LEAD SCORE MODEL

- BY NIKITA RAWAT

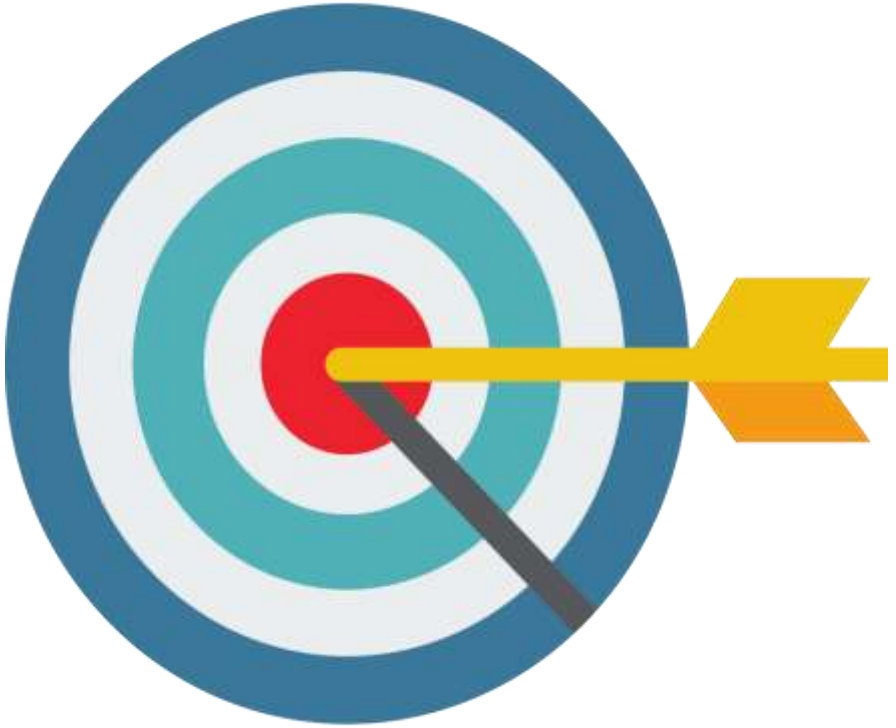
Problem Statement

- X Education company gets a lot of leads, but its lead-to-sale conversion rate is only 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The CEO, in particular, has given a ballpark estimate of the target lead conversion rate as being around 80%.

Aim Of The Project

- Build a model, 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.

- To identify the most promising leads, i.e., the leads that are most likely to convert into paying customers, so that the sales team can focus more on communicating with the potential leads rather than making calls to everyone and achieve better lead conversion rate.



Dataset Used for Model Building

- The dataset consists of leads converted from the past records of the company that contained approximately 9000 data points along with 37 variables.
- The target variable, in this case, is the column ‘Converted’, which tells whether a past lead was converted or not, where 1 means it was converted and 0 means it wasn’t converted.

Model Building Approach

- ❑ Importing the dataset.
- ❑ Exploring and understanding the data.
- ❑ Preparing the data for modelling.
- ❑ Selecting the logistic model for the target variable.
- ❑ Splitting the data to train and test dataset.
- ❑ Training the model on train dataset.
- ❑ Plotting the ROC curve to check the fit.
- ❑ Finding optimal cut-off point for converting conversion probability to converted.
- ❑ Testing the fit with test dataset.

Data Preparation

- Dropped variables where null value is greater than 45% .
- 'Select' values (assumed to have been result of not selecting the value by the customer) are taken as Null and filled with median.
- Boolean variable converted to Binary variables.
- Categorical Columns are converting into Dummy Variables Using One Hot Encoding.
- Total Rows and Columns for Analysis – 9204 & 64 ,respectively.

Model Building

- Splitting the Data into Training and Testing Sets (80:20 ratio), using train_test_split function from Scikit-learn library.
- Use RFE for Feature Selection, Running RFE with the 15 important variables affecting the target variable 'Converted' as output.
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- ROC curve is built which shows the trade-off between Sensitivity and Specificity and used to find the cut-off value.

Final Model Output

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.0099	0.190	-21.055	0.000	-4.383	-3.637
Total Time Spent on Website	1.4223	0.081	17.579	0.000	1.264	1.581
Lead Origin_Lead Add Form	2.7445	0.241	11.379	0.000	2.272	3.217
Tags_Busy	2.7691	0.275	10.054	0.000	2.229	3.309
Tags_Closed by Horizzon	8.3971	0.742	11.318	0.000	6.943	9.851
Tags_Lost to EINS	7.5527	0.550	13.742	0.000	6.476	8.630
Tags_Not Tagged	2.2689	0.187	12.150	0.000	1.903	2.635
Tags_Ringing	-1.4139	0.275	-5.133	0.000	-1.954	-0.874
Tags_Will revert after reading the email	6.3528	0.230	27.596	0.000	5.902	6.804
Tags_switched off	-1.6735	0.548	-3.052	0.002	-2.748	-0.599
Last Notable Activity_Modified	-1.0032	0.116	-8.684	0.000	-1.230	-0.777
Last Notable Activity_SMS Sent	2.2906	0.124	18.485	0.000	2.048	2.533

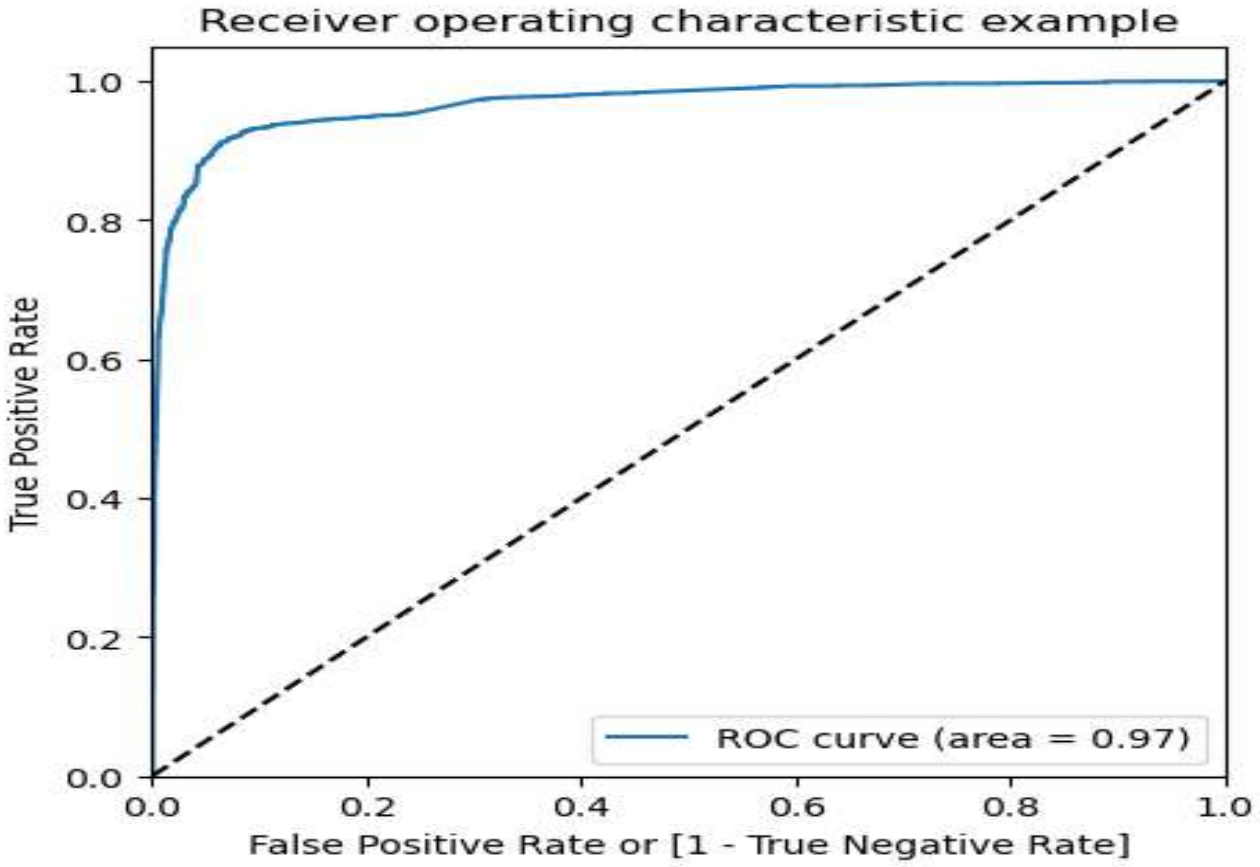
All p-values are below the 0.05 threshold, indicating statistical significance

VIF of the Final Model

	Features	VIF
7	Tags_Will revert after reading the email	1.75
10	Last Notable Activity_SMS Sent	1.67
9	Last Notable Activity_Modified	1.51
5	Tags_Not Tagged	1.44
0	Total Time Spent on Website	1.41
1	Lead Origin_Lead Add Form	1.41
3	Tags_Closed by Horizzon	1.32
6	Tags_Ringing	1.15
4	Tags_Lost to EINS	1.07
2	Tags_Busy	1.05
8	Tags_switched off	1.04

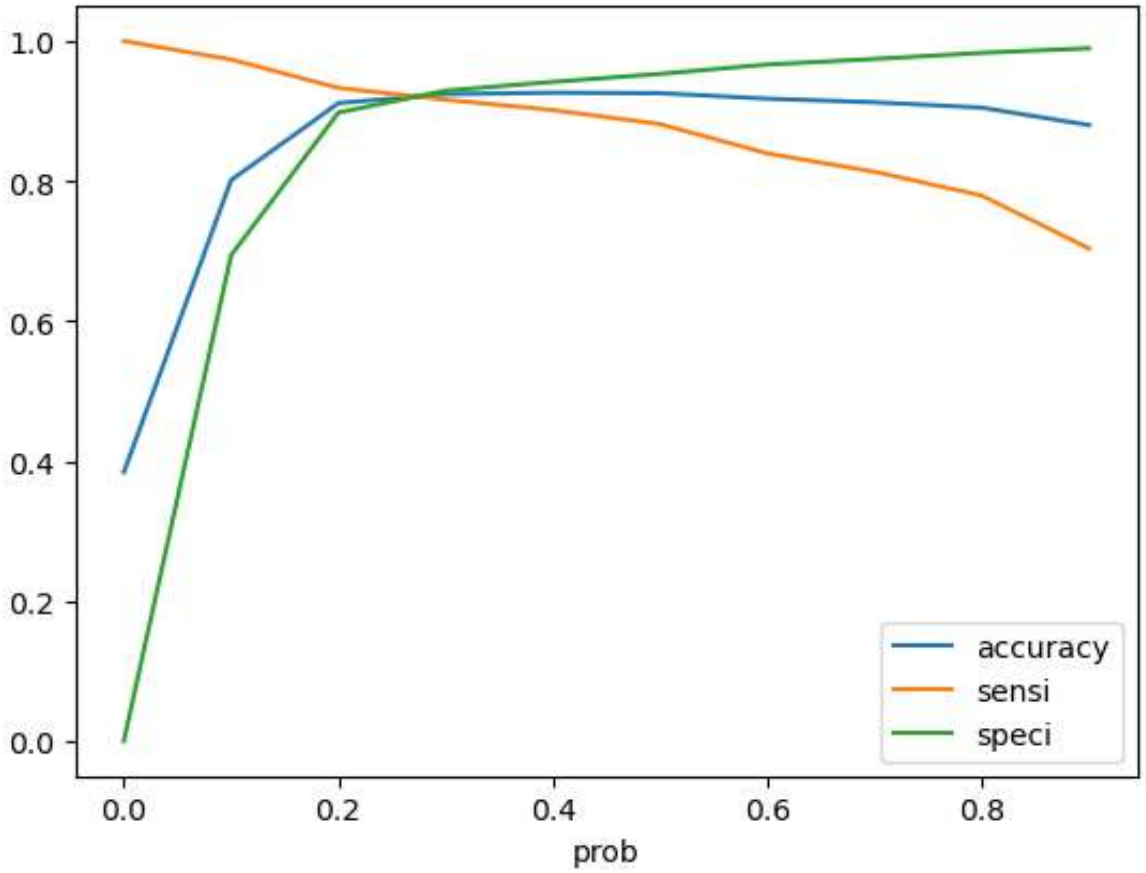
All VIF values are below 5, indicating no significant multicollinearity

ROC Curve



The ROC Curve rises sharply and stays close to the True Positive Rate (TPR) axis, indicating good model performance

Optimal Cut-off Point Selection



The optimal cutoff point is identified at a probability threshold of 0.3

Final Analysis at Cut-off Probability of 0.3

- Model Prediction On Train Dataset-
  - ✓Accuracy Score of Train Model is 92%
  - ✓Sensitivity of Train Model is 91%
  - ✓Specificity of Train Model is 92%
  - ✓F1 Score of Test Model is 90%

- Model Prediction On Test Dataset-
  - ✓Accuracy Score of Test Model is 92%
  - ✓Sensitivity of Test Model is 92%
  - ✓Specificity of Test Model is 93%
  - ✓F1 Score of Test Model is 91%

Conclusion

- Based on the original variable names, the most important variables influencing the conversion rate are: 'Tags', 'Lead Origin' and 'Last Notable Activity.

- Based on the model output, the top three variables impacting the conversion rate are - Tags stating that they were ‘Closed by Horizzon ’, ‘Lost to EINS ’ and ‘Will revert after reading the email ’.

- It is recommended that the company prioritize customers who are tagged as ‘Closed by Horisson’, ‘Lost to EINS’, or ‘Will revert after reading the email’, particularly those whose Lead Origin is the ‘Lead Add Form’ and whose Last Notable Activity is ‘SMS Sent’.

THANK YOU