Lead Scoring Case Study Summary

Problem statement: -

X Education sells online courses. Leads generated from various sources are captured. There is other metadata around lead that is captured for each lead. Team is assigned to nurture hot leads and convert such potential leads to confirmed opportunity (leads).

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution:

Effective way of working on the leads is to start with hot leads i.e. leads that have higher probability of getting converted. This will not only result in higher conversion ratio but also effective use of time. Time spent on nurturing hot leads can be increased whereas time spent on leads with low score (cold leads) can be minimized.

Determining hot and cold leads can be done by using a logistic regression model. Using the meta data provided for each lead, we will build a logistic regression model and assign lead score to each lead.

1. **Data Analysis**-

    a. There are columns with higher missing % in the data. Also, there are column where default value “Select” is populated. We will be initially considering this as missing values and apply same missing value treatment for such values.
    b. The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘not provided’ so as not to lose much data. Although they were later removed while making dummies. The treatment of null values and their percentages were also derived.
    c. Columns with too many NA values cannot be imputed and so it’s better to drop them and for those that follow a standard distribution will be imputed with median.
    d. Few columns have less than 2% NA values. We can afford to drop their respective rows altogether.

e. There are too many NA values with no logical way to impute in these columns so we will drop them entirely
f. Statistical analysis indicated that there isn't significant difference between median and mean for these columns and hence imputing with median should not create issues.
g. Other missing values will be treated as missing values since imputing might exaggerate the data.

2. **Data preparation-**

a. Boxplot and descriptive statistics indicate that there are outliers in the dataset. About 98% of the data has been retained after data cleaning.
b. We will not remove the outliers as this will help us in assigning lead score to all leads. Final review of model does indicate that metric (Accuracy, Sensitivity, Specificity) is good and hence we will not remove outliers.
c. Quick bivariate and univariate analysis indicates few categorical variables/numerical variables levels are critical for lead conversion. We will use this for conclusion.
d. Since logistic regression uses numerical data, we will convert categorical data using below technique.
    i. Dummy Variables – Categorical variable with low/moderate level will be treated using dummy variables.
    ii. The dummy variables were created and later the were encoded for the categorical scale.
    iii. Label Encoding – We will use label encoding for variables with higher level. This is to avoid drastic increase in data frame size.
e. Columns with no variance i.e., columns with single constant value will be dropped since they add no information/dimension for model building.
f. Quick heatmap indicates some correlation between variables. VIF will be further used during model building.

3. Model Building-

a. Split data into train and test dataset. We will train the train dataset and make predictions on test data set.
b. Train-Test split: The split was done at 70% and 30% for train and test data respectively.
c. Numerical data will be scaled using standard scaler to ensure data is on same scale and computationally efficient.
d. MinMaxScaler has been used for numeric variables
e. RFE–
    i. We will use 15 columns supported by RFE for regression
    ii. Build a Logistic Regression model using SKLearn for RFE
    iii. Criteria used for tuning the model (i.e. dropping variables).
        1. High p-value (variable not significant).
        2. High VIF (high collinearity with another variable).

f. ROC and AUC confirms that we have a decent model.
g. We will use below technique to final optimal cutoff value
    i. Plot accuracy, sensitivity, specificity
    ii. Plot recall, precision.
    iii. Since there is a tradeoff between sensitivity vs specificity it is important to find optimal cutoff.

4. **Model Selection and Lead Score**-

    a. Use the model build using the RFE technique for final prediction.
    b. This model gives best score and is easy to make suggestion and interpretations.
    c. We will assign Lead score to each lead using probability predicted by model (Lead Score = Predicted Probability * 100)
    d. Create a data frame to and plot conversion vs cutoff.
5. **Model Evaluation**:
    a. Finally, we have an overall accuracy of approx. 0.84 on our Logistic Regression model. That is, there is around 84.5% chance that our predicted leads will be converted. This meets the CEO's target of at least 80% lead conversion.

6. Conclusion

a. Following three variables are contributing the most towards the probability of a lead getting converted:
    i. Page_views_per
    ii. Lead_quality_not_sure
    iii. Total_time_spent

b. Again, based on the coefficient values, the following are the top three categorical/dummy variables that should be focused the most in order to increase the probability of lead conversion :
    i. lead_source_reference
    ii. lead_source_welingak_website
    iii. last_notable_activity_had_a_phone_conversation

7. Learnings

a. EDA is extremely important step prior to building model. Key insights from EDA helps in treating the data correctly.
b. 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 seems good and no outliers were found.
c. Data cleaning helps in building efficient model. Steps like missing value imputation, scaling, outlier treatment must be performed at minimum to ensure quality of data is not compromised.
    i. i. Missing value – Columns with higher percentage of missing value can be dropped whereas columns with lesser percentage can be imputed.
    ii. Outlier treatment – Outlier can impact model and result in less effective model. Hence care should be taken to treat outlier effectively. Care should also be taken to ensure that this does not result in significant loss of data.
d. RFE is efficient technique to identify key features to start building model.
e. Functions to perform repetitive steps can help in building a modular code. This also help in reusability of the code.
f. Understanding trade-off between sensitivity and specificity is key in determining ideal optimal cutoff for the mode.
g. Confusion metrics is good indicator to determine how model performs. Accuracy, sensitivity, specificity can be derived from confusion metrics.

Thus, this analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
4. When the last activity was:
    a. SMS
    b. Olark chat conversation
    5. When the lead origin is Lead add format.
    6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.