Steps Followed:

1. Read and clean the data:
    - Handling missing values
    - Standardizing
    - Handling Outliers
2. Understand the data (Numeric and Categorical Analysis)
3. Exploratory Data Analysis
4. Prepare the data for modeling:
    - Feature Engineering/Mapping categorical variables to integers
    - Dummy variable creation
    - Test-train split and scaling
5. Model Building:
    - Feature elimination based on correlations
    - Feature selection using RFE
    - Running our First Model
    - Manual feature elimination (using p-values and VIFs)
    - Creating Predictions
6. Model Evaluation:
    - Accuracy
    - Sensitivity and Specificity
    - Optimal cut-off using ROC curve
    - Precision and Recall
    - F-1 Score
7. Predictions on the test set
    - Precision and Recall to verify

Synopsis:
The process usually starts with loading the data in your python notebook and then cleaning the same. The major three steps under cleaning the data are, treating the missing value columns, standardizing these treated columns to the extent to provide easy and intuitive analysis outputs, and finally handling if there are any outliers in the numeric columns. Once the data is cleaned and structured, we perform the basic numeric and categorical analysis to understand what does the data is projecting. This eventually helps us to get a better direction for our model building procedure. Just before building the model, it is very important to understand the relationship between the variables in hand either individually or with each other. This step is usually performed under Exploratory Data Analysis, as Univariate and Bi-variate analysis give us a clear picture about how each variable is performing and impacting each other in the data frame. Finally, we build a Logistic Regression model using our binary response variable or you could say our target variable, computing the probabilities of each response to predict which independent variable is impacting the response negatively or positively. This model once created is done verified with a few model evaluation techniques to check the stability of the model which we will be discussing in detail below regarding our above case study. Let's start by observing how we started the above case before getting into the model evaluation part.

Missing Value Treatment:
Once the missing values were identified in the columns, we followed certain steps to treat them, followed by standardization techniques available. Below are the insights that we gather from this case and took the decision.

- First, we dropped the columns having missing values > 35%, also, they seem to be irrelevant for model building as they are some characteristics captured about the lead.
- There seem to be some columns with high missing values but high importance too, after cross-checking them with the value counts, we replaced their null value with 'Not Selected'.
- Also, there seems to be a label named 'Select' with high counts in a few of the columns i.e Missing Completely at Random (MCAR), which means that the particular person has not selected any option from the dropdown. We Identified those columns and replaced them with 'Not Selected' values too.
- Many binary response columns seem to be important for our model-building predictions. Hence, performed certain standardization techniques for easy and intuitive analysis.

Understanding the data (Numeric and Categorical Analysis)
Numeric Variable Insights:

- We witnessed some loyal and dedicated customers as total_visits has a huge jump from between 75% quantile and max value which is 251.
- The time spent on the website metric displays that the average time spent by a customer is around 8.11 mins, which is a good number as per industry standards.
- On the other hand, a good page view per visit should be more than 4 pages, well in this case it was below average. Hence, X Education might have to improve its website content.

Categorical Variable Insights:

- Google seems to be the top source in generating leads.
- Most of the leads have checked their email by opening it.
- Most of the leads belong to India.
- Many of them have not selected their specialization, source_reference.
- Most of them are unemployed.
- Better Career Prospects is the most top course selection criteria
- Most of the leads have not seen magazine, newspaper_article, x_education_forums, newspaper, digital_advertisement, or search about them on any browser.

Identifying the spread of Converted Leads:
Insight: The Conversion Rate was 38.54%

Exploratory Data Analysis:

Univariate Categorical Analysis:

- Most of the leads have originated from submitting a form on the Landing page followed by details fetched through API.
- Apart from Google ads, most of the leads have been sourced through direct traffic on the website, Olark Chart, Organic Search, and a few References.
- Apart from email, most of the leads were also active via SMS and Olark Chat conversations.
- The highest number of leads specialized in Finance Management followed by HR management and Marketing management.
- Many of them have been referred through an online search and then follower word of mouth and students of some schools.
- Apart from being unemployed, the next highest occupation of leads are working professionals and students followed by very few housewives and businessmen.
- Most of the leads are from Mumbai city followed by Thane & Outskirts.

Bivariate and Multivariate Analysis

- Few converted leads were originated from the Lead Add form which seems to be working apart from API and Lead Page Submission. We can work on the content of the form or make it more inquisitive to optimize it in the future.
- Facebook, blogs, PPC, and others are not working for us in converting leads.
- SMS seems to be much more active in conversion than email.
- Data seems to be skewed towards India.
- Most of the converted leads did not provide their source_reference, occupation_status, specialization, country, and city too.

Analyzing the characteristics of leads in numerical columns with respect to Converted ratio.

- Some loyal customers seem to get converted even under 5 visits on the website with less than 4 page_views_per visit and many seem to be still exploring even after several visits.
- Strange to see people spending more than 16 mins (more than 75% quantile) and not getting converted.

Visualizing correlation between Numeric Variables vs Numeric values and Conversion

- total_visits and page_views_per_visit was kind of obviously being detected highly correlated to each other but the interesting thing to identify is even after visiting more than 4 pages, some leads didn't get converted which means X Education need to improve the content on the pages.
- time_on_website seems to be positively correlated with conversion as the data was tightly bonded to each other on both ends i.e 0 and 1.
- We witnessed that potential leads spending less time on the website tend to not convert eventually.

Model Building:

After following the above steps we got our first model with the insignificant variables that signify high P-values and VIF values.

- In the first table computed, our key focus area is just the different coefficients and their respective p-values. As, there were many variables whose p-values were high, implying that that variable is statistically insignificant. So we eliminate some of the variables moving forward to build a better model.
- We eliminated a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we then used manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).
- As per Industry, a good VIF value should be equal to or less than 5.00. Our Logistic Regression Model 10 (logm10) seems to have VIF values under control and significant P-values. Out of all the variables computed from our final model, the top 7 variables, that contribute towards lead conversion and should be focused on are as follows

1. Total Time Spent on Website
2. Lead originated from Landing Page Submission & Lead Add Form (positively impacting)
3. Lead source from Reference (negatively impacting)
4. Last Activity and Last Notable Activity both for SMS Sent
5. Lead Activity from Olark Chat Conversation (negatively impacting)
6. Occupation Status as Working Professional
7. Lead profiles from Student of Someshcool (negatively impacting)

Model Evaluation on Train Set:

First, we calculated Accuracy based on the confusion matrix, the confusion matrix was as follows,

Predicted	not_converted	converted
Actual		
not_converted	**3545**	439
converted	739	**1672**

Hence the Accuracy = (3545+1672) / (3545+439+739+1672) = Approx 81.57%

Sensitivity = True Positives / (False Negatives + True Positives)
Number of actual Yeses correctly predicted / Total number of Yeses
= 69.34%

Specificity = True Negatives / (True Negatives + False Positives)
Number of actual Nos correctly predicted / Total number of actual Nos
= 88.98%

Finding Optimal Cutoff Point: Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
Insights of the ROC curve graph:

- We saw the probability thresholds are very low, the sensitivity is very high and specificity is very low. Similarly, for larger probability thresholds, the sensitivity values are very low but the specificity values are very high.
- At about 0.4, the three metrics seem to be very close with decent values. At the cut-off of 0.4, all the metric values (accuracy, sensitivity, specificity) are 0.81, 0.78, and 0.83 respectively and hence we choose 0.4 as the optimal cut-off point.
- But finally, by visualizing the graph, it showcases that at about 0.35, the three metrics intersect hence, we decided, **0.35 is the optimum point to take it as a cutoff probability.**
- Now, we could've chosen any other cut-off point as well based on which of these metrics we want to be high. If we want to capture the 'Converted' better, we could have let go of a little accuracy and would've chosen an even lower cut-off and vice-versa. It is completely dependent on the situation we're in. In this case, we just chose the 'Optimal' cut-off point to give you a fair idea of how the thresholds should be chosen.

Precision and Recall:

Precision**:** TP / (TP + FP). Precision is related to Predicted values, hence will focus on Predicted columns in the matrix (FP and TP). Precision' is the same as the 'Positive Predictive Value.
= 79.20%

Recall = TP / (TP + FN). The recall is related to Actual, hence will focus on Actual columns in the matrix (FN and TP). Recall' is the same as sensitivity.
= 69.34%

F-1 Score: The F1-score is useful when you want to look at the performance of precision and recall together.

F1_score = (2 * P * R) / (P + R)
= 73.94%

Model Evaluation on Test Set:

Final Model Summary of the Test Set:

1. Overall accuracy on Test set: 82.67%
2. Sensitivity of our logistic regression test model: 82%
3. Specificity of our logistic regression test model: 83%

Conclusion:
There is a good balance between Sensitivity and Specificity.
The metrics seem to hold on the test dataset as well. So, it looks like we have created a decent model for the converted dataset as the metrics are decent for both the training and test datasets.