Summary Report for Logistic Regression Model for Predicting Lead Conversions for X Education

Problem Statement: Here we are asked to build a logistic regression model to find the leads which are most likely to join the courses and classify them as hot leads so that the company can do targeted marketing on these leads. The aim is to increase the conversion rate from 30% to 80%.

First, we loaded the data file and examined the contents peripherally.

Then we proceeded to do the following steps:

- **Data Cleaning and Preparation**
  - Replaced 'Select' with null values.
  - Addressed null values and removed columns with greater than 40% null values.
  - Eliminated columns containing only one unique value.
  - Replaced missing values with 'not declared' in a few remaining columns.
  - Removed entries with null values because it was present in a very small percentage.
  - Binned columns in an appropriate manner.
  - Segmented the columns into categorical and continuous features.
- **Exploratory Data Analysis**
  - Did univariate analysis of categorical and continuous variables and drew inferences from the analysis.
  - Did outlier analysis and treatment of the numerical columns.
  - Categorical variables were analysed with respect to the target variable 'Converted' and inferences were drawn.
  - Numerical variables were analysed for correlation using pair plot and heatmap.
- **Model Creation**
  - Dummy variables were created for the categorical variables and the original columns were removed.
  - The dataset was split into train and test datasets in 70:30 Ratio.
  - Scaled the train dataset.
  - Insignificant variables were removed using Recursive Feature Elimination.
  - We built different models using statsmodels.api and assessed the significance of the features using p-values and multicollinearity using vif values.
  - Model 5 was finalized for prediction and we identified the features which were significant in the model prediction.
- **Model Prediction and Evaluation**
  - The model was evaluated on the training dataset and test dataset using metrics like accuracy, sensitivity, specificity from the confusion matrix for an arbitrary threshold value of 0.5
  - We plotted the ROC curve to check the effectiveness of the model and found the curve tending towards the upper-left and had an area under the curve of 0.96.
  - We obtained the optimum threshold value using the accuracy, sensitivity and specificity confluence for different threshold values which was 0.26.
  - Then we evaluated the model on the test dataset and got very good results.
  - Then we used the precision-recall tradeoff to determine the optimal threshold value and came up with a value of 0.35
  - We evaluated this on the test dataset and got very good results.

- **Lead Score**
  - We created the lead scores for the different entries and evaluated the 'hot leads' predictions using a threshold value of 35.
- **Results**
  - We were able to meet the overall required accuracy of 80% as mandated by the company and were able to identify the 'hot leads' with great precision.

Results on the Test Dataset with a Lead Score threshold of 35

Accuracy - 91.8%

Sensitivity/Recall - 88.7%

Specificity - 93.7%

Precision - 89.8%

Negative Predictive Value - 93.1%

- **Finalized Model**

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6338
Model:	GLM	Df Residuals:	6326
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1372.3
Date:	Sun, 16 Apr 2023	Deviance:	2744.6
Time:	12:09:27	Pearson chi2:	8.14e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.5905
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5771	0.097	-26.626	0.000	-2.767	-2.387
Total Time Spent on Website	3.4522	0.205	16.812	0.000	3.050	3.855
Lead Origin_lead add form	1.6819	0.392	4.289	0.000	0.913	2.450
Lead Source_welingak website	4.3478	1.094	3.974	0.000	2.203	6.492
Do Not Email_yes	-1.2829	0.224	-5.738	0.000	-1.721	-0.845
What matters most to you in choosing a course_not declared	-0.7321	0.113	-6.463	0.000	-0.954	-0.510
Tags_closed by horizzon	6.6048	1.012	6.528	0.000	4.622	8.588
Tags_lost to eins	5.9444	0.727	8.175	0.000	4.519	7.370
Tags_ringing	-3.6886	0.253	-14.566	0.000	-4.185	-3.192
Tags_switched off	-4.0762	0.607	-6.716	0.000	-5.266	-2.887
Tags_will revert after reading the email	4.3696	0.178	24.520	0.000	4.020	4.719
Last Notable Activity_sms sent	2.8408	0.122	23.261	0.000	2.601	3.080