Lead Scoring Case Study

Batch: DSC55

Team members

Girish B R

Lipsa Ray

Reshmi K V

Background

An education company named **X Education** sells **online courses** to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses. 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 Objective

- X education wants to 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
- X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.
- At times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Data Sets

- ‘Leads.csv’ dataset from the past with around 9000 data points.
- The 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.
- The target variable, 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.
- Leads Data Dictionary contains the description about each column.

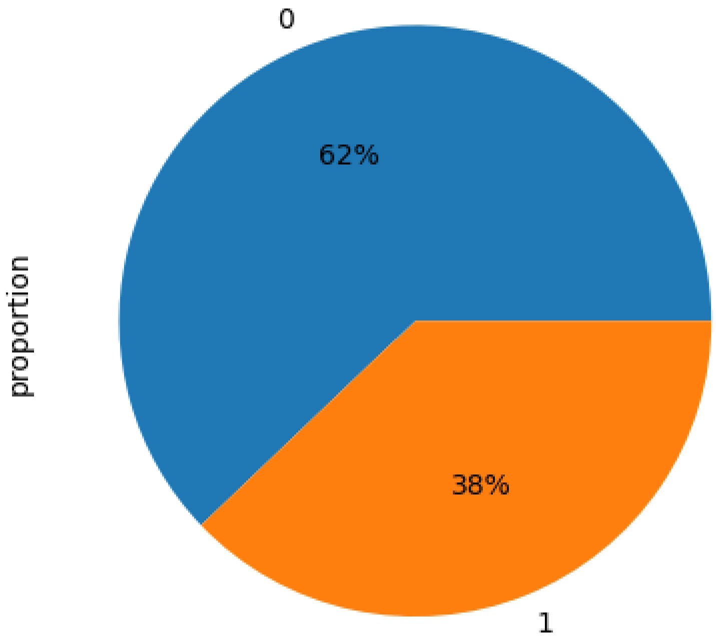
Methodology

- Import data
    - Reading the data
- Data cleaning
    - Impute or drop Null values, Standardise the data
- Exploratory Data Analysis
    - Check data imbalance, Univariate and Bivariate analysis
- Data preparation
    - Create dummy variables, Split train-test set, Feature scaling
- Model building
    - RFE for top 20 features, Manual feature reduction using p value and VIF
- Model Evaluation
    - Find the cut off, KPIs for training data
    - Predicts the model on test data
    - KPIs for test data, Assign lead score and get top features
- Recommendations

Data Insights

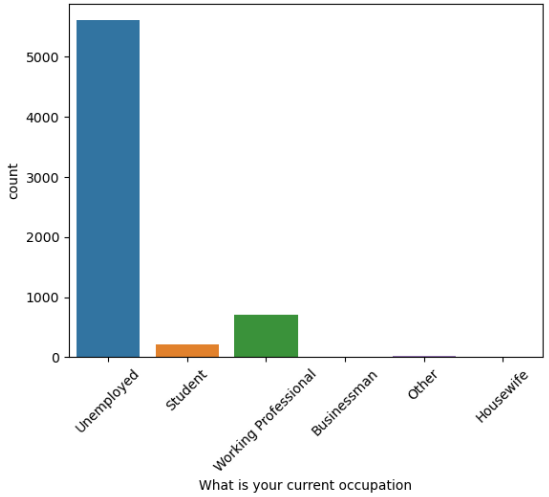
Data imbalance in target variable

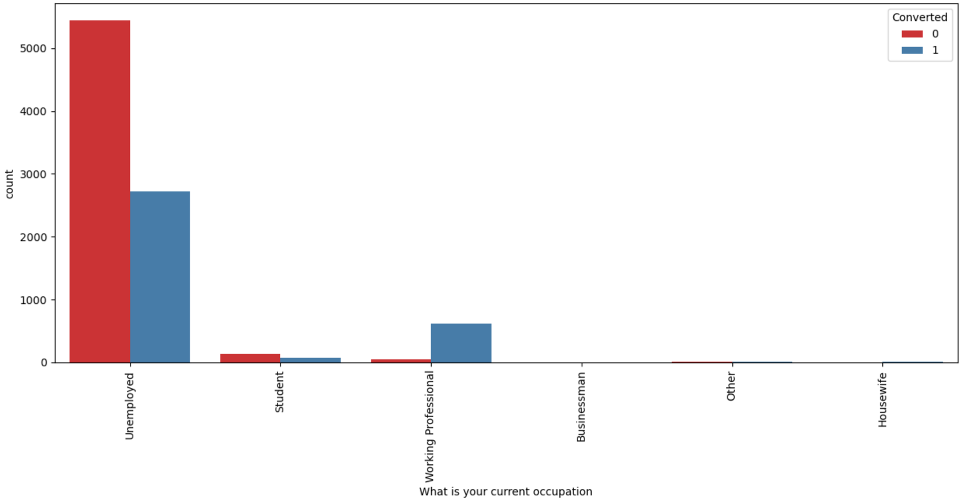
- **Data is imbalanced as the percentage of lead conversion is only 38% in the dataset.**



Occupation - EDA

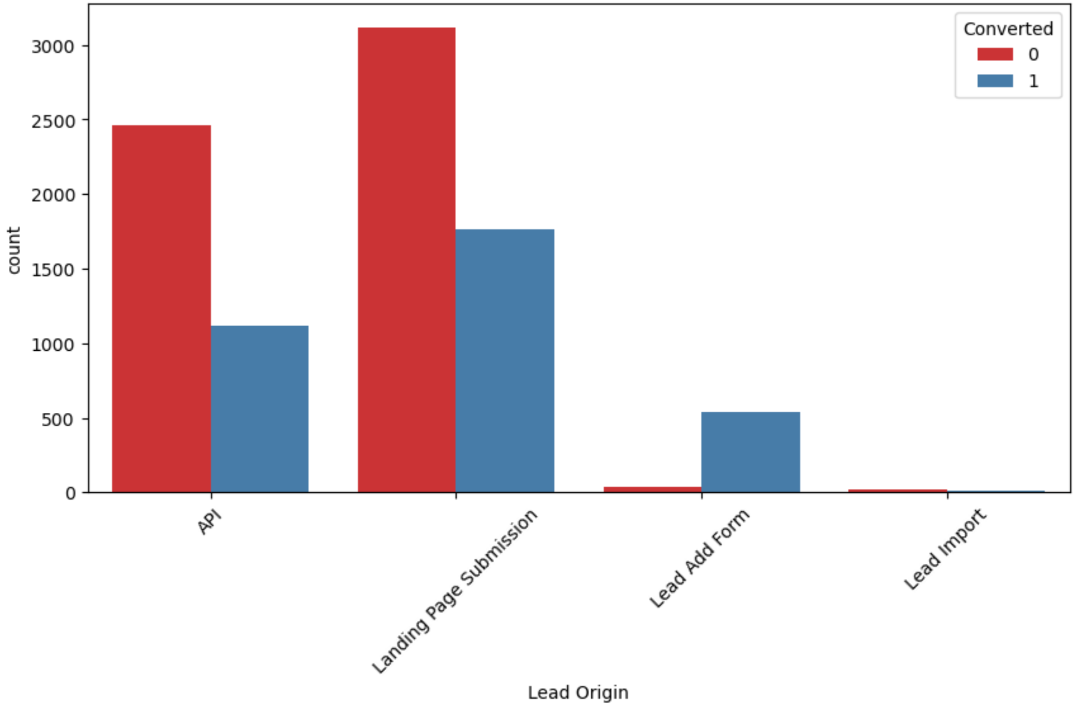
- **Most of the leads in the data are unemployed.**
- **The second most leads are working professional.**
- **The lead conversion rate is more for working professionals than unemployed leads.**





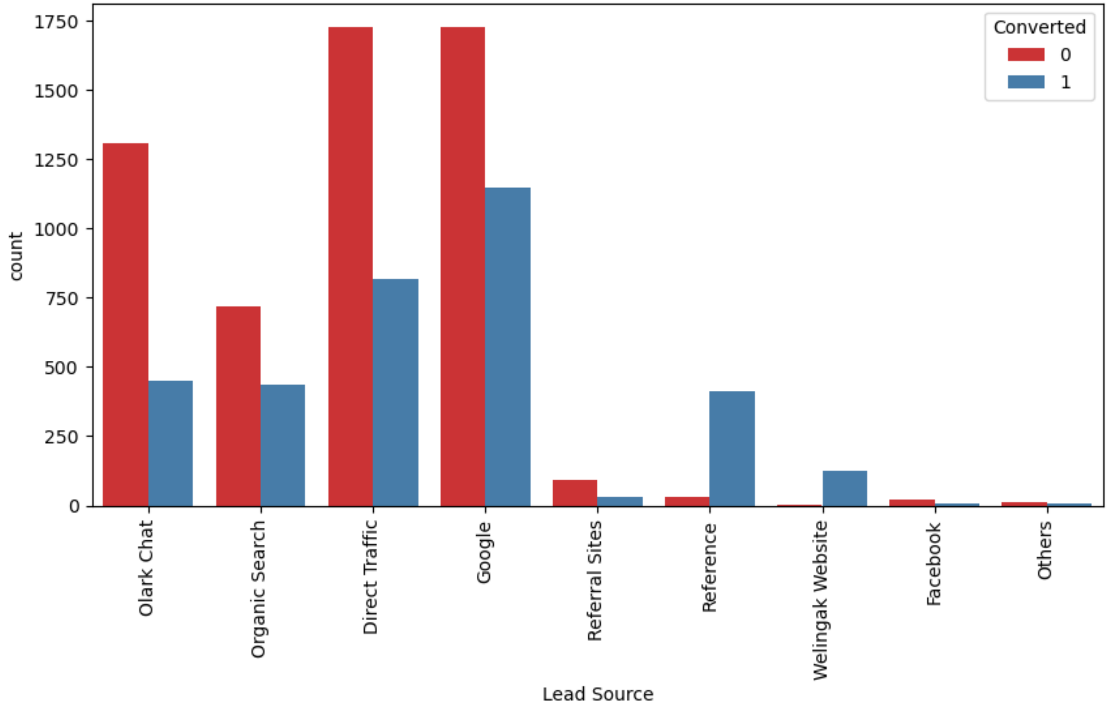
Lead Origin - EDA

- **The lead conversion rate is higher for Lead Add Form, when compared to other origin like landing page submission and API.**



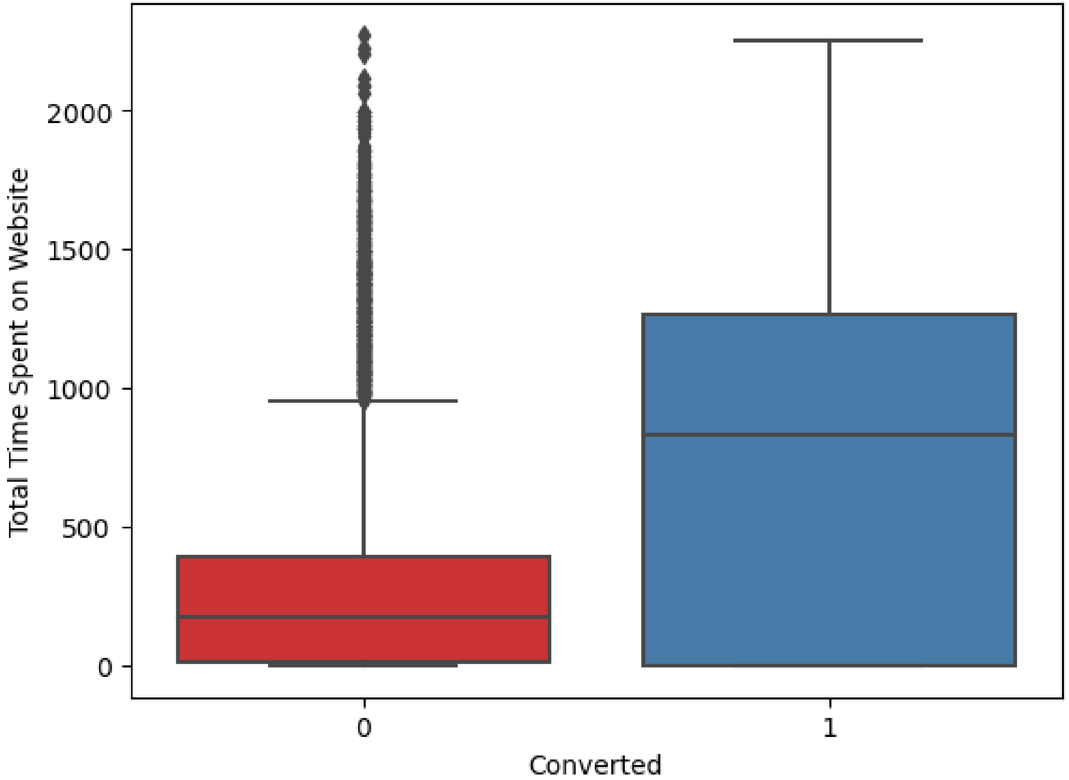
Lead source – EDA

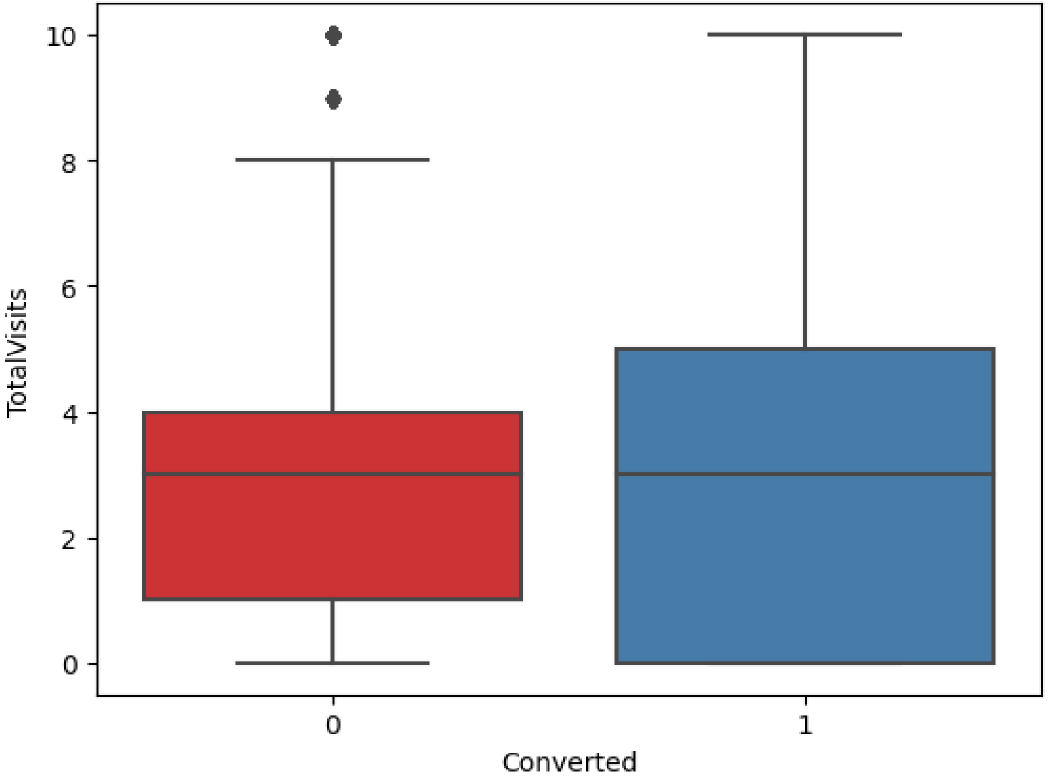
- **The maximum number of conversion is for ‘Google’ when comparing other lead source. But number of leads which did not get convert is higher than converted for this source.**
- **The ‘Reference’ lead source is having higher number of conversion than non converted leads.**
- **Also ‘Welingak Website’ lead source have higher number of conversion compared to non converted number.**



Total time spent on the site/ Total Visits - EDA

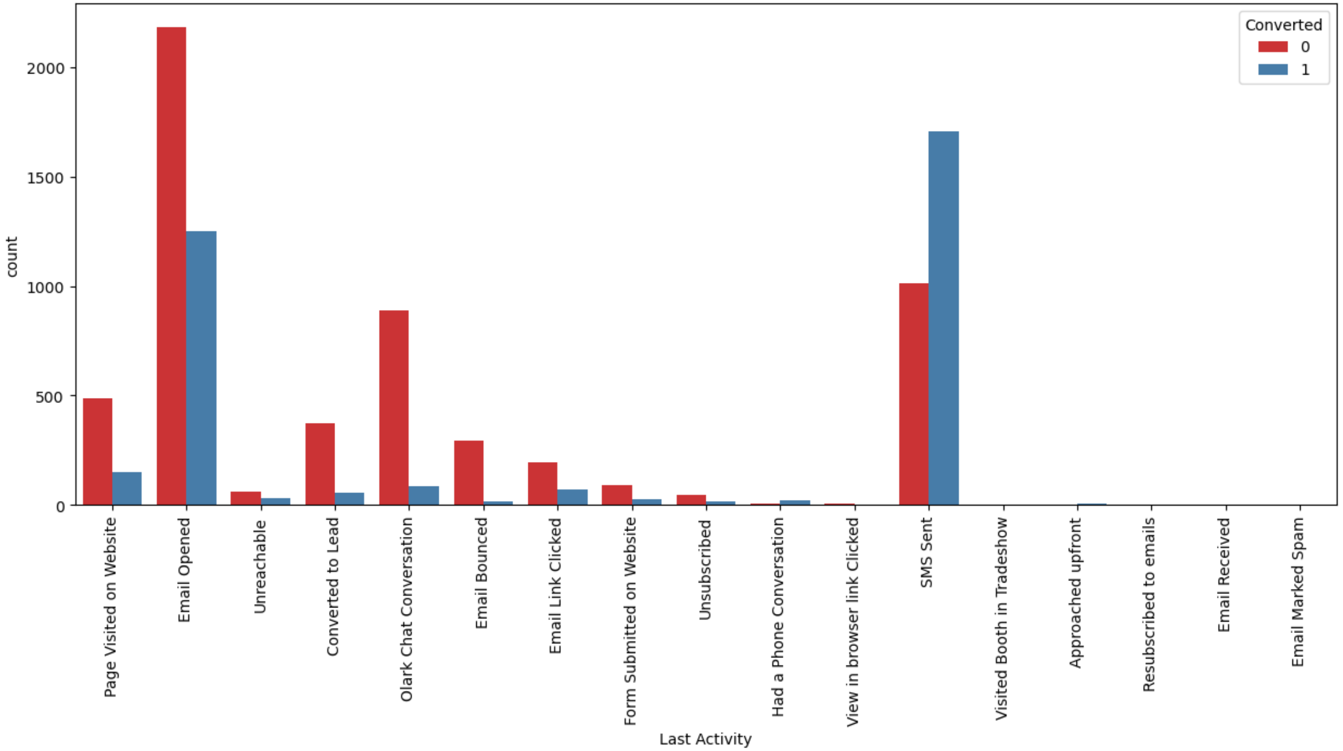
- **The total time spent by the leads who are converted is more when compared to non converted leads.**
- **The total number of visits is similar for both converted and non converted leads.**





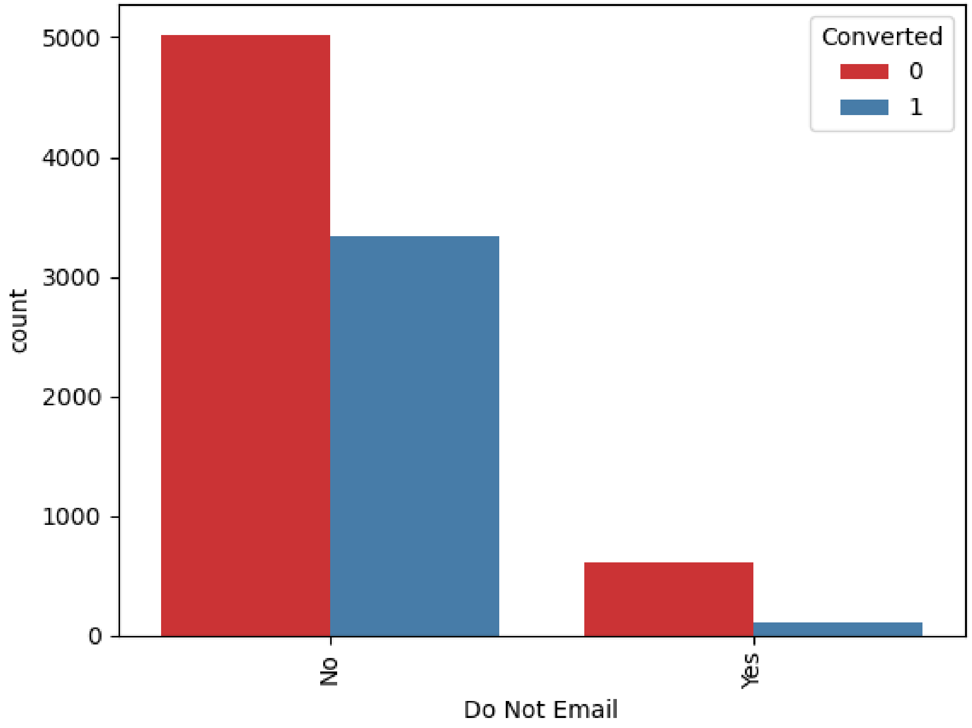
Last activity - EDA

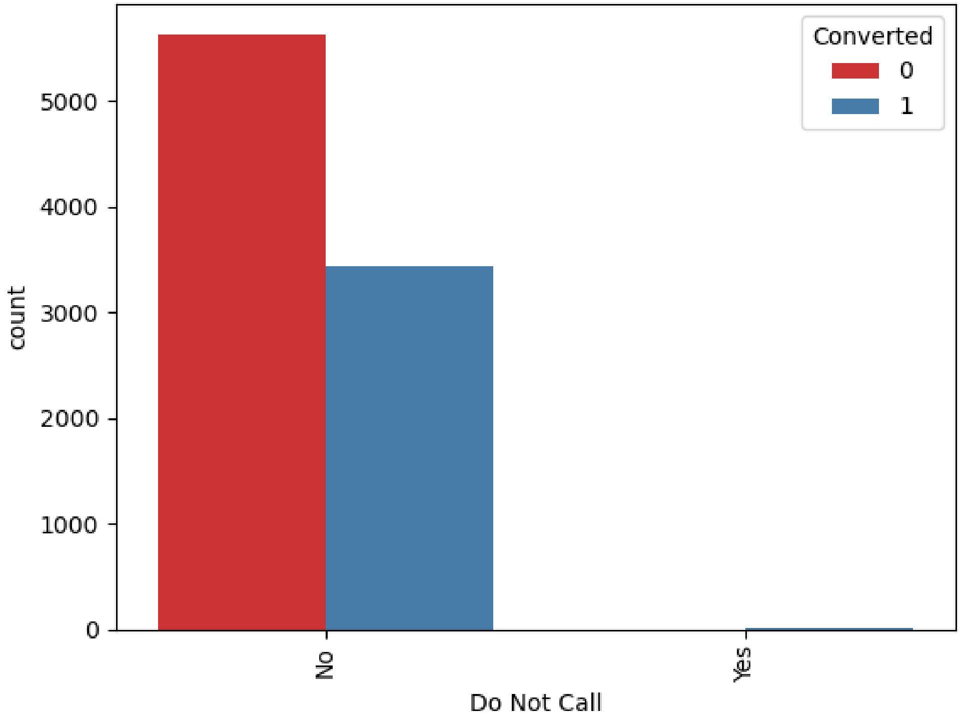
- **When the last activity of the lead is ‘SMS sent’, the rate of conversion is higher when compared to others.**
- **When the leads last activity is opening of email, then the rate of conversion is much lesser than rate of non conversion.**



Email/Call - EDA

- **Both converted and non converted leads not prefer calling or emailing them.**





Data Preparation

- **Conversion of binary variables to 1/0**
- **Created dummy variables for some of the categorical variables**
- **Split train and test data in the ratio 70:30**
- **Feature scaling using Standardized method**

Model Building

- **Using Recursive Feature Elimination method(RFE) selected the most important 20 features and reduced the computation time.**
- **By validating the P-Value and VIF(Variance Inflation Factor), eliminated the less significant variables, until we get a stable model.**

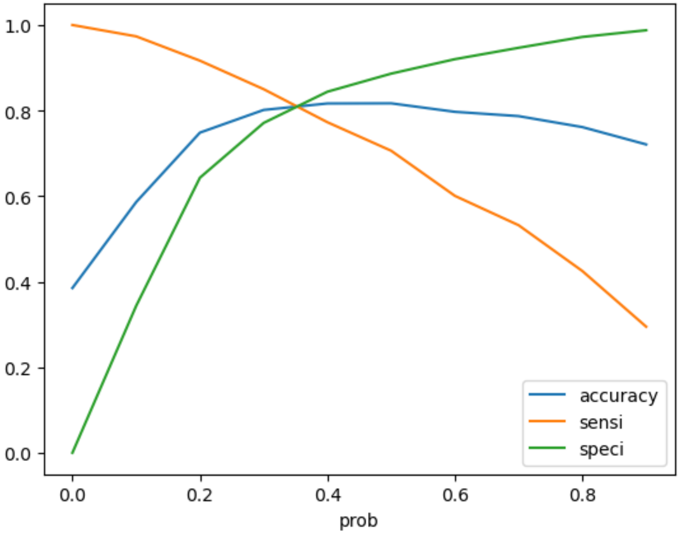
Model Evaluation

- **Got the Optimum cut off probability as .34 by plotting Accuracy, Sensitivity and Specificity**
- **There is a trade off between Precision and Recall, hence we can choose Sensitivity-specificity for test set evaluation.**

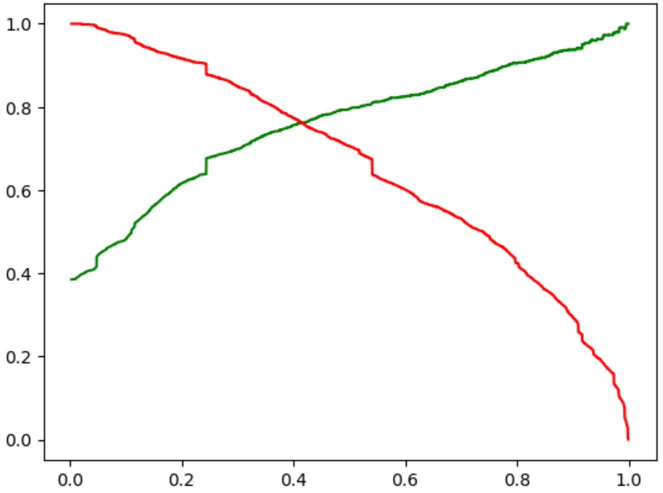
KPIs for train set

KPI	%
Accuracy	81
Sensitivity	81
Specificity	80
Precision	79
Recall	70

Accuracy-Sensitivity-Specificity



Precision-Recall Trade off



Model Evaluation

- **The calculated Accuracy, Sensitivity and Specificity for test data looks similar to that of train data**
- **Achieved a lead conversion rate of 80%**
- **The variable Lead Source Wellingak Website have highest coefficient value.**

KPIs for test set

KPI	%
Accuracy	80
Sensitivity	80
Specificity	80

Variables	Coeff Value
Lead Source_Welingak Website	5.811465
Lead Source_Reference	3.316598
What is your current occupation_Working Professional	2.608292
Last Activity_Other_Activity	2.175096
Last Activity_SMS Sent	1.294180
Total Time Spent on Website	1.095412
Lead Source_Olark Chat	1.081908

Recommendations

- **Overall, the X education can invest more fund towards the lead sources Welingak Website and Reference.Also effectively contact with Working Professionals also as there conversion rate is higher.**
- **During the intern higher periods,**
  - **The company should contact leads who are spending more time on ‘Welingak website’. Also, they can spend highly on advertising in ‘Welingak Website’.**
  - **The company can contact the reference provided by the learners to increase the lead conversion rate. They can provide discounts for learners who are providing reference that converts to lead.**
  - **They can contact a greater number of working professionals to improve the lead number.**
  - **Another variable the X education can focus on is the leads whose last activity is SMS to X education. The team can make calls to this leads for effective conversion rate.**
  - **Another strategy they can follow is contact the leads who are spending maximum time on the website.**
  - **Also, they can contact the leads who used Olark chat to increase the number of leads to be converted.**

Recommendations

- **When the company reaches its target for a quarter before the deadline,**
  - **The company can focus contacting only on high-impact lead sources like ‘Welingak Website’ and Reference.**
  - **They can tailor strategies for Working Professionals.**
  - **Engage with SMS and other activities effectively.**
  - **Optimize website experience for longer visits.**
  - **Leverage Olark Chat for real-time interactions.**
  - **Respect 'Do Not Email' preferences.**
  - **Monitor and adjust strategies for Landing Page Submission.**

Thank You