LEAD SCORING– Case study

Problem Statement

Education company – Lead Scoring to identify Hot Leads

From the customer data provided

Build logistic regression model to assign a lead score between 0 to 100

Identify top three variables in your model which contribute most towards the probability of a lead getting converted

Identify top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion

Strategy to employ and areas to concentrate on during time when the Sales team has more manpower

Strategy to employ and areas to concentrate on during time when the targets are met and sales team has lesser manpower

Data cleaning

1)We import the application data onto our notebook after importing the requisite libraries

2)We then check the dataset for the datatypes and the distribution of numeric and categorical data

3)We also conduct an outlier analysis on the continuous variables and see that the total visits and pages per view have outliers.

4)We then proceed to deal with the missing values

a. Calculate the percentage of data missing in each of the columns of the dataset

b. Delete all the columns with more that 40% data missing in them – This is done because when a large part of the data is missing from the column, we won't get any valuable insights for that column and it will just add noise to the dataset

c. We then club the categorical variables and assign them binary codes and delete the columns which have only one answer to the question as it will not contribute much

d. Next we deal with each column with missing data individually and either impute the missing data with median or mode, or in a few cases where the option 'Select' is present, include that as a null value in the calculation and delete or impute columns on the basis of the total percentage of actual null values

d. We also proceed to delete the prospect ID column as it serves the same purpose as lead number

5)After the data cleaning, we are left we 23 columns (37 initially) & all the original number of rows

Data Preparation – Scaling, Train Test split, Correlation

1) We treat the categorical variables using dummy variables and binary variables. We have already dealt with the categorical variables with binary data in the data cleaning stage. We create dummies for the remaining categorical variables

2) Next we split the data into X and y which has the feature variables and the response variables respectively

3) After that, we split the X and y data further into train and test sets

4) Then, we have three numerical variables which we scale using the Standard scaler and we have our X train data ready for modelling

5) We also realize that we have a lot of new columns due to the creation of dummy variables, so to make the modelling a little easier, we drop the columns which have a correlation of more than 0.95

Model Building

1) Import the required libraries for logistic regression modelling and run the same on the train dataset

2) Since we have a lot of categorical features, we use RFE analysis for selecting the top 15 which are contributing the most to the predictor variable

3) After we have the top 15 variables, we use statsmodels to check the p score and the constants for X train to determine which of the factors contributing highly to the final value. We also delete the columns which have a high p values and rerun the stats model till only variables with low p values remain

4) We next use the modified X train dataset to calculate the predicted converted rate values for y train dataset using the cut off as 0.5

5) We calculate the confusion matrix and the overall accuracy and see that the model has an accuracy of 92.36% which is very good

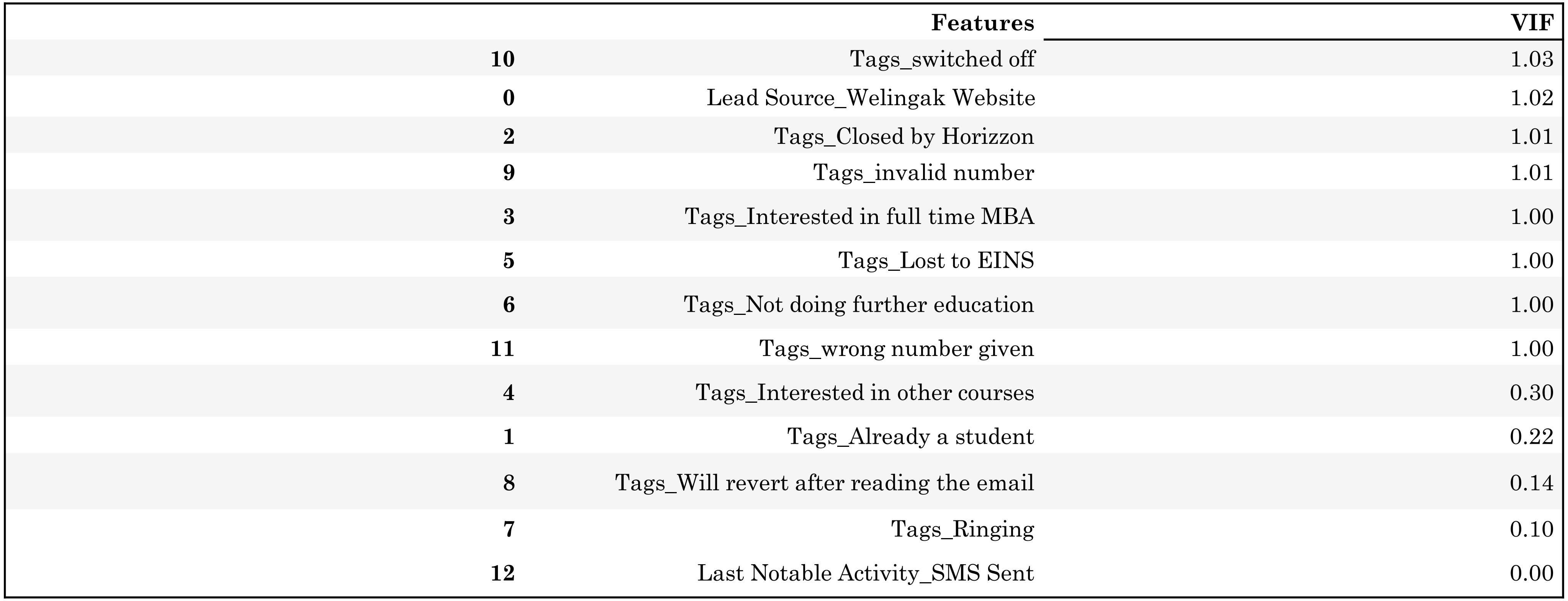
6) We run VIF analysis and see that the max VIF is 1.03. We need not delete any columns based on VIF but we do have two columns with high p values, so we delete those

Model Building

7) After dropping the columns with high p values, we run statsmodel again and see that none of the remaining variables have high p values

8) We calculate the predicted values on y train, and check the confusion matrix and accuracy. We see that the accuracy has dropped 0.1% and is at 92.26% which is good

9) We check the VIFs to see if the dropping of columns has impacted the scores but see that the max VIF is still 1.03 so we can proceed with this model



	Features	VIF
10	Tags_switched off	1.03
0	Lead Source_Welingak Website	1.02
2	Tags_Closed by Horizzon	1.01
9	Tags_invalid number	1.01
3	Tags_Interested in full time MBA	1.00
5	Tags_Lost to EINS	1.00
6	Tags_Not doing further education	1.00
11	Tags_wrong number given	1.00
4	Tags_Interested in other courses	0.30
1	Tags_Already a student	0.22
8	Tags_Will revert after reading the email	0.14
7	Tags_Ringing	0.10
12	Last Notable Activity_SMS Sent	0.00

Model Building

10) Using the confusion matrix we calculate the sensitivity and specificity

a) Sensitivity – 86.33%

b) Specificity – 95.93%

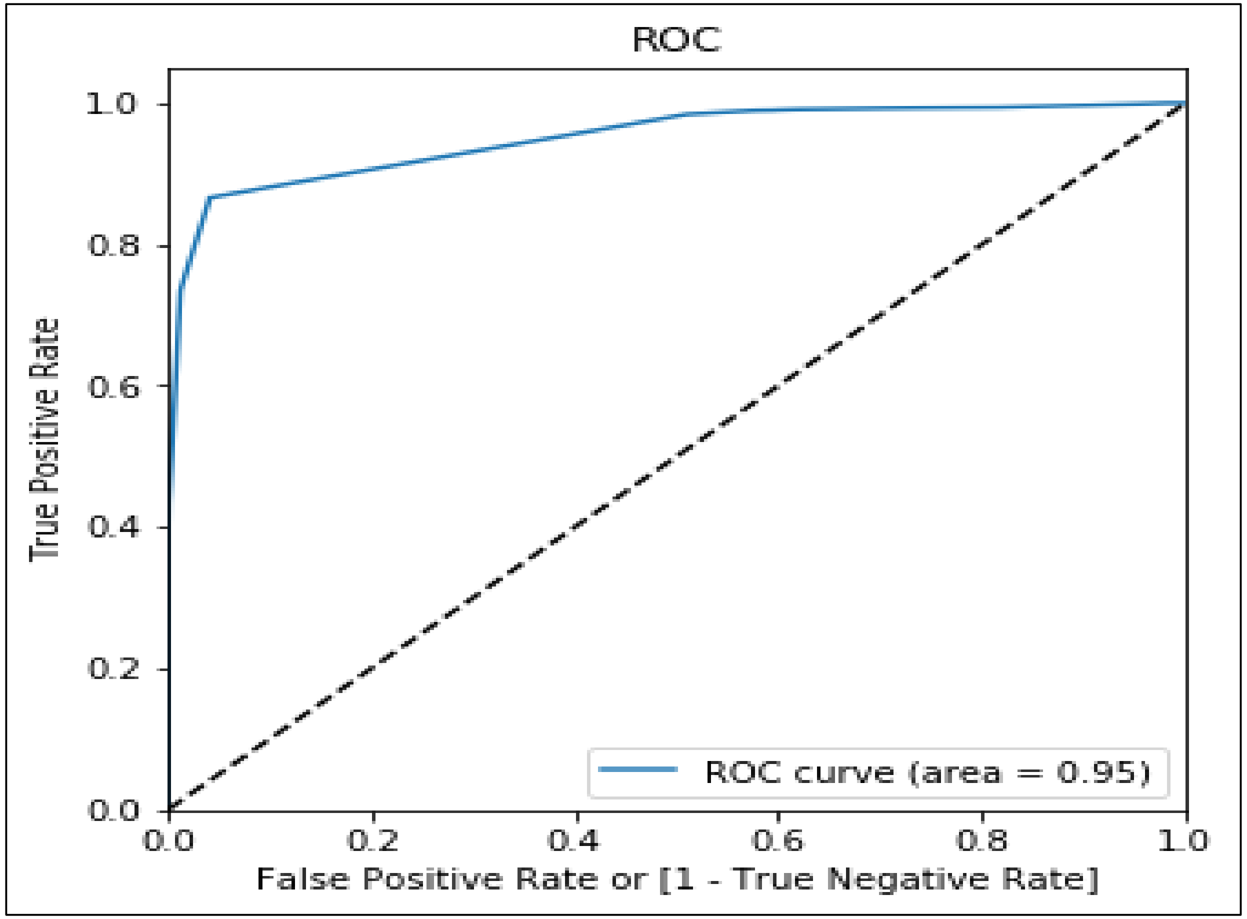
c) False conversion rate – 4.07%

d) Positive predictive value – 92.88%

e) Negative predictive value – 91.93%

11) ROC curve

We plot the ROC curve and can see that it is acceptable



Model Building

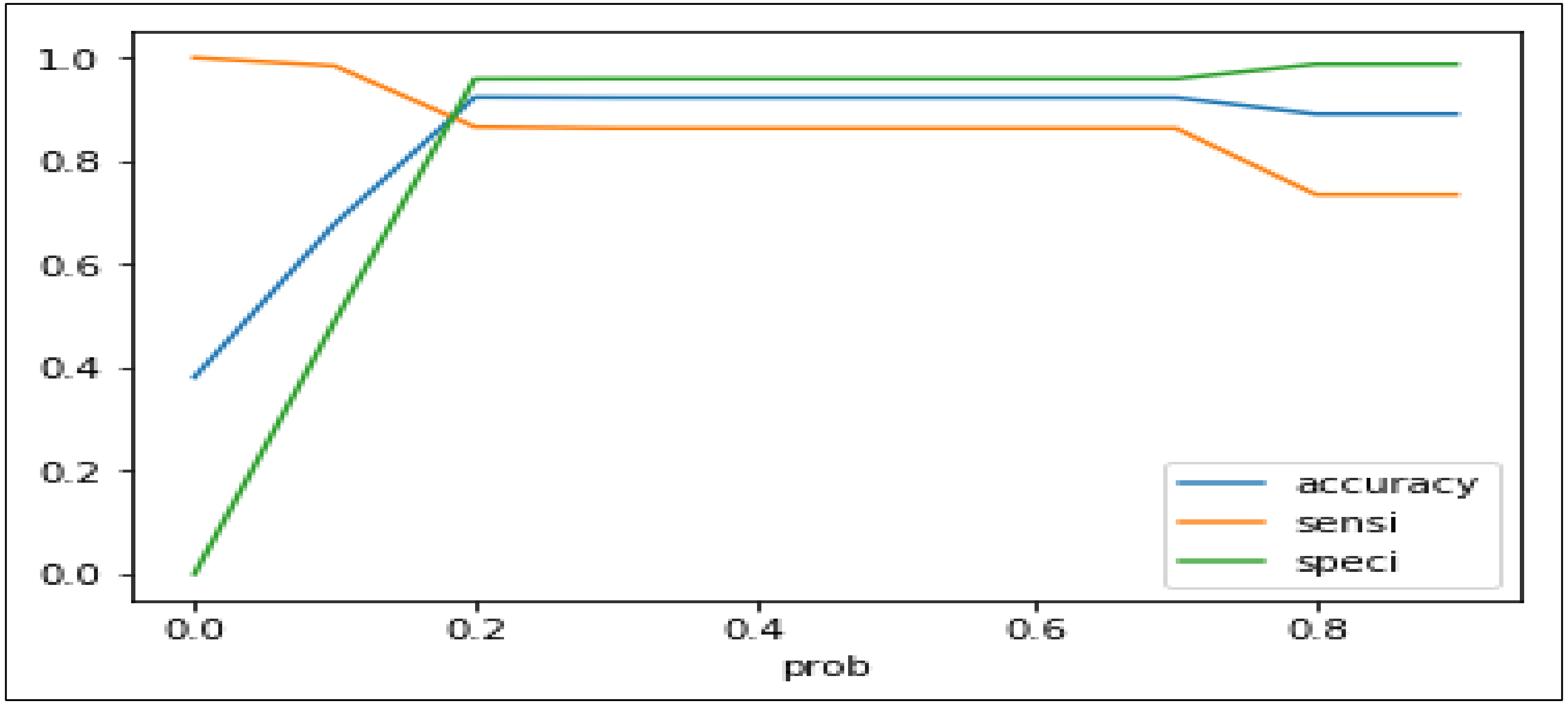
12) Optimal cut off point

a) We first create columns with different probability cut offs - 0, 0.1, 0.2 till 0.9

b) We calculate the accuracy, sensitivity, & specificity for each of the cut offs

c) Plotting of accuracy, sensitivity, & specificity for each of the probabilities

d) As per the graph, we see that the optimal cut off point is 0.2



Model Building

13) Using the cut off point as 0.2, we calculate the predicted value for y again and see that the overall accuracy is high at 92.36%

14) We also calculate the confusion matrix, sensitivity, and specificity again to see if changing the cut off from 0.5 to 0.2 has any effect on them

a) Sensitivity – 86.61%

b) Specificity – 95.90%

c) False conversion rate – 4.10%

d) Positive predictive value – 92.86%

e) Negative predictive value – 92.08%

15) We see that the values have changed marginally so we can proceed with this

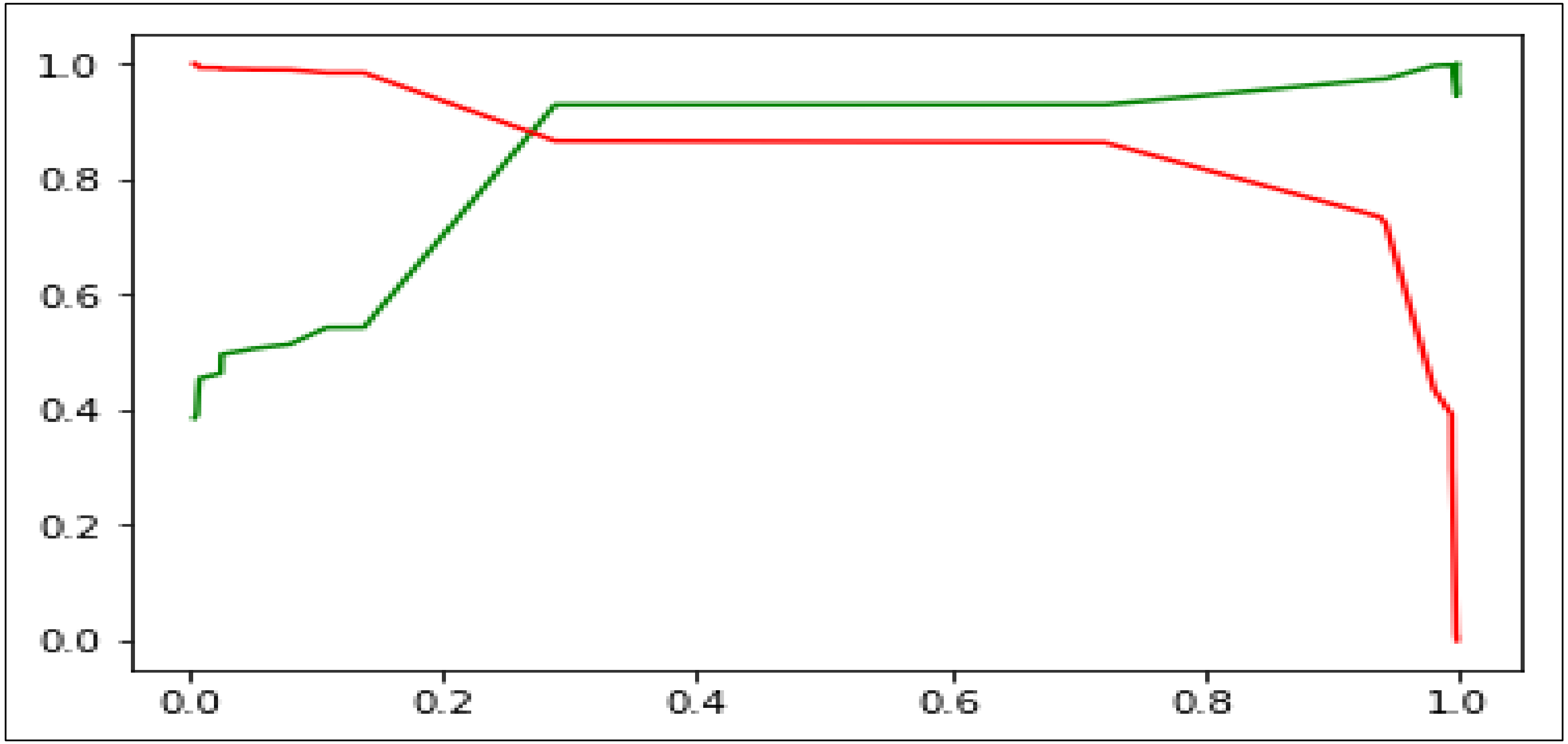
Precision and Recall

1) Using the confusion matrix, we calculate the precision and recall

   a) Precision – 92.88%

   b) Recall – 86.33%

2) Precision and Recall tradeoff - We plot the precision recall curve



Predictions on test set

1) We transform the X test in the same way we had run fit transform on X train

2) We add a constant to X test and use that to predict conversion probability values as y test dataframe

3) We concat the values of y test that we calculated from the model with the original y test dataset to compare the conversion rates

4) We calculate the final predicted value using cut off as 0.2 on the calculated conversion rate

5) We calculate the overall accuracy using the actual converted values from the dataset provided and the calculated conversion values using our model. We see that the accuracy is 92.46%

6) Using the confusion matrix, we calculate the sensitivity and specificity of the predicted values

a) Sensitivity – 88.03%

b) Specificity – 95.34%

c) False positive rate – 4.65%

Second Model

1) Since most of the features that the first model has considered belong to variable 'Tags', which is data inputted by the Sales team, we will also create a model that does not depend on this data. This is done for the following reasons

a) In order to reduce human effort dependency, we will create another model that does not consider any data that the Sales team has input themselves

b) This will help reduce model dependency on human effort as well as reduce the error in prediction due to human error

c) This model will be helpful even if the sales team doesn't input all of the data during sales calls

Second Model

2) We drop the column ‘Tags’ from the main database

3) Then we create dummies for categorical variables, separating the dataset in X and y with feature and response variable

4) We split the dataset into training and test to perform the modelling

5) Lastly, we standardize the numerical variables in X train dataset using scaler and drop the highly correlated variables

Model Building

6) We run the logistic regression model on X train and y train and perform RFE to get the top 15 variables influencing the response variable

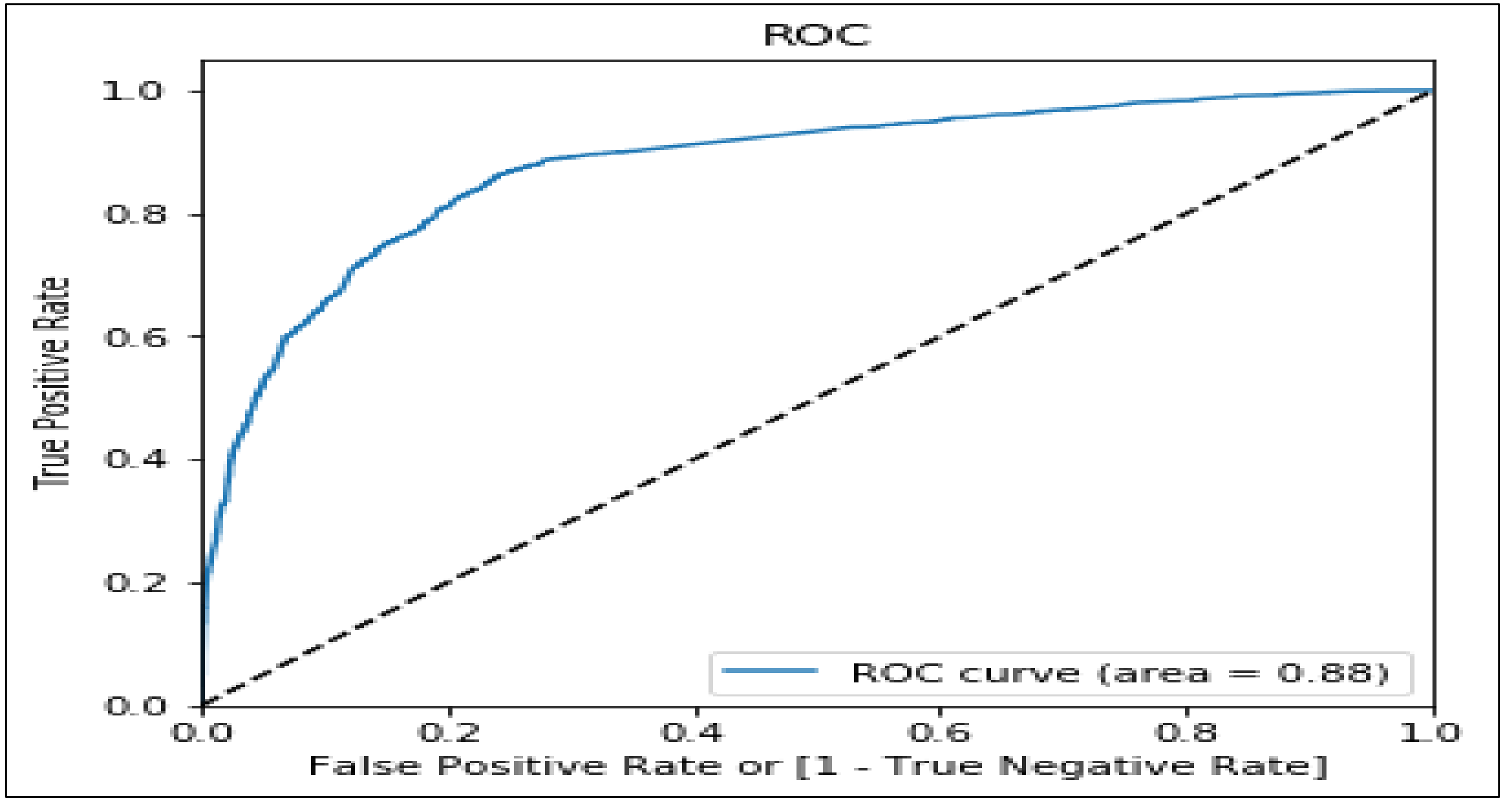
7) From the model results, we delete the columns with high p values and then calculate the predicted values for y train from modified X train

8) Conversion probability is calculated using the cut off value as 0.5

9) Accuracy for the model turns out to be decent at 80.64% and max VIF at 1.88

Second Model

10) We plot the ROC curve and it looks satisfactory



11) We transform the X test dataset and calculate the predicted values for y test data using the modified X test dataset

12) The predicted conversion probability values and the final predicted value from the model are compared with the actual data of the y test dataset and the accuracy is calculated to be 80.23%

Insights and Conclusion

1) The dataset provided had a lot of missing values, a lot of them pertaining to data captured manually. If the complete data was available, the analysis could have been more detailed

2) Most of the data could be ignored or statistically imputed but assumptions had to be made for missing value in the Tags column.

3) From the first model, we see the importance of the values in the **Tags** column – this data is captured during sales calls. It is imperative that the data be captured correctly because the model is highly dependent on this data to make predictions and generate a good database of Hot Leads which are worth pursuing

4) **Lead source** is also an important predictor of Hot Leads and the company must spend more resources in determining which sources lead to higher conversion of leads and concentrate their efforts and advertising budget on those sources

5) **Last notable activity** is also an important predictor and it gives insights into which customers are actively spending time reading about the company, visiting the website, and responding to emails. It also helps us weed out the disinterested customers in the initial stages as the people not responding to calls or mails will most definitely not be in the Hot Leads pool. So valuable time can be conserved and put to use in areas which will generate sales

Insights and Conclusion

6) The first model is highly dependent on sales calls entries and during the time the company has higher manpower in the sales team, it would be helpful to use this model to generate Hot Leads and also ensure that the sales calls data is being correctly captured which in turn will help strengthen the model

7) We have also made a **second model** in which we have removed the Tags data. This was done to see how the predictions for Hot Leads could be made only on the basis of automated data available.

8) Although the accuracy of the second model is lower, it is decent at 80% and it gives us valuable input on which areas to concentrate on during the leaner periods where the sales manpower is limited

9) **Last notable activity** clearly seems to one of the most important predictor as it has been captured in the second model as well

10) **Lead origin and Occupation** are important predictors in deciding whether the lead is worth pursuing or not. When the manpower is limited, we should concentrate on the working professionals as they would be interested in the course to further their career and also have the financial means to pay for it.