Summary Report

> *By Isaac Martin, Jai Ganesh G & Jaydeep Gokhale*

The objective of the Case study was to find the Lead Score indicating, how important a particular lead is for the company X Education. Using this score generated by our logistic model, the company wanted to contact those leads which has a high chance/probability of getting converted into a student.

The process that was followed in order to achieve good results:

1. Reading and understanding data:
    - In order to get accustomed with data, we went through the data dictionary to completely understand data.
    - We decided to drop score variables as they are present while prediction in real time.
    - It also included analysis of some variables, which resulted in dropping of Skewed variables (the variables which won't contribute to Modelling).
    - Replacing the "Select" data with np.nan
2. Missing Value Analysis:
    - We handled the missing data, by imputing some columns.
    - Removing those columns and rows with high percentage of missing values.
    - We managed to retain 9000 rows and 13 columns
3. Optimizing the Categorical Variables:
    - For the Categorical variables with large number of levels and having very low population, we decided to combine them into others for better analysis.
4. Outlier Treatment:
    - Outliers were found to be present in the field 'Total Time Spent on Website however, as per business understanding they could be playing a vital role in model building.
5. Data Preparation:
    - Dummy variable creation.
    - Correlation analysis and eliminating highly correlated variables.
    - Splitting Data into test and train.
    - Standard Scaling numerical variables.
6. Model Development:
    - We used RFE to choose the best 25 variables.
    - On the basis of high P value and high VIF we were able to bring down our variables from 25 to 15.
    - We used these final 13 variables to train our model.

- We were able to achieve good accuracy (79%), specificity (81%)and sensitivity (75%) on test data and almost similar on train data, which also proved that we had no overfitting.
- The top 5 contributing variables found to be:
    1. What Is Your Current Occupation
    2. Lead Source
    3. Lead Origin
    4. Total Visits
    5. Total Time Spent on Website

Recommendations:
Using the model, sales team can now optimize the lead calling and conversion process.

1. Calculate the conversion probability score for all new leads using a simple UI that we will develop at the front end.
2. The bare minimum cut off score for lead conversion is 0.30. Any lead with a score below this will not convert.
3. Depending on the business requirements, the sales team can decide on the value of cut off they want to work with. Probability of lead conversion improves with an increase in the score.
4. A lower cut off score will give more leads for calling but with a risk of reduction in conversion rate
5. A higher cut off score would give fine-tuned leads with a greater probability of conversion, so sales team can maximize efficiency and focus on other work