Lead Score Case Study

Case Study Objective:

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%.

The goal of the case study is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

APPROACH AND LEARNINGS FROM CASE STUDY:

1. **Data Collection and Data Cleaning :**
    a) Importing the data then cleaning it, checking if there are null values.
    b) The presences of "Select" value in many of the categorical columns are handled by converting them to "NaN" (null) values.
    c) Removed columns having more than 45% null values.
    d) Rest of the missing values have imputed with the maximum items in the columns.

2. **EDA (Data Visualizations) :**
    We begin by studying the dataset provided and do exploratory data analysis using statistical and visualization methods to both the

continuous and categorical data. The dataset has a larger section of categorical data which points us to apply more of a Logistic Regression Model post our Exploratory Data Analysis(EDA).

3. **<u>Dummy Variables :</u>**
   The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

4. **<u>Train-Test split :</u>**
   The split was done at 70% and 30% for train and test data respectively. Scaling will be done with the Standard Scaler.

5. **<u>Model Building :</u>**
   Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value(The variables with VIF< 5 and p-value< 0.05 were kept).

6. **<u>Model Evaluation :</u>**
   A confusion matrix was made. Later on the optimum cut off value(using ROC was sued to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. **<u>Prediction :</u>**
   Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. **<u>Precision- Recall :</u>**
   This method was also used to recheck and a cut off of 0.41 was found with precision around 73% and recall around 75% on the test data frame.

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. 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.