Summary Report – Lead Scoring Case Study

Problem Statement:

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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Solution:

1. Lead scoring case study has been done using logistic regression model to find the underlying factors that would convert the leads to be successful at higher rate.
2. Highest numbers of leads are found to be in India, with Mumbai being the highest with leads.
3. After taking a look at the data, we perform EDA to clean the data for building the model.
4. There are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. To get some useful data we have to make compulsory selection. Likewise, Customer occupation, Specialization, etc.
5. We also drop a few fields because of high presence of null values and they won't be of any use for the analysis.
6. The high number of total visits & Total time spent on platform may increasing the chances of lead to be converted.
7. The leads are joined course for Better Career Prospects, most of having Specialization from Finance Management. Leads from HR, Finance & marketing management specializations are high probability to convert.
8. Talking to last notable Activity, making improvement in customer engagement through email & calls will help to convert leads. As the leads which are opening email have high probability to convert, Same as Sending SMS will also benefit.
9. Most of leads current occupation is Unemployed, which means gave more focus on unemployed leads.

Learnings:

1. The process of performing EDA is very crucial step in model building as the insights from it helps in handling the data correct.
2. Data cleaning, a part of EDA is another very important step to make the data fit for analysis, like removing/retreating the null values, dropping unnecessary columns, Scaling.
3. RFE is an efficient technique to identify the key features to start building the model.
4. Functions to perform repetitive steps can help in building a modular code. This also help in reusability of the code.
5. Understanding trade-off between sensitivity and specificity is key in determining ideal optimal cutoff for the mode.
6. Confusion metrics is good indicator to determine how model performs. Accuracy, sensitivity, specificity can be derived from confusion metrics