Summary

Problem

An education company named X Education sells online courses to industry professionals. 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. The typical lead conversion rate at X education is around 30%.Now, although X Education gets a lot of leads, its lead conversion rate is very poor so they want to improve it.

Solution Proposed

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Steps Involved

- **Importing the data** : In this step we import all the necessary libraries and the converted CSV file of the data into a data frame.
- **Inspecting the data :** In this step we inspected the data for its shape, missing values, outliers and data types of the columns.
- **Data Preparation :** In this step we handled missing values in all the columns using various techniques like median or mode replacements and making a new category of the missing variables and in some case dropping the column if the missing values are higher than 40%. As there were no visible outliers in the data we concluded that the data is in proper form for applying EDA.
- **EDA and Data Modifications :** In this step we applied EDA on the given data and then modify the data based on our analysis such as reducing the categories In a column or dropping a column etc. We Also came out with some amazing insights from the data.
- **Preparing data for modeling** *:* In this step we will prepared the data for modeling, which includes Dummy variable creation for categorical variables, Test Train Split (30-70 split)and Standard Scaling for numerical variables.

- **Model building :** In this step we choose the top 15 variables using Recursive feature elimination technique(RFE) and them reduced the feature using manual approach using Logistic model from stats module with filter such as P-values < 0.05 and VIF < 5.
- **Making ROC curve :** After making model we checked the model using ROC curve in which we got an area under the curve of (AUC) = 0.89 which is good for a logistic regression model.
- **Finding Optimal Cutoff Point :** In this step we find the cut-off point for the model by plotting accuracy, sensitivity and specificity for various cut-offs. We concluded that 0.3 is a good cut-off for our analysis.We finally got a sensitivity of 0.84 for the training set. (In this case we take sensitivity as our metric as we want to reduce the false negative values).
- **9) Making predictions on the test set :** Finally we predict using the final model on the test set and achive sensitivity score of 0.83 on the test set.

Summary

It was found that the variables that mattered the most in the potential buyers are (In descending order):

- Lead Source_Welingak Website
- Lead Source_Reference
- What is your current occupation_Working Professional
- What is your current occupation_not selected
- Lead Source_Olark Chat
- Last Activity_SMS Sent
- Total Time Spent on Website
- Do Not Email
- Specialization_Select
- Specialization_Hospitality Management
- Lead Origin_Landing Page Submission
- Last Activity_Other activity
- TotalVisits_10+

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to buy their courses