Analysis & Recommendations

Lead Scoring Project

Important Variables in our ML Model

- Last Activity_Email Opened
- Last Activity_SMS Sent
- Last Activity_Olark Chat Conversation
- Tags_Will revert after reading the email
- What is your current occupation_Unemployed
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Last Notable Activity_Modified
- Total Time Spent on Website
- Lead Source_Welingak Website
- What is your current occupation_Student
- Tags_Positive other Tag
- Last Activity_Email Bounced
- Tags_Ringing

Our ML Models is a Logistic Regression model with the listed variables as the most relevant/impactful ones in the prediction of a lead being converted or not.

Actionable Variables

- **Top three categorical variables that we should focus on:**
  - **Tags**
  - **Lead Origin**
  - **Lead Source**

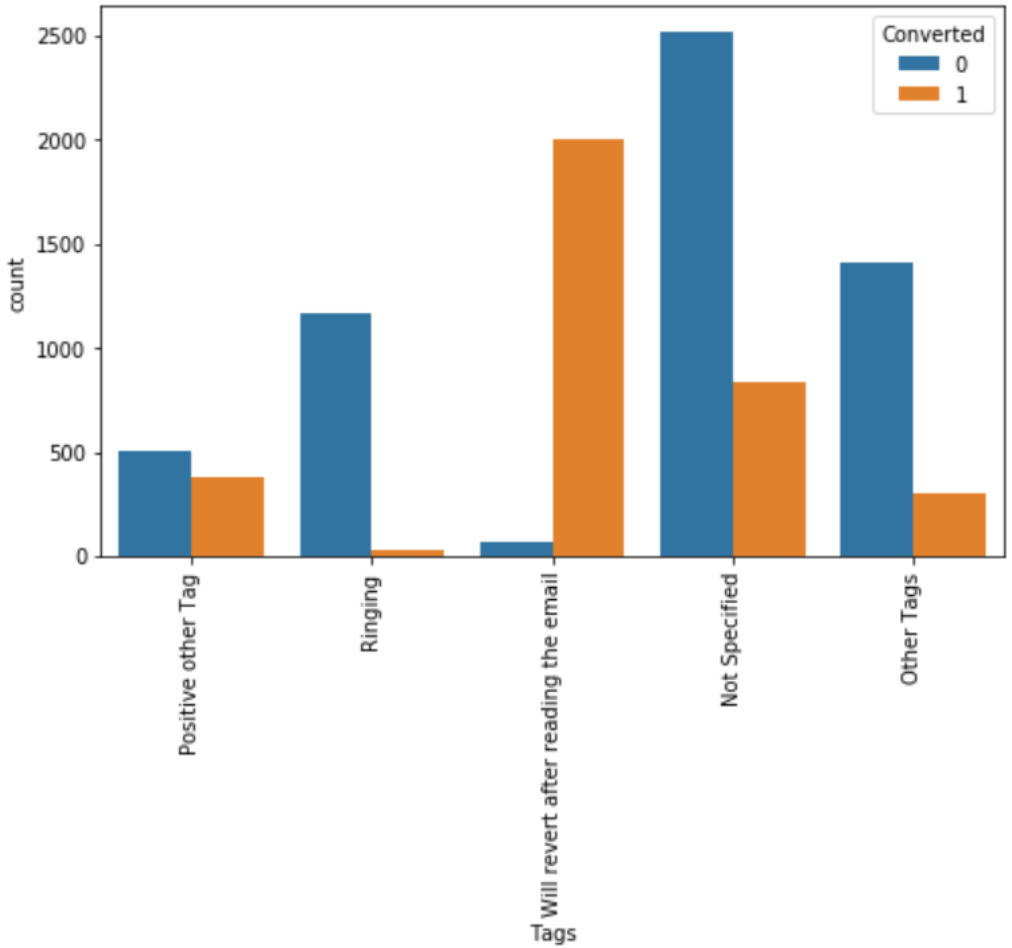
- **Numerical variable that needs to be focused**
- **Total time spent on website**

The next few slides will examine each of these actionable variables in greater detail – i.e., their analysis and **recommendations** based on that analysis.

Recommendations to Improve Conversion

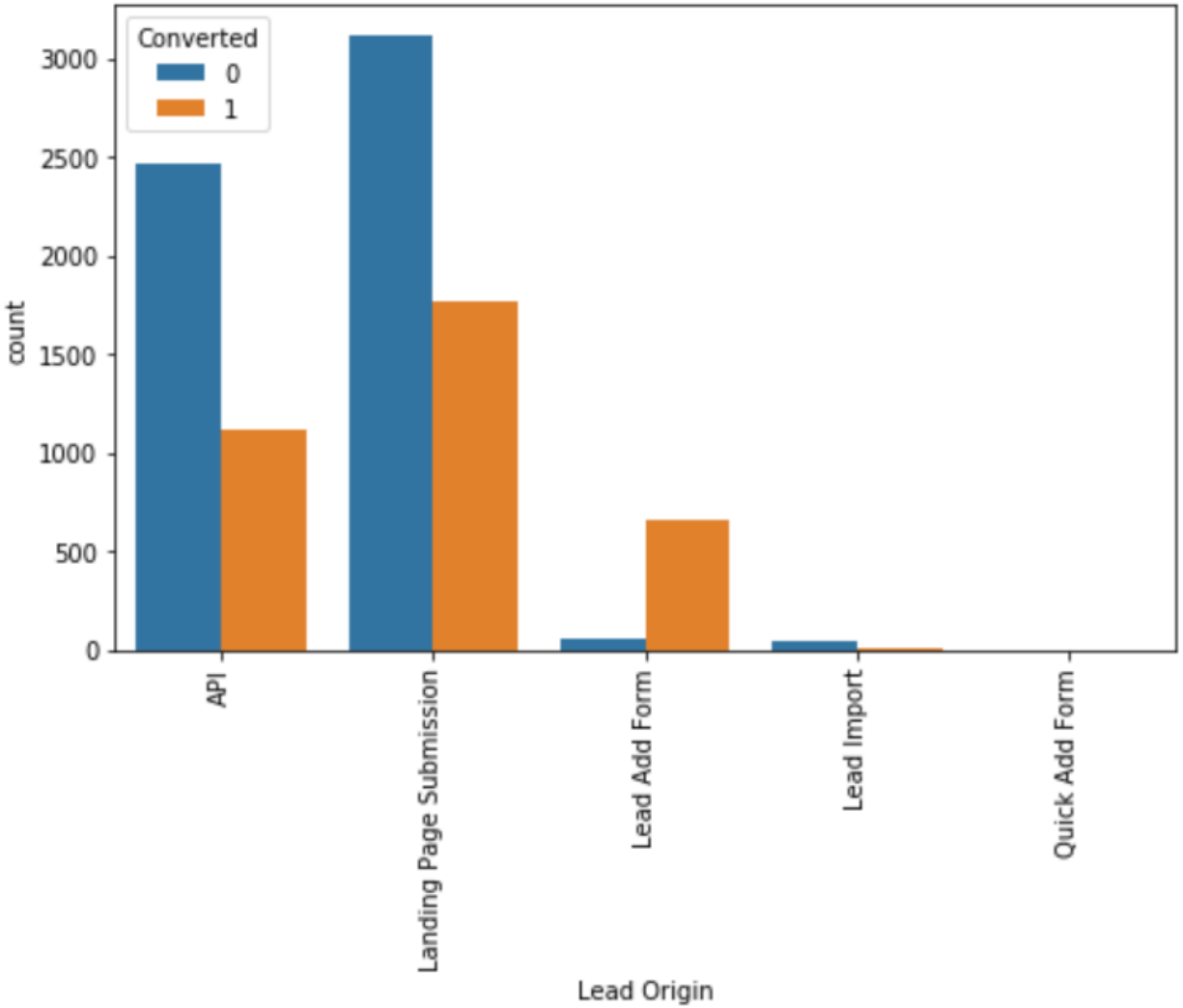
Tags

- Ringing up people does not seem to result in any significant conversions.
- However, people who say they will revert after reading our email, tend to convert the most often. Seems like sending them emails works much better than ringing them up.
- People tagged with “Positive other tags” belong to categories who are almost 50-50 in conversions. We can focus on these categories to raise conversion rates through incentives.
- People tagged with “Other tags” rarely tend to convert. Hence, we should not spend too many marketing resources on them.



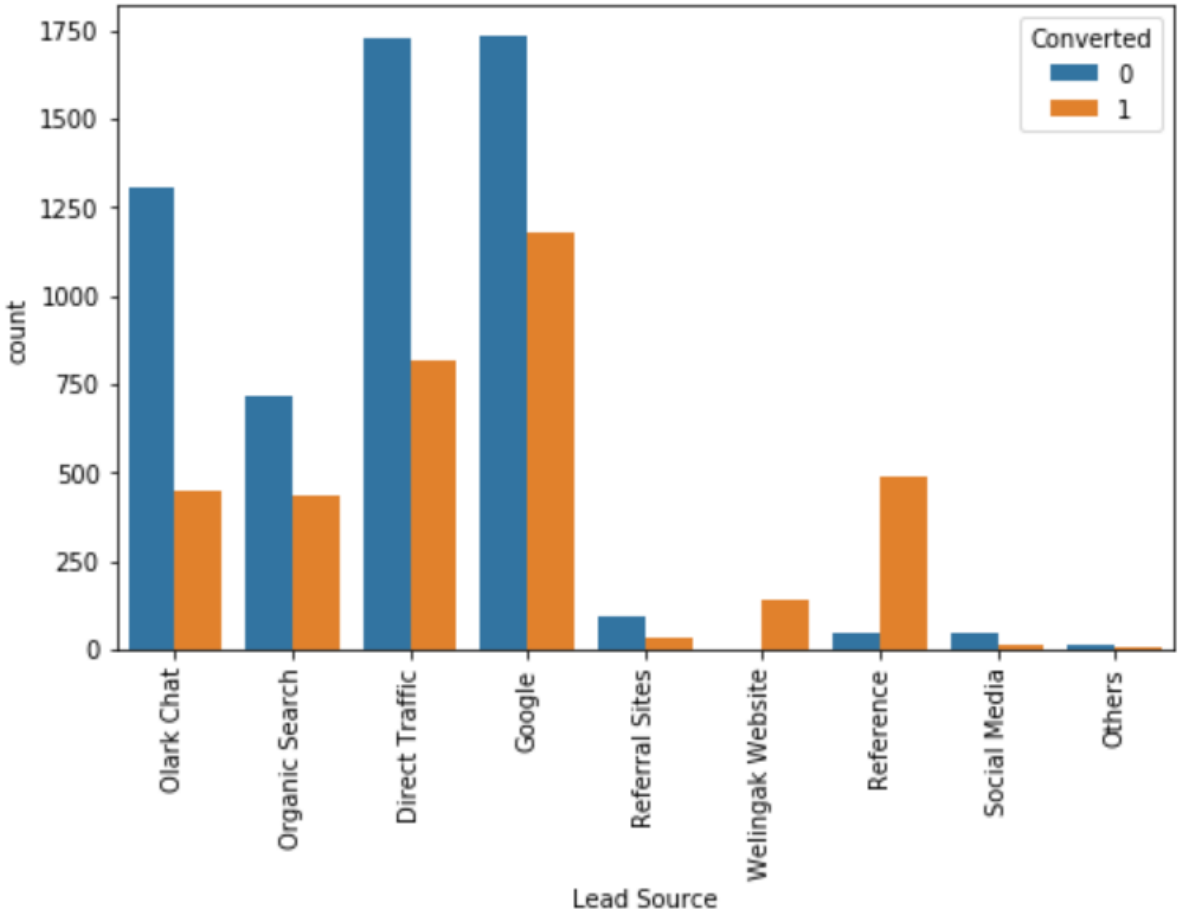
Lead Origin

- API and Landing Page submissions are the most prominent origins of Leads. However, they are currently resulting in about 30% conversion rate.
  - Hence, for these origins, we should focus on improving their conversion rate.
- The origin “Lead Add form” is almost a certain conversion.
  - Hence, for this origin, we should focus on increasing the volume of this origin.



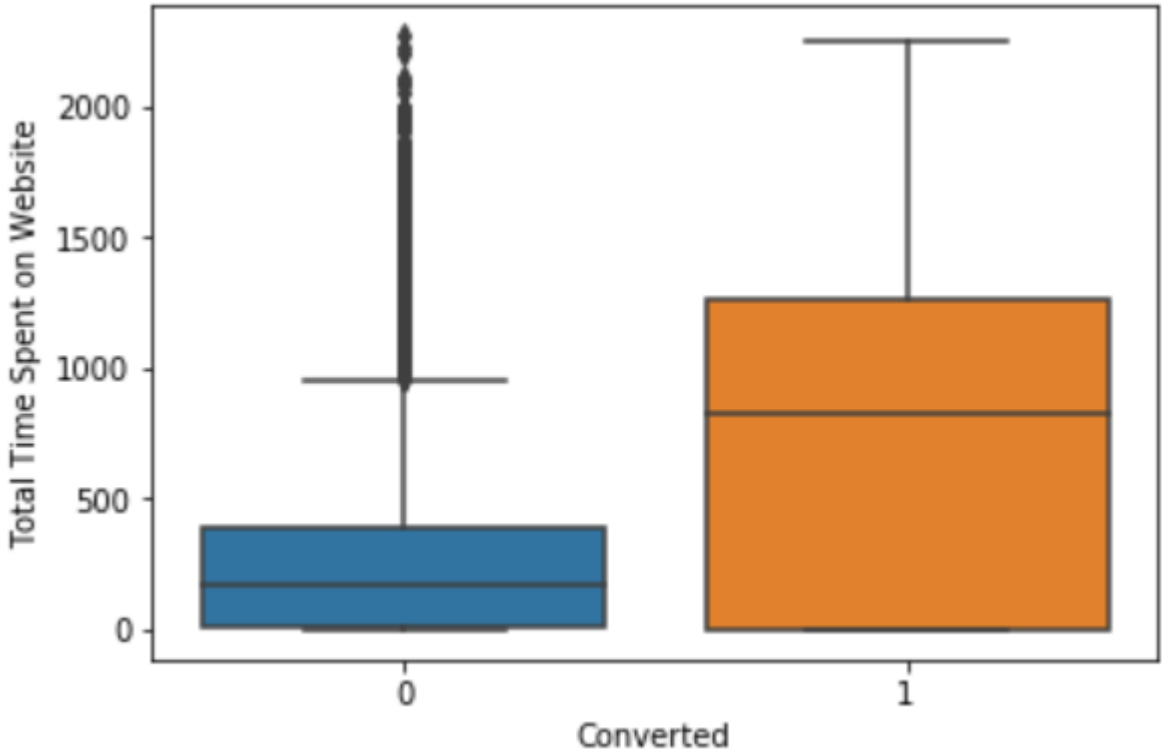
Lead Source

- **Direct Traffic, Google and Olark Chat is contributing more to get leads**
- **Conversion ratio is also good for leads in Google**
- **To improve overall lead conversion rate, focus should be on improving**
  - **lead conversion of olark chat, organic search, direct traffic, and google leads and**
  - **generate more leads from reference and welingak website.**



Total Time Spent on Website

- **Leads spending more time on the website are more likely to be converted.**
- **It's the total visit by a lead on the website. So attached plot says that the conversion ratio is good for the lead who visits the platform directly, its because the data in the website will be more related to the course.**
- **Website should be made more engaging to make leads spend more time.**



Appendix

Validity of the Logistic Regression Model (i.e., Is it a good model?)

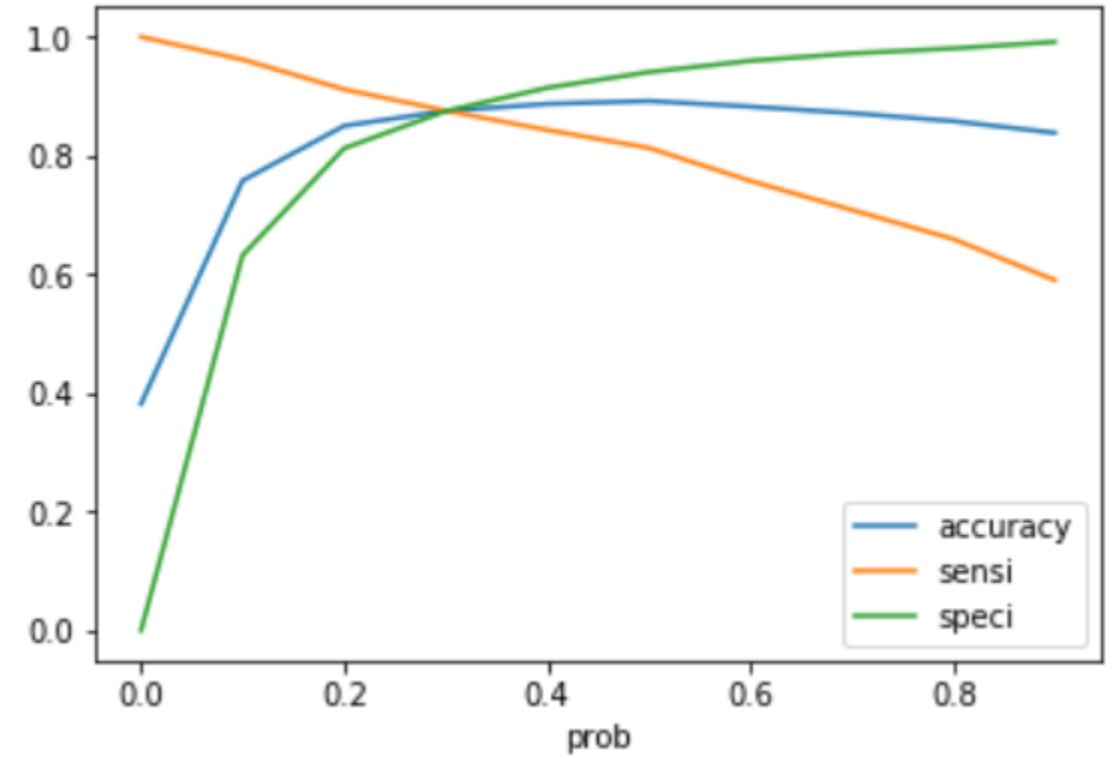
Evaluation of our Logistic Regression Model

- The evaluation can be evaluated using two alternative criteria.
    1. Specificity, Sensitivity and Accuracy of our model should be high.
    2. The ROC curve method (the area under the curve (AUC) should be high).

- Specificity refers to the ability to detect true negatives = TN / (TN + FP)
- Sensitivity (Recall) refers to the ability to detect true positives = TP / (TP + FN)
- Accuracy refers to ability to detect both true +ves & -ves = (TP + TN) / All
- Precision = TP / (TP+FP)

Selecting a good probability threshold for our logistic regression model.

- The cut-off value of the probability threshold determines which leads get categorized as “will convert” or “will not convert”.
- Different values of probability threshold will give different values of specificity, sensitivity and accuracy.
- We want all these parameters to be high … but they have opposing trends. Hence, the optimal values of probability is where these parameters intersect.
- In figure, we see that optimal value for our model is ~ 0.3.



Important Metrics Calculation of the Model

Based on selecting our probability threshold of 0.3, we get the following performance parameters:

Accuracy : **87.5 %**

Sensitivity : **87.5 %**

Specificity : **87.4 %**

Precision : **81.0 %**

Recall : **87.5 %**

As we can see with the above figures, the performance parameters are all in the high 80's (percentages). This can be considered a fairly accurate model.

ROC Curve of the Logistic Model

- It shows the tradeoff between sensitivity and specificity
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- **The above ROC curve has the area under curve (AUC) is 0.88 and the curve is towards the left which is good**