Lead Scoring Case Study

SUBMITTED BY

ANIL KUMAR M

Problem Statement:

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data Preparation

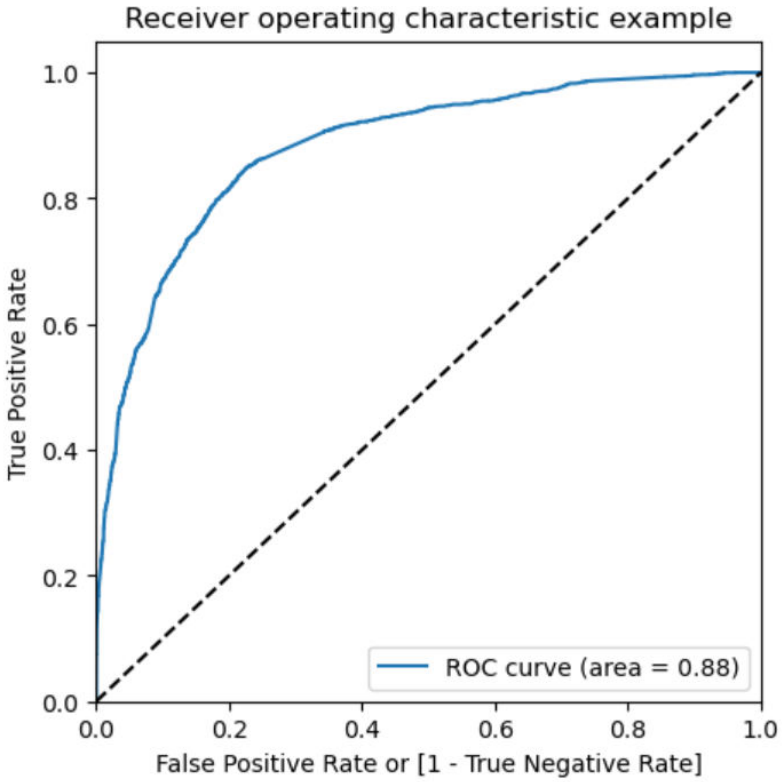
Following Step are performed to Prepare Data for Model Building

- Checked Statistical aspects of the dataframe
- Missed Values are checked and Appropriate columns are removed
- Data having “Select” values are handled
- Dropped Columns that are not necessary for analysis
- Continuous and Categorical Columns are detected
- Dummy Variables are created for categories in categorical columns
- Outlier Analysis is performed and values above 0.99 percentile are removed

Model Building and Evaluation

- Data is split into Train and Test Data
- Recursive Feature Elimination is done to select 15 relevant features
- Logistic Regression Models are build with the relevant feature variables
- Models are assessed and 6^{th} Model is selected as best fit Model.
- Optimum Probability is found to be 0.35
- Predicted converted values are found out with the Optimum Probability value
- ROC is applied and found the model to be 88% accurate
- Confusion Matrix and other Model features like accuracy,specificity,sensitivity,false positive rate And negative predictive values are calculated.
- Model is applied and Test data set and found to be matching with train Dataset.

Accuracy and Model Regression Results



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6351
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2649.5
Date:	Sat, 17 Feb 2024	**Deviance:**	5299.0
Time:	17:22:23	**Pearson chi2:**	6.87e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.3916
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3795	0.052	-26.668	0.000	-1.481	-1.278
Do Not Email	-1.7027	0.176	-9.670	0.000	-2.048	-1.358
Total Time Spent on Website	1.0904	0.040	27.430	0.000	1.013	1.168
Lead Origin_Lead Add Form	3.6609	0.197	18.627	0.000	3.276	4.046
Lead Source_Olark Chat	1.0972	0.103	10.635	0.000	0.895	1.299
Lead Source_Welingak Website	2.9627	1.030	2.877	0.004	0.944	4.981
Last Activity_Converted to Lead	-1.1421	0.208	-5.492	0.000	-1.550	-0.734
Last Activity_Olark Chat Conversation	-1.4028	0.159	-8.812	0.000	-1.715	-1.091
What is your current occupation_Working Professional	2.7690	0.187	14.810	0.000	2.403	3.135
Last Notable Activity_SMS Sent	1.5950	0.080	19.916	0.000	1.438	1.752
Last Notable Activity_Unreachable	2.0420	0.605	3.377	0.001	0.857	3.227
Last Notable Activity_Unsubscribed	1.7138	0.489	3.507	0.000	0.756	2.671

Conclusion:

- Model is built with 88% accuracy
- Hot Leads dataset is created with conversion rates greater than 80%
- Below Features are found to be most important

    I. “Lead Origin_Lead Add Form”
    II. “Lead Source_Welingak Website”
    III. “What is your current occupation_Working Professional”