Lead Scoring

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.

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- 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%.
- The company requires to build a model wherein it needs to assign a lead score 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%.
- **Objective:** 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.

Data Preparation

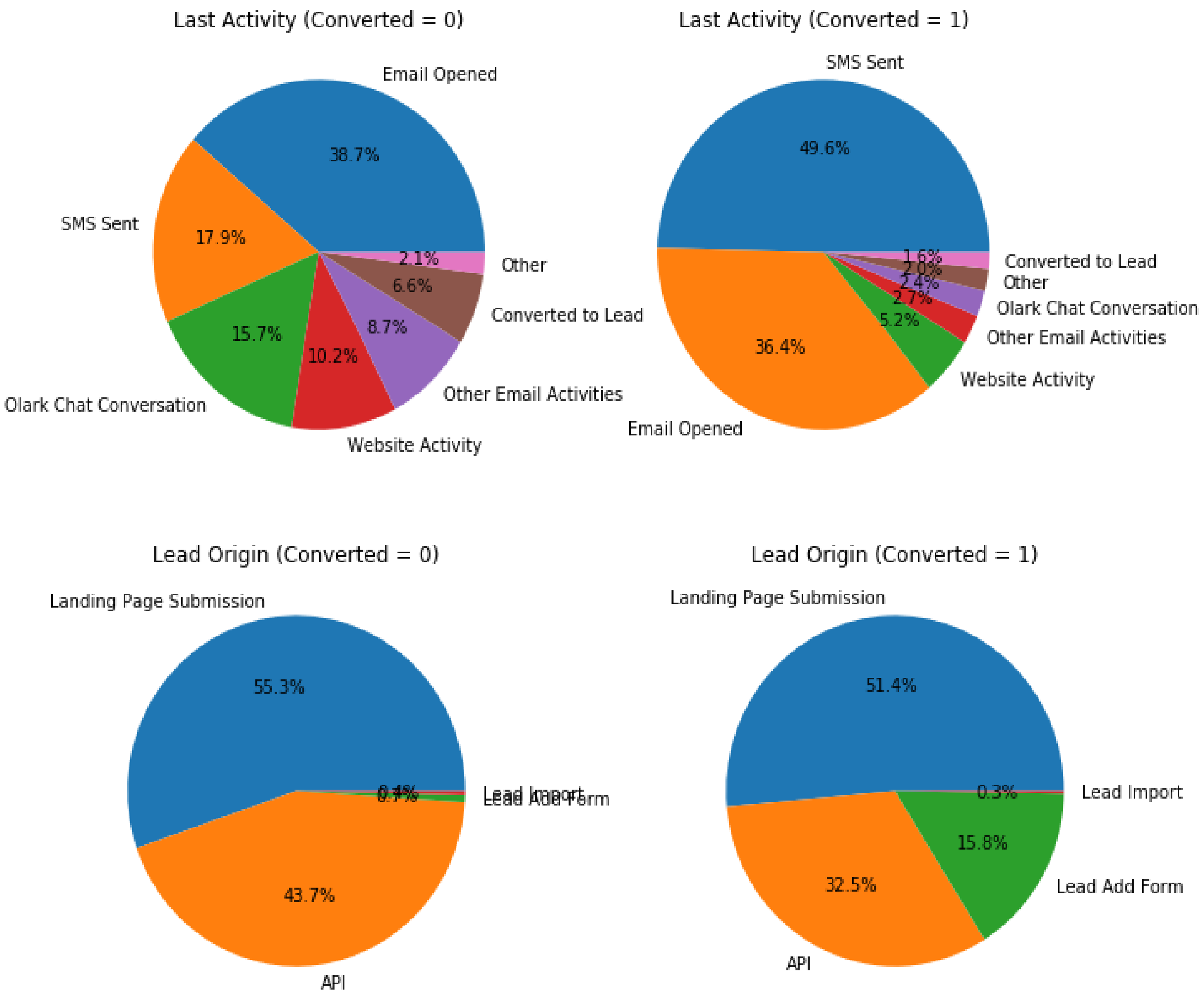
- All the scoring variables such as *Tags, Lead Quality, Lead Profile, Asymmetrique scores* are dropped as these are generated after handling the leads.
- Most of the features are found to be having data skewed towards one level and some of the features has more than 40% missing values (Select is also a missing value in some of the features) and all these features are dropped.
- *Page Views Per Visit, Total Visits & Total Time Spent on Website* are only the continuous columns after dropping unnecessary columns. The first two had missing values which are dropped as they are very less in numbers & outliers are treated with mode value.
- Below are the final set of features that will be used for modelling. All the categorical variables are converted to binary variables by creating respective dummy variables.

Binary Features	**Continuous Features**	**Categorical Features**
Converted	TotalVisits	Lead Origin
Total Time Spent on Website	Page Views Per Visit	Lead Source
A free copy of Mastering The Interview	Lead Number	Last Activity
		Specialization
		What is your current occupation
		City

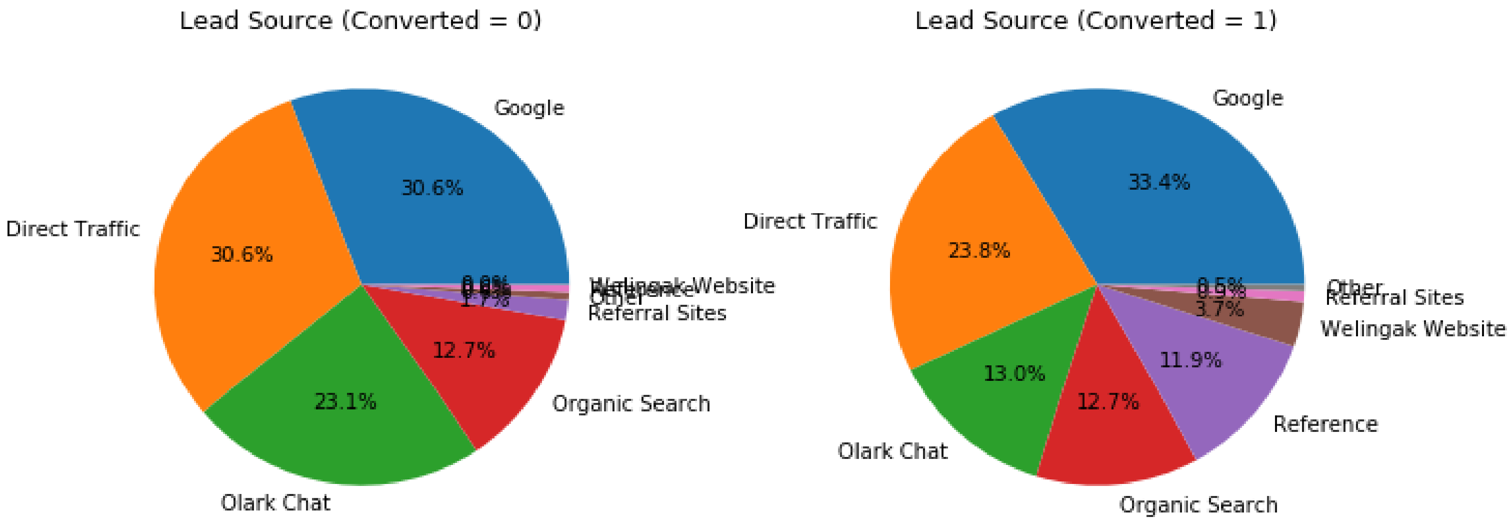
EDA

With these bivariate analysis of categorical variables, the following observations are evident:

- **SMS Sent** activity is potential lead candidate
- **Lead Add Form origin** is as well a potential lead candidate
- **Olark Chat source** is not a reliable lead candidate.



EDA

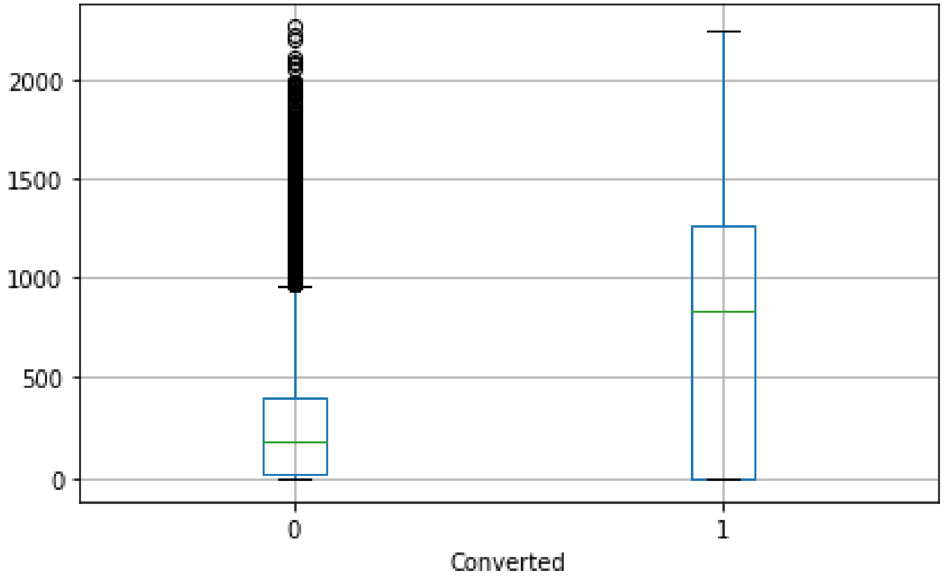


The final set of categorical variables are converted to dummy variables and the highly correlated dummy variables are also dropped.

EDA

When the bivariate analysis performed on the continuous features, it is observed that the Total Time Spent on Website influences the lead conversion. Higher the value, more chances of a lead conversion.

Total Time Spent on Website



The final dataframe combining the continuous features and dummy variables out of categorical features formed with 43 columns, which will be directly used for model building

Model Building

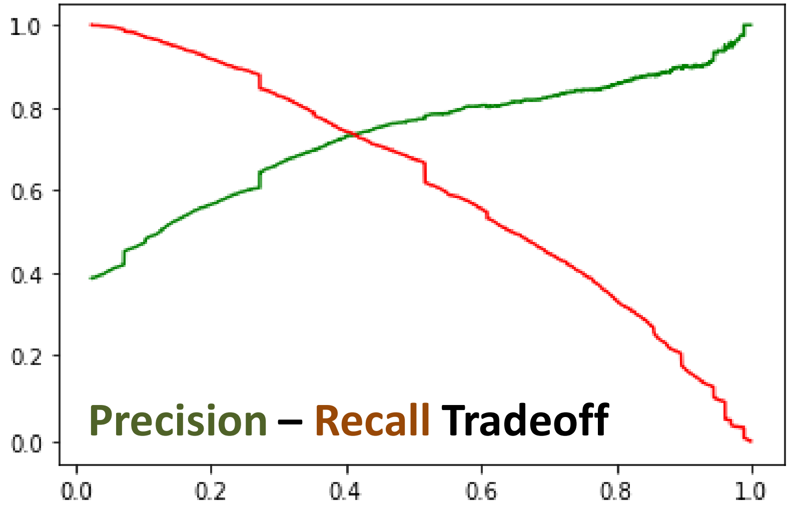
- To begin with, identified 20 variables with the help of RFE (Recursive Feature Elimination) method.

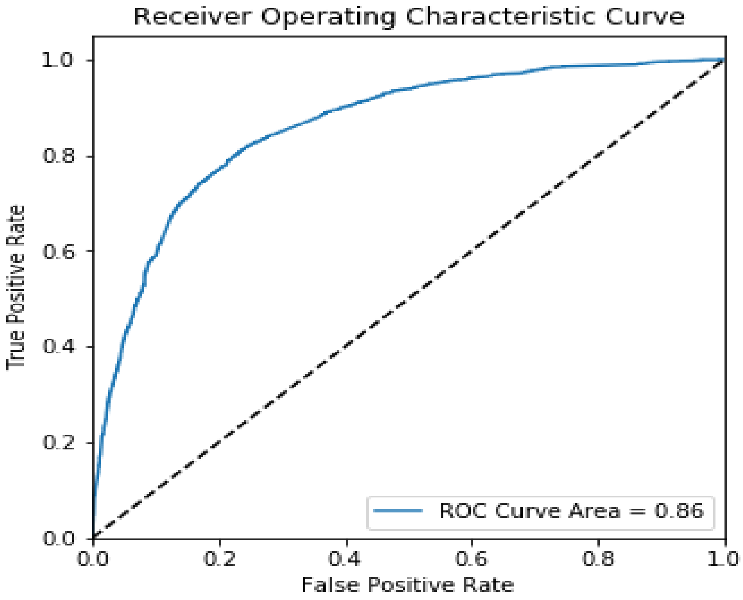
- After about 9 iterations of eliminating variables with high p-value (insignificant variables) and high VIF (highly correlated dependent variables), a final model is achieved with 12 variables and the variables are listed in the order of highest coefficients.

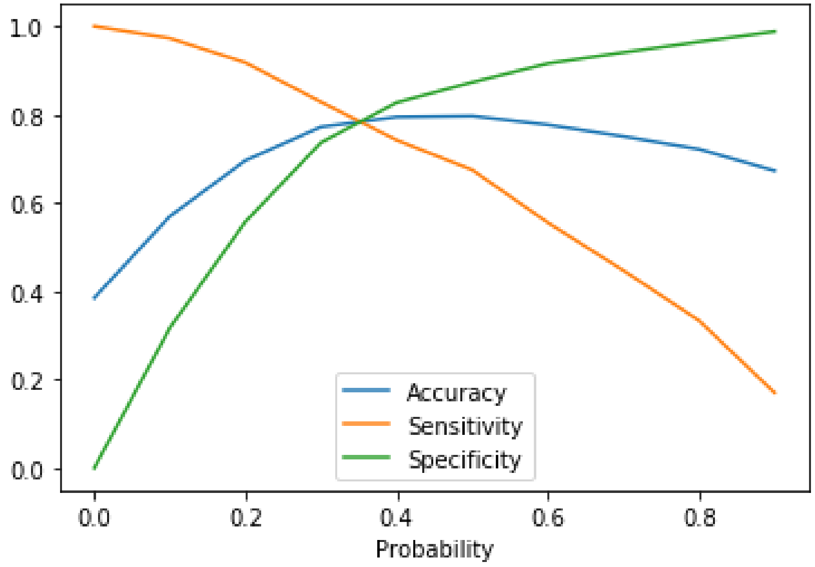
Feature	Coefficient
Lead Origin_Lead Add Form	4.16
Last Activity_SMS Sent	2.22
Lead Source_Welingak Website	1.69
Last Activity_Other	1.62
Lead Origin_Lead Import	1.52
Lead Source_Olark Chat	1.40
Last Activity_Email Opened	1.17
Total Time Spent on Website	1.15
Last Activity_Website Activity	0.54
Specialization_Travel and Tourism	-0.43
Last Activity_Olark Chat Conversation	-0.40
Specialization_Finance Management	-0.38

Identify the Probability Cut-off

- With the ROC area of 0.86, the model seems to be a good one.
- The Precision_Recall_Curve and Sensitivity_Specificity tradeoff curves are plotted to identify the optimal cut-off probability so as to keep all the metrics good.







Identify the Probability Cut-off

- 0.41 is found to be the optimal cut-off probability from Precision-Recall Trade-off.
- 0.35 is found to be the optimal cut-off probability from Sensitivity-Specificity trade-off.
- By the objective of the case study, we need to identify about 80% potential lead candidates. Hence, we should target high Sensitivity and should be around 80%. We will stick to the probability cut-off of 0.35.

Metric	Precision_Recall_0.41	Sens_Spec_0.35	Difference
Accuracy	79.59%	78.62%	0.98%
Sensitivity	73.71%	78.99%	-5.27%
Specificity	83.28%	78.39%	4.89%
Precision	73.41%	69.60%	3.82%
Recall	73.71%	78.99%	-5.27%
Negative Predictive Value	83.49%	85.62%	-2.13%
False Positive Rate	16.72%	21.61%	-4.89%

Predictions & Metrics on Test Data

- The final model is applied on the test data with a probability cut-off of 0.35 and the metrics seems to good, not much of drop compared to train data.
- The probability is also converted as lead score ranging from 0-100 for all the leads of train and test data.

Metric	Train Data	Test Data	Difference
Accuracy	78.62%	78.63%	-0.01%
Sensitivity	78.99%	77.76%	1.23%
Specificity	78.39%	79.12%	-0.74%
Precision	69.60%	67.99%	1.60%
Recall	78.99%	77.76%	1.23%
Negative Predictive Value	85.62%	86.18%	-0.56%
False Positive Rate	21.61%	20.88%	0.74%

Key Inputs to the Business

- The below features have high influence over the model as 9 out of 12 features in the final model are from these three features:
  - Lead Origin
  - Last Activity
  - Lead Source
- And especially the below category of each above feature seems to be very important in identifying the potential leads:
  - Lead Origin: Lead Add Form
  - Last Activity: SMS Sent
  - Lead Source: Welingak Website
- In a vital situation where business needs to target highly potential candidates for lead conversion, it is recommended to target with lead score over 60. Lead scores over 90 is very highly potential candidate for lead conversion.

THANK YOU