

Executive Summary: Credit Risk Classification Model Performance Analysis

In this analysis, we evaluated three machine learning models for credit risk classification: Random Forest (RF), XGBoost (XGB), and Deep Learning (DL). The dataset consisted of 200 samples, with approximately 70% belonging to class 0 (low-risk) and 30% belonging to class 1 (high-risk).

Model Performance Metrics:

1. Random Forest: The model demonstrated relatively high precision for low-risk samples (0.80) but lower precision for high-risk samples (0.52). Recall was higher for both classes, with a notable improvement in high-risk class recall (0.55 vs 0.78 for low-risk). The overall accuracy of the model was moderate at 0.715.
2. XGBoost: This model outperformed RF in terms of precision and recall for both classes, with a significant increase in high-risk class recall (0.56 vs 0.87 for low-risk). However, the overall accuracy was higher than RF but lower than DL at 0.78.
3. Deep Learning: This model achieved the highest precision for both classes, with a slight edge over XGBoost in high-risk class precision (0.65 vs 0.52). Recall was slightly lower compared to RF and XGBoost for both classes but still acceptable. The overall accuracy of the model was moderate at 0.705.

It is worth noting that while DL had a slightly higher weighted F1-score than XGBoost, it also had the lowest recall for high-risk samples. This suggests that the DL model may be biased towards predicting low-risk samples, which could result in more false negatives for high-risk cases.

Class Distribution and SMOTE Impact: The class distribution plots showed a significant imbalance between low-risk (70%) and high-risk (30%) samples. To address this issue, the dataset was resampled using the Synthetic Minority Over-sampling Technique (SMOTE). The post-SMOTE plot showed a more balanced distribution between the two classes.

Confusion Matrices: The confusion matrices for all three models demonstrated that all models had a higher tendency to predict low-risk samples, as shown by the larger number of false negatives in the high-risk class and smaller numbers of false positives in the low-risk class. The Deep Learning model had the highest number of false negatives among the three models.

Feature Importance: The feature importance plots for RF and XGBoost were similar, with credit history, loan amount, and income being the most important features in predicting credit risk. For DL, these top three features remained consistent, but other factors such as purpose of loan and employment status also played a significant role in model predictions.

Key Findings: - All models demonstrated relatively high precision for low-risk samples but struggled to accurately classify high-risk samples. - The Deep

Learning model had the highest overall accuracy, but it also had the lowest recall for high-risk samples and a tendency to overfit on the low-risk class. - Class imbalance in the dataset was addressed using SMOTE, leading to a more balanced distribution between classes.

Actionable Recommendations: 1. Optimize the Deep Learning model by adjusting hyperparameters or using ensemble methods to reduce overfitting and improve high-risk class predictions. 2. Evaluate the use of additional features such as employment status, purpose of loan, and other relevant factors that may impact credit risk. 3. Continuously monitor and retrain the models with new data to ensure their performance remains robust over time. 4. Implement strategies to mitigate false negatives in high-risk samples, such as more stringent underwriting guidelines or additional review processes for borderline cases.