Lead Score Case Study

SUBMITTED BY: Riya Jain

Problem Statement

- X Education provides online courses to Industry Professionals. Their conversion rate is about 30%. Now, the CEO wants to assign the leads being generated to be assigned with a lead score so that they can identify which leads are ‘hot’ and increase the conversion rate to around 80%.
- By identifying the most potential leads as hot leads, the CEO aims to make the process more efficient and improve the current “poor” conversion rate

Analysis Approach

- Read & Understand the Data
- Data Cleaning
- Data preparation, Pre-processing & Test-train split
- Feature selection & Building the Model
- Determining Optimal-Threshold
- Confusion Matrix & Performance Parameters
- Making Predictions

Understanding Data & EDA

After uploading and understanding the data at hand, the first step was to execute exploratory data analytics wherein the main aim was to clean it. While cleaning, we:

Treated the missing values

Conducted outlier analysis

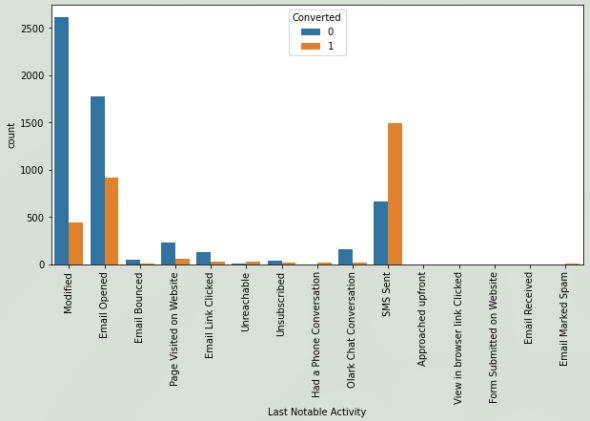
Performed Univariate & Bi-variate analysis

Also, Multivariate analysis

We started with (9260 X 37) samples and concluded EDA at (8144 X 20) samples for further analysis

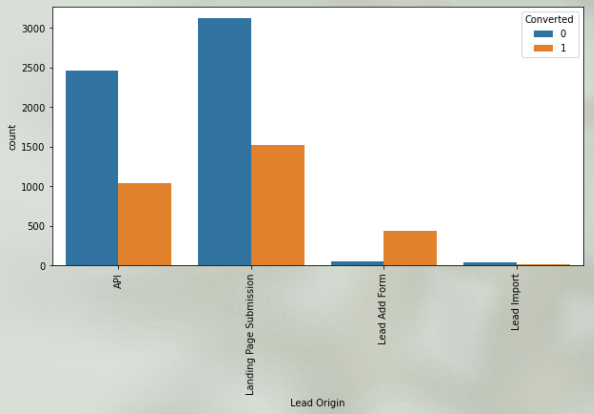
Last Notable Activity:

Conversions are way higher when SMS is sent to leads



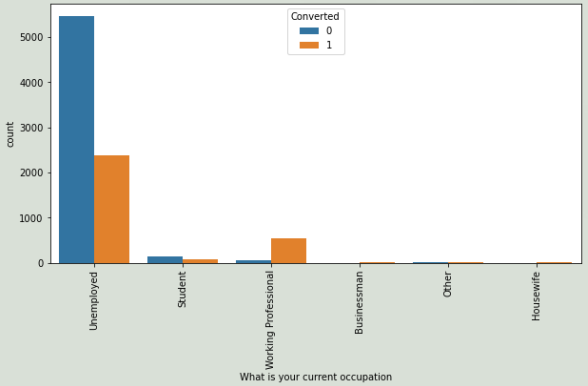
Lead Origin:

When leads are created through lead form, majority of them are converted



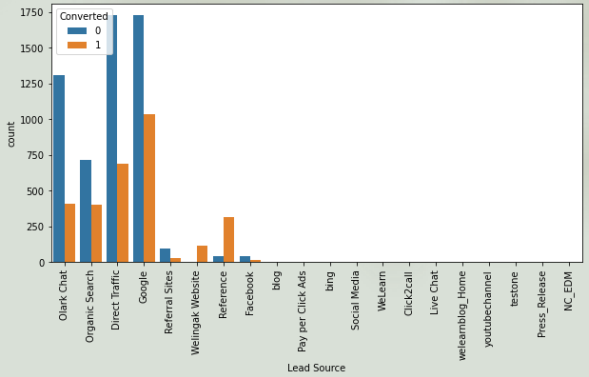
What is current occupation:

Majority of the working professionals that register, do end up enrolling for the course



Lead Source:

Leads coming in through ‘Reference’ & ‘Welingak website’ give the best conversions



Data preparation, Pre-processing & Test-train split

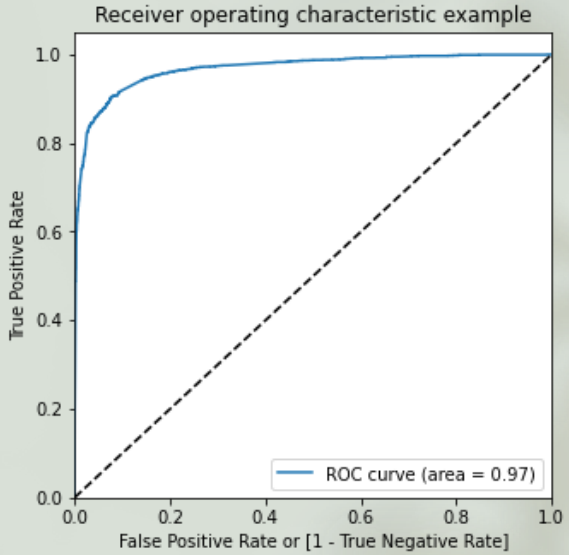
- The clean data had numerical and categorical columns.
- Numerical columns were scaled through the standard scaling procedure so that they are centered at 0 with a unit standard deviation
- Binary level categorical columns were replaced with 0 & 1
- Multi-level Categorical columns were treated with get_dummies()
- Data was split into X (independent variables) and y (target variable)
- Followed by splitting the data further X_train, X_test, y_train & y_test

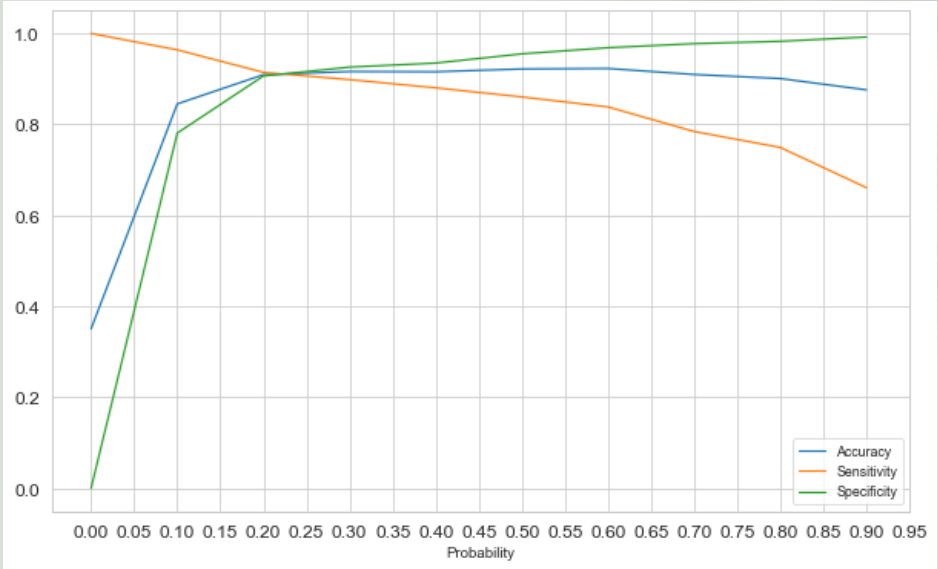
Feature selection & Building the Model

- Since there were 93 features to begin with, the initial 20 features were picked through Recursive Feature Selection,
- Followed by manually eliminating features on the basis of their significance (p value) & VIF values (multicollinearity)
- The process was repeated till all the features in the model had a p-value of less than 0.05 and VIF value under 5
- In the end, 16 features were shortlisted

Determining Optimal-Threshold

- We plotted the ROC curve, and
- Accuracy, Sensitivity & Specificity
- To conclude that the threshold value is 0.22





Confusion Matrix & Performance Parameters

Train Set

Sensitivity	90.8%
Specificity	91.1%
Precision	84.7%

Test Set

Sensitivity	91.2%
Specificity	90.8%
Precision	83.2%

Final Model

- Lead Score was calculated based on the probability of conversion multiplied by 100 and assigned to each sample
- Lead source- Welingak website, Olark chat & Lead add form give the best conversion
- Leads with tags- SMS sent & Will revert to emails are the ones that eventually enrol in the courses
- Switched off/Invalid numbers are ones that don’t but the courses, obviously