Lead Scoring Case Study using Logical Regression

Submitted by:

SUBHAM PAUL

PARTHIBAN SETHURAMAN

SAURABH SINGH

PROBLEM STATEMENT

▶ X Education sells online courses to industry professionals.

▶ X Education gets a lot of leads, but its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

▶ To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

▶ If they successfully identify these set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE

▶ X education wants us to build a model to give every lead a lead score between 0 –100. So that they can identify the Hot leads and increase their conversion rate as well.

▶ The CEO wants to achieve a lead conversion rate of 80%.

▶ They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

SOLUTION METHODOLOGY

-Data cleaning and data manipulation

-> Check and handle duplicate data.

-> Check and handle NA values and missing values.

-> Drop columns, if it contains a large number of missing values and are not useful for the analysis.

-> Imputation of the values, if necessary.

-> Check and handle outliers in data.

SOLUTION METHODOLOGY (contd.)

-Exploratory Data Analysis (EDA)

-> Univariate data analysis: value count, distribution of variables, etc.

-> Bivariate data analysis: correlation coefficients and pattern between the variables etc.

-> Feature Scaling & Dummy variables and encoding of the data.

-> Classification technique: logistic regression is used for model making and prediction.
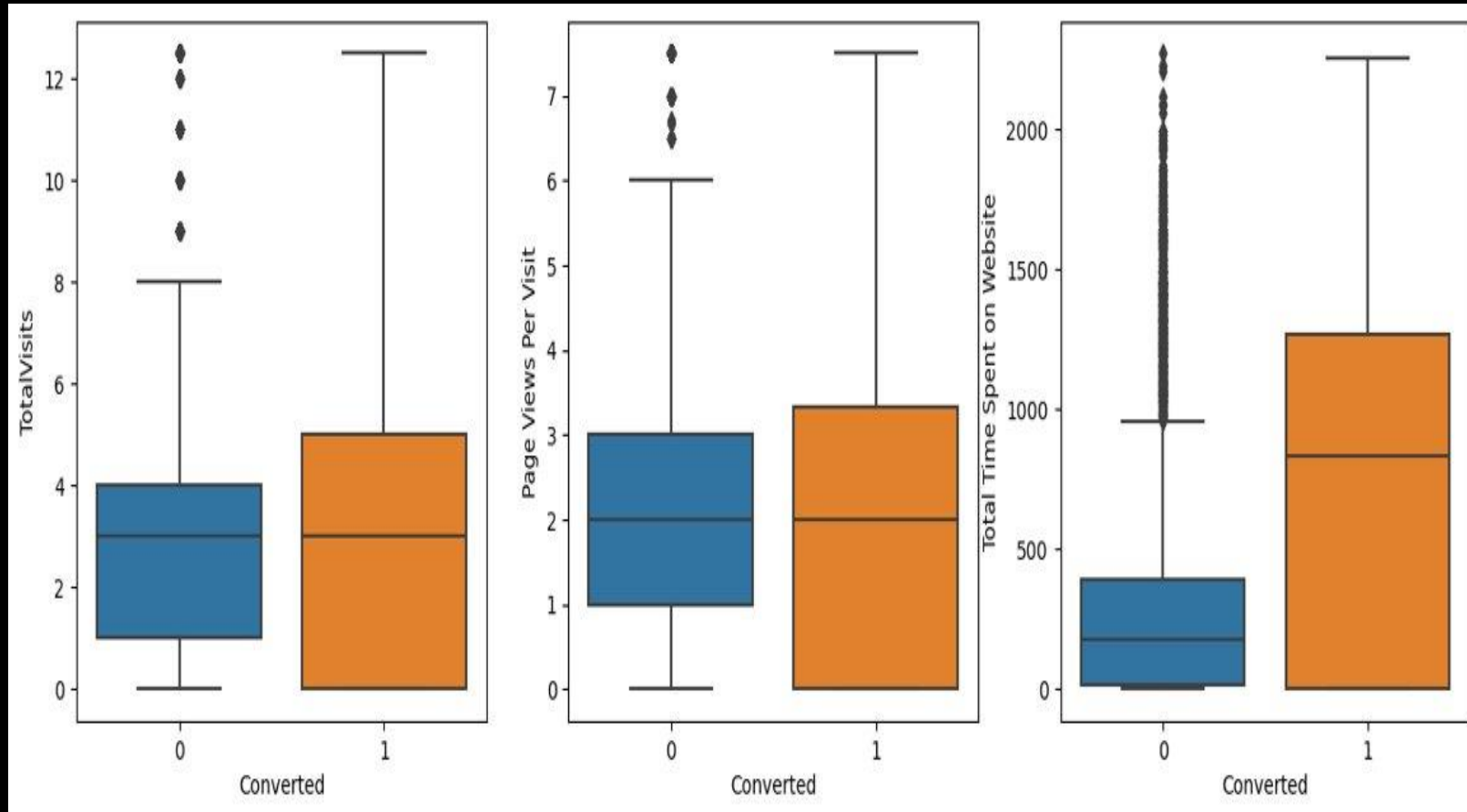
-> Validation of the model.

EXPLORATORY DATA ANALYSIS (EDA)
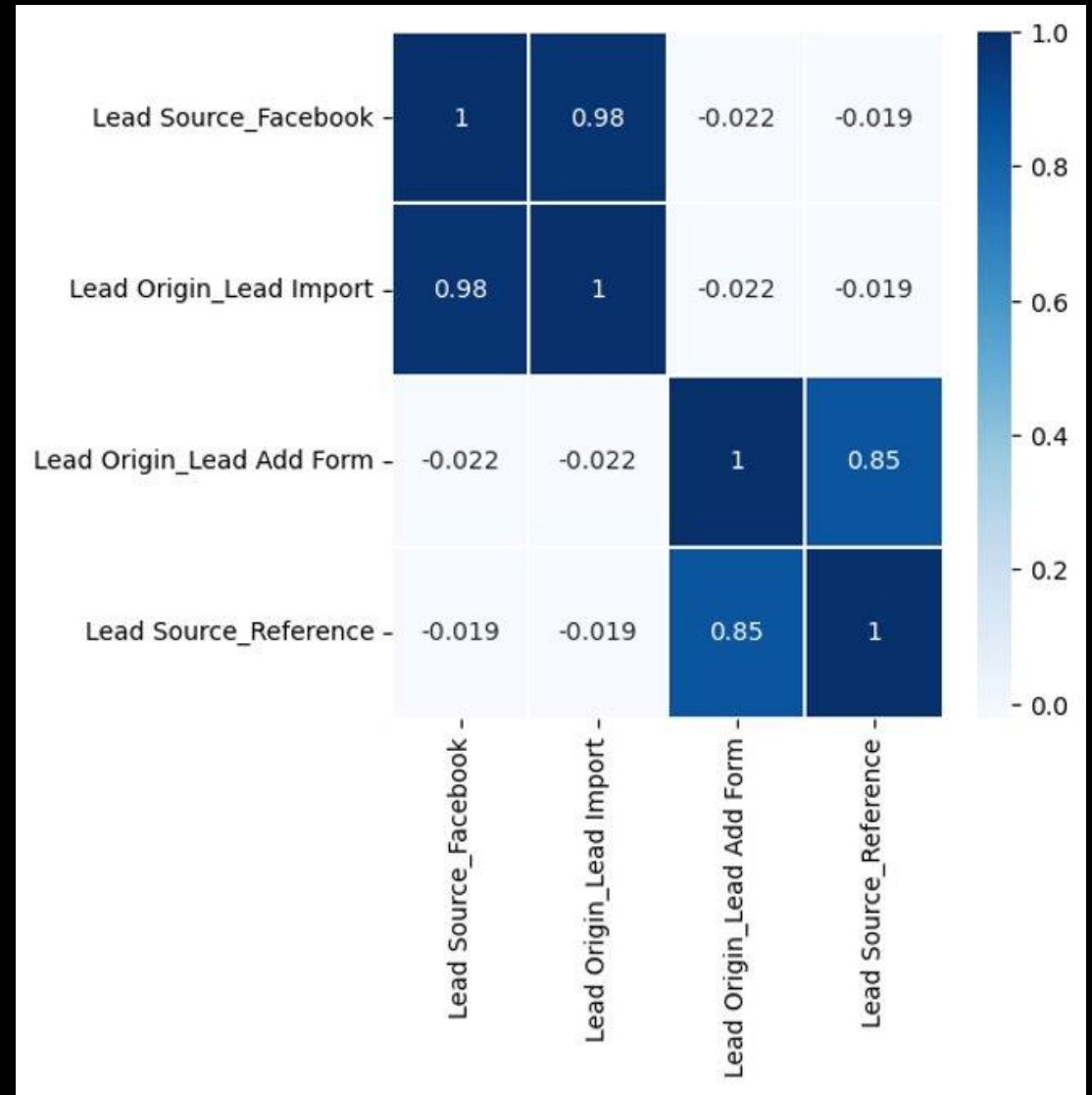
-Bivariate Analysis for Numerical Variables
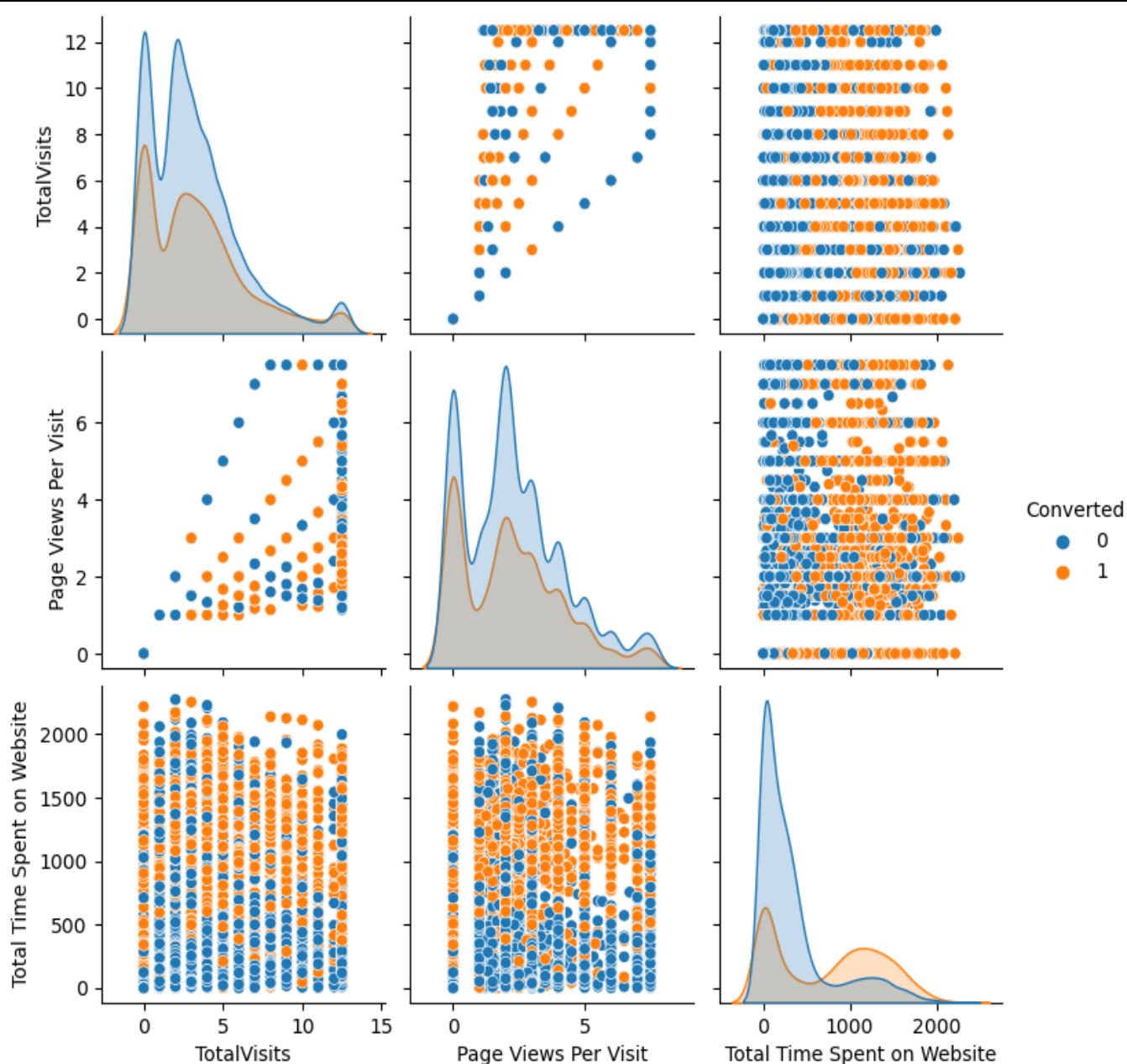
BOX PLOT

-Past Leads who spends more time on Website are successfully converted than those who spends less as seen in the boxplot

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

Receiver operating characteristic example

► Finding Optimal Cut off Point

► Probability where we get balanced sensitivity and specificity.

► From the second graph, it is visible that the optimal cut off is at 0.35.
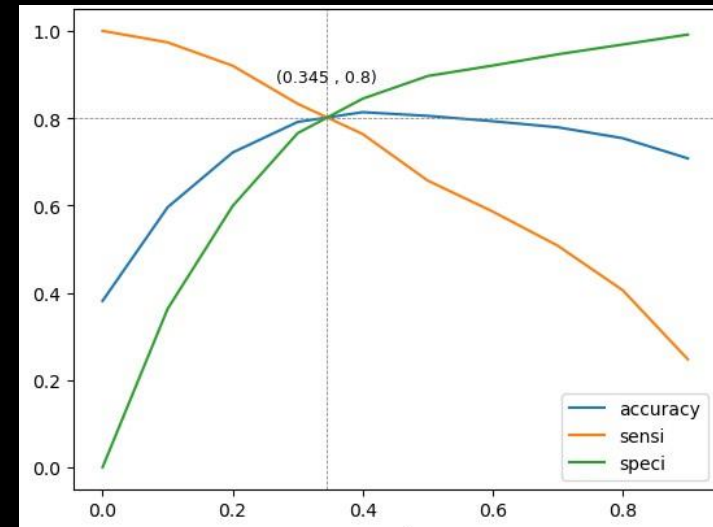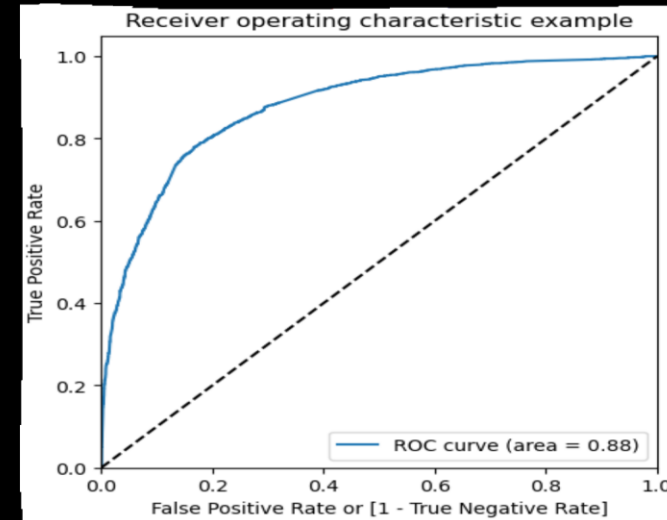
PREDICTION ON TEST SET

▶ 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 is the chance of conversion, low the lead score lower is the chance of getting converted.

CONCLUSION

▶ We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.

▶ Leads who spent more time on website, more likely to convert.

▶ Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with 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.

THANK YOU