Lead Scoring Case Study - Summary

Submitted By:
Kavita Vishwakarma
Kirthiga Rajenderan
Kartik Ukhalkar

The provided '**Leads**' dataset was first inspected and the meaning of the variables were analyzed from the '**Leads Data Dictionary**'. Once the dataset was inspected, data cleaning was done. The '**Select**' values in different columns were first **converted to null values**. This was done because the select values seem to be default values in the drop down menu of the company's website. During the process of data cleaning, missing values were taken care of by imputing with suitable values (*median for numeric value and mode for categorical value*) or were dropped if null values were around **40%** or more. Indexes were then reset after the cleaning was completed.

After the data cleaning process was complete **Univariate** and **Multivariate analysis** were carried out. It was found that the lead **conversion rate** was about **38.54%**. The **country** and **city** with **highest number** of leads and leads conversion were found to be **India** and **Mumbai** respectively. The **specialization** with the highest number of leads and leads conversion is '**Finance Management**'. Highest number of lead conversions was found to be having **Google** as a **lead source**. The **occupation** with the highest number of leads and leads conversion is '**unemployed**'.

While plotting **correlation** between the numeric variables it was found that conversion (Converted) is **correlated** with **Total website visits**, **Total time spent on the website**, and **inversely correlated** with **Page views per visit**.

After EDA was completed, categorical values were converted to **dummy variables** and the dataset was split into a train and test set (**70%** and **30%** respectively). The numeric variables were scaled using **StandardScaler**.

The first model was built using all the variables but most variables/features had high **p-value** which meant less significance. Therefore, in the second model features were selected in using **RFE** and in the third model (final model) the features selected by the RFE model were used but the features selected by RFE with high p-value were eliminated.

Using the final model, conversion **predictions** were made on the final train set and '**Lead Score**' were assigned. The **accuracy** of the model is found to be around 89%. After creating a confusion matrix, **sensitivity (86%)**, **specificity (91%)**, **precision (86%)**, and **recall (86%)** were calculated. After plotting **ROC curve** and precision and recall tradeoff, 0.3 was selected as the **optimal cutoff**.

After the final analysis, the following can be **recommended** to the company:

- The leads having high 'Lead Score' can be focused on more for better conversion rate.
- Marketing on Google, since the conversion rate from the traffic from Google is high.
- Encouraging existing converted leads for referrals by providing some incentives for the referrals.
- Since the number of leads is high in Mumbai as compared to other major cities, the company can increase marketing in the other cities as well to achieve more leads.
- The unemployed category can be focused on more and also individuals having Finance Management as specialization.
- Focus on the students can be minimized since the conversion rate is significantly low.