Summary of Lead Scoring Case Study

The following case study has been done for X Education as a way to find more leads on people who are more likely to convert into paying customers. In this study, we build a machine learning model, wherein we assign a “lead score” to people who visit the website. Observing and learning from their “behaviour” on the website, the model assigns the score and higher the score, higher the chance of a conversion to a paying customer.

A brief about the steps followed:

- Imported and had a brief inspection of the data.
- This was followed by cleaning up the data. The standard procedure was followed, dropping columns with 40% or more null values. The columns with lesser null values were either imputed with median of the whole column (numerical column) or imputed with mode of the whole column (categorical column).
- Next part was EDA. Visualizations were made to observe the data distribution and the outliers in said data are then dropped. Univariate analysis on the categorical columns and the numerical columns and also bivariate analysis on the categorical columns is done as well.
- Next, data prep was done by creating dummy variables for a few suitable variables. Then we did the train test split which is in 70-30 ratio respectively. Lastly for data prep we did scaling for a few variables using MinMaxScaler.
- Next, we did the feature elimination through RFE and got top 10 variables.
- Then came the logistic regression part, by building a model through GLM and dropped a few variables as the p-value was a bit high. We then proceeded with building a Variance Inflation Factor to check the collinearity between the variables.
- Then a confusion matrix to evaluate the model on accuracy, specificity, sensitivity, positive predictive value, negative predictive value. Also checked out the ROC curve between the false positive rate and true positive rate.

- Calculated the precision and recall for the model created. Upon which proceeded with plotting precision recall curve to find optimum cutoff. Based on the optimal cutoff, made predictions on the test data set. Finally created a data frame with lead number, probability, actual converted column, predicted converted column and lead score based on the probability.

- Final findings include:
    - Lead origin - API and Landing page submission has high hot leads
    - Lead Source - The number of Hot leads is higher in Direct Traffic and Google as compared to other categories.
    - Specialization - Most of the leads comes from Finance management but here Hot leads are lesser than Cold leads.
    - Last Activity - The number of hot leads are higher in SMS Sent. In Email Opened, cold leads are higher than hot leads.
    - Last Notable Activity - Same behaviour as Last Activity column
    - City - The response from Mumbai is high. It has maximum hot leads.
- Model Stats:
    - Accuracy: 84.72%
    - Sensitivity: 82.40%
    - Specificity: 86.15%
    - Positive Predictive Value: 78.57%
    - Negative Predictive Value: 88.82%