Executive Summary

Problem Statement:

The X Education company is facing challenges in lead conversion despite generating significant leads daily. The current lead conversion rate stands at a modest 30%, well below the desired target of 80%. The company seeks a solution to identify 'Hot Leads' – those with the highest potential for conversion – to prioritize resources towards engaging with leads most likely to convert into paying customers.

To address this issue, a team comprising Harsh Patel, Nikhil Jindal, and Vaibbhav Nadkarnni conducted an analysis using machine learning approach to overcome this issue. The primary objectives were to increase the lead conversion rate from 30% to 80% by identifying and prioritizing the most promising leads, build a model to assign lead scores, and deploy the model for future application.

Solution Summary:

The solution methodology involved data cleaning and manipulation, exploratory data analysis (EDA), dummy variable creation, binary variable encoding, feature scaling, building a logistic regression model, evaluating the model, creating a ROC curve, and presenting the model with conclusions and recommendations.

The EDA revealed several insights about the lead data. Leads from India formed the majority, and Mumbai city showed a high conversion rate. Leads from Google and Direct Traffic, as well as Welingak Website, Organic Search, and Reference, had high conversion rates. Landing Page Submission and API had the highest number of leads and conversion rate. Management specialization and unemployed customers showed the highest conversion rates. Leads who opted to receive emails and those who did not opt for a free copy of Mastering the Interview had higher lead counts and conversion rates.

After data conversion and handling outliers, a total of 8,953 rows and 64 columns were available for analysis. The next step was to divide the dataset into test and train dataset with 70% and 30% ratio. A logistic regression model was built, utilizing recursive feature elimination (RFE) for feature selection. The model achieved an accuracy of 93.11%, sensitivity of 94.18%, specificity of 92.50%, precision score of 87.71%, and recall score of 93.12%.

The ROC curve analysis revealed an area under the curve of 0.97, indicating an excellent model. The optimal cutoff point for balanced sensitivity and specificity was found to be 0.25.

Conclusion:

The key conclusions from the study are that variables essential for identifying potential buyers include Last Notable Activity, Lead Origin, Tags, Current Occupation, and Lead Source. Leads generated from Lead Add Form, those who revert after email, working professionals or unemployed individuals, and leads from Welingak Website, Olark Chat, Reference, Google, and Direct Traffic should be targeted as a priority.

By implementing the recommendations from this study, X Education can effectively prioritize and engage with the most promising leads, increasing the overall lead conversion rate and achieving their desired target of 80%.