Summary Report:

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%.

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach:

From the description we conclude that the above problem is the classification problem, hence we choose logistic regression to calculate the Lead rate.

1. Data Reading and Understanding:
   In this step we observe following things from data

   - Number of rows and columns
   - Data type of each columns
   - Checking how the data is spread.
   - Checking for duplicates, if any.

2. Data Cleaning:
Check for discrepancies in the dataset

- Checking for null values and imputing them with appropriate methods
    - Mode imputation for categorical columns.
    - Mean imputation for numerical columns, if there is no skewness in data
    - Median imputation for numerical columns, if there is skewness in the data.
- Checking for name correction.

3. Data Visualization/ Preparation and outliers Treatment:

- We performed univariate analysis on categorical column to see which columns makes more sense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- We have used IQR method to treat the outliers in the data set.
- In this step we also plotted the correlation matrix to identify the columns which are correlated.

4. Feature Scaling

At this stage our data was vary clean and no outliers. And also after train and test split as 70% and 30%. We convert all the categorical columns to numerical.

- Columns which have only two levels ‘yes’ and ‘no’ were converted to numerical using binary mapping
- Columns which have more than two levels were converted to dummies.

Now the data contained only numerical columns and dummy variables. Before proceeding for model building, we have rescaled all numerical columns by using standard Scaler method.

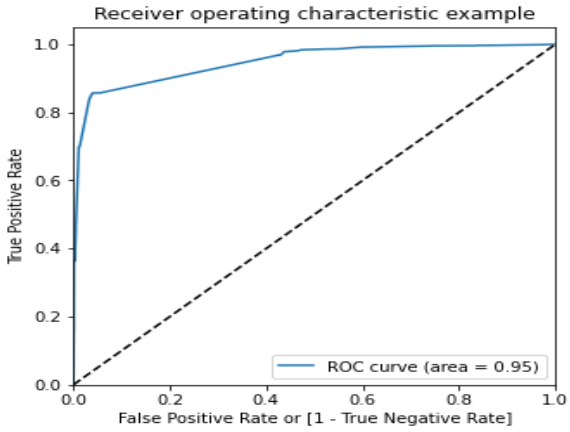
5. Model Building

- We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.
- In this step we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and vif values to be under 5. Variance inflation factor(vif) is used to treat the multicollinearity.
- We also calculated the metrics sensitivity, specificity, precision, recall and accuracy. We also plotted roc curve to find the area under the curve

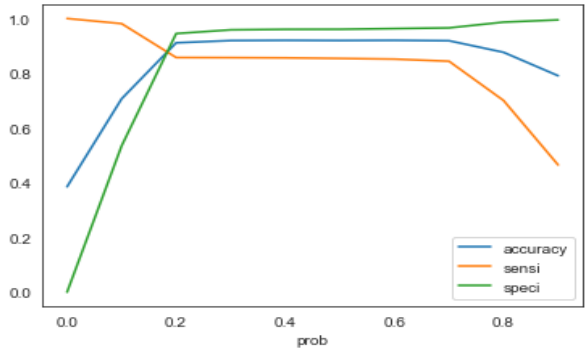
Generalized Linear Model Regression Results

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, 17 Oct 2022	Deviance:	3202.0
Time:	12:09:05	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
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
Do Not Email_Yes	-1.2835	0.212	-6.062	0.000	-1.698	-0.868
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



6. Finding optimal Threshold:



7. Prediction on Test set:

- After finalizing the optimum cutoff and calculating the metrics on train set, we predicted the data on test data set. Below are the observation

Dataset	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive Value	AUC
Train	0.9111	0.8573	0.9449	0.0550	0.9070	0.9135	0.9488
Test	0.9078	0.8412	0.9457	0.0542	0.8984	0.9126	0.9388

8. Assigning Lead Scores:
   - Using predicted probabilities to calculate Lead Scores **: Lead Score = Probability * 100**

9. Conclusion:
   - Feature having positive and negative impact on conversion probability after converting to Lead score