SUMMARY

Problem Statement:

X education is a company which sells online course to industry professionals. The company gets a lot of leads but the lead conversion for the company is very poor. They have assigned a team to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Solution:

Preparing and Cleaning Dataset:

- Given our dataset of over 9000 data points, we've identified numerous columns with a high percentage of missing values. As a rule, we're discarding columns with more than 30% missing data.

- Recognizing that our company exclusively offers online courses, we've opted to remove irrelevant variables such as City and Country.

- Prospect ID and Lead Number serve solely as record identifiers and hold no predictive value, thus we've removed them from our analysis.

- Columns exhibiting skewed data distributions have been excluded from our dataset due to their limited predictability.

- Following data cleaning procedures, we've determined a conversion rate of 48%.

Exploratory Data Analysis (EDA):

Univariate Analysis:

- It's hypothesized that the majority of leads originate from Landing Page Submissions, followed by those from API sources.

- Additionally, there's an observation suggesting that a significant portion of leads is sourced from unemployed individuals.

Bivariate Analysis:

- From the analysis of converted leads, it appears that those originated from Add Forms are more prone to conversion.

- Furthermore, demographics such as Working Professionals and Housewives exhibit higher conversion rates.

- The analysis also indicates that leads sourced from Live Chat, Reference, WeLearn, and the Welingak Website have a higher likelihood of conversion.

Model Building:

Here's a rephrased version of your statement:

- Categorical variables were transformed into dummy variables, and the dataset was divided into training and testing sets at a ratio of 70:30.
- Numerical features underwent scaling using MinMaxScaler to ensure uniformity in their ranges.
- Recursive Feature Elimination (RFE) was employed to identify the 15 most influential features in the dataset, enhancing the model's resilience.
- Following the initial model construction, the Variable Inflation Factor (VIF) and p-values were utilized to weed out statistically insignificant features.
- This iterative process ultimately resulted in a refined model consisting of 11 significant features, optimizing its predictive performance.
- We created a lead score (i.e. Conversion probability*100) to give a score between 0 and 100. A higher score indicates a hot lead having a higher probability of lead conversion

Model Evaluation:

- The area under the ROC curve was 86% which indicates this is a good model
- From the sensitivity and specificity tradeoff the optimal cutoff point was 0.44 and the metrics for the train set was

Accuracy	79.09%
Sensitivity	79.34%
Specificity	78.85%
Precision	77.71%
Recall	79.34%

Making Predictions on the Test Set:

- The metrics for predictions on the test set is as follows and they are very close to the training set.

Accuracy	78.95%
Sensitivity	77.71%
Specificity	80.10%
Precision	78.40%
Recall	77.71%

Conclusion:

The key features influencing the decision-making process are as follows:

1. TotalVisits

2. Total Time Spent on Website

3. Lead Origin categorized as Lead Add Form

4. Lead Source specifically identified as Welingak Website

5. Current Occupation indicating Unemployed status

6. Current Occupation indicating Student status

Learning:

- 1. TotalVisits
- 2. Total Time Spent on Website
- 3. Lead Origin, particularly through Lead Add Form submissions
- 4. Lead Source, with a notable emphasis on referrals from the Welingak Website
- 5. Current Occupation, notably the Unemployed category
- 6. Current Occupation, with a significant representation from the Student demographic.