LEAD SCORING CASE STUDY

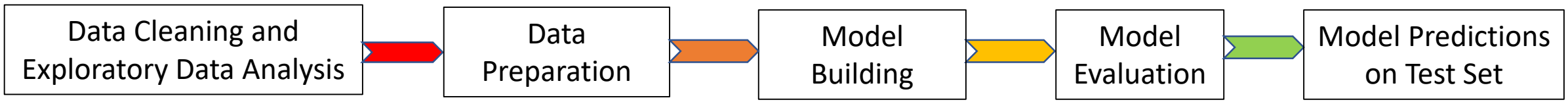
SURABHI KHARE & SRINATH GOPINATHAN

Problem Statement

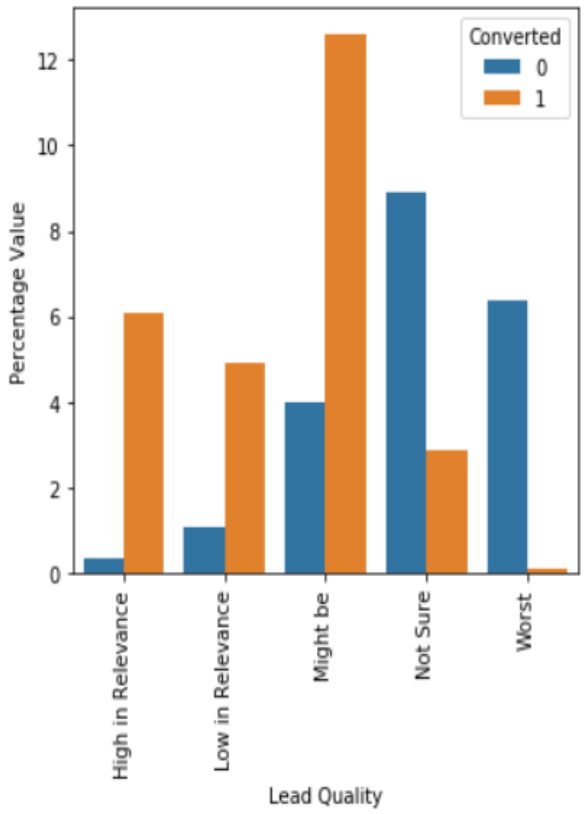
X Education sells online courses to industry professionals.

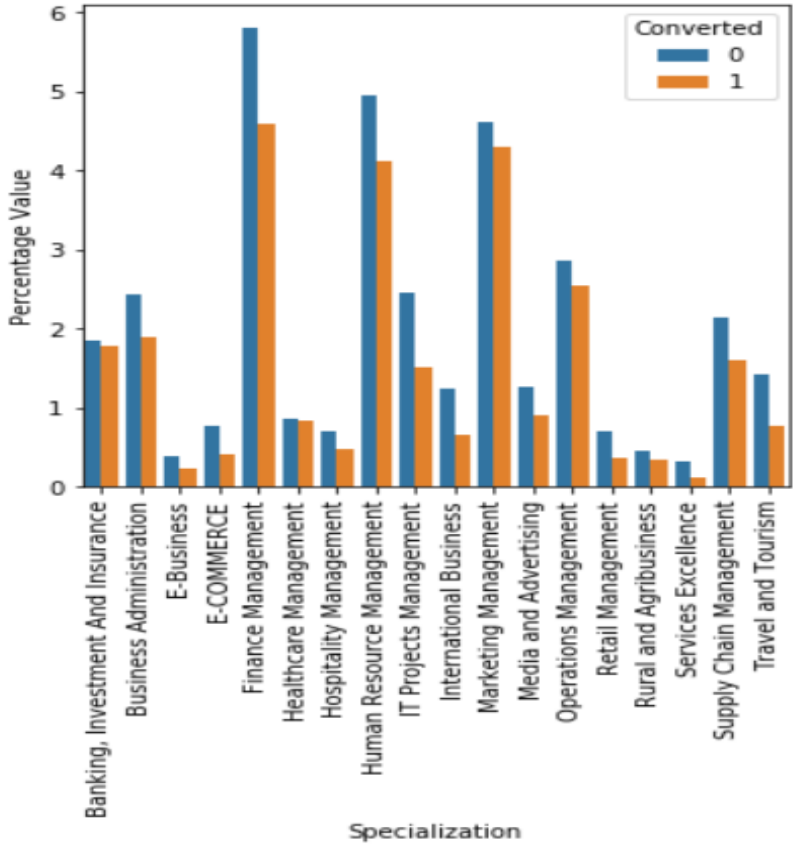
The goal is to build a logistic regression model to improve the conversion rates from 30% to 80% by targeting potential leads from the available data based on a scoring mechanism.

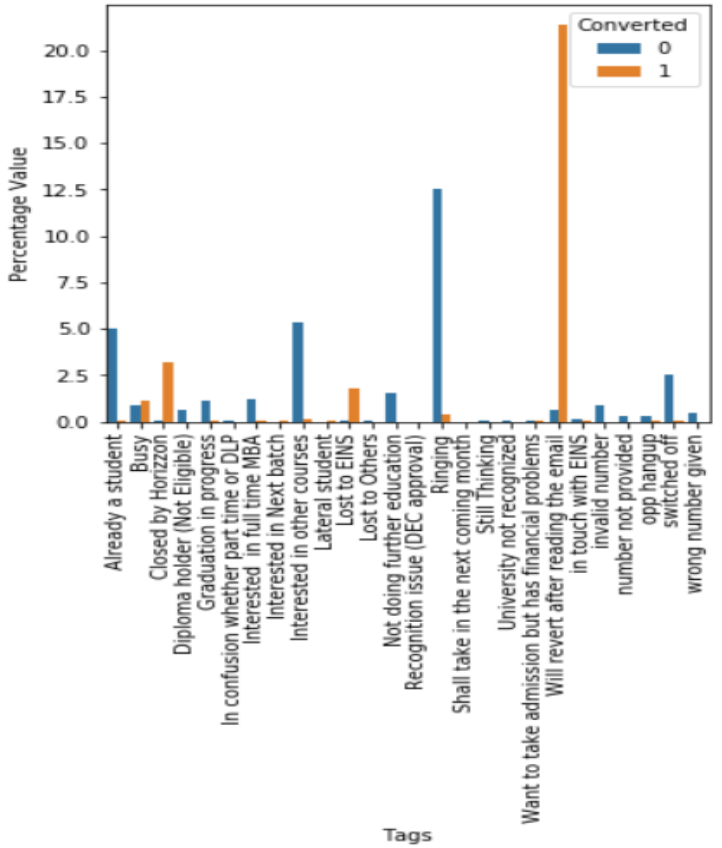
Analysis Approach



Data Cleaning and EDA

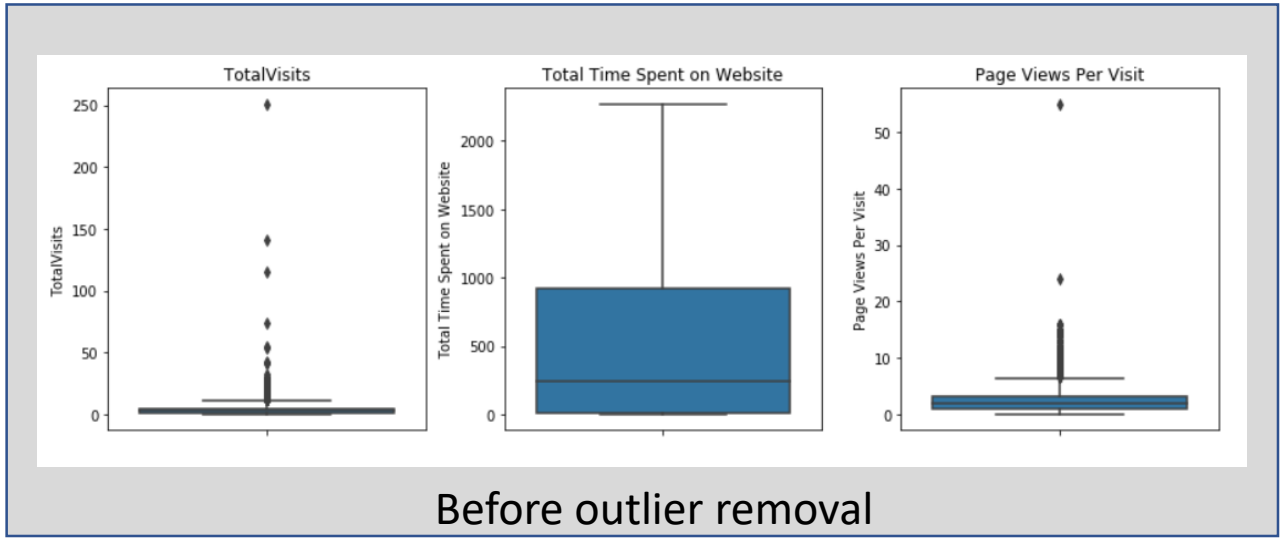




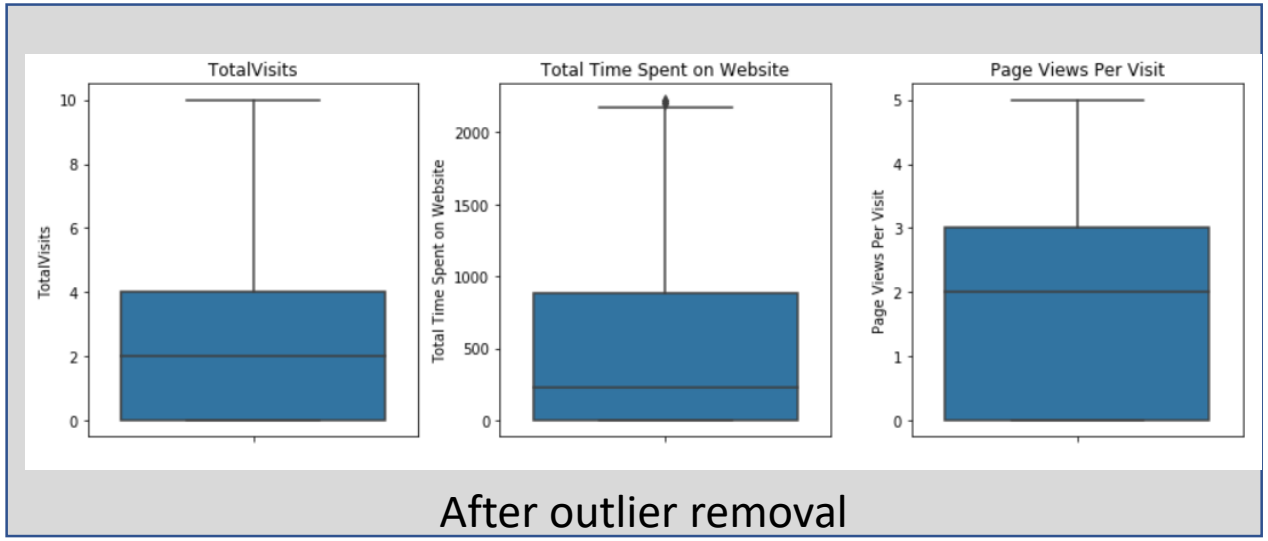


- Variables with more than 25% null values are dropped. However, variables with up to 50% missing values are considered for identification of any important variable.
- ‘Lead Quality’, ‘Specialization’ and ‘Tags’ come out to be as critical variables because of high variance. Missing values are imputed as ‘Others’.

EDA - Outlier Analysis



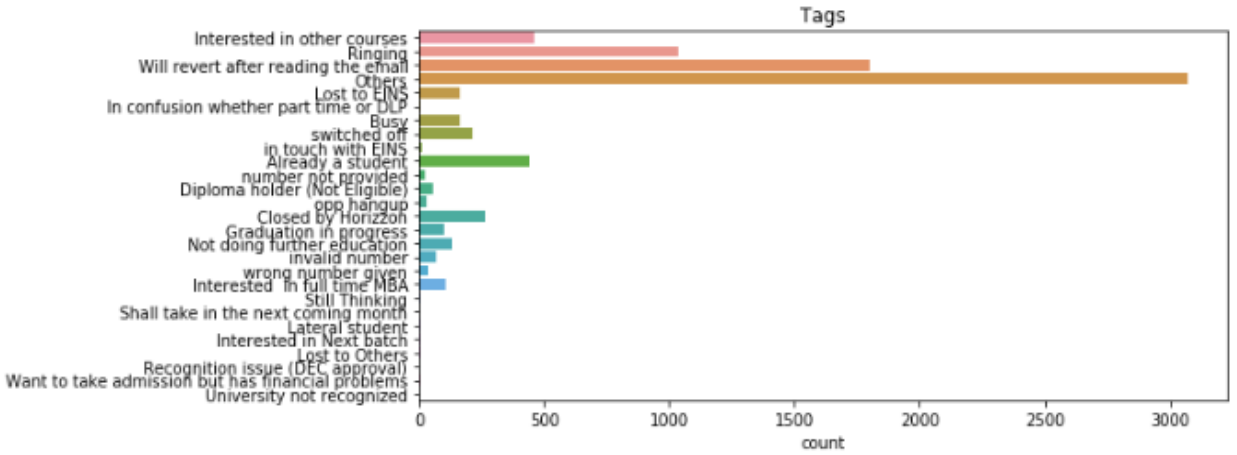
Before outlier removal

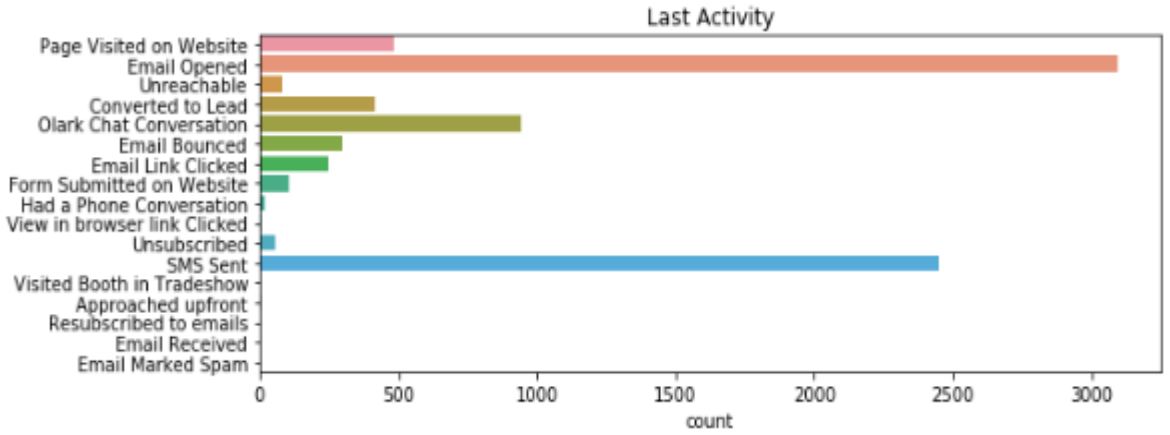


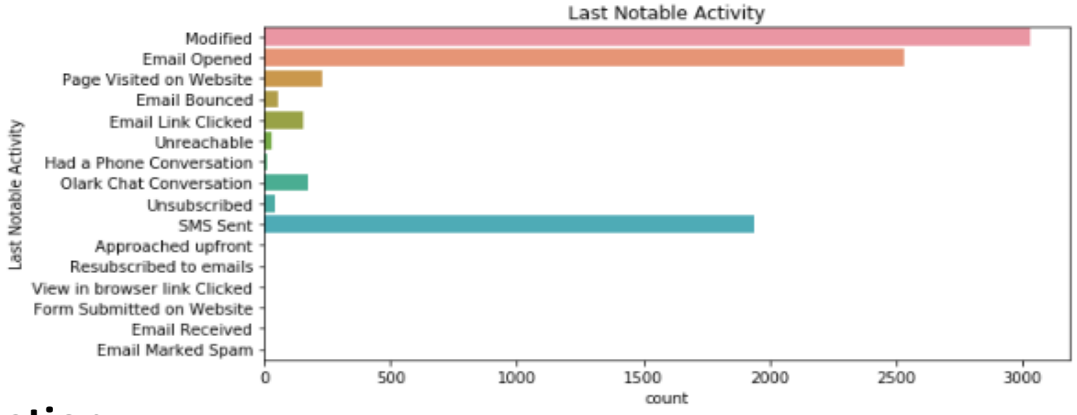
After outlier removal

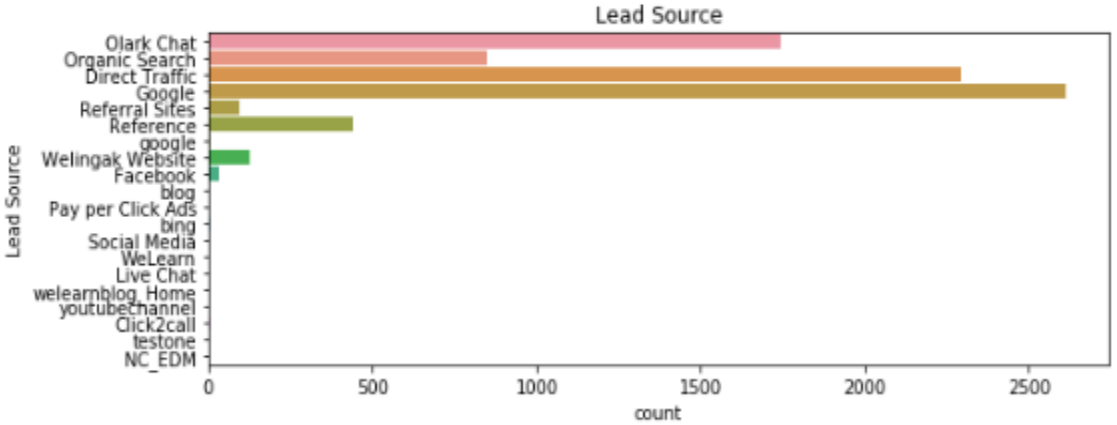
Outlier detection and removal is done using boxplots and 1.5*IQR method.

EDA - Observations and Assumptions









Observation :

- Some categorical levels with minimal counts are seen.

Assumptions:

- Categorical outliers with minimal count levels are removed.
- Variables with negligible variance are dropped.

Model Building

- Create the dummy variables, split the data into training / test sets and scale the data.
- Build the Logistic Regression Model using Recursive Feature Elimination (RFE) and statsmodels library.
- Variables deduction is recursively done using p-values and VIF values.

Dep. Variable:	Converted	**No. Observations:**	5595
Model:	GLM	**Df Residuals:**	5578
Model Family:	Binomial	**Df Model:**	16
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1111.8
Date:	Sun, 17 Nov 2019	**Deviance:**	2223.6
Time:	14:19:13	**Pearson chi2:**	7.84e+03
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1528	0.095	-22.756	0.000	-2.338	-1.967
Do Not Email	-1.2671	0.254	-4.985	0.000	-1.765	-0.769
Total Time Spent on Website	1.0195	0.064	15.977	0.000	0.894	1.145
Lead Origin_Lead Add Form	2.0549	0.446	4.612	0.000	1.182	2.928
Lead Source_Olark Chat	1.3914	0.157	8.878	0.000	1.084	1.699
Lead Source_Welingak Website	3.5901	0.857	4.191	0.000	1.911	5.269
Last Activity_Olark Chat Conversation	-1.5829	0.246	-6.441	0.000	-2.065	-1.101
Last Notable Activity_SMS Sent	2.6318	0.136	19.355	0.000	2.365	2.898
Tags_Already a student	-1.7325	0.618	-2.802	0.005	-2.944	-0.521
Tags_Closed by Horizzon	5.3600	0.759	7.065	0.000	3.873	6.847
Tags_Interested in other courses	-1.6901	0.368	-4.594	0.000	-2.411	-0.969
Tags_Lost to EINS	7.3047	1.170	6.242	0.000	5.011	9.599
Tags_Ringing	-3.4349	0.252	-13.617	0.000	-3.929	-2.941
Tags_Will revert after reading the email	4.4285	0.199	22.264	0.000	4.039	4.818
Tags_switched off	-3.7408	0.625	-5.989	0.000	-4.965	-2.517
Lead Quality_High in Relevance	1.0031	0.466	2.151	0.031	0.089	1.917
Lead Quality_Worst	-3.2060	0.921	-3.480	0.001	-5.012	-1.400

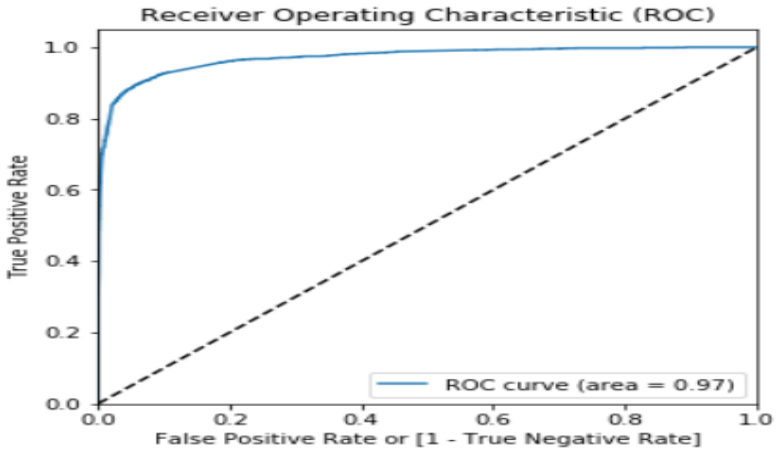
	Features	VIF
2	Lead Origin_Lead Add Form	1.92
12	Tags_Will revert after reading the email	1.73
3	Lead Source_Olark Chat	1.69
15	Lead Quality_Worst	1.63
7	Tags_Already a student	1.58
6	Last Notable Activity_SMS Sent	1.51
1	Total Time Spent on Website	1.46
5	Last Activity_Olark Chat Conversation	1.41
4	Lead Source_Welingak Website	1.37
14	Lead Quality_High in Relevance	1.36
8	Tags_Closed by Horizzon	1.30
9	Tags_Interested in other courses	1.09
11	Tags_Ringing	1.09
0	Do Not Email	1.05
10	Tags_Lost to EINS	1.03
13	Tags_switched off	1.03

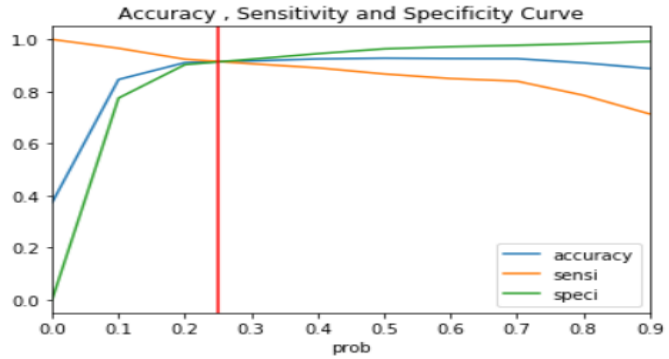
Note:
The final model doesn't have any high p-value or high VIF.
The p-values are less than 0.05 and the VIF values are less than 5.
Thus, this model is good to proceed with.

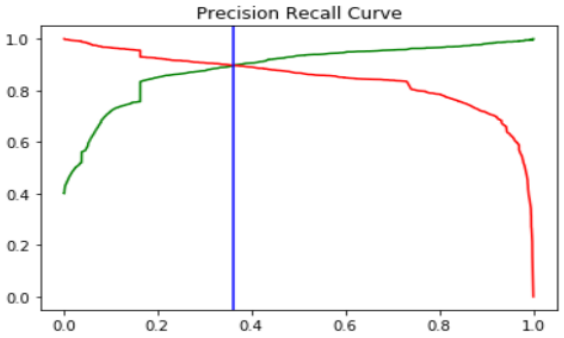
Model Evaluation

- Model evaluation is done using the ROC curve.
- Cut-off values are found using accuracy-sensitivity-specificity (0.24) and precision recall curve (0.36).
- The cut off obtained in Precision Recall Curve is chosen to optimize the precision of the model.
- The metrics obtained are similar in both the training and the test sets.

	Accuracy	Sensitivity	Specificity	Precision
Training Set	0.922967	0.89746	0.938141	0.896172
Test Set	0.927887	0.905274	0.942177	0.908207







BUSINESS RECOMMENDATION

TOP 3 VARIABLES THAT DETERMINE CONVERSION

TAGS

LAST NOTABLE ACTIVITY

LEAD SOURCE

ADDITIONAL AREAS FOR EXPLORATION

IMPROVE DIGITAL MARKETING STRATEGIES IN POPULAR SOCIAL MEDIA.

IMPROVE THE QUALITY OF OLARK CHAT CONVERSATION