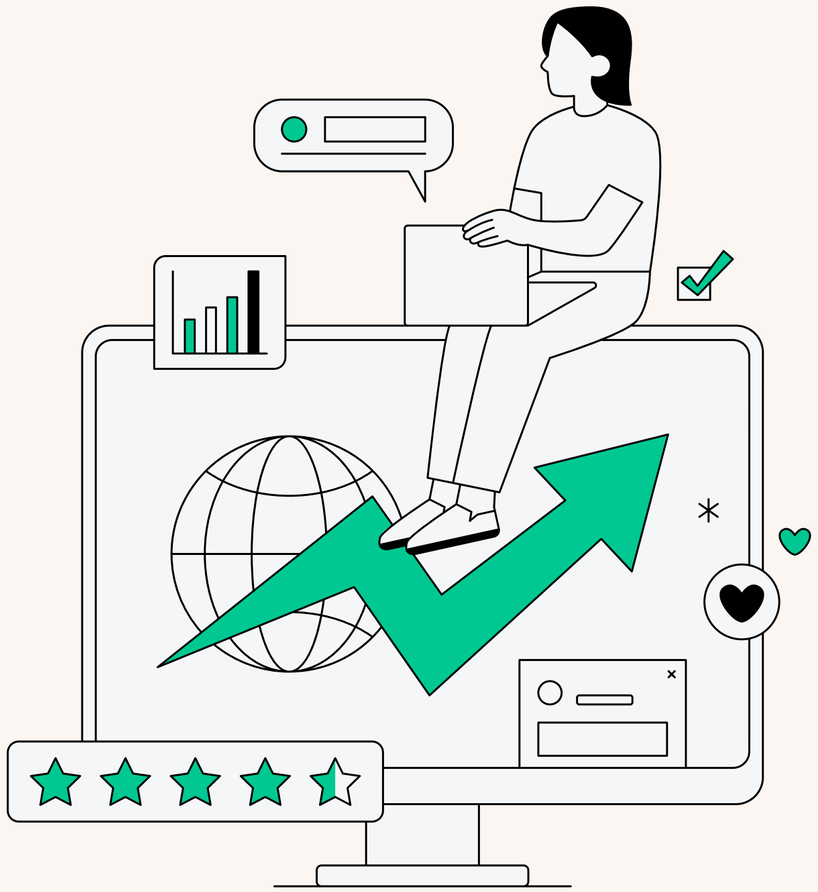
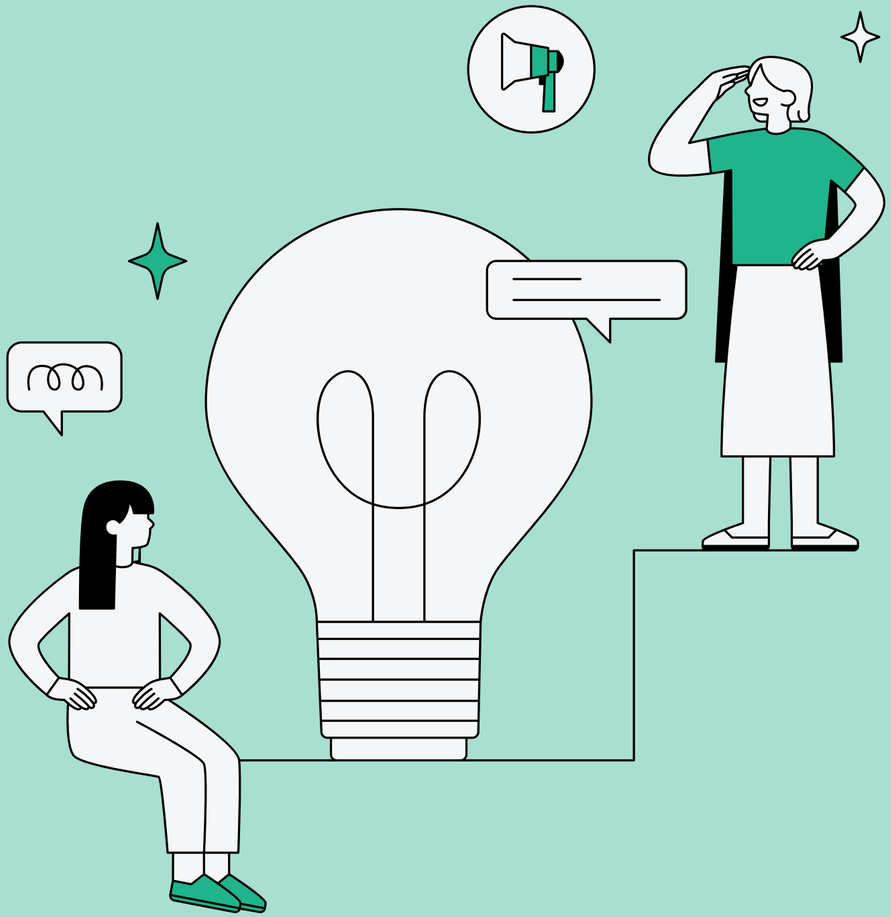
LEADS SCORING CASE STUDY

Presented by Papri Boyra



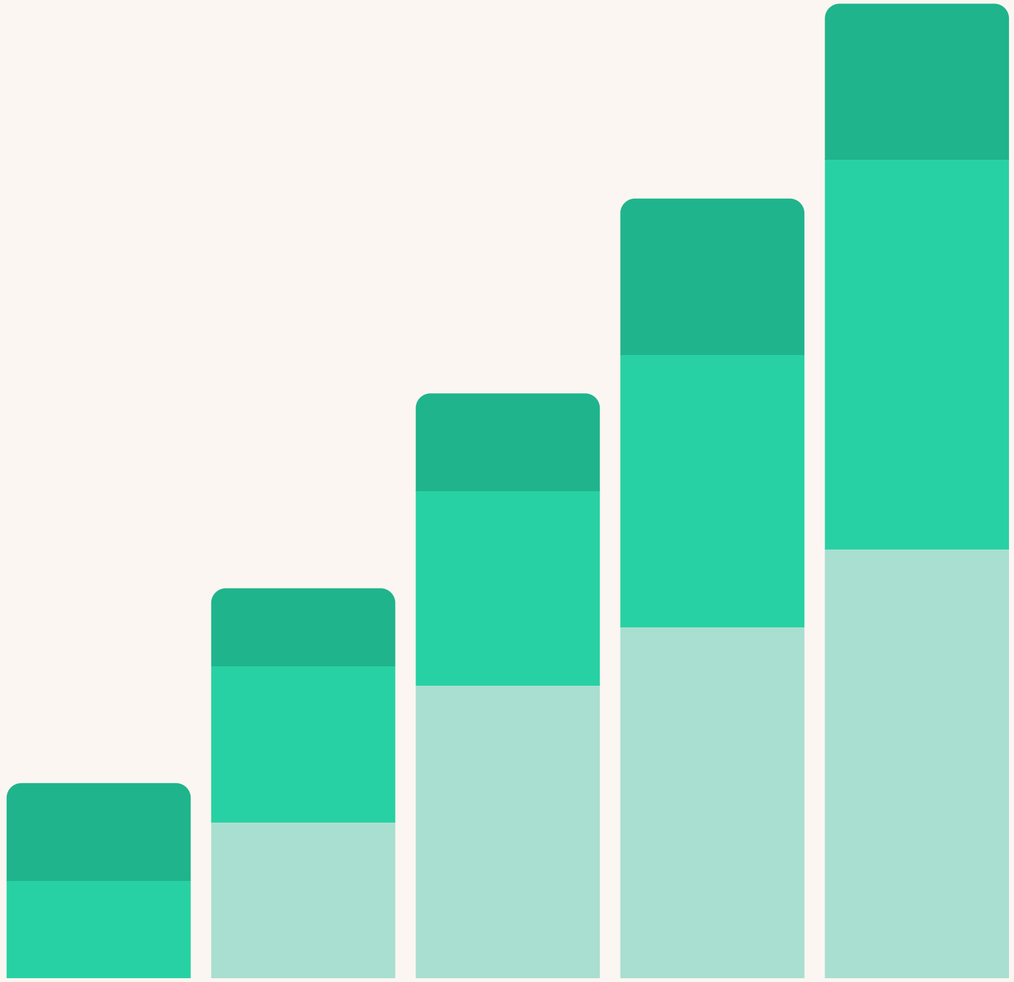
PROBLEM STATEMENT



- Business Challenge: X Education receives many leads daily from various marketing channels, but only about 30% convert into paying customers, resulting in a low lead-to-sale conversion rate.
- Sales Process: Once a lead fills out a form, the sales team follows up via calls and emails to try converting them, but most do not end up purchasing.
- Need for a Solution: To improve efficiency, the company wants to identify 'Hot Leads'—leads that are most likely to convert, allowing the sales team to focus their efforts better.
- Your Role: Build a logistic regression model that assigns a lead score (0–100) to each lead, representing the probability of conversion.
- Data Provided: A dataset of ~9000 leads with features like Lead Source, Time Spent on Website, etc., and a target column 'Converted' (1 or 0). You must clean the data, especially handling invalid entries like ‘Select’, which are treated as null values.

BUSINESS OBJECTIVE

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



SOLUTION METHODOLOGY

Data cleaning and data manipulation.

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

Exploratory Data Analysis (EDA)

1. Univariate data analysis: value count, distribution of variables, etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
3. Feature Scaling & Dummy variables and encoding of the data.
4. Classification technique: logistic regression is used for model making and prediction.
5. Validation of the model.
6. Model presentation.
7. Conclusions and recommendations.

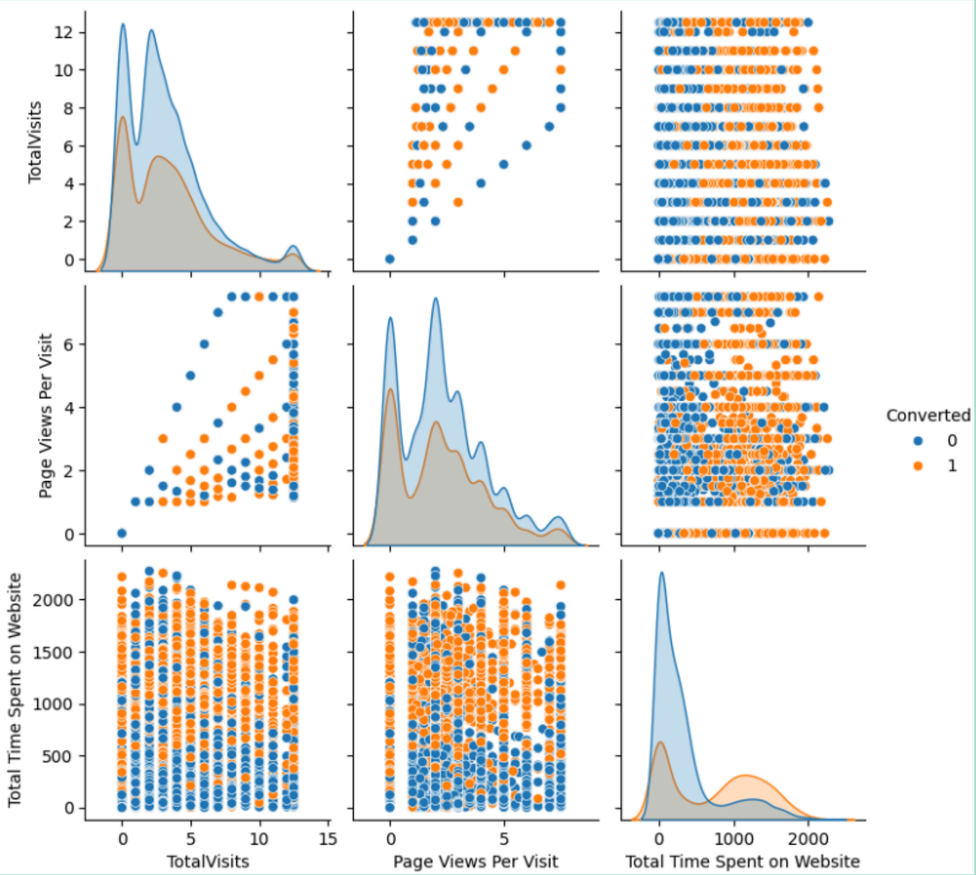
DATA MANIPULATION

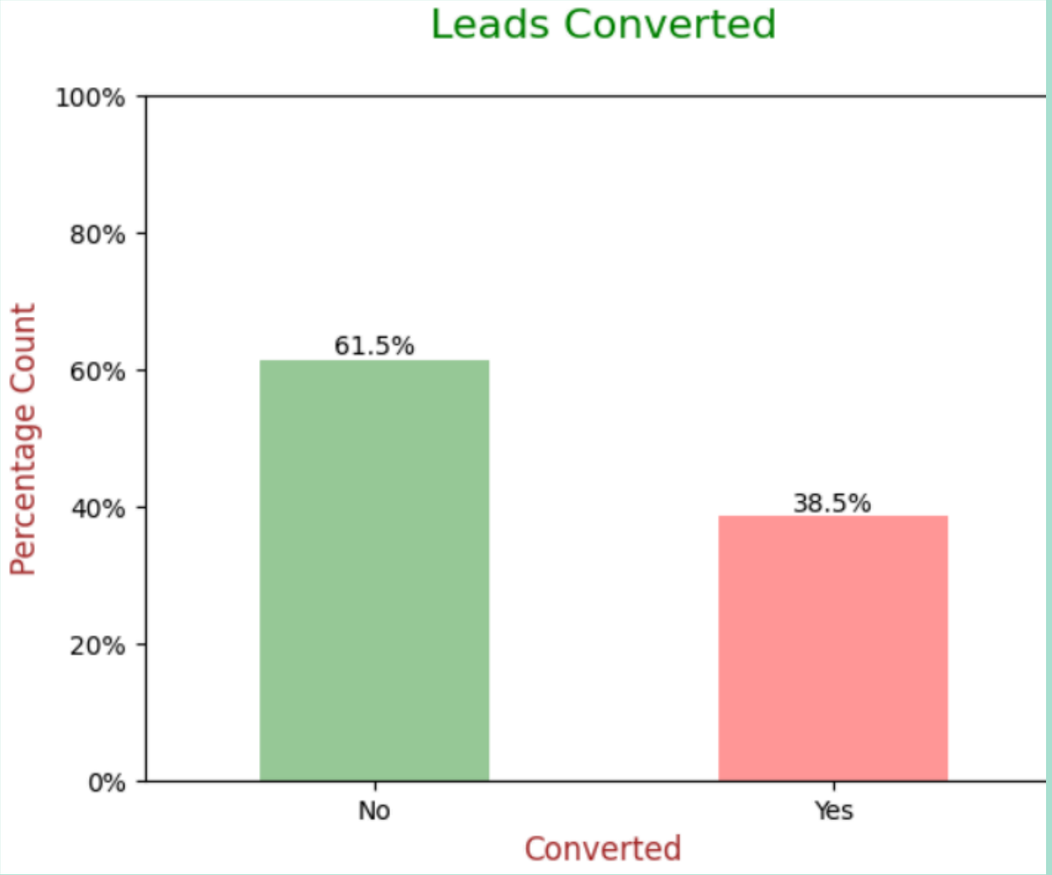
Additional Data Cleaning Insights

- The dataset initially contained 37 rows and 9,240 columns.
- Features with a single constant value such as “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” were dropped as they offered no predictive value.
- Identifiers like “Prospect ID” and “Lead Number” were also removed, as they do not contribute to the model's learning process.
- After analyzing value distributions in object-type variables, several features were found to have minimal variance and were dropped. These included:
- “Do Not Call”
- “What matters most to you in choosing a course”
- “Search”
- “Newspaper Article”
- “XEducation Forums”
- “Newspaper”
- “Digital Advertisement”
- Columns with over 35% missing values—such as “How did you hear about X Education” and “Lead Profile”—were also removed to maintain data quality and model reliability.

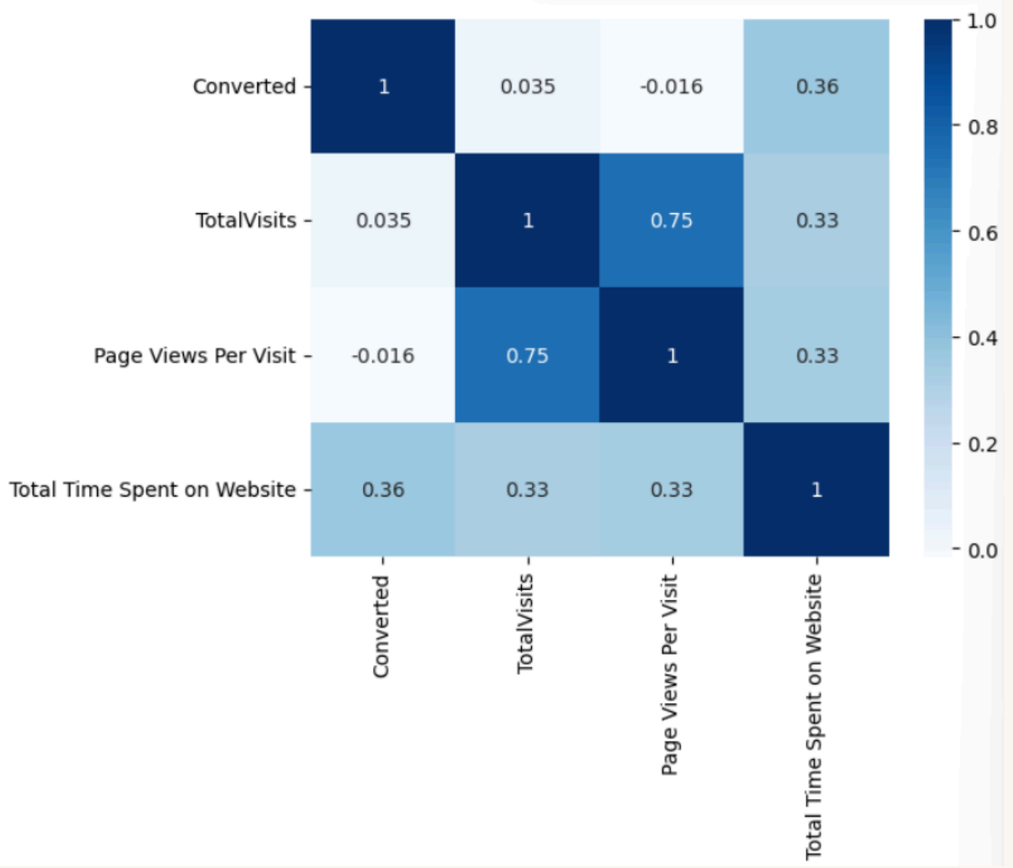
EXPLORATORYDATA ANALYSIS (EDA)

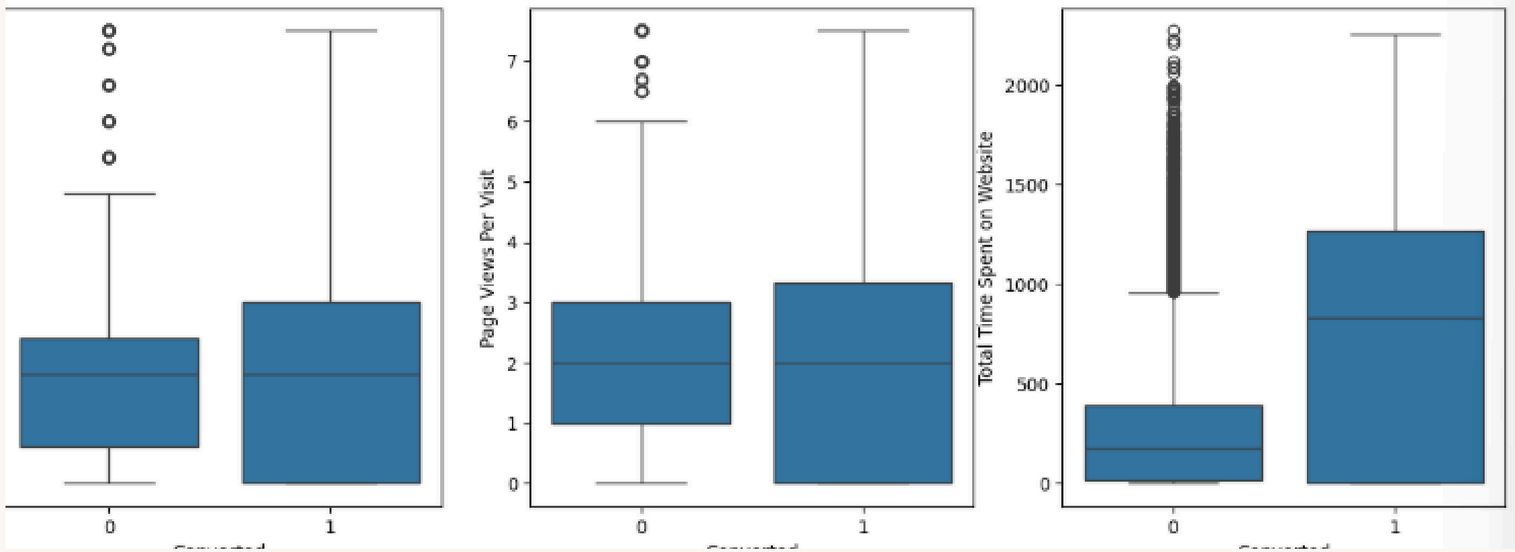
EDA revealed a significant class imbalance with only 38.5% leads converting. Key features like lead source, occupation, and website activity showed strong influence on conversion rates.





HEAT MAP

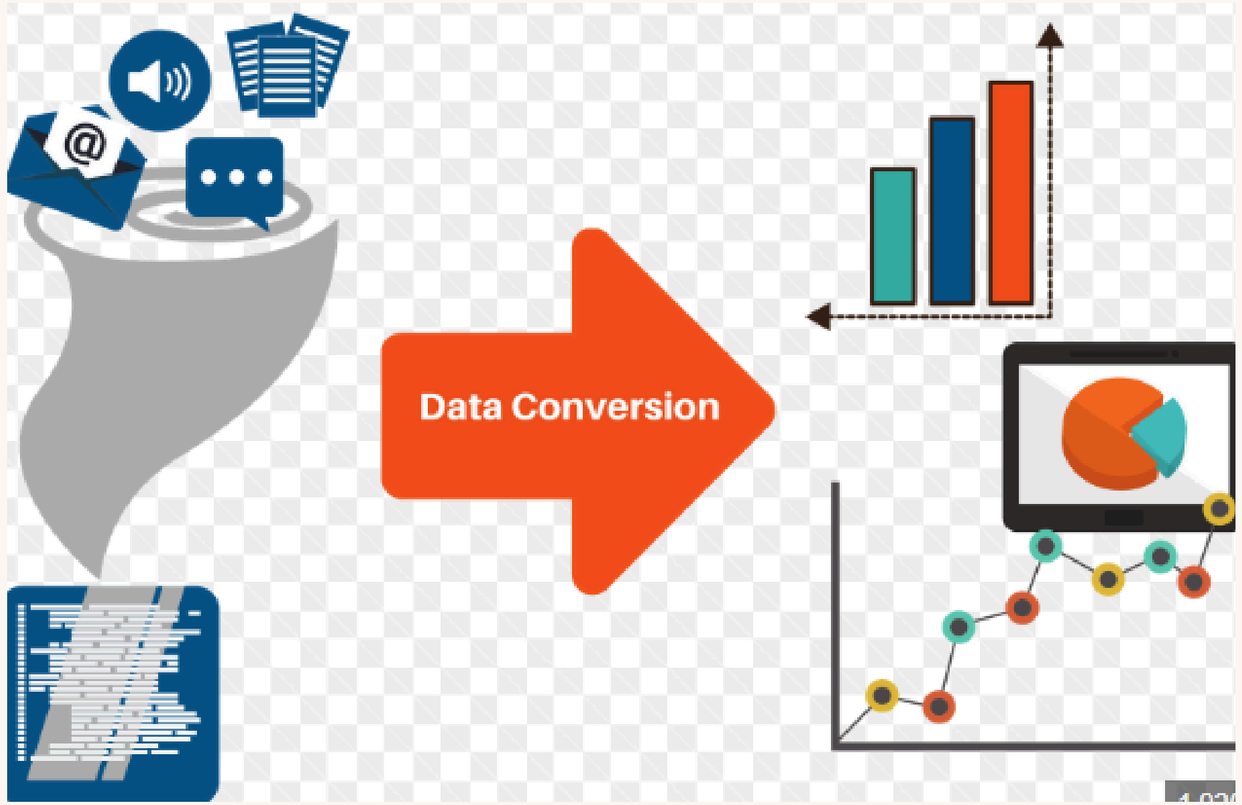




BOX PLOT

DATA CONVERSION

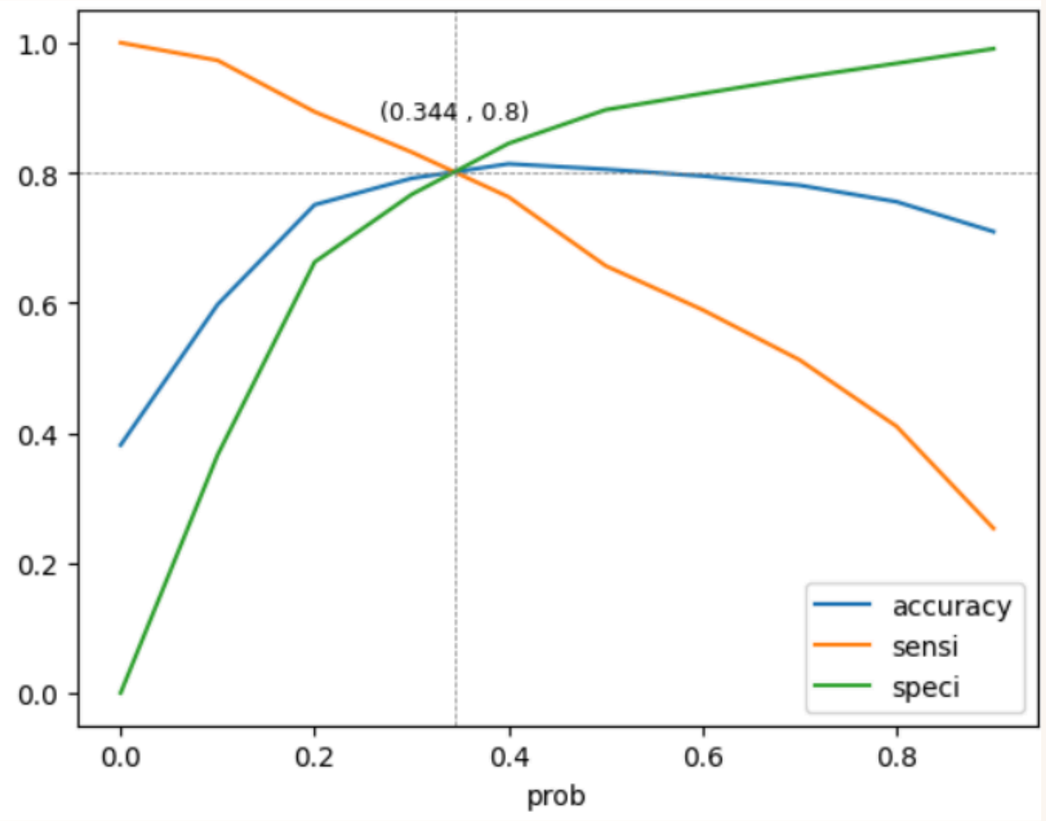
- The final dataset included 9,240 rows and 37 columns after cleaning and preprocessing.
- Numerical variables were normalized to bring them onto a common scale.
- Categorical (object-type) variables were converted into dummy variables using one-hot encoding.
- This transformation helped the model interpret and learn from non-numeric data effectively.
- The prepared dataset was then used for building and evaluating the lead scoring model.

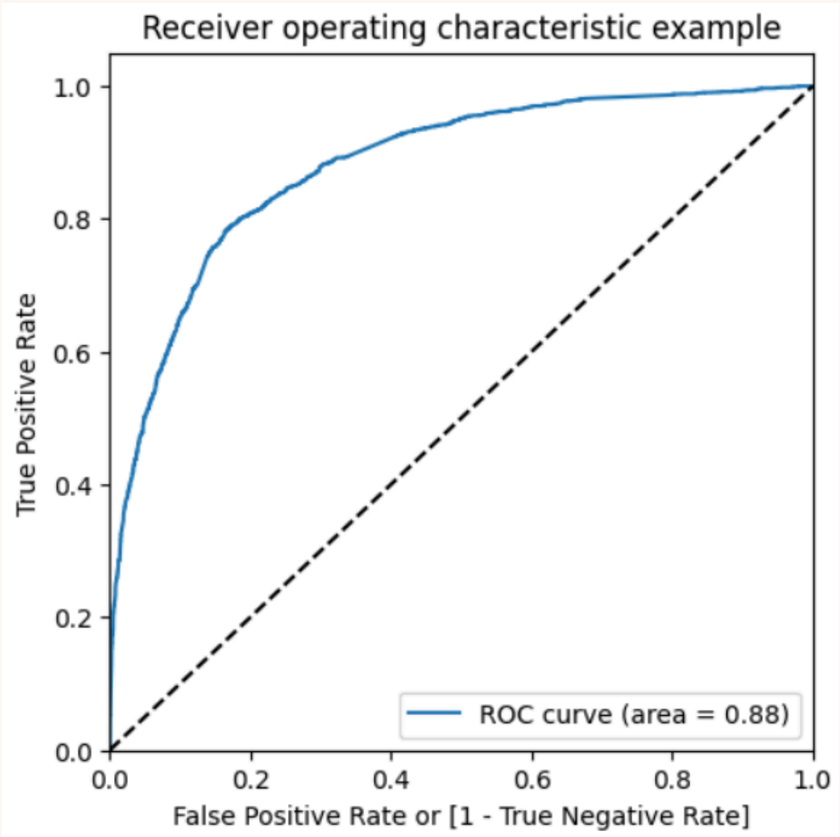


MODEL BUILDING

- The dataset was split into training and testing sets in a 70:30 ratio to ensure the model is trained on a majority portion while still being validated on unseen data.
- Recursive Feature Elimination (RFE) was performed to identify the top 15 most relevant features, simplifying the model without compromising performance.
- A manual feature reduction process followed, where variables with p-values greater than 0.05 and VIF scores above 5 were removed to retain only statistically significant, non-collinear predictors.
- The final regression model, built after several iterations, was both stable and interpretable, with 12 impactful variables retained.
- Predictions were made on the test dataset, and the performance metrics closely matched those of the training set, indicating good generalization.
- The model achieved an overall accuracy of 81%, with balanced sensitivity, specificity, and precision—meeting the business goal.
- The lead scoring system derived from the model helps prioritize leads with a higher likelihood of conversion, enabling better resource allocation.
- The approach not only boosts lead conversion efficiency but also aligns with the CEO's target of improving conversion rates to around 80%.

ROC Curve





- An optimal cut-off point was determined to balance sensitivity and specificity.
- The best cut-off probability was identified as 0.35.
- This value ensured improved prediction performance.
- It provided a balance between false positives and false negatives.
- The second graph confirmed 0.35 as the optimal threshold.

PREDICTION ON TEST SET

- Before predicting on the test set, the data was standardized to match the final training dataset's columns.
- Predictions were made on the test set, and the results were saved in a new dataframe.
- Model evaluation was performed by calculating accuracy, precision, and recall.
- The model achieved accuracy of 0.82, precision of 0.75, and recall of 0.75, indicating strong performance.
- These metrics confirm that the model is stable with acceptable levels of accuracy and recall.
- A lead score was assigned to the test dataset, where higher scores indicate a greater likelihood of conversion.
- The lead score helps prioritize high-potential leads, focusing efforts on those most likely to convert.
- By assigning scores, the model enables targeted marketing strategies and efficient resource allocation.
- Overall, the model’s performance and lead scoring system contribute to improving conversion rates and optimizing lead management.

CONCLUSION

- Key factors influencing conversions include total time spent on the website, number of visits, and lead sources like Google, Direct Traffic, Organic Search, and Welingak Website.
- Engagement activities such as SMS interactions and Olark chat conversations significantly increase conversion chances, showing that direct communication can build trust and drive decisions.
- Lead origin and current occupation (especially working professionals) also play a vital role in predicting conversions, indicating that professional leads are more likely to invest in courses.
- Focusing on these high-impact variables allows X Education to target the right audience, optimize its marketing spend, and enhance lead nurturing strategies.
- Additionally, understanding the factors that influence lead conversion helps prioritize efforts, ensuring that marketing resources are directed toward the most promising leads.
- With these strategies, X Education is well-positioned to achieve its 80% lead conversion goal, boosting course enrollments, improving resource efficiency, and driving long-term growth.

Thank you very much!

~ Presented by Papri Boyra