LEAD SCORE CASE STUDY

- Arpita Shirol

PROBLEM STATEMENT

UCTION:

tion company, X Education sells online courses to industry professionals. The company markets its courses on various websites and sea uch as Google

ple 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 their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals

se leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X educa 0%

S GOALS:

wishes to identify the most potential leads, also known as "Hot Leads"

any 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 n chance and customer with lower lead score have a lower conversion chance

n particular, has given a ballpark number for the lead conversion rate i.e. 80%

OVERALL APPROACH

EANING AND IMPUTING MISSING VALUES

TORY DATA ANALYSIS : UNIVARIATE , BIVARIATE and MULTIVARIATE ANALYSIS

E SCALING AND DUMMY VARIABLE CREATION

C REGRESSION MODEL BUILDING

EVALUATION : SPECIFICITY , SENSITIVITY, PRECISION and RECALL

SION AND RECOMMENDATION

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

5

RTING THE VARIABLE WITH VALUES YES/NO to 1/0s

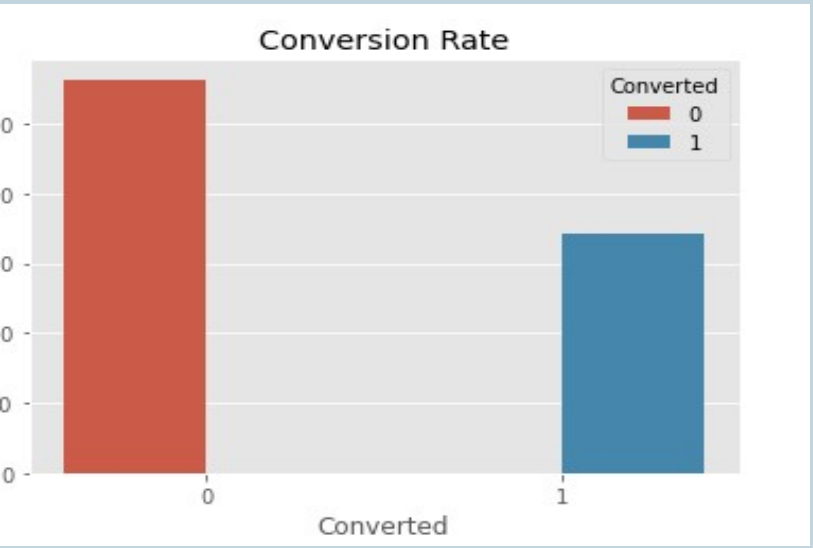
RTING THE 'SELECT' VALUES WITH NaNs

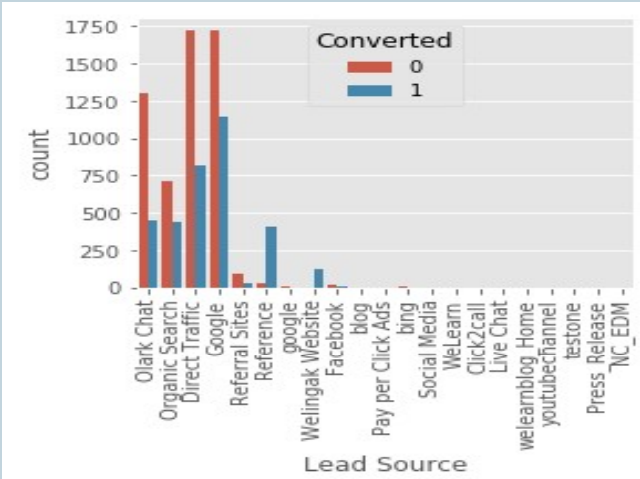
NG THE COLUMNS HAVING >40%OF NULL VALUES

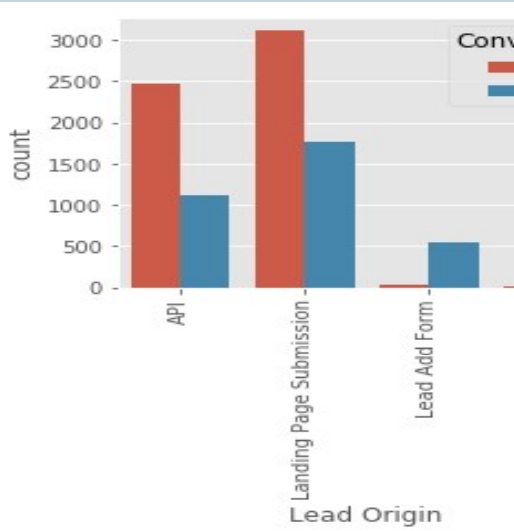
ING UNNECESSARY COLUMNS

ING THE ROWS AS THE NULL VALUES WERE < 2%

EXPLORATORY DATA ANALYSIS



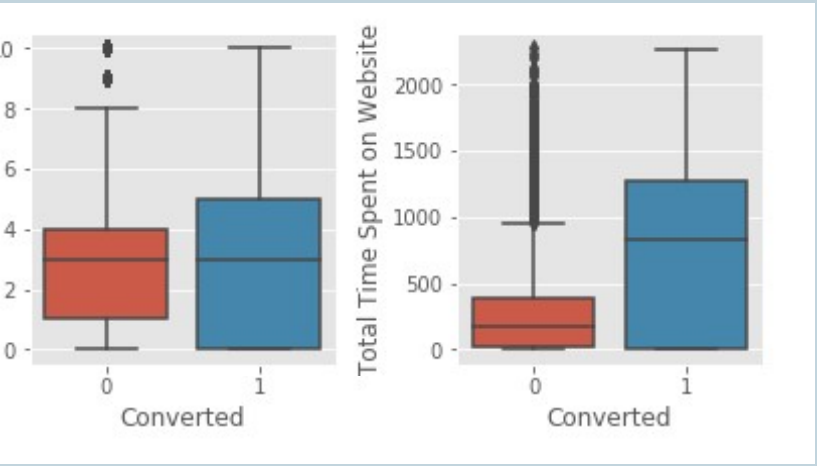




We have around 30% of Conversion Rate

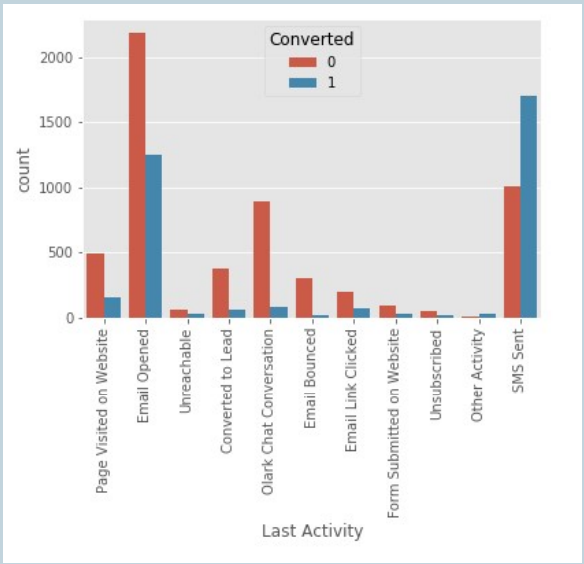
- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from th considerable
- The count of leads from the Lead Add Form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

EXPLORATORY DATA ANALYSIS



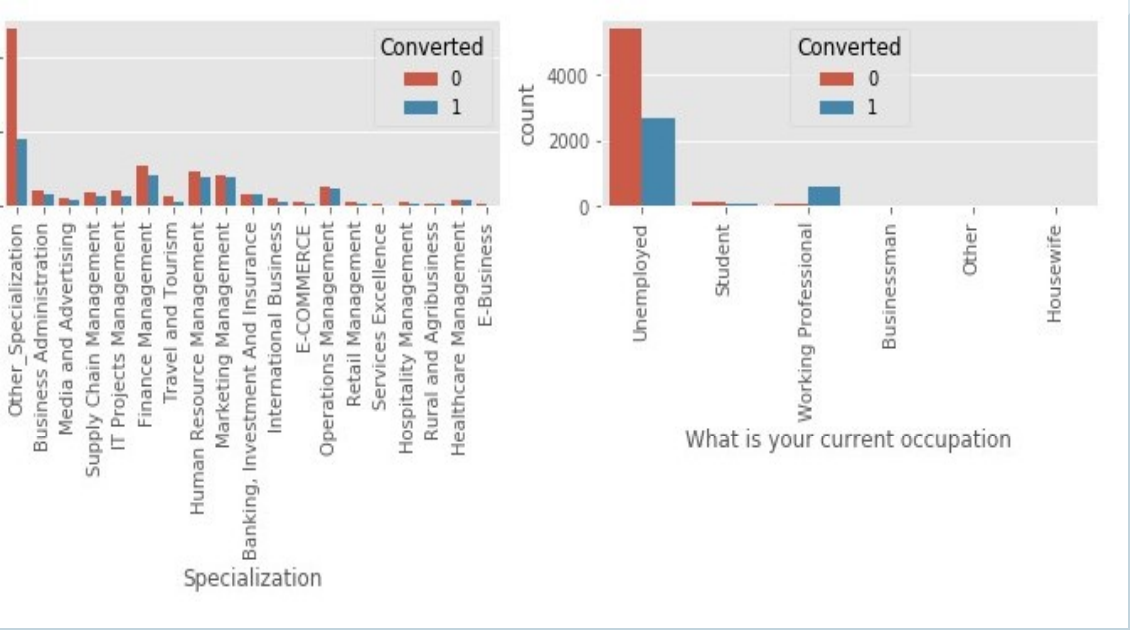
median of both the conversion and non-conversion are same and
e nothing conclusive can be said using this information

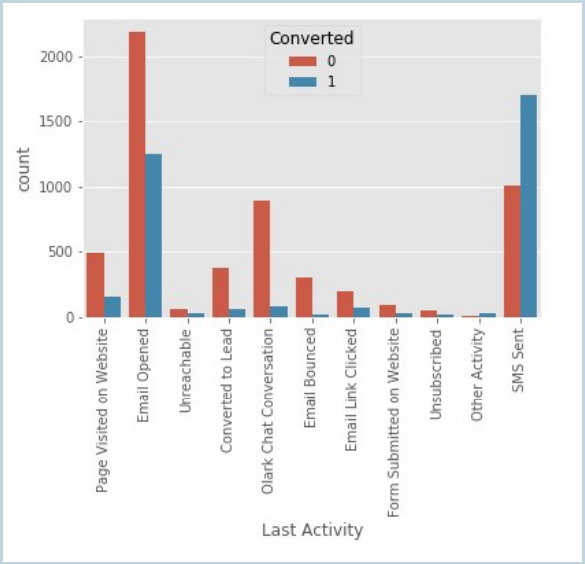
s spending more time on the website are more likely to get converted



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

EXPLORATORY DATA ANALYSIS





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

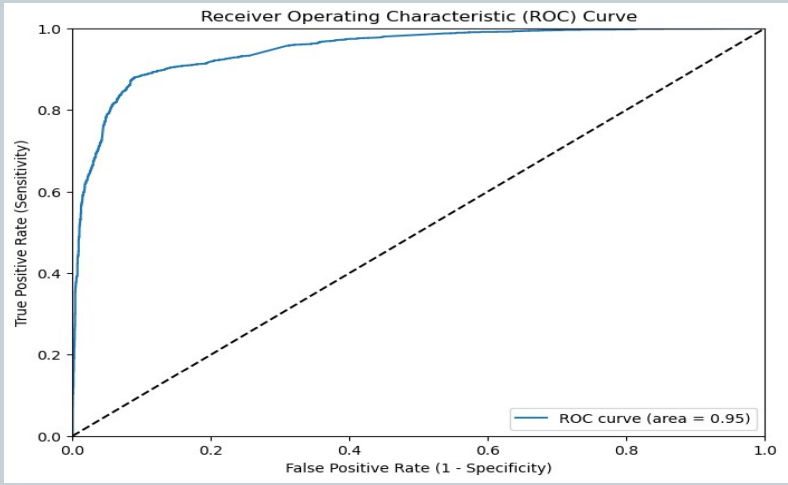
ng at above plot, we can say that working professionals have high conversion rate

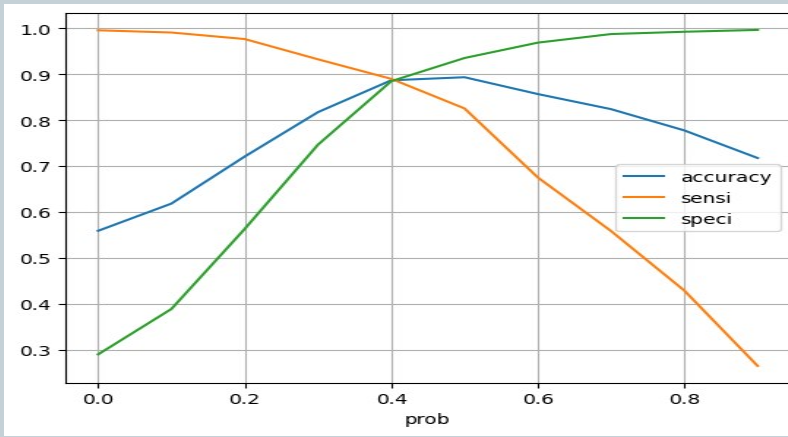
er of Unemployed leads are more than any other category

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

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 88.74%%





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

	prob	accuracy	sensi	speci
0.0	0.0	0.558905	0.995945	0.289605
0.1	0.1	0.618275	0.991079	0.388556
0.2	0.2	0.720779	0.976886	0.562969
0.3	0.3	0.817563	0.932685	0.746627
0.4	0.4	0.886982	0.890105	0.885057
0.5	0.5	0.893630	0.825629	0.935532
0.6	0.6	0.856988	0.675182	0.969015
0.7	0.7	0.824057	0.558394	0.987756
0.8	0.8	0.777675	0.428629	0.992754
0.9	0.9	0.717532	0.264396	0.996752

TRAIN DATA - CONFUSION MATRIX

PREDICTED / ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	3744	258
CONVERTED	430	2036

ACCURACY	88.74%
PRECISION	82.98%
SENSITIVITY	89.95%
SPECIFICITY	87.95%

MODEL PREDICTION

TOP FEATURES

Feature Importance	
const	-1.248649
Do Not Email	-1.180501
Lead Origin_Lead Add Form	0.908052
Lead Source_Welingak Website	3.218160
Last Activity_SMS Sent	1.927033
Tags_Busy	3.649486
Tags_Closed by Horizzon	8.555901
Tags_Lost to EINS	9.578632
Tags_Ringing	-1.771378
Tags_Will revert after reading the email	3.831727
Tags_switched off	-2.336683
Lead Quality_Not Sure	-3.479228
Lead Quality_Worst	-3.943680
Last Notable Activity_Modified	-1.682075
Last Notable Activity_Olark Chat Conversation	-1.304940

TEST DATA - CONFUSION MATRIX

PREDICTED / ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	1475	202
CONVERTED	110	985

ACCURACY	89.36%
PRECISION	88.7%
SENSITIVITY	82.56%
SPECIFICITY	93.55%

CONCLUSION

logistic regression model is used to predict the probabilIty of conversion of a customer.

le we have calculated both **sensitivity-specificity** as well as **Precision-Recall** metrics, we have considered optimal cu
basis of **sensitivity-specificity** for final prediction

d Score calculated shows the conversion rate of final predicted model is around **92% in test data**
pared to **95% in train data** In Business terms, this model has capability to adjust with the
ny's requirements in coming future

variables that contributes for lead getting converted in the model are:

gs_Lost to EINS

gs_Closed by Horizzon

ad Quality_Worst

ce Overall this model seems to be good