LEAD SCORING CASE STUDY

Prasasti Choudhury, Satadhriti Chakrabarty

DS C17 Group 2 February 2020

PROBLEM STATEMENT

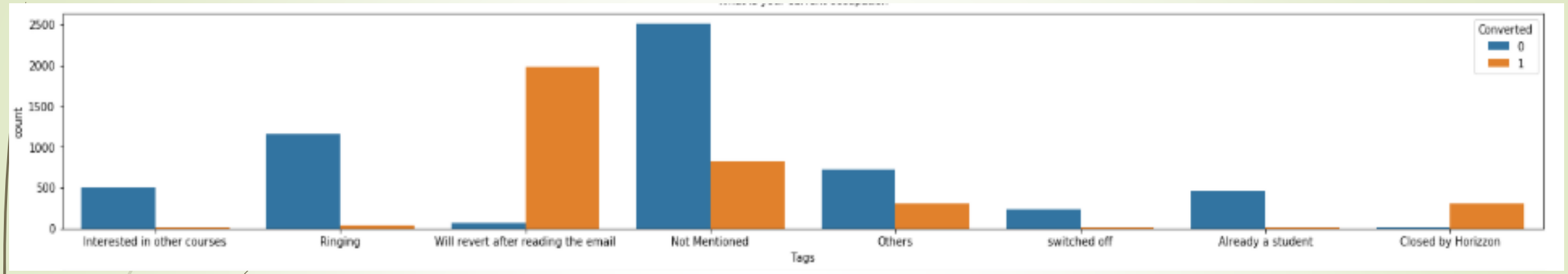
- Our company, X Education markets its online courses on several websites and search engines like Google.
- People are termed as a lead when they fill up a form providing their email address and phone numbers.
- The Sales team then contacts them via phone calls and mails and convince them to take up an online course through our platform.
- The typical conversion rate is only 30%.
- Now, our company wants to identify its potential customers so that the Lead conversion rate gets higher and effort and the cost that goes behind every lead gets optimized.
- We want to focus on our Hot Leads by engaging them in more conversations and offering discounts and explaining the benefits in detail.
- We need to generate a score which will help us identify these Hot Leads and target them aggressively instead of focusing on a large cohort.
- Our time and effort get valued in this way and remove the casual customers from our target. **Our objective is to build a robust model generating a Lead Score for every customer and providing the Sales team with a list of the Hot Leads.**

ANALYSIS APPROACH

- We perform some basic data cleaning steps like missing value imputation and grouping of fields with low counts in a particular variable.
- We drop some variables based on their skewness towards a category and then treat for outliers.
- We perform some EDA on the categorical and continuous variables and recommend some measures to try and increase the conversion rates based on those.
- We build a logistic regression model with 15 selected variables from RFE and remove some of them from the model based on their p-values and VIF.
- We get a model with 12 feature variables explaining our target variable, 'Converted'.
- As the next most important step of the analysis, we assign a Lead Score to each lead using the conversion probability.
- We check for the evaluation metrics like sensitivity, specificity, precision after we select an optimal cutoff point using ROC and then validate our results on the test set.
- The training and the test sets show results on similar lines and our model will serve as a key guide to X Education and their business goals.

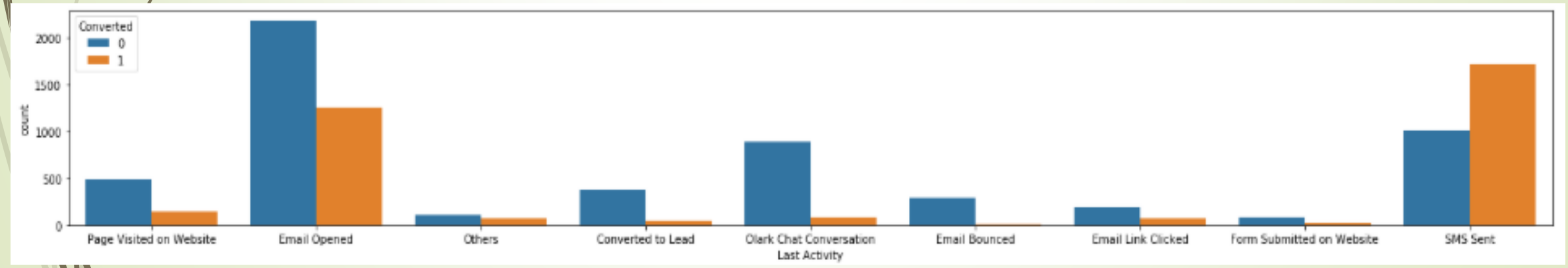
EXPLORATORY DATA ANALYSIS

Bar Plot for the "Tags" Categorical variable



From the plot it can be seen that when people say that they 'Will revert after reading the email', the conversion chance is generally higher

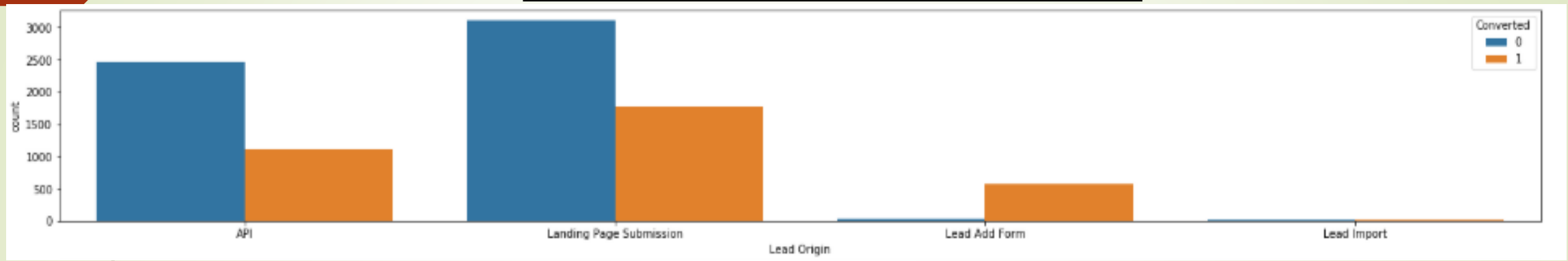
Bar Plot for the "Last Activity" Categorical variable



From the plot it can be seen that 'Email Opened' has the highest number of leads while 'SMS Sent' has the highest conversion as well as a good number of leads. We must definitely maintain this and engage more in text messages.

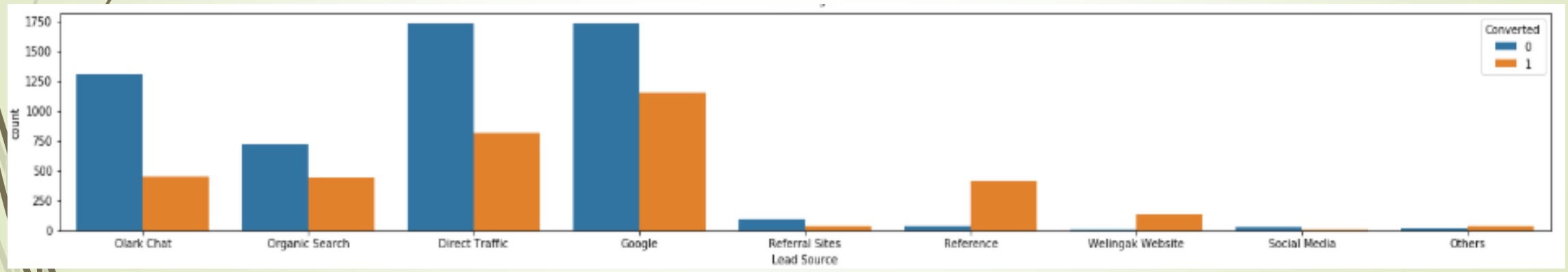
EXPLORATORY DATA ANALYSIS

Bar Plot for the "Lead Origin" Categorical variable



From the ploy it can be seen that 'API' and 'Landing Page Submission' have higher number of leads but the conversion rate is not good. In order to increase the conversion rate, we need to do something to increase conversion on these lines. 'Lead Add Form' has a very good conversion rate despite bringing in fewer leads. We must try to increase leads for this.

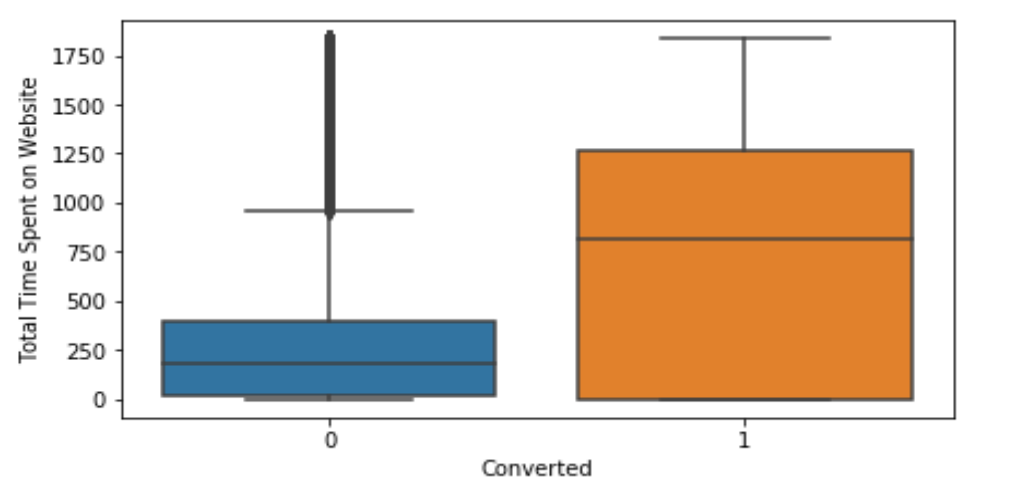
Bar Plot for the "Lead Source" Categorical variable



From the plot it can be seen that Performance wise, 'Google' does the best though the conversion rate is not satisfactory followed by 'Direct Traffic','Olark Chat' and 'Organic Search'. 'Reference' has a very high conversion rate though the number of leads generated through it is pretty low.

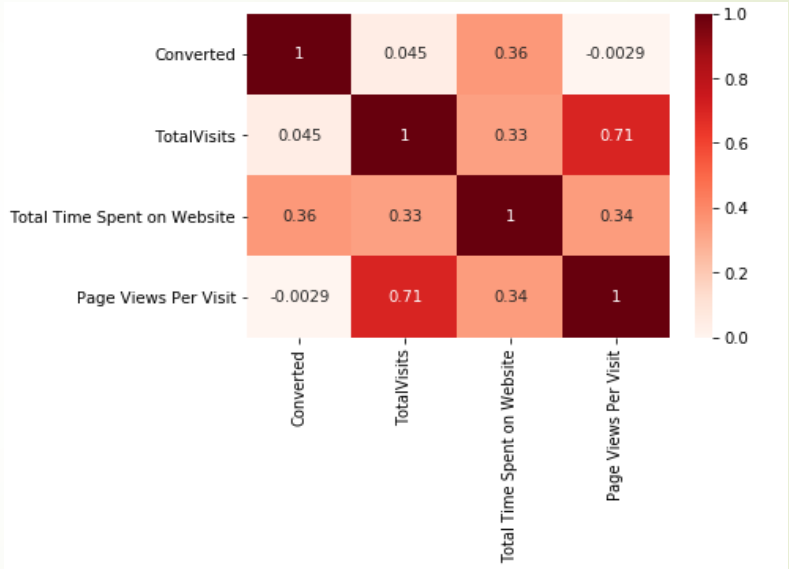
EXPLORATORY DATA ANALYSIS

Box plot for "Total time Spent on Website" Continuous variable



Clearly, people spending more time on Website have a higher tendency to get converted

Correlation matrix



From the heatmap, we can validate our statement that the target variable 'Converted' has the highest correlation with 'Total Time Spent on Website'

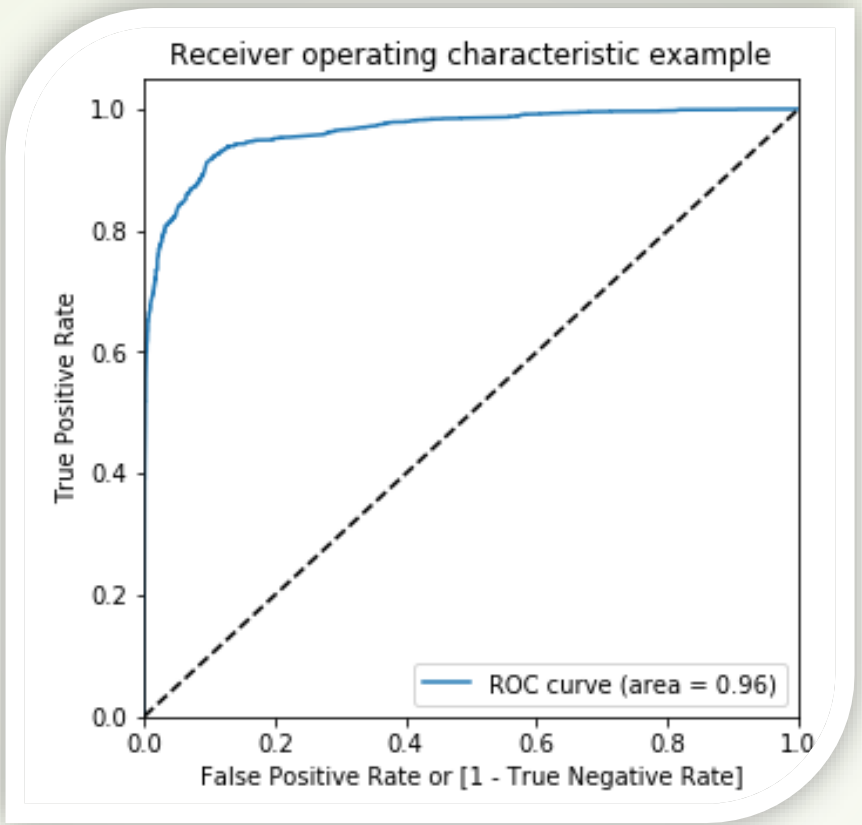
FINAL REGRESSION MODEL

- Final Equation from *Logistic Regression*:

Converted = -5.03 - 1.43 **Do Not Email** + 1.06 **Total Time Spent on Website** + 2.53 **Lead Origin_Lead Add Form** + 1.74 **Lead Source_Olark Chat** + 3.17 **Lead Source_Welingak Website** - 1.35 **Last Activity_Olark Chat Conversation** -1.24 **Specialization_Not Mentioned** + 9.02 **Tags_Closed by Horizzon** + 3.47 **Tags_Not Mentioned** + 3.47 **Tags_Others** + 7.5 **Tags_Will revert after reading the email** + 2.01 **Last Notable Activity_SMS Sent**

- Increasing certain features of '**Tags**' have a very **positive impact** on conversion of Leads. People who **spend more time on website**, **Lead Add Form** as the **Lead Origin**, **Olark Chat** and **Welingak Website** as the **Lead Sources**, **SMS sent** as the **Last Notable Activity** also **increase** the probability of Lead conversion. People having **Olark Chat Conversion** as the **Last Activity**, not mentioning **Specialization** and **not opting for Email** reduce the probability of Lead conversion.

ROC CURVE



- The ROC Curve shows us tradeoff between sensitivity and specificity.
- The Area Under ROC Curve or AUROC is 0.96, which is pretty high and bears testimony to the high accuracy and predictive power of our model.

EVALUATION METRICS

METRICS	TRAIN DATA	TEST DATA
Accuracy	90.4%	91.21%
Sensitivity	92%	92.13%
Specificity	89.4%	90.64%
Precision	84.2%	85.87%

- We have evaluated our model's prediction power on several metrics and all of them have come up with very good scores. Moreover, these metrics show us a similar results on both the Train and Test Data, implying the robustness and high predictive power of our model.

FINAL RESULT

RESULT STRUCTURE

Prospect ID	Converted	Conversion_Prob	Final_Predicted	Lead_Score
3504	0	0.005737	0	1
4050	1	0.997692	1	100
7201	0	0.367247	1	37
1196	0	0.005726	0	1
8219	1	0.115390	0	12

- We have generated **Lead Scores** for every Prospect ID, the unique identification key for a lead. We can use this Lead Score to identify a Hot Lead. A cutoff point, say around 80, should help us in the X Education to contact a Lead with the highest chances of getting converted. We can ignore the people with lower Lead Scores and save our Sales team's time and X Education's money.