Summary Report: Lead Score Case Study

Problem Statement:

X Education is an online education course provider company that provided various upskilling coursed to professionals and students. It markets its products on various platform to build lead funnel that are nurtured and converted to final sales. The challenge faced by X Education is that the lead conversation ratio is low and here is precisely where they want to build an predictive model to address this.

Dataset:

The data set provided contained about 9240 rows and 37 columns, this included information such as prospect ID, lead source, Total visit by prospect, Total time spent by prospect on the webpage, there geographical location, type of communication preferred by them, there profile and various other information gathered during the marketing stage.

Data Preprocessing:

Before conducting the analysis, the dataset underwent thorough data preprocessing steps to ensure its quality and suitability for the logistic regression model. **Missing values** were handled through imputation techniques, and **outliers** were identified and addressed appropriately to prevent any undue influence on the model. One peculiar observation happened to be noticed that is the **SELECT** attribute this appeared in 4 columns and added less interpretability of data hence considered as NULL and treated as disused above.

Exploratory Data Analysis:

The initial phase of the case study involved conducting exploratory data analysis (EDA) to gain insights into the distribution and relationships between variables. We plotted the box plots and bar graph for categorical variable against the target variable CONVERTED. Also, correlation between numeric variable was plotted and checked.

Logistic Regression Model Building:

The logistic regression model was built to predict the probability of a customer converted based on the independent variables. The model was trained using a portion of the dataset and evaluated on a separate validation set to ensure its predictive performance.

Model Evaluation:

To assess the model's performance, various evaluation metrics were employed, including accuracy, sensitivity and specificity and the receiver operating characteristic (ROC) curve. The model's ability to correctly classify prospects who converted and those who did not was analysed through these metrics. Additionally, a confusion matrix was generated to further examine the true positive, true negative, false positive, and false negative rates.

Results and Insights:

The logistic regression model built gave good metrices of accuracy, specificity & sensitivity in both Training & Test sets. A healthy AUC of 92% was seem in the ROC curve again pointing to the excellent model health. Future the model coefficient showed the FEATURES that added value to the predictions and the features that least adding value. Finally, the LEAD SCORE was arrived at by arranging the probabilities of conversion in the descending order and considered as HOT LEADS as desired by BUSINESS.