Brief Summary of Lead Scoring Case Study

X – Education Lead scoring case study looks at building a logistic regression model where basis the data provided in the form of dependent variables a binary output is defined. The logistic model itself will provide the probability value which basis the business scenario needs to be converted to appropriate binary value. With the focus of increasing the leads to give access to potential leads to X-Education company following activities are carried out,

- Data Analysis
  In this step the focus was to understand the types of column and the data itself. This step is carried out by analysing the information of the dataset and the data dictionary provided
- Data Cleaning
  As part of this step handling the data which was not in correct format like,
  - Removing 01.,02. & 03. from Asymmetrique Activity Index
  - Removing columns which has no variation of data as these would not provide any value for the model
  - Removing few columns like City & Country as these are not relevant from the business point of view considering their USP is providing online content irrespective of the location
  - Removing columns with very high (more than 30%) null values
  - Removing columns like tags as these are highly subjective and has high null values. Also this column can be better defined by Lead Quality and Lead Profile
- Outlier Analysis
  Outlier analysis focus on removing numeric value which are unusual to the data distribution
- Creating Dummy variables
  Converting categorical variables to numeric variables by appropriate measures either using Binary mapping or Dummy variables
- Scaling of the Data
  The numeric variable are brought to the same base using Standardisation
- Building Model
  Building a logistic regression model using RFE by selecting 20 features initially so that the company can have sufficient parameter with which it can work to ensure that they can convert maximum leads
  Post building the first model eliminating features that have high p-values (above 0.05) first one-by-one. Post that checking if there are any features having high VIF values (5 or above). After satisfying the above criteria using the finalised model to build the conversion probability. Calculating the optimum value where there is a trade-off between Accuracy, Specificity & Sensitivity so that an optimum value is selected without affecting the True positives or True Negatives. This would ensure allowing the company focus on the right dataset for lead conversion.