IIITB Data Science
Batch C44

Lead Score Case Study

A Bhargav Kumar.
Siddhesh Parab.
Divya Shah.

upGrad

Pioneering Excellence in Education, Research & Innovation

Problem Statement

- An Education Company, X Education sells online courses to industry professionals. The company markets its websites on various websites and search engines like Google.
- Once a users land on the website they might browse some courses, fill up forms or watch some introductory videos or chats with a chatbot. When a user fills up the form putting personal details like email ID and phone number, he/she is classified as a lead. Moreover, company also gets leads through past referrals.
- Once these leads are acquired, the company starts reaching out to them via different medium.
- The typical lead conversion rate at X Education is 30%.

Business Goals

- Build a logistic regression model to assign a lead score between 0 to 100 to each of the leads which be used a target. Higher score would mean lead is hot and can be converted whereas lower score means lead is cold that the customer will not convert.
- Based on model, suggest a hiring strategy for sales department for aggressive lead conversions.
- Suggest a strategy in order to minimize the rate of useless phone calls.

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 using Metrics
6. Conclusion and Recommendation

Problem Solving Methodology

Data Cleaning

- Reading Data
- Cleaning the data
- Null values treatment
- Outlier Treatment
- EDA

Data Preparation

- Splitting Data into train and test dataset
- Feature Scaling of Numerical Values

Model Building

- Feature Selection using RFE, VIF and p-value.
- Determining optimal model using Logistic Regression

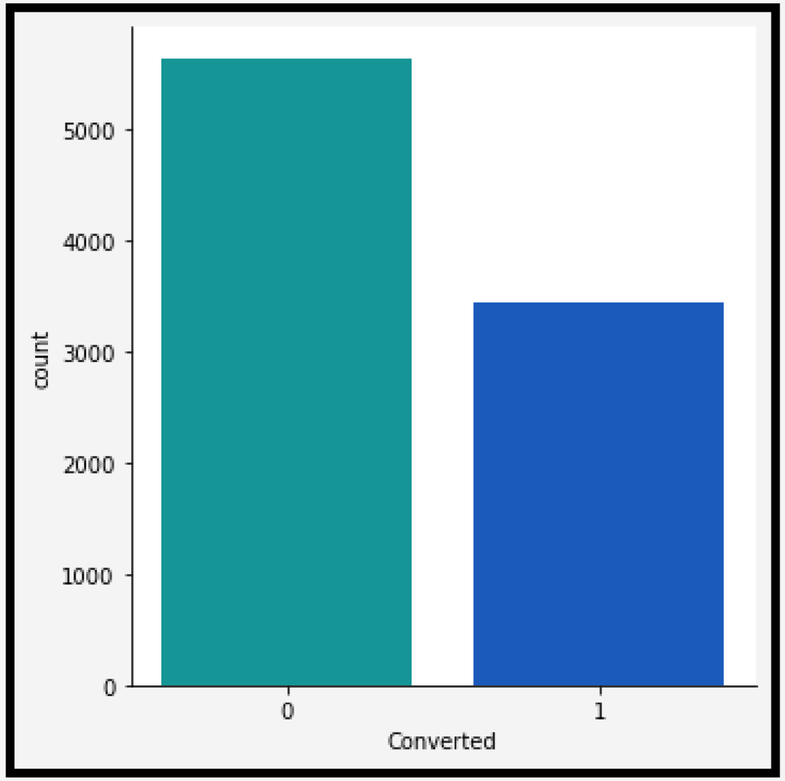
Model Evaluation

- Plotting ROC Curve, Optimal Cut-off and Precision-Recall Curve.
- Predictions Using final LR Model.

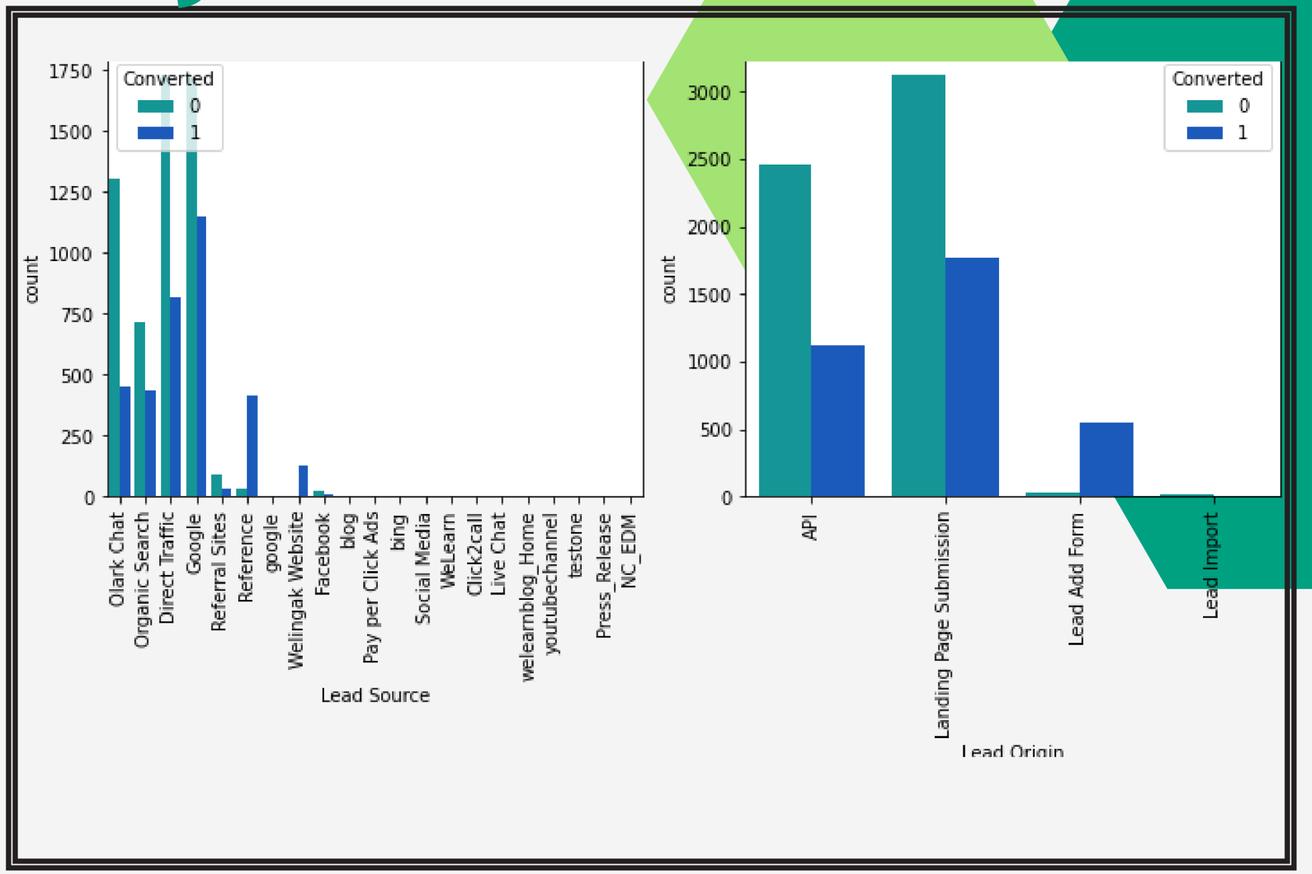
Data Cleaning

- Converting categorical variables having values Yes / No with 1 / 0
- Converting 'Select' values to 'Nan'
- Dropping columns having more than 70% of null values
- Dropping unnecessary columns
- Dropping rows where null values are less that 2%

Exploratory Data Analysis

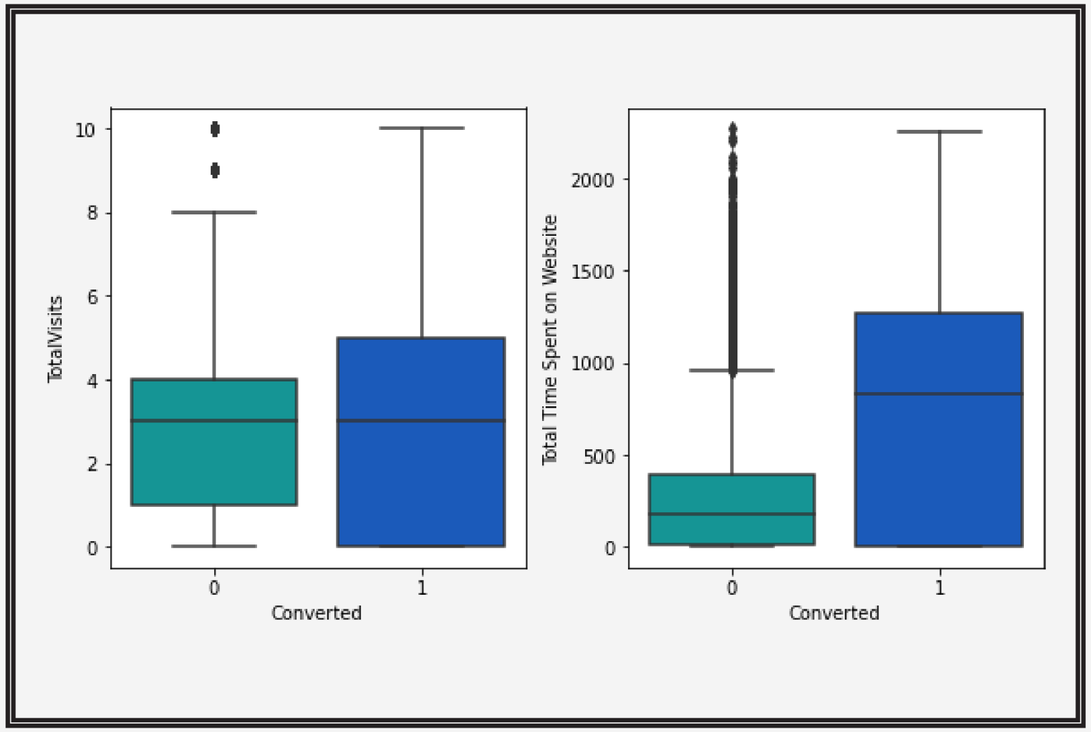


The conversion rate is around 30%.

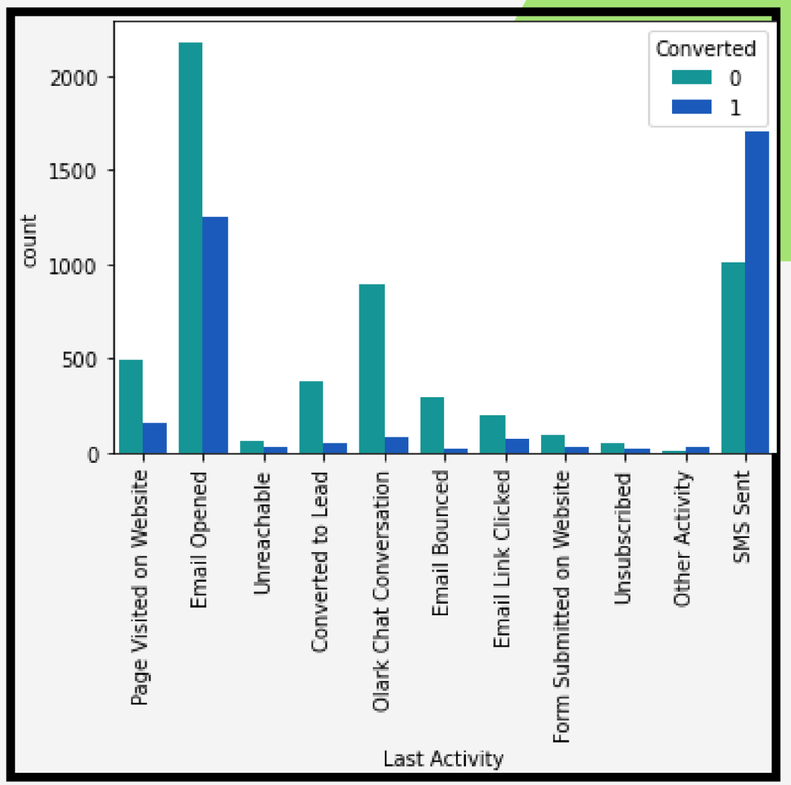


- The count of leads from Google and Direct Traffic is Maximum.
- The conversion rate of the leads from reference and Welingak Website is maximum.

Exploratory Data Analysis

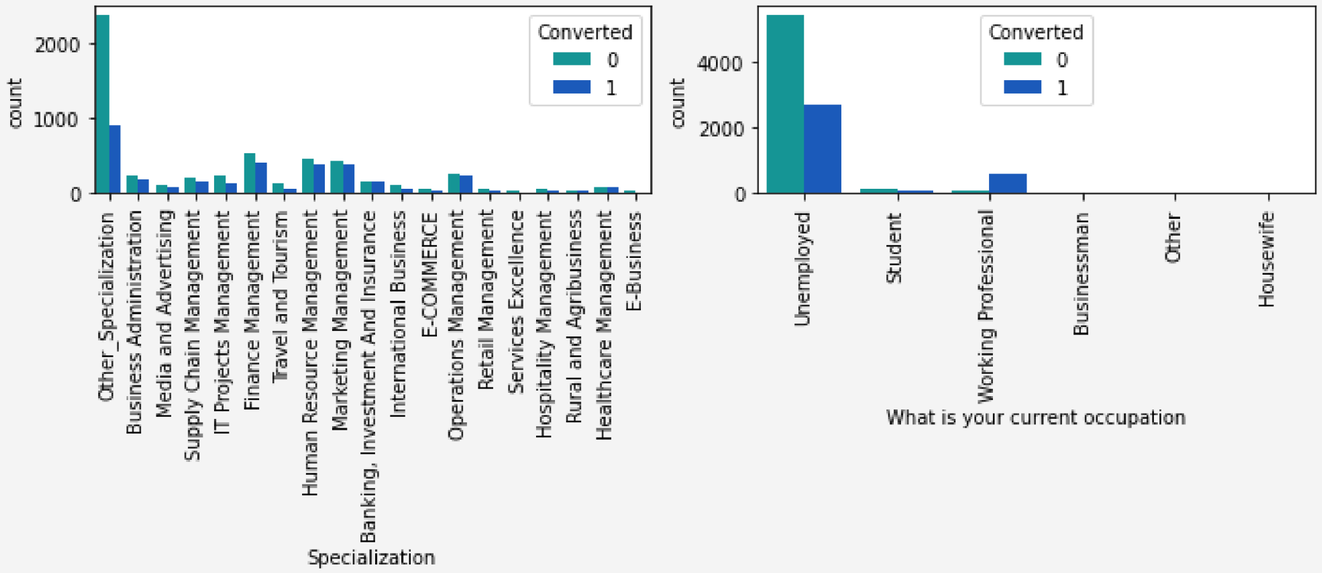


- The median of both converted and non-converted is same. Hence, nothing can be inferred from first sub-plot.
- Users spending more time on website are more likely to get converted.

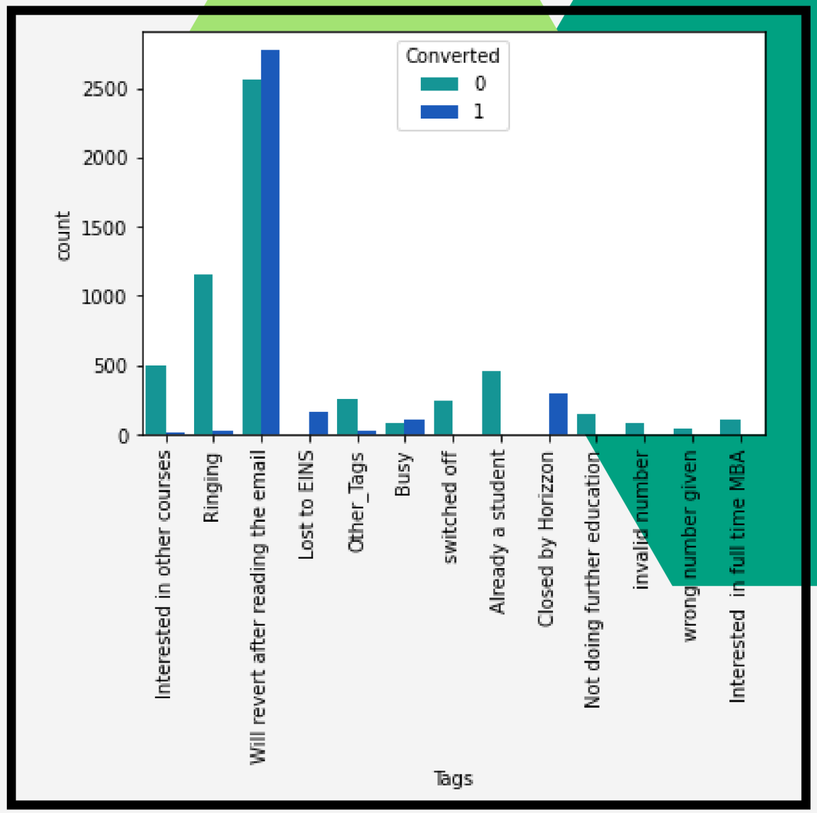


- The count of leads last activity as "Email Opened" is maximum.
- The conversion rate of "SMS Sent" is maximum.

Exploratory Data Analysis



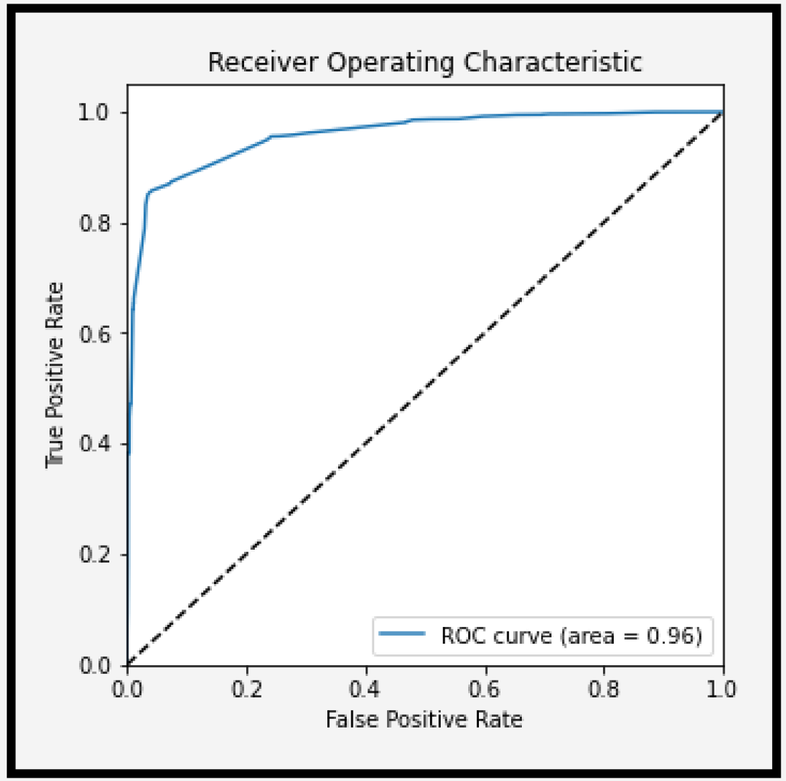
- By looking at the first sub-plot, no particular inference can be made about specialisation.
- Working professionals have high conversion rate.
- Count of unemployed leads is greatest among other categories.

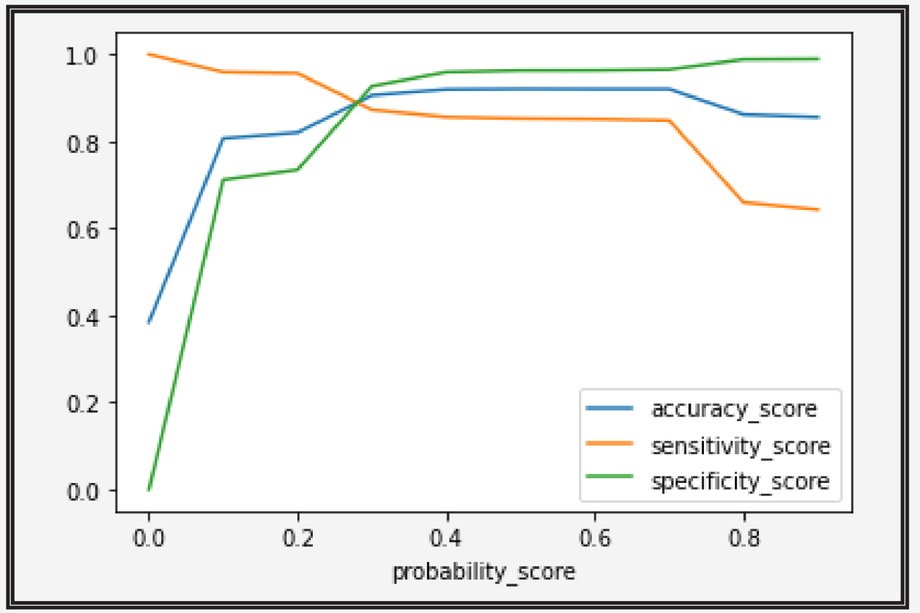


- "Will revert after reading mail" and "Closed by Horizzon" has high conversion rate.

Model Building

- Splitting data into test and training sets
- The train-test split ratio is 70:30.
- Using RFE to choose top 15 variables
- Remove variables with p-value > 0.05 and VIF > 5.
- Prediction on test dataset using ROC curve and optimal cut-off curve.
- Overall accuracy of model is 92%





Model Evaluation

- Calculated evaluation metrics from various probability cut-offs.
- As per the graph and chart it can be said that the optimal cut-off 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	0.807117	0.959526	0.711652	0.675785
0.2	0.2	0.820343	0.956664	0.734955	0.693333
0.3	0.3	0.905999	0.872445	0.927017	0.882183
0.4	0.4	0.919540	0.856092	0.959283	0.929427
0.5	0.5	0.920642	0.852821	0.963124	0.935426
0.6	0.6	0.920328	0.851594	0.963380	0.935759
0.7	0.7	0.920328	0.848324	0.965429	0.938914
0.8	0.8	0.861912	0.659853	0.988476	0.972875
0.9	0.9	0.856086	0.643500	0.989245	0.974010

Train Data Confusion Matrix

Prdicted / Actual	Not Converted	Converted
Not Converted	2987	918
Converted	124	2322

Accuracy	83.59%
Sensitivity	94.93%
Specificity	76.49%
Precision	71.66%

Top Features and Test Data Matrics

```
----------------------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
dtype: float64
```

Test Data Confusion Matrix

Prdicted / Actual	Not Converted	Converted
Not Converted	1303	431
Converted	71	918

Accuracy	81.56%
Sensitivity	92.82%
Specificity	75.14%
Precision	68.05%

Conclusion

- Final logistic regression model is built with **14** features.
- The model predicts the probability of the target variable having a certain value.
- A cutoff of the probability is used to obtain the predicted value of the target variable.
- Optimum cut off is chosen to be **0.27** i.e. any lead with greater than **27%** probability of getting converted is predicted as **Hot Lead** (customer will convert) and any lead with 0.27 or less probability of getting converted is predicted as **Cold Lead** (customer will not convert).
- Lead Score calculated shows the conversion rate of final predicted model is approx. **92%** in **test data** while **95%** in **train data.**
- The final model has a precision of 0.68, i.e. **68%** of predicted hot leads are **true hot leads**.
- Top 3 variables which contribute to leads getting converted are:
  Tags_lost to EINS
  Tags_Closed by Horizzon
  Tags_Will revert after reading mail
- In Business terms, this model has capability to adjust with company's future requirements.