

# Summary Report: Increasing Lead Conversion with Predictive Modeling

## Introduction:

In this assignment, the objective was to help X Education, an online education company, increase their lead conversion rate by identifying the most potential leads using a logistic regression model. The initial lead conversion rate was around 30%, and the company aimed to achieve a target conversion rate of 80%. This report summarizes the approach taken and the key learnings gathered throughout the analysis.

## Data Inspection and Handling Missing Values:

The dataset provided required thorough inspection to identify and handle missing values appropriately. Columns containing "Select" values were considered as null, and the "How did you hear about X Education" column, which had 75% null values, was dropped due to limited usefulness. Other columns with less than 40% missing values were imputed using mean, median, or mode values. For the "Specialization" feature, a new label, "No Specialization," was introduced to handle the 36% missing values without biasing other features.

## Data Preparation and Feature Engineering:

To improve the model's predictive performance, binary variables (Yes/No) were converted to 1/0 values. Categorical variables with multiple labels were transformed into dummy/indicator features. This process expanded the feature set and provided a more comprehensive representation of the data.

## Outlier Treatment and Feature Scaling:

Outliers were identified and treated in variables such as "TotalVisits," "Total Time Spent on Website," and "Page Views Per Visit." Mapping the values from 0 to 99th percentile helped remove extreme outliers and maintain data integrity. Feature scaling was applied to ensure that variables with different scales did not impact the model disproportionately.

## Correlation Analysis:

Correlation analysis was conducted to identify relationships between variables. A heatmap revealed only a few variables with significant correlation, which led to the decision of dropping highly correlated dummy variables. This step reduced multicollinearity and improved the model's interpretability and stability.

### **Model Building and Evaluation:**

Logistic regression was chosen as the modeling technique to assign lead scores. The model was trained on the prepared data and evaluated using accuracy and various performance metrics. The selected model achieved an accuracy of 79.89%, demonstrating its ability to predict lead conversion. Sensitivity, specificity, false positive rate, precision, and recall scores provided a comprehensive understanding of the model's performance.

### **Feature Selection and Model Refinement:**

Recursive Feature Elimination (RFE) was employed to select the most relevant features for lead scoring. Variables with high p-values and multicollinearity were iteratively dropped to refine the model. The final model demonstrated significance in the remaining features and showed no multicollinearity concerns.

### **Metrics Beyond Accuracy:**

The analysis went beyond accuracy to evaluate sensitivity, specificity, false positive rate, positive predictive value, and negative predictive value. These metrics provided insights into the model's ability to correctly identify positive conversions, minimize false positives, and predict lead conversion probabilities.

### **Lead Score Assignment and Interpretation:**

The final model's equation was derived to assign lead scores based on the selected features. The coefficients of the model provided insights into the importance of each feature in predicting lead conversion. The interpretation of coefficients helped understand the factors that significantly influenced a lead's conversion likelihood.

### **Business Implications and Recommendations:**

The analysis revealed the potential for X Education to improve their lead conversion rate by targeting "Hot Leads" identified through the logistic regression model. By focusing their sales efforts on leads with higher scores, the company can optimize their resources and increase efficiency. The recommendations include leveraging the lead scoring system to prioritize and personalize communication with potential leads, enhancing the chances of conversion.

### **Learnings and Conclusion:**

In conclusion, this report outlined the approach taken to increase lead conversion through predictive modeling. The logistic regression model successfully assigned lead scores and provided valuable insights into factors influencing conversion likelihood. The learnings gathered from this analysis emphasized the

importance of data cleanliness, feature engineering, model evaluation, and the use of performance metrics beyond accuracy. By implementing the recommendations and leveraging the lead scoring system, X Education can enhance their lead conversion rate and achieve their desired business objectives.