

Summary

The X education has appointed us to help them generate the most promising leads, to be precise, get leads that convert into paying customers. Also, we are given the task of assign scores to each leads. Higher the score higher the conversion chance. We are provided with data which tells us about how much time potential customer spent on website, times they visited the website and such.

The following steps are used:

- **Data Cleaning:** After importing the data, we checked for the null values, there were good chunk of null values, so we capped the columns to a limit of 40% null values. Also there were categories named select, that needs to be replaced with null values, or 'not mentioned'. To avoid losing data this replacement was done. Post this step dummies were created as an important step in model building. Due to high imbalance of data in some columns, some columns were dropped.
- **EDA:** A quick eda was done to explore trends among data, strong correlation was found among 'total visits' vs 'leads number'. It was also found that 'Management specialisations' have good conversion ratio. Also outlier treatment was done, potential outliers were capped upto 99%.
- **Dummy variables/Scaling:** Categorical variables were converted into dummy variables and scaling was done on dataset. Also, checks were done to reduce the dimensions of categories to make the model light-weight. Scaling is done on both train and test data set.
- **Train-Test split:** Train test split was performed and logistic regression was instantiated. It was performed in the ratio of 70/30 %.
- **Model Building:** For building a successful model, recursive feature elimination as well as variance inflation factor was used. With the help of user defined functions, we kept the limits of VIF at 5. After 11 iterations model was build, next step involved using metrics and threshold
- **Prediction :** Cut-off was kept at 0.3, with accuracy, recall and specificity well above 85%. On the test set also , model gave similar results.

Top three variables that the edtech company can focus on are:

- a. Tags_Closed by Horizon
- b. Tags_Lost to EINS
- c. Lead Source_Welingak Website

Also if there is a scenario where company need to make the lead conversion more aggressive, the model can be adjusted to low threshold to make it predict all conversions.

There is also another scenario where, company has attained its target before the quarter and now it needs to focus on new work, here the model can be tuned to attain high specificity, as specificity increases the model correctly predicts all the non conversions, so the cut-off value needs to be high to achieve this.
