# LEAD SCORING CASE STUDY

BY

Ameya Bataki

Sagar Chandrashekhar

# Problem statement

This case study is done for the Education X company which sells online courses to industry professionals. Education X markets their courses on several websites. Candidates land on there website, they brows website or watch some videos or fill the form on website. Once they fill the form and providing there email address or mobile number they are considered as lead. They also gets leads through referrals.
Then Employees from sales team starts making calls to these lead out of which some converts to join any course and most do not. There current lead conversion rate is 30%.

# Goal

Past data of leads is present. Using this data we have to prepare logistic regression model to predict hot leads and cold leads. Hot leads are one having high probability of conversion and cold lead with least probability.
We have to find driving factors for conversion and give 0-100 score each lead, highest score means high probability of conversion while low score means low chance of conversion.

# Approach for the case study

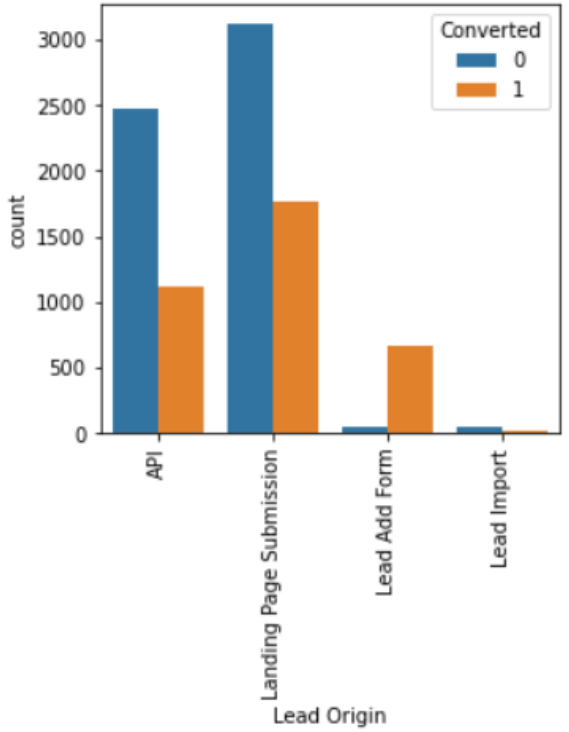
- Step1: Import all the required libraries
- Step2: load the dataset and do all the data cleaning processes

  (Outlier treatment, missing value imputation)
- Step3: Transform Categorical Variables (Creating dummy variables) and perform EDA
- Step4: Splitting test and train dataset
- Step5: Transform Numerical (Continuous) Variables : Scaling
- Step6: Using RFE to select top 15 variables for model building
- Step7: Creating and fitting logistic regression model
- Step8: Evaluating Model (P values, VIF, Sensitivity, Specificity, Accuracy)
- Step9: Making Prediction on Test data set

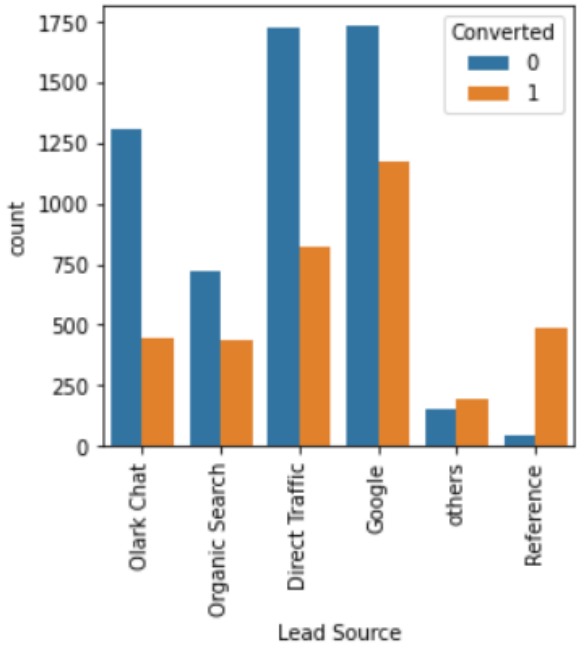
# EDA

- While exploring through data we found that many categorical variables are affecting target variable ‘Converted’
- There were many categorical columns which had very less count of some categories in it. So we clubbed them together as ‘other’ category
- There were many columns with missing values and category as ‘Selected’ which is equivalent to null value. So we removed columns with more than 45% missing values.
- There were outliers in numerical variables. We capped them using IQR method.
- We found many insights from the plots of variable we plotted. Some useful insights are shown on later slides with graphs
- In all below graphs 0 means not converted and 1 means converted.

- Lead Origin and Lead Source Count with respect to target variable

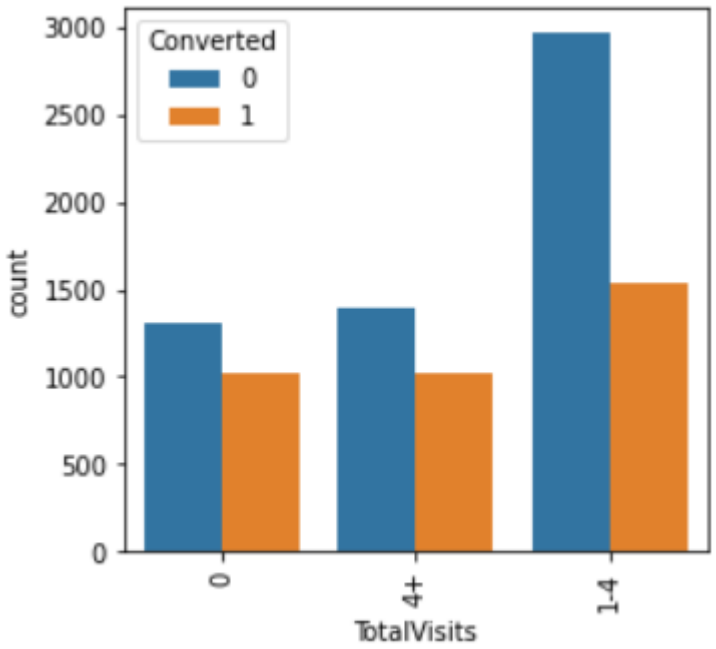



Lead origin from Lead Add Form shows high conversion rate

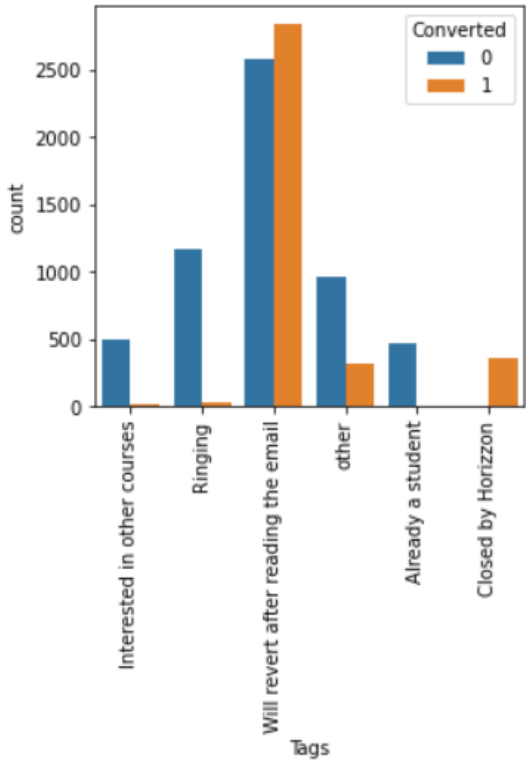



Leads coming through reference shows high conversion rate

- Total visits and Tags with respect to target variable

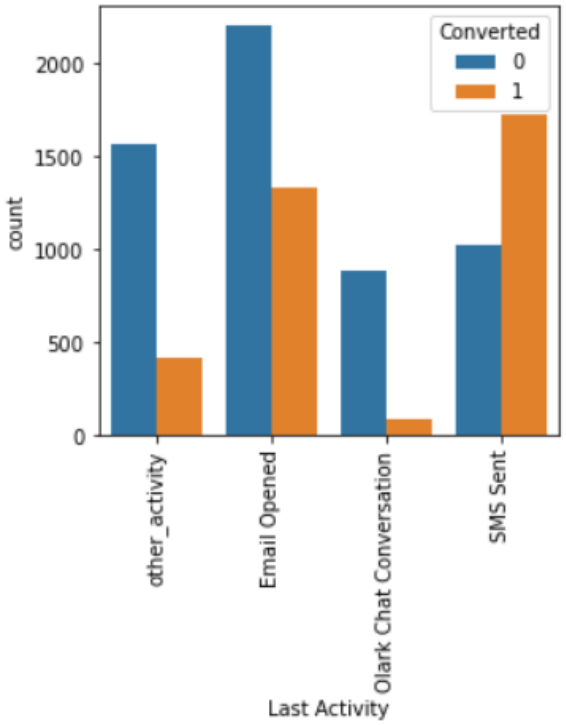


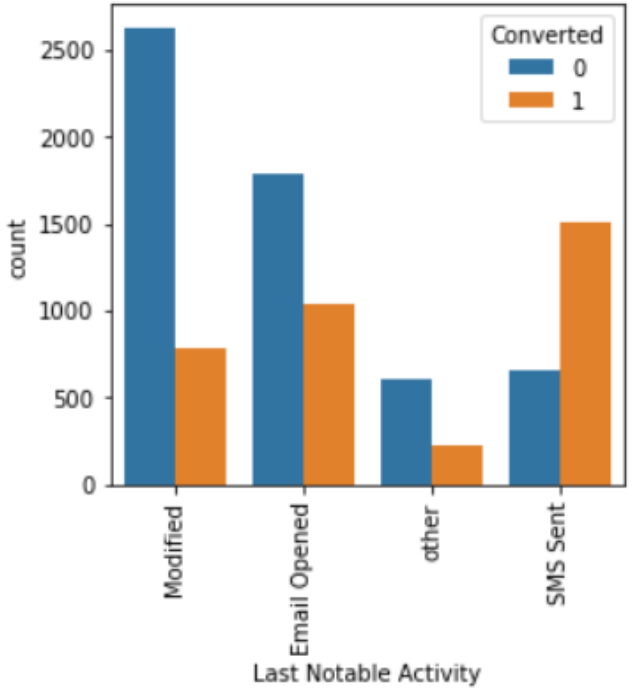

Total 4+ visits on page shows high conversion rate.




Closed by horizon tag shows highest conversion rate, while ‘will revert after reading the email’ also has high conversion rate

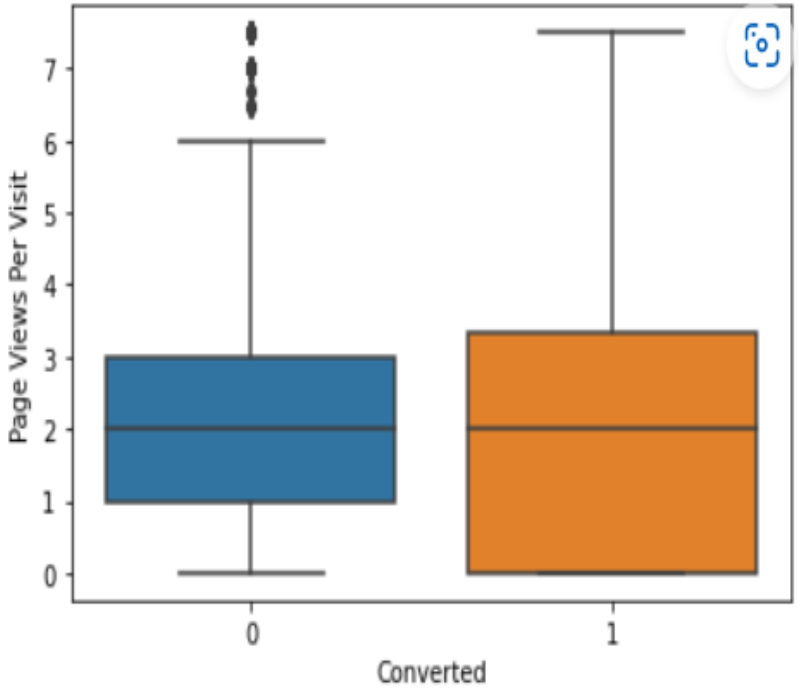
- Last Activity and Last notable activity with respect to target variable



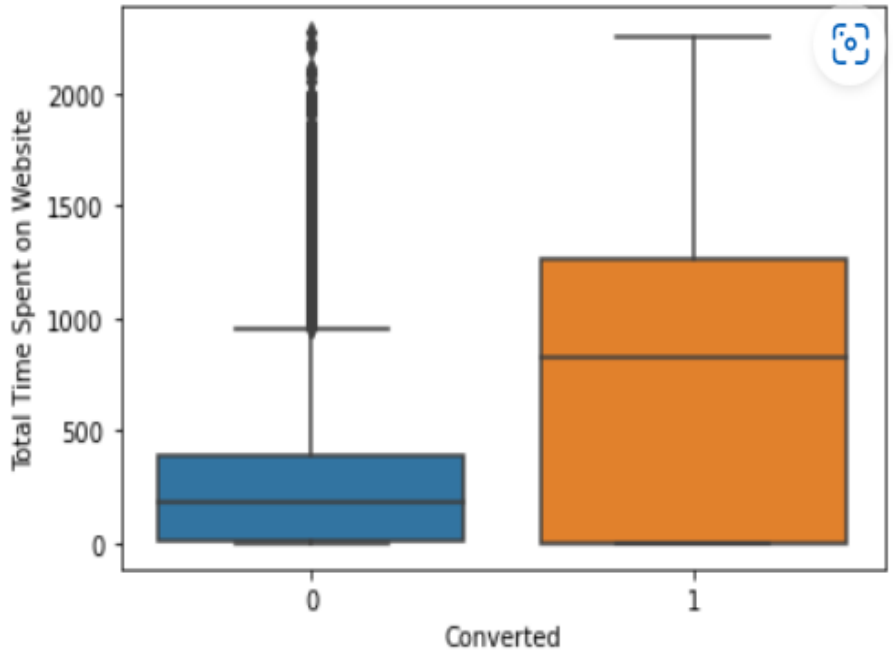




SMS sent is having high conversion rate in both Last activity and last notable activity

- Page views per visit and Total time spent on website with respect to target variable



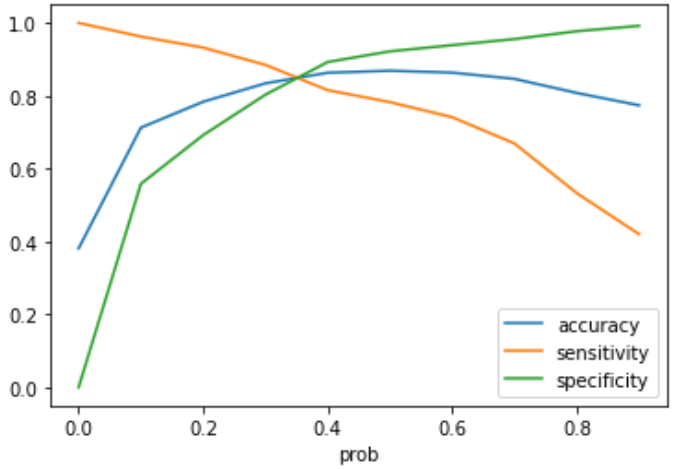

Page views per visit don’t have much relation with conversion as the median of both converted and non converted are near about same

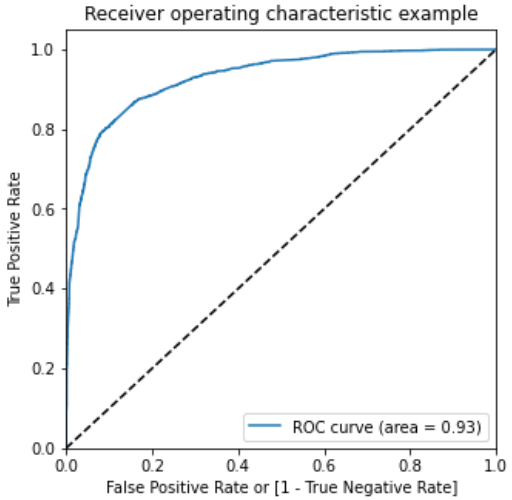



People who spend more time on website turns to convert more, we can see the median high for converted.

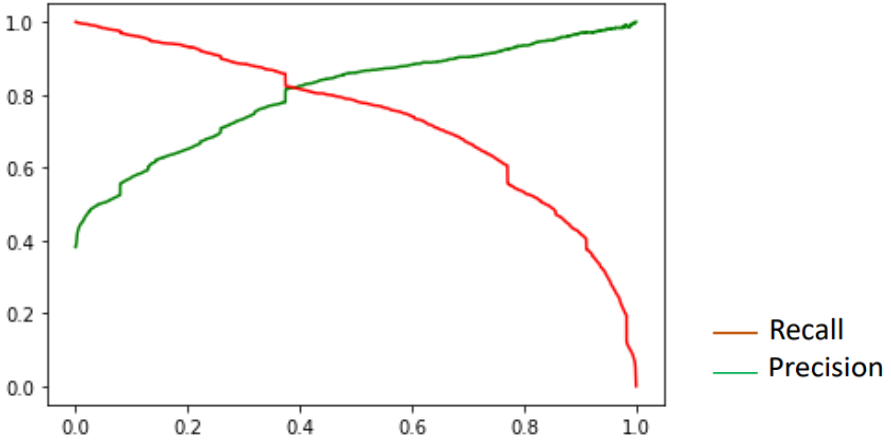
# Model Characteristics

**Evaluation on Train data:**
Accuracy : 85%
Sensitivity: 87%
Specificity: 84%
ROC curve area: 0.93




**Evaluation on Test data:**
Accuracy : 84%
Sensitivity: 87%
Specificity: 83%




**Conclusion:**
Model we build is good because it has shown high accuracy, sensitivity and specificity on both train and test dataset.

# Logistic Regression Model Summary

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -4.1944 | 0.317 | -13.237 | 0.000 | -4.815 | -3.573 |
| Total Time Spent on Website | 4.9656 | 0.200 | 24.824 | 0.000 | 4.574 | 5.358 |
| Lead Origin_Lead Add Form | 2.8386 | 0.229 | 12.416 | 0.000 | 2.391 | 3.287 |
| TotalVisits_1-4 | -0.9704 | 0.123 | -7.900 | 0.000 | -1.211 | -0.730 |
| TotalVisits_4+ | -0.6379 | 0.130 | -4.922 | 0.000 | -0.892 | -0.384 |
| Last Activity_Olark Chat Conversation | -1.3907 | 0.175 | -7.964 | 0.000 | -1.733 | -1.048 |
| Last Activity_other_activity | -0.7392 | 0.110 | -6.693 | 0.000 | -0.956 | -0.523 |
| Tags_Closed by Horizzon | 8.8169 | 1.055 | 8.358 | 0.000 | 6.749 | 10.884 |
| Tags_Ringing | -1.1235 | 0.378 | -2.976 | 0.003 | -1.864 | -0.384 |
| Tags_Will revert after reading the email | 3.6815 | 0.305 | 12.055 | 0.000 | 3.083 | 4.280 |
| Tags_other | 2.3176 | 0.317 | 7.308 | 0.000 | 1.696 | 2.939 |
| Last Notable Activity_Modified | -0.5395 | 0.097 | -5.577 | 0.000 | -0.729 | -0.350 |
| Last Notable Activity_SMS Sent | 1.7238 | 0.105 | 16.416 | 0.000 | 1.518 | 1.930 |

- Factors which converts leads into hot leads have positive coefficient in the model as we can see on summary. Like, Tags_Closed by Horizon, Total time spent on website etc. higher the correlation higher the chances of conversion.

- Variables with negative coefficient makes leads to not convert. variables such as Tags_Ringing, Last notable activity_modified etc.

# Finale Recommendations

Following are the top variables which are responsible for converting leads successfully. Sales team should focus on these variables while making calls to make efficient calling.
Also company should avoid wasting resources on these factors as they are performing good.

a) Tags_Closed by horizon: with positive coefficient of 8.81 these leads are mostly likely to convert.
b) Total Time Spent on Website: with positive coefficient of 4.96, leads who spend more time on website are more likely to convert.
c) Tags_Will revert after reading the email: with positive coefficient of 3.68 these leads are also likely to convert. Company should contact leads who have reverted to email.
d) Lead Origin_Lead add form: with coefficient 2.83 these are likely to convert

Following are the top variables which are responsible for not converting leads successfully. Sales team should not focus on these variables that much while making calls .
Also company can focus on these variables to improve the reaction of candidates.

a) Last Activity_Olark chat conversation: with coefficient of -1.39, we can say that students having questions goes to olark chat to seek answers and they are not getting it so company can focus on this point.
b) Tags_Ringing: with coefficient -1.12, to make efficient calling team should avoid calling to these leads
c) Total visits_1-4: with coefficient -0.97
d) Last Activity_other activity: With coefficient -0.73