Summary of Lead Scoring Case Study

Vishnu M Menon
menon.vishnum@gmail.com,
mob# 97312-88115

Dharmesh Bhuta
dharmeshbhuta85@gmail.com,
mob# 98197-38519

- Handling outliers by performing Exploratory Data Analysis and with the help of visualizations we removed them
- Handling null values in the form of “Select”: MODE, dropna, replace (parked them as “Others”)
- We found that few columns like: “Specialization”, “City”, “Country” had many unique values with low value_counts. We decided to club them under a single entity: “Others”. This made our analysis simple.
- Correlation Matrix and removed highly correlated variables (both positive and negative).
- Added dummy variables to the categorical variables and dropping the least favorable value manually. Instead of using drop_first, which might have removed one of the influential categories from the data set, hence, we decided to manually remove the least influential dummy variable as per the total count.
- We removed the variables which was added by the Sales team from the original data set, as it was modified by the Sales team later point of time. Those variables would have highly positively impacted our model which in turn would have given high accuracy, but greater count for “false positive”.
- Performed feature scaling for 3 variables
- First, we removed some of the highly correlated variables post analysis of the Heat Map before we started off with the Model Building
- Then we removed variables with no-correlation value
- Started off with model building using StatsModels and removed columns having “1” as their p-value. We had identified a lot of variables having values as “1” and closer to “1”. We decided to remove them before model building, as those variables might impact our analysis.
- Then we performed feature scaling using RFE on 20 columns, and used the supported column array for further model building
- Then we used the combination of p-value and VIF score to remove few more variables and built a final model with acceptable accuracy
- Then we performed ROC and got a value of area=0.83. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

- We identified the optimal probability cut-off point as 0.4, after performing iterations over multiple probability values ranging from [0.0 to 0.9] and plotted the derived extended evaluation matrices: accuracy, sensitivity & specificity
- Derived matrices:

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Train set	77.1	75.8	78.4
Test set	76.1	77.5	74.7