PRESENTATION AND RECOMMENDATION

Gordhan,
Shrestha and Kirti

AGENDA

Problem Statement

Business goals

Strategy

EDA

Model Building

Model Evaluation

Problem statement

X- Education is an organization which provides online courses for industry professionals. The company marks its courses on various websites like google.

X Education wants to select most promising leads that can be converted into paying customers.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Although the implementation process of the leads are not efficient in helping conversions.

BUSINESS GOALS

The company requires a model to be built for selecting most promising leads.
Lead score to be given to each leads such that it indicates how promising the could be. The higher the lead score the more promising the lead to get converted.

STRATEGY

- Import Data
- Clean and prepare the acquired data for further analysis.
- Performing Exploratory Data Analysis.
- Scaling features.

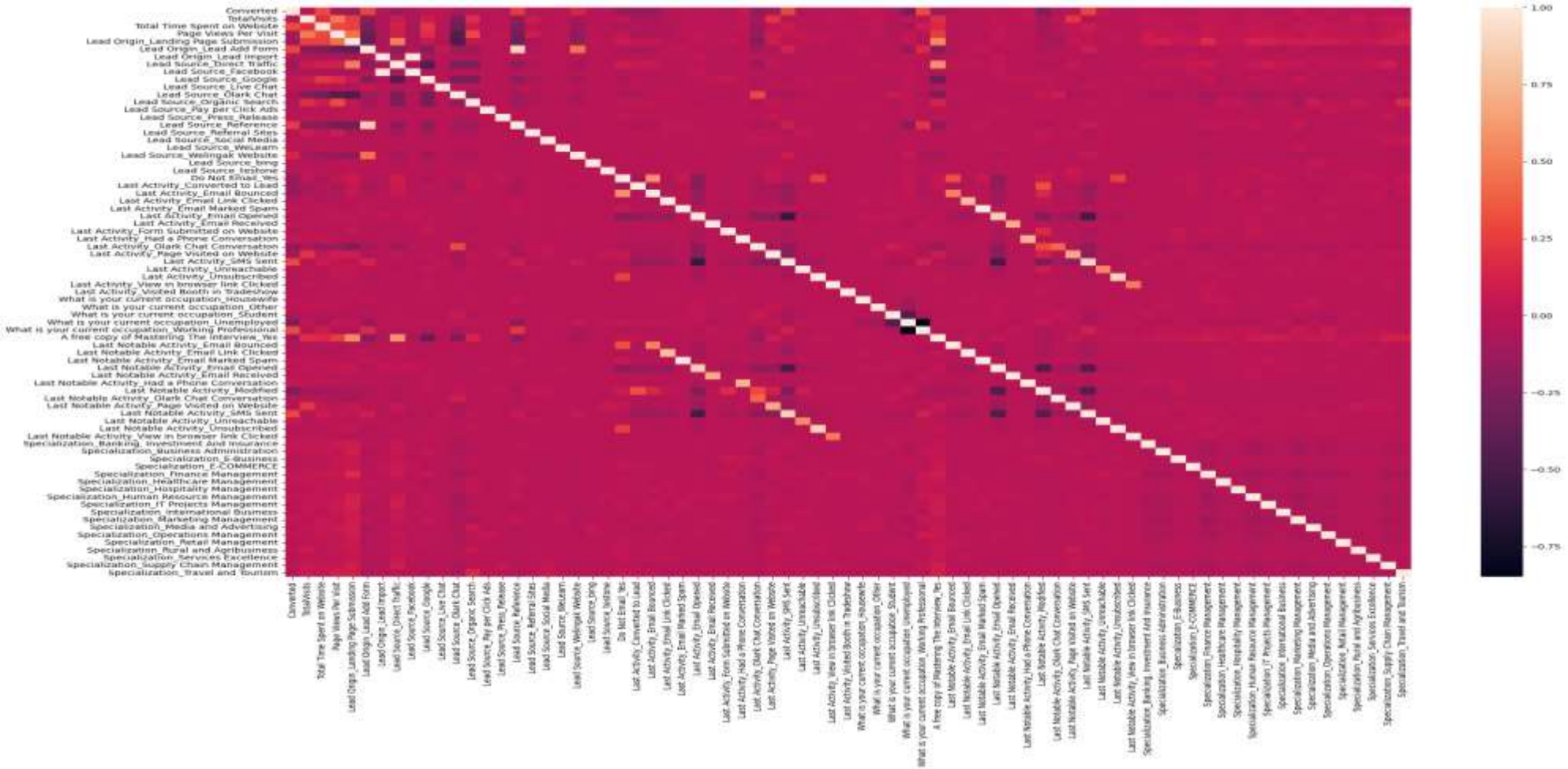
- Prepare the data for model building

- Build a logistic regression model
- Assign a lead score for each lead.

- Test the model on train set.
- Evaluate model by different measure and metrics.
- Test the model on test set.
- Measure the accuracy of the model and other metrics for evaluation.

EXPLORATORY DATA ANLYSIS

Heatmap
Showing corelation
Among various varia
bles in the dataset.



LOGISTIC REGRESSION MODEL ON X_TRAIN AFTER ADDING A CONSTANT

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4445
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2072.8
Date:	Tue, 03 Jan 2023	**Deviance:**	4145.5
Time:	01:56:25	**Pearson chi2:**	4.84e+03
No. Iterations:	22	**Pseudo R-squ. (CS):**	0.3660
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175

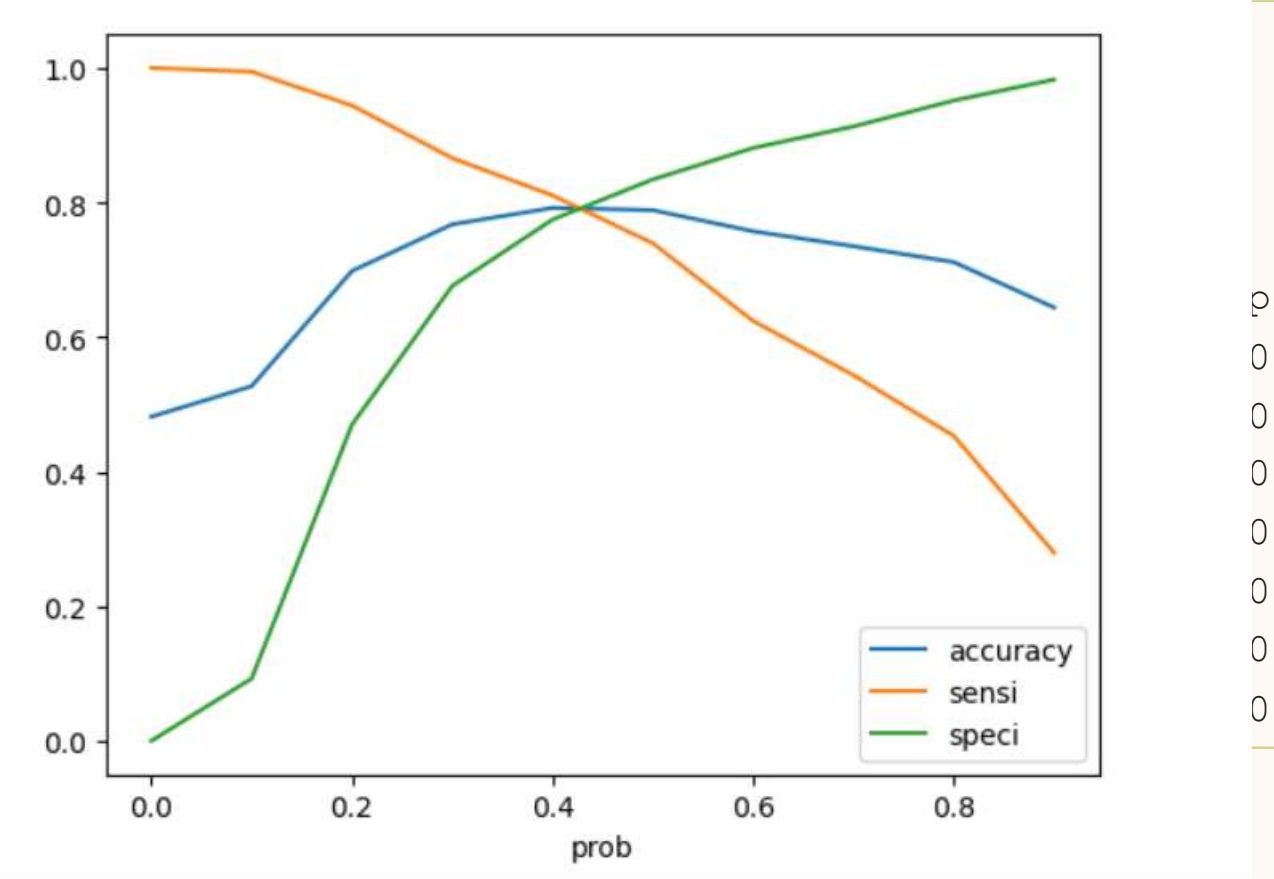
MODEL BUILDING

- Splitting into train and test set.
- Scale variables in train test.
- Build the first model
- Use RFE to eliminate less relevant variables.

- Build the next model.
- Eliminate variables based on high P-values.
- Check VIF values for all the existing columns.
- Predict using train set.

- Evaluate accuracy and other metric.
- Predict using test set.
- Precision and recall analysis on test predictions.

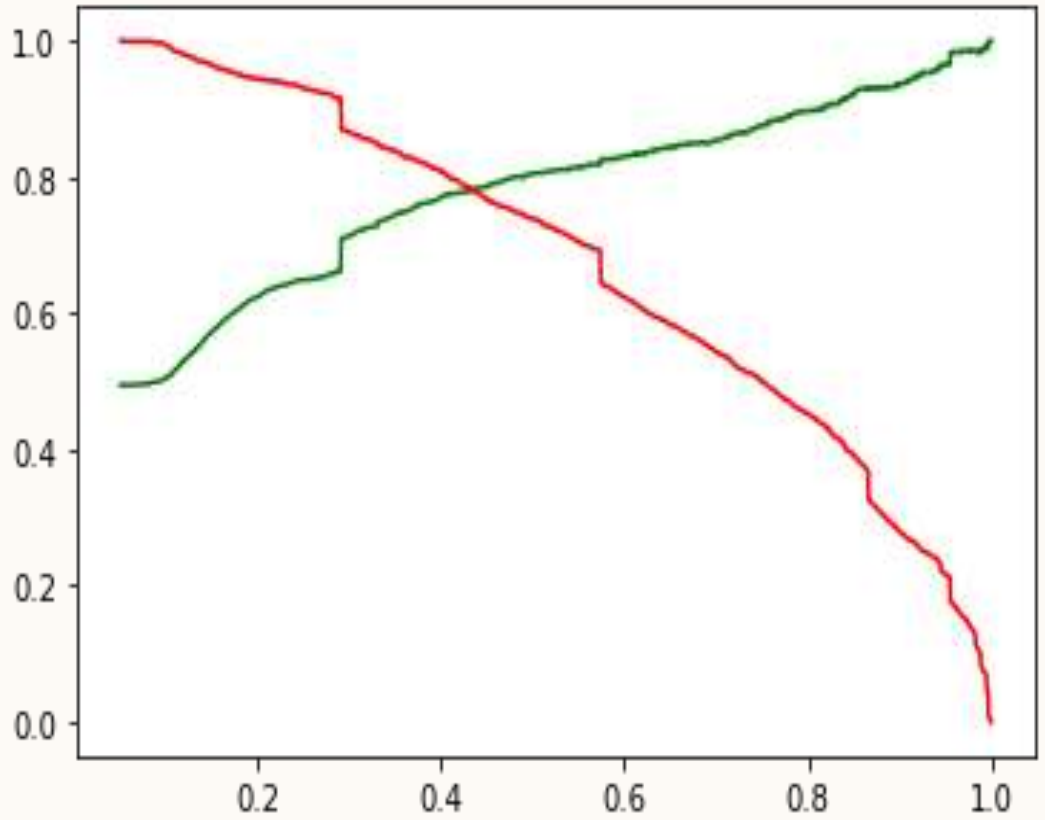
MODEL EVALUATION(TRAIN)



ACCURACY, SENSITIVITY AND SPECIFICITY

prob	accuracy	sensi	speci
0.0	0.481731	1.000000	0.000000
0.1	0.527012	0.994416	0.092561
0.2	0.698274	0.944160	0.469723
0.3	0.767541	0.865984	0.676038
0.4	0.791975	0.810610	0.774654
0.5	0.788612	0.739414	0.834343
0.6	0.757229	0.624011	0.881055

MODEL EVALUATION(TRAIN)



PRECISION AND RECALL TRADEOFF

Conversion_probability
0.300117
0.142002
0.127629
0.291558
0.954795

SUMMARY

EDA :

People spending higher than average time are promising leads, so targeting them should be a better choice.

Marketing management and human resource management have high conversion rates.

An alert message or information has high conversion rate.

Logistic Regression Model:

Model shows high close to 78% accuracy.

The threshold has been selected from accuracy, sensitivity, specificity measures and precision and recall curves.

Model shows 78% sensitivity and 79% specificity.

THANK YOU