Lead Score Case Study

Rupak Shah
Akshata Shetti

Contents

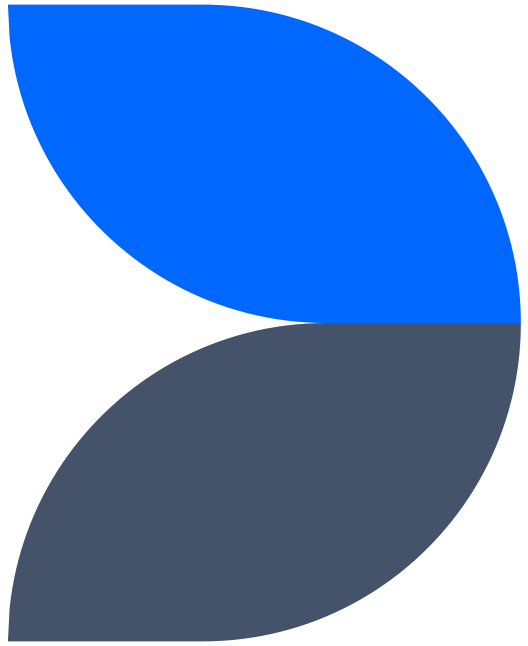
- ✓ Problem Statement
- ✓ Business Objective
- ✓ Solution Approach
- ✓ Summary

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For Ex., if the X Education acquire 100 leads in a day, only 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

- X Education wants to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires a model wherein a lead score will be assigned to each of the leads such that the customers with higher lead score have a higher conversion chance of being converted and the customers with lower lead score have a lower conversion chance.



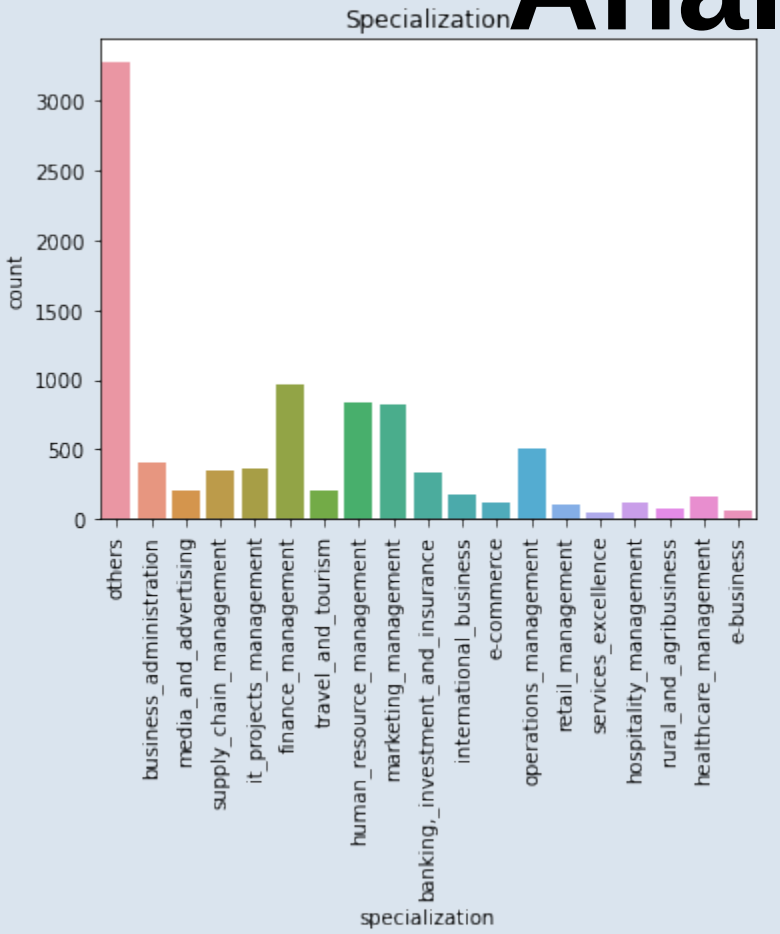
Approach Towards Solution

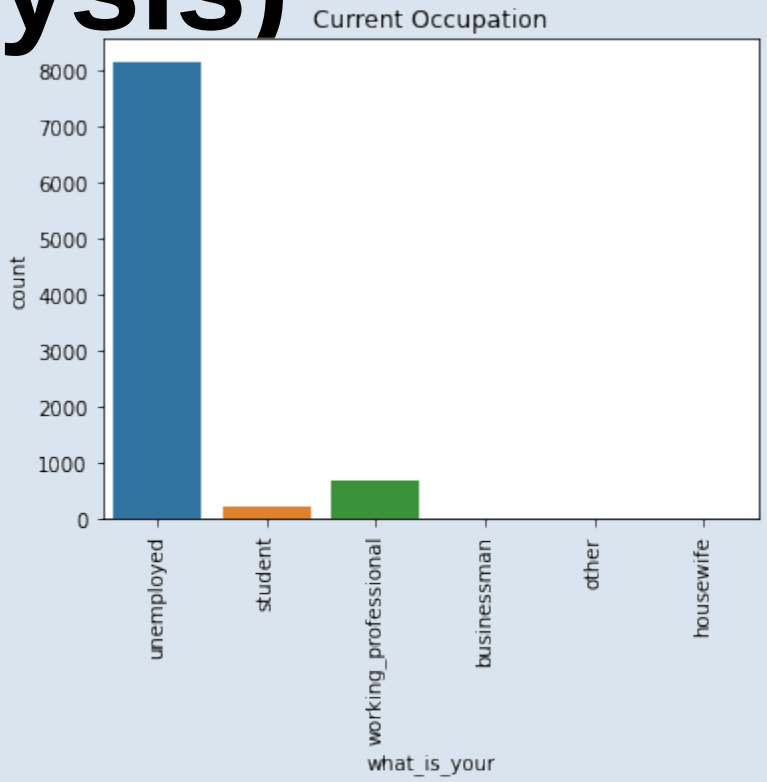
- Data Preparation
  - ✓ Checking & handling Null Values/ Duplicate Data
  - ✓ Reading & inspecting data
  - ✓ Dropping columns having large number of Null values and imputation of null values if required.
  - ✓ Checking & handling outliers.
- EDA
  - ✓ Univariate Analysis
  - ✓ Bi-variate Analysis
- Dummy Variables, Correlation Matrix Analysis & Encoding of the data.
- Train-Test Split, Feature Scaling
- Model building using Logistic Regression and RFE Technique, fine manual tuning using VIF score and P-values
- Evaluation of Model using Sensitivity, Specificity, Precision, Recall and F1_Score metrics
- Model Presentation
- Conclusions and recommendations

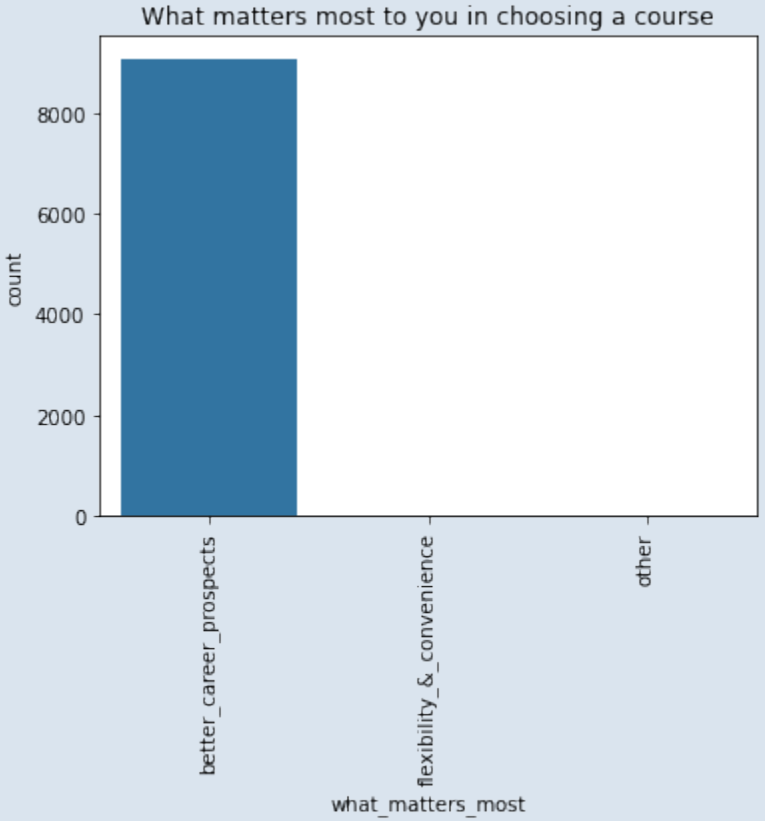
Data Cleaning & Manipulation

- Total Number of Rows are 9240 and columns are 37.
- Standardizing names of the columns by removing extra spaces.
- Replaced “Select” option with NaN.
- Dropped columns having for than 45% null values.
- Imputed remaining missing values of categorical variable through Mode.
- Outliers in Numerical Variable are removed by keeping to 99 Percentile Values
- Dropped columns which have very less significance as per business requirement.

EDA (Univariate Analysis)

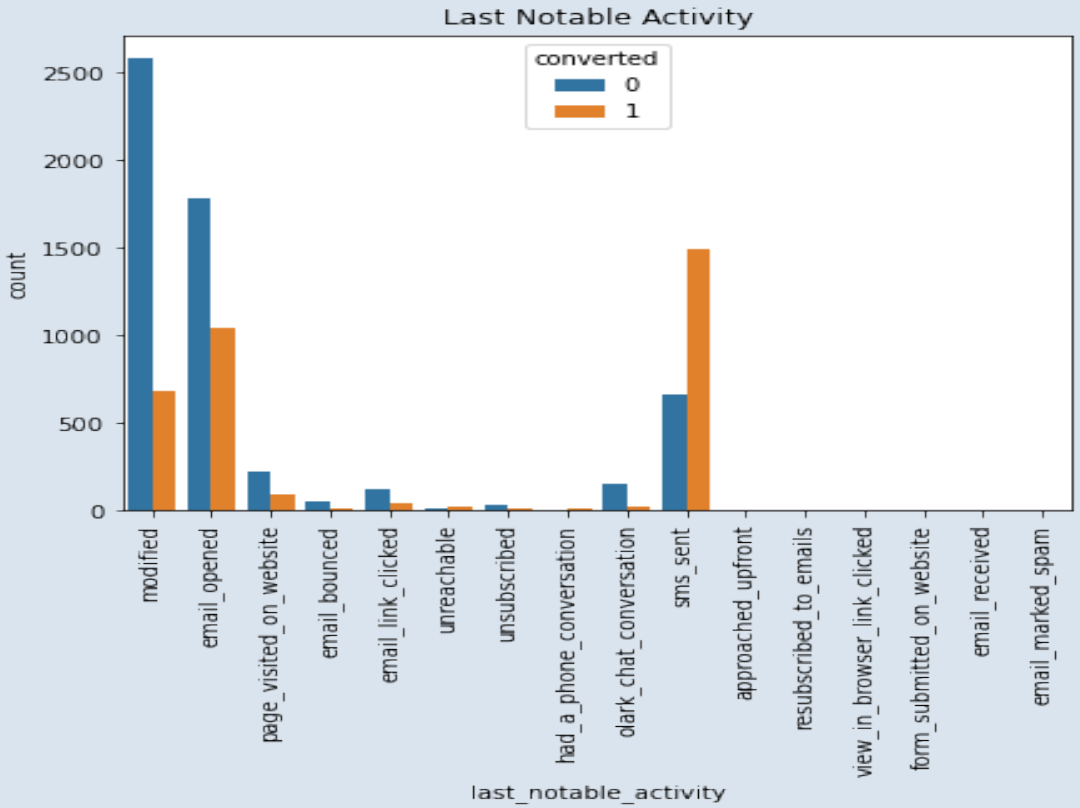


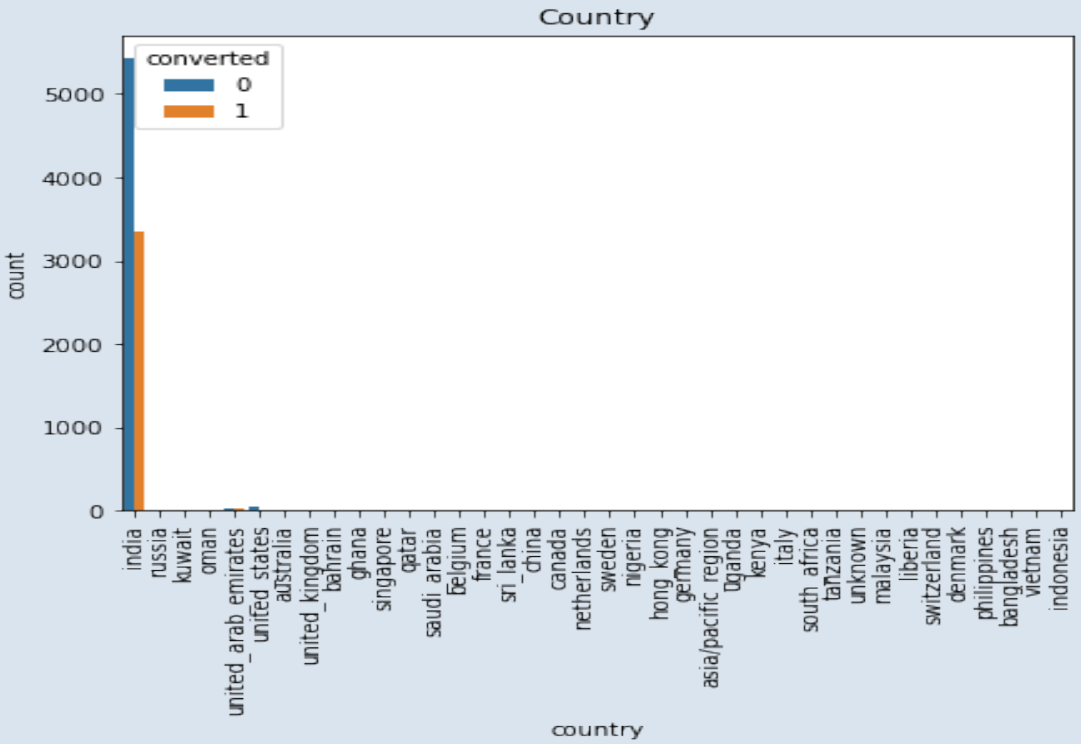




- In “Specialization column” we can see the majority counts is Others.
- In column "What_is_your_current_occupation" we can see that the majority of customers are Unemployed.
- In “What matters most” column the priority is better career prospects.

EDA (Univariate Analysis)

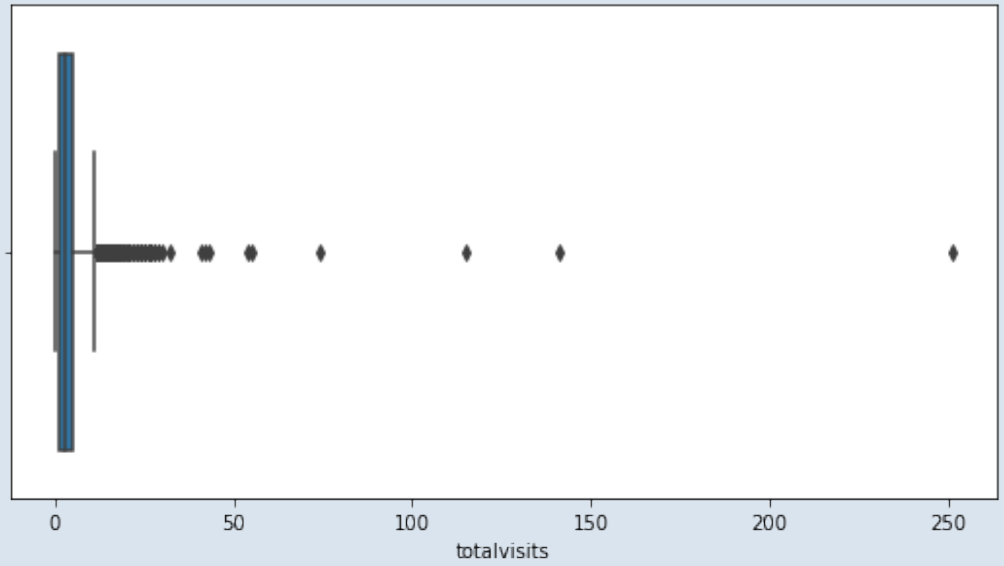


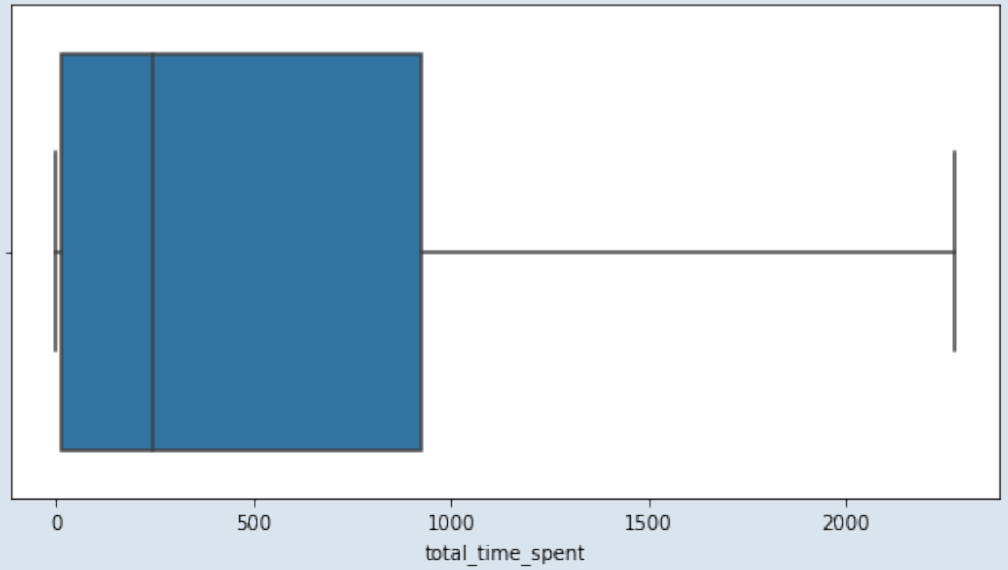


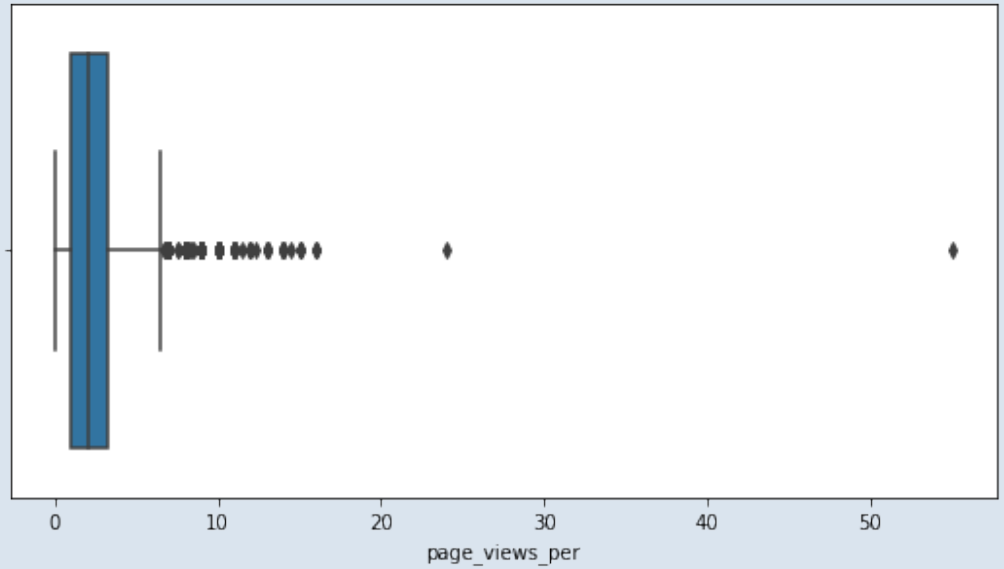
- Majority of the potential customers are from India and the column country has missing values so dropping this column as this will not be much helpful in Analysis.

- We removed outliers by keeping 99 percentile values.

EDA (Univariate Analysis)

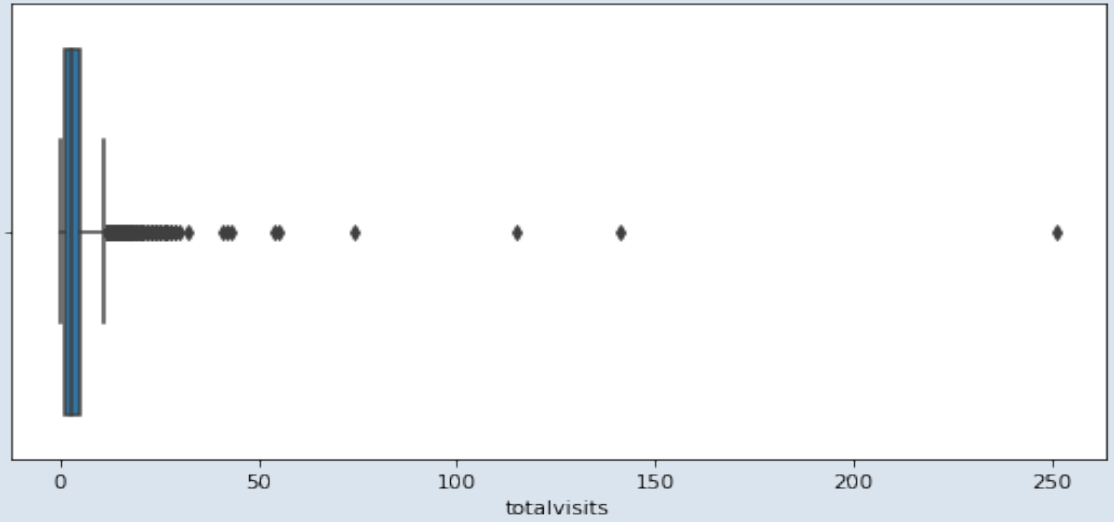


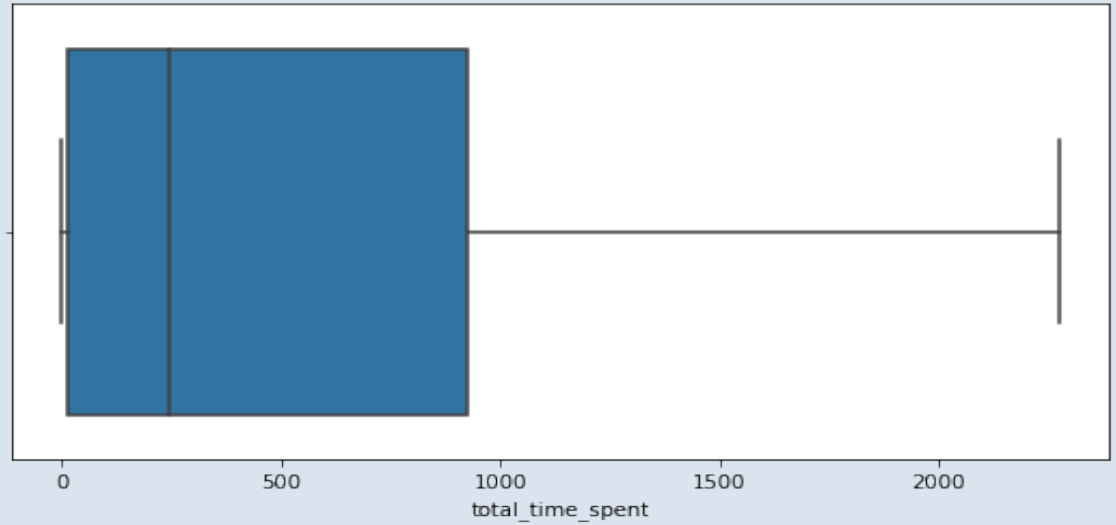


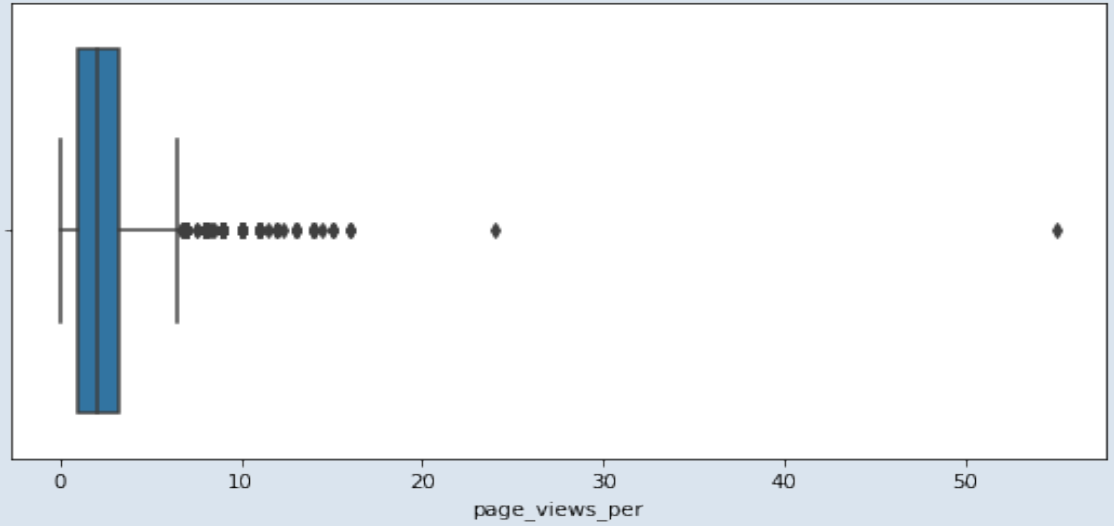


- **We can see that there are outliers in the column "Total_Time_Spent_on_Website”.**
- **We can see that there are outliers present in the column "Page_Views_Per_Visit”.**
- **We can see that there are outliers present in the column "TotalVisits”.**

EDA (Univariate Analysis)

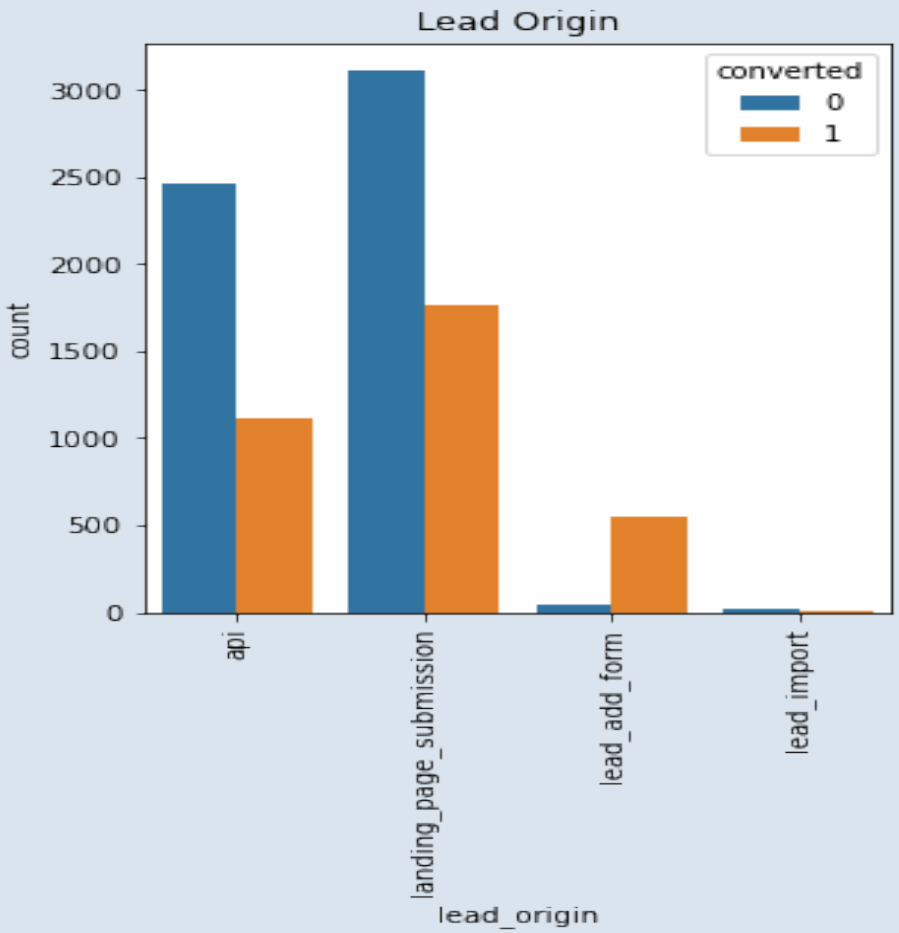


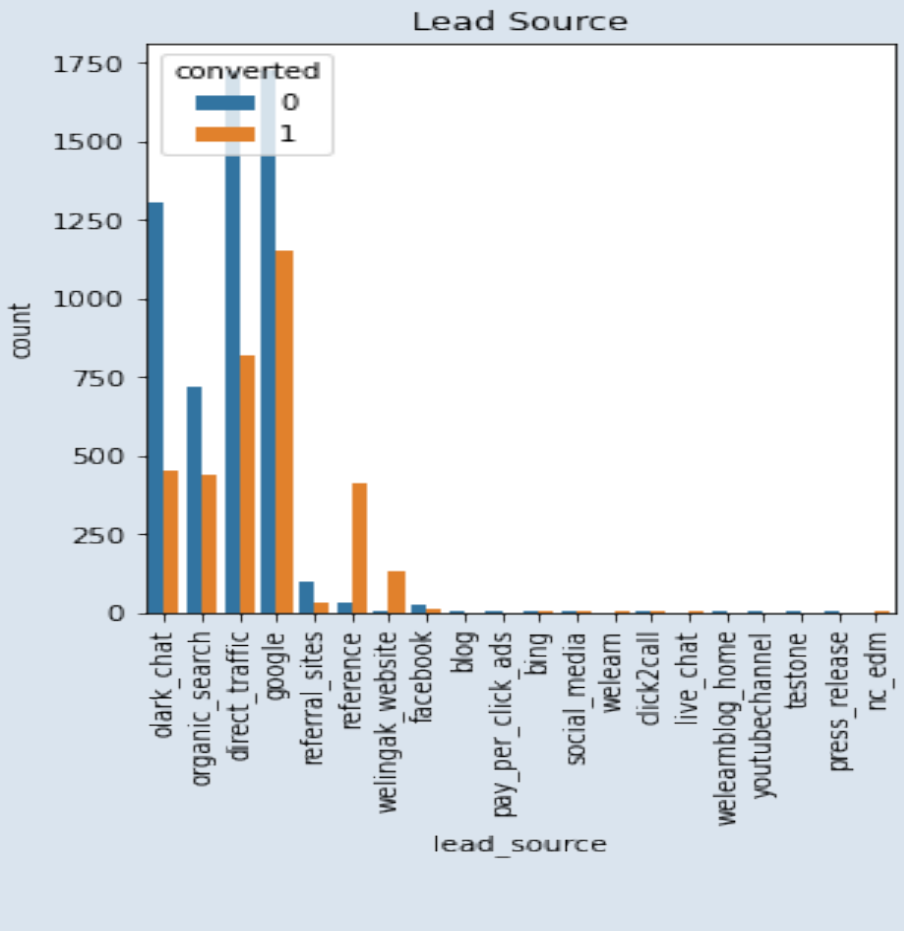


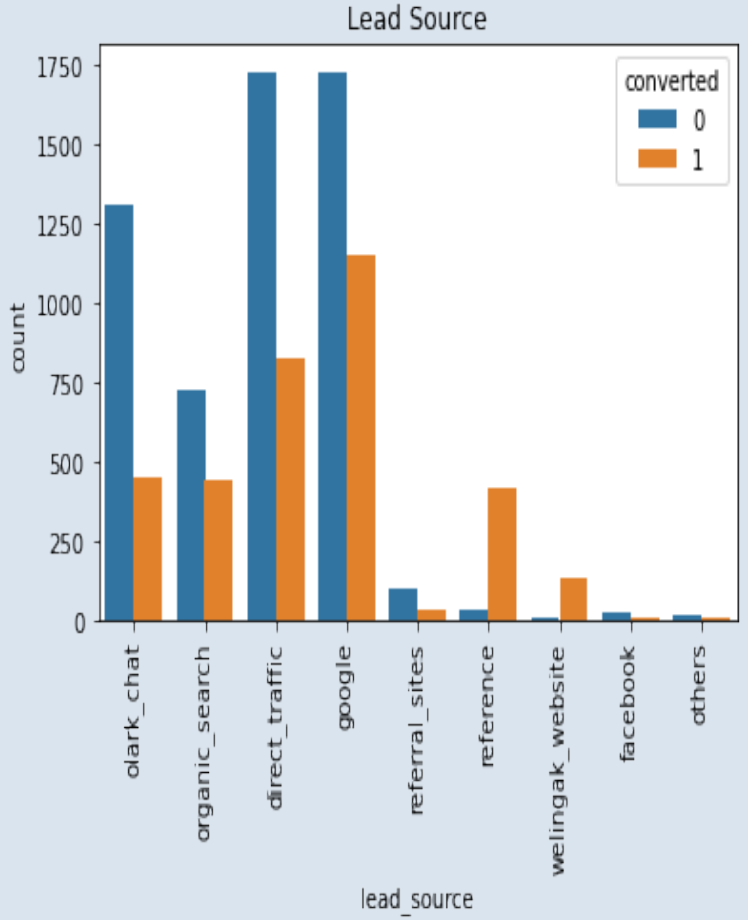


- There is no outliers present in the 'total_time_spent' column.
- Outliers in totalvisit and total_time_spent has been capped to 99 Percentile values.

EDA (Bivariate Analysis)

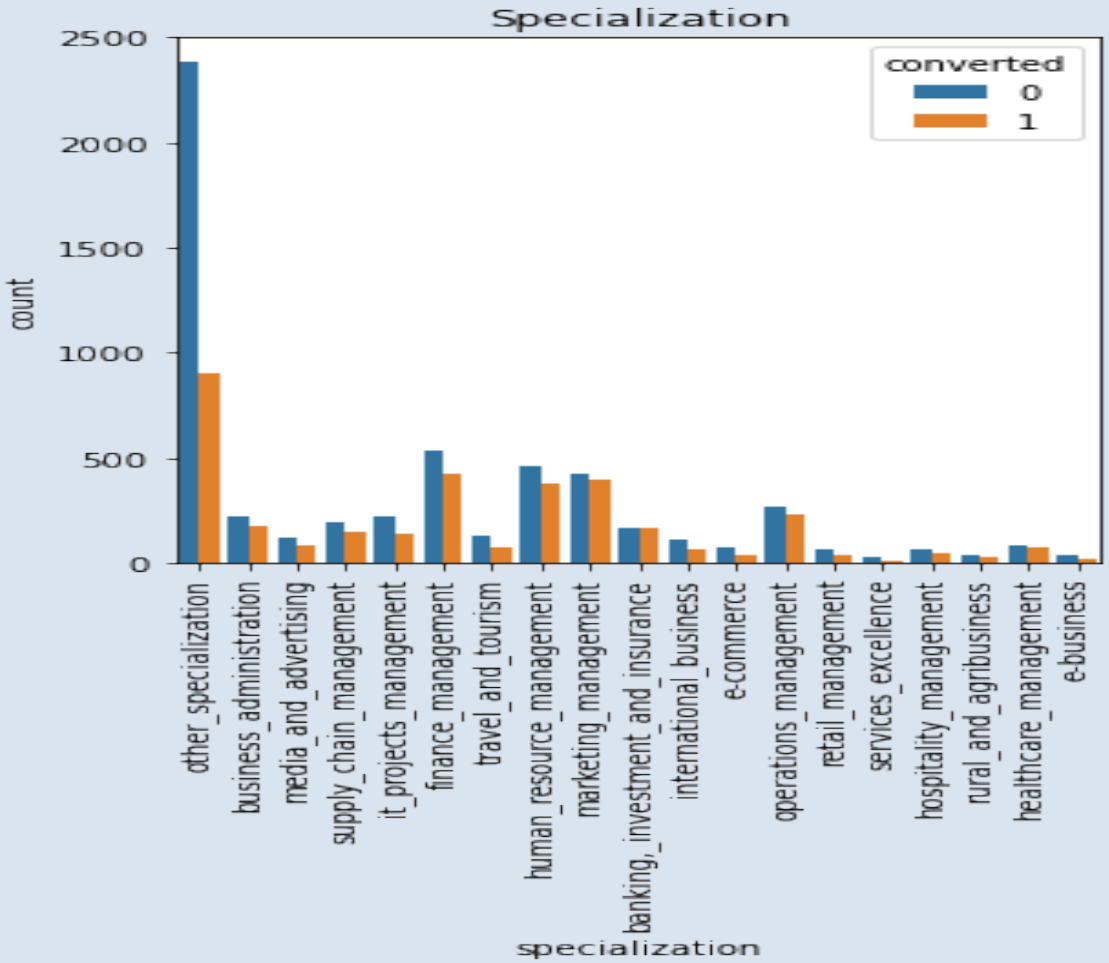


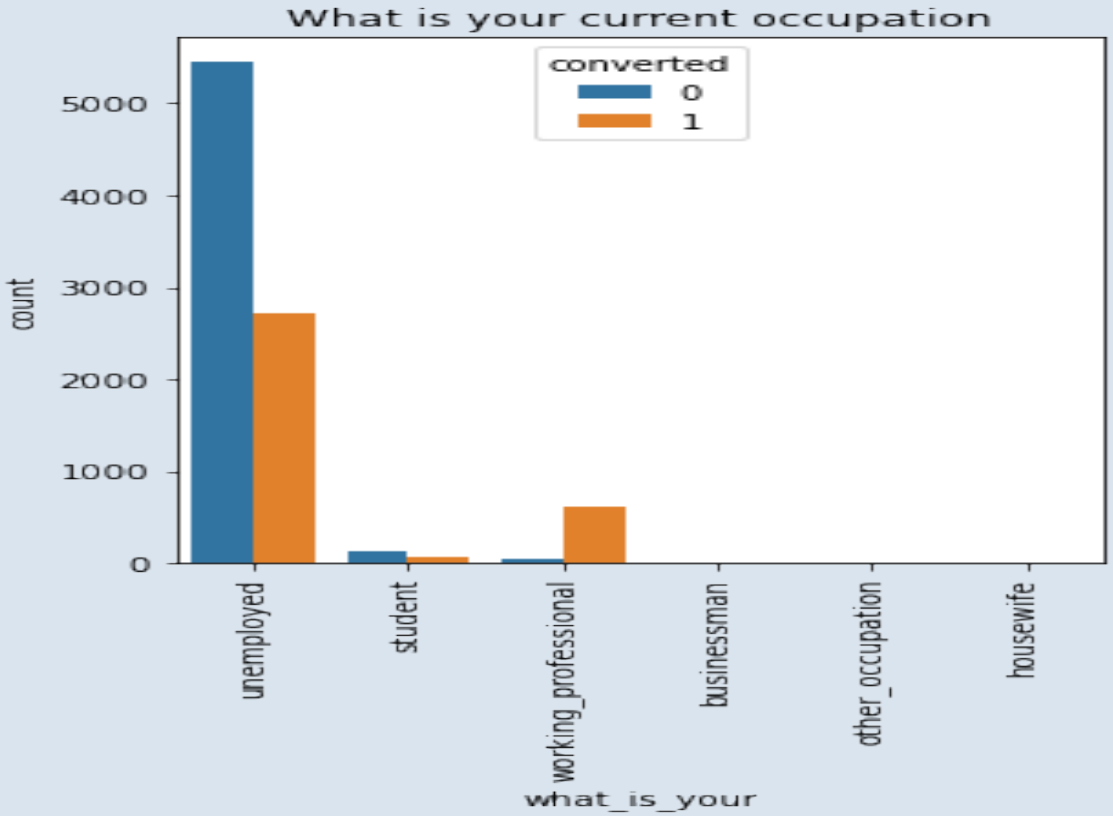




- Maximum number of leads are generated by Google and Direct Traffic
- Many parameters have less than 5 in counts, combined them as others
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead add form are not very high.
- Lead Import are very less in count.

EDA (Bivariate Analysis)





- Most number of leads are of Unemployed customers.
- Above figure shows that the leads of working professional have high conversion rate
- Most leads and conversion rate is high for Other_Specialization

Data Preparation

- Converted categorical columns "Do_Not_Email" and "A_free_copy_of_Mastering_The_Interview" with two levels(Yes and No) into numerical columns.
- Created dummy variables and dropped first and original column.
- Removed repetitive columns.
- Did Correlation Analysis on Features through Heat Map.
- Divided Dataset into train and test dataset.
- Utilized Min-max scaler for feature scaling and standardization.

Model Building

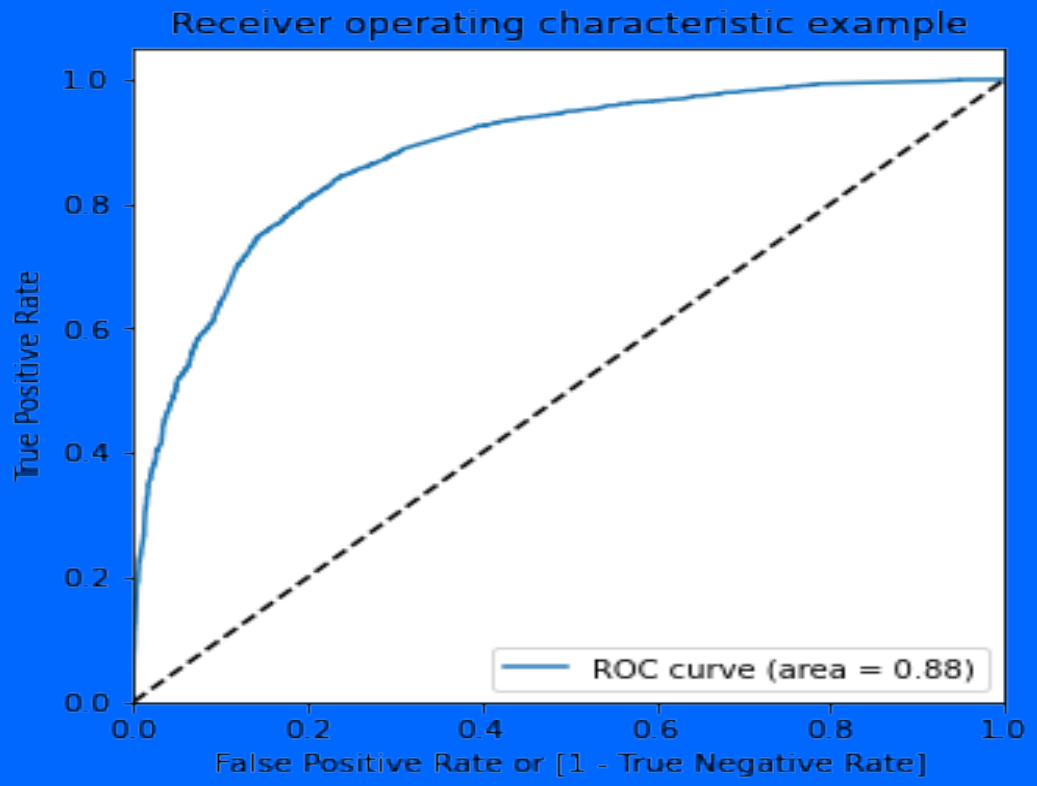
- The logistic regression model is built for Prediction of lead score.
- Used RFE technique for feature selection with 15 variable as output and fined tuned it manually by checking VIF and p-Values.
- Dropped columns “What_is_your_current_occupation_Housewife”, “Lead_Origin_Lead Add Form”, as these features had high p values/VIF.

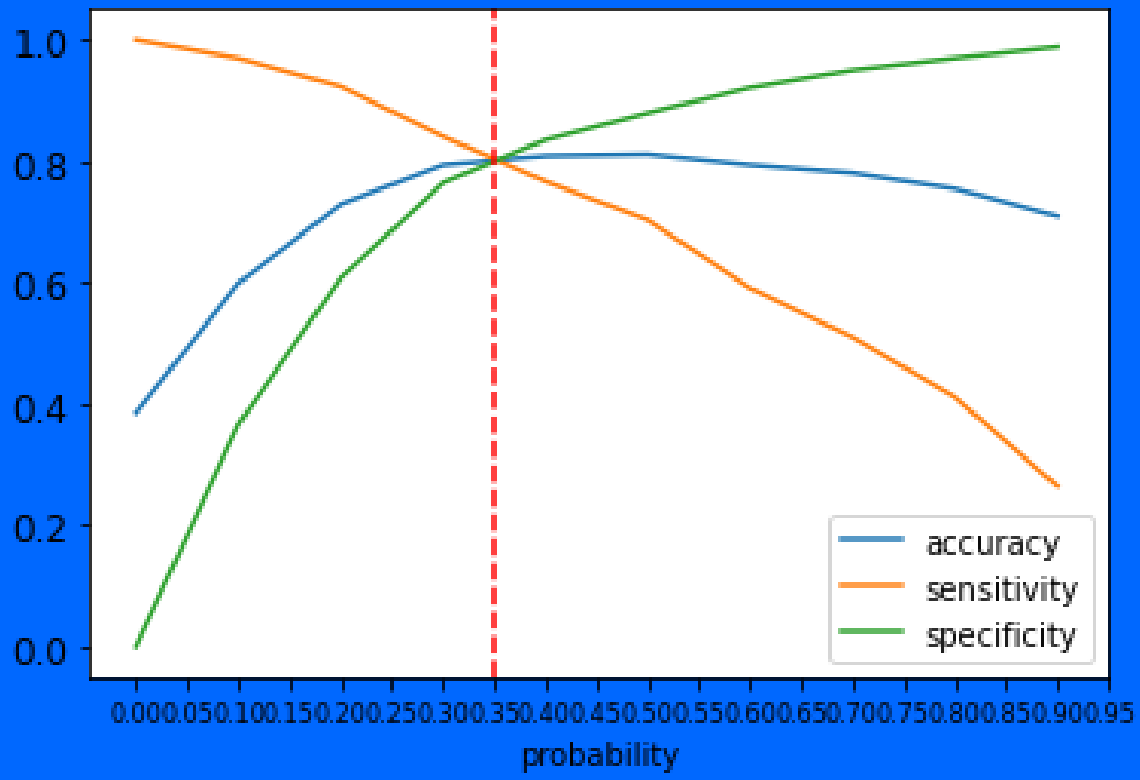
Model Evaluation

- **Evaluated the Model Performance by manually using 0.50 cut-off value.**
- **Using ROC Curve, trade off between Sensitivity and Specificity Parameters, established optimal cut-off value of 0.35.**
- **Ran Prediction for Train and Test Data Set---With the Current cut off of 0.35,**

Metrics Score on Train Data Set are	Metrics Score on Test Data Set are
Accuracy--80.40% Sensitivity--80.38% Specificity--80.41% Precision--72% Recall-80.38% True Positive rate--75.59% False Positive Rate--19.59% Positive Prediction Value--72%% Negative Prediction Value--86.74% F1_Score--75.95%	Accuracy--80.13% Sensitivity--80.18% Specificity--80.10% Precision--69.68% Recall-80.18% True Positive rate--80.18% False Positive Rate--19.89% Positive Prediction Value--69.68%% Negative Prediction Value--87.63% F1_Score--74.56%

ROC Curve





- Here figure shows that the area under the ROC curve is 0.88 which seems good.
- Found optimal cut-off point at 0.35

Summary

It was found that the variables that mattered the most in the potential buyers are (Almost In descending order) :

- The total time spend on the Website.
- When the lead source was:
  a. Welingak website
  b. Reference
  c. Olark chat
- When the last activity was:
  a. Phone Conversation
  b. SMS
  c. Olark chat conversation
- When their current occupation is as a working professional.
- When the lead origin is Landing page Submission

With all these Observations X Education will be easily able to identify potential leads and convert them into successful buyers.

Thank you