X EDUCATION

KrishnaPrabha Appuni
Swati Malladi

Problem Statement

Company wishes to identify 'Hot Leads'. This identification aims to increase the lead conversion rate. Hot leads are the most promising leads. They are most likely to convert into paying customers. Assign Lead Score to each lead. Score enables Sales Team to identify Hot Leads.

Target lead conversion rate is 80%.

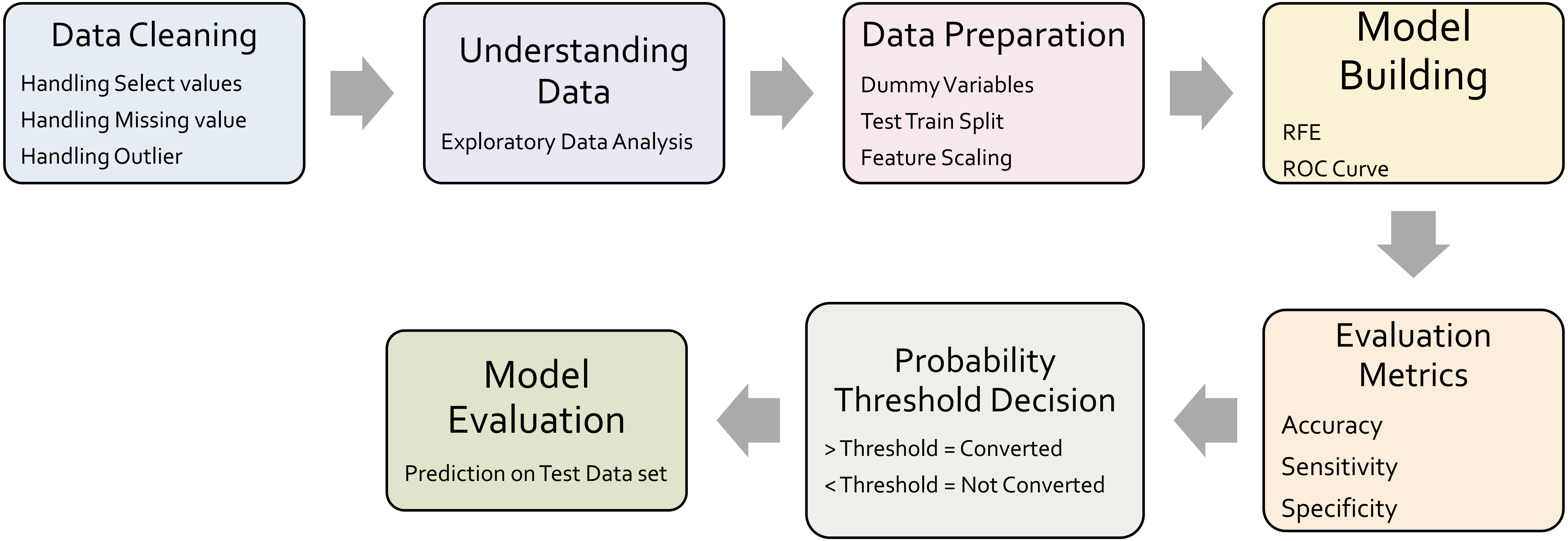
Objectives:

1) Identify Hot Leads
2) Assign Lead Score

Approach:

1) Build a Logistic Regression Model to predict conversion probabilities
2) Decide a threshold for Converted
3) Generate Lead Score from probabilities

Analysis Approach



Data Cleaning

- Handling Select values
- Handling Missing value
- Handling Outlier

Understanding Data

- Exploratory Data Analysis

Data Preparation

- Dummy Variables
- Test Train Split
- Feature Scaling

Model Building

- RFE
- ROC Curve

Evaluation Metrics

- Accuracy
- Sensitivity
- Specificity

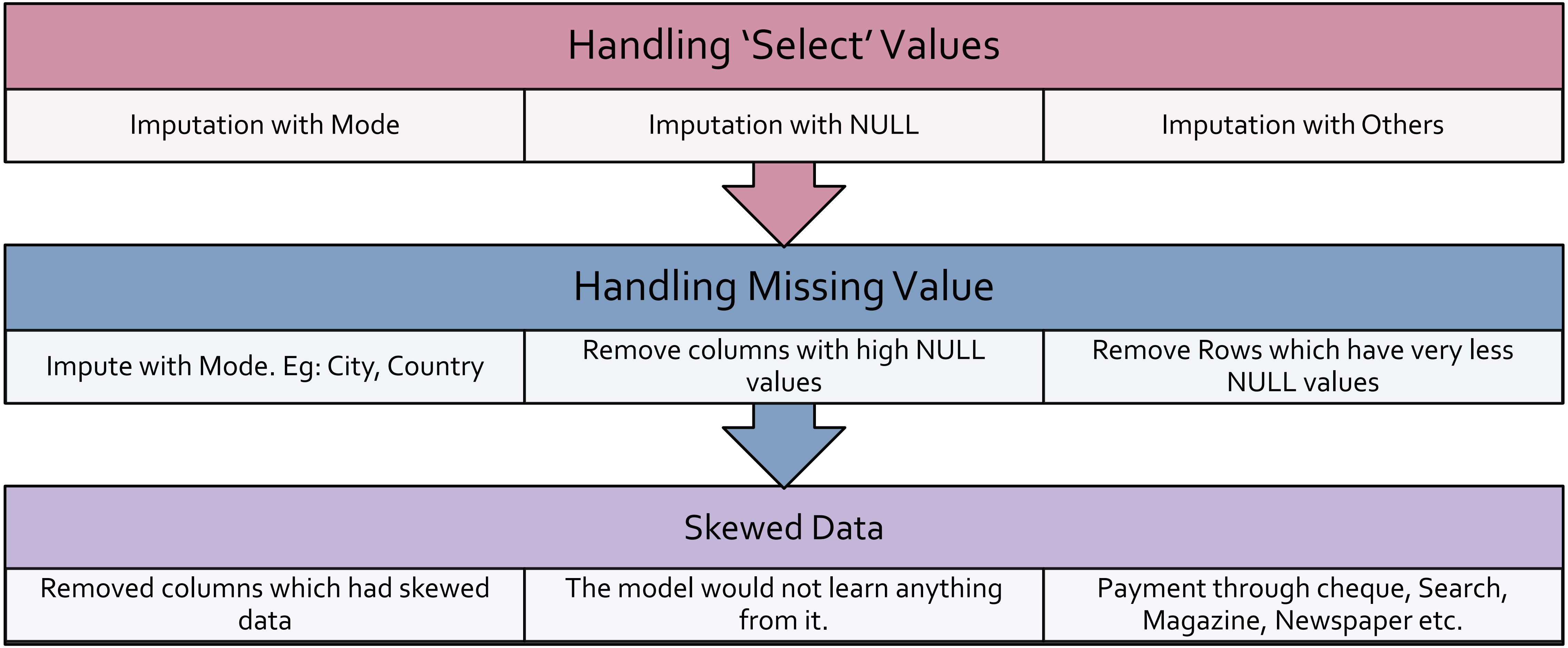
Probability Threshold Decision

- > Threshold = Converted
- < Threshold = Not Converted

Model Evaluation

- Prediction on Test Data set

Data Cleaning Process



Handling 'Select' Values

Imputation with Mode

Imputation with NULL

Imputation with Others

Handling Missing Value

Impute with Mode. Eg: City, Country

Remove columns with high NULL values

Remove Rows which have very less NULL values

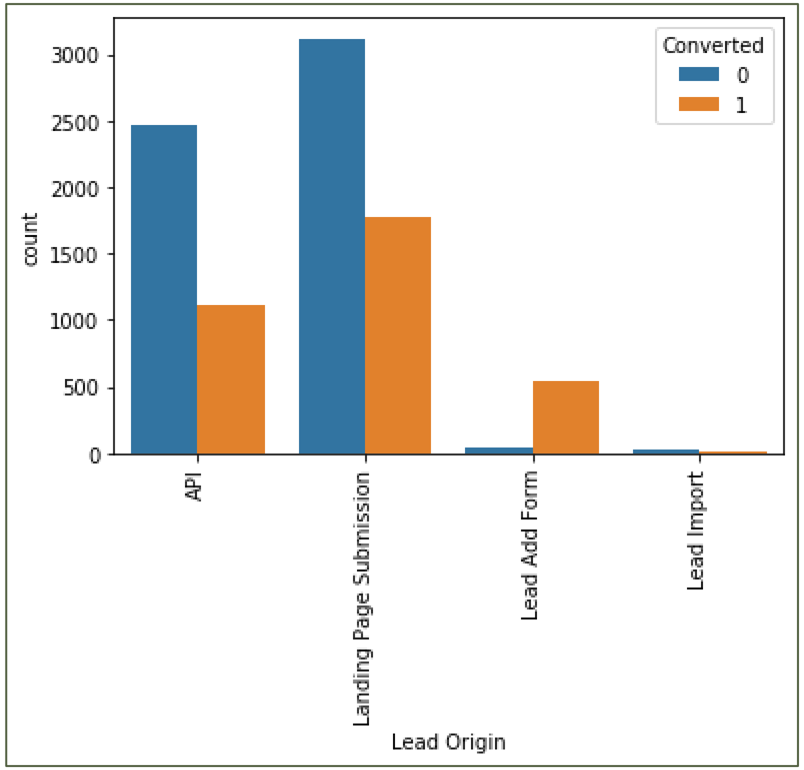
Skewed Data

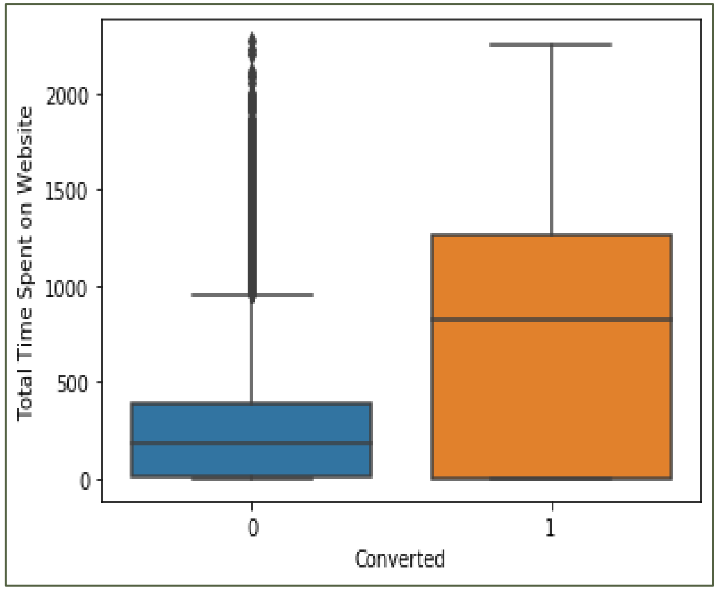
Removed columns which had skewed data

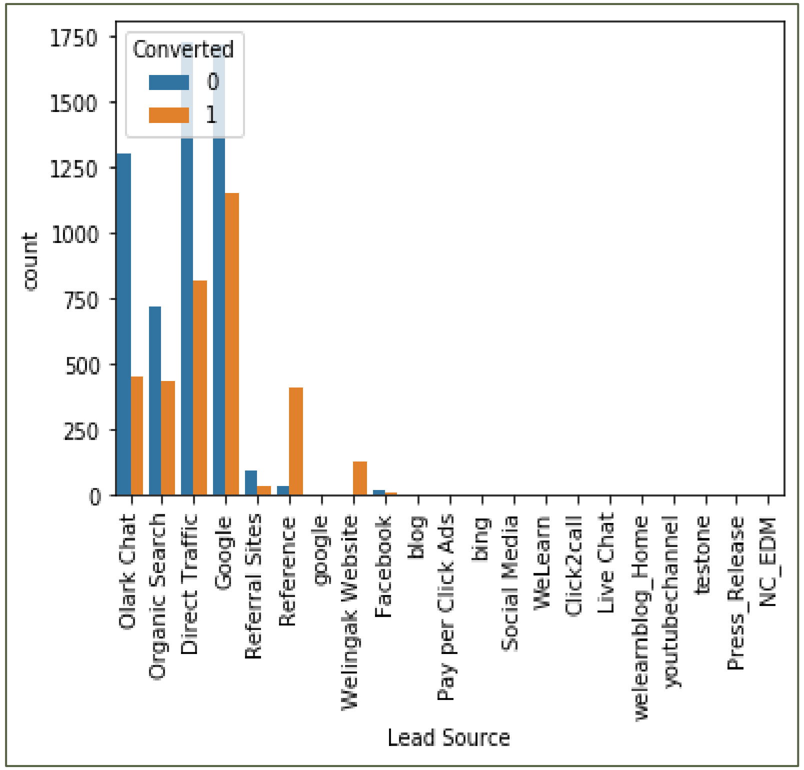
The model would not learn anything from it.

Payment through cheque, Search, Magazine, Newspaper etc.

Exploratory Analysis of Variables







Lead Origin

Lead conversion should be improved for API and Landing Page Submission. More number of leads should be generated from Lead Add Form.

Time Spent on Website

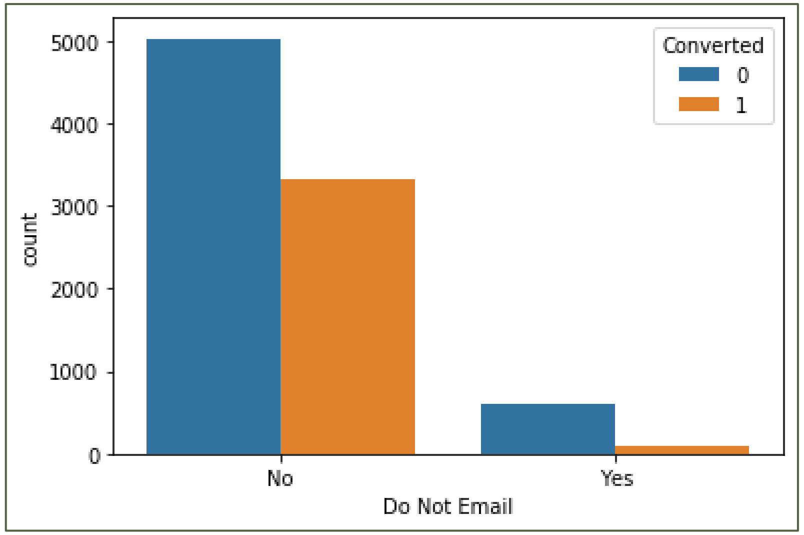
Interested people spend more time on website. Hence, website should be more user friendly and easy to navigate.

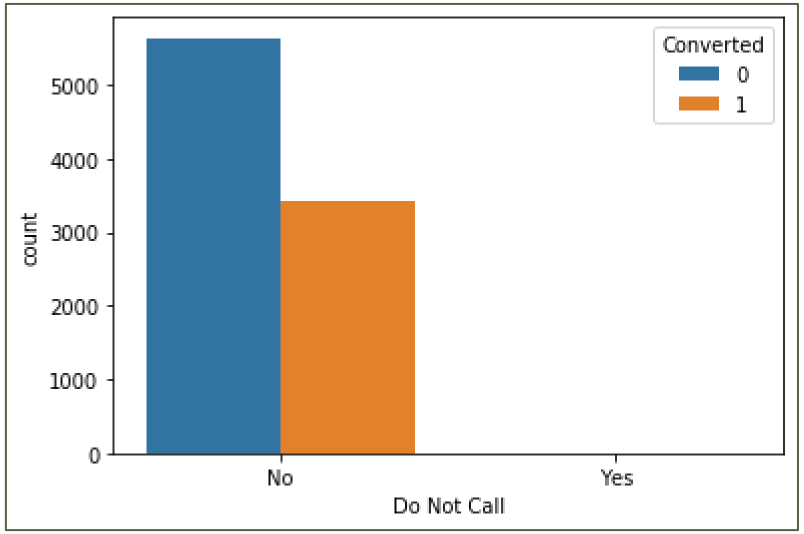
Lead Source

Improve lead generation of Reference and Welingak Website.

Improve lead conversion for Direct traffic, google, olark chat, organic search.

Exploratory Analysis of Variables





Total Visits

Interested people visit often. Median is same for Converted and Non converted.

Page Views per Visit

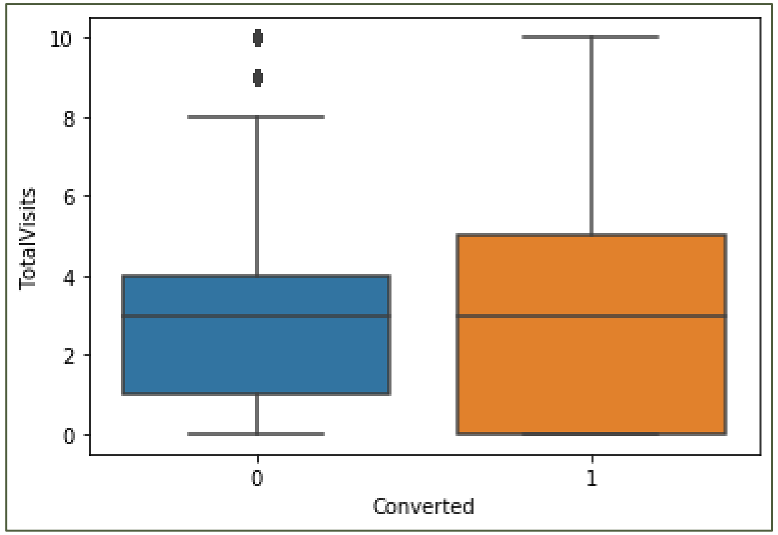
Interested people view more pages. Median is same for converted and Non converted.

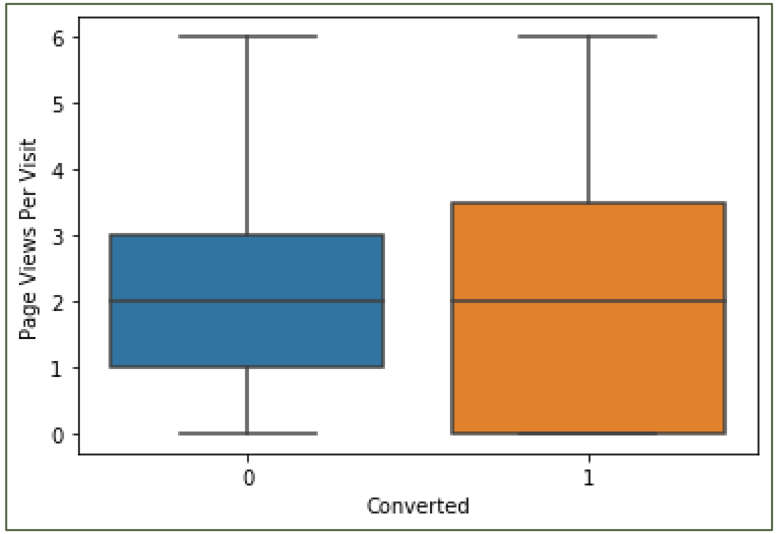
Do Not Email

Majority of leads do not want to subscribe for emails. However, this is not an indicator for Conversion.

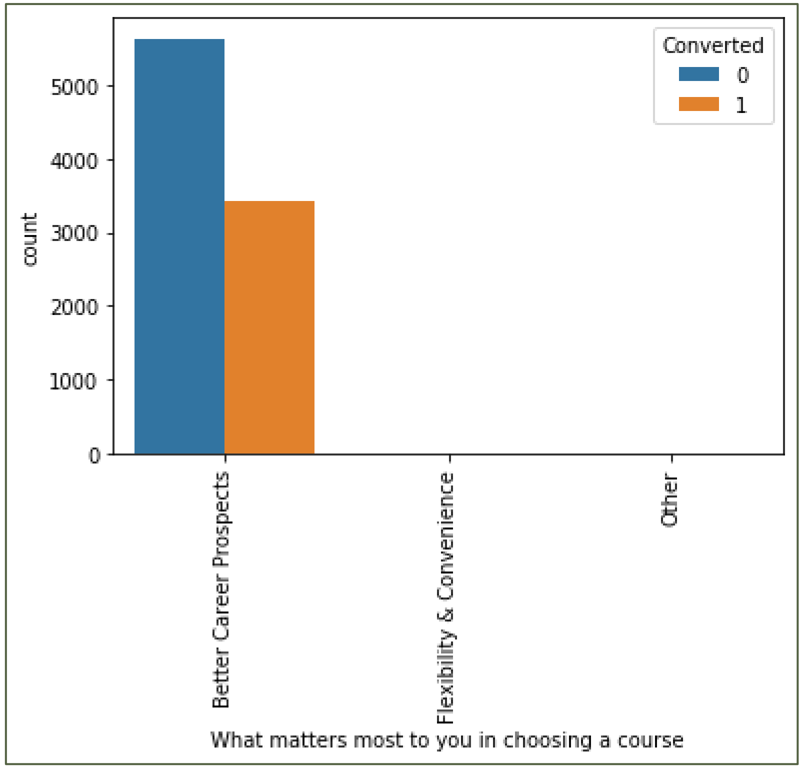
Do Not Call

All leads have said no to calls. Hence, it does not help in identifying Conversion pattern.

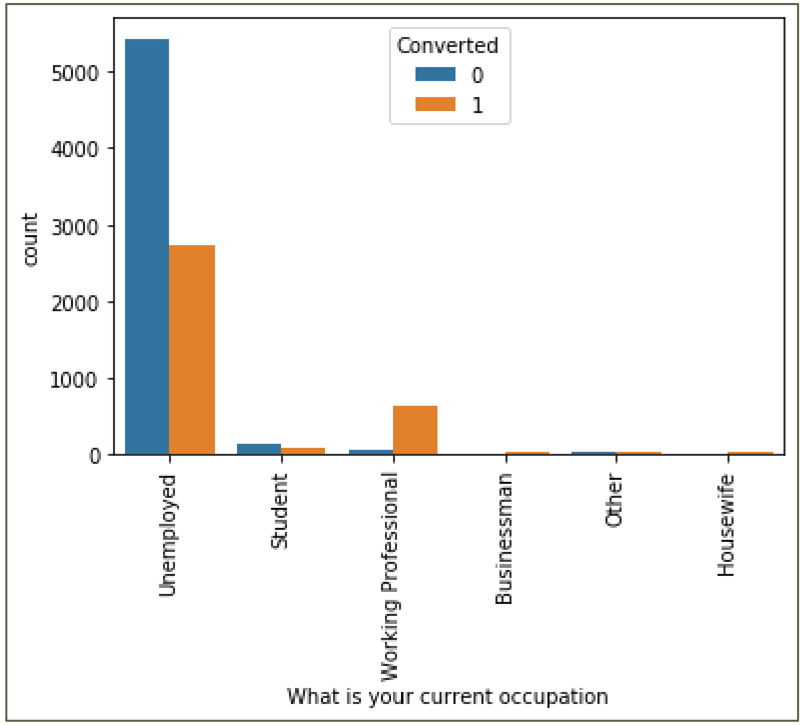




Exploratory Analysis of Variables



More lead generation for Working Professionals. High conversion for Unemployed.

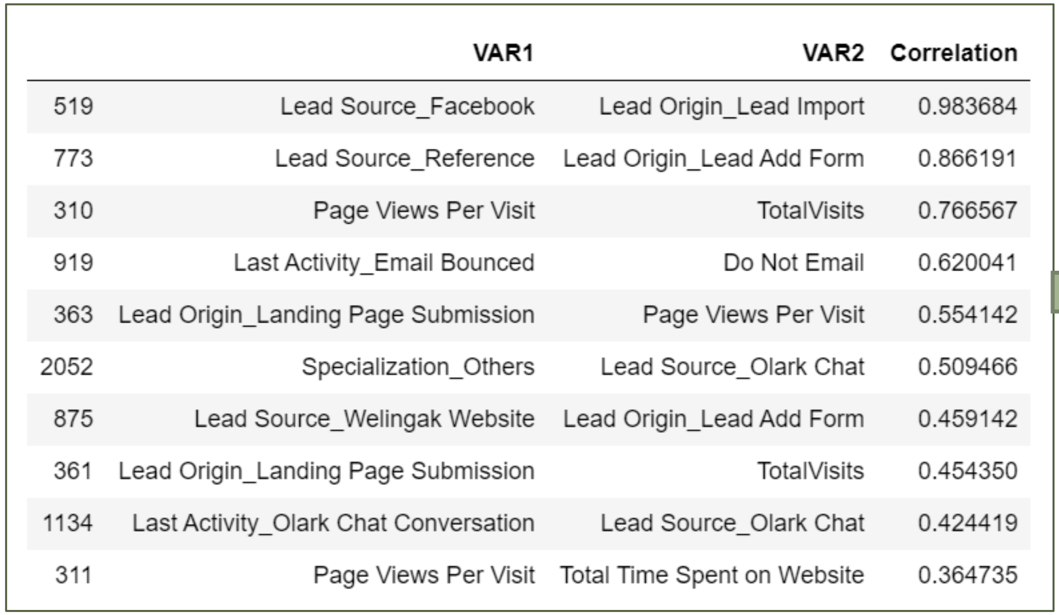


Leads are looking out for Better Career Prospects. The aim of product design team should also include Placements.

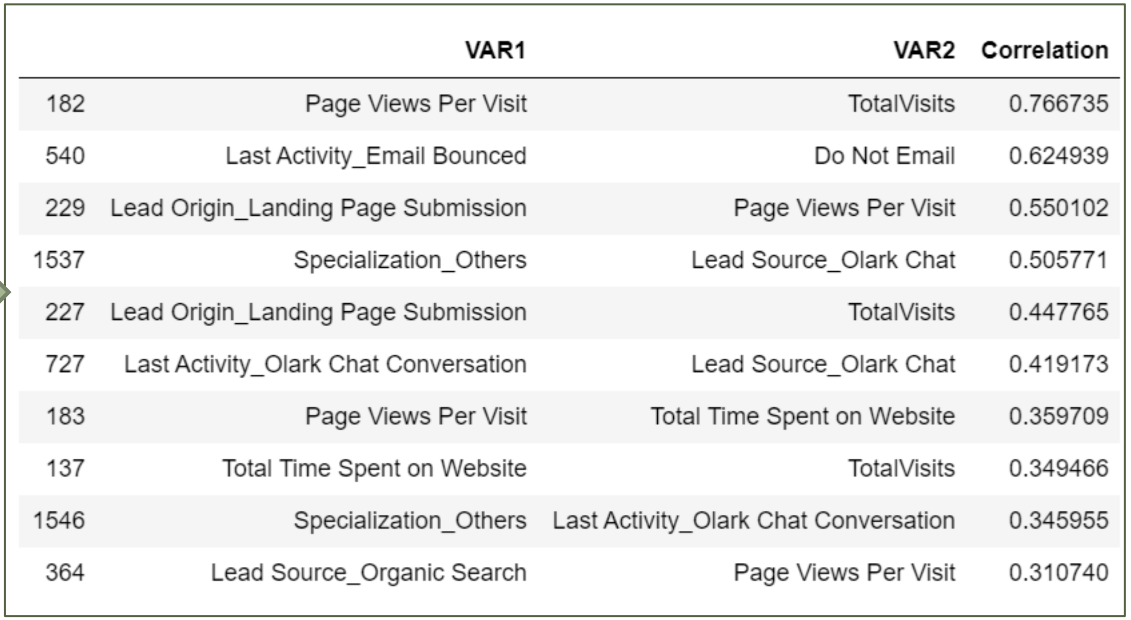
Unemployed people consist of majority leads. However, the conversion rate is less. On the contrary, not many Working professionals are leads but they have high Conversion rate.

Results of Model

Correlations



	VAR1	VAR2	Correlation
519	Lead Source_Facebook	Lead Origin_Lead Import	0.983684
773	Lead Source_Reference	Lead Origin_Lead Add Form	0.866191
310	Page Views Per Visit	TotalVisits	0.766567
919	Last Activity_Email Bounced	Do Not Email	0.620041
363	Lead Origin_Landing Page Submission	Page Views Per Visit	0.554142
2052	Specialization_Others	Lead Source_Olark Chat	0.509466
875	Lead Source_Welingak Website	Lead Origin_Lead Add Form	0.459142
361	Lead Origin_Landing Page Submission	TotalVisits	0.454350
1134	Last Activity_Olark Chat Conversation	Lead Source_Olark Chat	0.424419
311	Page Views Per Visit	Total Time Spent on Website	0.364735



	VAR1	VAR2	Correlation
182	Page Views Per Visit	TotalVisits	0.766735
540	Last Activity_Email Bounced	Do Not Email	0.624939
229	Lead Origin_Landing Page Submission	Page Views Per Visit	0.550102
1537	Specialization_Others	Lead Source_Olark Chat	0.505771
227	Lead Origin_Landing Page Submission	TotalVisits	0.447765
727	Last Activity_Olark Chat Conversation	Lead Source_Olark Chat	0.419173
183	Page Views Per Visit	Total Time Spent on Website	0.359709
137	Total Time Spent on Website	TotalVisits	0.349466
1546	Specialization_Others	Last Activity_Olark Chat Conversation	0.345955
364	Lead Source_Organic Search	Page Views Per Visit	0.310740

Before building the model, highly correlated dummy variables were removed. Hence, Lead Source_Facebook, Lead Origin_Lead Import, Lead Source_Reference, Lead Origin_Lead Add Form were removed before building model.

Results of Model

Coefficients

Features with Positive Coefficient
Total Time spent on website
Lead Source_Welingak Website
Last Activity_Email Link Checked
Last Activity_Email Opened
Last Activity_Other Activities
Last Activity_Page Visited on Website
Last Activity_SMS Sent
Last Activity_Unreachable
Last Activity_Unsubscribed
What is your current occupation_Working Professional

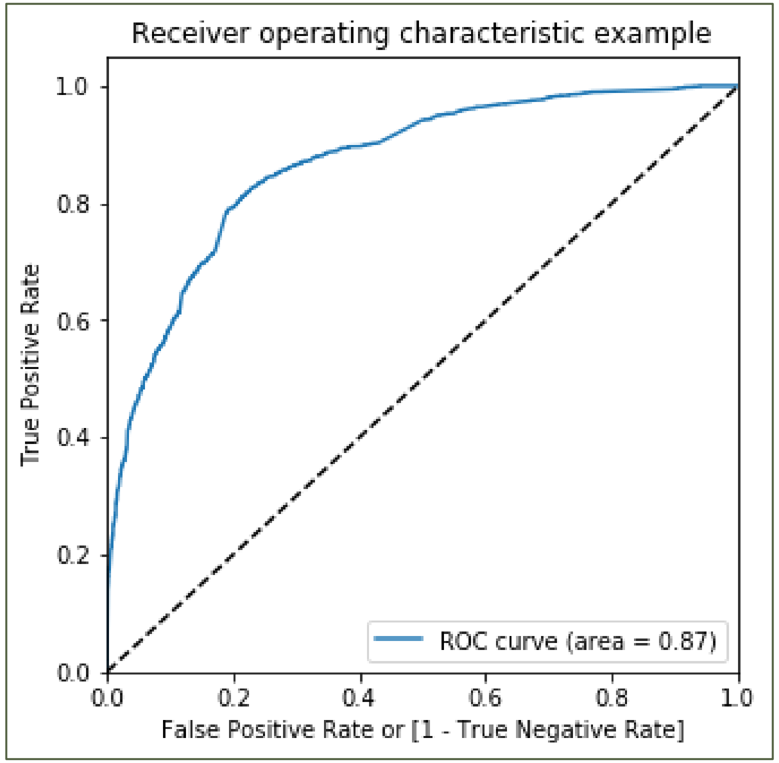
Features with Negative Coefficient
Do Not Email
Lead Origin_Landing Page Submission
Specialization_Others

Features with Positive Coefficient affect Lead conversion positively. If these Features will increase, then lead conversion will increase.

Features with Negative Coefficient affect Lead Conversion negatively. If these Features will decrease, then lead conversion will increase.

Results of Model

ROC Curve



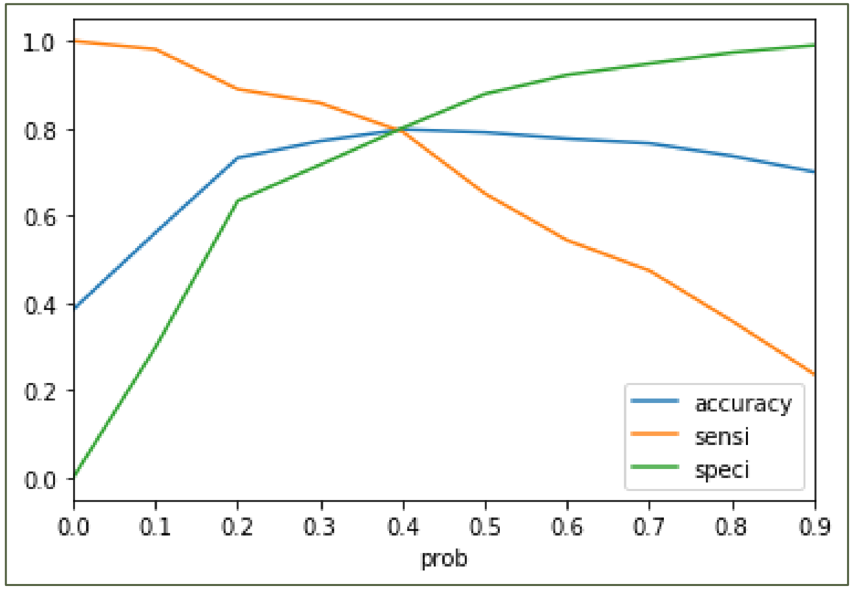
ROC Curve is a measurement of performance at various thresholds.
It is a probability curve.
It is plotted between TPR and FPR.

The ROC Curve is near 0.8. It is a good indication.
Higher the are under curve, better the model.
Area under ROC Curve is 0.87. It is indication of a good model.

It is a trade-off between Sensitivity and Specificity.

Results of Model

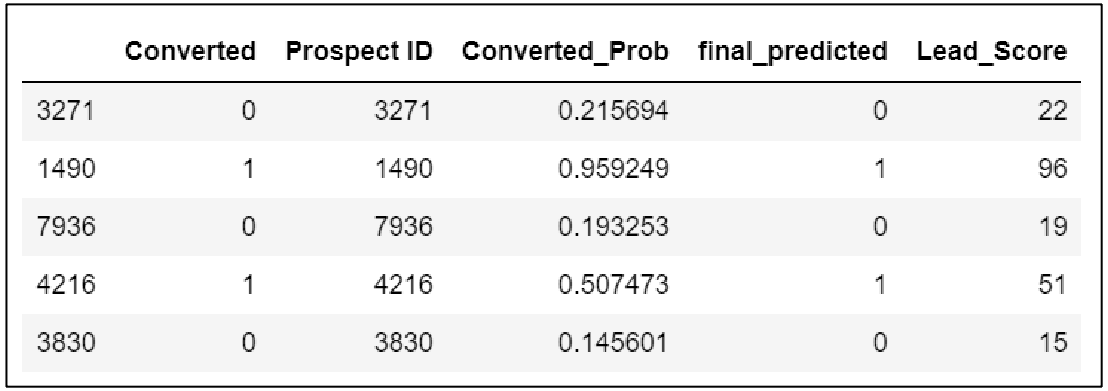
Optimal Probability Threshold



Here, accuracy, sensitivity and specificity were calculated for a set of data points.
These points were plotted onto the graph.
The intersection point of the three metrics should be at or above 0.8. Here we have intersection at 0.8.
The threshold value is determined at 0.4.

Results of Model

Lead Score



	Converted	Prospect ID	Converted_Prob	final_predicted	Lead_Score
3271	0	3271	0.215694	0	22
1490	1	1490	0.959249	1	96
7936	0	7936	0.193253	0	19
4216	1	4216	0.507473	1	51
3830	0	3830	0.145601	0	15

Lead score was generated on Test set for leads. This is generated by multiplying probability with 100. It depicts the conversion strength of a lead. The sales team can use this score for preparing a strategy of calls. Higher Lead Score will be called at first priority. These are the Hot Leads.

Evaluation Metrics

Confusion Matrix

Predicted/Actual	Not Converted	Converted
Not Converted	3127	778
Converted	507	1939

Decision for Threshold Value

0.4

Evaluation Metrics

	Formula	Train set	Test set
Accuracy	(TP+TN)/(TP+TN+FN+FP)	0.79	0.79
Sensitivity	TP/(TP+FN)	0.77	0.77
Specificity	TN/(TN+FP)	0.80	0.80
False Positive Rate	FP/(TN+FP)	0.19	0.19
Positive Predicted Value	TP/(TP+FP)	0.71	0.68

Accuracy is most intuitive. It answers, How many lead conversions did we correctly predict?

Sensitivity and specificity are inversely proportional to each other.
When specificity increases, sensitivity decreases.

The threshold value determines sensitivity and specificity. We have chosen a trade-off threshold value. Hence, sensitivity and specificity are acceptable.

Evaluation Metrics

	Formula	Train	Test
Negative Predicted Value	TN/(TN+FN)	0.86	0.86
Precision	TP/(TP+FP)	0.77	0.68
Recall	TP/(TP+FN)	0.65	0.77
F1 Score	2*(P*R)/(P+R)	0.70	0.72
Area Under Curve		0.87	

Precision is correctly positively labelled. It answers, how many those labelled as Hot Leads have actually Converted?

Recall is correctly positively labelled to who have converted in reality. It answers, of the people who converted, how many are correctly predicted by model?

F1 Score includes both precision and recall.
A good F1 Score means balanced Precision and Recall.
Such is the case here.

Accuracy would have been a good metric, if the FP and FN were close. Hence, for uneven distribution, F1 Score is best. We have F1 Score of 0.72

Here in this case, occurrence of False Negatives is acceptable. Hence, considering Recall. Recall value is 0.77 for Test dataset.

False positives are tolerable hence, specificity not taken into consideration.

Summary

- Lead Source_Welingak Website, Last Activity_Other Activities and Last Activity_SMS Sent contribute more towards the probability of leads getting converted.
- Positive coefficient features should be increased, negative coefficient features should be reduced.
- The p-values of all variables are less than 0.05. VIF values are less than 0.5.
- Threshold value is 0.4
- Accuracy of model is 0.79
- The lead Score of each customer is generated by multiplying probability with 100. This score can be used by sales team.
- Area under ROC Curve is 0.87. Hence, indication of a good model.
- The Optimal Probability Threshold helped in identifying correct value of threshold as 0.4.
- Lead Origin is an important feature. API, Landing Page Submission, Add Form all are balanced in data.

Summary

- Some methods in Lead source are promising. While some like Direct traffic, google search, organic search can be improved only by improving the digital marketing techniques of the company. Better the SEO, higher the lead generation.
- Majority people are either Unemployed or working professionals.
- Moreover, all types of leads are looking for a course with better career opportunities.
- Higher the time spent on the website, more the leads generated and higher the lead conversion.
- Last activity is a good indicator. 7 dummies from Last activity are part of the model built by RFE.
- Skewed features were not used in model building.
- Features with p value>0.5 were dropped after RFE
- Do Not Call was a skewed feature which is removed after RFE. Another model 2 was built to check changes. No changes in metrics seen.

Recommendations

- Need for a more robust questionnaire. This will help reduce the occurrences of Select in data.
- For API and Landing Page Submission, leads are generated but not converted. Expert sales team must handle this. Lead Add Form shows promising results for conversion. Hence, more leads to be generated from it.
- Lead generation from Reference and Welingak should be improved. They have high conversion. Direct traffic, google and organic search generate leads. Conversion can be improved. Investments and focus into digital marketing techniques required.
- Lead generation for working professionals should be improved. Conversion for Unemployed must be increased. This can be done by changing the design of course. Placements should be part of the course. People look into courses for better career opportunities.
- For leads with Higher Lead Score, effectively also called Hot Leads, X Education must have a different strategy for calling them. A thorough background check must be done for them. In order to cater them effectively. Also, best sales persons must be engaged with Hot Leads.

Recommendations

- Website should be made user-friendly and easy to navigate. The more the person spends time on our website, higher the chances of conversion.
- Websites similar to Welingak must be targeted. This website gives promising results. Hence, other platforms must be explored in this direction for advertising.
- Last activity plays a major role in deciding conversion of leads. Since, this variable was split into dummies and also we have 7 dummies in model, we can safely say that last activity should be monitored for successful lead conversion.

- Precision means if the model is Precise. 68% times the leads can be converted.
- Recall and Sensitivity means that if the lead has converted, the model can identify it 77% of times.
- ROC Curve are under it is 0.87. That is a good indication.
- False positive rate can be high. Because it is a hit and trial method. Some leads classified as Hot will not harm much.