LEAD SCORING CASE STUDY

Submitted By :
Siddharth Suman Rath
Rahul Thakur

Business Objective

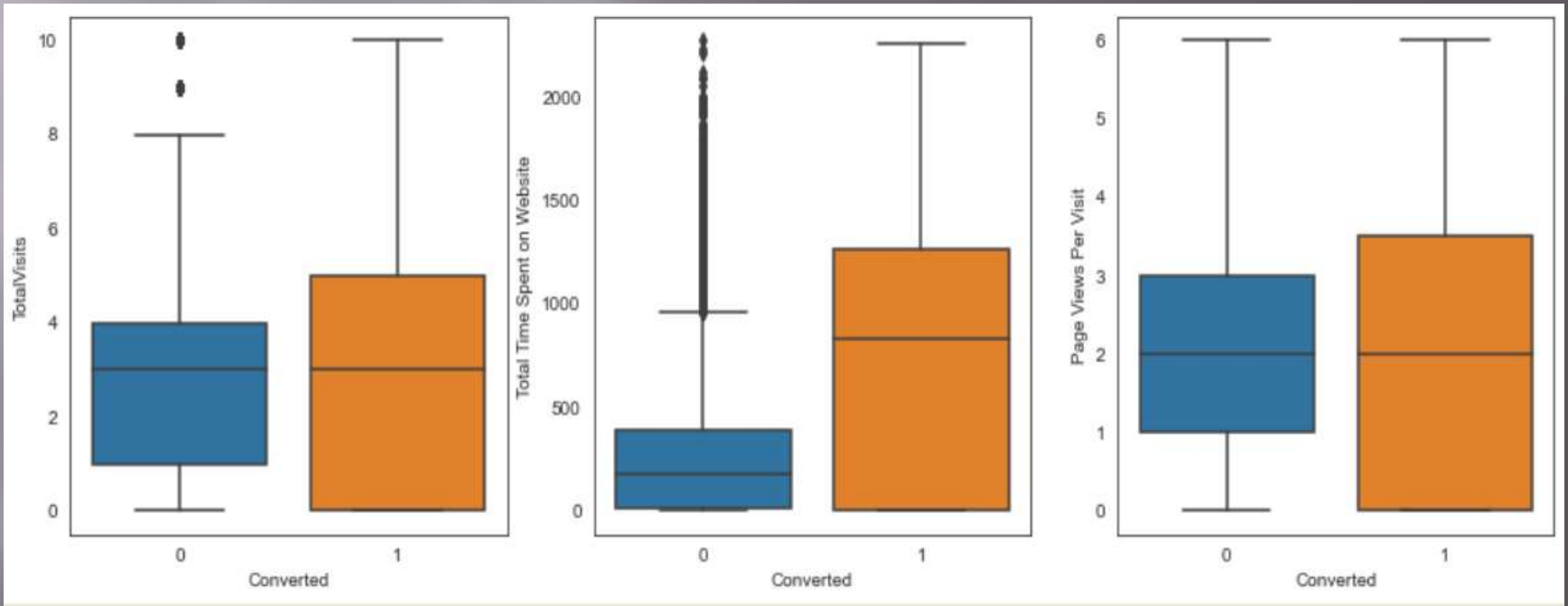
- To help X Education select most promising leads (Hot Leads)
- Hot Leads : The leads that are most likely to convert into paying customers
- Deployment of the model for the future use

Methodology

- Data cleaning and data manipulation
- EDA
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

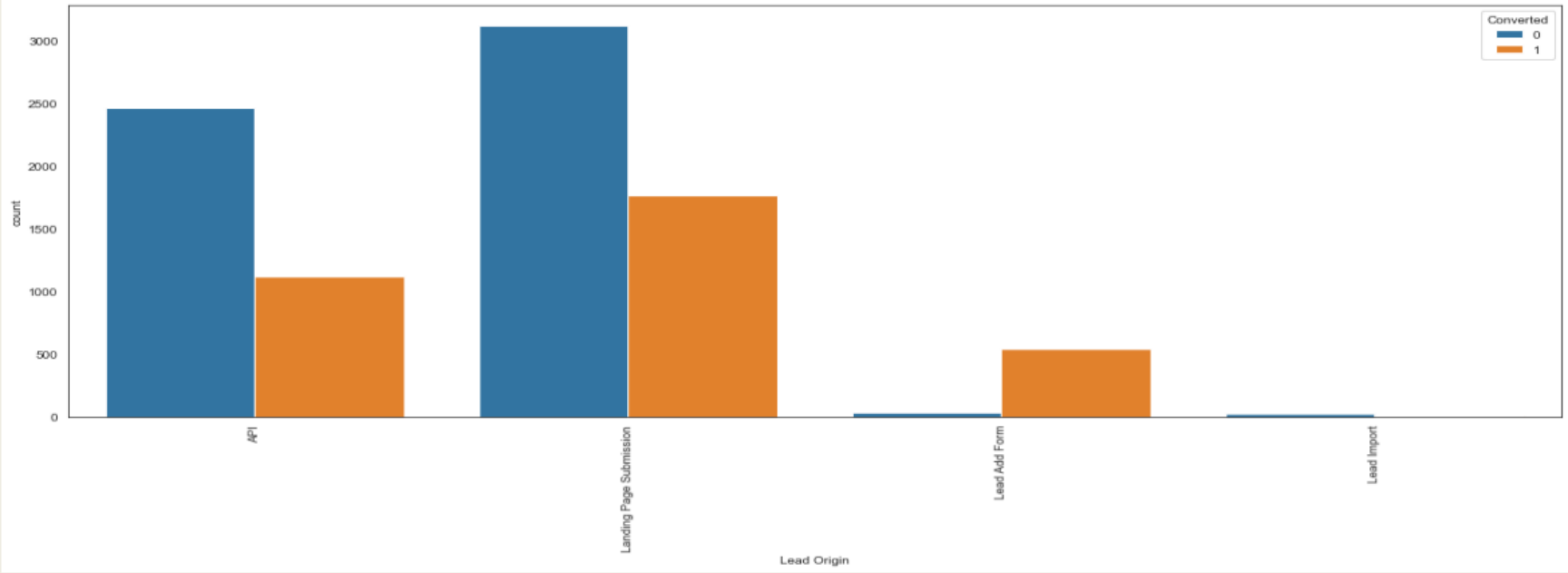
Data Visualisation

Numerical Variable



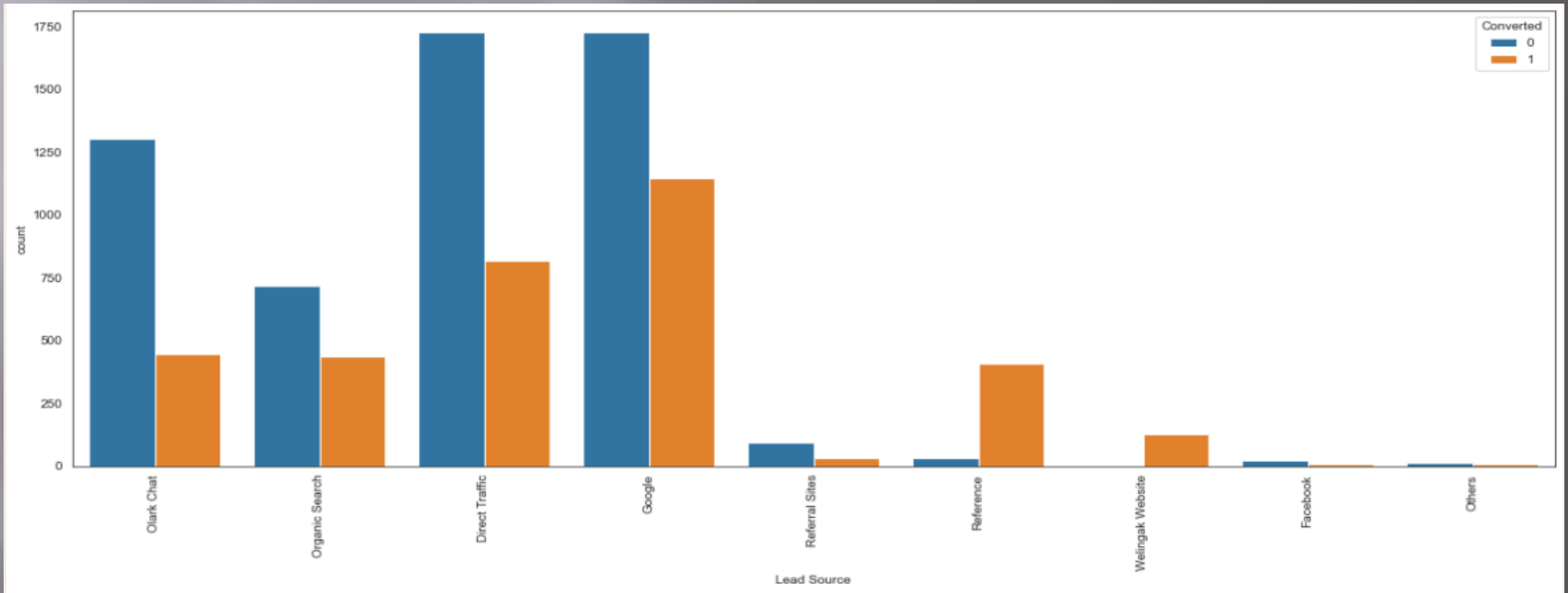
People spending more time on website are more likely to get converted.

Lead Origin

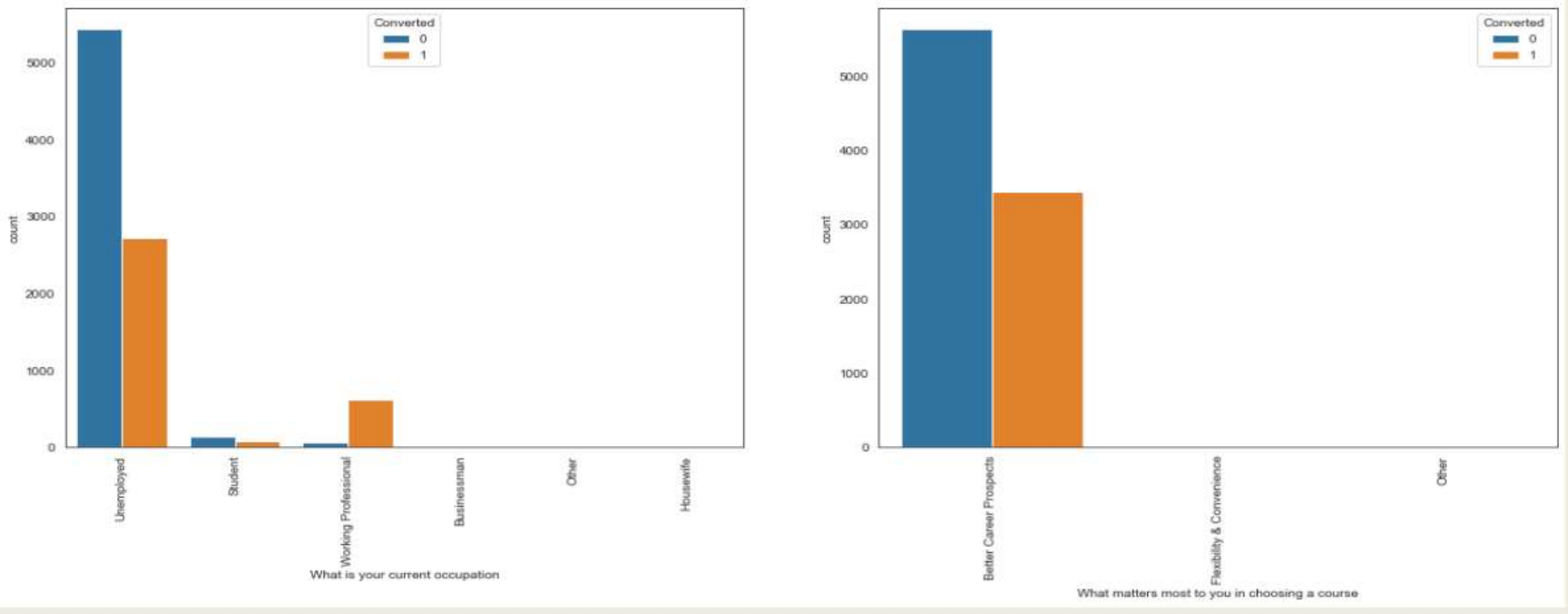


- ‘API’ and ‘Landing Page Submission’ generate the most leads but have less conversion rates, whereas ‘Lead Add Form’ generates less leads but conversion rate is great.
- Try to increase conversion rate for ‘API’ and ‘Landing Page Submission’, and increase leads generation using ‘Lead Add Form’.

Lead Source

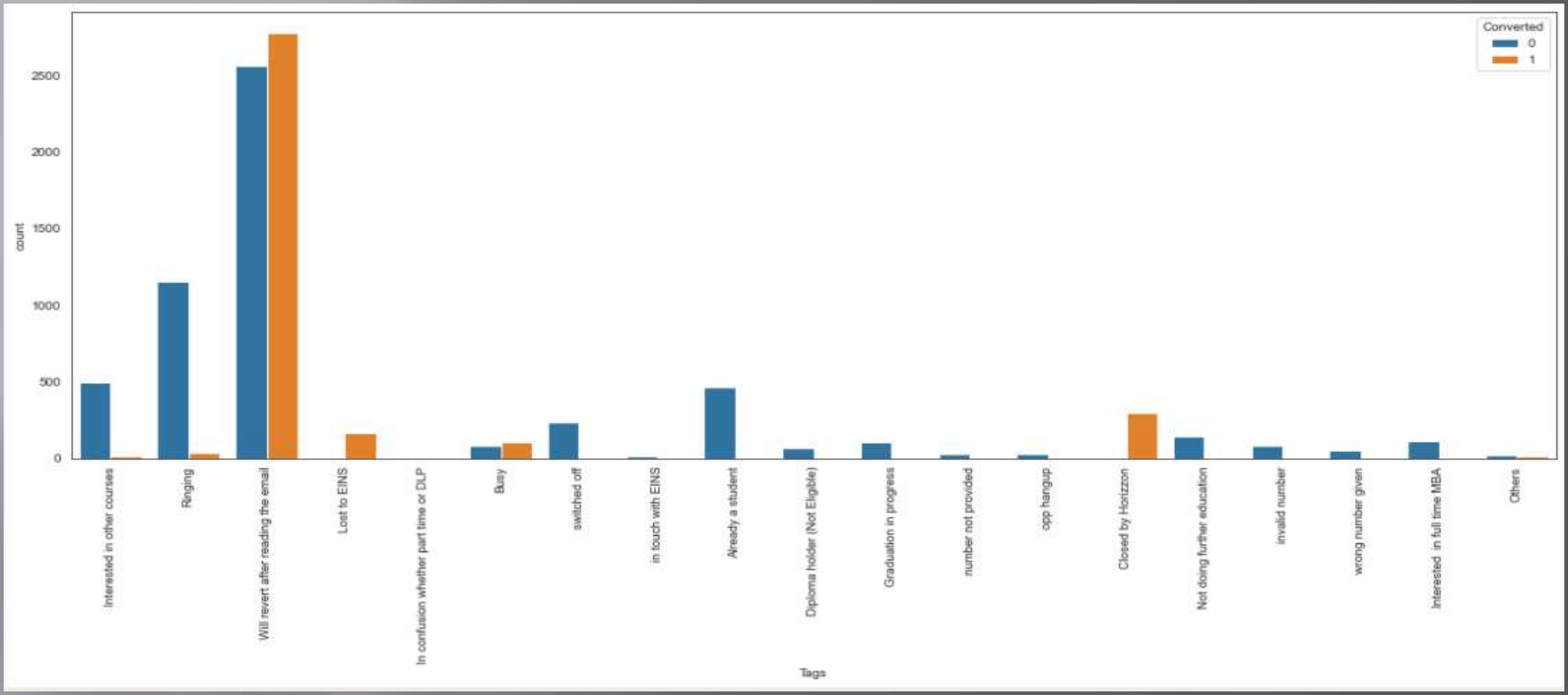

Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source

Current Occupation



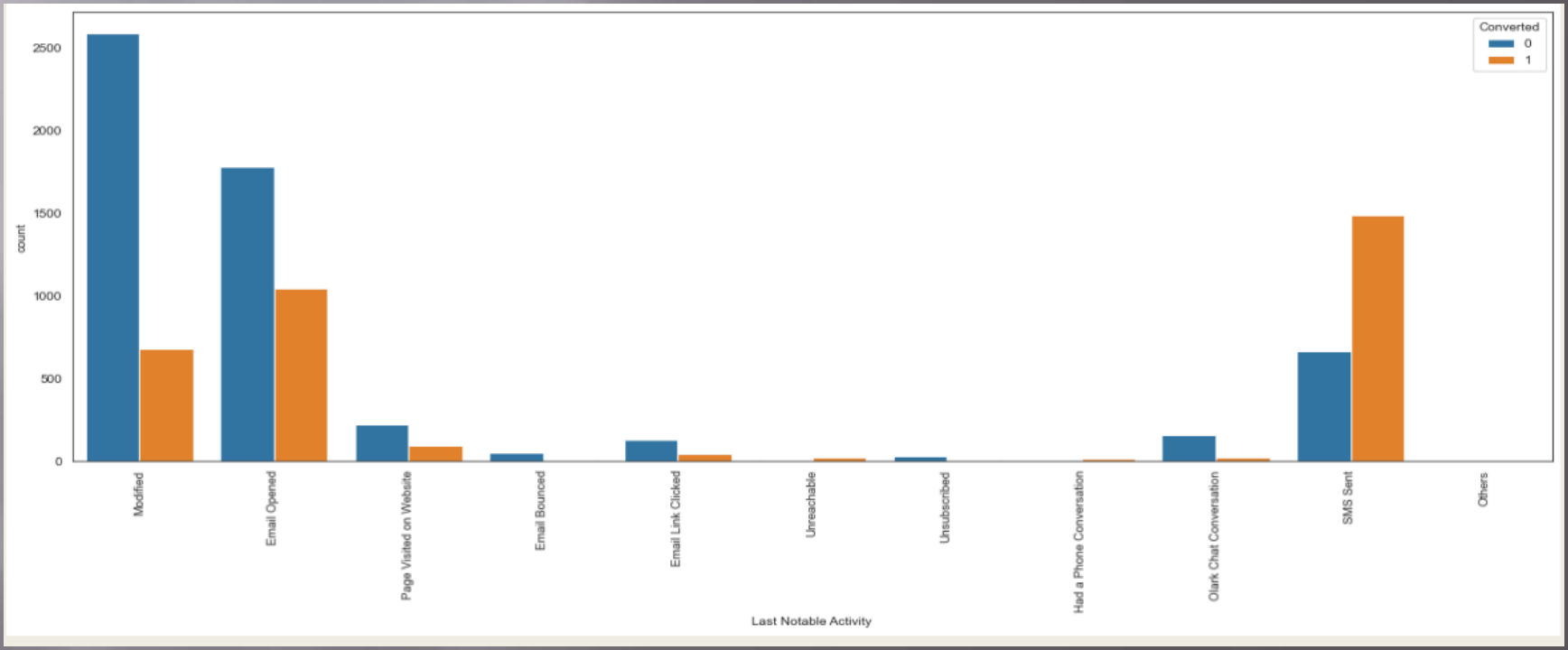
Working Professionals are most likely to get converted.

Tags



High conversion rates for tags ‘Will revert after reading the email’, ‘Closed by Horizon’, ‘Lost to EINS’, and ‘Busy’.

Last Notable Activity



Highest conversion rate is for the last notable activity ‘SMS Sent’.

Model Evaluation

Generalized Linear Model Regression Results

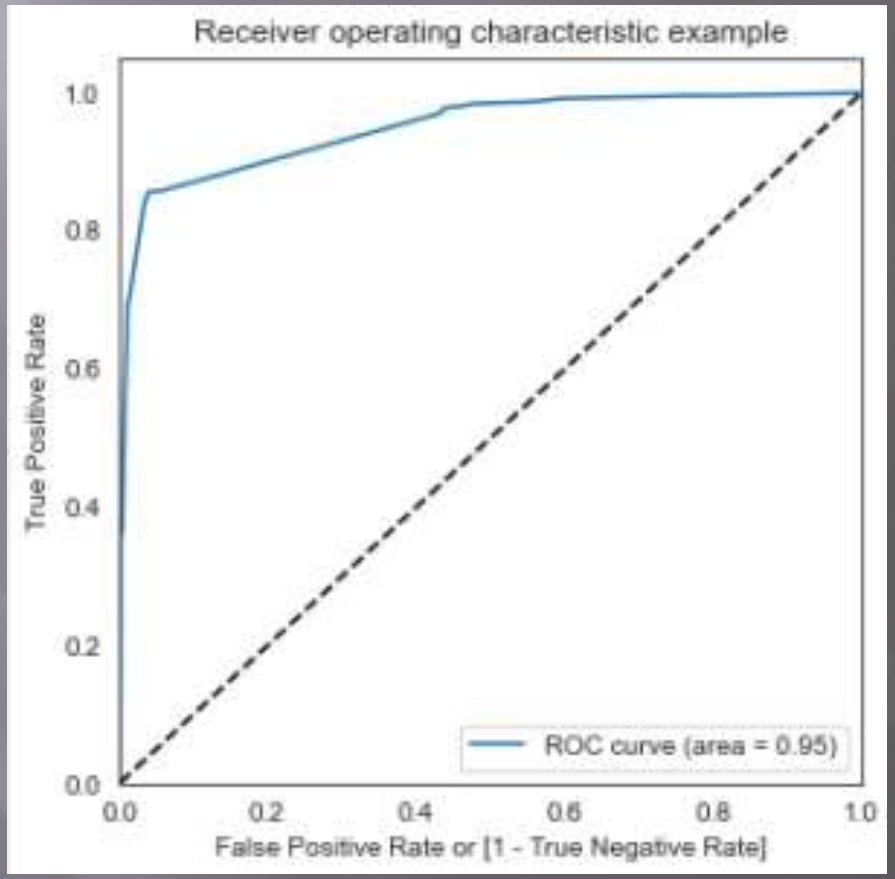
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6338
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1601.0
Date:	Mon, 18 May 2020	Deviance:	3202.0
Time:	02:23:54	Pearson chi2:	3.48e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9192	0.211	-9.080	0.000	-2.333	-1.505
Do Not Email	-1.2835	0.212	-6.062	0.000	-1.698	-0.868
Lead Origin_Lead Add Form	1.2035	0.368	3.267	0.001	0.482	1.925
Lead Source_Welingak Website	3.2825	0.820	4.002	0.000	1.675	4.890
Tags_Busy	3.8043	0.330	11.525	0.000	3.157	4.451
Tags_Closed by Horizzon	7.9789	0.762	10.467	0.000	6.485	9.473
Tags_Lost to EINS	9.1948	0.753	12.209	0.000	7.719	10.671
Tags_Ringing	-1.8121	0.336	-5.401	0.000	-2.470	-1.154
Tags_Will revert after reading the email	3.9906	0.228	17.508	0.000	3.544	4.437
Tags_switched off	-2.4456	0.586	-4.171	0.000	-3.595	-1.297
Lead Quality_Not Sure	-3.5218	0.126	-28.036	0.000	-3.768	-3.276
Lead Quality_Worst	-3.9106	0.856	-4.567	0.000	-5.589	-2.232
Last Notable Activity_SMS Sent	2.7395	0.120	22.907	0.000	2.505	2.974

All p-values are Zero

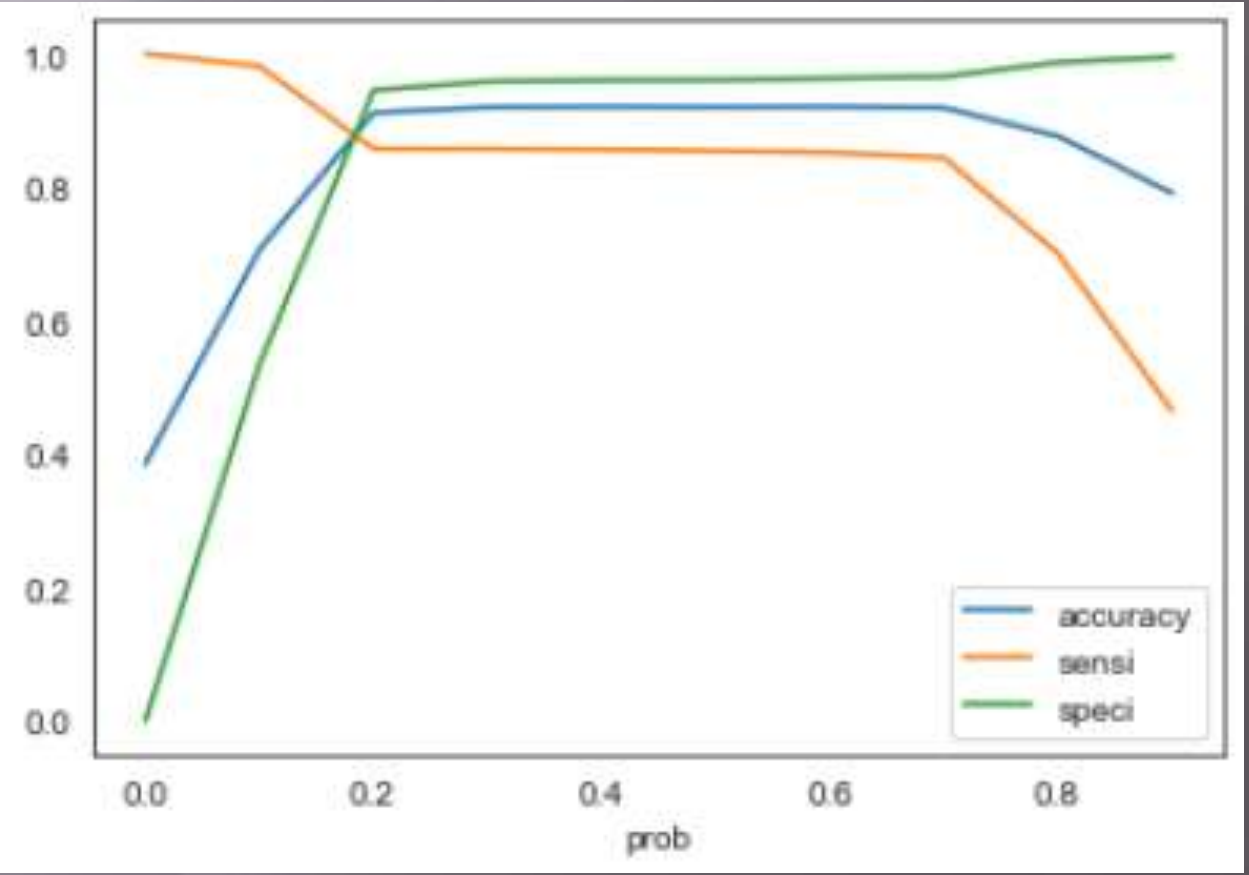
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 91%

ROC CURVE



Area under curve = 0.95

Optimum Threshold



Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values Optimal cutoff = 0.20

Features

- Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:

  I. Tags_Lost to EINS

  II. Tags_Closed by Horizzon

  III. Tags_Will revert after reading the email
- These are dummy features created from the categorical variable Tags.
- All three contribute positively towards the probability of a lead conversion.
- These results indicate that the company should focus more on the leads with these three tags

CONCLUSION

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
- When the last activity was:
    a. SMS
    b. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses

THANK YOU !