LEADS SCORING CASE STUDY

By

Saif Ali

PROBLEM STATEMENT

- X Education, an entity specializing in the provision of online courses to professionals within distinct industries, confronts a challenge characterized by an ostensible incongruity between the volume of leads it accrues and the ensuing inefficacious conversion rate. In elucidation, when X Education amasses, for instance, one hundred leads within a diurnal cycle, only a trifling quota, approximately thirty, eventuate in actualized conversions.
- To ameliorate the efficiency of this process, the corporation aspires to discriminate the leads exhibiting the highest proclivity for conversion, colloquially denominated as 'Hot Leads.' The corollary expectation is that by successfully segregating this subset of leads, the lead conversion rate will experience a discernible ascent. Such a phenomenon can be ascribed to the sales cadre's redirection of efforts, directing their communication endeavors towards these deemed potential leads, in contradistinction to disseminating their attention uniformly across the entire lead pool.

BUSINESS OBJECTIVE

- X Education aspires to ascertain the preeminent leads in their purview, therein prompting an endeavor to formulate a predictive model with the aptitude to discern 'Hot Leads.' Subsequent to the model's construction, there ensues an intention to effectuate its deployment, facilitating its utilization for forthcoming instances.

SOLUTION METHODOLOGY

Data cleaning and data manipulation

- Scrutinize and manage duplicate data entries.
- Examine and address the presence of NA (Not Available) values and gaps within the dataset.
- Eliminate columns replete with a substantial count of missing values that bear no significance for the analysis.
- Undertake imputation of missing values when deemed requisite.
- Identify and rectify outliers present within the dataset.

Exploratory Data Analysis (EDA)

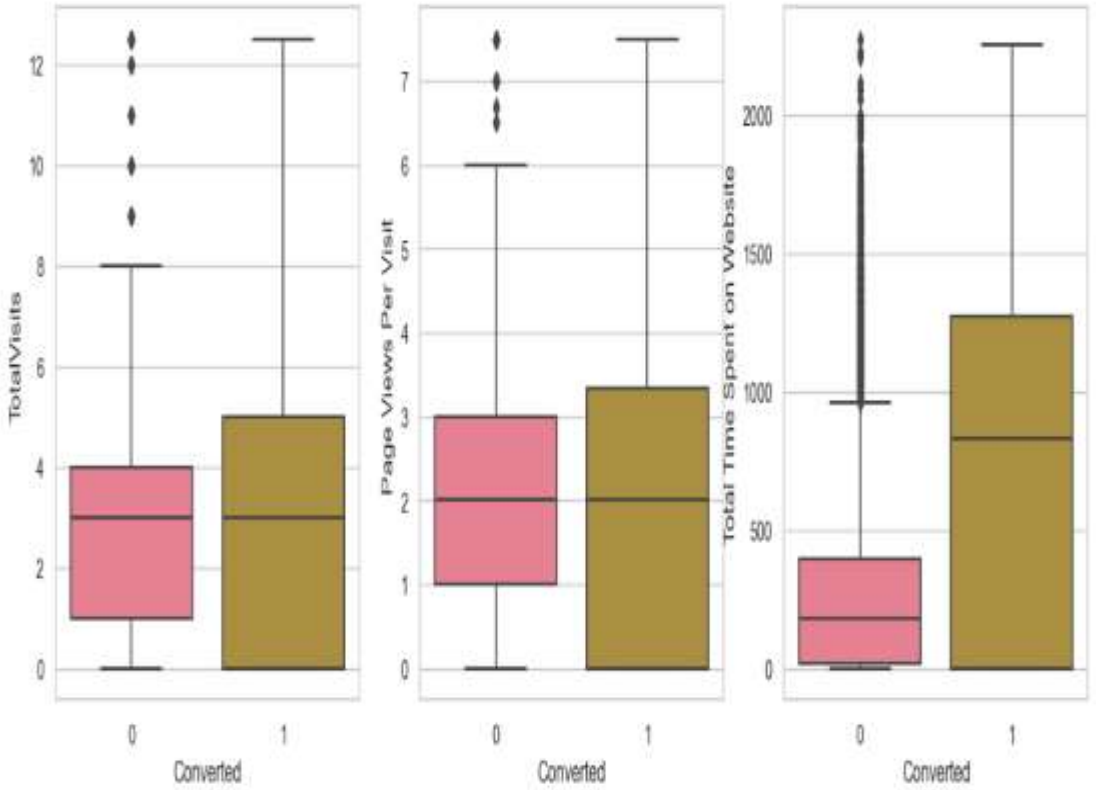
- Perform univariate data analysis, encompassing value counts and variable distributions.
- Execute bivariate data analysis, scrutinizing correlation coefficients and inter-variable patterns.
- Apply feature scaling, employ dummy variables, and conduct data encoding.
- Employ the logistic regression algorithm for model construction and predictive analytics.
- Validate the model's performance.
- Present the model.
- Formulate conclusions and proffer recommendations.

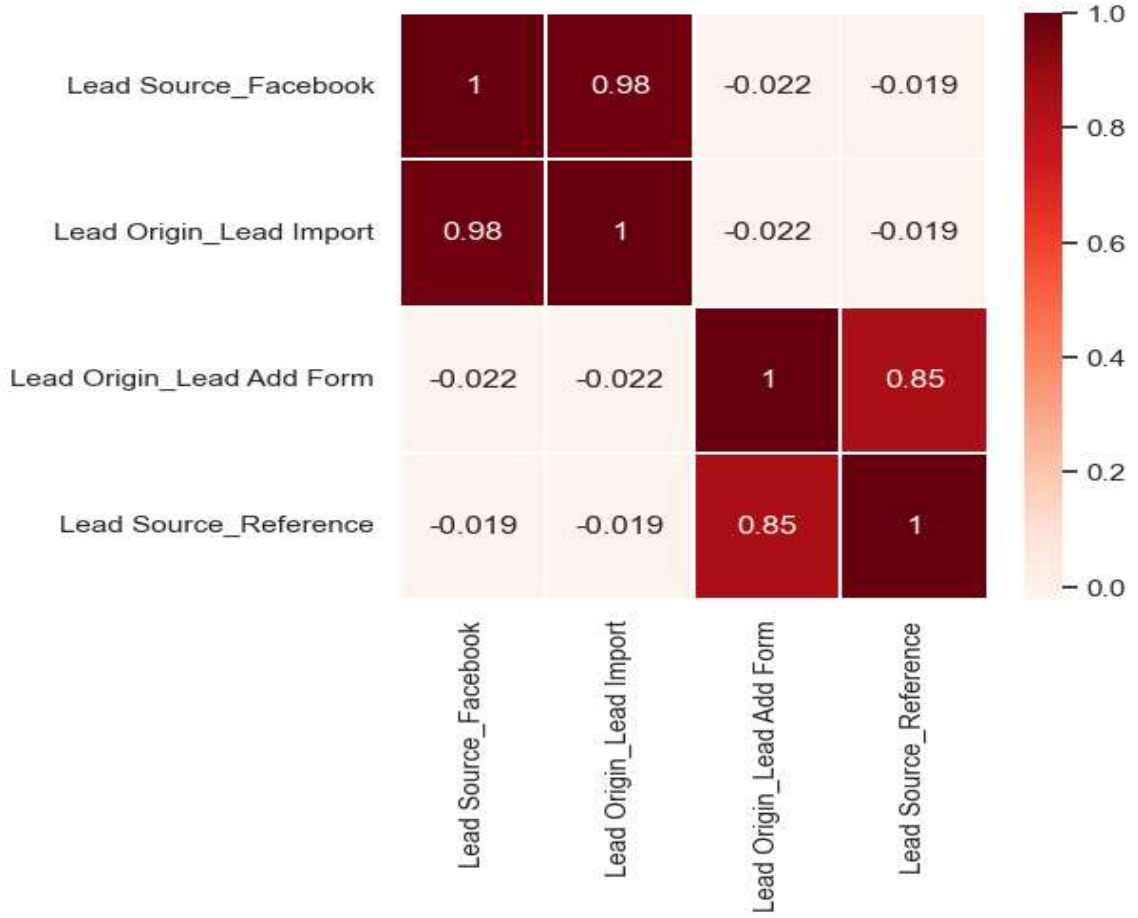
DATA MANIPULATION

- Initial dataset comprises 37 rows and 9240 columns.
- Features with singular values 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” were removed.
- "ProspectID" and "Lead Number" were excluded, deemed non-essential for analysis.
- Certain object type variables, e.g., "Do Not Call," "What matters most to you in choosing course," "Search," "Newspaper, Article," "XEducation Forums," "Newspaper," "Digital Advertisement," etc., displaying limited variance were dropped.
- Columns with over 35% missing values, such as ‘How did you hear about X Education’ and ‘Lead Profile,’ were removed.

EXPLORATORYDATA ANALYSIS (EDA)

BOX PLOT & HEAT MAP



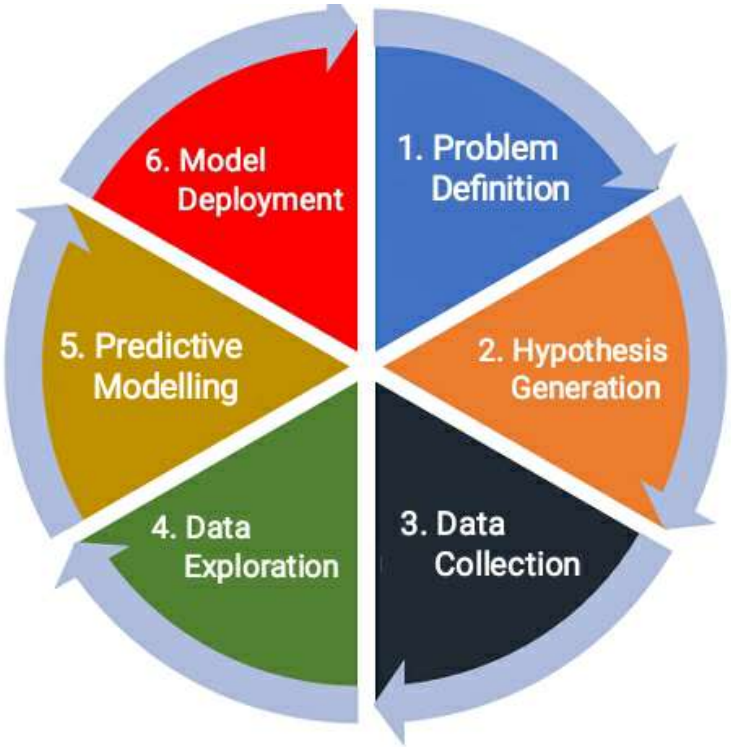


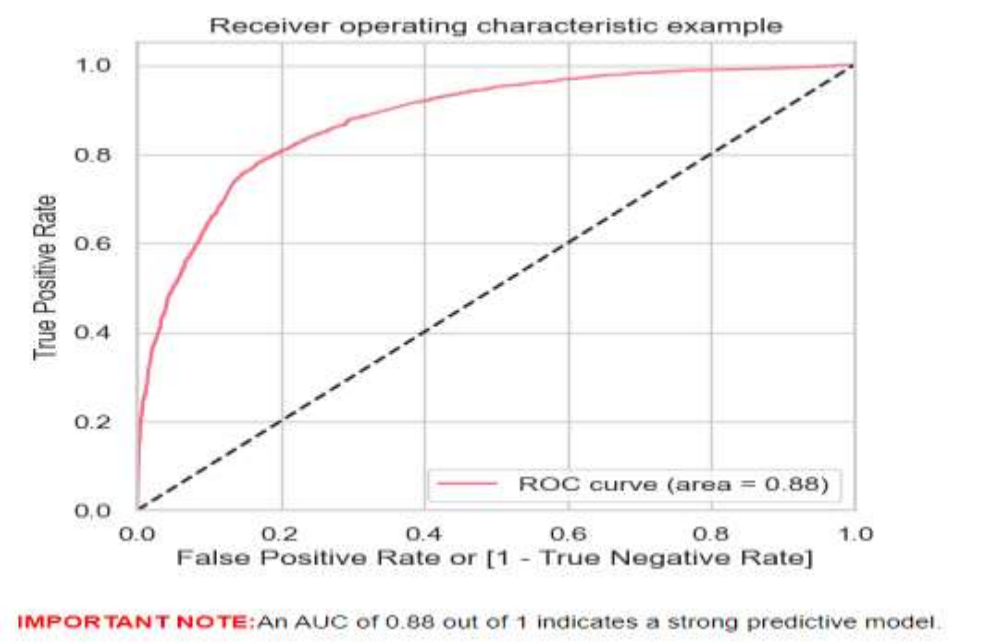
DATA CONVERSION

- Numerical variables have been normalized.
- Object type variables have been converted into dummy variables.
- Total rows for analysis: 9240
- Total columns for analysis: 37

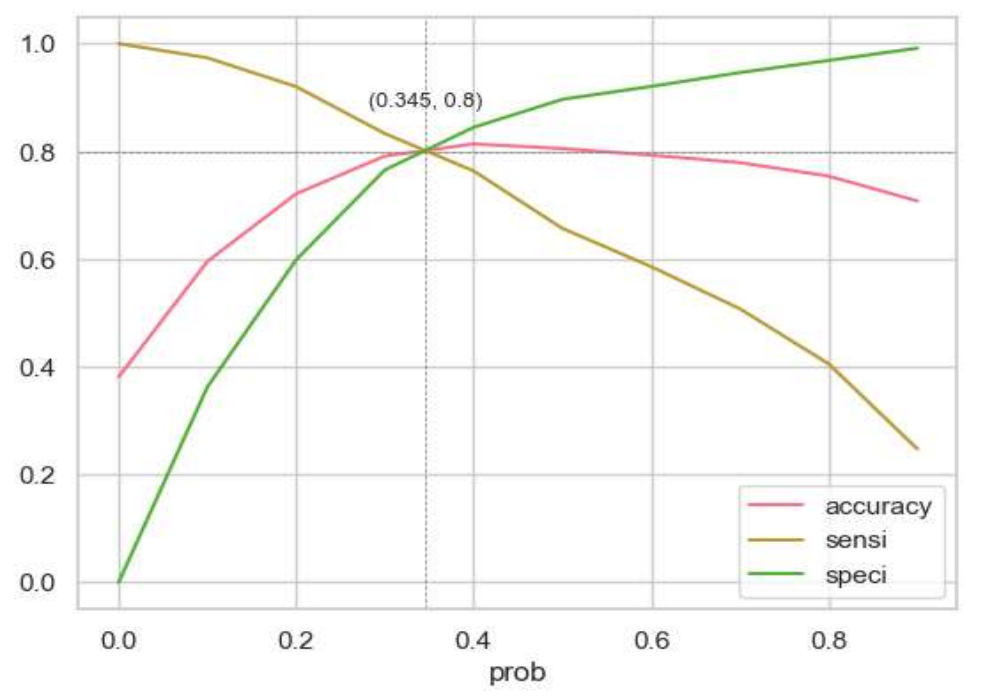
MODEL BUILDING

- The dataset has been divided into training and testing subsets.
- A 70:30 train-test split ratio has been employed.
- Recursive Feature Elimination (RFE) has been used for feature selection.
- RFE was executed with the objective of selecting 15 variables as output.
- The model has been constructed by eliminating variables with p-values exceeding 0.05 and variable importance (VI) values surpassing 5.
- Predictions have been generated on the test dataset.
- The model exhibits an overall accuracy of 81%.





IMPORTANT NOTE: An AUC of 0.88 out of 1 indicates a strong predictive model.



ROC CURVE

- Determining the Optimal Cut-off Point.
- The Optimal Cut-off Probability signifies the point where a balance between sensitivity and specificity is achieved.
- The second graph clearly indicates that the optimal cut-off point lies at 0.35.

PREDICTION ON TEST SET

- Before making predictions on the test set, standardization was applied, ensuring the presence of identical columns as the final training dataset.
- Predictions on the test set were executed, and the resultant predictions were stored in a new data frame.
- Subsequently, model evaluation was performed, encompassing accuracy, precision, and recall.
- The obtained scores were approximately 0.82 for accuracy, 0.75 for precision, and 0.75 for recall.
- These results indicate that the test predictions fall within an acceptable range.
- Furthermore, the stability and robustness of the model are evident, given the commendable accuracy and recall/sensitivity.
- A lead score was generated for the test dataset, which serves to distinguish hot leads, with a higher score indicating an increased likelihood of conversion and a lower score indicating a reduced chance of conversion.

CONCLUSION

The variables that exhibit the most significant influence on potential buyers, listed in descending order of importance, are as follows:

- The cumulative time spent on the website.
- The total number of visits.
- The lead source categories of Google, Direct traffic, Organic search, and Welingak website.
- The last activity types of SMS and Olark chat conversation.
- The lead origin categorized as Lead add format.
- Individuals with a current occupation as working professionals.

Considering these pivotal factors, X Education possesses a substantial probability of persuading nearly all prospective buyers to reconsider and make purchases of their courses, thus potentially enhancing their flourishing prospects.

THANK YOU!