Model Card – Loan Repayment Prediction

Model Details
- Developed by Xiang Gao at Lehigh University in 2023.
- XGBoost
- Pretrained for repayment prediction then fine-tuned with accuracy score for a balance between performance and efficiency.

Intended Use
- Intended to be used for predicting which people in the future are likely to repay their loan and those who are more likely to default.
- Particularly intended for applicants who report their property value.
- Not intended to make fully autonomous credit decisions.

Factors
- Given the effect of disparities in overall income and risk appetite on loan repayment, potentially relevant factors include groups for age, gender, race, ethnicity, and geographic tract.
- Evaluation factor include the race group, which comprises categories such as 2 or more minority races, American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White, as reported by the applicants.

Metrics
- Evaluation metrics include Acceptance Rate, False Positive Rate, False Negative Rate, and Calibration. These respectively assess the rate of predicted approvals, the probability of falsely accepting a negative case (agree to grant the loan to applicants who will default), the probability of falsely denying a positive case (refuse to grant the loan to applicants who will repay), and the alignment of predicted probabilities with actual occurrence rates.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the 0.5 decision threshold.

Training Data
- Derived from the publicly available HDMA dataset, including only home loan applications in Pennsylvania from 2019 with outcomes of either 'originated' or 'denied', omitting other result types. From this dataset, 70% is randomly allocated as the training data.

Evaluation Data
- From the above HDMA dataset, the remaining 30% is allocated as the testing data.

Ethical Considerations
- The HDMA dataset exhibits significant imbalance, especially within certain protected groups of people that have a limited number of loan applications, potentially resulting in lower predictive accuracy for these groups. Despite employing oversampling, it's challenging to address this issue effectively across all categories and protected features.
- The HDMA dataset was created and released by CFPB (Consumer Financial Protection Bureau), which collected the original information from the different financial institutions. Within this dataset, the loan approval outcomes may reflect biases and discriminatory practices by professionals within those financial institutions during the loan approval process. These biases and discriminatory practices may be embedded in the model during the training process.

Caveats and Recommendations
- The training and testing data are exclusively composed of Pennsylvania home loan applications from 2019, making the model potentially less effective for predicting the loan repayment in diverse geographic locations or across different time periods.

Quantitative Analyses