Lead Score Case Study

Submitted by :

Muthu Ram B O

Vikram Mylarapu

Abhishek Singh

AGENDA

- Problem Statement
- Dataset Analysis
- Primary goals
- Data outlook
- Summary

Lead Score Case Study for X Education

Problem Statement :

X Education sells online courses to industry professionals. 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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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.

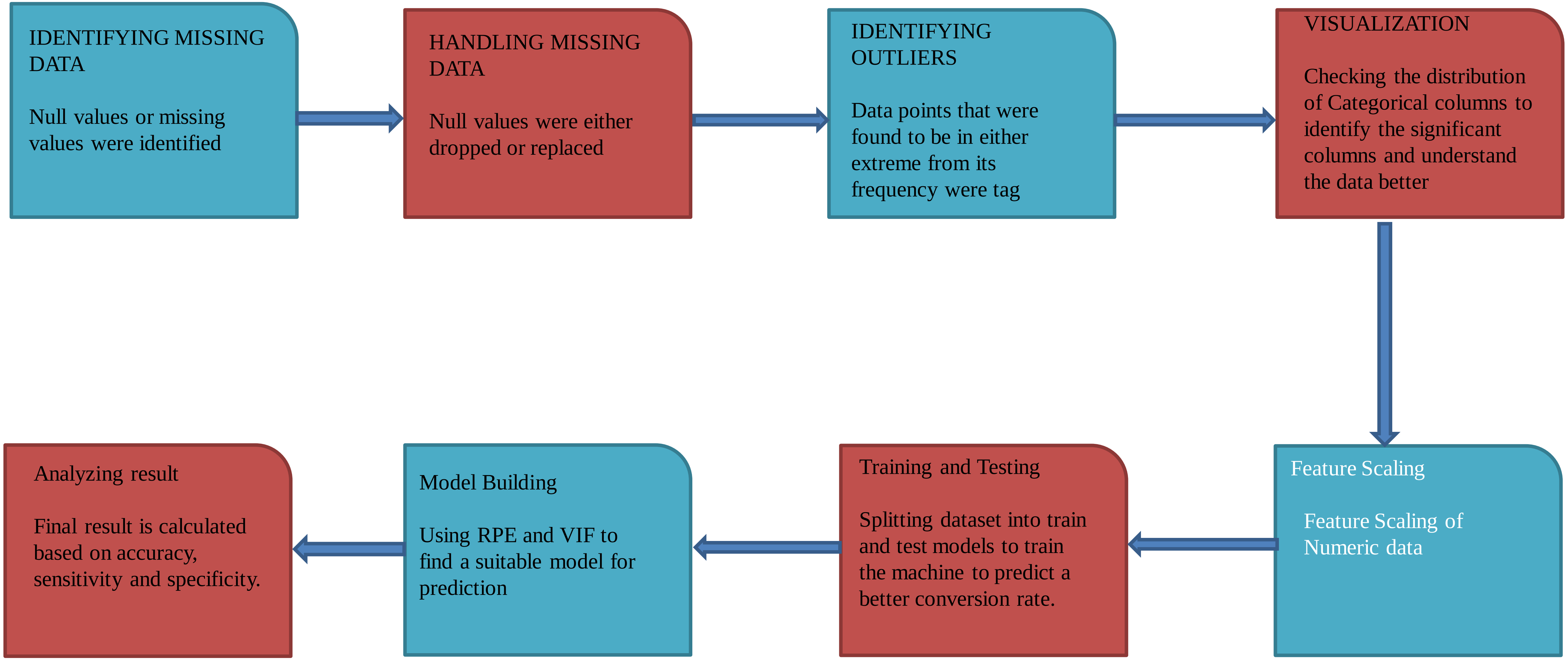
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Dataset

The dataset “Leads” contains:

- ✓ Information provided by clients interacting with course ads or website
- ✓ Contains 9240 rows of data and 37 attributes.
- ✓ Out of these 17 attributes were considered and among them also few were termed critical for analysis.

Dataset Analysis



IDENTIFYING MISSING DATA

Null values or missing values were identified

HANDLING MISSING DATA

Null values were either dropped or replaced

IDENTIFYING OUTLIERS

Data points that were found to be in either extreme from its frequency were tag

VISUALIZATION

Checking the distribution of Categorical columns to identify the significant columns and understand the data better

Feature Scaling

Feature Scaling of Numeric data

Training and Testing

Splitting dataset into train and test models to train the machine to predict a better conversion rate.

Model Building

Using RPE and VIF to find a suitable model for prediction

Analyzing result

Final result is calculated based on accuracy, sensitivity and specificity.

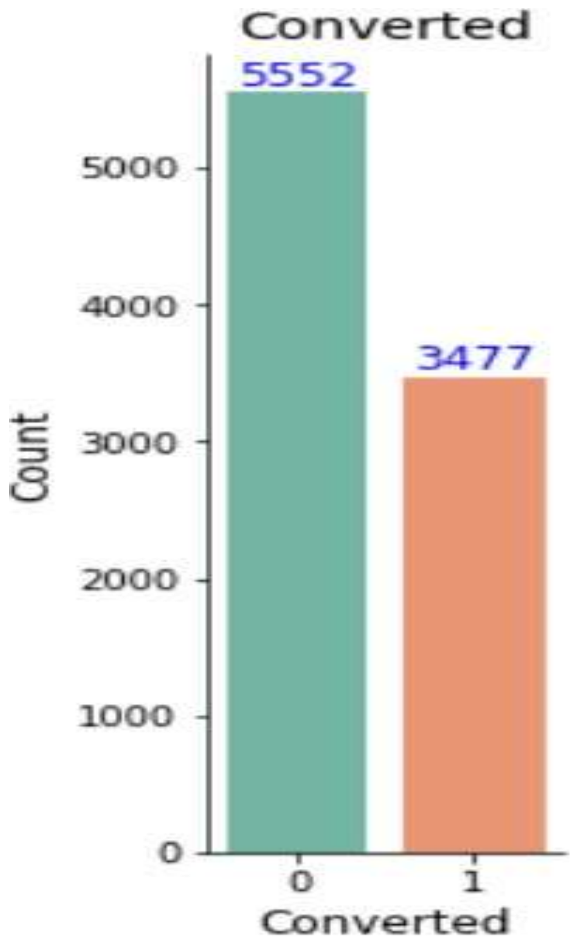
Primary Goals

- Identifying current conversion rate.
- Identifying outliers
- Identifying conversion rates of specific columns
- Identifying best models for training
- Model Evaluation
- Result

Identifying current conversion rate.

- In the graph, 0 represents not converted, 1 represents converted.
- Out of 9240 records, 3477 were converted.
- Therefore, we have around 39% Conversion rate in total.

.



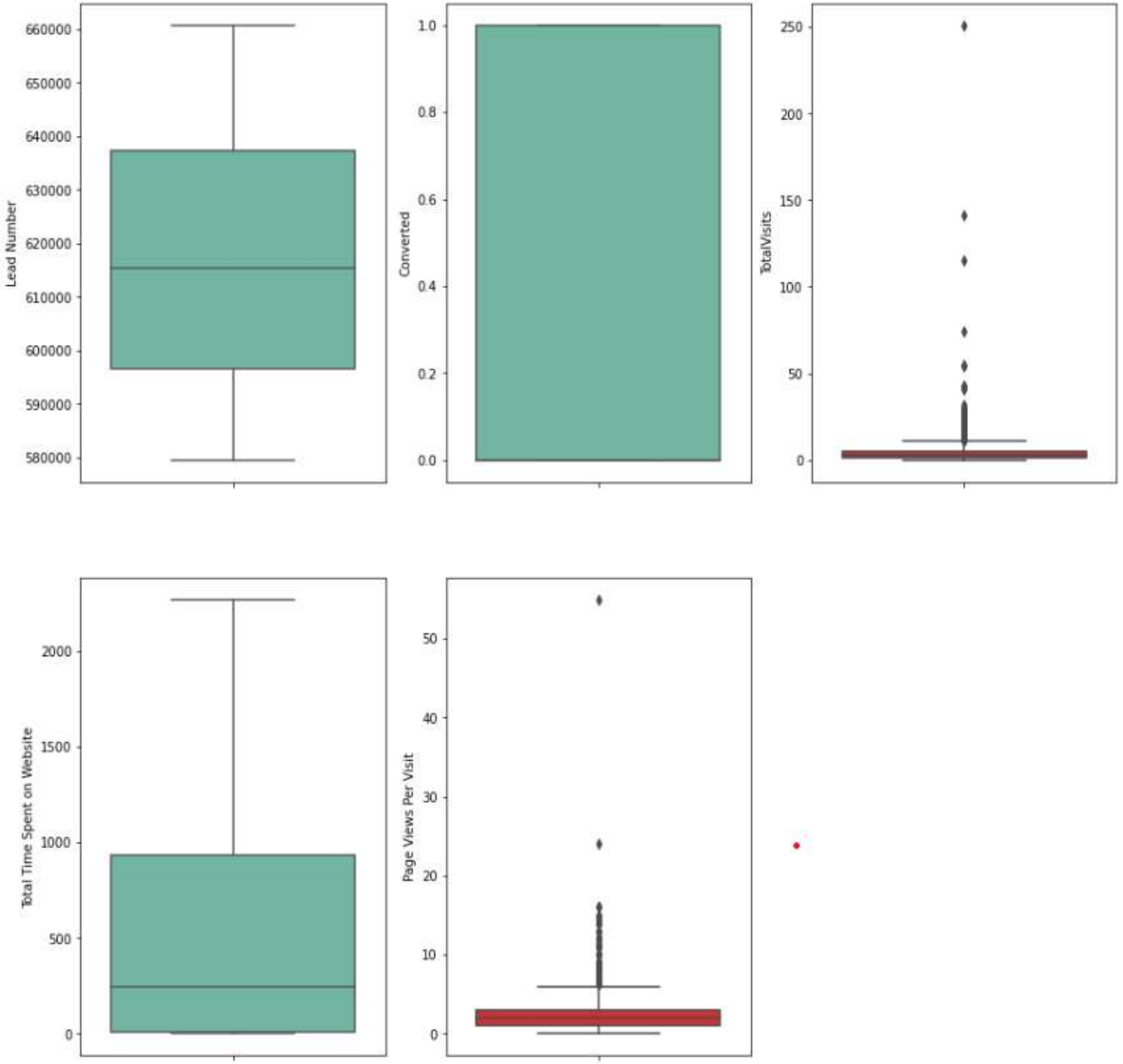
Identifying outliers

- When taking into account all the numerical fields for data analysis, it can be seen that two of the attributes have outliers; i.e. abnormal data present in the dataset.

- These skewed data was handled by removing data above a threshold values.

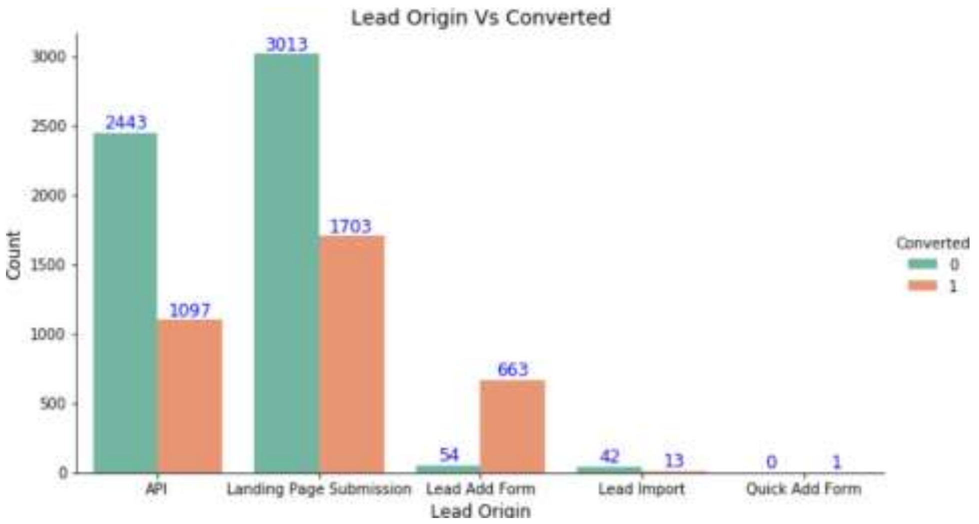
- Outliers were seen in columns: ‘Total Visits’ and ‘Page Views Per Visit’. Since these seem relevant with conversion of leads, handling the data makes more sense.

- After treating the outliers, the dataset became symmetrical and now it can be analysed for the conversion rates. And compared with total conversion rate.

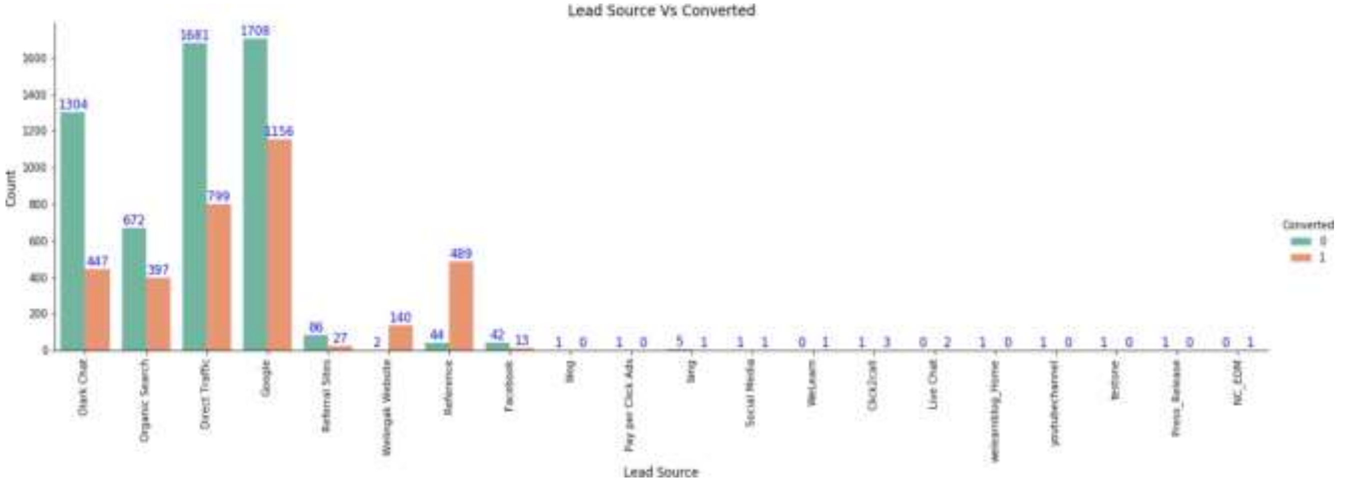


Identifying conversion rates of specific columns

- Lead Origin:
  - As can be seen in the graph, most number of lead originated as well as were converted from the landing page of the website.
  - Second most generated leads were from API, which were also converted the most.

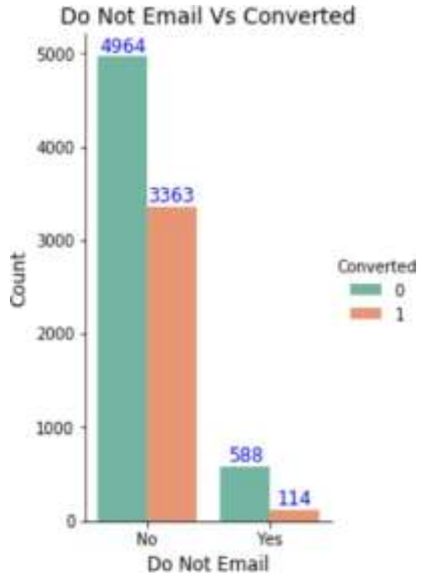


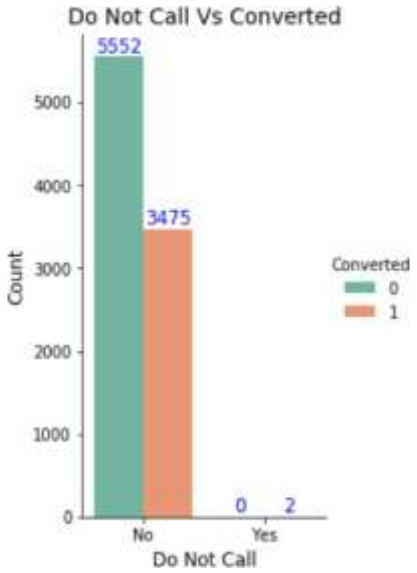
- Lead Source:
  - Most numbers of leads were from Google and among them maximum number of leads were converted.
  - Second largest conversion can be seen via direct traffic of clients on the website.



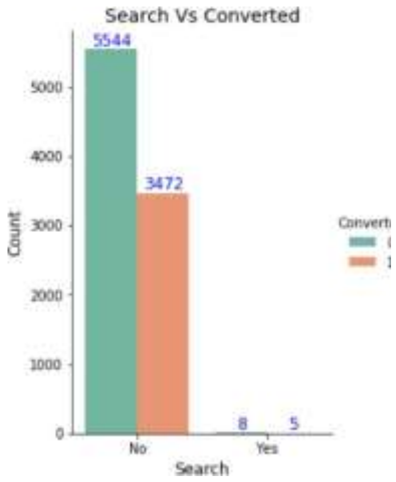
Identifying conversion rates of specific columns

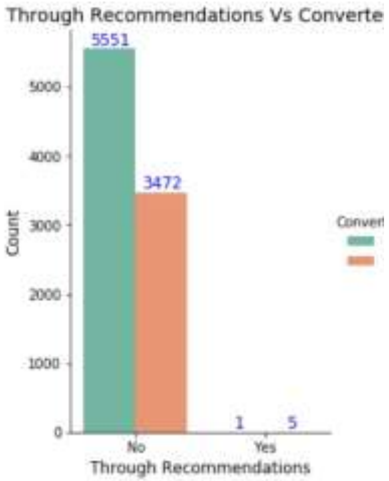
- Do not mail/Do not call:
  - A significant number of leads which opted ‘No’ for mails or calls were converted.
  - Approximately 20% of leads who chose ‘Yes’ for mails were converted, whereas 100% of leads who chose ‘Yes’ for calls were converted.

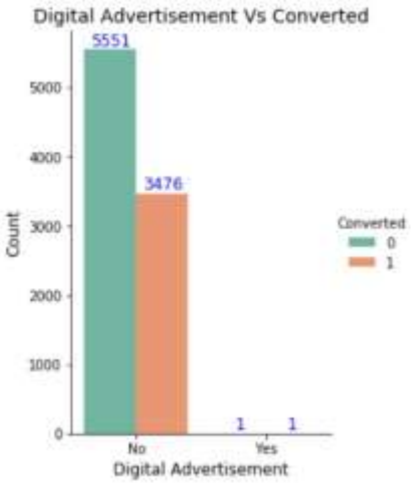




- Search/Recommendations/Digital Advertisement:
  - There is not a huge impact by these fields in the conversion rate.
  - As can be seen from the graph that there are very few selection in leads by these attributes. Although data spikes lie between 50-70%, still there can not be seen any significant changes in the conversion rate.

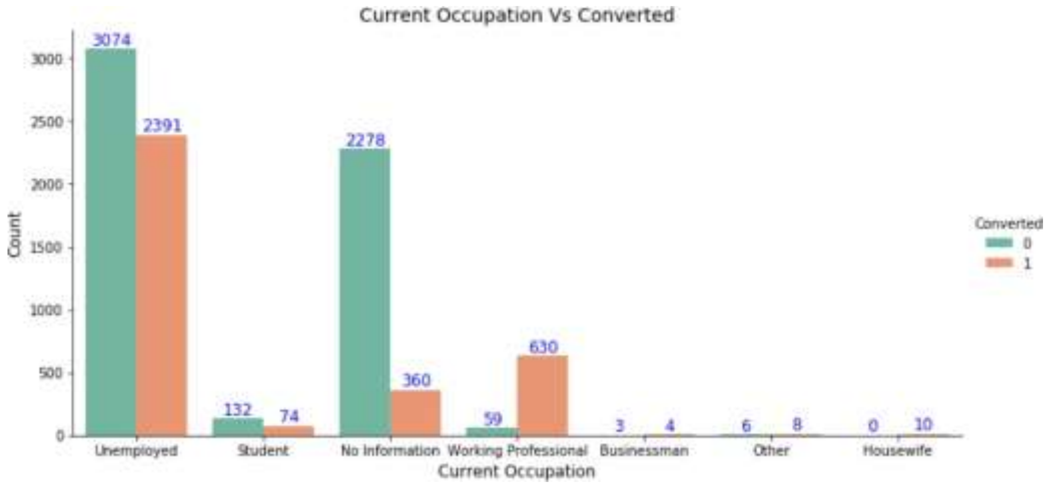




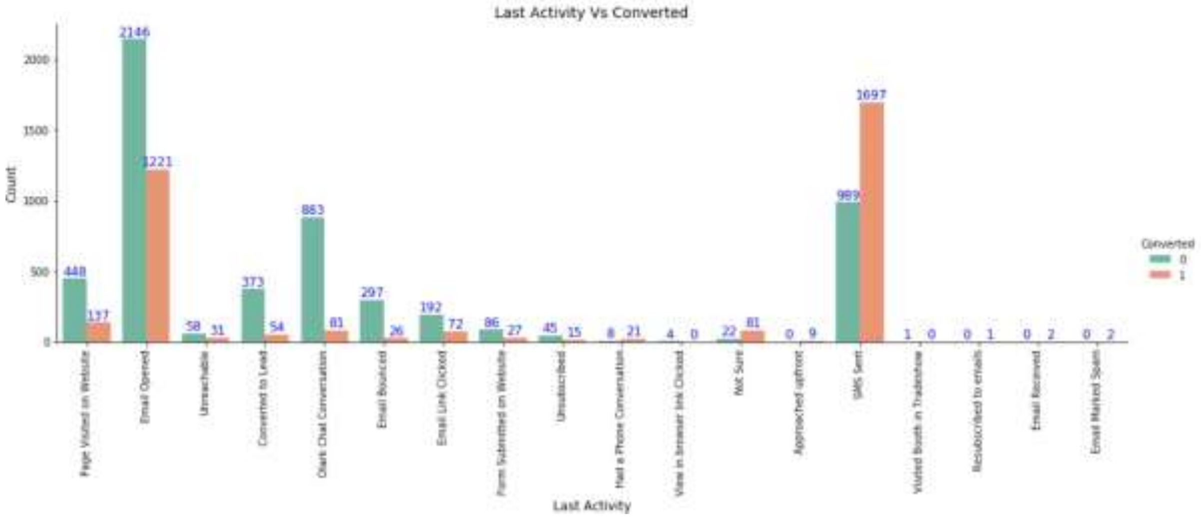


Identifying conversion rates of specific columns

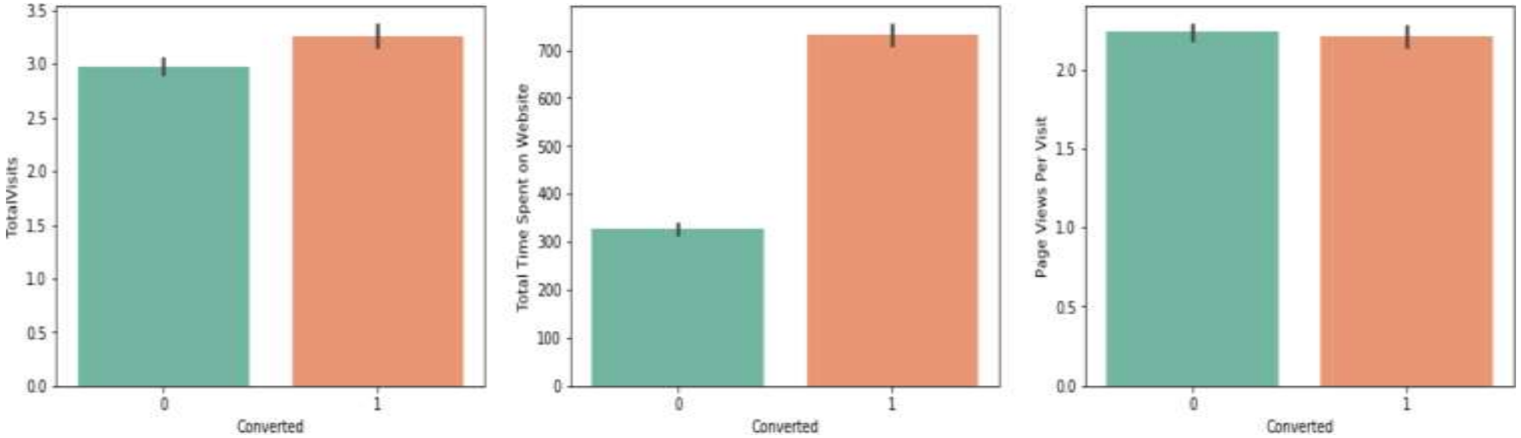
- Occupation:
  - As per the results, maximum conversion of leads belonged to unemployed category.
  - There were 90% conversion seen in working professional.
  - As well as, a significant 100% conversion rate can be seen in housewives category.



- Last Activity:
  - More SMS sent and mails opened, there were more number of conversion.
  - Page viewed on website follows this conversion rate.



Identifying conversion rates of specific columns : Conclusion



- It can be inferred that the conversion rates were high for ‘Total Visits’, ‘Total Time Spent on Website’ and ‘Page Views Per Visit’.
- Very clearly, the more time spent on the website leads to more conversion rate and can be understood as a critical field.
- Total visits and pages viewed per visit also are important attributes and can be considered for conversion rate.
- Further, there are other attributes outlined that have an impact on the rate of conversion of the lead.

Identifying best models for training

- To train the machine for better prediction, a suitable model has to be chosen. And after checking amongst various attributes, following were chosen to create a stable model for training.
- The attributes chosen are:

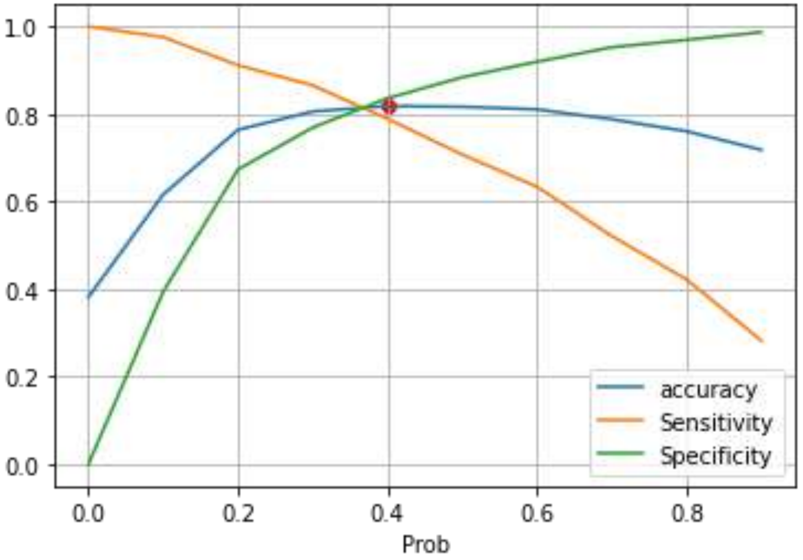
1. Do Not Email
2. TotalVisits
3. Total Time Spent on Website
4. LeadOrigin_Lead Add Form
5. LeadSource_Olark Chat
6. Lead Source_Welingak Website
7. LastActivity_Converted to Lead
8. LastActivity_Email Bounced
9. LastActivity_No Available
10. LastActivity_Olark Chat Conversation
11. CurrentOccupation_No Information
12. CurrentOccupation_Working Professional
13. LastNotableActivity_Email Link Clicked
14. LastNotableActivity_Email Opened
15. LastNotableActivity_Modified
16. LastNotableActivity_Olark Chat Conversation
17. LastNotableActivity_Page Visited on Website

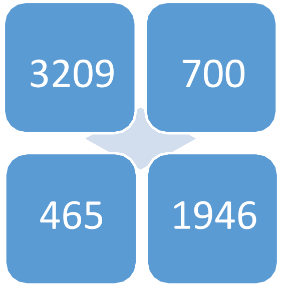
Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.36 based on Accuracy,
Sensitivity and Specificity.

It can be inferred from graph the following scores:

- Accuracy – 81.5%
- Sensitivity – 80.7%
- Specificity - 82 %
- False Positive Rate – 17.9 %
- Positive Predictive Value – 73.5 %
- Negative Predictive Value – 87%





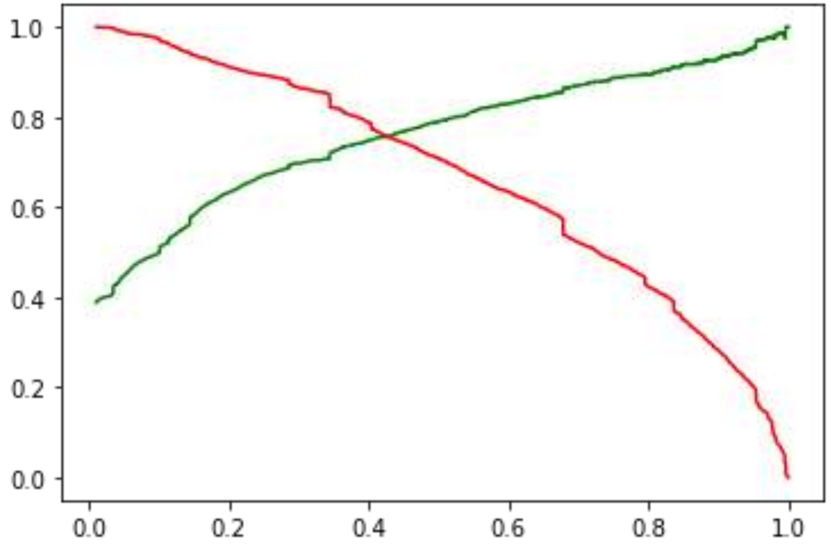
3209	700
465	1946

Confusion Matrix

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.36 based on Precision and
Recall

- Precision – 78.9 %
- Recall – 70.7 %



Confusion Matrix

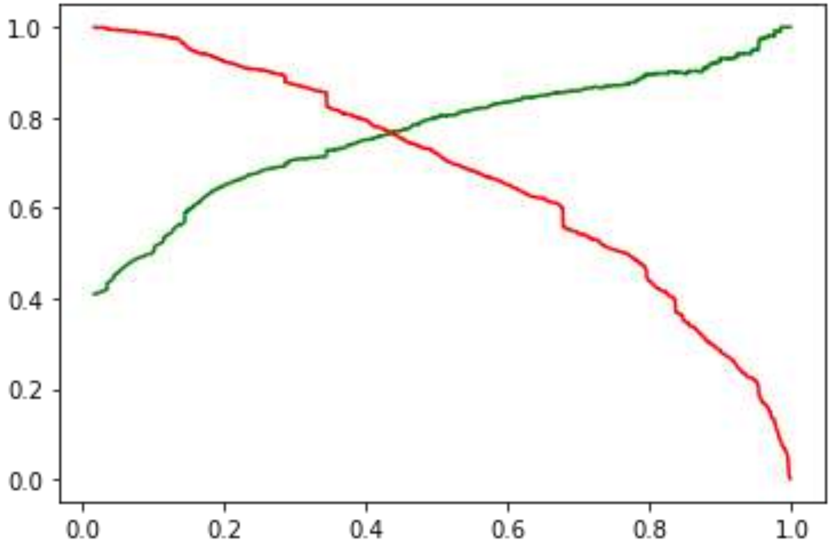
3455	454
706	1705

Model Evaluation – Sensitivity and Specificity on Test Dataset

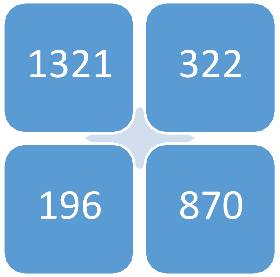
After training the model, the data was then used to predict and the overall metrics resulted are;

- Accuracy – 80.8%
- Sensitivity – 81.6%
- Specificity – 80.4 %
- Precision – 72.9%
- Recall – 81.6%

The graph shows the trade off between precision and recall.



Confusion Matrix



1321	322
196	870

Result

- 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. –
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 81% and 80% which are approximately closer to the respective values

  calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 79% (in train set) and 73% in test set
- Hence overall this model seems to be good.

Summary

- ✓ After understanding and learning from the dataset, the following conclusions can be made:
    - ✓ Conversion rates are higher in unemployed category, that means people are looking to learn new courses to land jobs and that can be considered by the company.
    - ✓ Another interesting point that came up was, every house-wife who showed interest in course was converted.
    - ✓ People who choose call and mail option were actively converted into positive leads.
    - ✓ A large population who spent more time browsing the website and searching for courses were converted in higher rate.
    - ✓ Lastly, the more phone calls and SMS/Email was sent as last activity, conversion rate of leads were more.

- ✓ These points can be considered by the company X and can increase the lead conversion rate drastically, as proven by the prediction model also.