

LEAD SCORING CASE STUDY

BY-JITEN RANDHIR KHAHALE and DIVYA SETHI

PROBLEM STATEMENT

X Education sells online courses to industry professionals.

X Education gets a lot of leads, its lead conversion rate is very poor.

For example

If they acquire 100 leads in a day but only about 30 of them are actually converted into a feasible output

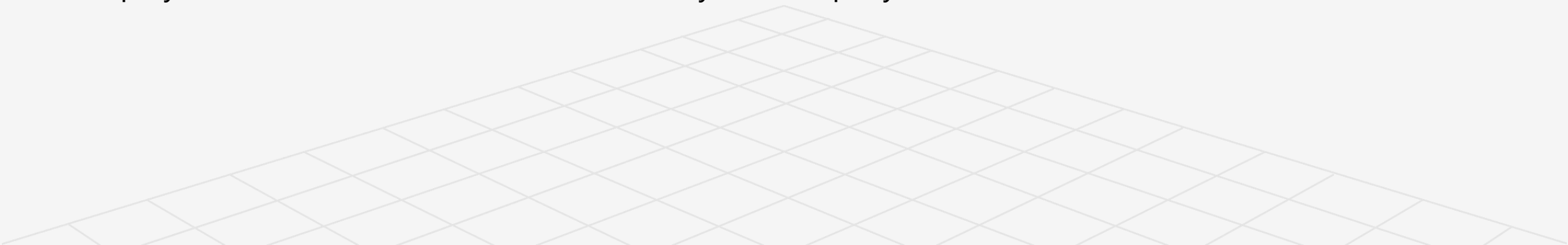
To make this process more efficient, the company wishes to identify the most Potential leads also known as 'Hot Leads'.

If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.



BUSINESS OBJECTIVE

X education wants to know most promising leads for which they want to

- Build a Model which identifies the "hot leads."
 - Create a model in such a way that the customers with "high lead score" have "higher conversion chance" and "low lead score" have "lower conversion chance"
 - The ballpark of the target lead conversion rate is around 80%.
 - The model should be able to adjust if the company's requirement changes in near future.
 - Deployment of the model for the future use by the company
- 

SOLUTION DERIVATION

➤ DATA CLEANING AND DATA MANIPULATION

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

➤ EXPLORATORY DATA ANALYSIS

- Univariate data analysis: value count, distribution of variable etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.

➤ FEATURE SCALING DUMMY VARIABLES AND ENCODING OF THE DATA

- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusion

DATA MANIPULATION

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply"
- Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.

After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are:

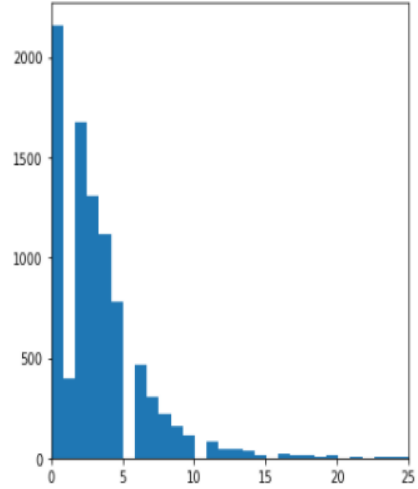
"Do Not Call" , "What matters most to you in choosing course", "Search","Newspaper Article", "X Education Forums", "Newspaper",

"Digital Advertisement"

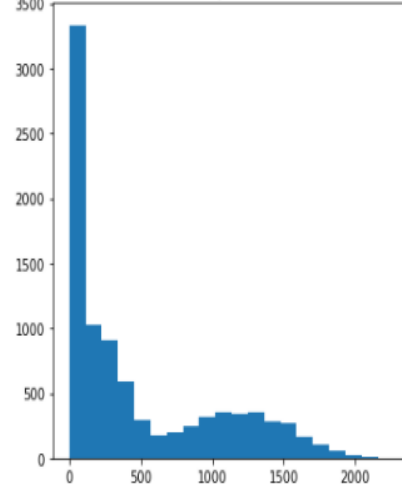
Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

EXPLORATORY DATA ANALYSIS

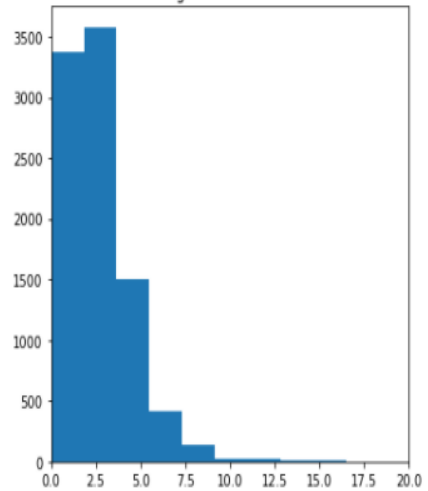
Total Visits



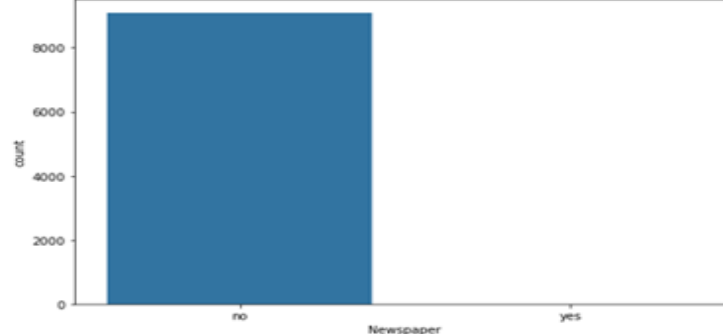
Total Time Spent on Website



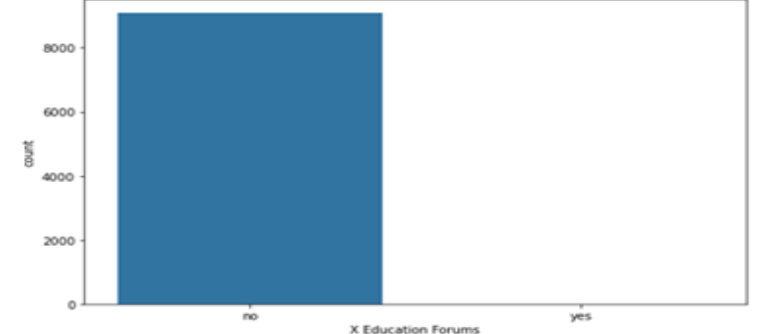
Page Views Per Visit



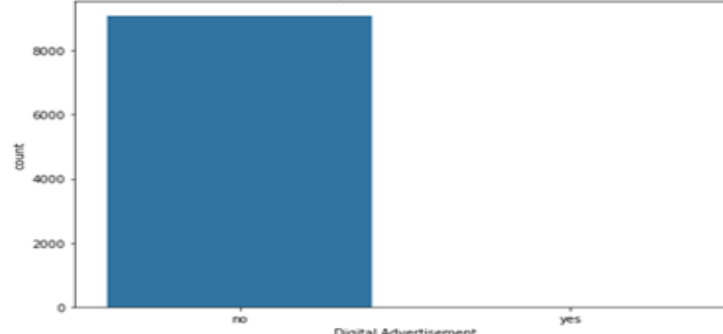
Newspaper



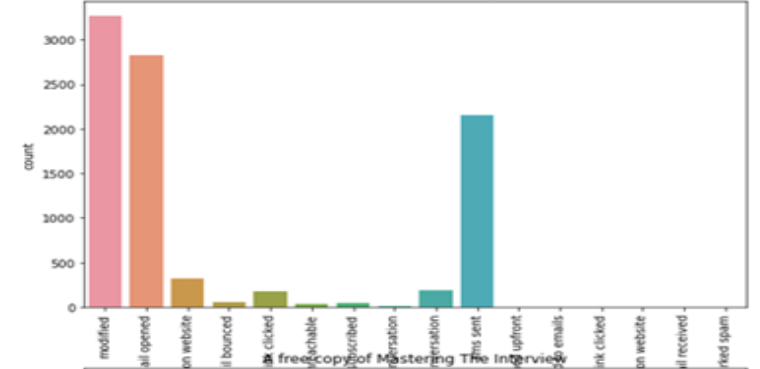
X Education Forums



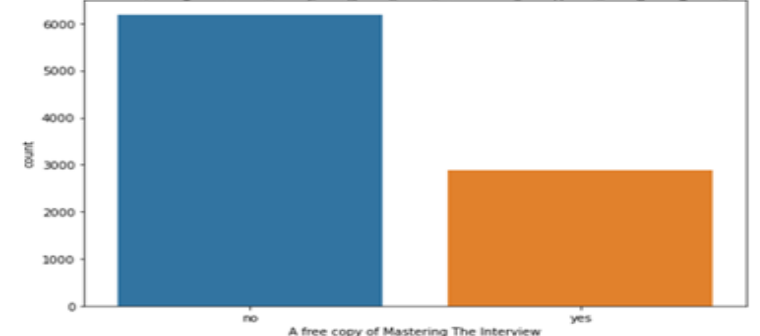
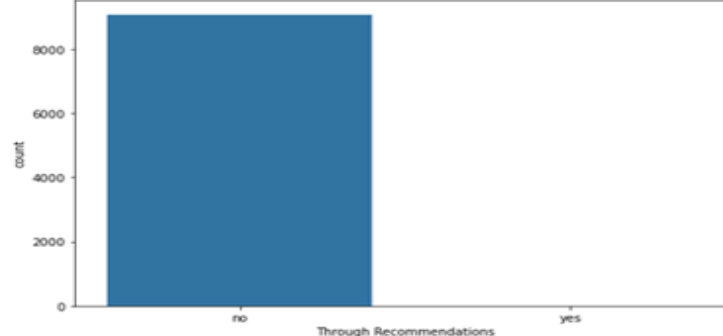
Digital Advertisement

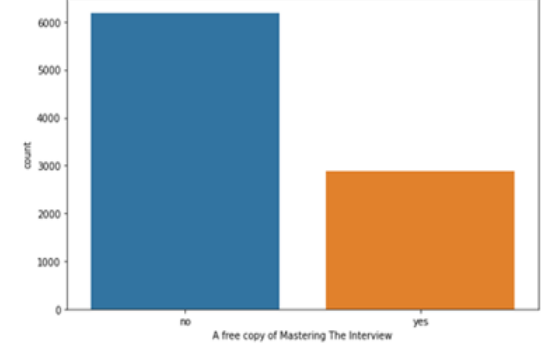
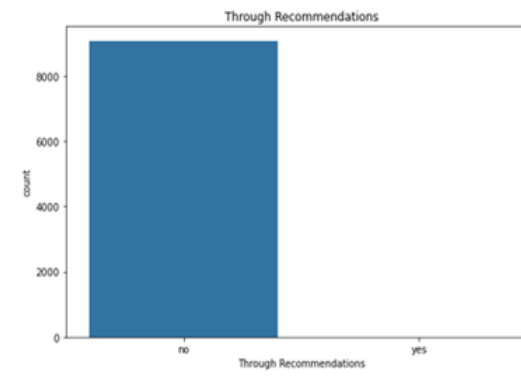
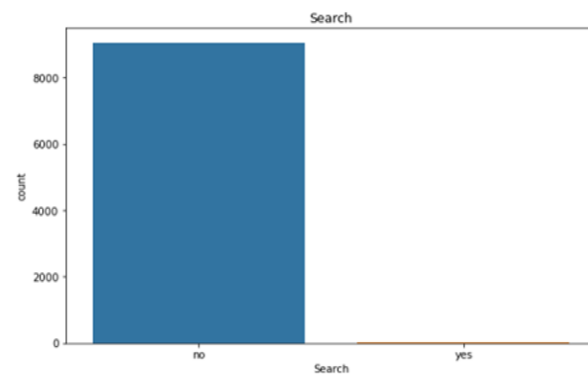
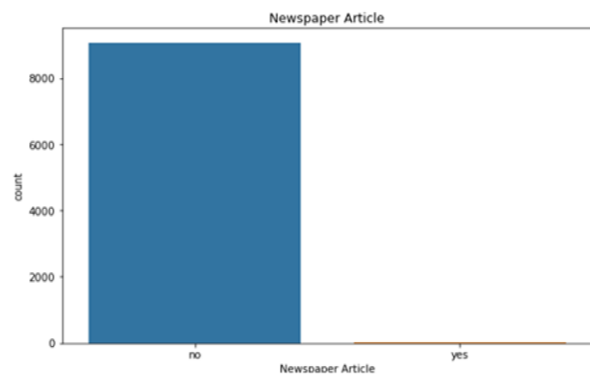
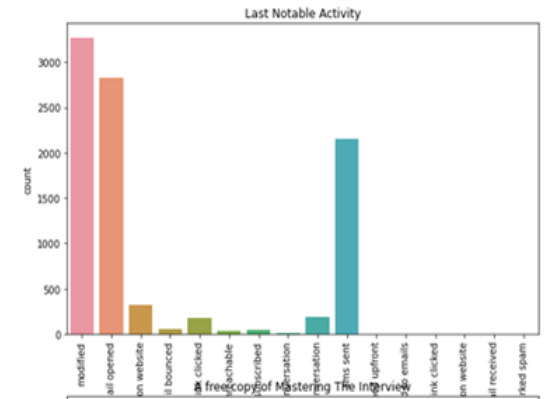
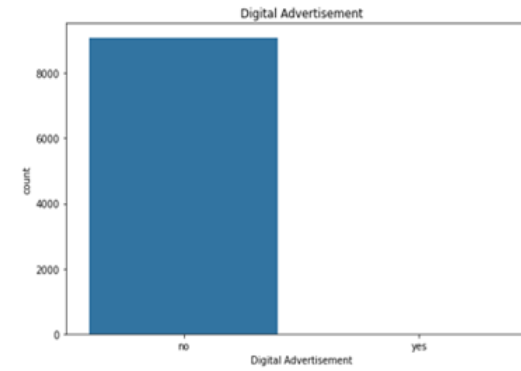
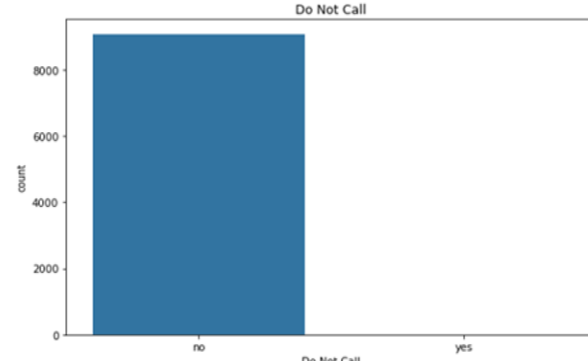
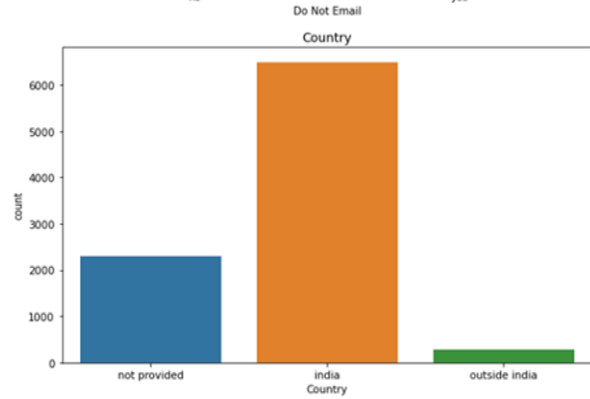
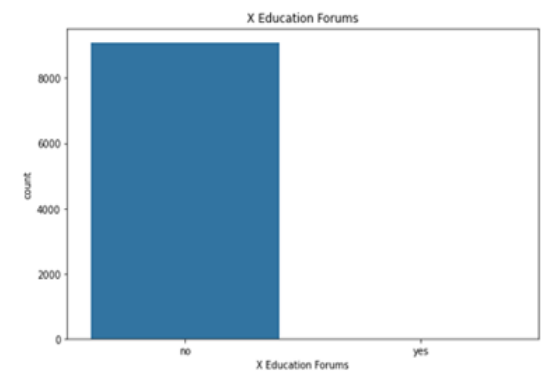
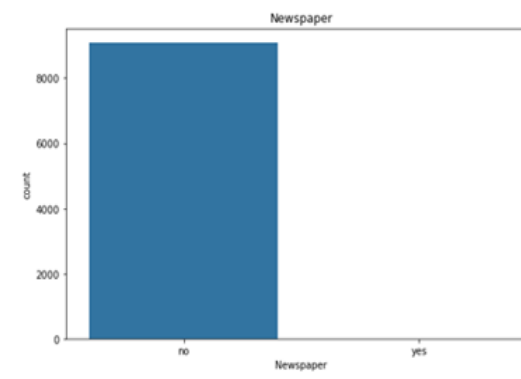
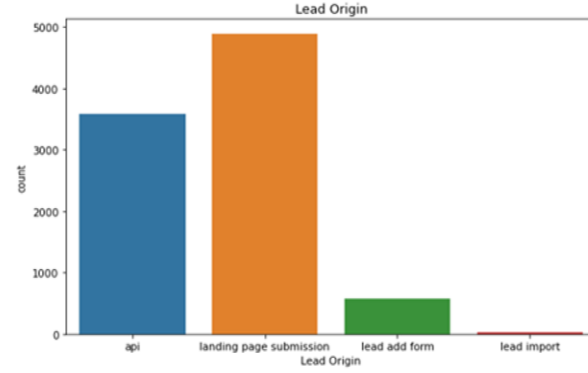
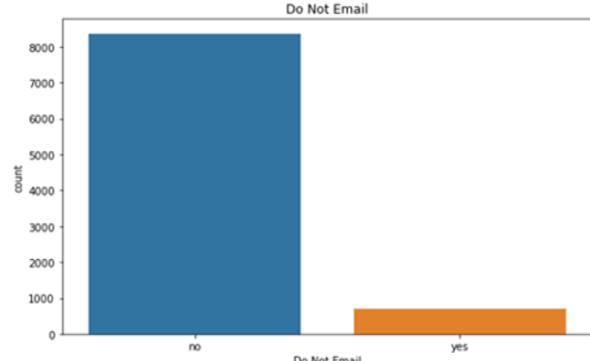


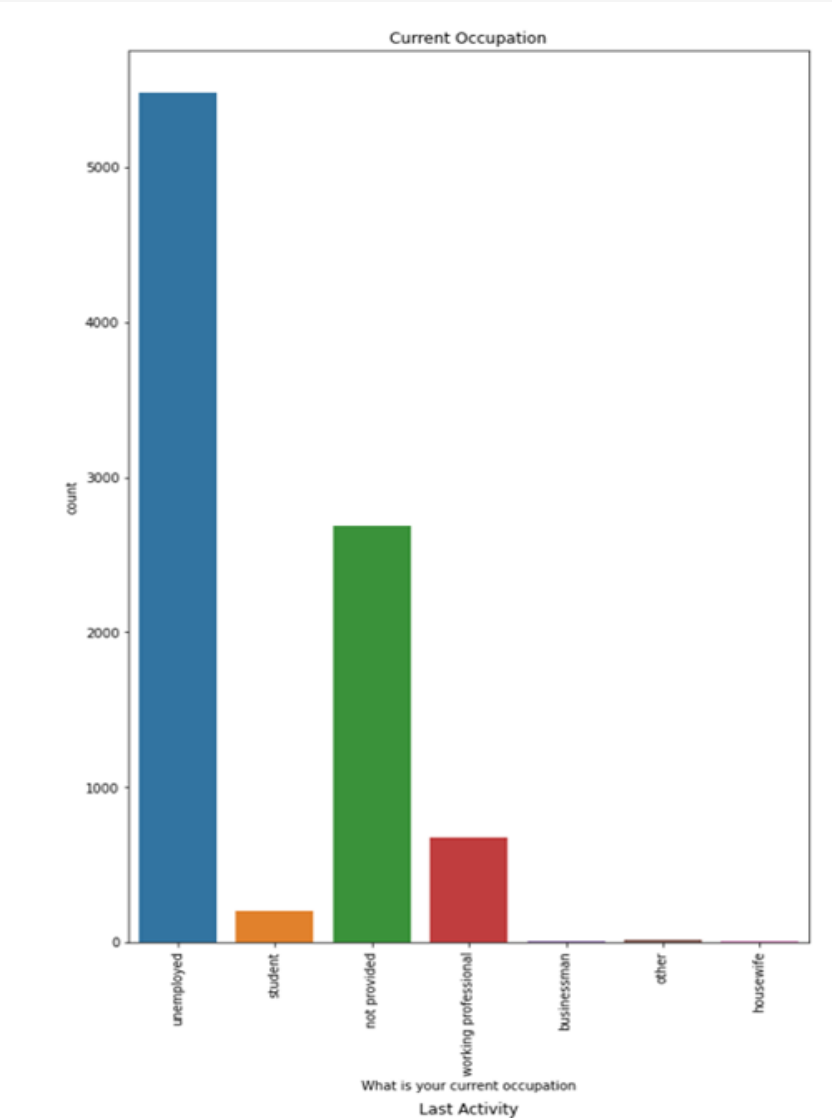
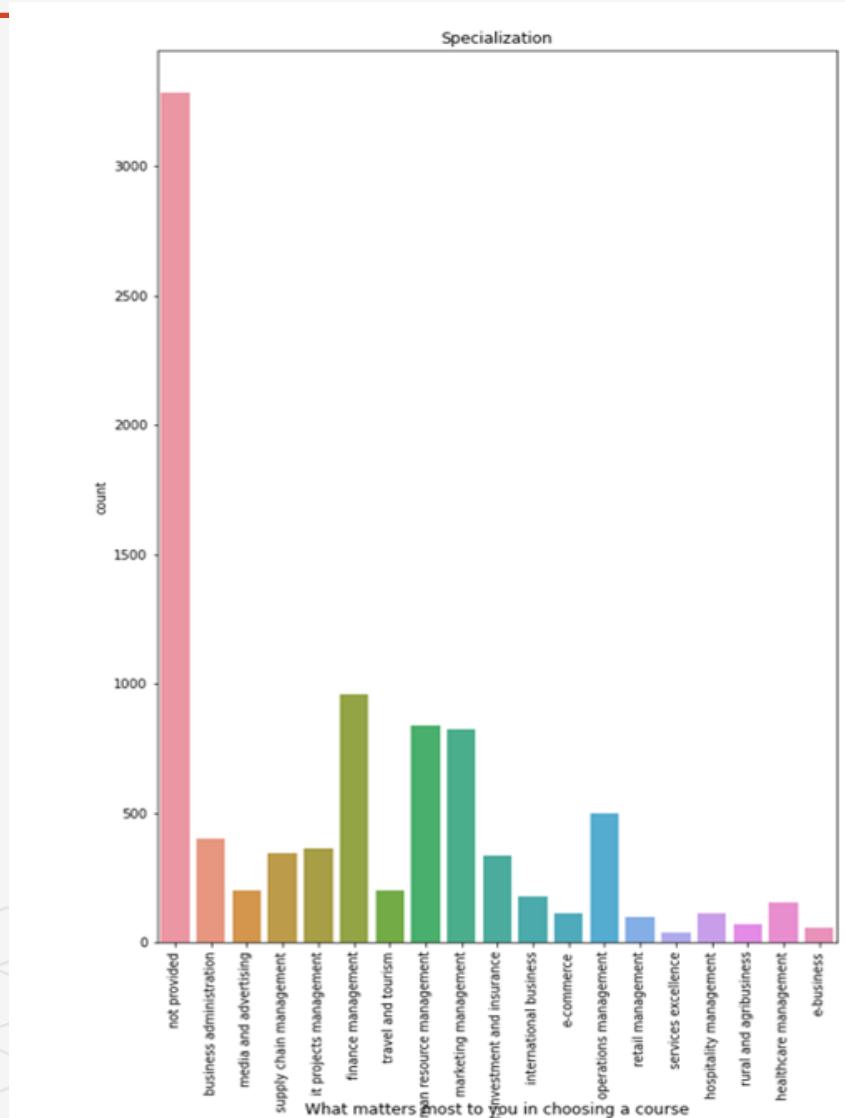
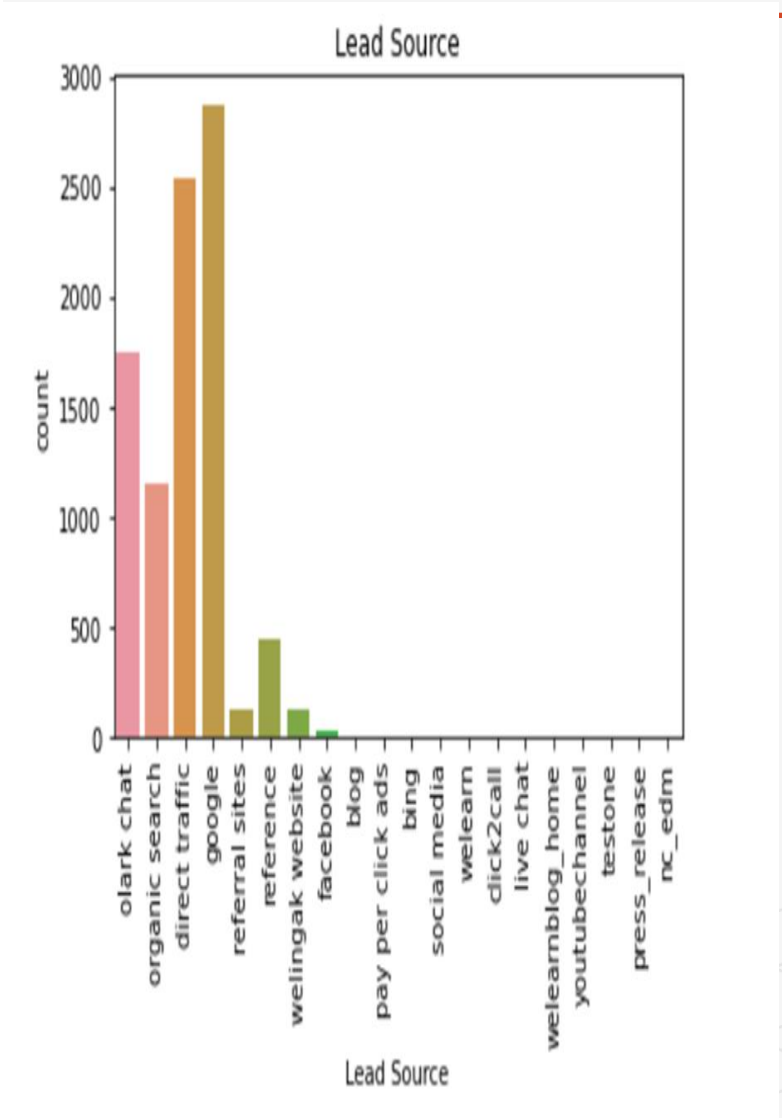
Last Notable Activity



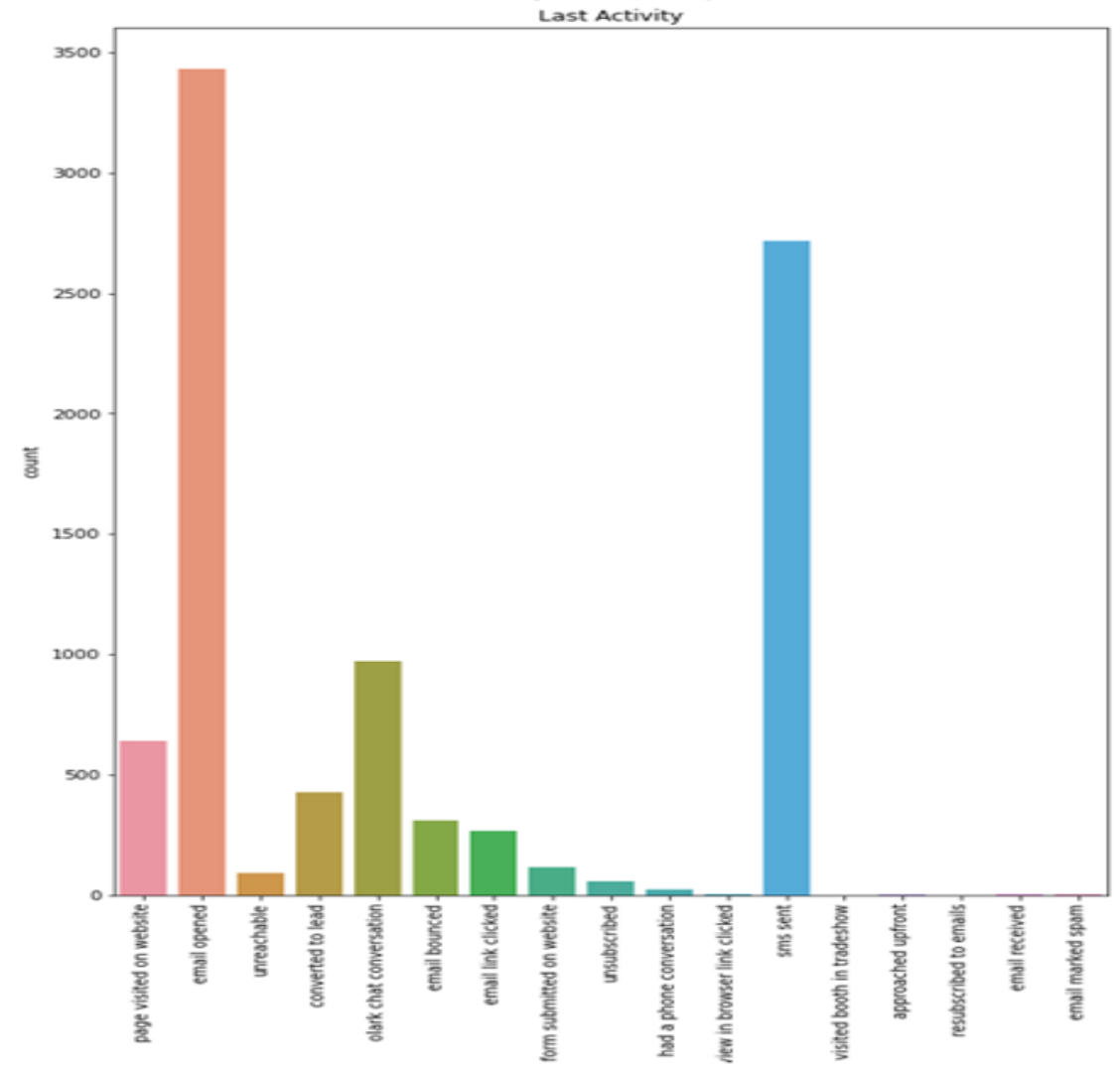
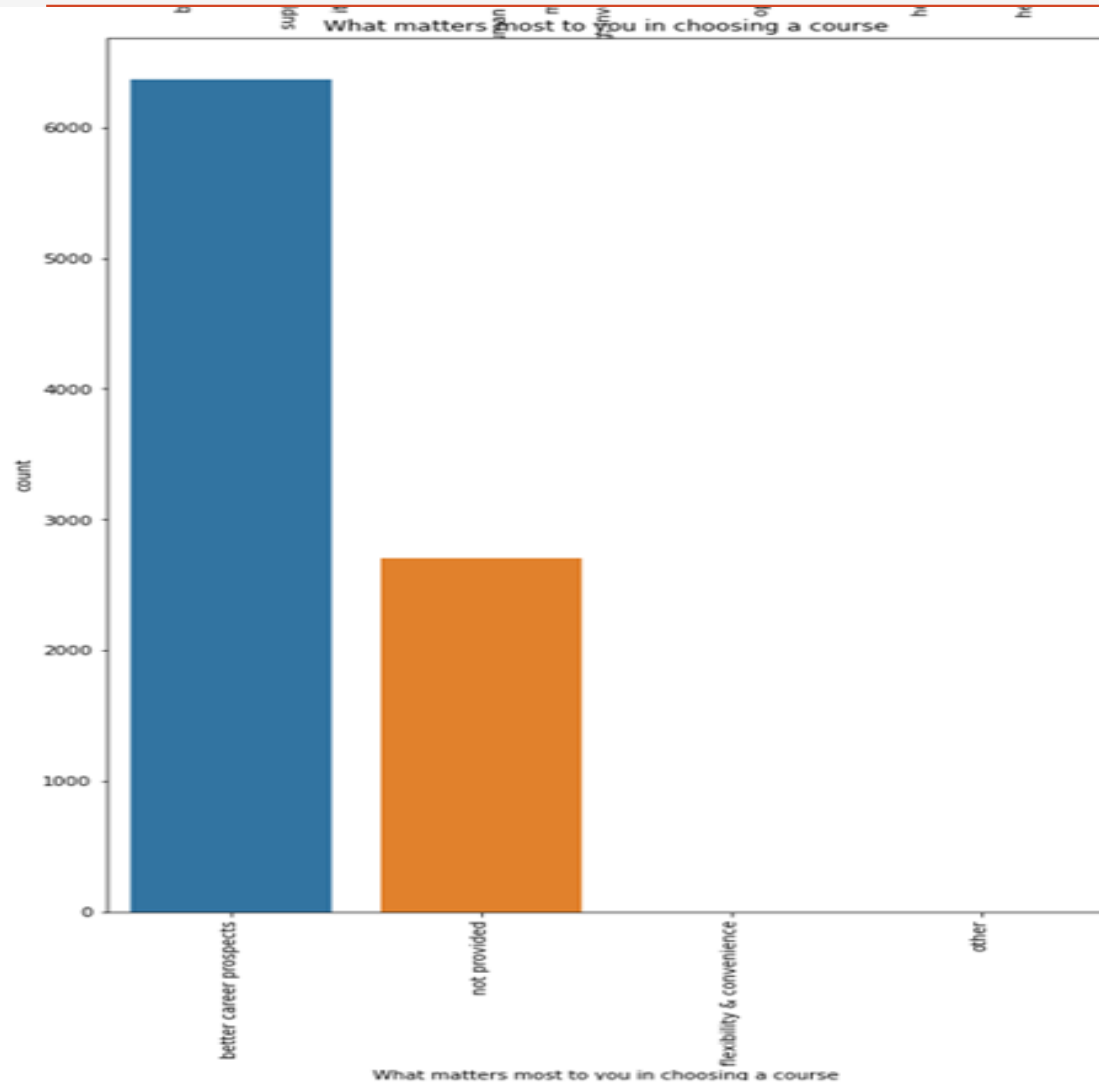
Through Recommendations








CATEGORICAL VARIABLE RELATION



MODEL BUILDING

Splitting the Data into Training and Testing Sets

The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

- Use RFE for Feature Selection
 - Running RFE with 15 variables as output
 - Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5 Predictions on test data set
 - Overall accuracy 81%
- 

FINAL MODEL VISUALIZATION

Generalized Linear Model Regression Results

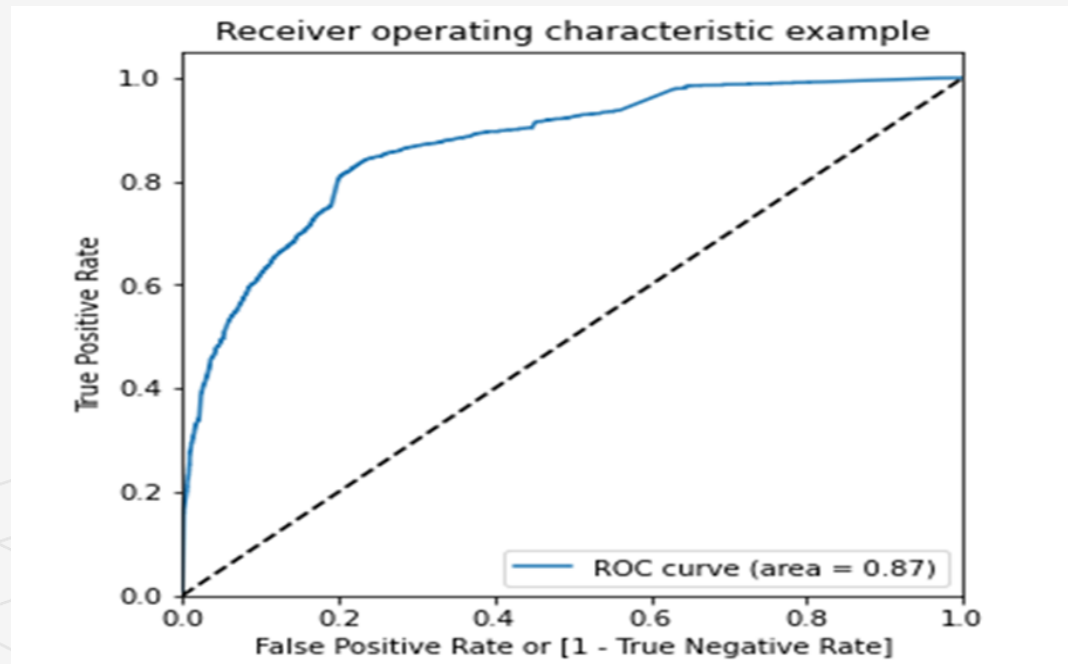
Dep. Variable:	Converted	No. Observations:	6533
Model:	GLM	Df Residuals:	6517
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2785.5
Date:	Mon, 06 Sep 2021	Deviance:	5571.0
Time:	22:02:10	Pearson chi2:	6.71e+03
No. Iterations:	20		
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-3.0298	0.092	-33.033	0.000	-3.210	-2.850
TotalVisits	1.7391	1.488	1.169	0.242	-1.177	4.655
Total Time Spent on Website	3.7351	0.142	26.258	0.000	3.456	4.014
Lead Origin_lead add form	3.3711	0.229	14.748	0.000	2.923	3.819
Lead Source_welingak website	2.5626	1.034	2.477	0.013	0.535	4.590
Do Not Email_yes	-1.5977	0.175	-9.121	0.000	-1.941	-1.254
Last Activity_had a phone conversation	0.8316	0.976	0.852	0.394	-1.082	2.745
Last Activity_sms sent	1.4222	0.071	19.918	0.000	1.282	1.562
Last Activity_unsubscribed	1.6555	0.441	3.751	0.000	0.790	2.521
What is your current occupation_housewife	22.4809	1.23e+04	0.002	0.999	-2.4e+04	2.41e+04
What is your current occupation_other	1.6534	0.686	2.412	0.016	0.310	2.997
What is your current occupation_student	1.0783	0.232	4.647	0.000	0.623	1.533
What is your current occupation_unemployed	1.1661	0.084	13.855	0.000	1.001	1.331
What is your current occupation_working professional	3.7048	0.200	18.487	0.000	3.312	4.098
Last Notable Activity_had a phone conversation	2.2567	1.505	1.500	0.134	-0.693	5.206
Last Notable Activity_unreachable	1.9086	0.528	3.616	0.000	0.874	2.943

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
90%	7.000000	1373.000000	5.000000
99%	17.000000	1839.000000	9.000000
max	251.000000	2272.000000	55.000000

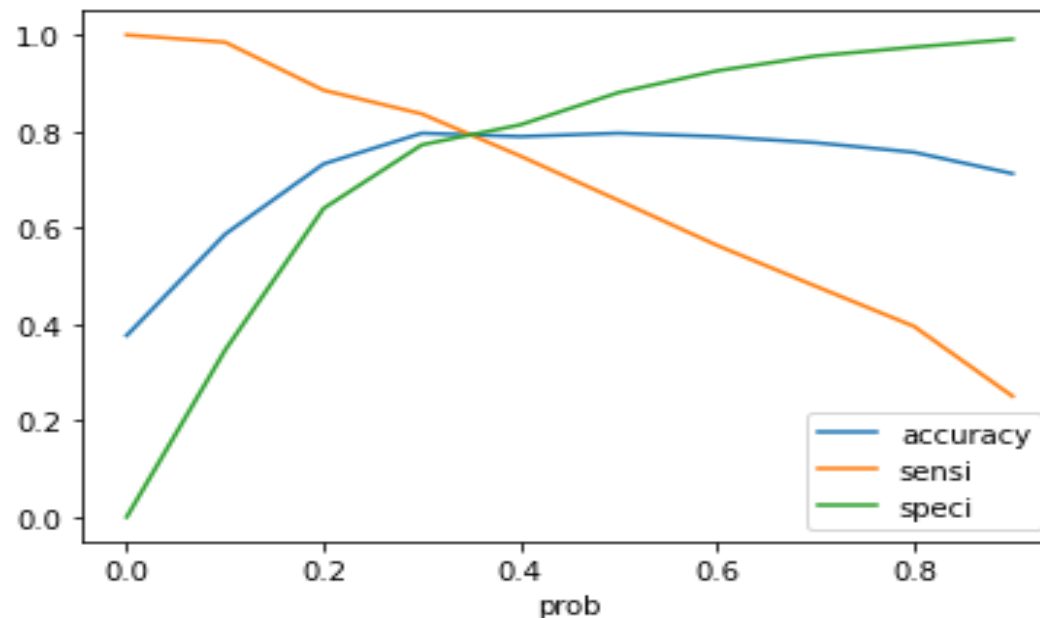
EVALUATING THE MODEL

- After building the final model making prediction on it (on train set), we created ROC curve to find the model stability with auc i.e area under the curve.
- As we can see from the graph plotted on the right side, the area score is 0.87 which is a great score.
- Our graph is leaned towards the left side of the border which means we have good accuracy.



FINDING THE OPTIMAL CUT OFF POINT

- Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each point and analyze which point to choose for probability cutoff.
- We found that on 0.4 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.
- To verify our answer we plotted this in a graph – line plot which is on the right side and we stand corrected that the meeting point is close to 0.4 and hence we choose 0.4 as our optimal probability cutoff.

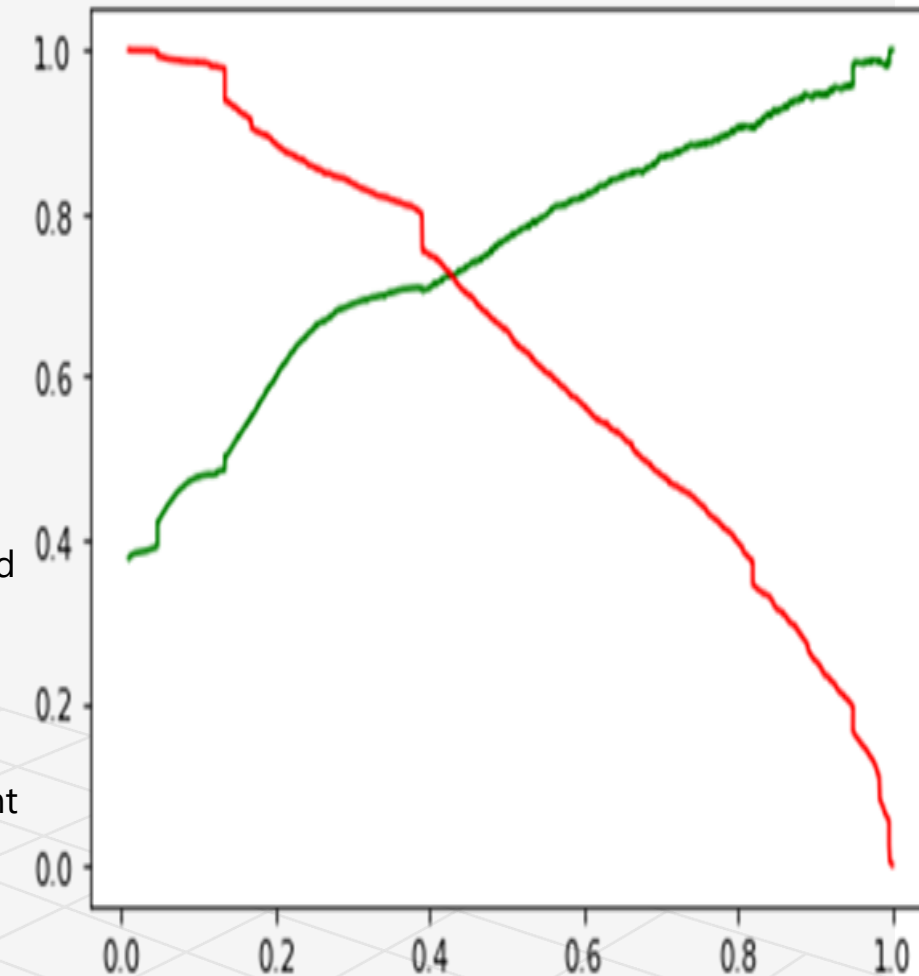


PREDICTION ON TEST SET

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision and recall.
- The accuracy score we found was 0.80, precision 0.81 and recall 0.79 approximately.
- This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

PRECISION AND RECALL

- We used this cutoff point to create a new column in our final dataset for predicting the outcomes.
- After this we did another type of evaluation which is by checking Precision and Recall
- As we all know, Precision and Recall plays very important role in build our model more business oriented and it also tells how our model behaves.
- Hence, we evaluated the precision and recall for this model and found the score as 0.76 for precision and 0.65 for recall.
- Now, recall our business objective - the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision i.e we get more relevant results - as many as hot lead customers from our model .
- We created a graph which will show us the tradeoff between Precision and recall.
- We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.5



CONCLUSION

➤ It was found that the variables that mattered the most in the potential course buyers

- The total time spend on the Website.
- Total number of visits.

➤ When the lead source was:

- Google search
- Direct traffic through various marketing activities
- Organic search
- Official website

➤ When the last activity was:

- SMS
- Olark chat conversation

➤ When the lead origin is Lead add format.

➤ When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses

VALUABLE INSIGHTS

- The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which
- contributes more towards the probability of a lead getting converted are :
- Last Notable Activity_Had a Phone Conversation
- Lead Origin_Lead Add Form and
- What is your current occupation_Working Professional