Lead Score Case Study

Submitted by:

Sonali

Prasanth Reddy Kumar Mura

Nitish Singh

Problem statement

- An education company, X Education sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google.
- Once 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. The typical lead conversion rate at X Education is around 30%.

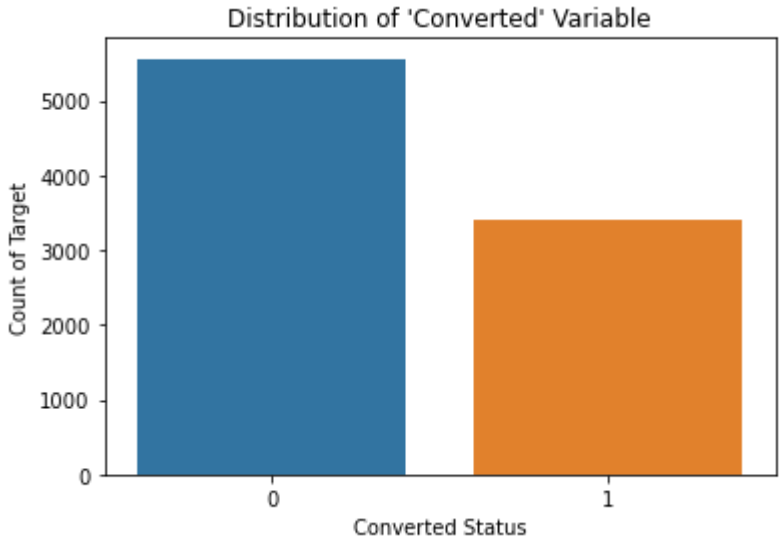
Business Goals

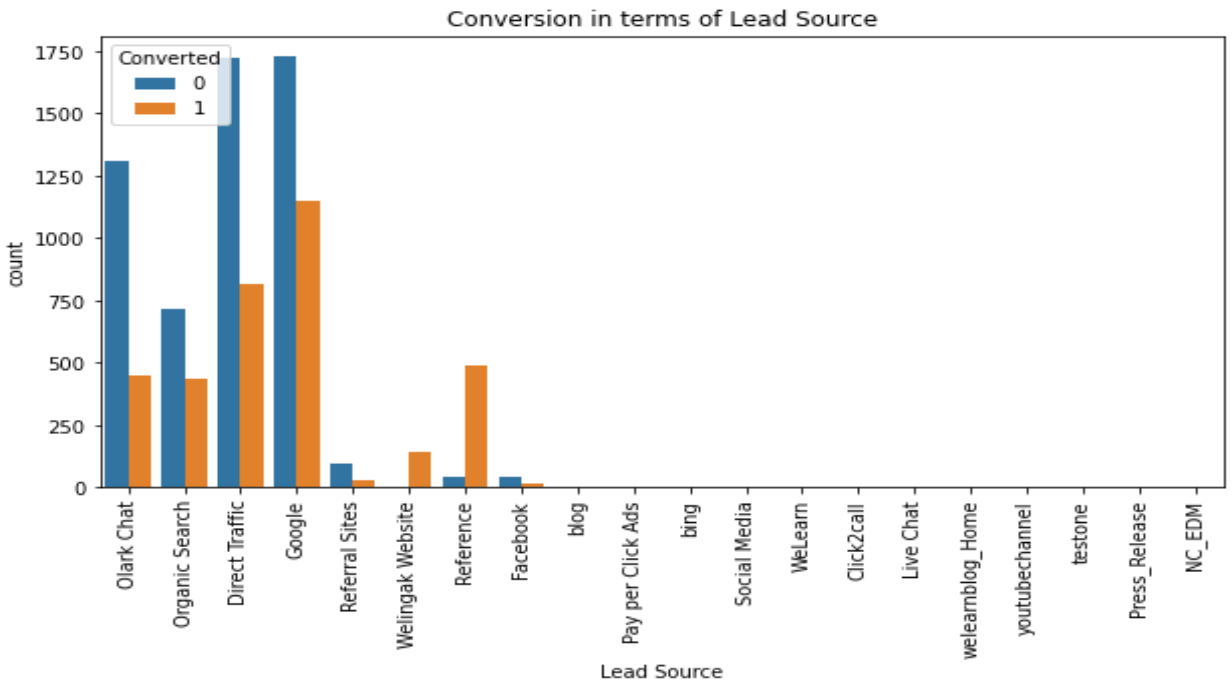
- Company wishes to identify the most potential leads, also known as “Hot Leads”.
- The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%.

Overall Approach

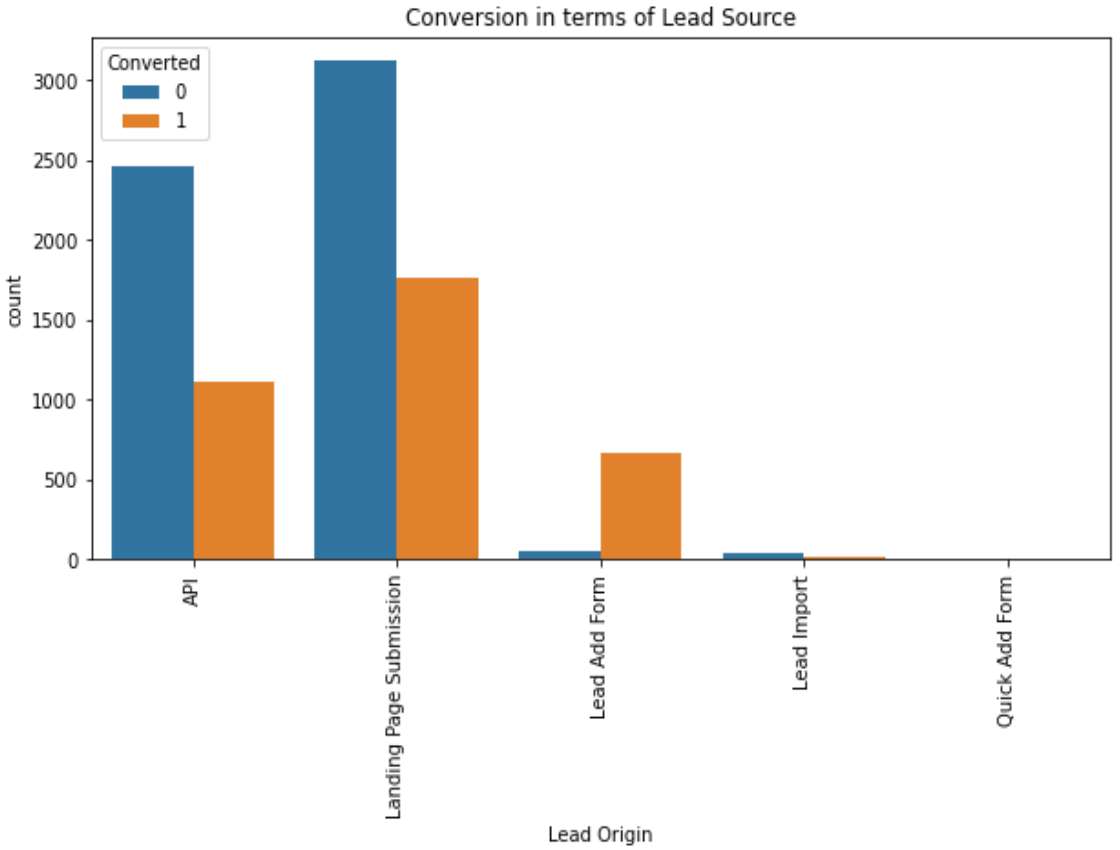
- Data cleaning and imputing missing values.
- Exploratory data analysis :univariate and bivariate analysis.
- Feature scaling and dummy variables creation.
- Logistic Regression model building.
- Model Evaluation :specificity, sensitivity, precision and recall.
- Conclusion and Recommendations.

Exploratory Data Analysis



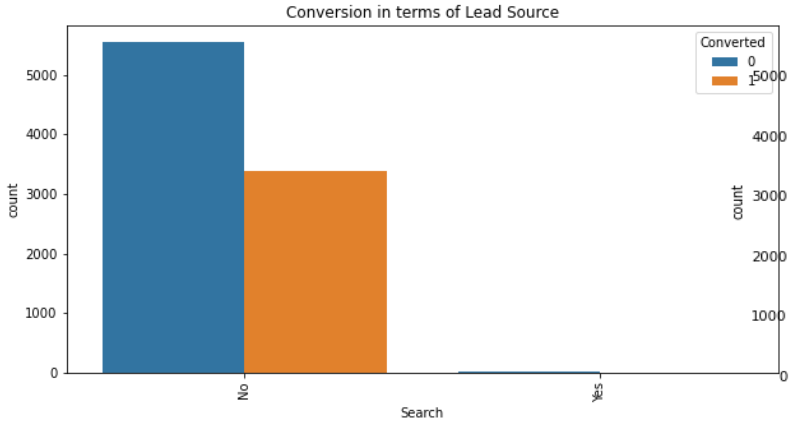


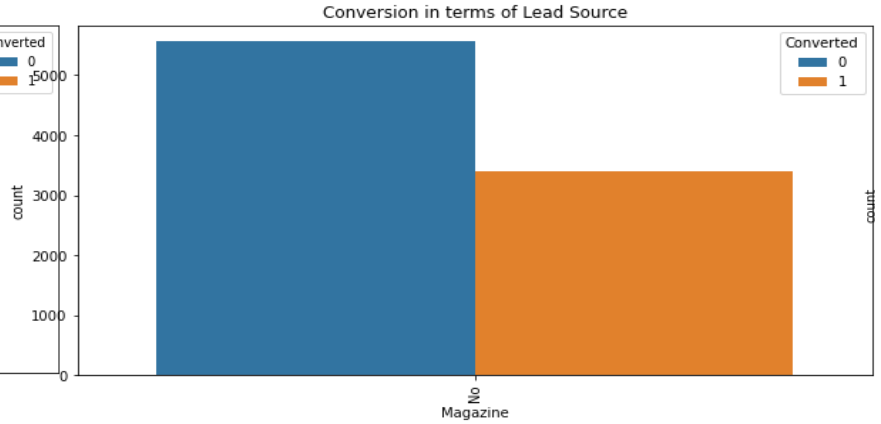
Exploratory Data Analysis

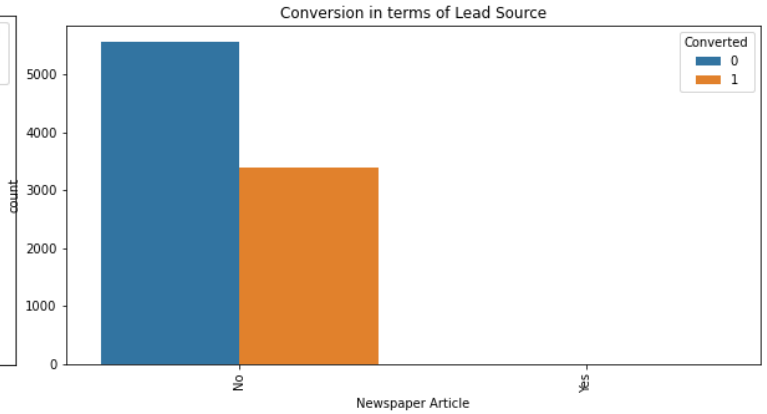


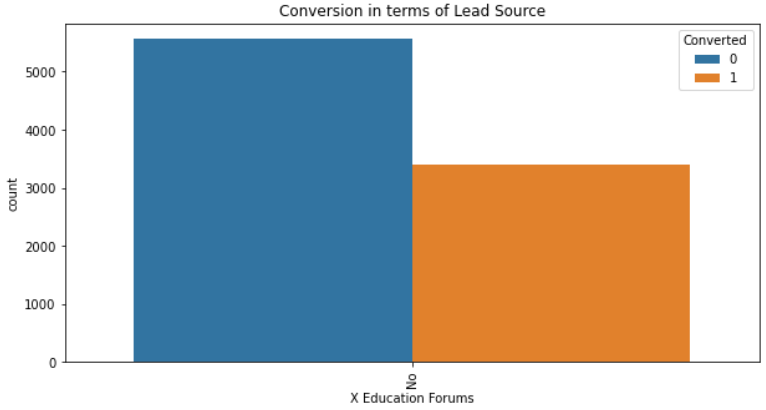
- We have a conversion rate of 37.92%.
- The count of leads from the Google and Direct Traffic is maximum.
- The conversion rate of the leads from Reference and Welingak Website is maximum.
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable.
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

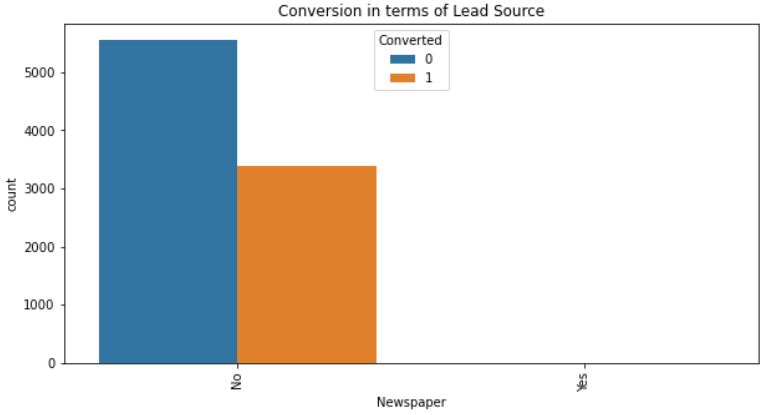
Exploratory Data Analysis

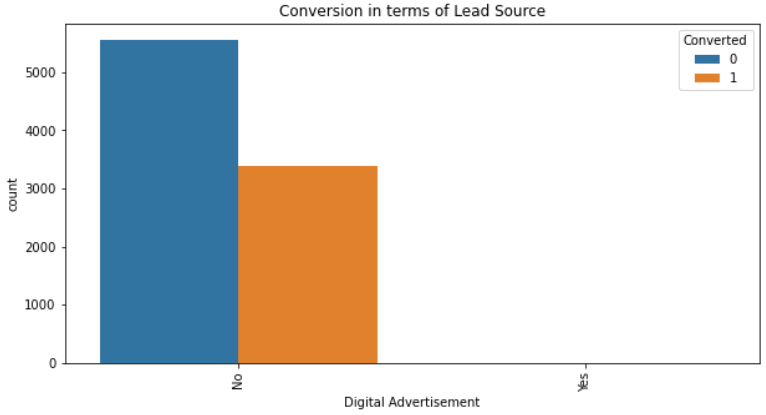




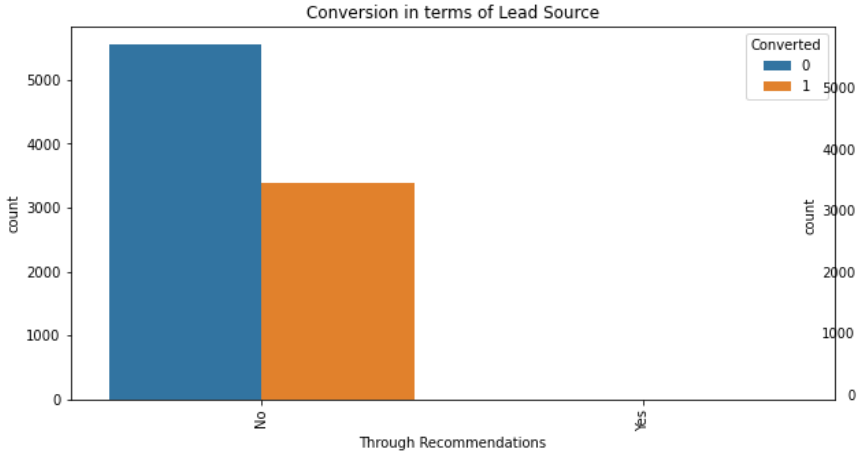


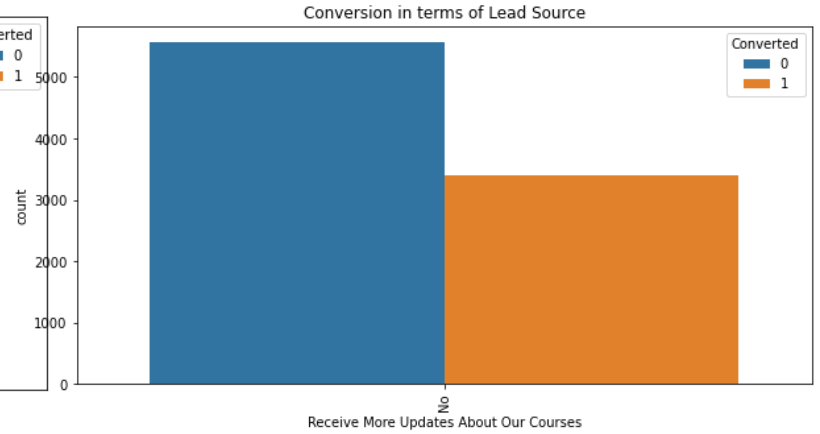


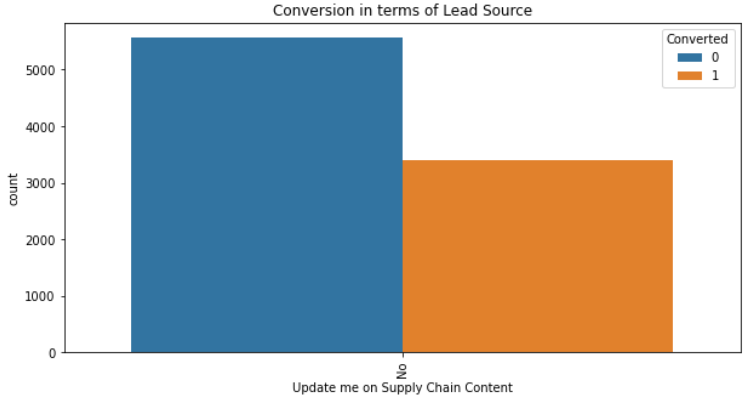


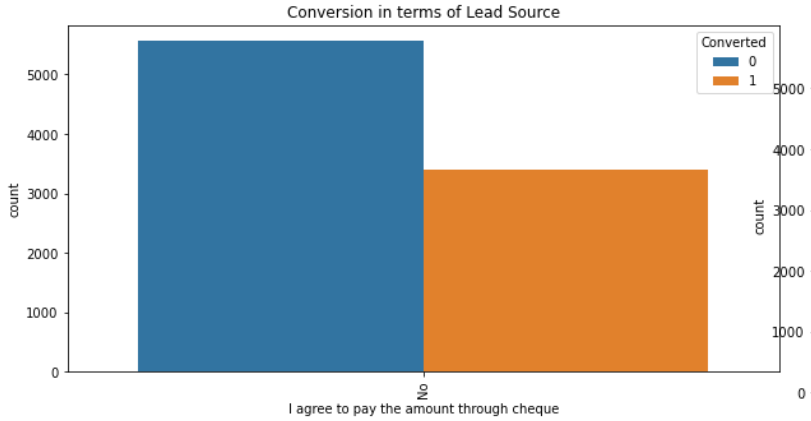


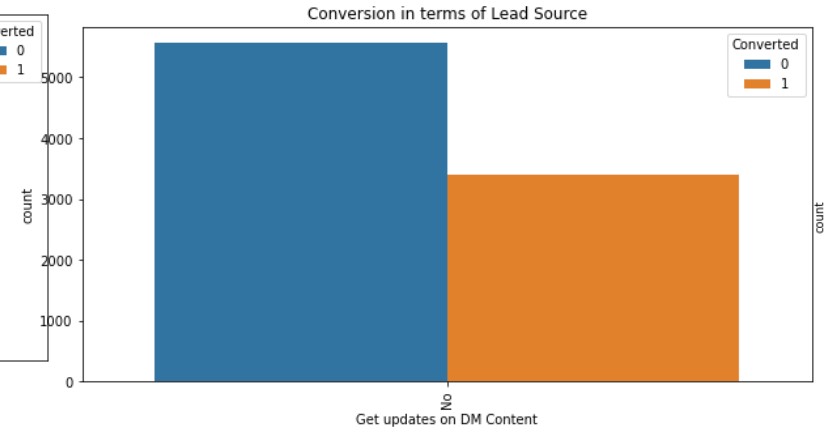
Exploratory Data Analysis

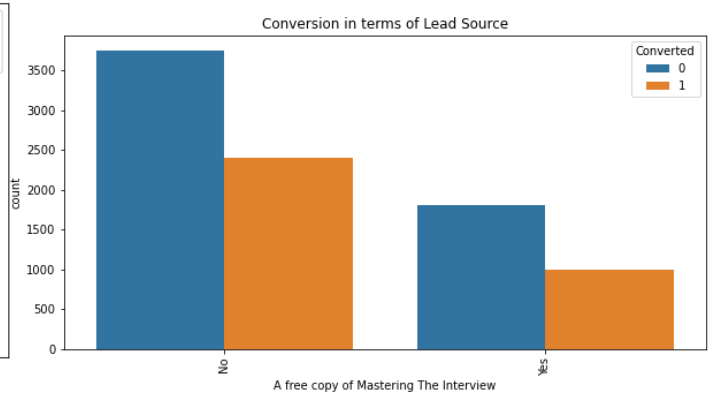












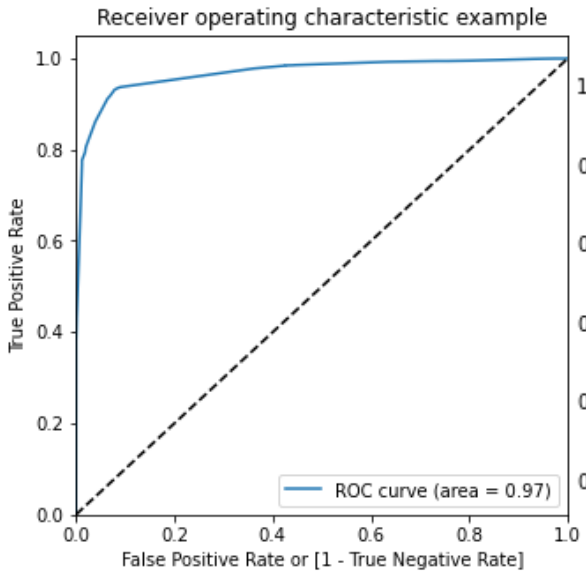
Exploratory Data Analysis

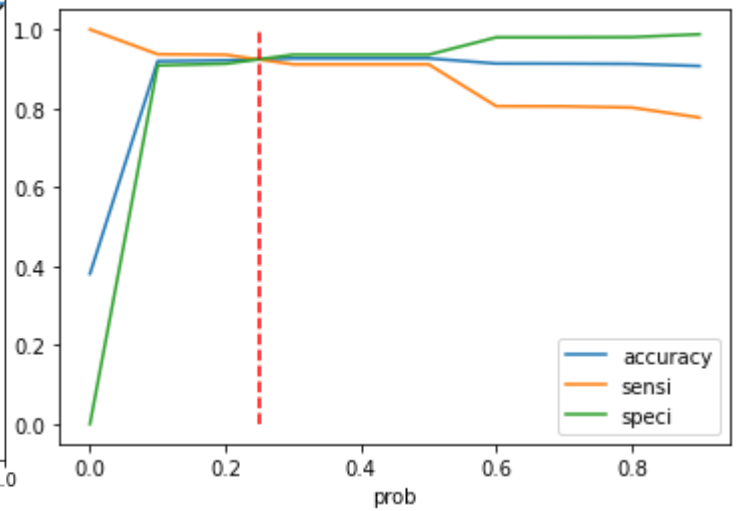
- From the graphs of the features on the previous slide most of the data is highly imbalanced hence they were dropped.
- “A free copy of Mastering The Interview” is the only one that doesn’t have a high data imbalance

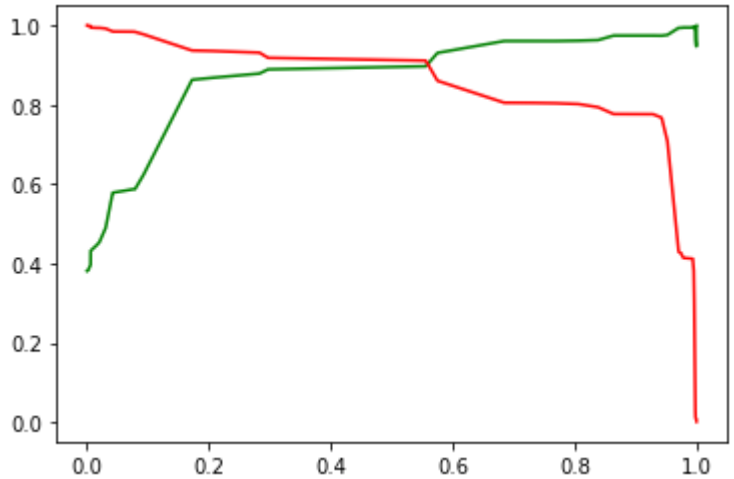
Model Building

- Splitting the data into test and training sets.
- Split Ratio is 70:30 for train and test respectively.
- Use RFE to choose top 15 variables.
- Build the model by removing the variables whose p-value score is > 0.05 and VIF is > 5.
- Used the model on the test dataset to predict outcomes.

Model Building







Precision Recall Trade-off

Model Prediction

Train data:

```
Accuracy: 92.51 %
Sensitivity: 91.15 %
Specificity: 93.58 %
```

Test data:

```
Accuracy: 93.07 %
Sensitivity: 94.05 %
Specificity: 92.48 %
```

Conclusion and Recommendations

- The logistic regression model is used to predict the probabililty of conversion of a customer.
- While we have calculated both sensitivity-specificity as well as Precision-Recall metrics, we have considered optimal cut off on the basis of sensitivity-specificity for final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 93.07 %, 94.05 % and 92.48 % respectively which are approximately closer to the respective values calculated using trained set. (92.51 %, 91.15 % and 93.58 %)
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:
  - Tags_ closed by horizon
  - Tags_ lost to EINS
  - Tags_Will revert after reading the email
- Hence, the overall this model seems to be good.