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. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

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 loosing 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%. The numeric values seems good and no outliers were found.

• Dummy variables/Scaling: The dummy variables were created and later on the dummies with 'not mentioned elements were removed. For numeric values we used the MinMaxScaler . 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: Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept. After 11 iterations model was build, next step involved using metrices and threshold

• Prediction : Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%. 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 Horizzon

b. Tags_Lost to EINS

c. Lead Source_Welingak Website

- Precision – Recall: This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame.

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 quarterand 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, sothe cut-off value needs to be high to achieve this.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX