LEAD SCORING CASE STUDY

PRESENTED BY :-

PARUL MISHRA
MANISH YAMSANI
NIKEETAA JAYARAAJ

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

GOALS OF THE CASE STUDY

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. X education wants to know most promising leads. For that they want to build a Model which identifies the hot leads .Deployment of the model for the future use .

PROBLEM SOLVING METHOD

- READING DATA : To acquire data from a source and place it into their volatile memory for processing.
- DATA CLEANING AND PREPARATION : Read the data from source , convert data into clean format which can be suitable for analysis , removing duplicate data , outliers treatment , Exploratory data analysis .
- SPLITTING THE DATA AND FEATURE SCALING : Splitting the data into train and test dataset , feature scaling of numerical variables .
- MODEL BUILDING : Feature selection using RFE ,VIF ,and p-value , Making predictions , Model evaluation , Determining optimal model using logistic regression .
- RESULT : Evaluate final predictions on test set .

DATA CLEANING & PREPARATION

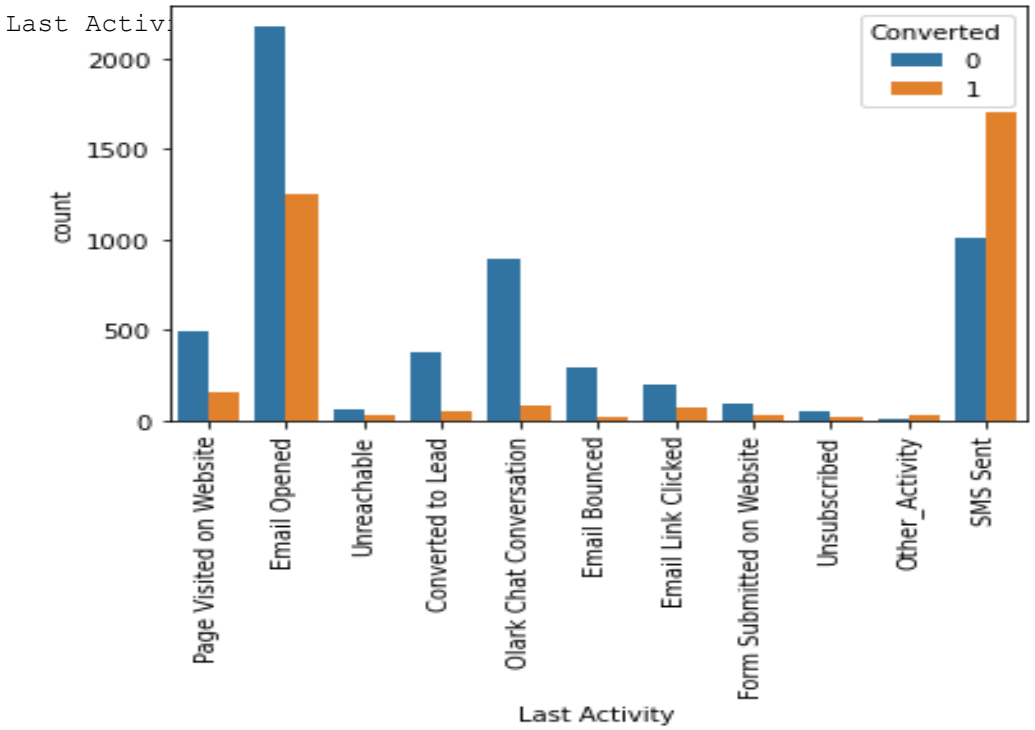
- Handling 'Select' values in some columns.
- Removing rows where a particular column has high missing values.
- Assigning a Unique Category to NULL/SELECT values.
- Imputing NULL values with Mode.
- Remove columns which has only one unique value.
- Outlier Treatment.
- Binary Encoding.

LIBRARIES USED

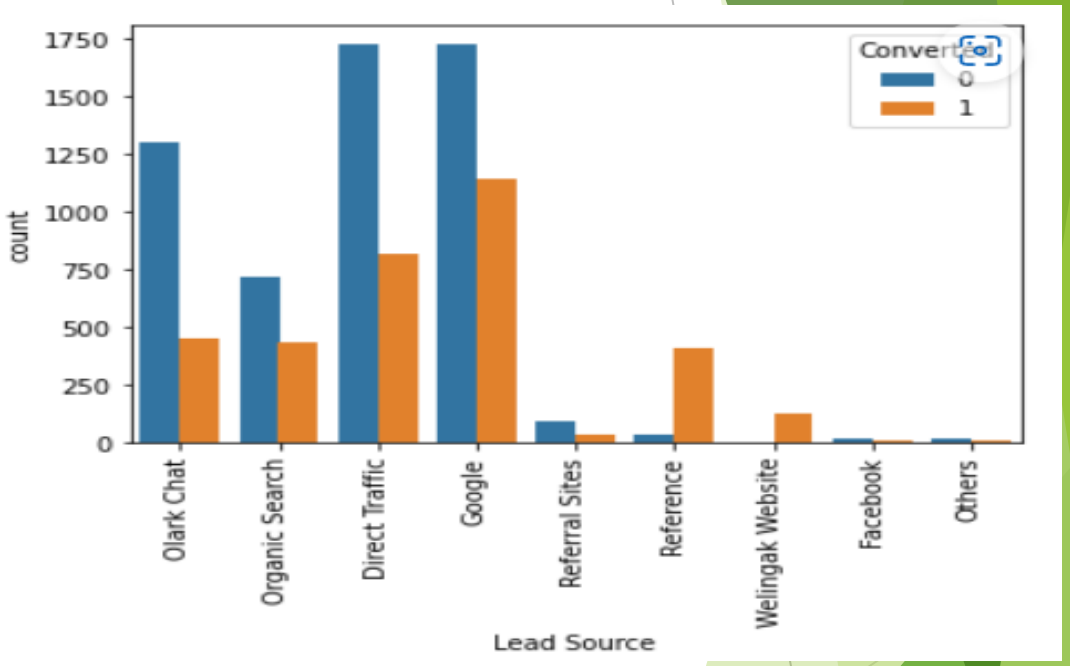
- Numpy
- Pandas
- Matplotlib
- SKLearn
- Seaborn
- Metrics

EXPLORATORY DATA ANALYSIS

Last Activity



Lead Source

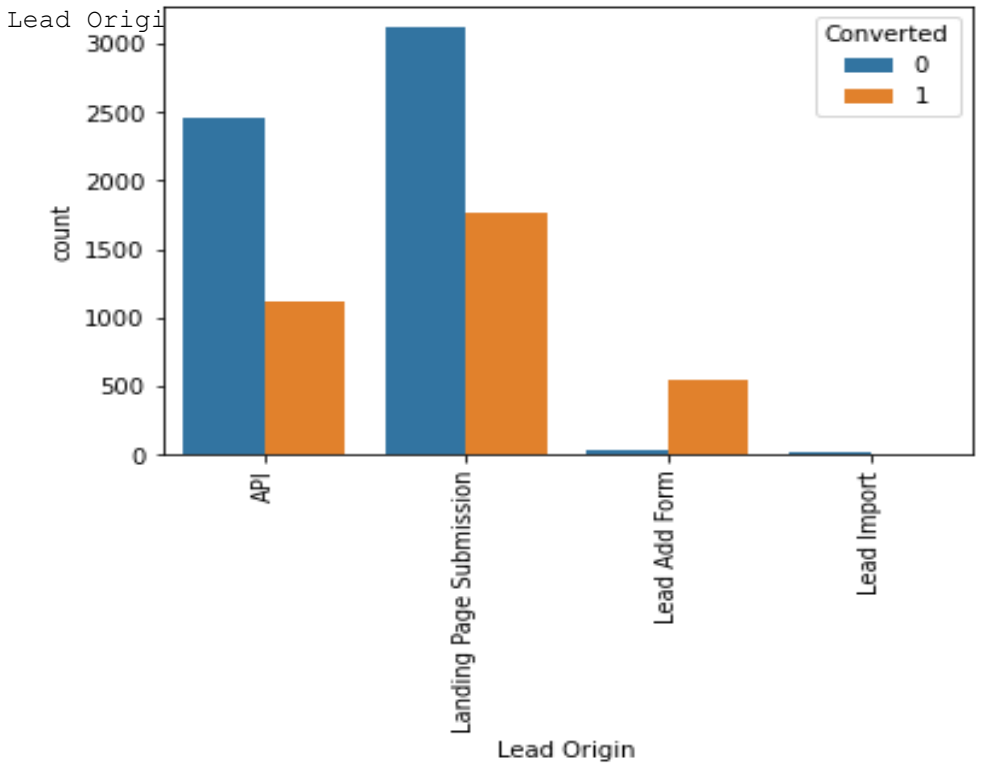


The count of leads last activity as 'Email Opened' is maximum.
The conversion rate of 'SMS Sent' is maximum .

The count of leads from GOOGLE and DIRECT TRAFFIC is maximum.
The count of conversion rate from Google is maximum.

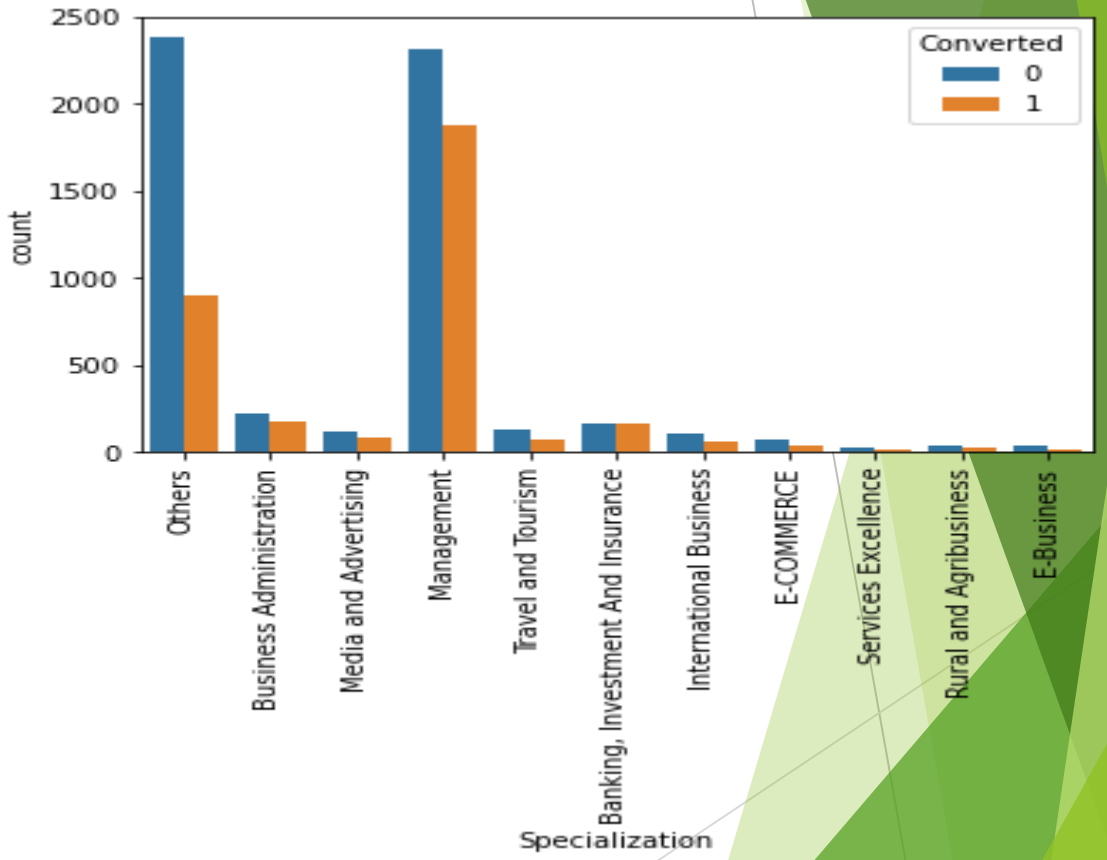
EXPLORATORY DATA ANALYSIS

Lead Origin



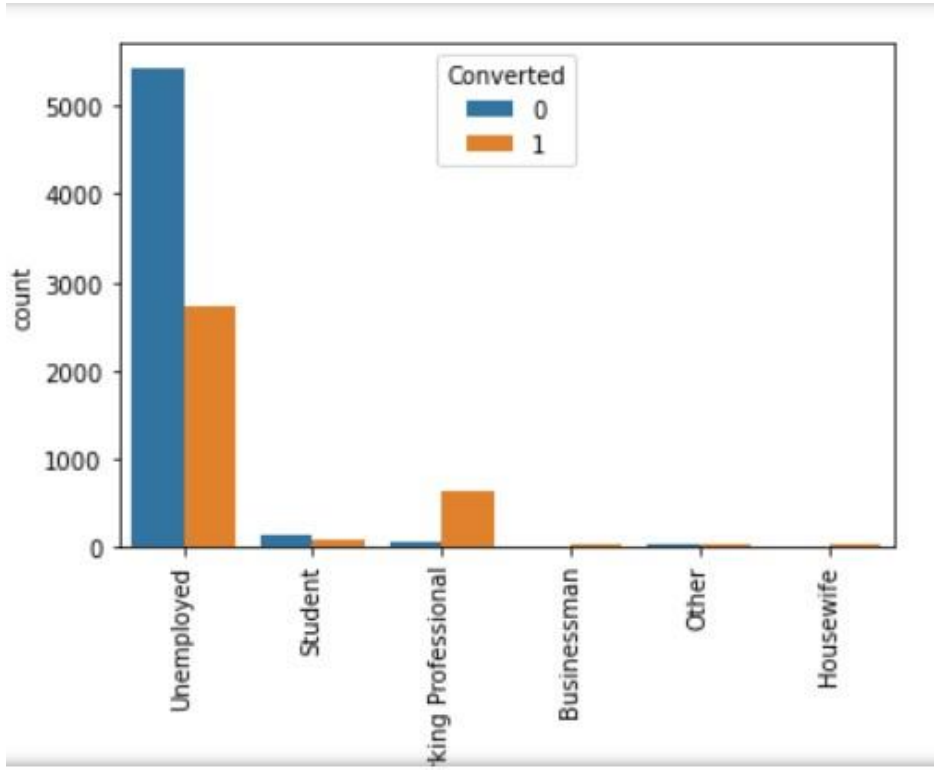
The count of leads origin as ‘Landing page submission’ is maximum.
The conversion rate of same attribute‘Landing page submission’ is maximum .

Specialization



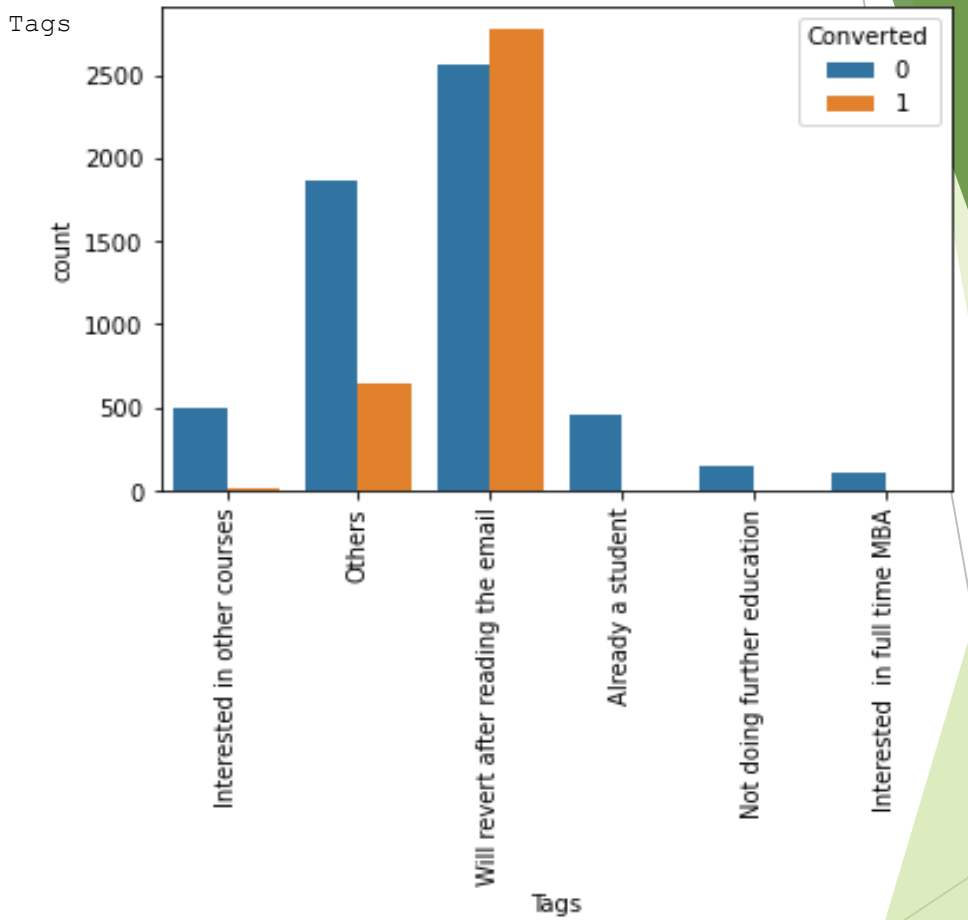
The count of specialization from management is maximum which are potential leads for conversion.

WHAT IS YOUR CURRANT OCCUPATION



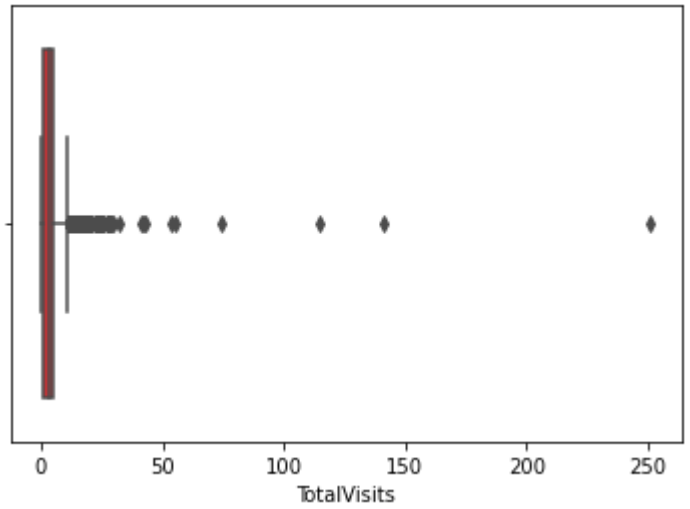
As per the above graph, the conversion of unemployed is higher, followed by **Working professionals**

Tags

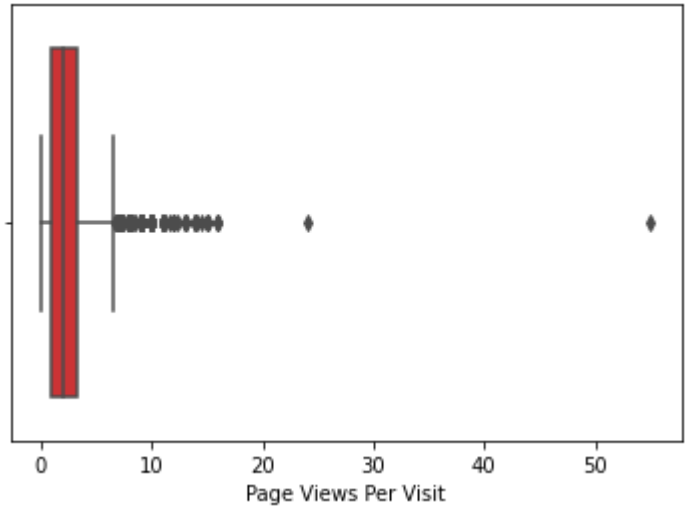


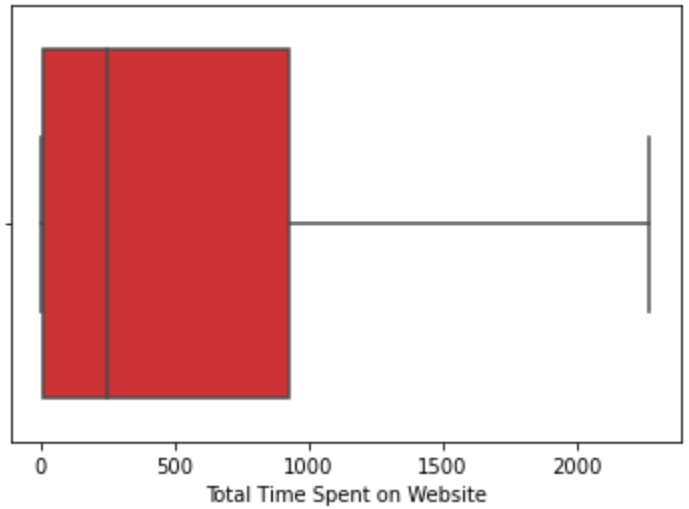
The Tag “Will revert after reading email” has a maximum count of conversion.

EXPLORATORY DATA ANALYSIS



The total number of visits are low





Users spend more time in websites

DATA CONVERSION

- Dummy variables creation is done for the below variable :
  'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization',
  'What is your current occupation', 'City', 'Tags',
  'A free copy of Mastering The Interview'.
- Test Train Split:- Dataset is split into 70% train set and 30% test sets.
- Feature Scaling:- It is done on 'Total Visits', 'Total Time Spent on Website', & 'Page Views Per Visit' by Standardization.

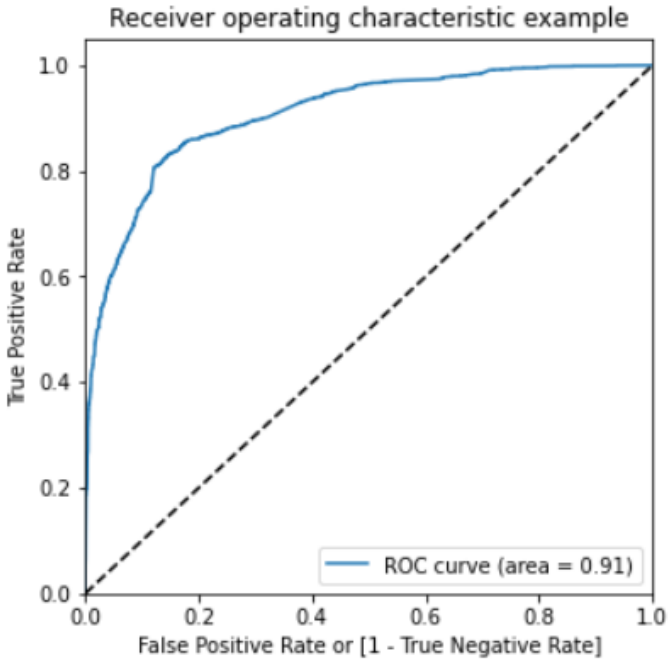
MODEL BUILDING

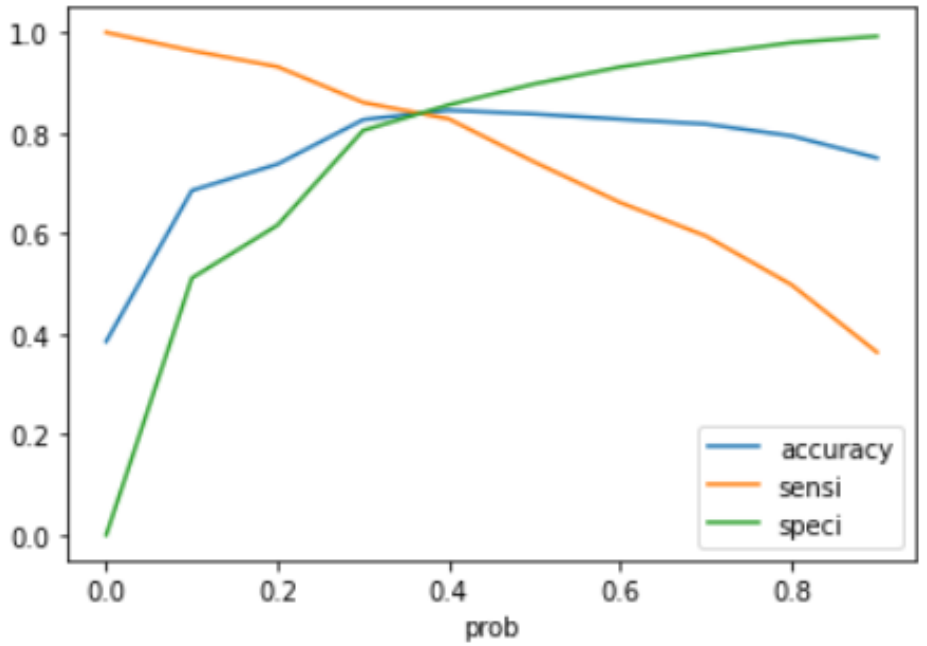
- Recursive Feature Elimination (RFE) method is used for feature selection.
- Top 15 variables are chosen using RFE.
- Model is built by removing variables with p-values greater than 0.05 and RFE values greater than 5.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.1441	0.297	-10.571	0.000	-3.727	-2.561
Do Not Email	-1.9247	0.198	-9.739	0.000	-2.312	-1.537
Total Time Spent on Website	0.9984	0.040	24.833	0.000	0.920	1.077
Lead Origin_Landing Page Submission	-1.6131	0.137	-11.809	0.000	-1.881	-1.345
Lead Source_Reference	3.1010	0.251	12.332	0.000	2.608	3.594
Lead Source_Welingak Website	5.3245	0.737	7.226	0.000	3.880	6.769
Last Activity_Olark Chat Conversation	-1.2077	0.169	-7.147	0.000	-1.539	-0.877
Last Activity_Other_Activity	2.3561	0.538	4.379	0.000	1.302	3.410
Last Activity_SMS Sent	1.2262	0.080	15.339	0.000	1.069	1.383
Last Activity_Unsubscribed	1.5650	0.541	2.893	0.004	0.505	2.625
Specialization_Others	-1.4701	0.138	-10.627	0.000	-1.741	-1.199
What is your current occupation_Working Professional	2.7197	0.235	11.581	0.000	2.259	3.180
Tags_Others	2.5523	0.292	8.736	0.000	1.980	3.125
Tags_Will revert after reading the email	4.1487	0.288	14.402	0.000	3.584	4.713

	Features	VIF
12	Tags_Will revert after reading the email	3.900
2	Lead Origin_Landing Page Submission	3.200
9	Specialization_Others	2.680
11	Tags_Others	2.360
7	Last Activity_SMS Sent	1.650
5	Last Activity_Olark Chat Conversation	1.340
3	Lead Source_Reference	1.320
10	What is your current occupation_Working Profes...	1.220
1	Total Time Spent on Website	1.200
0	Do Not Email	1.180
8	Last Activity_Unsubscribed	1.080
4	Lead Source_Welingak Website	1.070
6	Last Activity_Other_Activity	1.010

MODEL BUILDING





- An ROC curve demonstrates several things:
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by
- a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The value of ROC for our model is 0.91.
- Probability of balanced accuracy, sensitivity and specificity is found to be approximately 0.38

MODEL EVALUATION

Prediction on Train dataset.

Accuracy --0.84
Sensitivity --0.83
specificity --0.85
ROC --0.91

Prediction on Test dataset.

Accuracy --0.84
Sensitivity --0.81
specificity --0.85
ROC --0 .90

As per the findings, the accuracy is 84% and ROC is 91% hence the model is good.

RECOMMENDATIONS

The company should focus on the below features as they have more impact on the conversion of leads

- Lead Source from the Welingak website.
- Lead Source from Reference.
- When their current occupation is as a working professional.
- When the last activity was: SMS Sent.
- Total time spent on the Website.

The Model finds correct promising leads as well as the leads that have less chance of getting converted.

This article provided you with an in-depth understanding of what Lead Scores are, and why they're essential, along with a list of the key attributes that should ideally be included while creating a Lead Scoring framework for your business. Overall this model proves to be accurate.

THANK YOU