Lead Scoring Case Study

BY

VISHAL & ALEKHYA

Problem Statement

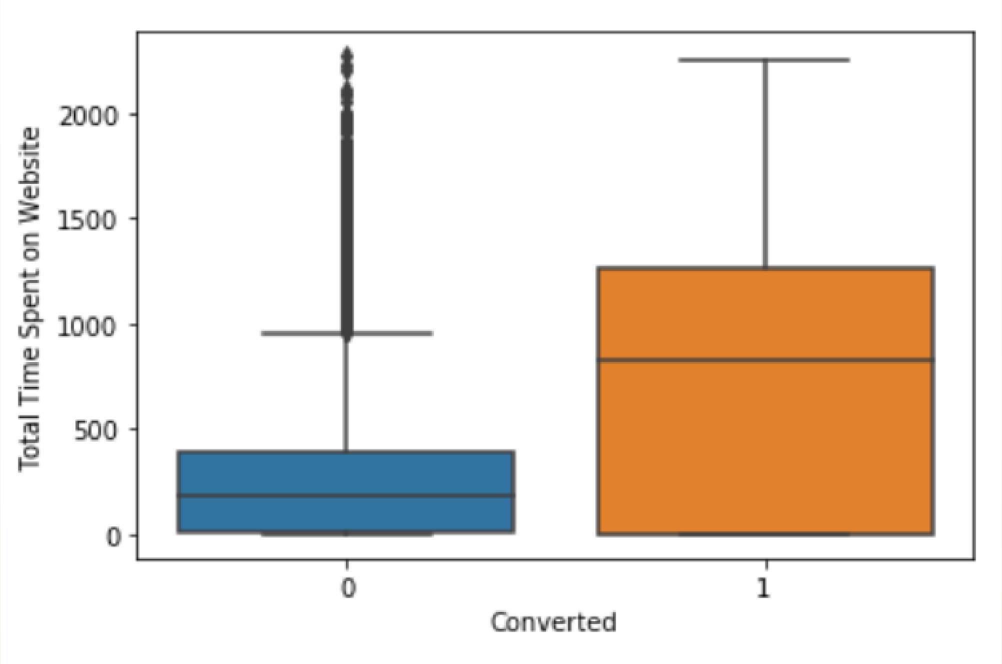
- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. On any given day, many professionals who are interested in the courses land on their website and browse for course, watch some videos or fill up a form for the course.
- The company classifies the people who filled the form providing their email address or phone number as lead. The sales team start making calls, writing emails, etc. to the leads acquired. In this process, some of the leads might get converted and the typical lead conversion rate at X education is around 30%. The company wishes to identify the most potential leads and maximize the lead conversion rate.

Analysis

- Imported and Cleaned the data set.
- Handled the missing values by dropping columns with high null percentages and imputed the columns with moderate null percentages with mode for categorical variables.
- Performed EDA using count plots, box plots etc. and dropped columns with high skewness and columns which are not so relevant for our goal.
- Created dummies for all the categorical variables.
- Performed Standard Scaling for all the numerical variables.

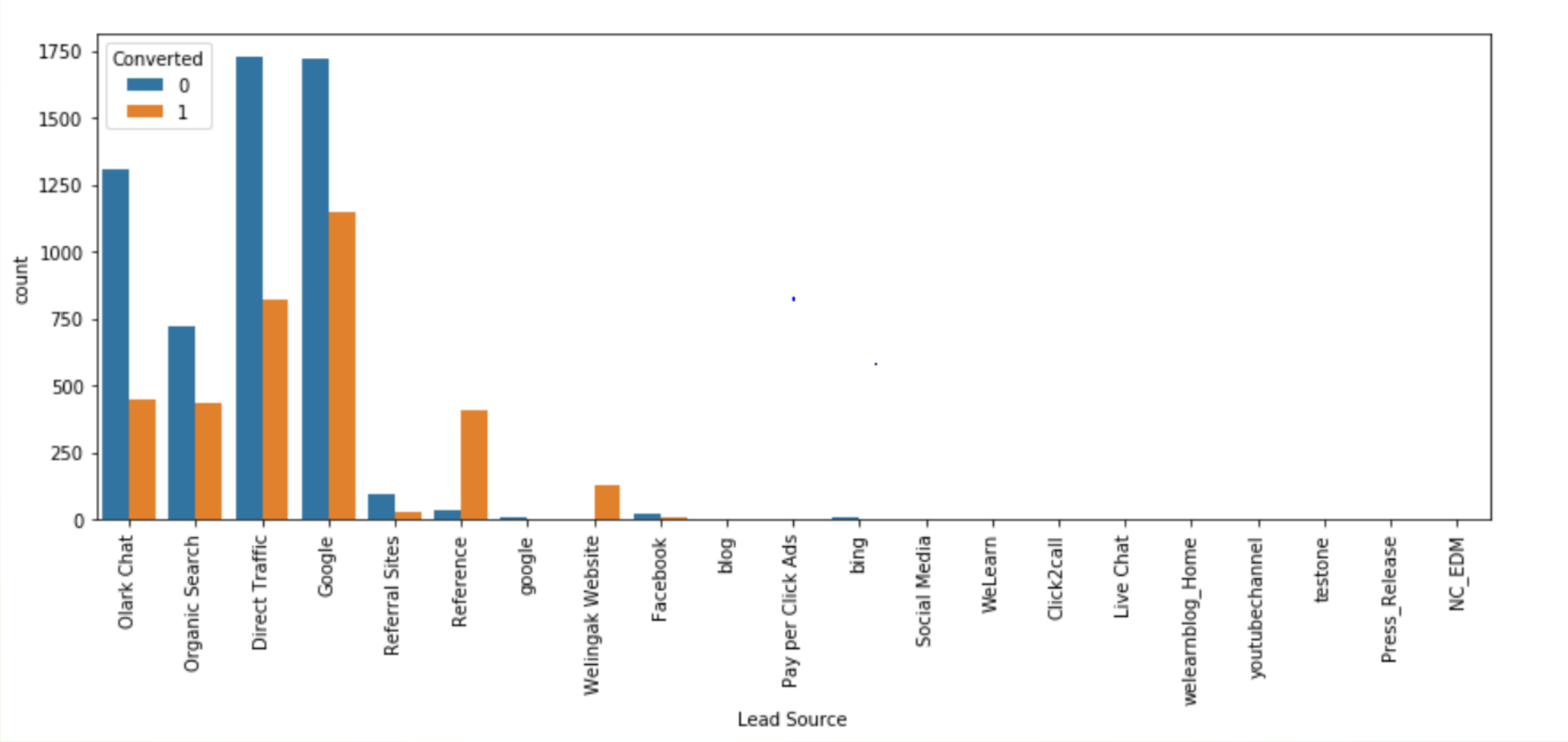
Visualizations through EDA

- Plotted a Box plot against Leads Converted vs Time spent on the websites.
- From the below plot, we could infer that spending more time on websites results in leads getting successfully converted.



- Plotted a count plot for Lead Source vs Count and had the following observations.

1. Most number of leads are from Direct Traffic and Google.
2. Welingak Website has best conversion ratio among all.

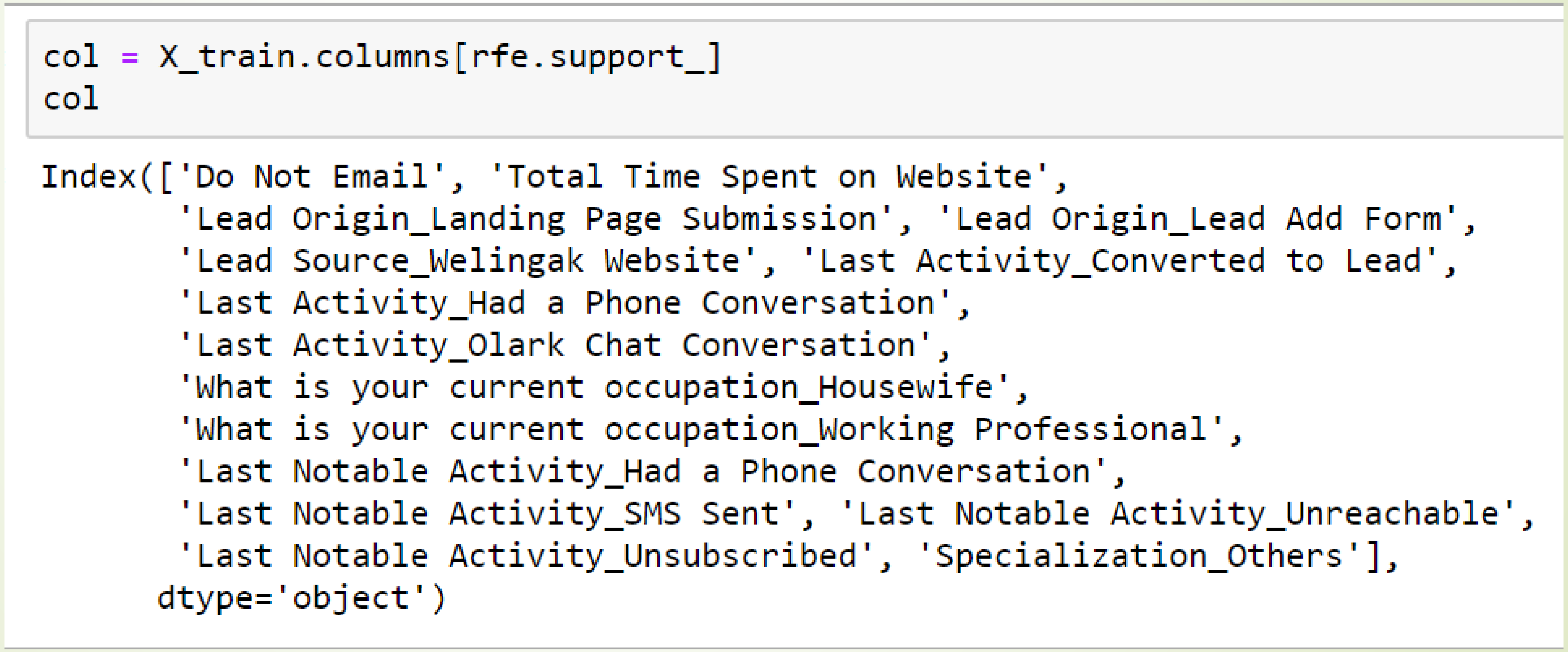


Model Building

- We split the data into train and test datasets.
- Feature Selection using RFE(selected top 15 variables)
- Stats Model (filtered to 10 variables)
- Added Converted probability column to the train data set.
- Created a Confusion Matrix with random cut off value as 0.5.
- Plotted ROC curve and found out the optimal cut-off of 0.35.
- Created another Confusion matrix with the above cut-off and then calculated Sensitivity, Specificity, Precision and Recall.
- Performed the same steps and made predictions on the test data set.

Feature Selection using RFE

- We have filtered out top 15 variables in the data set that contribute most towards the lead conversion rate using RFE method.

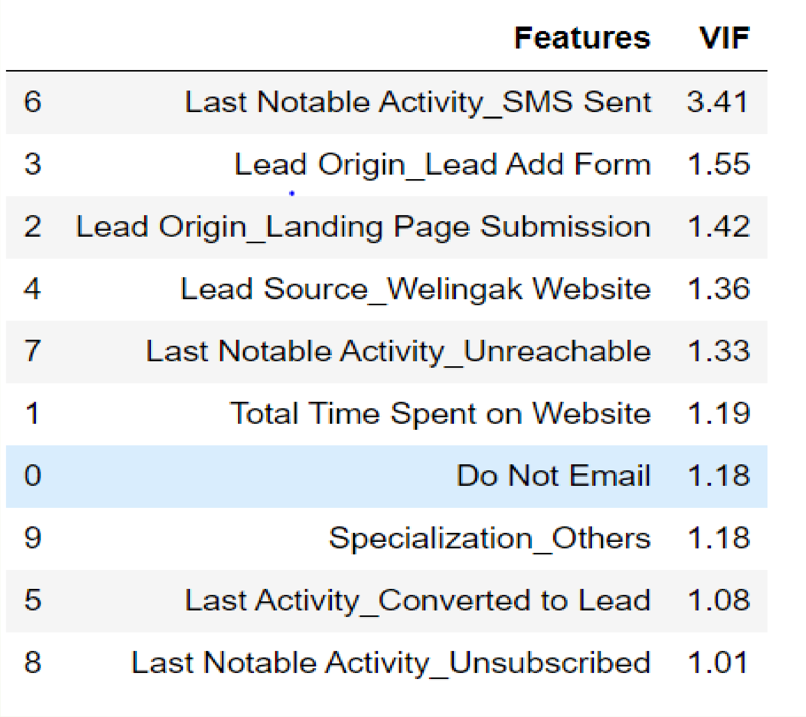


```
col = X_train.columns[rfe.support_]
col
```

```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Source_Welingak Website', 'Last Activity_Converted to Lead',
       'Last Activity_Had a Phone Conversation',
       'Last Activity_Olark Chat Conversation',
       'What is your current occupation_Housewife',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unreachable',
       'Last Notable Activity_Unsubscribed', 'Specialization_Others'],
      dtype='object')
```

Stats Model

- Using Stats Model, we have filtered the data set to 10 variables based on the
- p-values and VIFs obtained.

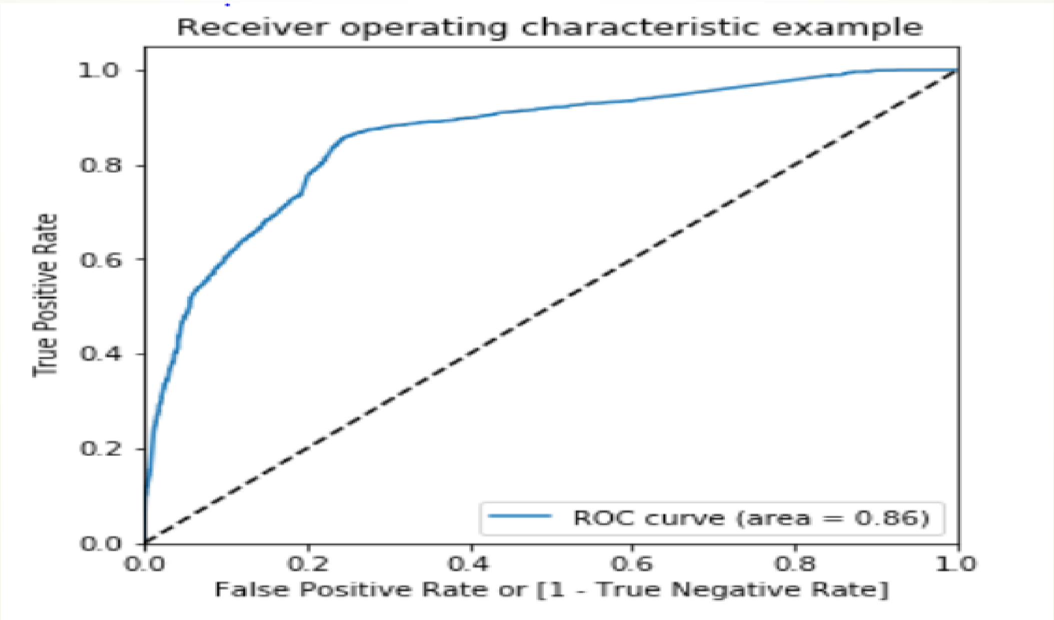


	Features	VIF
6	Last Notable Activity_SMS Sent	3.41
3	Lead Origin_Lead Add Form	1.55
2	Lead Origin_Landing Page Submission	1.42
4	Lead Source_Welingak Website	1.36
7	Last Notable Activity_Unreachable	1.33
1	Total Time Spent on Website	1.19
0	Do Not Email	1.18
9	Specialization_Others	1.18
5	Last Activity_Converted to Lead	1.08
8	Last Notable Activity_Unsubscribed	1.01

Plotting ROC Curve

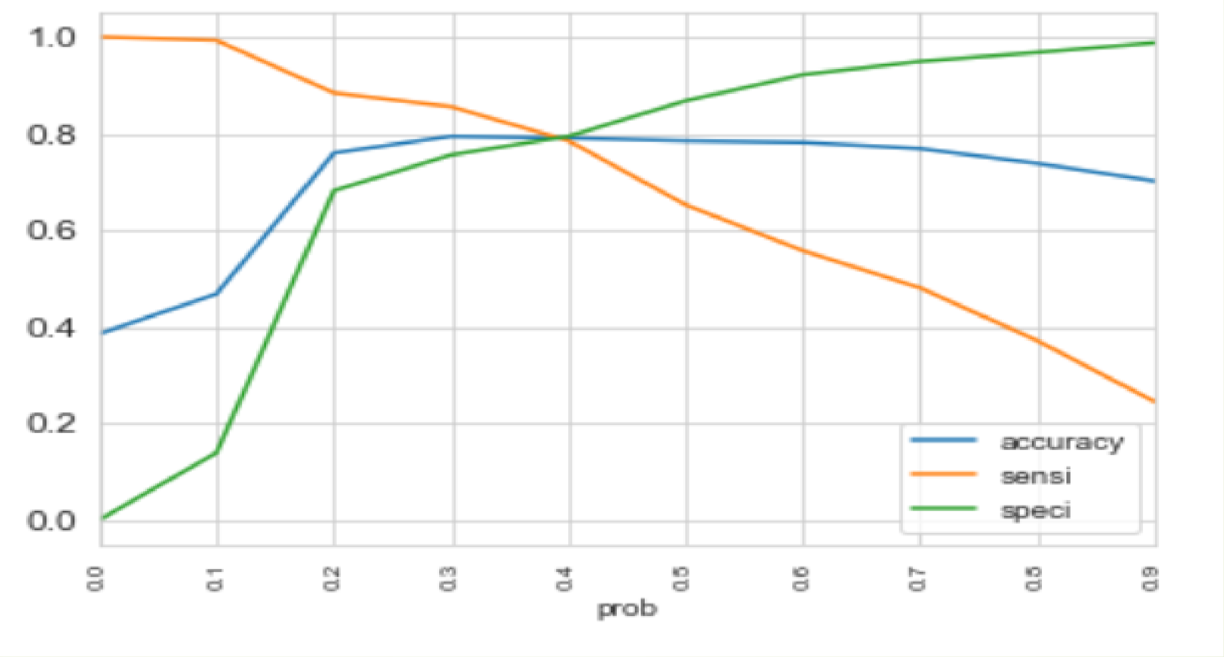
- The ROC curve demonstrates the following things.

1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Finding Optimal cut-off

- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity. From the below plot, we have considered 0.35 as the optimum point to take it as a cutoff probability and created a Confusion matrix with this cut-off and then calculated Sensitivity, Specificity, Precision and Recall.
- Sensitivity for this Logistic regression model was calculated to be around 81%.



Results and Conclusions

- After performing the analysis, the following variables contributed most towards increasing the lead conversion rate which was calculated to be around 81%.
  - Last Notable Activity_Unsubscribed
  - Last Activity_Converted to Lead
  - Specialization_Others
  - Total Time Spent on Website etc.
- By concentrating on the hot leads we found out using our model . The lead conversion increases and in turn helps in growth of the business.