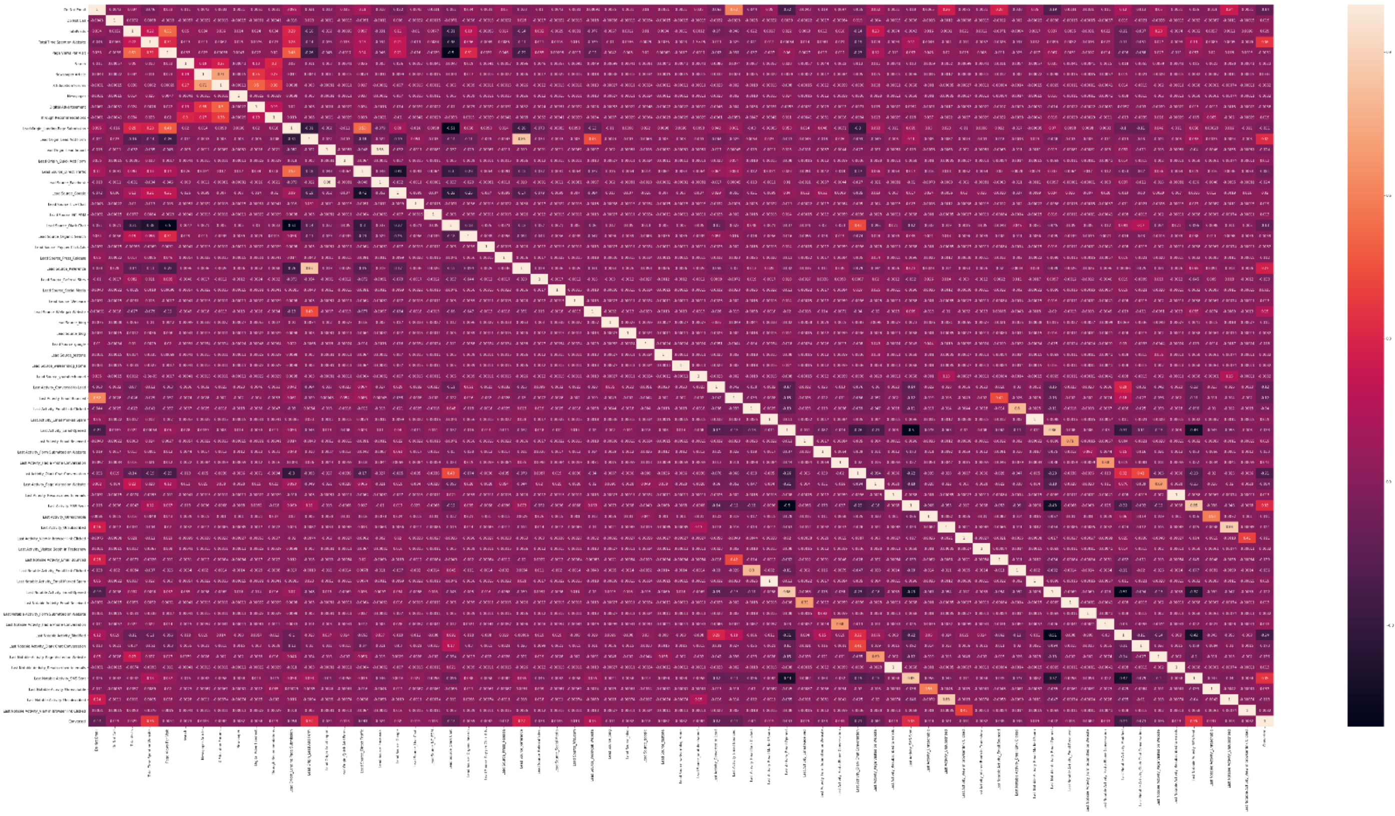
Authors: *Anish Mahapatra, Karthik Premanand*

Presentation for Executive, Chief Data Science officer.

Problem Statement:
The company X Education would like a model wherein lead score is to be mapped to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. Target lead conversion rate to be around 80%.

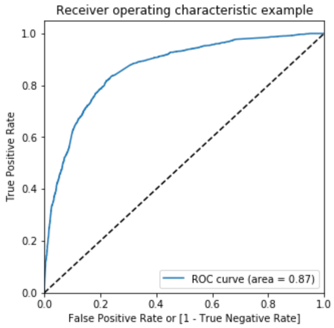
Data Analysis:

- There were 37 columns and 9240 rows present. Of this, multiple columns had “Select” as the data point, which indicated that the option was not selected and left at default. The columns that had "Select" are: Specialization, how did you hear about X Education, Lead Profile, City. These were converted to *NaN* before the analysis was performed.
- The level of the data was at Prospect ID / Lead Number level. There were certain columns that only had a **single value**. These columns were removed as well.
- The missing values were analyzed and the rows that had over 20-25% of missing value were discarded, unless they had a critical business implication. 13 columns were removed in this process.
- Outlier Analysis was performed on the data and the distribution on the numerical variables was analyzed. It was decided to keep the outliers as many outliers were present in certain features
- There were many categorical variables, on which one-hot encoding was performed to obtain dummy variables
- Correlation Analysis was performed on the dataset and the predictors that had extremely high correlation (multicollinearity) were excluded from the analysis to prevent a bias.

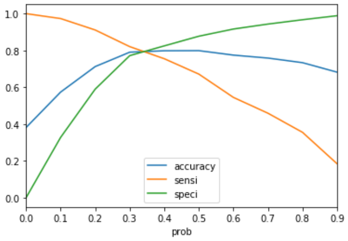


Modelling:

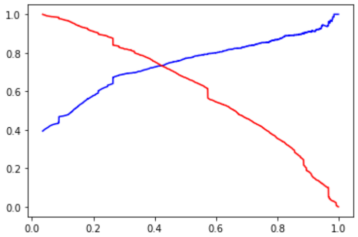
- A 70-30 train-test split was performed on the data and the numerical features were scaled using the StandardScaler() function.
- Recursive Feature Elimination was performed on the data and the top 15 variables were chosen after a couple of iterations.
- Model iterations were then performed keeping in mind the p-value and Variable Inflation Score. The threshold kept in mind was that p-value is supposed to be below 0.05 and the VIF score should not more than 10.
- Overall Accuracy of 80% was achieved on the training data set, post several model iterations.
- The ROC curve was then plotted to analyze the performance of the chosen model



- Sensitivity-Specificity analysis was performed to obtain an optimal probability of 0.3



- The Precision-Recall trade-off was analyzed via a plot as well to ensure that there was no bias in the model



- From this, we understand that the optimal threshold would be between 0.3 and 0.4.
- The model was tested on the test data to obtain the following evaluation metrics (**Test Data**):
  - **Accuracy**: 80.41%
  - **Sensitivity**: 74.79%
  - **Specificity**: 84.07%

Conclusion:

Hence, a successful model-building exercise was carried out and the sales representative shall be able to understand the leads to target to achieve 80% conversion for the leads pursued. If more leads are to be captured with lesser conversion rates, the threshold can be reduced and vice-versa.