LEAD SCORING CASE STUDY

PRESENTED BY:

SUDHANSHU DIXIT

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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

There are quite a few goals for this case study.

•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.

•There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

APPROACH:

1. Basic Data Check
2. Data Cleaning
3. Exploratory Data Analysis
4. Dummy Variables, Train-Test Split and Feature Scaling
5. Model Building
6. Prediction and Precision Recall

Basic Data Check and Data Cleaning

1. Loading the data in the python notebook and doing basic data checks such as shape of data, data types of the columns, null values check, percentage of null values in each column etc.
2. After that I started with data cleaning. But before that I changed all ‘Select’ options in each column to ‘NaN’ as suggested in the business problem.
3. I then removed all the columns that had null values > 40%.

Exploratory Data Analysis

1. I started with univariate analysis for numerical features that included removing outliers as they could skew the results.
2. Post univariate analysis I then went to bivariate analysis with the target variable ‘Converted’ using correlation in order to find strong positive correlation to know the feature that impacted the target variables the most.
3. After numerical columns, I then went to categorical variables where I removed various columns that had data imbalance.
4. Null values in the categorical feature columns were imputed with either new category such as ‘Others’ or ‘Not specified’ or with highest occurring category feature in the respective column. These imputations were done keeping various factors in mind.
5. Following this, many categories of the categorical features that had very low count were merged with others or a new category was created in order to replace all these new categories into 1.
   REASON: This was done because, firstly they were less impactful due to very negligible count and secondly, they would **have led to creatin of lot of additional columns when creating dummy variables.**

Dummy Variables, Train-Test Split and Feature Scaling

1. I then created dummy variables for all categorical columns.
2. I then split the data in to train data that was 70% of the data and test data that comprised of 30% of the data.
3. For the training data, I then used StandardScaler (standard scaler) for feature scaling of numerical features/columns.

Model Building, Prediction and Precision Recall

1. I then built the 1st Logistic regression model and evaluated it based on p-value. I then deleted columns that had higher p-values and re-built the model.
2. Once all p-values were under a specified range of hypothesis, I then checked VIF or variation inflation factor. VIF number if greater than 5 was noted for any column then that column was dropped and 1st and 2nd step was repeated.
3. 2nd step is repeated till VIF is all under a specified range.
4. Prediction was the done with training data and confusion matrix and other metrics such as accuracy of model, sensitivity and specificity was noted.
5. I then optimum cut off value using ROC curve was used.

Thank you!