Module 20 - Analysis Report

OVERVIEW

Purpose of this Report is to analyze the past data to make an accurate model to determine if loans would be successful or default. The factors the models looked at to predict if a loan would be successful or not are loan size, interest rate, borrower income, debt to income, number of accounts, derogatory_marks, and total debt.
The csv file had over 70,000 loans that could be grouped into over 5,000 different types (value_count). Each factor had unique values ranging from 4 (derogatory_marks) to 4692 (interest_rate).

The data was put in a linear regression model (fits a straight line or surface that minimizes the discrepancies between predicted and actual output values). We looked at the status of the loan (0 = in good status, 1 = in bad status) on the 7 factors listed above. The model made predictions of the loan status based on the 7 factors and those predictions (X_test) were compared to the actual results (Y_test). The comparisons were put into a confusion matrix and a classification report was made to view the results of the predictions.

Results for the Linear Regression model from Confusion Matrix

	Precision	Recall	F1-Score	Support
Low Risk Loan	1.00	0.99	1.00	18765
High Risk Loan	0.84	0.94	0.89	619
Accuracy			0.99	19384
Macro Avg	0.92	0.97	0.94	19384
Weighted Avg	0.99	0.99	0.99	19384

Summary

Low Risk loans performed the best. It achieved the best accuracy score and F1 score. The model based on the data in the Confusion Matrix was able to correctly identify over 18,000 low risk loans. High Risk Loans had a very good precision (0.84) and F1-score (0.89). The

majority of the loans in low risk were in good status (18655 versus 110). The majority of the high risk loans were in bad status (583 versus 36). This model it is more important to determine the bad status of the loans than the good status. At the current state I would not recommend this model. Even though .84 is a good score, I would suggest the model be improved in efficiency before recommending it. 16% chance to be wrong on high risk loans does not sound like a number that should be used for lending. I would suggest using a standard scaler to see if that would improve the predictions (by having the factors equally weighting against each other in the rows). Also I would suggest having the predictions for loans that had derogatory marks and those that did not have them in different models. I think this model makes a great starting point for a prediction model that needs to be refined before its used.