Lead Scoring Case Study

Name:- Divit Karmiani
Email ID:- divitkarmiani1998@gmail.com

Problem Statement

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.

Task - Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Analysis Approach

Data Inspection & Cleaning → Outlier Treatment → EDA → Data Prep → Model Building

Data Inspection & Cleaning

- Data Inspection is a very necessary step in order to get to know the data and its various traits
- There will be a lot of null and unnecessary data in the dataset and this will skew the model results. Thus, it is very important to clean the data

Outlier Treatment

- Since outliers affect cluster centers a lot, the higher range outliers that needs to be treated must be capped using soft capping to a value at 99th percentile

EDA

- EDA - The primary aim with exploratory analysis is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis

Data Prep

- Converting some binary variables (Yes/No) to 1/0
- Creating Dummy variables
- Train-Test split (70-30)
- Feature Scaling

Model Building

- Creating the final model based on 13 variables and increasing the Recall by finding the optimum threshold value

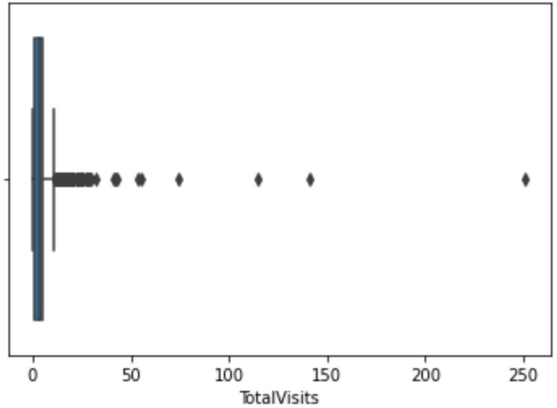
Data Inspection & Cleaning

- Data Inspection is done for getting an overall view of the data and checking for any abnormalities in the data before feeding it to the model
- One the data is inspected, the necessary changes to the data needs to be made which leads to Data Cleaning

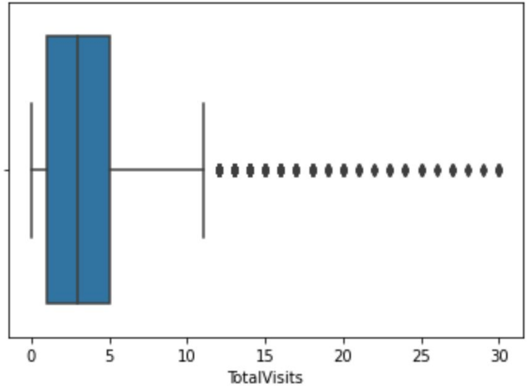
- Data Cleaning methods:
  - Prospect ID & Lead Number are two variables that are just indicative of the ID number of the contacted people & can be dropped
  - Checked and removed duplicates
  - The response from the Google form recorded as “Select” was replaced by Null values for the below mentioned columns:
    - Specialization
    - How did you hear about X education
    - Lead Profile
    - City
  - Dropped the columns that have Null values of >45%
  - Missing value handling:
    - Dropped missing values in numeric variables
    - Replaced missing values of categorical variables as a different category
    - Impute missing values of categorical variables as Mode

Outlier Treatment - Total Visits

- Having visits greater than 100 is a very extreme case and very rare. Since the values are too extreme and there are only 3 records, we can drop these records

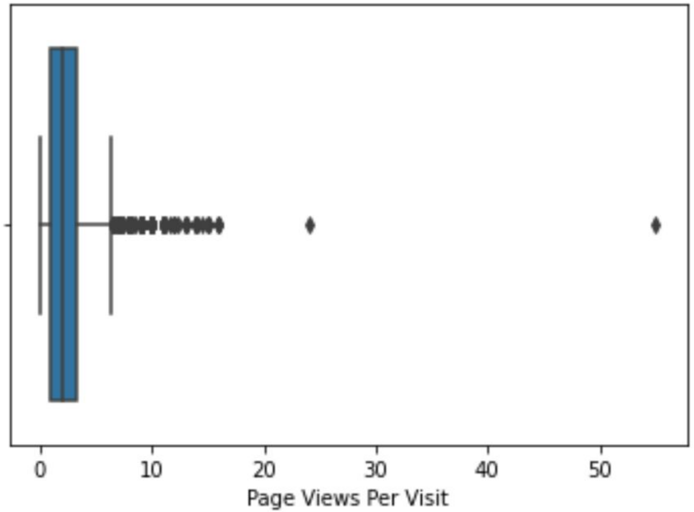


- We can cap the values to 30



Outlier Treatment - Page views per visit

- For these 2 records the total page views comes to be
  1. 30*55
  2. 24*24
- These are a bit extreme. Since these are only 2 records, we can safely drop it



Exploratory Data Analysis

- Standardization of values
  - Reduced number of categories for some features

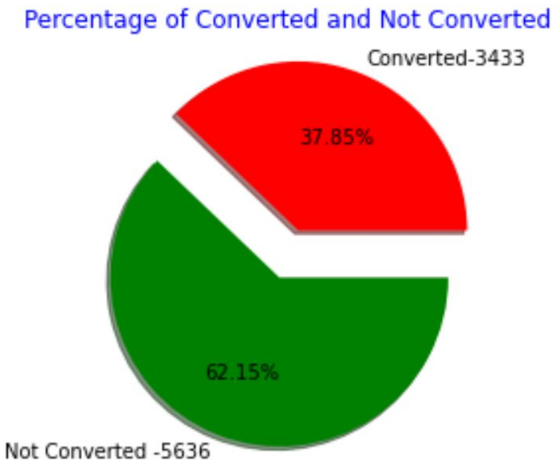
- Data Imbalance
  - We will also check which columns can be dropped due to the imbalanced nature of the columns
    - Combined the variables with less than 1% values as Others
    - Do not email, Do not call - Columns are imbalanced. Though these feature might be very important but the resulting model will always favor the higher category. Since we can't sample data again, and other techniques might be an overkill, we will drop this column
    - Dropped many columns that has imbalanced values

Final Columns

#	Column	Non-Null Count	Dtype
0	Lead Origin	9074 non-null	object
1	Lead Source	9074 non-null	object
2	Converted	9074 non-null	int64
3	TotalVisits	9074 non-null	float64
4	Total Time Spent on Website	9074 non-null	int64
5	Page Views Per Visit	9074 non-null	float64
6	Last Activity	9074 non-null	object
7	Specialization	9074 non-null	object
8	Current Occupation	9074 non-null	object
9	Tags	9074 non-null	object
10	City	9074 non-null	object
11	A free copy of Mastering The Interview	9074 non-null	object
12	Last Notable Activity	9074 non-null	object

Exploratory Data Analysis - Learnings

Checking Data Imbalance



Numeric Variables

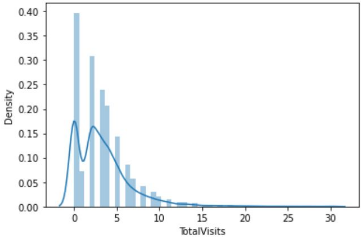
- Multiple peaks are observed in all the numeric variables
- No definite pattern found between numeric variables
- Total time spent on website
  - Those who have spent more time on website have converted more.

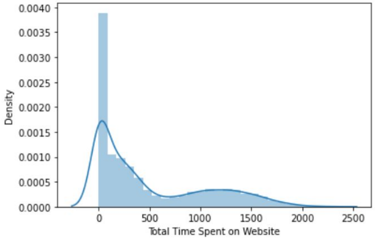
Categorical Variables

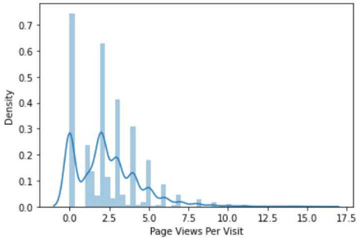
- Lead Add Form had a very high conversion rate
- People referred or on Welingak Website were likely to be converted more than any other source
- If an SMS was sent most recently, they were likely to be converted more than any other last activity
- Those opting for Management courses were likely to be converted
- Thane and Outskirts had a good conversion rate

Exploratory Data Analysis - Univariate

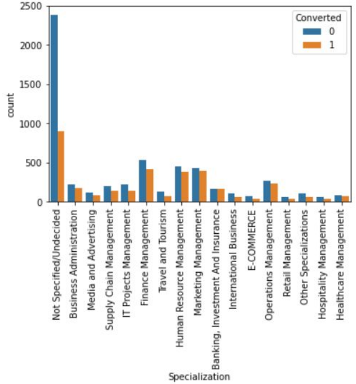
Numerical

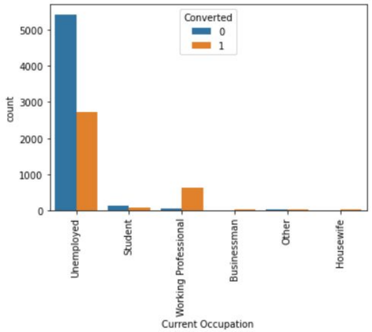






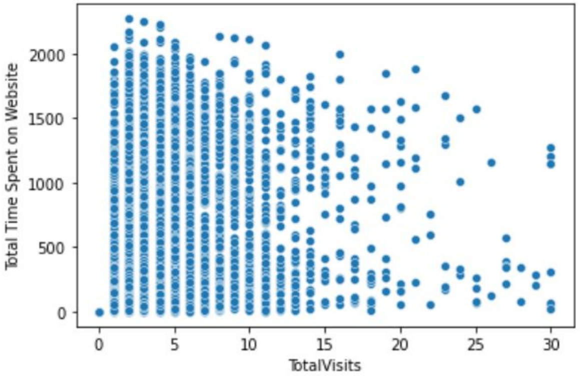
Categorical

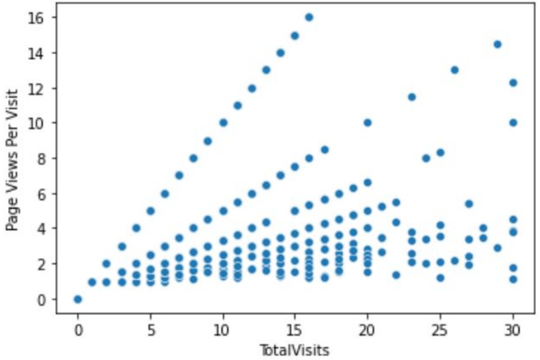


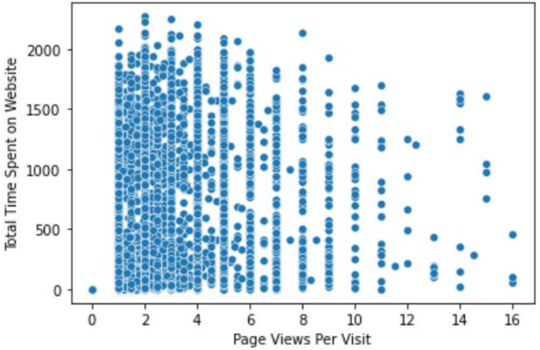


Exploratory Data Analysis - Bivariate 1

Numerical Variables

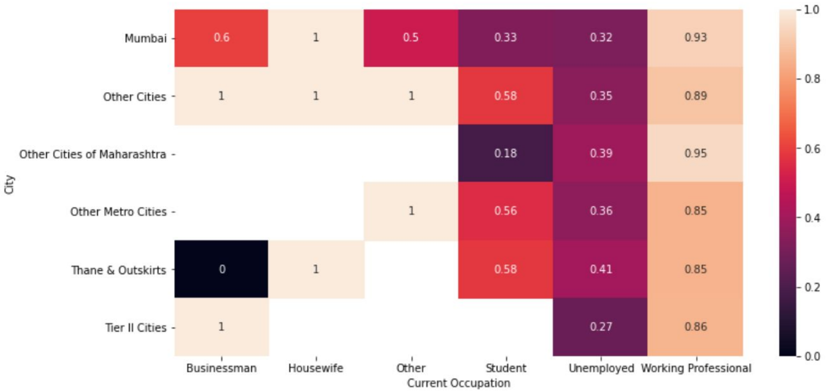


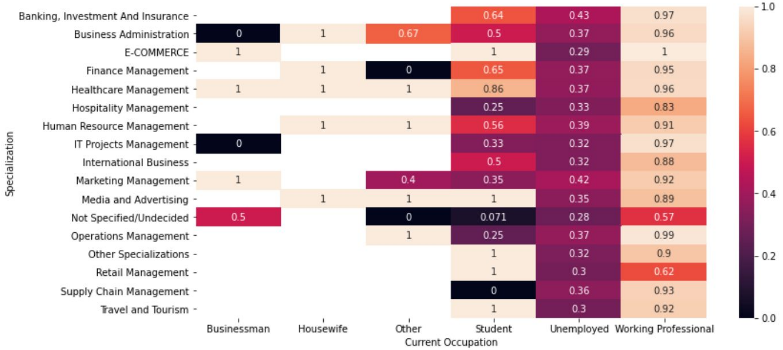


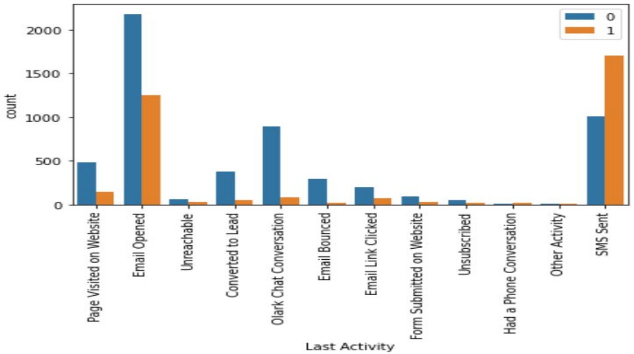


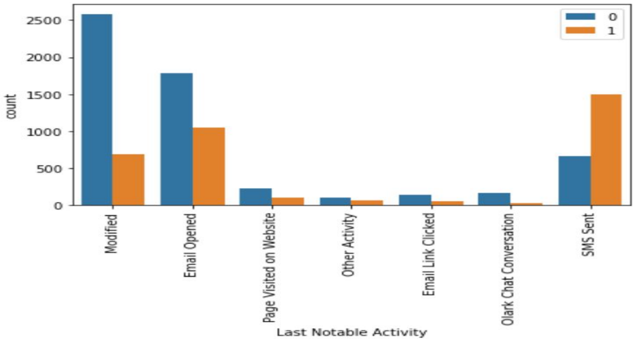
Exploratory Data Analysis - Bivariate 2

Categorical Variables

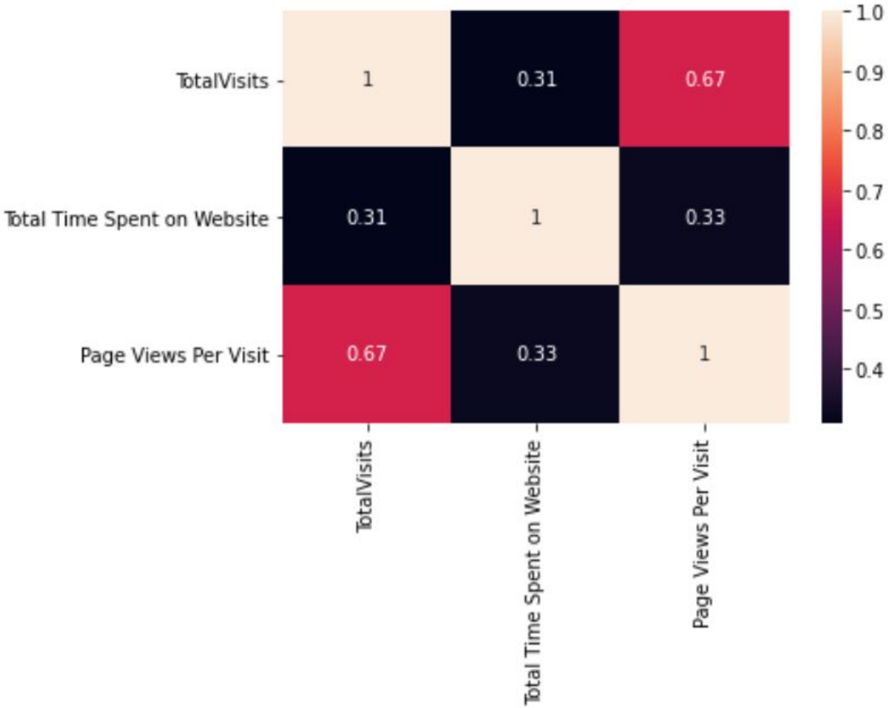








Correlation



Not so high correlation between numeric variables

Modeling - 1

- The total number of columns in the data is 64, I reduced it to 15 using RFE (Recursive Feature Elimination)
- I built the model using these 15 variables
- Out of these 15, 2 of them had high p-value (>0.05), so they were dropped and now we have 13 variables
- The final model was built with 13 variables and the predicted values were found on the Train dataset
- Then I took a random threshold value of 0.5 and calculated accuracy
  - Accuracy - 92.4%
- Calculation of VIF

22]:

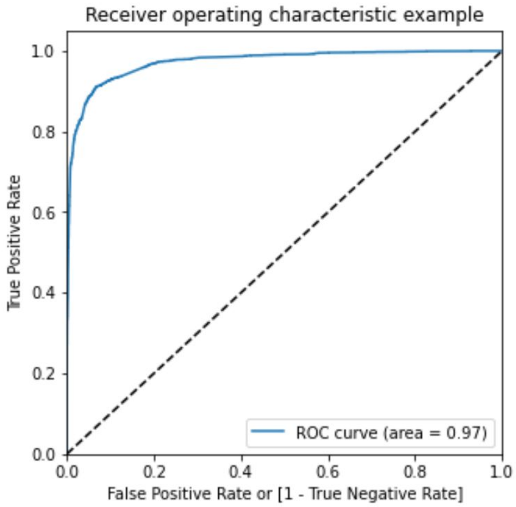
	Features	VIF
1	Lead Origin_Landing Page Submission	2.64
13	Tags_Will revert after reading the email	2.40
11	Tags_Not Specified	2.35
6	Last Activity_SMS Sent	2.22
5	Last Activity_Email Opened	2.08
2	Lead Origin_Lead Add Form	2.01
3	Lead Source_Olark Chat	1.89
12	Tags_Ringing	1.57
0	Total Time Spent on Website	1.46
4	Lead Source_Welingak Website	1.37
8	Tags_Closed by Horizzon	1.30
7	Tags_Busy	1.13
9	Tags_Interested in other courses	1.12
10	Tags_Lost to EINS	1.12
14	Tags_switched off	1.11

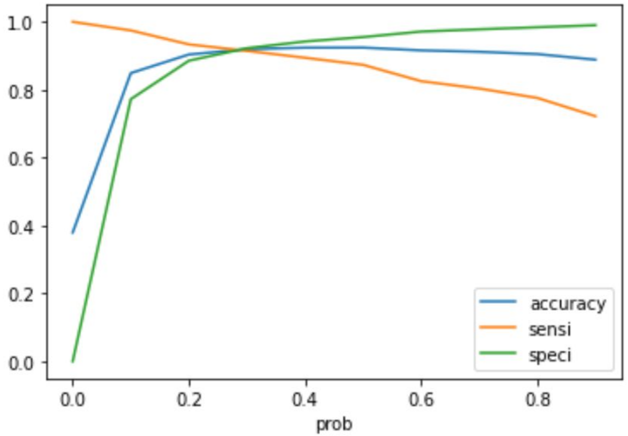
VIF looks good

VIF value gives the dependency of one variable on the other. As all the features had low VIF values, no features will be dropped; they do not depend highly on each other

Modeling - 2

- ROC curve was drawn for optimal cut-off for Sensitivity and Specificity which turned out to be 0.3

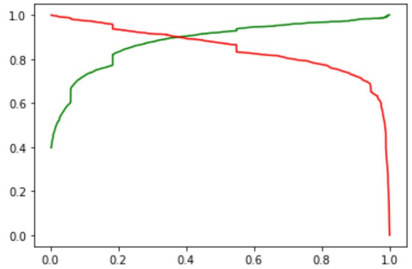




From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Modeling - 3

- In this use case, the Recall should be atleast 80%, we first find the optimum cut-off for Precision-Recall Framework as follows:



- The cut-off is - 0.377
- The Precision and Recall value for this cut-off for the Test Set is:
  - Precision - 89.54%
  - Recall - 89.27%
- Accuracy - 92.02%

Final Results

- The results on the Test Set are as follows:
    a. Precision - 89.54%
    b. Recall - 89.27%
    c. Accuracy - 92.02%
- **We prepared the final dataset by we appending the Lead Score in the range 0-100**