LEAD SCORING

Ridheesh Bhan, Piyush Mandal & Ramyaseetha

PROBLEM STATEMENT

- ✓ X Education sells online courses to professionals in the sector.
- ✓ Although X Education receives a large number of leads, its lead conversion rate is quite low. For example, if they get 100 leads in a day, only around 30 of them will be converted.
- ✓ To make this process more effective, the organization wants to identify the most promising prospects, commonly known as "Hot Leads."
- ✓ If they are successful in identifying this set of prospects, the lead conversion rate should increase because the sales staff will now be focusing on connecting with the potential leads rather than calling everyone.

BUSINESS OBJECTIVE

- ✓ X education is looking for the most promising leads.
- ✓ They wish to create a Model that detects hot leads for this purpose.
- ✓ The model is being deployed for future use.

SOLUTION METHODOLOGY

Data cleaning and data manipulation

- ✓ Check for and deal with duplicate data.
- ✓ Check for and handle NA and missing values.
- ✓ Drop columns that have a substantial number of missing data and are no longer useful for the analysis.
- ✓ If necessary, the values are imputed.
- ✓ Check for and handle data outliers.

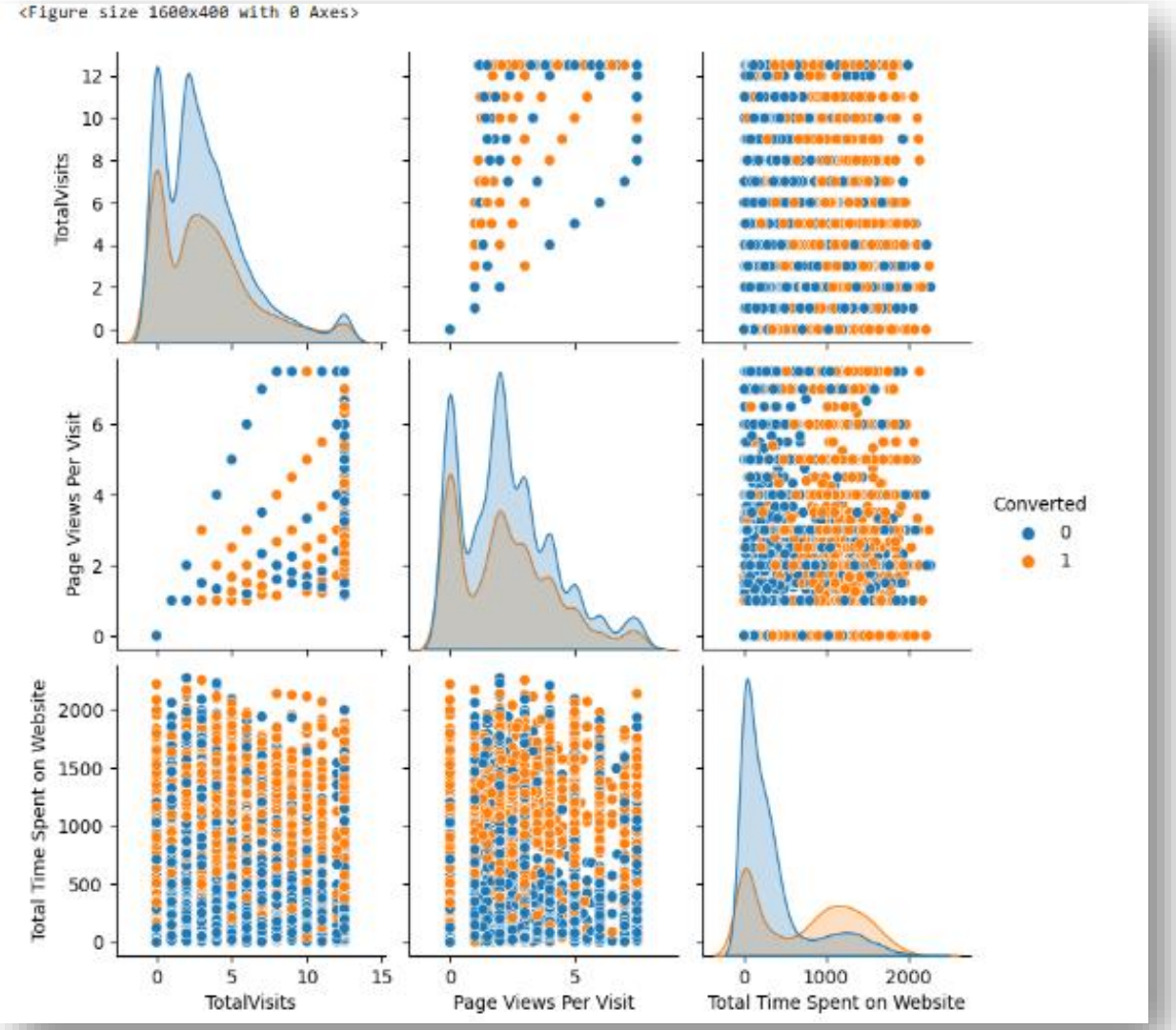
Exploratory Data Analysis (EDA)

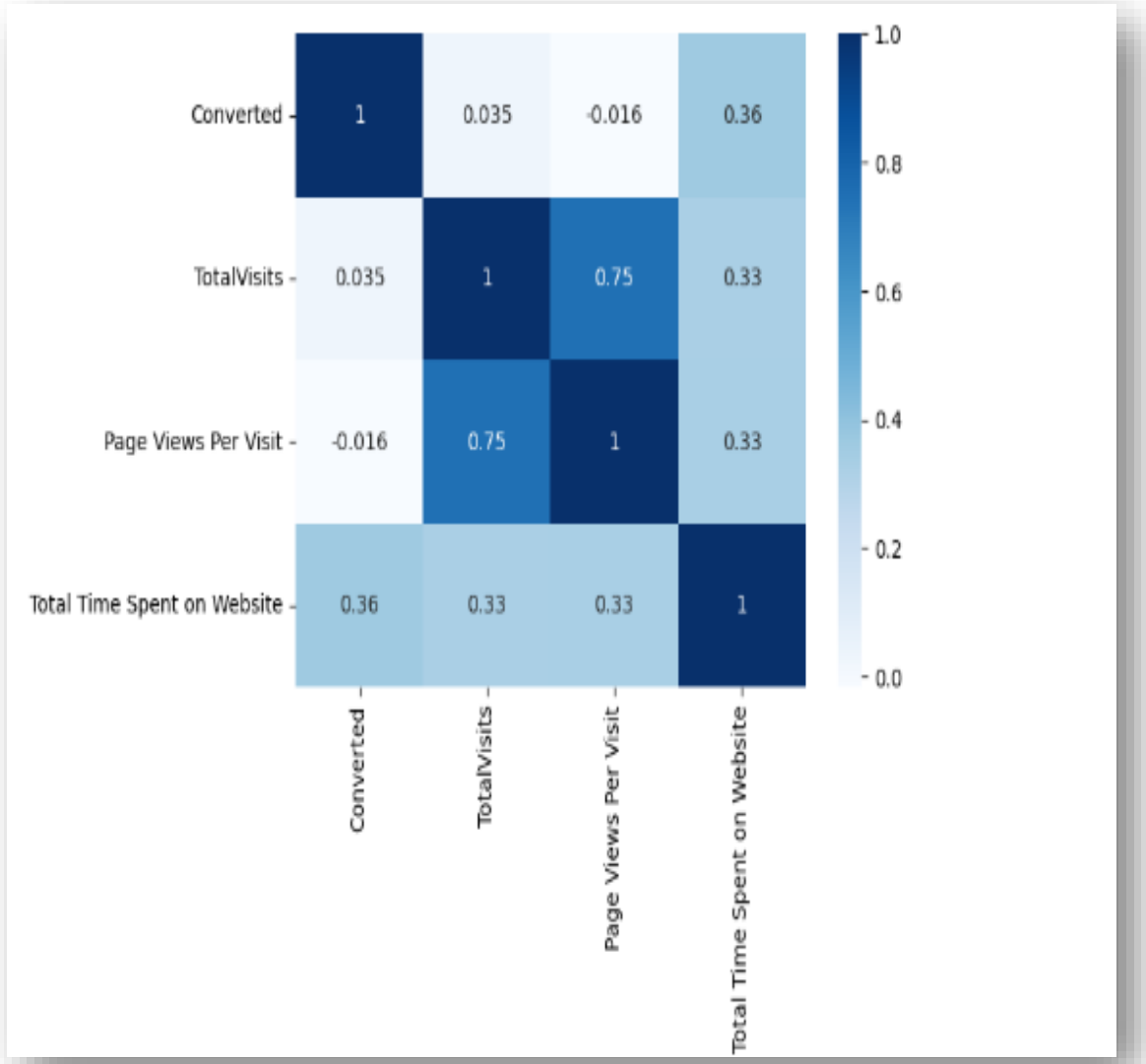
- ✓ Univariate data analysis: value count, variable distribution, and so on.
- ✓ Bivariate data analysis: correlation coefficients, pattern between variables, and so on.
- ✓ Data encoding and feature scaling, as well as dummy variables.
- ✓ The logistic regression classification approach is used for model creation and prediction.
- ✓ The model is validated.
- ✓ Conclusions and recommendations are included in the model presentation.

DATA MANIPULATION

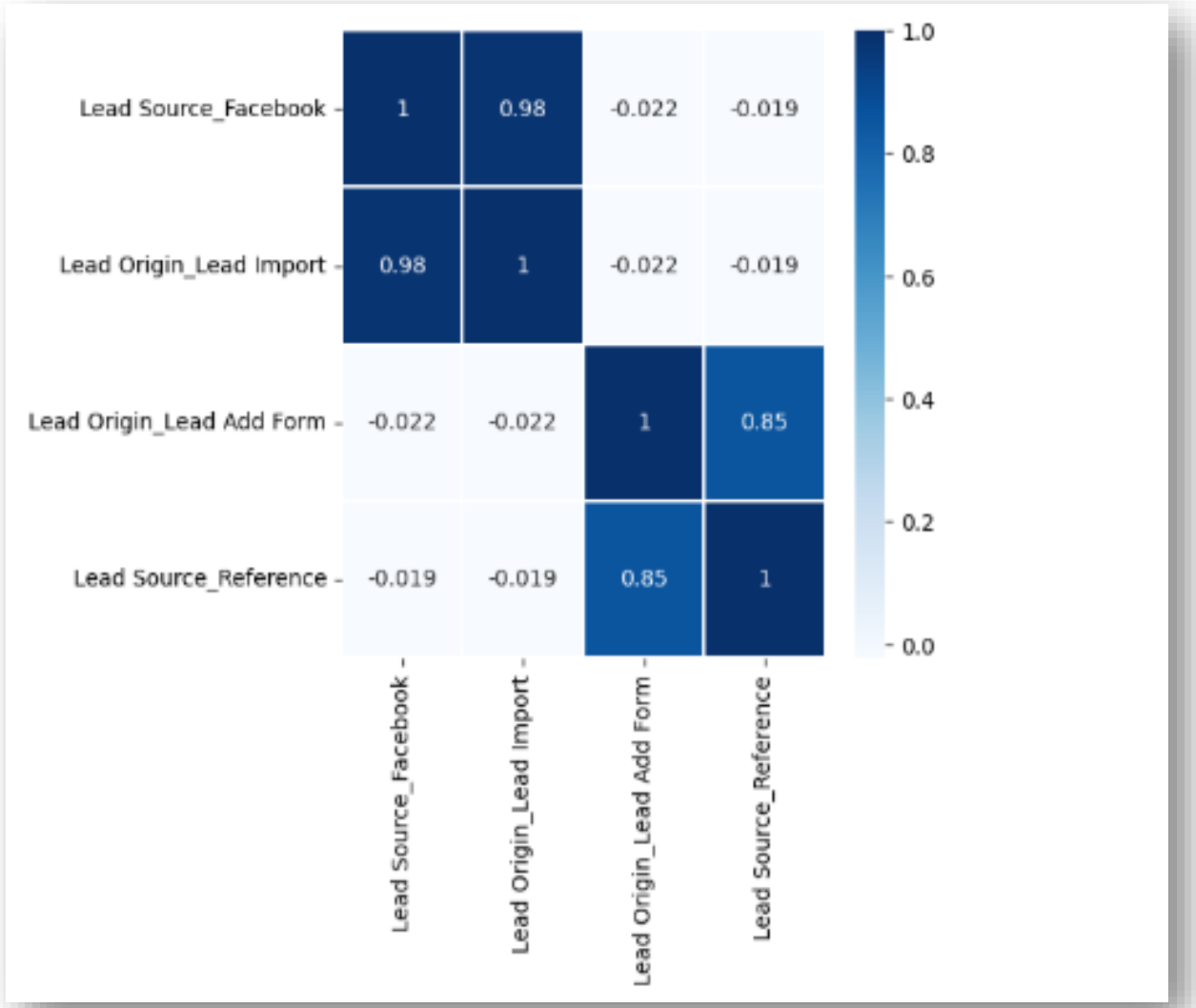
- ✓ Total Number of Rows=37, Total Number of Columns =9240.
- ✓ Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update my supply”
- ✓ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- ✓ Removing the “Prospect ID” and “Lead Number” which are not necessary for the analysis.
- ✓ After checking for the value counts for some of the object type 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”, “XEducation Forums”, “Newspaper”, “Digital Advertisement” etc.
- ✓ Dropping the column shaving more than 35% as missing values such as ‘How did you hear about X Education’ and ‘Lead Profile’.

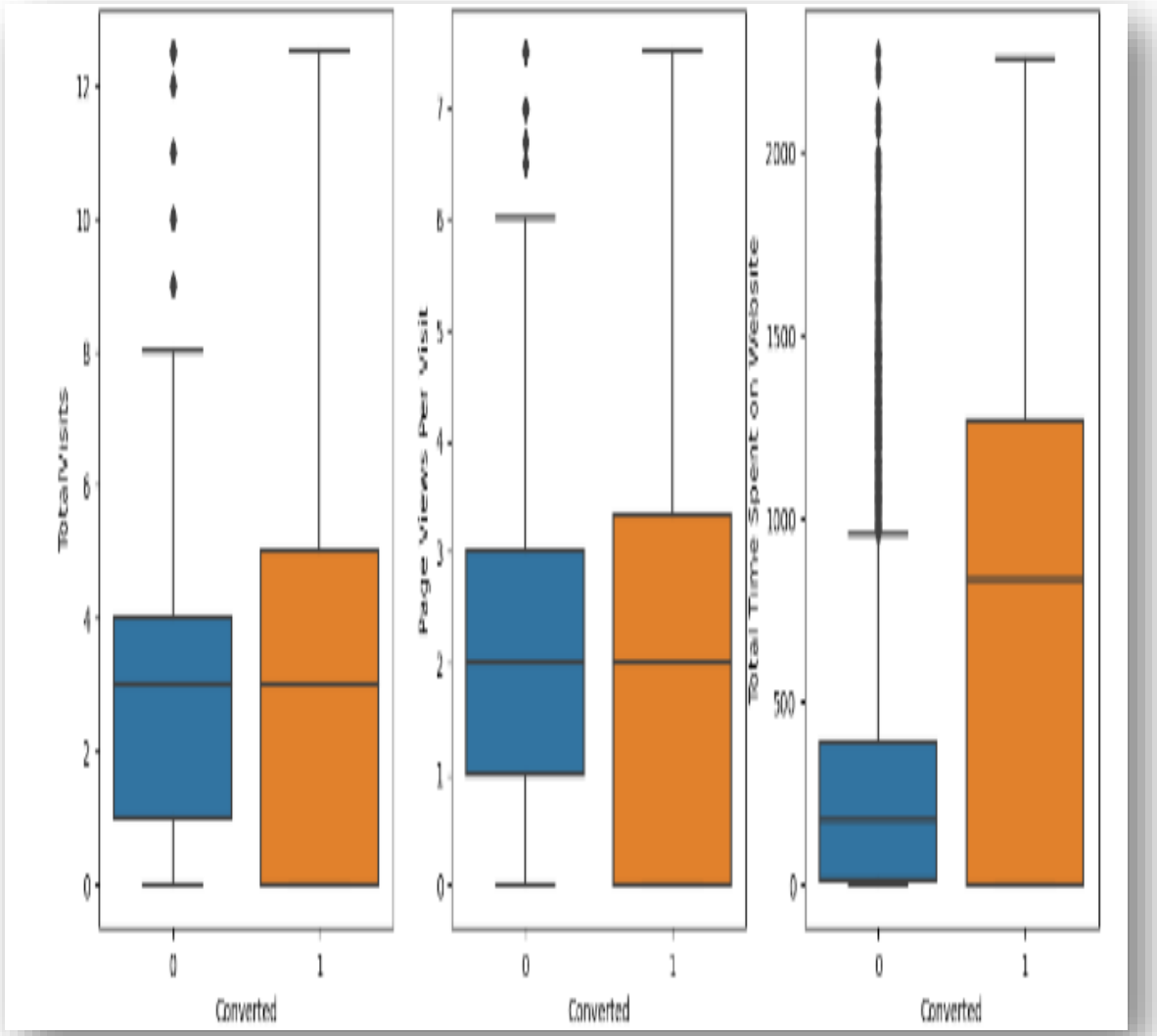
EXPLORATORY DATA ANALYSIS (EDA) & HEAT MAP





HEAT MAP & BOX PLOT





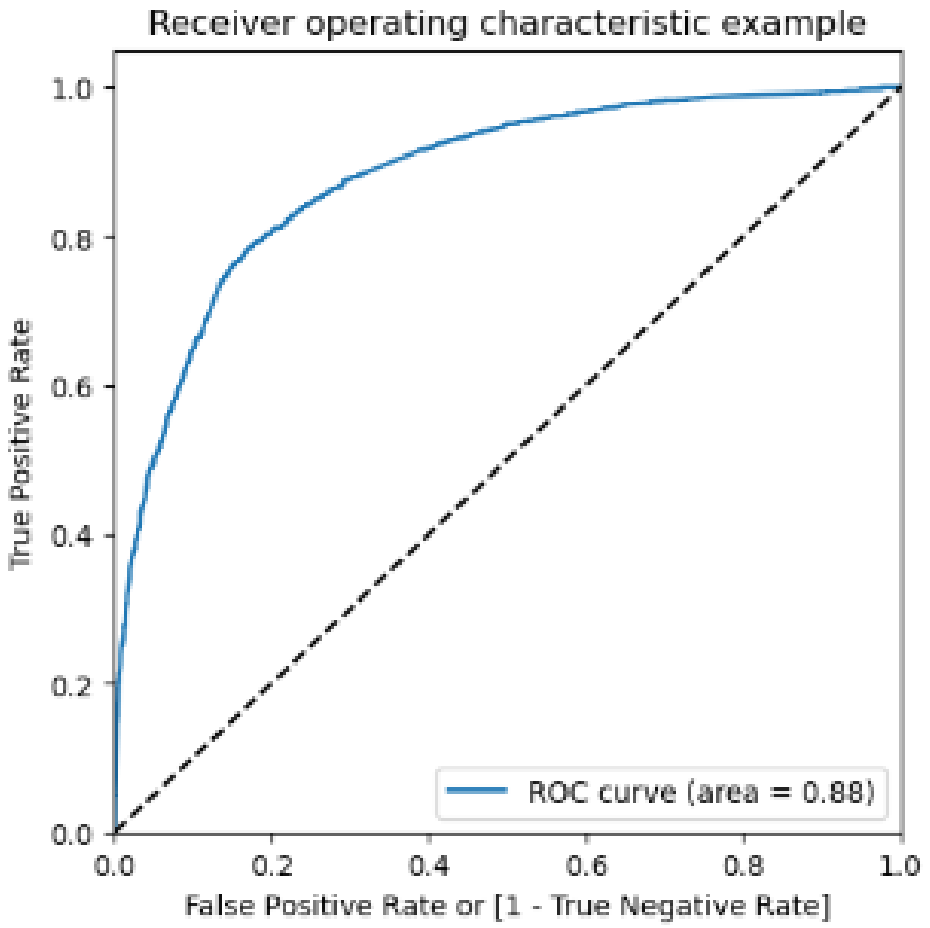
MODEL BUILDING

- ✓ 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 vi value is greater than 5
- ✓ Predictions on test data set
- ✓ Overall accuracy 81%

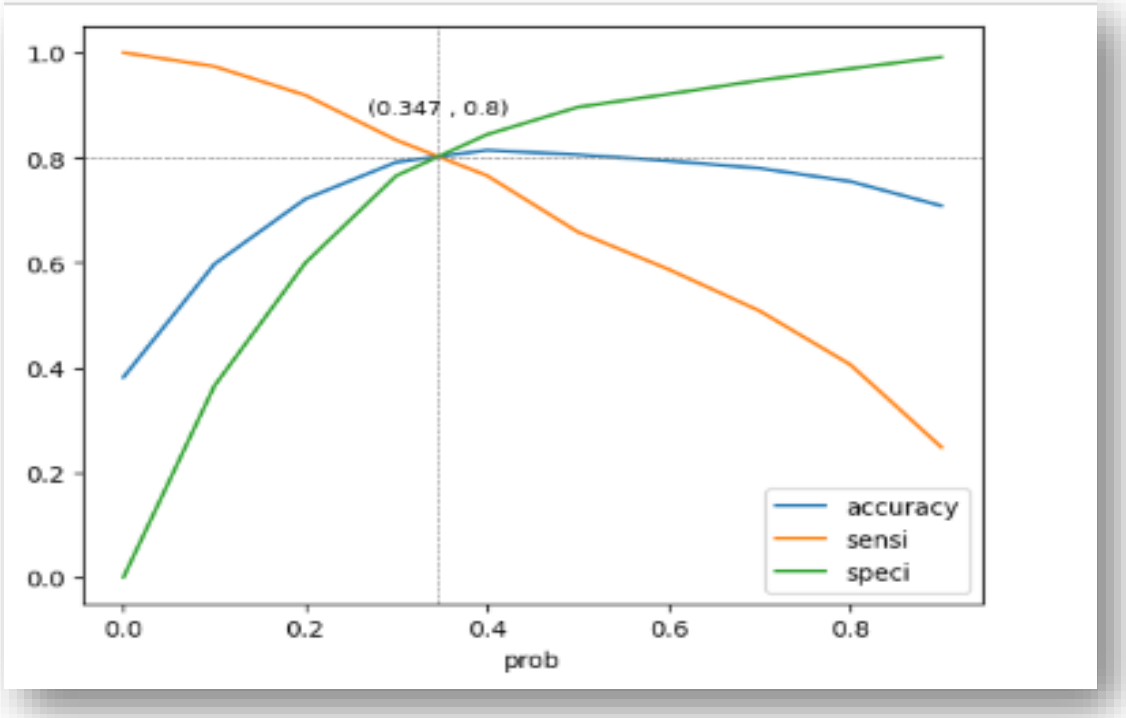
DATA CONVERSION

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

ROC CURVE



- ✓ Finding 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 0.35



PREDICTION ON TEST SET

- ✓ Prior to making predictions on the test set, it is necessary to standardize the test set and ensure that the final training dataset and the test set have identical columns.
- ✓ Subsequent to completing the aforementioned step, we commenced prediction on the test set, and the resultant prediction values were stored in a fresh dataframe.
- ✓ Following this, we conducted model evaluation, encompassing accuracy, precision, and recall calculations.
- ✓ The computed accuracy score was 0.82, with precision and recall both approximately at 0.75.
- ✓ This indicates that our test predictions exhibit accuracy, precision, and recall scores within an acceptable range.
- ✓ Furthermore, this underscores the stability of our model, characterized by commendable accuracy and recall/sensitivity.
- ✓ A lead score is generated using the test dataset to identify hot leads; a higher lead score corresponds to an increased likelihood of conversion, whereas a lower lead score suggests a diminished likelihood of conversion.

CONCLUSION

The variables most relevant to potential buyers were identified as follows (in descending order):

- ✓ The total time spent on the Website.
- ✓ The total number of visits.
- ✓ The lead sources: Google, Direct traffic, Organic search, Welingak website.
- ✓ The last activities: SMS, Olark chat conversation.
- ✓ The lead origin as Lead add format.
- ✓ The current occupation as a working professional.

Considering these factors, X Education has the opportunity to thrive, given the substantial likelihood of persuading nearly all potential buyers to reconsider and purchase their courses.