Lead Scoring Case Study

To Improve efficiency in process of targeting leads that are more likely to convert into customer, by using previous data about the process.

Steps in our Analysis-

- Getting data
- Data Preprocessing
- EDA
- Model Training
- Selecting threshold probability
- Evaluating model metrics – Accuracy, Sensitivity and Specificity.
- Performance on test data

1. Getting data

- We had data set with 9240 data points and 37 feature info. Most of the features were categorical and a few were numerical, we can draw insights based on univariate and bivariate analysis as part of EDA

2. Data preprocessing

- We selected features that has less than 35% missing values.
- Removed feature rows that has ‘Select’ value – as it is not very useful to analysis.
- By common sense – Removed not so useful features like City, Country, etc.
- Removed features that are highly skewed, to a category level. For ex – Newspaper, Magzine, etc
- Finally we left with 6343 data points, with no null values

EDA

- Univariate analysis of data by using **sweetviz** library, revealed the features, that have to be removed, in data preprocessing step.
- Also, associations among features were looked at
- Important features were – Total time spent, Total visits, etc

Model Training

- We did train test split at 70:30 ratio.
- Trained a base logistic regression model in sklearn and got train and test accuracy as 81 and 76 percent (by using all 75 features).
- We then selected top 15 RFE selected features to make our model.
- We then manually trained model in statsmodel and monitored p-value and VIF of feature, such that p-value < 0.05 and VIF < 5 for significant features.
- Finally, we had 12 features, that had significant p-value and VIF in our model.

Selecting threshold

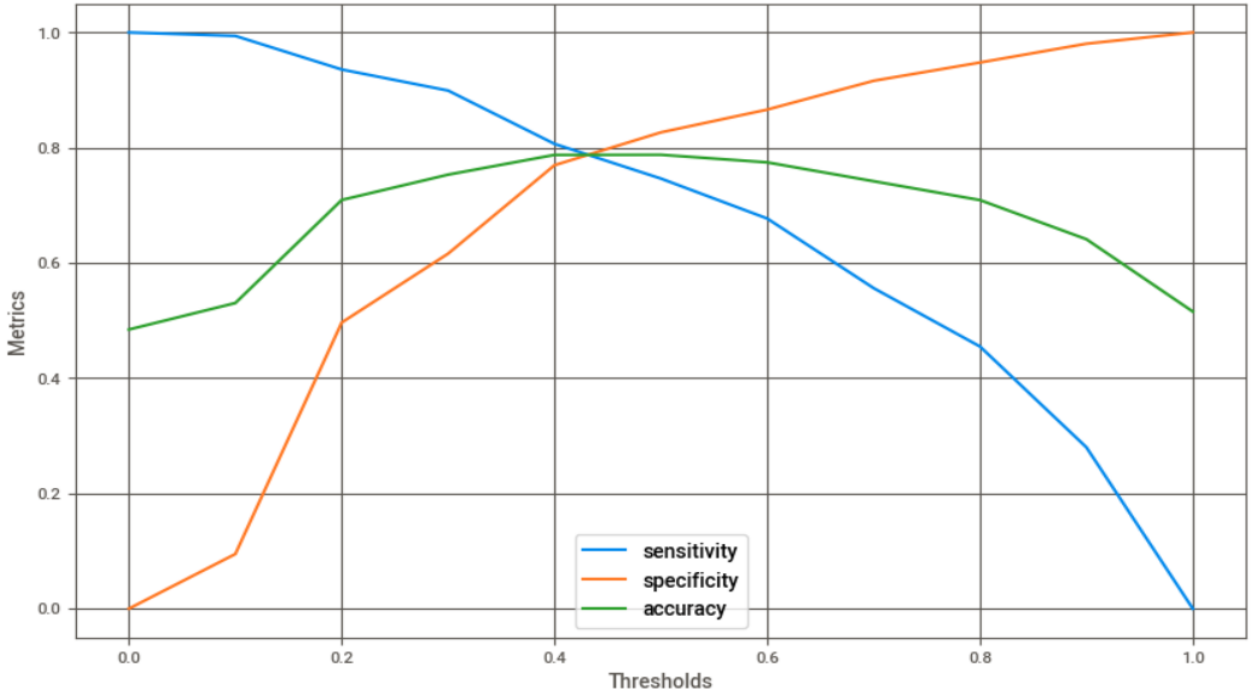
- Finding the threshold probability for prediction, is a subjective question as per business requirement.
- Assuming, we want to minimize false negative (to prevent loss of potential customer not getting converted) – we want a model with high sensitivity (TPR).
- Also, specificity should be high, so that we don’t end up targeting customers which are not going to convert easily. As it will end up eating the efficiency of the sales team.

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Performance metrics

- Our model we balanced sensitivity, specificity and accuracy to arrive at threshold value of 0.43 by plotting the metrics at different values of threshold probability.

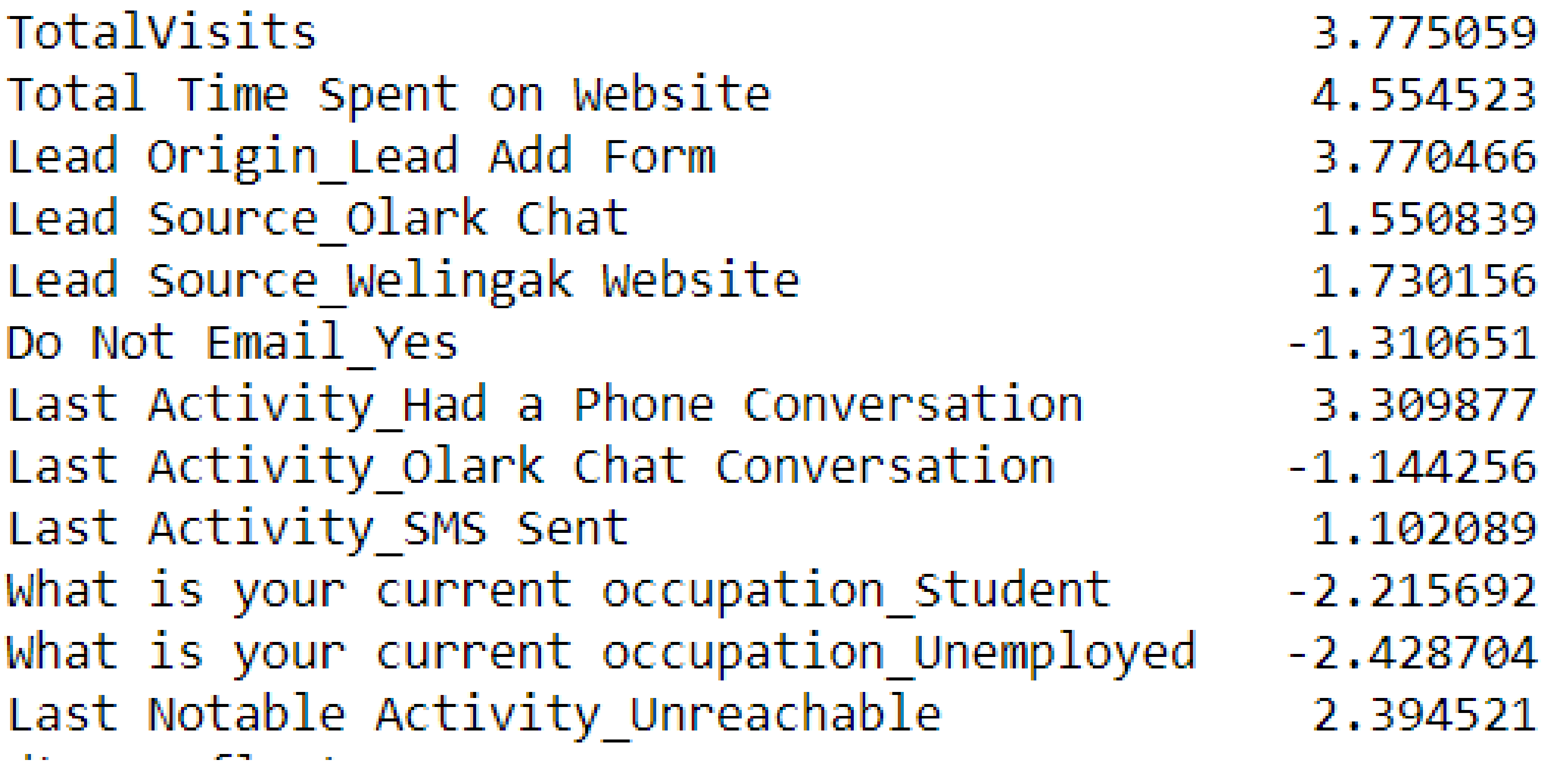


Performance on test data

- Test accuracy – 78%
- Test Sensitivity – 78%
- Test Specificity – 78%

Important features

- Feature importance is in than rank of absolute value of coefficient of feature in weight vector.



```
TotalVisits                                  3.775059
Total Time Spent on Website                  4.554523
Lead Origin_Lead Add Form                    3.770466
Lead Source_Olark Chat                       1.550839
Lead Source_Welingak Website                 1.730156
Do Not Email_Yes                            -1.310651
Last Activity_Had a Phone Conversation       3.309877
Last Activity_Olark Chat Conversation       -1.144256
Last Activity_SMS Sent                       1.102089
What is your current occupation_Student     -2.215692
What is your current occupation_Unemployed  -2.428704
Last Notable Activity_Unreachable            2.394521
```

Important features-

- Top features –

Feature Name	Coefficient value
TotalVisits	3.775059
Total Time Spent on Website	4.554523
Lead Origin_Lead Add Form	3.770466