Lead Scoring Summary

The objective was to build a logistic regression model for lead scoring, improving X Education's lead conversion rate.

Below approach was taken to work through this assignment -

1. Data Exploration and Preparation

Our first step was to thoroughly explore the dataset. We conducted a detailed Exploratory Data Analysis (EDA) to understand the distribution, relationships, and characteristics of each feature. This process also included identifying and addressing missing values and null categories, such as 'Select', which did not provide any meaningful information.

We discovered that several columns had varying levels of missing data, ranging from 1% to as high as 50%. Additionally, some columns had multiple categories, while others contained outliers, such as 'Total Visits'. Columns with more than 30% missing data were deemed unfit for analysis and were removed. Furthermore, any columns with constant values were dropped, as they added no variability or predictive power to the model.

For the remaining columns, we ensured proper imputation: missing numerical values were filled using mean or median imputation, and categorical values were replaced with the mode central measure of the column to preserve the dataset's integrity. This thorough data cleaning process set a solid foundation for accurate and reliable model building.

2. Model Building and Evaluation

After cleaning and preparing the dataset, we split it into training and testing sets (70-30 split) to build the logistic regression model. To ensure consistency, we scaled all the features before beginning the model building. We used Recursive Feature Elimination (RFE) to remove less relevant features systematically.

During RFE, we eliminated features with high p-values that did not significantly influence the target variable ("Converted"). After several iterations, we refined the model, achieving a Variance Inflation Factor (VIF) below 5 and a p-value threshold of 0.05, which indicated multicollinearity and statistical significance were within acceptable limits. We then proceeded to make predictions.

Our first model performed well, with an Area Under the Curve (AUC) of 89%. Despite attempting further improvements, the second iteration showed a decline in performance. Therefore, we reverted to the first model to fine-tune the cut-off threshold. We identified an optimal cut-off value of 0.3, where sensitivity (true positives / (true positives + false negatives)) was maximized without compromising accuracy and specificity.

Finally, we applied the model to the test data and evaluated the metrics, which confirmed the model's robustness. The evaluation metrics met our expectations, and we finalized the model based on these results.

3. Insights and Learnings

Analyzing the model coefficients, we identified that the most influential variables are:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6250	0.059	-27.523	0.000	-1.741	-1.509
Total Time Spent on Website	1.0987	0.040	27.153	0.000	1.019	1.178
Lead Profile_Potential Lead	1.6803	0.099	16.980	0.000	1.486	1.874
Lead Profile_Student of SomeSchool	-1.7175	0.445	-3.859	0.000	-2.590	-0.845
Lead Origin_Lead Add Form	3.4766	0.228	15.269	0.000	3.030	3.923
Lead Source_Live Chat	1.2261	0.106	11.553	0.000	1.018	1.434
Lead Source_Welingak Website	2.4586	0.755	3.257	0.001	0.979	3.938
Do Not Email_Yes	-1.4498	0.165	-8.776	0.000	-1.774	-1.126
Last Activity_Converted to Lead	-1.2386	0.216	-5.736	0.000	-1.662	-0.815
Last Activity_Had a Phone Conversation	1.5660	0.648	2.417	0.016	0.296	2.836
Last Activity_Olark Chat Conversation	-1.4237	0.166	-8.581	0.000	-1.749	-1.098
Last Activity_SMS Sent	1.2716	0.076	16.670	0.000	1.122	1.421
What is your current occupation_Working Professional	2.5692	0.190	13.515	0.000	2.197	2.942

These variables are strong indicators of lead engagement and intent, revealing that engaged behaviour is crucial for conversion.

4. Conclusion

Based on the model results,

1. Although the initial dataset contained numerous features that could potentially influence lead conversion, multiple iterations of model development helped identify the most impactful ones.
2. Leads with a high VIF score are more likely to convert.
3. Improving customer engagement by focusing on Lead Origin_Lead Add Form, can significantly enhance conversion rates.