Lead Score Case Study

BY- AARADHANA, SONI AND SOURASIS

PROBLEM STATEMENT

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 websites, 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 lead. Moreover, the company also gets leads through the referrals.

Once the leads are acquired, employees from the sales 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 knowns as “Hot Leads”

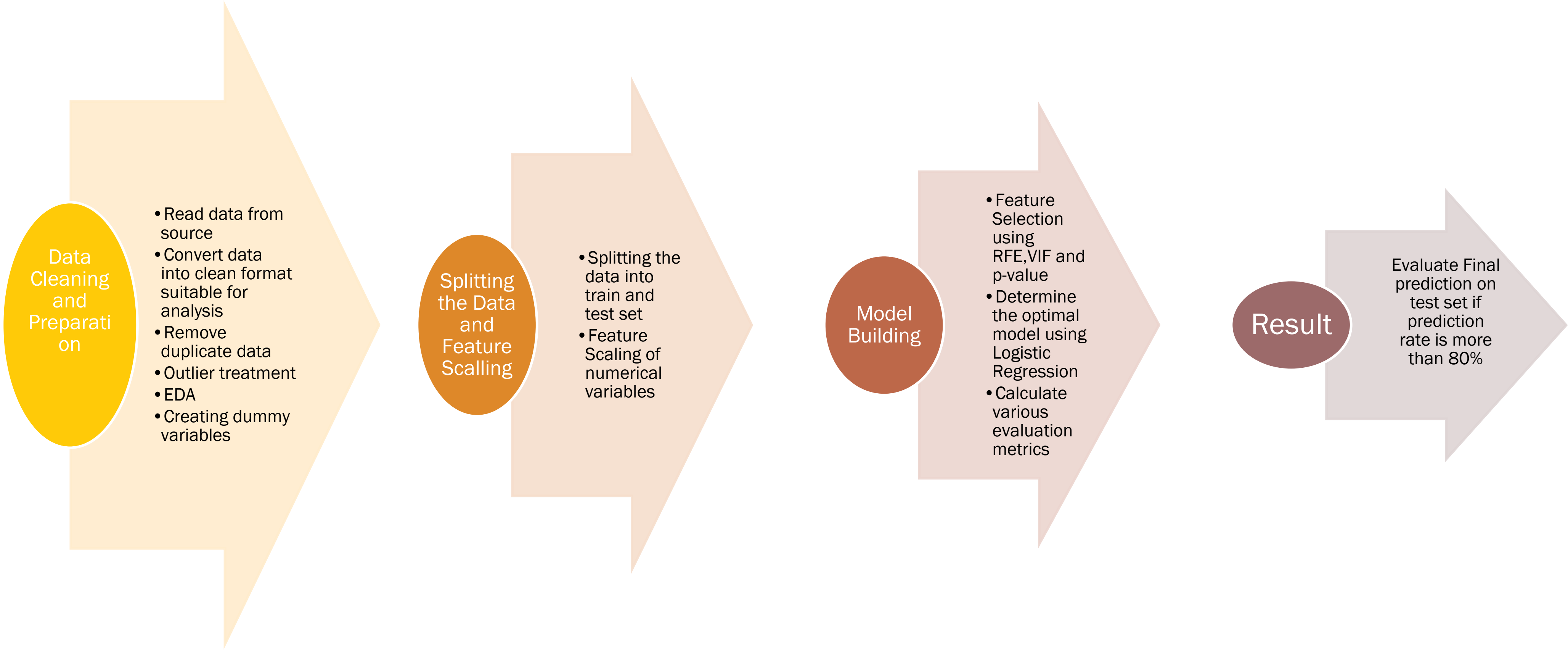
The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance.

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
3. FEATURE SCALING AND DUMMY VARIABLE CREATION
4. LOGISTIC REGRESSION MODEL BUILDING
5. MODEL EVALUATION: ACCURACY, SENSITIVITY and SPECIFICITY
6. CONCLUSION AND RECOMMENDATION

PROBLEM SOLVING METHODOLOGY



Data Cleaning and Preparati on

- Read data from source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier treatment
- EDA
- Creating dummy variables

Splitting the Data and Feature Scalling

- Splitting the data into train and test set
- Feature Scaling of numerical variables

Model Building

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

Result

Evaluate Final prediction on test set if prediction rate is more than 80%

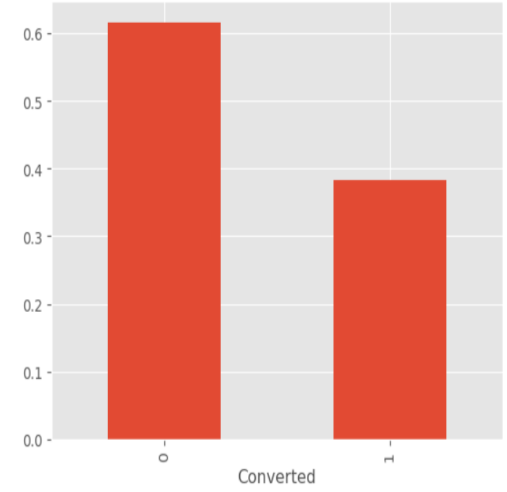
DATA CONVERSION

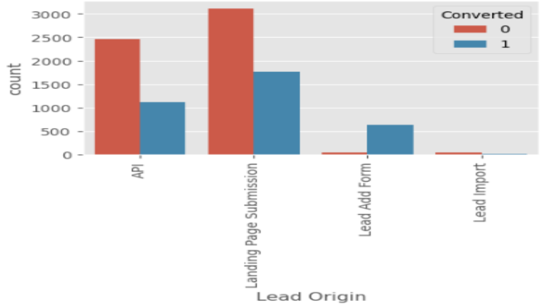
1. CONVERTING THE VARIABLES WITH VALUES YES/NO TO 1/0s
2. CONVERTING THE ‘Select’ VALUES WITH NaNs
3. DROPPING THE COLUMNS HAVING >70% OF NULL VALUES
4. DROPPING UNNESSARY COLUMNS
5. MODEL EVALUATION: ACCURACY, SENSITIVITY and SPECIFICITY
6. IMPUTING ALL THE NULL VALUE TO MEDIAN AND MODE

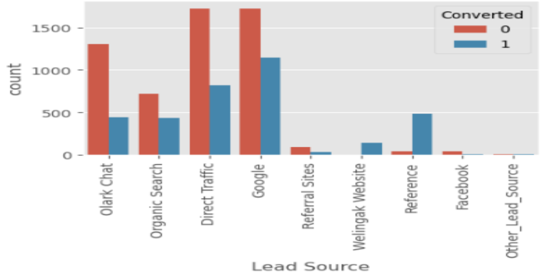
EXPLORATORY DATA ANALYSIS

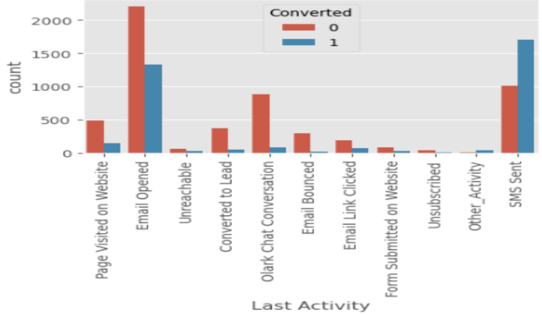
We have around 38% conversation rate data with us

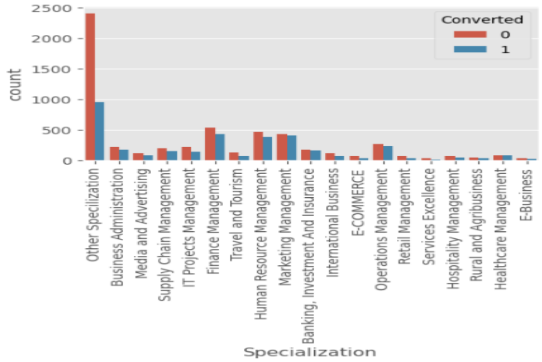
- The count of leads from Add Form is pretty low but the conversion rate is very high.
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable
- The conversion rate of the leads from Reference and Welingak Website is maximum
- The count of leads from Google and Direct Traffic is maximum
- The conversion rate of SMS sent as Last activity is maximum
- Can't find any references for specialization as maximum leads and conversions doesn't have tags.





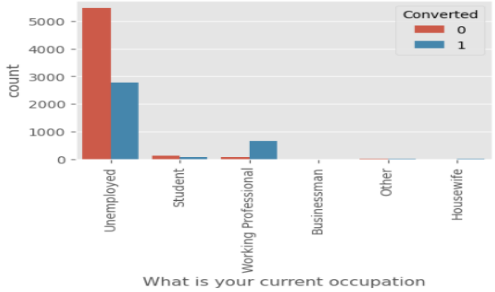


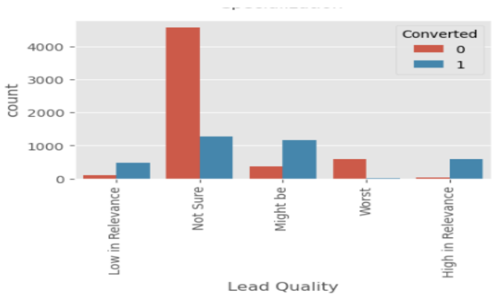


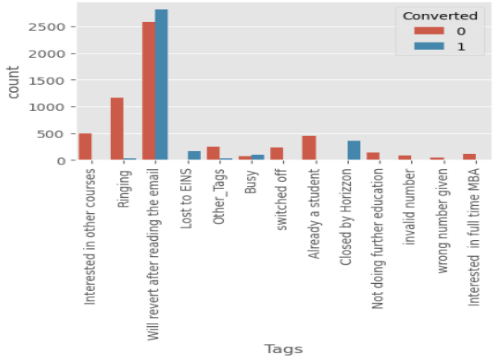


EXPLORATORY DATA ANALYSIS

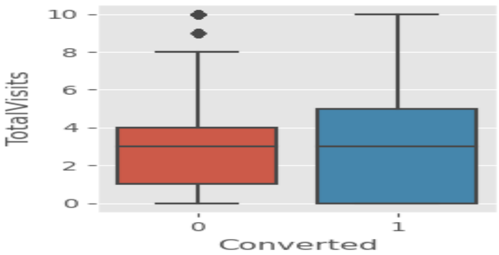
- ➢ Working professional has high conversion rate
- ➢ ‘Will revert after reading the email’ and ‘Closed by Horizon’ has high conversion rate

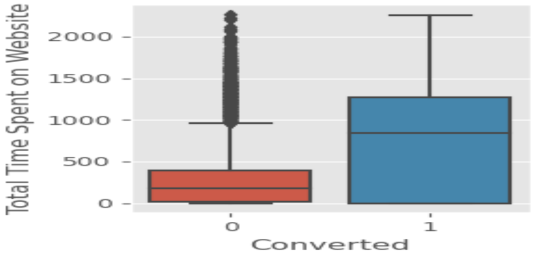


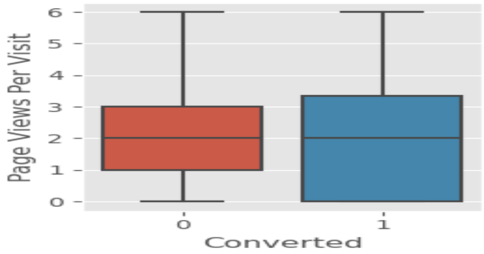




- ➢ Total view/Page Views per visit: The median of both conversion and non-conversion are same and hence nothing conclusive can be said using this information.
- ➢ User spending more time on website are more likely to be converted.

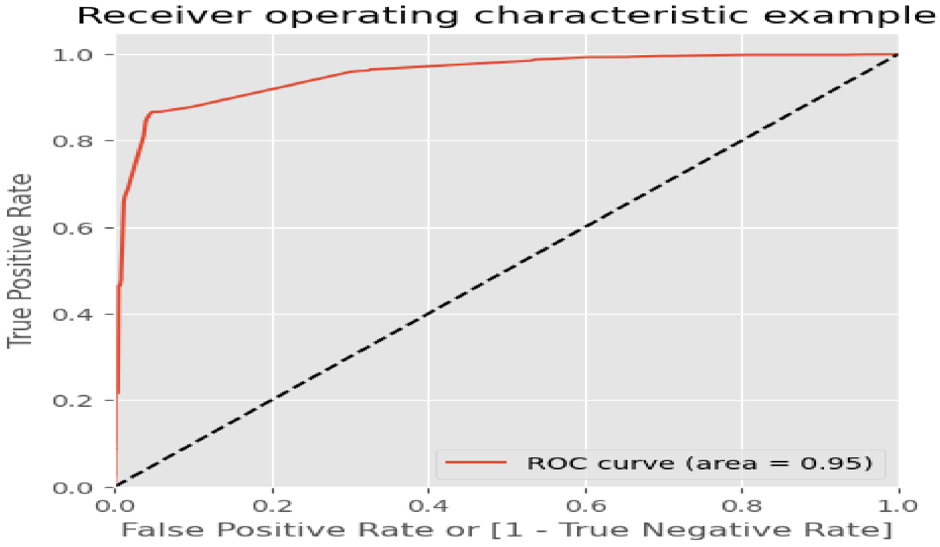


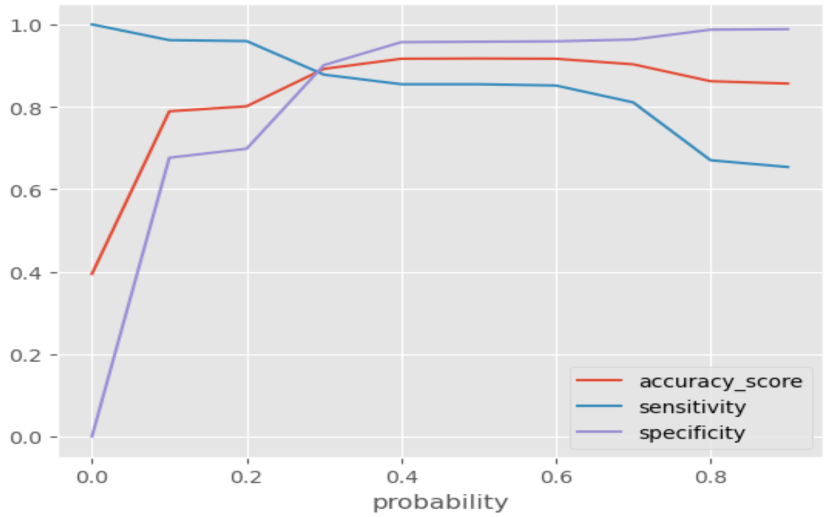




MODEL BUILDING

- Splitting the data into TEST and TRAIN sets
- We have chosen the TRAIN_TEST split ration as 70:30
- Using RFE to choose Top 15 Variables
- Build Model by removing variables whose p-value >.05 and VIF>5
- Plotted ROC curve and visualize AUC is 95%
- From train set accuracy score, sensitivity and specificity we got cutoff as .27
- Prediction on Test Dataset with cutoff .27
- Overall accuracy is 89% on Test set

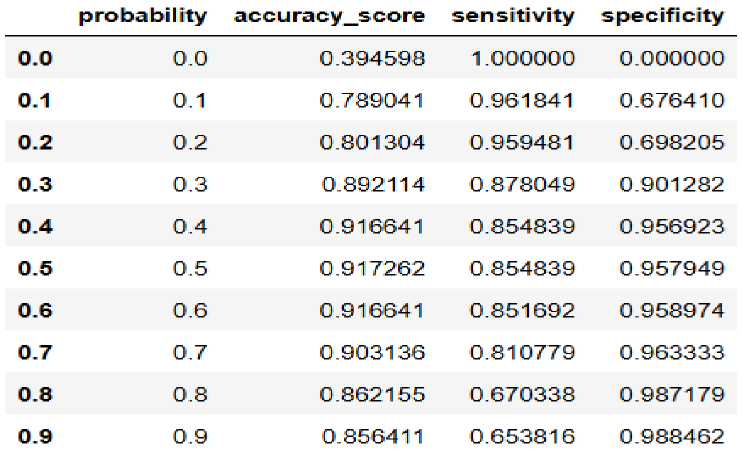




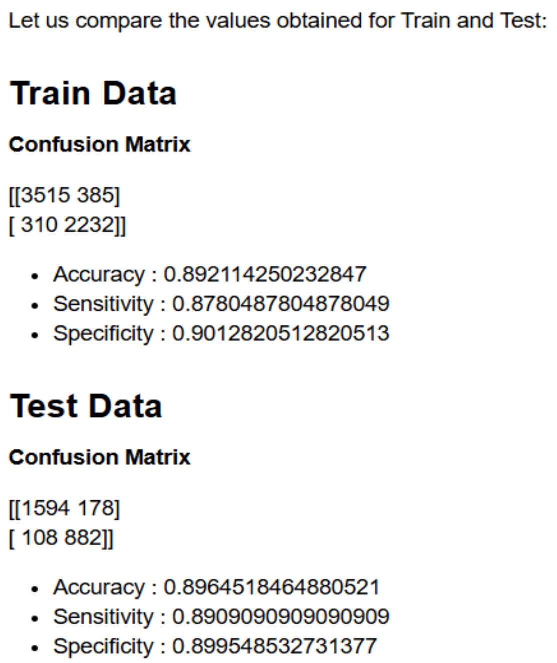
From the curve above, 0.27 is the optimum point to take it as a cutoff probability.

MODEL EVALUATION AMD PREDICTION

- Created confusion matrix's to calculate TPR, FPR, Sensitivity, Specificity and confusion Matrix
- Calculated ACCURACY, SENSITIVITY and SPECIFICITY for various probability cutoffs from .1 to .9
- As per the graph and looking at the others scores, it can be seen that the optimal points is .27



	probability	accuracy_score	sensitivity	specificity
0.0	0.0	0.394598	1.000000	0.000000
0.1	0.1	0.789041	0.961841	0.676410
0.2	0.2	0.801304	0.959481	0.698205
0.3	0.3	0.892114	0.878049	0.901282
0.4	0.4	0.916641	0.854839	0.956923
0.5	0.5	0.917262	0.854839	0.957949
0.6	0.6	0.916641	0.851692	0.958974
0.7	0.7	0.903136	0.810779	0.963333
0.8	0.8	0.862155	0.670338	0.987179
0.9	0.9	0.856411	0.653816	0.988462



Let us compare the values obtained for Train and Test:

Train Data

Confusion Matrix

```
[[3515 385]
 [ 310 2232]]
```

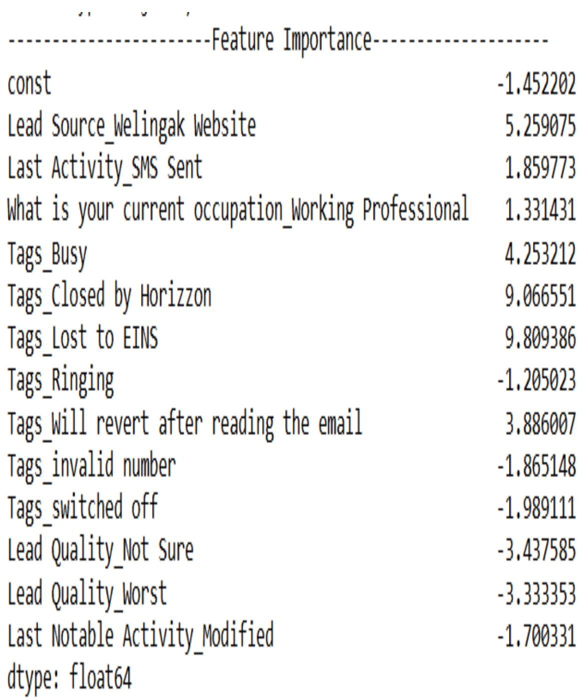
- Accuracy : 0.892114250232847
- Sensitivity : 0.8780487804878049
- Specificity : 0.9012820512820513

Test Data

Confusion Matrix

```
[[1594 178]
 [ 108 882]]
```

- Accuracy : 0.8964518464880521
- Sensitivity : 0.8909090909090909
- Specificity : 0.899548532731377



```
---------------------Feature Importance------------------
const                                                 -1.452202
Lead Source_Welingak Website                           5.259075
Last Activity_SMS Sent                                1.859773
What is your current occupation_Working Professional  1.331431
Tags_Busy                                             4.253212
Tags_Closed by Horizzon                               9.066551
Tags_Lost to EINS                                     9.809386
Tags_Ringing                                         -1.205023
Tags_Will revert after reading the email              3.886007
Tags_invalid number                                  -1.865148
Tags_switched off                                    -1.989111
Lead Quality_Not Sure                                -3.437585
Lead Quality_Worst                                   -3.333353
Last Notable Activity_Modified                       -1.700331
dtype: float64
```

CONCLUSION

- The logistic regression model is used to predict the probability of conversion of a customer.
- We have considered optimal cutoff on the basis of sensitivity-specificity
- Lead score calculated shows the conversion rate of final predicted model is around 89% in train and test data.
- Hence Overall this model seems to be good

Thank you