Lead Scoring Case Study

Summary Report

Submitted By:

Vaishali Papneja

Sushasree Vasudevan Suseel Kumar

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%.

Objective:

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%.

Approached Methodology:

1. **Data Understanding and Data Cleaning:**
    - Once we imported the data, we realized that there were many columns which has the value as 'Select'. As 'Select' value would have been given by default to any field if customer doesn't select anything. So, we converted all such values to NULL.
    - **Missing Values Check:** Then we dropped all such columns which had >70% missing data. Then we imputed rest of the missing values with the help of other features.
    - **Duplicate Data Check:** We realized that as we have Prospect ID and Lead Number present which is unique to each customer. So, we dropped Prospect ID and made Lead Number as index of our dataset. After doing so, we got duplicate data in our dataset, and then we dropped all duplicated records.
    - **Features with No/Very less variation:** As these columns cannot add any benefit to our analysis, therefore we dropped them.
    - **Outlier Treatment:** we also found that variable such that Total Visits and Page Views per Visits have outliers in them. So, we treated them using IQR method.
2. **Data Visualization:**

- In this step, we created few derived metrics for continuous variables and studied patterns followed by customers who are converted or not converted.
- Result of our analysis:
    - *Conversion Ratio in our dataset currently is 40%.*
    - *Customers with High Lead Quality tend to convert into our hot lead and customers with worst quality almost never a potential customer. So, we should not waste our time dealing with worst quality customers. Instead should focus on other customers.*
    - *If lead source is from "**'WeLearn' or 'Welingak Website'**, there are highest chances of it becoming our lead.*
    - *Conversion Rate is highest if number of visits, a customer make to website is more.*
    - *Conversion rate is high if customer want to be notified via E-mails.*

3. **Data Preparation for Modelling:**
    - **Feature Scaling:**
        - First, for all – yes/no feature variables, we encoded them using OneHotEncodig.
        - For rest of the categorical variables, we created dummies.
        - Then we scaled our numerical features.
    - Then, we plotted a heatmap and removed all redundant columns.
    - **Train-Test Split:** Then we divided our dataset into train and test datasets with Split Ratio 70:30.
4. **Modelling:**
    - Here, we selected top 10 features using RFE and created GLM models until we find an optimum model (with features having pValues < 0.05 and VIF<5)
    - We found below important features for lead conversion:
        - ***Last Notable Activity_SMS Sent (Positively Impacted)***
        - ***Tags_Will revert after reading the email (Positively Impacted)***
        - ***Total Time Spent on Website (Positively Impacted)***
        - ***Tags_Ringing (Negatively Impacted)***
        - ***Tags_switched off (Negatively Impacted)***
        - ***Lead Quality_Worst (Negatively Impacted)***
        - ***Tags_Closed by Horizzon (Positively Impacted)***
        - ***Tags_Lost to EINS (Positively Impacted)***
5. **Calculating Lead Score:** Later we assigned lead score to each customer. To get this, we multiplied their conversion probability predicted by model with 100.
6. **Conversion Rate Cut-off:** We found that after keeping cut-off as 5% , we can achieve our target of 80% conversion rate.