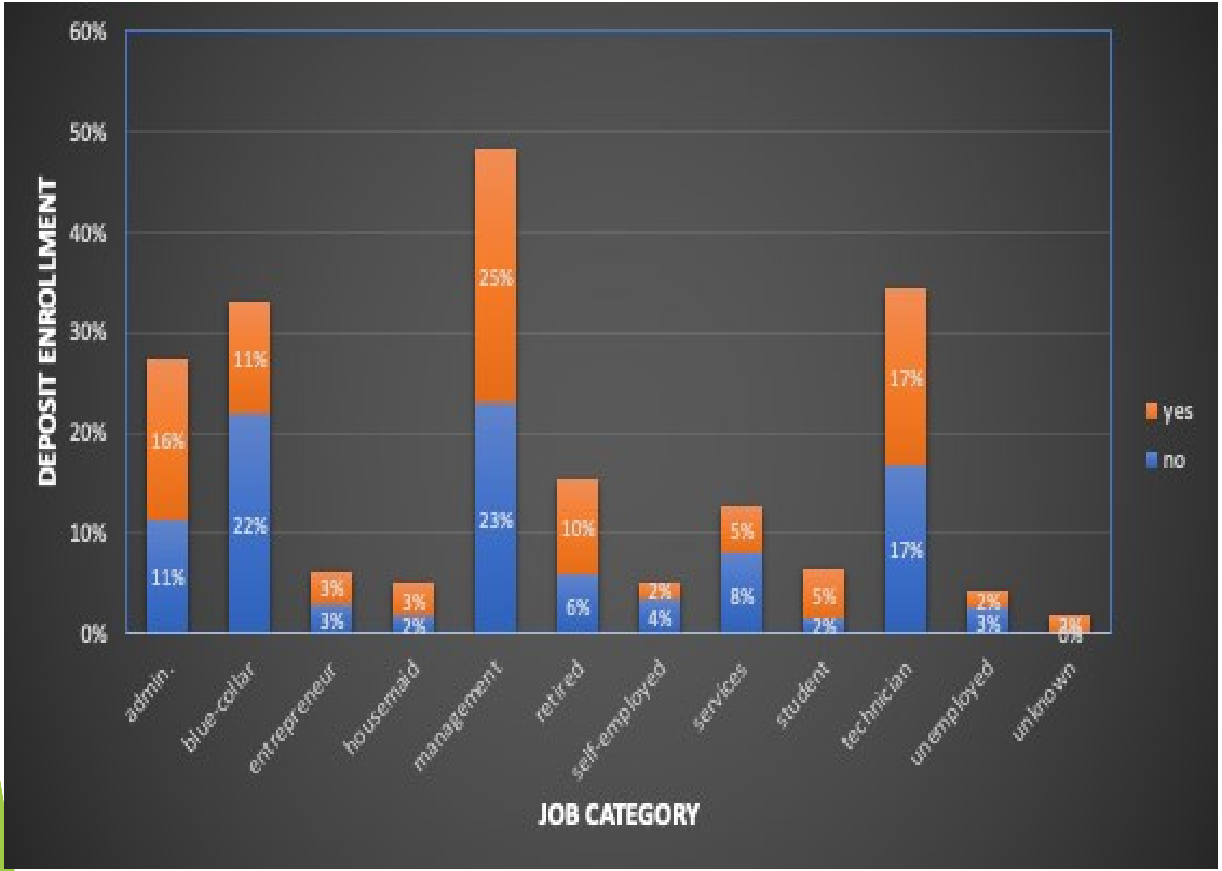
Summary

- The given chart shows the Distribution of Customer Age captured for the data set
- Maximum customers fall in the 27-35 years of age helping Business teams run targeted Marketing campaigns



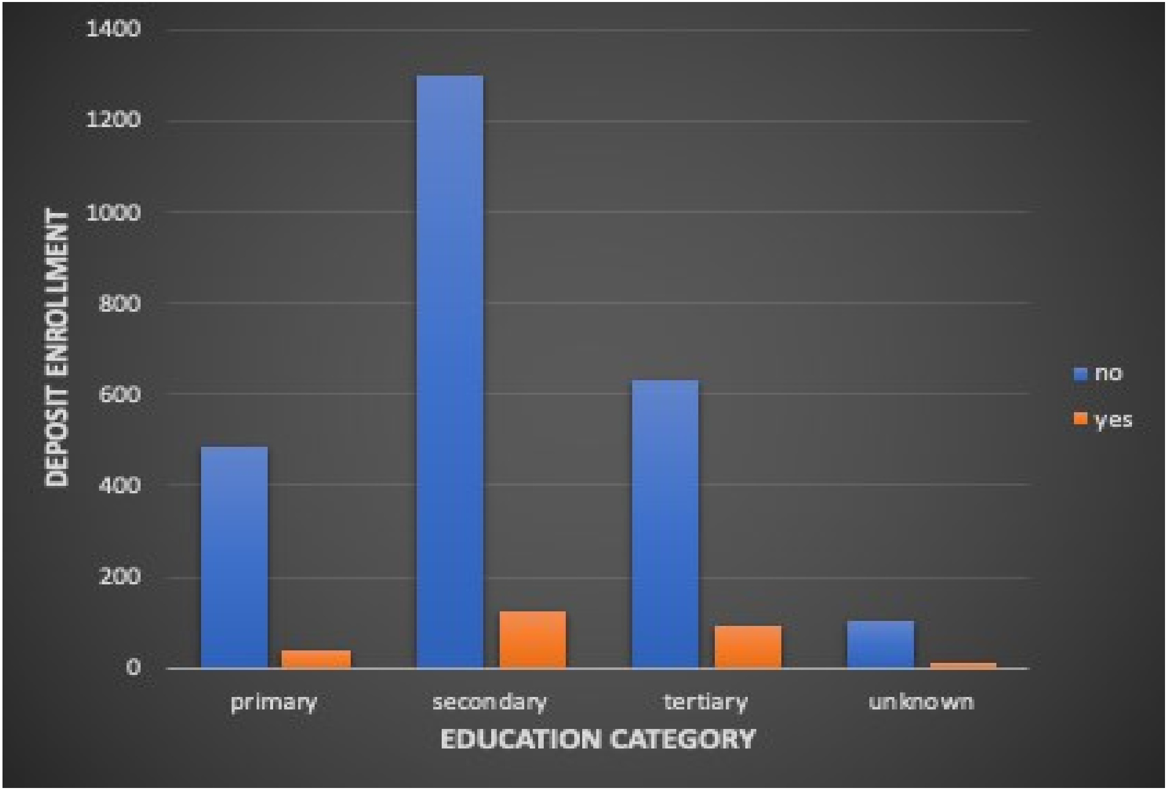
Summary

- The given Pie chart shows the distribution of Deposit Enrollment based on Marital Status.
- The chart shows that Married people have the highest enrollment for Term Deposits.



Summary

- This chart shows distribution according to job categories who did not enroll for Term Deposit in the previous campaign but enrolled after attending the current campaign.
- Data shows that the Management category had the highest percentage of customers who did not enroll in Term Deposit even after attending the current campaign.



Summary

The given Pivot chart shows the distribution of Married people enrolled for Deposit based on their Education Category helping Business teams understand their customers and gaps to be filled

Association Between variables
Part B

Part B

- Explanatory and dependent variables
- Correlation Analysis
- Logistic Regression Analysis
- Interpretation of results
- Evaluation of findings
- Predictions
- Business Use

Objective of Regression Analysis

In this section, we did the Logistic regression. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Our data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, to access if the product (bank term deposit) would be (‘yes’) or not (‘no’) subscribed. Thus, our dependent variable is named y(term deposit enrollment) in the dataset.

Selection of independent variables for the multiple regression model - Table of Correlation Coefficients

The variables that are used to explain or predict the y(term deposit enrollment) variable are called the explanatory variables. we carry out the correlation analysis to highlight the significant independent variable and include those variables in the regression model.

	balance	*age*	*campaign*	*pdays*	*previous*	*duration*	*Yout*
balance	1						
age	0.083820142	1					
campaign	-0.009976166	-0.00515	1				
pdays	0.009436676	-0.00889	-0.09314	1			
previous	0.026196357	-0.00351	-0.06783	0.577562	1		
duration	-0.015949918	-0.00237	-0.06838	0.01038	0.01808	1	
Y(term deposit enrollment)	0.017905098	0.045092	-0.06115	0.104087	0.116714	0.401118	1

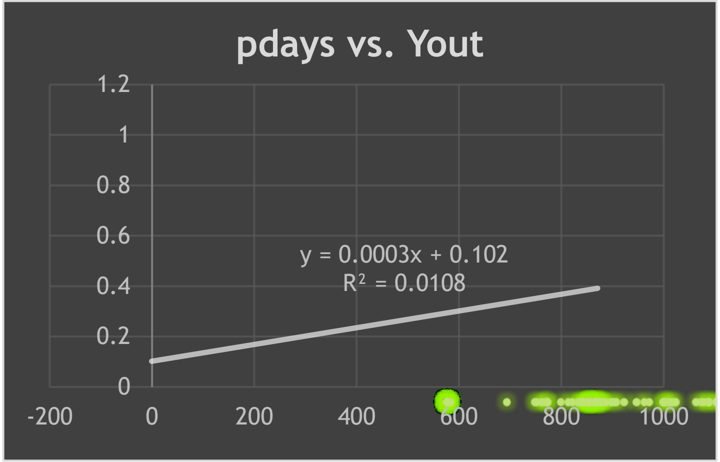
Selection of independent variables for the multiple regression model - Table of Correlation Coefficients

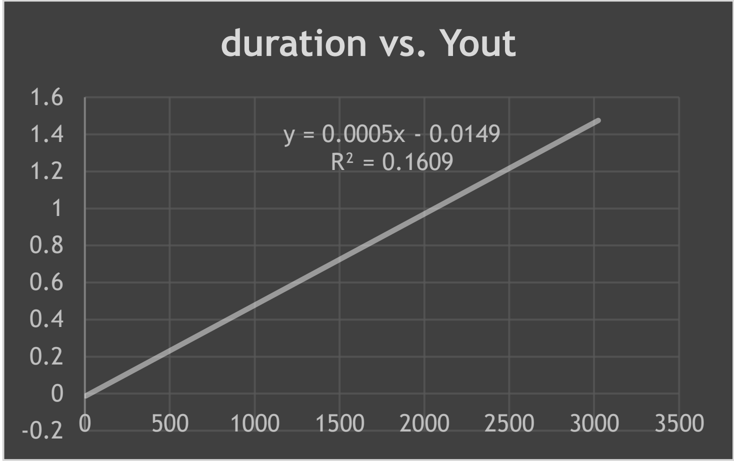
Order of independent variables inputted as referenced from the correlation table:

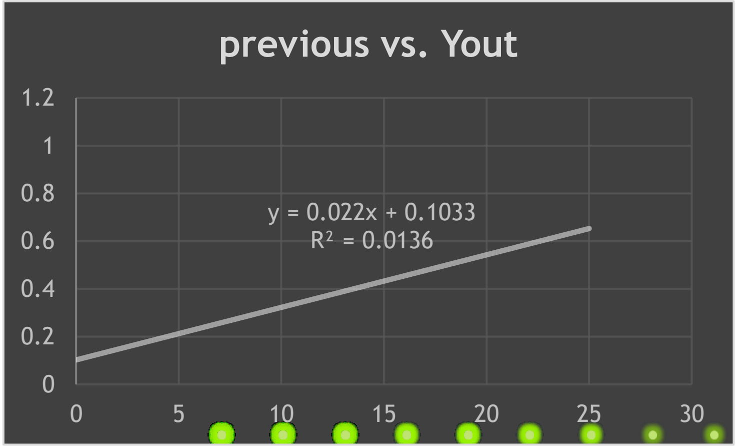
1. duration
2. previous
3. pdays
4. age
5. balance
6. Campaign

Based on the degree of correlation, we included duration, pdays and previous variables as explanatory variables in the regression model

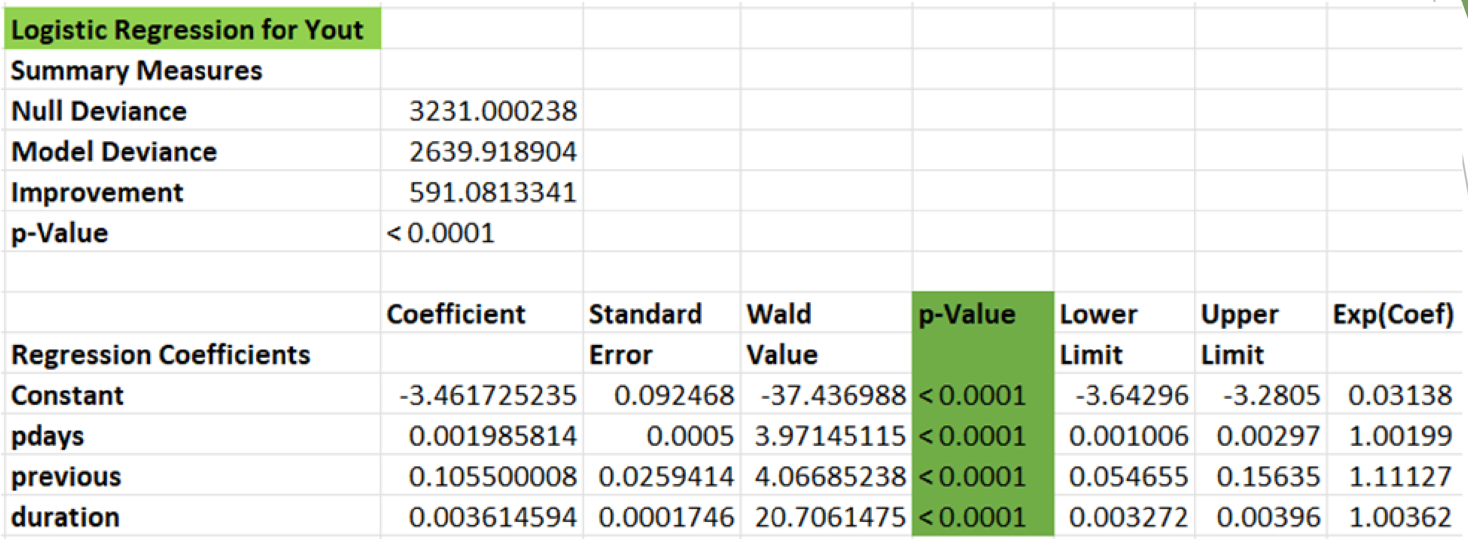
Scatterplots of Explanatory Variables:







Regression Analysis Interpretation:



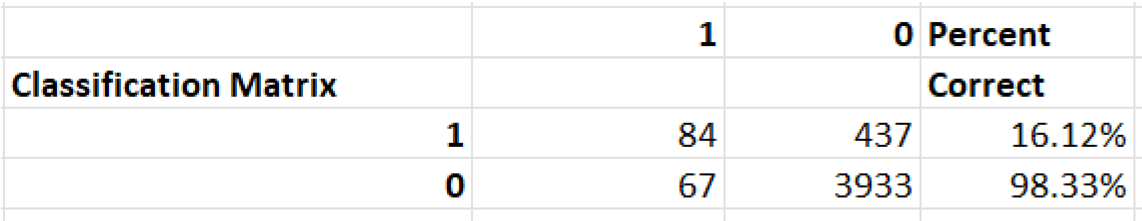
Logistic Regression for Yout							
Summary Measures							
Null Deviance	3231.000238						
Model Deviance	2639.918904						
Improvement	591.0813341						
p-Value	< 0.0001						
	Coefficient	Standard	Wald	p-Value	Lower	Upper	Exp(Coef)
Regression Coefficients		Error	Value		Limit	Limit	
Constant	-3.461725235	0.092468	-37.436988	< 0.0001	-3.64296	-3.2805	0.03138
pdays	0.001985814	0.0005	3.97145115	< 0.0001	0.001006	0.00297	1.00199
previous	0.105500008	0.0259414	4.06685238	< 0.0001	0.054655	0.15635	1.11127
duration	0.003614594	0.0001746	20.7061475	< 0.0001	0.003272	0.00396	1.00362

The p-values of all three co-efficient are <0.0001 which suggests that all the three explanatory variables considered in the regression analysis are statistically significant in explaining the dependent variable y(term deposit enrollment).

In Logistic regression, we first calculate Logit and then the Probability of the outcome using the regression coefficients. Calculation of logit is given below:

Logit = -3.46172 + 0.001985 pdays + 0.105550 previous + 0.003614 duration.
Calculation of Probability is given below: -
Probability = EXP(Logit)/(1+EXP(Logit))

Based on the value of previous, pdays and duration of each customer we have calculated Logit and probability. The performance of the Logistic regression model is shown below.



	1	0	Percent
Classification Matrix			Correct
1	84	437	16.12%
0	67	3933	98.33%

From this, we can interpret that our model is able to predict enrollment outcomes with 16.12% accuracy. i.e., 16.12% of historical records with y(term deposit enrollment) as 1 have a probability
of >50%.

Prediction using the regression equation

Here, we have calculated the probability of two different customers to check whether they might enroll for the term deposit.

1. Duration = 2769 seconds, previous = 0, pdays = -1.

For this customer, the Logit = 6.545 and Probability = 0.998. As the duration of the call is high the probability of enrolling for the deposit is higher. i.e., 99%

This data is a historical record, and the customer did enroll for the deposit.

2. Duration = 3025 seconds, previous = 0, pdays = -1

For this customer, the Logit = 7.470 and Probability = 0.999. As the duration of the call is high the probability of enrolling for the deposit is higher. i.e., 99%

In this case, the model suggests that the customer will enroll for the term deposit, however, the customer did not enroll.

In conclusion, I would say that there are certain non-linearities that are not being considered in the regression analysis.

Conclusion:

Based on the analysis performed throughout the project, we can say that the regression model could be used to predict the outcome of the Term Deposit Enrollment and accordingly provide inputs to the Bank Telemarketing team to segment and target the audience accurately.

Also, the model is quite rigid and cannot model adequately complex nonlinear relationships. Hence the accuracy of the Logistic regression in this case is lesser.

As a future scope, we can work on implementing decision trees (DTs), neural networks (NNs) or support vector machines to better understand the prospective bank customer and save hefty contact costs for the banking client.

Thank You ☺