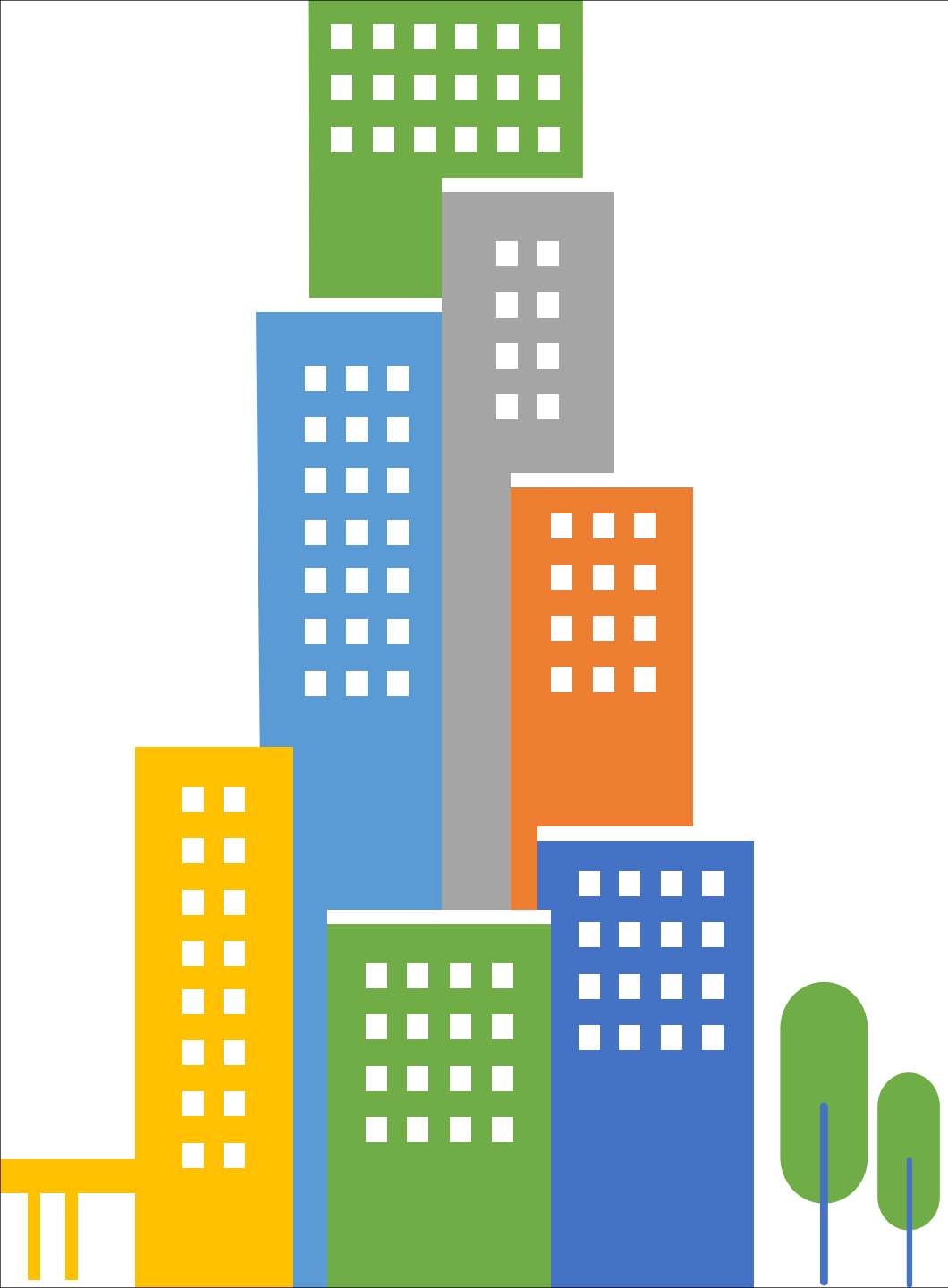


LEAD SCORE CASE STUDY

Presentation by :
1. Tejas Guptha
2. Sana Monin
3. Ganesh Rathod

Outline of the presentation

01. Problem Statement & Objectives

02. Methodology

03. Assumptions

04. EDA

05. Model Building & Evaluation

06. Conclusions & Recommendations

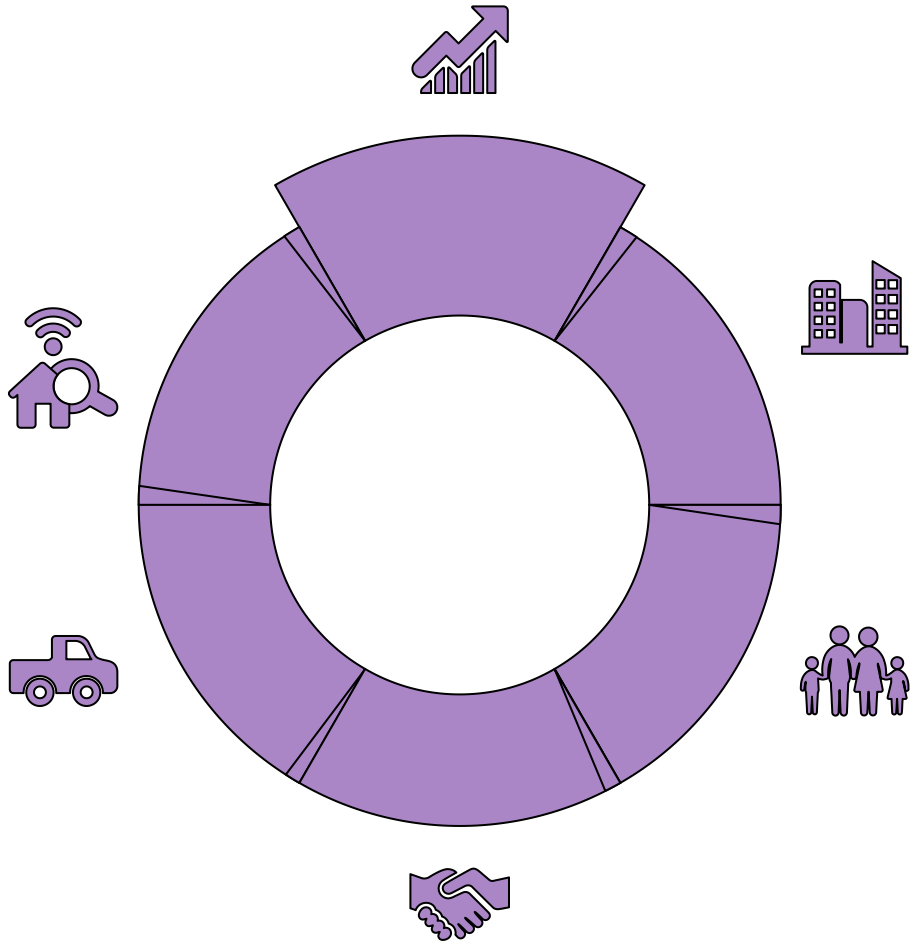
01 Problem Statement & Objectives

01 To create a logistic regression model to provide each lead a lead score between 0 and 100 that the business may use to target potential prospects.

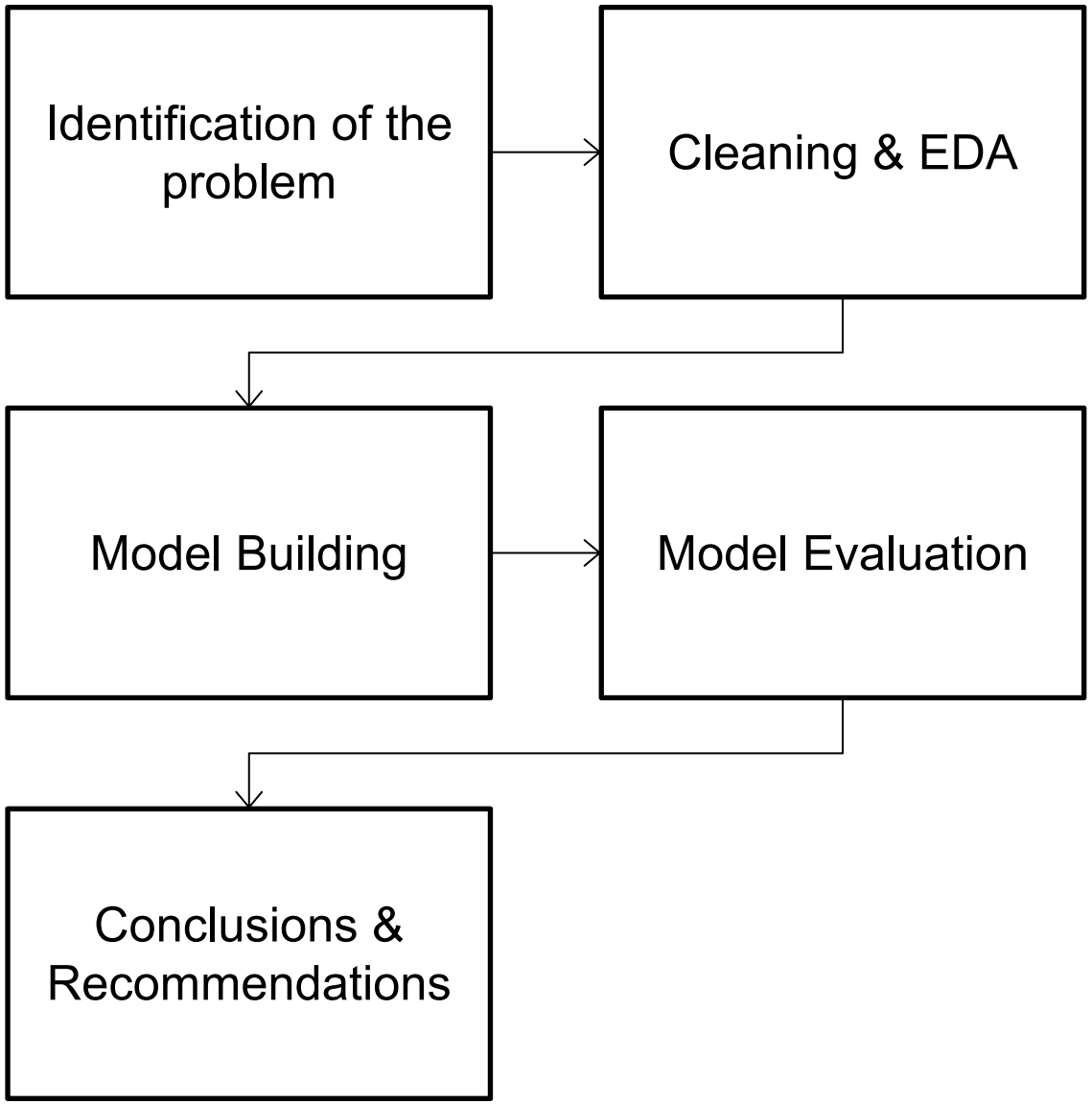
02 Applying the concept of EDA to carry out the analysis.

03 To evaluate the logistic regression model

04 To provide valuable suggestions to the company for better business



02 Methodology



03 Assumptions made in the study

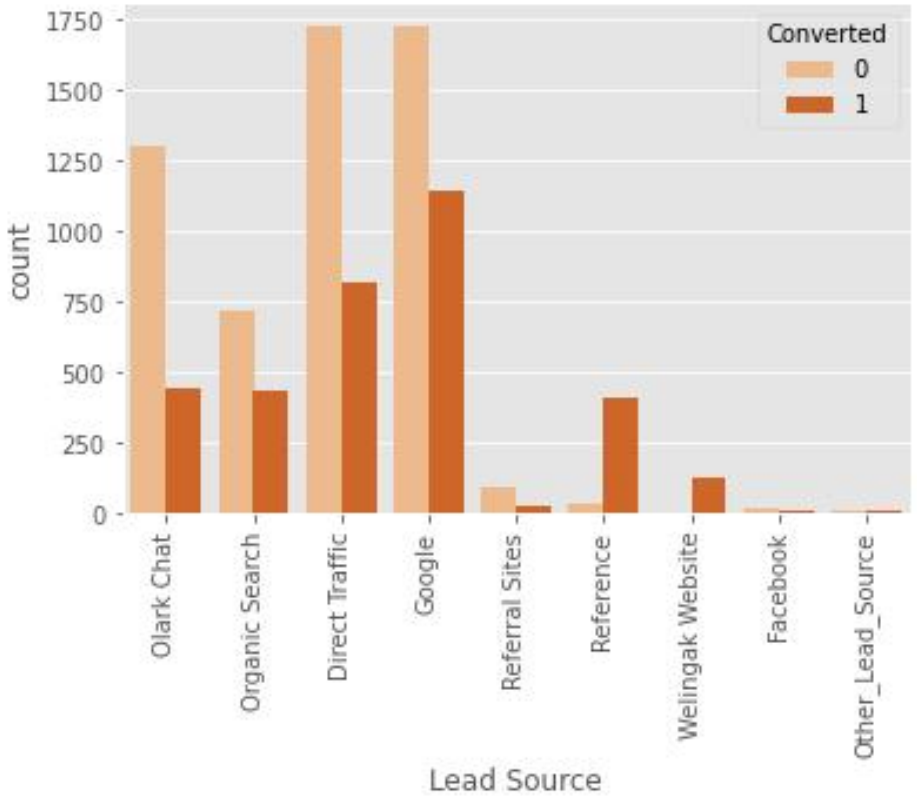
The data provided by the company is genuine and free from errors.

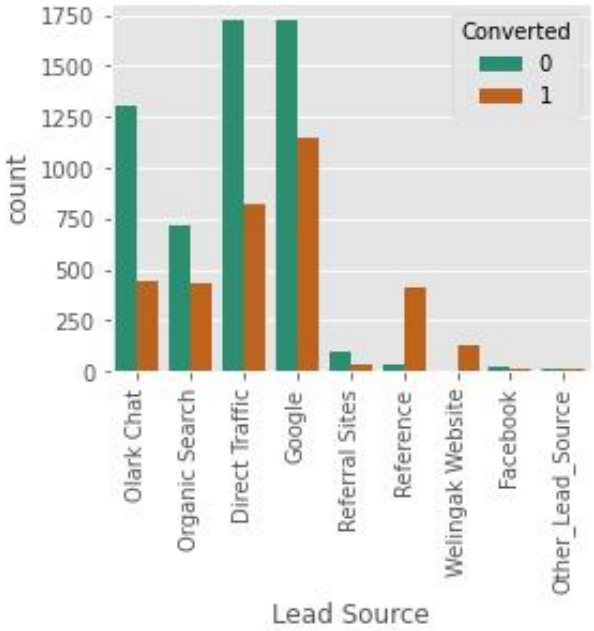
A column/feature with more than 40% nulls are dropped.

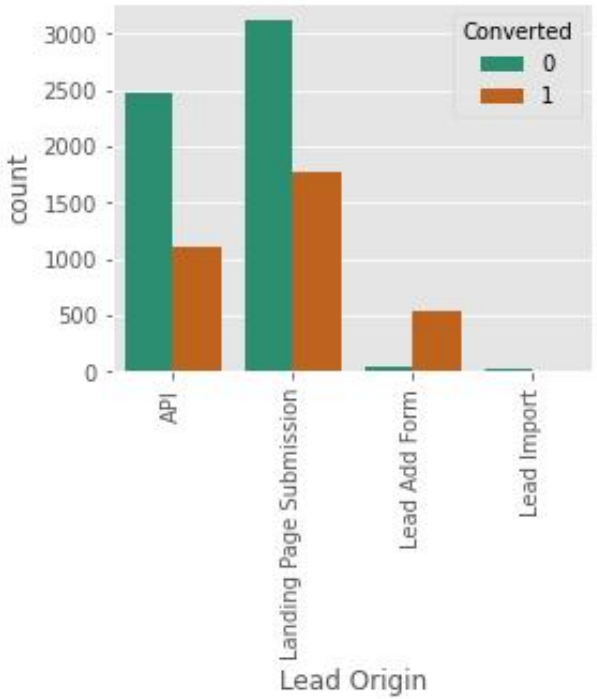
RFE approach was used to find the top 15 features

Major insights are included in the presentation. IPYNB file may be referred for detailed analysis

04 EDA



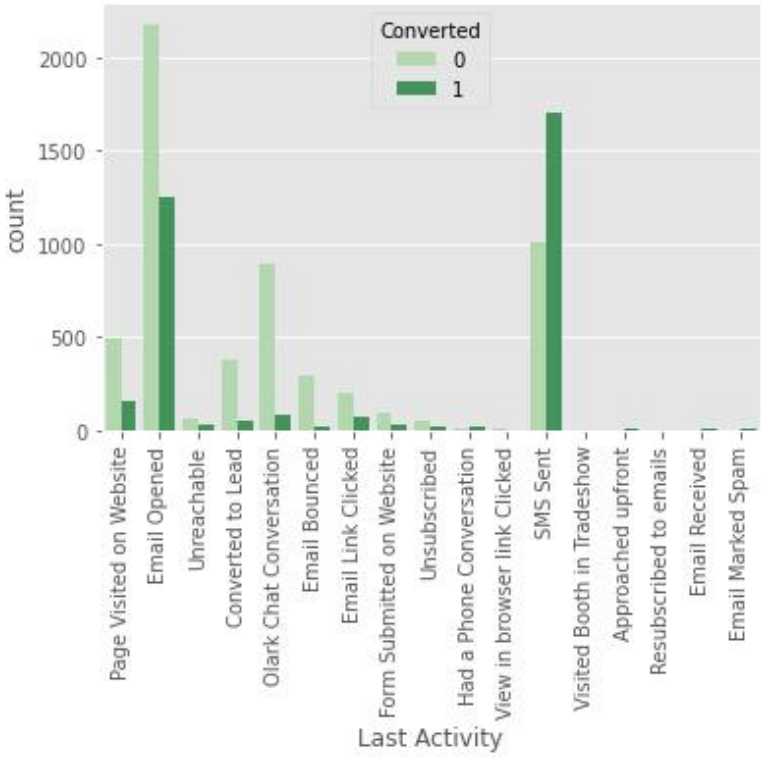


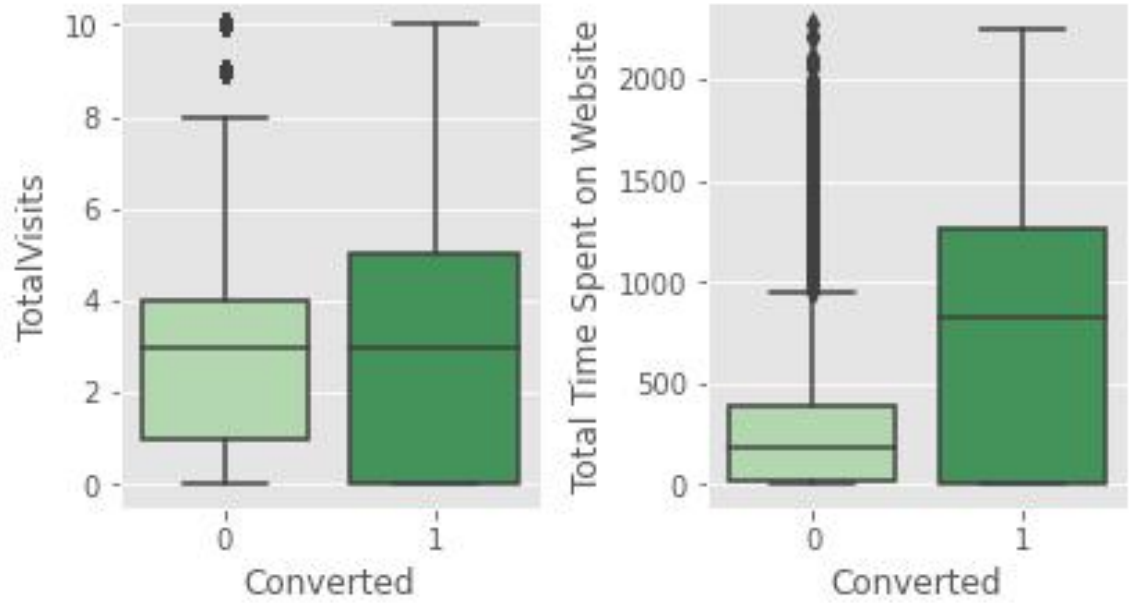


- **Looking at above plots, we can infer that we have maximum leads from Google and Direct Traffic.**
- **The conversion rates of leads from Welingak website and reference is maximum**
- **Leads from landing page submission are considerate however have less conversion rate**

EDA

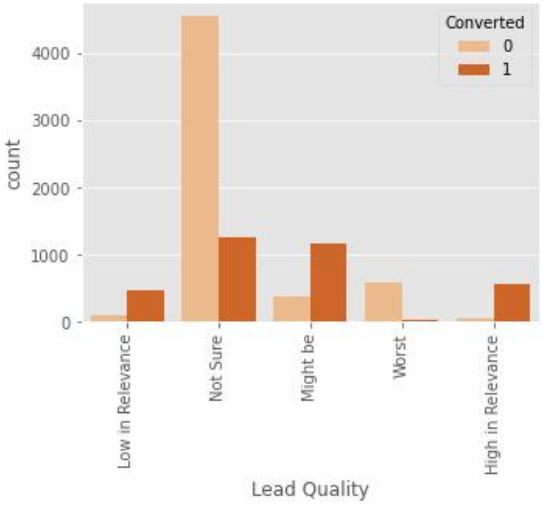
- **We have around 30% conversion rates.**
- **The median from conversion and non-conversion is same, hence its inconclusive**
- **The more time spent on website by user leads to potential conversion.**
- **As for last activity, ‘email opened’ and ‘SMS’ is maximum**

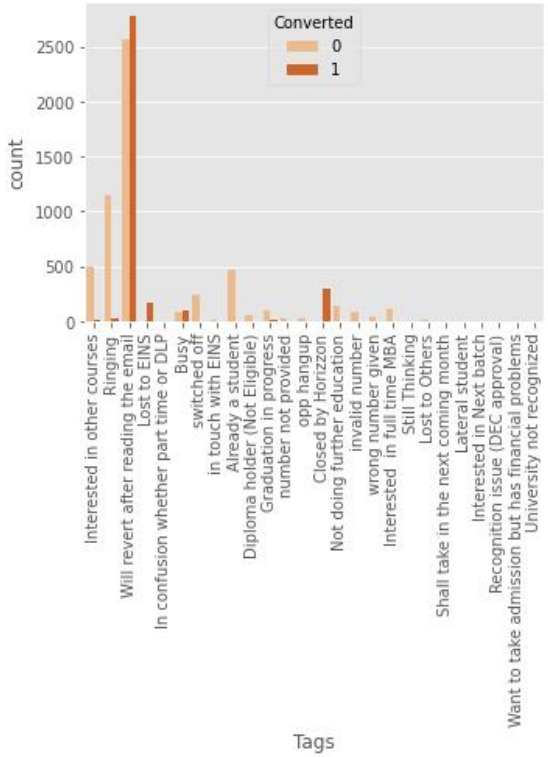


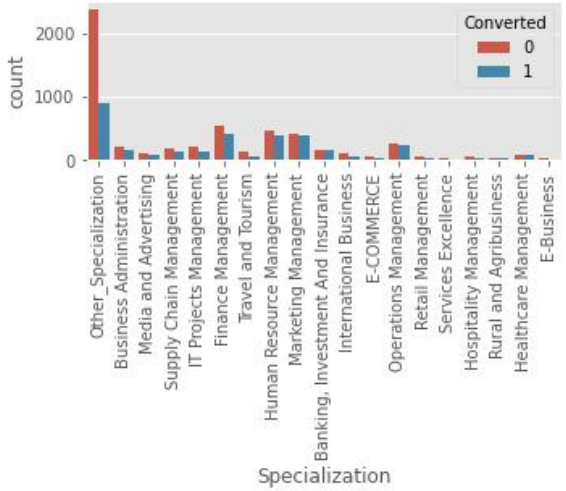


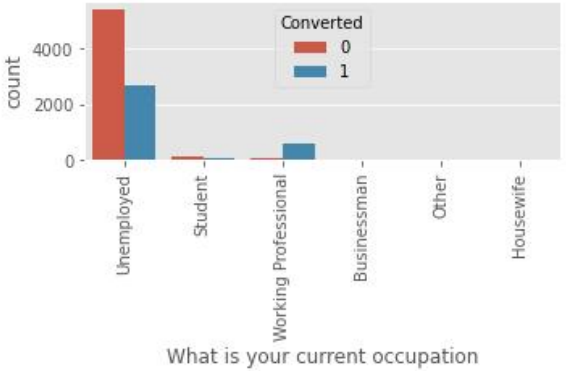
EDA

- **No specific inference can be drawn from specialization.**
- **Working professionals have higher rate of conversion**







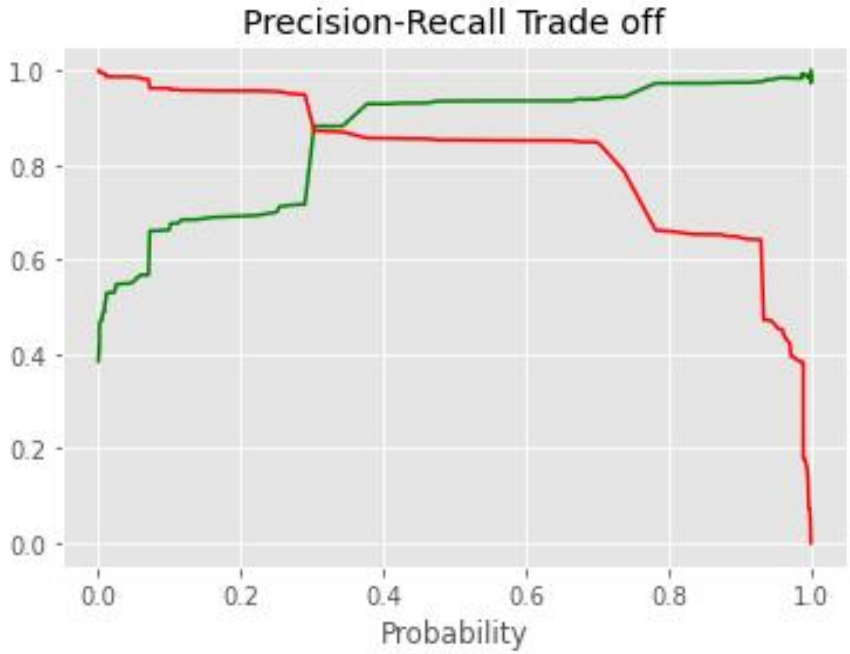


05 Model Building

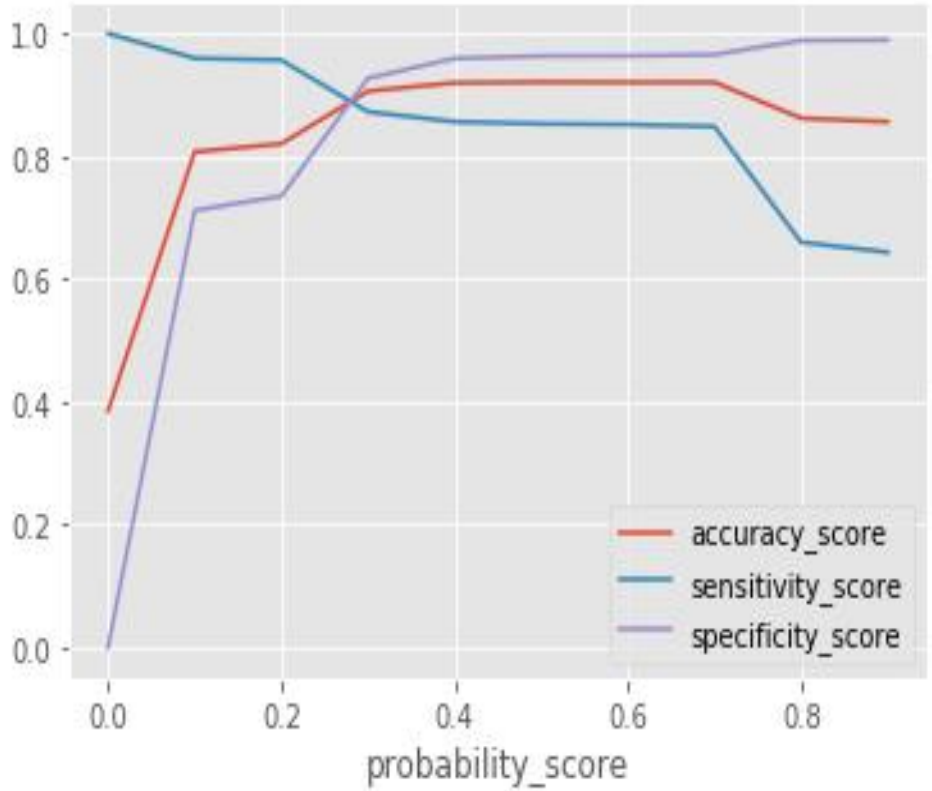
1. Dummy variables were created for the numerical data.
2. Standard scaler was used to scale the data.
3. RFE approach was adopted to select the top 15 influencing features.
4. The data was split in the ratio of 70% as train data and 30% as test data.
5. The final model was decided based on the p-value and the VIF of the features. It was ensured in the final model that all the features have a p-value less that 0.05 and VIF less than 5.

Model Evaluation

- In Sensitivity-Specificity-Accuracy plot 0.27 probability looks optimal. In Precision-Recall Curve 0.3 looks optimal.
- We are taking 0.3 is the optimum point as a cutoff probability and assigning Lead Score in training data.



Model Evaluation-Sensitivity & Specificity on Train Data Set



- Accuracy:0.899
- Sensitivity:0.854
- Specificity:0.924
- Precision: 0.865

Precision and Recall on Train dataset

- In Precision-Recall Curve 0.3 looks optimal.

- Precision = 79%
- Recall = 65%

Final Model

- The final model has Sensitivity of 0.854, this means the model is able to predict 85% customers out of all the converted customers, (Positive conversion) correctly.
- The final model has Precision of 0.86, this means 86% of predicted hot leads are True Hot Leads.
- We have also built an reusable code block which will predict Convert value and Lead Score given training, test data and a cut-off. Different cutoffs can be used depending on the use-cases (for eg. when high sensitivity is required, when model have optimum precision score etc.)

06 Conclusions & Recommendations

- The logistic regression model predicts the probability of the target variable having certain value.
- Optimum cut off value is chosen to be 0.3
- The final logistic model is build with 14 features.
- Tags_lost to EINS (coefficient factor = 9.578632)
- Lead quality_worst (coefficient factor = -3.943680)
- Tags_closed by horizon (coeficient factor = 8.555901)
- The final model has sensitivity of 0.928, this means the model is able to predict 92% customers out of all the
- The final model has Precision of 0.865 this means 86.5% of predicted hot leads are True Hot Leads.

Thank You