Initial Results Summary

Strahinja Nakic - 500809487

CIND 820-DAH

Professor Tamer Abdou, PhD

Toronto Metropolitan University

July 14, 2025

After cleaning and preparing the German Credit dataset, I trained three basic classification models to predict whether a borrower is likely to be a good or bad credit risk. The models used were Logistic Regression, Decision Tree, and Naive Bayes. Before modeling, all categorical variables were converted into numerical format using label encoding, and the data was scaled using StandardScaler to improve model performance, especially for logistic regression which had previously struggled to converge. The dataset was then split into training and testing sets using an 80/20 ratio, and 5-fold cross-validation was added to validate model performance.

Of the three models tried, Logistic Regression performed the best with 76 percent accuracy and 0.864 recall score, which means that it was the best-performing model at classifying borrowers that were actually good credit risks. Naive Bayes provided an equally good performance with 0.807 recall and corresponding accuracy. Decision Tree was the least accurate but still provided a good baseline to work against. All three models were evaluated with accuracy, recall, confusion tables, and cross-validation scores in order to obtain a more stable picture of their performance, considering that class imbalance is already an existing issue with credit risk data. The overall results lend credence to the hypothesis that with even simple, interpretable models, it's possible to achieve dependable insights into creditworthiness, especially if coupled with good data preparation. The results lend credence to the method espoused both in the abstract and the literature review, wherein the use of simpler models was due to their interpretability and conformity with compliance-oriented work settings like risk and AML analysis. This project phase verifies that it's possible with a machine learning approach to provide relevant predictions and forms the basis of further analysis or model refinement during subsequential phases.