Lead score case study report

Business Objective:-

Increasing the Lead Conversion rate from around 30% to around 80%

- Current Lead conversion is around 30%
- Building the right model to identify and classify the most potential leads tagged as ”Hot Leads”
- The conversion rate from the “Hot Leads” should be around 80%

Solution Approach:-

- Data Cleaning / Data inspecting
- Exploratory Data Analysis
- Preparing the data for modelling
- Splitting the data into train-test
- Model Building and evaluation on training dataset
- Select the optimal cut off
- Model building and evaluation on testing dataset

Reading , Understanding and Manipulate the data

- We have 9240 rows and 37 columns in the given dataset for analysis.
- Converted ‘Select’ values to Null (select means user did not select any values in dropdown).
- Before proceeding with data analysis first we dropped columns which had more than 40 % of missing values.
- Dropped columns which had only one unique value.
- Drop all the columns which had High skewed distribution of data.
- Outlier Treatment:- Treated outliers in ‘TotalVisits’ and 'Page Views Per Visit' columns , removed top 1% of data entries rest for other columns no huge outliers were observed.
- Handled “NULL” values across different variables, with respective mean/mode.
- Dummy variables creation for categorical columns
- Splitted the data into Training and Testing datasets and performed feature scaling.
- Dropped the original columns for which the dummy variables have been created.
- As the number of columns were too high , correlation matrix was not feasible so didn’t included one. Highly correlated variables will be handled while variable selection using RFE.

Logistic Regression Model

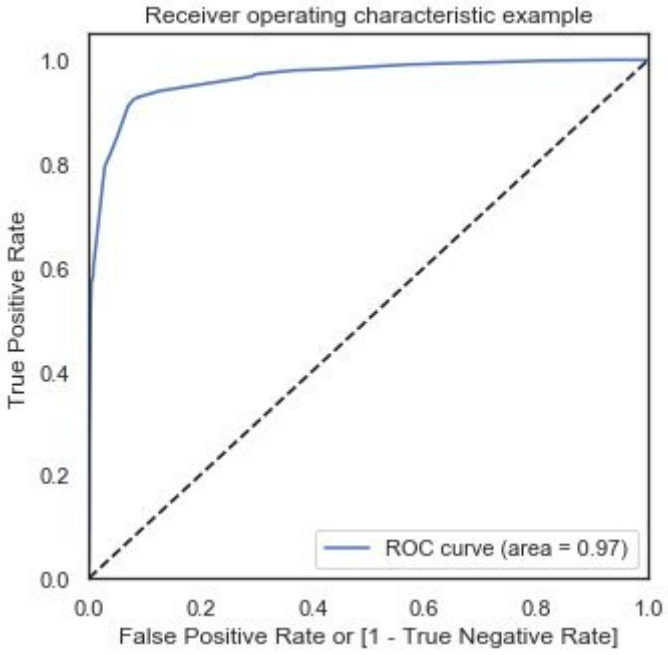
1. After EDA, Logistic Regression Model is built in python using **GLM()** function under statsmodel library.
2. The model contained all the variables, some of which had insignificant coefficients
3. Such variables are removed using Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
4. Manual approach based on VIFs and p values.
5. The final variables with their respective values.
6. Significant p-values near to zero
7. VIFs < 3

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.7812	0.235	-28.843	0.000	-7.242	-6.320
Lead Origin_Lead Add Form	1.4916	0.360	4.139	0.000	0.785	2.198
What is your current occupation_Student	3.9369	0.450	8.746	0.000	3.055	4.819
What is your current occupation_Unemployed	3.7139	0.123	30.149	0.000	3.472	3.955
What is your current occupation_Working Professional	5.0913	0.295	17.245	0.000	4.513	5.670
Last Notable Activity_Modified	-1.3943	0.114	-12.210	0.000	-1.618	-1.170
Last Notable Activity_Olark Chat Conversation	-1.8883	0.397	-4.753	0.000	-2.667	-1.110
Last Activity_Email Bounced	-2.3618	0.407	-5.804	0.000	-3.159	-1.564
Last Activity_SMS Sent	2.0274	0.116	17.458	0.000	1.800	2.255
Tags_Already a student	-2.3312	1.024	-2.277	0.023	-4.338	-0.325
Tags_Busy	2.3710	0.275	8.635	0.000	1.833	2.909
Tags_Closed by Horizzon	8.2370	0.758	10.862	0.000	6.751	9.723
Tags_Lost to EINS	10.3913	0.714	14.556	0.000	8.992	11.790
Tags_Ringing	-1.5644	0.283	-5.529	0.000	-2.119	-1.010
Tags_Will revert after reading the email	4.8084	0.207	23.266	0.000	4.403	5.213
Tags_switched off	-2.6280	0.741	-3.549	0.000	-4.079	-1.177

	Features	VIF
10	Tags_Closed by Horizzon	1.25
1	What is your current occupation_Student	1.15
14	Tags_switched off	1.12
9	Tags_Busy	1.11
6	Last Activity_Email Bounced	1.09
11	Tags_Lost to EINS	1.06
5	Last Notable Activity_Olark Chat Conversation	1.04
3	What is your current occupation_Working Profes...	0.55
0	Lead Origin_Lead Add Form	0.43
8	Tags_Already a student	0.27
2	What is your current occupation_Unemployed	0.16
4	Last Notable Activity_Modified	0.13
12	Tags_Ringing	0.12
7	Last Activity_SMS Sent	0.10
13	Tags_Will revert after reading the email	0.06

Model Evaluation and optimisation

ROC Curve:-



Roc Curve demonstrates:-

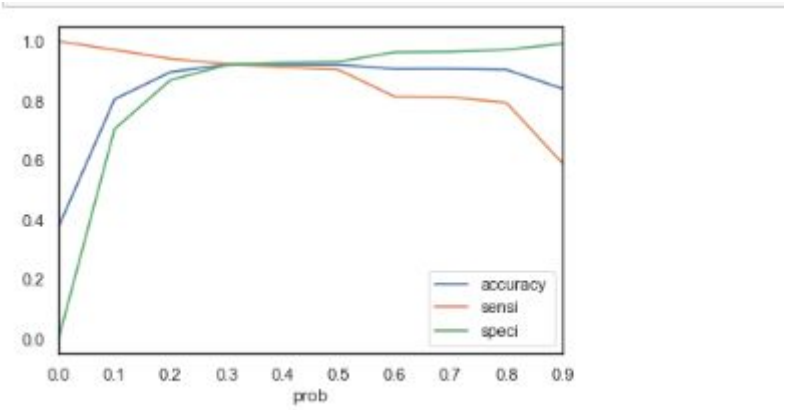
ROC Curves which show the trade off between the True Positive Rate (TPR) and the False Positive Rate (FPR).

Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test.

For our model, ROC curve is towards the upper left corner, and area under the curve is more as displayed in Fig

Finding optimal cut off

• Plotting accuracy, sensitivity and specificity for various Probabilities in fig.



• Cut-Off point is 0.3, where all three coincide, resulting:

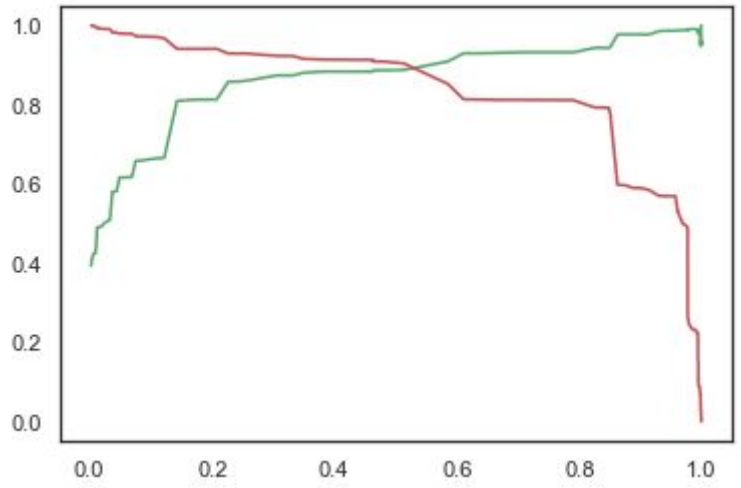
sensitivity = 0.8141106886354035
specificity = 0.9630667345899134
false positive rate = 0.0369332654100866
positive predictive value = 0.930019305019305
Negative predictive value = 0.8957592987443733
Accuracy = 0.9070395677737169

Final model building

• As per business requirement, we have chosen 0.3 as a cut- off value, which gives better result for both accuracy and precision
• Accuracy: 0.9266123054114158

Business Evaluation :

sensitivity = 0.9331395348837209
specificity = 0.9225690276110444
false positive rate = 0.07743097238895558
Precision = 0.8818681318681318
Recall = 0.9570361145703612



precision-recall tradeoff occur due to increasing one of the parameter(**precision** or **recall**) while keeping the model same. This is possible, for instance, by changing the threshold of the classifier.

CONCLUSION:

The model is prepared for prediction of the converted leads. The probability values are generated by the model. The cut-off decided for the model is 0.3. All leads whose probability is generated above this threshold value can be classified as Hot Leads.