LEAD SCORING: CASE STUDY

APPROACH

The project was approached in the below order:

1. Importing libraries
2. Reading and inspecting the dataset
3. Performing data cleaning
4. Exploratory Data Analysis
5. Preparation of data for building the model
6. Model Building
7. Model Evaluation
8. Business Recommendations

DATA INSPECTION

- The dataset had:
  - 9240 rows
  - 37 columns
- Out of the 37 columns, 9 columns were populated by the sales team.
- Missing values could be found in 17 columns, initially.
- Many columns had:
  - Single category
  - Skewed categories
- The column with unique identifiers ('Prospect ID' and 'Lead Number') were all unique and had no duplicate values

DATA CLEANING

- Dropped the identifier columns ('Prospect ID' and 'Lead Number') since it wouldn't help in analysis or modelling.
- Dropped 5 columns which had just 1 category each since it won't help in differentiated between CONVERTED or NON-CONVERTED leads.
- Dropped 9 columns which had extremely skewed categories.
- The category 'Select' was rightfully replaced as NULL value in all columns.
- Columns with > 39% values missing were dropped.

DATA CLEANING - IMPUTATION

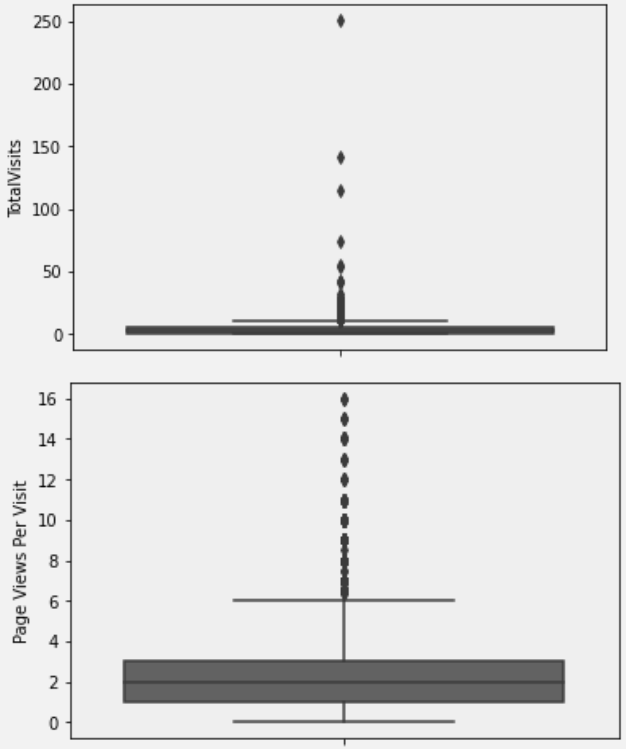
- In '**Specialization**' column:
  - Missing values were imputed with 'Unknown' as the category.
  - The other categories which belonged to either 'Business' or 'Management' categories, were grouped accordingly.
  - Categories with <3% observations were grouped together in 'Others' variable.
- In '**What is your current occupation**' column:
  - Missing values were imputed with 'Unknown'
  - Rest of the groups were divided into – Employed, Unemployed & Student, accordingly.

DATA CLEANING - IMPUTATION

- In '**Lead Source**' column:
  - 'Google' and 'google' were combined since they were segregated due to capitalizations.
  - Missing values were imputed with the mode of the column.
  - Categories with <10% observations were combined into a single category 'Others'
- Median was used to impute the numerical features below (due to presence of extreme values):
  - '**TotalVisits**'
  - '**Page Views Per Visit**'

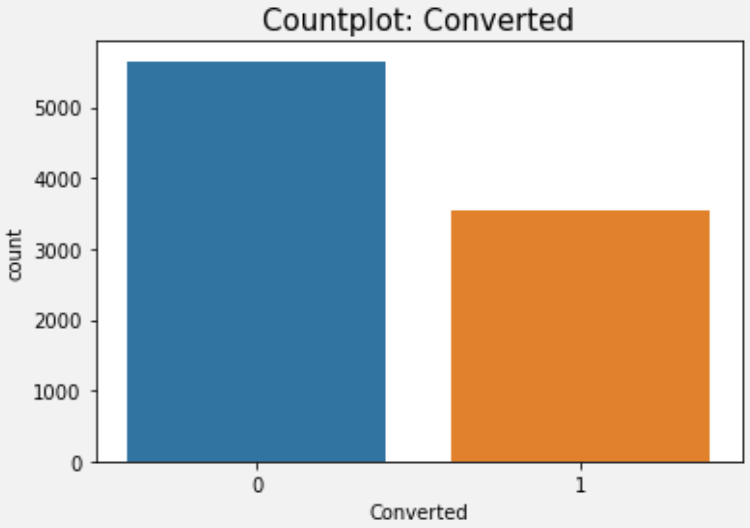
OUTLIER ANALYSIS

- Outliers were spotted in 2 variables:
  - '**TotalVisits**'
  - '**Page Views Per Visit**'
- The outliers were treated by dropping the top 0.5 percentile of rows as per '**TotalVisits**' column.
- The rest of the outliers were not removed since the existence of those values was possible more often and could draw some meaning.
- We had 99.5% rows remaining after complete cleaning.



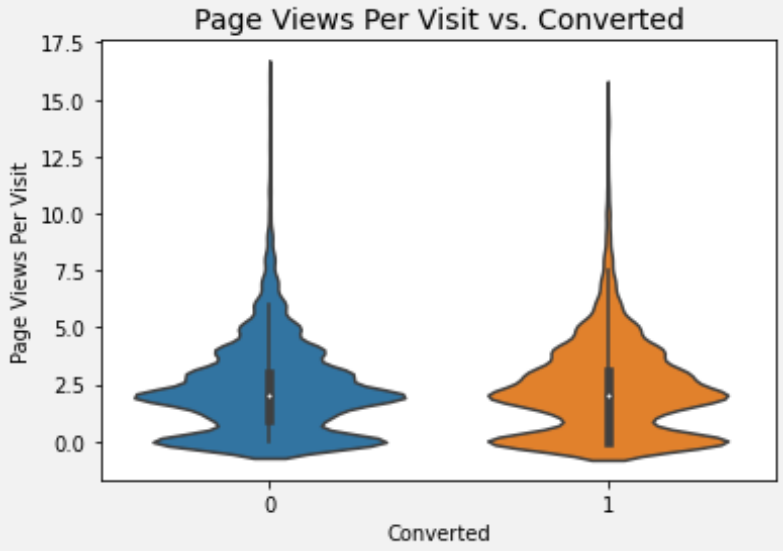
TARGET VARIABLE - IMBALANCE

- The dataset had slight imbalance in the target variable '**Converted**'
- Non-converted leads had 61.5% and converted leads had 38.5% data
- The conversion ratio based on the dataset was 62.7%



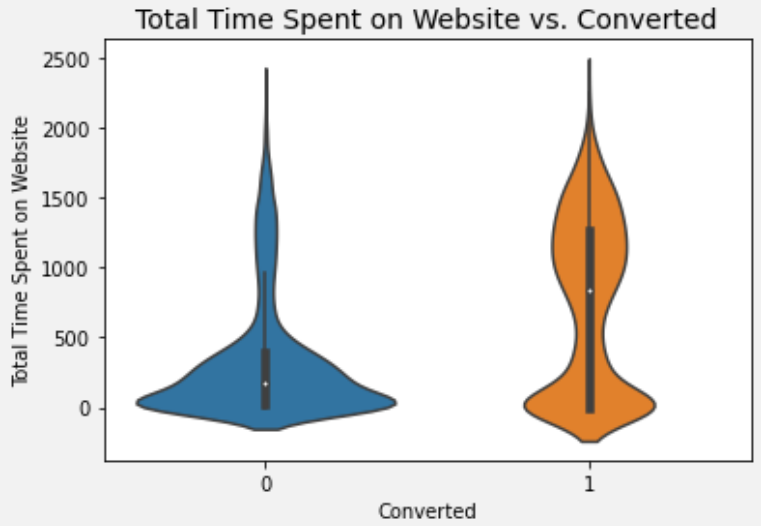
EDA: **PAGE VIEWS PER VISIT** VS. **CONVERTED**

- The distribution for both the Converted values can be seen high at 0 and 2.
- It can also be observed that 2 page views has slightly lower distribution in converted leads than in non-converted leads.



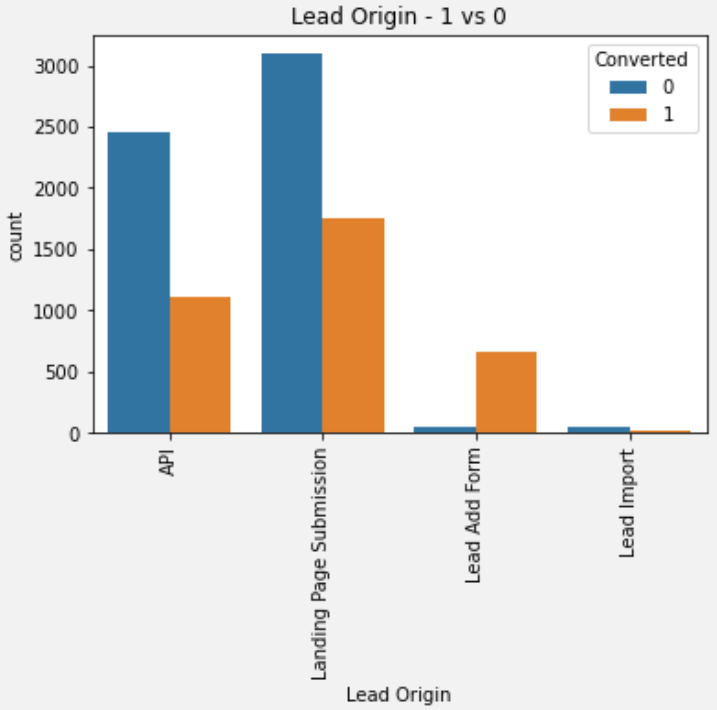
EDA: **TOTAL TIME SPENT ON WEBSITE** VS. **CONVERTED**

- Most of the non-converted leads have 0 time spent on the website
- The distribution at 0 is much more higher for non-converted leads than converted leads.



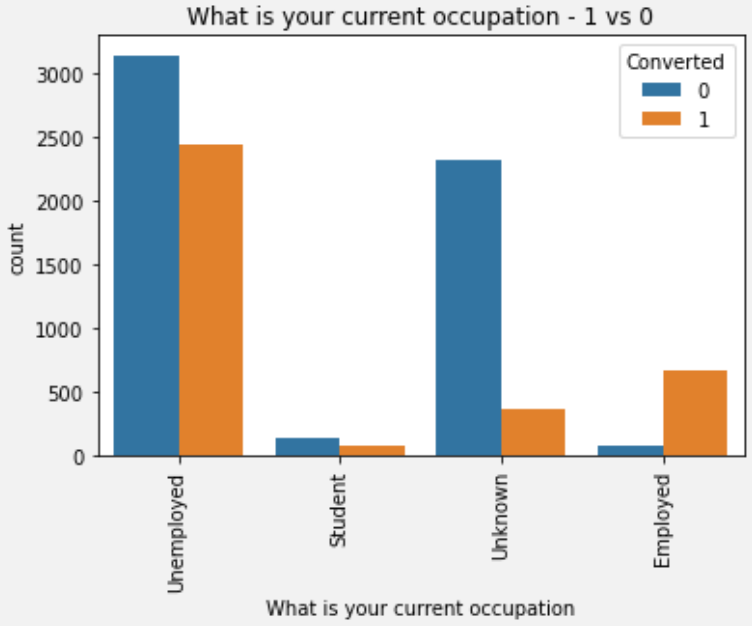
EDA: **LEAD ORIGIN** VS. **CONVERTED**

- The non-converted leads have much higher observations in ‘API’ and ‘Landing Page Submission’ than converted leads.
- Although ‘Lead Add Form’ seems to have higher converted leads.



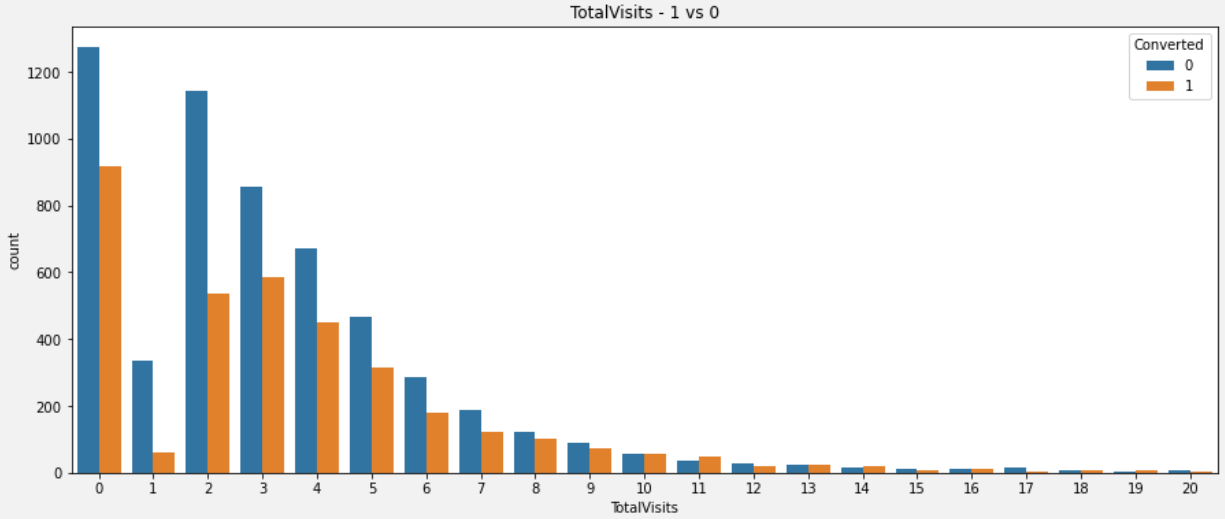
EDA: **WHAT IS YOUR CURRENT OCCUPATION** VS. **CONVERTED**

- In the case where the current occupation is Unknown there seems to be extremely high observations of non-converted leads than converted leads.
- Employed seems to be a pretty significant factor for identifying converted leads
- Unemployed category has a ratio of non-converted leads higher than converted leads by a small fraction.



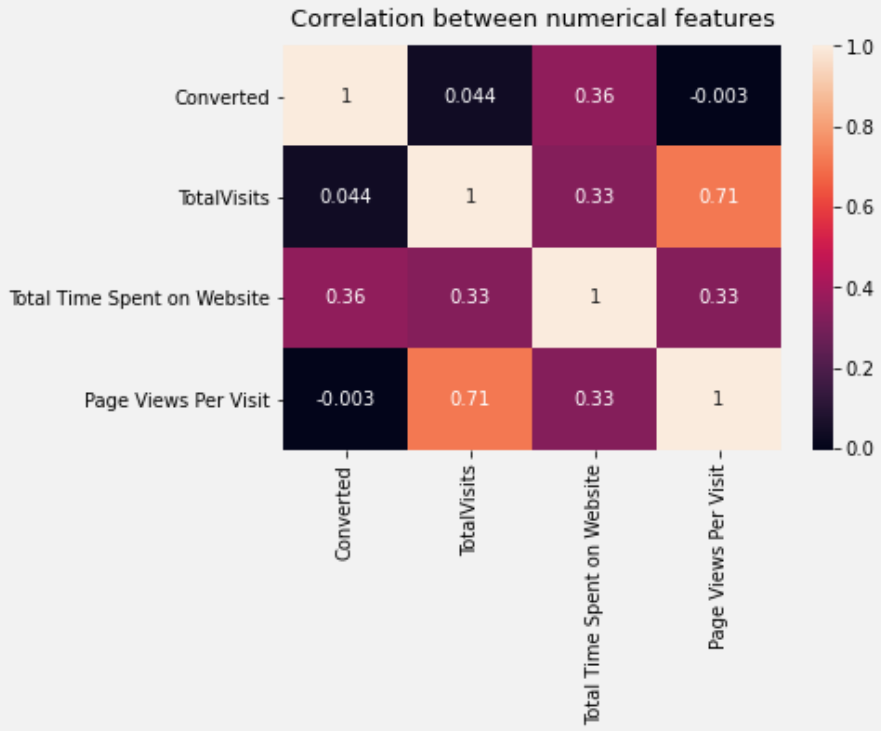
EDA: **TOTAL VISITS** VS. **CONVERTED**

- The ratio of conversion seems to be almost similar in all **TotalVisits** below 9 values except 1 and 2, it seems that no. of observations of non-converted leads are much higher than converted leads.
- A slight increase can be seen in converted leads as the count of **TotalVisits** increases.



EDA: **CORRELATION MATRIX**

- All the features seem to have acceptable correlation values and multicolllinearity, except:
  - **TotalVisits**
  - **Page Views Per Visit**
- Decision of which of the above columns to drop will be taken based on the p value and VIF value during model building.



DATA PREPARATION

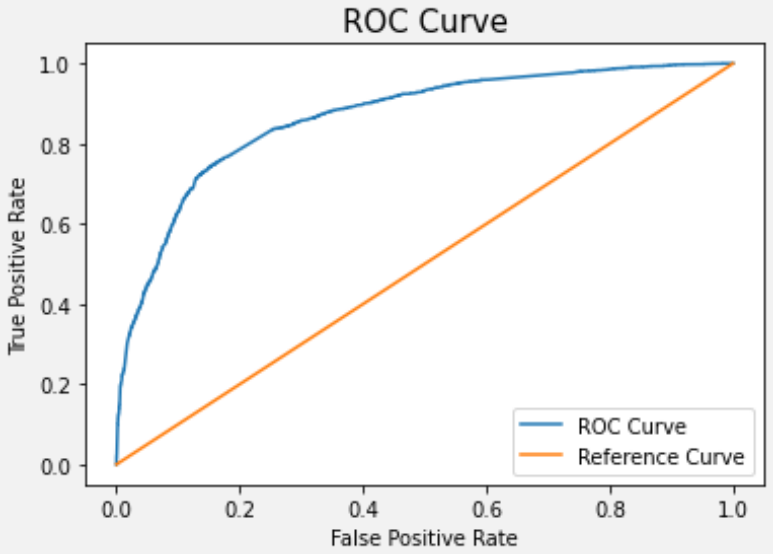
- Columns which were populated by sales team were dropped
- Columns with binary categories (yes and no) were replaced with numerical values (1 and 0, respectively).
- Dummies were made out of rest of the categorical columns by which we created new column for each category.
- Original categorical columns were dropped after joining the dummy columns to the dataset.
- The data was then divided as independent(X) and dependant(y) datasets.
- The X and y datasets were then divided further as training and testing datasets with training dataset having 70% and test having the rest of the 30%.
- The numerical columns in the training test was then scaled by normalizing the values.

MODEL BUILDING

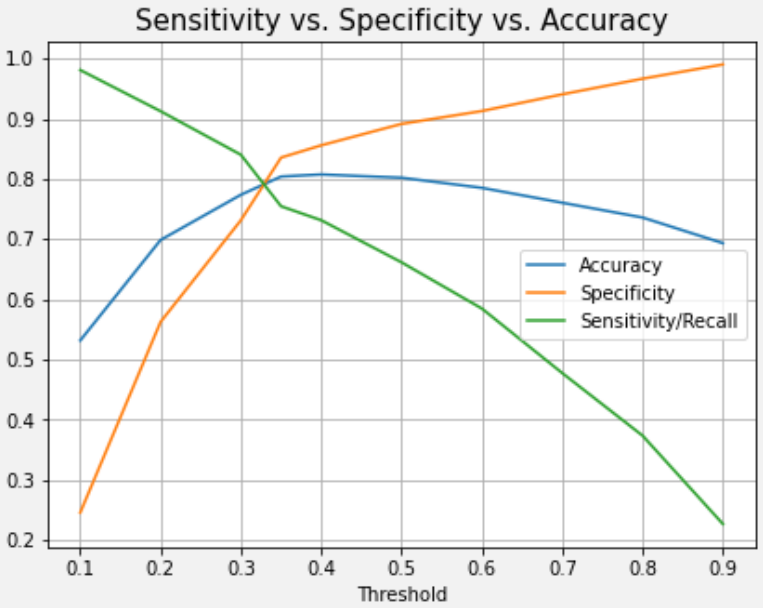
- Out of the final 19 columns in our dataset, we selected 15 columns with the help of RFE (Recursive Feature Elimination)
- We built a **logistic regression** model with the help of **statsmodels** library.
- By checking the model summary, the interpretation of columns significane was done with the help of *p-values*.
  - If a column had high *p-value*, it was dropped and the model was rebuilt and summary checked again.
- If all *p-values* were in an acceptable range, *VIF* (Variance Inflation Factor) was checked.
- When both *p-value* and *VIF* had acceptable values, the model was to be evaluated.

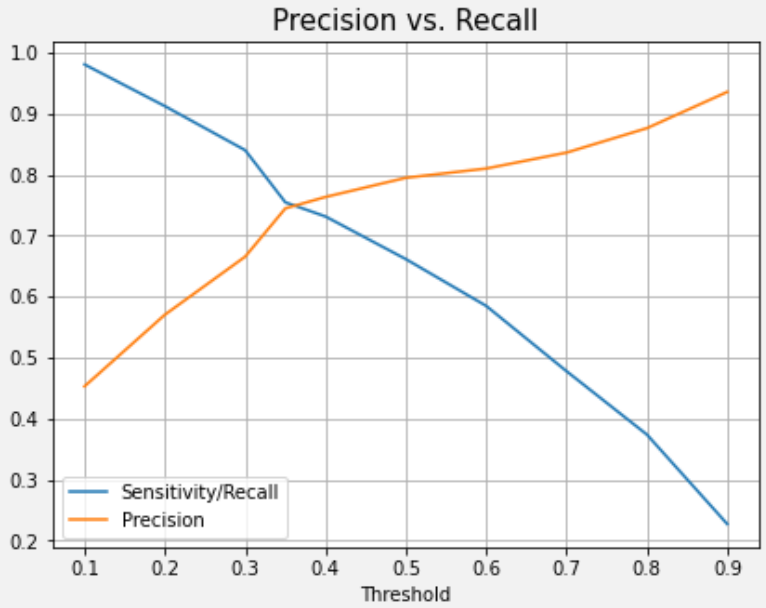
MODEL EVALUATION

- The quality of the model was decided in a general sense with the help of ***ROC AUC score***
  - In our case, we had *ROC AUC Score* of **0.87**, which is considered to be a good score
- Based on the recommendations from the **CEO of X-Education**, we needed to have a conversion rate of around **80%**.
- Since we can afford having *False Positives*, but wanted to reduce the number of *False Negatives*, it was decided that the major metric to be considered for optimal threshold was **Sensitivity / Recall.**



MODEL EVALUATION - SCORES



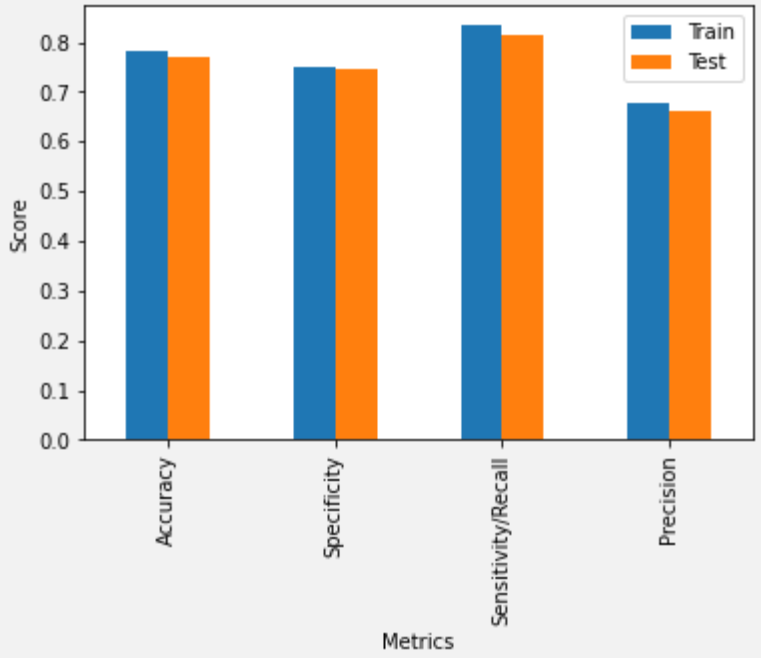


Based on the above plots, and taking a good sensitivity into consideration, the optimal cut-off threshold was decided to be **0.32**

MODEL BUILDING – FINAL SCORES

After building the model, and predicting for the testing set, below are the scores on which lead scores were predicted.

	Accuracy	Specificity	Sensitivity/Recall	Precision
Train	0.78	0.75	0.83	0.68
Test	0.77	0.75	0.81	0.66



BUSINESS RECOMMENDATIONS

The features with **POSITIVE** effect on the final prediction are:

- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Olark Chat
- Lead Origin_Lead Add Form
- Total Time Spent on Website

The features with **NEGATIVE** effect on the final prediction are:

- What is your current occupation_Unknown
- What is your current occupation_Student
- What is your current occupation_Unemployed
- Do Not Email
- Specialization_Unknown

BUSINESS RECOMMENDATIONS

The below features which have the **HIGHEST** effect on the Final Prediction:

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

By keeping the above conditions in check and checking the 'Lead Scores' as predicted by the model, the business can make stronger decisions for pursuing leads and confirm if a lead is either ***HOT*** or ***COLD***.

THANK YOU!