Model Card – Loan Approval Predictions

Model Details

- **Person Developing Model**: XYZ
- **Model Date**: December 5, 2024
- **Model Version**: v4
  - ➢ v1- It was the first version of the model, developed with simple data cleaning and data handling techniques. However, the preparation of data was not satisfying, making a model that would not be so accurate. The features selected did not fully correspond to the target variable, which directly affected the model's predictability.
  - ➢ v2 - In this version, I conducted a full value audit, meaning checking the model from different perspectives by selecting the features in such a way as to identify more appropriate variables with regard to the target. This greatly improved the performance, increasing the accuracy of the model from 30% to approximately 93%, showing great improvement.
  - ➢ v3 - In an attempt to let the model be non-biased, I used three fairness metrics based on research studies. To monitor the performance of the model on various demographic groups, the same were used to analyze the model. With this, I was able to look at the model's effectiveness more equitably among groups that no demographic group was being treated inequitably by the model's outcomes.
  - ➢ v4 - Based on the previous analysis, I enhanced the model by addressing feature selection and possible biases. I created three more models, this time with balanced demographic representation of race, sex, and ethnicity in order to ensure fairness. This approach allowed me to free the model from any bias and check whether its predictions were fair for different demographics.
- **Model Type**: The model is a logistic regression model intended for binary classification. It employs the binomial family to estimate the probability of a binary outcome, such as loan approval versus denial.
  - ➢ Logistic regression was chosen because of its interpretability and efficiency in modeling the relationship between predictors and the target variable (binary classifier), `action_taken_binary`.
  - ➢ The data were also balanced across the demographic groups for each demographic group to develop three different models. This was done with specific purposes of testing bias or whether the model predictions differed for these groups.
  - ➢ Also, the performance of the model was done using the ROC curve, which gives a notion about the balance between sensitivity and specificity.

- **Information about Training Algorithms, Parameters, Fairness Constraints, or Other Applied Approaches, and Features**:
    - Training Algorithm: The model was developed using logistic regression for binary classification, specifically utilizing the binomial family. The model has been optimized using maximum likelihood estimation (MLE).
    - Fairness Constraints: To address fairness, data balancing techniques were implemented to reduce potential biases related to demographic factors such as race, gender, and ethnicity, promoting a more equitable representation during the training of the model.
    - Features: The model incorporated the following features-
        - Loan Details: `conforming_loan_limit`, `loan_amount`, `loan_type`, `loan_purpose`, `hoepa_status`
        - Applicant Financials: `income`, `debt_to_income_ratio`, `applicant_credit_score_type`
        - - Demographics: `derived_sex`, `derived_ethnicity`, `derived_race`
        - - Outcome Variable: `action_taken` (converted into a binary variable to indicate whether a loan was approved or denied)
    - The reason I chose these features is because I believe all applicants deserve to be treated in a fair and consistent manner by the lenders throughout the credit granting process, irrespective of the applicant's personal characteristics, and thus should be assessed on their creditworthiness and financial qualifications rather than discriminatory factors.
- **Paper or Other Resource for More Information**:
    - https://dl.acm.org/doi/10.5555/3648699.3649011
    - https://www.experian.com/blogs/insights/fair-lending-and-machine-learning-models/
    - Public HMDA - LAR Data Fields | HMDA Documentation
- **Citation Details**
    - .XYZ (2024). Fair Loan Prediction: A logistic regression model incorporating fairness metrics. The Journal of Machine Learning Research.
- **License**: xyz License
- **Where to Send Questions or Comments About the Model**: contact@xyz.org

Intended Use

- **Primary Intended Uses**: The model is specifically designed to predict loan approval for housing by concentrating on the financial aspects of an individual, including income, loan amount, and other pertinent financial metrics.
- **Primary Intended Users**: This model is intended for bank lenders, mortgage companies, and other stakeholders involved in evaluating loan eligibility.
- **Out-of-Scope Use Cases**:
  - ➢ Predicting approval for loans unrelated to housing or other financial products.
  - ➢ Utilizing non-financial factors (such as demographic or personal traits) to determine loan approval.
  - ➢ Making decisions based solely on the model's output without any human intervention.
  - ➢ Applying the model in areas or financial situations that differ significantly from the dataset's context (for example, outside of Georgia or in different regulatory settings).

Factors

- **Relevant Factors**:
  - ➢ Financial attributes The main attributes are income, loan amount, debt-to-income ratio, credit score, etc..
  - ➢ Demographic groups: Representation and fairness across sex, ethnicity, and race to avoid discrimination and biases.
  - ➢ Data quality: The completeness, accuracy, and reliability of the financial data provided by applicants are paramount.
- **Evaluation Factors**:
  - ➢ **Performance across financial attributes**: How accurate the predictions are for varying ranges of debt-to-income-ratio, loan amount, and credit score to ensure relevance to a wide range of applicants.
  - ➢ **Performance across demographic groups**: For testing the fairness of the model, the performance of different demographic groups was assessed using a variety of metrics: the confusion matrix, ROC curve, and approval rates from both the original and three balanced models. The confusion matrix allowed us to have an insight into true positives, false positives, true negatives, and false negatives for each demographic group, thus providing an ability to conduct detailed error analysis. The ROC curve showed the balance of sensitivity and specificity within these groups, reflecting a good reliability of predictive performance. Furthermore, approval rates were analyzed between the original and balanced models to determine if the interventions enhanced fairness without unduly benefiting or

disadvantaging specific groups. This method directly tackles the fairness metric known as demographic parity, ensuring that the likelihood of loan approval is consistent across different demographic categories such as gender, ethnicity, and race. By balancing the dataset and reassessing the models, the analysis aimed to uphold both fairness and predictive accuracy, with a strong emphasis on reducing bias against underrepresented groups.

- **Error analysis**: Monitoring false positives (approvals that should be denials) and false negatives (denials that should be approvals) to maintain trust in the system.

Metrics

- **Model Performance Measures**:
  - TPR: Proportion of actual approved cases that were correctly predicted as approved
  - FPR: Proportion of actual denied cases that were incorrectly predicted as approved
  - TNR: Proportion of actual denied cases that were correctly predicted as denied
  - FNR: Proportion of actual approved cases that were incorrectly predicted as denied
    1. Original Model
        - i) Accuracy: 96.11%
        - ii) True Positive Rate: 95.65%
        - iii) False Positive Rate: 2.71%
        - iv) True Negative Rate: 97.28%
        - v) False Negative Rate: 4.34%
    2. Race Balanced Model
        - i) Accuracy: 97.22%
        - ii) True Positive Rate: 96.11%
        - iii) False Positive Rate: 0.59%
        - iv) True Negative Rate: 99.41%
        - v) False Negative Rate: 3.89%
    3. Ethnicity Balanced Model
        - i) Accuracy: 95.44%
        - ii) True Positive Rate: 94.63%
        - iii) False Positive Rate: 2.48%
        - iv) True Negative Rate: 97.52%
        - v) False Negative Rate: 5.37%
    4. Sex Balanced Model
        - i) Accuracy: 95.53%
        - ii) True Positive Rate: 94.66%
        - iii) False Positive Rate: 2.31%

iv) True Negative Rate: 97.69%

v) False Negative Rate: 5.34%

- The performance metrics of the original model and the three models balanced on Race, Ethnicity, and Sex do not really vary in their key indicators. True positive rates, false positive rates, true negative rates, and false negative rates maintain the same pattern in these different model configurations, which is indicative of a robust model that has no biases. The small variations in these metrics, typically less than 2 percentage points, indicate that the original model does not exhibit significant demographic bias. Moreover, the performance of the balanced models is very close to that of the original model, further supporting the fact that the model retains predictive accuracy and fairness across different demographic groups. This consistency strongly indicates the model's reliability and equitable performance, showing that it delivers balanced predictions regardless of race, ethnicity, or sex.

Evaluation Data

- **Datasets**: The evaluation data comes from the Loan Application Register (LAR) dataset, which is provided by the Federal Financial Institutions Examination Council (FFIEC). This dataset includes comprehensive loan-level information, such as applicant demographics, loan features, and outcomes, making it very useful for financial modeling. It is publicly accessible and complies with government regulations, ensuring transparency and adherence to data privacy standards. More details about the dataset can be found here.
- **Motivation**: The dataset was chosen because it is a valid source from the government, which means that it meets the legal requirements and is ethical in terms of data collection. Being public also ensures that the data protects privacy, hence trustworthy and secure to work with. Besides, the completeness of the dataset supports good modeling for loan approval and financial behavior analysis, which is very close to the objective of this project.
- **Preprocessing**: Cleaning and preprocessing of the dataset were done to make it ready for analysis. The important steps involved in this were the selection of relevant financial attributes, including debt to income ratio and loan amount, which are vital predictors for loan approval. Highly collinear features were identified by performing a correlation analysis that enhances the robustness of the model. Missing values were handled using a simple na.omit method post feature selection to avoid redundant data loss. The dataset was later balanced to handle fairness issues and, therefore, reduce any potential biases in the model.

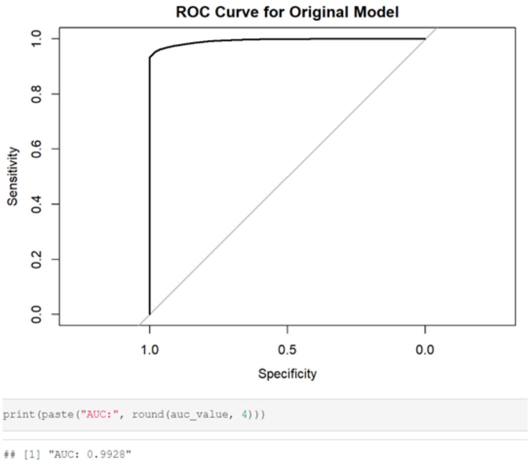
Training Data

The training data were prepared in a 70-30 train-test split, with 70% of the dataset allocated to model training and 30% for evaluation. This was performed on a stratified basis based on the target variable (`action_taken_binary`) to ensure consistent class proportions in both sets. Feature selection and handling of missing values were done for the training data, similar to that of the evaluation data. This method allows the model to be trained on a representatively sampled dataset while preserving much of the integrity of the distribution, reducing potential biases.

Quantitative Analyses

- **Unitary Results**:

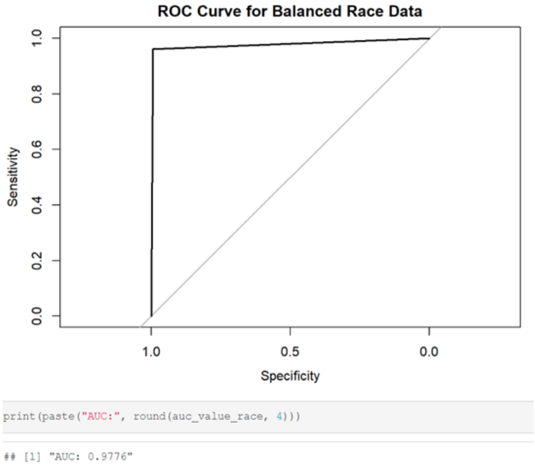
➢ ROC Curve



```
print(paste("AUC:", round(auc_value, 4)))
```

```
## [1] "AUC: 0.9928"
```

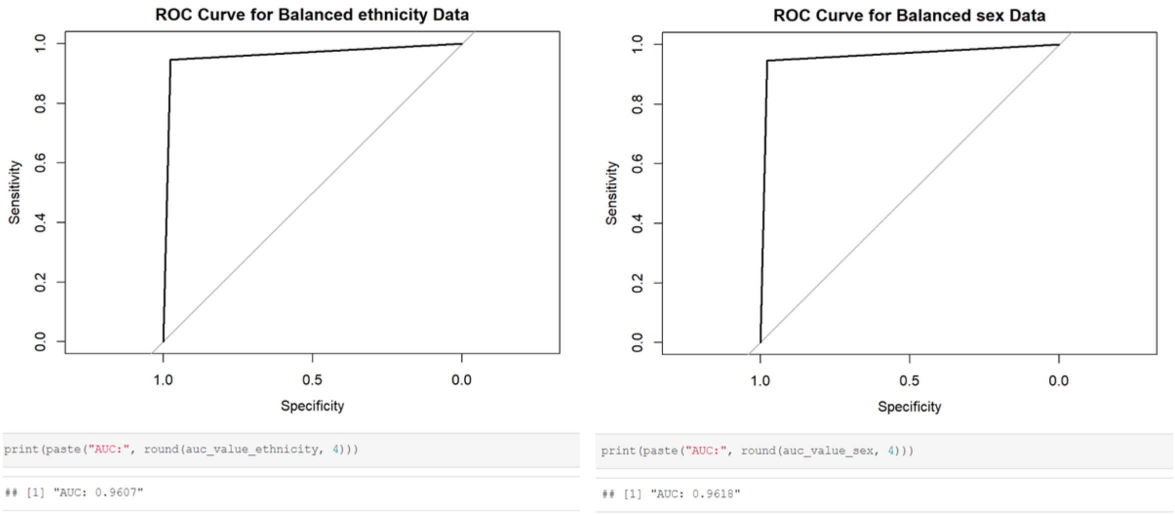
(a)



```
print(paste("AUC:", round(auc_value_race, 4)))
```

```
## [1] "AUC: 0.9776"
```

(b)



ROC Curve for Balanced ethnicity Data

```
print(paste("AUC:", round(auc_value_ethnicity, 4)))
```

```
## [1] "AUC: 0.9607"
```

ROC Curve for Balanced sex Data

```
print(paste("AUC:", round(auc_value_sex, 4)))
```

```
## [1] "AUC: 0.9618"
```

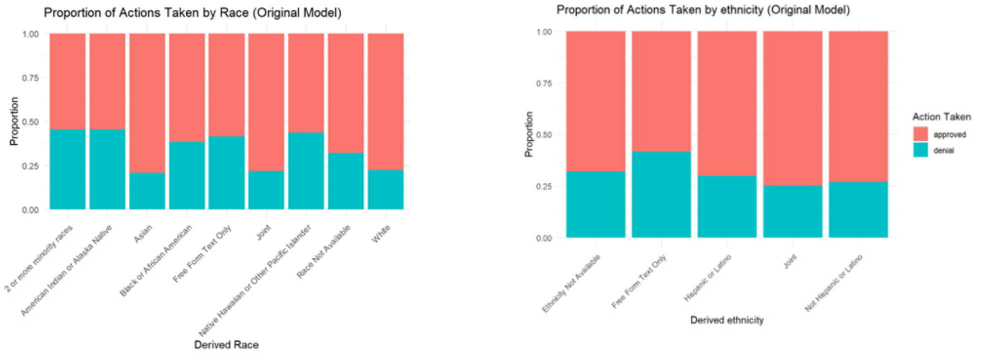
(c)

(d)

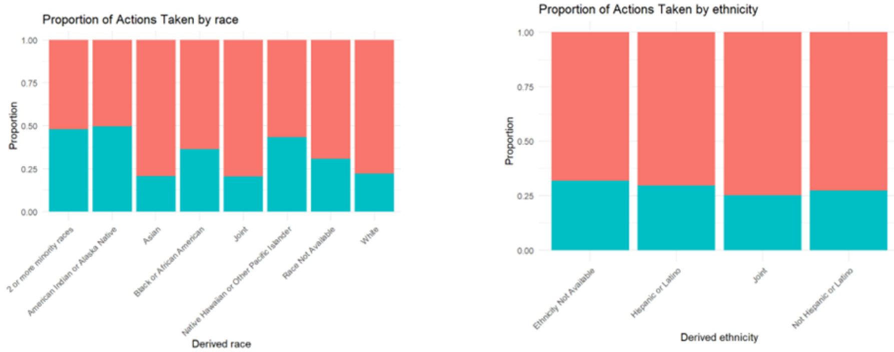
The most important conclusion from these results is that all three balanced models, namely the ones for race, ethnicity, and sex, are producing results very close to that of the original model. All AUC values are larger than 0.95 and the ROC curves are much alike for the different model configurations. This suggests that the original model does not have significant demographic bias, since these balanced models do not have major improvements over it. This consistency in performance across these subgroups indicates that the model should be able to make decent predictions independent of race, ethnicity, or sex.

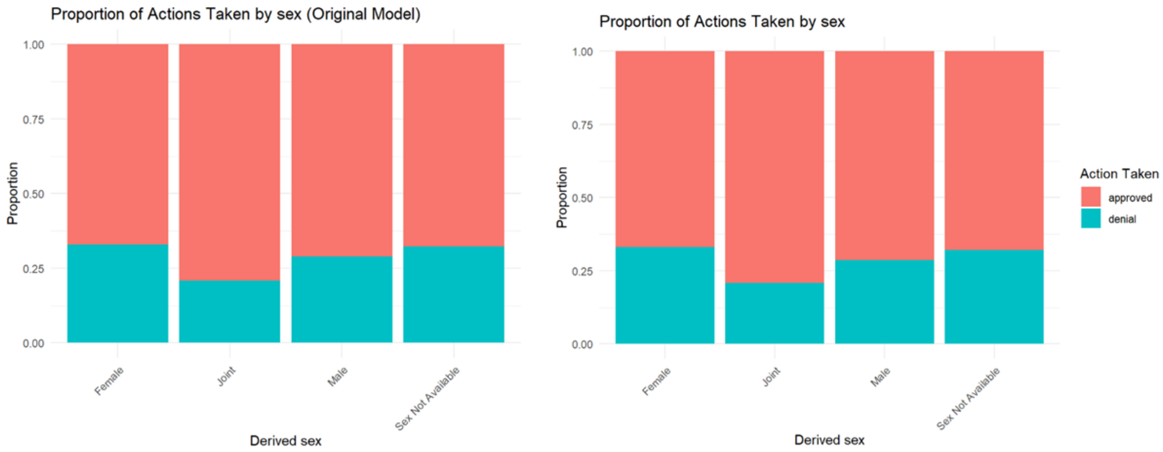
- Approval Rates

(e)



(f)





(g)

Overall, the original model seems to operate fairly and without bias among various demographic groups, as shown by the similar rates of approved and denied actions. The balanced models do not indicate any significant problems that weren't already evident in the outputs of the original model. This implies that the original model is well-calibrated and does not show major demographic biases, despite some minor differences in proportions among specific groups. The comparisons with the balanced models further support the fairness and equity of the original model.

Ethical Considerations

- **Bias and Fairness**: Several demographic groups were tested in model performance to reveal the existing biases. We then did an analysis of several metrics of hit rate, false positive rate, and false negative rate to identify disparities in outcomes according to gender, ethnicity, and race. We then employed multiple techniques for balancing the dataset and enhancing fairness by reducing bias. Explicitly, we test several fairness measures with the intent that all groups have the same outcome.
- **Privacy**: Our data set, for the protection of data, comes from public and government-provided sources that follow the best-established standards, legally and ethically. We never use personal data. Strict respect for all the norms that guarantee privacy, and the protection of single individual rights is observed.

Caveats and Recommendations

- **Caveats**: Most of the model predictions are based on financial characteristics and do not take into consideration all minor factors that may affect loan approval, such as a change in economic or political status.
- **Recommendations**: Any model performance may break for test distributions significantly different from the training data sets, for instance, in regional economies. These can be improved by regular checks to remove biases in the fairness of outcomes for different populations. Besides, regular recycling with new data maintains performance and adaptation of changes in economic and social life.