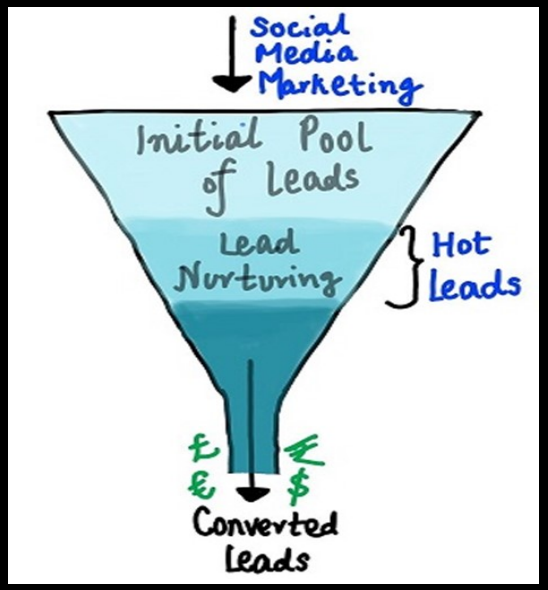
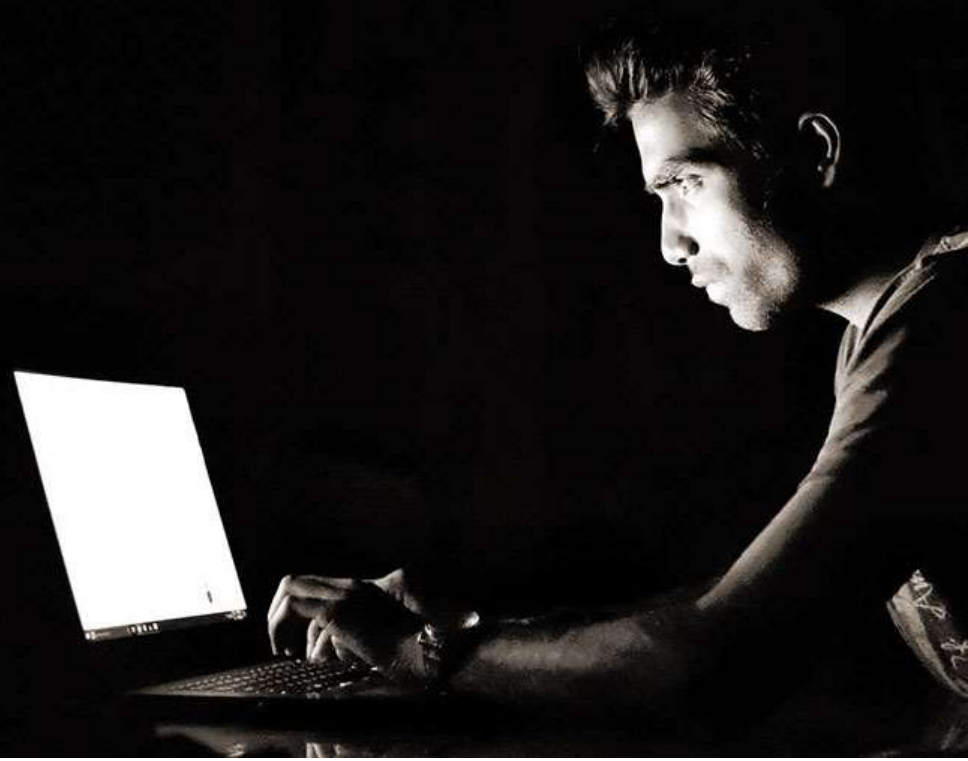
LEAD SCORING CASE STUDY

Submitted by:
VANDANA N S
RAJAT PRATAP SINGH

Problem Statement

The X Education company requires you to build a logistic regression model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

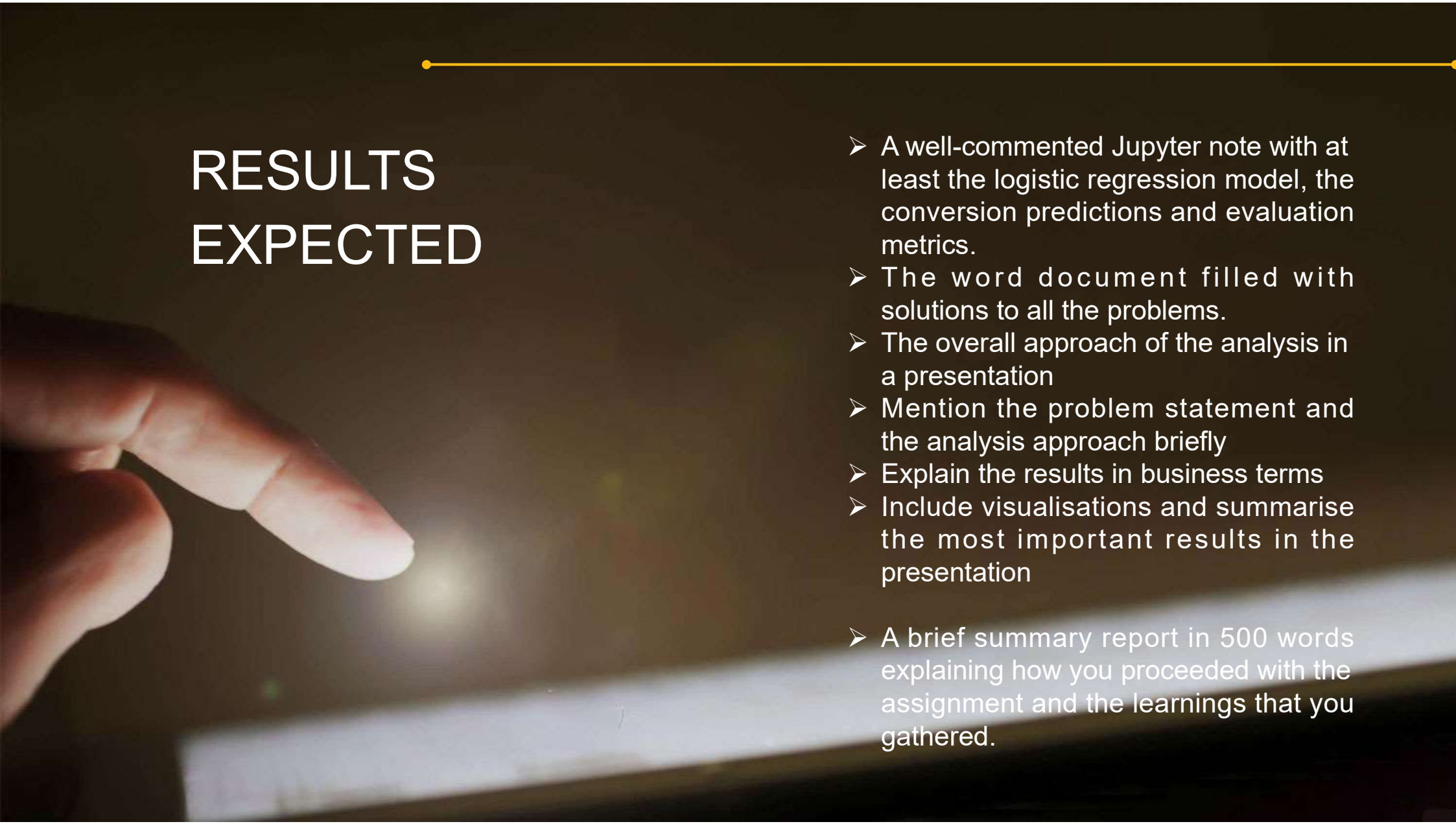




Agenda

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.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.



RESULTS EXPECTED

- A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics.
- The word document filled with solutions to all the problems.
- The overall approach of the analysis in a presentation
- Mention the problem statement and the analysis approach briefly
- Explain the results in business terms
- Include visualisations and summarise the most important results in the presentation

- A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Roadmap

Created train and test set by splitting the original cleaned data set after treating

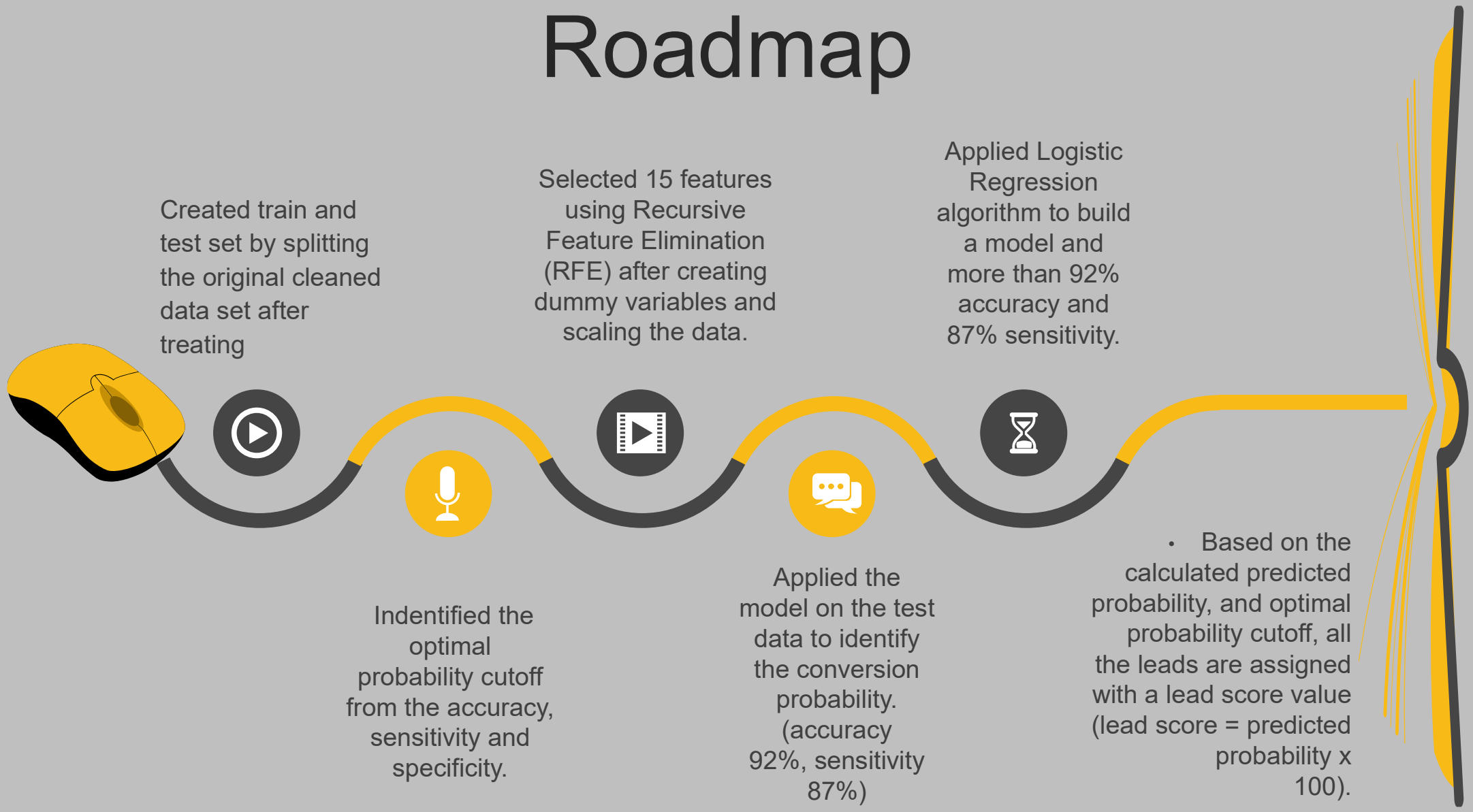
Indentified the optimal probability cutoff from the accuracy, sensitivity and specificity.

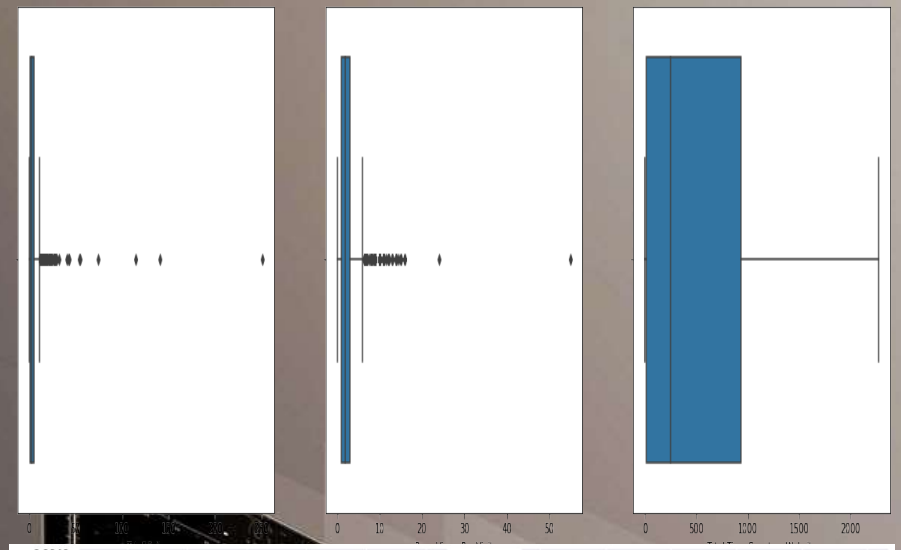
Selected 15 features using Recursive Feature Elimination (RFE) after creating dummy variables and scaling the data.

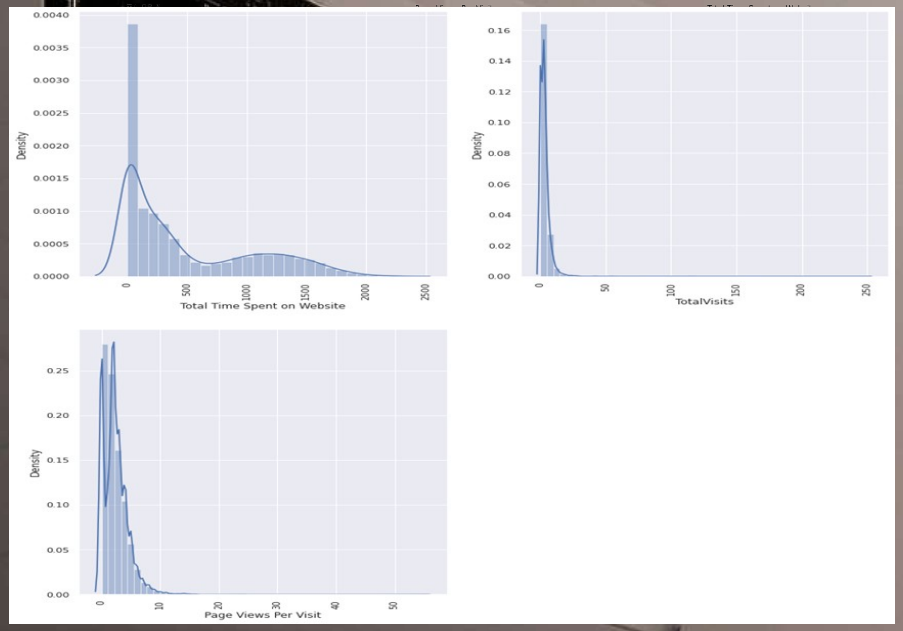
Applied the model on the test data to identify the conversion probability. (accuracy 92%, sensitivity 87%)

Applied Logistic Regression algorithm to build a model and more than 92% accuracy and 87% sensitivity.

- Based on the calculated predicted probability, and optimal probability cutoff, all the leads are assigned with a lead score value (lead score = predicted probability x 100).

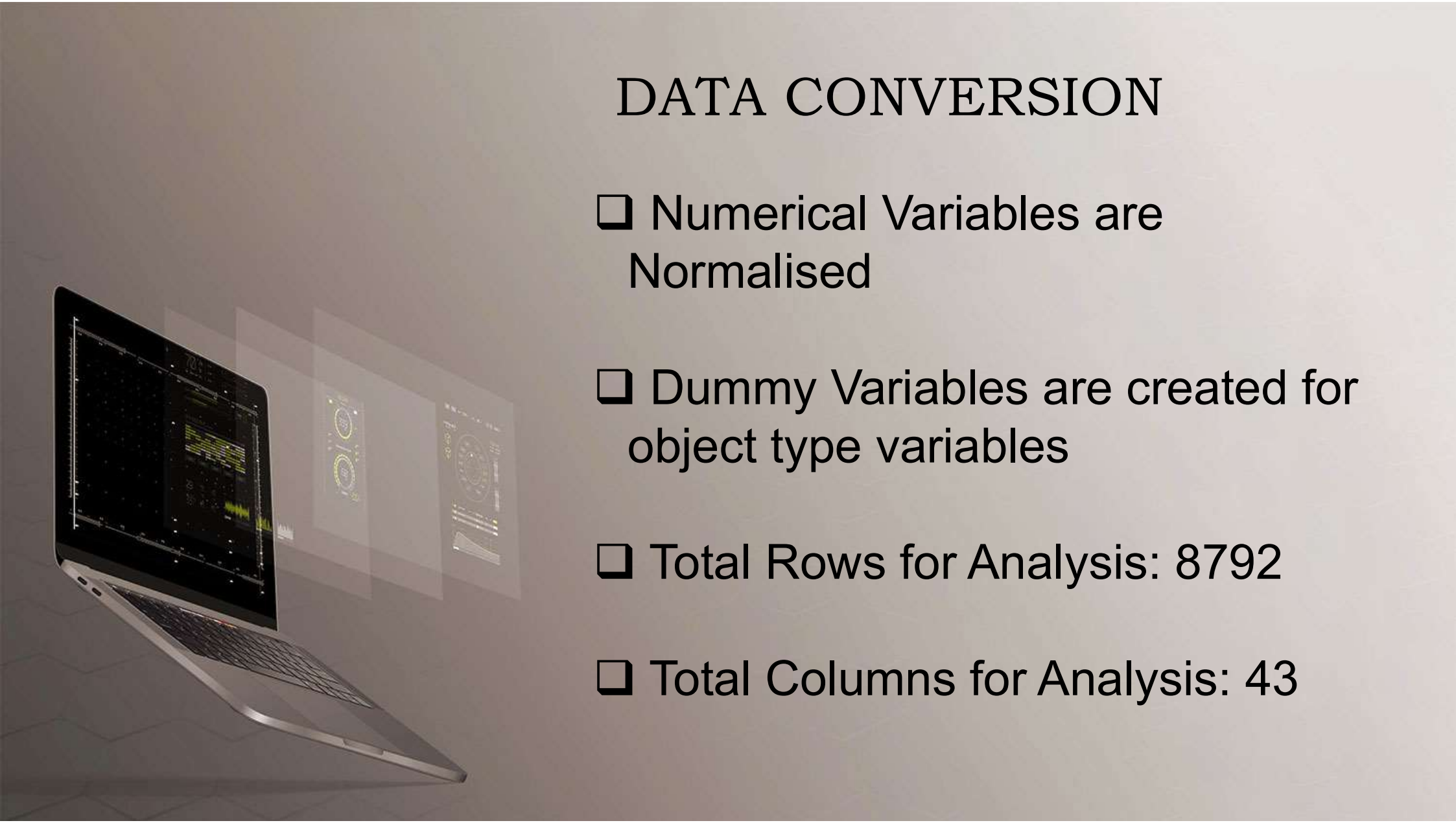






OUTLIERS HANDLING

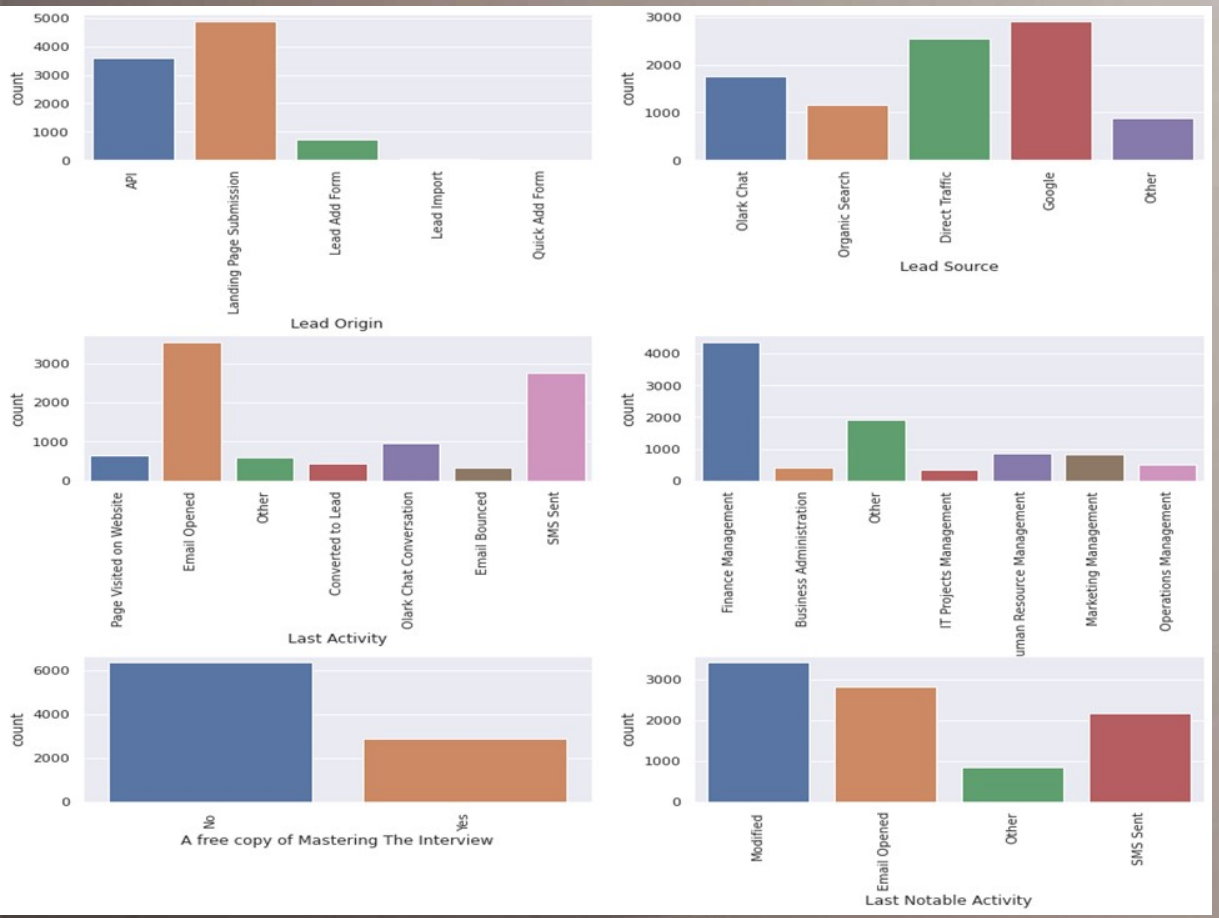
- Continuous Variables are not in Normal distribution
- Outliers are there in Total Visits and Page Views Per Visit
- We can see that the total visits have more values is between 0-50 and page views per visits 0-20



DATA CONVERSION

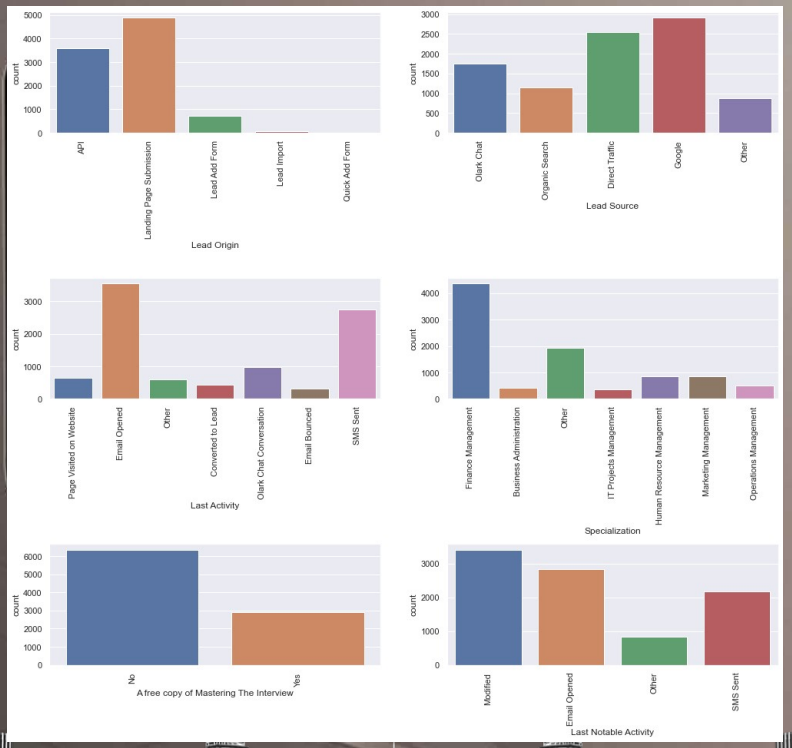
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

EDA



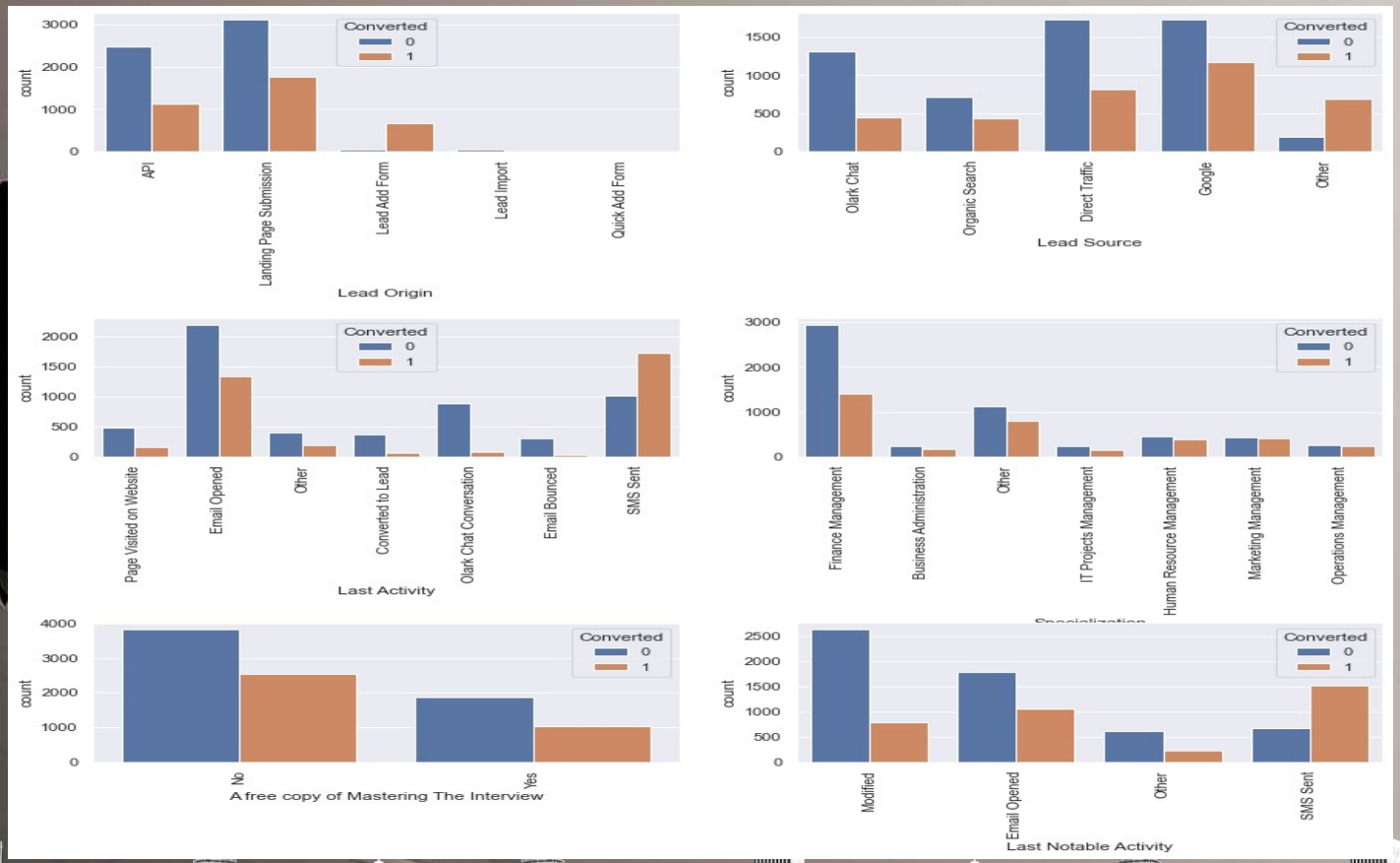
- Lead Source we can see that Direct Traffic and Google are the two main source for Leads
- Email Opened and SMS Sent in Last Activity is high we can see
- Finance Management Specialization is the most chosen one in the specialization category

Univariate Analysis

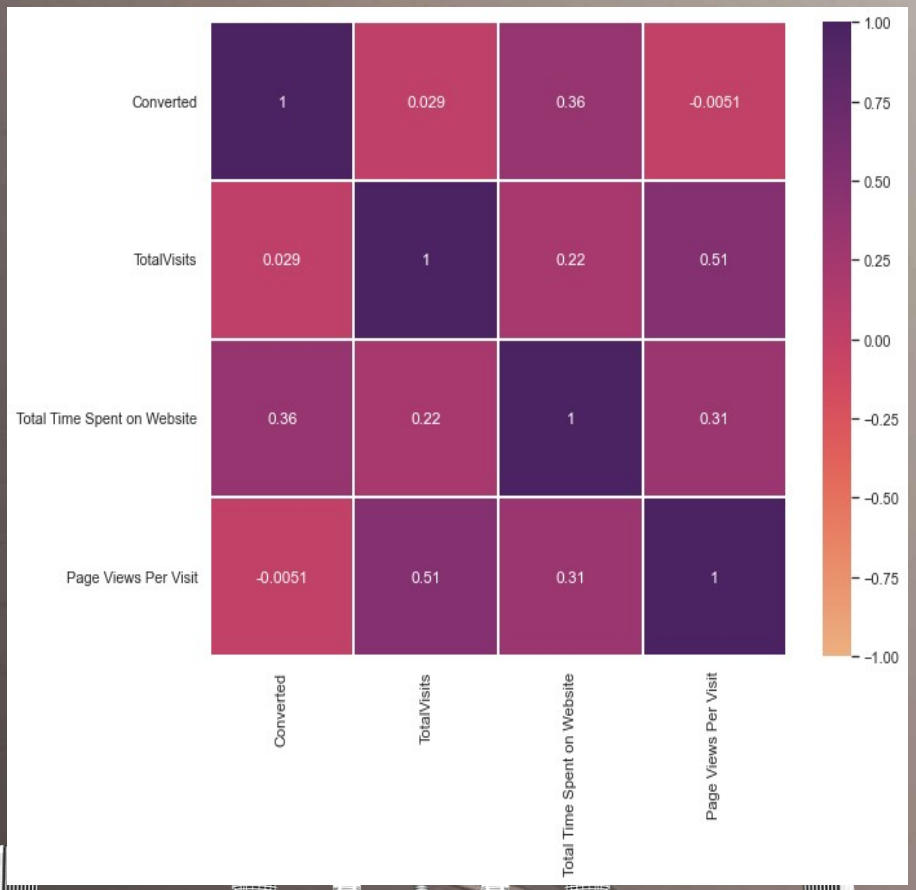


- Lead Origin
- Lead Source
- Last Activity
- Specialization
- A free copy of Mastering The Interview
- Last Notable Activity

Bivariate Analysis



Correlation

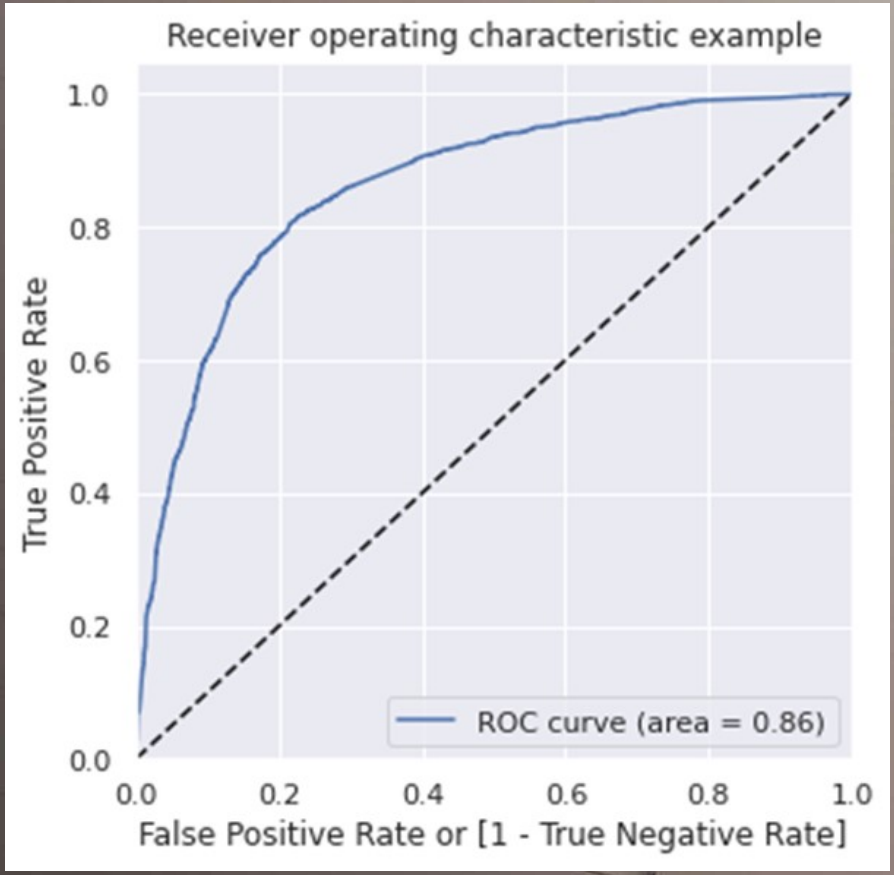


- Total visits
- Converted
- Total time spent on website
- Page views per visit

ML MODEL

- *Splitting the Data into Training and 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 greater than 0.05 and VIF value is greater than 5*
- *Predictions on test data set*
- *Overall accuracy 81%*

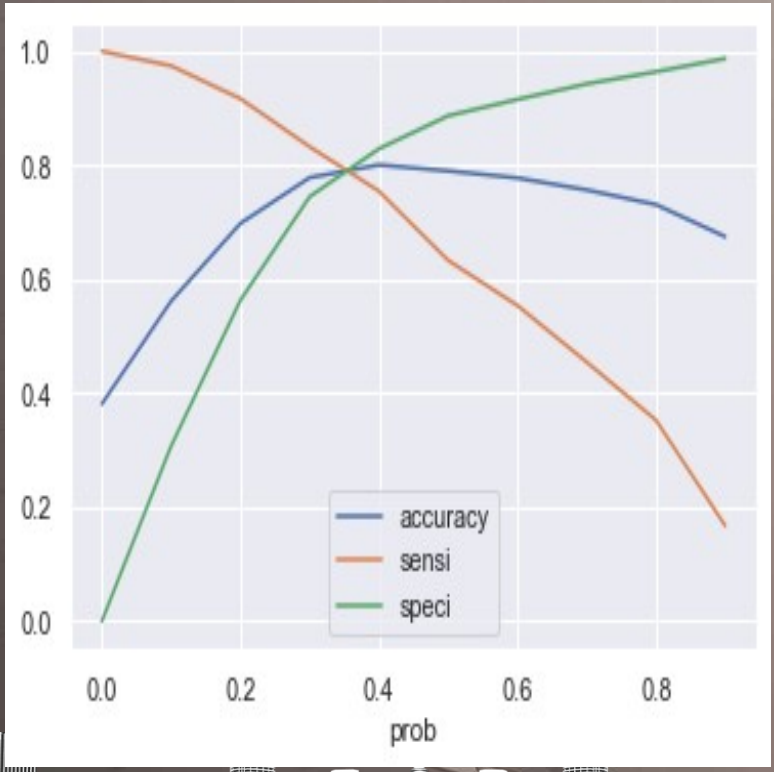
ROC Curve



- The ROC curve shows that the 96% of the area is under the curve.
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

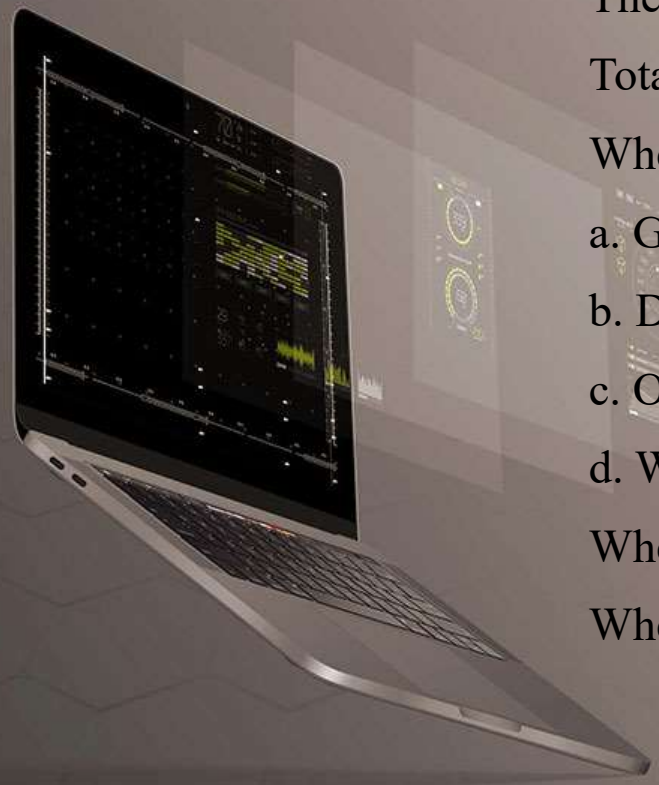
Optimal Probability Cutoff

- *Optimal probability cutoff is identified as 0.35 for better accuracy of the classification of lead conversion.*



- *With 0.35 cutoff the model has*
  - *Accuracy : 78%*
  - *Sensitivity : 77%*
  - *Specificity : 78%*

Conclusion



Below are the points via which we can a get a understanding between a hot lead and a cold lead

The total time spend on the Website.

Total number of visits.

When the lead source was from the below sites:

a. Google

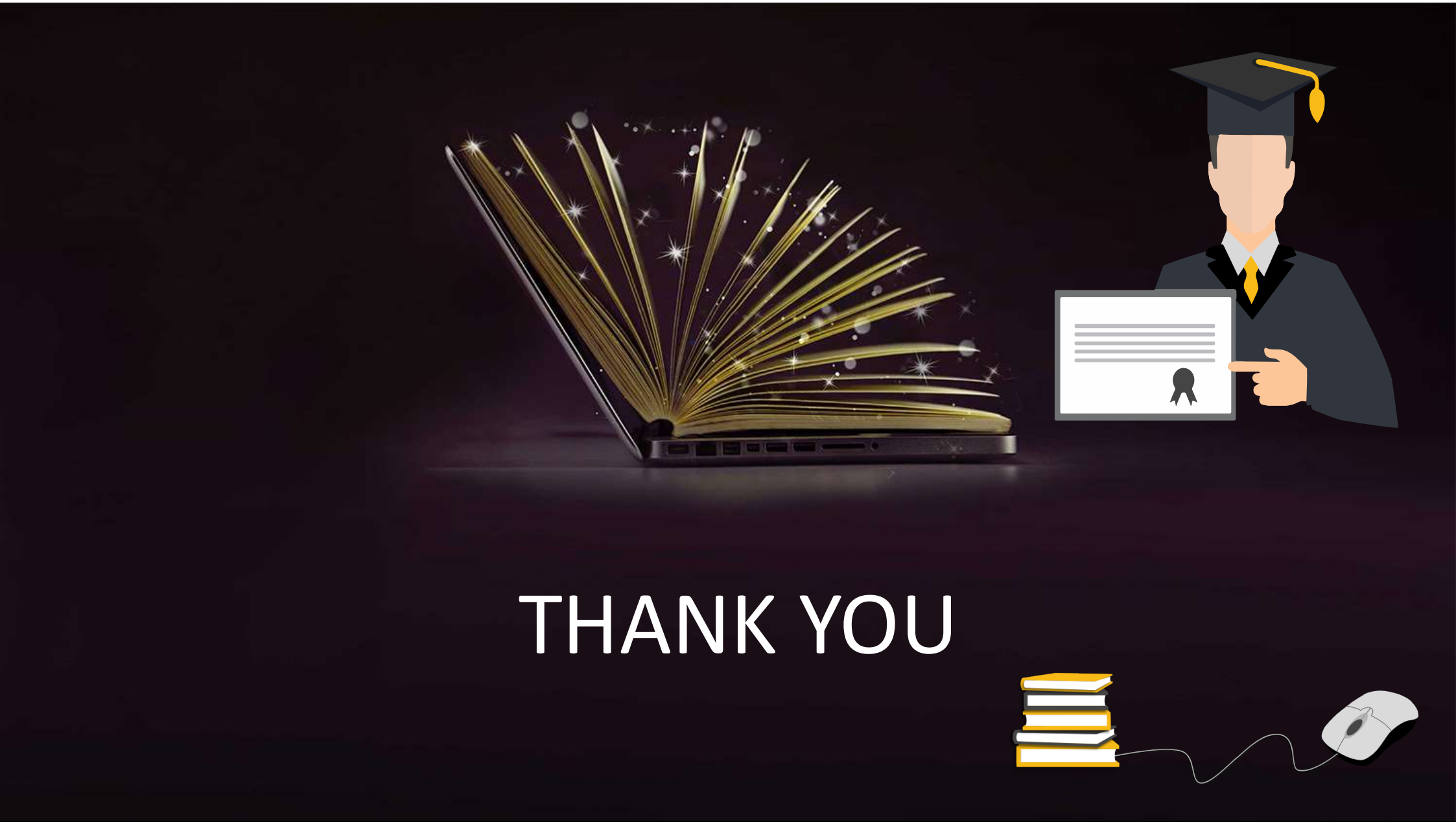
b. Direct traffic

c. Organic search

d. Welingak website

When the lead origin is Lead add format

When their current occupation is as a working professional.


THANK YOU