Bias Analysis in Training Dataset and Mitigation Strategies

Tran Quoc Thai
Student ID: 2370759

April 9, 2025

1 Identifying Potential Biases

The training dataset consists of features such as Age, CreditScore, and Education to predict the RiskLevel. Biases in the dataset can arise due to:

- **Class Imbalance:** The dataset contains an unequal distribution of RiskLevel labels, with more instances of one class than another, which can lead to a biased model favoring the majority class.
- **Feature Distribution Skew:** The distribution of CreditScore and Education levels may not be representative of the broader population, leading to models that generalize poorly.
- **Missing Data:** The presence of missing Education values in both training and test data introduces bias, as predictions for individuals with missing values may be systematically different.
- **Age Representation:** If certain age groups are overrepresented, the model may be biased toward the characteristics of those age groups, reducing fairness across different demographic segments.

2 Bias Mitigation Strategies

To address these biases, we propose the following methods:

1. **Reweighting and Resampling:** To correct class imbalance, we can either upsample the minority class or assign higher weights to underrepresented classes during training. This ensures that the model does not favor one class disproportionately.
2. **Data Imputation and Augmentation:** Missing values in the Education feature can be imputed using statistical methods such as mean/mode

imputation or predictive imputation using regression models. Additionally, synthetic data generation techniques can be applied to balance the feature distributions.

These strategies help to improve the fairness and generalization ability of the model, ensuring that predictions are not systematically biased due to dataset limitations.