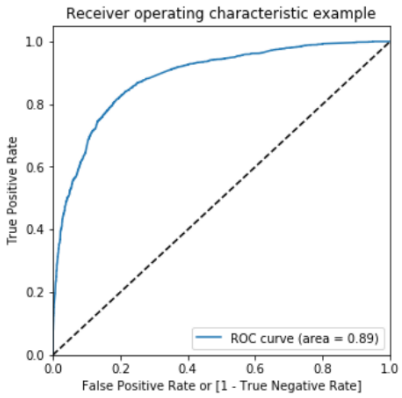
Summary of Lead Scoring Assignment – 18.04.2023

Submitted by – Bhavani Redhivari, Deepa & Kausik Rana

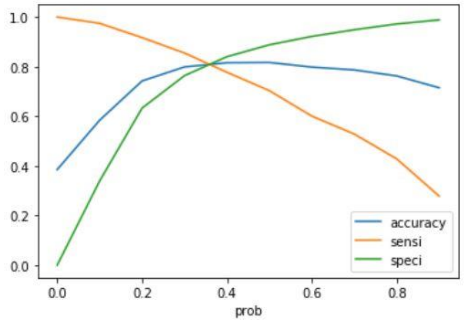
By understanding the problem statement, we straightly approached a logistic regression model on the past dataset provided by the company named X Education to handle the business problem.

Our approaches are:

1. Import and inspect the data set: Used necessary python libraries to get the data in note book and inspected.
2. Data Cleaning: During inspection found outliers and null values are present in dataset. Checked uniqueness of dataset and dropped as required. During treatment of null values, found select which is as good as null. Initially we straight away removed the columns which contain more than 70% null values which do not make any sense. Later, we treated null values of columns one by one and settled those as required based on EDA analysis. After taking care of null values, we found some outliers present and the same is treated by dropping them. Finally, we remained with 98.2 of data.
3. Exploratory Data Analysis: We did EDA to clean dataset and dropped least significant columns. Merged minimal data points as others. Dropped columns as required.
4. Data Preparation: After cleaning, we started data preparation. Here, we converted binary levels of categorical variables to 0 & 1 and created dummies for multi levels of categorical variables.
5. Split of dataset: We split the dataset into train and test after features scaling. We choose MinMax scaling here. Highly correlated columns are dropped.
6. Building of models and RFE treatment: We built logistic regression model by using linear_model with all available columns. As all columns are not significant to predict the model, we referred RFE to choose top 15 features. We rebuilt the model again with 15 features. Checked significance & multi-collinearity by using P-value (<0.05) & VIF (Variance Inflation Factor) (<2).
7. Confusion Metrics and model accuracy: Calculated confusion metrics and evaluated accuracy, sensitivity and specificity with a default cut-off value of 0.5. Accuracy, sensitivity and specificity are 81%, 70% and 89%.



8. ROC curve: Calculated TPR & FPF
9. and plot them. Graph is more in upper left as area under the curve is more implies model is good.
10. Finding of optimal cut-off: Plotted accuracy, sensitivity and specificity with the probability of data point to find-out optimal cut-off.



Here the optimal cutoff is near 0.38 0r 0.39.

Created confusion metrics of train dataset optimal cutoff and we found the accuracy of the model is 81%, sensitivity is 82% and specificity is 81% which improves the model accuracy.

11. Testing model on test dataset: Tested model on testing data set and found accuracy, sensitivity and specificity are 81%, 80% and 81% which are near to training accuracy.
12. Assigning lead score, feature importance and final conclusion: We created leads score column based on predicted variables for the company sales representative to focus. We created features importance as well to predict most likely conversion lead for the company sales representatives.

    Total Time Spent on Website 100.000000
    Lead Origin_Lead Add Form 64.832962
    What is your current occupation_Working Professional 58.721321

    The above top three variables are really important to consider for convert leads most likely.