X Education Lead Scoring Model

Prepared by:
Chandrashekhar Purandare & Maganuru Basavaraju

2nd Mar 2020

Executive Context

- X Education - an education company which sells online courses to industry professionals.
- X Education markets its courses on several websites and search engines like Google.
- X Education also gets leads through past referrals.
- Though, X Education gets a lot of leads, its lead conversion rate is very poor at ~30%
- The CEO of X Education wants to overcome this challenge and set the business goal of targeting lead conversion rate of around 80%

Business Goals

- Current lead conversion rate of X education is low at around 30% which needs to be improved.
- Categorize the Leads into hot and cold by assigning a lead score to each of the open leads to improve the Lead Conversion Ratio.
- The CEO of X Education wants to increase lead conversion rate to around 80%

Leads Dataset Overview

- Leads dataset contains 37 columns and 9240 rows.
- Leads dataset contains information about origin and source of the lead, lead profile, Occupation of Lead and Domain Specialization.
- Lead data contains geographic details like country and city of the lead.
- Lead data contains clickstream data such as Total Visits, Total Time Spent on Website and Page Views Per Visit.
- Information about where the lead had seen the add such as Search, Magazine, Newspaper, X Education Forums and Digital Advertisement.
- Lead preferences such as receive more updates about the courses, update on the Supply Chain Content, updates on the DM Content, do not email and do not call.
- Finally, dataset contains the target (class) variable “Converted”, which indicates if lead is converted or not.

Analysis Methodology

Exploratory Data Analysis

- Load and Visualize Data
- Statistical Summary Analysis
- Data Cleansing
- Outlier analysis & treatment
- Univariate and Bivariate Analysis

Data Preparation

- Covert Binary Variables
- One-hot Encoding
- Train-Test Split
- Feature Scaling

Model Training

- Feature selection using RFE
- Manual Feature Selection using p-Value and VIF

Model Validation

- Accuracy
- Confusion Matrix
- Sensitivity and Specificity
- False Positive Rate
- Positive & Negative predictive value
- ROC Curve
- Optimal Cutoff Point
- Precision and Recall Trade-off

Prediction on Test Data

- Feature Scaling
- Prediction on Test Data
- Model Evaluation using Test Data

Generate Lead Score

- Merge Train & Test Data
- Assign Lead Score
- Export Results to CSV

Missing Data Summary

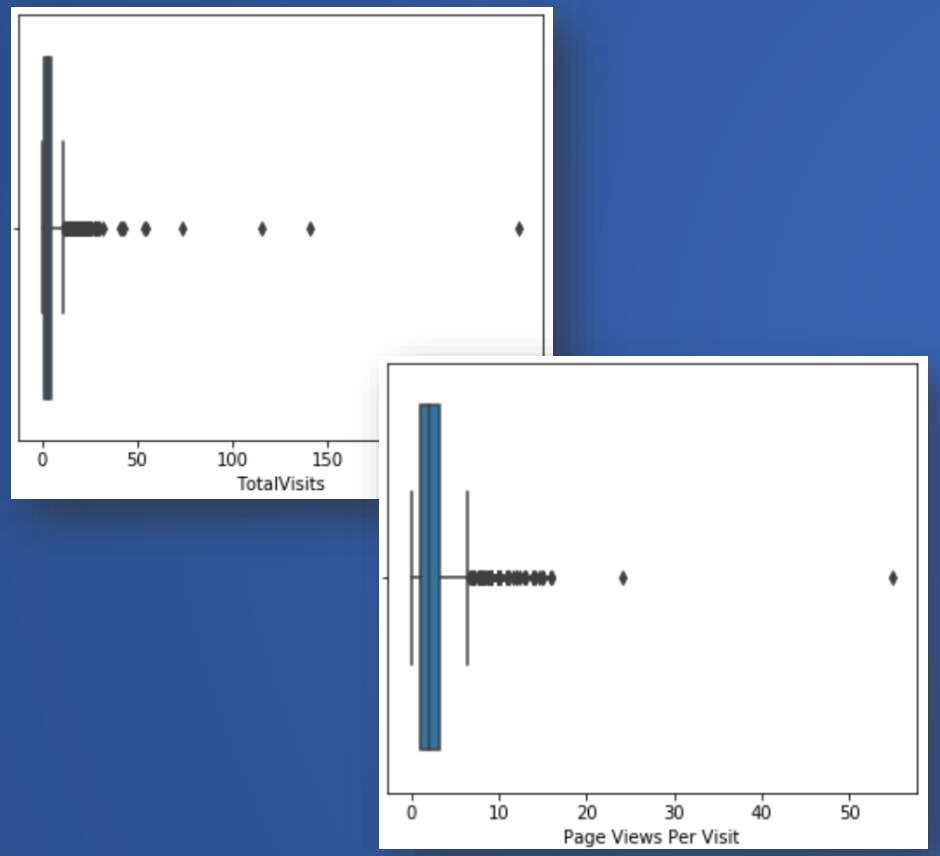
Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	36.58
How did you hear about X Education	78.46
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0.00
Magazine	0.00
Newspaper Article	0.00
X Education Forums	0.00
Newspaper	0.00
Digital Advertisement	0.00
Through Recommendations	0.00
Receive More Updates About Our Courses	0.00
Tags	36.29
Lead Quality	51.59
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
Lead Profile	74.19
City	39.71
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65
I agree to pay the amount through cheque	0.00
A free copy of Mastering The Interview	0.00
Last Notable Activity	0.00

Original Dataset

Prospect ID	0.0
Lead Number	0.0
Lead Origin	0.0
Lead Source	0.0
Do Not Email	0.0
Do Not Call	0.0
Converted	0.0
TotalVisits	0.0
Total Time Spent on Website	0.0
Page Views Per Visit	0.0
Last Activity	0.0
Country	0.0
Specialization	0.0
What is your current occupation	0.0
What matters most to you in choosing a course	0.0
Search	0.0
Magazine	0.0
Newspaper Article	0.0
X Education Forums	0.0
Newspaper	0.0
Digital Advertisement	0.0
Through Recommendations	0.0
Receive More Updates About Our Courses	0.0
Tags	0.0
Lead Quality	0.0
Update me on Supply Chain Content	0.0
Get updates on DM Content	0.0
City	0.0
I agree to pay the amount through cheque	0.0
A free copy of Mastering The Interview	0.0
Last Notable Activity	0.0

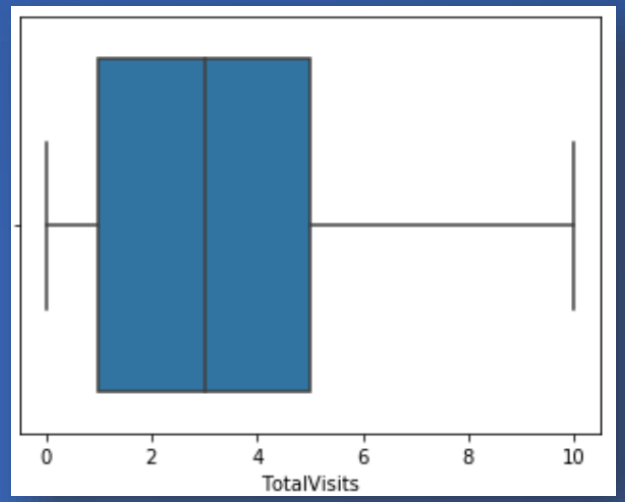
Transformed Dataset

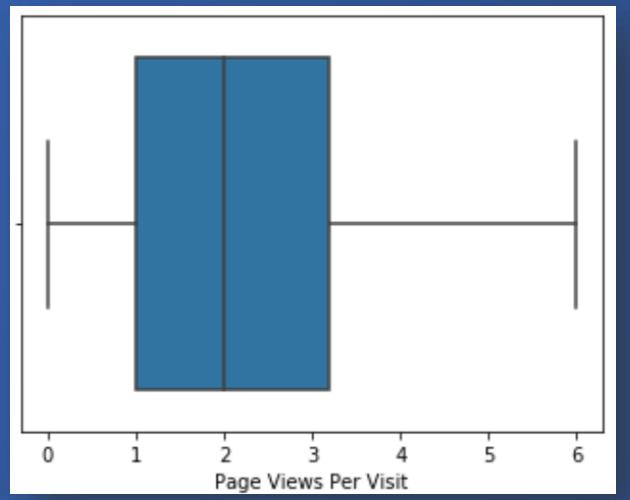
Outlier Analysis



Original Data

Outliers Removal

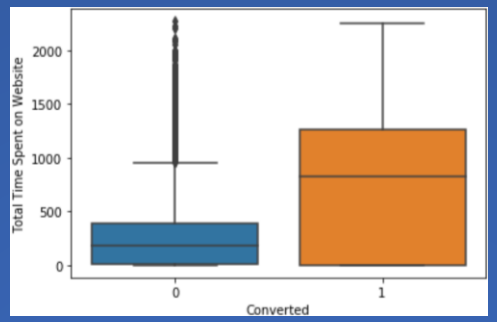


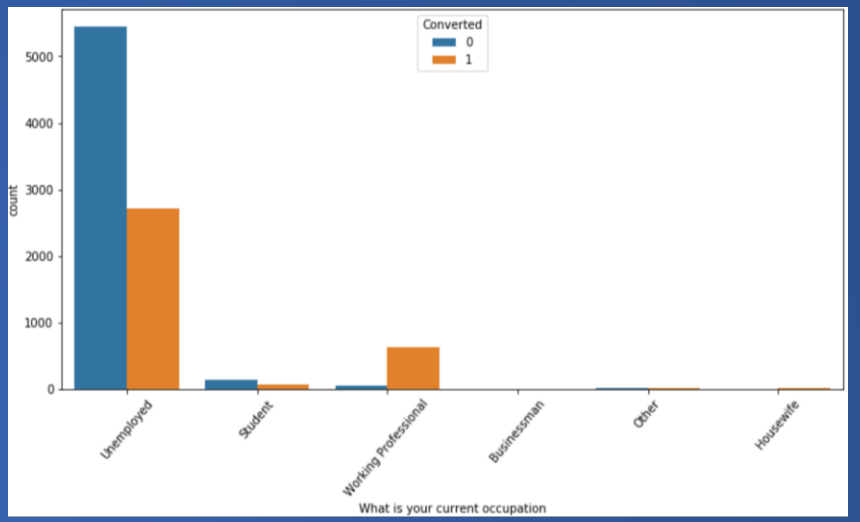


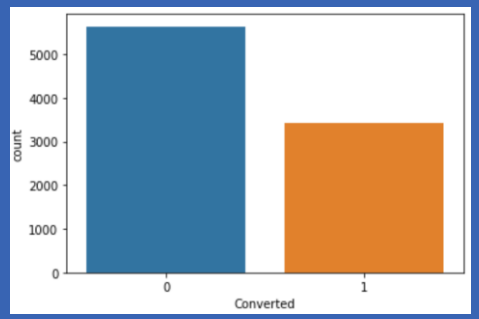
Transformed Data

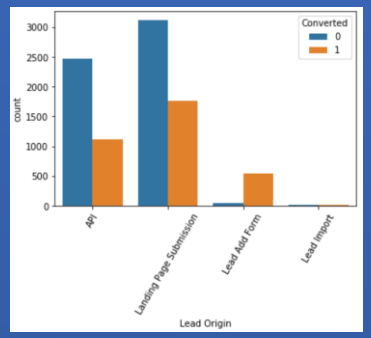
Univariate Analysis

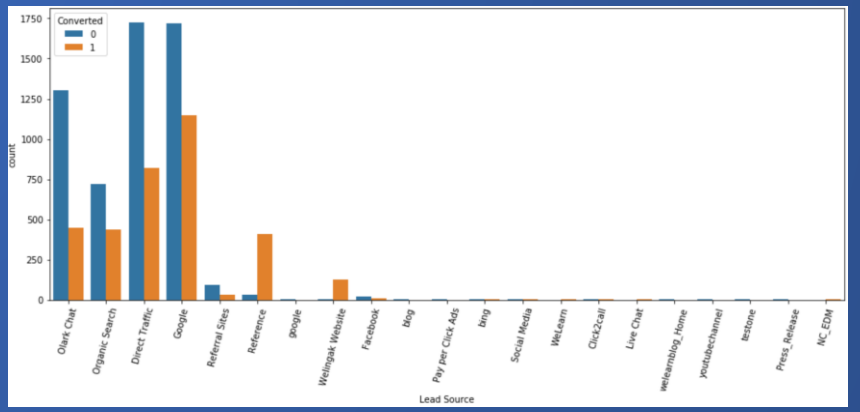
- Focus on improving lead conversion of API and Landing Page sources. Since, the Lead Add Form has the highest number of lead conversion, more leads should be generated using Lead Add Form.
- Focus on improving lead conversion of Direct Traffic, Google, Olark Chat and Organic Search. Since, the Reference and Welingak Website has the highest number of lead conversion, more leads should be generated using Reference and Welingak Website.
- Make web site more intuitive and engaging, so that leads can spend more time going through the web site contents and know about online courses.











Final LR Model Summary

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6337
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1588.8
Date:	Sun, 01 Mar 2020	**Deviance:**	3177.6
Time:	21:26:14	**Pearson chi2:**	3.08e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

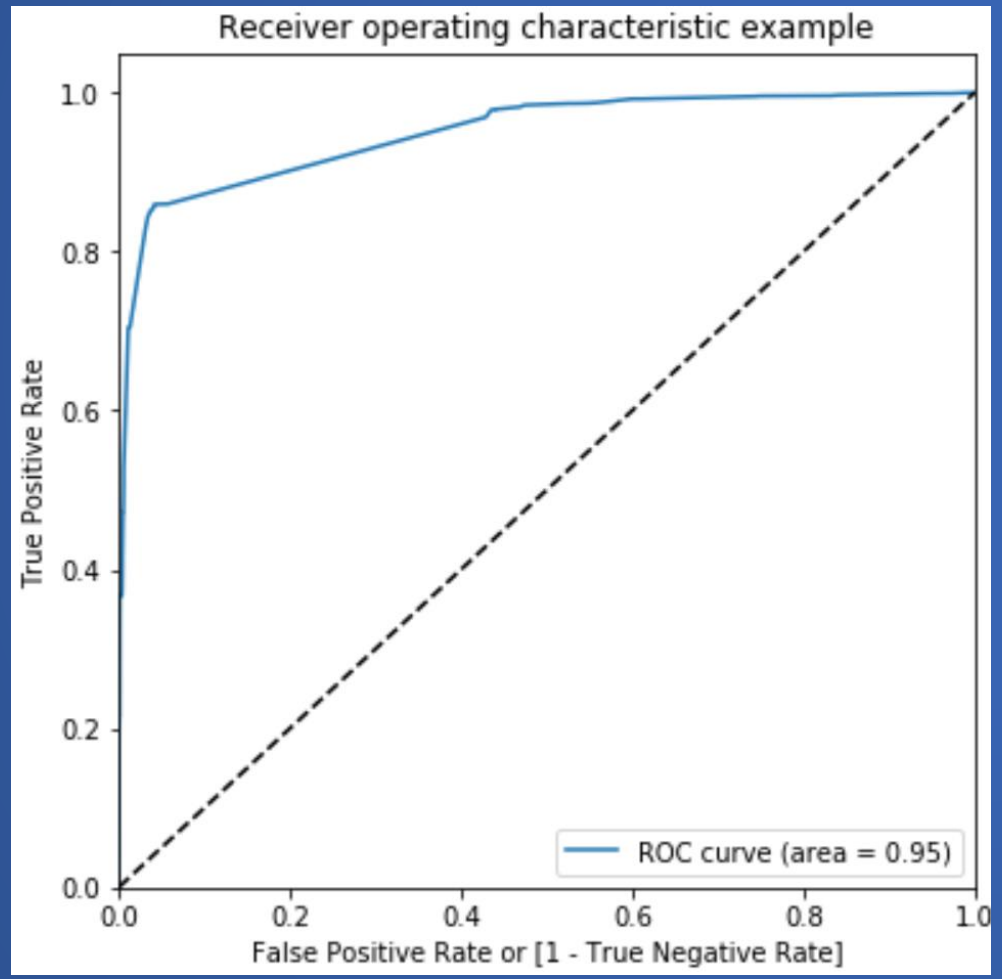
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0888	0.216	-9.654	0.000	-2.513	-1.665
Do Not Email	-1.3012	0.212	-6.134	0.000	-1.717	-0.885
Lead Origin_Lead Add Form	1.0894	0.363	3.001	0.003	0.378	1.801
Lead Source_Welingak Website	3.4138	0.818	4.173	0.000	1.810	5.017
What is your current occupation_Working Professional	1.3403	0.291	4.602	0.000	0.769	1.911
Tags_Busy	3.8040	0.330	11.532	0.000	3.157	4.450
Tags_Closed by Horizzon	7.9562	0.763	10.433	0.000	6.461	9.451
Tags_Lost to EINS	9.1785	0.754	12.177	0.000	7.701	10.656
Tags_Ringing	-1.6947	0.337	-5.036	0.000	-2.354	-1.035
Tags_Will revert after reading the email	3.9665	0.229	17.311	0.000	3.517	4.416
Tags_switched off	-2.2882	0.587	-3.900	0.000	-3.438	-1.138
Lead Quality_Not Sure	-3.3406	0.128	-26.026	0.000	-3.592	-3.089
Lead Quality_Worst	-3.7624	0.850	-4.426	0.000	-5.428	-2.096
Last Notable Activity_SMS Sent	2.7406	0.120	22.847	0.000	2.506	2.976

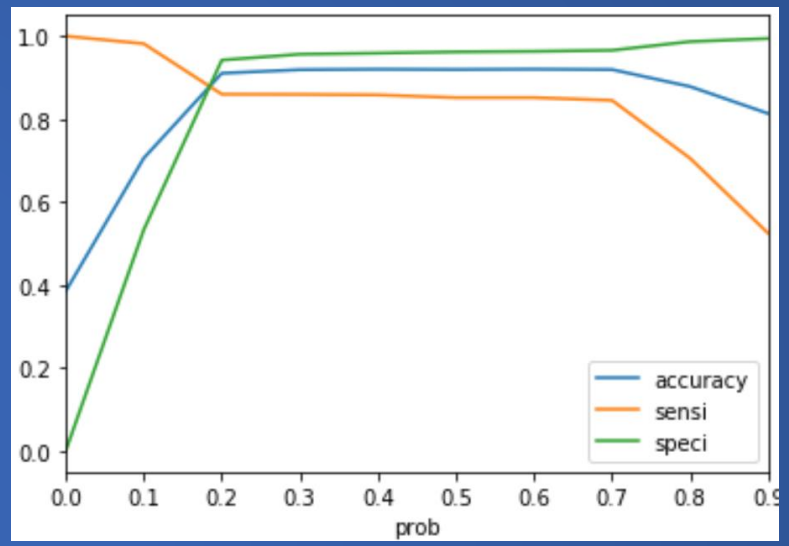
StatsModels Summary

	Features	VIF
0	const	12.69
9	Tags_Will revert after reading the email	2.48
8	Tags_Ringing	1.84
12	Lead Quality_Worst	1.67
11	Lead Quality_Not Sure	1.53
2	Lead Origin_Lead Add Form	1.52
6	Tags_Closed by Horizzon	1.39
3	Lead Source_Welingak Website	1.35
4	What is your current occupation_Working Professional	1.22
10	Tags_switched off	1.19
13	Last Notable Activity_SMS Sent	1.18
5	Tags_Busy	1.16
7	Tags_Lost to EINS	1.12
1	Do Not Email	1.02

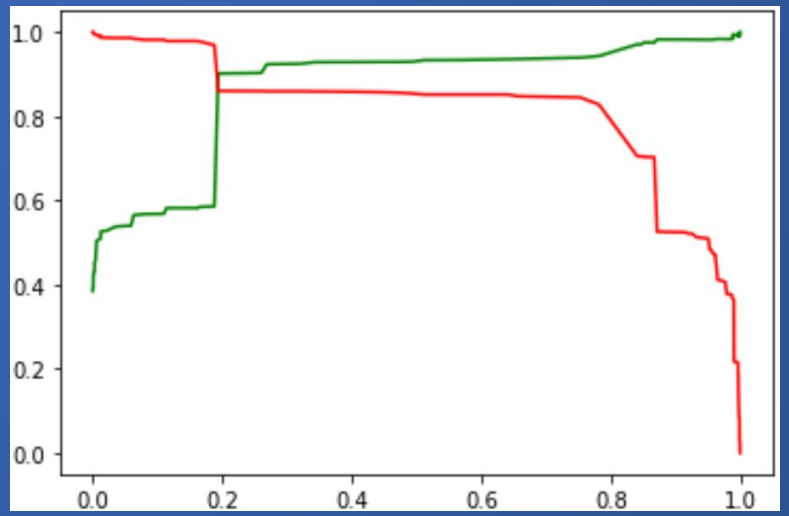
Final Features with VIF

ROC Curve, Optimal Cut-off & Precision Recall Trade-off





Optimal Probability Cut-off is 0.2



Precision and Recall Trade-off

Model Evaluation (Train Data)

Confusion Matrix

Actual	Predicted	
	No (Converted)	Yes (Converted)
No (Converted)	3679	226
Yes (Converted)	343	2103

Accuracy	Sensitivity	Specificity	+ve Predictive Value	-ve Predictive Value
0.91	0.86	0.94	0.90	0.91

False Positive Rate	Precision Score	Recall Score	F1 Score	ROC Area
0.06	0.90	0.86	0.86	0.95

Model Evaluation (Test Data)

Confusion Matrix

Actual	Predicted	
	No (Converted)	Yes (Converted)
No (Converted)	1635	99
Yes (Converted)	155	834

Accuracy	Sensitivity	Specificity	+ve Predictive Value	-ve Predictive Value
0.91	0.84	0.94	0.89	0.91

False Positive Rate	Precision Score	Recall Score	F1 Score	ROC Area
0.06	0.9	0.86	0.86	0.95

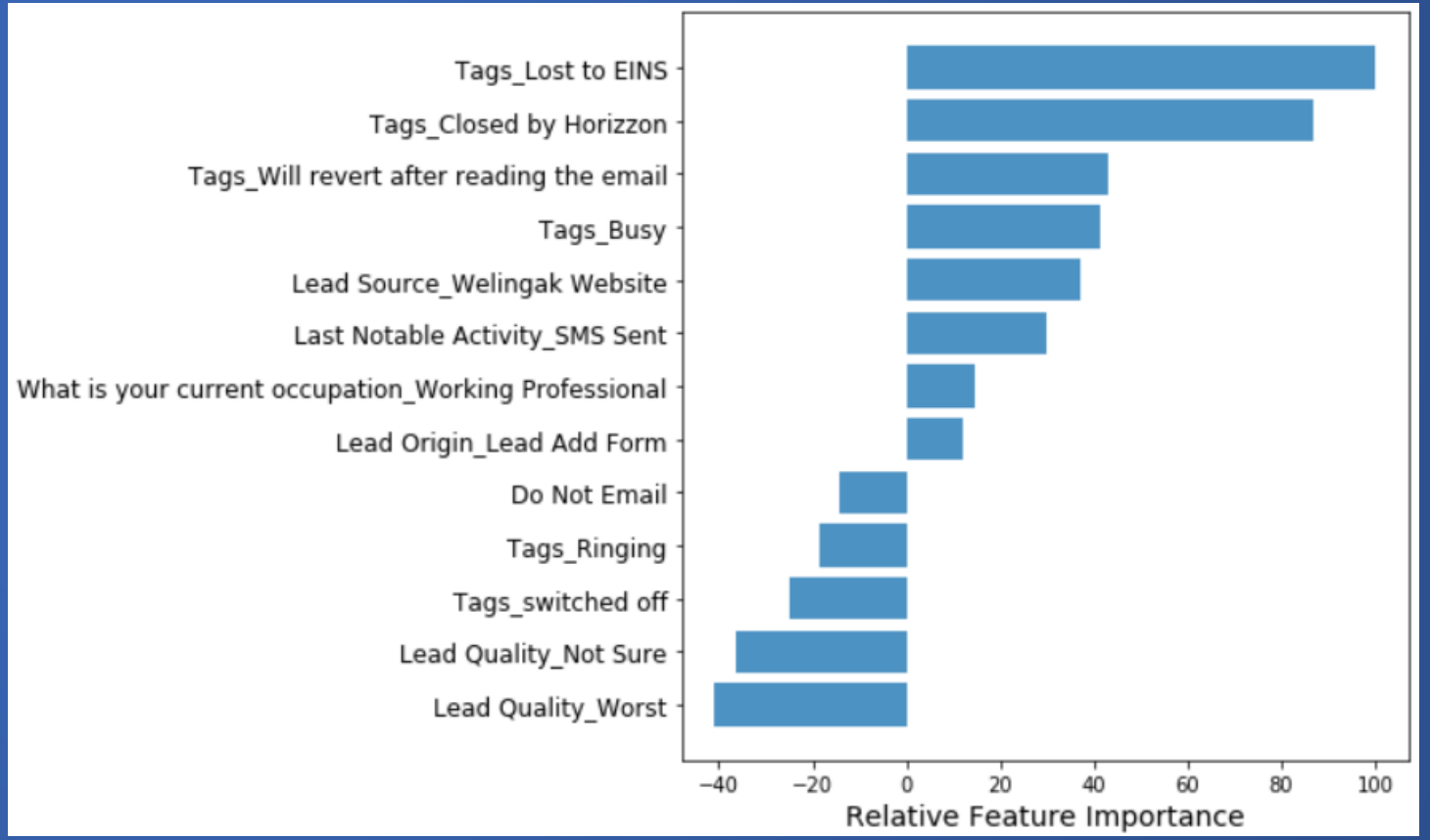
Features contribution to Lead Conversion

The lead conversion probability increases when there is increase in the below feature value.

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Tags_Will revert after reading the email
- Tags_Busy
- Lead Source_Welingak Website
- Last Notable Activity_SMS Sent
- What is your current occupation_Working Professional
- What is your current occupation_Unemployed

The lead conversion probability increases when there is decrease in the below feature value.

- Lead Quality_Worst
- Lead Quality_Not Sure
- Tags_switched off
- Tags_Ringing
- Do Not Email



Final Recommendations

Below three features contribute most to increase the probability of lead conversion.

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tags_Will revert after reading the email

Recommendations to manage seasonality in the LR Model.

- We have considered the probability threshold value of 0.2 based on statistical evaluation. However, based on the business needs, we can either increase or decrease the probability threshold value which will impact the Sensitivity and Specificity of LR model.
- High Sensitivity will ensure that almost all leads who are likely to convert are correctly predicted whereas high Specificity will ensure that leads that are on the edge case are not selected for follow-up.
- So, we should increase the Sensitivity by adjusting probability threshold value during period of 2 months when X Education hires interns and want to be more aggressive in lead conversion. On the other hand, we should increase Specificity, when company reaches its target for a quarter before the deadline and wants the sales team to focus on some new work.

Thank you!