Lead Scoring Case Study

Group Members -
1. **Viswanath Venkatraj**
2. **Vaishnavi Sampath**
3. **Vivek Patil**

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, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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 that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use**.**

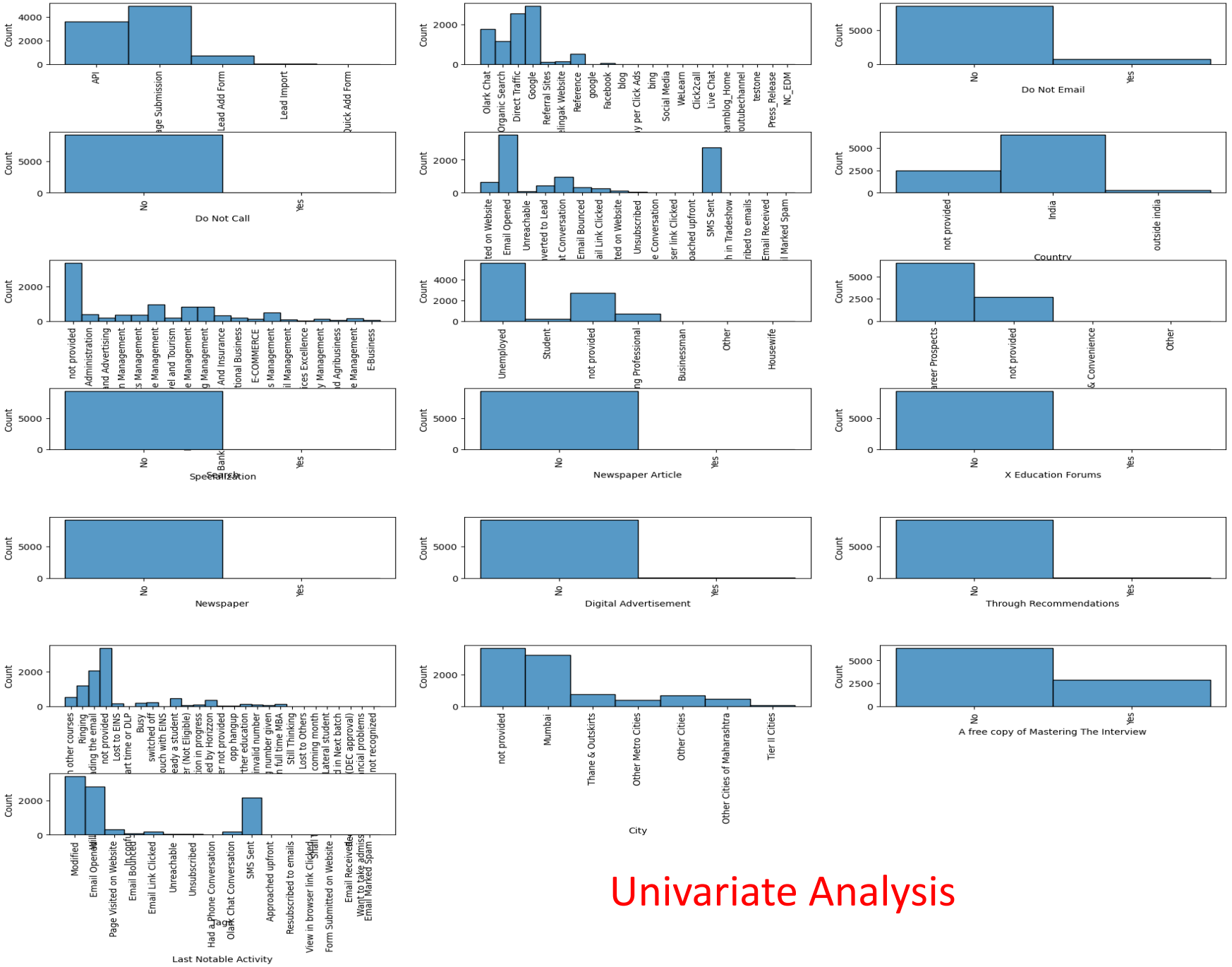
Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large number of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. 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.
- Conclusions and recommendations.

Data Manipulation

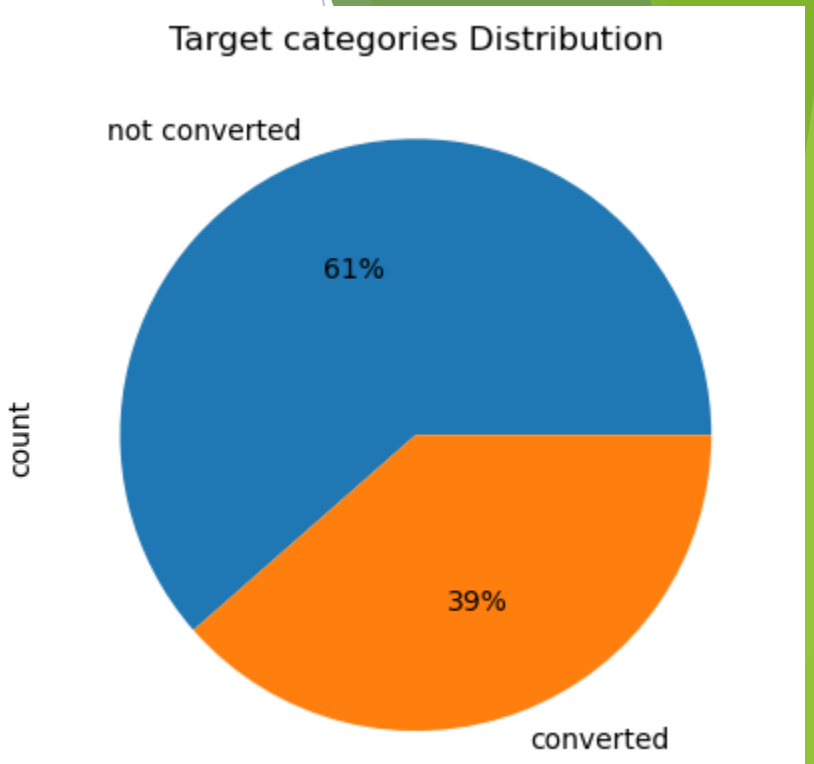
- Total Number of Rows =9240, Total Number of Columns =37.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, ”Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the categorical variables, we find some of the features which has not 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” etc.
- Dropping the columns having more than 40% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’, ‘Lead Quality’.

EDA



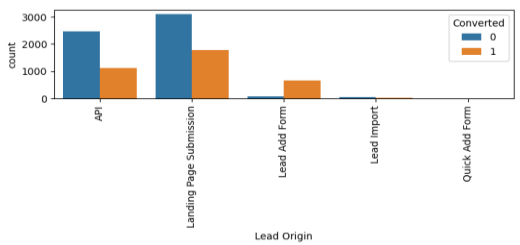
Univariate Analysis

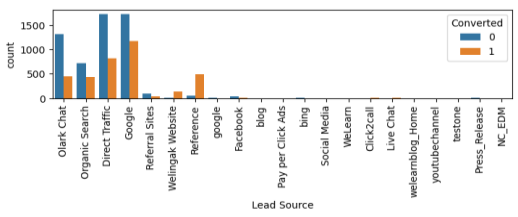
Correlations between numerical columns

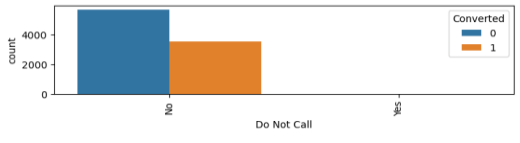


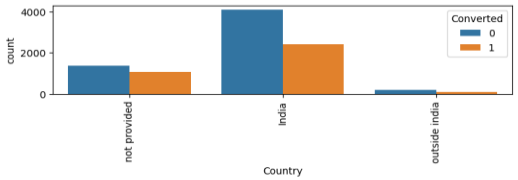
Ratio of data imbalance

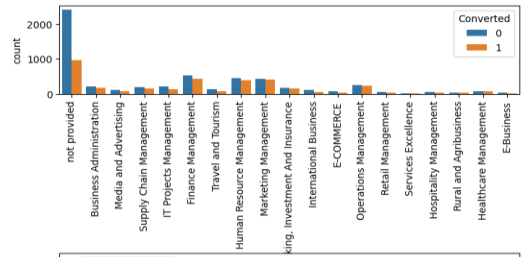
Bivariate Analysis

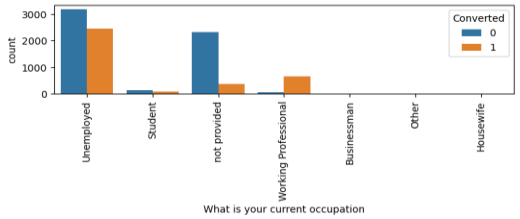


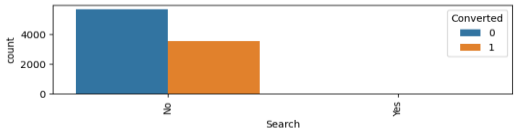


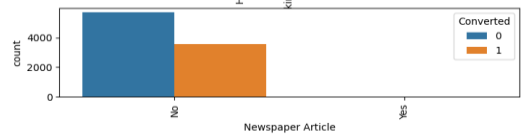


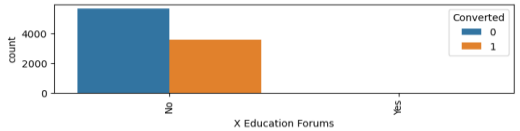


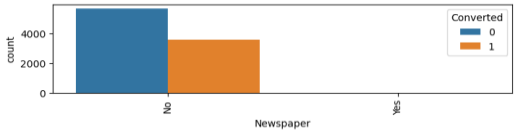


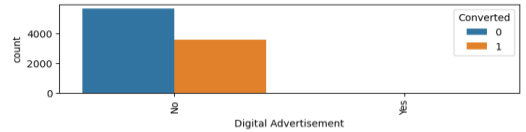


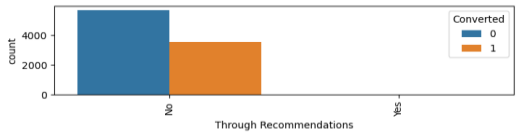


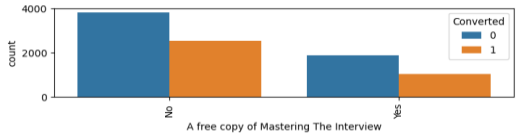


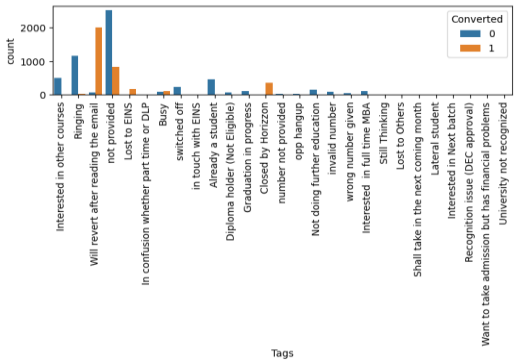


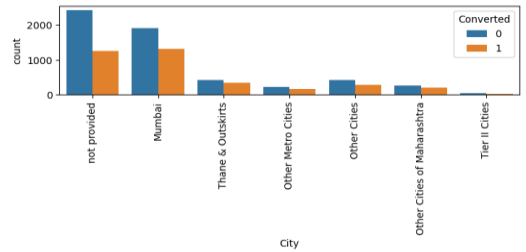












Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for categorical variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 88

Model Building

- The first basic step for regression is to perform a train-test split, we have chosen 70:30
- Running RFE with 15 variables as output
- Building models by removing the variables whose p-value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data
- Overall accuracy of 91%

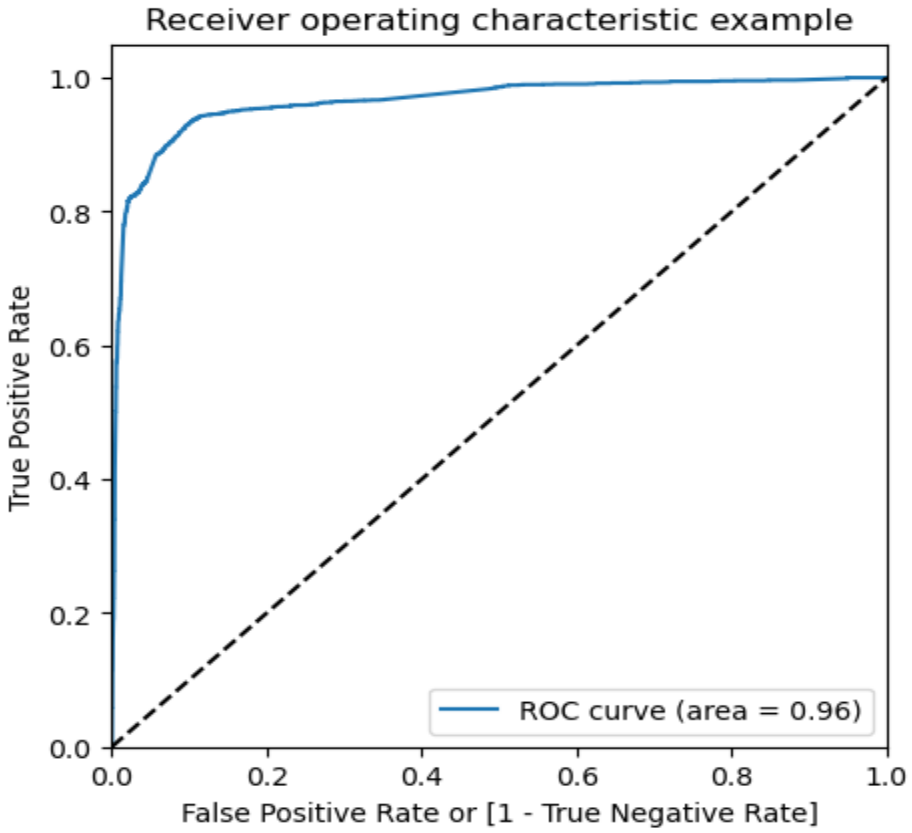
Generalized Linear Model Regression Results

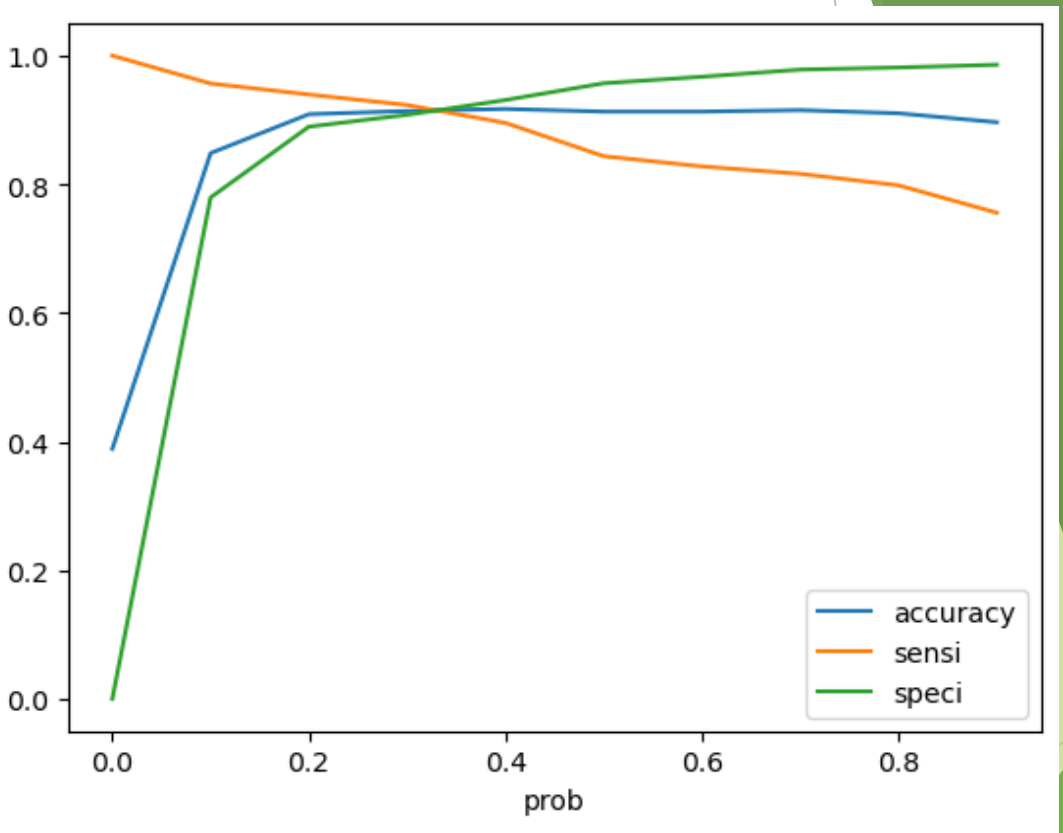
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6457
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1387.2
Date:	Sun, 14 Jan 2024	Deviance:	2774.4
Time:	20:08:53	Pearson chi2:	7.57e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.5964
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.5198	0.157	-28.750	0.000	-4.828	-4.212
Total Time Spent on Website	3.8438	0.209	18.385	0.000	3.434	4.254
Lead Origin_Lead Add Form	1.5445	0.342	4.513	0.000	0.874	2.215
Lead Source_Welingak Website	2.9097	1.069	2.721	0.007	0.814	5.006
Do Not Email_Yes	-1.5074	0.224	-6.737	0.000	-1.946	-1.069
What is your current occupation_not provided	-2.5641	0.135	-19.042	0.000	-2.828	-2.300
Tags_Busy	3.6945	0.239	15.480	0.000	3.227	4.162
Tags_Closed by Horizzon	7.8659	0.735	10.701	0.000	6.425	9.307
Tags_Lost to EINS	8.1630	0.589	13.850	0.000	7.008	9.318
Tags_Will revert after reading the email	6.8715	0.209	32.826	0.000	6.461	7.282
Tags_not provided	4.4336	0.180	24.607	0.000	4.080	4.787

Final model summary

ROC Curve





- **Finding Optimal Cut off Point →**
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics(87% and 91% respectively), we have considered the optimal cut off(0.35) based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 91%, 91% and 91% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 91% Hence overall this model seems to be good.