**Summary Report – Lead Scoring Case Study**

**Problem Statement :**

X Education sells online courses to industry professionals. Company markets its product on multiple websites and search engines such as google. once people landed on website, they may fill up form for enquiry, they may watch videos or they may browse the courses on website. if these people filling form with their Email ID and Phone numbers they are considered as lead also the leads can be taken from past referrals.

Once the leads are acquired company employees start making calls and Email to these people. Through this process leads get converted and while most do not. The current lead conversion rate for X education is 30% of acquired leads. Now company wants to focus on only those potential lead which can be converted so the employees can invest time to nurture these leads only. Hence company wants us to assign score on scale of 1-100 to each lead basis on information collected and on basis of this score predict whether lead will get converted or not.

**Step by Step approach for solution :**

Step 1 : Reading and understanding the given dataset.

Read and understand the data set by inspection.

**Step 2 : Data Preparation**

1. As per understanding taken from problem statement, replacing 'Select' with 'NAN' as 'Select' showing only when people are selecting nothing for that option
2. Missing value treatment: Dropping column which are having more than 30 % null values.
3. Dropping variable which are ‘zero variance variable’ mean variable means variable which contain very little information and mostly consist of a single value.
4. Imputation of data: imputing median and mode for numerical and categorical variable respectively where null value percentage is less than 30 %

**Step 3 : Visualization**

1. Checked data imbalanced of target variable.
2. Univariate analysis: Checked count distribution of categorical variable using hist plot and taken inferences from same. Also, checked distribution of numerical variable using boxplot.
3. Bivariate analysis: Checked distribution of data in numerical variables with target variable using boxplot.
4. Multivariate analysis: Compared numerical variable with target variable using heatmap

**Step 4 : Transformation of data:**

1. Changing binary variable from ‘Yes’ & ‘No’ to 0 & 1
2. Dummy variable creation: Created dummy variables for all categorical variable and dropped repeated and redundant variables.
3. Outlier treatment: Removed statistical outliers from numerical variable using capping.

**Step 5 : Train- Test Split**

1. Split the data set into train and test data set with proportion of 70 and 30 respectively.
2. Separated dependant and independent variable into `X’ and ‘y’

**Step 6 : Standardization of data**

Converted all variable of train and test data set to same scale using standardization method

**Step 7 : Looking at correlation**

Checked correlation between all variable and dropped variables which are highly correlated to each other’s.

**Step 8 : Feature selection using RFE**

Using recursive feature elimination, we selected top 15 variables which are highly significant.

**Step 9 : Model building**

1. Added constant to train and test data set using stats model library.
2. Created generalize linear model and using the statistical summary, recursively looking at p-values of all variables and dropping variables which shows insignificancy by having p-value greater than threshold value (i.e.0.05)
3. Finally, we arrived at 11 most significant variables which are having p-value and VIF within threshold limit.

**Step 10 : Evaluation**

1. Prediction made on test data set using final model
2. Calculated the sensitivity and specificity from confusion metrics.

**Step 11 : Plotting ROC curve and Finding optimal cut-off**

1. Plotted ROC curve to find out Area under the curve.
2. Made prediction for 0.1 to 0.9 probability and checked for closer accuracy, sensitivity and specificity on these probabilities to decide optimal cut off probability.
3. Plot trade-off between probabilities of accuracy, sensitivity and specificity to get optimal cut off probability.

**Step 12 : Making predictions on test data set**

1. Made predictions on test data set using optimal cut off probability.
2. Assigned lead score on basis of Conversion probability of customer.

**Step 13 : Evaluation on test data set**

Calculated Accuracy, Sensitivity and Specificity on Test data set which is 0.78 , 0.83 and 0.75 respectively.

**Conclusion :**

1. The Conversion rate calculated on basis of lead score on final prediction is 83 % and which meets the ballpark set by client CEO that conversion rate should be around 80 %
2. Good percentage of sensitivity shows that our model is good enough to predict most promising leads
3. Below are the top 3 features which are contributing more to get probability of lead conversion.

* Lead source
* Last activity
* Total time spent on website