International Institute of Information Technology Bangalore

upGrad

Lead Scoring Case Study

By : Mahesh Kumar P & Pritha Tyagi

International Institute of Information Technology Bangalore

upGrad

Business Problem

- 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.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- For example, if say, they acquire 100 leads in a day, only about 30 of them are converted.
- The objective is to build a model to identify the hot leads and achieve lead conversion rate to 80%.

International Institute of Information Technology Bangalore

upGrad

Business Objective

“The Business Objective Is To Build A Logistic Regression Model To Identify The Hot/Potential Leads And Achieve The Lead Conversion Rate To 80%.”

International Institute of Information Technology Bangalore

upGrad

Goal of The Case Study

1.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.

2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

International Institute of Information Technology Bangalore

upGrad

Understanding Dataset

- We got a file named "Leads.csv" provided with a leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- To learn more about the dataset we got the data dictionary.
- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
- Another thing that to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

International Institute of Information Technology Bangalore

upGrad

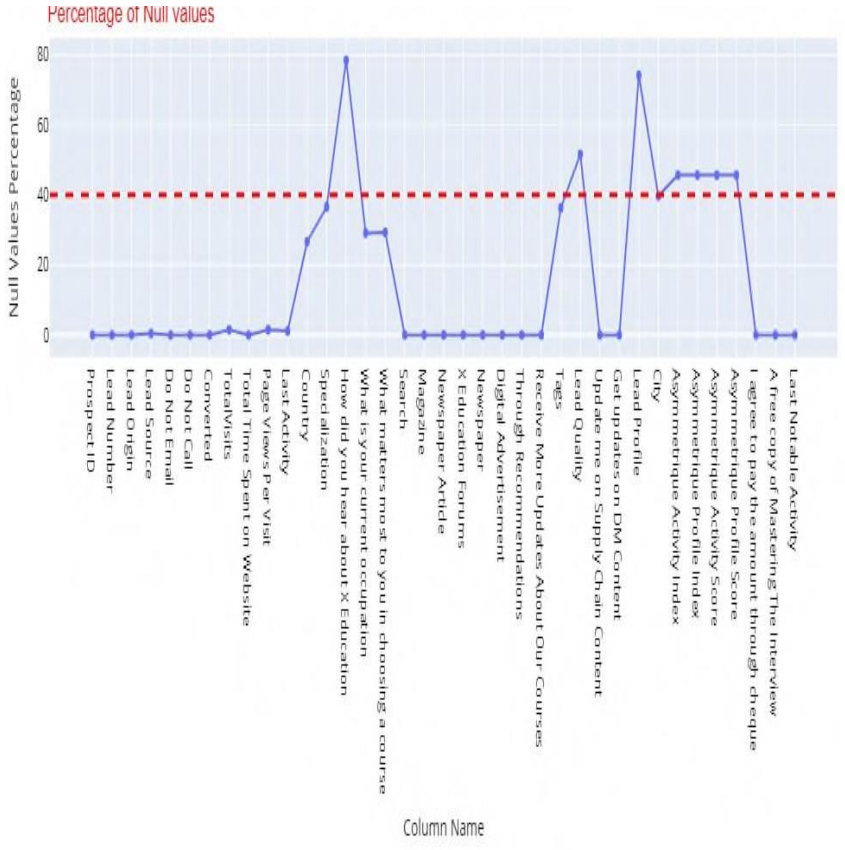
Strategy For Data Cleaning

a. Check percentage of null values in columns and drop the columns which have more than 45% missing values.

b. Also, some of the variables are created by the sales team once they contact the potential lead. These variables will not be available for the model building as these features would not be available before the lead is being contacted. We will drop these columns too.

c. Some of the columns have only 1 category. These columns will not add any value to the model and can be deleted.

d. Some of the columns have one of the value as "Select" These should be considered as null values. Data Value needs to be updated for these columns



International Institute of Information Technology Bangalore

upGrad

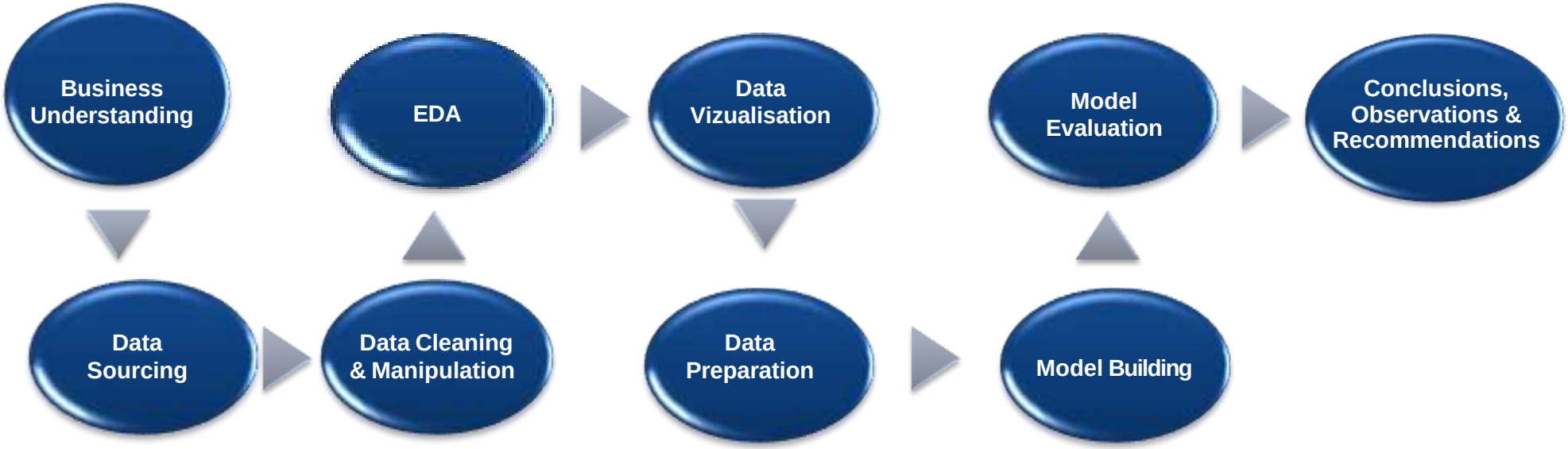
Data Knowledge

- Dataset used : "**Leads.csv** "
- Total number of customers present : **9240**
- Total number of features : **37**
- Model used : **Logistic Regression**
- After initial analysis, we see that there are multiple factors that influence conversion rate.
- The target column in our dataset : **"Converted" .**
- We need to reduce the features to maximize the conversion rate.
- Current Conversion Rate = **38.53%**

International Institute of Information Technology Bangalore

upGrad

Technical Approach For Solving Business Problem



International Institute of Information Technology Bangalore

upGrad

Data Cleaning

Handling 'Select' variable

- "Select" variable indicates that the user has not selected any option.
- We impute the same with null values.

Dropping Score and Activity variables

- Score and Activity :

  This is the data that is obtained after contact with the lead. So we need to remove them.
- Score variables:

  Tags, Lead Quality, Lead Profile, Activity Index, Activity Score and Profile Score.
- Activity variables:

  Last Notable Activity

Treating Categorical data

- High Data Imbalance – Columns having high data imbalance must be removed. For e.g. : Category A has 98% , and Category B has 2% - This data is irrelevant to our analysis as one category is overpowering the other.
- In other categorical columns where there are columns with small percentages should be removed

Dropping column with high null values

- Columns having null values greater than 40% does not have meaning to the data, hence we drop these columns
- For Specialization, we consider the column where people have not selected any value into one more column known as Not Specified and we use this for model building.

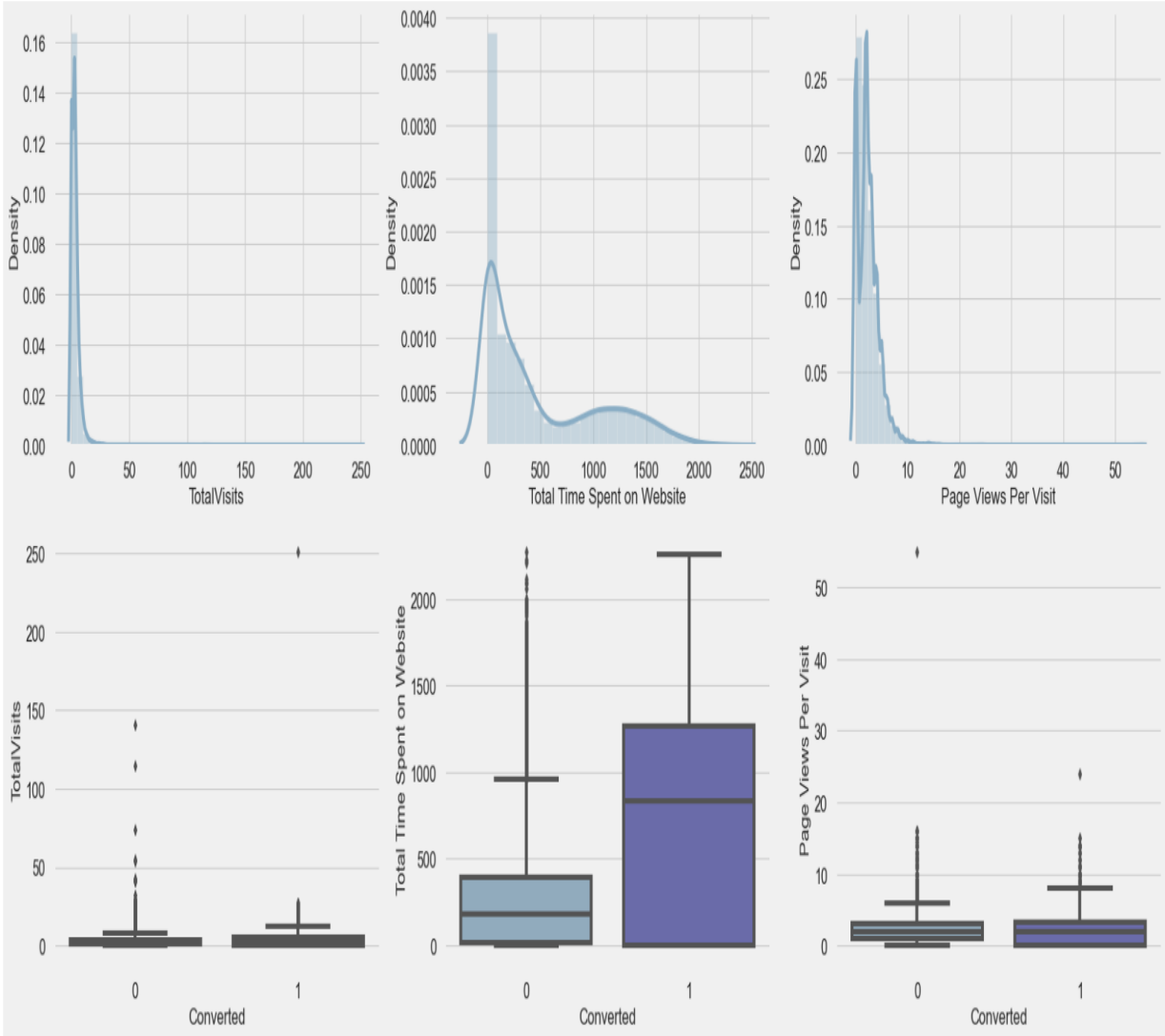
Final list of features

- Lead Origin
- Lead Source
- Do Not Email
- Do Not Call
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Last Activity
- Specialization
- Occupation
- Search
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Recommendation
- Free Copy

International Institute of Information Technology Bangalore

upGrad

EDA : Numerical Data



Total Visits

- The max probability for Total Visits is found to be around 15-20. It increases initially but decreases further.
- The average total visits for both converted and non converted people is found to be the same.
- Has some outliers which needs to be treated.

Total Time Spent On The Website

- The probability of time spent is found to be high for time between 0-300 seconds and decreases further.
- The mean is found to be higher in case of Converted people rather than non-converted people.

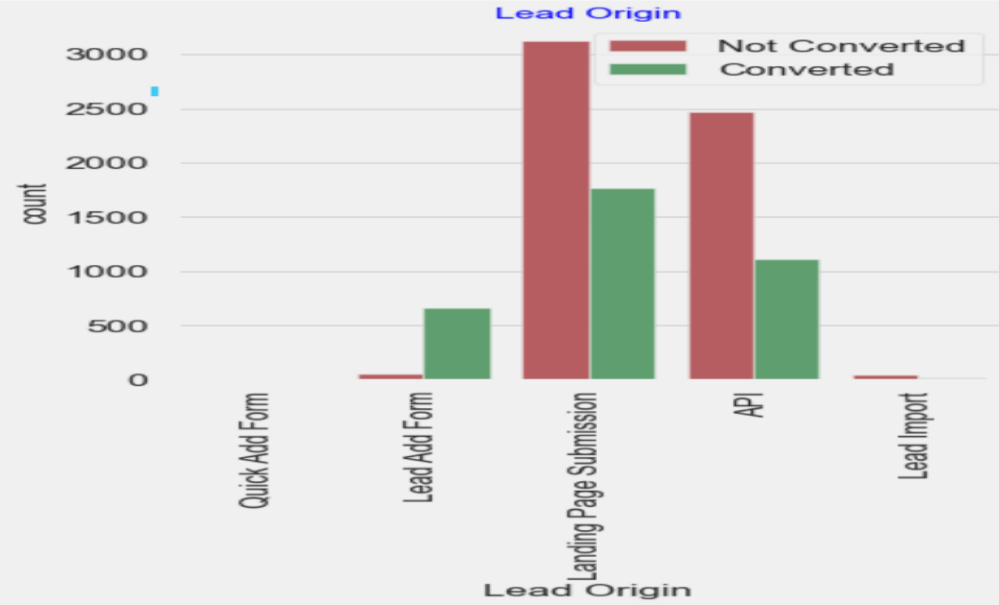
Page Views Per Visit

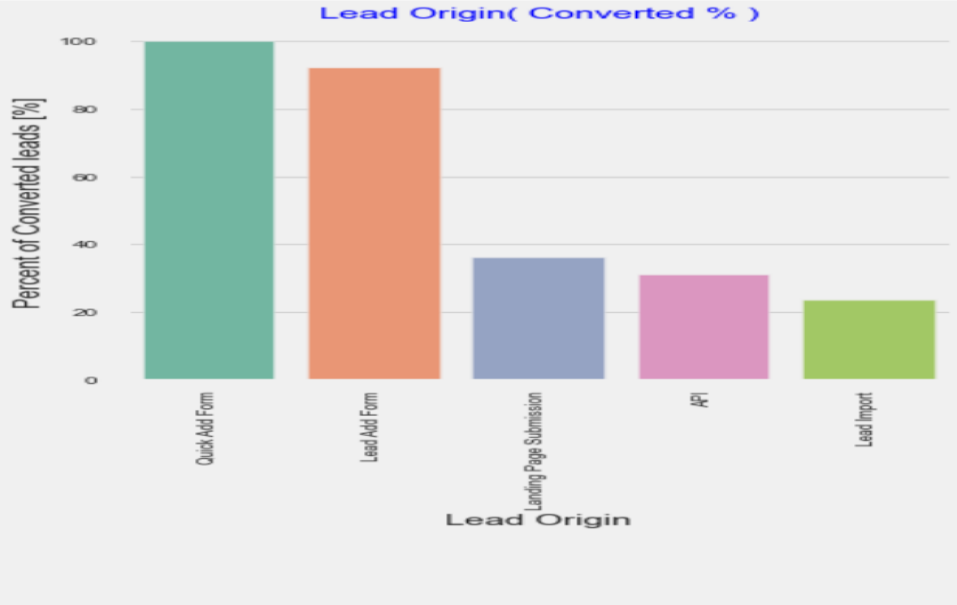
- The max probability for Page Views Per Visit is found to be around to be 3-5.
- The average page views for both converted and non converted is found to be the same.
- Has some outliers which needs to be treated.

International Institute of Information Technology Bangalore

upGrad

EDA : Categorical Data





Most Leads originated from submissions on the landing page and around 38% of those are converted followed by API, where around 30% are converted.

Even though Lead Origins from Quick Add Form are 100% Converted, there was just 1 lead from that category. Leads from the Lead Add Form are the next highest conversions in this category at around 90% of 718 leads.

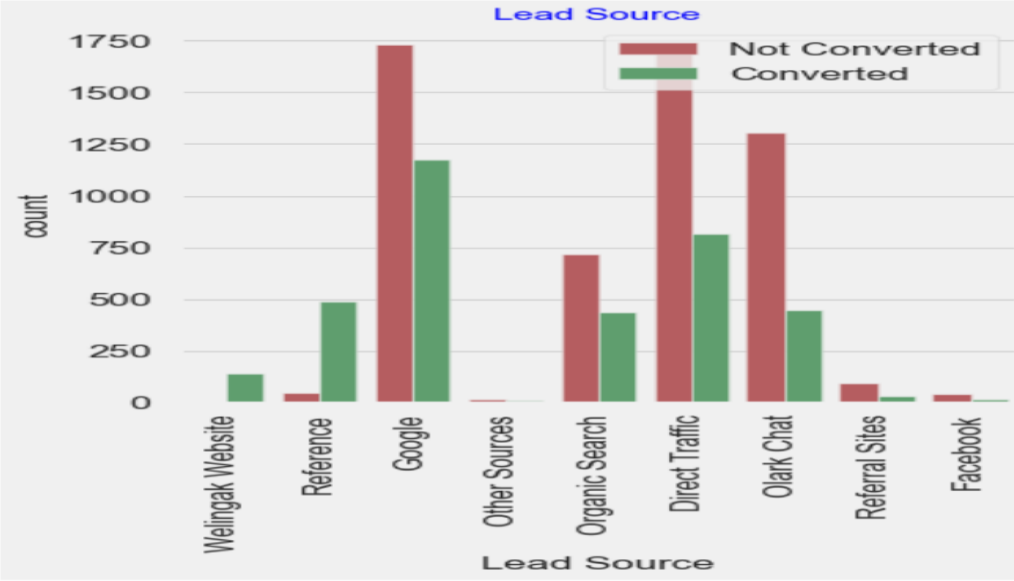
Lead Import are very less in count and conversion rate is also the lowest

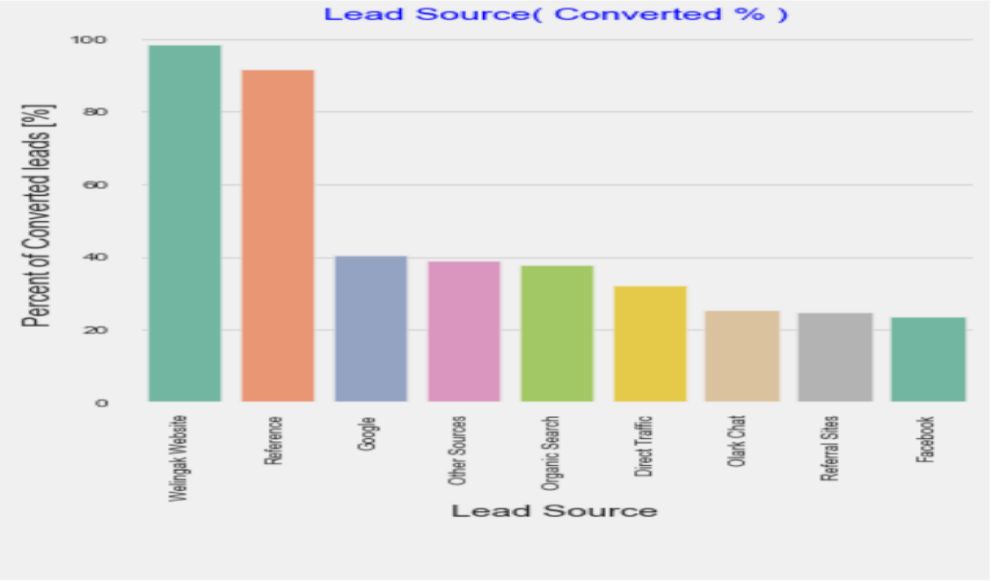
To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form..

International Institute of Information Technology Bangalore

upGrad

EDA : Categorical Data





We combined smaller lead sources as 'Other Sources'

The source of most leads was Google, and 40% of the leads converted, followed by Direct Traffic,Organic search and Olark chat where around 35%, 38% and 30% converted respectively

A lead that came from a reference has over 90% conversion from the total of 534.

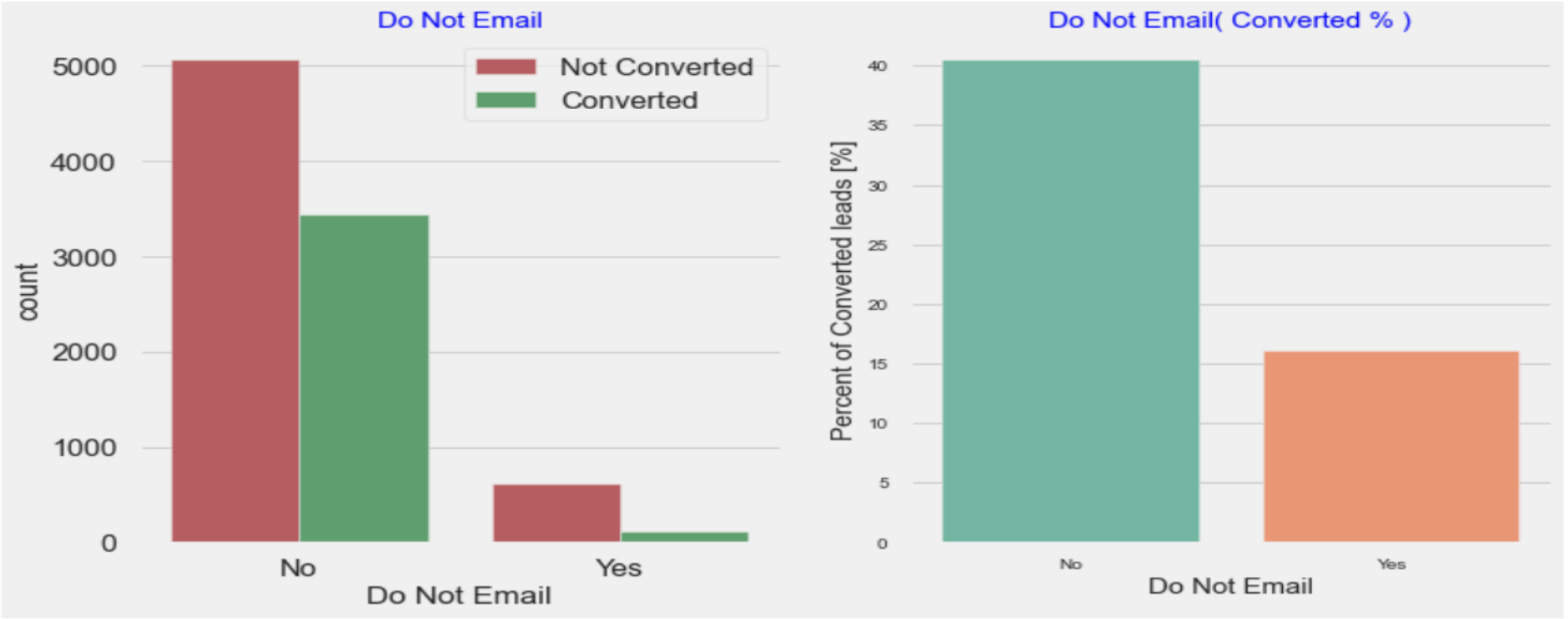
Welingak Website has almost 100% lead conversion rate. This option should be explored more to increase lead conversion

To increase lead count, initiatives should be taken so already exitsing members increase their referrals.

International Institute of Information Technology Bangalore

upGrad

EDA : Categorical Data



Majority of the people are ok with receiving email (~92%)

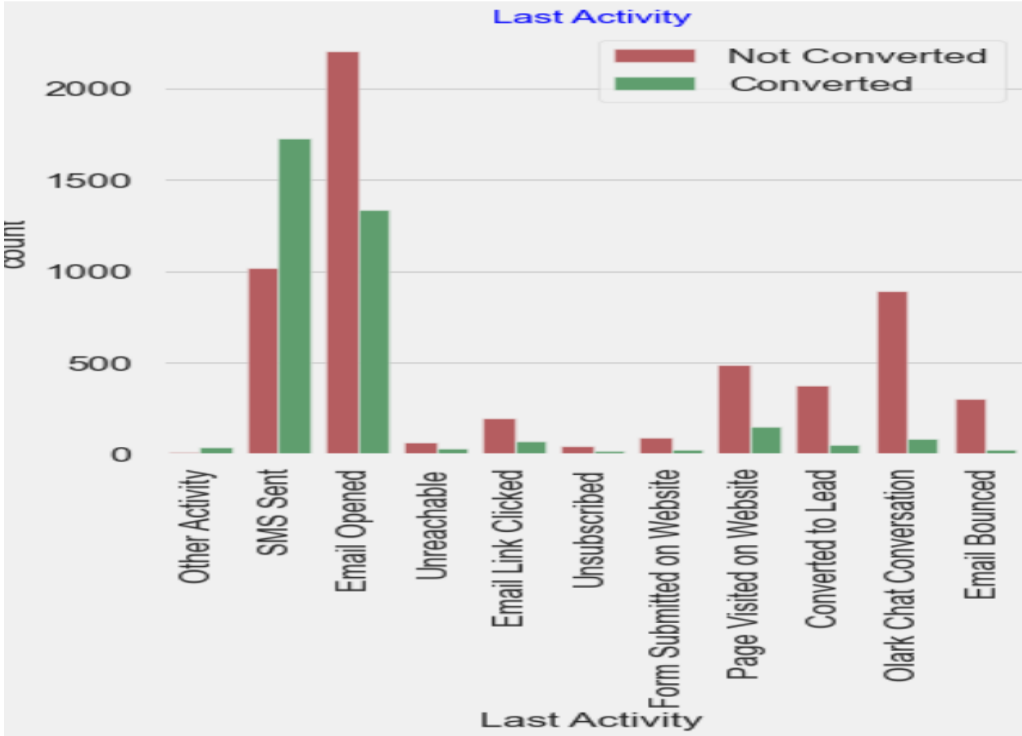
People who are ok with email has conversion rate of 40%

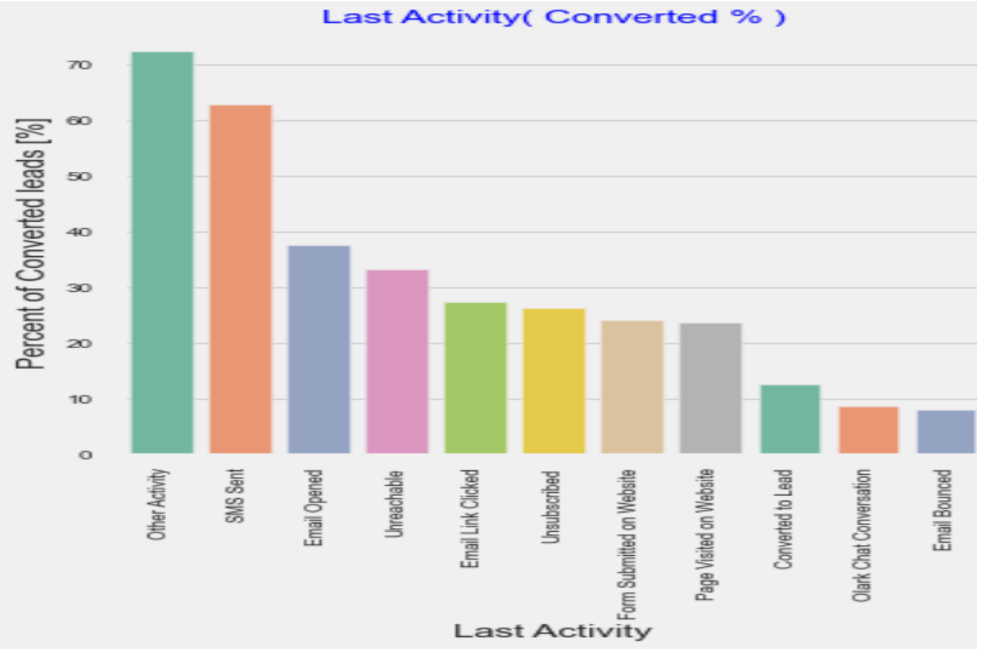
People who have opted out of receive email has lower rate of conversion (only 15%)

International Institute of Information Technology Bangalore

upGrad

EDA : Categorical Data





We combined smaller lead sources as Other Activity

Most of the lead have their Email opened as their last activity

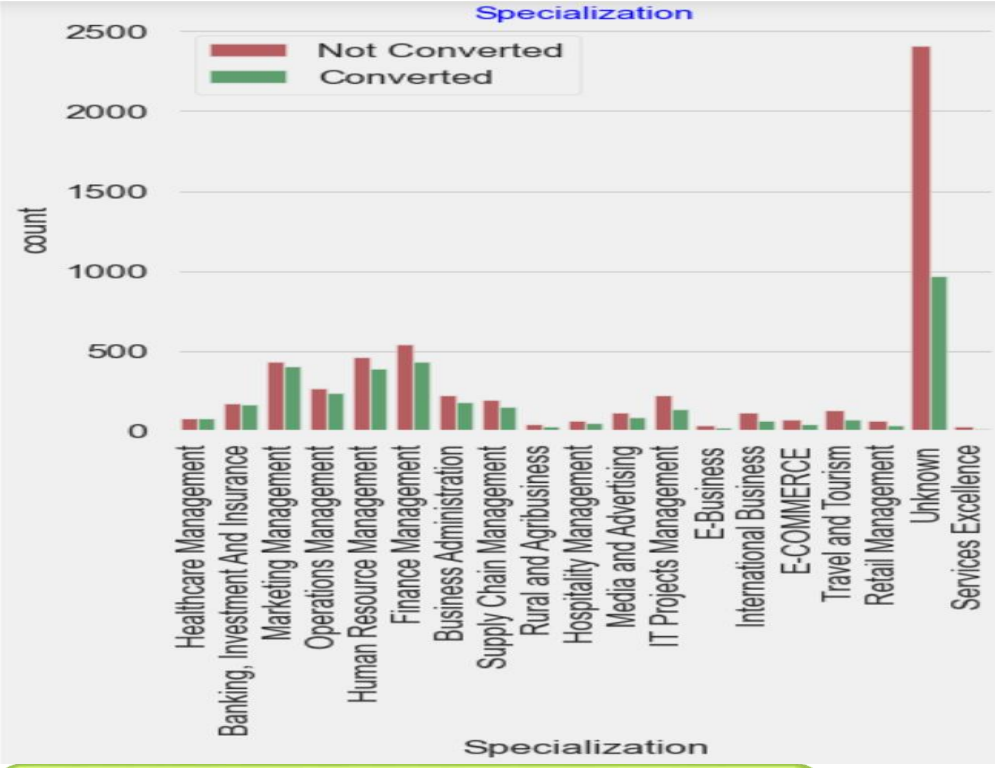
After combining smaller Last Activity types as Other Activity, the lead conversion is very high (~70%)

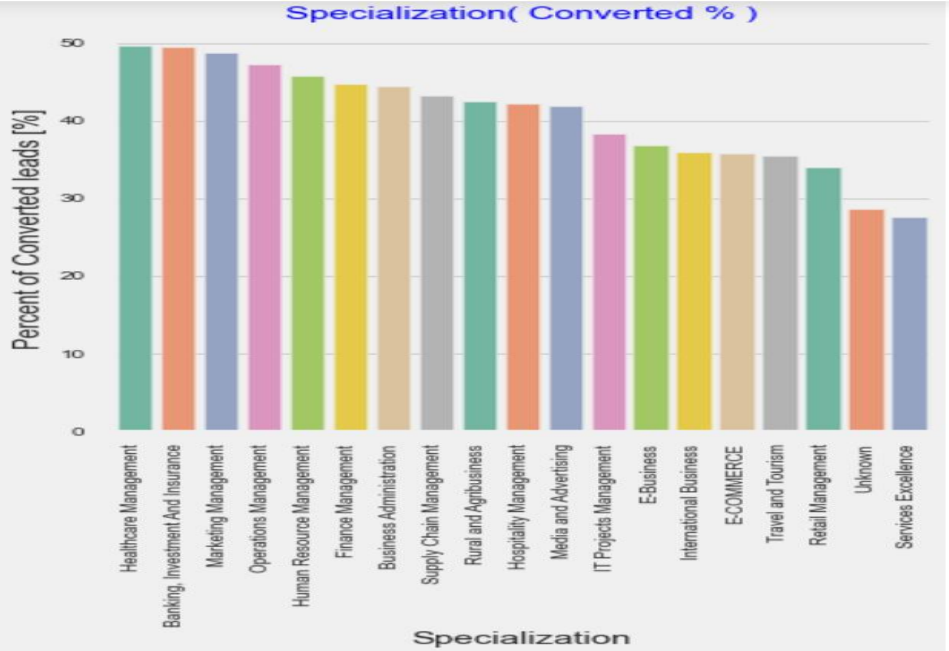
Conversion rate for leads with last activity as SMS Sent is almost 60%

International Institute of Information Technology Bangalore

upGrad

EDA : Categorical Data





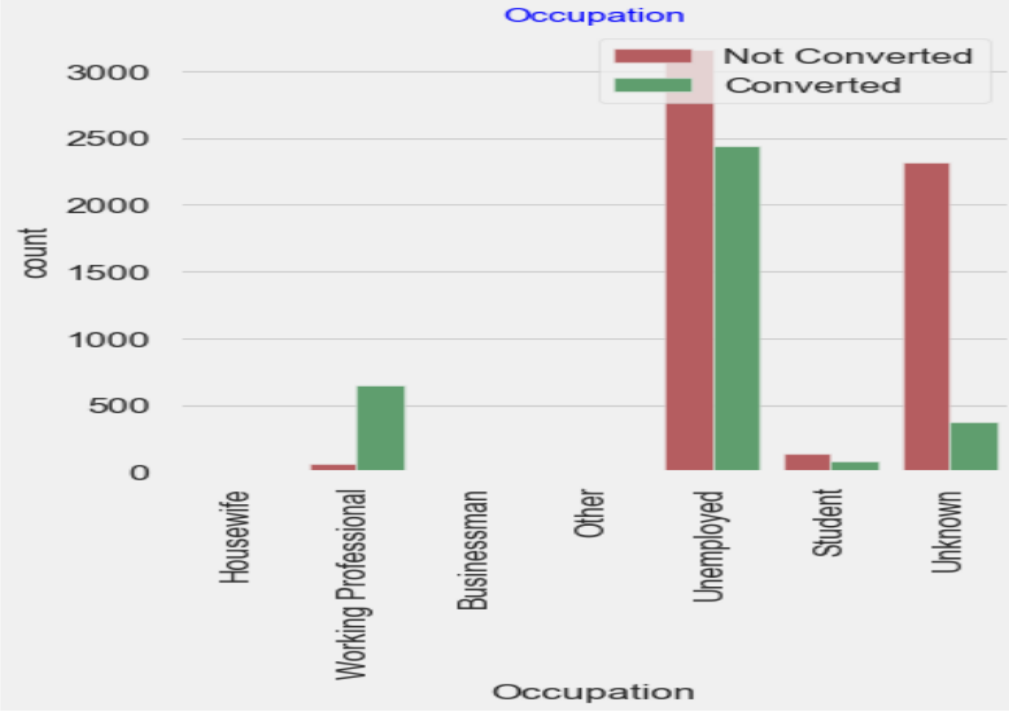
Most of the leads have not mentioned a specialization and around 28% of those converted

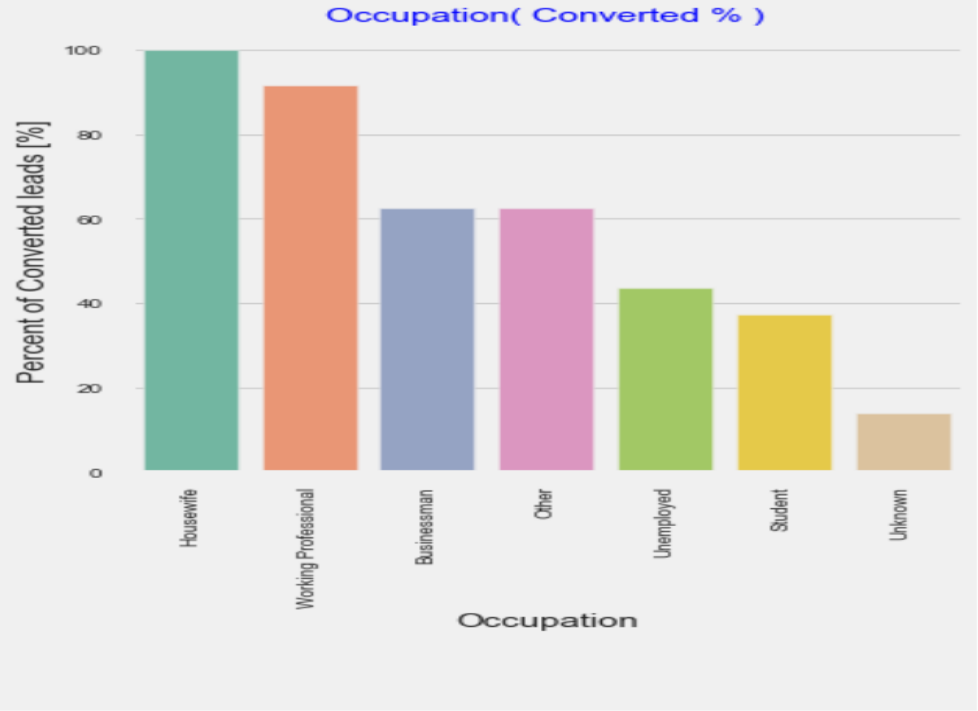
Leads with Finance management and Marketing Management - Over 45% Converted

International Institute of Information Technology Bangalore

upGrad

EDA : Categorical Data





Though Housewives are less in numbers, they have 100% conversion rate

Working professionals, Businessmen and Other category have high conversion rate

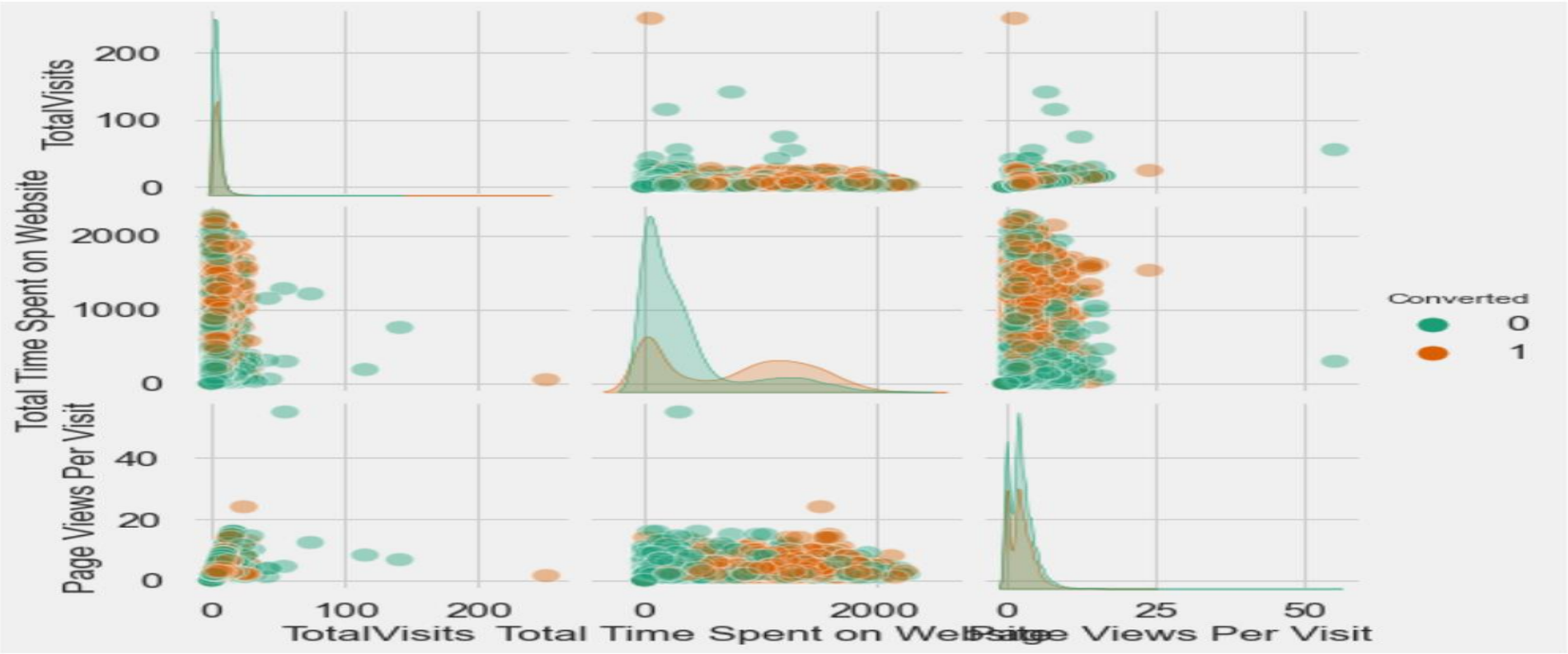
Though Unemployed people have been contacted in the highest number, the conversion rate is low (~40%)

We cannot combine smalled value categories as their conversion rate is very different. Combing them may provide wrong predictions.

International Institute of Information Technology Bangalore

upGrad

EDA : Bivariate Analysis



Data is not normally distributed.

International Institute of Information Technology Bangalore

upGrad

Data Preprocessing

- Number of features after scaling and dummy variable creation : 51
- Target Variable : Converted o Libraries used: StandardScaler()
- Columns that are not considered : Lead Number and Prospect ID (these variables do not help in model building)
- The steps are as follows:

Outlier Treatment:

- Though outliers in TotalVisits and Page Views Per Visit shows valid values, this will misclassify the outcomes and consequently create problems when making inferences with the wrong model. Logistic Regression is heavily influenced by outliers. So lets cap the TotalVisits and Page Views Per Visit to their 95 th percentile due to following reasons:

Data set is fairly high number 95th and 99th percentile of these columns are very close and hence impact of capping to 95th or 99th percentile will be the same

Binary Mapping:

- "A Do Not Email contains values in terms of Yes/No , we convert these to 1/0 so it converts into numerical values and helps in model building.

Dummy Variable Creation:

- We need to create dummy variables for all the categorical columns as they enable us to use a regression equation on multiple groups.
- List Of Columns Converted into Dummy :
- Lead Origin', 'Lead Source', 'Occupation', 'Last Activity', 'Specialization'

Test Train Split:

- Division of data into test data and train data to check the stability of the model.
- We have randomly sampled 70% of the data as the test data and 30% of the data as test data.
- Random State = 100

Scaling:

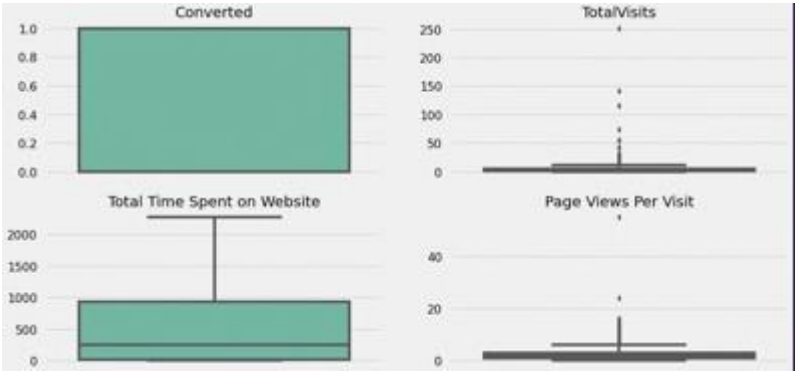
- Division of Train Data into X and Y where X has all the features and Y has the target variable – Converted.
- We perform scaling to normalize the data within a particular range
- Technique : Standard Scaler

International Institute of Information Technology Bangalore

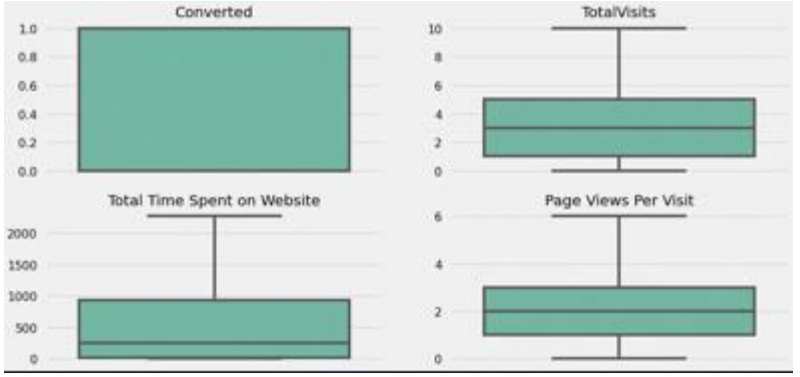
upGrad

Outlier Treatment

Though outliers in TotalVisits and Page Views Per Visit shows valid values, this will misclassify the outcomes and consequently create problems when making inferences with the wrong model. Logistic Regression is heavily influenced by outliers. So lets cap the TotalVisits and Page Views Per Visit to their 95 th percentile due to following reasons:



- Data set is fairly high number.
- 95th percentile and 99th percentile of these columns are very close and hence impact of capping to 95th or 99th percentile will be the same.



International Institute of Information Technology Bangalore

upGrad

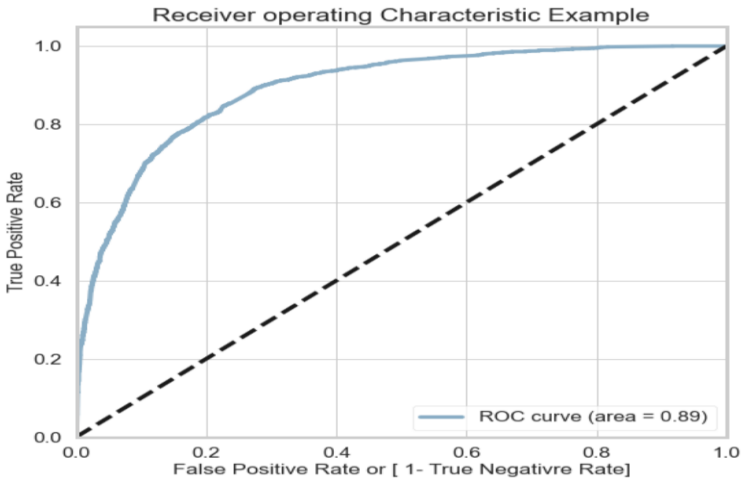
Model Building

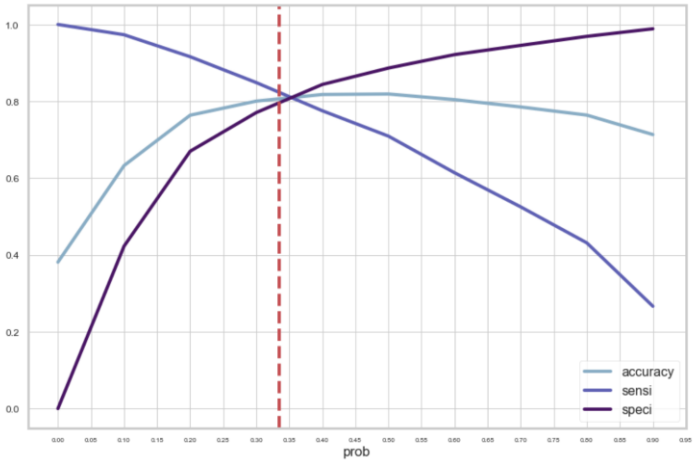
Model – I to V: Basic Model1	• We build a basic model using 20 features. Since it is not efficient we perform RFE to obtain a model with • Top – 20 features. There are so many variables with high p-values and VIF value, we need to remove them. • We will remove **'Occupation_Housewife'** feature due to high P-value of 0.999
Model – II	• We removed **'Occupation_Housewife'** feature and build model 2 with 19 features. • Since p –value is 0.209 for **'Specialization_Retail Management'** due to high P-value we need remove this feature
Model -III	• We removed **'Specialization_Retail Management'** feature and build model 3 with 18 features. • Again We have high P-value of 0.204 for **'Lead Source_Facebook'**
Model IV	• We removed **'Lead Source_Facebook '** feature and build model 3 with 18 features. • Again We have high P-value of 0.174 for **Specialization_Rural and Agribusiness**
Model – VIII : The Final Model	• We removed **'Specialization_Rural and Agribusiness'** feature due to high P-Value of 0.174. All the parameters have VIF values below 3, which indicates that features are not Multi-colinear in nature. • Final model has 16 features in total.

International Institute of Information Technology Bangalore

upGrad

ROC Curve And Optical Cut-Off Probability





- ROC Curve represents how much the model is able to distinguish between the classes.

- AUC – Area under the curve represents that it is distinguishing the 1's and 0's correctly.

- On plotting the ROC curve for our data we see that, AUC is around 0.89 which means at around 89% of the times, the model is able to distinguish the 1's as 1's and 0's as 0's.

- AUC of 0.89 is found to be very stable model.

- When we plot the sensitivity, accuracy and specificity of the model together, the optimal cut off point is found to be at 0.35. This means that at 35% probability, the sensitivity and specificity are found to be balanced.

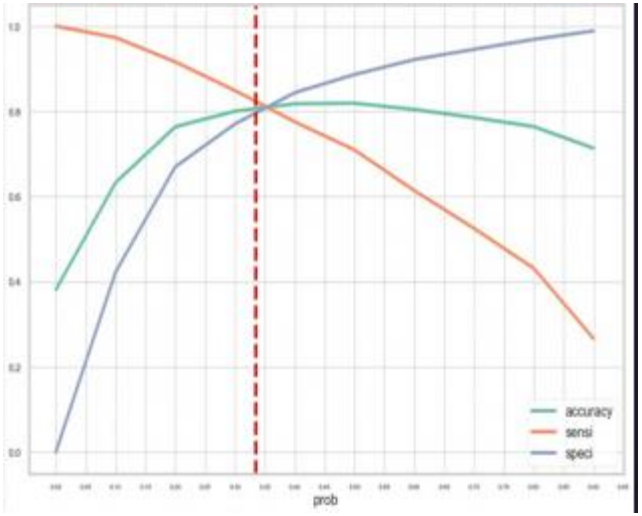
- With probability = 0.35 , we predict y-values with X- Train, in such a way that, any conversion prob > 35% is said to be converted to a lead.

International Institute of Information Technology Bangalore

upGrad

Model Evaluation –Train Set

Optimal cut-off



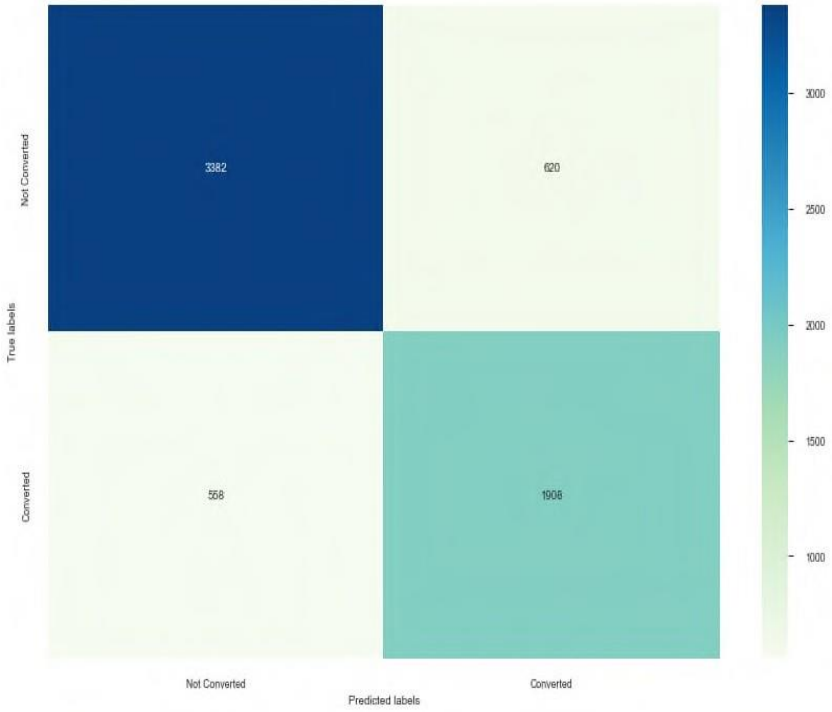
Model Performance

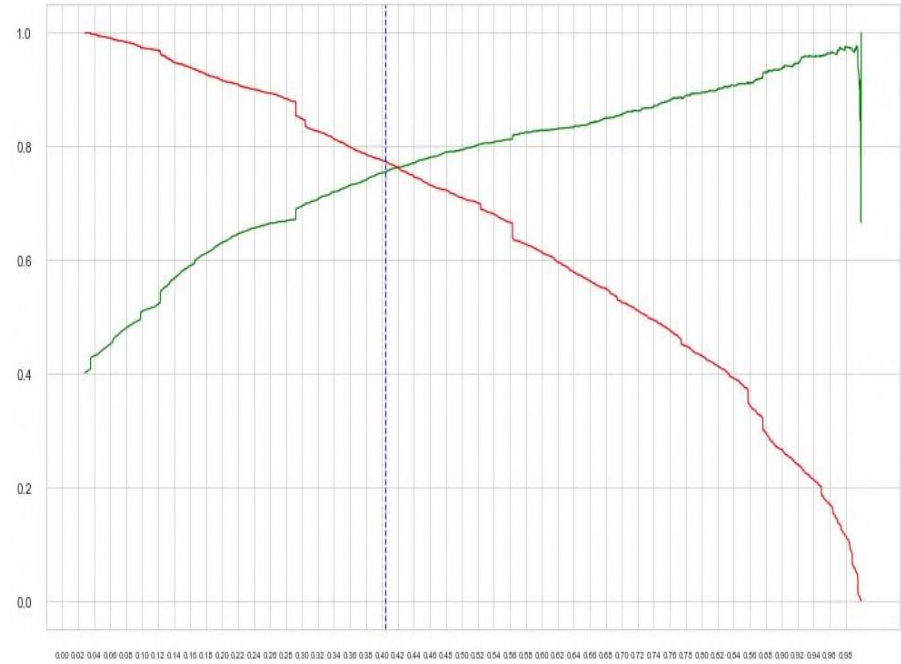
Model Accuracy value is	:	80.71 %
Model Sensitivity value is	:	81.79 %
Model Specificity value is	:	80.03 %
Model Precision value is	:	71.63 %
Model Recall value is	:	81.79 %
Model True Positive Rate (TPR)	:	81.79 %
Model False Positive Rate (FPR)	:	19.97 %
Model Poitive Prediction Value is	:	71.63 %
Model Negative Prediction value is	:	87.71 %

International Institute of Information Technology Bangalore

upGrad

Precision Recall Trade Off





International Institute of Information Technology Bangalore

upGrad

Model Performance Test

Train Set

Model Accuracy value is : 80.71 %

Model Sensitivity value is : 81.79 %

Model Specificity value is : 80.03 %

Model Precision value is : 71.63 %

Vs

TEST Set

Model Accuracy value is : 80.7 %

Model Sensitivity value is : 81.83 %

Model Specificity value is : 79.96 %

Model Precision value is : 72.73 %

The sensitivity value on Test data is 81.83% vs 80.29% in Train data. The accuracy values is 80.7%.

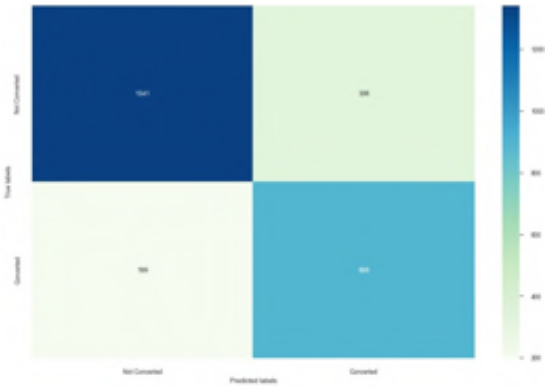
It shows that model is performing well in test data set also and is not over-trained.

International Institute of Information Technology Bangalore

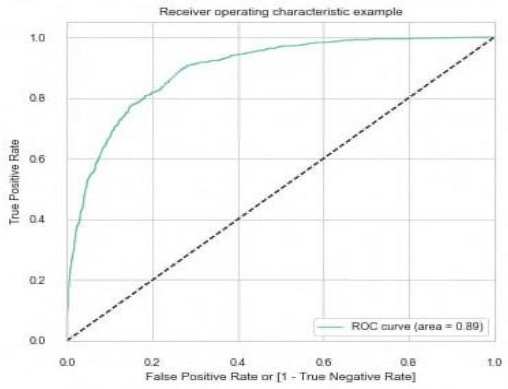
upGrad

Model Evaluation Test -Set

Confusion matrix



ROC Curve



Model Performance

Model Accuracy value is	:	80.7 %
Model Sensitivity value is	:	81.83 %
Model Specificity value is	:	79.96 %
Model Precision value is	:	72.73 %
Model Recall value is	:	81.83 %
Model True Positive Rate (TPR)	:	81.83 %
Model False Positive Rate (FPR)	:	20.04 %
Model Poitive Prediction Value is	:	72.73 %
Model Negative Prediction value is	:	87.08 %

International Institute of Information Technology Bangalore

upGrad

Lead Score And Conversion Rate

- Conversion Rate is the number of customers who are converted to leads and interested in the course.
- Before model building the Conversion Rate was found to be 38.53% .
- After model building, the conversion rate is increased to 81.7% .
- Hence we can conclude that our final model has served to the business purpose.

Steps taken to assign a lead score variable for all customers.

1) Train the data with the model.

- Run the model on the entire Leads dataset.
- Do not divide into Test and Train and run the obtained LR model on the entire data frame

2) Predict the Probability using Cutoff

- Predict the Conversion probability for all the customers using the cutoff value = 0.35.
- Create a new data frame and store Conversion_Probability and actual converted values in this.

3)Adding Lead Scorefor all variables

- Create a new column called Lead Score.
- Convert the probability score into Lead Score by multiplying by 100 and store it in this column.

4) Calculate the conversion rate

- Once we obtain the complete model result on the data, we filter only the leads as predicted by the model.
- Calculate the Conversion Rate using this filtered result.

International Institute of Information Technology Bangalore

upGrad

Hot Leads

- Hot leads are people who have a high probability to be converted as a Lead and thus needs to be identified. They have a higher conversion rate.

- The leads whose lead score is greater than 35% are considered as potential leads. The conversion rate is around 81%. When we increase this threshold from 35% to 95% we get Hot Leads.

- Conversion Rate for hot leads is increases from 73% to 96%. This means they have a 96% probability of getting converted to a lead.

- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

International Institute of Information Technology Bangalore

upGrad

Conclusion

Equation :-

-0.7911 * const + 0.1944 * Do Not Email - 1.1811* Time Spent on Web site +1.0651 * Lead Origin_Landing Page Submission -1.0227 * Lead Origin_Lead Add Form + 2.8029 * Lead Source_Olark Chat-1.0093 * Lead Source_Welingak Website 2.4629 * Occupation_Unknown - 1.0818 * Occupation_Working Professional +2.3996 * Last Activity_Email Opened + 0.7288 * Last Activity_Olark Chat Conversation - 0.6068 * Last Activity_Other Activity +2.2419 * Last Activity_Unreachable +0.8487 * Last Activity_Unsubscribed +1.3906 * Last Activity_SMS Sent -1.8672 * Specialization_Hospitality Management – 0.9951 * Specialization_Unknown -0.9785

From our model, we can conclude following points :

- ❑ The customer/leads who fills the form are the potential leads.
- ❑ We must majorly focus on working professionals.
- ❑We must majorly focus on leads whose last activity is SMS sent or Email opened.
- ❑It's always good to focus on customers, who have spent significant time on our website.
- ❑It's better to focus least on customers to whom the sent mail is bounced back.
- ❑ If the lead source is referral, he/she may not be the potential lead.
- ❑If the lead didn't fill specialization, he/she may not know what to study and are not right people to target. So, it's better to focus less on such cases.

International Institute of Information Technology Bangalore

upGrad

Inferences & Recommendations

LEAD GENERATION

Major indicators that a lead will get converted to a hot lead

1. **Lead Origin_Lead Add Form : A lead sourced from Lead Origin_Lead Add Form is more likely to get converted.**
2. **Occupation_Working Professional :- Working professionals are more likely to get converted.**
3. **Lead_Source_Welingak website : A lead sourced from Welingak Website is more likely to get converted.**
4. **Last Activity_SMS Sent :A lead having SMS sent previously are more likely to get converted.**
5. **Lead Source_Olark Chat :A lead sourced from Olark Chat is more likely to get converted**

International Institute of Information Technology Bangalore

upGrad

Inferences & Recommendations

LEAD GENERATION

Major indicators that a lead will NOT get converted to a hot lead:

1. **Last_Activity_Olark chat conversation : Customer who had olark chat conversion, are less likely to get converted into hot leads.**
2. **Lead Ongin_Landmg Page Submission : Customer who hadLead Ongin_Landmg Page Submission, are less likely to get converted into hot leads .**
3. **Do Not Email :Customer who choose Do Not Email, are less likely to get converted into hot leads .**

RECOMMENDATIONS

The company should use a leads score threshold of 34 to identify "Hot Leads" as at this threshold, Sensitivity Score of the model is around 81% which is as good as CEO's target of 80%

International Institute of Information Technology Bangalore

upGrad

Thank You