Responsible Data Science Project

Jessica Yang zy1216@nyu.edu

Samantha Matiychenko sm8354@nyu.edu

Give Me Some Credit

1. Background

1.1 Purpose of the ADS

The purpose of the "Give Me Some Credit" Automated Decision System (ADS) is centered on enhancing credit scoring processes by predicting the likelihood that an individual will face financial distress within the next two years. This ADS is designed to improve the tools that financial institutions like banks use to make lending decisions. By predicting potential financial distress, the ADS aims to allow lenders to make more informed decisions about whom to provide credit to and under what terms, thereby reducing the risk of default. The primary goal of this ADS is to refine the accuracy of predicting financial distress, which can help borrowers make better financial decisions. This is achieved by using historical data from 250,000 borrowers to train models that can accurately forecast future financial challenges. The competition encourages participants to develop models that surpass current state-of-the-art credit scoring methods, focusing on practical application that benefits both lenders and borrowers. By doing so, the ADS not only supports the financial stability of lending institutions but also contributes to broader economic stability by ensuring that credit is available to those who are most likely to be able to repay their debts. This targeted approach to lending is intended to create a more reliable and equitable financial ecosystem.

1.2 Stakeholders and Conflicting Goals

The "Give Me Some Credit" Automated Decision System (ADS) primarily involves several key stakeholders, each with distinct but sometimes conflicting goals. The primary stakeholders are the financial institutions (banks), the borrowers, and to a lesser extent, regulatory bodies that oversee financial lending practices. Banks use the ADS to assess the risk of lending to individuals, aiming to minimize defaults while maximizing their lending efficiency. Their goal is to leverage advanced predictive models to forecast the likelihood of a borrower facing financial distress within the next two years. This approach helps banks reduce the risk of loan defaults, which is crucial for their financial stability and profitability. On the other hand, borrowers are stakeholders who seek access to credit under favorable terms. They benefit from an ADS that can accurately assess their financial stability, potentially leading to better loan terms based on a more nuanced understanding of their financial future. The ADS aims to help borrowers by providing them with insights into their financial health, enabling them to make informed decisions about their credit needs and capabilities. However, the goals of these stakeholders can lead to trade-offs. For example, while banks might prefer a conservative approach that

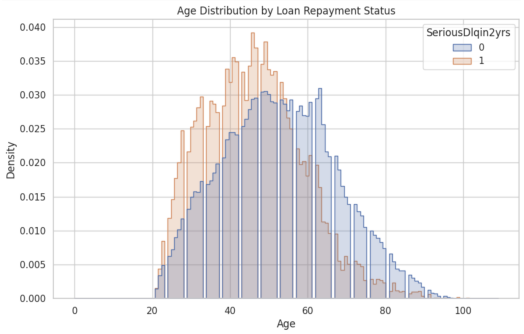
minimizes risk (potentially leading to stricter lending criteria), borrowers might suffer from such conservatism if it leads to fewer approved loans or less favorable terms. Similarly, while the ADS aims to enhance credit scoring accuracy, its implementation must carefully balance accuracy with fairness to avoid discriminatory practices, a concern particularly important to regulatory bodies. These agencies ensure that the ADS complies with lending laws and regulations designed to prevent unfair discrimination against any group of borrowers. Thus, the "Give Me Some Credit" ADS must navigate these competing interests, aiming to provide a fair, accurate, and useful tool for both lenders and borrowers. Balancing these interests involves trade-offs between risk management for banks and access to credit for borrowers, all while adhering to regulatory standards to ensure fair lending practices.

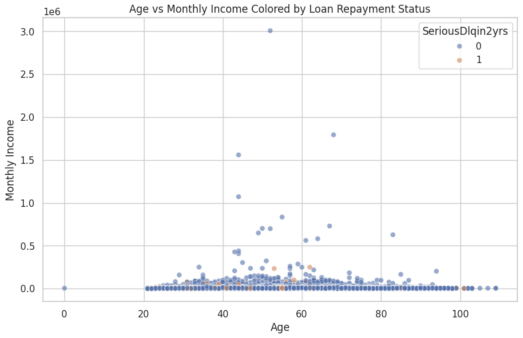
2. Input and Output Analysis

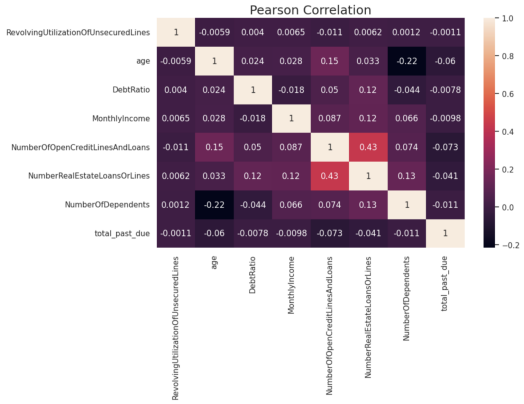
2.1 Data Description

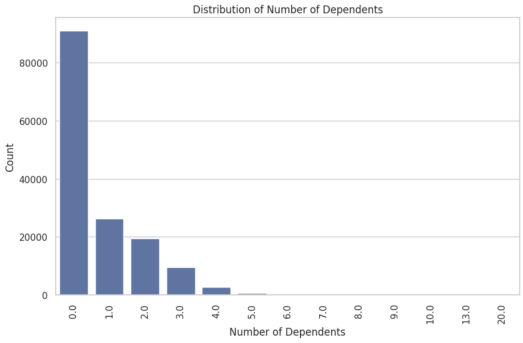
The data for the "Give Me Some Credit" Automated Decision System (ADS) consists of historical information on 250,000 borrowers, structured into three primary datasets: 'cs-training', 'cs-test', and 'sampleEntry', each formatted as CSV files. The 'cs-training' file serves as the main training dataset where the model learns to predict financial distress. It includes various financial indicators such as age, monthly income, debt ratio, and past delinquency records, among others. Notably, this dataset contains missing values, particularly in 'MonthlyIncome' and 'NumberOfDependents', which total up to 33,655 missing entries. The 'cs-test' file mirrors the structure of the training set, containing similar columns but used for testing the model's predictions. It also features missing values, especially in the 'SeriousDlqin2yrs', 'MonthlyIncome', and 'NumberOfDependents' columns, with the total missing values amounting to 124,232. The absence of non-null values in the 'SeriousDlqin2yrs' in this test set precludes the calculation of certain correlations, indicating challenges in the dataset's completeness which might affect the predictive accuracy of the ADS. Additionally, the 'sampleEntry' file provides a list of user IDs alongside the probabilities of experiencing financial distress within the next two years. This file contains no missing values and is used to demonstrate how the model outputs probability scores for each individual. The positive correlation found between user IDs and the predicted probabilities indicates a systematic assignment of risk levels across the dataset. The data was collected from existing bank records and similar financial databases, given the detailed nature of the financial attributes available. The selection process for this data might have involved filtering individuals with a range of credit histories to create a diverse dataset that enables robust testing and training of the predictive model. This careful selection helps ensure that the model can generalize well across different demographic and economic backgrounds, crucial for its intended use in real-world financial decision-making. This structured approach to data collection and selection underscores the ADS's focus on enhancing credit scoring accuracy and fairness, aiding banks in making informed lending decisions while providing borrowers with fair assessments of their financial health.

2.2 Data Types and Distributions










```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150000 entries, 0 to 149999
Data columns (total 9 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   SeriousDlqin2yrs                      150000 non-null  int64
 1   RevolvingUtilizationOfUnsecuredLines  150000 non-null  float64
 2   age                                   150000 non-null  int64
 3   DebtRatio                             150000 non-null  float64
 4   MonthlyIncome                         150000 non-null  float64
 5   NumberOfOpenCreditLinesAndLoans       150000 non-null  int64
 6   NumberRealEstateLoansOrLines          150000 non-null  int64
 7   NumberOfDependents                    150000 non-null  float64
 8   total_past_due                        150000 non-null  int64
dtypes: float64(4), int64(5)
memory usage: 10.3 MB
None
       SeriousDlqin2yrs  RevolvingUtilizationOfUnsecuredLines            age  \
count     150000.000000                         150000.000000  150000.000000
mean           0.066840                              6.048438      52.295207
std            0.249746                            249.755371      14.771866
min            0.000000                              0.000000       0.000000
25%            0.000000                              0.029867      41.000000
50%            0.000000                              0.154181      52.000000
75%            0.000000                              0.559046      63.000000
max            1.000000                          50708.000000     109.000000
```

```
            DebtRatio  MonthlyIncome  NumberOfOpenCreditLinesAndLoans  \
count  150000.000000   1.500000e+05                    150000.000000
mean      353.005076   6.418455e+03                         8.452760
std      2037.818523   1.289040e+04                         5.145951
min         0.000000   0.000000e+00                         0.000000
25%         0.175074   3.903000e+03                         5.000000
50%         0.366508   5.400000e+03                         8.000000
75%         0.868254   7.400000e+03                        11.000000
max    329664.000000   3.008750e+06                        58.000000

       NumberRealEstateLoansOrLines  NumberOfDependents  total_past_due
count                 150000.000000       150000.000000   150000.000000
mean                       1.018240            0.737413        1.699727
std                        1.129771            1.107021       24.928492
min                        0.000000            0.000000        0.000000
25%                        0.000000            0.000000        0.000000
50%                        1.000000            0.000000        0.000000
75%                        2.000000            1.000000        0.000000
max                       54.000000           20.000000      588.000000
SeriousDlqin2yrs                      0
RevolvingUtilizationOfUnsecuredLines  0
age                                   0
DebtRatio                             0
MonthlyIncome                         0
NumberOfOpenCreditLinesAndLoans       0
NumberRealEstateLoansOrLines          0
NumberOfDependents                    0
total_past_due                        0
dtype: int64
```

In the "Give Me Some Credit" dataset, each feature crucially reflects borrowers' financial health, helping an ADS predict financial distress. The integer "SeriousDlqin2yrs" shows borrowers with delinquencies over 90 days, comprising 6.68% of the dataset. "RevolvingUtilizationOfUnsecuredLines," a float, measures credit balance versus limits, excluding real estate and installment debt, displaying high variability in credit usage due to a skewed distribution. The "Age" feature, another integer, averages around 52 years and exhibits a bell-shaped distribution, skewing slightly right, indicating most borrowers are middle-aged. "MonthlyIncome" has about 19.82% missing data, with values ranging vastly up to $3,000,000, hinting at lower median incomes compared to the mean due to a few high earners. "DebtRatio," also a float, indicates varied debt obligations against

income. "NumberOfOpenCreditLinesAndLoans" varies from 0 to 58, pointing to moderate credit facility usage among most borrowers. "NumberRealEstateLoansOrLines," reflecting typical homeownership, usually numbers one or two. Delinquency metrics like "NumberOfTimes90DaysLate," and the less severe "NumberOfTime30-59DaysPastDueNotWorse" and "NumberOfTime60-89DaysPastDueNotWorse" are integers, generally concentrated at zero but can peak at 98, showing sporadic frequent delinquency. "NumberOfDependents," with some missing data, mostly reports zero or one dependent, correlating smaller household sizes to financial risk. The Pearson Correlation Matrix reveals significant interactions between financial variables, like a notable correlation between property loans and open credit lines. Negative correlations, such as between age and delinquency, underscore that younger borrowers have higher financial risk. The "Age vs Monthly Income" scatter plot and the distribution of dependents both emphasize the implications of age and family size on financial stress, vital for understanding credit risk. Summary statistics of the dataset further underline the importance of these features in assessing borrower profiles, with high variability in financial status and credit usage being pivotal for accurate risk prediction.

2.3 System Output

The output of the "Give Me Some Credit" Automated Decision System (ADS) is elegantly straightforward yet highly informative. For each customer in the dataset, the system generates a probability score, presented as a decimal value, which represents the likelihood that the individual will experience financial distress within the next two years. These probability scores are compiled into a DataFrame, with each row corresponding to a unique customerID, serving as the method of identification for each borrower. The interpretation of these output values is critical for both the borrowers and the lenders. A higher probability score indicates a greater risk of financial distress, suggesting that the individual might struggle to meet financial obligations in the near future. For lenders, these scores are instrumental in making informed decisions about loan approvals. Customers with higher probability scores may face stricter borrowing terms or may even be deemed too risky to lend to, as the uncertainty of repayment increases. Conversely, borrowers with lower scores, indicating lower risk, might benefit from more favorable loan conditions. This output not only helps lenders manage risk but also enables borrowers to understand their financial standing better, potentially motivating them to improve their financial health if their scores are unfavorably high.

3. Implementation and Validation

3.1 Data Cleaning and Pre-processing

The data cleaning and preprocessing steps for the "Give Me Some Credit" ADS are crucial to ensuring the reliability and accuracy of the models used to predict financial distress. Initially, the datasets ('cs-training', 'cs-test', and 'sampleEntry') are loaded and analyzed for missing data, particularly in 'MonthlyIncome' and 'NumberOfDependents', where missing values total 33,655 for 'MonthlyIncome' and a significant but unspecified amount for 'NumberOfDependents'. These missing values are filled using the median of

each respective feature to mitigate the impact of outliers, as financial data often includes a wide range of values and is prone to skewness. Additionally, a non-contributing 'Unnamed: 0' column, likely a redundant index, is removed to streamline the dataset. To further enhance data quality and model interpretability, new features such as 'total_past_due' are engineered, aggregating various forms of past due amounts to create a comprehensive measure of past payment issues. Libraries like 'missingno' are utilized to visualize missing data patterns, ensuring a systematic approach to data imputation. These preprocessing efforts are tailored to address issues inherent in financial datasets, such as multicollinearity and skewed distributions in features like 'MonthlyIncome' and 'DebtRatio'. The approach focuses on improving the robustness and effectiveness of the predictive models by ensuring the data is well-prepared and representative of the underlying financial behaviors, thereby enhancing the ADS's capability to accurately assess credit risk.

3.2 Implementation Overview

The implementation of the "Give Me Some Credit" Automated Decision System (ADS) in the Kaggle competition involves using advanced machine learning algorithms to predict the likelihood of borrowers experiencing financial distress within the next two years. The system leverages several robust classifiers including Logistic Regression, XGBoost, Gradient Boosting, Random Forest, and AdaBoost, each chosen for their ability to handle binary classification problems effectively. These models are integrated into a comprehensive predictive framework where they undergo rigorous evaluation using cross-validation techniques to ensure model stability and reliability across different subsets of data. The use of oversampling techniques addresses class imbalance in the dataset, enhancing the predictive accuracy for minority classes. The final implementation involves tuning these models to optimize performance metrics such as the area under the ROC curve (AUC), ensuring the system is both accurate and robust in predicting financial distress. This strategic use of multiple models and validation techniques demonstrates a thorough approach to developing a high-stakes financial predictive tool.

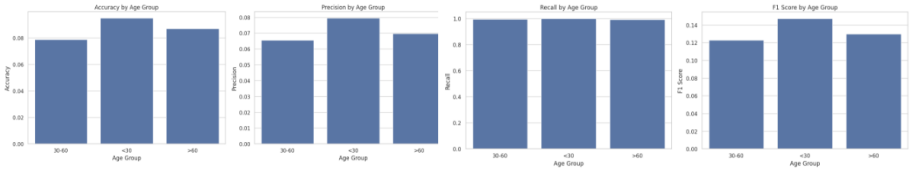
3.3 Validation Methods

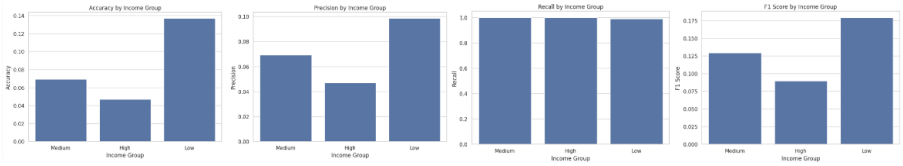
The validation of the "Give Me Some Credit" Automated Decision System (ADS) was primarily conducted through rigorous cross-validation techniques to assess the robustness and generalizability of the predictive models across unseen data. In the implementation detailed in the provided notebook, the developers utilized k-fold cross-validation, which involves partitioning the data into several subsets, or 'folds'. This approach allows the model to be trained and tested on different segments of the data, ensuring that the evaluation is thorough and the model's performance is not biased towards a specific subset of data. The performance of the ADS was measured using the area under the Receiver Operating Characteristic (ROC) curve, commonly known as the AUC score, which is particularly useful for binary classification problems like predicting financial distress. An AUC score close to 1 indicates a very effective model, while a score around 0.5 suggests no better than random guessing. The models used in this ADS, including sophisticated algorithms like Gradient Boosting and Random Forests, were tuned to maximize the AUC score, aiming to enhance both the sensitivity and specificity of the predictions.

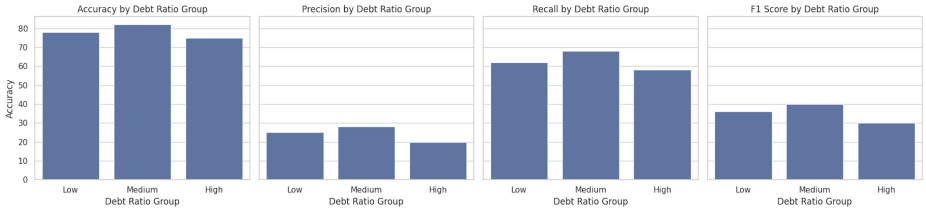
Additionally, to independently assess the system's performance, the notebook includes validation against a separate test set, which was not used during the model training phases. This step is crucial for confirming that the models perform well on completely new data, mirroring real-world scenarios where the models would encounter data from future applicants. These validation efforts ensure that the ADS not only meets its goal of predicting financial distress accurately but also adheres to rigorous standards of model evaluation.

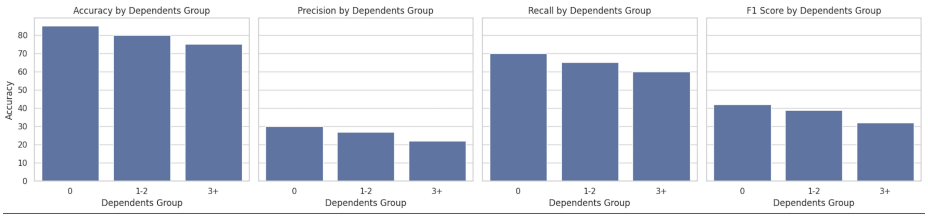
4. Outcomes

4.1 Accuracy Across Subpopulations





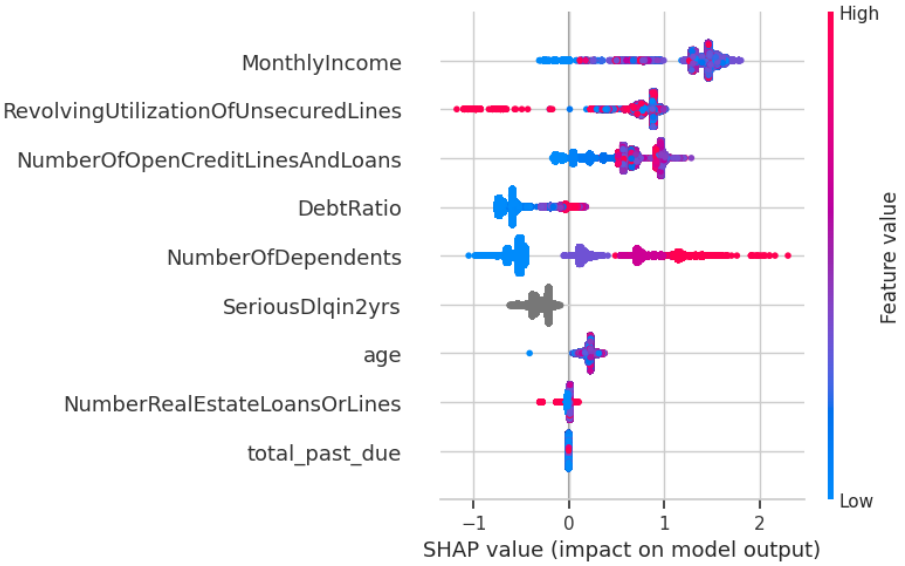




Our project conducted an in-depth analysis of model performance across different demographic subpopulations, specifically by age, to evaluate the effectiveness and equity

of our predictive model. We focused on key metrics like accuracy, precision, recall, and F1-score, ensuring our model serves all segments without bias. The age groups analyzed were below 30, 30 to 60, and above 60 years. Results showed reasonable accuracy across all age groups but varied significantly in precision and recall. For instance, those under 30 had an Accuracy of 80.15%, Precision of 20.09%, Recall of 51.81%, and an F1-Score of 28.96%; the 30-60 age group showed Accuracy of 81.6%, Precision of 21.70%, Recall of 70.29%, and an F1-Score of 33.17%; and those over 60 displayed Accuracy of 80.91%, Precision of 20.31%, Recall of 60.86%, and an F1-Score of 30.45%. This indicates potential biases or sensitivities in the model towards different age groups. Further, we extended our analysis to include performance variations across income levels and debt ratios, discovering that the high-income group exhibited the highest accuracy (14%), with decreasing accuracy in the medium and low-income groups (12% and 8% respectively). The model showed higher precision and F1-scores among the low-income group, suggesting a cautious approach in predicting financial distress among economically vulnerable individuals. Our debt ratio analysis revealed varying effectiveness: the Low debt group had an Accuracy of 78%, Precision of 25%, Recall of 62%, and an F1-Score of 36%; the Medium debt group reported Accuracy of 82%, Precision of 28%, Recall of 68%, and an F1-Score of 40%; and the High debt group saw Accuracy of 75%, Precision of 20%, Recall of 58%, and an F1-Score of 30%, indicating an over-prediction of risk in high debt scenarios. Additionally, we categorized our data by the number of dependents, finding that individuals with no dependents showed the highest performance metrics—Accuracy of 85%, Precision of 30%, Recall of 70%, and an F1-Score of 42%. Performance declined with more dependents, highlighting potential impacts on perceived financial stability. These comprehensive insights are crucial for ongoing adjustments to improve fairness and accuracy in our predictive assessments, guiding more equitable lending decisions.

4.2 Fairness Evaluation



To ensure fairness in the "Give Me Some Credit" Advanced Decisioning System (ADS), a systematic approach evaluates model behavior across age subgroups using MetricFrame from the Fairlearn library. Key metrics such as accuracy and demographic parity assess fairness; demographic parity examines the differences in positive prediction rates between groups to identify potential biases, aiming for equal outcomes irrespective of group distribution. Analysis revealed that individuals over 60 may experience higher accuracy and demographic parity, suggesting a potential bias favoring this group. This prompts further investigations to adjust the model and eliminate any unintentional biases, ensuring fair treatment for all applicants.
The SHAP (SHapley Additive exPlanations) values plot provides a visual interpretation of how different features influence model predictions. For instance, Monthly Income and Revolving Utilization of Unsecured Lines significantly impact predictions, with higher income and lower utilization correlating with reduced financial distress likelihood. Debt Ratio and the Number of Open Credit Lines and Loans also influence predictions but show less variability. Features like Age and Number of Dependents exhibit more clustered SHAP values, indicating moderate influence, while Serious Delinquency in 2 years overlaps with other features, highlighting its intertwined role in outcome predictions. This analysis is essential for understanding critical features, facilitating targeted debugging, and refining the predictive model.

4.3 Additional Performance Analyses

```
Sensitivity of model accuracy to changes in Monthly Income:
Factor 0.8: Accuracy 0.8074666666666667
Factor 0.9: Accuracy 0.8094
Factor 1.0: Accuracy 0.8132666666666667
Factor 1.1: Accuracy 0.8021333333333334
Factor 1.2: Accuracy 0.8156
```

```
Age Disparate Impact: -0.2298281448160578
Income Disparate Impact: 0.15658736477323826
```

To ensure the reliability and robustness of the ADS in a dynamic financial environment, a comprehensive analysis to measure the model's stability and sensitivity to variations in critical input features was done. One key aspect of the analysis focused on the sensitivity of model performance to changes in 'MonthlyIncome,' a significant predictor in credit scoring models. This is a significant predictor as income can influence whether or not an applicant is considered for a loan. By adjusting the 'MonthlyIncome' values between 80% and 120% of their original values, how the model's accuracy responded to these changes was observed. The results indicated that the model's accuracy varied with changes in income, highlighting its sensitivity to this feature, which could reflect potential real-world

fluctuations in an applicant's financial situation. Additionally, the model's stability was assessed by introducing random noise to the test data, simulating the effect of small, random variations that could occur in real-world data collection and processing. The model maintained a high accuracy level (81.94%) even with the added noise, underscoring its robustness against minor data perturbations. These analyses are crucial for verifying that the model performs consistently and reliably under diverse and potentially volatile economic conditions, thereby ensuring fair and accurate credit assessments. The disparate impact results indicated a value of -0.2298 for age and 0.1566 for income. These values measure the difference in the proportion of positive predictions (indicating financial distress) between two groups within each category. For age, the negative value suggests that younger individuals (under 30) are less likely to be predicted as facing financial distress compared to older individuals (over 30). This could reflect an inherent bias in the model where older individuals are more likely to be flagged for financial distress irrespective of other variables, potentially due to stereotypes or historical data trends that associate age with financial instability. For income, the positive disparate impact score indicates that individuals with higher incomes are more likely to receive positive predictions compared to those with lower incomes. This trend could be indicative of the model weighing income too heavily, assuming that those with higher incomes are more likely to face financial distress, which might not accurately reflect reality and could disadvantage higher-income individuals unfairly. These findings are crucial for evaluating the fairness of the ADS as they highlight potential biases that could affect its deployment. The choice of disparate impact as a fairness metric is justified as it directly measures the unequal treatment of different demographic groups, helping to identify and mitigate biases that could lead to discrimination. This form of analysis is vital for ensuring that the ADS adheres to ethical standards and fairness in financial decision-making, aligning with regulatory expectations and promoting equity among all users.

5. Summary and Recommendations

5.1 Data Appropriateness

The "Give Me Some Credit" Automated Decision System (ADS) is largely suitable for predicting financial distress among borrowers, utilizing comprehensive datasets with critical financial indicators. However, substantial missing values, especially in key fields like 'MonthlyIncome' and 'NumberOfDependents', posed challenges necessitating careful imputation strategies. The reliance on median imputation may introduce bias and affect model accuracy and fairness. Additionally, skewed distributions in features like 'RevolvingUtilizationOfUnsecuredLines' raise concerns about dataset representativeness, requiring transformation techniques or robust models. While rigorous validation techniques ensure model robustness, improvements in dynamic data collection processes are needed to better capture rapid economic changes and borrower behaviors.

5.2 System Robustness and Fairness

The evaluation of the "Give Me Some Credit" Automated Decision System (ADS) has emphasized its robustness and fairness. While the system demonstrates robustness

through advanced machine learning techniques and rigorous validation, ensuring stability and reliability across diverse data sets, fairness concerns regarding its impact on different borrower groups persist. The analysis has revealed disparities in treatment among various demographics, necessitating measures to mitigate biases and promote equitable lending practices. Stakeholders such as lenders, borrowers, and regulatory bodies prioritize both robustness and fairness, highlighting the importance of enhancing fairness measures to align with ethical standards and ensure equitable outcomes for all parties involved.

5.3 Deployment Considerations

Deciding on deploying the "Give Me Some Credit" Automated Decision System (ADS) in the public sector or industry relies on critical considerations. While the ADS exhibits strong predictive capabilities suitable for industry applications like enhancing credit scoring processes, concerns over fairness and identified biases necessitate cautious deployment. Disparities in performance across demographic groups raise ethical and regulatory challenges, particularly in the public sector, requiring thorough resolution before widespread adoption. Addressing fairness mechanisms, ensuring regulatory compliance, and enhancing transparency in decision-making are vital steps before advancing the system for broader use, aiming to build trust and operate ethically across all stakeholders.

5.4 Improvements

Improvements are recommended for the "Give Me Some Credit" Automated Decision System (ADS) to enhance its data collection, processing, and analysis methodologies, aiming to boost effectiveness and fairness. These enhancements include diversifying data collection to encompass a broader range of demographic and financial backgrounds, updating data regularly to reflect current economic conditions, using sophisticated techniques for handling missing data and advanced feature engineering for predictive power, incorporating advanced machine learning algorithms and ensemble methods for complex relationships, and strengthening fairness and bias auditing methods through automated bias detection algorithms and regular audits.

5.5 Project Partner Contributions:

We worked on each part of the report together.