Summary Report

Lead Score Case Study

Objective of the case study was to select hot leads from the pool of leads gathered to concentrate only on the leads predicted 1 (due to high probability score) and monetize them by closing the sell.

Approach:

Data Cleaning:

- Dropping columns with null values greater than 45%
- Dropping categorical columns having single level or if skewed to a level
- Dropping columns not available at the time of prediction example Tags
- Imputing null values for columns having lesser than 45 % null values as appropriate

Outlier Treatment:

Removing outliers beyond 99 % percentile to avoid abrupt increase in values

Dummy Variable creation:

Creating dummy variables for categorical columns and proceed to modelling.

Modelling:

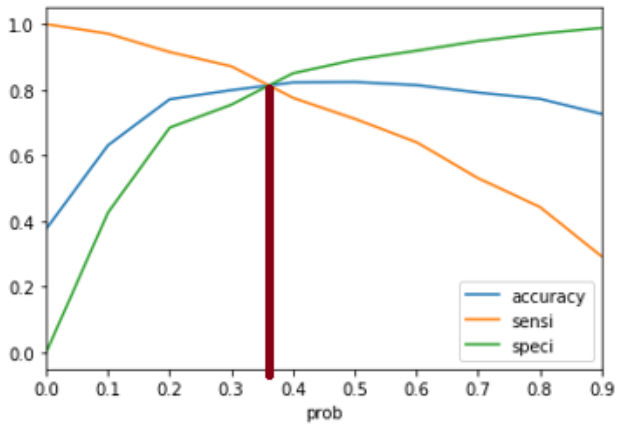
RFE was used to select 15 columns based on their ranking using scikit learn library in python for modelling.

Output of RFE is fed into logistic regression model from stats model api to further look at p values of coefficients to select those features which are significant and not selected by chance or if unstable due to collinearity present among columns.

VIF score is calculated after running the model and those features with VIF score more or equal to 5 are dropped one by one and running the model to keep an eye on coefficients p value to see if are stable now that is are having p value lesser than 0.05.

Model Evaluation:

Final model is evaluated for both class 1 and 0 for recall and specificity more than 80% with help of sensitivity /specificity plot to get cut off probability of 0.34 for predicting class 1/hot lead.



Learnings:

- Importance of optimal feature selection using business intuition as well as statistical methods
- Finding threshold probability using sensitivity /specificity curve