Assignment 2 Report

Simrat Kaur Randhawa
20031112
srandhaw@stevens.edu
https://www.kaggle.com/simratkr

During my first assignment, the **feature engineering process was extensive**, requiring significant time and computational effort. Given that experience, for this assignment, I **focused more on the computational aspects**, ensuring that the models were optimized for performance and efficiency.

The final submission includes **two regression models**:

1. **Linear Regression:** A baseline model to establish interpretability and performance benchmarks.
2. **Lasso Regression:** Introduced to **reduce overfitting** and handle potential multicollinearity by applying L1 regularization, thereby performing feature selection.

For **predicting loan status**, I implemented **logistic regression** but opted for **random sampling over SMOTE**. The dataset is highly complex, and after careful evaluation, I determined that SMOTE (Synthetic Minority Over-sampling Technique) might introduce **unwanted noise**, potentially **distorting decision boundaries**. Instead, **random sampling provided a better balance** between minority and majority classes without over-amplifying synthetic patterns.

Additionally, the models were trained using a **rolling window approach instead of an expanding window**. After conducting multiple trials, I observed that the rolling window approach **led to more stable and generalizable results** by ensuring that the model continuously adapted to new data while maintaining a fixed training period. This approach helped mitigate the effects of **concept drift** and **seasonal trends** in loan data, ultimately improving predictive accuracy.

Overall, the models and methodologies were carefully selected based on majorly multiple test runs.

Generative AI help:

In this assignment I used ChatGPT's help for a few debugs and aside from that the project was done in visual studio.