# Leads Scoring Case Study Summary

## Shraavan Sridhar and Shruthi Manjrekar

## Data Understanding

- Checked the shape of the data set.
- Used df.describe() to check numerical metrics and df.info() to check column data types.
- Performed Null/duplicates inspection. Found no duplicates but various columns having nulls were observed.

## Data Cleaning

- Some columns were having label as 'Select' which means the customer has chosen not to answer this question. Hence, we changed those labels from 'Select' to null values.
- Columns with less than 2 unique values were removed as they were insignificant for our analysis.
- Columns having more than 40% null values have been removed.
- For numerical columns, their null values were imputed with the median.
- Categorical null values were replaced with 'Unknown' tag as imputing those with a value will skew the data.
- Dropped irrelevant variables like magazine, Get update on dm content, Lead Numbers, Asymmetric score and Tags. Data outliers were also removed.
- Prospect_ID, Lead_Number was not adding any information hence we have removed columns irrelevant for model building.

## Exploratory Data Analysis

- Performed data imbalance check on target variable- “Converted”. It had lower converted (38.5%) records as compared to those which were not converted (61.5%).

- Performed EDA for categorical and numerical variables against the target variable ‘converted’. The results are logged in the presentation.

## Data Transformation

Performed Outlier treatment for Total visits and page viewed per visit.

## Model Building

- Split the dataset into train and test. We split the data into train and test dataset with a 70:30 ratio.
- Performed Feature Scaling using MinMaxScalar().
- Created Logistic Regression model using RFE we eliminated irrelevant fields from the model and selected top 15 feature variables, followed by manual feature reduction to reach at 13 variables by checking VIF, P-Value (VIF<5 and p-value <0.05).

## Model Validation

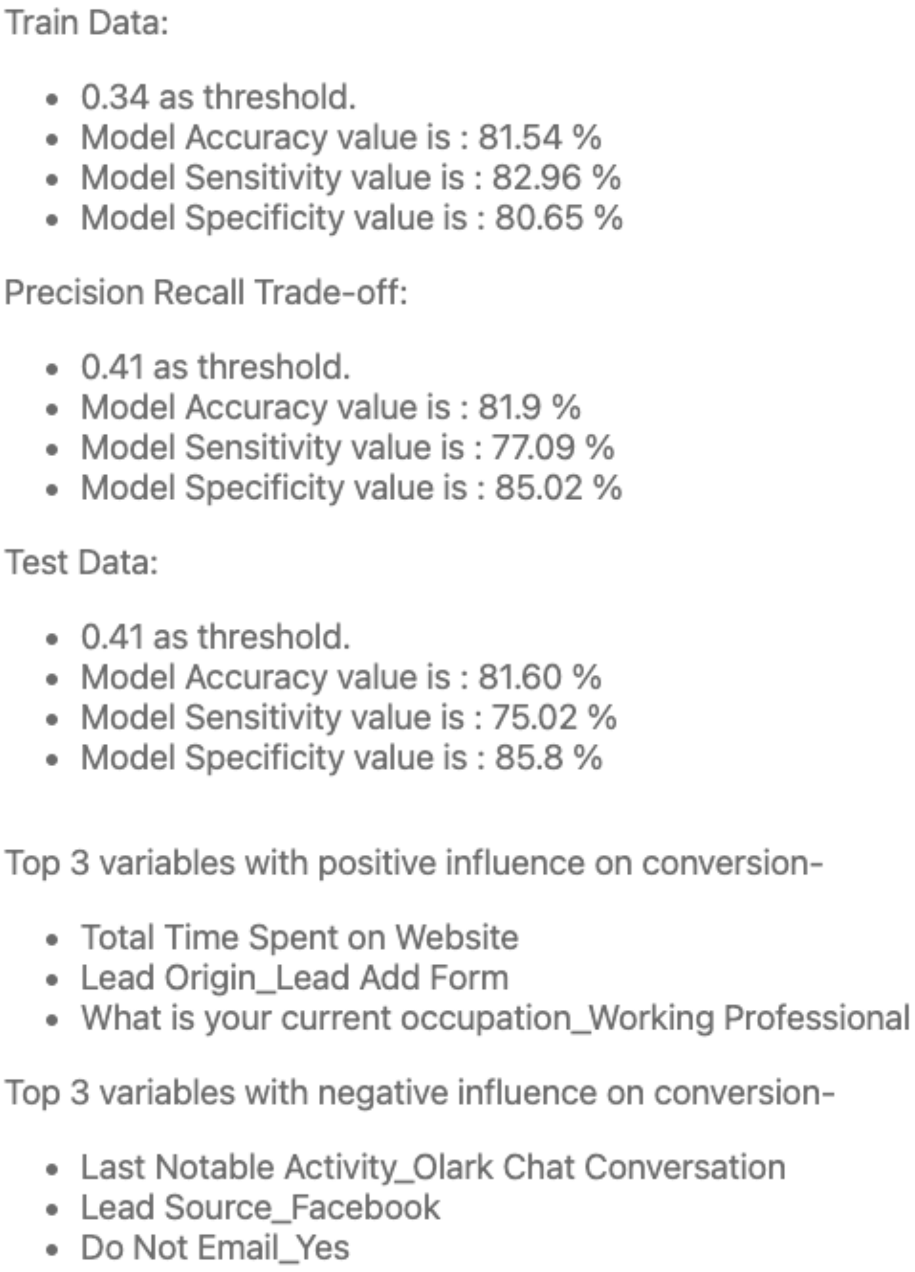
- Performed probability predication.
- Checked the optimal probability cut-off by finding points and checking the accuracy, sensitivity and specificity and found one convergent points (at 0.34).
- Checked confusion matrix, Accuracy, Sensitivity, and Specificity ranged in 80% (acceptable range). ROC curve (0.90 area under the curve)
- Performed Precision-Recall Trade off that gave cut off 0.41.
- Chose the same columns from the above train model for test prediction.
- Applied the precision trade off cut off to the test data and arrived at a final model.
- Created Confusion matrix, ROC curve on Test Model. Test set is having 0.89 as area under curve which is similar to train data.
- Assigned Lead Score on the training data and test data.

## Final Suggestions/Recommendations

- The total time spent on the website heavily influences lead conversion. More time the lead spends on website, more likely is the conversion.
- Leads generated through Add_forms will see greater conversions.
- Working professionals tend to take up the course more often than not.
- Olark chat conversations, Facebook leads, the ones who asked not be emailed have a negative influence on conversion and one needs to exercise caution while approaching these leads.
- If the lead engages activity through SMS, he/she is more likely to be converted.
- For leads through references, the conversion is more than non-conversion.

- X Education company can target conversions through References and Welligak Website leads as they see higher conversions.
- Target leads with higher lead score as they have higher probability to get converted.

## Final Model Metrics



Train Data:

- 0.34 as threshold.
- Model Accuracy value is : 81.54 %
- Model Sensitivity value is : 82.96 %
- Model Specificity value is : 80.65 %

Precision Recall Trade-off:

- 0.41 as threshold.
- Model Accuracy value is : 81.9 %
- Model Sensitivity value is : 77.09 %
- Model Specificity value is : 85.02 %

Test Data:

- 0.41 as threshold.
- Model Accuracy value is : 81.60 %
- Model Sensitivity value is : 75.02 %
- Model Specificity value is : 85.8 %

Top 3 variables with positive influence on conversion-

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional

Top 3 variables with negative influence on conversion-

- Last Notable Activity_Olark Chat Conversation
- Lead Source_Facebook
- Do Not Email_Yes