

Summary Report

This analysis was performed for X Education to explore ways to increase enrolment in their course from industry professionals. The provided data offered crucial insights into how potential customers engage with the website, including their time spent on the site, the channels through which they access it, and the conversion rates.

For the above understanding following steps were used:

1. Cleaning Data

- Columns with more than 37% missing values were removed. For categorical columns, value counts were examined to determine the best course of action. Columns that provided little value were also removed.
- Numerical categorical data were imputed with the mode, and columns with only one unique customer response were dropped.
- Additional steps included handling outliers, correcting invalid data, grouping low-frequency values, and mapping binary categorical values

2. EDA

- Data imbalance was assessed, revealing that only 38.5% of leads were converted.
- Univariate and bivariate analyses were conducted for both categorical and numerical variables.
- The amount of time spent on the website was found to have a positive effect on lead conversion.

3. Data Preparation

- Dummy variables (one-hot encoding) were created for categorical features. The data was split into training and test sets.
- Feature scaling was applied using standardization. Several columns were dropped due to high correlation with each other.

4. Model Building

- RFE was applied to reduce the number of variables to 20, making the dataframe more manageable.
- Subsequently, the remaining variables were manually removed based on their VIF values and p-values, retaining only those with a $VIF < 5$ (signs of multicollinearity) and a p-value < 0.05 (stable values).

5. Model Evaluation

- A confusion matrix was created, and a cut-off point of 0.5 was selected based on accuracy, sensitivity, and specificity plots.

- Prediction was done on the test data frame and with an optimum cut off as 0.5 with accuracy of 89%.
- Recall was 83% on the test data frame efficiently predicting hot and cold leads.

As per the analysis, the variables that were of utmost importance to the the potential buyers are listed below:

- **Tags_Closed by Horizzon and Tags_Lost to EINS:** Both of these tags likely represent stages or interactions in the sales process that strongly correlate with successful lead conversion suggesting that leads lost to a competitor (EINS) may still have a higher chance of conversion if they re-engage.
- **LastNotAct_Had a Phone Conversation:** provide valuable insight into the likelihood of converting a lead, with phone conversations having a positive effect on the conversion rate.
- **Total Time Spent on Website:** more time spent on the website, the more likely the lead is to convert.
- **LastNotAct_SMS Sent:** if an SMS has been sent in the "Last Notable Activity," the likelihood of conversion increases.
- **CurOccu_Working Professional:** Being a "Working Professional" increases the likelihood of lead conversion.

Recommendations:

1. Re-engage Leads Lost to Competitors

- Reconnect with leads marked as lost to EINS, as they still show potential for conversion. Use personalized outreach or special offers to win them back.

2. Prioritize Phone Conversations

- Encourage more phone conversations with leads, as they significantly improve conversion rates. Train sales reps to provide personalized, high-quality interactions.

3. Enhance Website Engagement

- Improve website content and user experience to keep visitors engaged longer. Offer interactive content and personalized recommendations to boost time spent on site.

4. Leverage SMS Follow-ups

- Implement automated SMS follow-ups with leads, as they increase conversion chances. Keep messages short, personalized, and valuable, such as reminders or special offers.

5. Target Working Professionals

- Tailor marketing efforts to working professionals, highlighting time-efficient courses or certifications that align with their career advancement goals.

6. Refine Lead Scoring

- Enhance lead scoring by incorporating key factors such as phone conversations, website activity, and occupation to better prioritize high-potential leads.