SUMMARY – LEAD SCORING CASE STUDY IIIT-B

BY ANANTAKUMAAR VR & SAMBRIT SAHA

PROBLEM STATEMENT: -

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- 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.

SUMMARY: -

1. **Data Cleaning:**
   - First step to clean the dataset we choose to remove the redundant variables/features.
   - The data set was partially clean except for a few null values and the option 'Select' has to replace with a null value since it did not give us much information.
   - Dropped the high percentage of Null values more than 40%.
   - Checked for number of unique Categories for all Categorical columns.
   - From that Identified the Highly skewed columns and dropped them.
   - Treated the missing values by imputing the favourable aggregate function like (Mean, Median, and Mode).
   - Detected the Outliers.
2. **Exploratory Data Analysis:**
   - A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values are good but outliers are present.
   - Performed Univariate Analysis for both Continuous and Categorical variables.
   - Performed Bivariate Analysis with respect to Target variable.
3. **Dummy Variables:**
   - The dummy variables are created for all the categorical columns.
4. **Scaling:**
   - Used Standard scalar to scale the data for Continuous variables.
5. **Train-Test Split:**
   - The Spit was done at 70% and 30% for train and test the data respectively.

6. **Model Building:**
   - By using Recursive Feature Elimination. It gives the relevant variables. Later the irrelevant features were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value 0.05 were kept).
7. **Model Evaluation:**
   - A confusion matrix was made. Later on, the optimum cut-off value by using ROC curve was used to find the accuracy, sensitivity and specificity which came to be around 80%.
8. **Prediction:**
   - Prediction was done on the test data frame an optimum cut-off as 0.35 with accuracy, sensitivity and Specificity of almost 80%.
9. **Precision-Recall:**
   - The method was also used to recheck and a value of 0.35 was taken as final cut-off for Precision-Recall Curve.
10. **Conclusion:**

    The variables that important the most in the potential buyers are as follows:
    - The total time spent on the Website.
    - When the lead source was: - Olark Chat, Google and Direct Traffic
    - When the last activity was: - SMS and Olark chat conversation
    - When Lead origin is Lead Import.