LEAD SCORE CASE STUDY SUMMARY

Summary contains the analysis about the X Education company which is an online course seller. Company needs to know the factors which can identify the potential buyers, so that sales team can target them. All the information related to the past buyers and their conversion is provided.

Followed below steps to build efficient model:

Data Cleaning: There were many SELECT values which is considered as missing. If the missing values counts is more than 40% in a column, then drop the column. Dropping columns with unique value like ProspectID and leadID. Dropping columns with only one value (ex-Magazine). Imputing missing values with mode for categorical data and median for continuous data.

EDA: Created the list for categorical and continuous data type separately for easy graph plotting and analysis. There were many columns having 99% of single value, dropped them (ex-Do not call). Outliers were not found.

Dummy variables: For logistic regression model all categorical data are split into dummy variables to change the values to numeric form (1,0).

Train –Test data: Data is split into train and test dataset by 70:30 ratio. Train and test data are normalised using standard scaler.

Feature selection and Model building: Using hybrid approach to select feature. RFE is used to select 15 features and then manually features are dropped by analysing p-value (less than 0.05) and VIF (less than 5). Finally, 11 features are obtained in Model5.

Final Features are:

1. Lead Origin_Lead Add Form
2. Do Not Email_Yes
3. Last Activity_Olark Chat Conversation
4. What is your current occupation_Working Professional
5. Tags_Busy
6. Tags_Closed by Horizzon
7. Tags_Lost to EINS
8. Tags_Ringing
9. Tags_Will revert after reading the email
10. Tags_in touch with EINS
11. Last Notable Activity_SMS Sent

Model evaluation: Predicting from train data and evaluating Accuracy, Recall and Precision. Plotting ROC curve to check the model. Checked the metrices at cutoff value 0.35 and 0.38. We got Optimum cutoff at 0.35. As stated, the TPR is maintained above 80%

Prediction on test set: Normalising data in test set. Predicting from test data and evaluating Accuracy, Recall and Precision. On comparing the metrics from test and train data set were almost same depicting the model efficiency.

Key learnings to identify the hot leads are:

- When the customer is a working professional, it has high chance of conversion.
- When the Lead origin is Lead add form.
- Last activity is identified as Olark Chat Conversation and SMS sent.
- When customer has permitted for email.
- When customer is tagged as ‘lost to EINS’, ‘closed by horizzon’,’Will revert after reading mail’, ‘In touch with EINS’.
- Through EDA we can see the ‘time spent in websites’, ‘total visits’, ‘lead source as Google’ seems to give fruitful result.

X Education company can refer to above factors and go for potential buyers.