Lead Scoring Case Study

Sumneet Khanna | Anoop
DS C59

AGENDA

01 PROBLEM STATEMENT

02 GOALS OF THE CASE STUDY

03 ANALYSIS APPROACH

04 MODEL SELECTION

05 MODEL PERFORMANCE

06 BUSINESS RECOMMENDATIONS

INTRODUCTION

PROBLEM STATEMENT

X Education aims to improve lead conversion rates using lead scoring. The company seeks to identify hot leads and optimize sales efforts

GOALS OF THE CASE STUDY

- **Build a logistic regression model**

  The model can assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- **Handle other related problems**

  There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future

ANALYSIS APPROACH

DATA PREPROCESSING:

- Libraries Used: NumPy, Pandas, Sklearn, statsmodel, matplotlib,seaborn
- Columns with only single values such as all ‘NO’ , with data imbalance removed.
- NULL VALUES removed and imputed accordingly.
- Label Encoding performed to convert Yes/No columns to 1/0

EDA:

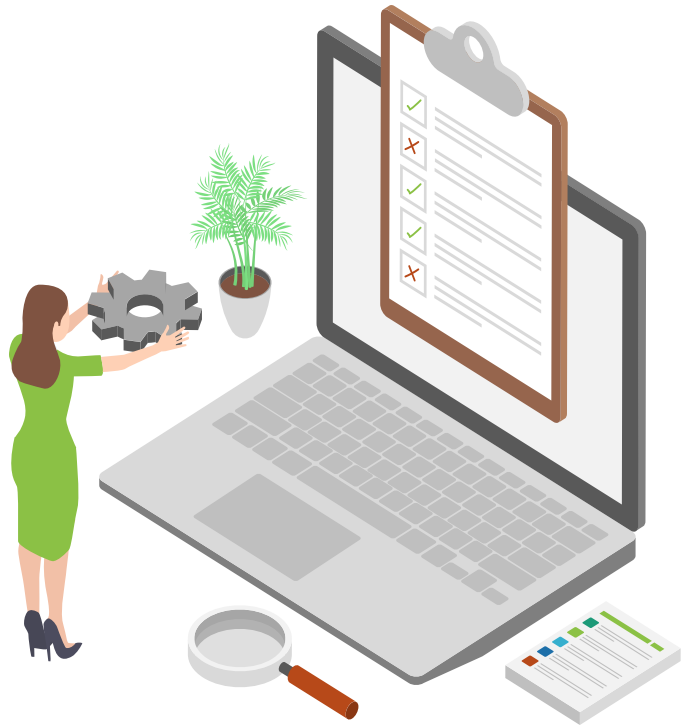
- Univariate Analysis conducted, Correlation matrix and heatmap drawn
- Outliers detected and dealt with accordingly

Model Building and Selection(Logistic Regression)

- Used pd.dummies to convert categorical variables to numerical form
- Model building started with train and test split. RFE helped select features
- ROC curve made and optimal cut off point searched

Model Evaluation

- Confusion matrix created
- Predictions made on the test set and model evaluation done via Sensitivity, Specificity, Precision, recall and F1 score
- Finally, a Lead score assigned



MODEL SELECTION

MODEL SUMMARY WITH COEFFICIENTS AND P-VALUES

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7029	0.116	6.058	0.000	0.476	0.930
Do Not Email	-1.2852	0.152	-8.442	0.000	-1.584	-0.987
Total Time Spent on Website	1.1419	0.039	28.967	0.000	1.065	1.219
Lead Origin_Lead Add Form	2.3360	0.200	11.702	0.000	1.945	2.727
Lead Origin_Lead Import	-1.2804	0.495	-2.587	0.010	-2.250	-0.311
Lead Source_Welingak Website	2.2212	0.742	2.992	0.003	0.766	3.676
Specialization_Not Specified	-0.3836	0.086	-4.460	0.000	-0.552	-0.215
What is your current occupation_Not Specified	-1.3720	0.085	-16.235	0.000	-1.538	-1.206
What is your current occupation_Working Professional	2.2455	0.171	13.143	0.000	1.911	2.580
Page Views Range Per Visit_1-2	-1.2610	0.136	-9.284	0.000	-1.527	-0.995
Page Views Range Per Visit_2-3	-1.4349	0.127	-11.294	0.000	-1.684	-1.186
Page Views Range Per Visit_3-4	-1.2088	0.137	-8.794	0.000	-1.478	-0.939
Page Views Range Per Visit_4-5	-1.2802	0.151	-8.477	0.000	-1.576	-0.984
Page Views Range Per Visit_5-10	-1.3294	0.146	-9.095	0.000	-1.616	-1.043
Page Views Range Per Visit_10-20	-1.2866	0.369	-3.483	0.000	-2.010	-0.563

VIF

	Features	VIF
5	Specialization_Not Specified	1.53
6	What is your current occupation_Not Specified	1.52
2	Lead Origin_Lead Add Form	1.40
4	Lead Source_Welingak Website	1.27
1	Total Time Spent on Website	1.23
9	Page Views Range Per Visit_2-3	1.21
7	What is your current occupation_Working Profes...	1.17
8	Page Views Range Per Visit_1-2	1.16
10	Page Views Range Per Visit_3-4	1.11
0	Do Not Email	1.10
12	Page Views Range Per Visit_5-10	1.08
11	Page Views Range Per Visit_4-5	1.07
3	Lead Origin_Lead Import	1.01
13	Page Views Range Per Visit_10-20	1.01

- Model 2, satisfied the basic checks with all p-values below 0.05 and VIF of all features below 2.

MODEL PERFORMANCE

```
Sensitivity_Train: 83.38670936749399
Specificity_Train: 73.2205956587582
Precision_Train: 66.2531806615776
Recall_Train: 83.38670936749399
Accuracy_Train: 0.7715170278637771
```

Sensitivity_Train (Recall_Train): Approximately 83.39%, this indicates that the model correctly identifies around 83.39% of the actual positive cases in the training data.

Specificity_Train: The model correctly identifies around 73.22% of the actual negative cases in the training data.

Precision_Train: This indicates that around 66.25% of the positive predictions made by the model are correct.

Accuracy_Train: Accuracy measures the overall correctness of the model's predictions. In the training set, the accuracy is approximately 77.15%.

```
Sensitivity_Test: 83.89830508474576
Specificity_Test: 72.1734036321031
Precision_Test: 65.22693997071742
Recall_Test: 83.89830508474576
Accuracy_Test: 76.67027807872878
```

Sensitivity_Test (Recall_Test): Approximately 83.90%, this indicates that the model correctly identifies around 83.90% of the actual positive cases in the testing data**.**

Specificity_Test: Approximately 72.17%., this indicates that the model correctly identifies around 72.17% of the actual negative cases in the testing data.

Precision_Test: Approximately 65.23%, this indicates that around 65.23% of the positive predictions made by the model are correct.

Accuracy_Test: In the testing set, accuracy is approximately 76.67%.

MODEL PERFORMANCE- CONTINUED

73.39%

F1-Score provides a balanced assessment of a classification model's performance. It considers both false positives and false negatives.
An F1-score of approximately 73.39% t suggests that the model achieves a good balance between precision and recall. A higher F1-score indicates better overall performance of the model in terms of both false positives and false negatives.

BUSINESS RECOMMENDATIONS

01

The top three variables contributing most to the probability of lead conversion are Total Time Spent on Website, Lead Origin (Lead Add Form), and Working Professional Occupation. These variables indicate strong engagement with the website, active participation through lead forms, and a target demographic with higher purchasing power and intent, respectively, all of which positively influence the likelihood of conversion.

02

The top three categorical/dummy variables that significantly influence lead conversion probability are Lead Origin (Lead Add Form), Lead Source (Welingak Website), and Specialization (Not Specified). Emphasizing lead generation through the lead add form, optimizing the website experience, and tailoring marketing strategies to address unspecified specialization can effectively boost conversion rates by capturing and engaging potential leads more effectively.

03

During the internship period, X Education should prioritize leads predicted as high conversion potential by the model. By focusing on personalized phone calls, automated email workflows, and tailored incentives, they can effectively engage these leads and maximize conversion rates. This proactive strategy ensures efficient utilization of resources and enhances the likelihood of converting potential leads into paying customers.

04

During periods of early target achievement, X Education should refine lead scoring, automate lead nurturing, and utilize strategic engagement channels to minimize unnecessary phone calls. By prioritizing high-value leads, optimizing resource allocation, and focusing on alternative communication methods, they can maintain efficiency and explore new growth opportunities effectively.

CONCLUSION

01 Focus on “HOT LEADS" - those with high conversion potential - identified through variables like total time spent on the website, lead origin, and occupation.

02 By strategically focusing on these leads and tailoring proactive engagement strategies, such as personalized phone calls and targeted communication, X Education can significantly improve its lead conversion rates.

03 This approach empowers X Education to capitalize on valuable opportunities, optimize resource allocation, and drive sustainable growth in its business endeavors.

THANK YOU