Lead Scoring

Problem Description

- X Education is an education company that offers online courses to industry professionals
- The company advertises its courses on various websites and search engines, including Google
- X Education wants to improve its lead conversion rate by identifying the most promising leads - those who are most likely to become paying customers

Approach

Based on the problem description, it has been determined that the problem is a classification problem. As a result, logistic regression was chosen as the method for calculating the lead rate. The following steps were taken to solve the problem Data Reading and Understanding

- Number of rows and columns
- Data types of each columns
- Checking first few rows how data looks
- Checking how the data is spread
- Checking for duplicates, if any

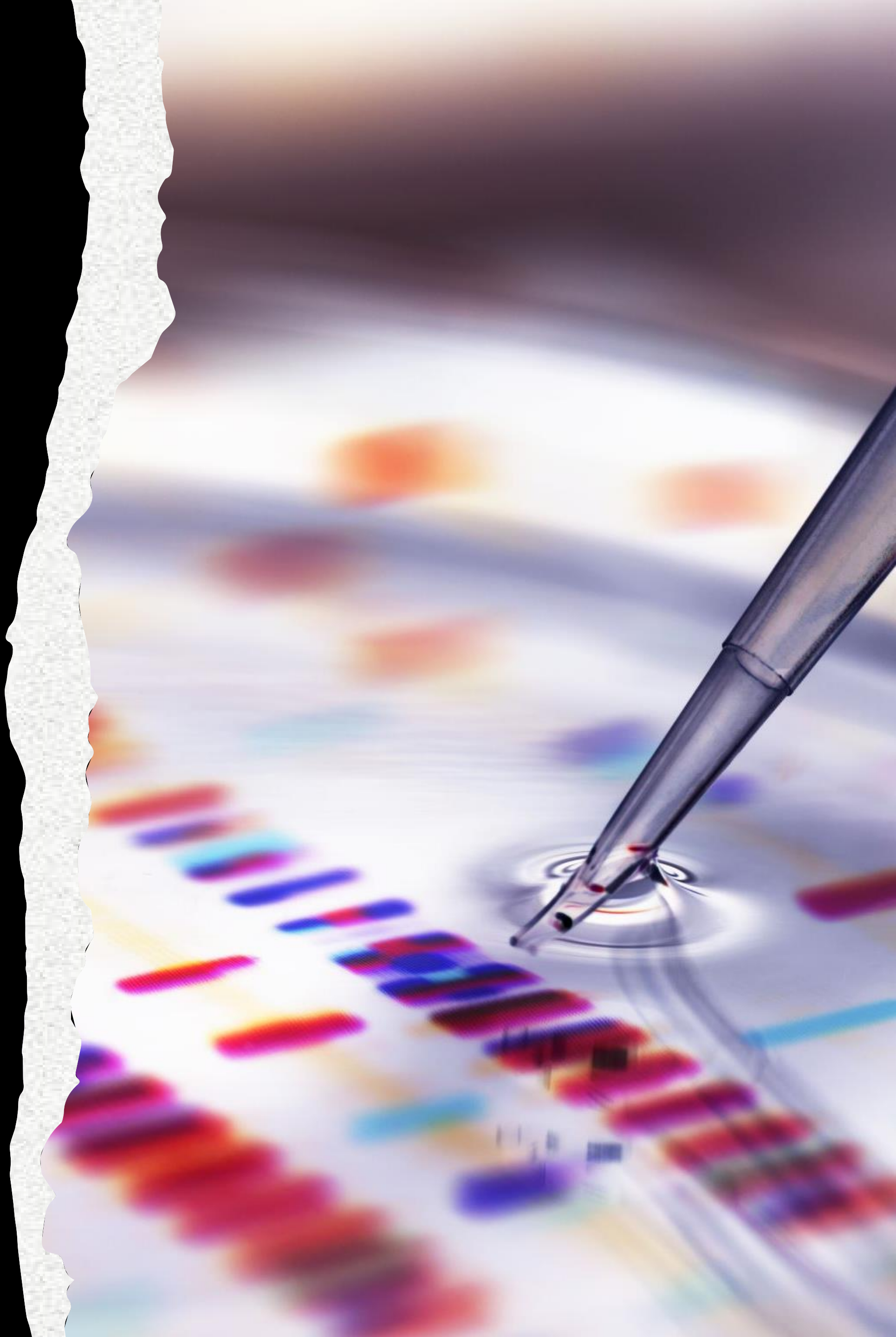
Data Cleaning

- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods
  - We used mode imputation for categorical columns
  - We used mean imputation for numerical columns, if there is no skewness in data
  - We used median imputation for numerical columns, if there is skewness in the data

Data Visualization and Outliers Treatment

- We conducted a univariate analysis on categorical columns to determine which columns were most relevant and removed those with near-zero variance
- We also performed a bivariate analysis on categorical columns to see how they varied with respect to the Converted column
- We used the IQR method to treat outliers in the dataset and plotted a correlation matrix to identify correlated columns

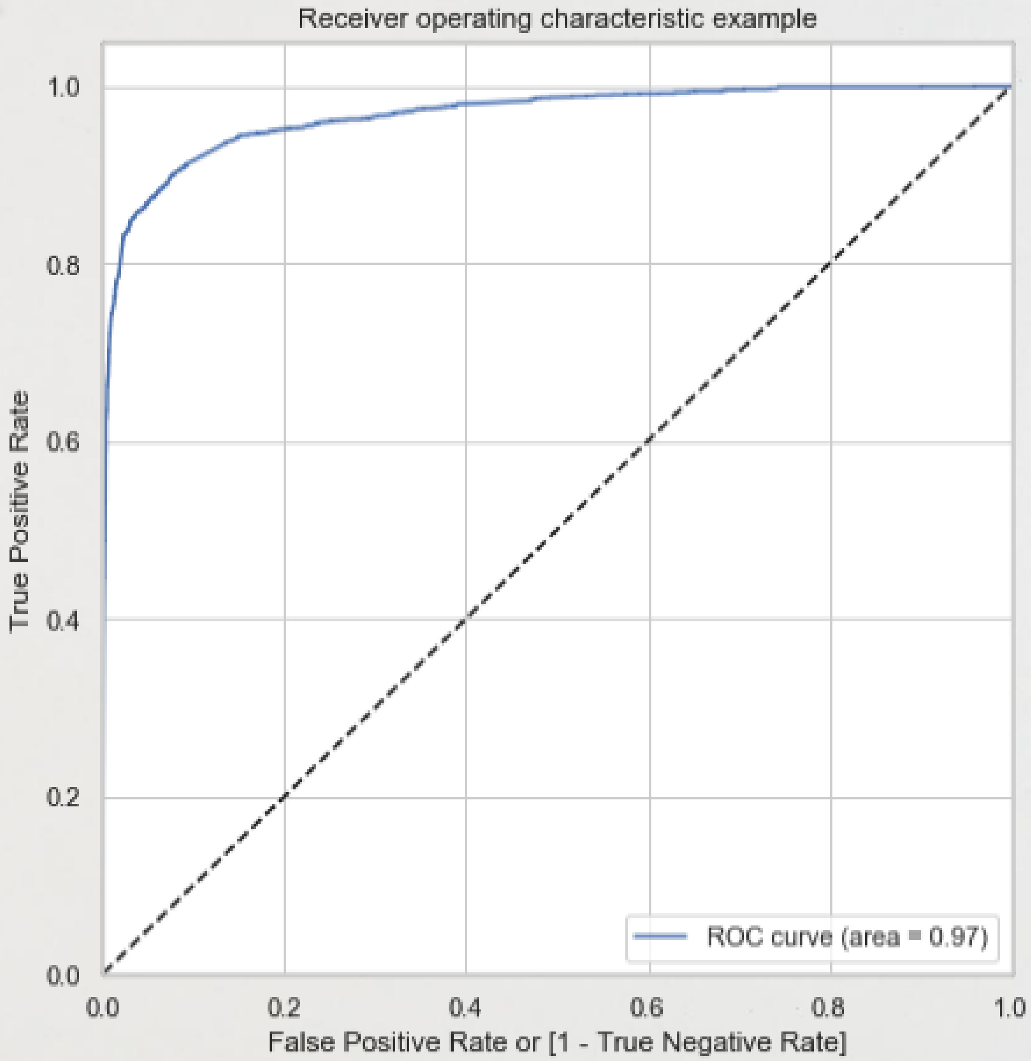
Feature Scaling

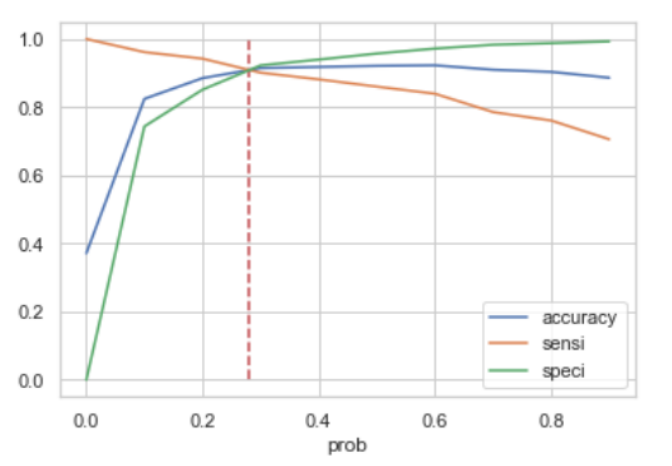


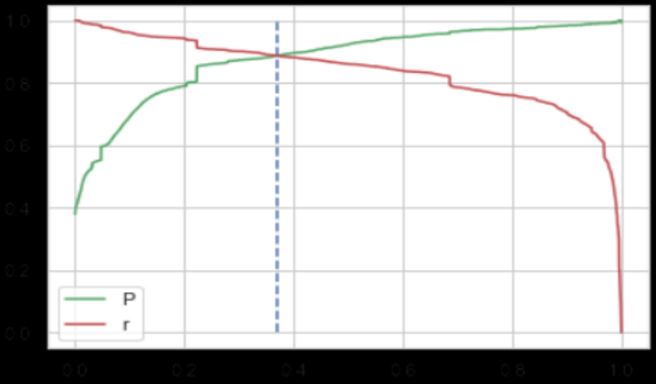
At this point, our data was clean and free of outliers. Since logistic regression requires numerical input parameters, we converted all categorical columns to numerical values. Columns with only two levels, “Yes” and “No,” were converted to numerical values using binary mapping. Columns with more than two levels were converted to dummy variables using the pd.get_dummies function. The resulting data contained only numerical columns and dummy variables. Before building the model, we rescaled all numerical columns using the standard scaler method

Model Building

- We utilized the Recursive Feature Elimination technique to remove attributes and build a model on the remaining attributes
- RFE uses model accuracy to determine which attributes and combinations of attributes are most predictive of the target attribute
- After creating a stable model, we predicted probabilities on the train set and created a new “predicted” column with a value of 1 if the probability was greater than 0.5, and 0 otherwise







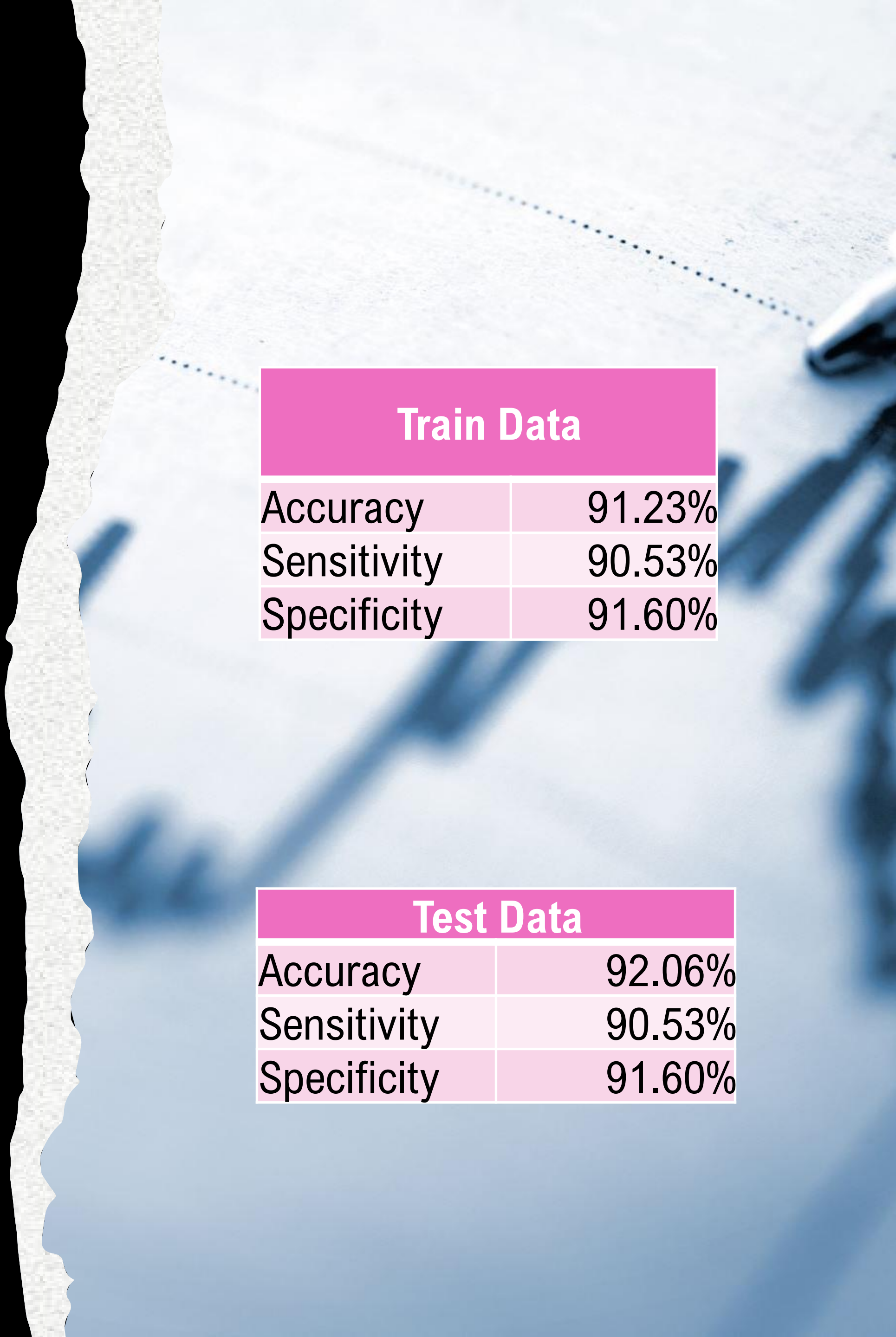
Model Evaluation on Train Set

- We calculated three metrics - accuracy, sensitivity, and specificity - for probabilities ranging from 0.0 to 0
- To make predictions on the train dataset, we determined the optimum cutoff to be 0.28 by finding the intersection of sensitivity, specificity, and accuracy
- For the test dataset, we considered the optimum cutoff to be obtained from the precision-recall graph of the train dataset

Predictions on Test Set

After determining the optimum cutoff and calculating the metrics on the train set, we made predictions on the test dataset

Here are our observations: It seems that the model performed well on both the train and test datasets, with high accuracy, sensitivity, and specificity values

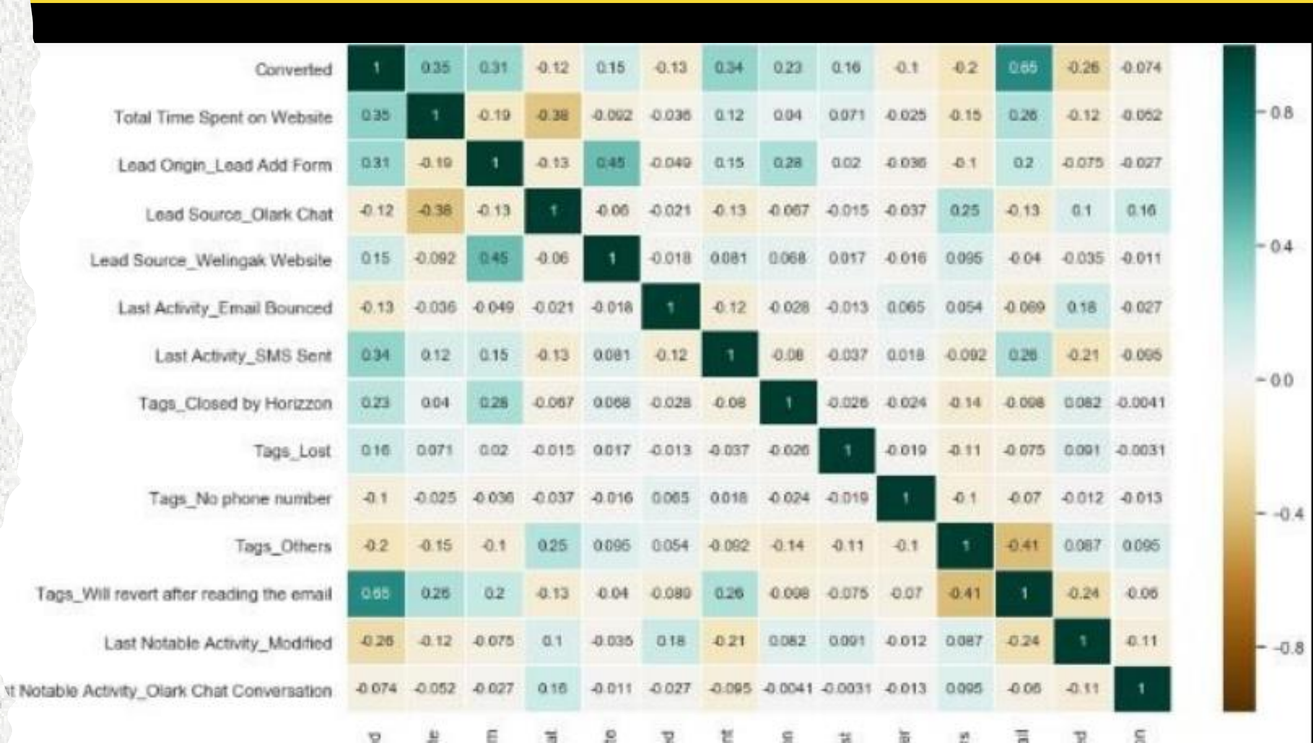


Train Data	
Accuracy	91.23%
Sensitivity	90.53%
Specificity	91.60%

Test Data	
Accuracy	92.06%
Sensitivity	90.53%
Specificity	91.60%

Final Observations

- The model appears to accurately predict the conversion rate, which should enable the education company to select the most promising or “hot” leads
- There are certain variables that help drive hot leads


Converted 1 0.35 0.31 -0.12 0.15 -0.13 0.34 0.23 0.16 -0.1 -0.2 0.65 -0.26 -0.074
Total Time Spent on Website 0.35 1 -0.19 -0.38 -0.092 -0.036 0.12 0.04 0.071 -0.025 -0.15 0.26 -0.12 -0.052
Lead Origin_Lead Add Form 0.31 -0.19 1 -0.13 0.45 -0.049 0.15 0.28 0.02 -0.036 -0.1 0.2 -0.075 -0.027
Lead Source_Olark Chat -0.12 -0.38 -0.13 1 -0.06 -0.021 -0.13 -0.067 -0.015 -0.037 0.25 -0.13 0.1 0.16
Lead Source_Welingak Website 0.15 -0.092 0.45 -0.06 1 -0.018 0.081 0.068 0.017 -0.016 0.095 -0.04 -0.035 -0.011
Last Activity_Email Bounced -0.13 -0.036 -0.049 -0.021 -0.018 1 -0.12 -0.028 -0.013 0.065 0.054 -0.089 0.18 -0.027
Last Activity_SMS Sent 0.34 0.12 0.15 -0.13 0.081 -0.12 1 -0.08 -0.037 0.018 -0.092 0.26 -0.21 -0.095
Tags_Closed by Horizzon 0.23 0.04 0.28 -0.067 0.068 -0.028 -0.08 1 -0.026 -0.024 -0.14 -0.098 0.082 -0.0041
Tags_Lost 0.16 0.071 0.02 -0.015 0.017 -0.013 -0.037 -0.026 1 -0.019 -0.11 -0.075 0.091 -0.0031
Tags_No phone number -0.1 -0.025 -0.036 -0.037 -0.016 0.065 0.018 -0.024 -0.019 1 -0.1 -0.07 -0.012 -0.013
Tags_Others -0.2 -0.15 -0.1 0.25 0.095 0.054 -0.092 -0.14 -0.11 -0.1 1 -0.41 0.087 0.095
Tags_Will revert after reading the email 0.65 0.26 0.2 -0.13 -0.04 -0.089 0.26 -0.098 -0.075 -0.07 -0.41 1 -0.24 -0.06
Last Notable Activity_Modified -0.26 -0.12 -0.075 0.1 -0.035 0.18 -0.21 0.082 0.091 -0.012 0.087 -0.24 1 -0.11
t Notable Activity_Olark Chat Conversation -0.074 -0.052 -0.027 0.16 -0.011 -0.027 -0.095 -0.0041 -0.0031 -0.013 0.095 -0.06 -0.11 1
0.8
0.4
0.0
-0.4
-0.8