LEAD SCORING CASE STUDY

DONG NGUYEN

OUTLINE

1. Problem statement
2. Objective
3. Strategy
4. Conclusion and Recommendations

PROBLEM STATEMENT

- Customers can purchase online courses from X Education, a company that provides education.
- X Education receives a lot of leads, but it has an extremely low lead conversion rate.
- The business wants to identify the most promising leads, often known as "Hot Leads," in order to increase the efficiency of this process.

OBJECTIVE

- Create a model in which the company give each lead a lead score so that leads with higher lead scores have a better chance of converting, while leads with lower lead scores have a lesser chance of converting.
- Particularly the CEO has provided a rough estimate of the desired lead conversion rate.

Strategy

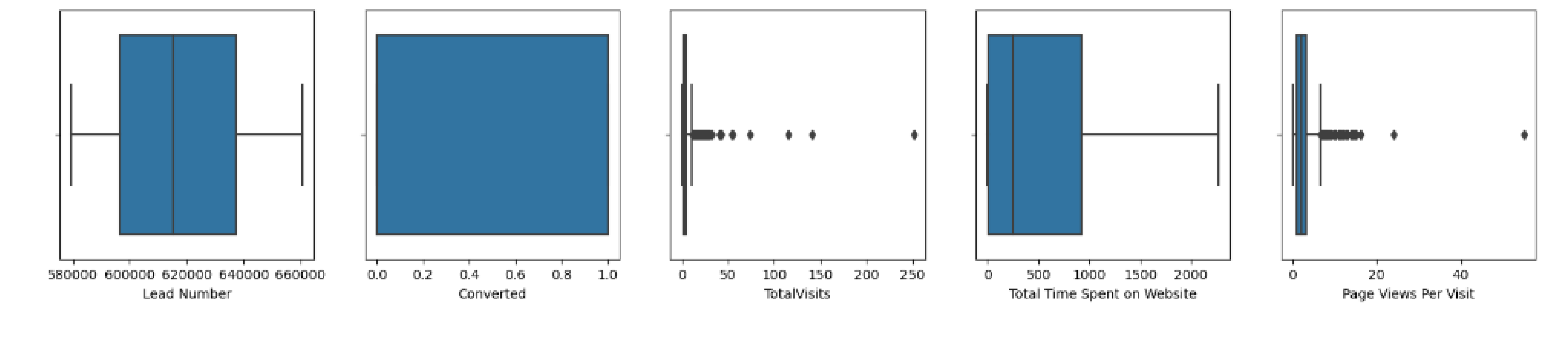
1. Data cleaning
2. Exploratory Data Analysis
3. Data Preparation
4. Model Building
5. Model Evaluation

DATA CLEANING

Treat ‘Select’ as missing values since Select means customers did not choose any options.

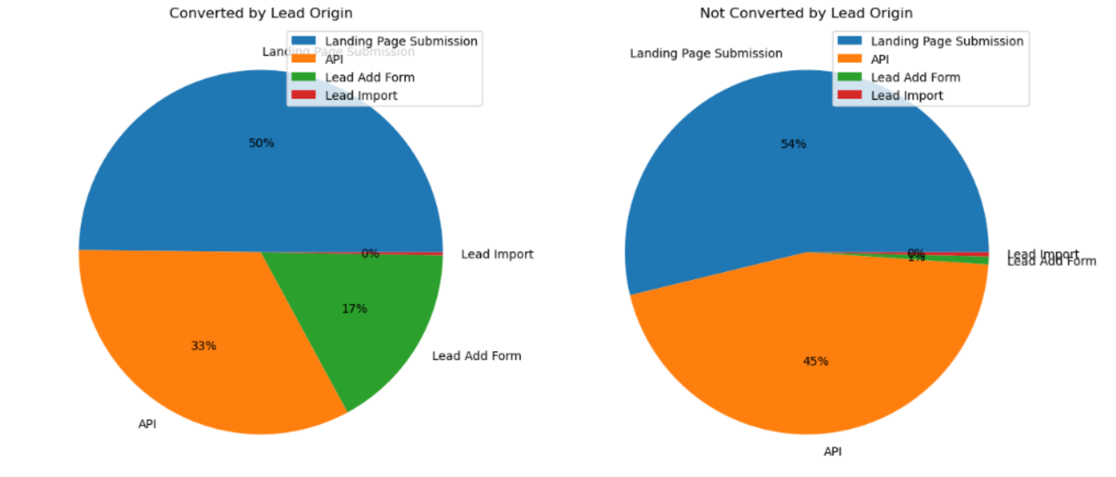
Delete the columns that have more than 40% of missing values. For the columns that have less than 40% of missing values, impute the missing values with mode. For the columns have less than 5% of missing values, drop the missing values.

DATA CLEANING

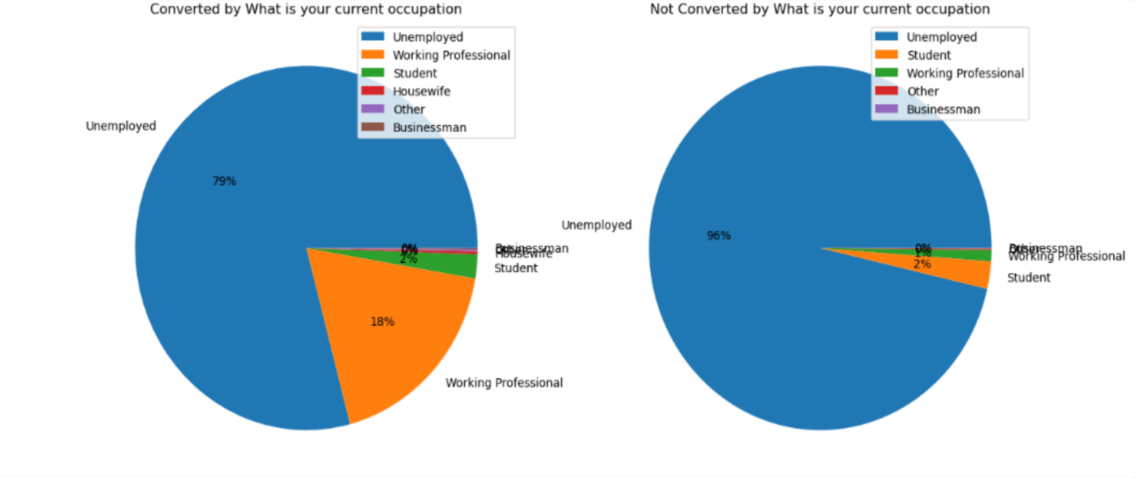


Detect outliers for the continuous variables and replace them with NULL value. Then drop them since these NULL values are less than 3%.

EXPLORATORY DATA ANALYSIS


Converted by Lead Origin

Not Converted by Lead Origin


Converted by What is your current occupation

Not Converted by What is your current occupation

The better a lead has been converted, the higher the proportion of Lead Add Form and the lower the proportion of API.

The percentage of working Professional is higher when a lead has been converted.

DATA PREPARATION

Converting some binary variables (yes/no) to 0/1

Create dummy features (one-hot encoded) for categorical variables with multiple levels.

Drop the repeated variables

Split the data into train set and test set in the ratio of 70:30.

Feature scaling for continuous variables using a standard scaler.

MODEL BUILDING

Build the model and select features using RFE.

Drop the features with P-values higher than 0.5

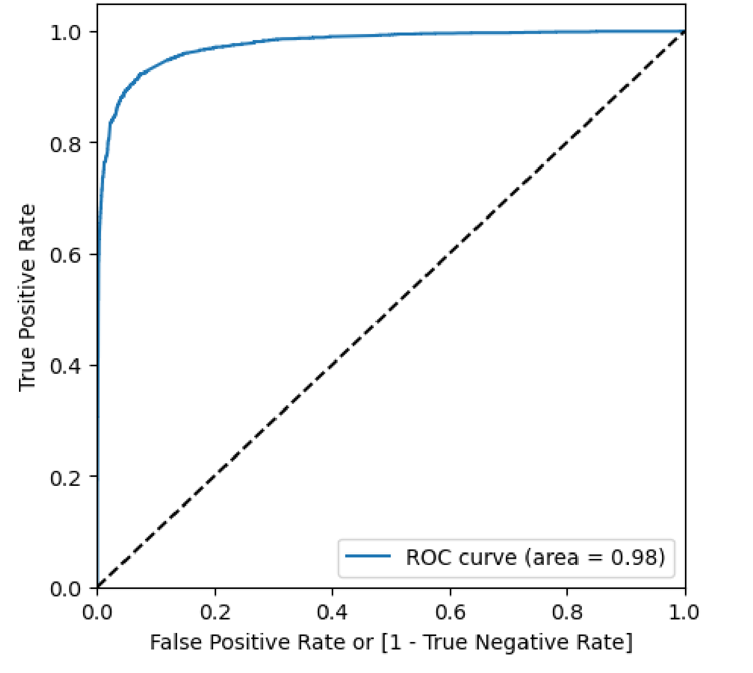
Check the VIFs score for the features and drop them one by one until all the VIFs are below 5.

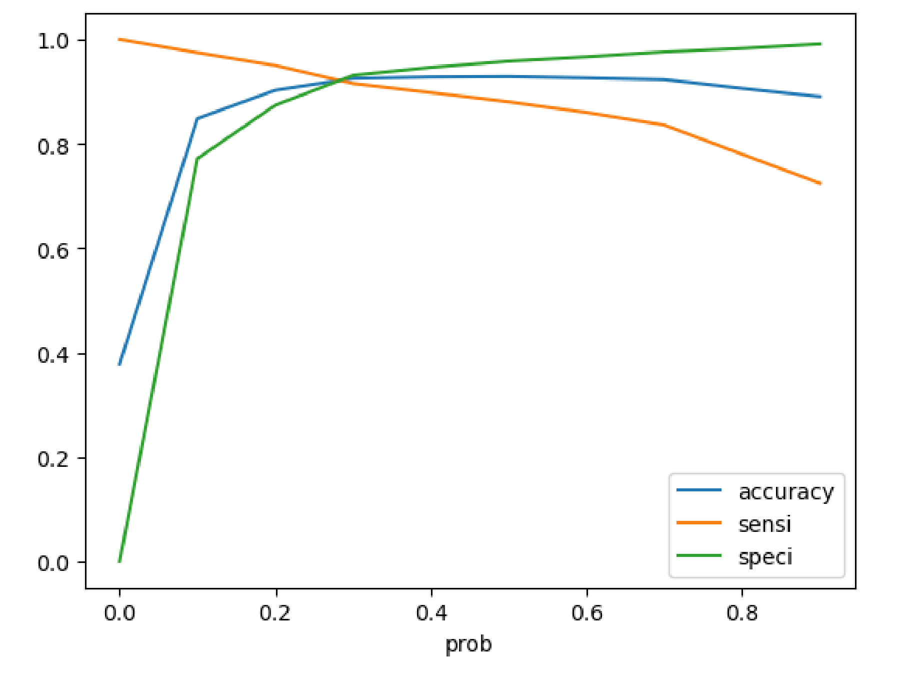
Predict the train set.

Make the predictions on the test set.

Comparing the values obtained for the Train set & Test set

MODEL BUILDING





The area under the ROC curve = 0.98, hence our model is a good one.

From the curve above, 0.30 is the optimum point to take it as a cutoff probability.

MODEL EVALUATION

Train set: Accuracy: 92.4 % Sensitivity: 91.5 % Specificity: 93.0 %

Test set: Accuracy: 92.8 % Sensitivity: 90.8 % Specificity: 94.0 %

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. We have succeeded in achieving our objective of estimating the desired lead conversion rate to be around 92%. We should be able to provide the CEO with the confidence to make decisions based on this model.

RECOMMENDATION

Since customers are more likely to convert, the X Education company should call the leads with a Tag of ‘Closed by Horizzon’, ‘Lost to EINS’, ‘Will revert after reading the email’, ‘Want to take admission but has financial problems’.

The company should make calls to the leads coming from the lead sources "Welingak Websites“ and whose Last Activity is ‘Had a Phone Conversation’ as these are more likely to get converted.

The cut-off point in this model is the optimal point to get maximum sensitivity and specificity. Depending on the business situation, X Education can adjust the cut-off point to achieve their target.