Summary – Lead Scoring Case Study

Contributors: Darshith P K, Nimisha Kashyap, Sakshi Priyadarshi

The goal of the case study was 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.

Steps involved were:

- Data cleaning and manipulation processes which included :
    - Identification and resolution of duplicate data.
    - Handling of NA values and missing data.
    - Elimination of columns containing a significant amount of missing values and deemed irrelevant for analysis.
    - Imputation of missing values if required.
    - Identification and handling of outliers in the dataset.
- Perform Exploratory Data Analysis (EDA) which included:
    - Univariate analysis such as value counts and variable distributions.
    - Bivariate analysis including correlation coefficient calculations and examination of patterns between variables.
    - Implement feature scaling, creation of dummy variables, and encoding of the data for further analysis.
    - Utilize classification techniques, particularly logistic regression, for model development and prediction.
    - Validate the model to ensure its reliability and accuracy.
    - Present the finalized model along with relevant insights.
    - Draw conclusions and provide recommendations based on the analysis conducted.

Data Manipulation:
The dataset provided had a total of 37 rows 9240 columns. Features having single values were excluded (please refer to .ipynb file and ppt for the same), unnecessary columns were dropped. Upon further examining, it was observed that value counts for some object type variables lacked sufficient variability and were thus dropped. Columns with more than 35% missing values were dropped.
Normalization was applied to the numerical variables and object types variables were encoded into dummy variables.

Model Building:
The data was divided into training and testing sets. The initial step involved traintest split at a ratio of 70:30. RFE was used for feature selection. The model was constructed by excluding variables having pvalue more than 0.05 and VIF greater that 5.

Summary:

There were lot of leads generated initially but only few of them came out as paying customers later. In the middle stage, we need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get the higher lead conversion. First, we need to sort out the best prospects from the leads generated. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' contributes most towards the probability of a lead getting converted. After that, we must keep a list of leads handy so that they can be informed about new courses, services, job offers and further higher studies. Monitor each lead carefully so that we can tailor the information sent to them. Carefully providing information about the courses, jobs in the market, how the courses will help them in landing their dream job, or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects. Focus on converted leads. Hold question-answer sessions with leads to extract the right information. Make further enquiries and appointments with the leads to determine their intention to join the courses.