LEADS SCORING ASSIGNMENT PRESENTATION

DS C55 Batch:

By:

Nanda Banakar

S Tumheen Brahman

Nagarjuna Karnati

PROBLEM STATEMENT **OF THE CASE STUDY**

- An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor. or 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.

- 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%.

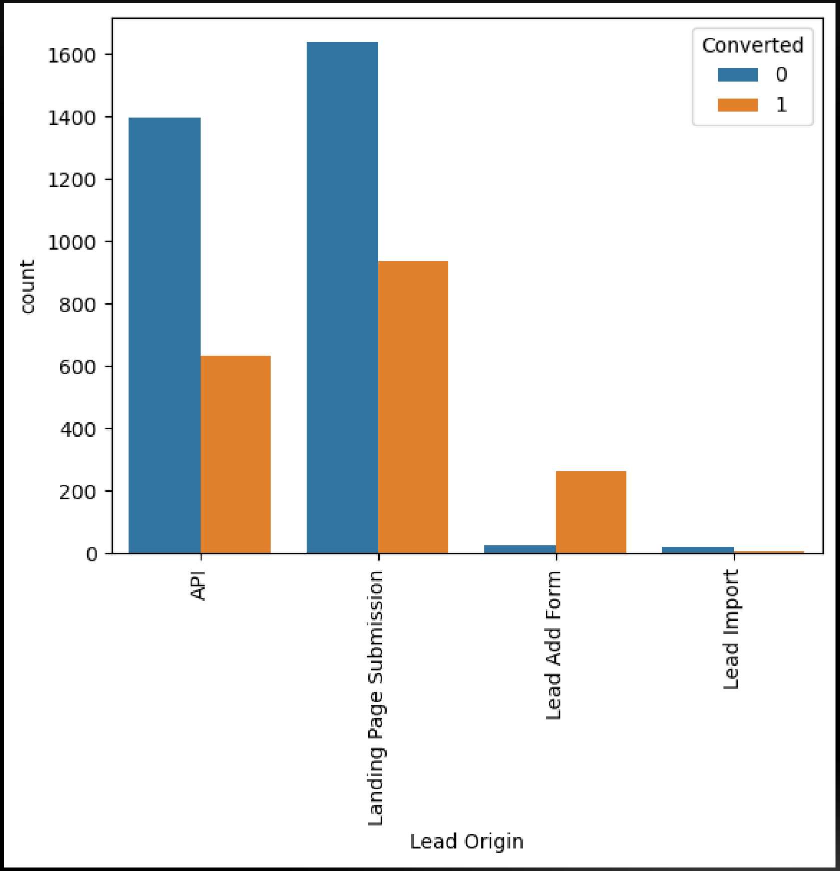
GOALS OF THE CASE STUDY

- Build a logistic regression model to 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

DATA LOADING & DATA CLEANING

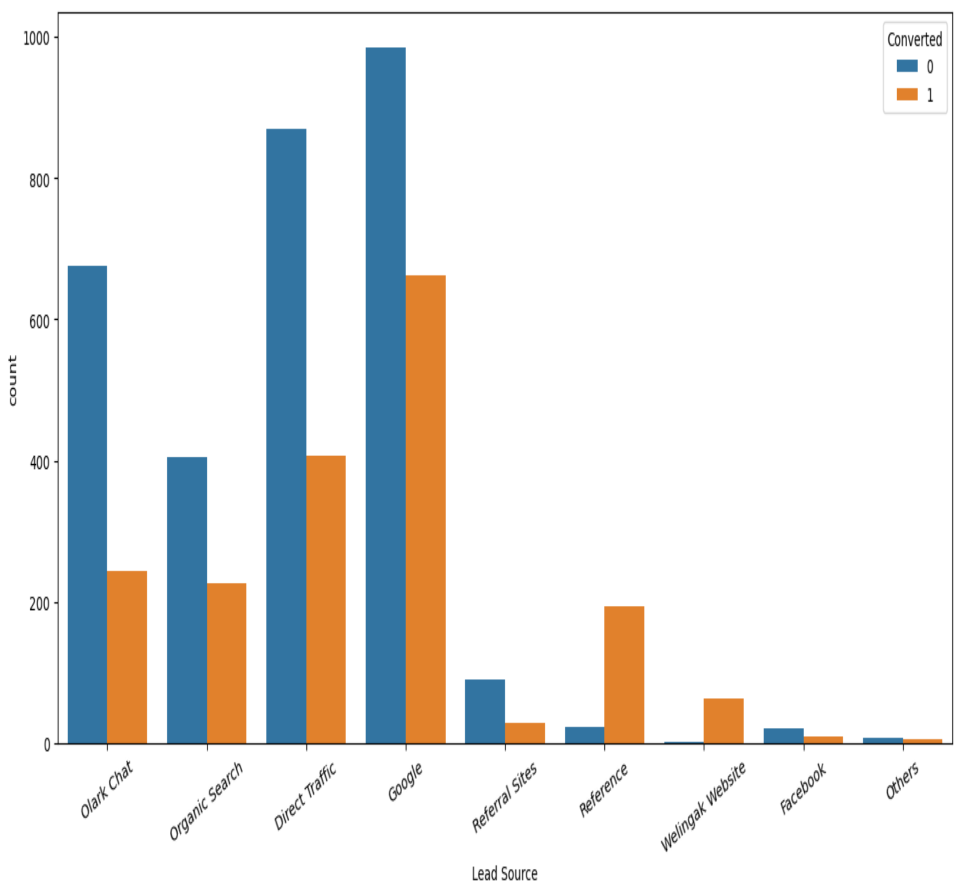
- The data received Leads.csv file has total of 9240 rows and 37 columns.
- As we can observe there are select values for many column. This is because customer did not select any option from the list, hence it shows select. Select values are as good as NULL. So these Select values are replaced by NaN.
- As column ‘Lead quality’ is based on the intuition of employee, so if left blank we can impute 'Not Sure' as NaN safely.
- For ‘City’ column around 60% of the data is Mumbai so we have imputed Mumbai in the missing values.
- It maybe the case that lead has not entered any specialization if his/her option is not available on the list, may not have any specialization or is a student. Hence we have made category "Others" for missing values.

EDA - UNIVARIATE ANALYSIS



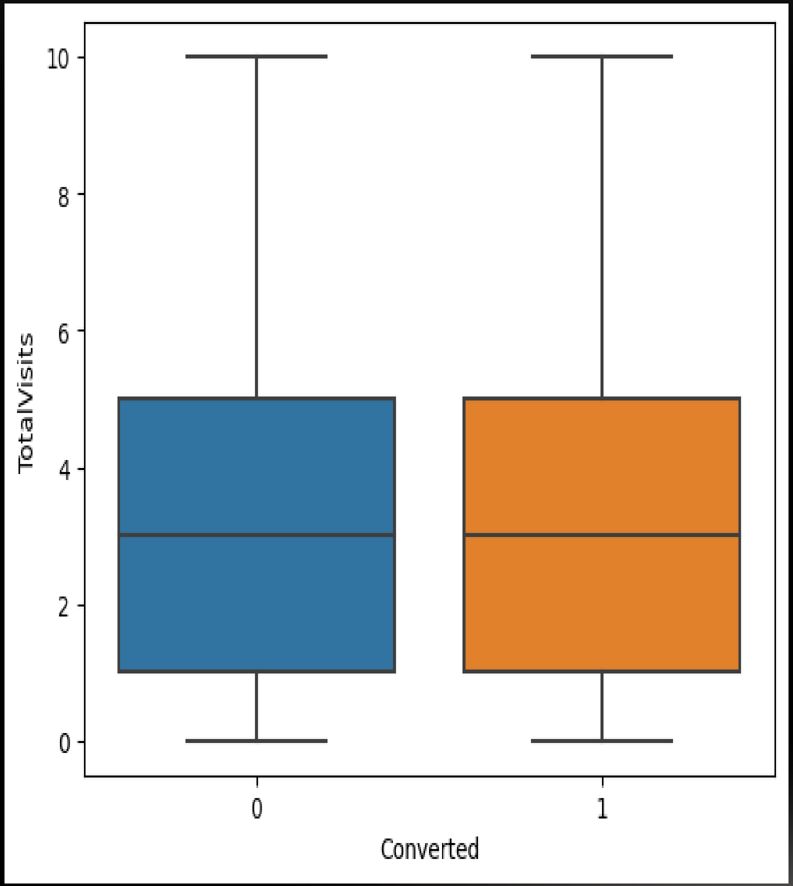
Inference on Lead Origin variable analysis:

- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- **To improve overall lead conversion rate, the focus will be on improving lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.**



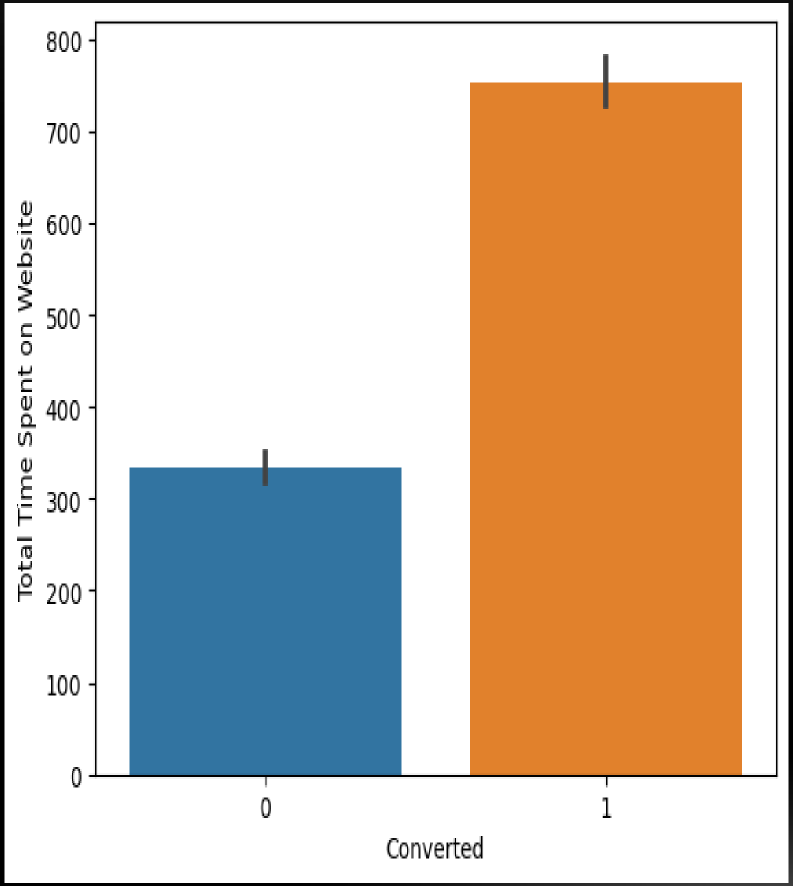
Inference on Lead Source variable analysis:

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.



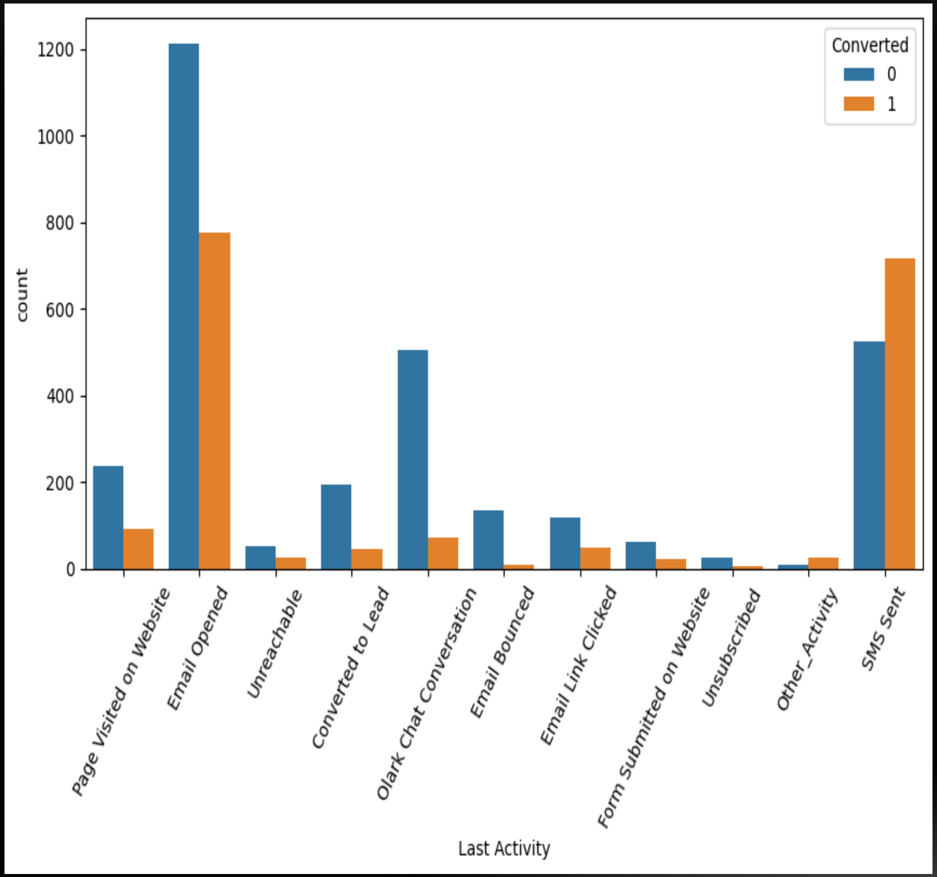
Inference on Total Visits variable analysis:

- Median for converted and non-converted leads are the same , hence Total visits are of no importance for further analysis.



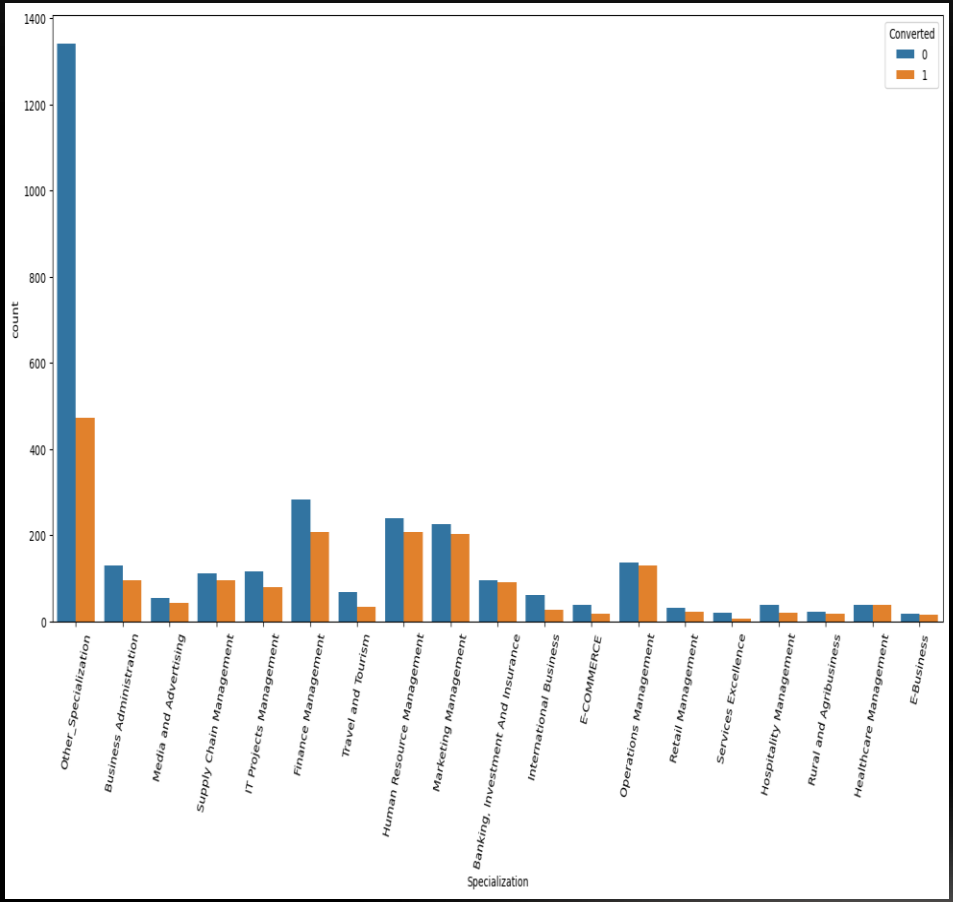
Inference on Total Time Spent on Website variable analysis:

- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.



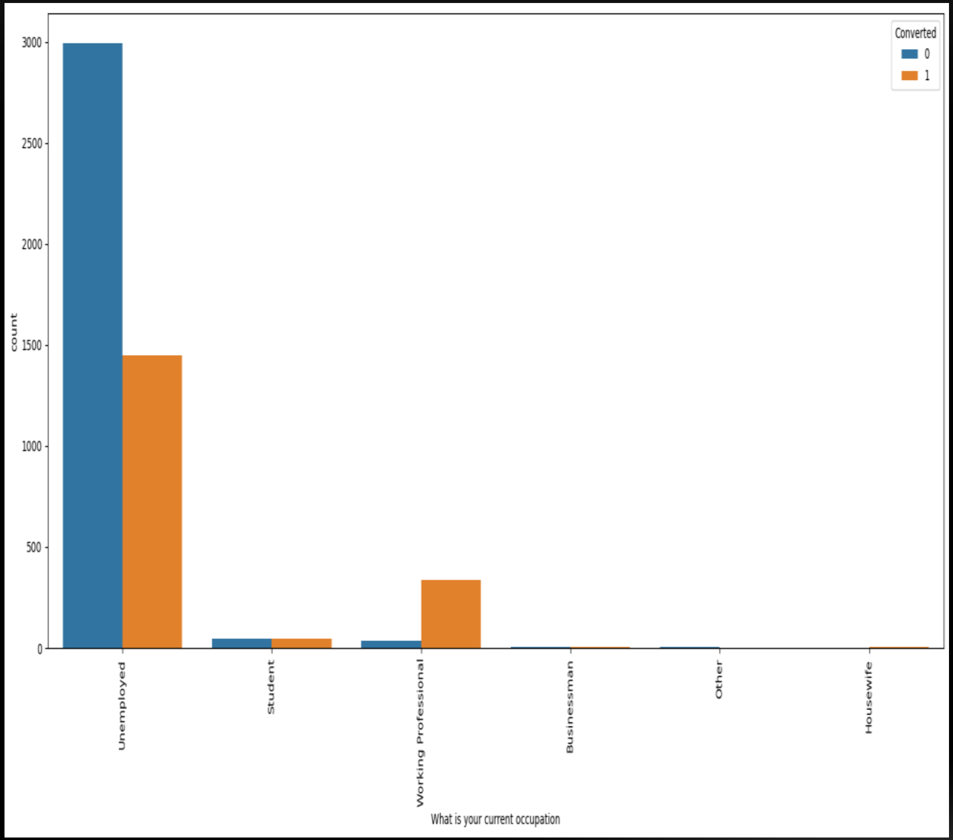
Inference on Last Activity variable analysis:

- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 60%.



Inference on Specialization variable analysis:

- Focus should be more on the Specialization with high conversion rate.
- We can focus here on Management related Professions.

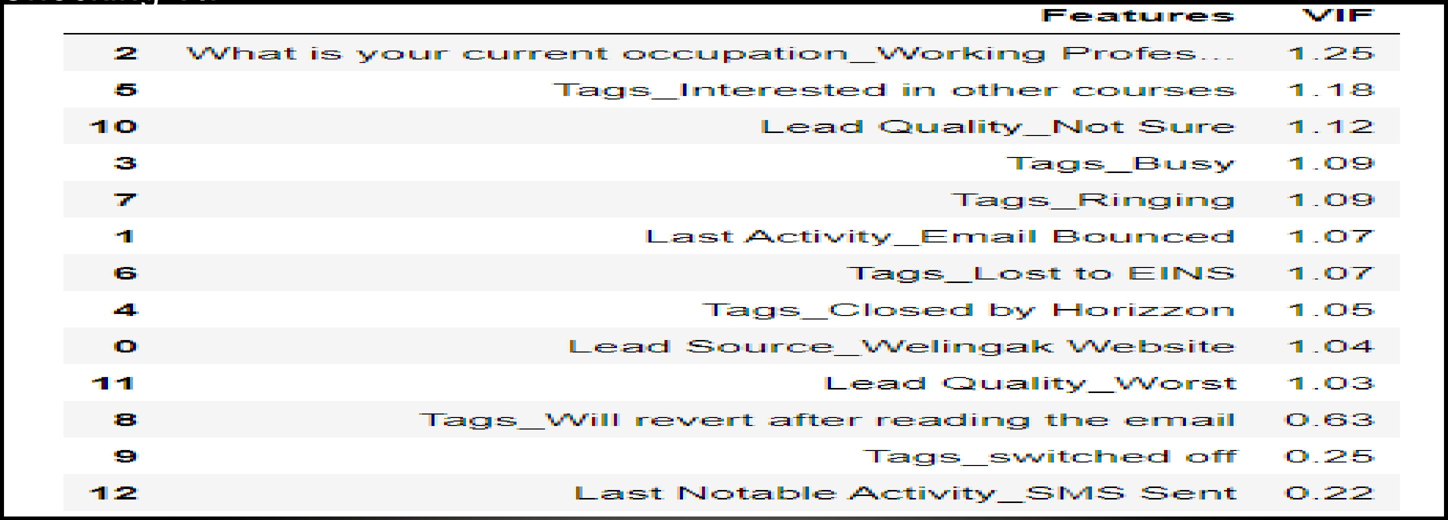


Inference on What is your current occupation variable analysis:

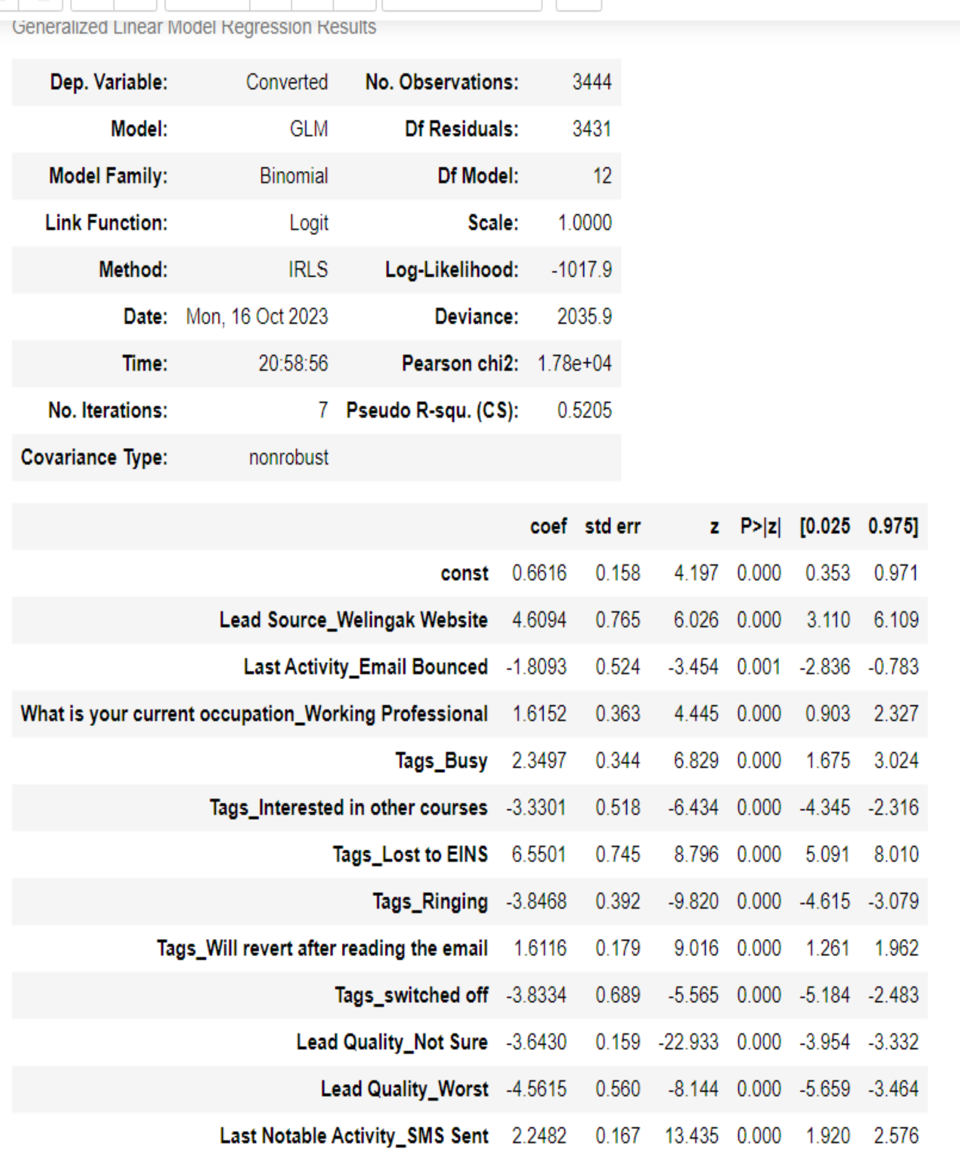
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers but has around 30-35% conversion rate.

MODEL BUILDING

- Feature scaling is done by dividing dataset into test data and train data.
- RFE (coarse tuning) Recursive feature elimination reduces model complexity by removing features one by one until the optimal number of features is left. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model.
- Checking VIF



	Features	VIF
2	What is your current occupation_Working Profes...	1.25
5	Tags_Interested in other courses	1.18
10	Lead Quality_Not Sure	1.12
3	Tags_Busy	1.09
7	Tags_Ringing	1.09
1	Last Activity_Email Bounced	1.07
6	Tags_Lost to EINS	1.07
4	Tags_Closed by Horizzon	1.05
0	Lead Source_Welingak Website	1.04
11	Lead Quality_Worst	1.03
8	Tags_Will revert after reading the email	0.63
9	Tags_switched off	0.25
12	Last Notable Activity_SMS Sent	0.22



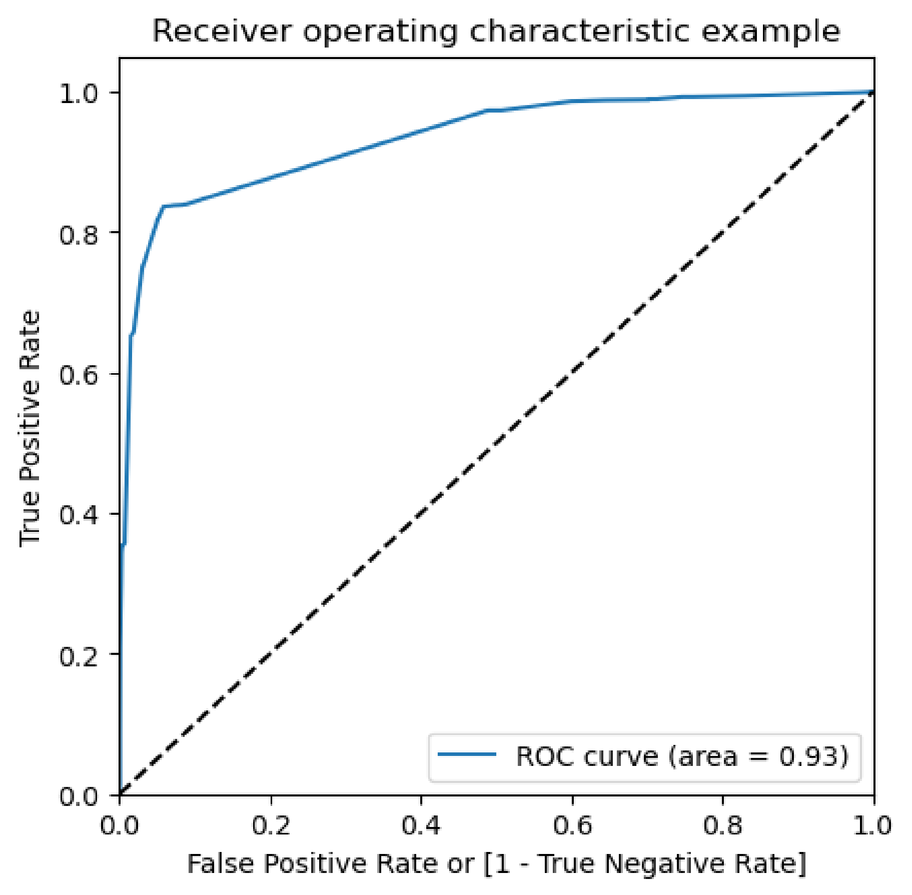
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	3444
Model:	GLM	**Df Residuals:**	3431
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1017.9
Date:	Mon, 16 Oct 2023	**Deviance:**	2035.9
Time:	20:58:56	**Pearson chi2:**	1.78e+04
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.5205
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.6616	0.158	4.197	0.000	0.353	0.971
Lead Source_Welingak Website	4.6094	0.765	6.026	0.000	3.110	6.109
Last Activity_Email Bounced	-1.8093	0.524	-3.454	0.001	-2.836	-0.783
What is your current occupation_Working Professional	1.6152	0.363	4.445	0.000	0.903	2.327
Tags_Busy	2.3497	0.344	6.829	0.000	1.675	3.024
Tags_Interested in other courses	-3.3301	0.518	-6.434	0.000	-4.345	-2.316
Tags_Lost to EINS	6.5501	0.745	8.796	0.000	5.091	8.010
Tags_Ringing	-3.8468	0.392	-9.820	0.000	-4.615	-3.079
Tags_Will revert after reading the email	1.6116	0.179	9.016	0.000	1.261	1.962
Tags_switched off	-3.8334	0.689	-5.565	0.000	-5.184	-2.483
Lead Quality_Not Sure	-3.6430	0.159	-22.933	0.000	-3.954	-3.332
Lead Quality_Worst	-4.5615	0.560	-8.144	0.000	-5.659	-3.464
Last Notable Activity_SMS Sent	2.2482	0.167	13.435	0.000	1.920	2.576

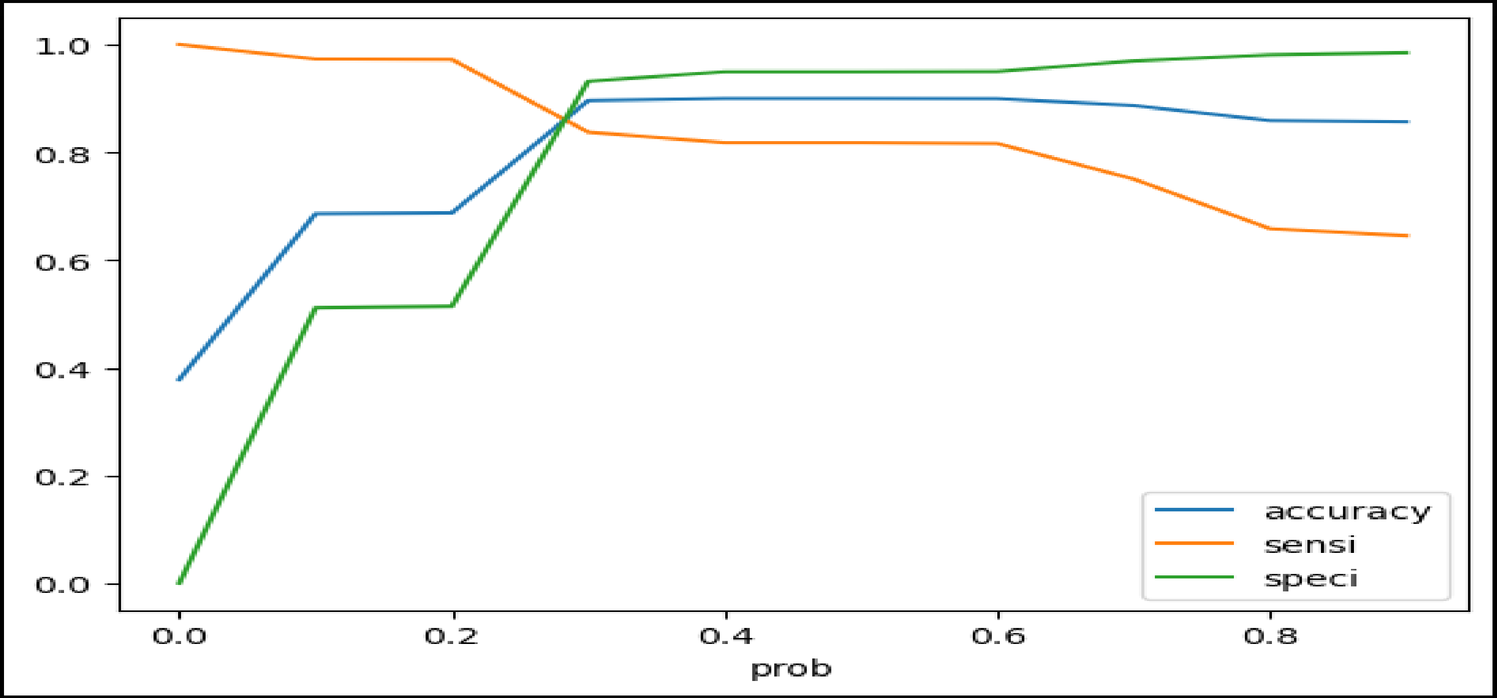
- P-value of all the predictor variables are less than < 0.05 ; indicates - statistically significant
- VIF values less than 5, it suggests that there is no strong multicollinearity among the independent variables in regression model. This is a good sign because high multicollinearity can make it challenging to interpret the impact of individual predictors.
- Hence, we can consider this as our final trained model

PLOTTING THE ROC CURVE



- **An ROC curve demonstrates several things:**
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

FINDING OPTIMAL CUTOFF POINT



CALCULATING ACCURACY SENSITIVITY SPECIFICITY ON TEST & TRAIN DATA

- After running the model on the Train Dataset these are the figures we obtain:

  Accuracy : 89.98%

  Sensitivity : 81.80%

  Specificity : 94.96%

- After running the model on the Test Dataset these are the figures we obtain:

  Accuracy : 83.37%

  Sensitivity : 85.52%

  Specificity : 91.58%

CONCLUSION

- The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost similar to train set.
- No concerns w.r.t Over - fitting. Model is in stable state.
- However, in terms of exactly predicting the lead conversion it can still do better. Since, the recall or sensitivity slightly in lower end
- **Below are the top three variables in model which contribute most towards the probability of a lead getting converted.**

  `Total Time Spent on Website`:

  - Positive contribution

  - Higher the time spent on the website, higher the probability of the lead converting into a customer

  - Sales team should focus on such leads

CONCLUSION

* `What is your current occupation - Student`:

- Negative contribution

- If the lead is already a student, chances are they will not take up another course which is designed for working professionals.

- Sales team should not focus on such leads

* `Lead Source Reference`:

- Positive contribution

- If the source of the lead is a Reference, then there is a higher probability that the lead would convert, as the referrals not only provide for cashbacks but also assurances from current users and friends who will mostly be trusted

- Sales team should focus on such leads

CONCLUSION

- Do not focus on unemployed leads. They might not have a budget to spend on the course
- Do not focus on students, since they are already studying and would not be willing to enroll into a course specially designed for working professionals, so early in the tenure