LEADS SCORING CASE STUDY

SACHIN KUMAR SAHOO
SAGNIK SAHA
SAHIL SHARMA

PROBLEM STATEMENT

- X Education specializes in offering online courses to professionals within various industries.
- Despite attracting a significant number of leads, X Education faces challenges in effectively converting these leads. To illustrate, out of 100 leads acquired in a day, only approximately 30 of them ultimately become customers.
- In an effort to streamline and optimize this process, the company is keen on pinpointing the leads with the highest potential, often referred to as "Hot Leads."
- By successfully identifying this select group of promising leads, X Education anticipates a considerable improvement in its lead conversion rate. This enhancement will result from the sales team's increased focus on engaging with these high-potential leads, as opposed to expending efforts on reaching out to everyone indiscriminately.

BUSINESS OBJECTIVE

- X Education is interested in identifying the most prospective leads in their pool.
- To achieve this, they aim to develop a model that can effectively pinpoint "hot leads."
- They also plan to implement and integrate this model for future use, ensuring its ongoing utility in lead identification.

SOLUTION METHODOLOGY

Cleaning and Manipulating Data:

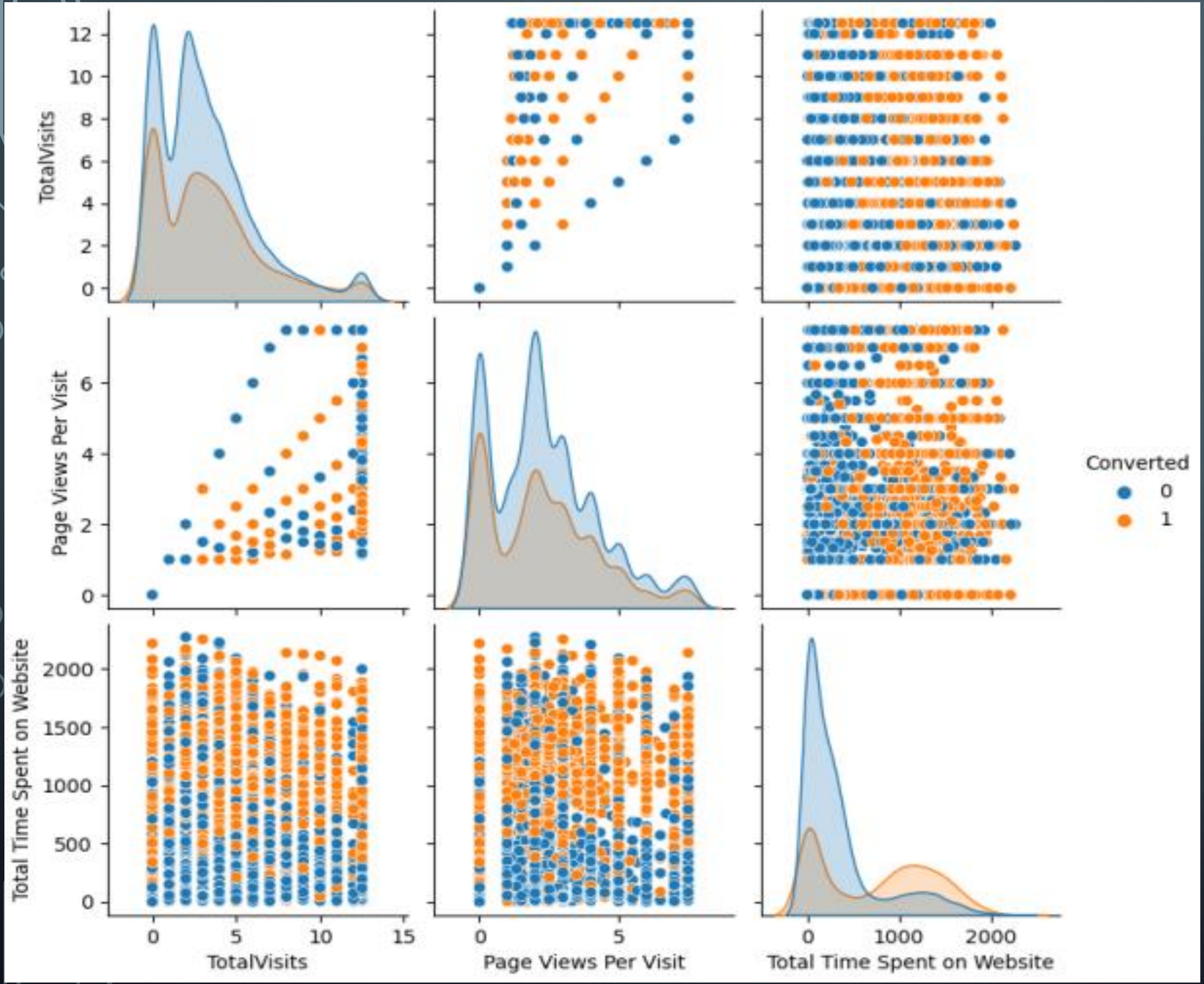
- This involves the process of refining and modifying the dataset.
- Detecting and Managing Duplicate Data: Identifying and addressing instances where the same data appears more than once.
- Handling NA and Missing Values: Dealing with null or missing data points within the dataset.
- Column Removal: Eliminating columns that contain a significant number of missing values and are not pertinent to the analysis.
- Data Imputation: Filling in missing data values if necessary.
- Outlier Detection and Treatment: Identifying and addressing outliers within the dataset.

Exploratory Data Analysis (EDA):

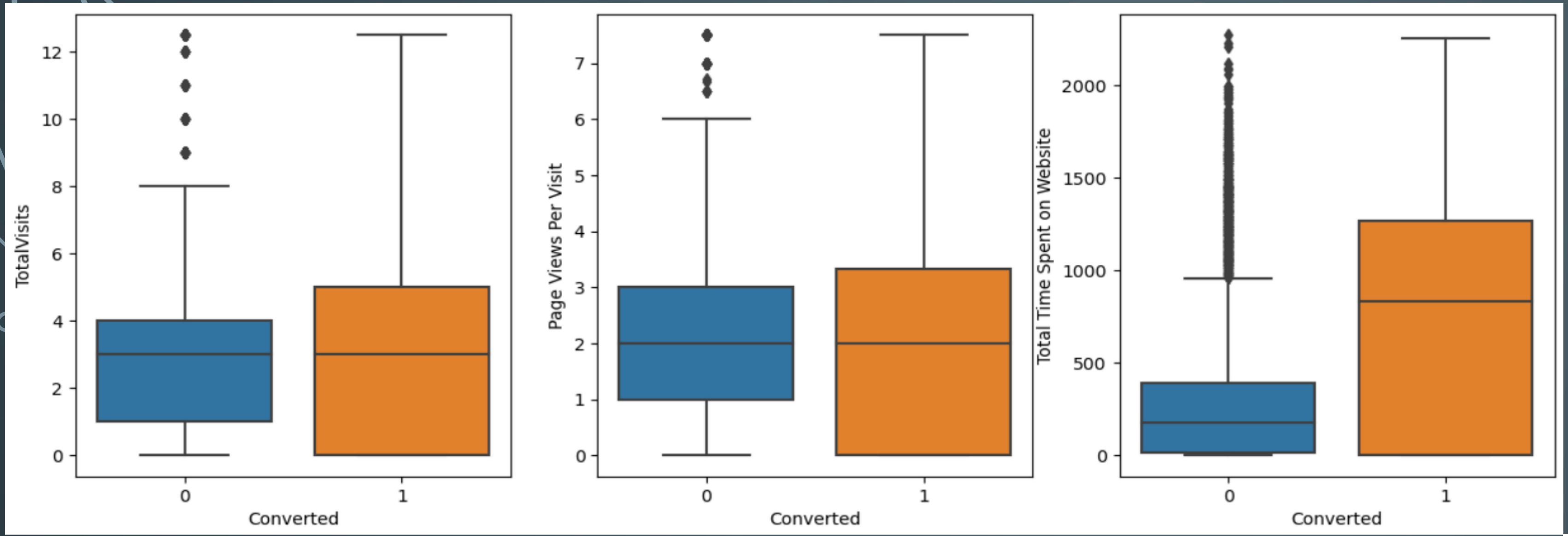
- The examination of data to uncover patterns, distributions, and insights.
- Univariate Data Analysis: Analyzing individual variables, including value counts and variable distributions.
- Bivariate Data Analysis: Exploring relationships and correlations between pairs of variables.
- Feature Scaling and Encoding: Standardizing features and encoding categorical data.
- Classification Technique: Utilizing logistic regression as the primary method for building and predicting with the model.
- Model Validation: Assessing the model's accuracy and reliability.
- Model Presentation: Communicating the model's results and findings.
- Conclusions and Recommendations: Summarizing the outcomes and offering suggestions based on the analysis.

DATA MANIPULATION

- The dataset contains 37 rows and 9240 columns.
- Certain single-value features like "magazine," "receive more updates about our courses," "update my supply," "chain content," "get updates on dm content," and "i agree to pay the amount through cheque" have been removed as they don't provide meaningful variation.
- Unnecessary columns "prospectid" and "lead number" have been excluded from the analysis.
- During the examination of object-type variables, some features were identified with limited variance and thus eliminated. these features include "do not call," "what matters most to you in choosing course," "search," "newspaper, article," "xeducation forums," "newspaper," and "digital advertisement," among others.
- Columns with a high percentage of missing values, such as 'how did you hear about x education' and 'lead profile,' exceeding 35%, have also been dropped from the dataset.

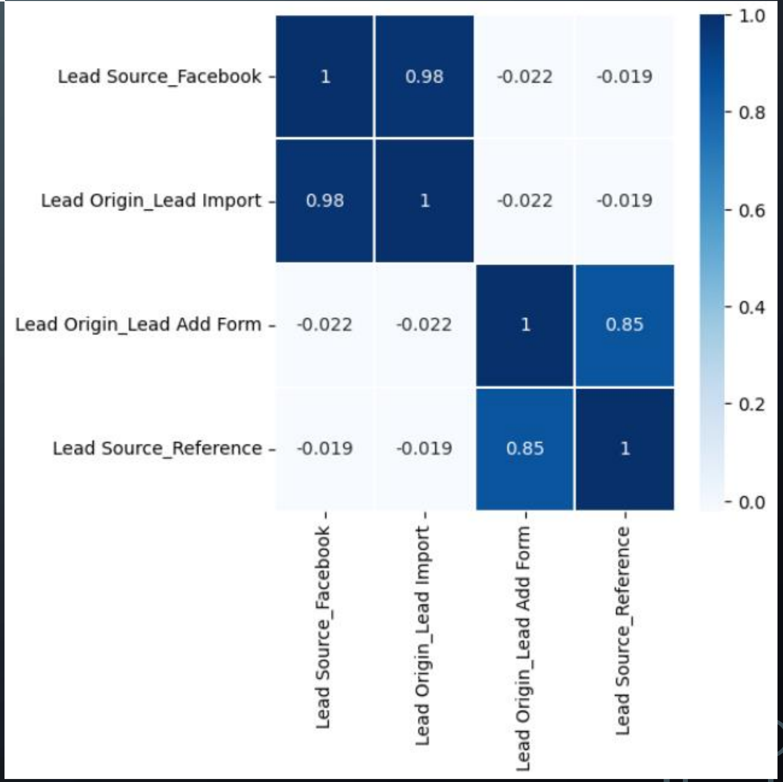


EXPLORATORY DATA ANALYSIS (EDA)



BOX PLOT

HEAT MAP



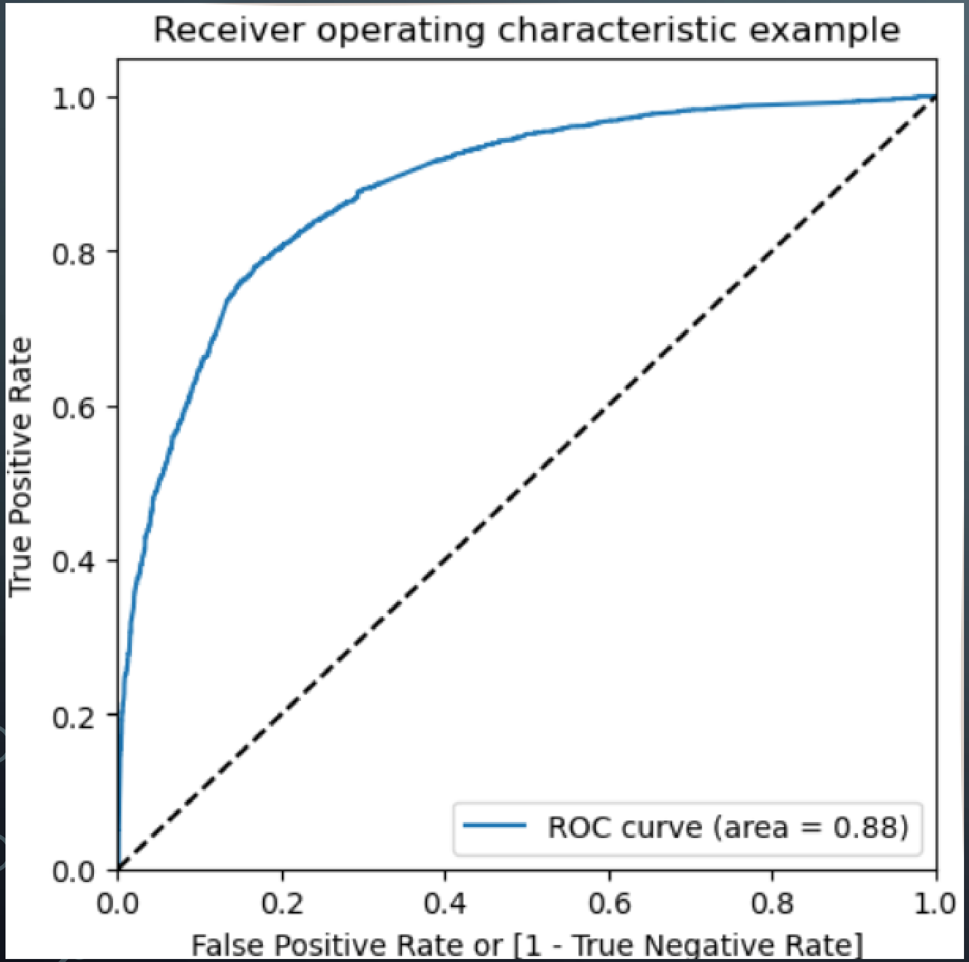
DATA CONVERSION

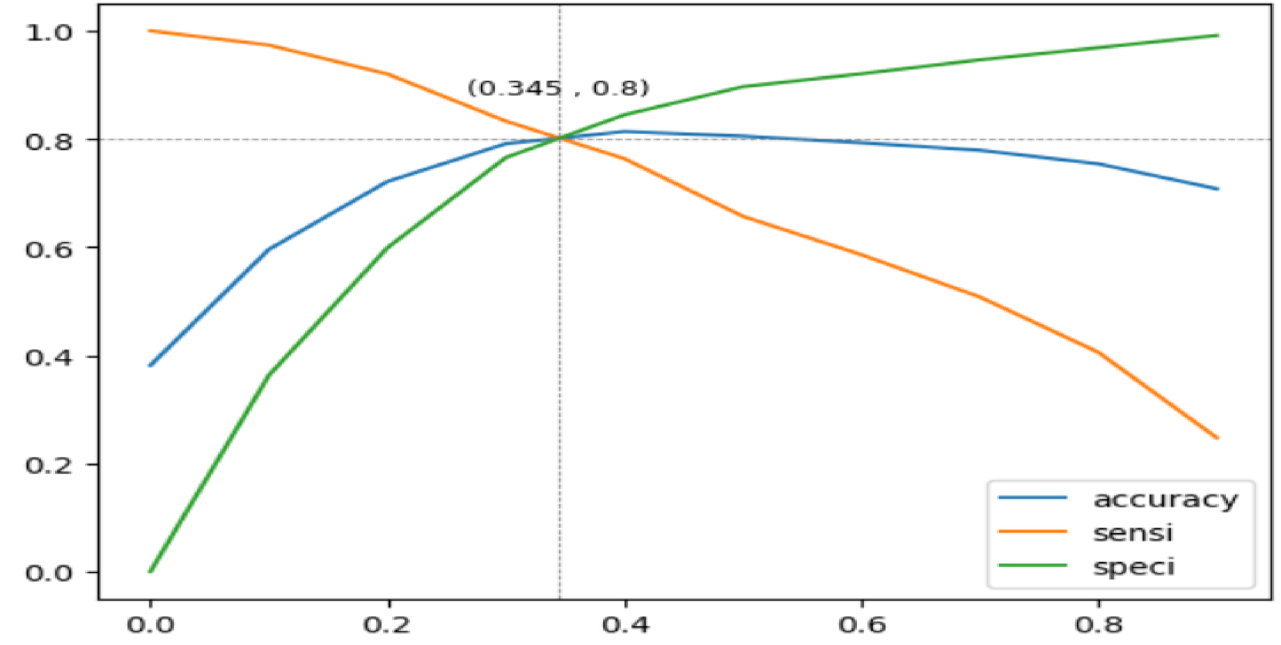
- Numerical variables are normalized
- Dummy variables are created for object type variables
- Total rows foranalysis: 9240
- Total columns for analysis: 37

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%

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

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation i.e. finding the accuracy, precision, and recall.
- The accuracy score we found was 0.82, precision 0.75, and recall 0.75 approximately.
- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (in descending order) :

- The total time spent on the website.
- Total number of visits.
- When the lead source was: google direct traffic organic search welingak website
- When the last activity was: sms olark chat conversation
- When the lead origin is lead add format.
- When their current occupation is as a working professional. keeping these 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.