Summary

The main objective of this analysis is to increase the lead conversion rate of the sales team of an education company:- X - Education by prioritising the successful identification of the leads who are more likely to be converted into a student. The company believes that if they pick out these set of leads, the lead conversion rate should go up as well.

Solution methodology for the analysis:-

1. **Data cleaning**:-
    - The dataset had a set of “score” variables given to each lead by the sales team based on their communication with them. These variables were dropped.
    - The variables with highly skewed data were also removed as they won’t contribute a lot to our analysis.
    - Some categorical variables had a value of “select” in them, these were treated as missing data and were imputed with NaN value in order to identify how many of them were present in the dataset.
    - Variables with % missing values greater than 40 were dropped as they would not have contributed to the analysis a lot.
    - Outliers in the numerical variables outside of the 99% quantile were removed as we had sufficient amount of data to work with.
    - Categorical variables which were well distributed but had classes with fewer occurrences were clubbed together to form a single class.

2. **EDA**:-
    - Univariate analysis of the categorical variables suggest:-

        - People who don’t want a copy are more likely to opt for the course.
        - A call is more likely to be successful if the user has opted for the email service.
        - People who have learned about the course through reference have the highest conversion rate.
    - Univariate analysis of numerical variables suggests that most people spend less than 250 minutes on the website.
    - Bi-variate analysis of the variables:-
        - People who landed on the website via Organic search, Google, referral sites and spend a lot of time on the website are most likely to opt for the course.

3. **Data preparation**:-
    - Categorical variables are dummyfied
    - Dataset is split into train and test data set.
    - Numeric variables with high values are standardised.
4. **Model Building**:-
   Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-values(The variables with VIF<5 and p-value<0.05 were kept.)
5. **Predictions**:-
   Sensitivity was chosen as the main evaluation metric of the model and a cut-off which maximised it while maintaining a decent score for the accuracy and specificity was chosen. The cut-off of 0.3 was used to make predictions on the test data set.
6. **Model Evaluation**:- The sensitivity, specificity and accuracy for the final model with cut-off as 0.35 came out to be 0.83,0.78 and 0.8 respectively on the test dataset.

It was identified that the Top -3 KPI’s which X-education should focus on, in order to identify whether a lead will turn into a student or not are -

- Total time spent on the website by the leads.
- Leads who are working professionals.
- The leads which come across their product via the Welingak website.