

LEAD SCORING CASE STUDY

Om Shirke
Ishita Kothari
Tarun Rishishwar

BACKGROUND

- X Education , an education company sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites 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

BACKGROUND

- When these people fill up a form providing their email address or phone number, they are classified to be a lead
- 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%

PROBLEM STATEMENT

- X Education gets a lot of leads but its lead conversion rate is very poor
- 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

PROBLEM SOLUTION

Leads Clustering

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

Focus Communication

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

Increase conversion

Since we focussed on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.



IMPLEMENTATION

Loading &
Observing the past
data provided by
the Company

Univariate, Bivariate,
and Heatmap for
numerical and
categorical columns

Performing pre-
requisites for RFE
and Logistic
Regression



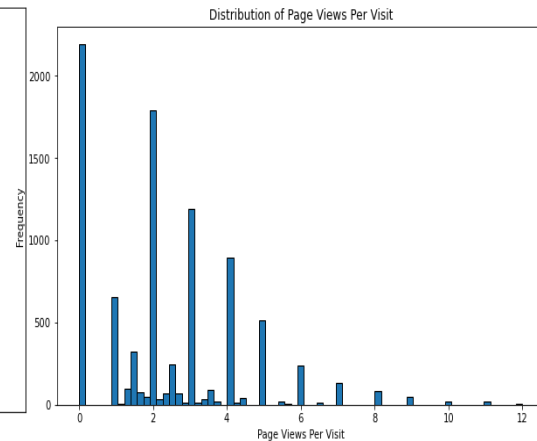
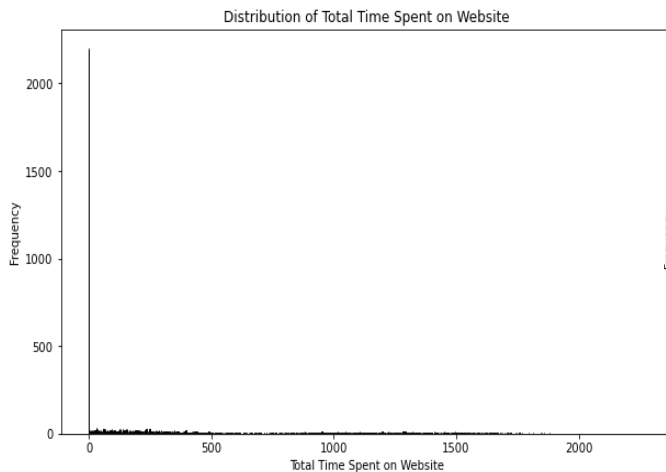
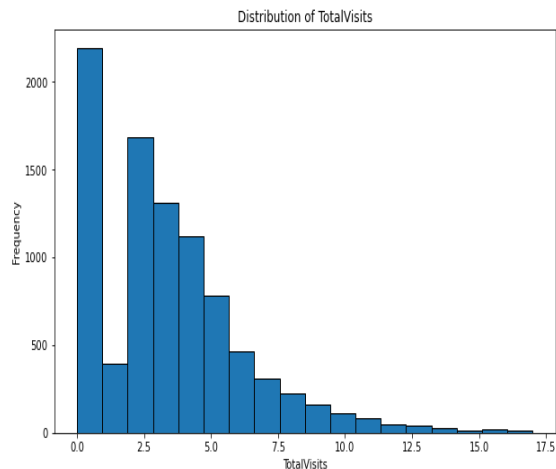
Duplicate removal, null
value treatment,
unnecessary column
elimination, etc.

Outlier Treatment,
Feature-
Standardization



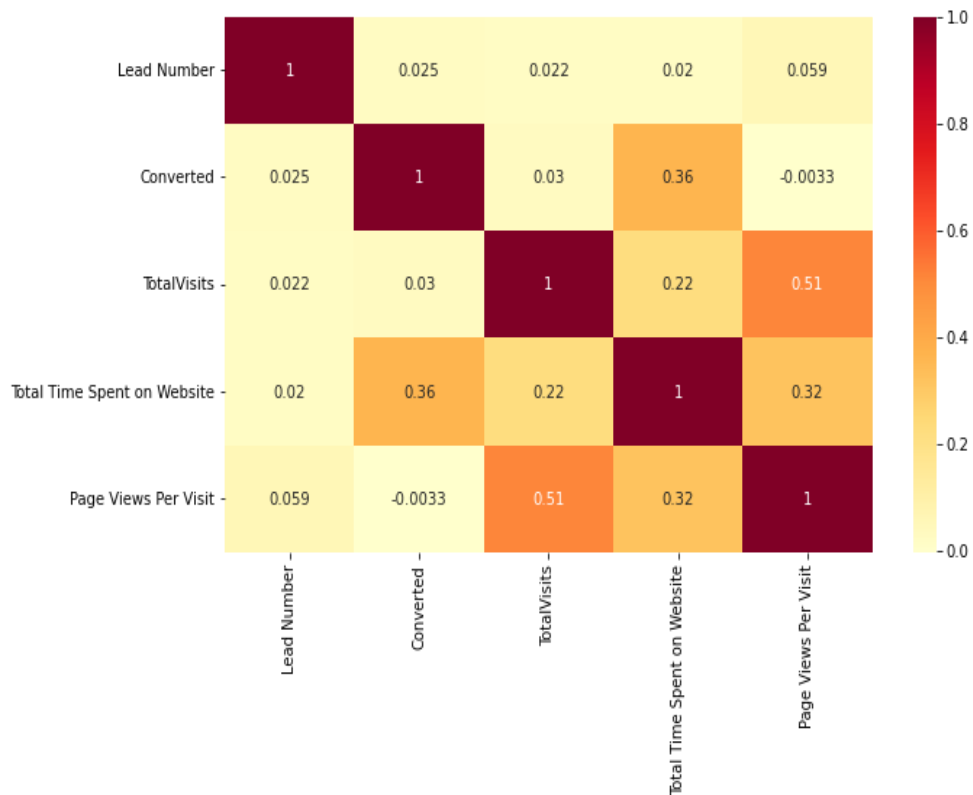
PLOTS (VISUALIZATION)

EDA : NUMERICAL COLUMNS



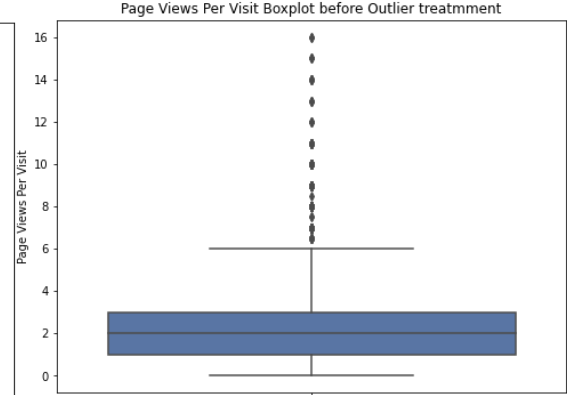
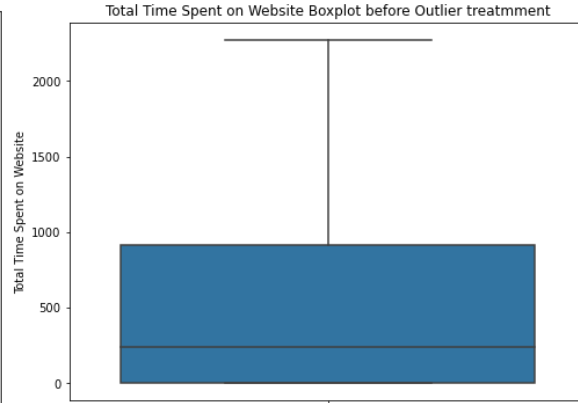
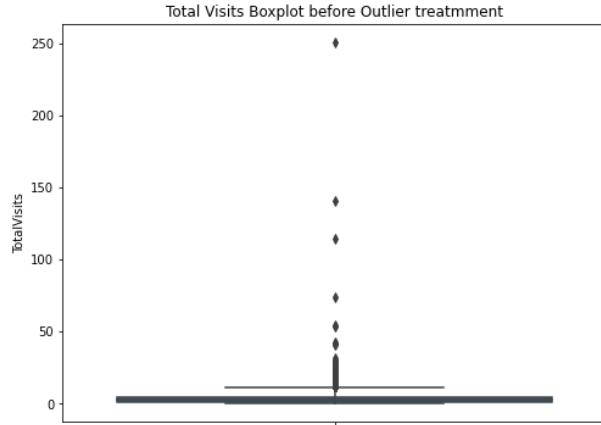
We can observe that the data has High peaks and is skewed. There might be a possibility of outliers.

EDA : HEATMAPS



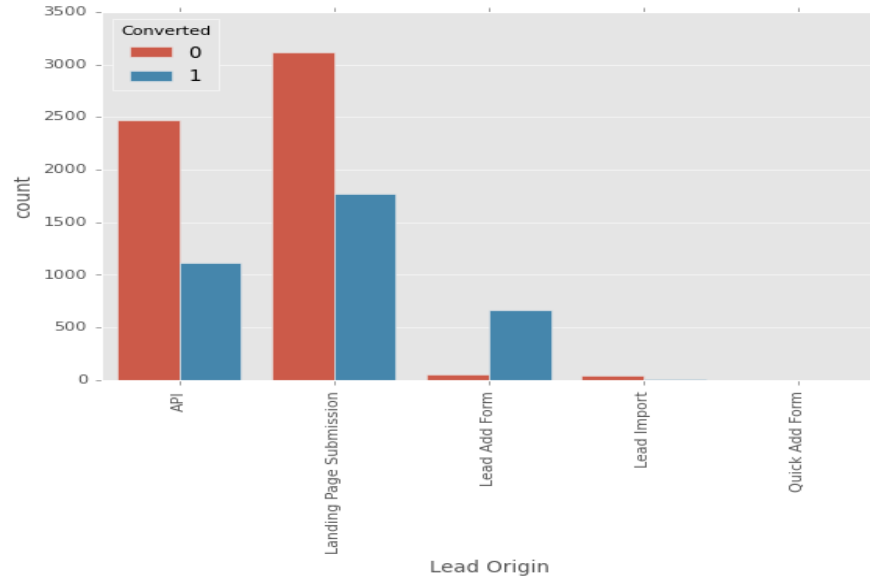
There is no No significant correlation to drop the columns

EDA : OUTLIERS



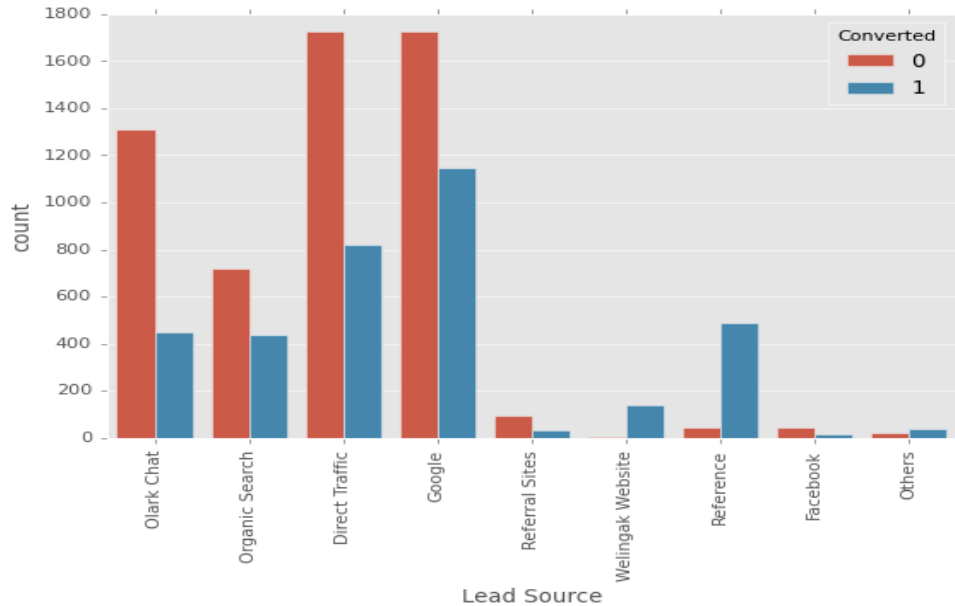
There are upper limit outliers in the total visits and page views per visit columns when analysing the box plots and the statistics.

EDA : CATEGORICAL COLUMNS



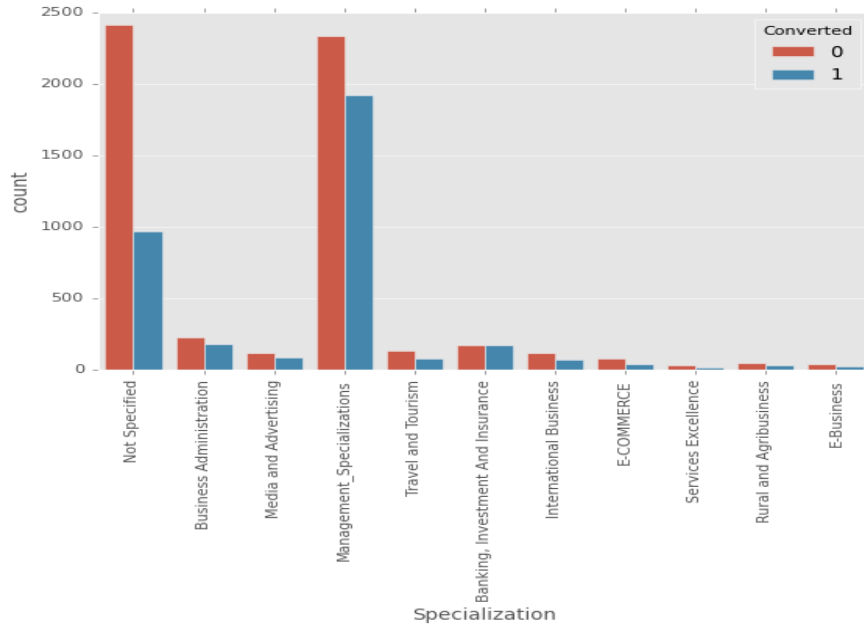
EDA plots in categorical column showing Lead origins with highest conversion through landing page submission

EDA : CATEGORICAL COLUMNS



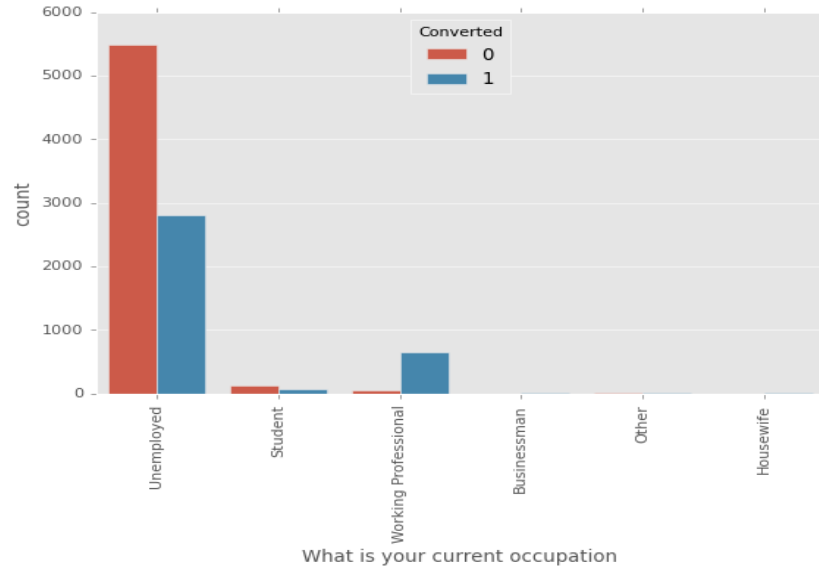
EDA plots in categorical column showing Lead source with highest conversion through Google

EDA : CATEGORICAL COLUMNS



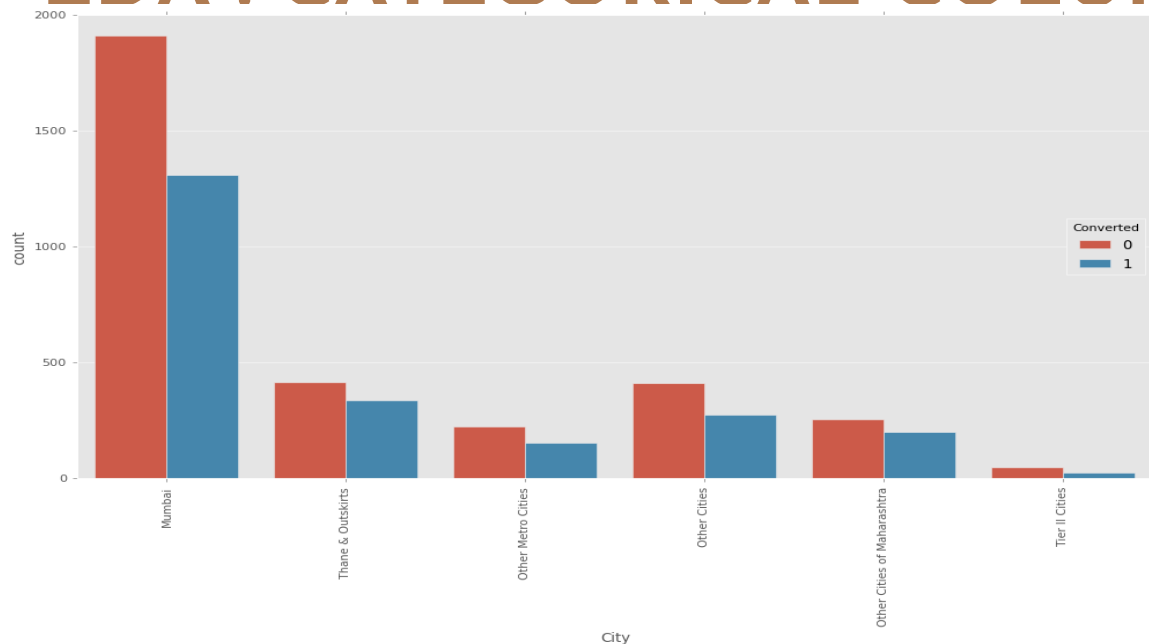
EDA plots in categorical column showing Lead specialisation with highest conversion through management specialisation

EDA : CATEGORICAL COLUMNS



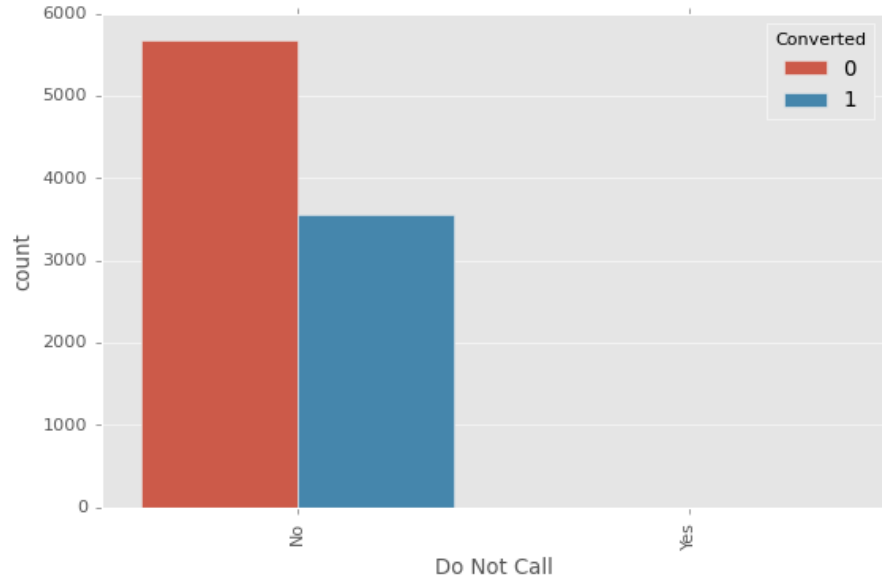
EDA plots in categorical column showing occupation with highest conversion through unemployment

EDA : CATEGORICAL COLUMNS



EDA plots in categorical column showing city with highest conversion from Mumbai

EDA : CATEGORICAL COLUMNS



EDA plots in categorical column showing do not email preference with highest opting for no.



MODEL BUILDING



MODEL BUILDING

- After EDA, Logistic Regression Model is built in python using GLM() function, under statsmodel library.
- The model contained all the variables, some of which had insignificant coefficients.
- Such variables are removed using Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
- Manual approach based on VIFs and p values.
- The final tally of variables with their respective values Significant p-values near to 0.00 and VIFs < 5.

MODEL BUILDING

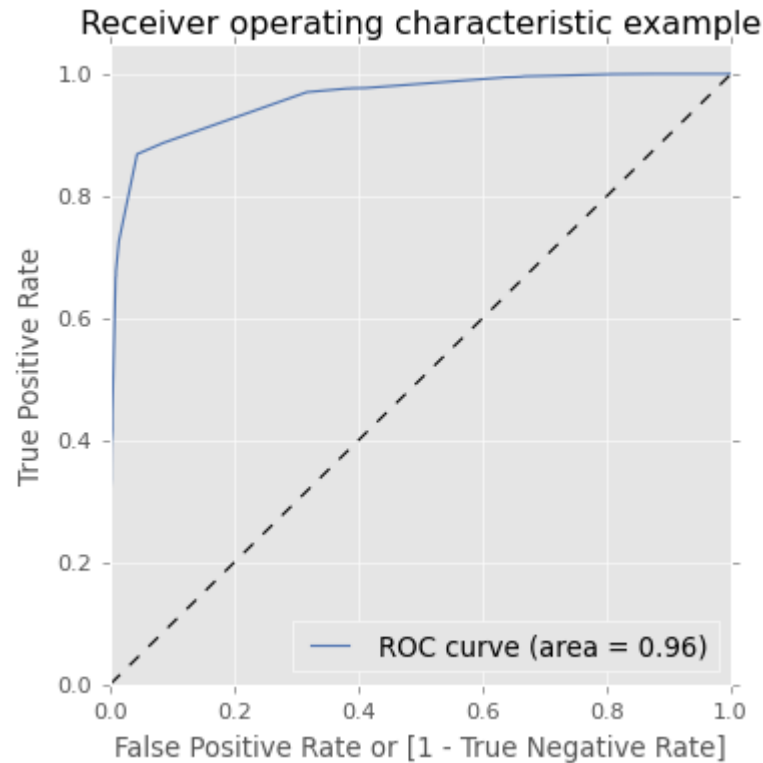
	coef	std err	z	P> z	[0.025	0.975]
const	-1.3877	0.072	-19.371	0.000	-1.528	-1.247
Lead_Source_Welingak Website	4.2366	0.753	5.628	0.000	2.761	5.712
Last_Activity_SMS Sent	2.3473	0.109	21.606	0.000	2.134	2.560
Tags_Already a student	-3.1927	0.713	-4.475	0.000	-4.591	-1.794
Tags_Closed by Horizon	7.2392	0.722	10.027	0.000	5.824	8.654
Tags_Interested in other courses	-1.7795	0.378	-4.702	0.000	-2.521	-1.038
Tags_Lost to EINS	6.8087	0.726	9.382	0.000	5.386	8.231
Tags_Not doing further education	-2.5128	1.025	-2.451	0.014	-4.522	-0.504
Tags_Ringing	-3.5446	0.234	-15.117	0.000	-4.004	-3.085
Tags_Will revert after reading the email	4.7863	0.174	27.444	0.000	4.445	5.128
Tags_invalid number	-3.9394	1.025	-3.842	0.000	-5.949	-1.930
Tags_number not provided	-24.2047	2.85e+04	-0.001	0.999	-5.59e+04	5.58e+04
Tags_opp hangup	-2.0847	0.802	-2.600	0.009	-3.656	-0.513
Tags_switched off	-4.5507	0.721	-6.312	0.000	-5.964	-3.138
Tags_wrong number given	-24.2599	2.04e+04	-0.001	0.999	-4.01e+04	4e+04
Last_Notable_Activity_Modified	-1.9238	0.121	-15.909	0.000	-2.161	-1.687

The VIF values seem fine but some p-values are 99 %. So removing 'Tags_number not provided' and tags_wrong number given' one by one.



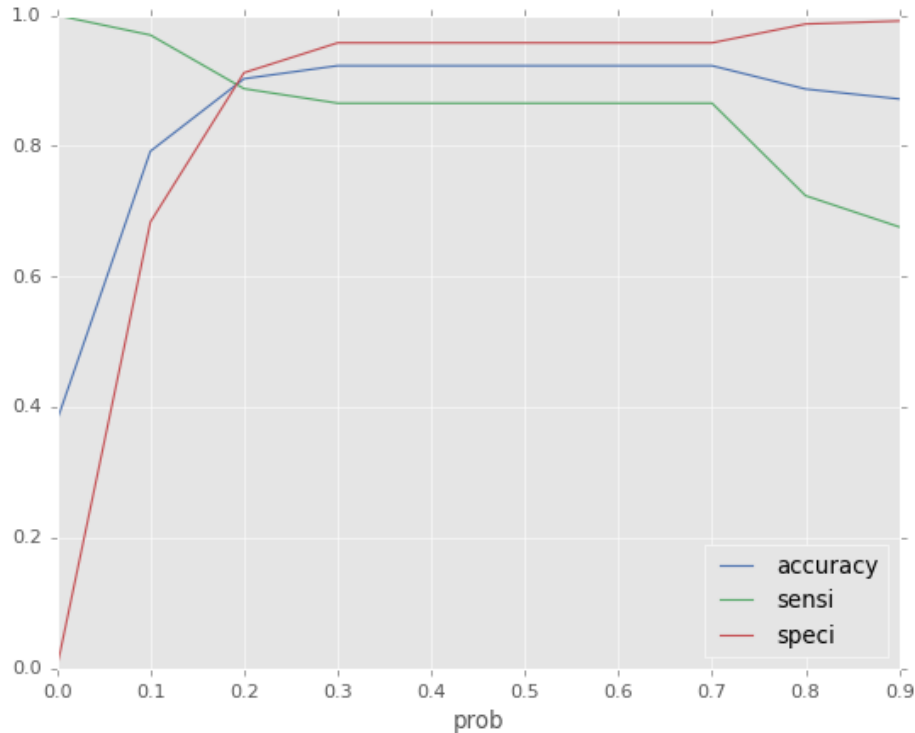
MODEL EVALUATION

ROC CURVE



ROC Curve demonstrates tradeoff between sensitivity and specificity. Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test. Closer the curve comes to 45° diagonal of the ROC space, the less accurate the test. For our model, ROC curve is towards the upper left corner, and area under the curve is more as displayed in figure. Thus, our model is an optimal choice to move forward with the analysis.

OPTIMAL CUT - OFF



Plotting the accuracy, sensitivity and specificity for various probabilities

From the graph it is visible that the optimal cut off is at 0.2.



CONCLUSION

MODEL ANALYSIS

Overall accuracy on Test set: 90%(approx)

Sensitivity of our logistic regression model: 89%(approx)

Specificity of our logistic regression model: 91%(approx)

LOGISTIC REGRESSION MODEL CONCLUSION

Our Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA, with 90 % Accuracy on Test Set, 89 % Sensitivity and 91% Specificity.

We can vary these parameters by varying the cut-off value and thus predict Hot leads based on scenarios like availability of extra resources and vice-versa.

RECOMMENDATION

The factors that the company must watch out for since they contribute the most to the probability that a lead will convert:

1)**Tags_Closed by Horizzon**: When the current status of the lead is that he/she is closed by Horizzon, that's when there is the most chance that a lead will convert.

2)**Tags_Lost to EINS** : When the current status of the lead is that he/she has lost to ENIS, there is a very high chance that the lead will convert.

RECOMMENDATION

- Now, since we selected a optimal cut off point as 0.2. When we run the model on new data, X Education Company can check the lead score and only pursue the leads that have a lead score above 20.
- These are the required HOT leads.
- Since, We got a Sensitivity of over 88%, our conversion rate for HOT leads will also be around 88% based on past data.
- Additionally, We got our leads conversion rate well over the required target of 80



THANK YOU