LEAD SCORE CASE STUDY

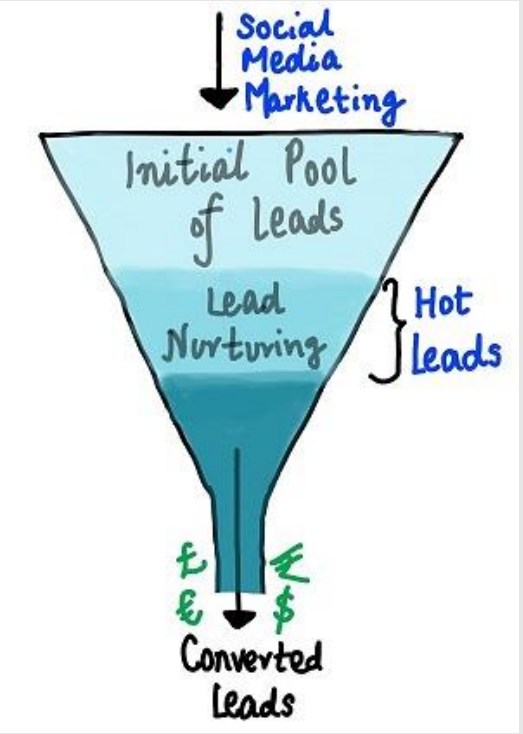
By Samanyu Ghose

Problem Statement

An education company by the name X Education sells online courses to the industry professionals. The company market its courses on several websites and search engines.

Once a visitor land on its websites, they need to fill up form for course or watch some videos, fill e-mail address or phone number to be considered as leads.

Now typically the lead conversion ratio is 30% which is considered very low. So in order to convert more leads, the company wants to identify potential leads which can be considered as 'Hot Leads' and can be followed up rigorously by the sales team rather than making calls to every one. The CEO, in particular, has given a target lead conversion rate be around 80%.



To solve this issue, the company wants to know :-

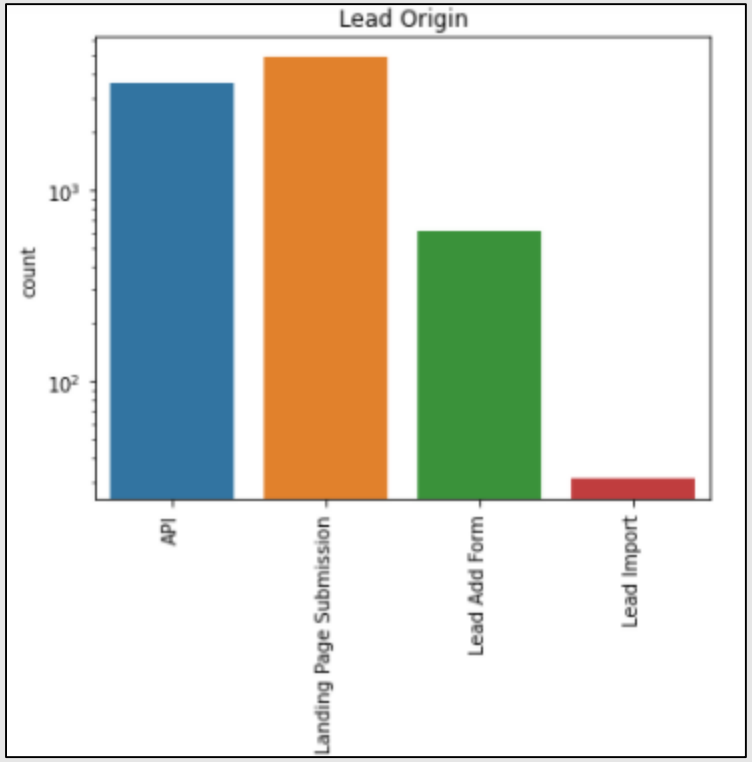
1. Which are the Top 3 variables which contribute most towards the probability of lead getting converted
2. What are the Top 3 categorical/dummy variables which should be focussed the most in order to increase the probability of lead conversion

Analysis Approach

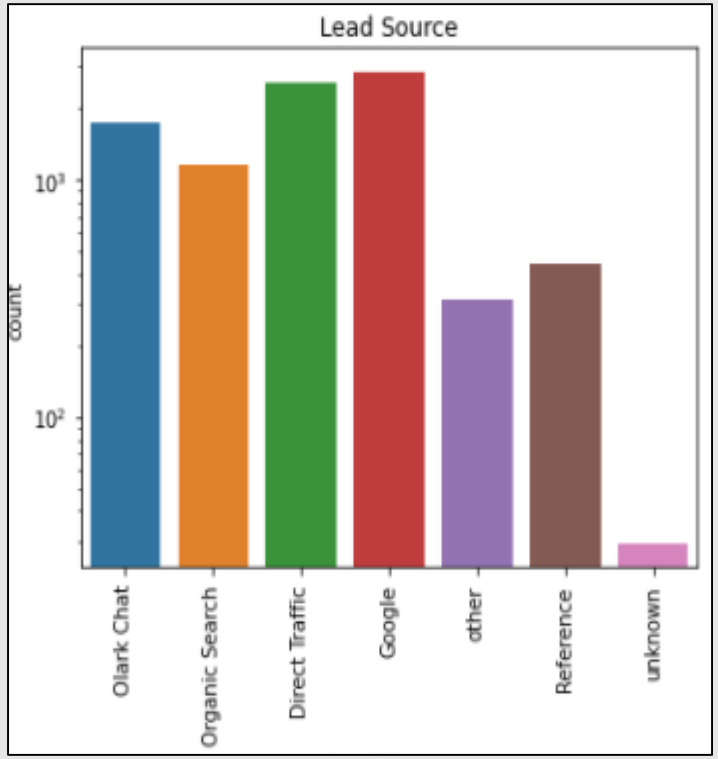
- Data Understanding – From the dataset, we can check that there are 9240 rows and 37 columns out of which 7 are numerical columns and 30 categorical columns

- Data Cleaning –
  a. First we have dropped the Sales generated and unique value generated columns.
  b. Then we have taken out the columns where there is value named as 'Select' and we have replaced this with NAN so that it does not effect our analysis.
  c. Then we have Then we have dropped columns having 45% Null values.
  d. After that we have checked the skewed columns and dropped highly skewed columns with 90% as one value. In some of the columns where the percentage value of two or more categorical variable is very less, we have combined them and replaced them with a value 'other'
  e. We have also dropped rows where the percentage of missing values is more than 70%
  f. Then at last we have imputed numerical columns with the median value as there are outliers observed in those columns
  g. Then we checked percentage of columns and rows and we have found 9103 rows and 9 columns with a 98% of records retained
- EDA – We have performed Univariate , Bivariate analysis and Outlier analysis
- Then we did Train-Test Split and scaled the data through Standardized method
- Data Modelling – Then we have started doing data modelling through mixed analysis i.e first through RFE and then taken manual approach to find the optimal set of data through GLM and VIF to reduce the number of variable less than equal to 15. We have found 10 variable for the data set after modelling
- Model Evaluation – We did model evaluation by taking conversion probability of >0.35 and taken out the Accuracy ,Confusion, Sensitivity, Specificity before Plotting the ROC curve, after taking the optimal cut off value of ROC curve at 30%. We have found our sensitivity value after plotting the ROC curve as **77.8%** which quite good as per the sensitivity matrix. Also we did Precision and Recall trade off and made predictions on Test dataset where we have found Sensitivity as **77.5%**
- Also we have taken out the F1 score to find the harmonic mean between Precision and Recall and we have found F1 score as **73%**

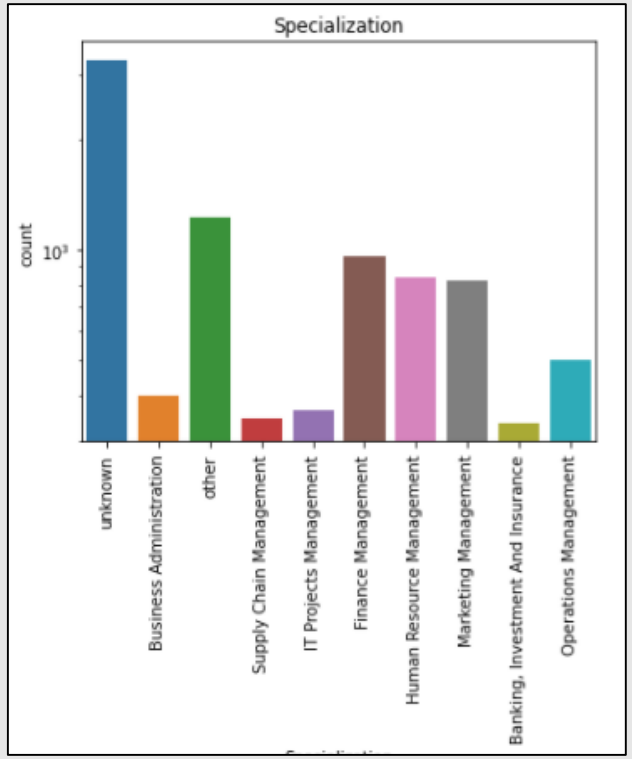
Visualization of Categorical Data



Lead Origin



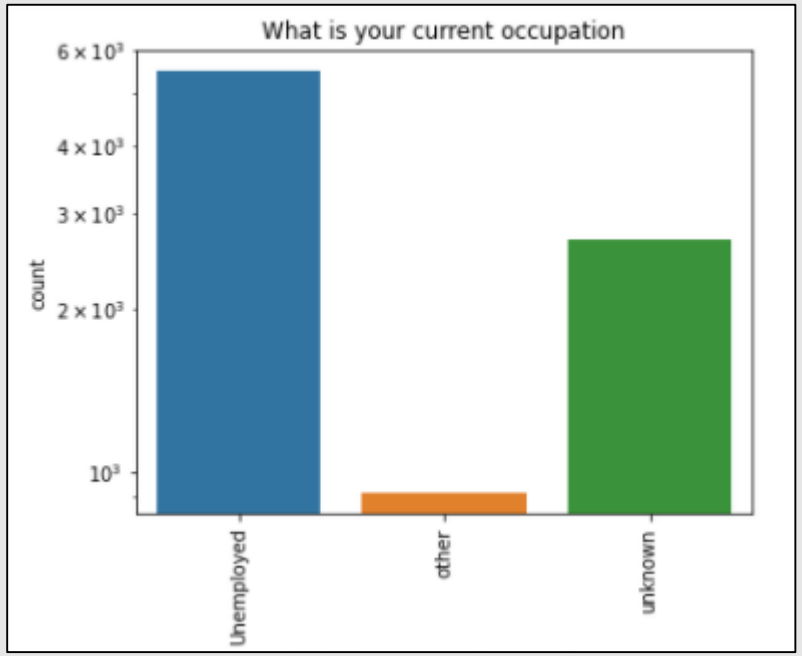
Lead Source



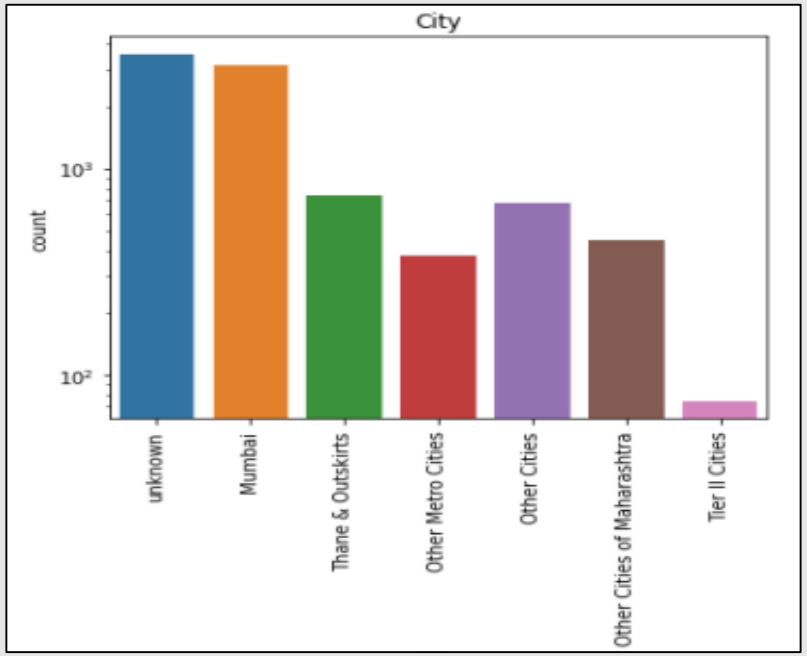
Specialization

a. In 'Lead Origin', the 'Landing Page Submission' is the highest
b. In 'Lead Source', 'Google' has the highest source
c. In 'Specialization', the 'unknown' value is the highest

Visualization of Categorical Data (Continued)



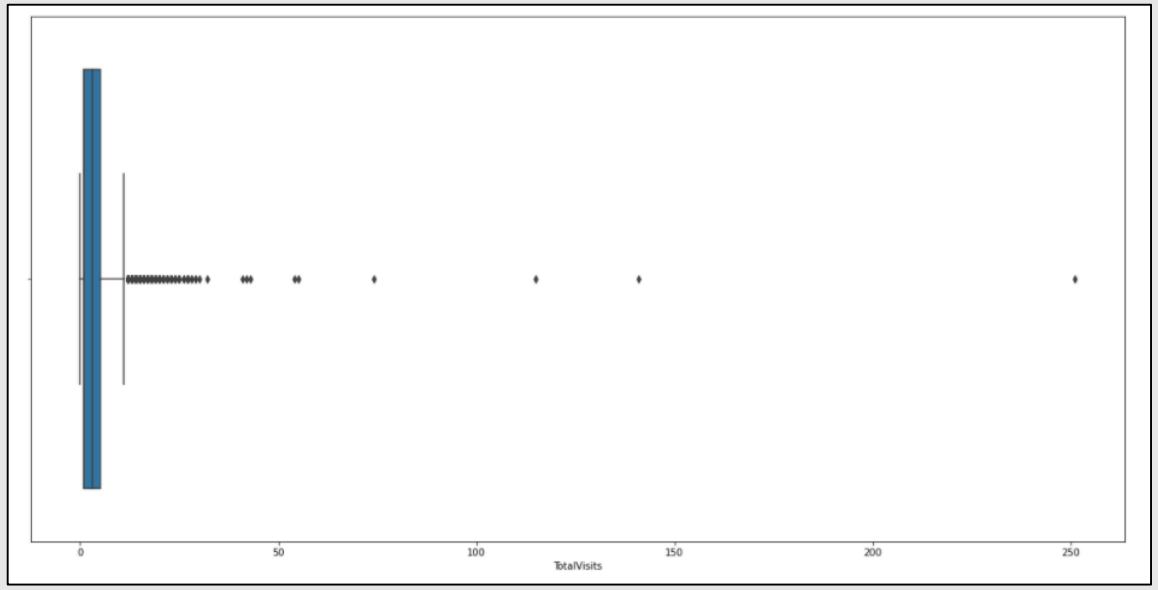
What is your current occupation

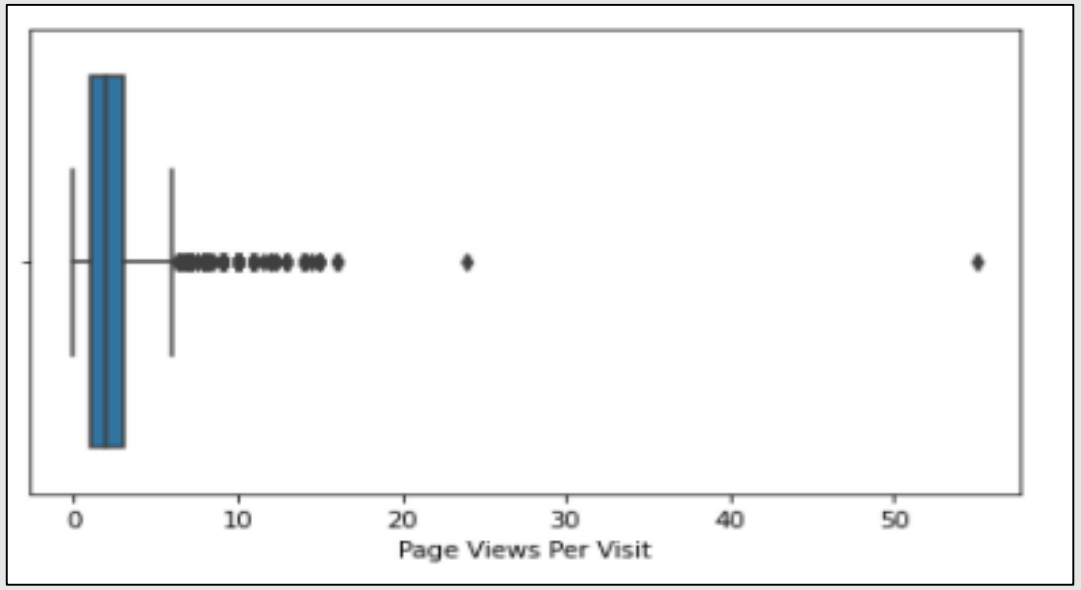


City

a. In 'What is your current occupation', 'Unemployed' is the highest
b. In 'City', the lowest number of customer comes from 'Tier II Cities' and the 'unknown' value is the highest.

Outlier analysis on numerical columns





TotalVisits

Page Views Per Visit

a. There are number of outliers in 'Total Visits' and ‘Page Views Per Visit’

GLM Regression Result and VIF score

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6361
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2934.5
Date:	Sun, 06 Dec 2020	Deviance:	5868.9
Time:	12:08:23	Pearson chi2:	6.33e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3689	0.135	-2.725	0.006	-0.634	-0.104
Total Time Spent on Website	1.0796	0.038	28.571	0.000	1.006	1.154
Lead Origin_Landing Page Submission	-0.6228	0.120	-5.177	0.000	-0.859	-0.387
Lead Origin_Lead Add Form	4.0647	0.229	17.764	0.000	3.616	4.513
Lead Source_Google	0.3621	0.087	4.150	0.000	0.191	0.533
Lead Source_Olark Chat	1.1705	0.137	8.526	0.000	0.901	1.440
Lead Source_Organic Search	0.2467	0.111	2.227	0.026	0.030	0.464
Lead Source_other	0.5034	0.215	2.340	0.019	0.082	0.925
Specialization_unknown	-0.9089	0.112	-8.141	0.000	-1.128	-0.690
What is your current occupation_other	1.4338	0.119	12.070	0.000	1.201	1.667
What is your current occupation_unknown	-1.2412	0.084	-14.804	0.000	-1.406	-1.077

GLM Regression

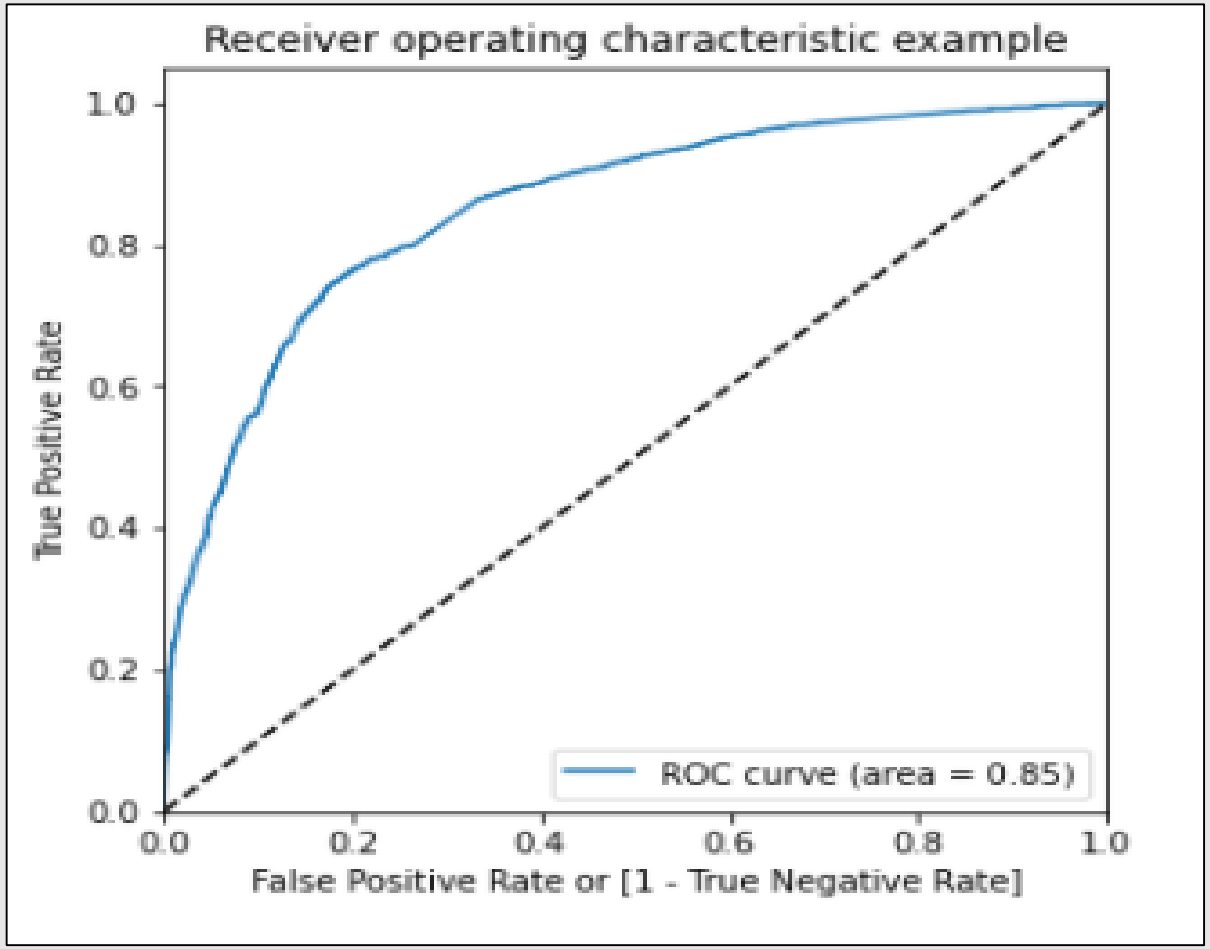
	Features	VIF
7	Specialization_unknown	2.76
4	Lead Source_Olark Chat	2.20
1	Lead Origin_Landing Page Submission	1.79
3	Lead Source_Google	1.77
9	What is your current occupation_unknown	1.59
5	Lead Source_Organic Search	1.31
0	Total Time Spent on Website	1.28
2	Lead Origin_Lead Add Form	1.28
6	Lead Source_other	1.23
8	What is your current occupation_other	1.18

VIF Score

Here we have GLM result where the P value of all variables is less than 0.05 and also the VIF score of all variable is less than 5 which is the ideal condition to go with

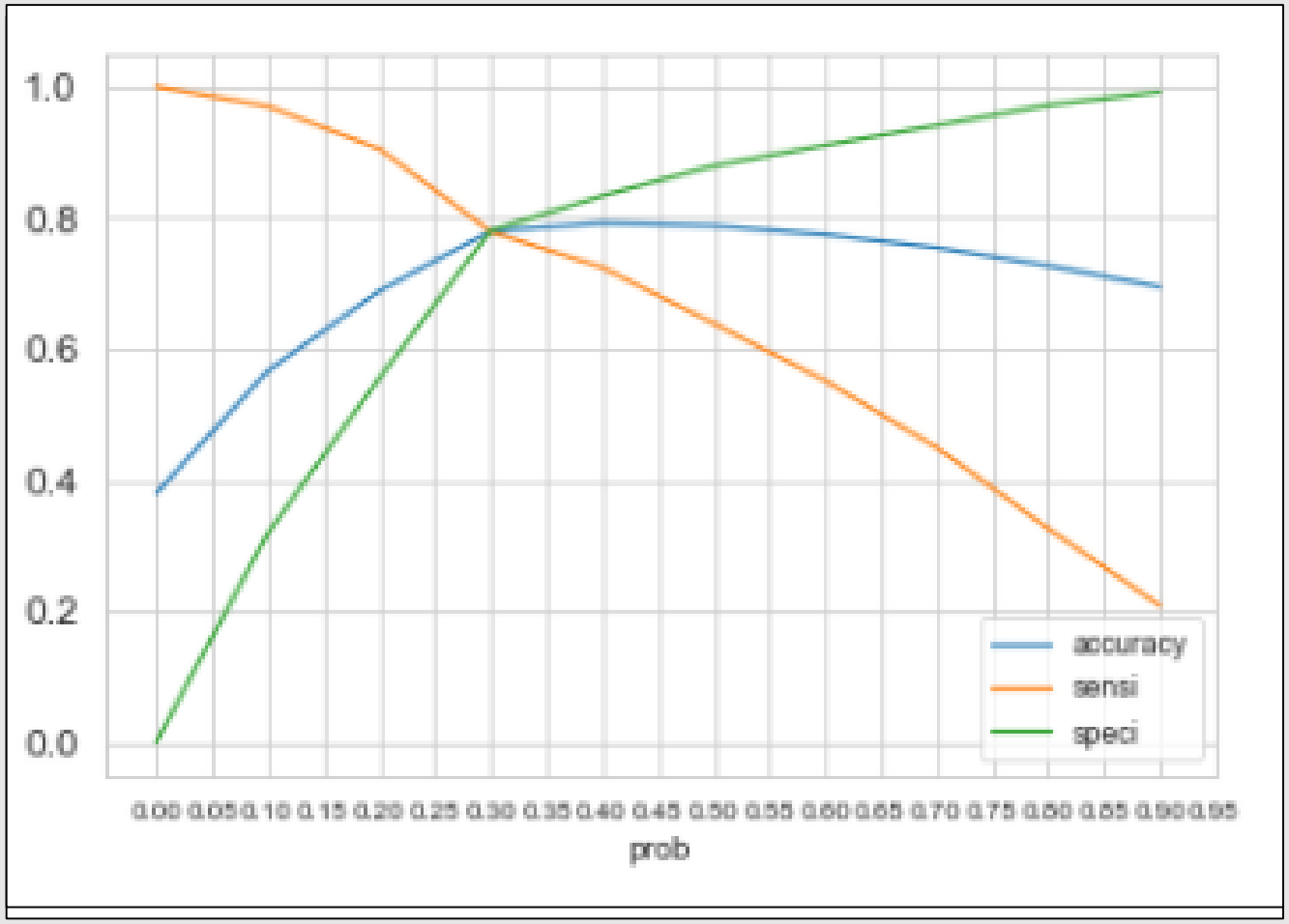
ROC Curve

- On taking threshold of positive rates by operating values, we can see that the model stand by with 85% cut-off value



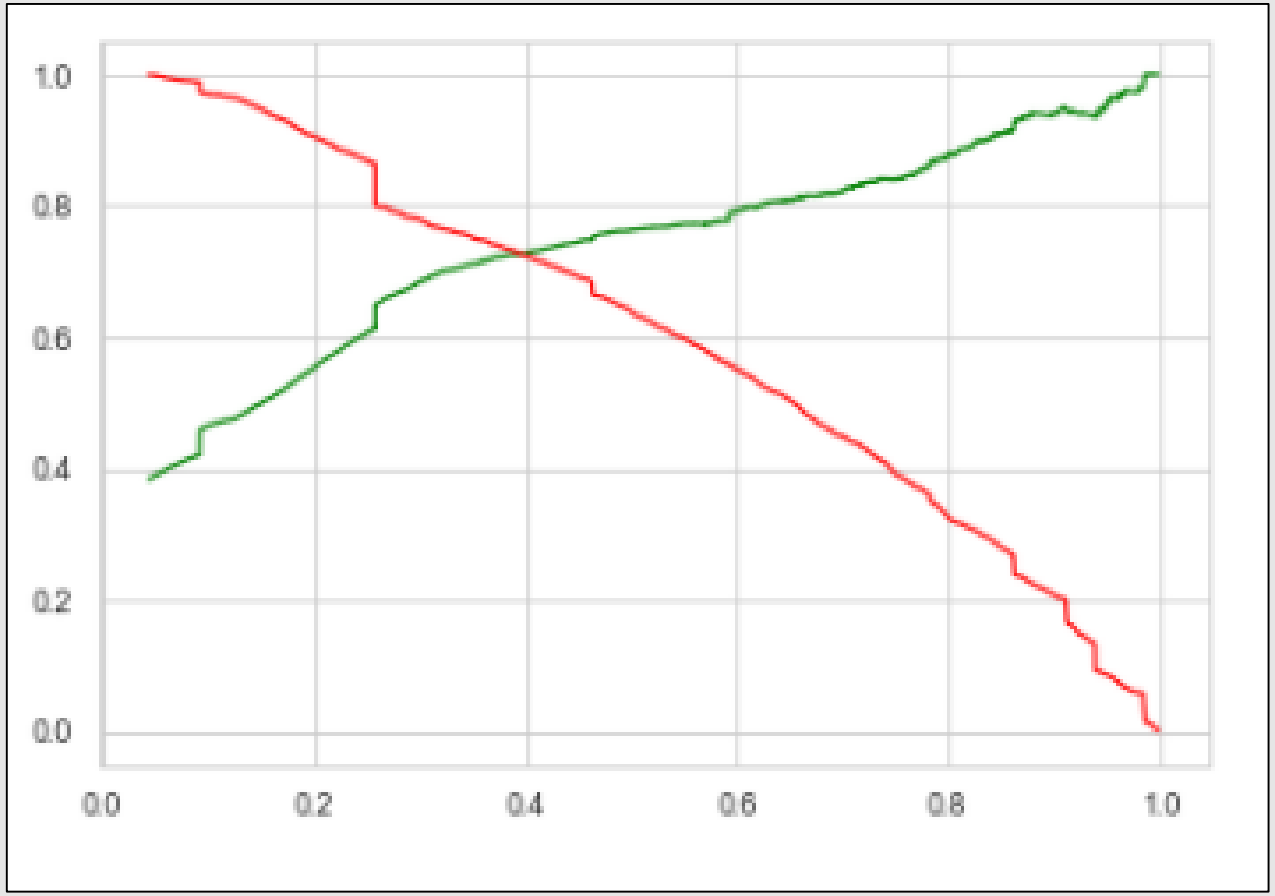
ROC Curve

- Based on the ROC curve, we came to a conclusion of 30% as cut-off value



Precision and Recall

- Based on Precision and Recall, we can take around 38% as cut-off



Conclusion

- On doing analysis , we have found that the sensitivity percentage on train data set is 77.8% that means the potential leads which we have found after analysing the data is around 78% which is good percentage to go with as per the requirement
- Also the F1 score is of 73% which is also an acceptable value to go with
- Also we have found 77.5% sensitivity of Test data set