Lead Scoring - Case Study

By - Shashwat Joshi

Business Question

The X Educations sells online courses but there current conversion rate is only around 35%, they want to refine their techniques and make efficient changes to increase this conversion rate. They want to identify the leads and focus on them and try converting them and wants to understand the factors that can affect conversion rate.

Objective

- The X Educations wants to have a score for the potential leads so that they can focus on converting these leads into customers.

- The company needs a model where they can assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO has a benchmark target lead conversion rate of approx 80%.

Data Provided

- 1. 'lead.csv' contains all the information of the potential leads and the factor that can affect the leads.

- 3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Data Understanding - Lead Data

- The application data consists of 37 column and 9240 rows.
- Apart from the target variable, each column is a independent variable which provides an attribute about the potential lead

The first step in understanding the data is to do a sanity check

To do the sanity check we need to:

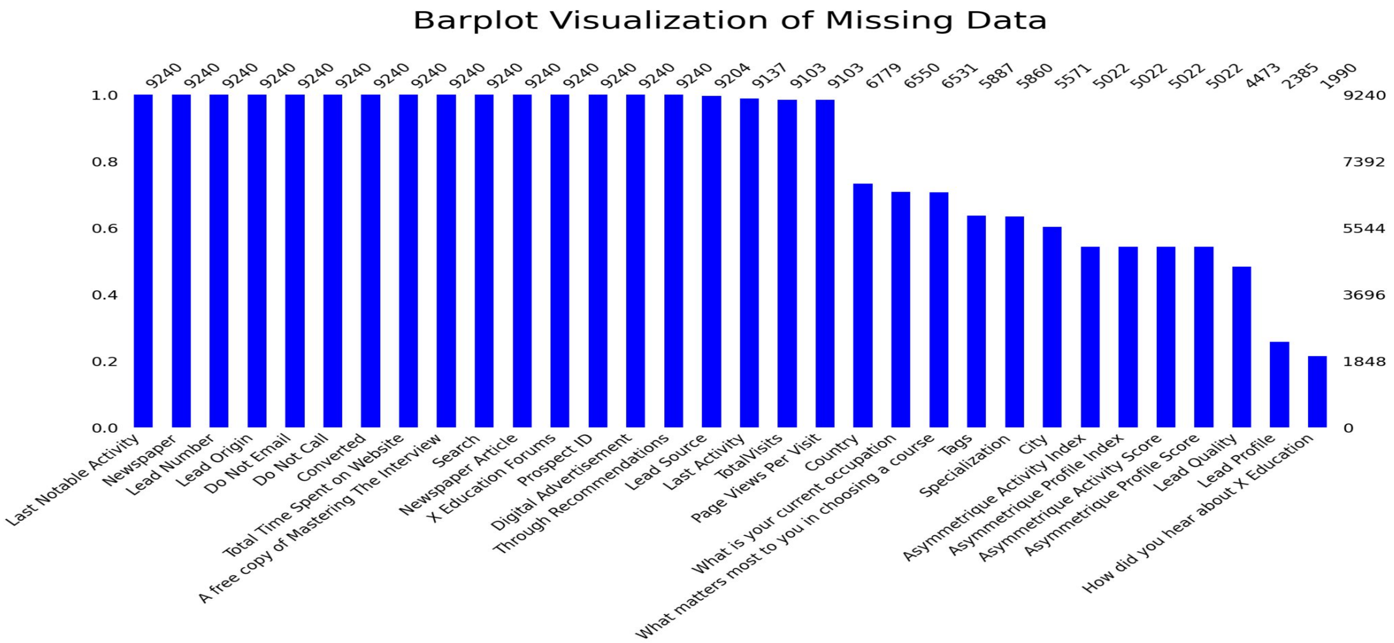
- Check if there are any columns with null values
- If there are columns with null values, we need to check the percentage of null values in that column, if the number is significant, we can drop the column altogether otherwise we can use multiple imputation techniques to fill the missing values
- We can check if the data has outliers, if yes, according to business sense, we can choose to keep or not to keep the outliers
- We can drop certain columns if we feel that they are not needed in the analysis
- We can transform columns to understand them better

Identifying the null values and dropping insignificant columns

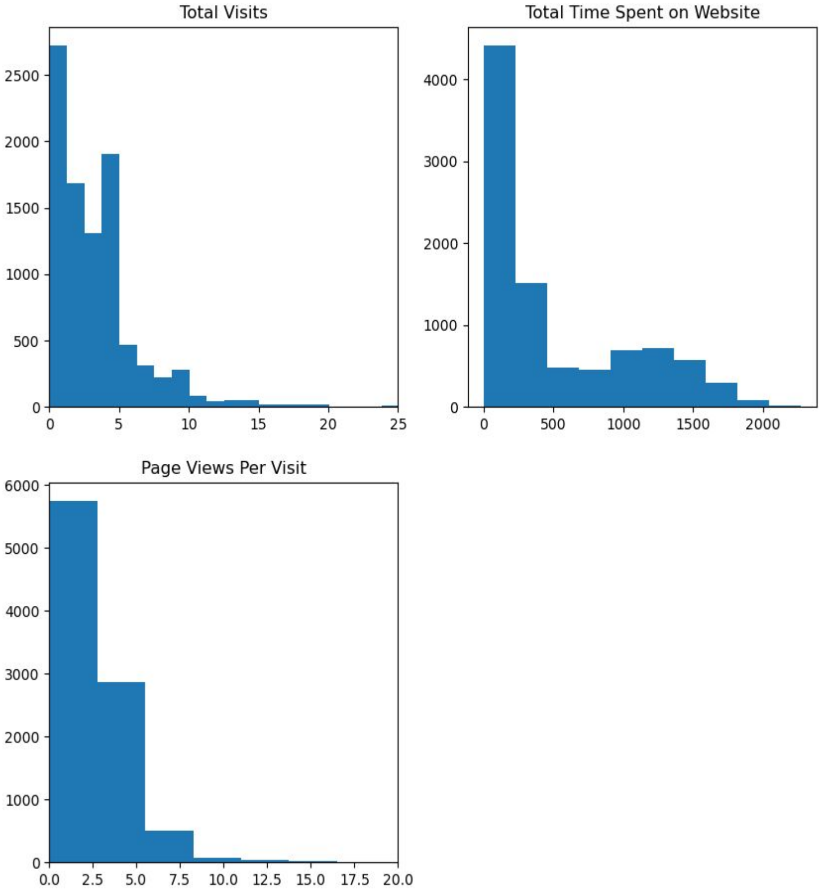
- We see that there are multiple columns which have significant null values
- Dropping the null values which have more than 40% of the data missing
- After dropping the null values we still see that there are some columns which still have some null values
- For the variables which only have a single variable in it, we drop them
- We also drop columns like City, Country, Source ID and lead number which are irrelevant for oue analysis

Data Imputation

- We check that some categorical columns has a variable named “Select”, we place it with data not available
- For the numerical columns we replaced the missing data with their modes



EDA



Numerical Variables

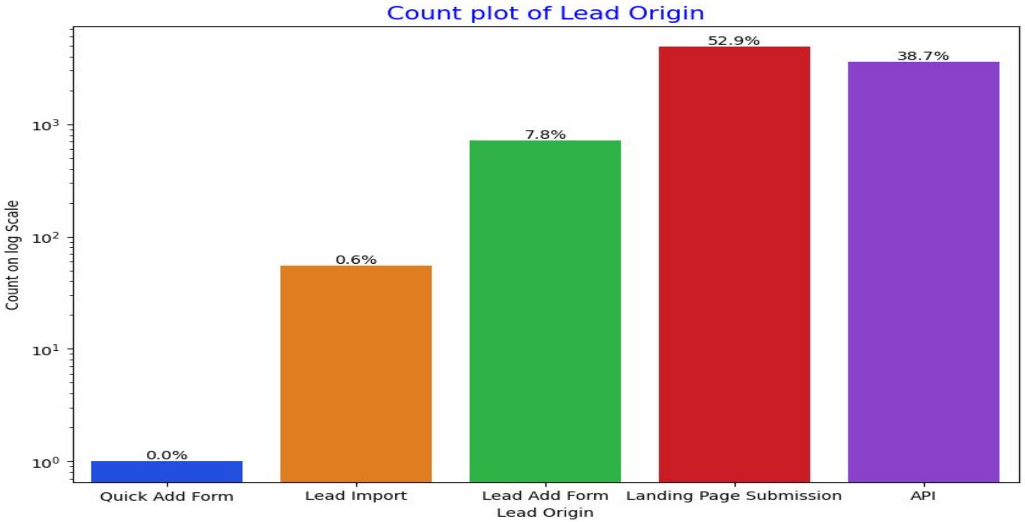
- We only have 3 categorical variables in the dataset
- We make a histogram plot for them
- We see that most of the potential leads visit the website less than 5 times and spend less than 500 units of time and few around 2 pages of the content

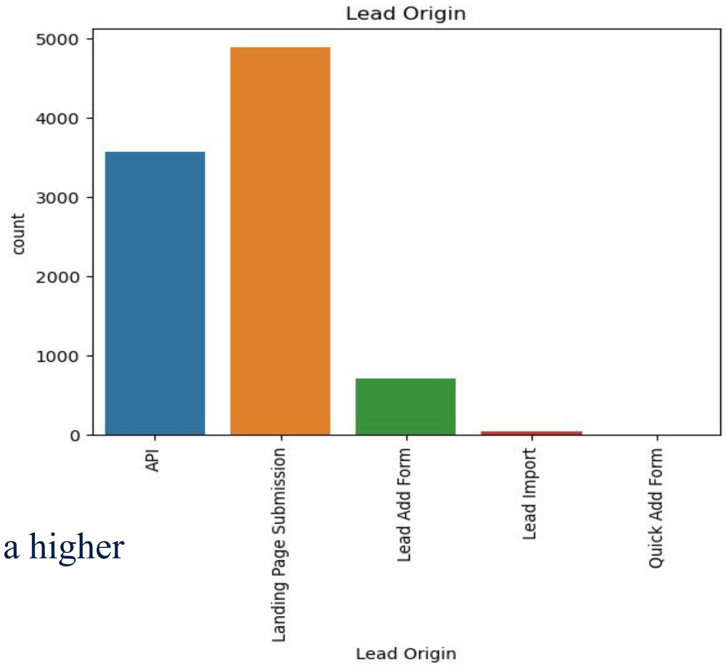
Analysing the Data :

- Univariate Analysis - Univariate analysis is basically the simplest form to analyze data. Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summarise it, and then find some pattern in the data
- We make dummy variables for each variable in the categorical columns

Let’s do univariate analysis on the following categorical columns:

1. **Lead Origin**





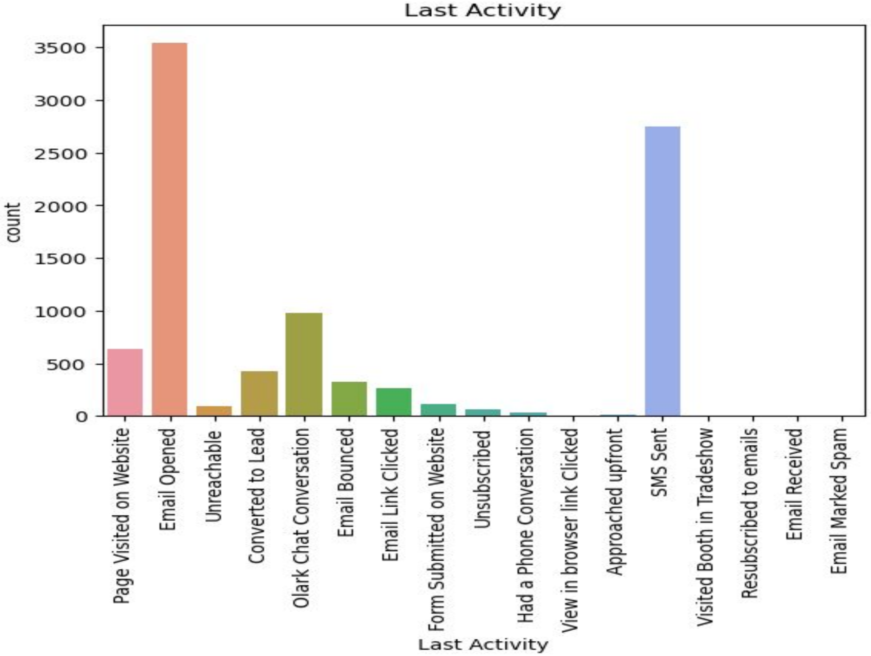
- Most leads originate from the landing page submission and that has a higher conversion rate followed by API
- Lead import and Quick add form have done poorly

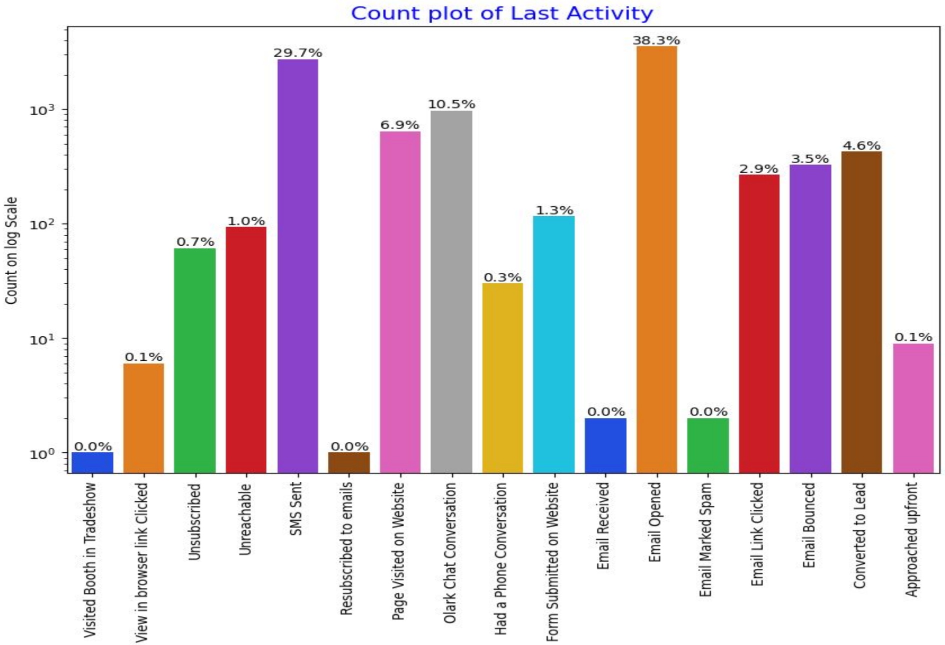
Analysing the Data : Application Data

2. LAST ACTIVITY

- Majority of leads last activity was email opend which also had a high conversion followed by SMS sent

- Other forms of last activity leads to a dead end

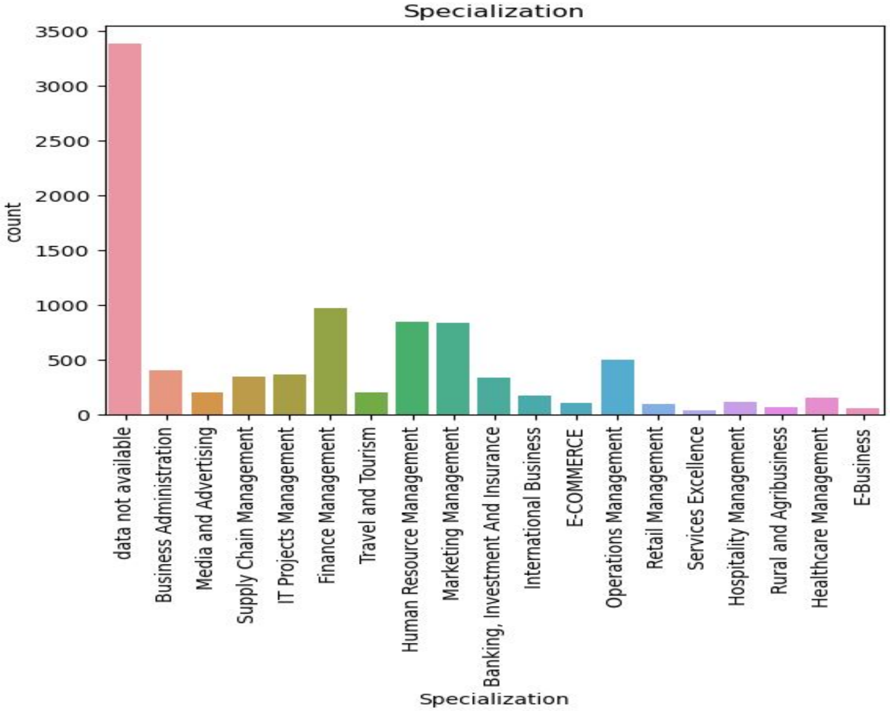


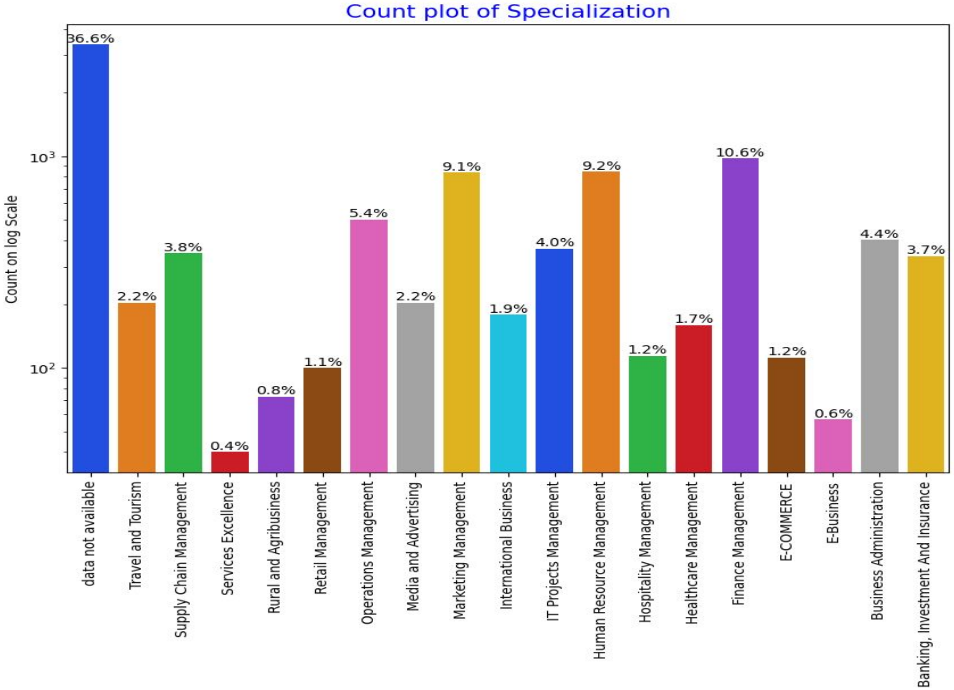


Analysing the Data : Application Data

3. SPECIALIZATION

- At most of the places data is not available, other then that people are trying for courses for management and it has a higher conversion rate then others

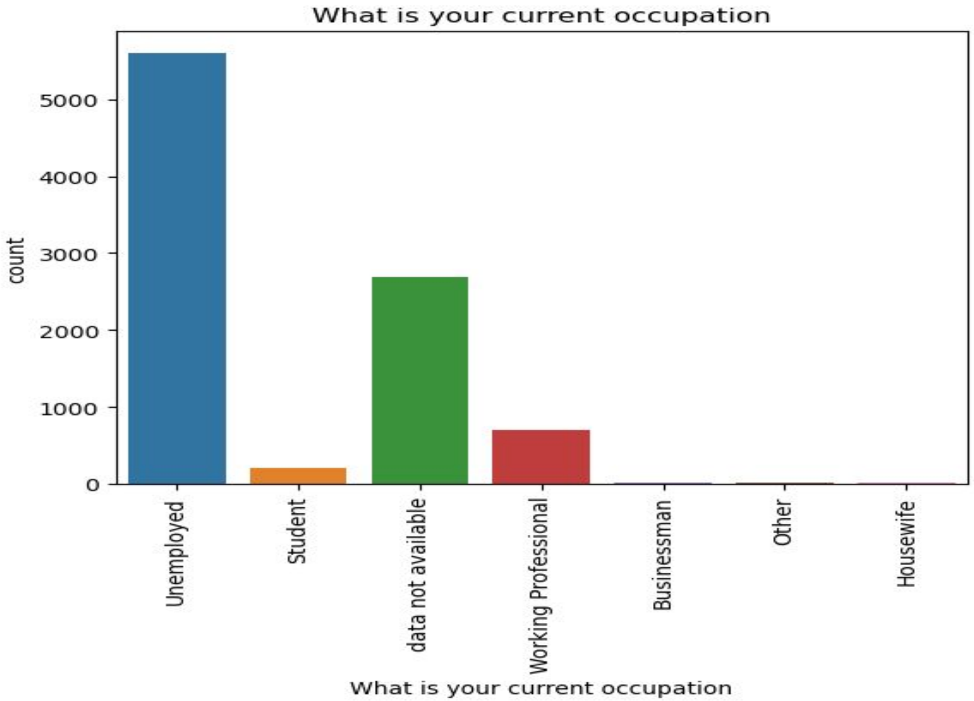


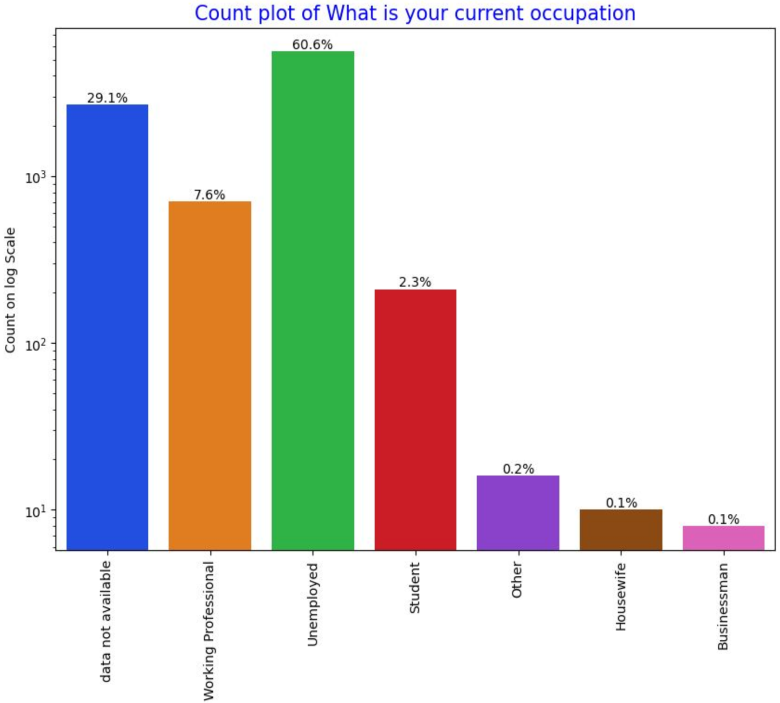


Analysing the Data : Application Data

4. What is your current occupation

- Most of the people who applies for a course are unemployed and have a higher chance of conversion



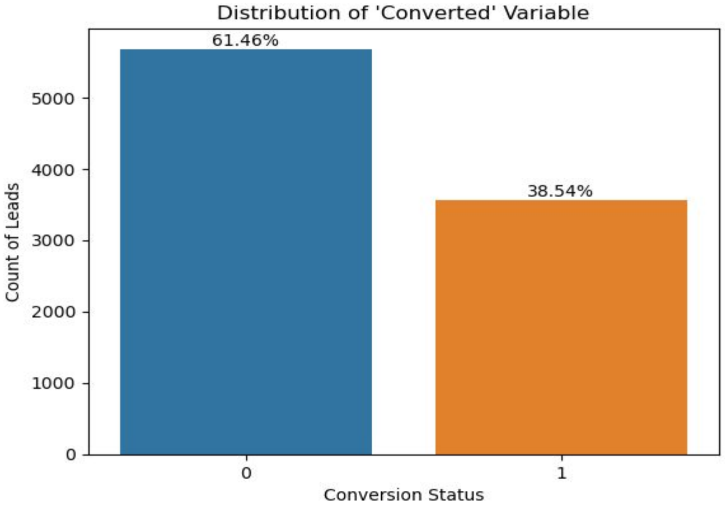


Analysing the Data : Application Data

5. Rest of the columns

- The following columns have the Data which is highly skewed. So, we drop them.
- Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Newspaper Article', 'Digital Advertisement', 'Through Recommendations'', ‘What matters most to you in choosing a course', 'Lead Profile', ''A free copy of Mastering The Interview, “tags”

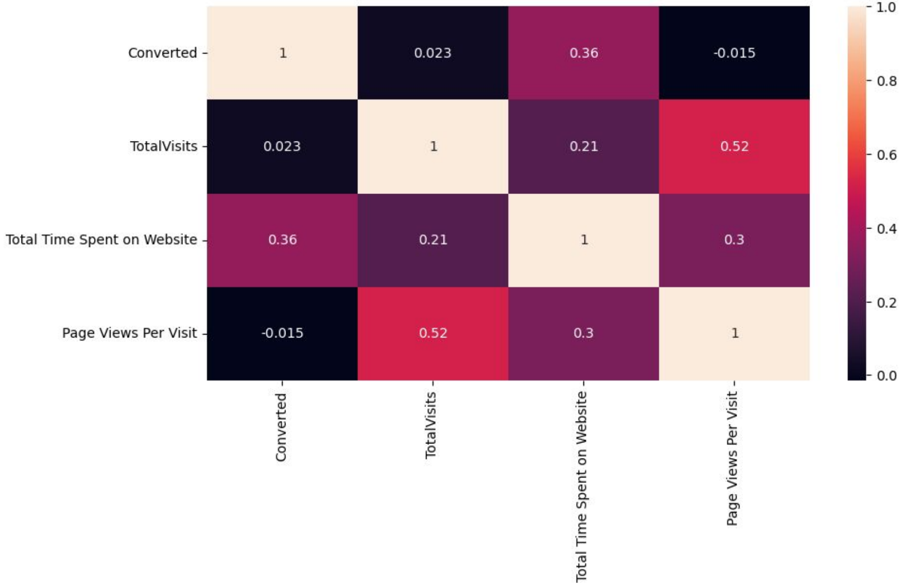
Checking for Data Imbalance



- A classification data set with skewed class proportions is called imbalanced. Classes that make up a large proportion of the data set are called majority classes. Those that make up a smaller proportion are minority classes.
- In this case, the dependent variable is fairly balanced between converted and unconverted leads

Analysing the Data : Application Data

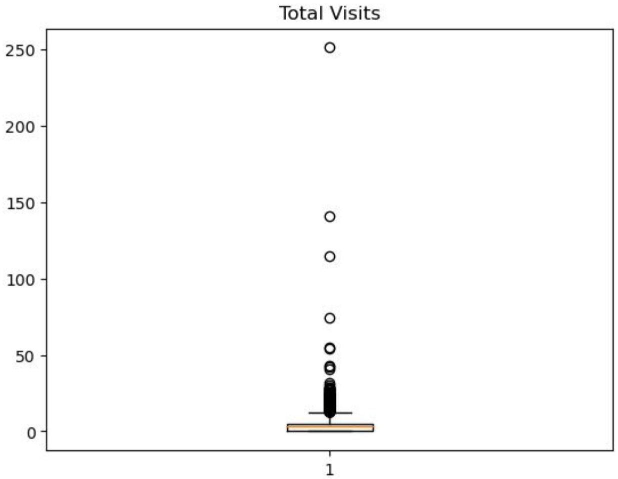
Bivariate Analysis - Bivariate analyses are conducted to determine whether a statistical association exists between two variables, the degree of association if one does exist, and whether one variable may be predicted from another

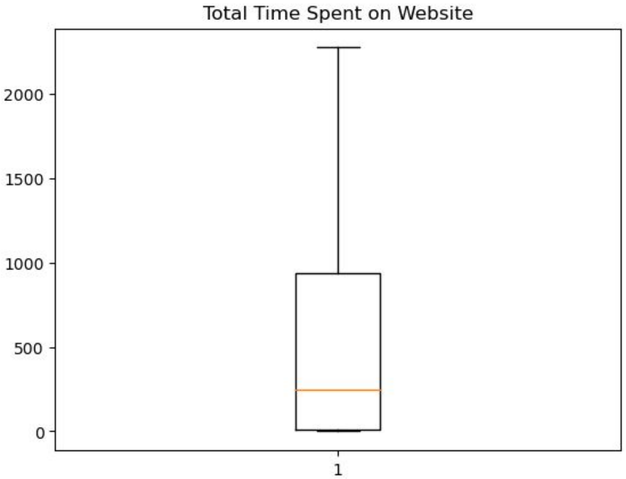


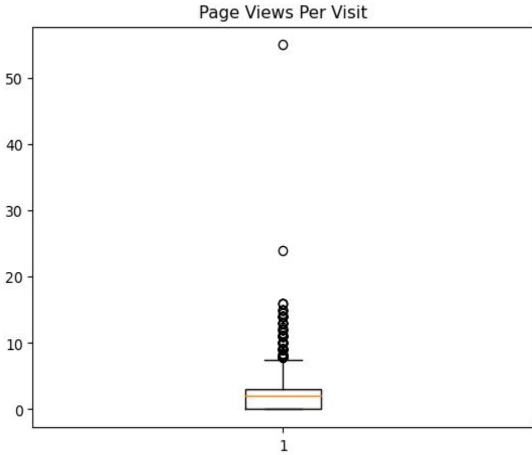
- We checked for correlation among the numeric variables and didn’t see any variables with a high correlation

Analysing the Data : Application Data

Checking outliers







- We checked the outliers for the numerical variables
- Though we saw some outliers in the total visits and page views per visit but we cant remove or cap them because these are true data and important for our analysis

Model Setup

Train and Test Split

- We split the data into train and test with a train size of 0.7

Feature Scaling

- We use Min Max Scaler to scale our numerical columns in the range of 0 to 1

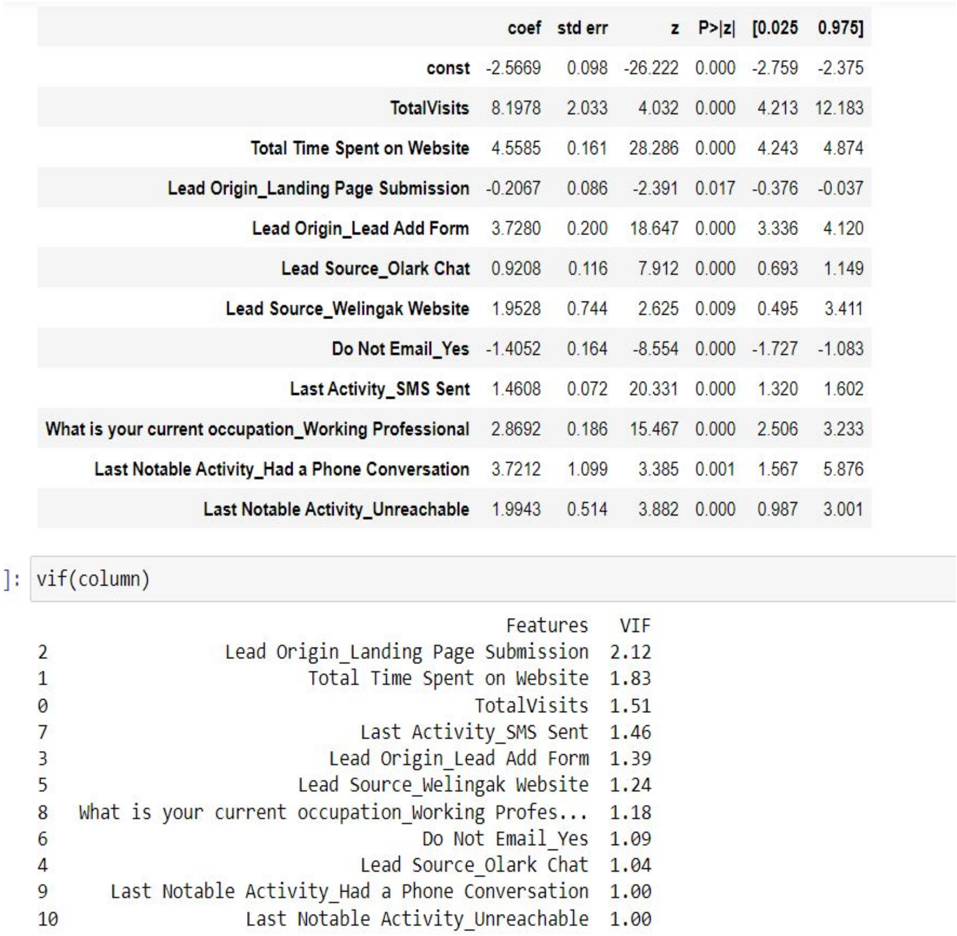
Feature Selection

- We use RFE to select top 15 variables to be used in the model

Algorithm Selection

- For this analysis we are using Logistic Regression Algorithm from the sklearn library

Modelling



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5669	0.098	-26.222	0.000	-2.759	-2.375
TotalVisits	8.1978	2.033	4.032	0.000	4.213	12.183
Total Time Spent on Website	4.5585	0.161	28.286	0.000	4.243	4.874
Lead Origin_Landing Page Submission	-0.2067	0.086	-2.391	0.017	-0.376	-0.037
Lead Origin_Lead Add Form	3.7280	0.200	18.647	0.000	3.336	4.120
Lead Source_Olark Chat	0.9208	0.116	7.912	0.000	0.693	1.149
Lead Source_Welingak Website	1.9528	0.744	2.625	0.009	0.495	3.411
Do Not Email_Yes	-1.4052	0.164	-8.554	0.000	-1.727	-1.083
Last Activity_SMS Sent	1.4608	0.072	20.331	0.000	1.320	1.602
What is your current occupation_Working Professional	2.8692	0.186	15.467	0.000	2.506	3.233
Last Notable Activity_Had a Phone Conversation	3.7212	1.099	3.385	0.001	1.567	5.876
Last Notable Activity_Unreachable	1.9943	0.514	3.882	0.000	0.987	3.001

```
]: vif(column)
```

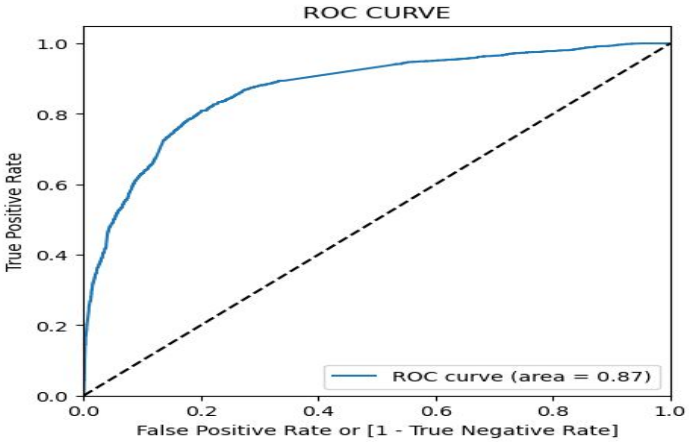
```
                                          Features   VIF
2              Lead Origin_Landing Page Submission  2.12
1                      Total Time Spent on Website  1.83
0                                      TotalVisits  1.51
7                           Last Activity_SMS Sent  1.46
3                        Lead Origin_Lead Add Form  1.39
5                     Lead Source_Welingak Website  1.24
8   What is your current occupation_Working Profes...  1.18
6                                 Do Not Email_Yes  1.09
4                           Lead Source_Olark Chat  1.04
9     Last Notable Activity_Had a Phone Conversation  1.00
10               Last Notable Activity_Unreachable  1.00
```

- If the P value for a variable is more than 0.05 or its VIF is more than 5, we remove those variables from our model

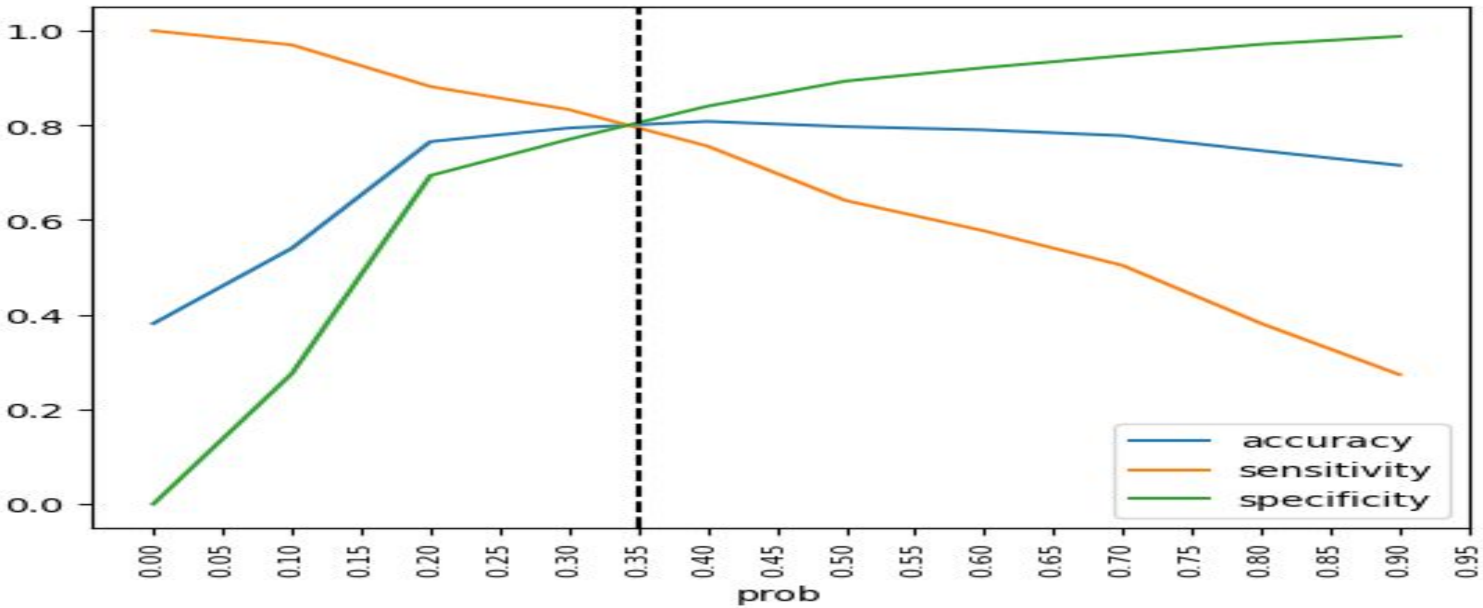
- Once we get the model to satisfy the above condition, we go for model Prediction and Evaluation

Model Evaluation - Predicting on Train set

- We make a Y_Pred variable using y_train_pred = logm.predict(X_train_sm)
- We rename is to Conversion_Prob, when its greater than 0.5 we put it as 1 else 0
- We find the accuracy, which here comes out to be 80%
- We with use of confusion matrix, calculate the specificity and sensitivity
  - Sensitivity (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive.
  - Specificity (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative.
- We creare an ROC curve, An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.



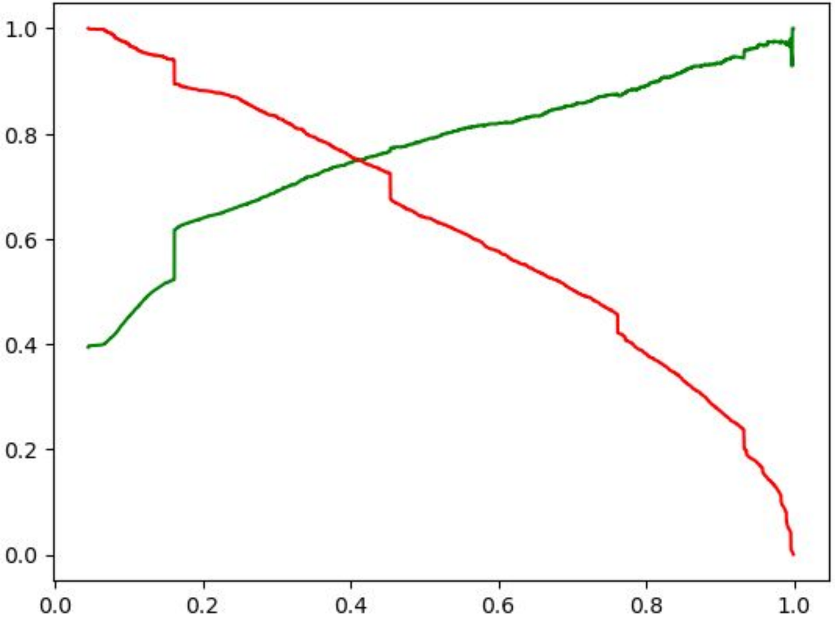
Model Evaluation - Predicting on Train set



- We plot the accuracy. specificity and sensitivity together to get the optimum range of the cutoff, in thi case it comes out to be 0.35
- We change the Conversion_Prob with the new cutoff we got

Model Evaluation - Predicting on Test Set

- X_test = X_test[column]
- X_test_sm = sm.add_constant(X_test)
- y_test_pred = logm.predict(X_test_sm)
- We do the predictions on the test set using the above code and again calculate the Sensitivity and Specificity



- We do a precision and Recall tradeoff to get the new cutoff of 0.41
- We again calculate the Sensitivity and Specificity

Conclusion

- ## Top Predictors to our model
  - Total Visits
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
  - Last Notable Activity_Had a Phone Conversation
  - What is your current occupation_Working Professional

- ## Negative influencers to our model
  - Do not email yes
  - lead origin_landing page submission

Recommendations

- We need to drive people on our portal and make them spend more time on it
- We can give them a demo or a trail course for them to interact with our portal
- We need to make a phone conversation with the leads and understand what are they looking for in a course and explain them the benefits of taking up our courses
- We need to focus on working professionals more, giving them courses to upgrade their skill set
- When people phone do not email at yes that means that they are not much interested in our courses and we shouldn't focus on them much