Lead Score Case Study

Submitted By:

Satheesh Gopalan
Shubhanshi Jain

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. 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 Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you 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%.

Strategy

- Data Sourcing for analysis.
- Cleaning and preparing the data
- Exploratory Data Analysis.
- Feature Scaling.
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and using RFE for elimination of columns with higher VIF value and calculating the Lead Score to understand Hot leads for better conversion.
- Evaluating the model by using different metrics like Accuracy Score, Specificity ,Sensitivity , Precision and Recall.
- Applying the best model on Test data based on the Sensitivity and Specificity Metrics.

Problem solving methodology

Data Sourcing , Cleaning and Preparation

1. Reading the Data from Source.
2. Converting data into clean format suitable for analysis 3. Removing duplicate data .
4. Outlier Treatment for imputation
5.Exploratory Data Analysis
6. Feature Standardization

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

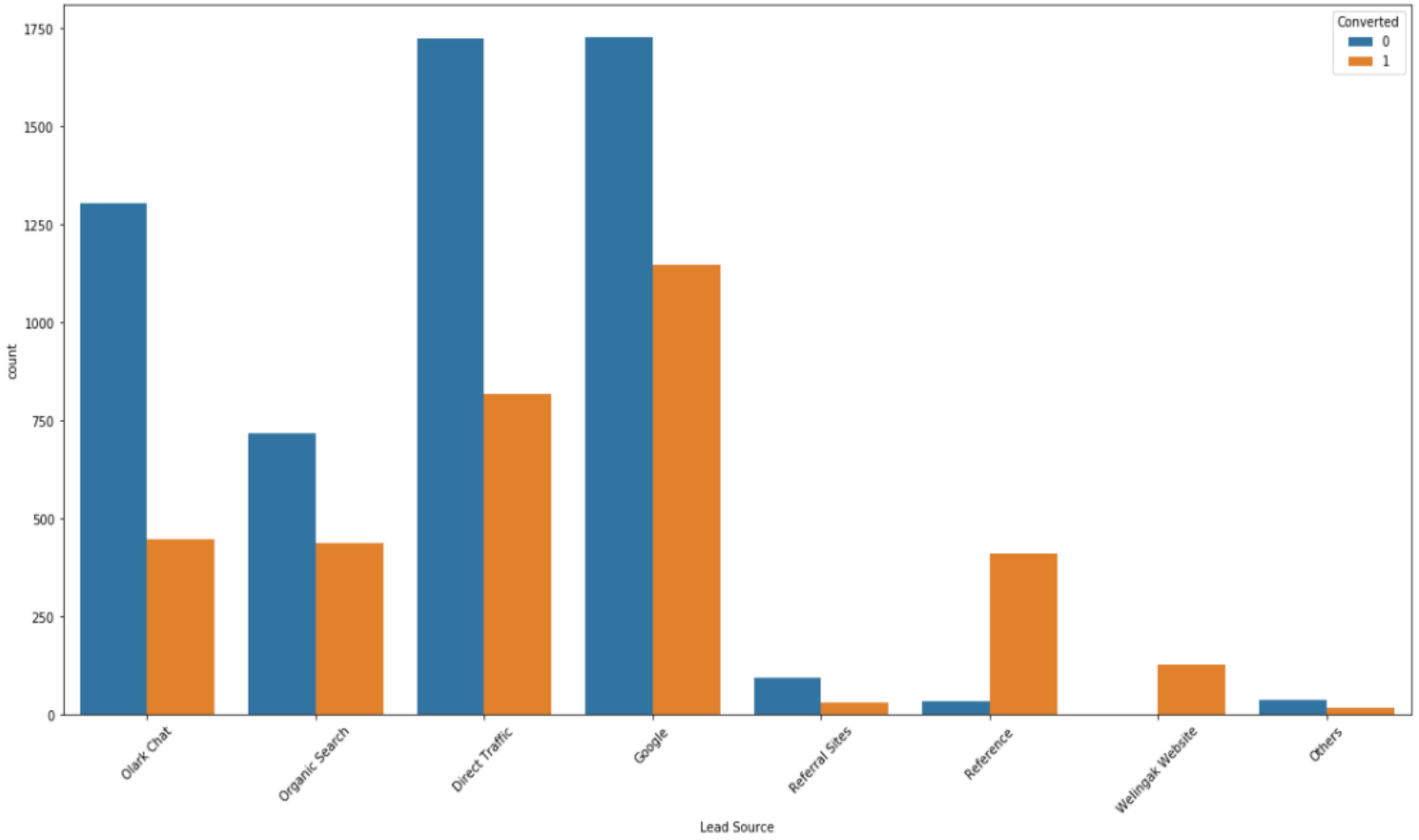
Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

- Determination of the lead score and checking if the target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis

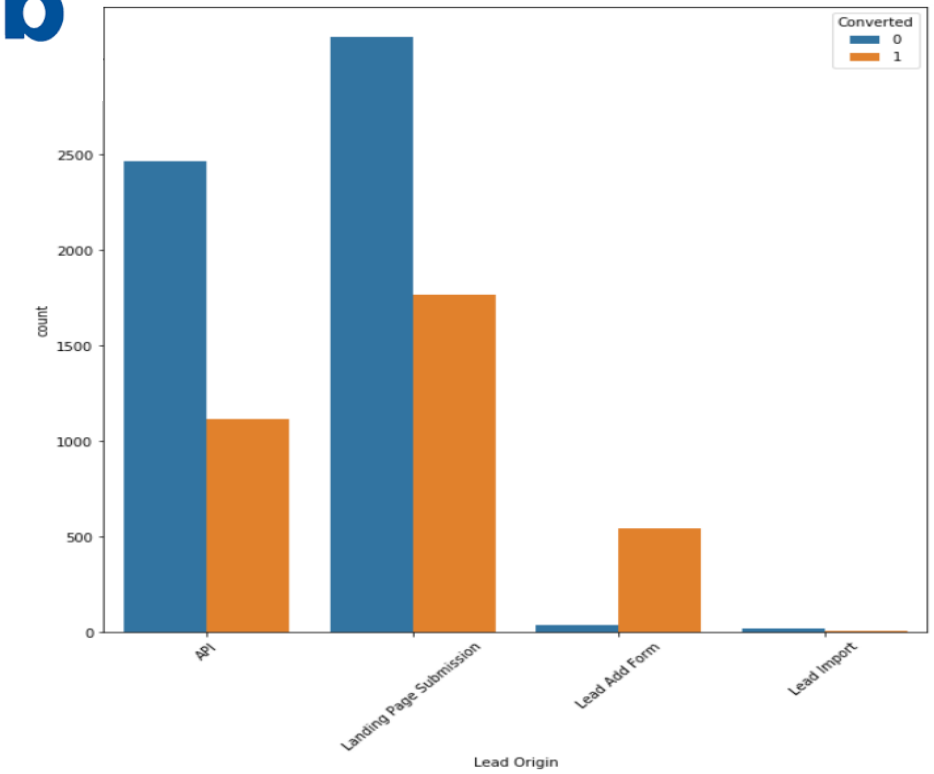


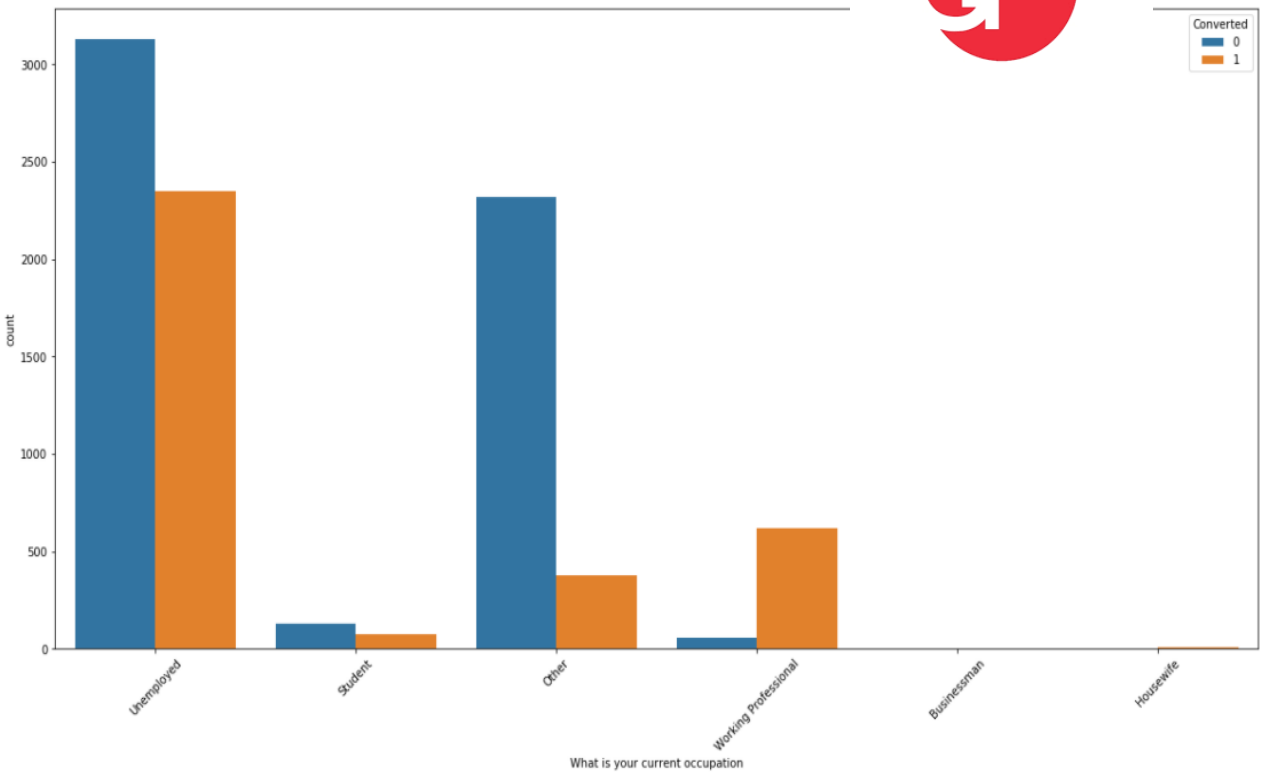
Remarks -

Lead Source Vs Converted :

1. Leads from Google & Direct Traffic are the highest.

2. Leads from References & Welling Ak Website are low but conversion rate is high.



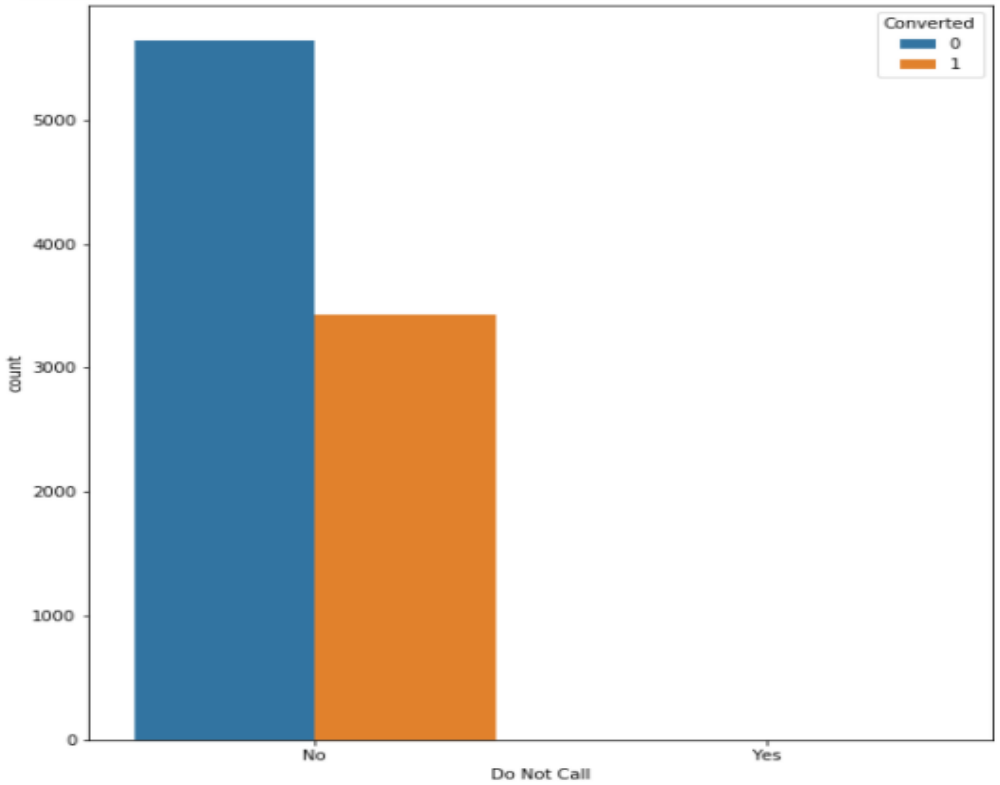


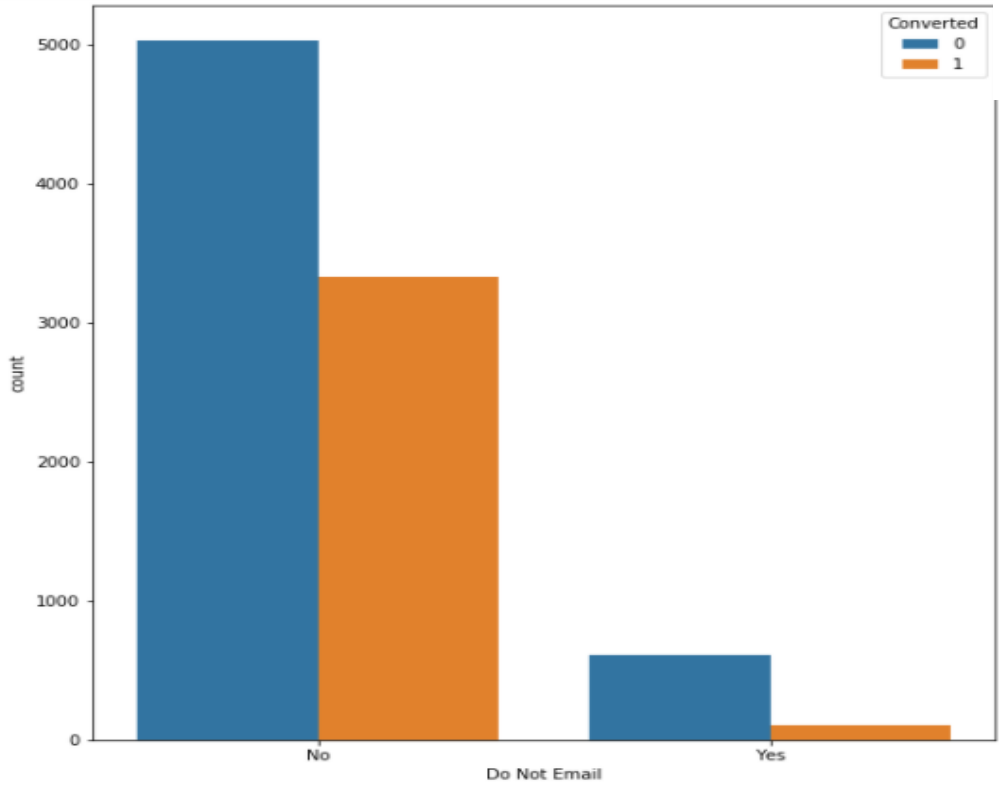
Remarks

1) Landing page Submission & API generate most of the leads but only has roughly half of the leads converted.
2) Leads generated from 'Lead Add Form' is less but conversion rate is high.
3) Leads generated from 'Lead Import' is very less.

Remarks

Majority Conversions happend for people who are unemployed and Working professionals.





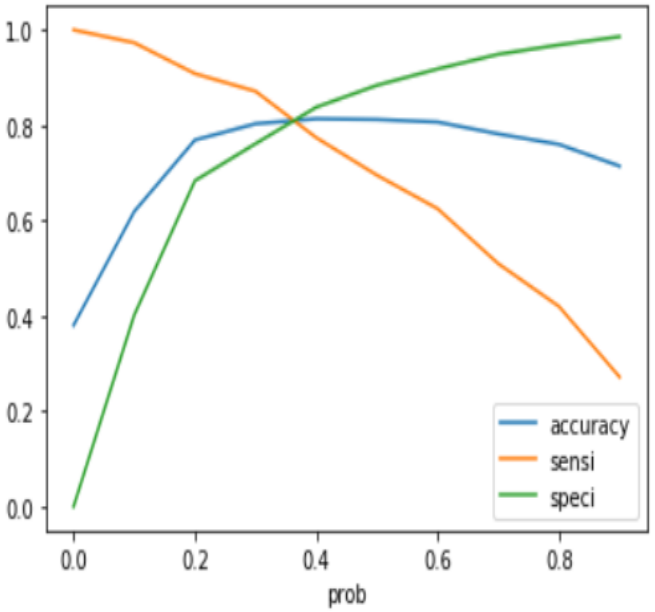
Remarks –

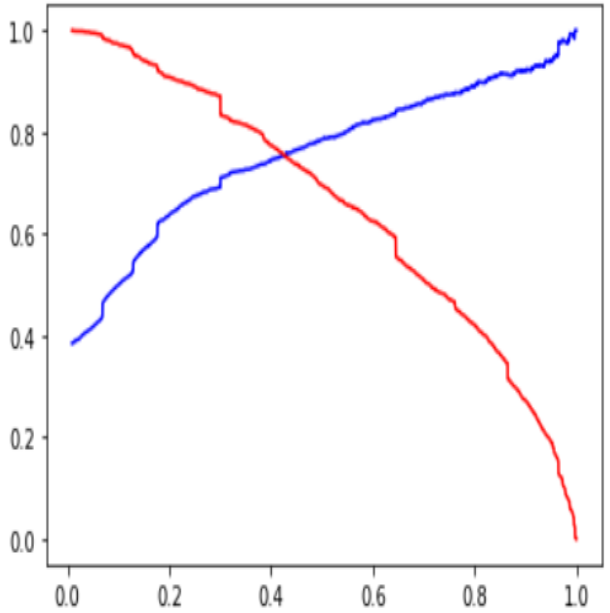
Major Conversions happened from Emails Sent and Calls made

List Of Variables impacting the conversion rate

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- Last Activity_Email Opened
- Last Activity_Others
- Last Activity_SMS Sent
- What is your current occupation_Housewife
- What is your current occupation_Other
- What is your current occupation_Working Professional
- Last Notable Activity_Others
- Last Notable Activity_SMS Sent
- Last Activity_Others

Model Evaluation - Sensitivity and Specificity on Train Data Set





3172	699
467	1908

- Accuracy Score – 81.33 %
- Sensitivity – 80.33 %
- Specificity – 81.94 %
- Precision – 73.18 %
- Recall – 80.33 %

From the curve above, 0.37 is the optimum point to take it as a cutoff probability.

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Accuracy Score – 80.20 %
- Sensitivity – 79.07 %
- Specificity – 80.87 %
- Precision – 70.93 %
- Recall – 79.07 %
- Conversion Rate of test Data – 41.37 %

1362	322
208	786

Confusion Matrix

Conclusion

- We have checked both Sensitivity-Specificity as well as Precision and Recall Metrics and have considered the optimal cut off as 0.37 based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 80.20%, 79.07% and 80.87% which are approximately closer to the respective values calculated using trained dataset.
- The top 3 variables that contribute for lead getting converted in the model are :

1. Last Activity
2. What is your current occupation
3. Total Time Spent on Website

Hence overall this model seems to be good.