Lead Score - Summary Report

We have been provided with data from Education company X, which sells online courses to working professionals. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead-conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

For the analyzing purpose we have been provided with **9240** data points and **37** features, with most features having data type as object.

There are a few variables which have level as ***'Select'***, meaning the corresponding data has not been entered by the customer. Such entries need to be considered as missing values. It has been found that nearly **6 features** though useful **have more than 30% missing value**, leading to their **removal. ,'Receive More Updates About Our Courses', 'Magazine', 'I agree to pay the amount through cheque', 'Update me on Supply Chain Content', 'Get updates on DM Content'** these features had only 0 as entries, so we have dropped these columns as they didn't contribute to the analysis.

The missing values of **categorical features** have been filled by **mode** value. As most of the **numerical data** had **skewed distribution** their missing values are filled by **median** values of the respective features.

Some of the features such as '**countries', last notable activity', 'last activity' and 'lead source'** had many sub features. We did **feature engineering** and grouped those sub-features under possible common heading. These along with other categorical and binary features are then converted **dummy variables** for further analysis through the model. In the numerical features **'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit'** were treated for the outliers.

The data is then checked for **class imbalance**, the value **being 1.59** though a moderate class imbalance is **not an extreme** one. So, we proceed without any treatment.

To treat for multicollinearity, we check for **correlation** value and **removed** those that have values **above 0.6 and below -0.6**.

In the model building stage, the **train data is standardized** for easy calculation and the model is built using logistic regression. Then we use **RFE to reduce the features to 15**. With the features obtained we again look for p value and VIF to eliminate features having higher value. **We stop once the VIF value is less than 3 and p value less than 0.01**.

$$\frac{1}{e^{-1.0960 - 1.4806 * \text{Do Not Email} + 1.0643 * \text{Total Time Spent on Website} - 0.2735 * \text{Lead Origin}_{\text{Landing}}\text{Page Submission} + 3.6710 * \text{Lead Origin}_{\text{Lead}}\text{Add Form} + 0.8153 * \text{Lead Source}_{\text{Internal}}\text{Sources} - 1.4425 * \text{Last Activity}_{\text{Chat}}\text{Interaction} - 0.9807 * \text{Last Activity}_{\text{Lead}}\text{Generation} + 0.6125 * \text{Last Activity}_{\text{Other}} + 2.6150 * \text{What is your current occupation\_Working Professional} + 1.5933 * \text{Last Notable Activity\_Phone or SMS}}}$$

The prediction on the test dataset was done with **probability 0.5** for which the **accuracy** was **nearly 81%.** Then we did a **ROC** to check the trade off between specificity and sensitivity. Which was around **87%** indicating the model build to be good.

Then we tried for different probability and found that **0.375** would be more appropriate probability to balance between **sensitivity and specificity and accuracy**.

	Train set prediction	Test set prediction
Accuracy	80%	81%
Sensitivity	81%	84%
specificity	79%	79%

We then tried with **precision and recall** for which the probability was calculated **as 0.4**

	Train set prediction	Test set prediction
Accuracy	80%	81%
Precision	73%	74%
Recall	76%	80%

Even though Specificity-Sensitivity and precision-recall give same accuracy precision-recall is more suitable because,

. Precision is crucial when it's important to minimize false positives, i.e., when you want to ensure that most of the predicted leads marked as "converted" are indeed customers.

. Recall is important when missing potential customers (false negatives) is costly, i.e., when you want to catch as many true conversions as possible, even at the risk of some false positives