Lead Scoring Optimization

By

Arnab Saha

Problem Statement

X-Education is an online education company which provides online courses to industry professionals.

The company markets its courses on several websites and search engines like Google.

The company wants to capture most promising leads that can be converted to paying customers.

Although the company gets a lot of leads, its lead conversion rate is very poor.

Current lead conversion rate is only 30%.

Business Goal

- The company wants us to build a Logistic Regression Model.
- The Model should be able to assign a lead score between 0 and 100 to each of the leads.
- This lead score 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.
- A lower score would mean that the lead is cold and will mostly not get converted.
- The CEO of the company, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data Set

We have been provided with a leads dataset from the past with around 9000 data points.

This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

The target variable, in this case, is the column Converted.

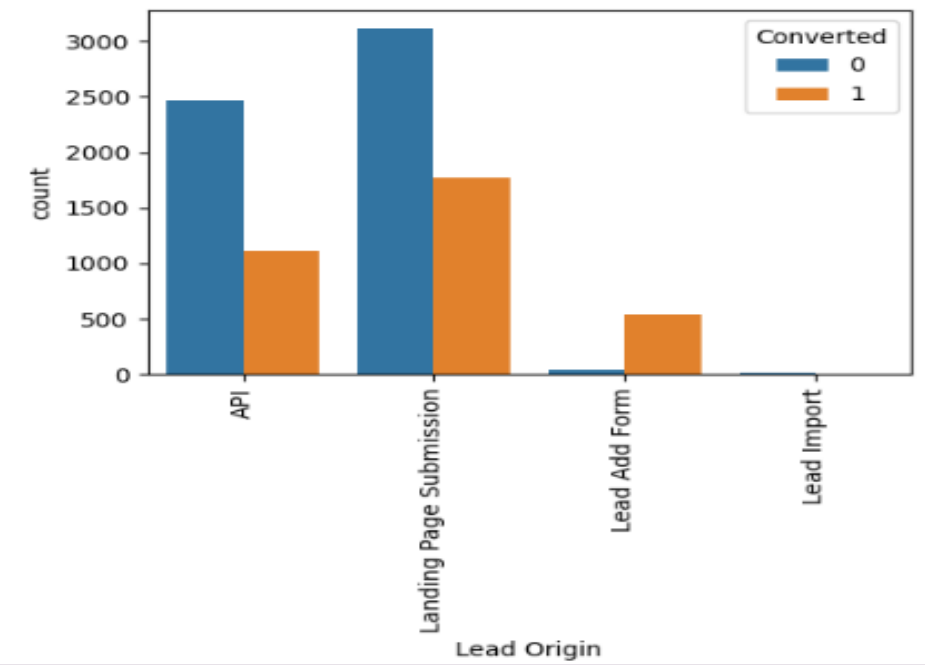
It tells whether a past lead was converted or not.

Where 1 means it was converted and 0 means it wasn't converted.

Model Building Strategy

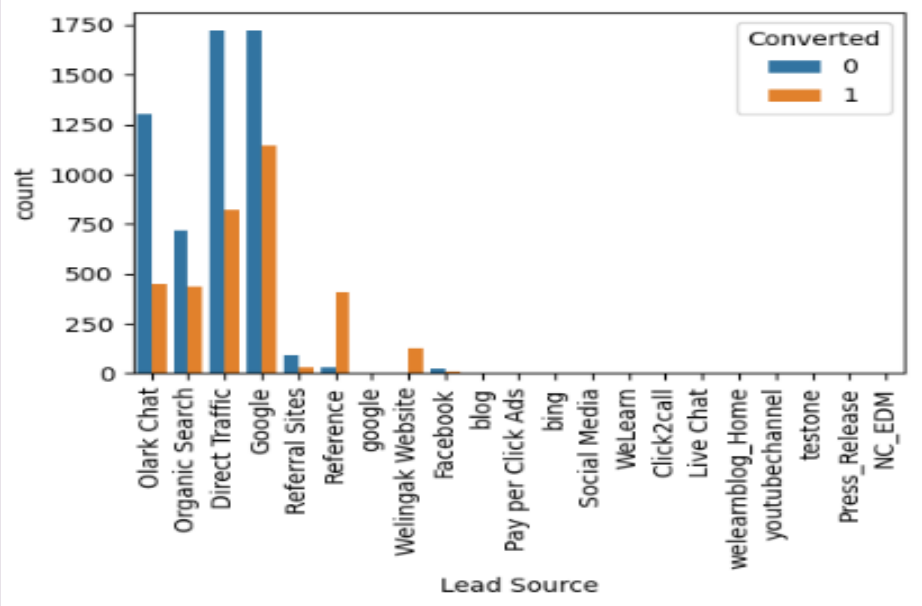
Exploratory Data Analysis Findings

- Every Numeric and Categoric features are visualized one by one with our target variable to see what is their influence on the target variable.
- Some of the important findings are shown



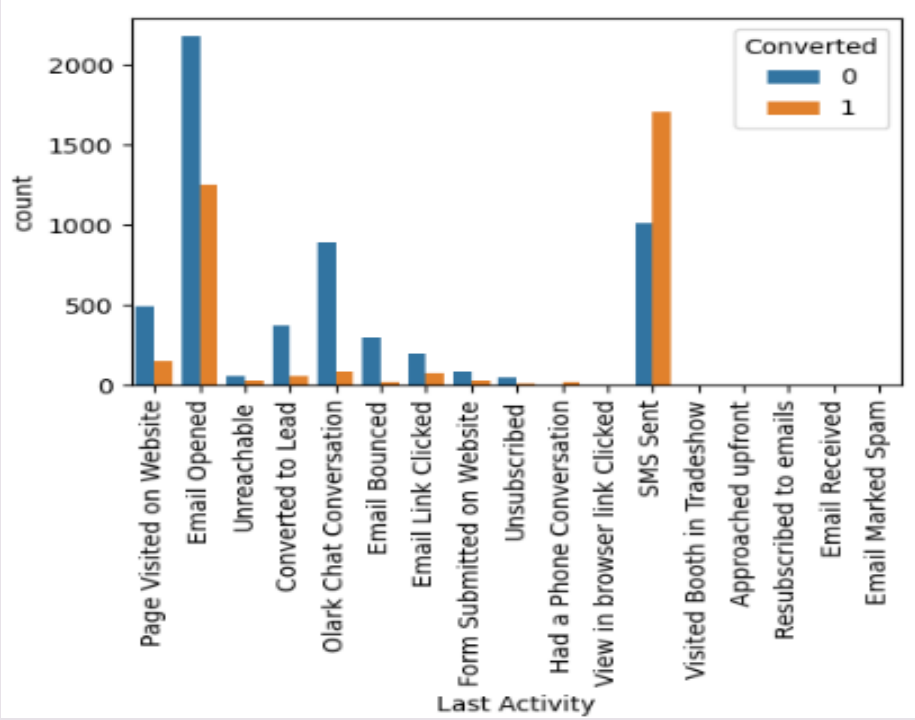
Lead origin VS Converted

- API and Landing page submission have 30-40% lead conversion rate but count of lead originated from them are less.
- Lead add form has more than 90% conversion rate.



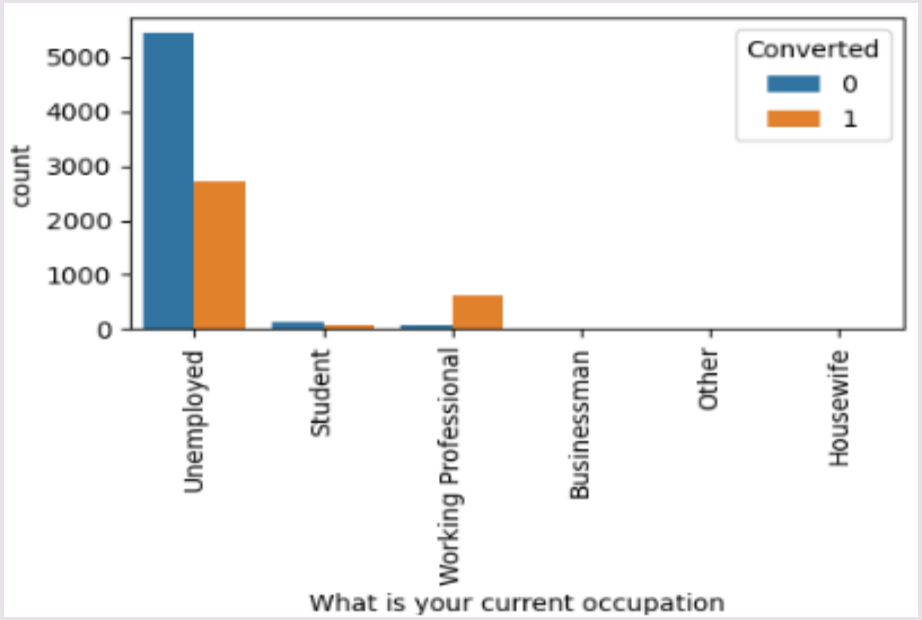
Lead Source VS Converted

- Google and Direct Traffic generates maximum number of leads.
- Conversion rate through reference and welingak website is high.



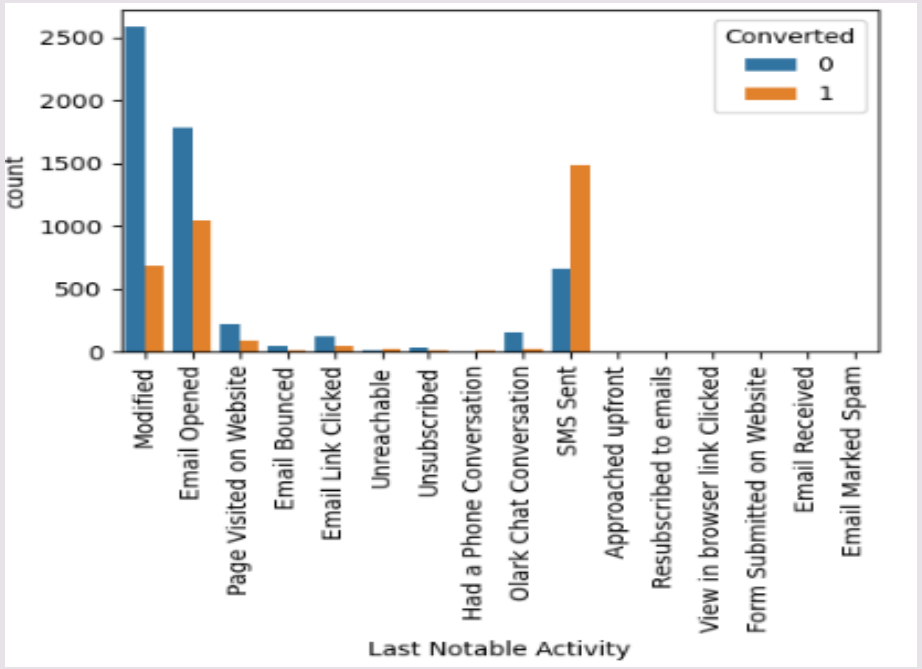
Last Activity VS Converted

- Most of the leads have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS sent is high.



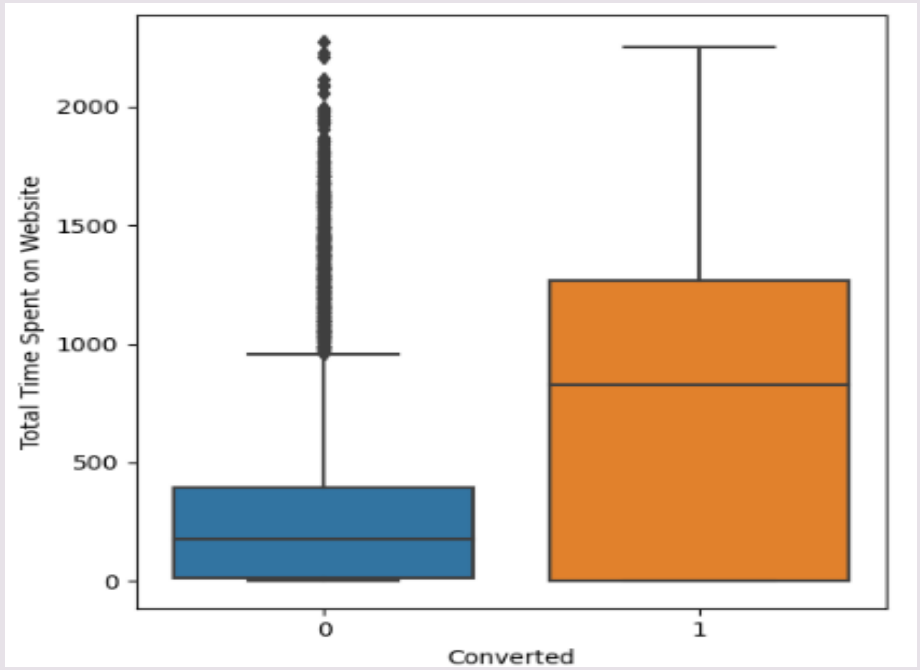
Current Occupation VS Converted

- Working professionals opting for the courses have high chances of conversion.
- Unemployed leads are the most in numbers but have only 30-35% conversion rate.



Last Notable Activity VS Converted

- Most of the leads have SMS sent as their last notable activity.
- Conversion rate for leads with last notable activity as Email opened and Modified is around 30-35% only.



Total Time Spent on Website VS Converted

- Total Time Spent on Website has good impact on lead conversion.
- Leads spending more time on website tends to convert more.

Model building

- Splitting the Data in Train and Test Set.
- Re-Scaling of the features.
- Use RFE to eliminate less relevant features.
- Build the first model.
- Eliminate variables based on high P-value.
- Check VIF value for all the existing features.
- Predict using Train Data set.
- Evaluate accuracy and other metrices for Train set.
- Predict using Test Data Set.
- Evaluate accuracy and other metrices for Test set.

Model Evaluation (Train)

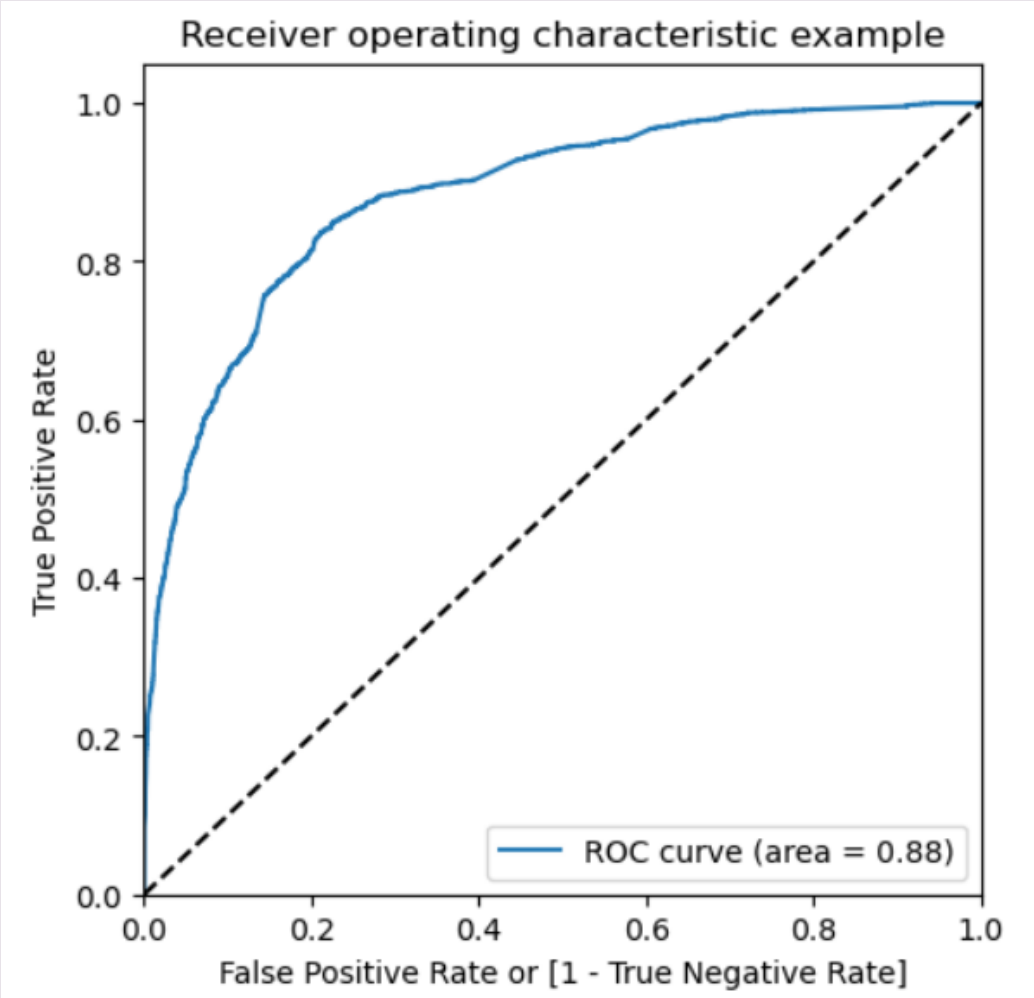


Fig : ROC Curve

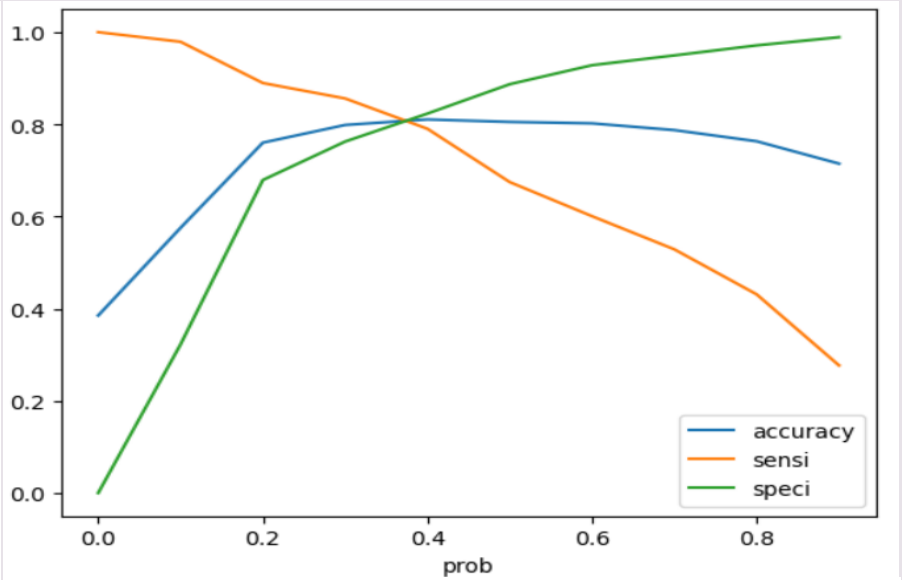


Fig : Accuracy, Sensitivity and Specificity Trade-off

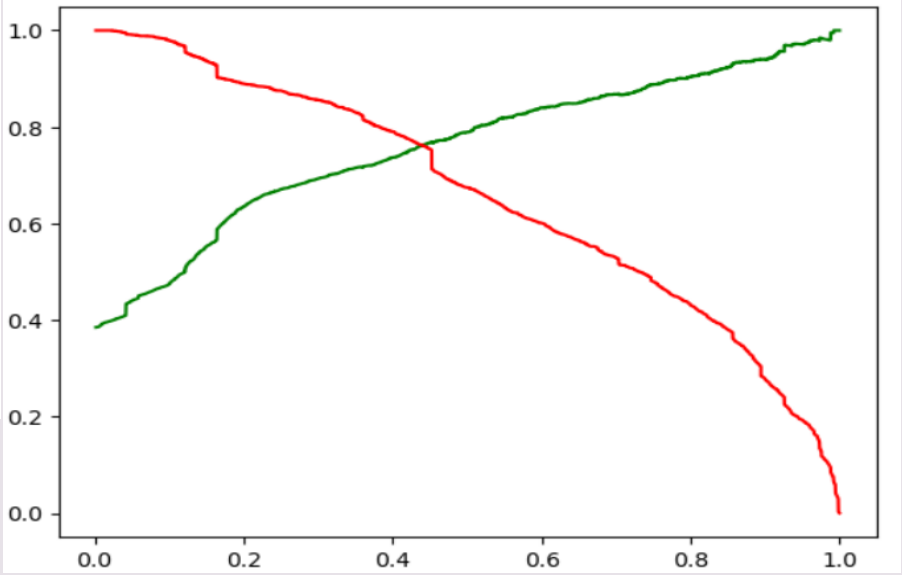


Fig : Precision-Recall Trade-off

Model Evaluation Results

Model performance on the Train dataset, We got below evolution metrices -

- Accuracy : 80.5%
- Sensitivity: 80.9%
- Specificity: 80.3%

Model Performance on the Test dataset, we got below evaluation metrices -

- Accuracy : 79.8%
- Sensitivity: 78.6%
- Specificity: 80.5%

Summary

Thus, we can say that our model is able to achieve a lead conversion rate of approx. 80%.

And thus, successfully achieved the X-Education company CEO's goal of getting a ball pack of the target lead conversion rate to be around 80%.

The Model seems to predict the conversion rate very well and we should be able to give the company confidence in making good calls based on this model, to get a higher lead conversion rate of around 80%.

Additionally a lead score is assigned to each leads, and a list of potential leads also provided for ease of company's team.

Recommendations

- The company should contact the leads coming from "Lead Origin" as "Lead Add Form" as there is a high chance of lead conversion. On the other hand, leads coming from "Lead Origin" as "Landing Page Submission" is seen to have less chances lead conversion. So, company should improve these areas to attract more leads.
- The company should contact the leads with "occupation" as "Working Professionals" as there is a high chance of lead conversion.
- The company should contact the leads coming from "Lead Source" of "Welingak Website" as there is a high chance of lead conversion. And should improve "Lead Sources" of "Direct Traffic" and "Organic Search" since they have less chances lead conversion rate.
- The company should contact the leads having "Last Activity" as "Had a Phone Conversation" as there is a high chance of lead conversion.
- The company should contact the leads having high "Total Time Spent on Website" as there is a high chance of lead conversion.
- The company should avoid contacting leads having "Specialization as Other" , "Last Activity as Converted to Lead" , "Last Activity as Olark Chat Conversation", "Last Activity as Email Bounced" , "Last Notable Activity as Olark Chat Conversation", "Last Notable Activity as Email Opened", "Do Not Email", "Last Notable Activity as Page Visited on Website", "Last Notable Activity as Email Link Clicked", "Last Notable Activity as Modified" since they are less prone to conversion.
- And improving in these areas will attract more leads and corresponds to a higher lead conversion.

Thank You