Lead Scoring Case Study

- Charu Rathi
- Davlesh Katta
- Krishnakumar Thankappan

Contents

- **Problem statement**
- **Problem approach**
- **EDA**
- **Correlations**
- **Model Evaluation**
- **Observations**
- **Conclusions**

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as 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%.** Now, this means 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

- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

Approach

- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE VIF and p- values)
- Model Evaluation
- Making predictions on test set

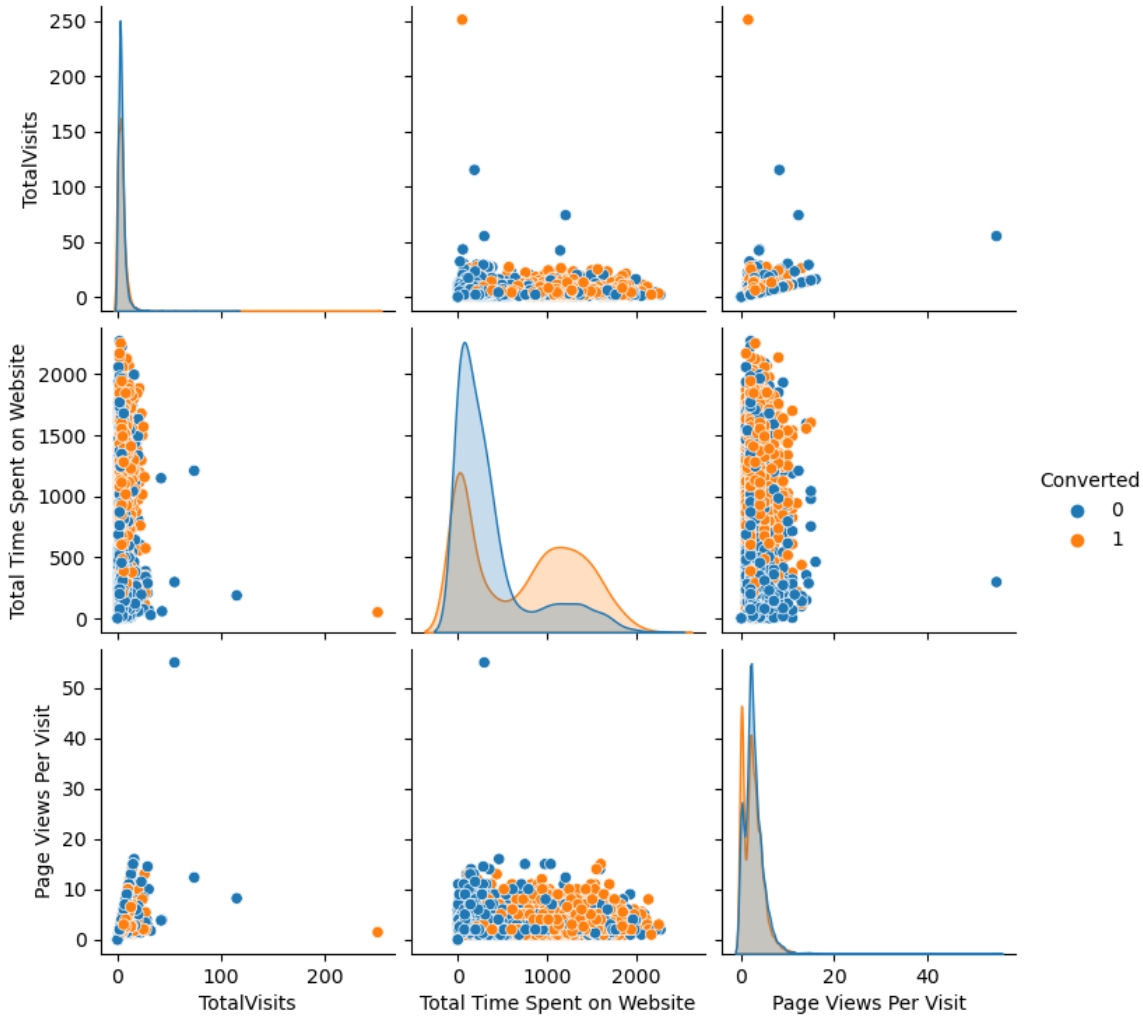
Reading and studying data

- Using
- .shape
- .columns
- .describe()
- .info()

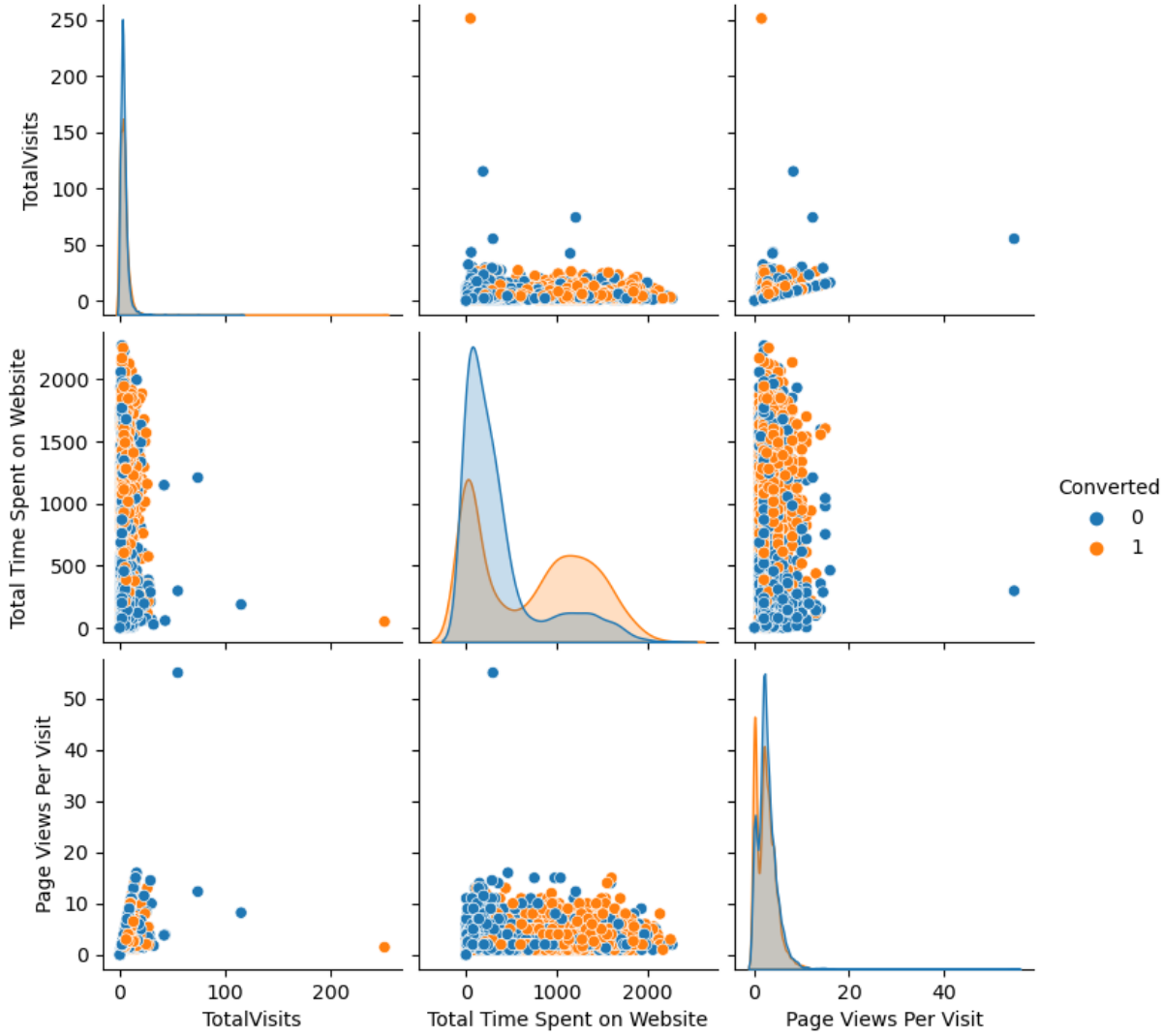
Data cleaning

- Dropping columns with more than 3000 missing values
- Dropping city and country (not relevant)
- Dropping columns with “Select”
  - ['Lead Profile’], ['How did you hear about X Education’]
- Dropping columns with only one value for all data points
- Dropping null rows for columns
  - ['What is your current occupation’], ['TotalVisits’], ['Lead Source’], ['Specialization']

Data preparation



Data preparation



Dummy variable creation

- To deal with categorical variables
- 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'What is your current occupation','A free copy of Mastering The Interview', 'Last Notable Activity’
- 'Specialization'

Train- Test Split and Scaling

- 70:30 ratio
- Using Min max scaler

Correlations

- Done but not much useful

Model building

- On train data
- Using RFE to select 15 variables

Model building

- First model and p values

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

Top VIF values

	Features	VIF
2	Lead Origin_Lead Add Form	84.19
4	Lead Source_Reference	65.18
5	Lead Source_Welingak Website	20.03
11	What is your current occupation_Unemployed	3.65
7	Last Activity_Had a Phone Conversation	2.44
13	Last Notable Activity_Had a Phone Conversation	2.43
1	Total Time Spent on Website	2.38
0	TotalVisits	1.62
8	Last Activity_SMS Sent	1.59
12	What is your current occupation_Working Profes...	1.56
3	Lead Source_Olark Chat	1.44

Further models dropping variables based on p value and VIF

- Model 2: dropped 'Lead Source_Reference’
- Model 3: dropped 'Last Notable Activity_Had a Phone Conversation’
- Model 4: dropped 'What is your current occupation_Housewife’
- Model 5: dropped 'What is your current occupation_Working Professional’

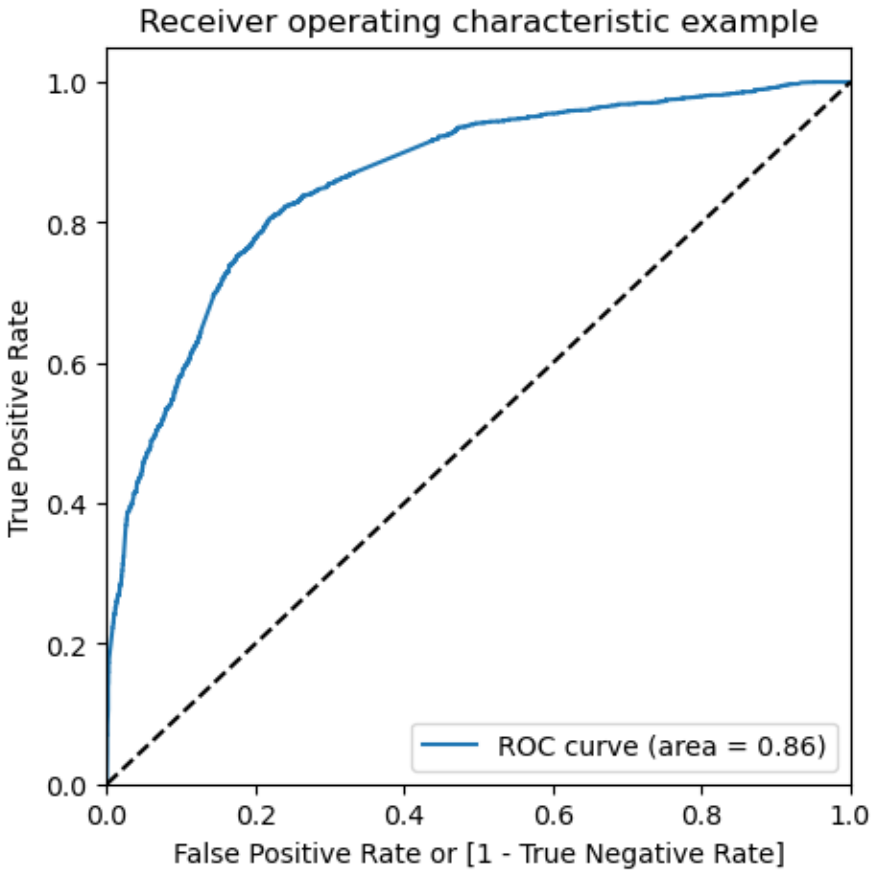
Final model

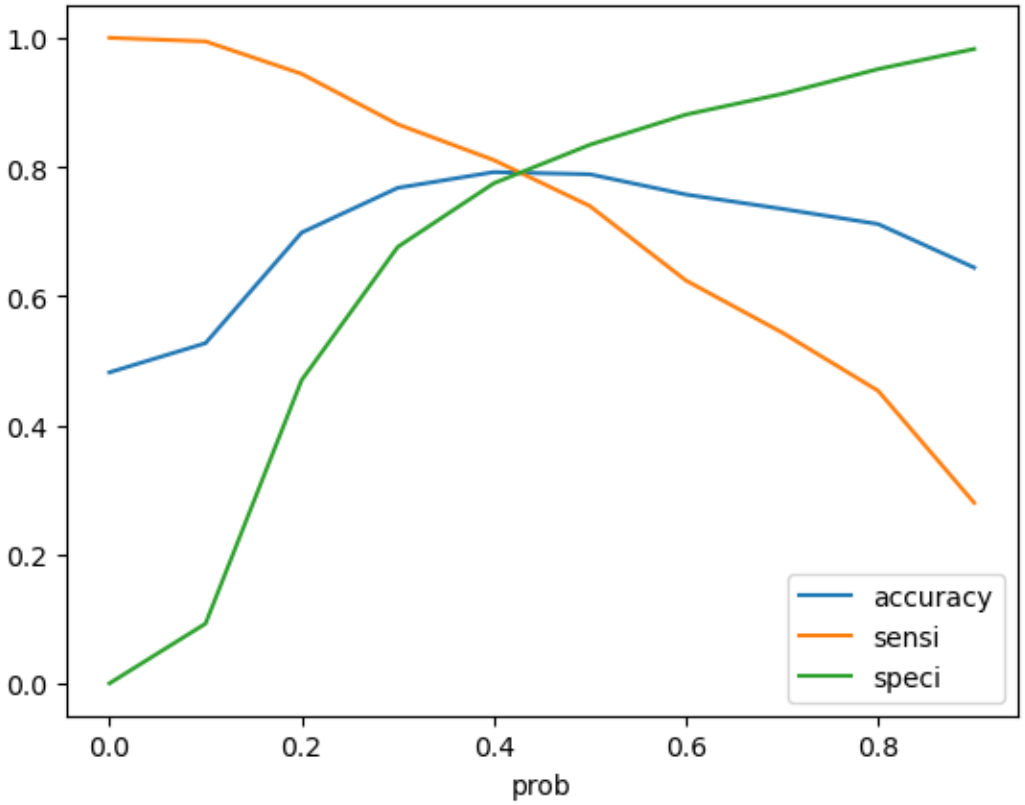
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

Model evaluation: Initial, on train set, 0.5 cut off

- Accuracy: 0.78
- Sensitivity: 0.73
- Specificity: 0.83

Model evaluation optimal cut off





Optimal cut off: 0.42

Evaluation matrices based on 0.42 cut off

- Accuracy: 0.79
- Sensitivity: 0.79
- Specificity: 0.78

Test Set predictions

- Accuracy: 0.78
- Sensitivity: 0.77
- Specificity: 0.78

Comparing Train set to Test set

Train set

- Accuracy: 0.79
- Sensitivity: 0.79
- Specificity: 0.78

Test set

- Accuracy: 0.78
- Sensitivity: 0.77
- Specificity: 0.78

Final features list

Total Time Spent on Website
TotalVisits
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Welingak Website
Do Not Email_Yes
What is your current occupation_Student
Last Activity_Had a Phone Conversation
Last Notable Activity_Unreachable

Precision Recall

Train set

- Accuracy: 0.78
- Precision : 0.78
- Recall: 0. 77

Test set

- Accuracy: 0.78
- Precision : 0.78
- Recall: 0. 78

Conclusion

- The test data values are not much different from the train data.
- Using Accuracy, Sensitivity, Specificity, Precision and Recall
- All values nearly 80 %
- So, the model is good for deployment

- Initially, identify the most promising prospects among the generated leads by evaluating metrics such as 'TotalVisits,' 'Total Time Spent on Website,' and Lead Source' These factors play a significant role in determining the likelihood of a lead converting.