LEAD SCORE CASE STUDY

By

Kiran Vutukuri

Tony C

Case Study Background

- X education is an education company that sells online courses
- Courses marketed on websites like Google
- The company website provided details of the courses
- Interested customers may browse the courses or fill up a form for the course or watch some videos
- When persons fill up a form with email address or phone number, they are classified to be a lead
- Sales team communicate to leads through calls, SMS and emails
- Leads who finally join the course are considered converted
- The typical lead conversion rate at X Education is around 30%

Problem Statement and Solution

- **Current Problem**
  - Lead conversion rate is very poor for X Education
- **Problem Statement**
  - To make Lead conversion process more efficient,
  - Identify the most potential leads as ‘Hot Leads’
  - Enable sales team to focus on most promising leads
  - Increase the lead conversion rate

Assumptions

- There will a unique record for each lead

Problem Statement and Solution

- **Objective of Solution**
  - Build a Machine Learning model to assign a lead score
  - Customers with higher conversion chance should be given higher lead score
- **Criteria for selection of Hot Leads**
  - Model should identify hot leads based on the lead score.
  - The more accurate the selection of the hot leads, the more chances of higher conversion ration.
  - Target of 80% conversion rate with high accuracy in obtaining hot leads.
- **Target Variable**
  - Converted

Approach to Solution

- Data Gathering
- Reading & Understanding the data
- Data Cleaning
- Performing EDA
- Splitting the data into test & train dataset
- Prepare the data for modelling
- Model building
- Model evaluation-specificity & sensitivity or precision recall
- Making predictions on the test set.
- Model improvement
- Final Model
- Lead score calculation

Data Analysis

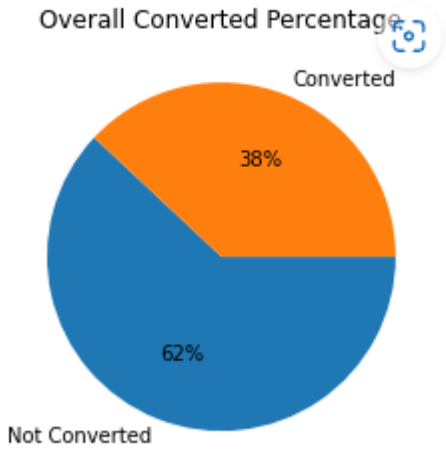
- **Final % records retained for Model Creation**
  - 9130 records from 9240 – 98.52%

- **Data Treatment and Cleaning**

Data Treatment Step	Sample Feature list	Action and criteria
Handling Missing value	Do Not Email/Call, Country, Search, Magazine	Dropping features when > 50%
Handling data imbalance	Occupation	Dropping features, value > 85%
Combining Categorical Values	Specialization, Tags, Lead Source	Combine values when less than 5%
Outlier treatments	TotalVisits, Total Time Spent on Website, Page Views Per Visit	Capping maximum values to 99 percentile

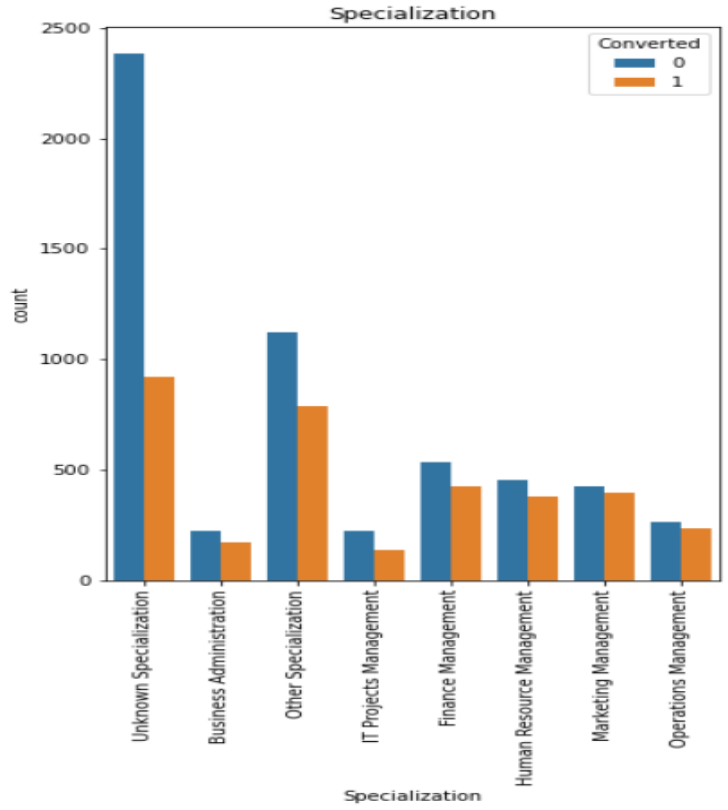
EDA – Univariate Analysis

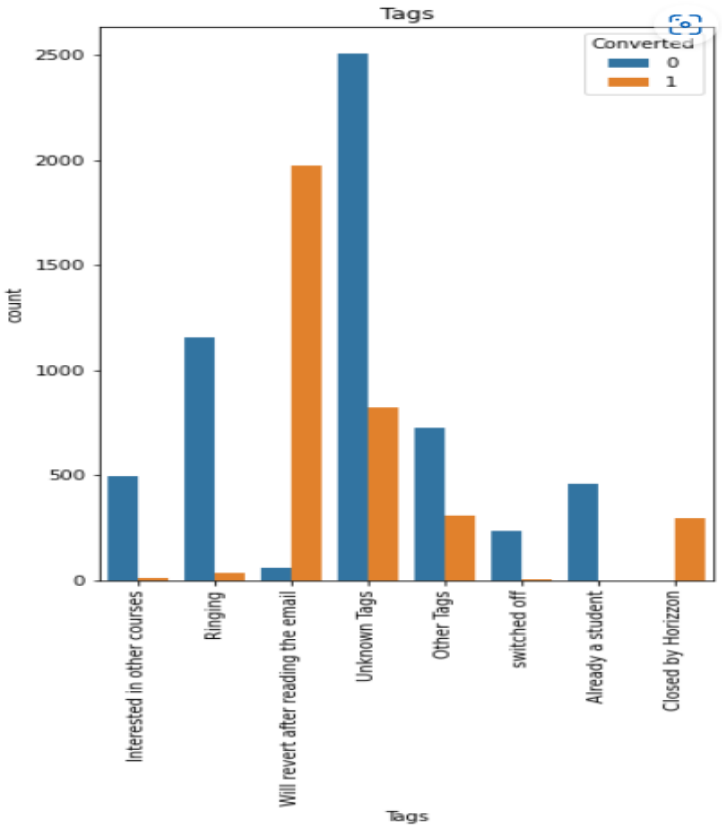
- **Conversion Rate**



EDA – Bivariate Analysis

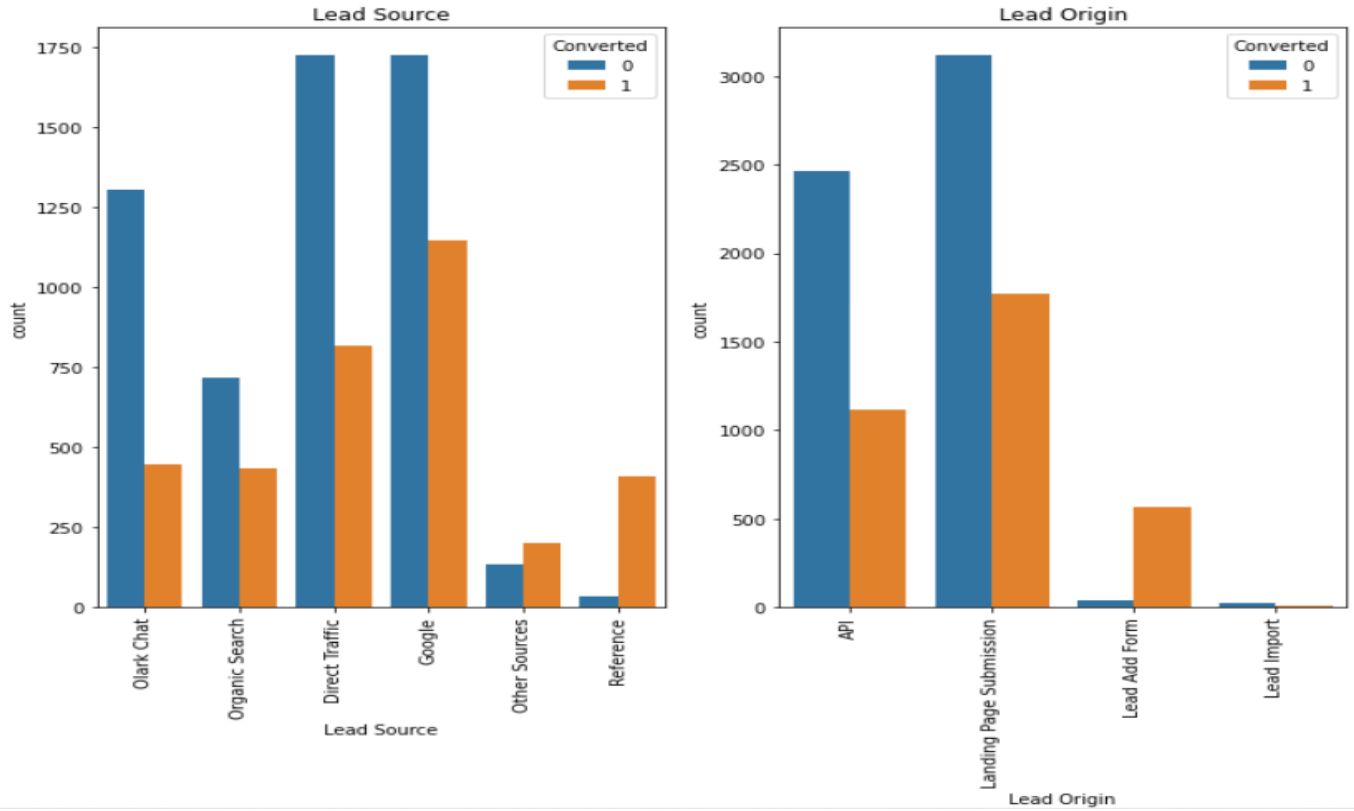
- Specialization And Tags





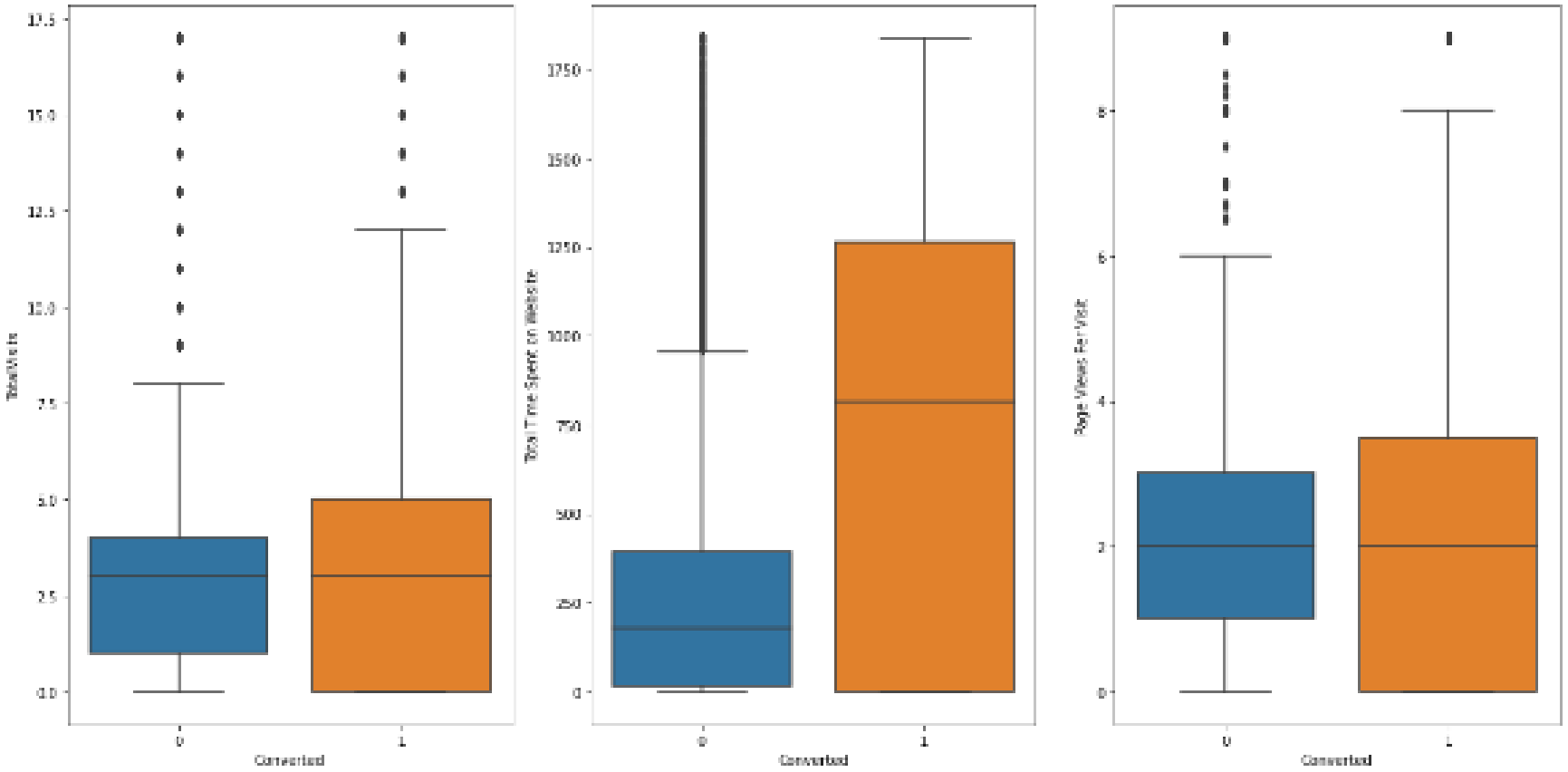
EDA – Bivariate Analysis

- Lead Source and Lead Origin



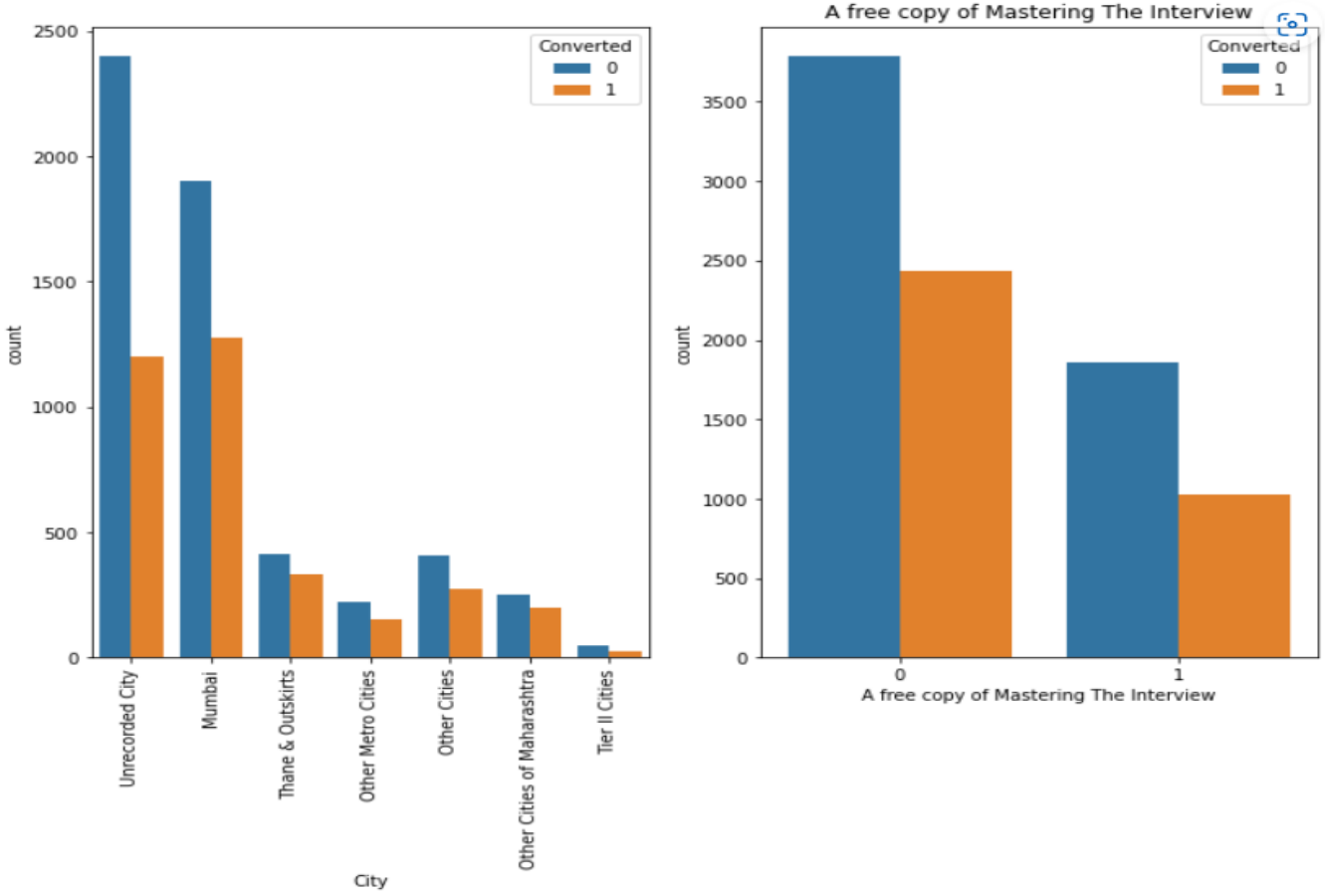
EDA – Bivariate Analysis – Box plots

- Total visits , Total Time Spent on Website and Page Views Per visit



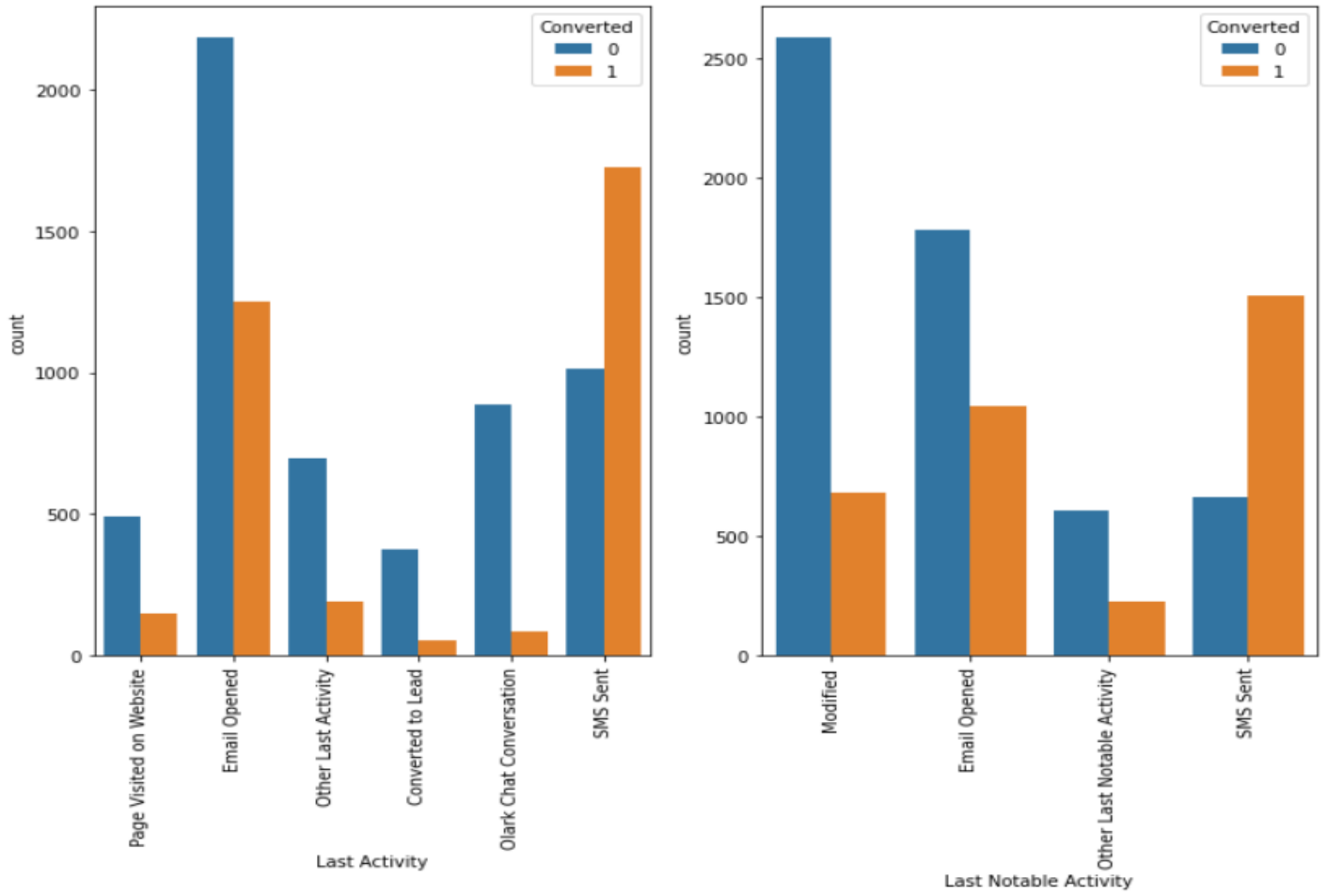
EDA – Bivariate Analysis

- City and "A free copy of Mastering The Interview"



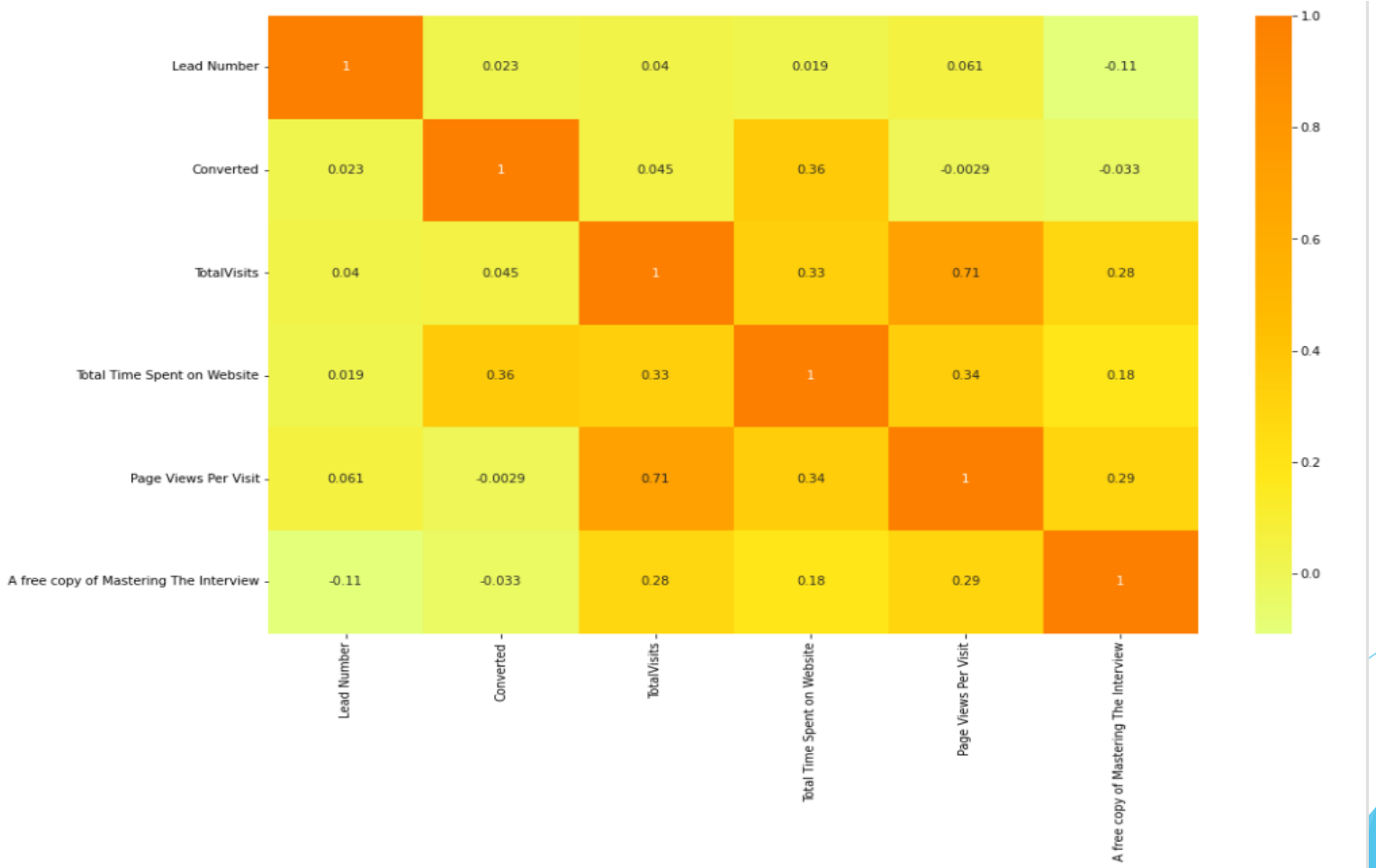
EDA - Bivariate Analysis

- Last Activity and Last Notable Activity



EDA – Heat Map for Correlation

- Variables are not highly correlated with each other but there is multicollinearity among some features.



Data Preparation for Model creation

- Dummy Variable Creation
  - Lead Origin
  - Lead Source
  - Specialization
  - Tags
  - City
  - Last Activity
  - Last Notable Activity
- Test Train Test Data Set
  - 70 / 30 % of train and test
- Scaling of features with Continuous Values
  - MinMax Scaler

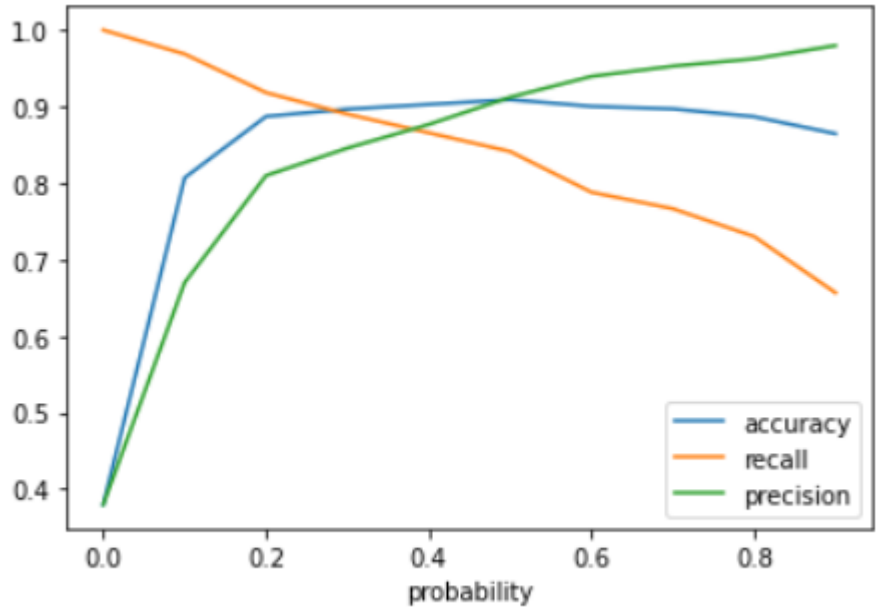
Logistic Regression Model Creation

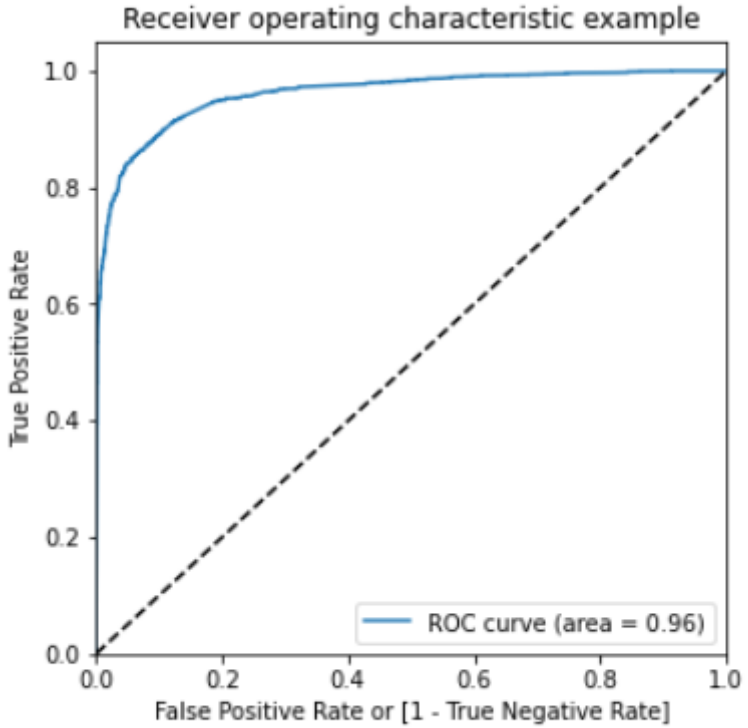
- Approach to Model Creation
  - RFE for key feature selection
  - Logistic Regression
  - VIF for dropping highly corelated features
- Most important features of final Model with Coefficient values
  - Tags_Closed by Horizzon 8.392270
  - Tags_Will revert after reading the email 6.938861
  - Lead Origin_Lead Add Form 5.461889
  - Total Time Spent on Website 3.666934
  - Tags_Other Tags 2.831843
  - Lead Source_Reference -2.822079
  - Last Notable Activity_SMS Sent 2.690800
  - Tags_Unknown Tags 2.505002
  - Lead Source_Olark Chat 1.340468

Model Evaluation - Recall & Precision

For Cutoff of 0.4 probability

- Accuracy - 90%
- Sensitivity/Recall - 87%
- Precision - 88 %





Lead Score

- Lead Score Calculation
  - Lead score = Conversion Probability * 100

```
============ Final Lead Score with Lead Number and Conversion Probability ============
```

:

	Lead Number	Converted	Converted_Prob	Lead score
0	0	0	0.007523	1.0
1	1	0	0.009960	1.0
2	2	1	0.991561	99.0
3	3	0	0.001616	0.0
4	4	1	0.974904	97.0
5	5	0	0.084929	8.0
6	6	1	0.993185	99.0
7	7	0	0.084929	8.0
8	8	0	0.070391	7.0
9	9	0	0.075119	8.0

Recommendations

- The three most important features/variables for conversion
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email
  - Lead Origin_Lead Add Form
- Lead score indicates most promising leads. High lead score – hot leads
- For overall best conversion results, lead score cut-off 40