LEAD SCORE CASE STUDY

PROBLEM STATEMENT

- 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%

OVERALL APPROACH

1. DATA CLEANING AND IMPUTING MISSING VALUES
2. EXPLORATORY DATA ANALYSIS : UNIVARIATE , BIVARIATE and MULTIVARIATE ANALYSIS
3. FEATURE SCALING AND DUMMY VARIABLE CREATION
4. LOGISTIC REGRESSION MODEL BUILDING
5. MODEL EVALUATION : SPECIFICITY , SENSITIVITY, PRECISION and RECALL
6. CONCLUSION AND RECOMMENDATION

PROBLEM SOLVING METHODOLOGY

DATA CLEANING AND PREPARATION

- Read data from source
- Convert data into clean for mat 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

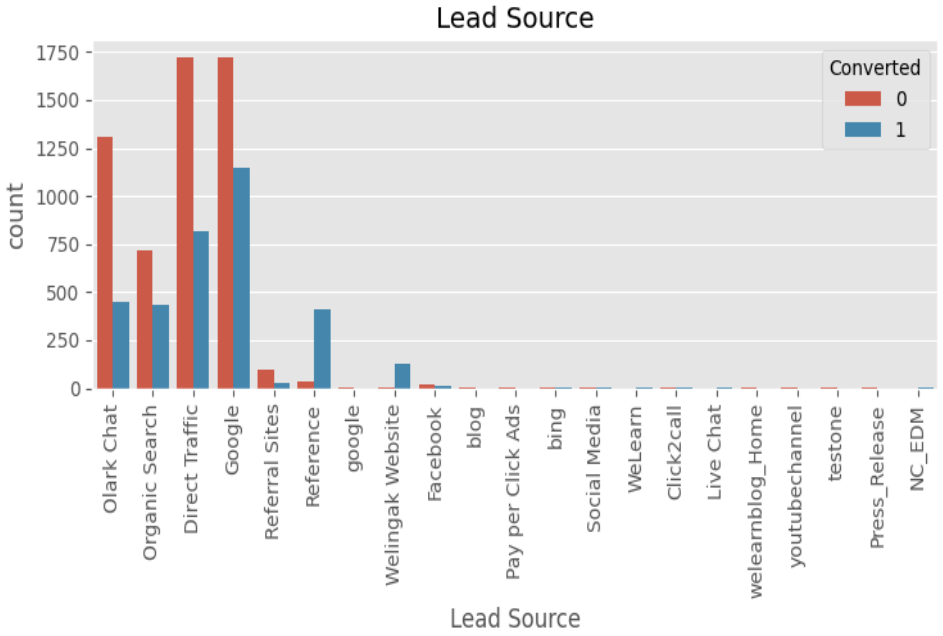
2. CONVERTING THE ‘SELECT’ VALUES WITH NaNs

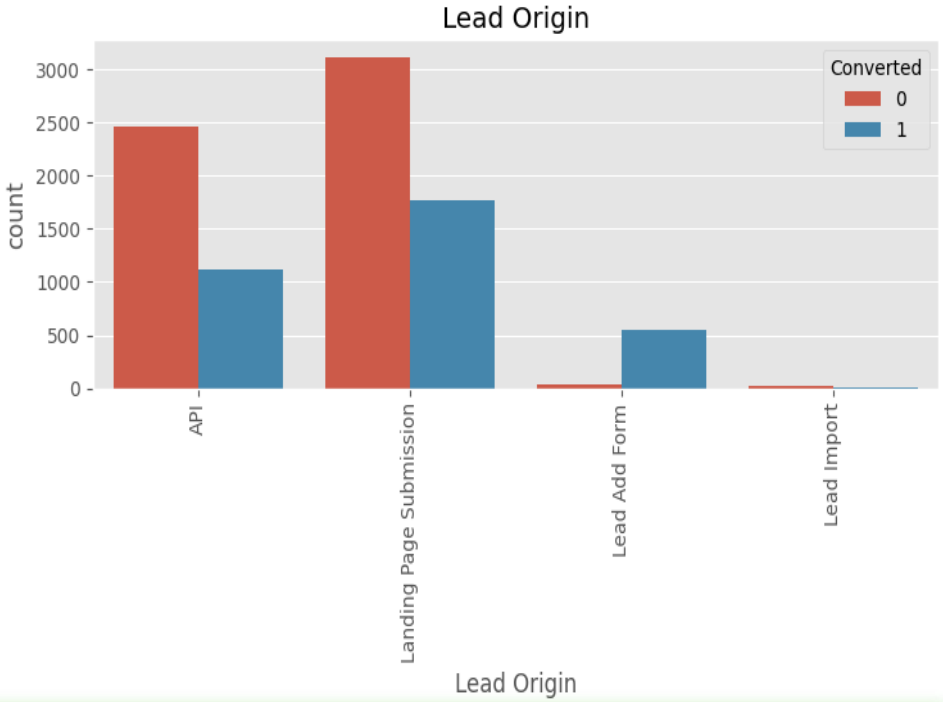
3. DROPIING THE COLUMNS HAVING >70% OF NULL VALUES

4. DROPPING UNNECESSARY COLUMNS

5. DROPPING THE ROWS AS THE NULL VALUES WERE

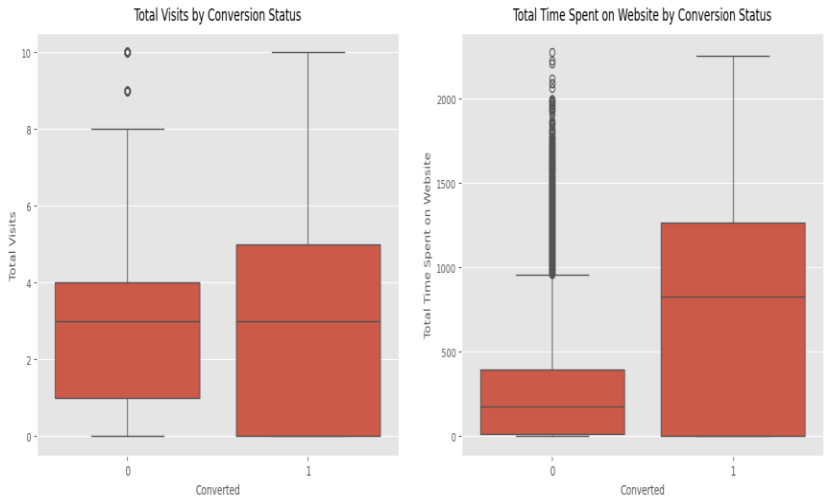
EXPLORATORY DATA ANALYSIS

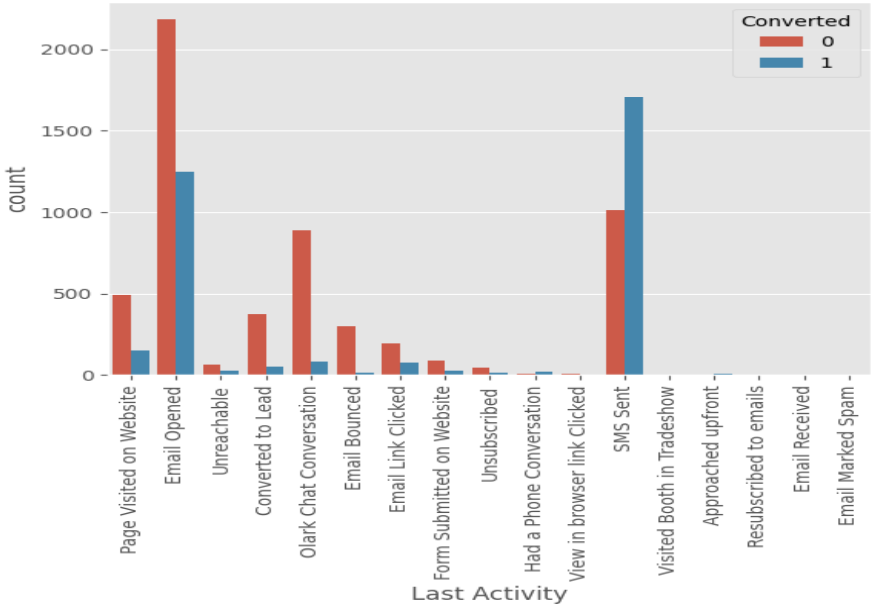




- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

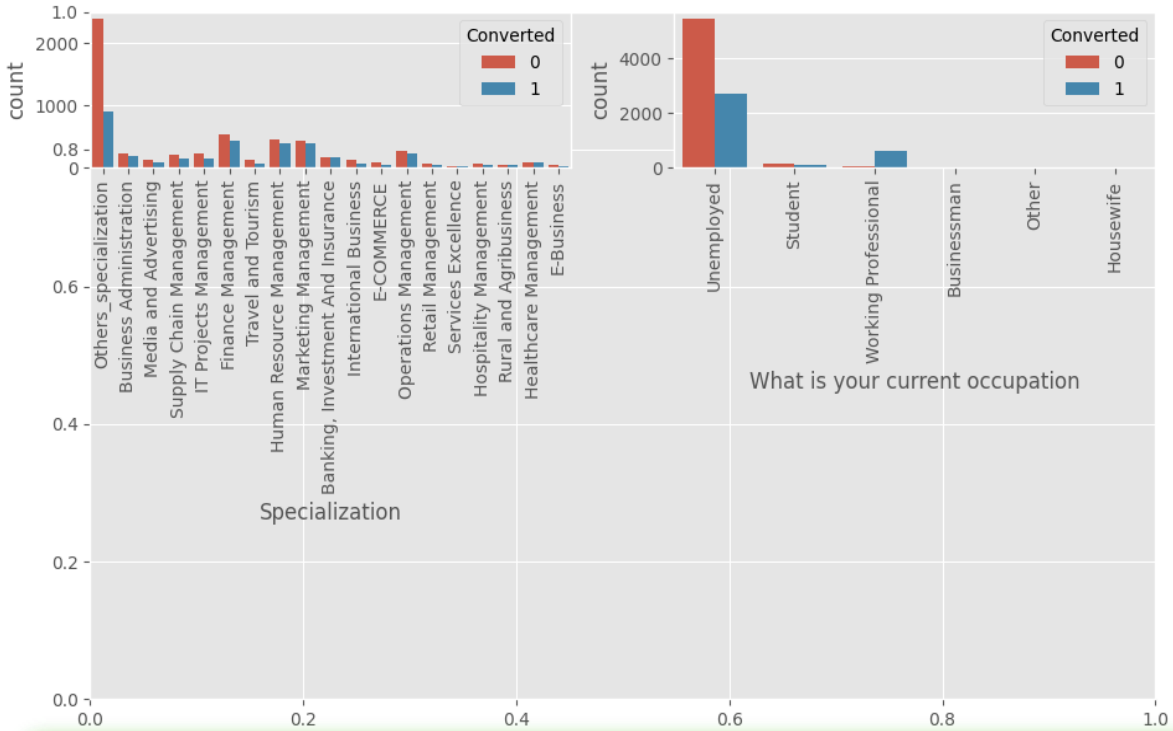
EXPLORATORY DATA ANALYSIS

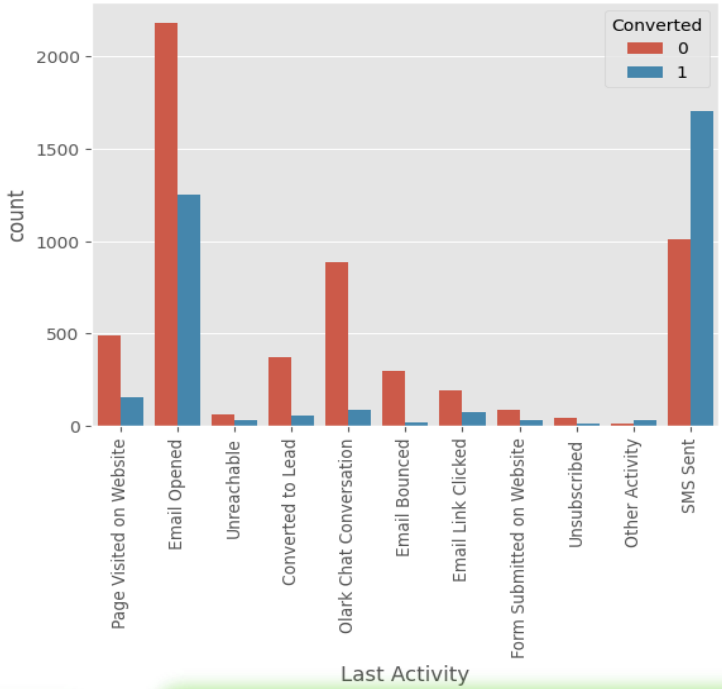




- The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information
- Users spending more time on the website are more likely to get converted

- The count of lead's last activity as "Email Opened" is maximum
- The conversion rate of SMS sent as last activity is maximum





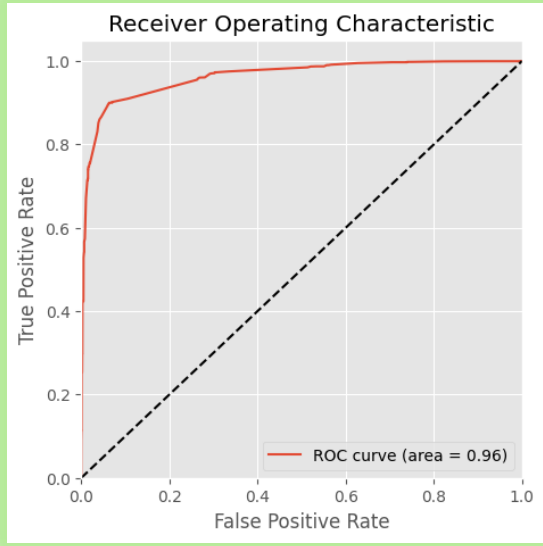
- Looking at above plot, no particular inference can be made for Specialization
- Looking at above plot, we can say that working professionals have high conversion rate
- Number of Unemployed leads are more than any other category

- 'Will revert after reading the email' and 'Closed by Horizzon' has high conversion rate

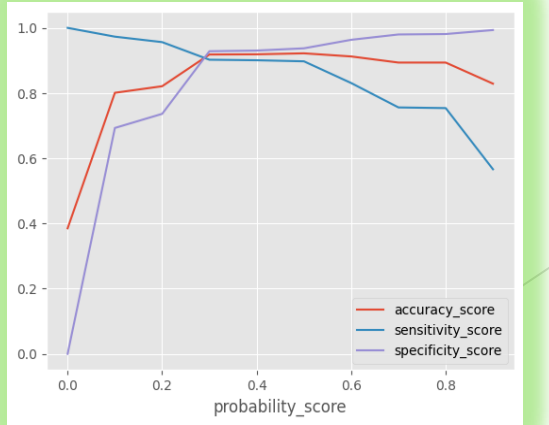
MODEL BUILDING

- 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
- PREDICTIONS ON TEST DATASET
- BUILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE > 0.05 AND VIF > 5
- OVERALL ACCURACY IS 92.0 %

ROC CURVE



OPTIMAL CUT-OFF



MODEL EVALUATION

- CALCULATED ACCURACY, SENSITIVITY AND SPECIFICITYFOR 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 0.27

	Probability_Score	Accuracy_Score	Sensitivity_Score	Specificity_Score	Precision_Score
0.0	0.0	0.385136	1.000000	0.000000	0.385136
0.1	0.1	0.800819	0.973017	0.692958	0.664990
0.2	0.2	0.820973	0.956255	0.736236	0.694271
0.3	0.3	0.918281	0.902289	0.928297	0.887415
0.4	0.4	0.918910	0.900654	0.930346	0.890101
0.5	0.5	0.921902	0.897383	0.937260	0.899590
0.6	0.6	0.912297	0.830744	0.963380	0.934253
0.7	0.7	0.893560	0.755928	0.979770	0.959025
0.8	0.8	0.893560	0.753884	0.981050	0.961418
0.9	0.9	0.828846	0.566231	0.993342	0.981573

TRAIN DATA - CONFUSION MATRIX

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	3621	284
CONVERTED	239	2207

ACCURACY	91.77%
PRECISION	88.60%
SENSITIVITY	90.23%
SPECIFICITY	92.73%

TOP FEATURES

```
----------------------Feature Importance--------------------
const                                                -1.791204
Total Time Spent on Website                           1.395719
Lead Origin_Lead Add Form                             1.311767
Lead Source_Welingak Website                          3.066837
Last Activity_SMS Sent                                1.810108
Tags_Busy                                             3.598185
Tags_Closed by Horizzon                               8.653719
Tags_Lost to EINS                                     9.532070
Tags_Ringing                                         -1.626147
Tags_Will revert after reading the email              3.890666
Tags_switched off                                    -2.312673
Lead Quality_Not Sure                                -3.364760
Lead Quality_Worst                                   -3.949331
Last Notable Activity_Modified                       -1.740279
Last Notable Activity_Olark Chat Conversation        -1.470152
```

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	393	1341
CONVERTED	34	955

ACCURACY	49.50%
PRECISION	41.59%
SENSITIVITY	96.56%
SPECIFICITY	22.66%

CONCLUSION

●The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.

● Here, the logistic regression model is used to predict the probabilty of conversion of a customer.

● Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert)

● Our final Logistic Regression Model is built with 14 features.

● Features used in final model are ['Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_switched off', 'Lead Quality_Not Sure', 'Lead Quality_Worst', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation']

● The top three categorical/dummy variables in the final model are 'Tags_Lost to EINS', 'Tags_Closed by Horizzon', 'Lead Quality_Worst' with respect to the absolute value of their coefficient factors.

● 'Tags_Lost to EINS', 'Tags_Closed by Horizzon' are obtained by encoding original categorical variable 'Tags'. 'Lead Quality_Worst' is obtained by encoding the categorical variable 'Lead Quality'.

- Tags_Lost to EINS (Coefficient factor = 9.532070)
- Tags_Closed by Horizzon (Coefficient factor = 8.653719)
- Lead Quality_Worst (Coefficient factor =-3.949331)
- The final model has Sensitivity of 0.9656, this means the model is able to predict 96.56% customers out of all the converted customers, (Positive conversion) correctly.
- The final model has Precision of 0.4159, this means 41.59% of predicted hot leads are True Hot Leads.
- We have also built an reusable code block which will predict Convert value and Lead Score given training, test data and a cut-off. Different cutoffs can be used depending on the use-cases (for eg. when high sensitivity is required, when model have optimum precision score etc.)