Data Analysis – Lead Scoring Project

1. Introduction

Initially the project had 36 features. However after performing dummy, the project had 87 features. The aim is to reduce the features and take only the ones that are important for building a model. Too many features can impact the accuracy of the model.
Secondly, there were too many columns that had one data point in majority and had to be dropped. The data comprised of the word "Select' which simply meant missing values and had to be categorized as missing and treated as a separate category to prevent data loss.

The correlations have been extracted between different features with correlation martic and corr() method.

Methods such as RFE and VIF has been used to select the topmost features.

The top 4 variables that contribute the most towards the probability of a lead getting converted are:

- Lead source
- What is your current occupation?
- Do Not email
- Last activity

2. MethodsProblem Definition

The purpose is to extract high conversion leads. When people fill up a form providing their email address or phone number, including referrals, they are classified to be a lead. Most leads are lost during the nurturing process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%, which the company wants to increase. The company aims to improve their lead nurturing process and identify hot leads i.e. those leads which are likely to convert.

3. Experimental Evaluation

3.1 Methodology

- Missing data is categorized in to type 1(dropped columns), type 2 (categorizing as Missing) and type 3(removing with rows)treated accordingly. A threshold of 40% i.e. 3600 is applied and columns are deleted. The others with above 2% missing values are categorized as 'Missing' label category. The columns with majority of only 1 data int value are removed as well as those columns which are not required are removed.
- Used min-max scalar method to transform the data and remove outliers and skewness.
- Used the RFE and VIF feature reduction method.
- After treating the data, there are only 12 features which were than increased to 87 columns with dummy features.
- Used the confusion matrix, to find out the precision, sensitivity, accuracy, etc

- 0.38 was the cut-off point found for the threshold from the ROC curve to categorize data.

3.2 Results

- we can see more converted leads as the total time spent on the website increases (above 600 minutes for converted leads). They typically have 1 to 6 page views per visit on the website

- Most converted leads come from people who have opted out of email than the ones who have opted for it. The majority of the people opt out of email, hence it's not advisable to focus much on emails.

- Landing Page Submission has the highest conversion rate and the lead ad form is the lowest conversion. Interns can focus on making landing page forms easy and API-based conversions and work on improving lead ad form conversions

- There are more conversions for Total time spent on the website than non-converted leads. Around 1300 conversions and 400 non converted leads

- The highest converted leads and total visits for Google, Direct Traffic and Reference. Reference has the most converted leads and negligible non converted leads out of all the categories. We can see the highest non-converted leads and total visits for Direct Traffic.

- Customized Communication Scripts: Provide interns with customized communication scripts for different segments of high-probability leads. Tailor the messaging based on the lead's characteristics and behavior, such as occupation, etc. Most converted leads are unemployed and the motive behind taking up a course is job prospects as shown in graphs and data visualization below.

- As we can see most of the leads(converted and non-converted have less than 40 total visits on the website. Most of the leads have less than 20 total visits. We can see that most converted leads have less than 20 Total visits and spend more than 600 minutes on the website.

- Spending time on freebies, such as infographics or ePDFs should be productive. People are currently not getting converted with infographics or ebooks . Either one has to change the topic to a more relevant one or find another way to convert

- Most converted leads come from people who have opted out of email than the ones who have opted for it. The majority of the people opt out of email, hence it's not advisable to focus much on emails.

- Landing Page Submission has the highest conversion rate and the lead ad form is the lowest conversion. One can focus on making landing page forms easy and API-based conversions and work on improving lead ad form conversions

- There are more conversions for Total time spent on the website than non-converted leads. Around 1300 conversions and 400 non-converted leads

- As we can see, the highest converted leads and total visits for Google, Direct Traffic and Reference. Reference has the most converted leads and negligible non converted leads out of all the categories. We can see the highest non-converted leads and total visits for Direct Traffic. The interns should focus on Google, Direct Traffic and Reference as lead sources and try to nurture them into the funnel.

- One should focus on SMS as most converted leads are active from SMS ad the most non-converted ones from email.

- Customized Communication Scripts: Provide interns with customized communication scripts for different segments of high-probability leads. Tailor the messaging based on the lead's characteristics and behavior, such as occupation, etc. Most converted leads are unemployed and the motive behind taking up a course is job prospects as shown in graphs and data visualization below.

- Spending time on freebies, such as infographics or ePDFs should be productive. People are currently not getting converted with infographics or ebooks . Either one has to change the topic to a more relevant one or find another way to convert

- we can see more converted leads as the total time spent on the website increases (above 600 minutes for converted leads). They typically have 1 to 6 page views per visit on the website

Insights from the correlation matrix:

- There is a high correlation between the Lead origin as lead import and the lead source as Facebook.

- There is a high correlation between the Last Activity of the users which was **Unsubscribed** and the **lead source** which was **Facebook**. It means most people who are unsubscribing are from Facebook. Interns need to focus on Facebook advertising with customized ads and specialized content.

- High correlation between the lead origin in **Ad form category and lead source**, reference category. It simply means the converted leads from **ad form** are **references**.

- High correlation between **Last Activity** column in the **Email Opened** category and **Last Notable Activity** column in the **Email Opened** category. It simply means the customers whose last activity was checking email were most likely to check emails or be active on email as compared to other platforms.

- High correlation between the **Last Activity** column in the **SMS Sent** category and **Last Notable Activity** column in the **SMS Sent category**. It simply means the customers whose last activity was SMS Sent were most likely to check SMS or be active on SMS as compared to other platforms.

- High correlation between **Last Activity** column in the **Email opened** category and Last Activity in the **Email Link Clicked** category. It simply means the customers whose last activity was **Email Link Clicked** were most likely to check emails or be active on emails as compared to other platforms.

- High correlation between the **Last Activity** column in the **Had a Phone Conversation** category and the **Last Notable Activity** column in the **Had a Phone Conversation** category. It simply means the customers whose last activity was **having a phone conversation** were most likely to interact on calls as compared to other platforms.

- High correlation between the **Last Activity** column in the **Email Received** and the **Last Notable Activity** column in the **Email Received** category. It simply means the customers whose last activity was **receiving an email** were most likely to interact on emails as compared to other platforms.

Test Data

Sensitivity	80%
Specificity	82%
False Positive	18%
Positive predictive value	71%
Negative Predictive value	87%
Accuracy	80.08%

Train Data:

Accuracy is 81.04%

Sensitivity	78%
Specificity	83%
False Positive	16%
Positive predictive value	74%
Negative Predictive value	86%
Accuracy	81.04%