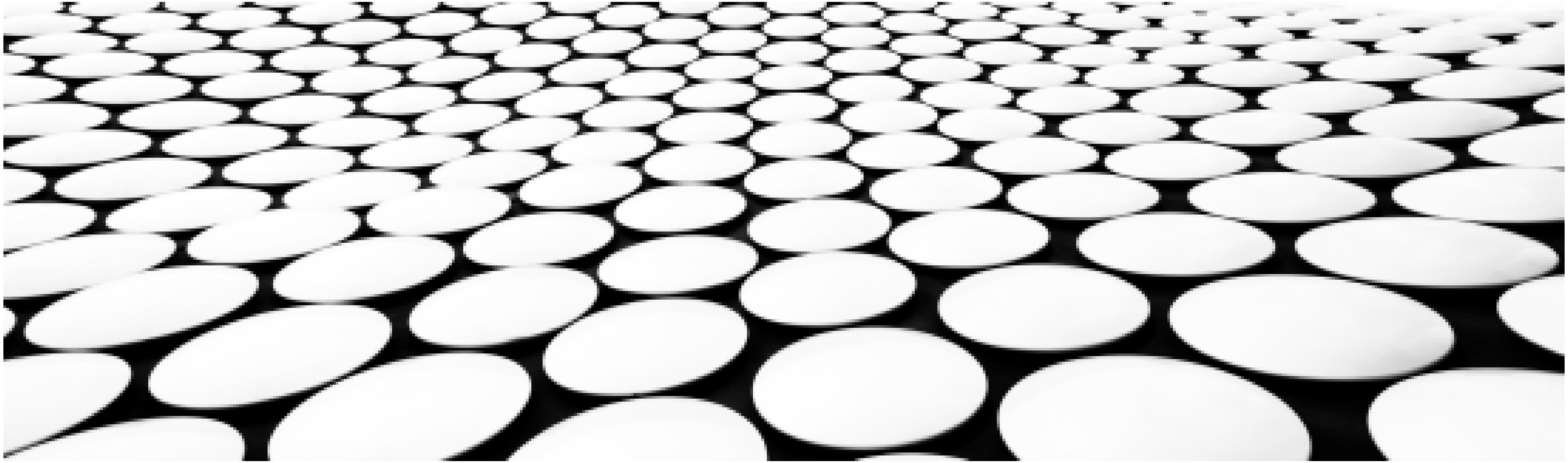
LEAD SCORE – LR MODEL

ANUPRIYA, ANURADHA, ANWESHA



PROBLEM STATEMENT AND OBJECTIVE

- X Education company, sells online courses to industry professional. The company captures leads from its website, search engines, and past referrals.
- The typical lead conversion rate at X Education is around 30%. The CEO wants to improve the lead conversion rate to 80%..
- Additionally, the company requires a model to 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.

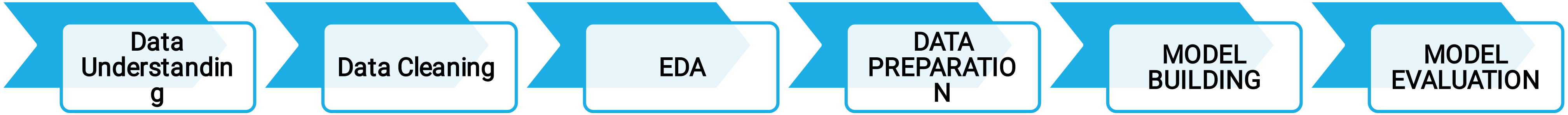
Objective:

- The objective of this project is to create a logistic regression model within a Python-based Jupiter notebook.

- This model will assign lead scores ranging from 0 to 100 to each lead, facilitating X Education in its efforts to enhance lead conversion rates and boost sales.

- A higher score will indicate a "hot" lead with a greater likelihood of conversion, allowing the sales team to prioritize communication with these leads for a more efficient and effective sales process.

STEPS TAKEN, METHODOLOGY



Data Understandin g

Data Cleaning

EDA

DATA PREPARATIO N

MODEL BUILDING

MODEL EVALUATION

PROVIDED DATA – WHAT IS IT

- A dataset containing over 9,000 data points (9240 rows * 37 columns) has been provided.
- Among these attributes are Lead Source, Total Time Spent on Website, Total Visits, Last Activity, and others, which may or may not prove useful in determining lead conversion.
- The target variable is the 'Converted' column, where 1 signifies a converted lead and 0 indicates an unconverted lead.
- While there are columns with a substantial number of null or missing values, there are no duplicate rows in the dataset.

DATA CLEANING

- 'Select' level in specific categorical variables was replaced with 'NaN' to indicate null values, signifying instances where customers made no selections.

- Columns with more than 40% missing values were eliminated from the dataset, including 'Prospect ID' and 'Lead Number' due to their unique values in each row.

- Columns with constant values across all rows were also removed.

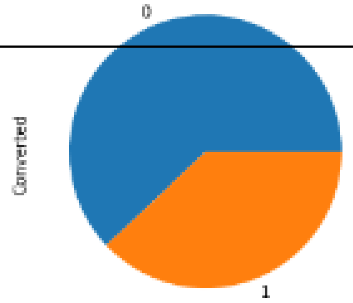
- Addressing missing values in categorical columns depended on category distribution and data balance. Imputation was employed for evenly distributed categories, while columns were dropped if imputation would lead to skewed data.

- Numerical columns with missing values were imputed using the mode after assessing their distribution.

- Finally after the cleaning we came with 9240 ROWS and 30 COLUMNS

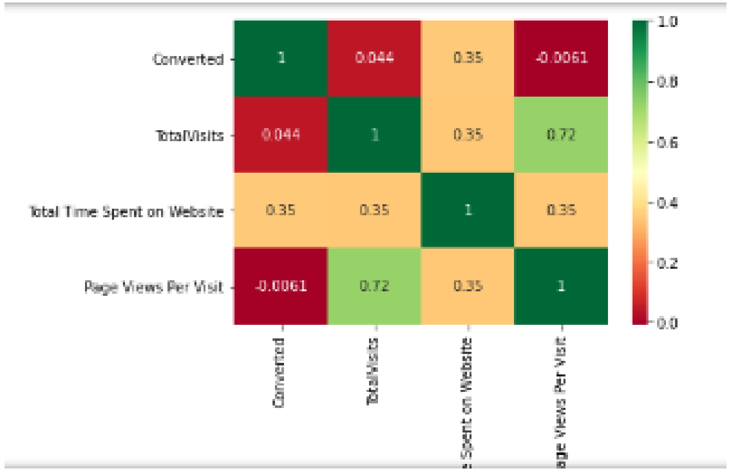
EDA

Univariate Numerical Analysis on Target variable 'Converted' gave 38.02% as converted as shown



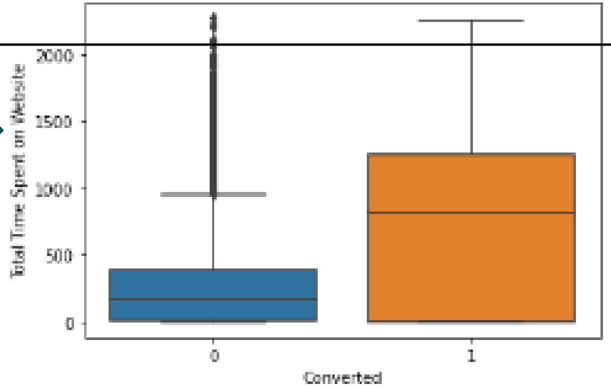
Only **38.02%** of whole data are being **converted**

Multivariate Analysis



The target variable "Converted" has a medium correlation with variable "Total time spent on website".
Total visits and Page views per visit have a high correlation.

Bivariate analysis on Total Time Spent on Website vs Converted



- The leads spending more time on website are more likely to get converted
- The website should be improved to increase the conversion rate

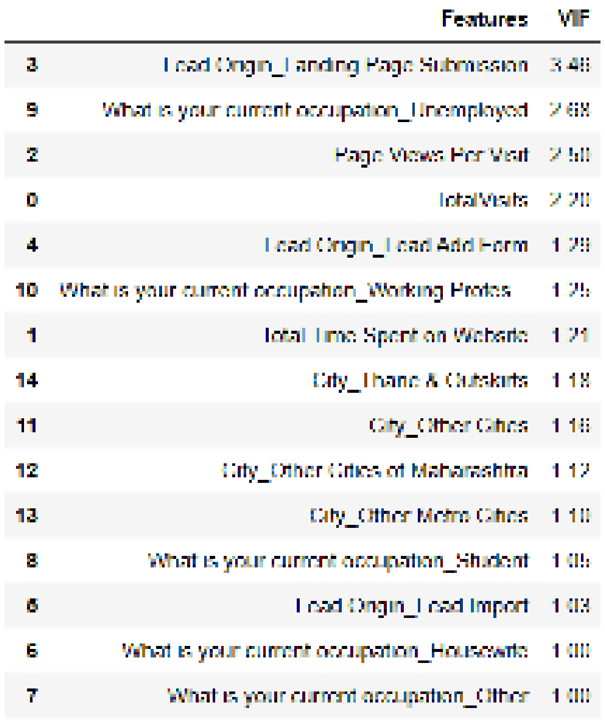
DATA PREPARATION

- Binary mapping of “do Not Email', 'A free copy of Mastering The Interview'” was chosen
- Creation of dummy variables for categorical attributes: 'Lead Origin,' 'Lead Source,' 'Last Activity,' 'Specialization,' and 'Occupation.’
- Dataset division into Train and Test sets, following a 70:30 ratio.
- Feature scaling was executed utilizing the Standard Scaler method.

MODEL BUILDING AND EVALUATION

MODEL BUILDING:-

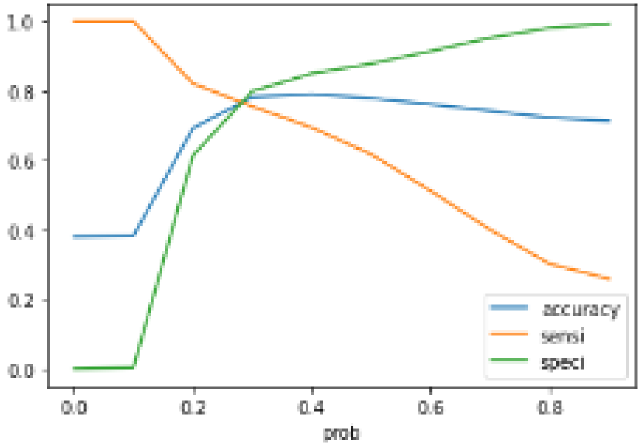
- Selected 15 features using RFE.
- Variables with p-value> 0.05 were dropped.
- Total 3 models were built with 14 variables having GOOD VIF



	Features	VIF
3	Lead Origin_Landing Page Submission	3.48
9	What is your current occupation_Unemployed	2.68
2	Page Views Per Visit	2.50
0	TotalVisits	2.20
4	Lead Origin_Lead Add Form	1.29
10	What is your current occupation_Working Profes	1.25
1	Total Time Spent on Website	1.21
14	City_Thane & Outskirts	1.18
11	City_Other Cities	1.16
12	City_Other Cities of Maharashtra	1.12
13	City_Other Metro Cities	1.10
8	What is your current occupation_Student	1.05
5	Lead Origin_Lead Import	1.03
6	What is your current occupation_Housewife	1.00
7	What is your current occupation_Other	1.00

MODEL EVALUATION:-

- Confusion matrix with cutoff value 0.3 was selected based on ROC cur
- Since the precision-recall view gave performance metrics around 75%, we preferred sensitivity-specificity score.
- Evaluation metrics for train & test are very close to around 80%.



CONCLUSION & RECOMMENDATION

The top three variables in our model contributing towards the probability of a lead getting converted?

1. **Total Time Spent on Website:**
   - ✓ Positive contribution
   - ✓ Higher the time spent on the website, higher the probability of the
   - ✓ lead converting into customer
   - ✓ Sales team should focus on such leads
2. **Lead Source Reference:**
   - ✓ Positive contribution
   - ✓ If the source of the lead is a Reference, then there is a higher probability that the lead would convert, as the referrals not only provide for cashbacks but also assurances from current users and friends who will mostly be trusted - Sales team should focus on such leads

3. **What is your current occupation Student:**
   - ✓ Negative contribution
   - ✓ If the lead is already a student, chances are they will not take up another course which is designed for working professionals.
   - ✓ Sales team should not focus on such leads

THANK YOU