LEAD SCORE CASE STUDY

PRASANJIT NAYAK NITHIN DHONGADE & NITISH KUMAR

PROBLEM STATEMENT

INTRODUCTION:

An education company named x education sells online courses to industry professionals.

On any given day, many professionals who are interested in the courses land on their website and browse for courses

Once people land on the websites, they might browse the courses or fill up a form for the course.

When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the compony also gets leads through past referrals

Once these leads are acquiring, employees from the sales team start making call, writing emails.

The typical lead conversion rate at x education is around 30%

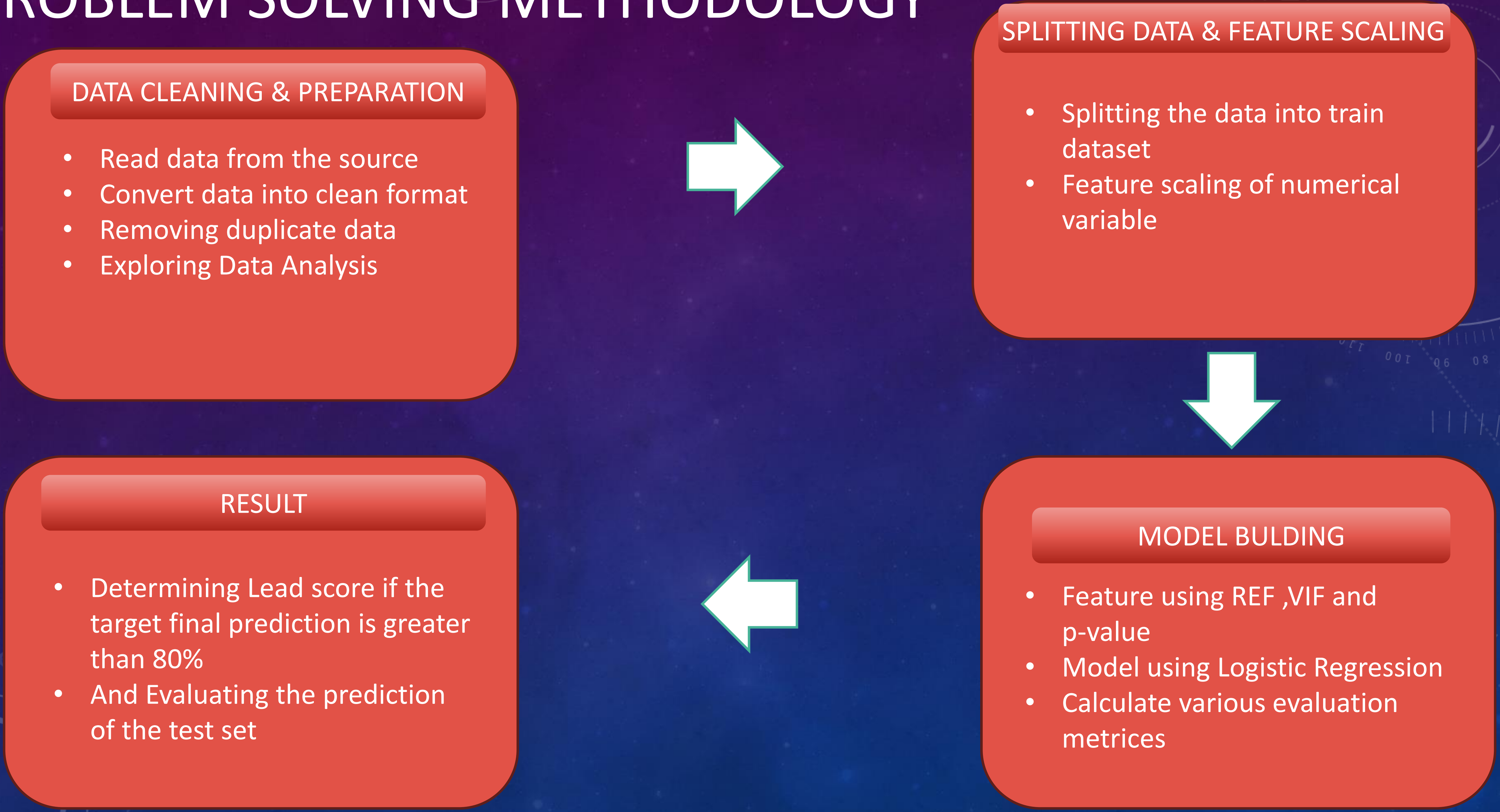
BUSINESS GOALS:

Build a logistic regression model to assign lead scores (0-100) based on conversion likelihood, ensuring adaptability for future changes like feature adjustments and hyperparameter tuning for improved targeting

OVERALL APPROACH

1. DATA CLEANING
2. EXPLORATORY DATA ANALYSIS: UNIVARIATE,BIVARIATE and MULTIVARIATE ANALYSIS
3. SCALING AND DUMMY VARIABLE CREATION
4. LOGISTIC REGRESSION MODEL BULDING
5. MODEL EVALUATING : SPECIFICITY, SENSITIVITY and RECALL
6. CONCLUSION AND RECOMMENDATION

PROBLEM SOLVING METHODOLOGY



DATA CLEANING & PREPARATION

- Read data from the source
- Convert data into clean format
- Removing duplicate data
- Exploring Data Analysis

SPLITTING DATA & FEATURE SCALING

- Splitting the data into train dataset
- Feature scaling of numerical variable

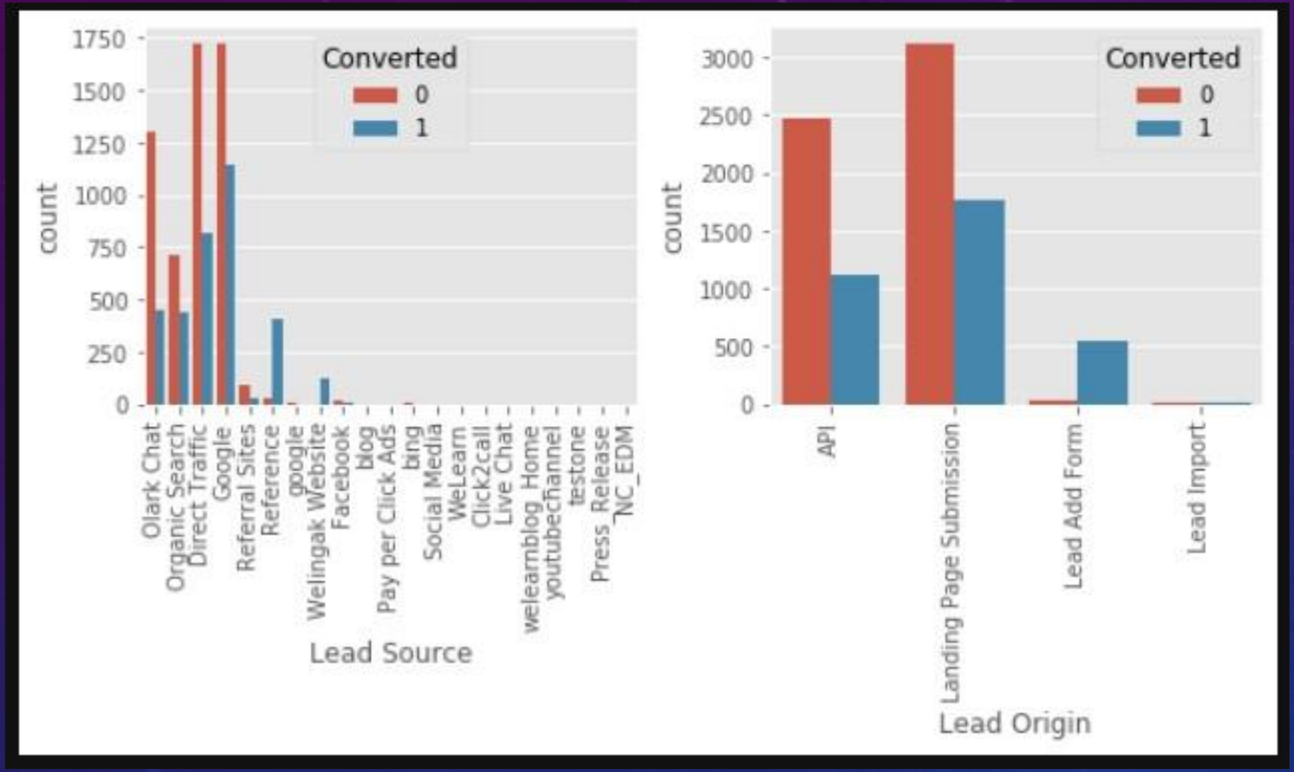
MODEL BULDING

- Feature using REF ,VIF and p-value
- Model using Logistic Regression
- Calculate various evaluation metrices

RESULT

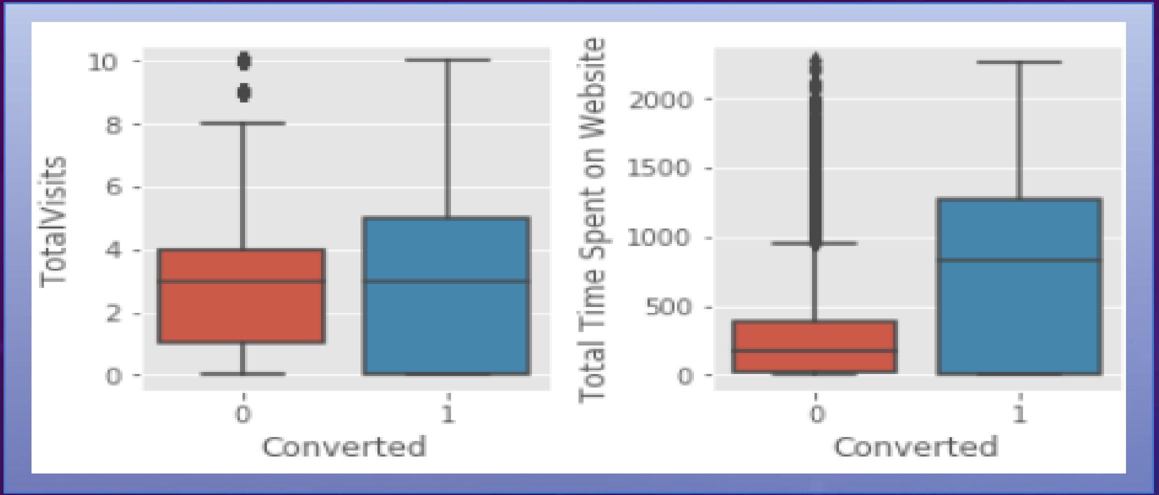
- Determining Lead score if the target final prediction is greater than 80%
- And Evaluating the prediction of the test set

EXPLORATORY DATA ANALYSIS

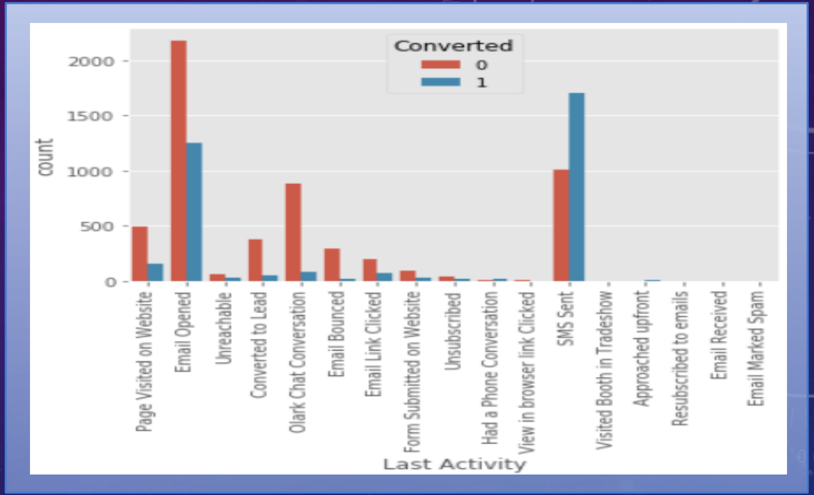


- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum
- API and landing page Submission has less conversion rate(-30%)but counts of the leads from them are considerable
- The count of leads from the Lead Add from is pretty low but conversion rate is very high

EXPLORATORY DATA ANALYST

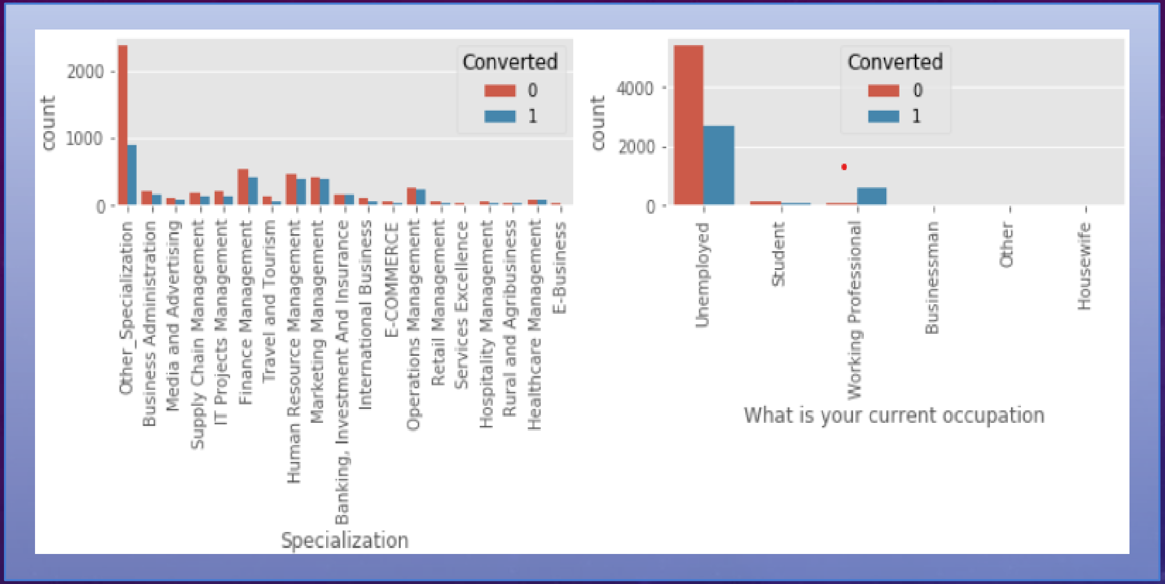


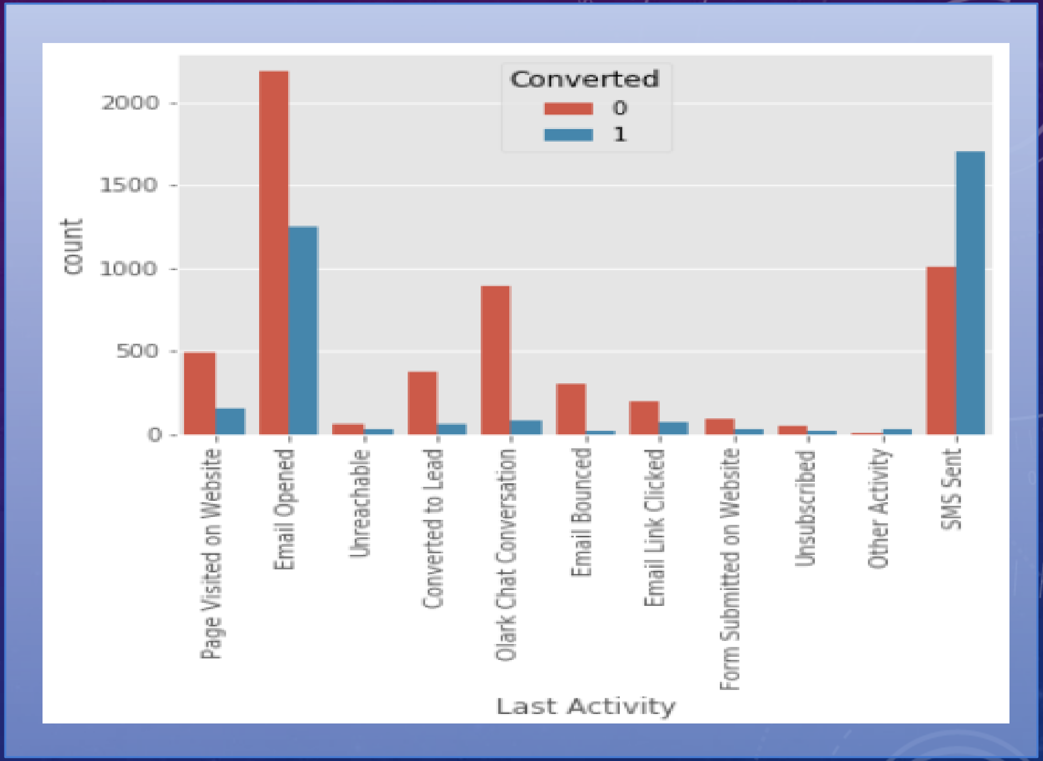
- The median of both the conversion and non-conversion are same and hence nothing conclusive can be use
- User spending more time on the website are more likely to get converted



- The count of leads last activity as ‘Email Opened ’ is maximum
- The conversion rate of SMS sent as last activity is maximum

EXPLORATORY DATA ANALYST





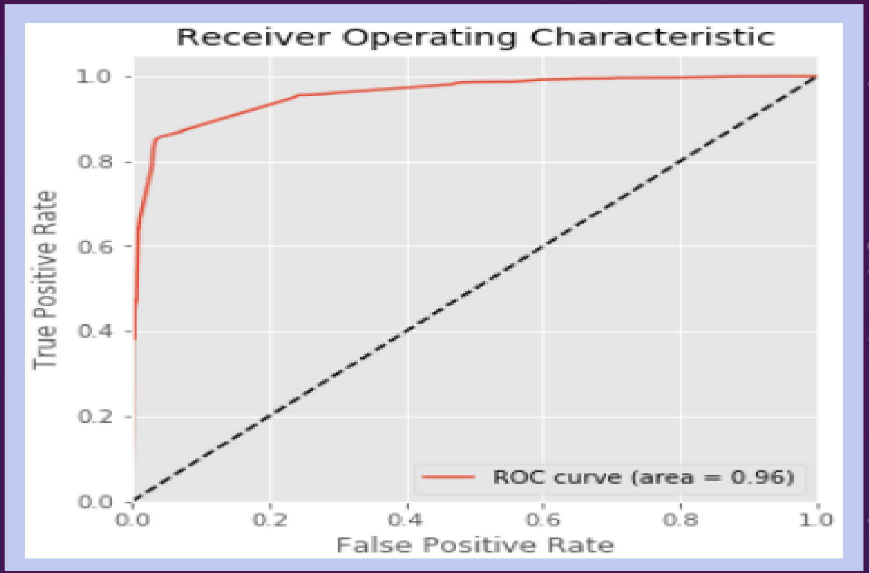
- At the above plot, no particular interface can be made for specialization
- Looking at the above plot, we can say that working professional have high convention rate
- Number of unemployed leads are more than any other category

- “will revert reading the email” and “Closed by Horizzon” has high conversion rate

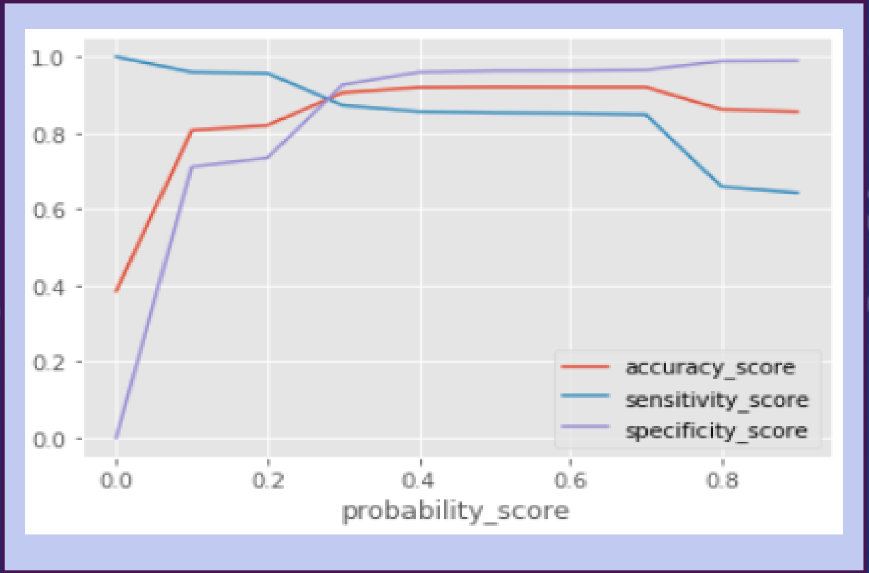
MODEL BULDING

- Splitting the data into test and training sets
- We have chosen the train_testsplit ratio as 70:30
- Using RFE to choose top 15 variables
- Building model by moving the variable whose p-value > 0.05 and vif>5
- Prediction on test dataset
- Overall accuracy is 92.0%

ROC CURVE

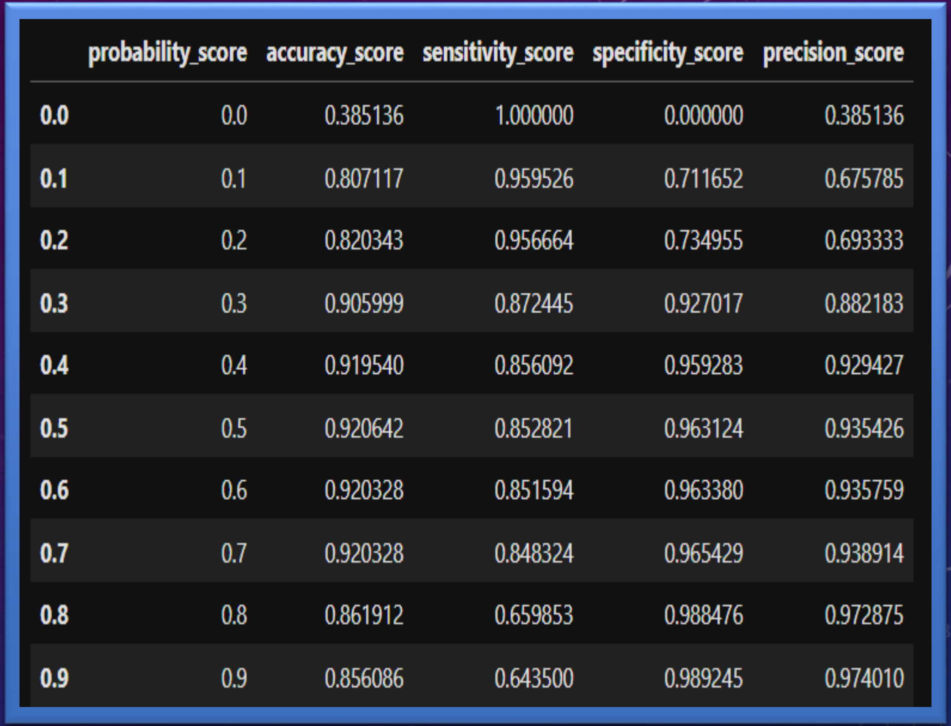


OPTIMAL CUT-OFF



MODEL EVALUATING

- Calculated accuracy, sensitivity and specificity for various cutoff from 0.1 to 0.9
- As per the graph and looking at the other score, it can be seen that the optimal point is 0.27



	probability_score	accuracy_score	sensitivity_score	specificity_score	precision_score
0.0	0.0	0.385136	1.000000	0.000000	0.385136
0.1	0.1	0.807117	0.959526	0.711652	0.675785
0.2	0.2	0.820343	0.956664	0.734955	0.693333
0.3	0.3	0.905999	0.872445	0.927017	0.882183
0.4	0.4	0.919540	0.856092	0.959283	0.929427
0.5	0.5	0.920642	0.852821	0.963124	0.935426
0.6	0.6	0.920328	0.851594	0.963380	0.935759
0.7	0.7	0.920328	0.848324	0.965429	0.938914
0.8	0.8	0.861912	0.659853	0.988476	0.972875
0.9	0.9	0.856086	0.643500	0.989245	0.974010

PREDECTED ACTUAL	NOT CONVERTED	CONVERTED
• NOT CONVERTED • CONVERTED	• 2987 • 918	• 918 • 2322

CONCLUSION

The logistic regression model is used to predict the probability of conversion of a customer.

While we have calculated both sensitive-specification as well as Precision-Recall metrices, we have consider optimal cut on the basic of sensitivity-specification for final prediction

Lead Score calculate shows the conversion rate of final prediction model in around 92% in test data as compared to 95%% in train data

In Business terms, this model has capability to adjust with the company’s requirements in coming future

TOP variables that contributed for lead getting converted in the model are:

- Tags_lost to EINS
- Tags_Closed by Horizzon
- Lead Quality_Worst

Hence Overall this model seems to be good