SUMMARY REPORT LEAD SCORING CASE STUDY

1. EDA
2. CLEANING DATA, HANDLING MISSING VALUES
3. FEATURE SELECTION
4. DRAWING INFERENCES
5. MODEL TUNING
6. ASSIGNING LEAD SCORES

The Case study is divided among 5 sub sections where the first section starts at EDA. We started with the basic check of null values, standard deviations of various numerical features to get an intuition on the distribution of the data, also to get an idea on the presence of outliers. From the Inferences drawn on EDA, we concluded that the data distribution between the categories of the target variable is imbalanced, to some extent where the ratio of the categories was 3:2 that is 60% not converted leads and 40%converted leads. We then compared the conversion percentage among all the different features and multiple sub categories. This helped us to draw insights on how the market looks like, which are target consumers should be, which Specialization or country to focus on, what are the indicators that the lead is getting converted or not. Our EDA ended with all the possible univariate, bivariate and multivariate analysis.

We had certain columns which were having null values shown as select, now select was possibly the category which was coming first in the drop down and that could be the reason that leads which didn't fill it by default the values got replaced by select automatically. Later we replaced those values by null values and did a mode imputation. There were columns like Asymmetric which are not created during the leads generation but instead are created possibly after presales or during lead evaluation, we dropped such metrics since it creates a bias in our model, due to the reason that these categories are created after lead is contacted or evaluated but our main business objective is to assess the

lead quality before any point of contact is made or lead is allotted to different sales executive. Hence, we dropped such columns, also we dropped columns having null values greater than 40%.

For feature selection we did it in two parts, for some of the features we are having category which are showing very less frequency of the data, in that case we merged some of the categories which were showing less than 1 or 2% frequency into one single category to reduce the overall dimension and also to not have any bias. We also used WOE and IV for all the categorical variables to see the importance of the variables and do further selection, also we combined it with RFE to further reduce the dimension.

Further Model tunning was performed using VIF, p value etc, Also Metrics like AUC, ROC, sensitivity, specificity etc was taken into account to select the optimum probability.

Using a simple multiplication by 100 with the probability and converting the fractions into discrete integers. We got discrete scores between 0 to 100.

Through this case study, we gained the knowledge on the Sales domain or sector of any corporate industry, what are the metrics which influence the customers, also from technical perspective we got an exposure on how to perform various tasks related to data mining, feature selection, engineering, cleaning, modelling and optimising.