LEAD SCORING CASE STUDY

Presented by:
Surya Nag S
Husna Maliakkal
Sushil Patil

LEAD SCORING PROCESS

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- When these 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 say only 30%.
- 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.

BUSINESS OBJECTIVE

- X Education wants to know most promising leads.
- For that they want to build a model which identifies the hot leads.
- Deployement of the model for the future use.

SOLUTION METHODOLOGY

Data cleaning & Data manipulation

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains a large number of missing values and are not useful for the analysis
- Imputation of the values, if necessary
- Check and handle outliers in data.

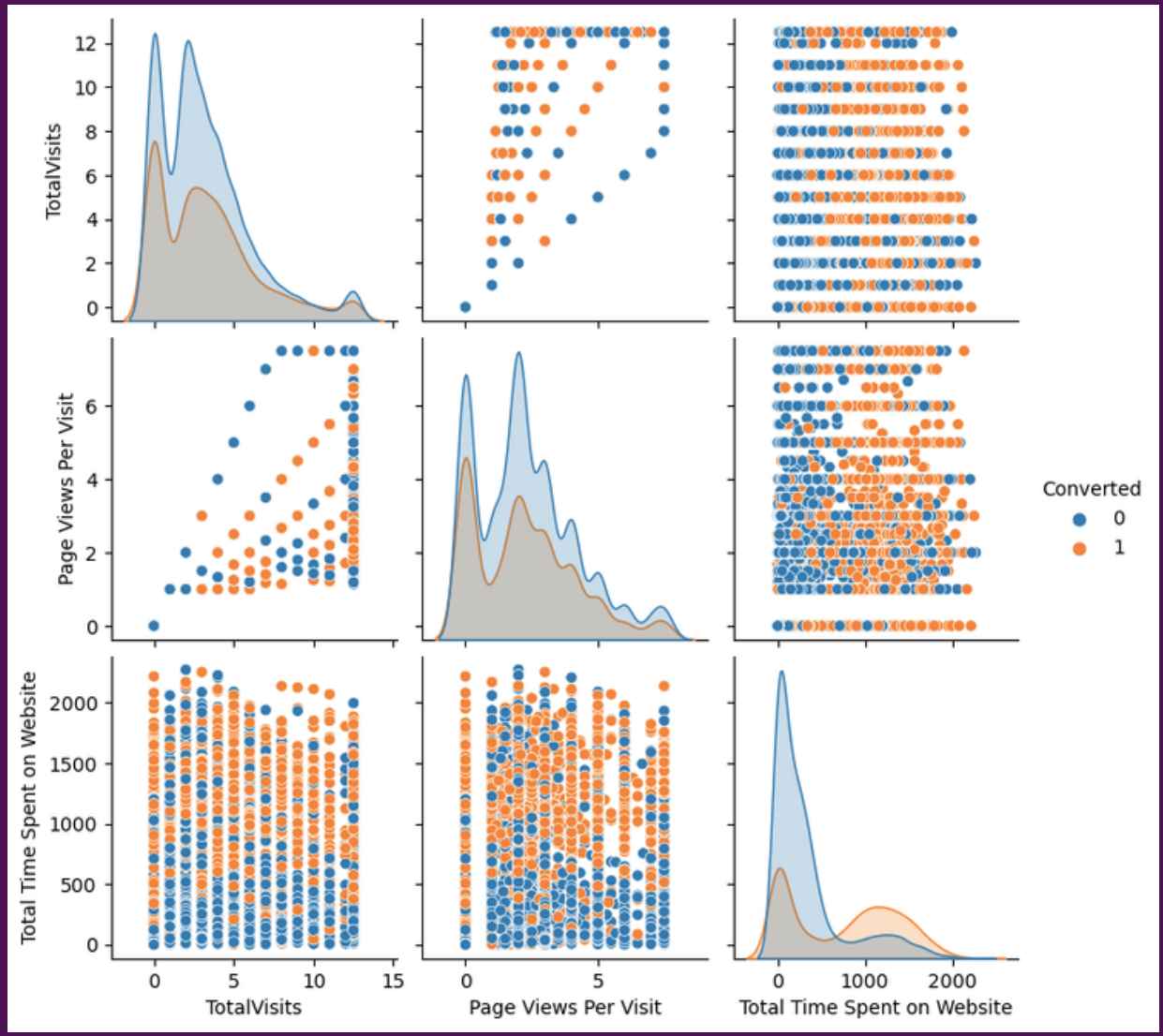
Exploratory Data Analysis

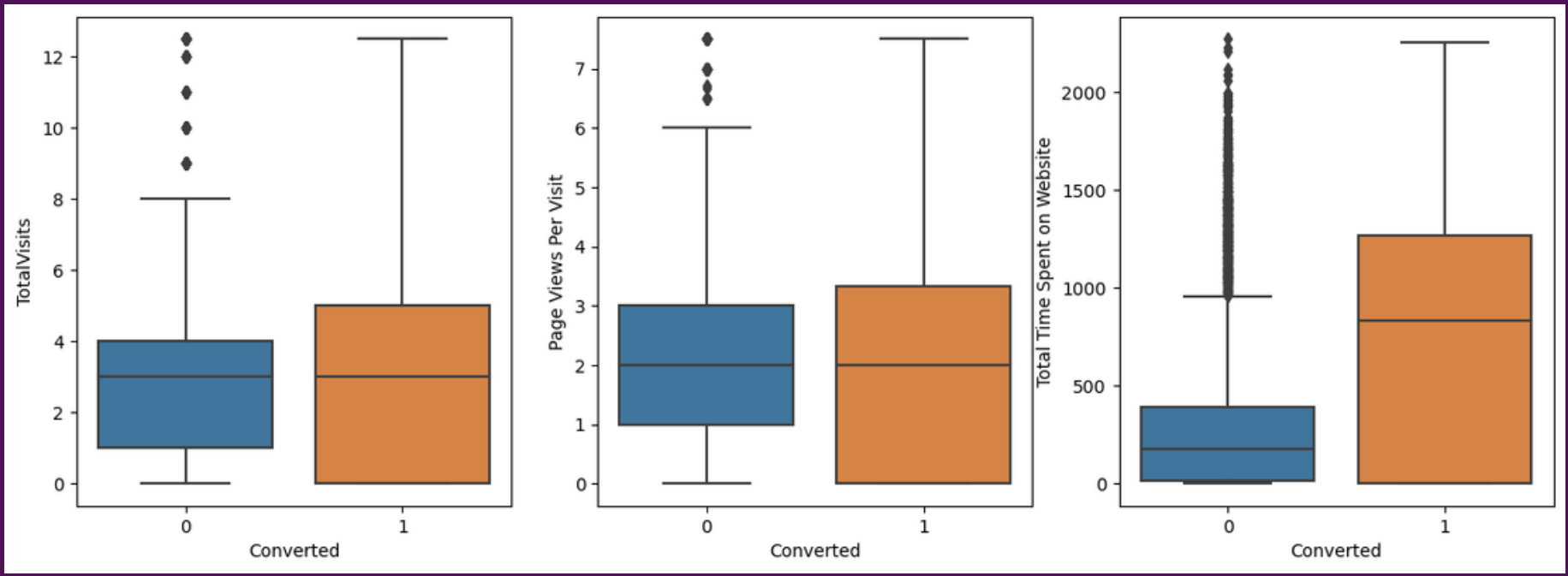
- Univariate data analysis: value count, distribution of variables, etc.
- Bivariate data analysis: Correlation coefficients and pattern between the variables ect.
- Feature scaling & Dummy variables and encoding of the data.
- Classification technique: Logistic regression is used for model making & prediction.
- Validation of the model.
- Model presentation.
- Conclusion and recommendations

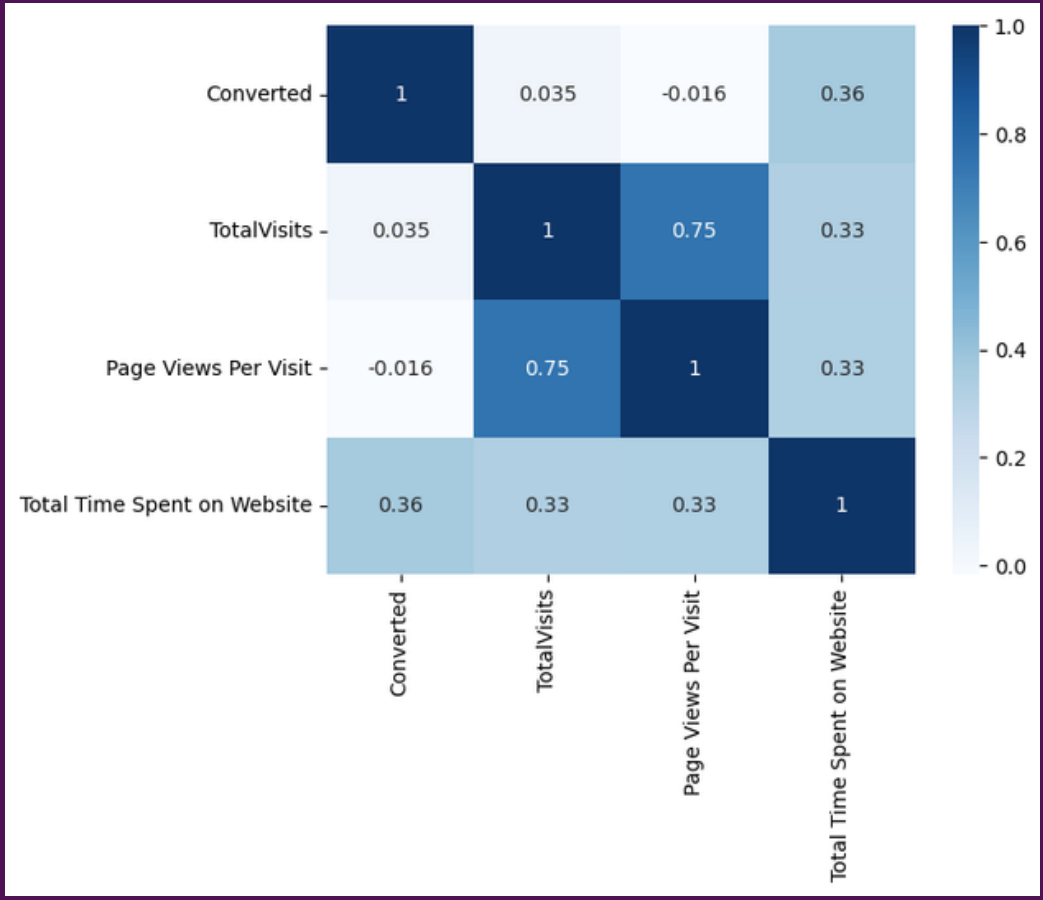
DATA MANIPULATION

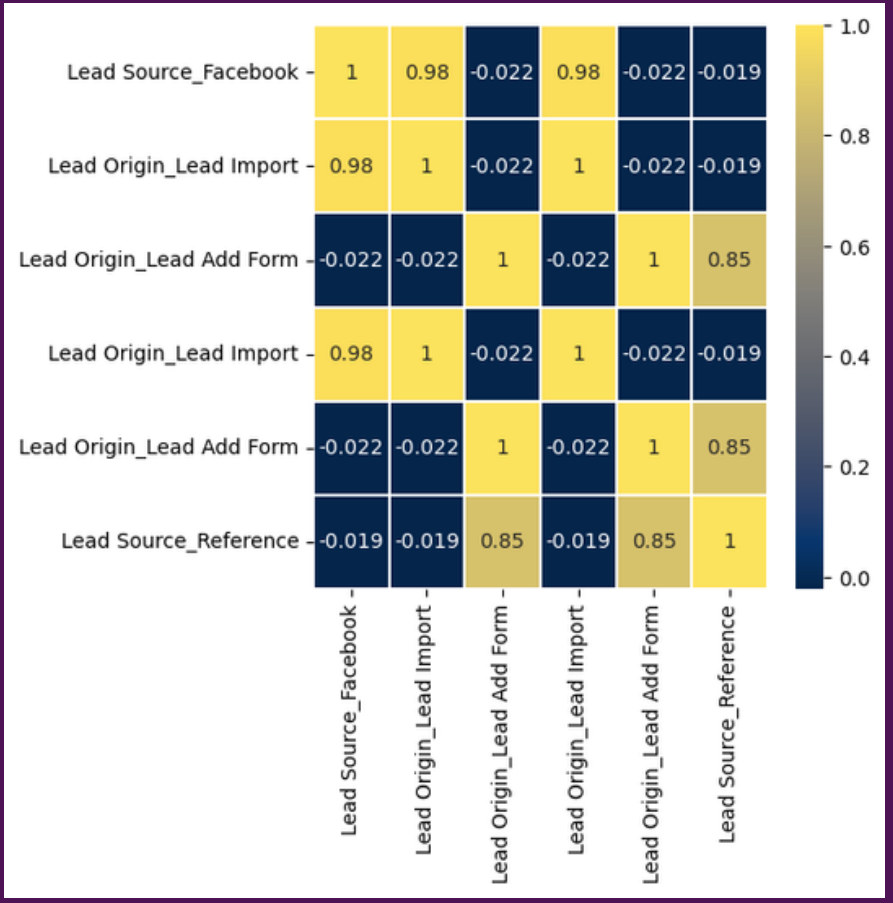
- Total Number of rows = 9240, Total no. of columns = 37.
- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update my supply","Chain content"," Get updates on DM content", "I agree to pay the amount through cheque" etc. were removed.
- Also removed the "Prospect ID and "Lead Number" which are not necessary for the analysis.
- After checking for the value counts for some of the object type type variables, we find some of the features which have enough variance, which have dropped, the features are:"Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", " X Education Forums", "Newspaper", "Digital Advertisement" etc.
- Dropping the column shaving more than 40% as missing values such as "How did you hear about X Education" and "Lead Profile"

EXPLORATORY DATA ANALYSIS (EDA)



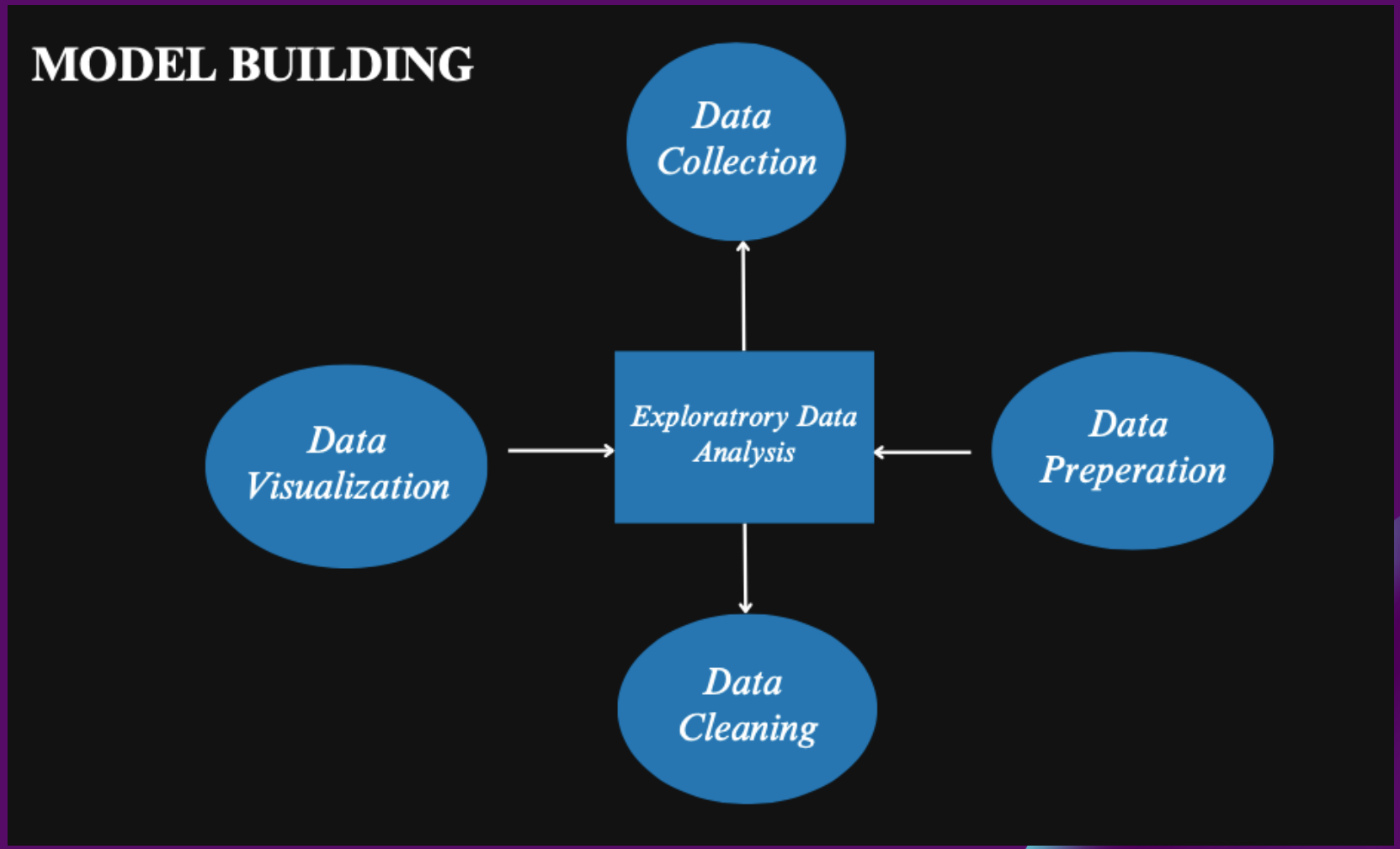

TotalVisits
0
2
4
6
8
10
12
0
1
Converted
Page Views Per Visit
0
1
2
3
4
5
6
7
0
1
Converted
Total Time Spent on Website
0
500
1000
1500
2000
0
1
Converted


Converted
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1
0.035
-0.016
0.36
0.035
1
0.75
0.33
-0.016
0.75
1
0.33
0.36
0.33
0.33
1
1.0
0.8
0.6
0.4
0.2
0.0


Lead Source_Facebook
Lead Origin_Lead Import
Lead Origin_Lead Add Form
Lead Origin_Lead Import
Lead Origin_Lead Add Form
Lead Source_Reference
1
0.98
-0.022
0.98
-0.022
-0.019
0.98
1
-0.022
1
-0.022
-0.019
-0.022
-0.022
1
-0.022
1
0.85
0.98
1
-0.022
1
-0.022
-0.019
-0.022
-0.022
1
-0.022
1
0.85
-0.019
-0.019
0.85
-0.019
0.85
1
1.0
0.8
0.6
0.4
0.2
0.0

DATA CONVERSION

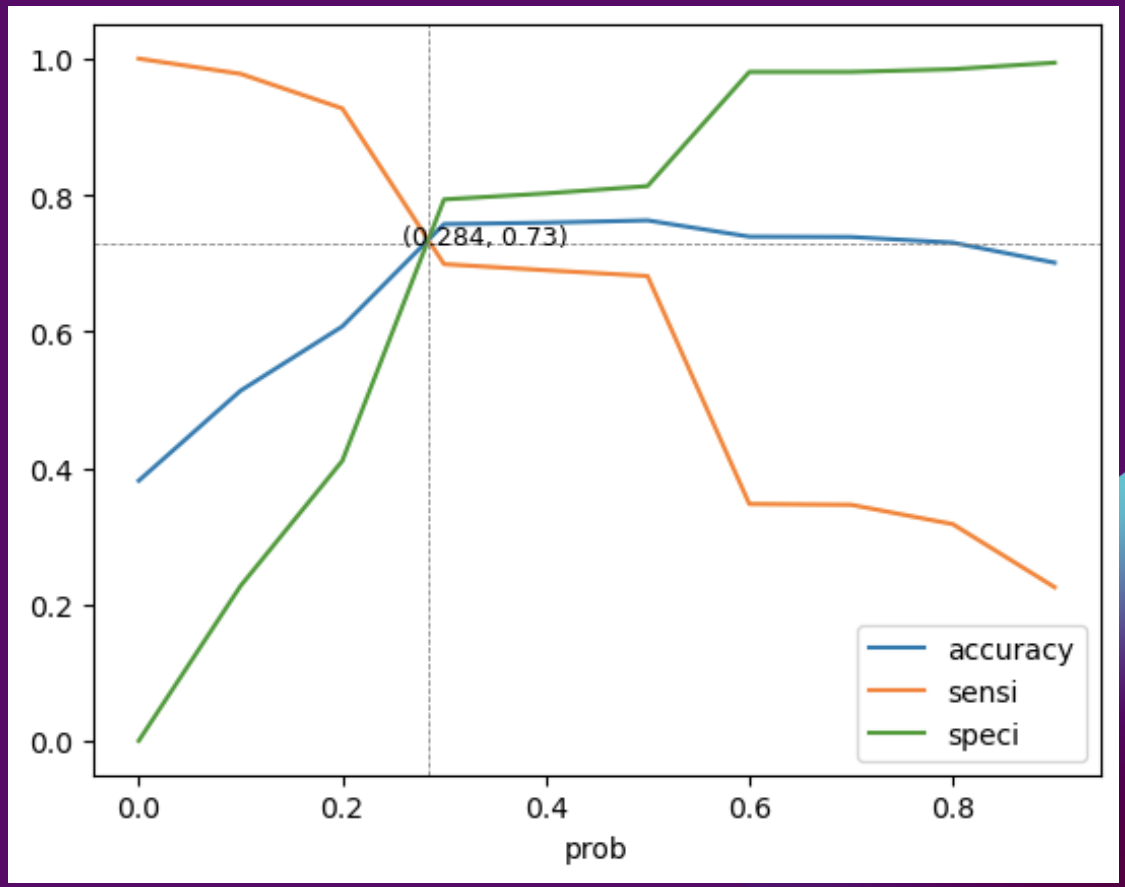
- Total Rows for Analysis- 37
- Total Columns for Analysis- 9240
- Numerical Variables are normalized
- Dummy Variables are created for object type variables


MODEL BUILDING
Data Collection
Data Visualization
Exploratrory Data Analysis
Data Preperation
Data Cleaning

MODEL BUILDING

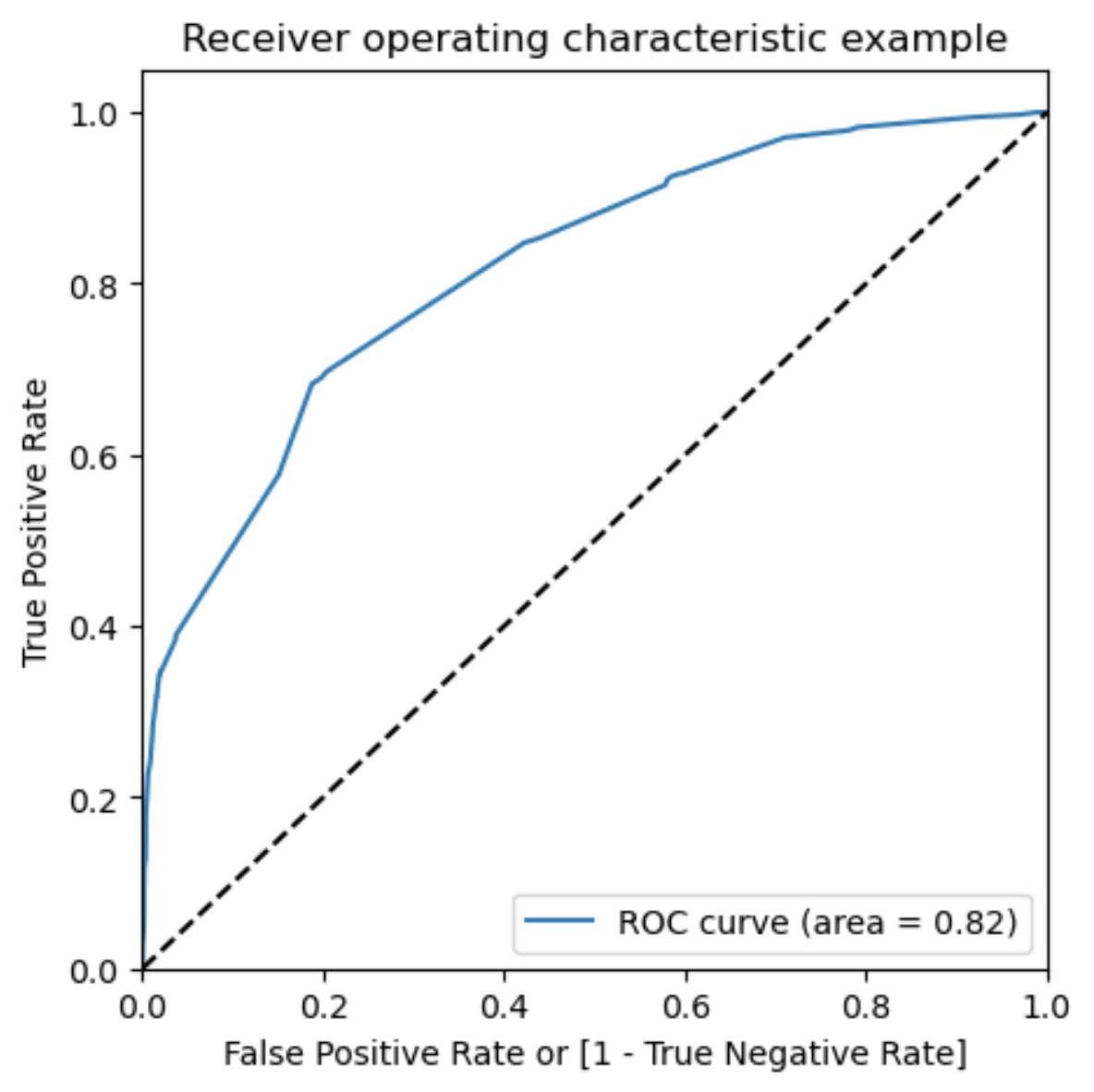
- Splitting the Data in to Training & Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose P-value is grater than 0.05 and VIF value is greater than 5
- Predictions on test data set.
- Over all accuracy 81%.

OPTIMAL CUT OFF POINT



- Optimal cut-off probability is that probability where we get balanced sensitivity and specificity
- From the second graph it is visible that the optimal cut off is at 0.284

ROC CURVE


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.82)

PREDICTION ON TEST SET

- The evaluation matrix are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.
- For the Test Set:

  Accuracy: 76.28%
  Sensitivity: 68.13%
  Specificity: 81.31%

- These metrics are very close to train set, so out final model logm4 is performing with good consistency on both Train & Test Set.
- This shows that our test prediction is having accuracy, precision and recall scores in an acceptable range.

CONCLUSION

- Top 3 Features that contributing Positivesly to pretending hot leads in the model are:

1. Lead Source_Welingak Website
2. Current_occupation_Working Professional
3. Lead Source_Reference

- Keeping these features in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

RECOMMENDATIONS

- Focus on features with positive co-efficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimizing communication channels based on lead engagement impact.
- Engage working professionals with tailored messaging.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have conversion rate and will have better financial situation to pay higher fees too.

THANK YOU