Lead Score Case Study

GROUP MEMBERS:

1. ABHILASH ARYA
2. BHARATH N
3. CHANDAN PAL

Problem Statement

- X Education Institute which sells online courses to individual.
- They get good number of leads of potential candidates who might be interested in taking the course
- Currently of all leads acquired through all means they only have 30% of conversion rate
- To get more conversion leads company wants to create a method to identify Hot Leads or candidate who has higher chances of taking the course

Case Study Objective:

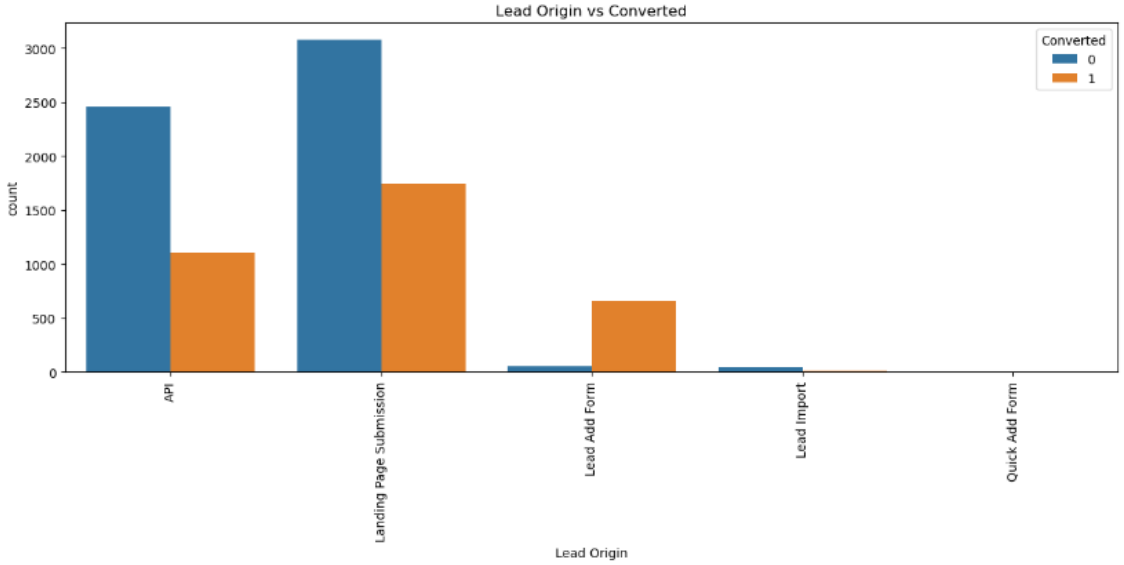
- The objective of this case study is to build a logistic regression model for the company to identify hot leads
- Scalable model to adjust the possible new feature in the model to identify the hot leads

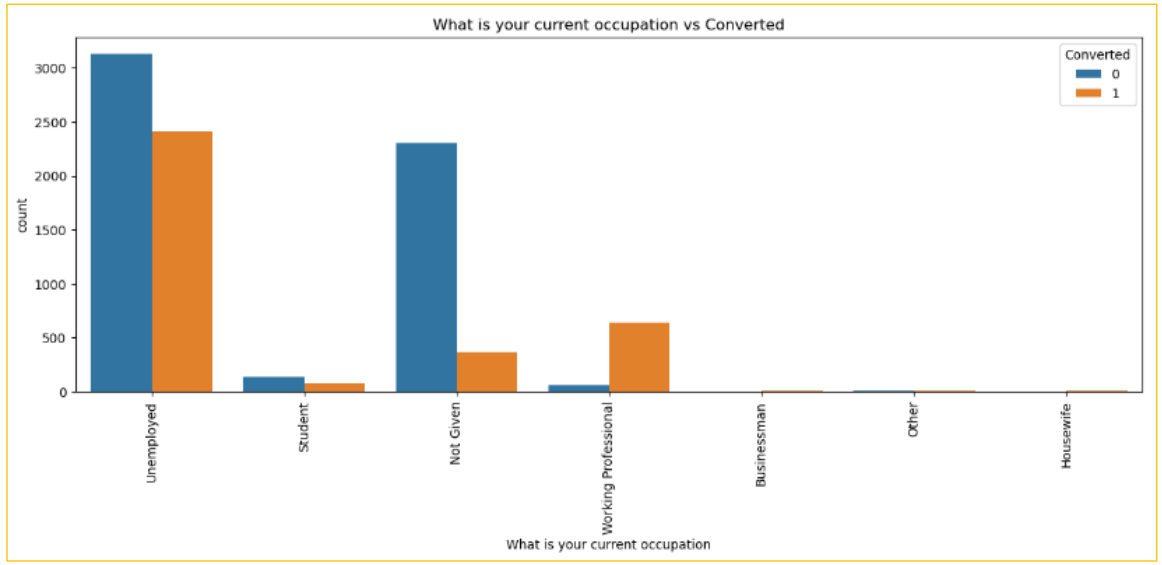
Analysis Approach

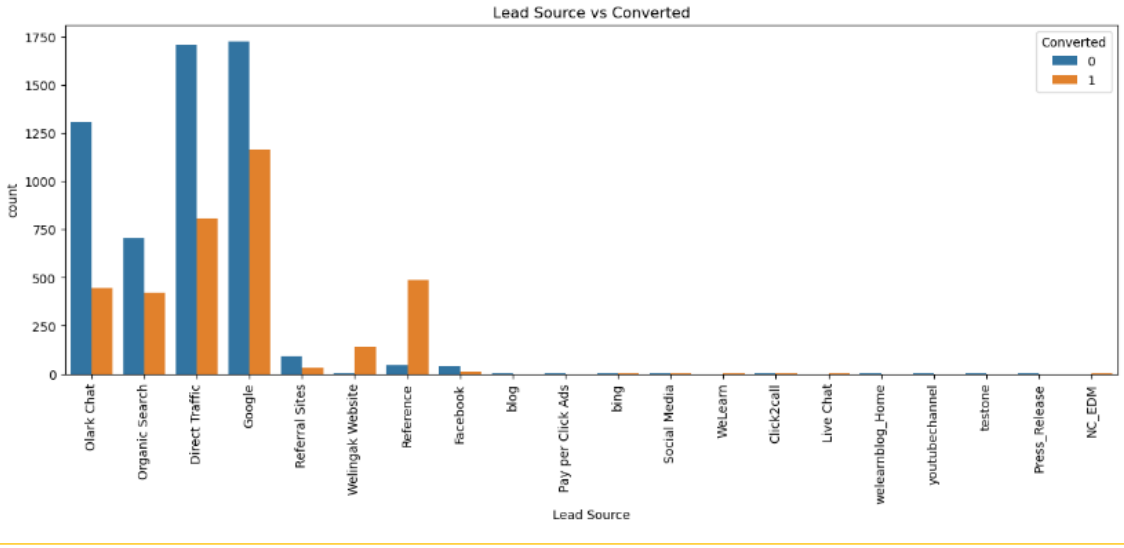
Below are the step we took to analyze the date and to build and test the model

1. We started with Data cleaning and preparation step
    1. Check and handled the missing/Null values by either removing or imputing values in the blanks
    2. Handled the outliers by removing them
    3. Removing the imbalanced columns
2. We performed the EDA analysis to understand the data better by performing Univariate and multivariate analysis
3. We then moved on to dummy variable creation for categorical features and feature scaling of numerical variable
4. We then split the data in test and train
5. We then moved on to creation of logistic regression model. We used RFE and manual feature selection method to come up to final model which can be used for prediction.
6. We then created prediction using the model.
7. We analyzed the model prediction using various metrics like accuracy, sensitivity, specificity and ROC curves.
8. Once's we had satisfaction result of above metrics we then used the model for prediction on test data and analyzed the same metric for test data as well

EDA







Highest conversion rate is from:

- Lead from direct website submission and lead add from
- Lead source of google and Direct traffic
- Lead with Last activity as SMS and Email response
- Leads of people who are either unemployed or working professionals
- Leads of people who want better career prospects

Model Building

- After Data Preparation we started with splitting data in Test and Train dataset
- We first build model with all the features, since there were too many feature after dummy variables and most of them were insignificant to the model as per P-value analysis
- We Use REF approach to filter out the feature and only kept top 15 features
- Then using these features we build model and later removed some of the insignificant features manually one by one by analyzing P-value and VIF values.
- We remove any feature which is having >.05 of P-value and more than 5 VIF

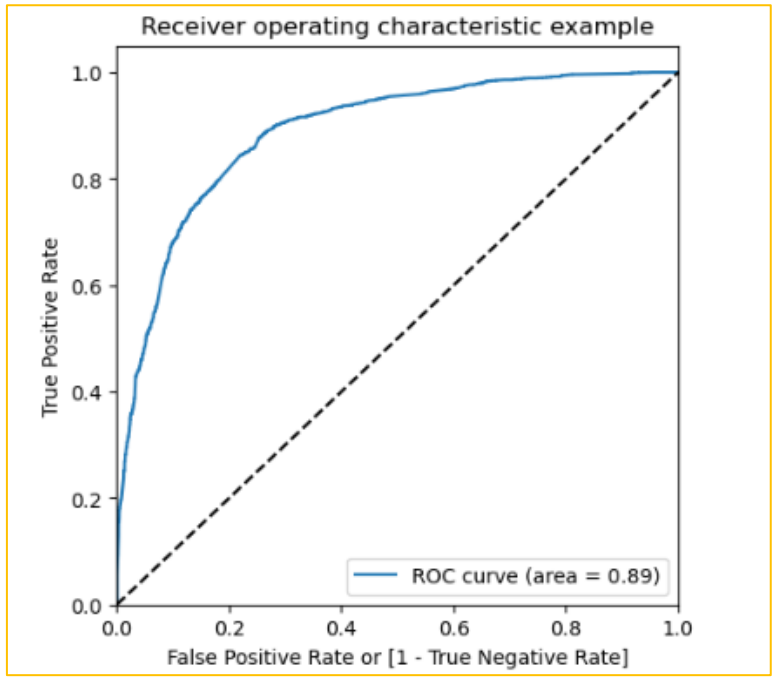
Generalized Linear Model Regression Results

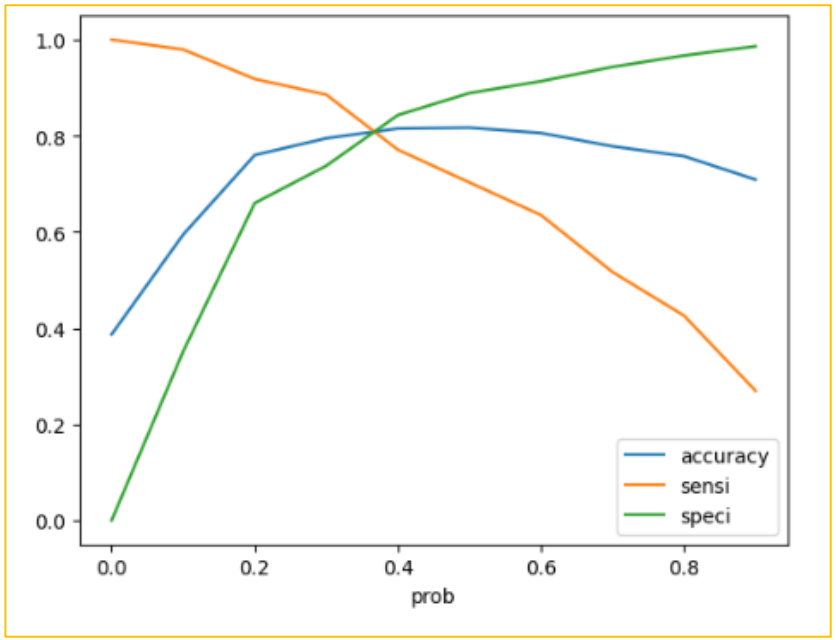
Dep. Variable:	Converted	**No. Observations:**	6409
Model:	GLM	**Df Residuals:**	6396
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2633.8
Date:	Sun, 15 Dec 2024	**Deviance:**	5267.7
Time:	13:14:33	**Pearson chi2:**	6.96e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4012
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0889	0.056	-19.334	0.000	-1.199	-0.979
Total Time Spent on Website	1.1403	0.041	27.910	0.000	1.060	1.220
Lead Origin_Lead Add Form	1.9855	0.206	9.649	0.000	1.582	2.389
Lead Source_Facebook	-1.1206	0.415	-2.702	0.007	-1.933	-0.308
Lead Source_Welingak Website	1.9558	0.750	2.607	0.009	0.486	3.426
Last Activity_Converted to Lead	-1.3767	0.227	-6.075	0.000	-1.821	-0.933
Last Activity_Email Bounced	-1.9858	0.312	-6.362	0.000	-2.598	-1.374
Last Activity_Olark Chat Conversation	-1.3852	0.166	-8.326	0.000	-1.711	-1.059
Last Activity_SMS Sent	1.1515	0.074	15.464	0.000	1.006	1.297
Country_Not Given	1.5105	0.106	14.307	0.000	1.304	1.717
What is your current occupation_Working Professional	2.3903	0.183	13.090	0.000	2.032	2.748
What matters most to you in choosing a course_Not Given	-1.3183	0.087	-15.107	0.000	-1.489	-1.147
Last Notable Activity_Had a Phone Conversation	3.6001	1.119	3.217	0.001	1.406	5.794

Model Evaluation

- After predicting the values using the final model, we then evaluated our model using various metrics.
- Accuracy: 82%
- Sensitivity: 70%
- Specificity: 89%
- Positive predictive value: 80%
- Negative predictive value: 83%
- ROC: .89
- We plotted accuracy, sensitivity and specificity on single curve and found Optimal cutoff point at 0.4





- After evaluating model on train data, we then move on to predict the values on the test data as well and found the below metric with cutoff value of 0.4
- Accuracy: 81%
- Sensitivity: 75%
- Specificity: 84%

Conclusion

By forming the logistic regression model and looking at the result of various metric on test and train data we can conclude that model is performing quite well. It is consistent on Test and Train data which shows that model is not overfitting or underfitting. Below are the main features that can help the institute to target the potential customer to increase conversion rate.

Institute can target user who

- Have spend more time on Website
- Lead origin from Lead add form
- lead source from welingak website
- last activity as SMS response
- User who are working profession
- User who responded to phone conversations