Leads Scoring Case Study

Study by:
Lasya Nallamalli
Pranav Kohli
Kushagra Srivastava

Problem Statement

X organization provides online courses for industry professionals. The courses are marketed at different search platforms and the lead generated through these 3rd parties are pursued. CEO’s objectives:

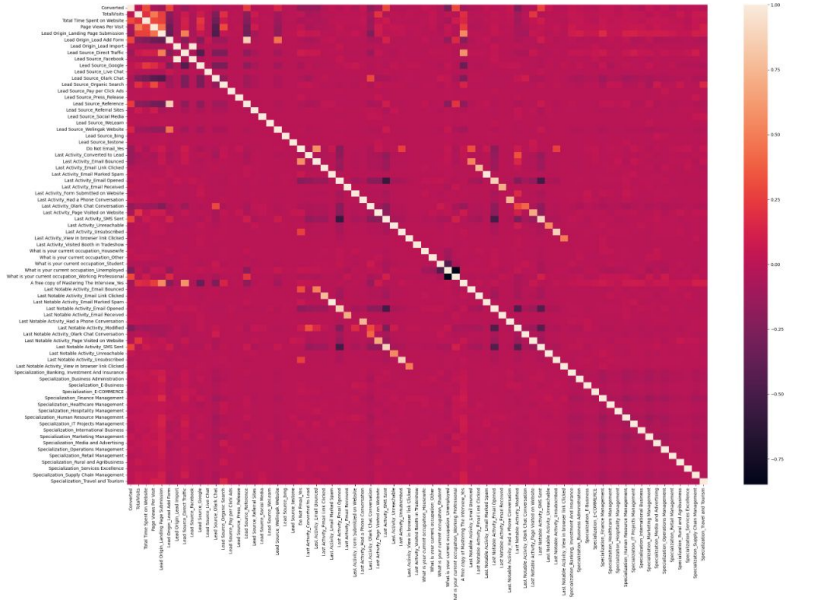
- Identify sources and journey of promising leads
- Increase lead conversions rate from the existing 30% to the desired 80%

Analysis Approach

- Started with cleaning and imputing the data to avoid outliers and missing values
- Splitting the dataset in train:test split of 70:30
- Evaluating correlation of existing parameters
- Model building
- Model accuracy and precision check
- Identifying most useful features for business decisions

EDA & Data Preparation

EDA- Observations
Country with most leads- India
Mumbai & other cities in Maharashtra with most leads
Max Lead Source- Google
‘Better Career Prospects’ is the main driver for many people
Most of the people are ‘Unemployed’ in leads
Lead origins are majoly form landing page



Lead Origin	
Landing Page Submission	4886
API	3580
Lead Add Form	718
Lead Import	55
Quick Add Form	1

Name: count, dtype: int64

Lead Source	
Google	2868
Direct Traffic	2543
Olark Chat	1755
Organic Search	1154
Reference	534

What matters most to you in choosing a course	
Better Career Prospects	6528
Flexibility & Convenience	2
Other	1

What is your current occupation	
Unemployed	5600
Working Professional	706
Student	210
Other	16
Housewife	10

EDA & Data Preparation

- Country and columns are dropped as most of the leads are from India and major cities in Maharashtra including Mumbai
- Prospect ID Lead Number, as they are unique identifiers of no use
- As "Lead Profile" and "How did you hear about X Education" have a lot of rows of value Select which is of no use, those can be dropped
- As 'What matters most to you in choosing a course' have most of the values as 'Better Career Prospects', we can drop the column
- The column 'What is your current occupation' has lot of null values, let's remove the rows having null values.
- Dropping the null values rows in the column 'TotalVisits' & ‘LeadSource’
- Post this - **Dummy Variables** are created for categorical variables and added to the dataframe

Logistic Regression Model Building Steps

Model 1 made through RFE
Features= 15

Model 1

Model 2

Checking for high P & High VIF values; Dropping variable: Lead Source_Reference feature to get Model 2

No feature with High VIF. The high p value feature was dropped: Last Notable Activity_Had a Phone Conversation

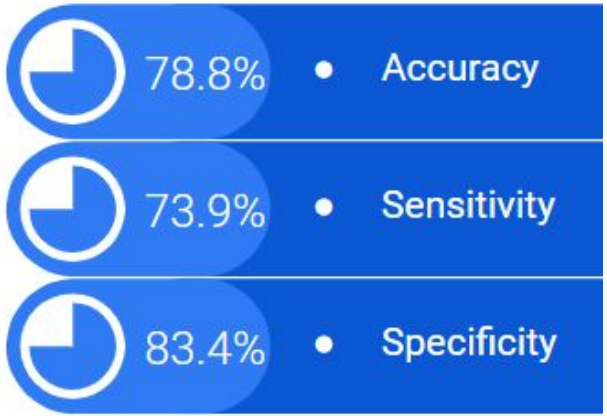
Model 3

Model 4

Similarly in next iteration, What is your current occupation_Housewife was dropped

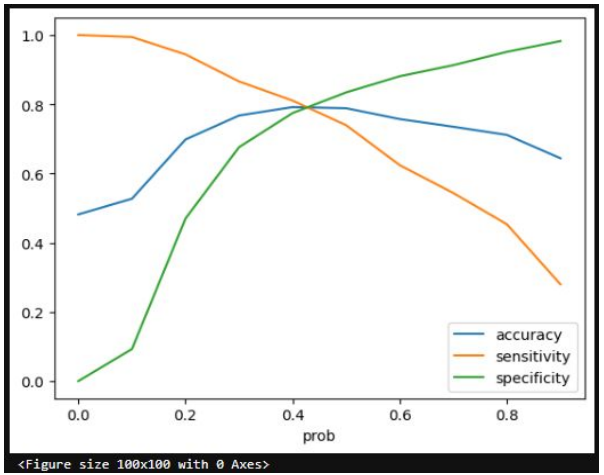
Final model obtained after dropping current occupation_Working Professional

Model 5



Plotting the Receiver Operating Characteristics

0.42 is the optimal value of the three metric to balance the accuracy, sensitivity and specificity



Summary of the observations

Data set reduced from (9240, 37) to (6373, 75) after initial cleaning.

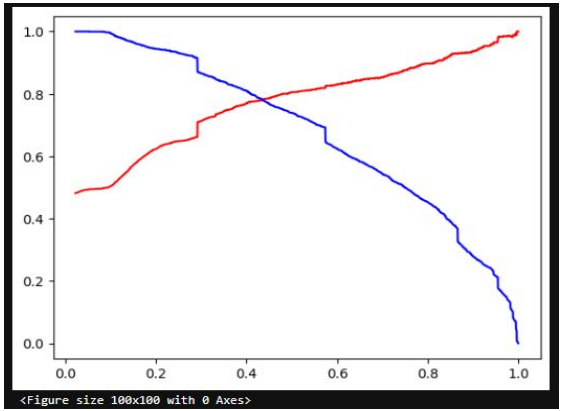
TRAIN DATA:

Accuracy : 78.8
Sensitivity: 73.9
Specificity: 83.4

TEST DATA:

Accuracy : 78.4
Sensitivity: 77.9
Specificity: 78.9

Precision Vs Recall Balance



0.44 is the value where Precision & recall are getting balanced

TEST DATA with 0.44 cutoff

Accuracy : 78.4
Precision : 77.9
Recall : 78.9

Recommendations

1. The company should be focussing more on the leads which have more number of visits and the time they spent there, and Lead origin should be tracked for instance max is from Google and Welingak site.
2. Landing page submission can also lead to higher conversions
3. Marketing management, Human resource management have higher conversion
4. Maximum leads are from unemployed people and max conversion is for working professional