LEAD SCORING CASE STUDY

SHREYAS PATIL

KAJAL TOTAWAR

RAHUL SUPOLIA

CONTENTS

- Problem Statement
- Business Objectives
- Data Set
- Solution Methodology
- Data preparation and EDA
- Model Building
- ROC/AUC Curve
- Conclusion

PROBLEM STATEMENT

- X Education sells online courses to industry professionals. 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. The typical lead conversion rate at X education is around 30%.
- The company wants to identify the potential leads, also called Hot Leads.
- If company is able to identify this segment, then their sales team could call these leads , thus leading to better conversion rate.

BUSINESS OBJECTIVES

- X education wants to select the most promising leads.
- X education wants to build a model to identify the hot leads.
- X educations wants to deploy the model for future use as well in case requirement changes

DATA SET

We have been provided with a leads dataset from the past with around 9000 data points:

- *This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.*
- *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 data dictionary is also provided for more details on data.*

SOLUTION METHODOLOGY

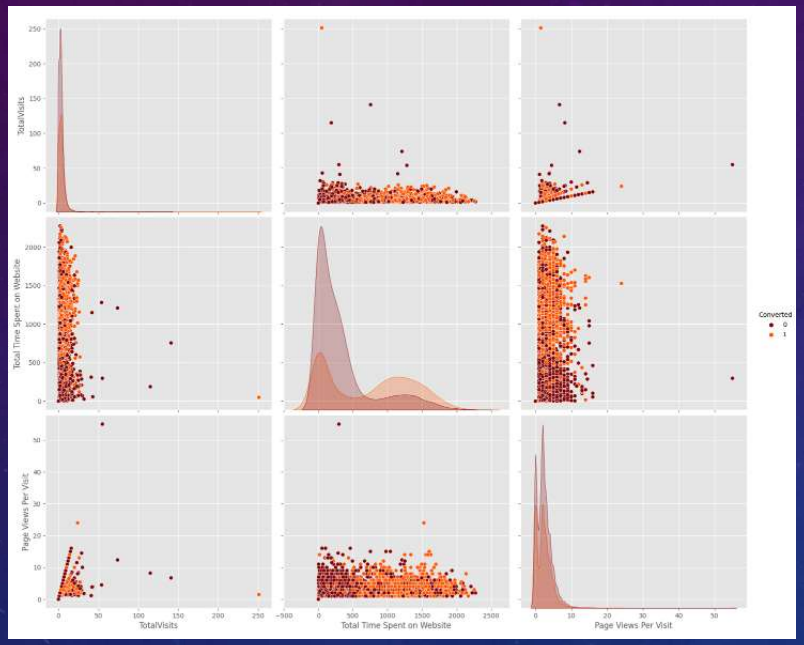
Below steps were used:

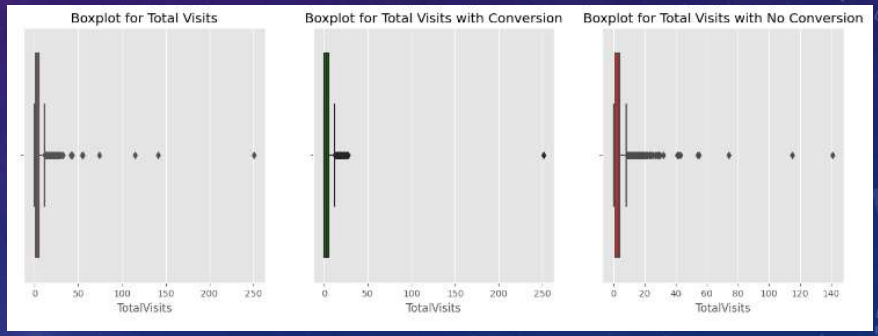
1. Data Cleaning
2. EDA
3. Train- Test split
4. Model Building
5. Model Evaluation
6. Model Presentation
7. Conclusion

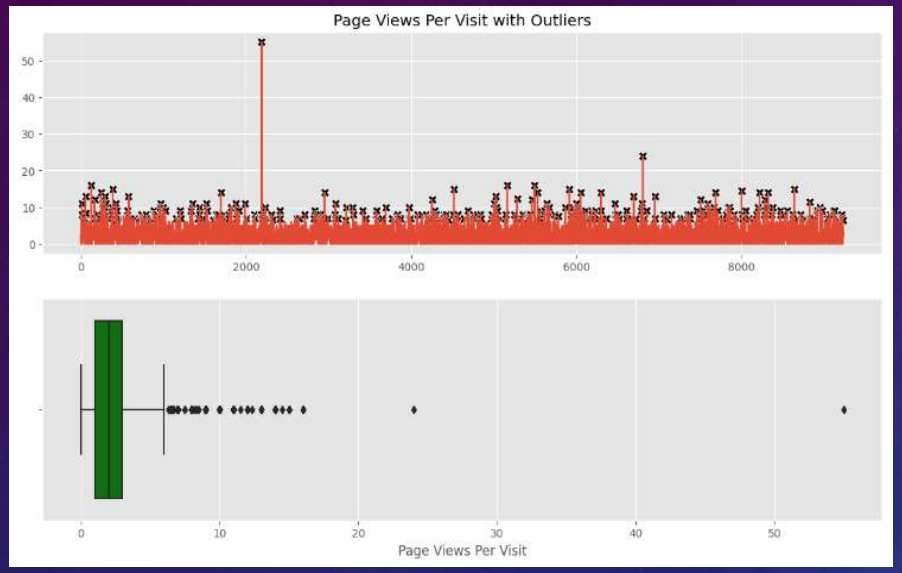
DATA PREPARATION AND EDA

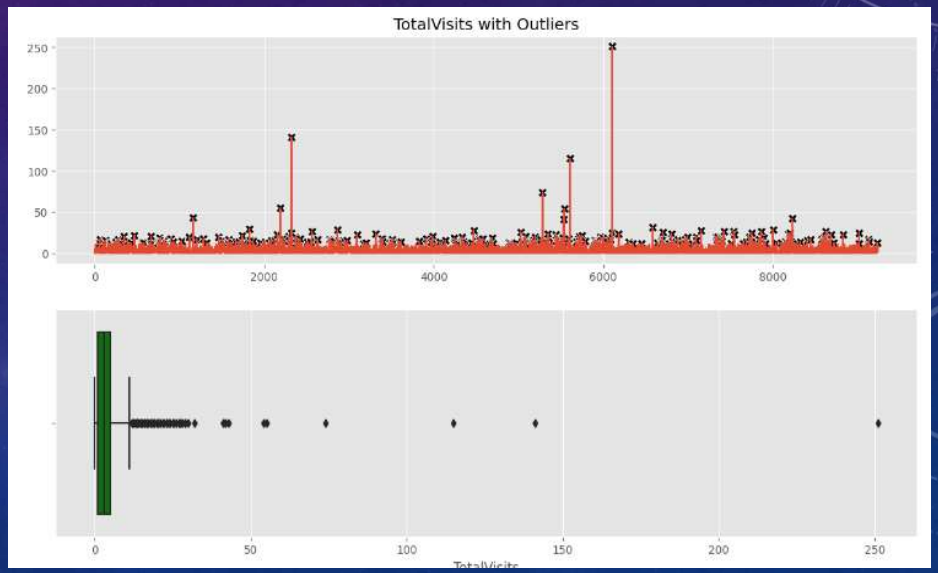
- Total Number of Rows =37, Total Number of Columns =9240.
- 4 columns with float values, 3 with integer value, and 30 with categorical value.
- Dropping ‘Asymmetrique Activity Score’, and ‘Asymmetrique Profile Score’ from numerical dataframe.
- Dropping from ‘Lead Quality’, ‘Asymmetrique Profile Index’, ‘Asymmetrique Activity Index’ categorical dataframe as null values are exceeding 40%.
- Created a new variable for those values which are either 'Select' or 'NaN' for these 4 columns.
- Imputed the frequent value in case of any null values present in our data.
- Addressed the outliers.

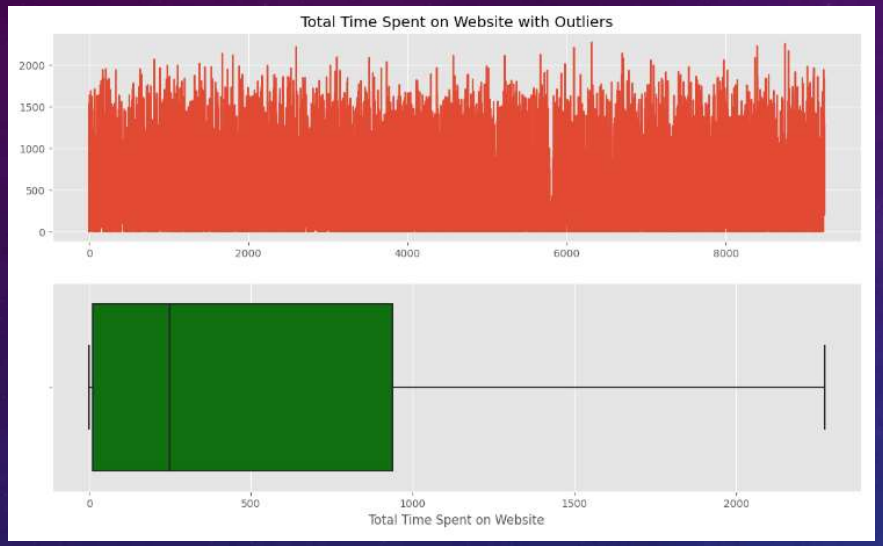
EDA

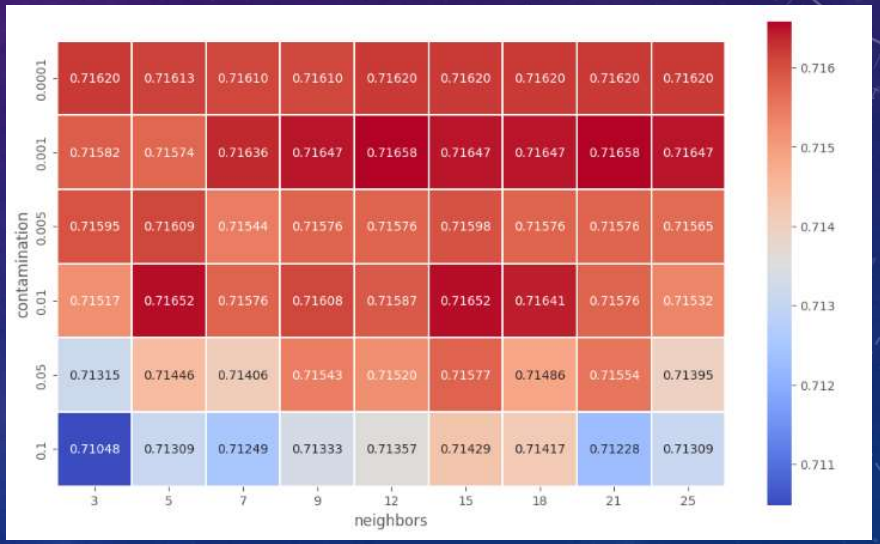

TotalVisits
Total Time Spent on Website
Page Views Per Visit
Converted
0
1


Boxplot for Total Visits
Boxplot for Total Visits with Conversion
Boxplot for Total Visits with No Conversion
TotalVisits
TotalVisits
TotalVisits


Page Views Per Visit with Outliers
50
40
30
20
10
0
0
2000
4000
6000
8000
0
10
20
30
40
50
Page Views Per Visit


TotalVisits with Outliers
250
200
150
100
50
0
0
2000
4000
6000
8000
0
50
100
150
200
250

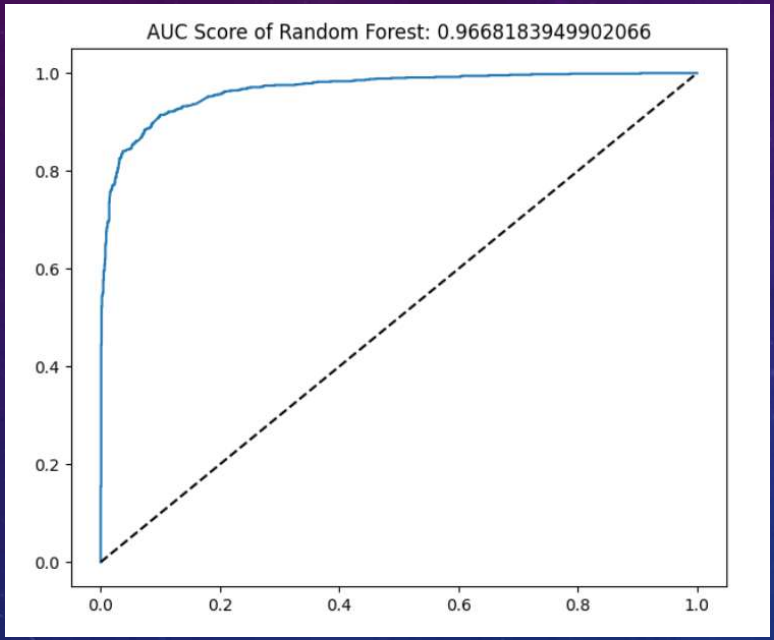

Total Time Spent on Website with Outliers
2000
1500
1000
500
0
0
2000
4000
6000
8000
0
500
1000
1500
2000
Total Time Spent on Website

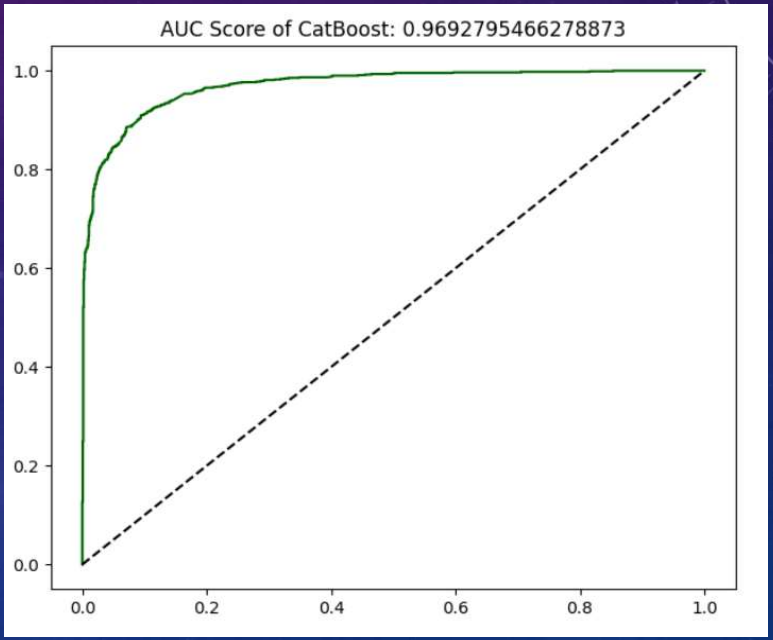

contamination
0.0001
0.001
0.005
0.01
0.05
0.1
0.71620 0.71613 0.71610 0.71610 0.71620 0.71620 0.71620 0.71620 0.71620
0.71582 0.71574 0.71636 0.71647 0.71658 0.71647 0.71647 0.71658 0.71647
0.71595 0.71609 0.71544 0.71576 0.71576 0.71598 0.71576 0.71576 0.71565
0.71517 0.71652 0.71576 0.71608 0.71587 0.71652 0.71641 0.71576 0.71532
0.71315 0.71446 0.71406 0.71543 0.71520 0.71577 0.71486 0.71554 0.71395
0.71048 0.71309 0.71249 0.71333 0.71357 0.71429 0.71417 0.71228 0.71309
3
5
7
9
12
15
18
21
25
neighbors
0.716
0.715
0.714
0.713
0.712
0.711

MODEL BUILDING

- Splitting dataset into training and testing sets.
- The split was done at 75% and 25% for train and test data respectively.
- Based on the results, we chose the RandomForest Classifier, GradientBoosting, LightGBM & Catboost on which we tested the other metrics to see in depth performance of these 4 models based on several different metrics to choose the best model for our analysis.
- Predictions on test dataset.
- Hyperparameter tuning done with a focus on Random Forest and CatBoost models for enhanced results
- Overall accuracy comes to 91%

PLOTTING ROC/AUC CURVE





CONCLUSION

- It was found that the variables that mattered the most in the potential buyers:
  1. The total time spend on the Website.
  2. Total number of visits.
  3. When the lead source was:
     a. Google
     b. Direct traffic
     c. Organic search
     d. Welingak website
  4. When the last activity was:
     a. SMS
     b. Olark chat conversation
  5. When the lead origin is Lead add format.
  6. When their current occupation is as a working professional.