CASE STUDY ON LEAD SCORING

BY,

KRISHNA PRASAD PONNUR RAJENDRAN

SANTOSH KUMAR K

MITALI PENSIA

TABLE OF CONTENT

INTRODUCTION TO X EDUCATION COMPANY

- An education company named X Education sells online courses to industry professionals.
- Many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these 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. Through this process, some of the leads get converted while most do not.

PROBLEM STATEMENT

- The typical lead conversion rate at X education is around 30%.
- Now, 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'.
- 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.

OBJECTIVE OF THE CASE STUDY

- Select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Build a model wherein one can assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

GOALS OF THE CASE STUDY

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Predict the model and test the model so that the model would be able to adjust to if the company's requirement changes in the future

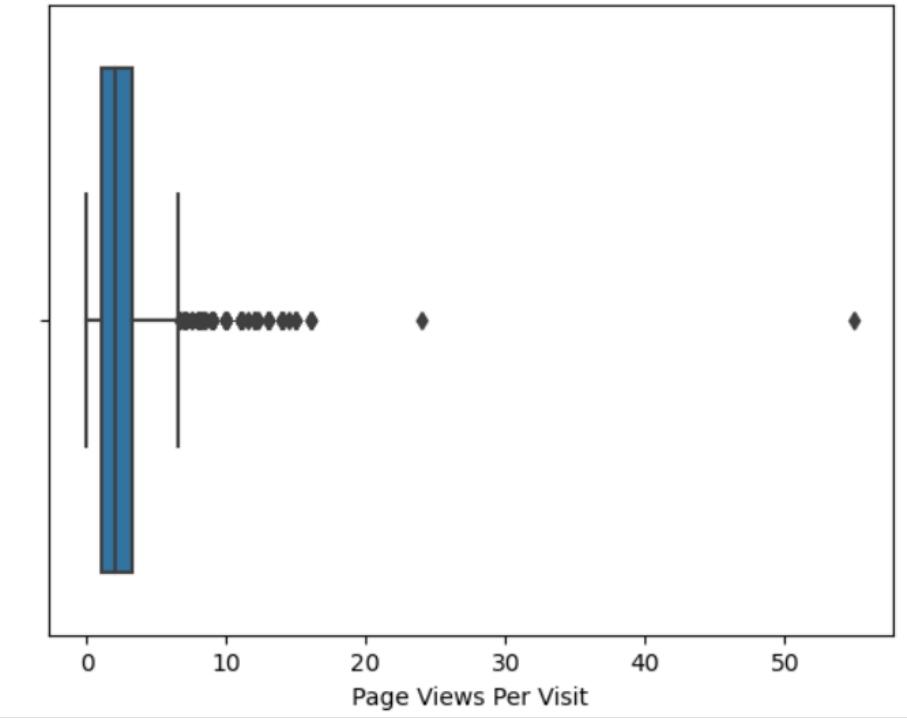
STEPS IN BUILDING THE MODEL

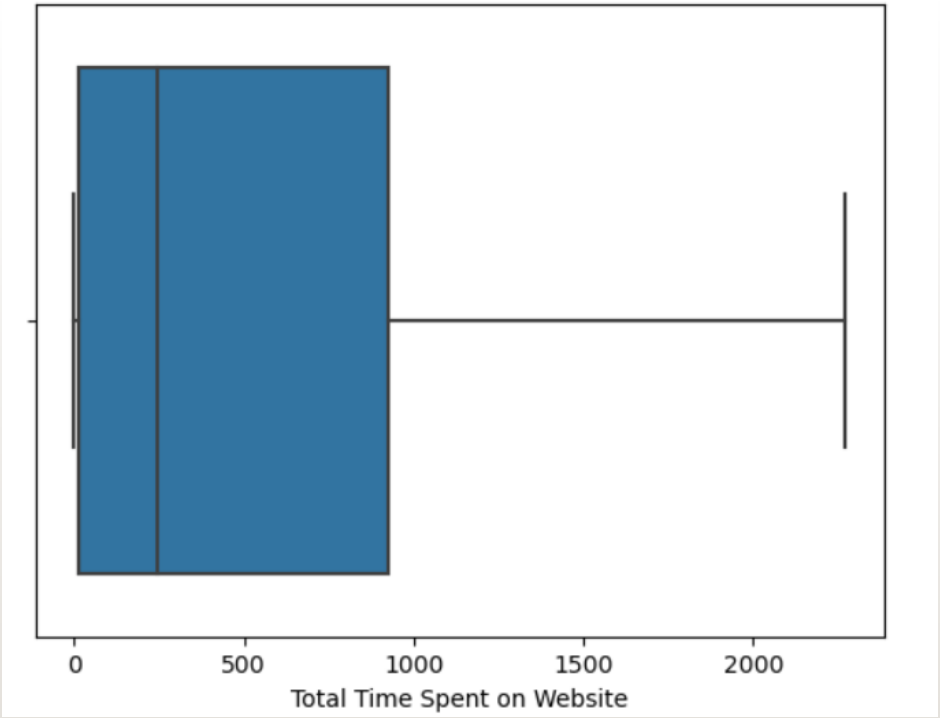
- Data Cleaning
- EDA
- Data Preparation
- Model Building

DATA CLEANING

- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- A level called 'Select’ for categorical variables which is same as null values, is handled by deleting it after the split is done.
- Handling Missing values
  - Deleting the columns with over 40% null values
  - Deleting the columns which has only single values
  - Deleting the columns which has no use for modelling like lead Number, Propect ID
  - Converting few of the null values to unknown
  - Deleting the rows for which the missing values are less than 2%.

EDA – UNIVARIATE ANALYSIS





EDA - CORRELATION

- The correlation of the variables was no found and variables are independent of each other

DATA PREPARATION

- Train- test Split
- Feature Scaling using MinMaxScaler

	Do Not Email	TotalVisits	Total Time Spent on Website	Page Views Per Visit
7113	0	0.026087	0.272887	0.0625
4749	0	0.000000	0.000000	0.0000
7986	0	0.000000	0.000000	0.0000
1281	0	0.034783	0.433099	0.1250
7346	0	0.034783	0.419014	0.0625

MODEL BUILDING

- Logistic Regression Model for binary classification, as it showed promise in handling the target variable, "Converted".
- RFE was preformed to select the most important columns.
- Manual Feature Reduction process was used to build models by dropping variables with p value greater than 0.05.
- The third Model was good with less p-value and VIF less than 5.

MODEL BUILDING

Model - 3

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3892	0.097	-24.659	0.000	-2.579	-2.199
Total Time Spent on Website	3.1863	0.208	15.342	0.000	2.779	3.593
Lead Source_Welingak Website	4.8604	0.735	6.615	0.000	3.420	6.301
Last Activity_SMS Sent	1.9380	0.104	18.587	0.000	1.734	2.142
Specialization_Unknown	-0.6983	0.134	-5.197	0.000	-0.962	-0.435
Tags_Already a student	-3.5072	0.715	-4.902	0.000	-4.909	-2.105
Tags_Closed by Horizzon	6.3382	0.718	8.832	0.000	4.932	7.745
Tags_Diploma holder (Not Eligible)	-2.8315	1.046	-2.707	0.007	-4.882	-0.781
Tags_Interested in other courses	-2.2841	0.356	-6.418	0.000	-2.982	-1.587
Tags_Lost to EINS	5.0618	0.596	8.488	0.000	3.893	6.231
Tags_Not doing further education	-3.2432	1.017	-3.190	0.001	-5.236	-1.251
Tags_Ringing	-3.0502	0.219	-13.923	0.000	-3.480	-2.621
Tags_Will revert after reading the email	4.5280	0.181	25.079	0.000	4.174	4.882
Tags_invalid number	-3.6531	1.060	-3.447	0.001	-5.730	-1.576
Tags_switched off	-3.6761	0.599	-6.141	0.000	-4.849	-2.503

VIF

	Features	VIF
0	Total Time Spent on Website	1.69
11	Tags_Will revert after reading the email	1.56
2	Last Activity_SMS Sent	1.51
10	Tags_Ringing	1.13
5	Tags_Closed by Horizzon	1.04
8	Tags_Lost to EINS	1.04
1	Lead Source_Welingak Website	1.03
3	Specialization_Unknown	1.03
13	Tags_switched off	1.03
4	Tags_Already a student	1.02
7	Tags_Interested in other courses	1.02
9	Tags_Not doing further education	1.01
12	Tags_invalid number	1.01
6	Tags_Diploma holder (Not Eligible)	1.00

MODEL EVALUATION

Train Set

```
Confusion matrix:
[[3778  158]
 [ 397 2018]]

ACCURACY:
0.9126121870571563
```

```
Sensitivity (Recall): 0.8356107660455486
Specificity: 0.9598577235772358
False Positive Rate (FPR): 0.04014227642276423
Positive Predictive Value (Precision): 0.9273897058823529
Negative Predictive Value (NPV): 0.9049101796407185
```

Test Set

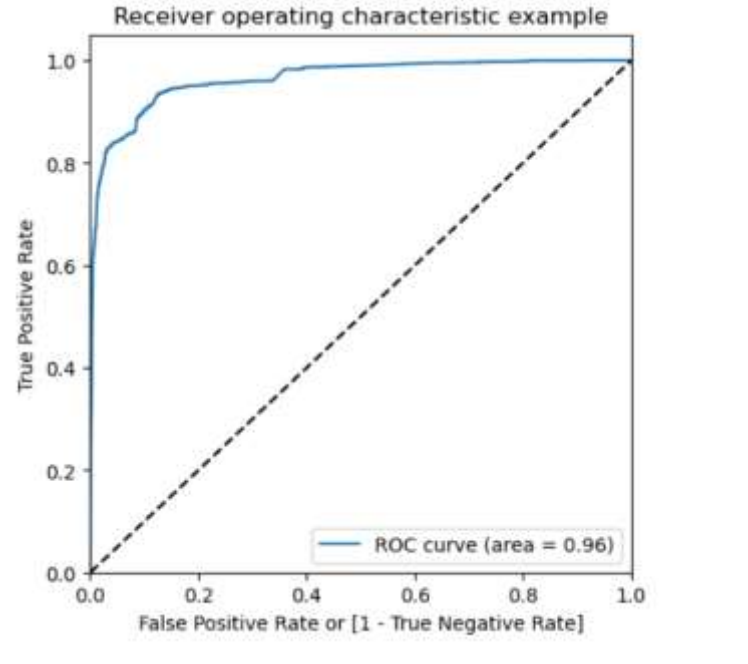
```
Confusion matrix:
[[1579  123]
 [ 147  873]]

ACCURACY:
0.9008082292432035
```

```
Sensitivity (Recall): 0.8558823529411764
Specificity: 0.9277320799059929
False Positive Rate (FPR): 0.07226792009400705
Positive Predictive Value (Precision): 0.8765060240963856
Negative Predictive Value (NPV): 0.9148319814600232
```

MODEL EVALUATION

- ROC Curve



PROBLEM SOLUTION

- Based on the Final Model suggestion can be provided to the company as:

1. Identify the Hot leads based on the high probability of conversion and spend efforts on them.
2. Send out automated Emails, SMS and Chatbots to the Leads.
3. Notify the Leads with the latest offers and updates of the company.
4. Improve the quality of the Interaction with the leads by increasing more sales team.