Module 20 Report Credit Risk Analysis through Machine Learning

Overview of the Analysis
* Purpose: Given a Dataset of 77,536 customer's financial/loan information (loan size, interest rate, borrower income, debt to income ratio, number of accounts, derogatory marks, total debt and loan status), train and evaluate a ML model to predict healthy and at high-risk loans.

* The model will use "loan size, interest rate, borrower income, debt to income ratio, number of accounts, derogatory marks and total debt values to train a Logistic Regression model and Loan status as target value

* The proceess used to analyze the data and train and test the model was the following:

  1. Import modules and a CSV file with customer information and create a data frame in jupyter notebook.
  2. review dataframe and data for null-values and categoorical data and perfom appropiate encoding if nesseary (None required).
  3. Performed correlation analysis and review for imbalance (imbalance in dataset towards Healthy Loans was present).
  4. Scale/fit the data and seperate the training columns from the target values (Loan_status).
  5. Split_train_test the data and perform logistic model training, save the model and test.
  6. Generate a confusion Matrix and report analysis.

Results
* Machine Learning Model - Logistic Regression
This table summarizes the performance of the Logistic Regression model for two classes:

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     15008
           1       0.89      0.93      0.91       500

    accuracy                           0.99     15508
   macro avg       0.94      0.97      0.95     15508
weighted avg       0.99      0.99      0.99     15508
```

- **Class 0 Healthy Loan** (Majority class, with 15,008 samples)
- **Class 1 High-Risk Loan** (Minority class, with 500 samples)

* Description of Model Accuracy scores:

** Accuracy**
- **0.99**

* Description of Model Precision scores:

Precision
- **Class 0:** 1.00 - The model perfectly predicted all instances of Class 0 (no false positives).
- **Class 1:** 0.89 - When the model predicts Class 1, 89% of those predictions are correct.

* Description of Model Recall scores:

Recall

- **Class 0:** 1.00 - The model correctly identified all actual Class 0 samples (no false negatives).
- **Class 1:** 0.93 - Out of all actual Class 1 samples, 93% were correctly identified, while 7% were misclassified as Class 0.

* Descrption of Model f-1 Score
F1-Score
- **Class 0:** 1.00 - 100% classification.
- **Class 1:** 0.91 - A good balance between precision (0.89) and recall (0.93).

Summary
- The model perfectly predicted on **Class 0** (likely due to its dominance in the dataset).
- For **Class 1**, the model is quite good (F1-score = 0.91), though **precision (0.89)** is slightly lower than **recall (0.93)**, meaning the model makes some false positive errors (misclassifying Class 0 as Class 1).
- Since the dataset is highly **imbalanced** (Class 0 has more samples than Class 1), the **high performance on Class 0 might indicate a bias toward the majority class**.
This bias can impact the results and create a situation in denying eligeble borrows a loan (More likely) or putting the lending institution at risk by providing loans to high risk borrowers.

I would still recommend the Logistic Regression model as a predictive tool for Healthy and High-risk Loans. It had a high accuracy overall and high scores on all other metrics.

** Suggestions for Improvement**
Resampling, either oversampling Class 1 or undersampling Class 0 to balance the bias could help.