Summary

Problem statement:

In the lead score case study client was having the lead but very few leads (approx. 30%) was converted successfully. This conversion percentage was small for the client wanted a solution to convert the maximum leads into the paying customer based.

Insights:

1. City column exhibits 39.71% missing values. Filling in missing data with 'Mumbai' could significantly skew the dataset, leading to potential model bias. Hence, it's advisable to `drop the City column`.

2. With 36.58% missing values, the Specialization column demonstrates an even distribution of selections. In this scenario, creating an additional category labelled `'Others'` is more appropriate than imputation or dropping.

3. Tags indicate the current status of leads and contain 36.29% missing values. Since this information may not be useful for modelling, it's `recommended to drop the Tags column`.

4. This variable, with 29.32% missing values, sees 'better career prospects' selected by 99.95% of customers, indicating significant skewness. Thus, it's unlikely to `provide meaningful insights`.

5. What is your current occupation: Imputing missing values with 'Unemployed', the most prevalent occupation, seems appropriate given X Education's context of selling online courses.

6. Country: Around 96% of customers are from India, making it impractical to impute missing values with 'India'. Hence, `dropping the Country column` is recommended.

7. Last Activity: "Email Opened" is the most frequent activity, and only 1.11% of values are missing. Hence, imputing missing values with 'Email Opened' is a reasonable strategy.

8. Lead Source: "Google" is the most common source, and only 0.39% of values are missing. Therefore, imputing missing values with 'Google' is a logical 'Google' is a reasonable strategy.

Insights Univariate:

1. Here's a breakdown of features from variables predominantly present (both Converted and Not Converted included)
2. Lead Origin: "Landing Page Submission" accounts for 53% of customers, while "API" accounts for 39%.
3. Current_occupation: Approximately 90% of customers are categorized as Unemployed.
4. Do Not Email: Around 92% of people have opted out of receiving emails about the course.
5. Lead Source: Combined, Google and Direct Traffic represent 58% of the lead sources.
6. Last Activity: The majority, constituting 68% of customers, are engaged in activities such as SMS Sent and Email Opened.
7. Lead Origin: Approximately 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%. The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.
8. Lead Origin: Approximately 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%. The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.
9. Lead Source: Google has an LCR of 40% out of 31% of customers, while Direct Traffic contributes a 32% LCR with 27% of customers, which is lower than Google. Organic Search also yields a 37.8% LCR, but only 12.5% of customers come through this Lead Source. Reference has an LCR of 91%, but there are only around 6% of customers through this Lead Source.

To increase our Lead Conversion Rates:

1. Emphasize features with positive coefficients to tailor marketing strategies effectively.
2. Implement targeted campaigns to attract high-quality leads from the most successful lead sources.
3. Craft personalized messages aimed at engaging working professionals effectively.
4. Optimize communication channels based on their impact on lead engagement.
5. Allocate more budget to advertising on the Welingak Website for increased visibility.
6. Offer incentives or discounts for successful lead referrals to encourage more referrals.
7. Aggressively target working professionals due to their high conversion rates and potentially better financial situations for higher fees.