Summary

Analysis is done for an education company named X Education where business demand is to find the most potential leads, also known as Hot Leads in an efficient manner. The given dataset provided a lot of information about professionals who are interested in the courses, how much they visit the site, the time they spent on site & how they reached the site in the first place.

The business objective includes building a Logistic Regression model to predict whether a professional will convert or not, given the data & to assign a Lead Score value between 0-100 based on the conversion chances.

Below are steps we took to achieve the objective:

1. Data Understanding & Feature Selection:

- We started with basic steps like loading the dataset, checking the shape of the dataset, checking numeric feature descriptions, etc.
- Dropped a few redundant features initially since they contained a unique value for each customer & were not necessary.
- We renamed features to shorter strings using snake-case nomenclature for better readability since original feature columns had spaces & were large strings.
- Dropped some features having only a single value or only 2 values. Some features with few values were also highly skewed to a single value. We dropped such features.
- We checked for null values & dropped features with null values> 30%. We did some imputations for the rest depending on the type of value the feature column had.
- Then we checked numeric columns for collinearity using a heat-map. We also checked for outliers in numeric columns using heat-map, took the necessary steps to handle outliers.
- We did some visualization for their relevance for the model building & completed the feature selection process with it.

2. Data Transformation:

- We changed features with binary variable (Yes/No) to 0 & 1.
- Features having multi-category were handled by one-hot encoding.
- For model building, we split the dataset into train-test. Train dataset was used for model building & test dataset was used to evaluate the built model.
- Before model building, we also scaled numeric features using Standard scaler.

3. Model Building & Evaluation:

- For Model building, we used a mix of automated approach & manual approach.

- We used the ScikitLearn library to get the top 15 relevant features. We started building the model with the selected features & dropped features in further steps depending on their statistical significance from p-Value & multicollinearity using VIF value. We selected features with p-Value <0.05 & VIF <5.

- We got the desired result for accuracy, specificity & sensitivity

4. Conclusion:

- Since we got a similar score, we consider the final model to be a good model without any overfitting issues. In business terms, our model is stable and has good accuracy. It'll also adapt to companies requirement changes made in coming future.
- A few top features for a good conversion rate are:
    i. `ls_not_act_Had a Phone Conversation`
    ii. `lead_src_Welingak Website`
    iii. `lead_org_Lead Add Form`

At the end, we achieved the business agenda of analysis & provided necessarysteps that can be followed by the company to achieve its need.