Lead Score Summary

Problem Understanding:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%

Problem statement:

Currently X Education lead conversion is only around 30%, they are looking for conversion rate of 80%.

Steps followed:

1. The past 9240 leads data given is loaded into the system.
2. **Exploratory Data Analysis** is performed. In univariate analysis boxplot and histplot is plotted on numerical columns **TotalVisits, Total Time Spent on Website, Page Views Per Visit.** Categorical plot is plotted for categorical columns. In Bivariate analysis boxplot is plotted between numerical and categorical columns. Box plot of target column **Converted** and categorical columns are plotted. Pair plot between numerical columns are plotted. Heatmap is also plotted to check the correlation between different columns.
3. **Data Cleaning:** Columns with missing values more than 3000 is dropped. Similarly columns of location is dropped since the courses will take place online. Other columns with data imbalance such that only one value is present are dropped. ID columns are also deleted. Then finally null value rows are dropped.
4. **Preprocessing**. For categorical columns dummy variables where created and dropped the original column.
5. **Train Test data split:** the preprocessed data is split into train (70%) and test (30%) data with random state.
6. **Scaling:** Min Max Scaler is used to scale numerical values.
7. **Feature Selection:** RFE (Recursive Feature Elimination) is used to select top 15 columns required to build a logistic regression model.
8. **Logistic regression Model building:** Using GLM the model is built, the statistical features and VIF (Variance Inflation Factor) is checked.
9. **Dropping columns:** Columns which has high P-value and high VIF are dropped one by one.
10. **Predicted values on Train set:** From the final model probability of lead conversion is predicted, the lead is classified as Converted or not based on probability, if the probability is greater than 0.5 then lead is classified as converted else not converted. Accuracy of the model is calculated, which came to 79%

11. **Determine Optimal Cut off:** for different probability accuracy, sensitivity and specificity are calculated and plotted. From the plot cut-off probability is determined as 0.42 and accuracy is 79%.
12. **Making prediction on the test set:** the test values are predicted by using the final model which was built earlier. If the predicted value is greater than 0.42 then the lead is converted else the lead is not. Accuracy is determined to be 78%. Sensitivity around .77 and specificity around .78.
13. **Precision and Recall:** for different probability, precision and recall are calculated and plotted. From the plot cut-off probability is determined as 0.44
14. X Education Company needs to focus on following key aspects to improve the overall conversion rate:
    - Improve the user engagement on their website to help in higher conversion
    - Increase on sending SMS notifications since this helps in higher conversion
    - Get TotalVisits increased by advertising etc. since this helps in higher conversion
    - Improve the Welingak Website since this is affecting the conversion negatively