Lead Scoring Case Study using Logistic Regression

DS C55

Neha Jethale
Dheeraj Jain
Jayantha S

Content :

1. Problem Statement
2. Business Objective
3. Overall Approach
4. EDA
5. Correlation
6. Model Building
7. Model Evaluation
8. Conclusion
9. Recommendations

Problem Statement :

- X education sells online courses to industry professionals. X education gets a lots 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 us to build a model and assign a lead score between 0-100 which can used a potential leads, higher the score means hot lead.
- The CEO wants to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well.

Overall Approach :

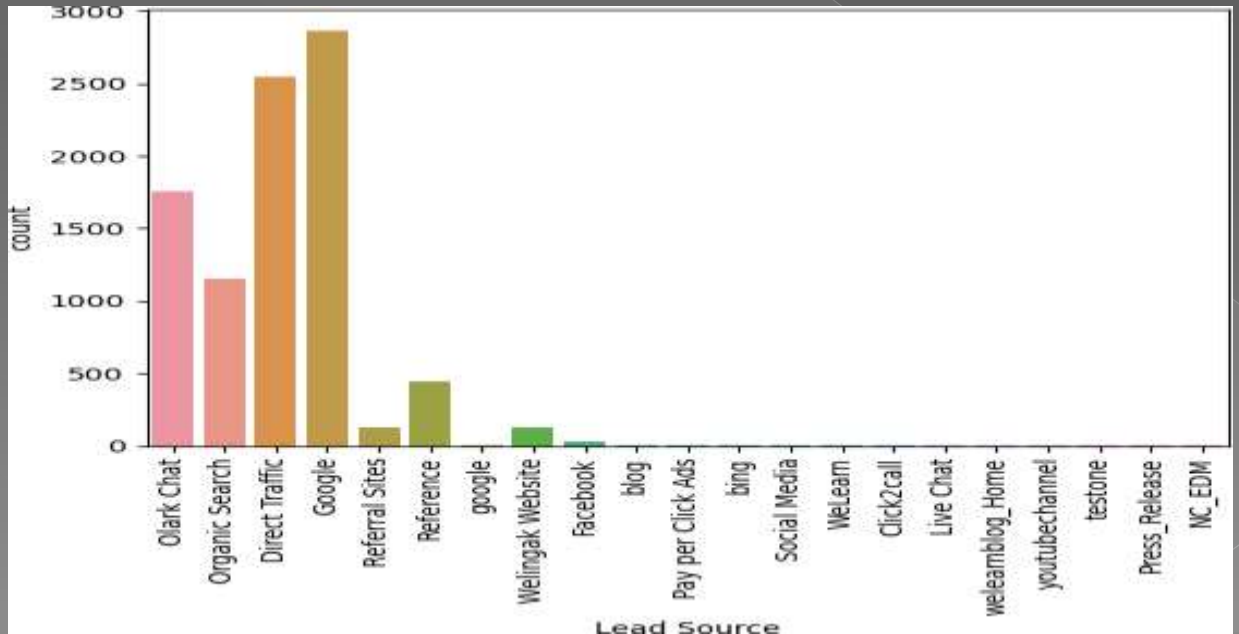
1. Importing the data and inspecting the data frame
2. Data preparation
3. EDA
4. Dummy creation
5. Test-Train split
6. Feature scaling
7. Model building
8. Model evaluation
9. Making prediction on test set
10. Conclusion
11. Recomendations

EDA :
Univariate Analysis
Bivariate Analysis

Univariate Analysis :

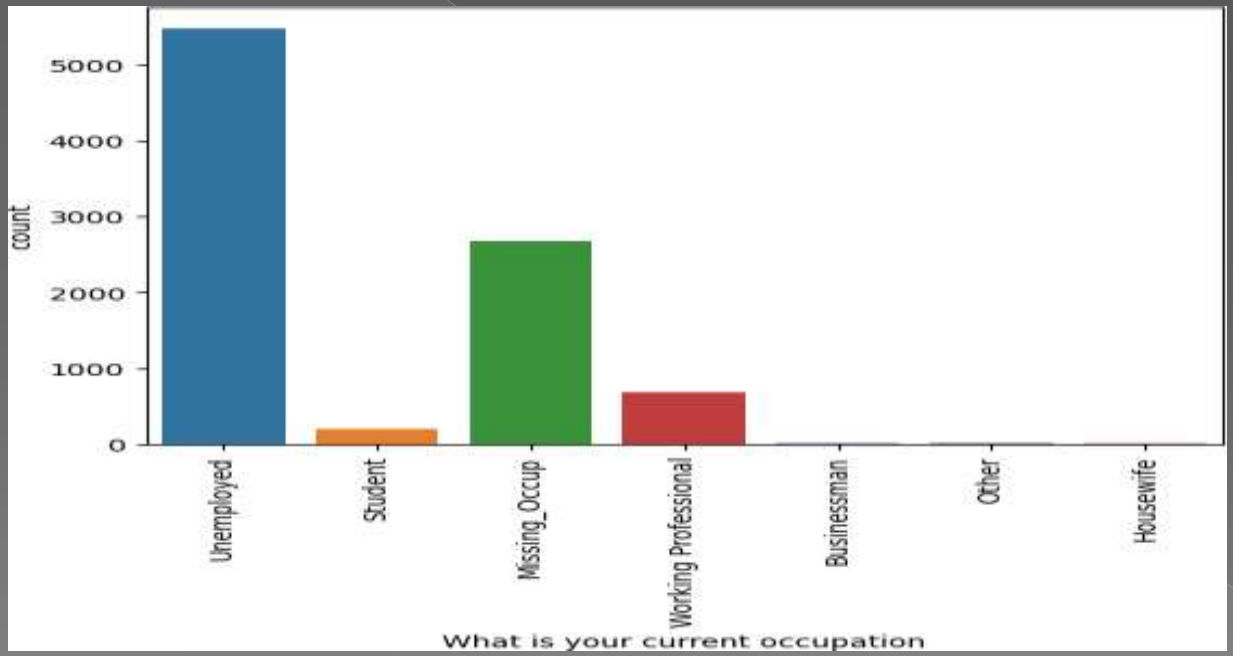
LEAD SOURCE

- Google is the main source of getting the leads



WHAT IS YOUR CURRENT OCCUPATION

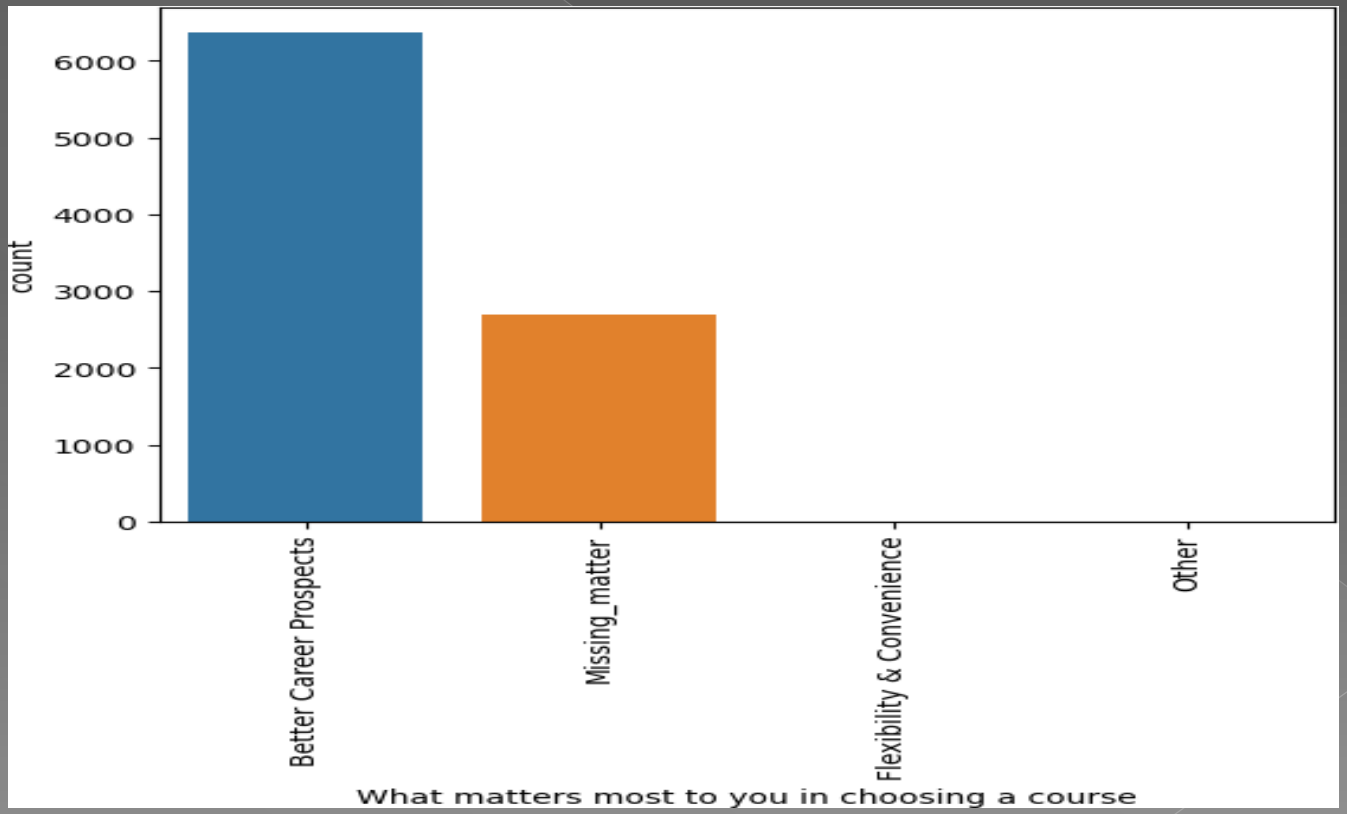
- Majority of the leads are unemployed



Univariate Analysis

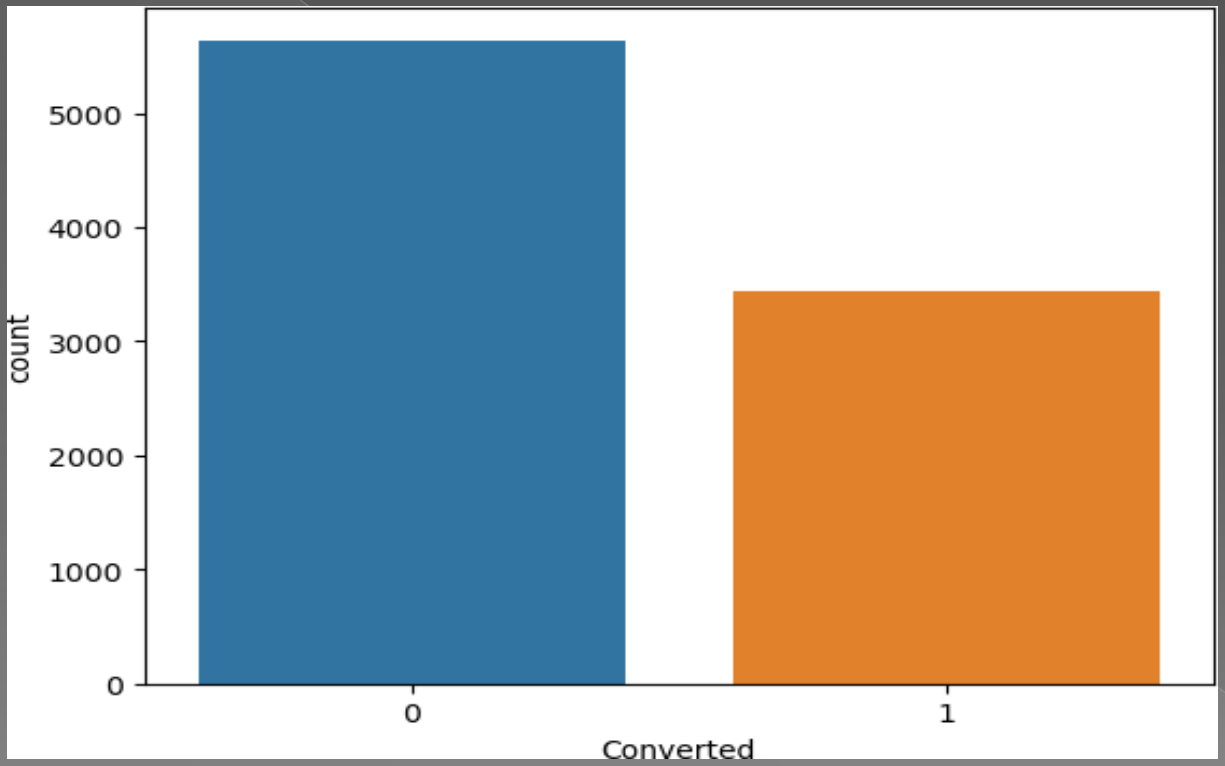
WHAT MATTERS MOST TO YOU IN CHOOSING A COURSE

- Most of the enquiries are looking for the better career prospects out of the course



CONVERTED

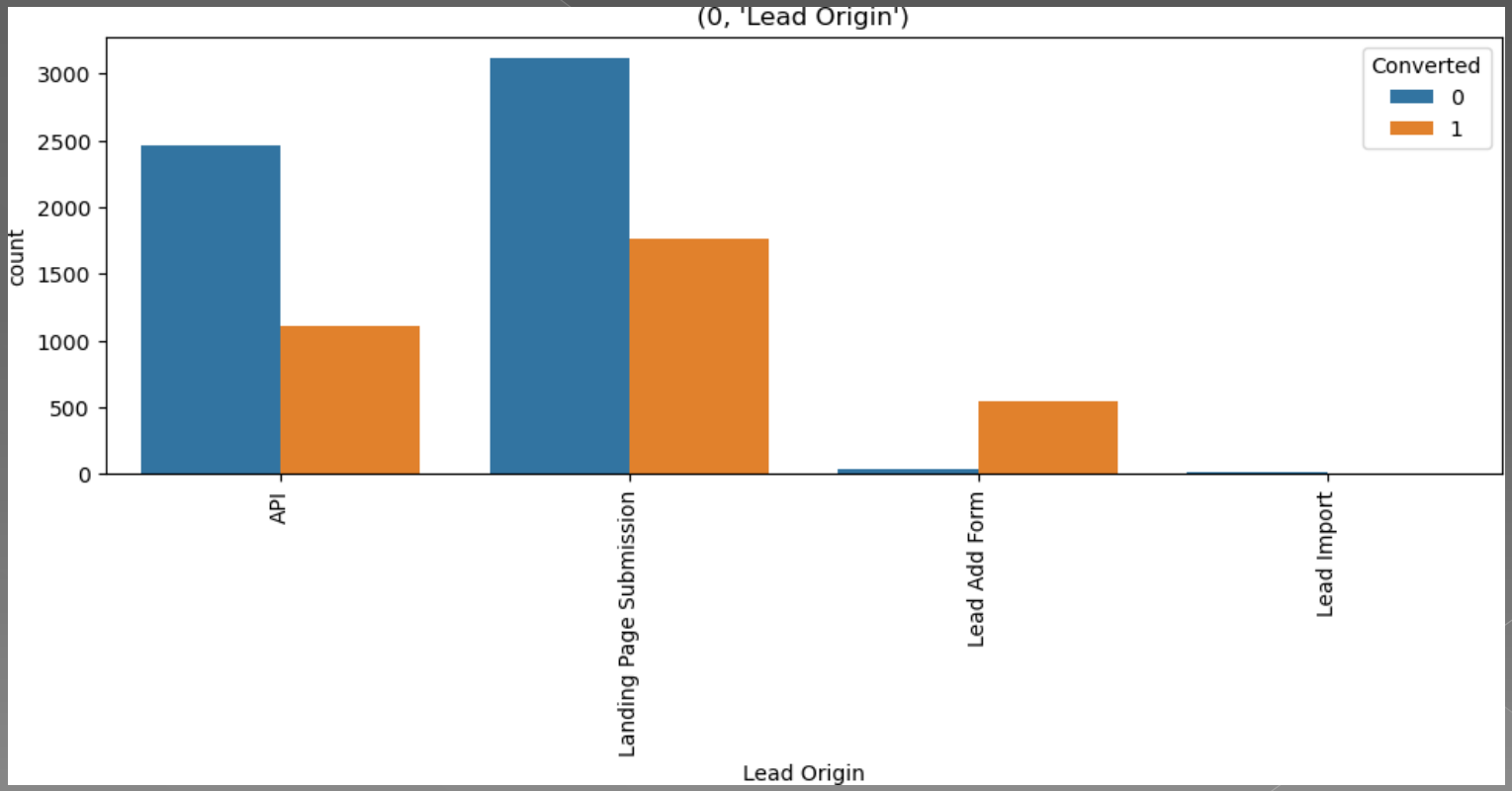
- More than 50% is the rate of conversion of the leads into orders



Bivariate Analysis

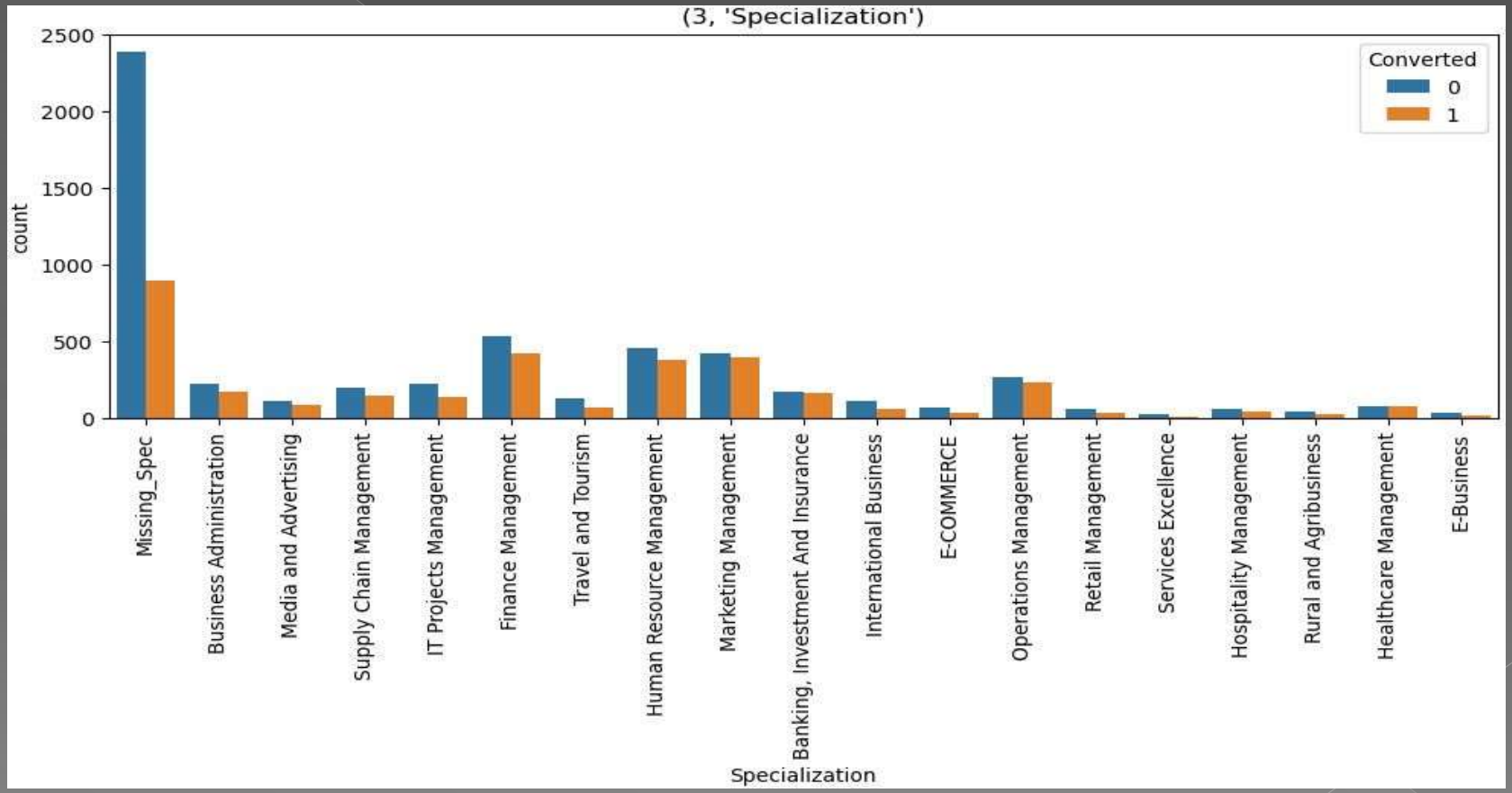
LEAD ORIGIN

- Maximum numbers of leads are converted from landing page submission



SPECIALIZATION

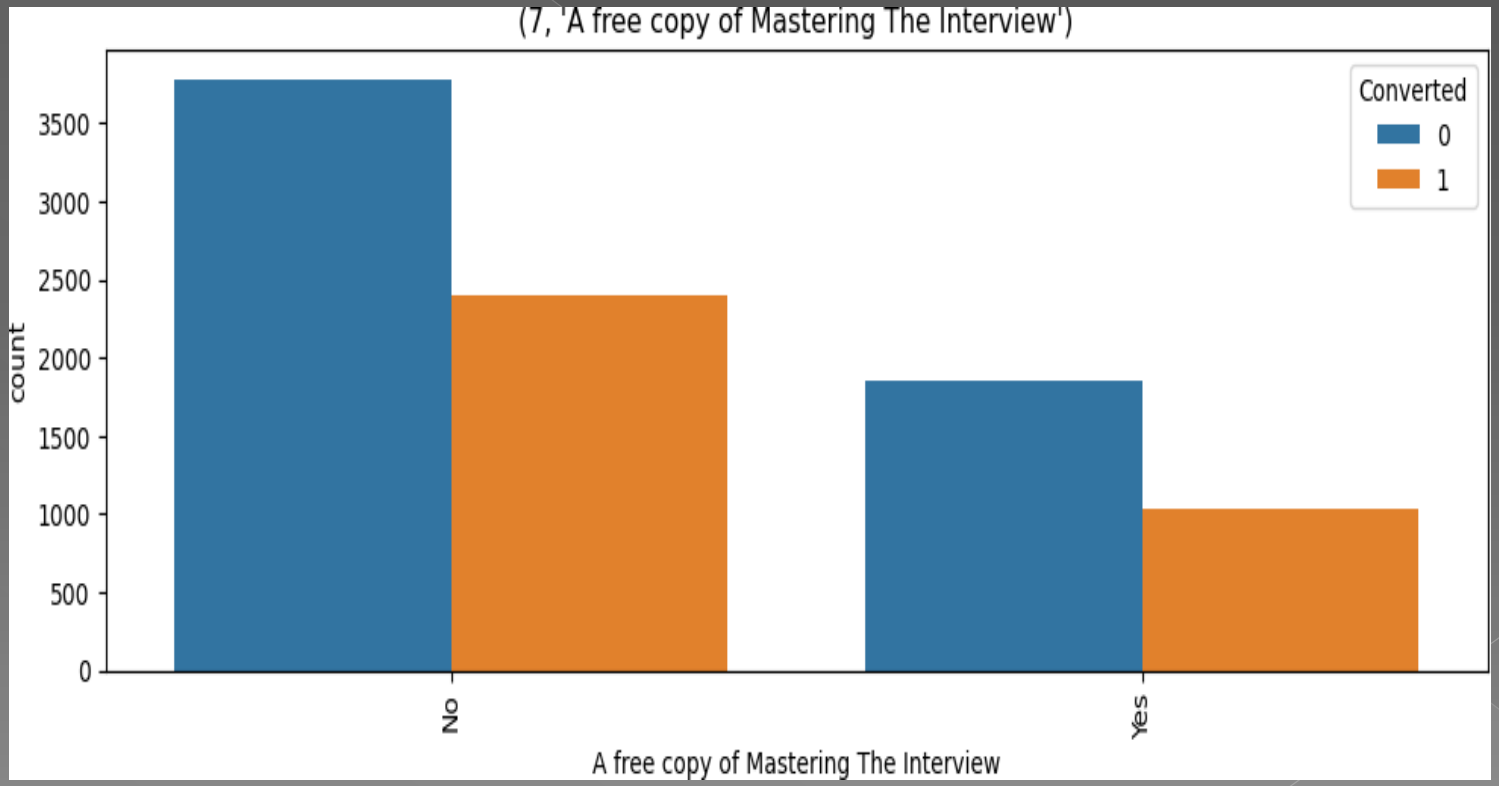
- Maximum numbers of lead conversion are into a finance management specialization



Bivariate Analysis

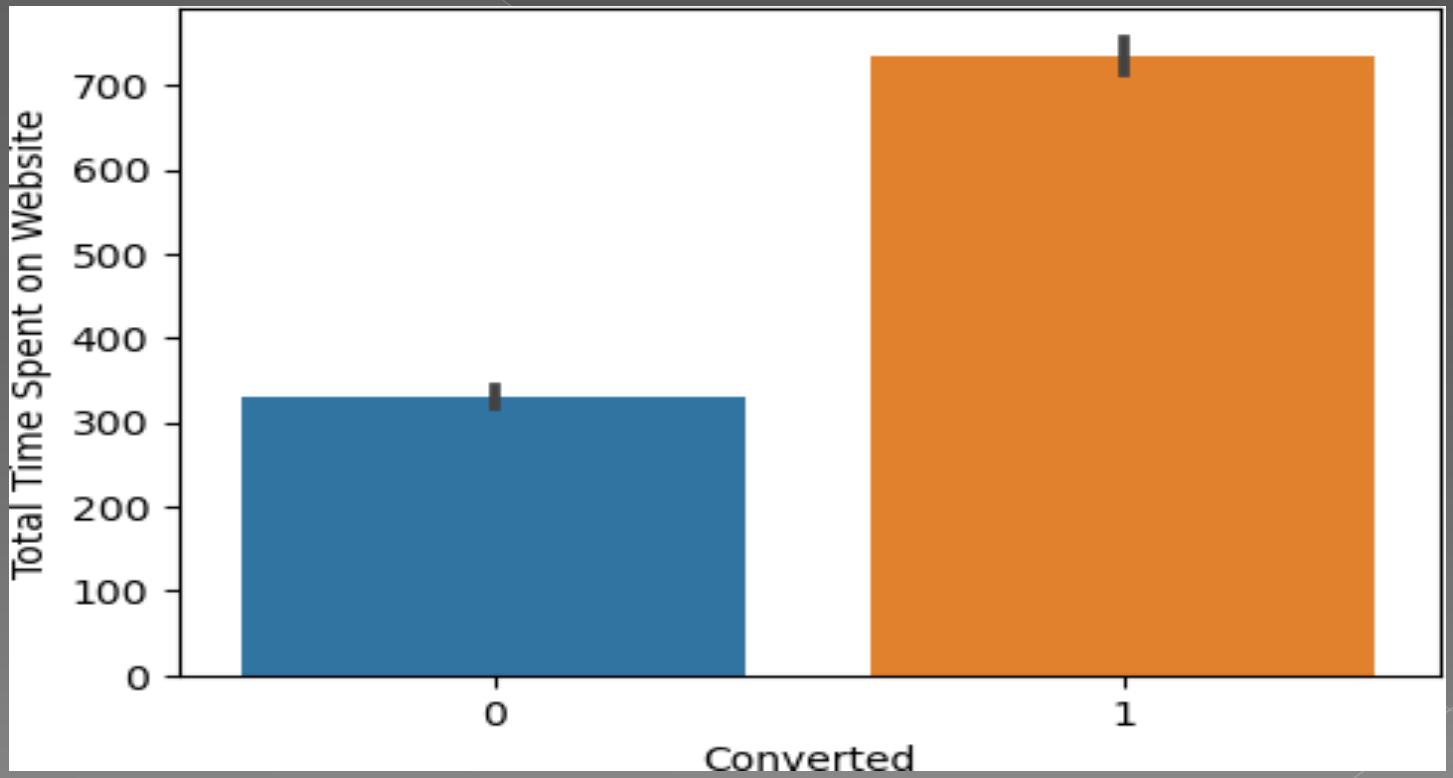
A FREE COPY OF MASTERING INTERVIEW

- Conversion rate is high on leads who do not want a free copy of Mastering Interviews



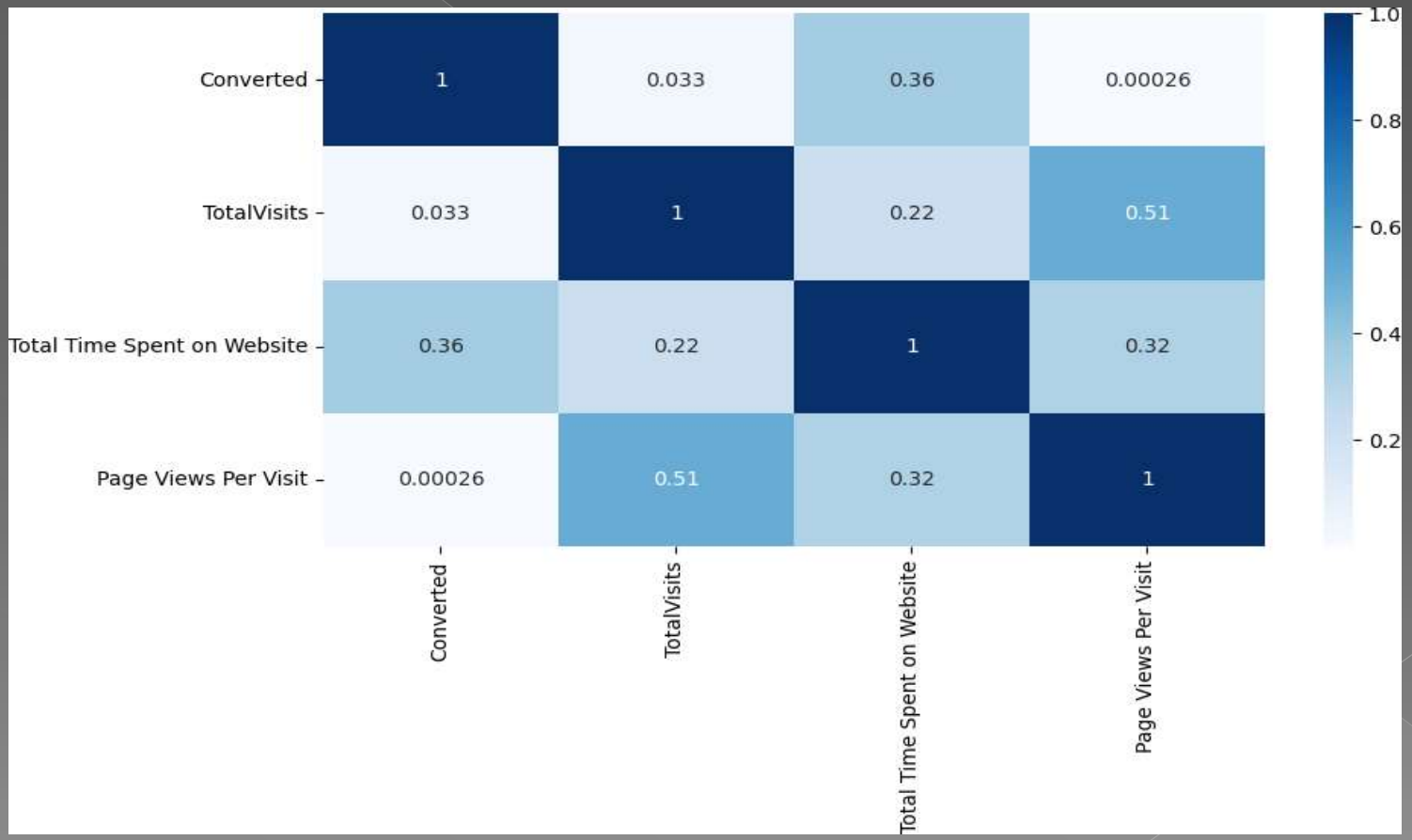
TOTAL TIME SPENT ON WEB SITE

- Conversion rate is high for the leads who spent maximum time on web site



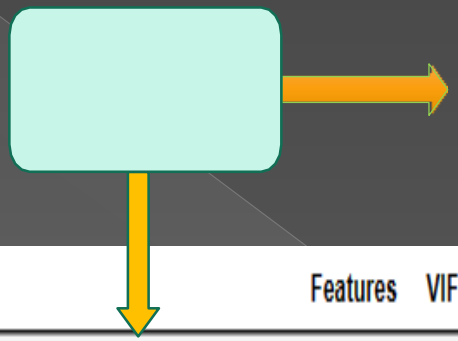
CORRELATION :

There is no High Correlation in any of the numeric variables so all these numeric variables play a vital role.



Model Building :

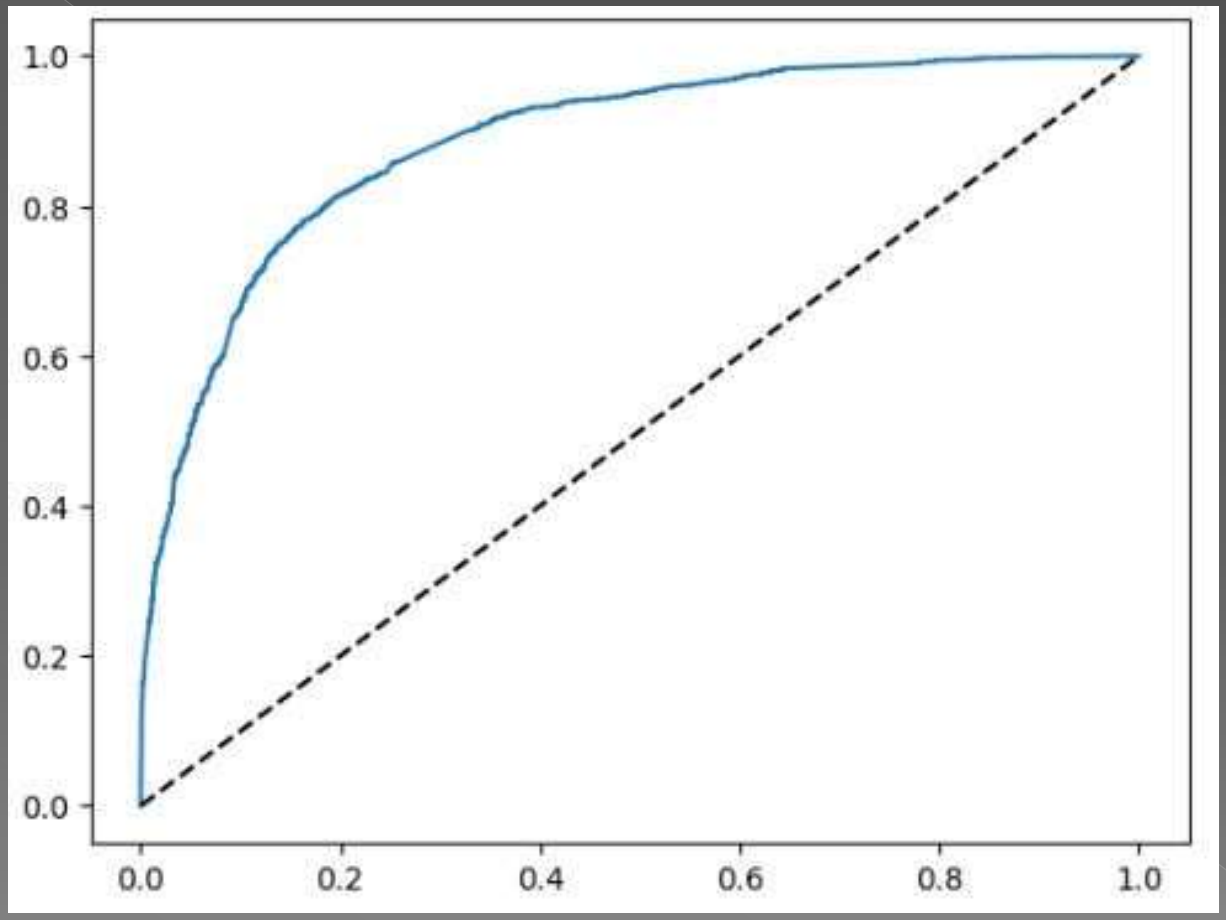
- Splitting in to train and test set
- Feature scaling
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables bases on high p-value and VIF values
- Predict using train set
- On the basis of confusion matrix evaluate accuracy and other metric
- Predict using test set
- On the basis of confusion matrix evaluate accuracy and other metric.

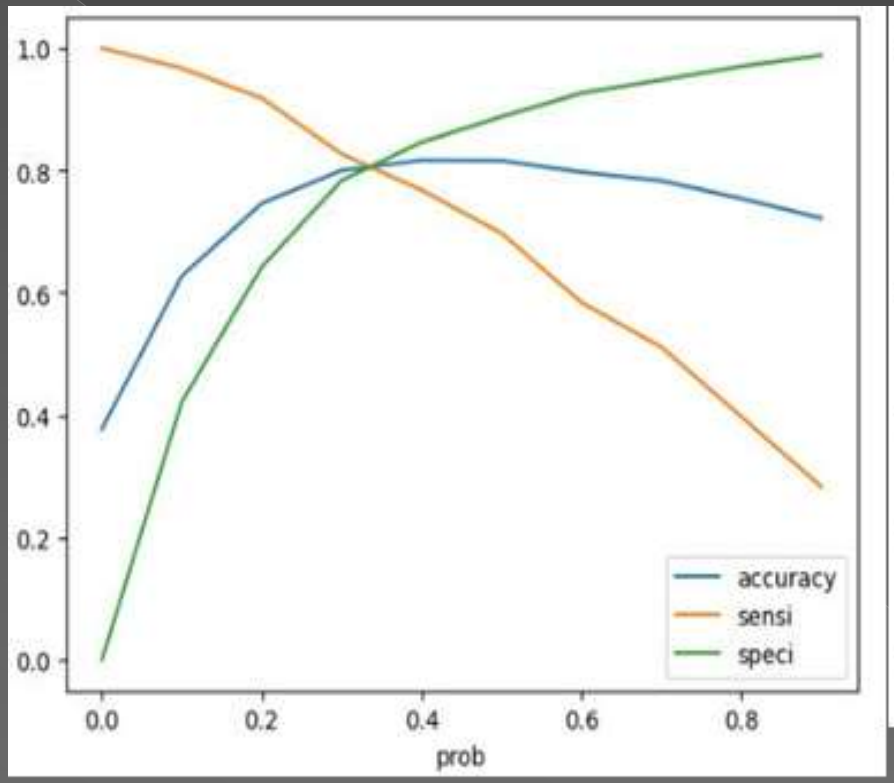


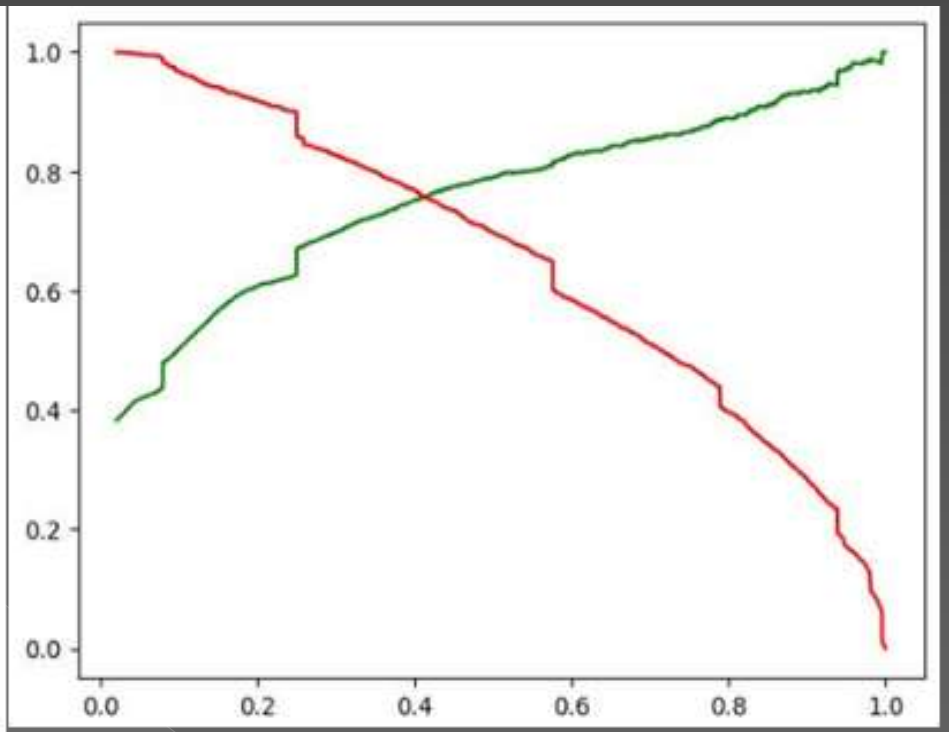
	Features	VIF
0	TotalVisits	2.72
2	Lead Origin_Landing Page Submission	2.66
1	Total Time Spent on Website	2.08
8	Last Activity_SMS Sent	1.52
3	Lead Origin_Lead Add Form	1.50
10	What matters most to you in choosing a course_...	1.46
5	Lead Source_Welingak Website	1.33
4	Lead Source_Olark Chat	1.27
9	What is your current occupation_Working Profes...	1.19
6	Last Activity_Email Bounced	1.06
7	Last Activity_Had a Phone Conversation	1.01
11	Last Notable Activity_Unreachable	1.01

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5027	0.116	-21.632	0.000	-2.730	-2.276
TotalVisits	1.5705	0.250	6.285	0.000	1.081	2.060
Total Time Spent on Website	4.6251	0.168	27.529	0.000	4.296	4.954
Lead Origin_Landing Page Submission	-0.3274	0.091	-3.617	0.000	-0.505	-0.150
Lead Origin_Lead Add Form	3.8176	0.241	15.858	0.000	3.346	4.289
Lead Source_Olark Chat	1.3998	0.132	10.596	0.000	1.141	1.659
Lead Source_Welingak Website	2.5940	1.033	2.512	0.012	0.570	4.618
Last Activity_Email Bounced	-1.4436	0.328	-4.404	0.000	-2.086	-0.801
Last Activity_Had a Phone Conversation	2.7378	0.834	3.281	0.001	1.102	4.373
Last Activity_SMS Sent	1.4067	0.075	18.740	0.000	1.260	1.554
What is your current occupation_Working Professional	2.5599	0.190	13.475	0.000	2.188	2.932
What matters most to you in choosing a course_Missing_matter	-1.3599	0.088	-15.410	0.000	-1.533	-1.187
Last Notable Activity_Unreachable	2.0065	0.575	3.487	0.000	0.879	3.134

ROC Curve with Optimal Cutoff







1. The area under the curve of the ROC is 0.88 which is quite good.
2. Optimal Cutoff Probability is 0.35 for Sensitivity and Specificity.
3. Optimal Cutoff Probability is 0.4 for Precision and Recall.

Model Evaluation – Accuracy, Sensitivity, Specificity, Train Set and Test Set

As per the Confusion Matrix of Train Set

3155	708
407	1817

Accuracy – 81.01

Sensitivity – 79.9

Specificity – 81.6

As per the Confusion Matrix of Test Set

1347	302
231	779

Accuracy – 79.95

Sensitivity – 77.12

Specificity – 81.68

Model Evaluation –Precision and Recall on Train Set and Test Set

As per the Confusion Matrix of Train Set

3268	595
544	1797

Accuracy – 81.64

Precision – 75.12

Recall – 76.76

As per the Confusion Matrix of Test Set

1391	258
259	751

Accuracy – 80.55

Precision – 74.43

Recall – 74.35

Conclusion :

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 79%, 77% and 81% 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 80%
- Hence overall this model seems to be good.

Recommendations :

- Focus on features with positive coefficients to target marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimizing communication channels based on lead engagement impact.
- Engage working professionals with tailored messaging or with the use of email.
- Incentives/discounts for providing reference that convert to lead, encourage them to providing more references for further benefits.
- Working professionals to be aggressively targeted as they have conversion rate and will have better financial situation to pay higher fees too.
- Finding out the current occupation and guiding them accordingly towards that course and giving them rewards for making other join will also be in the best interest of the organization
- Targeting people who are unemployed and explaining them the latest trends and the value of the course will also help assist them in for improvement and open doors for opportunity for them.