SUMMARY

PROBLEM:

An education company called X Education which sells online courses to industry professionals. The company needs our help in selecting the most potential leads, i.e. the leads that are most likely to get converted into taking up there courses and become a paying lead. The company requires a model which contains a lead score assigned to each of the leads such that the customers with higher lead score would have a higher conversion chance and the customers with lower lead score would have a lower conversion chance. The CEO of the company, has specifically given a ballpark of the target lead conversion rate to be around 80%.

SOLUTION:

These were the steps followed:

1. **Importing, reading and understanding the data:**
    - The data was imported
    - The data was read
    - Then we understood its variables, its statistical info, its data types, etc.
2. **Data Cleaning:**
    - Firstly we transformed the select values to null values as no info was given.
    - We checked for variables that had high percentage of null values (30% and above) and dropped them or imputed them based on importance of the variable.
    - We also checked variables having null values below 30% and performed various imputations or dropped them depending on importance.
    - We also checked skewness, variables with unique values and dropped them which were highly skewed.
    - We also did outlier treatment and capped them at 99% as most outliers were above 99% quantile.
3. **Data Analysis:**
    - We did univariate analysis of the categorical and the numerical variable and found various insights on it.
    - Then we did Bivariate analysis by relating the categorical and numerical variables with the target variable (target variable is Converted) and found various insights on it.
    - During the univariate and bivariate we dropped variables that were felt to be redundant or just had a single value.
4. **Data Preparation:**
    - We started by converting the binary variables to 0 and 1.
    - Then we created dummies for the remaining categorical variables.
    - We then split the data into train and test.
    - We scaled the feature using the Min-Max scaler

5. **Model Building:**
    - We did feature selection using RFE (Recursive Feature Elimination) and selected the top 15 important features.
    - We then build a model using logistic regression.
    - We did manual elimination of the variables having high p-value (above 0.05) and high VIF (Variation inflation factor above 5).
    - We finally arrived at fair model having significant p-value and VIF.
6. **Model Evaluation:**
    - We plotted the confusion matrix which contained the True positive, True negative, False positive and False negative.
    - We calculated the accuracy, sensitivity and specificity of the model.
    - It was low based on our cut off, so we calculated the optimal cut off using ROC curve and the threshold cut off came out to be 0.32.
    - We rechecked the accuracy, sensitivity, specificity which came out fair enough (78%, 82%, and 75%).
    - We also calculated the % of correctness of the model which came out to be 82%.
7. **Model Prediction**
    - We made predictions on the model using the test set.
    - We checked the accuracy, sensitivity and specificity of the model and it was almost similar to out train set (78%, 83% and 75%), hence concluding the model was good.
    - We cross checked using the metrics precision and recall.
    - We also calculated the various lead probabilities and assigned a score to each of them ranging from 0 to 100, i.e. 100 meaning a highly convertible lead and 0 meaning a no converting lead.

Inferences:

- More the leads visits the website more he is to convert and pay.
- The time spent by the lead on website also plays an important role, higher the time then higher is the conversion rate.
- Leads attained via Google, directly enquiring, organic search, add form, welingak website, referral sites and Facebook have strong relation with respect to conversion, i.e. it has inverse relation except welingak website leads, so this needs to be looked respectively.
- Lead which are unemployed, are students, working professionals also have a higher conversion rates.
- So, focussing on these features majorly would help increase the company’s conversion rate.