LEAD SCORE CASE STUDY

Team Members: Rajdeep Das, Ratish Moondra, Saniya Javed Shaikh

Problem Statement:

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score 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%.

Goals and Objectives:

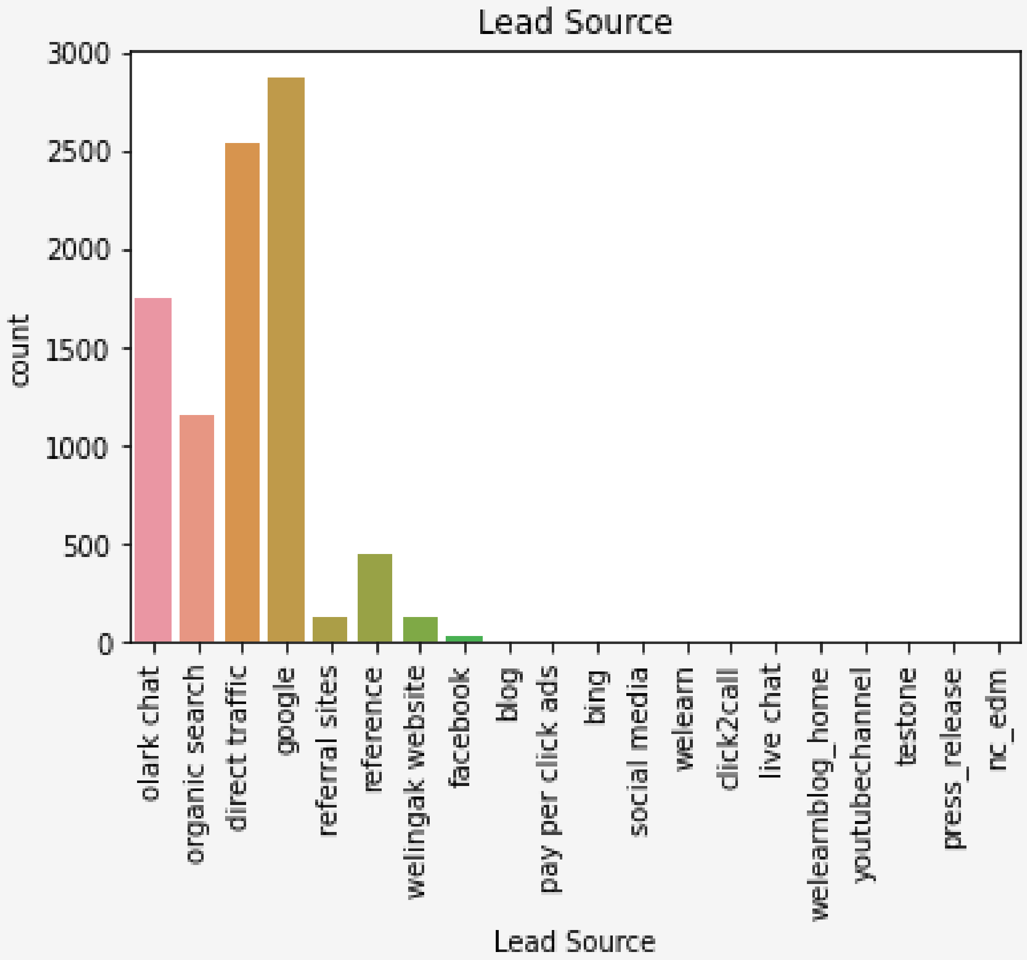
Build a logistic regression model to assign a lead score between 0 and 100 to

- Each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Solution Methodology:

- Step 1: Loading and cleaning of data – Check for duplicate data, NA values, missing values and treating the accordingly. Drop Columns if necessary as it may contain large amount of missing value or those column that doesn't lead to analysis. Check for outliers in the data.
- Step 2: EDA – Univariate analysis and bivariate analysis to understand the correlation between the variables.
- Step 3: Creating of dummy variables, feature scaling and encoding of data.
- Step 4: Model Building.
- Step 5: Creating prediction from the model created.
- Step 7: Model evaluation using the metrics like sensitivity, specificity, presession & recall.
- Step 8: Prediction on test set.
- Step 9: Precession and recall trade-off.
- Step 10: Making the final prediction of the test set and measuring the evaluation metrics.

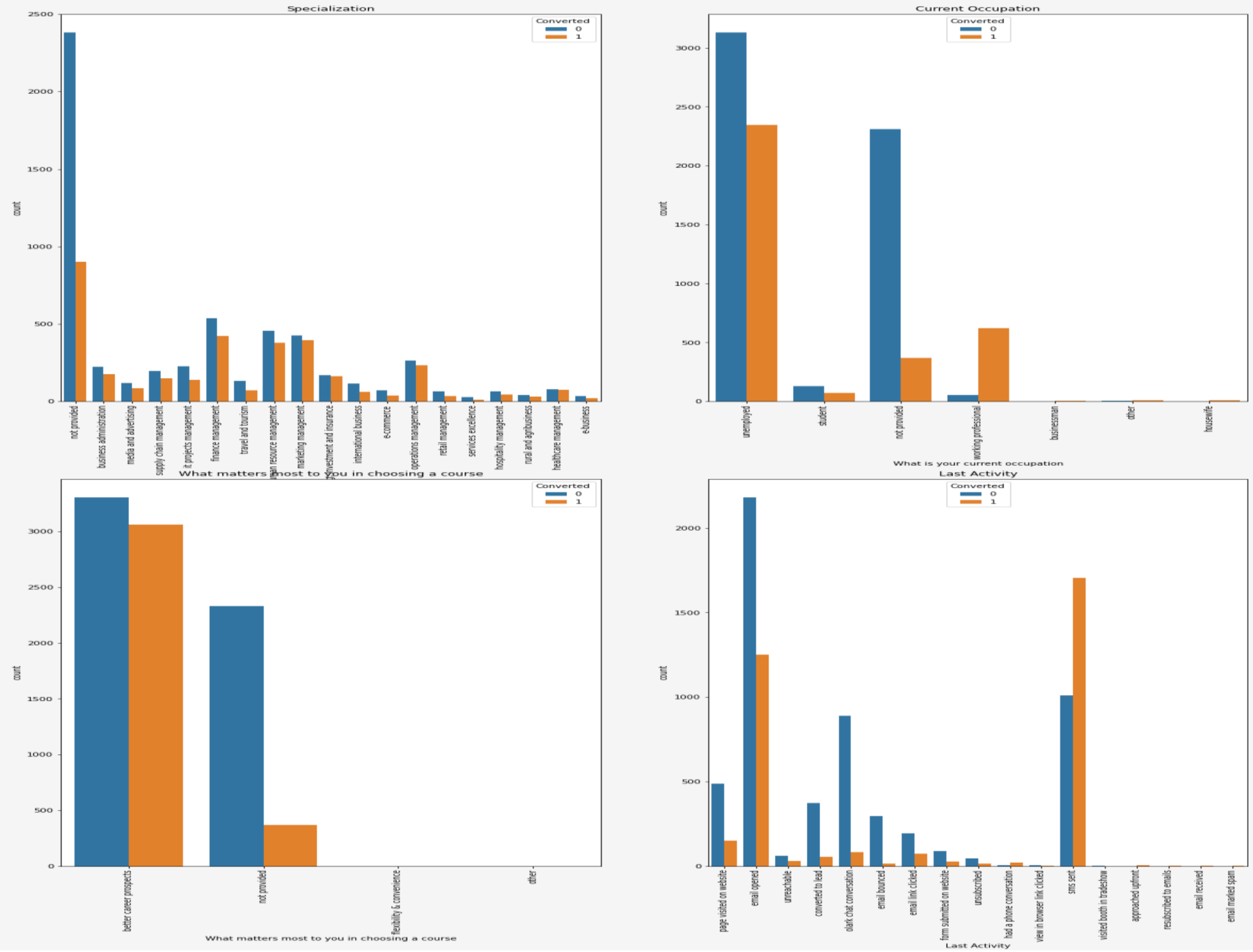
EDA Analysis:



Inferences:

Lead are generated primarily through google followed by direct traffic, olark chat and organic search.

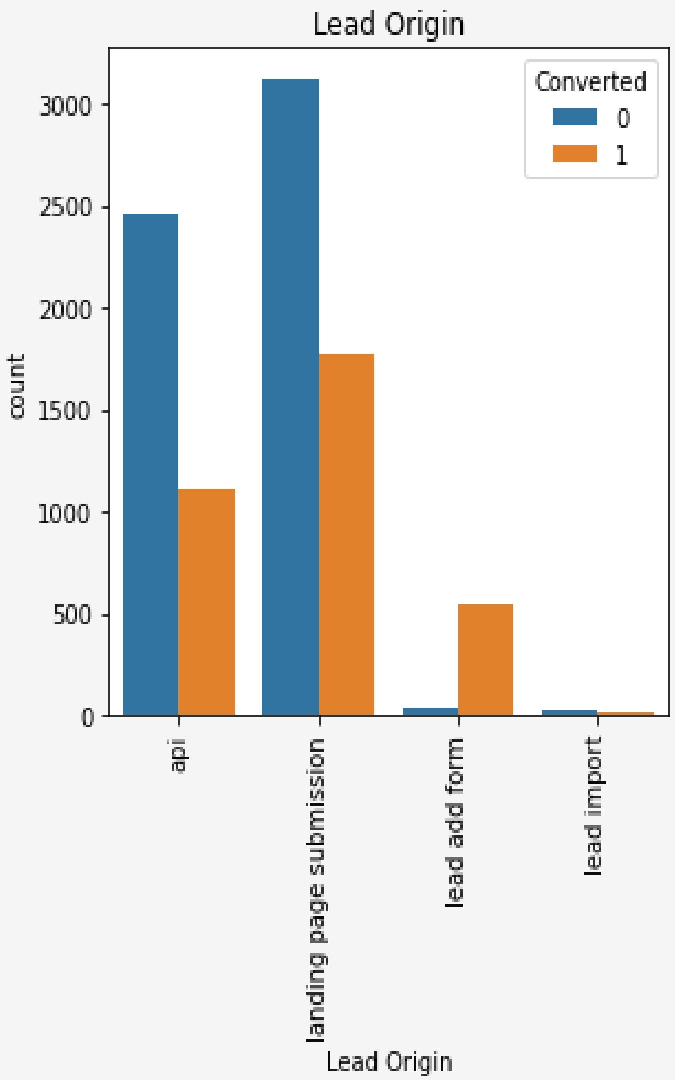
EDA Analysis:

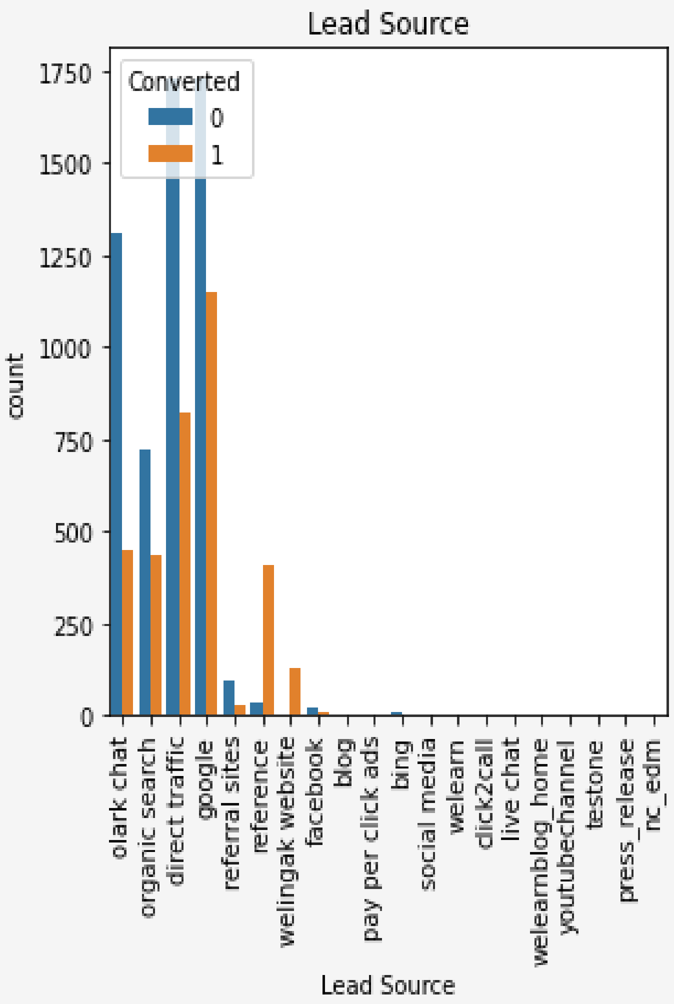


Inference:

1.Maximum lead are unemployed and looking for upskilling to get employable.
2.Most of the leads are looking for better career prospects by enrolling them for upskilling courses.
3.Important communication can be done through email and SMS as this mode are used maximum used by the learners.
4.There are patches of specialization where Conversion ratio is high.

EDA Analysis:





Inference:

1.Maximum lead are converted to sales from 'landing page submission' followed by api.

2.Lead form advertisement are not getting converted to sales.

3.Lead from google & direct traffics are getting converted the most as leads from this source can be considered hot as learners are interested in upskilling.

Model Building:

- As there were quite a number of features we use RFE to select the top 15 variables for us and we have build the model using those 15 variables.

- Then we used VIF score and P Value to eliminate the features and build the model. The final model had 13 features and we dropped two features.

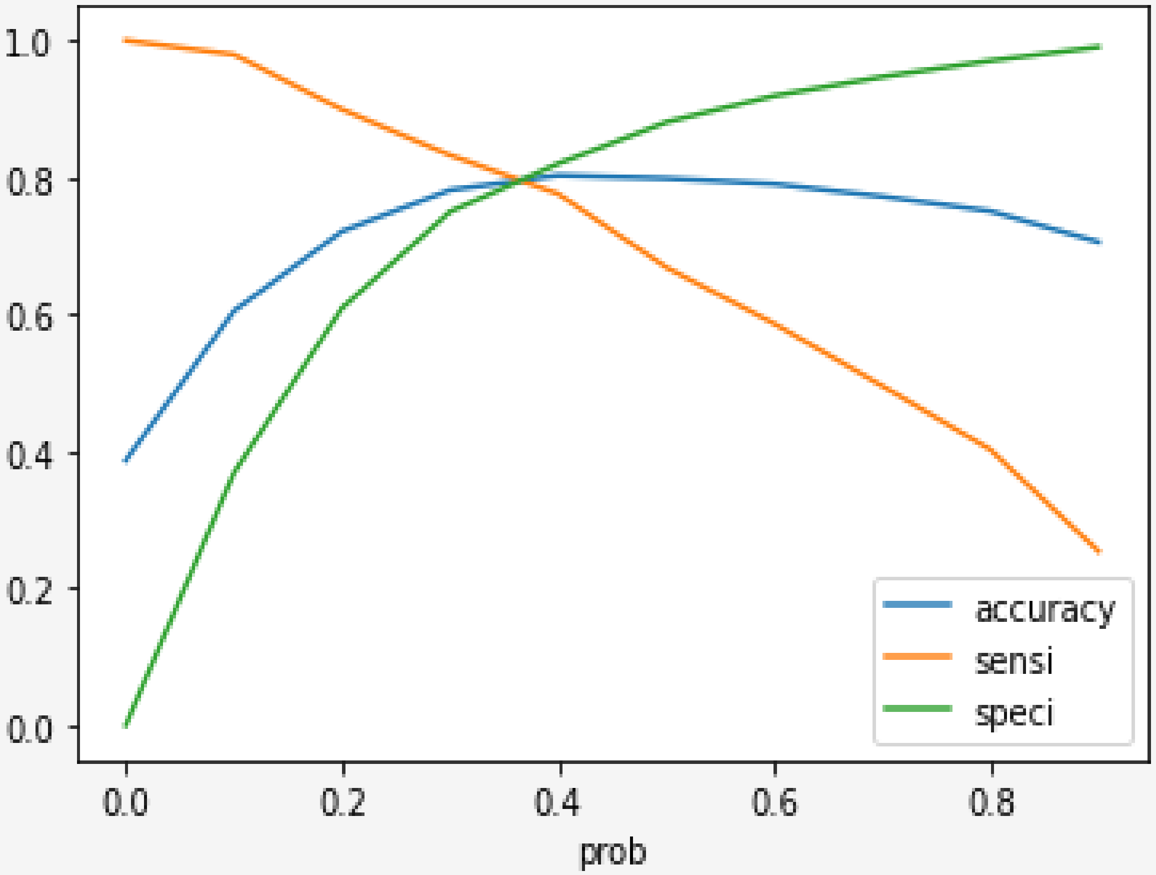
- We selected initial cut off od 0.5 and we go accuracy of 79% with sensitivity of 66% and specificity of 88%.

- We than use ROC curve to find the optimal cutoff. The optimal cutoff was 0.35 and the accuracy we achieved was 79%, with sensitivity 80% and specificity of 79%.

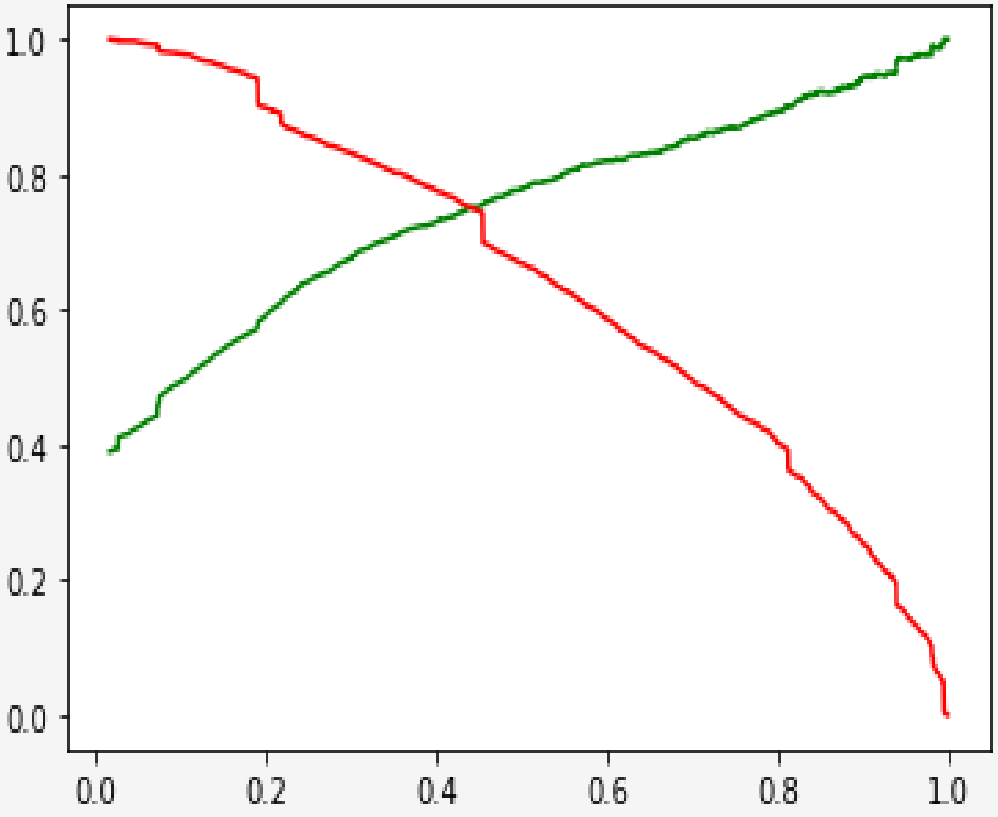
- We than used precesion and recall tradeoff and got the new optimal cut off as 0.41. With new optimal cutoff we got accuracy of 80%, we have precesion of ~73% and recall of ~77%.

- On test set we got accuracy of 81% with precesion of 71% and recall of 78%.

ROC Curve:



The optimal cut off using ROC cure was 0.35



The optimal cut using precesion recall trade of was 0.41

Conclusion:

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits. ▪ When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.