LEAD SCORING CASE STUDY…

CASE STUDY PARTNERS:

Rubi Bhakshi

Roshan Bhosale

PROBLEM STATEMENT

- An education company , X Education sells online courses to industry professionals. The Company markets its courses on various websites and search engines like Google
- Once the 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 forms providing their email address or phone number ,they are classified to be lead. Moreover also gets 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%

OVERALL APPROCH USED

1. DATA CLEANING AND IMMPUTING MISSING VALUES
2. EXPLORATORY DATA ANALYSIS
3. FEATURE SCALING AND DUMMY VARIABLE CREATION
4. LOGISTIC REGRESSION MODEL BUILDING
5. MODEL EVALUATION
6. CONCLUSION AND RECOMMENDATION

PROBLEM SOLVING METHODOLOGY

STEP 1: DATA CLEANING AND PREPARATION

- READ DATA FROM SOURCE
- CONVERT DATA INTO CLEAN FORMAT SUITABLE FOR ANALYSIS
- REMOVE DUPLICATE DATA
- OUTLIER TREATMENT
- EXPLORATORY DATA ANALYSIS

STEP 2: SPLITTING THE DATA AND FEATURE

- SPLITTING THE DATA INTO TRAIN AND TEST DATASET
- FEATURE SCALING OF NUMERICAL VARIABLE

STEP 2: MODEL BUILDING

- FEATURE SELECTION USING RFE, VIF AND P-VALUE
- DETERMINE OPTIMAL MODEL USING LOGISTIC REGRESSION
- CALCULATE VARIOUS EVALUATION METRICS

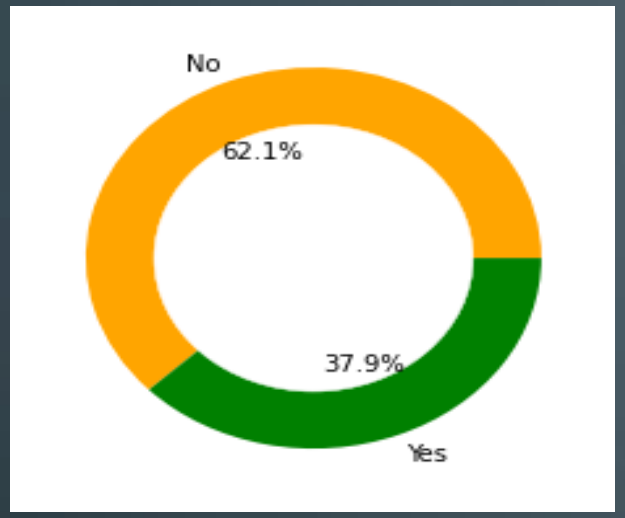
STEP 4: RESULT

- DETERMINE LEAD SCORE AND CHECK IF TARGET FINAL PREDICTION IS
- EVALAUATE FINAL PREDICTION ON TEST SET

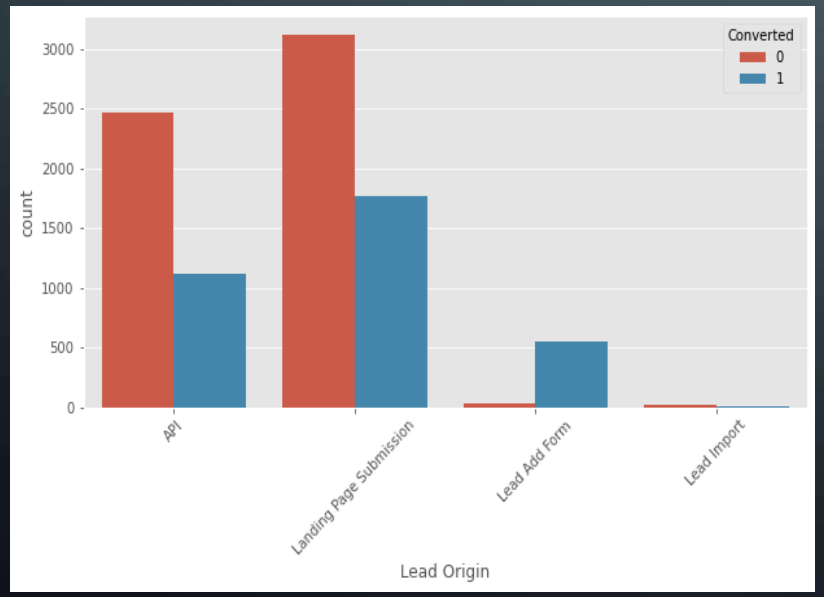
DATA CONVERSION

1. CONVERTING THE VARIABLE WITH VALUES YES/NO TO 1/0
2. CONVERTING THE 'SELECT' VALUES WITH 'NAN'S'
3. DROPPING THE COLUMNS HAVING >70% OF NULL VALUES
4. DROPPING UNNECESSARY COLUMNS
5. DROPPING THE ROWS AS THE NULL VALUE WERE <2%

EXPLORATORY DATA ANALYSIS



In the lead conversion ration, 37.9% has converted to leads where as 62.1% did not convert to a lead. So it seems like a balanced dataset.

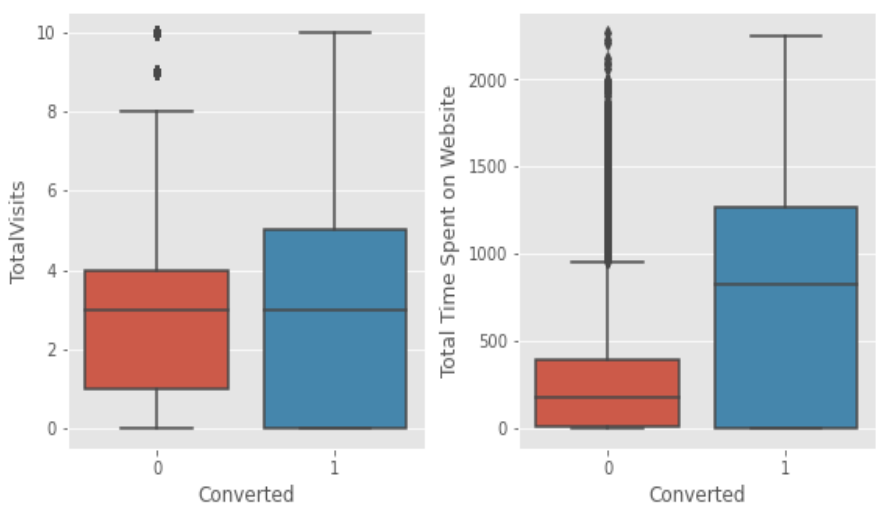


API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable

The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

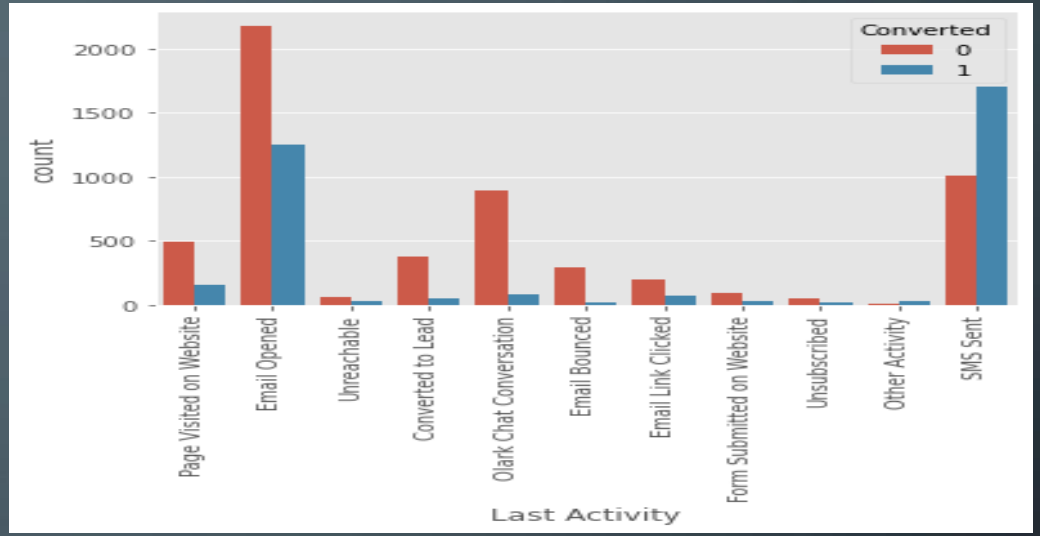
Lead Import has very less count as well as conversion rate and hence can be ignored

EXPLORATORY DATA ANALYSIS



- The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information

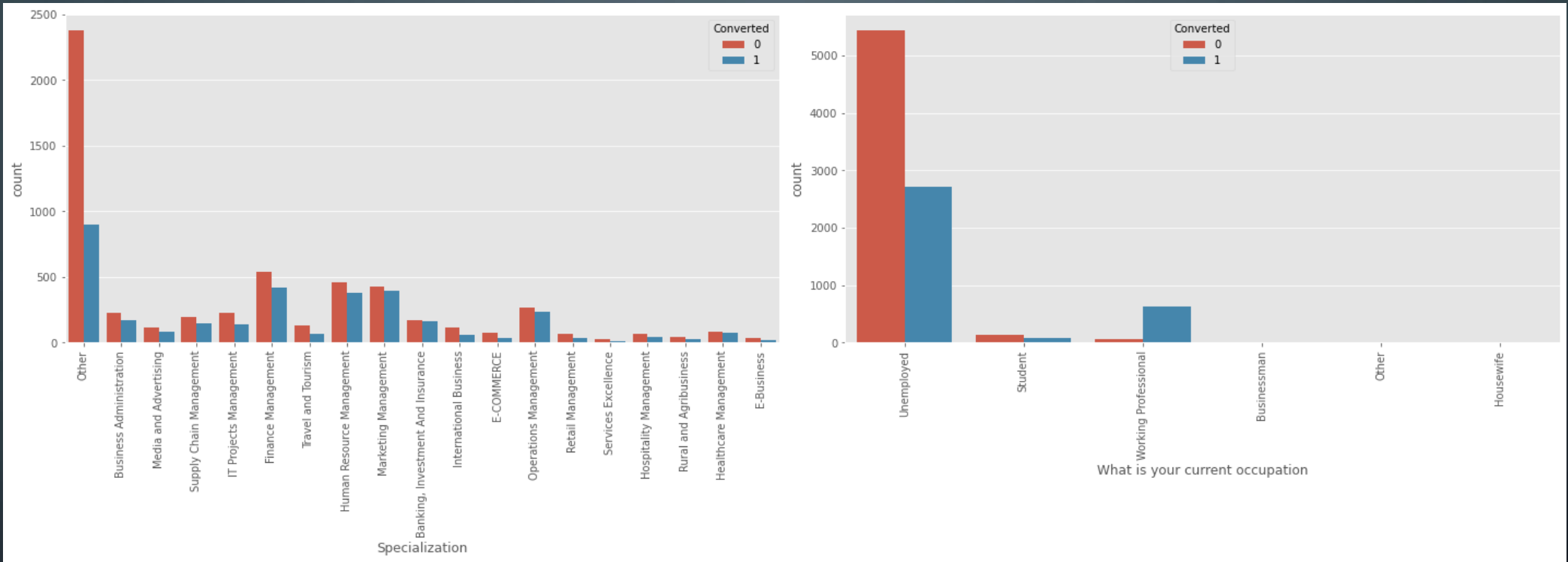
- Users spending more time on the website are more likely to get converted



- The count of lst activity as "Email Opened" is max

- The conversion rate of SMS sent as last activity is maximum

EXPLORATORY DATA ANALYSIS



Looking at above plot, no particular inference can be made for Specialization

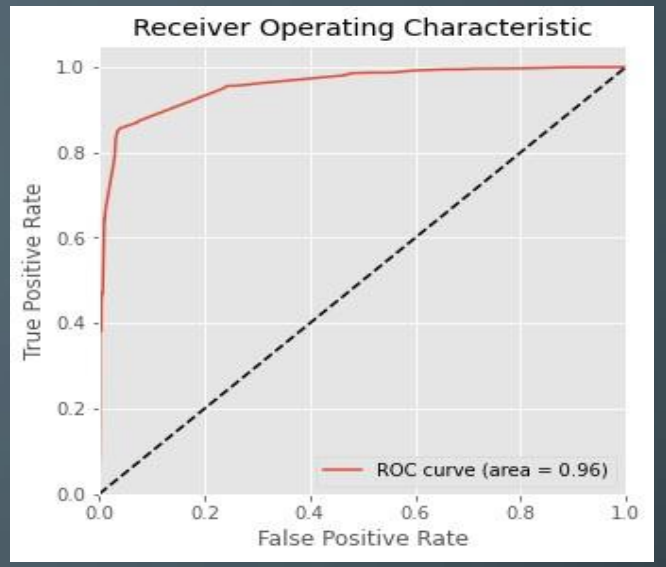
Looking at above plot, we can say that working professionals have high conversion rate

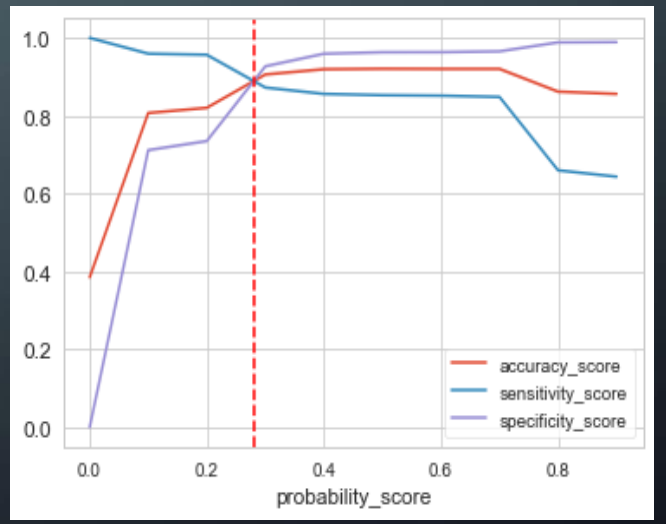
Number of Unemployed leads are more than any other category

To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads

MODEL BUILDING

- SPLITTING THE DATA INTO TEST AND TRAINING SETS
- WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO 70:30
- USING RFE TO CHOOSE TOP 15 VARIABLES
- BUILD MODEL BY REMOVING THE VARIABLE WHOSE P-VALUE > 0.05 AND VIF>5
- PREDICTIONS ON TEST DATASET
- OVERALL 92%





MODEL EVALUATION (TRAIN DATASET)

- CALCULATED ACCURACY, SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITY CUTOFF FROM 0.1 TO 0.9
- AS PER GRAPH AND LOOKING AT THE OTHER SCORES IT CAN BE SEEN THAT THE OPTIMAL POINT IS 0.28

	probability_score	accuracy_score	sensitivity_score	specificity_score
0.0	0.0	0.385136	1.000000	0.000000
0.1	0.1	0.807117	0.959526	0.711652
0.2	0.2	0.820343	0.956664	0.734955
0.3	0.3	0.905999	0.872445	0.927017
0.4	0.4	0.919540	0.856092	0.959283
0.5	0.5	0.920642	0.852821	0.963124
0.6	0.6	0.920328	0.851594	0.963380
0.7	0.7	0.920328	0.848324	0.965429
0.8	0.8	0.861912	0.659853	0.988476
0.9	0.9	0.856086	0.643500	0.989245

TRAIN DATA – CONFUSION MATRIX

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
CONVERTED	124	2322

ACCURACY 83.59%

SPECIFICITY 76.5%

MODEL PREDICTION (TEST DATASET)

TEST DATA – CONFUSION MATRIX

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
CONVERTED	71	918

Tags_Lost to EINS	9.58
Tags_Closed by Horizzon	8.56
Tags_Will revert after reading the email	3.83
Tags_Busy	3.65
Lead Source_Welingak Website	3.22
Last Activity_SMS Sent	1.93
Lead Origin_Lead Add Form	0.91
Do Not Email	-1.18
Last Notable Activity_Olark Chat Conversation	-1.30
Last Notable Activity_Modified	-1.68
Tags_Ringing	-1.77
Tags_switched off	-2.34
Lead Quality_Not Sure	-3.48
Lead Quality_Worst	-3.94
dtype: float64	

ACCURACY	**81.56%**
SPECIFICITY	75.14%

CONCLUSION

- THE LOGISTIC REGRESSION MODEL IS USED TO PREDICT THE PROBAILITY OF A CUSTOMER
- WHILE WE HAVE CALCULATED BOTH SENSITIVITY- SPECEIFICITY AS WELL AS PRECISON-RECALL METRICS ,WE HAVE CONSIDERED OPTIMAL CUT OFF ON THE BASIS OF SENSITIVITY- SPECIFICITY FOR FINAL PREDICTION
- LEAD SCORE CALCULATED SHOWS THE CONVERSION RATE OF FINAL PREDICTED MODEL IS AROUND 92% IN TESTDATA AS COMPARED TO 94%IN TRAIN DATA
- IN BUISNESS TERMS. THIS MODEL HAS CAPABILITY TO ADJUST WITH COMPANY'S REQUIREMNETS IN COMING FUTURE
- TOP FEATURES

1. TAGS_LOST TO EINS
2. TAGS_CLOSED BY HORIZZON
3. TAGS_WILL REVERT AFTER READING THE EMAIL

HENCE OVERALL THIS MODEL SEEMS TO BE GOOD…