Lead Scoring Case Study

Submitted by:

Reetuparna Ghosh

Vijay Aggarwal

Lead Score Case Study for X Education

Problem Statement :

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if say, they acquire 100 leads in a day, only about 30 of them are converted.

- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

- If they successfully identify this set of leads, 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.

Business Goals

- X Education needs help in selecting the most promising leads.

- The company needs a model where a lead score is assigned 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%.

Problem Solving Methodology

- **Data cleaning and data manipulation.**

1. Check and handle the duplicate data, NA and missing values.
2. Drop the columns if it contains large amount of missing values and have no use for the analysis.
3. Imputation of the values.
4. Check and handle outliers in data.

- **EDA**

1. Univariate data analysis: By checking value count, distribution of variable etc.

2. Bivariate data analysis: correlation coefficients and pattern between the variables etc. Feature Scaling & Dummy Variables and encoding of the data.

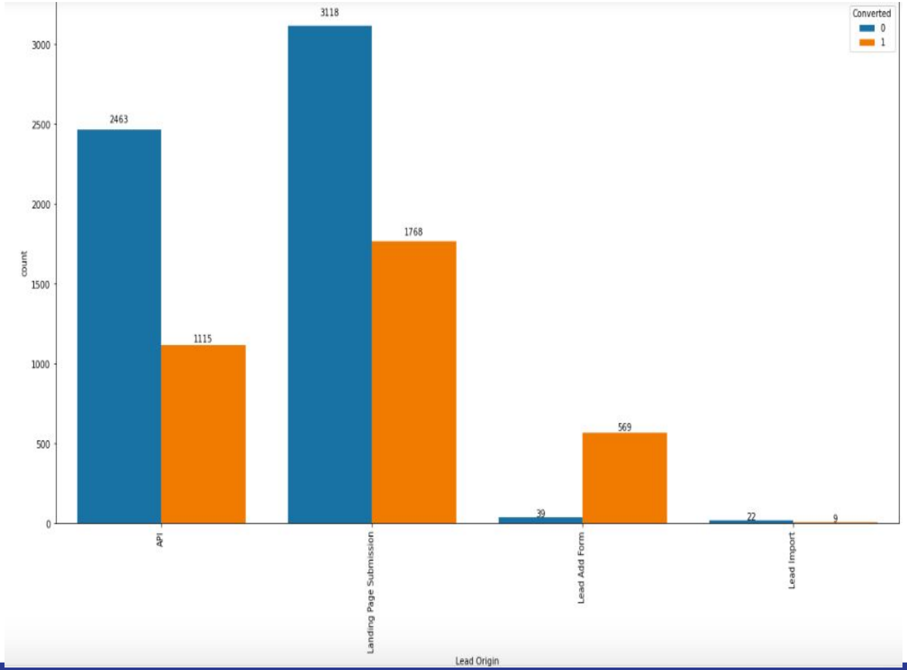
- **Classification technique:** Making a logistic regression model for prediction.
- **Model Validation.**
- **Model presentation.**
- **Conclusions and recommendations.**

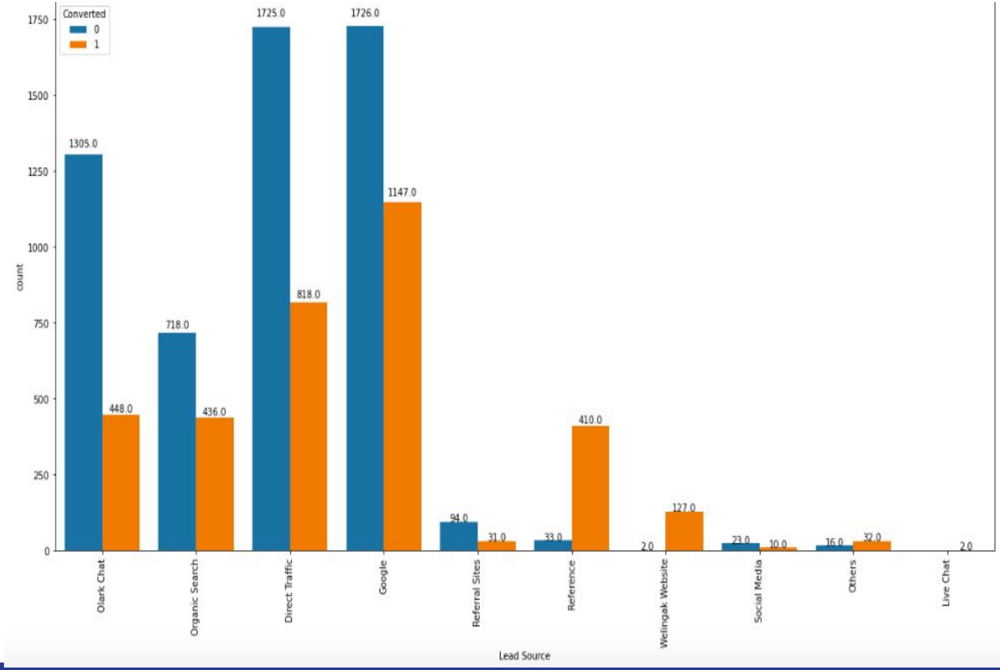
EDA

Visualising Categorical Variables

Lead Origin & Lead Source Count

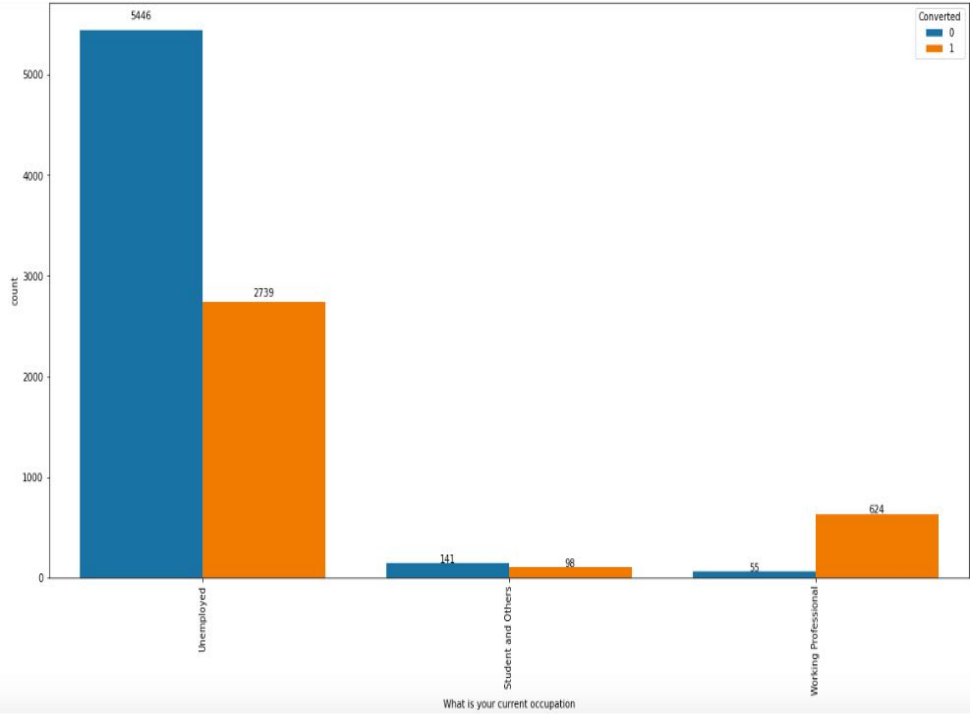
- ❖ Lead Add Form has more than 90% conversion rate but count of lead are low and Lead Import are very less in count as compare to others.
- ❖ Google and Direct traffic leads to the most conversion rate in Lead source.

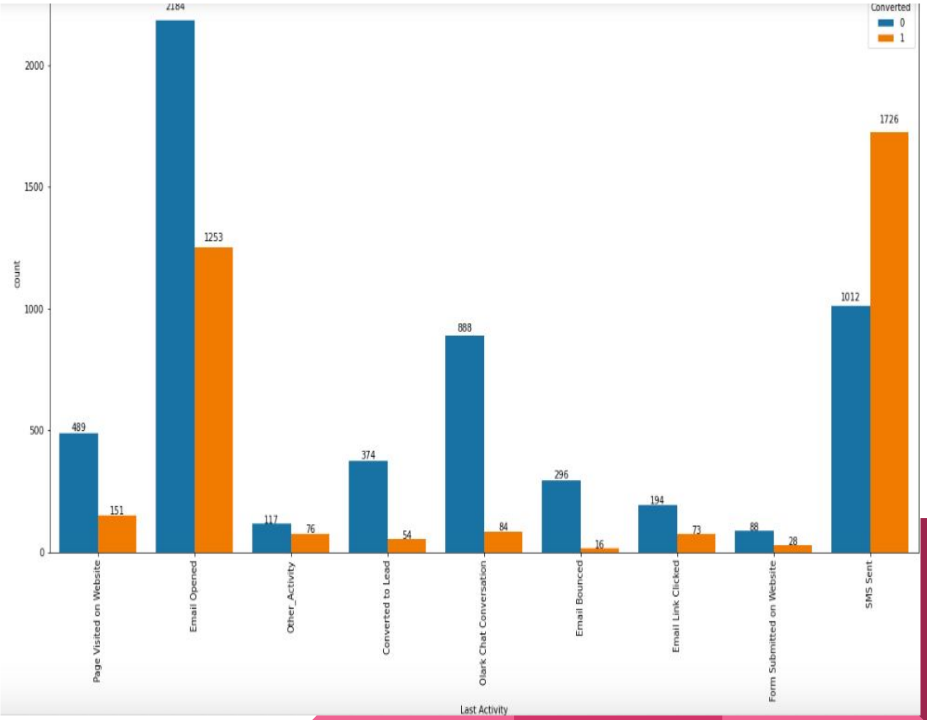




What is your current occupation & Last Activity Count

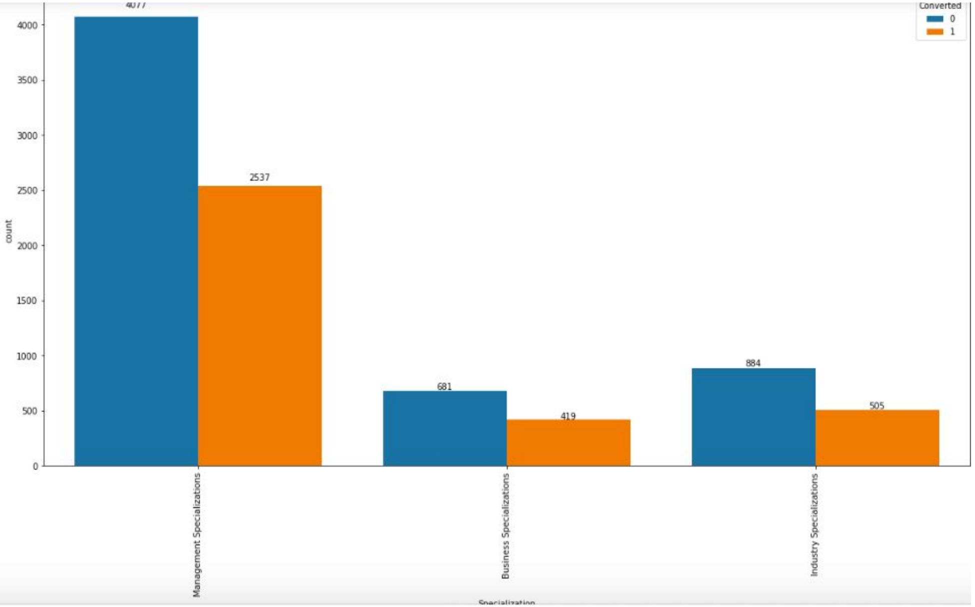
- ❖ Most of the conversion happened with the customer who are unemployed.
- ❖ Last Activity value of SMS Sent and Email opened had more conversion.

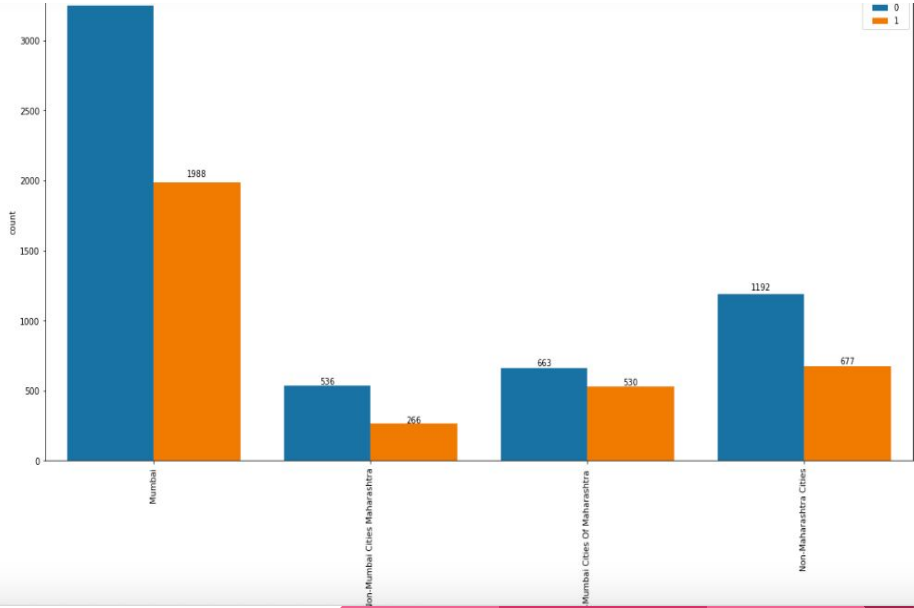




Specialization & City Count

Customers who are in management specialization and from Mumbai location having most conversion lead rate.



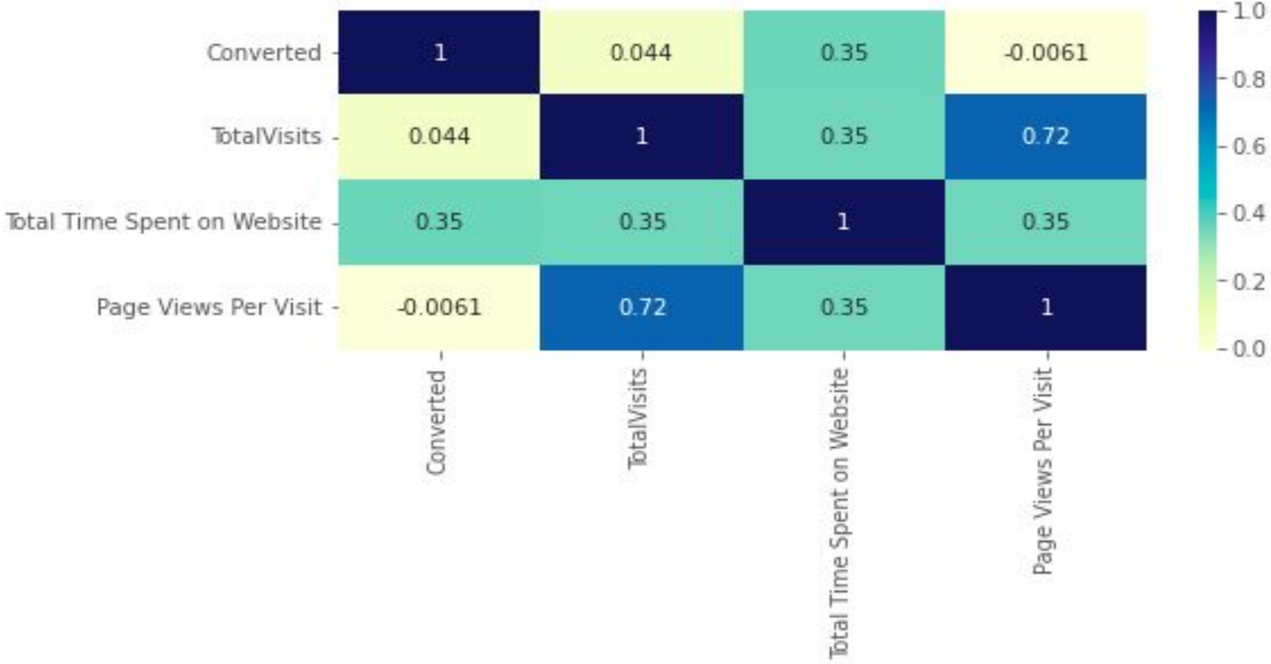


Visualisation of Numeric Variables

Correlation Of Numeric Variables

Correlated Variables- Converted', 'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit'

We can see that the variables are not highly correlated with each other but still there is multicollinearity among some features.



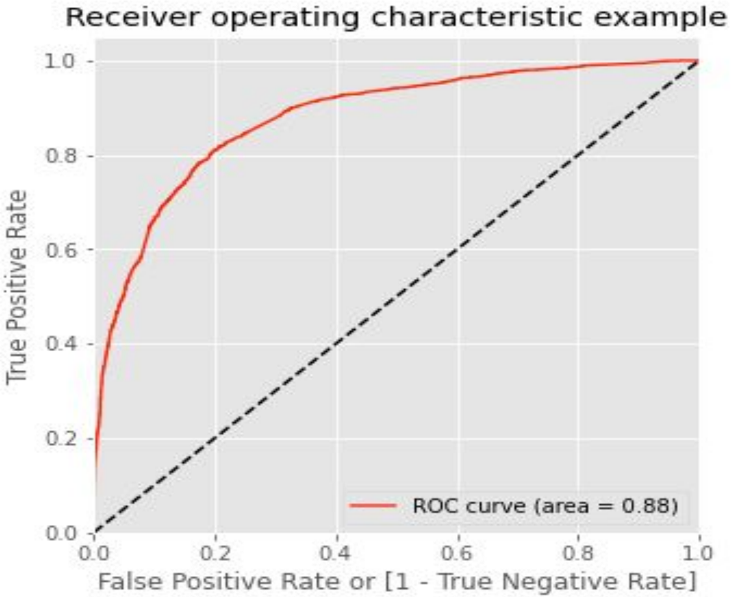
Model Building

- Splitting the data into Train and Test set.
- First step for regression is to perform train-test split, for that we have chosen 70:30 ratio.
- Use RFE to eliminate less relevant variables.
- Building models by removing the variables whose p-value is greater than 0.05 and
- VIF value is greater than 5.
- Prediction on dataset.
- Precision and Recall analysis based on the predictions.

Model Evaluation

ROC Curve:

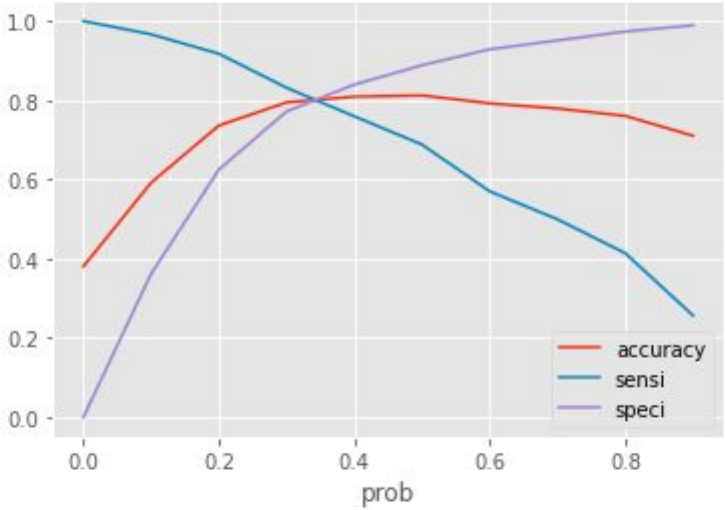
ROC Curve value should be close to 1 considers a good model and graph shows a good value of 0.88 which indicates a good predictive model.



Checking Accuracy ,Sensitivity, Specificity, Precision and Recall on Train set

The graph depicts an optimal cut off of 0.35 based on Accuracy, Sensitivity and Specificity.

- ❖ Accuracy - 80%
- ❖ Sensitivity - 80 %
- ❖ Specificity - 80%
- ❖ False Positive Rate - 19 %
- ❖ Positive Predictive Value - 72 %
- ❖ Positive Predictive Value – 87%
- ❖ Precision- 72%
- ❖ Recall- 80%



Checking Accuracy ,Sensitivity, Specificity, Precision and Recall on Test set

- Accuracy - 80%
- Sensitivity - 80 %
- Specificity - 80%
- Precision- 71%
- Recall- 80%

Conclusion

- Google and Direct traffic showed the most conversion rate in Lead source.
- Most of the conversion happened for customers who are unemployed.
- People who are in management specialization and from Mumbai location have more conversion lead rate.
- The conversion rates are high for Total Visits, Total Time Spent on Website and Page Views Per Visit.
- The Model shows highest accuracy rate around 80%.
- The Threshold has been selected from the accuracy, sensitivity, specificity measures the precision and recall curve.
- Lead score shows the conversion rate on the final predicted model is around 80% in both train and test set.
- Overall this model proves to be accurate.

Recommendation

- Company should focus on Lead score which are greater than 80% to expedite the conversion rate.
- Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.
- Area to be focus on – Increasing the conversion rates for the categories generating more leads.Generating more leads for categories having high conversion rates.
- Based on various business needs, modify the probability threshold value for identifying potential leads.