IDC4251C Module 9 Project

In this project we will analyze a financial dataset to predict high credit risks using logistic regression.

Load the CSV data file 'credit_risk_data.csv' from the repository in a new PowerBI workbook and transform the data so the CreditRisk column is represented as whole numbers. Note that this file may contain missing and/or invalid data that you will need to clean (remove any observations that contain this type of data).

Create a scatter chart as you did for Module 8 which represents the data with the Credit Score on the X-axis, Debt-to-Income Ration on the Y-axis, and the credit risk value (1 for high risk, 0 for not low risk) as the legend. Include a slicer to filter on the risk values. Here is my chart:

Remember that for logistic regression to work successfully, our data classifications should be linearly separable. Based on the above chart, do you think this data set fulfills this assumption?

After creating your scatter chart, insert a Python visual on Sheet 2 and load the Python script included in the GitHub repository. This script will run a logistic regression which will fit our data to a logistic curve.

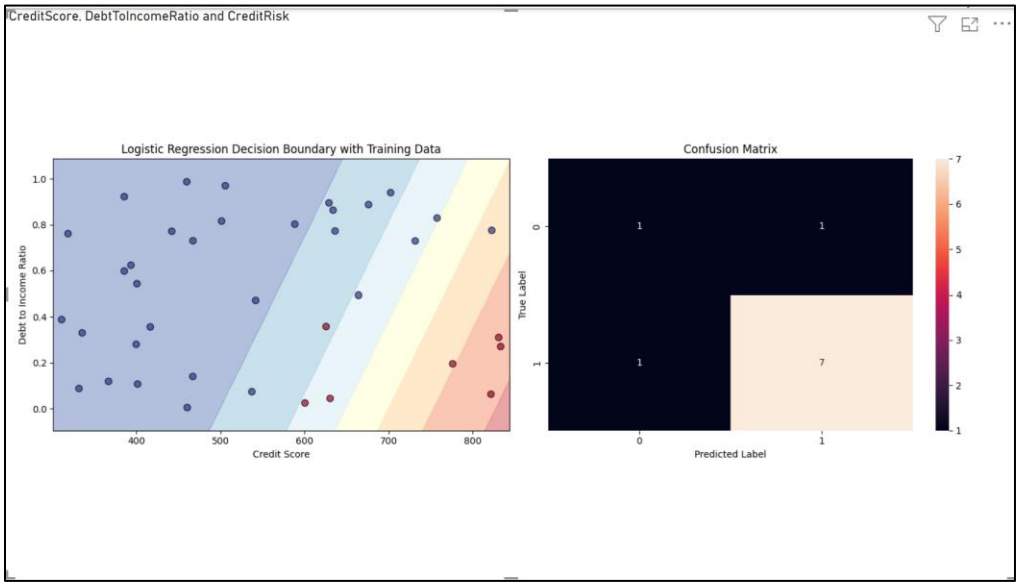
The python code will create two plots:

The first is a contour plot showing the logistic regression decision boundary based on the training data with Credit Score and Debt to Income Ratio. This plot represents the probability of being classified as high risk, with different colors indicating the probability from the logistic regression model. The scatter plot overlays the actual training data points, colored by their true classification labels.

The second is a confusion matrix of the test set predictions. It provides a visualization of the model's predictions, highlighting the number of true positives, true negatives, false positives, and false negatives.

The value in each cell of the confusion matrix represents the number of predictions relative to its position in the matrix, e.g. the number 1 in the False Positives cell (top right means that there was one instance where the model predicted the positive class (indicating high credit risk in our context) when the actual class was negative (indicating low credit risk).

Relative to the analysis, this single instance of FP means that one customer or data point was flagged as a high risk by the model, even though they were actually a low risk according to the ground truth in the data. In practical terms for a credit risk model, this could mean denying a loan to an individual who was actually creditworthy, which can be a significant error depending on the application's sensitivity to false positives.



In a financial context, particularly in credit scoring and lending, it's crucial to minimize such errors because they can result in lost business opportunities (if creditworthy customers are turned away) or customer dissatisfaction. Therefore, even a single false positive might warrant further investigation to understand why the model made that prediction and to see if the model can be improved.

Submit your completed PowerBI workbook to the GitHub Classroom repository.