DATA SCIENCE

Lead Scoring Case Study

MOHD ZAID

ASHISH DIXIT

AGUM MALIK

Hello !

Let's get started with our findings on Lead Score Case Study

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature
Mohd Zaid
Ashish Dixit
Agum Malik

CERTIFICATE

This is to certify this Presentation along with Jupyter Notebook entitled "**LEAD SCORING CASE STUDY**" which is submitted by **Mohd Zaid, Ashish Dixit** and **Agum Malik** in partial fulfillment of the requirement for the award of **Executive Post Graduate in Data Science**, to The International Institute of Information Technology, Bangalore, is a record of the candidates' own work carried out by them under our supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Supervisor Name

upGrad.com

DISCLAIMER

This presentation report is summary based on the research done on **Lead Scoring Case Study** given by **IIIT-B** in association with **upGrad.com**.

While every effort is made to ensure the accuracy and completeness of information contained within this summary presentation report, the students of upGrad, **MOHD ZAID, ASHISH DIXIT** and **AGUM MALIK** has prepared this report with utmost honesty and sincerity, maintaining the dignity of the data analysis research report profession, without involving in any plagiarism and genuinely maintained the integrity of the original work.

The students **MOHD ZAID, ASHISH DIXIT** and **AGUM MALIK** takes no responsibility and assumes no liability for any error/omission or accuracy of the information as their work is totally based on the observations and analysis done on the report **LEAD SCORING** which was allotted to them by IIIT-B in association with upGrad. Recipients of this summary report should rely on their own judgments and conclusions from relevant sources before making any critical comment and final evaluation of this honest work done by these dedicated students.

INTRODUCTION

1

INTRODUCTION

The Case Study of Lead Scoring aims at understanding the meaning of Logistic Regression Model which refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

This Case Study given by upGrad in coalition with IIIT-B, has the following objectives:

- To understand and learn to Logistic Regression in a real business scenario.
- To develop a basic understanding of targeting potential leads in product based firms.
- To understand the use of Data so as to increase the score that means the lead is hot.

BUSINESS UNDERSTANDING

2

BUSINESS UNDERSTANDING

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

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 OBJECTIVES

3

BUSINESS OBJECTIVES

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Main points:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

SOLUTION METHODOLOGY

4

BASIC STEPS PERFORMED

1. Data cleaning and data manipulation.
   - Check and handle duplicate data.
   - Check and handle NA values and missing values.
   - Drop columns, if it contains large amount of missing values and not useful for the analysis.
   - Imputation of the values, if necessary.
   - Check and handle outliers in data.

2. EDA
   - Univariate data analysis: value count, distribution of variable etc.
   - Bivariate data analysis: correlation coefficients and pattern between the variables etc.

3. Feature Scaling & Dummy Variables and encoding of the data.
4. Classification technique: logistic regression used for the model making and prediction.
5. Validation of the Model.
6. Model Presentation.
7. Conclusions and Recommendations.

DATA PREPARATION

5

DATA PREPARATION

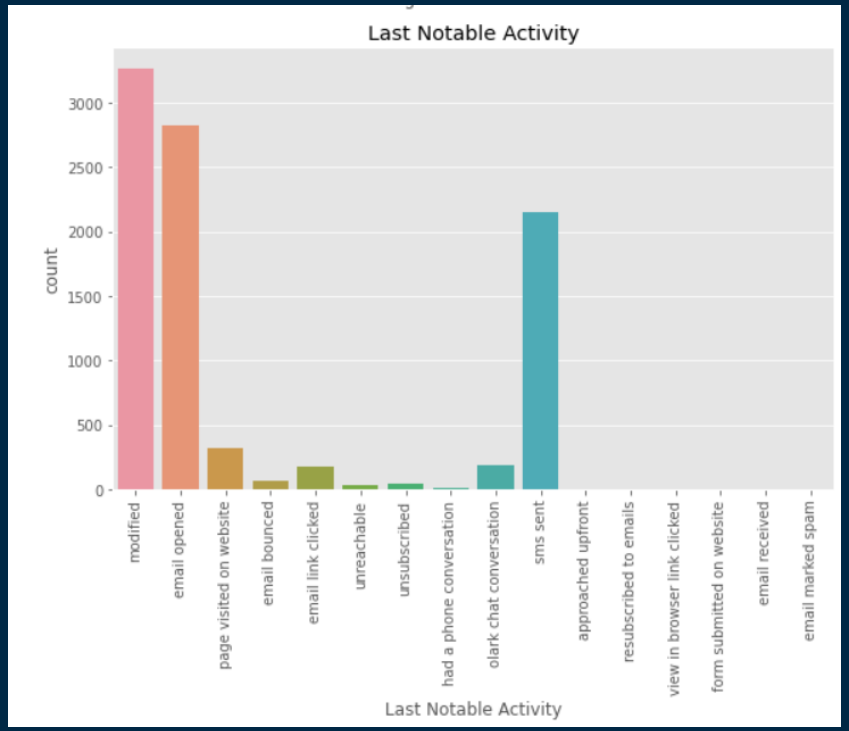
- The data we are provided with has total Number of Rows are 37 and total Number of Columns are 9240.
- There are some single value features present like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”.
- We have dropped redundant columns like ‘last_activity’, 'get_updates_on', 'i_agree_to‘, and 'update_me_on' etc.
- We have replaced all 'select' values with NaN Value.
- We have replaced the value "wrong number given" with the value "invalid number”.
- We have removed the columns ‘prospect_id’ and 'lead_number' which is not necessary for the analysis.
- Dropping the columns having more than 35% as missing value such as ‘how_did_you’ and ‘lead_profile’.
- We have categorised Country in binary as 'India' and 'Other’.
- In 'lead_source' column, we have classifed Values with frequency less than 30 as "others”.
- We have classified NA Values of Column "lead_quality", which has almost 52% NA values, as "not_sure“.

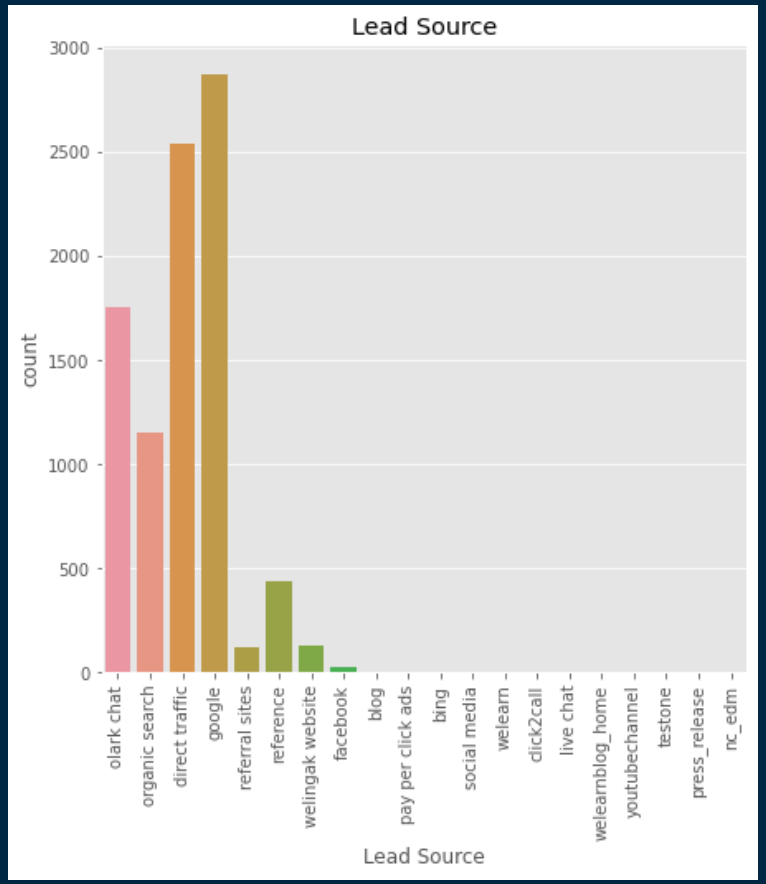
EDA

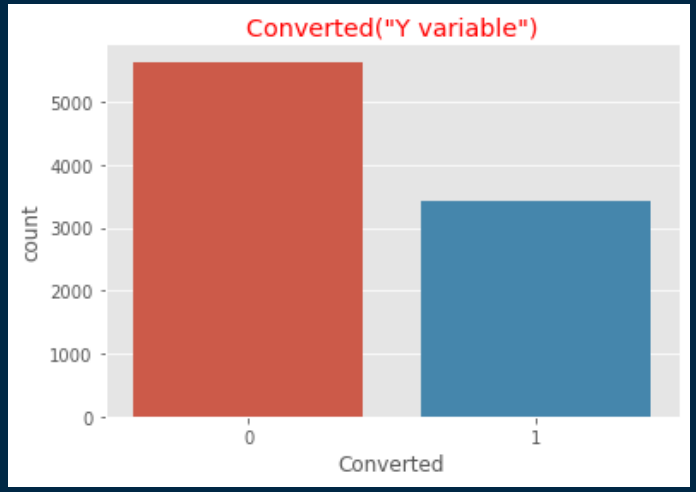
UNIVARIATE ANALYSIS

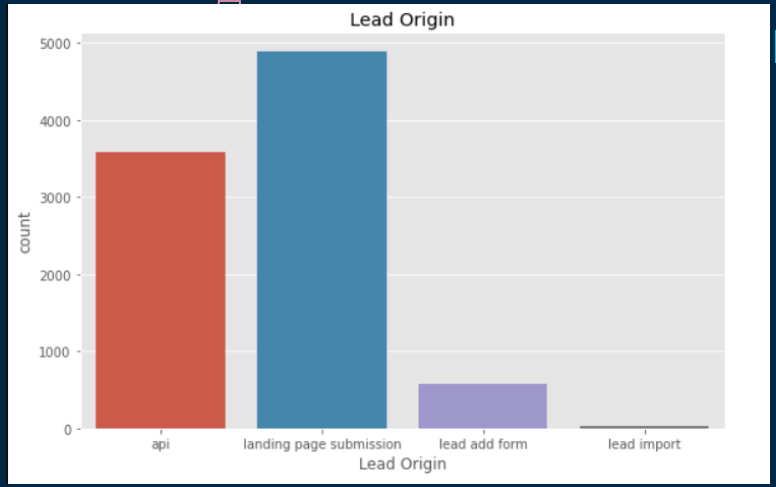
6

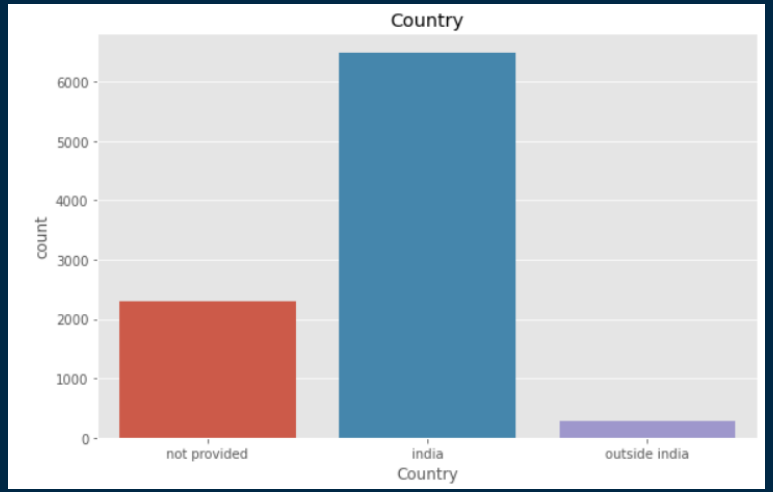
Categorical Variable

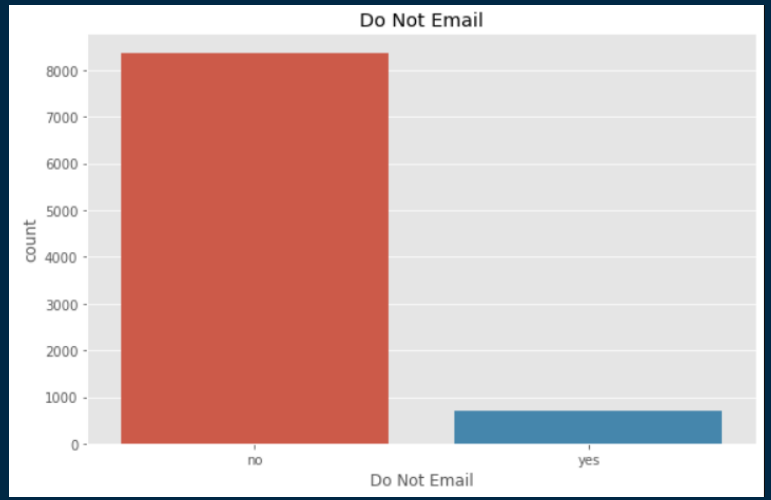





Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin


Country
count
6000
5000
4000
3000
2000
1000
0
not provided
india
outside india
Country

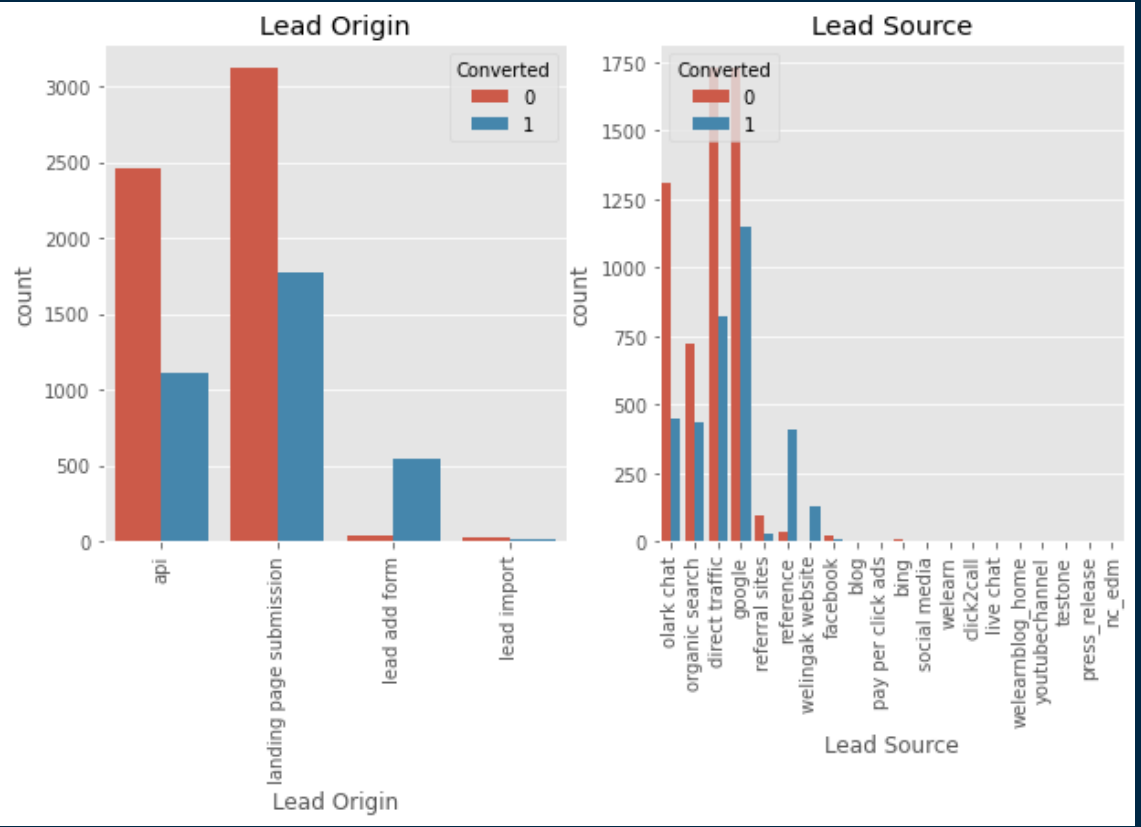

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

EDA

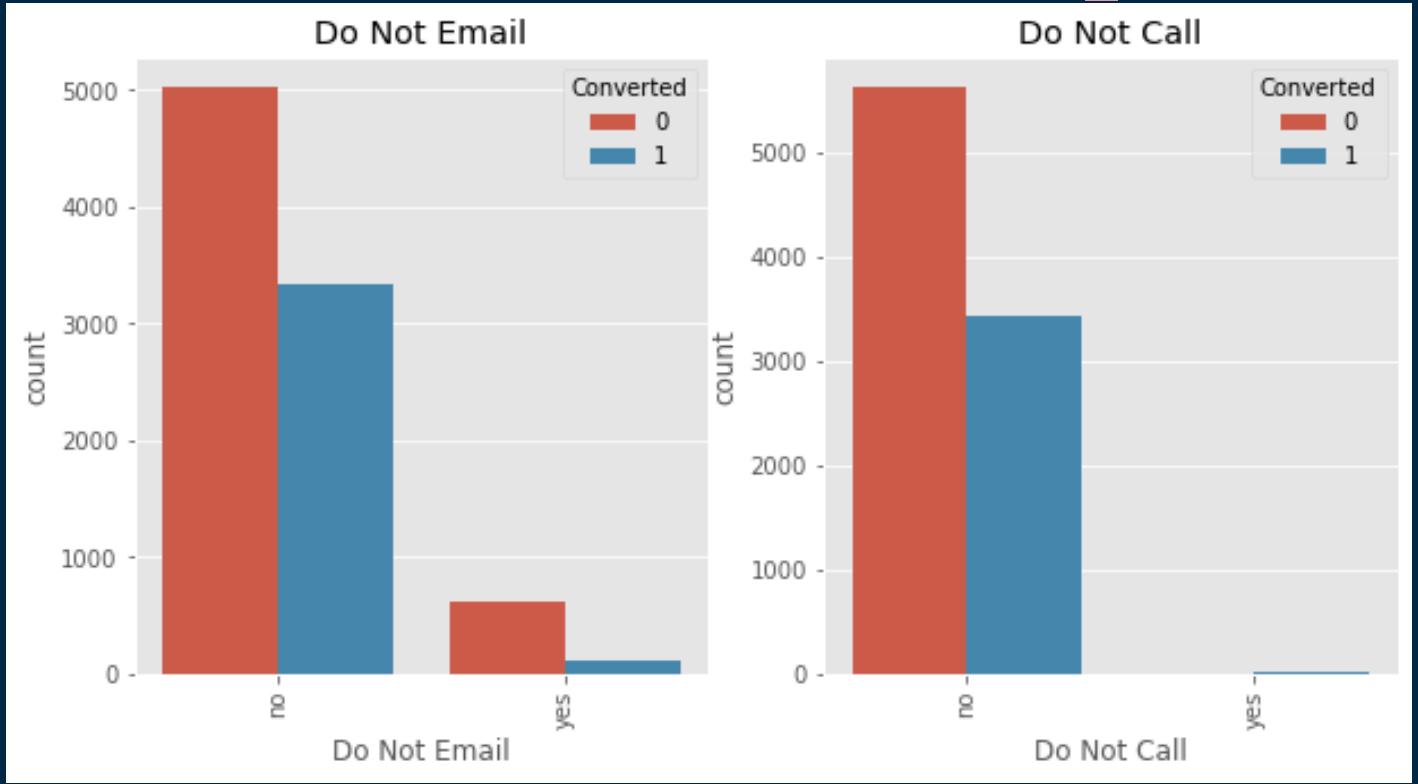
BIVARIATE ANALYSIS

7

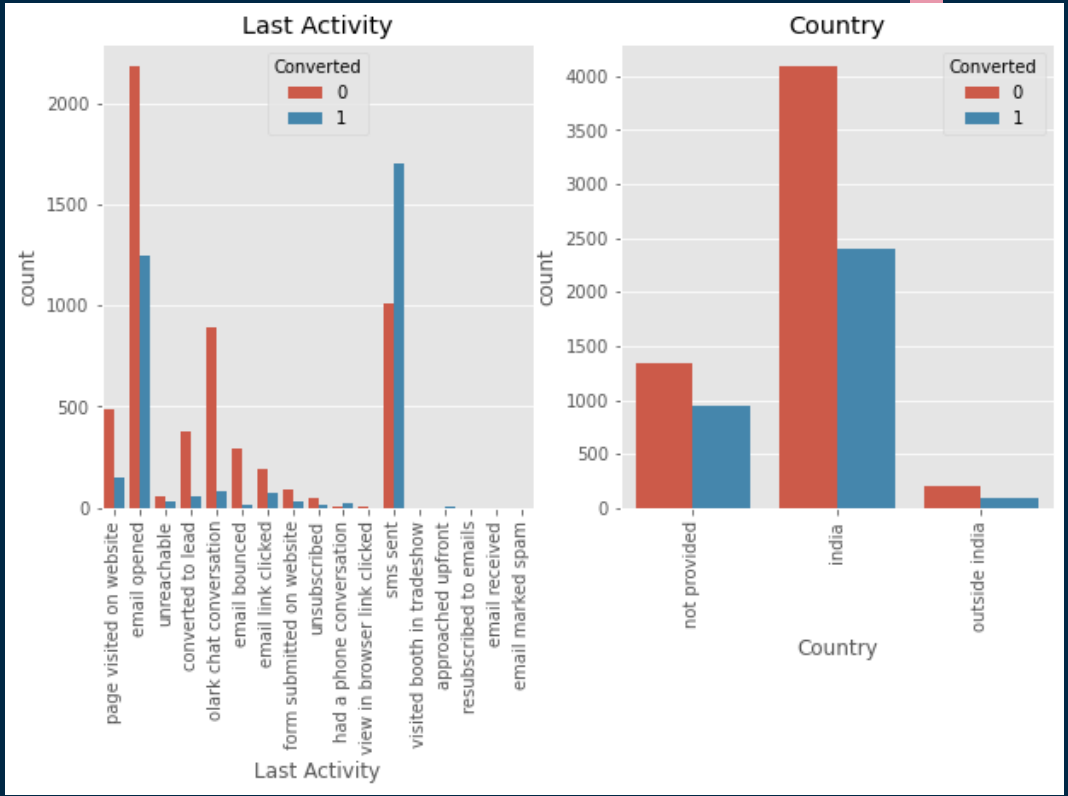
Lead Origin and Lead Source with respect to Leads who got Converted (Target 1) and those who didn't (Target 0).



Do Not Email and Do Not Call with respect to Leads who got Converted (Target 1) and those who didn't (Target 0).



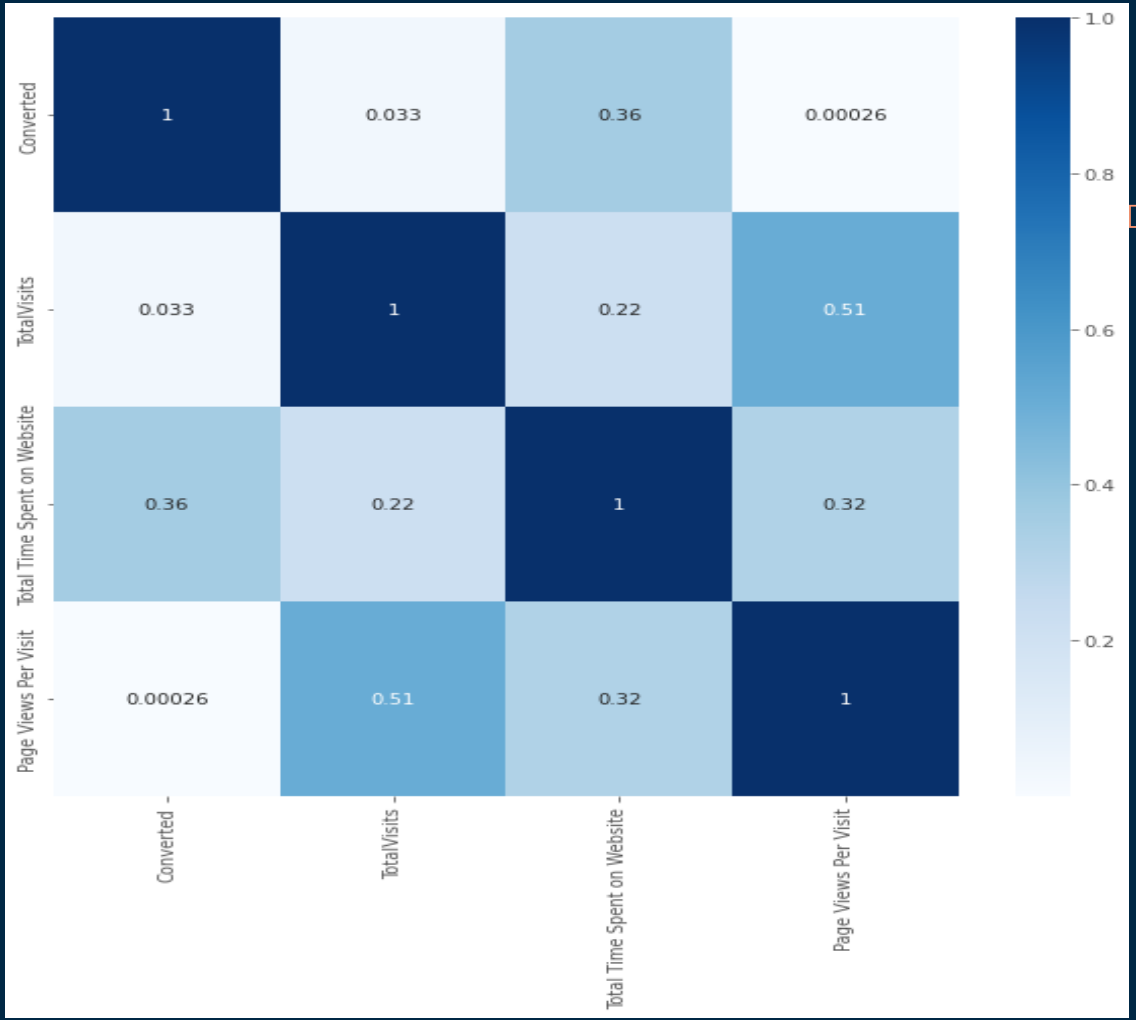
Last Activity and Country with respect to Leads who got Converted (Target 1) and those who didn't (Target 0).



TOP CORRELATION

8

Top Correlation among the Variables



DATA CONVERSION AND MODEL BUILDING

9

Data Conversion:

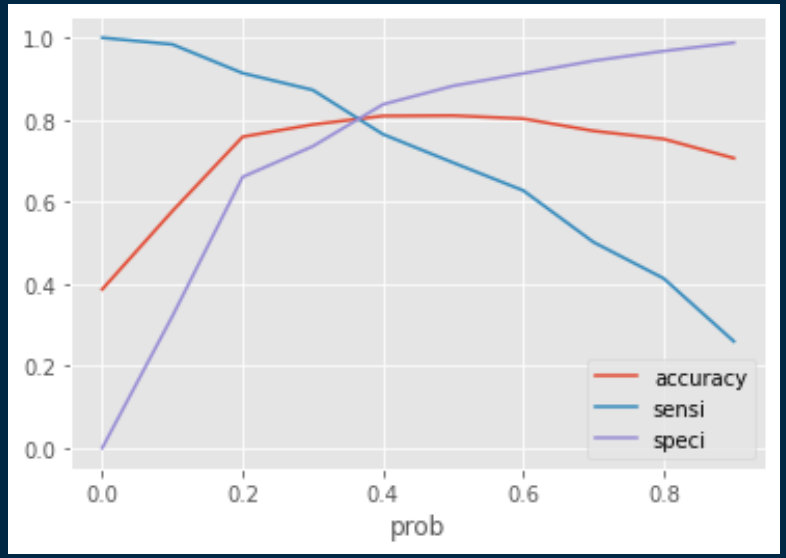
- Here in it, the Numerical Variables are Normalised.
- Dummy Variables were created for object type variables.
- The Total Number of Rows we had for Analysis is 9074.
- The Total Number of Columns we had for Analysis is 81.

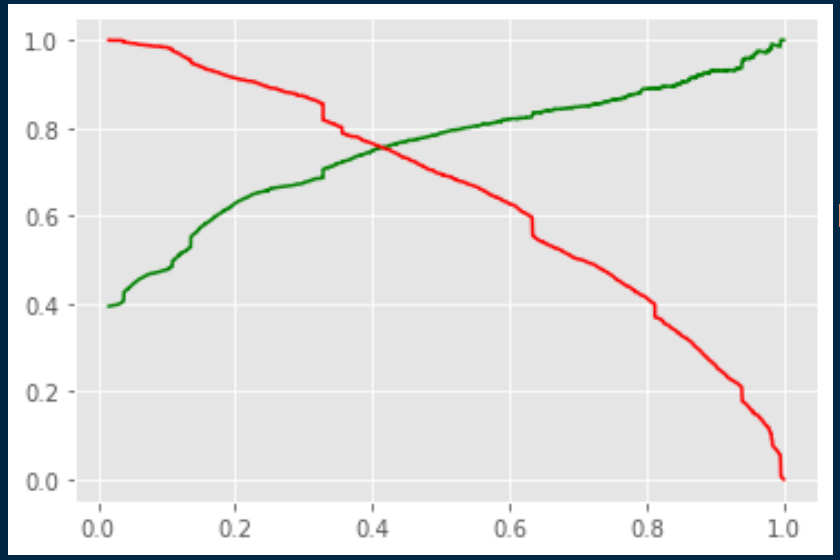
Model Building:

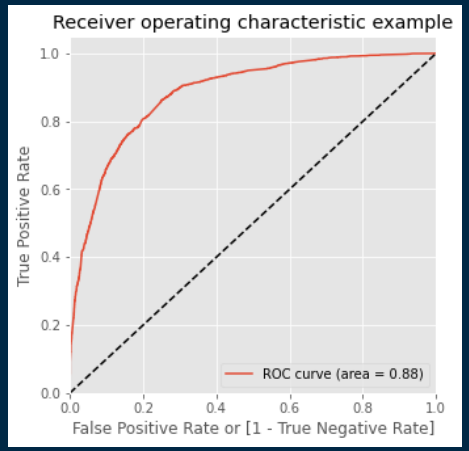
- We did the splitting of the Data Set into Training and Testing Sets.
- The first basic step for regression is to perform a train-test split. Hence, we have chosen 70:30 ratio.
- We used RFE for Feature Selection.
- We ran RFE with 15 Variables as output.
- We build the Model by removing the variables whose p-value is greater than 0.05 and VIF value is greater than 5.
- We did the predictions on the test data set.
- The Overall accuracy we found is 81%.

ROC CURVE

10







At Cut-off = 0.35

- Accuracy - 81%
- Sensitivity - 81%
- Specificity - 80%

At Cut-off = 0.41

- Precision -73%
- Recall - 75%

CONCLUSION

11

We found that the variables that mattered the most in the potential buyers are as follows (In the Descending Order) :

1. The Total Time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website

4. When the last activity was:
    a. SMS
    b. Olark chat conversation

5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.
7. While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
8. Accuracy, Sensitivity and Specificity values of Test Set are around 81%, 81% and 80% respectively which are approximately closer to the respective values calculated using trained set.

27,388

Whoa! That’s a big number, aren’t we proud?

We did analysis on this much Rows from all Data Frames

THANK YOU