Lead Scoring case study summary

This analysis is done for X Education company to find ways to get more industry professionals to join their courses. The data provided lots of information to like time spent by the leads, leads visits on site, how they reached site and conversion rate etc.
This report is structured into four sections :

- Introduction
- Data and methods
- Results
- Implications
- Conclusions

1. Introduction
    - An X education company sells online courses to industry professionals. Although it gets a lot of leads, its lead-to-scale-conversion rate is very poor. The purpose of the project is to get increase in leads conversion rate.
    - Developed a lead scoring model that can identify high-quality leads based on a set of criteria and metric.
2. Data and Methods

    2.1 Data cleaning
    - The data was cleaned by dropping unnecessary columns and imputing missing values with mode or 'unknown'. Outliers were removed from numerical columns using boxplots.

    2.2 Exploratory Data Analysis
    - Exploratory data analysis was performed to identify patterns and trends in the data.

    2.3 Data Preparation
    - Data was prepared by binary conversion of data and creating of dummy variables for categorical columns, and scaling numerical features using standard scaler method.

    2.4 Model Building
    - Feature selection was performed using RFE and manual selection based on p-values and VIF values.

    2.5 Model evaluation on train dataset
    - The model was evaluated on the training dataset using various performance metrics such as accuracy, sensitivity, specificity, and confusion matrix. The optimal cut-off point was determined using sensitivity.
    - The lead score was generated using the logistic regression model's predicted probability of conversion, and ROC curve was used to evaluate model performance with an AUC of 0.89.

    2.6 Making Prediction on test dataset
    - Predictions were made on the test dataset after applying the same preprocessing steps as the training dataset.
3. Results
    - The top5 significant features based on the logistic regression coefficients and VIF values were 'lead Origin lead add-form : 2.8', 'occupation working professional : 2.4', 'last activity SMS sent : 1.9', 'lead source Welingak Website : 2.5', 'Total time spent on website :1.06'.

- The VIF values for these features range from 1.21 to 2.97, indicating that they are not highly correlated with other predictors in the model.
- The confusion matrix showed that the model correctly predicted 3251 didn't converted and 1990 leads.
- The accuracy of the model is 81%, which means that it correctly predicted the target variable 81% of the time.
- Sensitivity is 80%, which is a good indication that the model is able to identify the majority of actual positive cases.
- Specificity is 81%, which indicates model has correctly predicted actual negative cases.
- The precision shows that out of all the predicted positive cases, only 72% were actually positive.

4. Conclusion
    - Prioritize 'reference' and 'welingak website' for improved conversion rates.

Focus on leads generated through 'landing page submission' and API, which had a higher co nversion rate. Consider optimizing the features 'do not email', 'total time spent on website', ' lead origin landing page submission', 'lead origin add form', 'lead source olark chat', 'last ac tivity email opened', 'last activity olark chat conversation' to improve overall lead conversio n rates. Prioritise leads generating from 'working professional'.