LEAD SCORE CASE STUDY

Govardhana Reddy P, Varsha Bijish Banu, Vatsla Shukla

Problem Statement :

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines likeGoogle.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When thesepeople 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. Through this process, some of theleads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score havea 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%.

OVERALL APPROACH

DATA CLEANING AND IMPUTING MISSING VALUES .

EXPLORATORY DATA ANALYSIS

FEATURE SCALING AND DUMMY VARIABLE CREATION .

SPLITTING THE DATA INTO TRAIN TEST

LOGISTIC REGRESSION MODEL BUILDING AND CALCULATION OF LEAD SCORE

MODEL EVALUATION USING DIFFERENT METRICS

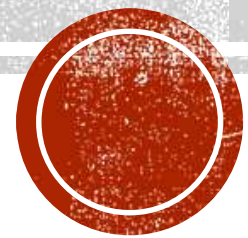
CONCLUSION AND RECOMMENDATION

BUSINESS UNDERSTANDING

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%

DATA CLEANING AND IMPUTING MISSING VALUE

- CONVERTING THE VARIABLE WITH VALUES YES/NO to 1/0s FOR BUILDING
- CONVERTING THE ‘SELECT’ VALUES WITH NULLS
- DROPIING THE COLUMNS HAVING >60% OF NULL VALUES .
- DROPPING UNNECESSARY COLUMNS

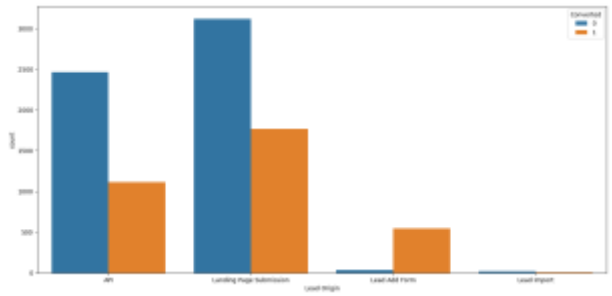


EXPLORATORY DATA ANALYSIS

LEAD ORIGIN:

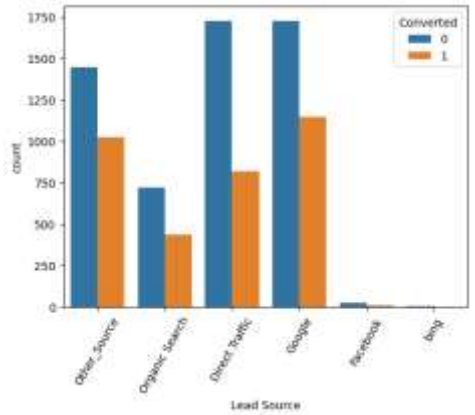
Conversion rates of Landing Page Submission needs to be improved.

Number of API leads needs to be increased



LEAD SOURCE:

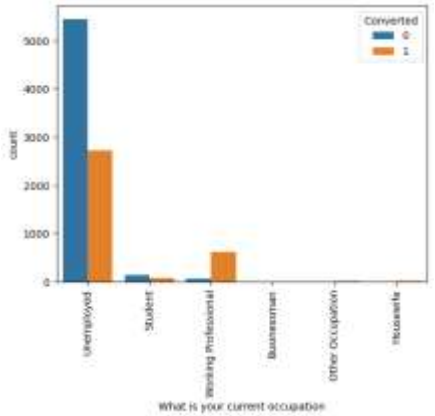
- Direct Traffic and Google have highest Conversion rates
- The conversion rates of direct traffic needs improvement.



OCCUPATION:

Unemployed leads conversion rate is low.

Working professionals have a higher conversion rate.



LAST ACTIVITY:

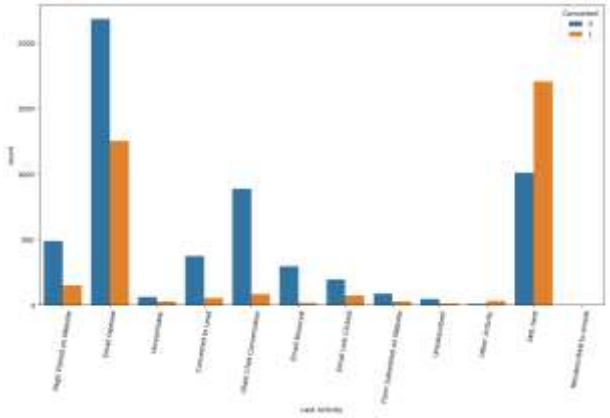
- Leads whose Last Activity is 'SMS Sent' has very high chances of Lead Conversion.
- 'Email opened' activity has high lead.

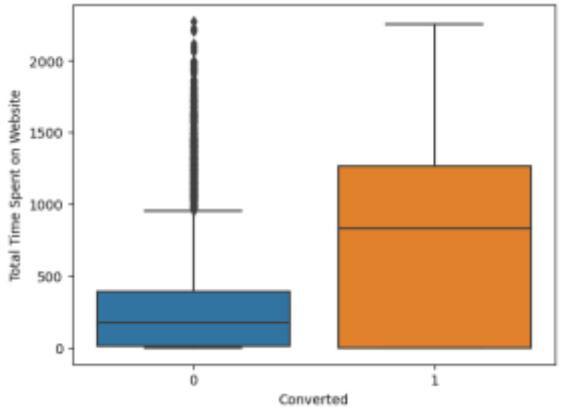
TOTAL TIME SPEND IN WEBSITE:

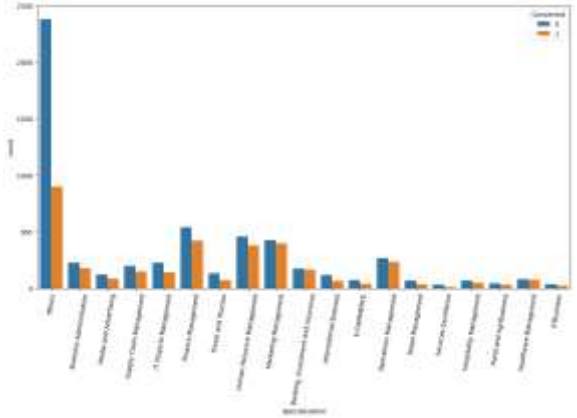
Leads who spend more time on the website have higher conversion rates

SPECIALISATION:

Marketing management, Banking, Health care has high conversion rates.







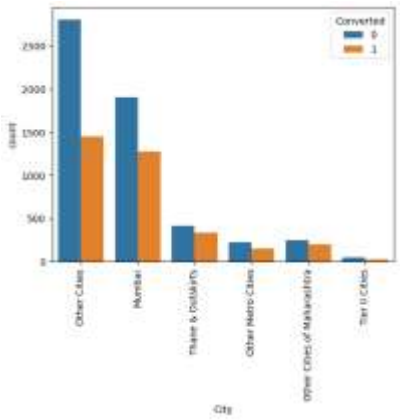
CITY:

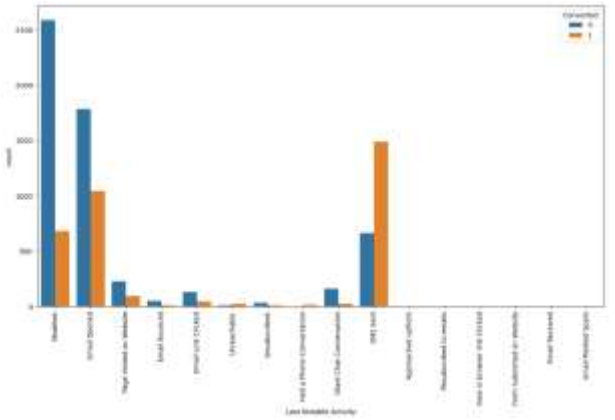
- Most of the Leads are from Mumbai
- 'Other Cities of Maharastra' has good Conversion rate.

LAST NOTEABLE ACTIVITY:

SMS Sent has high conversion rates.

Email opened has moderate conversion.





DATA PREPARATION:

9

- BINARY LEVEL CATEGORICAL COLUMNS WERE MAPPED TO 1/0 TO MAKE THEM COMPATIBLE WITH THE LOGISTIC REGRESSION MODEL.
- DUMMY FEATURES WERE CREATED FOR CATEGORICAL VARIABLES SUCH AS LEAD ORIGIN, LEAD SOURCE, LAST ACTIVITY, SPECIALIZATION, AND CURRENT_OCCUPATION USING HOT LEAD ENCODING,.
- THE TRAIN AND TEST SETS WERE SPLIT IN A 70:30 RATIO TO TRAIN THE MODEL AND EVALUATE ITS PERFORMANCE ON UNSEEN DATA.
- FEATURE SCALING WAS PERFORMED USING THE MINMAX SCALER METHOD TO ENSURE THAT ALL FEATURES WERE ON THE SAME SCALE.
- CORRELATED PREDICTOR VARIABLES, SUCH 'PAGE VIEWS PER VISIT' ', SPECIALIZATION_OTHERS' WERE DROPPED TO AVOID MULTICOLLINEARITY ISSUES, KEEPING THE THRESHOLD OF 0.7 +/-

MODEL BUILDING

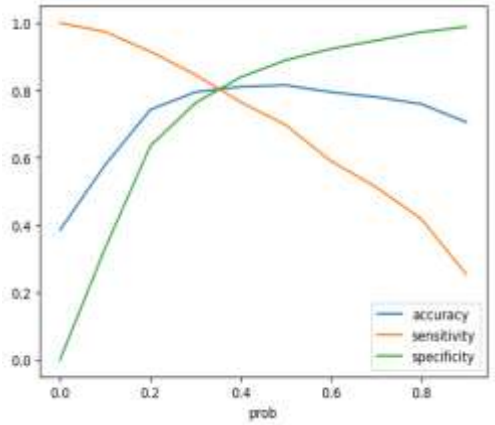
- SPLITTING THE DATA INTO TEST AND TRAINING SETS
- WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30
- USING RFE TO CHOOSE TOP 15 VARIABLES
- BUILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE > 0.05 AND VIF > 5
- PREDICTIONS ON TEST DATASET OVERALL ACCURACY IS 80%

VARIABLES IMPACTING THE CONVERSION RATE

- Lead Source - Other Source
- Last Activity - Olark Chat Conversation
- Last Notable Activity - Modified
- Lead Origin-Lead Add Form
- Total Time Spent on Website
- Last Activity - SMS Sent
- Do Not Email
- Current occupation-Working Profession
- Last Activity - Unsubscribed
- Last Activity - Other Activity
- Lead Origin - Lead Import
- Last Notable Activity - Unreachable

MODEL EVALUATION - SENSITIVITY AND SPECIFICITY ON TRAIN DATA SET

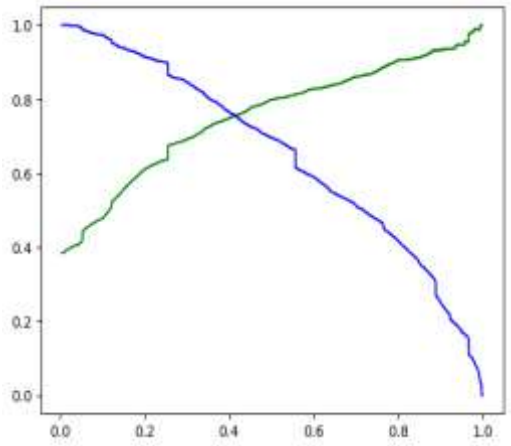
12



- The graph depicts an optimal cut-off of 0.37 based on Accuracy, Sensitivity and Specificity.
- Accuracy - 81%
- Sensitivity - 70%
- Specificity - 88 %
- False Positive Rate - 11 %
- Positive Predictive Value - 80 %
- Negative Predictive Value – 83%

MODEL EVALUATION- PRECISION AND RECALL ON TRAIN DATASET

13



- The graph depicts an optimal cut off of 0.4 based on Precision and Recall

- Precision : 80%
- Recall : 70%

CONCLUSION

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 77% and 82% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated shows the conversion rate on the final predicted model is around 81% (in train set) and 80% in test set
- The top 3 variables that contribute for lead getting converted in the model are
  - Total time spent on website
  - Lead Add Form from Lead Origin
  - Working Professional from Current Occupation
- We have also populated the Hot Leads where the lead conversion rate is high.
- The overall final model is good.