Lead Scoring Case Study

Hemant Kokane

Shriraj Prabhugaonkar

Dona Maria Joseph

Problem Statement

An education company named X Education sells online courses to industry professionals. When people fill up a form providing their email address or phone number, they are classified to be a lead. Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Assumptions

- Variables with more than 40% missing values were EXCLUDED.
- Missing values of categorical data-type were replace by MODE.
- Missing values of numerical data-type were replace by MEDIAN.
- Outliers were cap to 99% percentile.
- Columns with minimum variation in data-levels were DROPED.
- 7 : 3 ratio is used to split data into train and test.
- Model build by using RFE elimination method.
- Threshold for P-value and VIF are 0.05 and 5 respectively.
- Model is evaluated on the basis of Recall-Score and ROC Curve.

Overall Approach

Data Understanding And Cleaning

Data Analysis(univariate & bivariate)

Data Preparation

Model Building

Model Evaluation

Summary

Missing Values Before And After Imputation

```
lead_data.isna().sum()
```

```
Lead Origin                                       0
Lead Source                                      36
Converted                                         0
TotalVisits                                     137
Total Time Spent on Website                       0
Page Views Per Visit                            137
Last Activity                                   103
Country                                        2461
Specialization                                 3380
What is your current occupation                2690
What matters most to you in choosing a course  2709
Tags                                           3353
City                                           3669
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```

Before Imputation

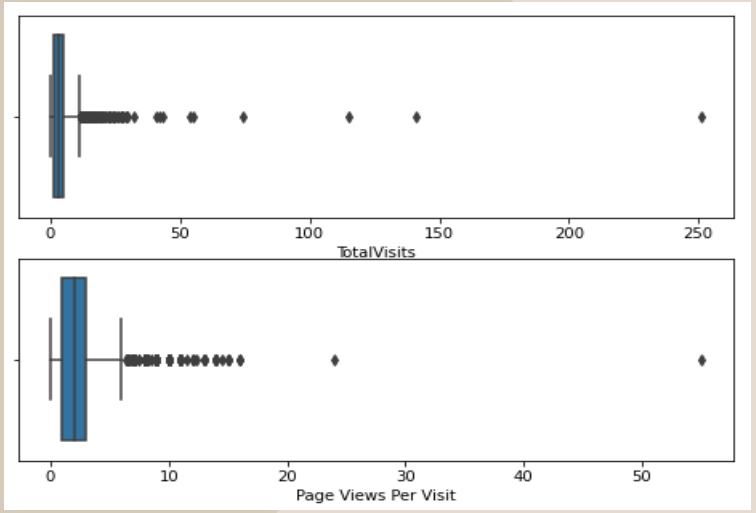
```
lead_data.isna().sum()/9240*100
```

```
Lead Origin                                    0.0
Lead Source                                    0.0
Converted                                      0.0
TotalVisits                                    0.0
Total Time Spent on Website                    0.0
Page Views Per Visit                           0.0
Last Activity                                  0.0
Country                                        0.0
Specialization                                 0.0
What is your current occupation                0.0
What matters most to you in choosing a course  0.0
Tags                                           0.0
City                                           0.0
A free copy of Mastering The Interview         0.0
Last Notable Activity                          0.0
dtype: float64
```

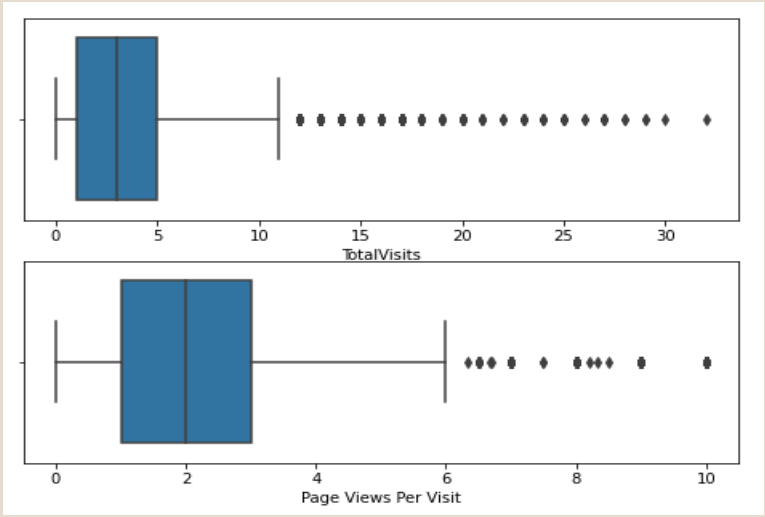
After Imputation

- Missing values of Categorical data were imputed by using MODE.
- Missing values of Numerical data were imputed by using MEDIAN.

Outliers Treatment



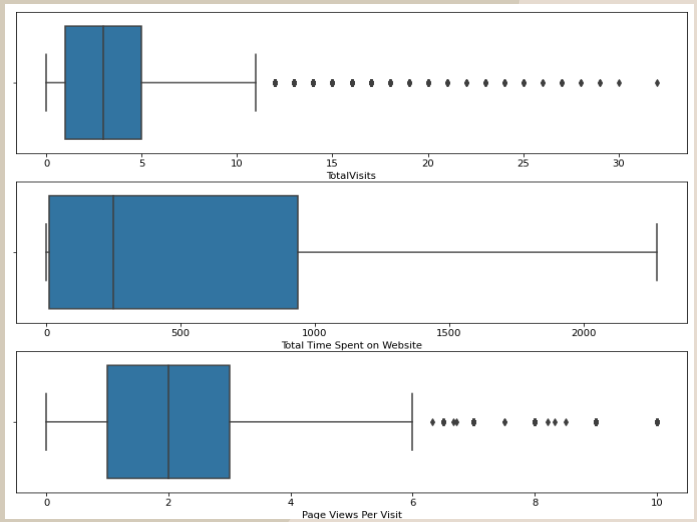
Before Capping



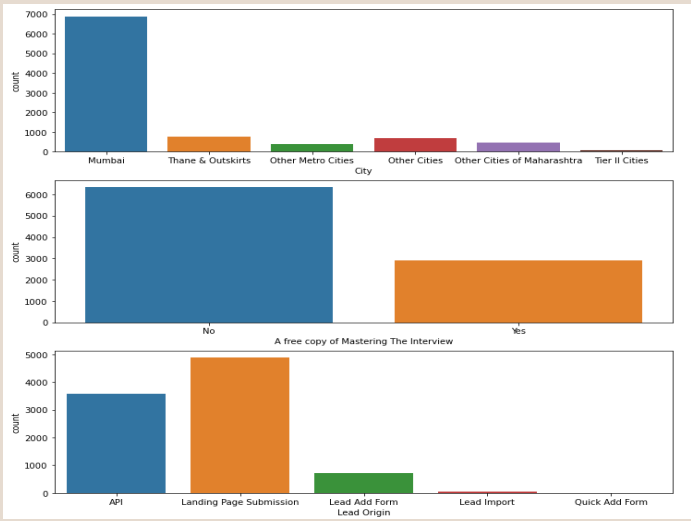
After Capping

- Values greater than 40 of "Total Visits" were cap to 99%.
- Values greater than 10 of "Page Views Per Visit" were cap to 99%.

Univariate Analysis



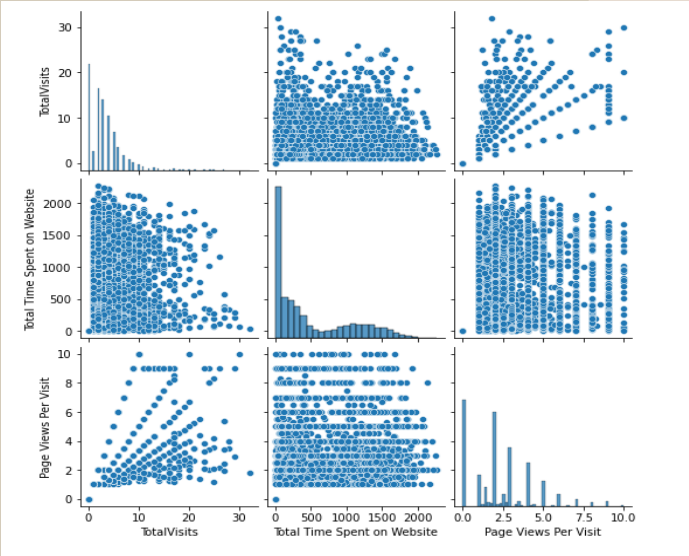
Numerical Data



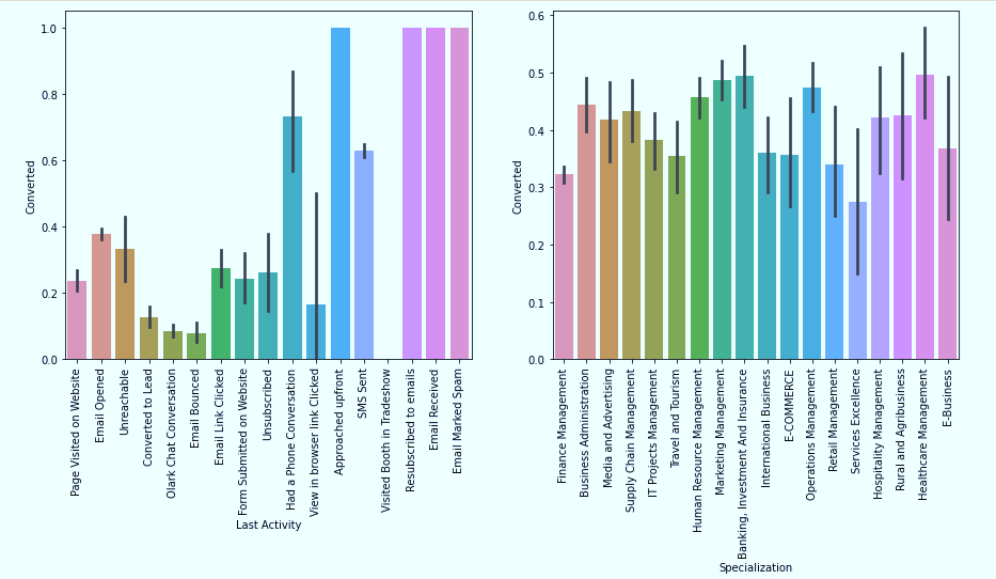
Categorical Data

- Above boxplot shows how data distributed in percentile term.
- Count plot shows frequency of each data level of categorical data.

Bivariate Analysis



Numerical Data



Categorical Data

- Pairplot shows relation between numeric variables. There is strong relation between “Total visits” and “Pages views per visits”
- Bar plot shows mean of Target in each data level.

Data Preparation

- Dummy Feature Creation : We can not fit model on categorical levels. For that we have to convert such levels or data points in to binary form. And dummy feature creation do same thing by using pd.get_dummies.

- Splitting The Data : We use train_test_split from sklearn.model_selection to split data into train and test data.

- Rescaling the data : MinMaxScaler() is used to scale the data. So we can build model on uniform data.

Model Building

- Following libraries were used to build logistic model
  1. from sklearn.feature_selection import RFE
  2. from sklearn.linear_model import LogisticRegression

- RFE : Recursive Feature Elimination, or RFE for short, is a feature selection algorithm. Technically, RFE is a wrapper-style feature selection algorithm that also uses filter-based feature selection internally. works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains.

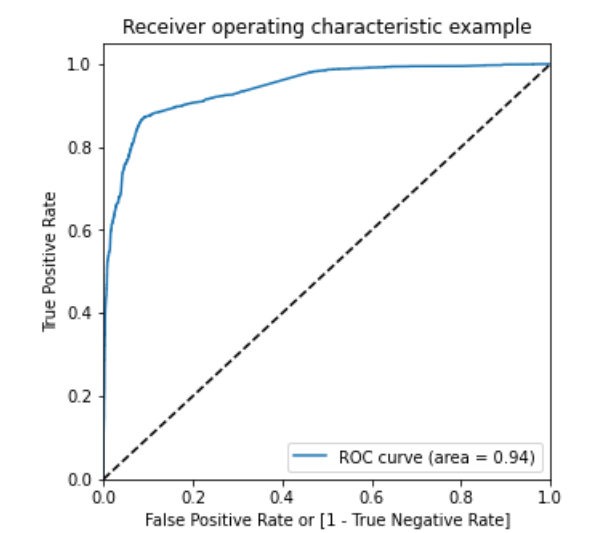
P- Value and VIF

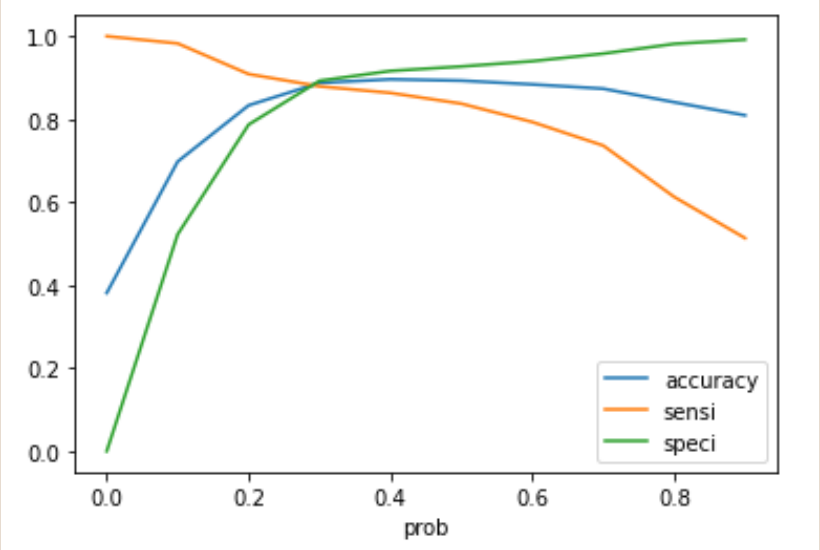
	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.2450	0.211	-24.908	0.000	-5.658	-4.832
Total Time Spent on Website	4.3512	0.183	23.755	0.000	3.992	4.710
Lead Origin_Lead Add Form	3.3527	0.221	15.189	0.000	2.920	3.785
Last Activity_Email Bounced	-1.8876	0.328	-5.759	0.000	-2.530	-1.245
What is your current occupation_Working Professional	2.7254	0.245	11.120	0.000	2.245	3.206
Tags_Busy	2.8123	0.299	9.407	0.000	2.226	3.398
Tags_Closed by Horizzon	8.5589	1.030	8.312	0.000	6.541	10.577
Tags_Lost to EINS	8.0980	0.750	10.799	0.000	6.628	9.568
Tags_Ringing	-1.4091	0.303	-4.645	0.000	-2.004	-0.815
Tags_Will revert after reading the email	3.4974	0.201	17.358	0.000	3.103	3.892
Tags_in touch with EINS	3.3956	0.822	4.133	0.000	1.785	5.006
Tags_switched off	-1.3352	0.561	-2.381	0.017	-2.434	-0.236
Last Notable Activity_Had a Phone Conversation	4.0679	1.365	2.981	0.003	1.393	6.743
Last Notable Activity_SMS Sent	2.6022	0.115	22.616	0.000	2.377	2.828

	Features	VIF
8	Tags_Will revert after reading the email	1.85
0	Total Time Spent on Website	1.83
12	Last Notable Activity_SMS Sent	1.52
1	Lead Origin_Lead Add Form	1.37
5	Tags_Closed by Horizzon	1.31
3	What is your current occupation_Working Profes...	1.19
7	Tags_Ringing	1.15
4	Tags_Busy	1.05
6	Tags_Lost to EINS	1.04
10	Tags_switched off	1.04
2	Last Activity_Email Bounced	1.03
9	Tags_in touch with EINS	1.00
11	Last Notable Activity_Had a Phone Conversation	1.00

- From above chart we can see that P-Value of all variables are less than 0.05. And VIF of all variables are less than 5.
- This is standard assumption that P-Value should be less than 0.05 and VIF should be less than 5.

ROC And Optimal Threshold





- ROC :The ROC Curve is a useful diagnostic tool for understanding the trade-off for different thresholds and the ROC provides a useful number for comparing models based on their general capabilities.
- Area under ROC curve is 0.94. Optimal probability is 0.3(approx.)

Model Evaluation

1. Confusion Matrix

Train Data

	FN	FP
TN	3572	430
TP	298	2168

Test Data

	FN	FP
TN	1483	194
TP	114	981

2. Recall

Recall_Train = 2168/(2168+298) = **0.88** & Recall_Test = 981/(981+114) = **0.89**

3. F1 Score

F1_Train = 0.86 & F1_Test = 0.86

Summary

We build a logistic regression model which assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. For that we did univariate and bivariate analysis on given data, Missing value imputation, Outliers treatment, Dummy feature creation, Splitting and Scaling of data, Model building and Evaluation.

Finally we test model by using recall score. Recall score for train data was 0.88 and for test data was 0.89.

thank you

Hemant Kokane

Shriraj Prabhugaonkar

Dona Maria Joseph