LEAD SCORING CASE STUDY

Deepak Thambi | Sneha Mathew

PROBLEM STATEMENT

X Education who is an edtech company sells online courses to industry professionals. Though they get a lot of leads, their lead conversion rate is quite poor. The typical lead conversion rate at X education is around 30%.The company has approached us to improve the lead conversion rate by identifying leads which have the highest potential to get converted.

The company wants us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

SOLUTION APPROACH

DATA CLEANING

Check and handle missing, duplicate and outlier data by either dropping or imputing the values.

EDA

Perform Univariate and Bivariate data analysis to identify pattern between variables and the target variable.

SCALING

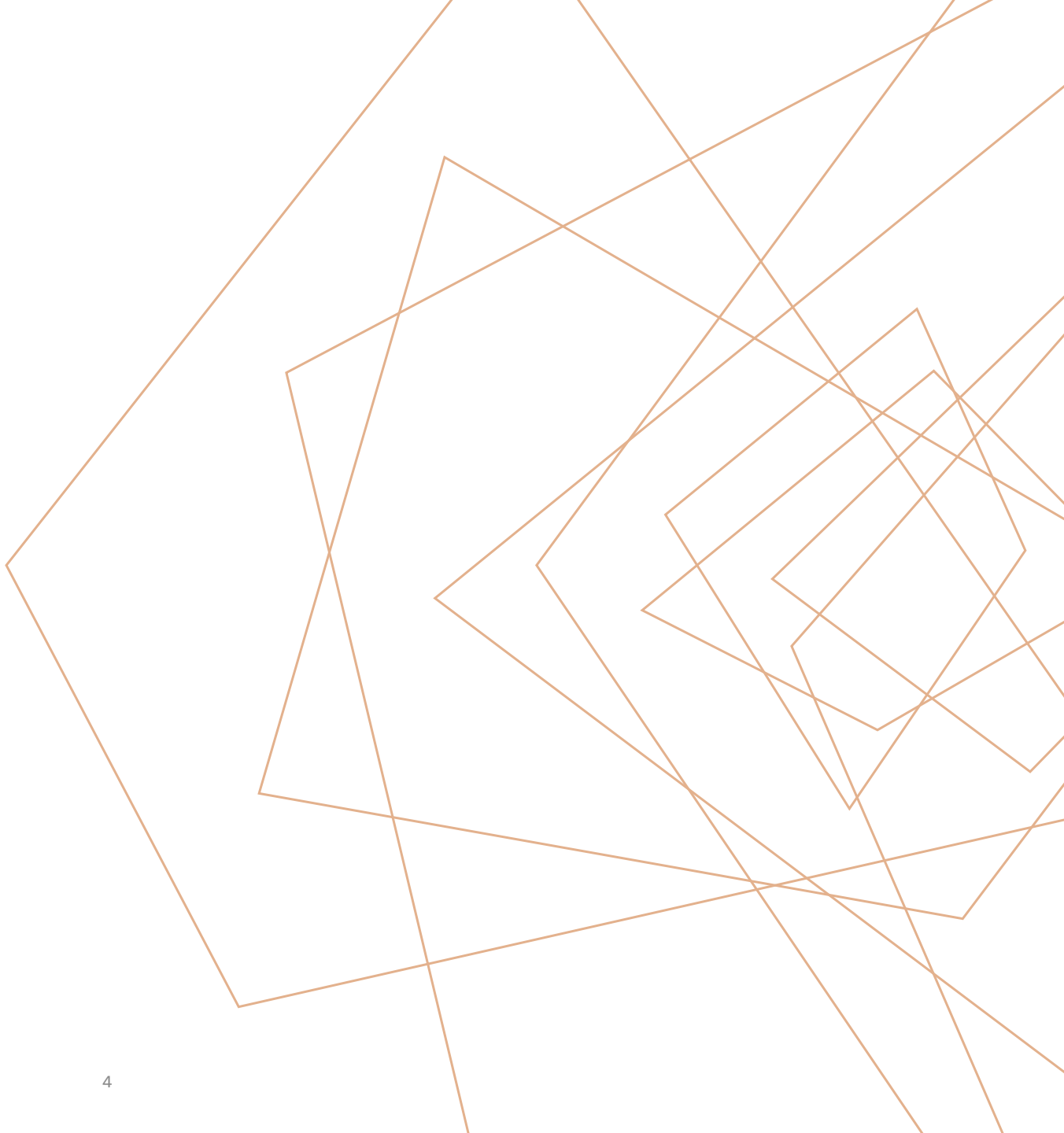
Scale the variables using MinMaxScaler method. Additionally create dummy variables.

CLASSIFICATION

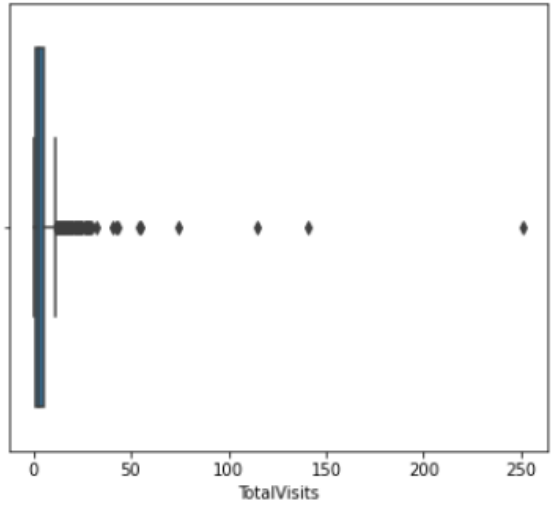
Logistic Regression will be used for model building and making prediction.

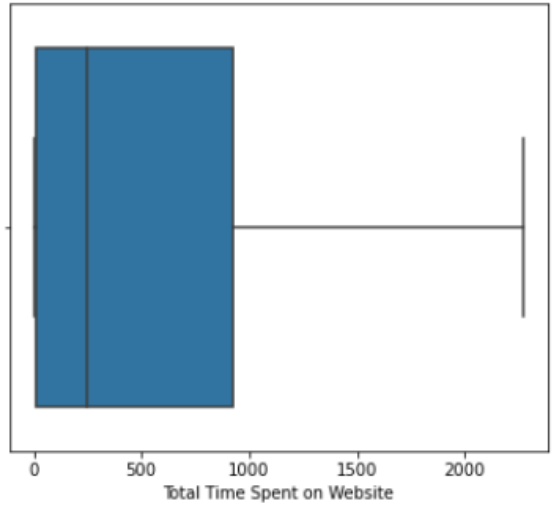
DATA CLEANING

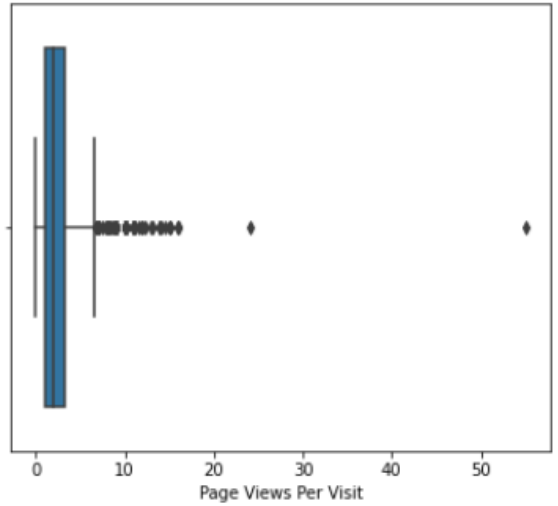
- There are 9240 rows and 37 columns in the data set.
- All the data was converted to lower case.
- Replaced value as 'Select' with ‘NaN’ as it looks like there was an option and nothing was selected.
- Columns that have 40% or more null values were dropped.
- Removed prospect id column since the values are unique for all.
- Dropped columns that have a single unique value.
- To handle columns which had less than 40% missing values, to prevent data loss, imputed them as ‘No info’.



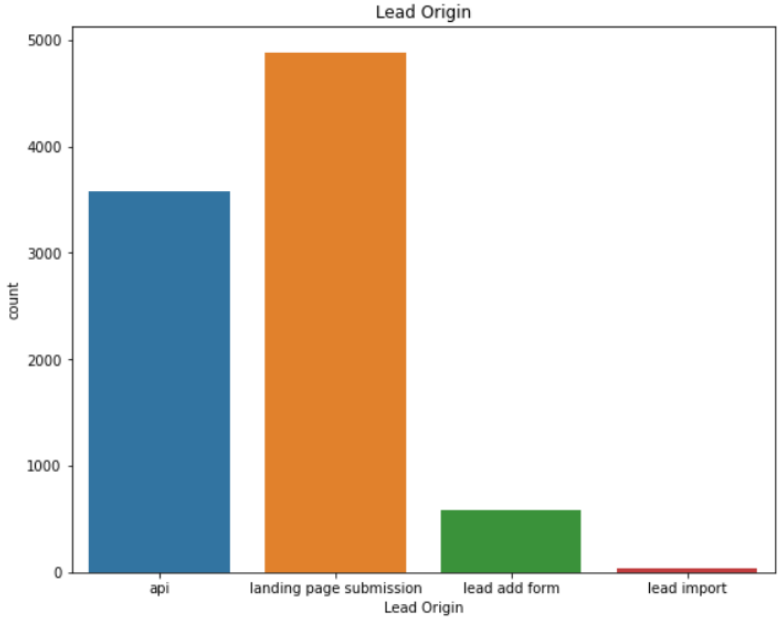
EXPLORATORY DATA ANALYSIS | OUTLIER CHECK

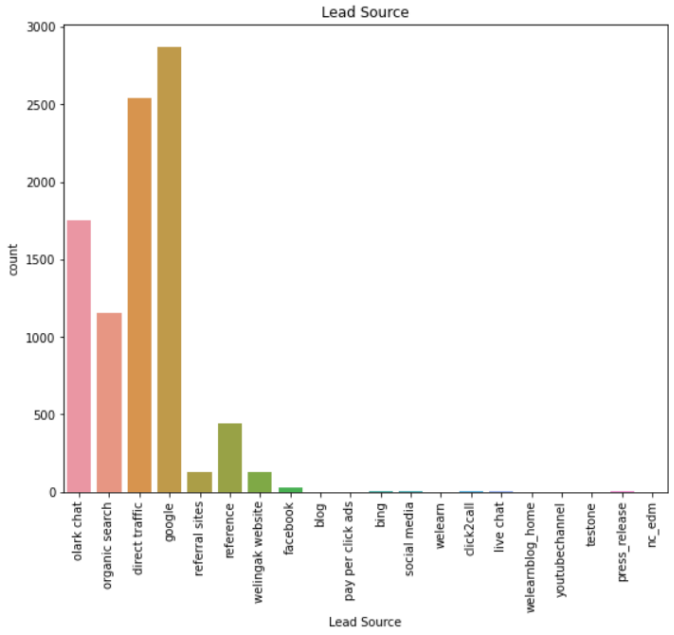


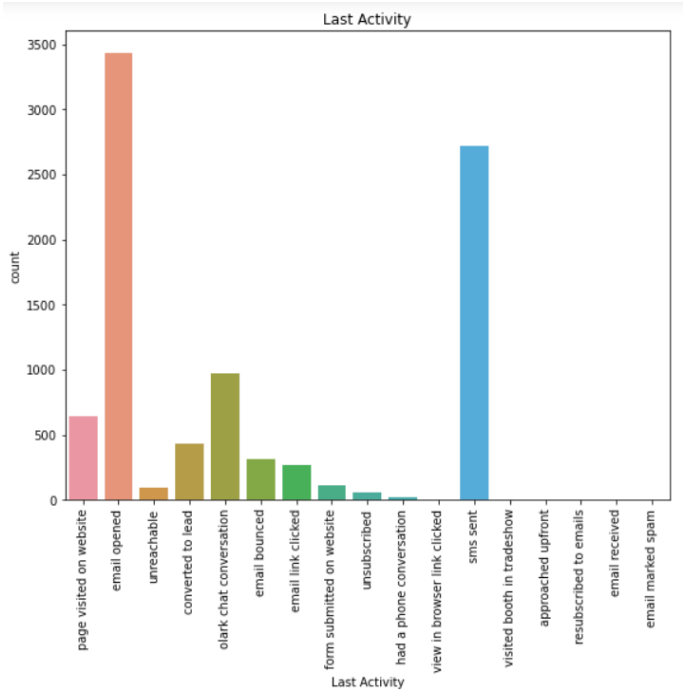




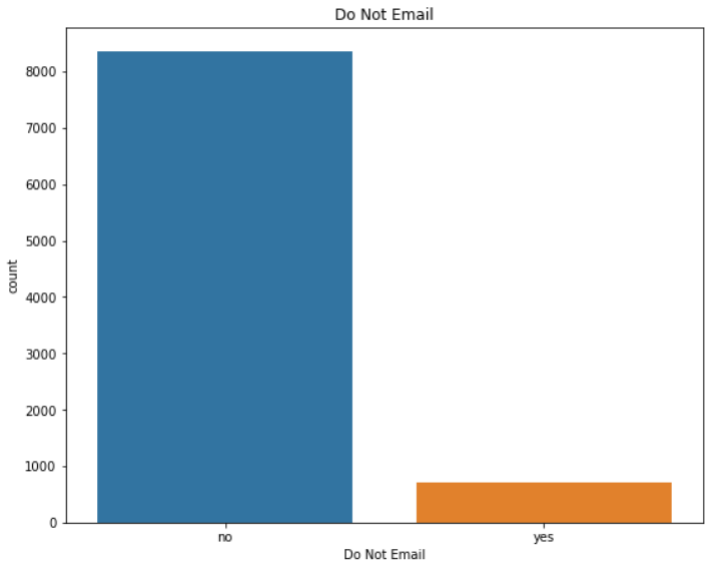
EXPLORATORY DATA ANALYSIS | UNIVARIATE

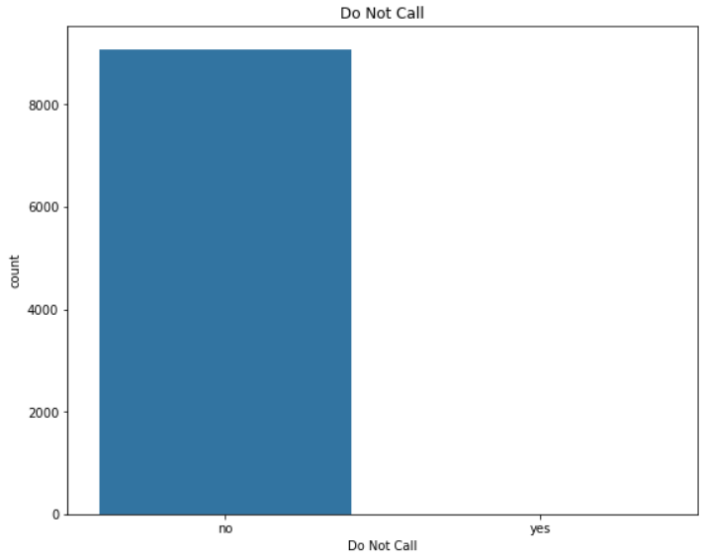


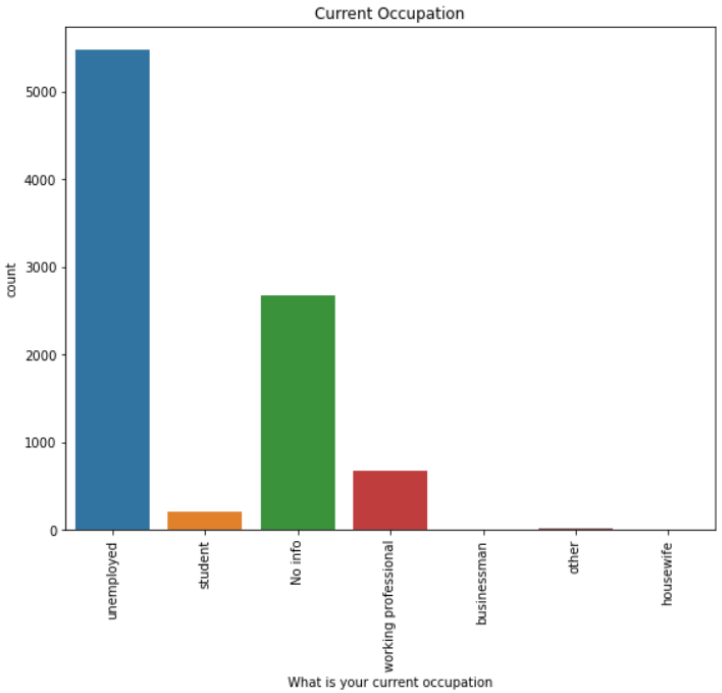




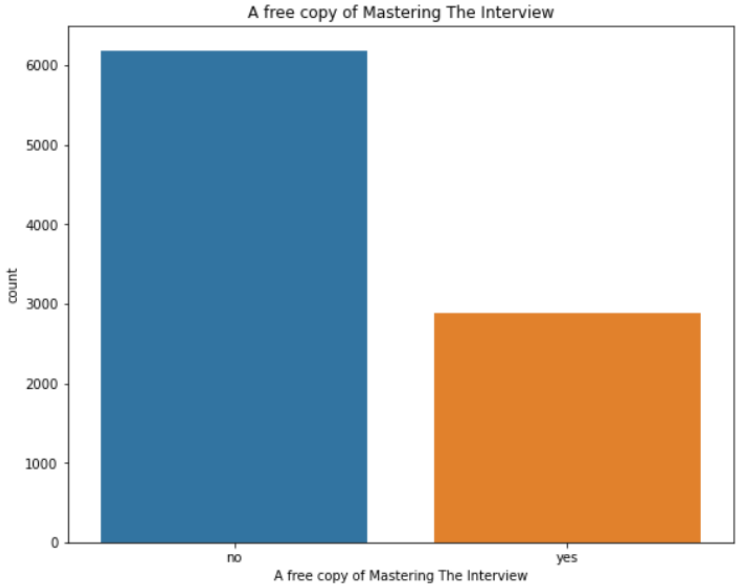
EXPLORATORY DATA ANALYSIS | UNIVARIATE

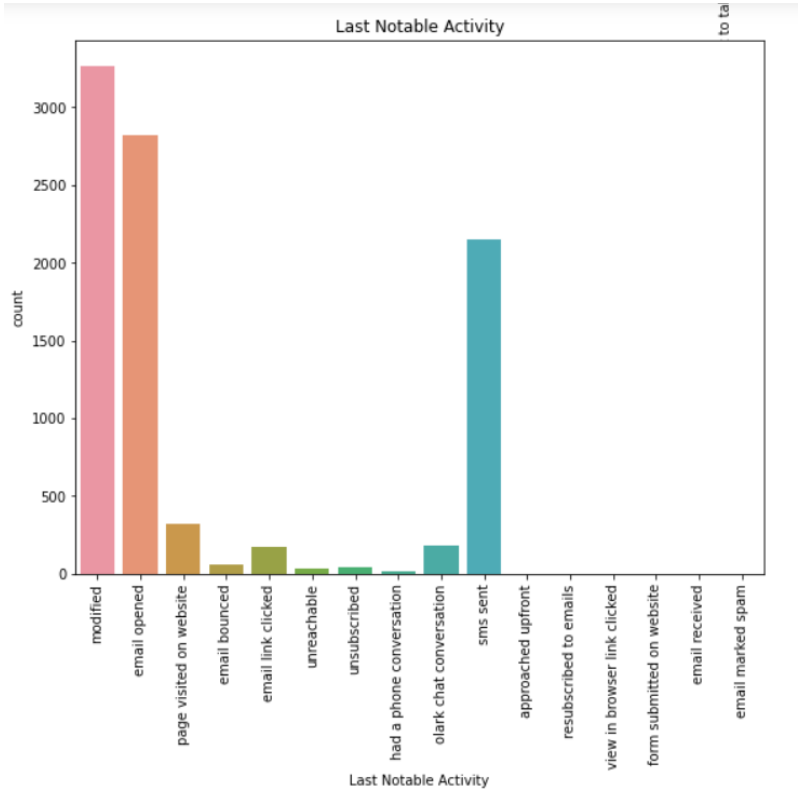


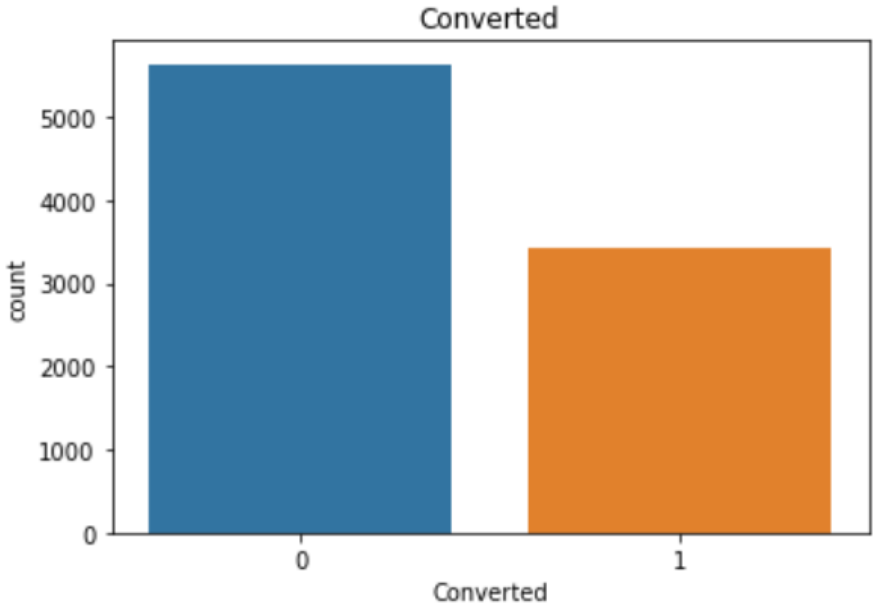




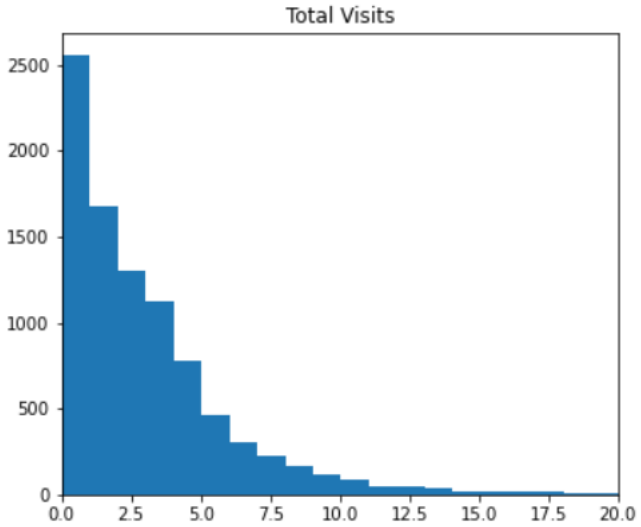
EXPLORATORY DATA ANALYSIS | UNIVARIATE

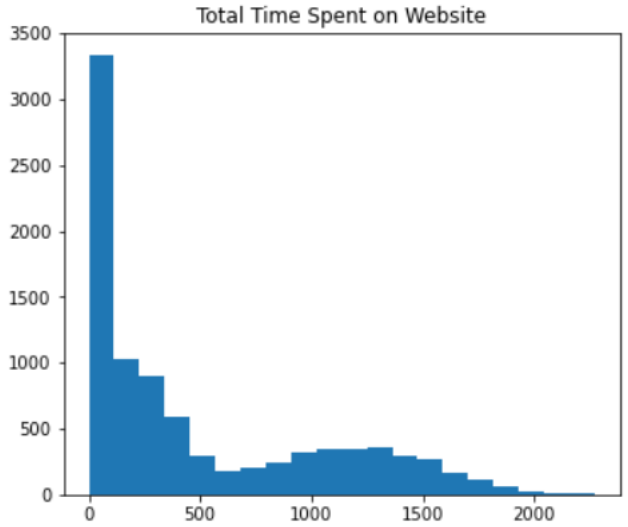


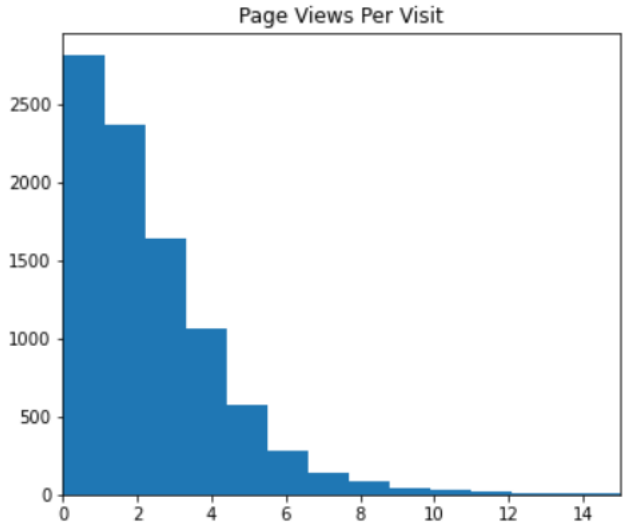




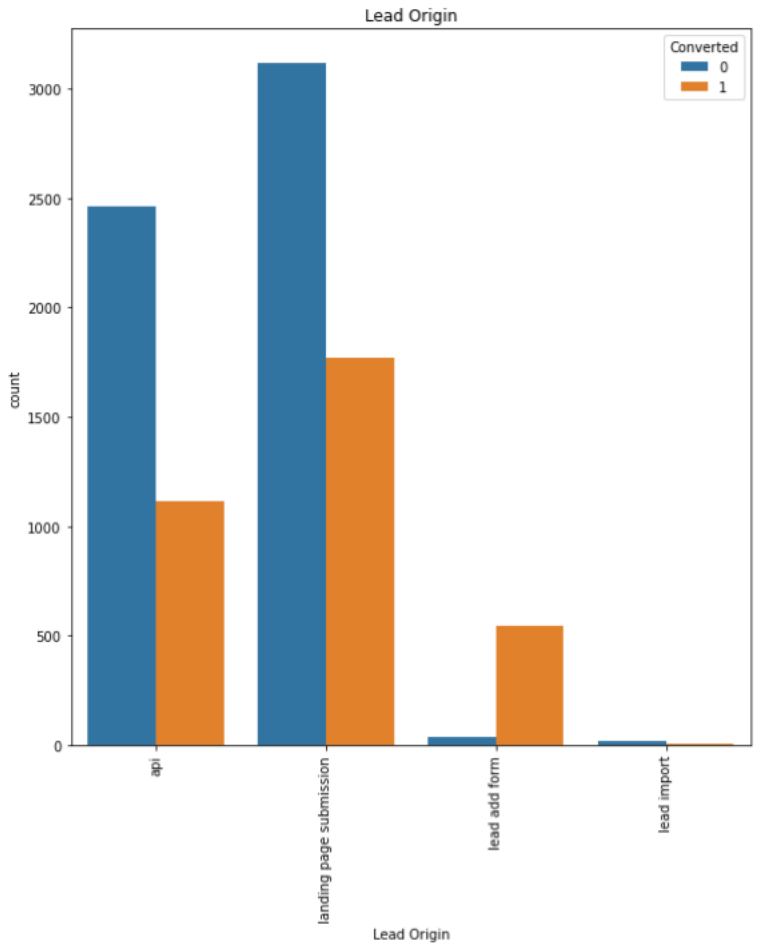
EXPLORATORY DATA ANALYSIS | UNIVARIATE

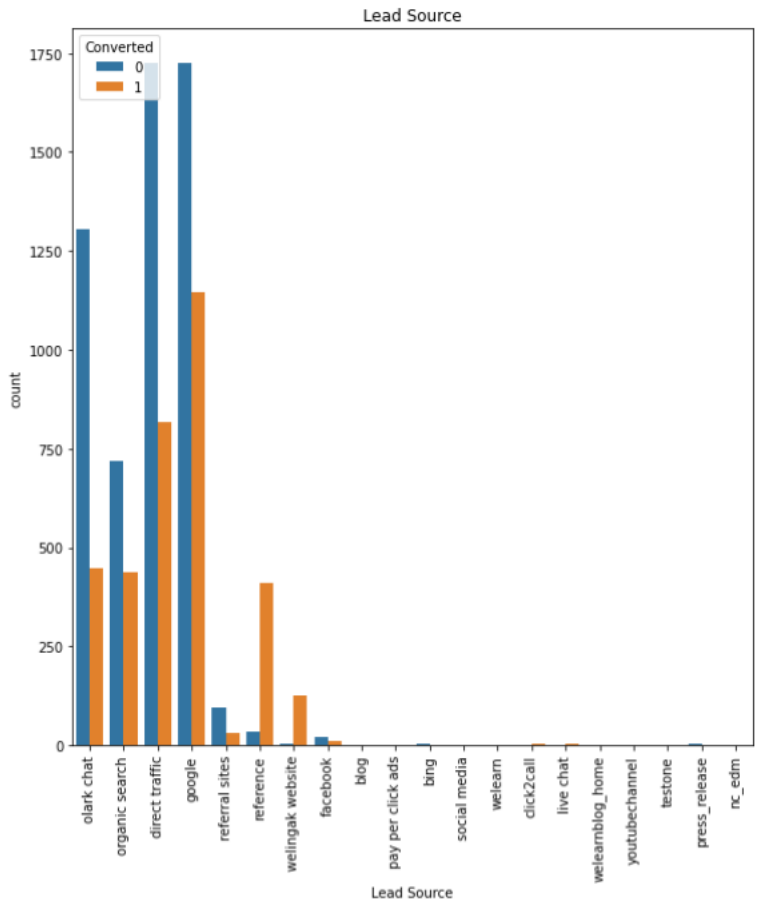




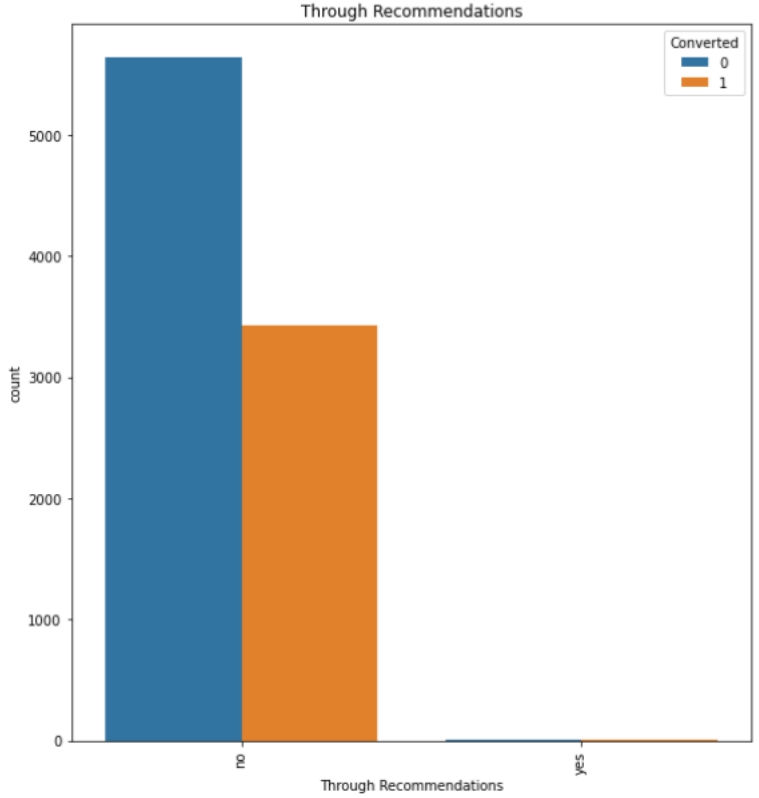


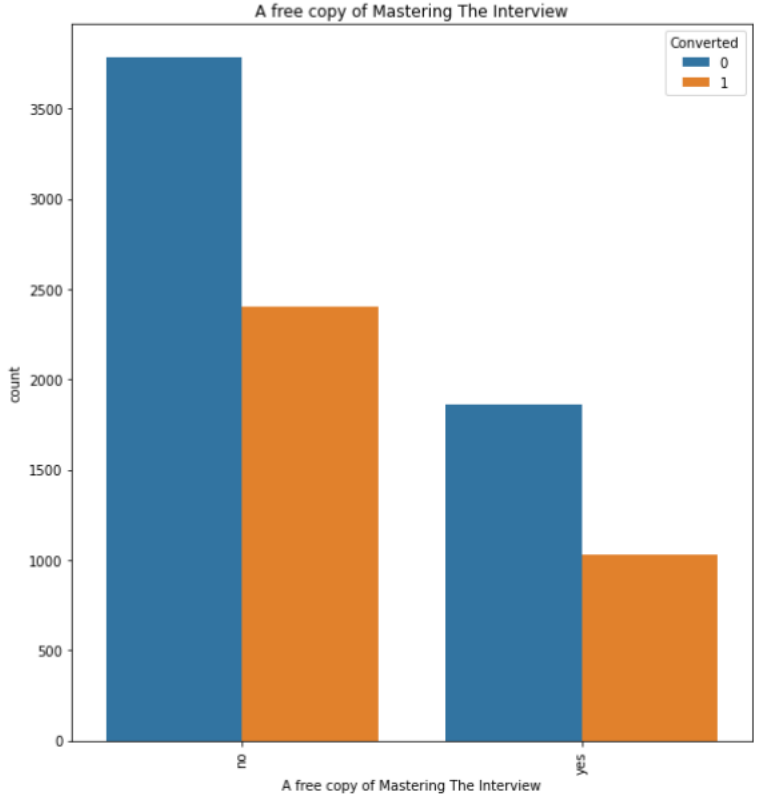
EXPLORATORY DATA ANALYSIS | BIVARIATE



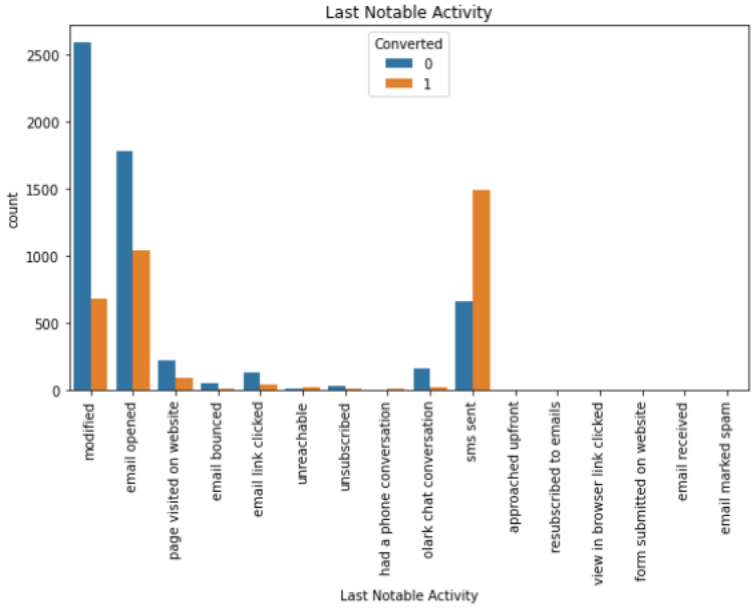


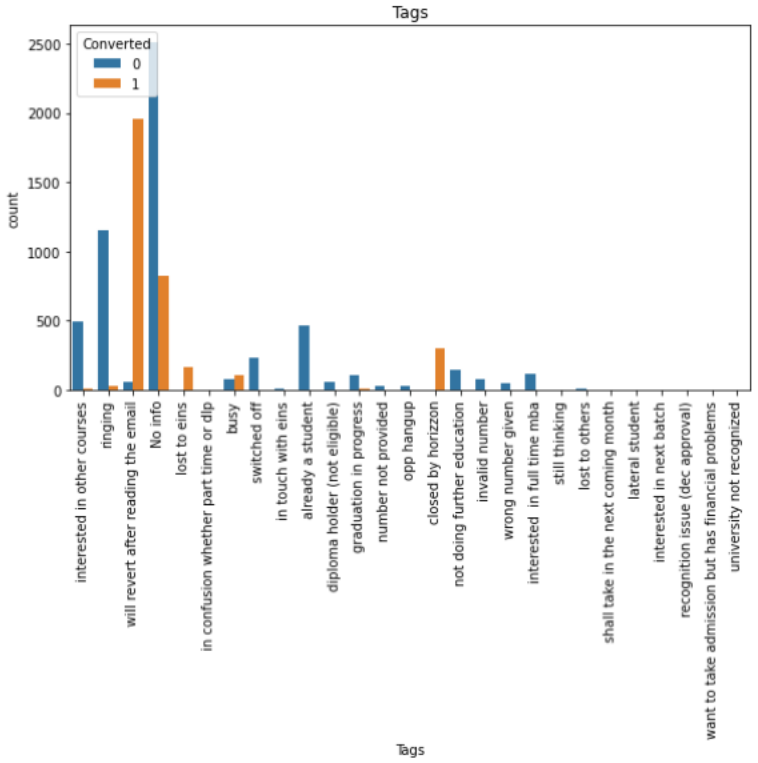
EXPLORATORY DATA ANALYSIS | BIVARIATE





EXPLORATORY DATA ANALYSIS | BIVARIATE





EXPLORATORY DATA ANALYSIS | HEATMAP

DATA CONVERSION

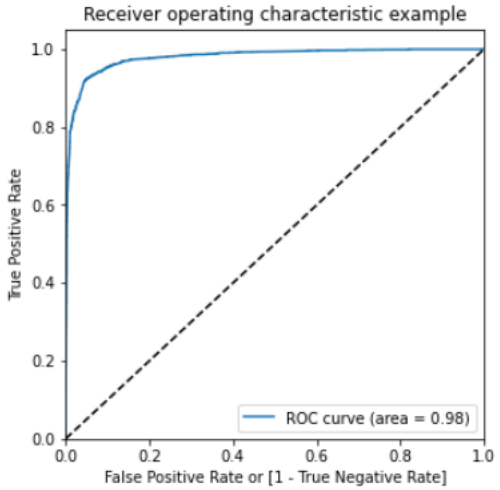
Numerical variables are normalized and dummy variables were created for object type variables (categorical).

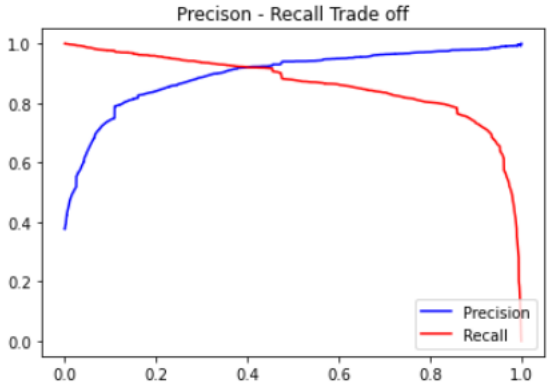
MODEL BUILDING

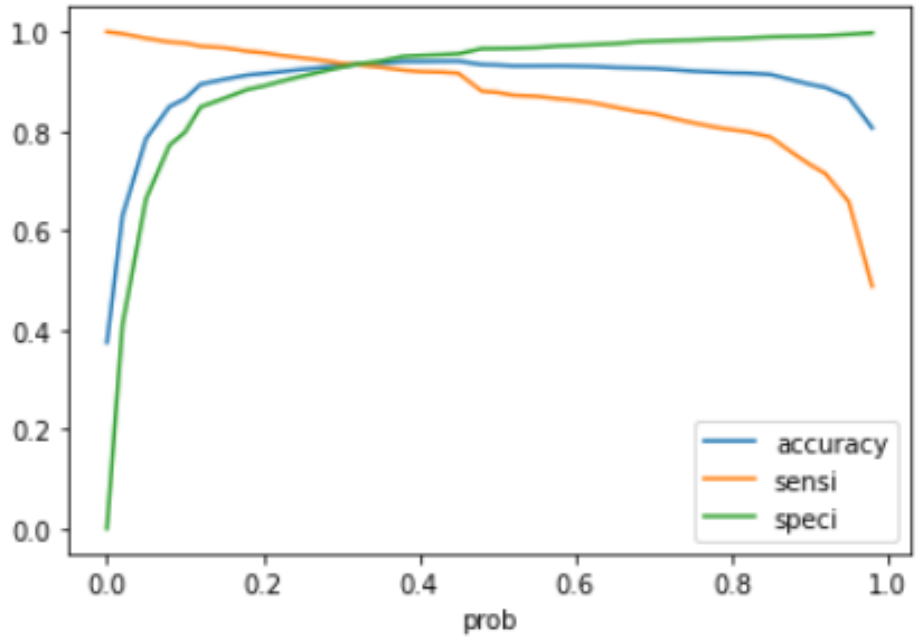
- The data was split into test and train. Test size was decided as 33% and train size as 67%.
- The numeric columns were then scaled.
- Optimal number of features was estimated to be 28 which gives 93.7% accuracy.
- RFE was used to arrive at those 28 features.
- Features with VIF more than or equal to 5 was dropped. Additionally features with p value greater than 0.05 were dropped.

ROC CURVE

```
Sensitivity (Recall): 0.8779631255487269
Specificity: 0.9663246514075243
Precision: 0.9398496240601504
F-Score: 0.9078529278256923
```








```
Optimum cut-off value is: 0.32
```

CONCLUSION

- Prediction was done on test data with accuracy, sensitivity, specificity, precision and F-score at approx. 90%. Precision-Recall cut-off was found to be ~ 0.45.
- Some of the features that matter the most which company X should leverage are:

1. Total time spent on the website.
2. Total number of visits.
3. Last activity was SMS.
4. Source is Welingak website
5. Country is Germany
6. Tags closed by horizon