Lead Scoring Group Case Study

By -
Harmeet Singh
Himanshu Kumar
Mohammed Aseer Shaik

Problem Statement

X Education sells online courses to industry professionals.

The company gets a lot of leads, its lead conversion rate is very low near about 30 percent.

To make this process more efficient, the company's CEO wishes to identify the most potential leads, also known as 'Hot Leads' in order to increase the lead conversion rate to 80 percent.

If these leads are identified successfully, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Steps

- ➔ Data Cleaning and Manipulation
- ➔ EDA
- ➔ Data Preparation
- ➔ Model Building
- ➔ Model Validation
- ➔ Conclusion

Data Cleaning and Manipulation

Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Dropped the “Prospect ID” and “Lead Number” which are not necessary for the analysis as they only mark the customer as unique so they are as good as index.

Dropped some columns which had high biases towards some particular values such as Country.

After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

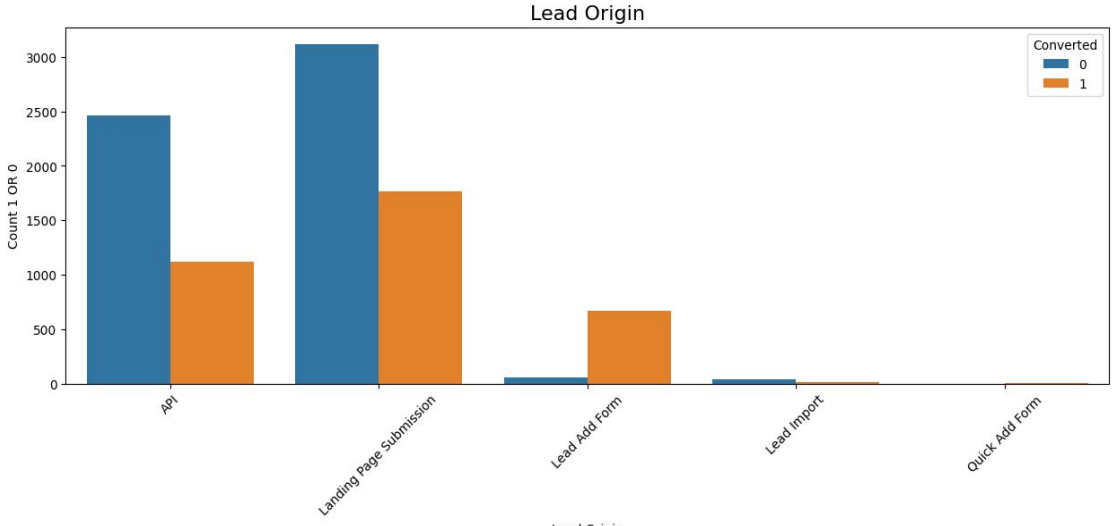
EDA

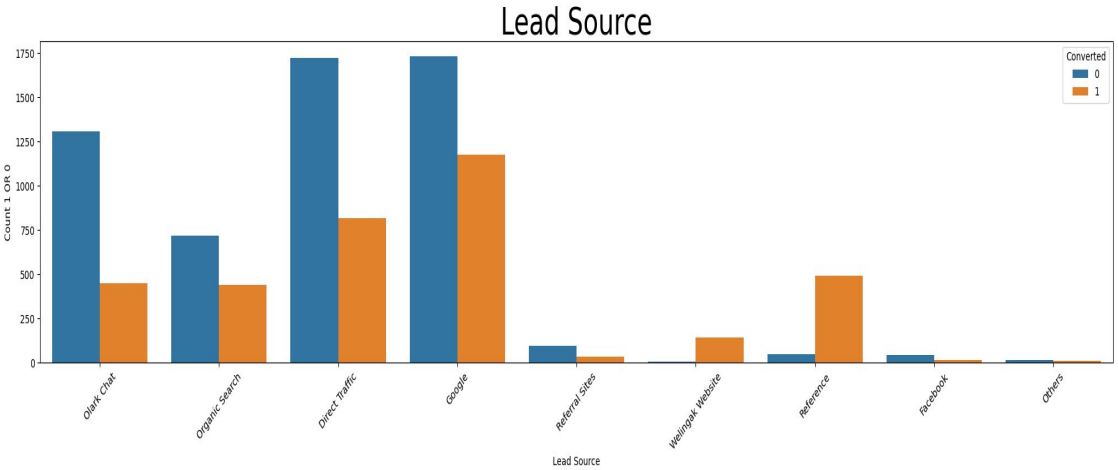
- ➔ Univariate Analysis
- ➔ Bivariate Analysis
- ➔ Outlier Correction

Univariate and Bivariate Analysis

API, Lead Import and Landing page submission have high non converted leads. So, we can check what changes can be done in lead add form in order for it to be more efficient.

Google and Direct traffic generates maximum leads but the conversion rate is low. 'Reference' and 'Welingak Website' leads have high conversion rate.





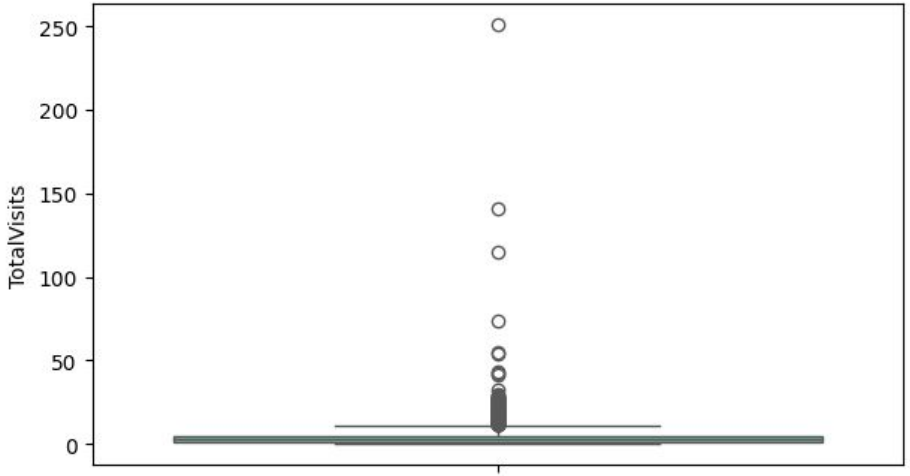
'TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of .72.

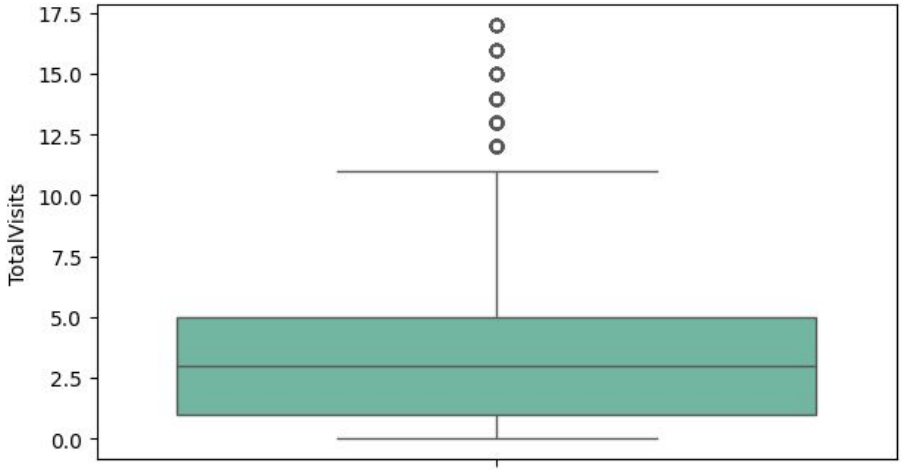
There is also correlation present between Time spent on website and Converted variables.

There is a high negative correlation between 'Page Views per Visit' and 'Converted'.

Outlier Treatment

As it is evident in these two numerical column variable's graphs, outliers were present in them i.e TotalVisits and Page views per visit. Removed those beyond Q3.





Data Preparation

- Normalised Numerical Variables
- Created dummies for Categorical variables
- Dimensions of the data set after the preparation - 56 X 9090

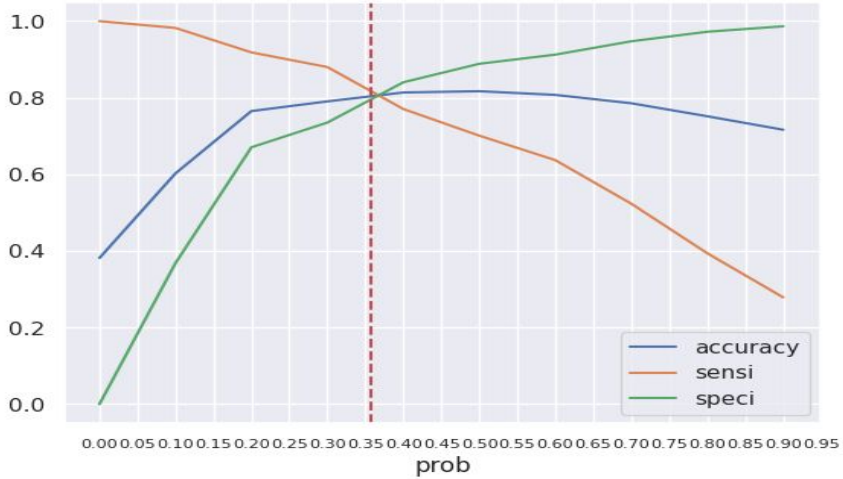
Model Building

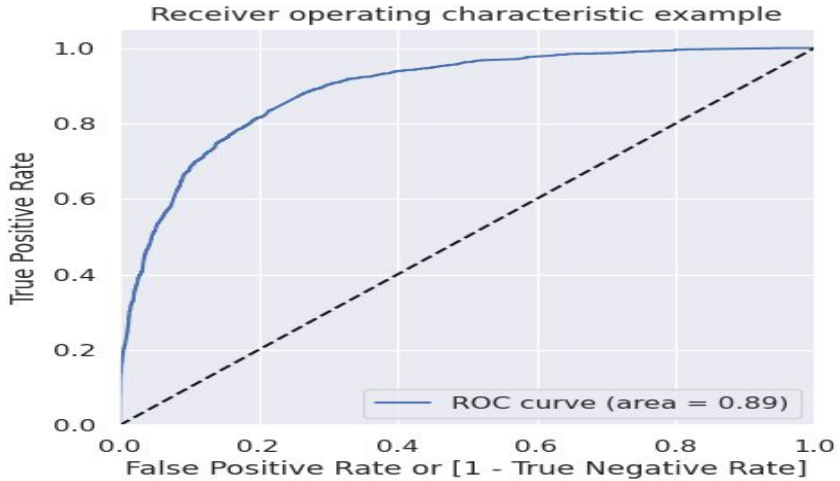
- ➔ Splitting data into training and test sets
- ➔ Using RFE to select top 15 features
- ➔ Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.

ROC Curve and Optimal Cutoff

Found the optimal cutoff point which is 0.35 using the first graph

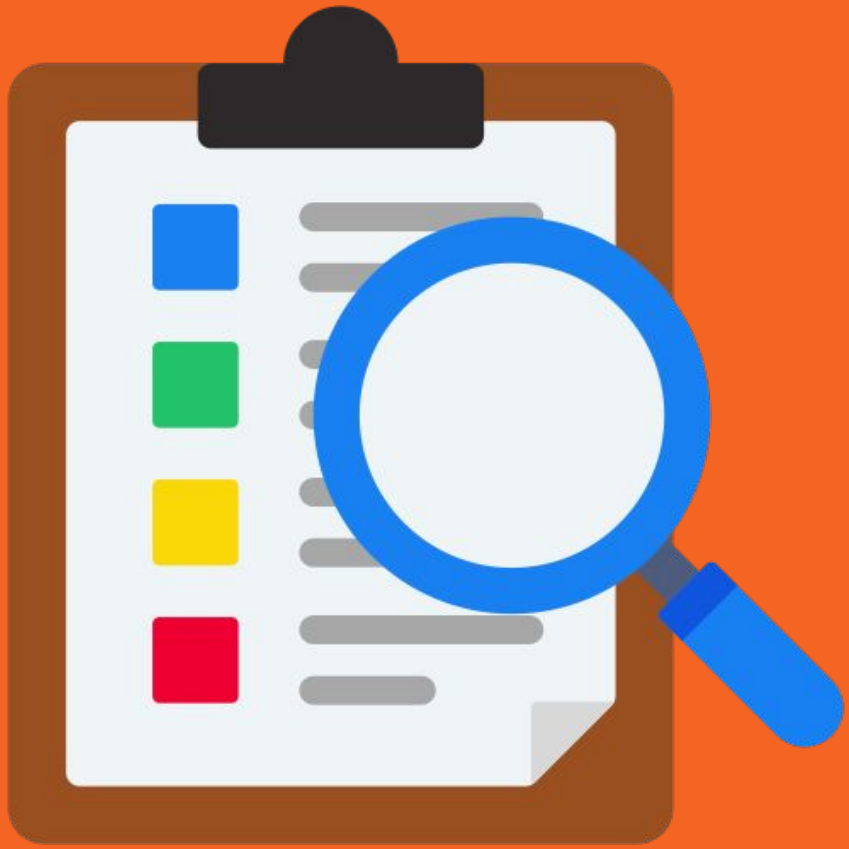
ROC curve's value is 0.89 which is an indicator model's predictive power.





Model Evaluation

- **Calculated measures like specificity, sensitivity, precision and recall for both training and test sets.**
- **For Test Set -** Sensitivity : 80%, Specificity : 80%, Precision: 72%, Recall: 79% approx.
- **For Training Set -** Sensitivity : 80% Specificity : 81% Precision: 72% Recall: 80% approx.



Conclusion

Based on the analysis the potential leads can be generated through -

- ➔ **Total time spent on website.**
- ➔ **When the lead source was - Google, Direct traffic, Organic search, Welingak website.**
- ➔ **When the last activity was - SMS, Olark chat conversation.**
- ➔ **When the lead origin is Lead add format.**
- ➔ **When their current occupation is as a working professional.**