Model Summary

X-Education, A leading online platform who sales online courses to working professional, wants to make sure that they call only leads whit a higher probability of conversion to optimize resources, and to adhere this process, they have been collecting data on multiple parameters.

As the requirement here is to predict if a lead is hot or cold, we can state this problem as binary classification problem.

We have used these parameters to create a statistical model to continently identify most probable leads, based on a scoring mechanism.

After getting the data into a data frame, primary analysis has been performed on the data set to identify the distribution of data in each feature, if there are any skewness in the features and if there are missing values in the data.

After carefully assessing each and every feature we selected the features that had more relevance in terms of successful analysis and we decided to keep out the highly skewed features which had a clear dominance of an individual subcategory. We imputed missing values for the features with a high number of missing values and we decided to remove the datapoints for features with low number of missing values.

Moving on with the analysis, we checked the relation between categorical features and conversion and the summary of those analysis is clearly mentioned in the presentation.

After that, we grouped low frequency subcategories in a single subcategory and then created dummy variables for those categories, we split the data into 4 parts (X_train, X_test, y_train, y_test) and standardized the values of the numeric columns in the training dataset. This concludes the data cleaning and data preprocessing steps performed at this analysis at a high-level. After this we had 36 variables in hand to continue with the next step.

After this we went on with feature selection which involved use of RFE and manual feature elimination process. At first, we built a logistic regression model using SK Learn framework, then we applied RFE to get the most significant 15 variables which represents the variability of the dependent variable, we created a new data set with the above mentioned 15 variables after adding in the constant term, and then, we initialized logistic regression algorithm from ‘Statsmodels.api’ and fitted the data into it.

From the summary of the learned parameters we got a pretty clear idea of the variables which had a higher P value and was less significant for that reason, but, we also complemented this understanding of manual feature elimination process using VIF, variables with a high VIF score and a high P value(VIF more than 5 and P value more than 0.05) is removed from the dataset one at a time ordered in a descending manner, after removing one feature with high VIF or high P value we rebuilt the model and did the same checks again and repeated the process until we got read of all the variables with higher p values or a high VIF. Then we used confusion matrix to get Sensitivity, Specificity, Accuracy and precision scores of the model for a rando threshold.

Then, we use ROC to determine a suitable threshold based on the business problem and used this model for further predictions with the threshold that meets the busyness needs.
A score is then assigned to each datapoint to mark leads with high conversion rate.