LEAD SCORING CASE STUDY

Presenters:

Maitrai Bahuguna

Viraj Bhosale

Shivam Chouskey

INTRODUCTION:

- An education company, X Education sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30% .

BUSINESS GOALS:

- Company wishes to identify the most potential leads, also known as “Hot Leads” The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance The CEO, in particular, has given a ballpark number for the lead conversion rate i.e., 80%

APPROACH

1. Importing Data
2. Inspecting the Data frame
3. Data Preparation
   - Encoding Categorical Variables
   - Handling Null Values
4. EDA
   - univariate analysis
   - outlier detection
   - checking data imbalance
5. Dummy Variable Creation
6. Test-Train Split
7. Feature Scaling
8. Looking at Correlations
9. Model Building
   - Feature Selection Using RFE
   - Improvising the model further inspecting adjusted R-squared, VIF and p-vales.
10. Buildfinal model
11. Model evaluation with different metrics Sensitivity, Specificity

PROBLEM SOLVING METHODOLOGY:

DATA CLEANING AND PREPARATION:

- Read data from source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier treatment
- Exploratory data analysis

SPLITTING THE DATA AND FEATURE SCALING:

- Splitting the data into train and test dataset
- Feature scaling of numerical variables

MODEL BUILDING:

- Feature selection using RFE, VIF and p-value
- Determine optimal model using Logistic Regression
- Calculate various evaluation metrics

RESULT:

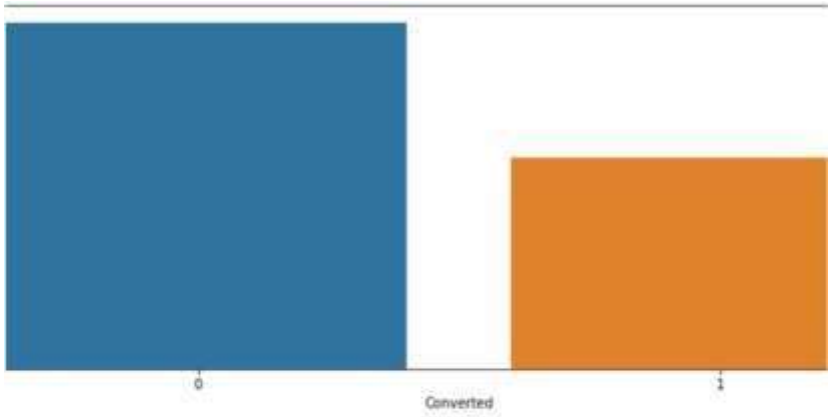
- Determine Lead score and check if target final prediction is greater than 80% conversion rate
- Evaluate final prediction on test set.

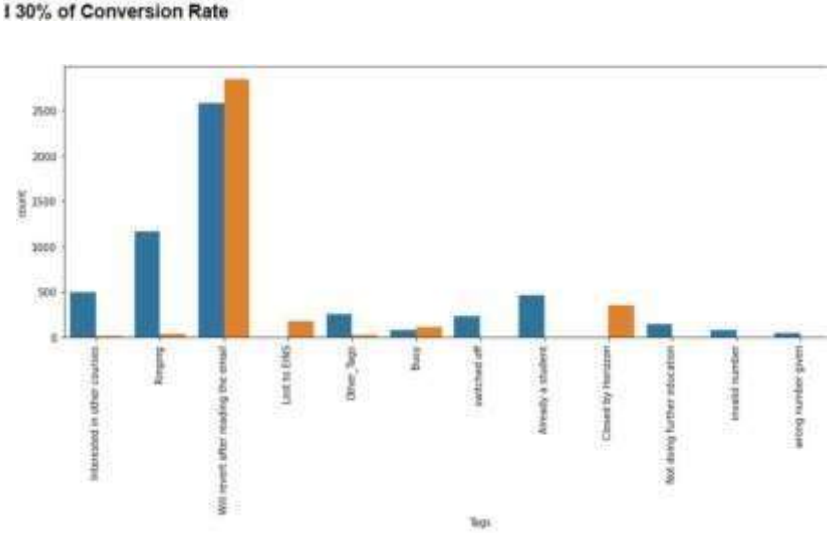
DATA CONVERSION:

1. CONVERTING THE VARIABLE WITH VALUES YES/NO to 1/0s
2. CONVERTING THE 'SELECT' VALUES WITH NaNs
3. DROPIING THE COLUMNS HAVING >70% OF NULL VALUES
4. DROPPING UNNECESSARY COLUMNS
5. DROPPING THE ROWS AS THE NULL VALUES WERE <2%.

EXPLORATORY DATA ANALYSIS:

- The figure 1 is of Tags variable with hue set with target variable.
- We can see that conversion rate of leads more in closed by Horizzon and EINS.
- People with response of will revert on reading a mail is more generic way of reverting to the calls.
- The figure 2 is showing the conversion rate in total as the we can see that about 30% of people are being converted.

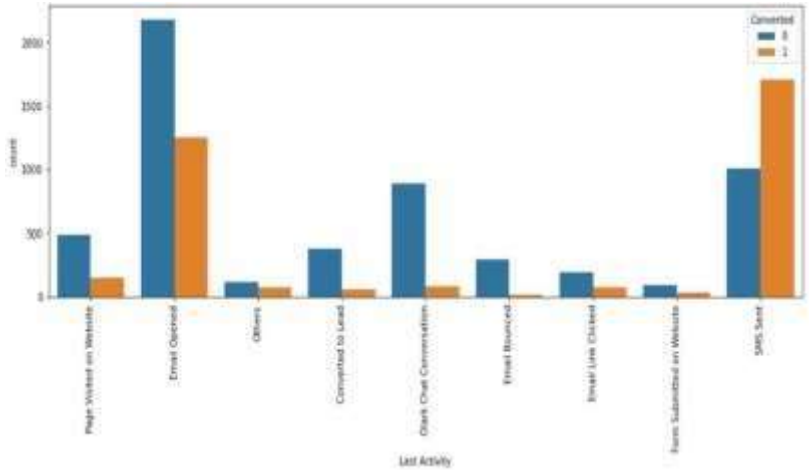


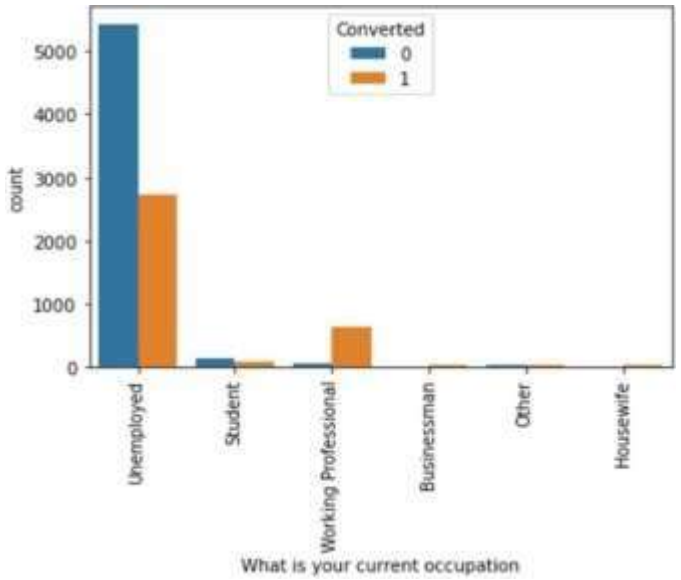


Last Activity and Occupation:

- Conversion rate is more in working professionals and total leads are more in unemployed people

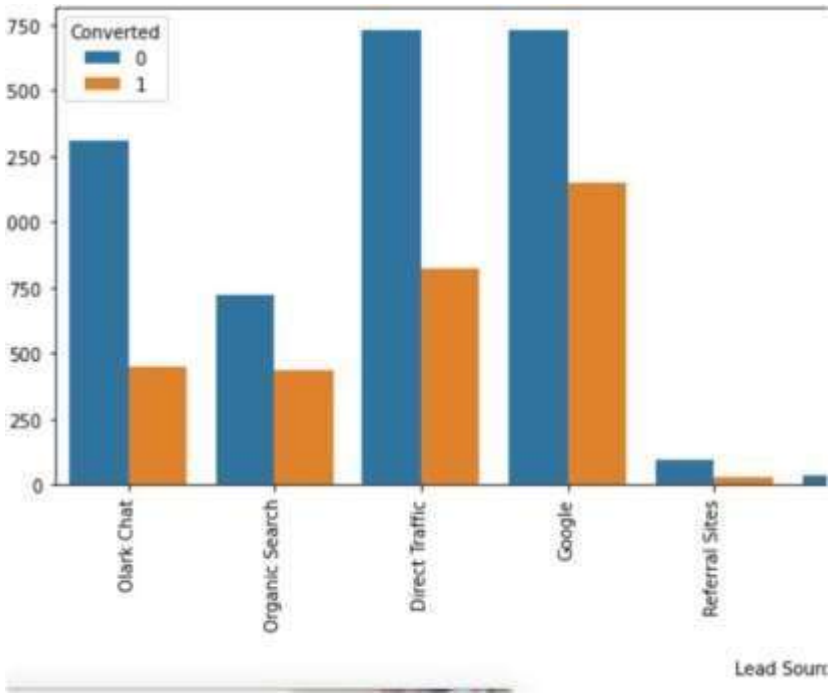
- last activity was:
  - a. SMS
  - b. Olark chat conversation





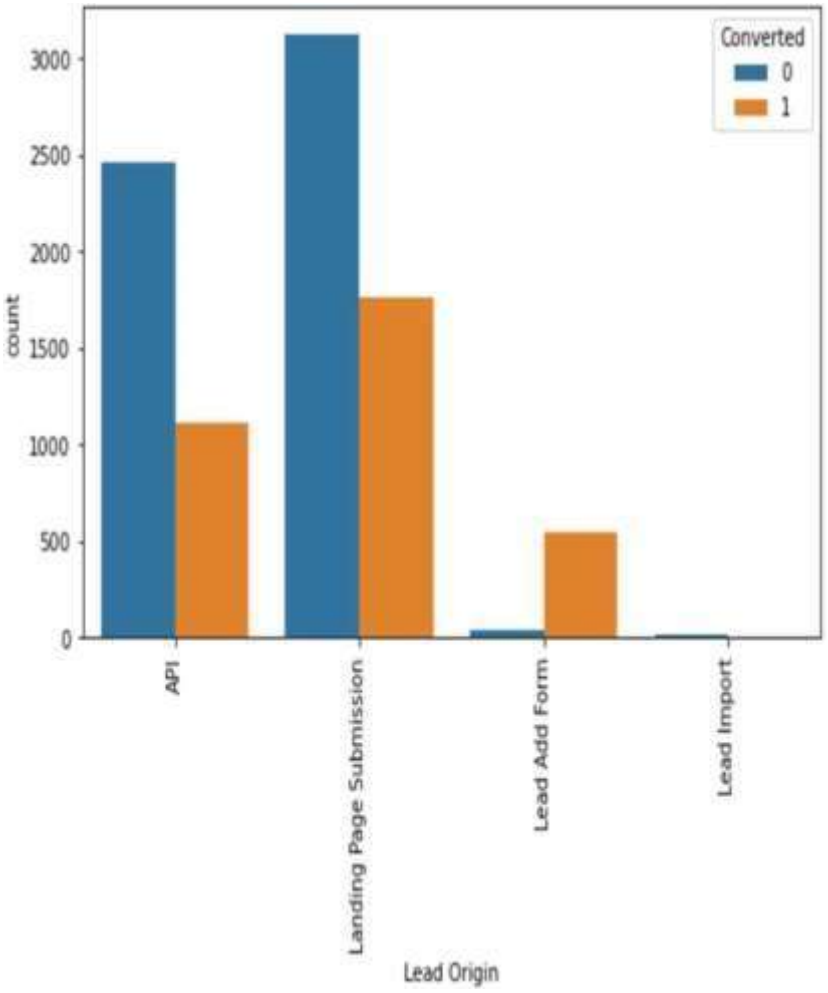
LEAD SOURCE:

- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum
- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'



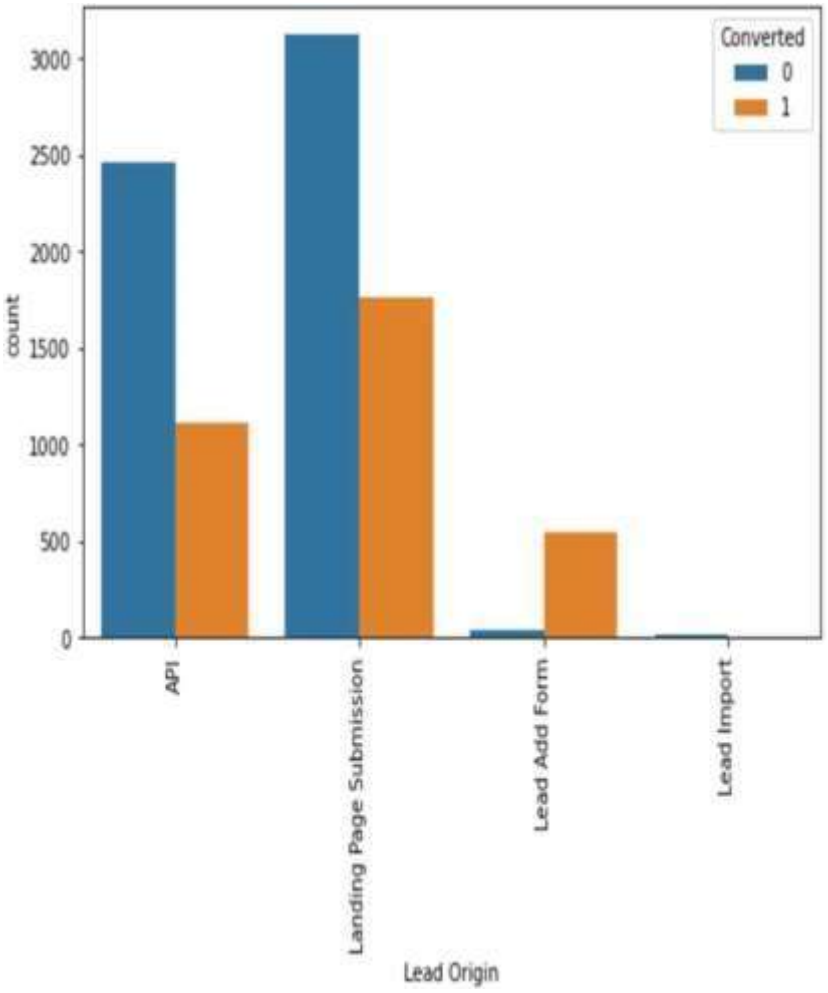
Lead origin:

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



Lead origin:

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



Model Building:

SPLITTING THE DATA INTO TEST AND TRAINING SETS

WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30

USING RFE TO CHOOSE TOP 15 VARIABLES

BUILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE > 0.05 AND VIF > 5

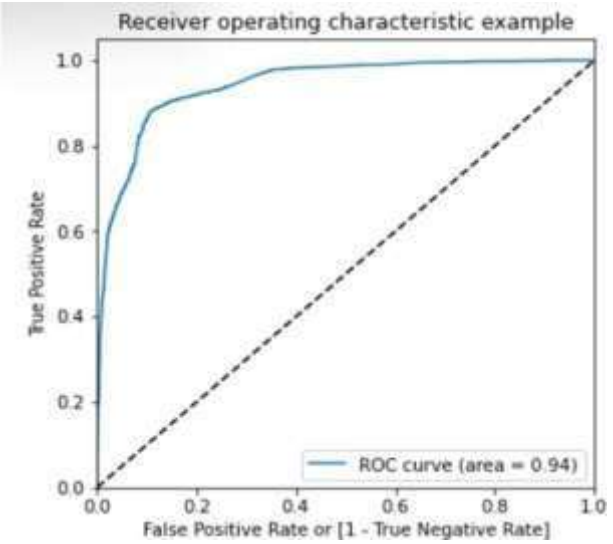
PREDICTIONS ON TEST DATASET Ø OVERALL ACCURACY IS 92.0 %

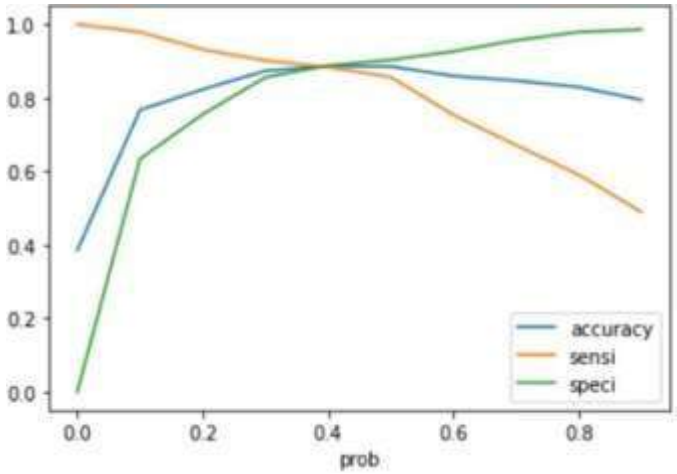
Fig 1: ROC Curve:

- The ROC curve has the area 0.94 which is good as the ROC curve should be within 1.

Fig2:Optimum Cut-off:

- Optimum cut-off is of 0.4 approx.





Model Evaluation:

- CALCULATED ACCURACY, SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITY CUTOFFS FROM 0.1 TO 0.9
- AS PER THE GRAPH AND LOOKING AT THE OTHER SCORES, IT CAN BE SEEN THAT THE OPTIMAL POINT IS 0.4

Predicted Actual	Not Converted	Converted
Not Converted	3341	564
Converted	242	2204

	prob	accuracy	sensi	speci
0.0	0.0	0.385136	1.000000	0.000000
0.1	0.1	0.766651	0.978741	0.633803
0.2	0.2	0.822863	0.931316	0.754930
0.3	0.3	0.873091	0.901063	0.855570
0.4	0.4	0.886002	0.883483	0.887580
0.5	0.5	0.884900	0.856092	0.902945
0.6	0.6	0.859392	0.753066	0.925992
0.7	0.7	0.846481	0.670891	0.956466
0.8	0.8	0.829161	0.590352	0.978745
0.9	0.9	0.794206	0.488962	0.985403

With Probability cut-off of 0.4 we have precision and recall rates as 79.62% and 90.10%

Observation

Accuracy:87.30%

Sensitivity:90.10%

Specificity:85.55%

Model Prediction:

- The following figures shows the Top features and the metrics of the model .
- Top Variables are:
  1. **Tags_Lost to EINS**
  2. **Lead Origin_Landing Page Submission**
  3. **Tags_Closed by Horizzon**

Predicted Actual	Not Converted	Converted
Not Converted	1469	265
Converted	109	808

Top Features

Total Time Spent on Website	1.0485
Lead Origin_Landing Page Submission	-1.9721
Lead Origin_Lead Add Form	2.5556
Lead Source_Welingak Website	1.9362
Tags_Busy	3.4359
Tags_Closed by Horizzon	8.7423
Tags_Lost to EINS	8.4432
Tags_Will revert after reading the email	4.1791
Do Not Email_Yes	-1.8025
Last Activity_Olark Chat Conversation	-1.3341
Last Activity_Others	1.3348
Last Activity_SMS Sent	1.8520
Specialization_others	-1.8333

With Probability cut-off of 0.4 we have precision and recall rates as 79.62% and 90.10%

Observation

Accuracy:87.30%

Sensitivity:90.10%

Specificity:85.55%

Conclusion:

1. The logistic regression model is used to predict the probability of conversion of a customer.
2. Optimum cut off is chosen to be 0.41 i.e., any lead with greater than 0.4 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.41 or less probability of converting is predicted as Cold Lead (customer will not convert)
3. Total Features used as 13 in number
4. When the following factors are met the conversion rates are high:
    1. Lead source was:
        a. Google
        b. b. Direct traffic
        c. c. Organic search
        d. d. Welingak website
    2. last activity was:
        a. SMS
        b. b. Olark chat conversation
    c. 6.lead origin is Lead add format 3.current occupation is as a working professional.