GROUP 3

DATA BIAS PROJECT

Seth Phillips, Yousaf Khaliq, Clay Skiles, John Allard

Topic: Credit risk

Dataset: “Give me some credit”

https://www.kaggle.com/c/GiveMeSomeCredit

Objectives:

- Observe bias in "Revolving Utilization Of Unsecured Lines"
- Observe bias in "Age"

Process:

To prepare the data:

- I first label-encoded the age variable in groups (18 to 24 = 1, 25 to 34 = 2, etc.) as the age data is spread out as a qualitative numerical variable.
- Each Revolving Utilization quantitative variable was converted to a percentage.
- Revolving Utilization outliers were eliminated from the dataset.
- A count of all actual delinquents was taken with respect to each age group.

In order to find the bias in Age, the first step is to identify the differences between actual delinquencies and the age bins. Based on Figure 1, it is apparent that the individuals sampled in the younger age groups are more likely to be cited as a potential serious delinquent as 10.45% in the 0-age bin were labeled by a human as a delinquent compared 2.79% in the age 5 bin (ages 65-74). This is evidence of historical bias as it is deemed that younger individuals are more likely to be frivolous with money, which is later to be challenged by the logistic model implemented.

age	delinq_yes	count	delinq_percent
0	217	2075	10.457831
1	1933	17165	11.261288
2	2585	28563	9.050170
3	2850	36776	7.749619
4	1659	34228	4.846909
5	553	19823	2.789689
6	229	11370	2.014072

Figure 1: Comparison of delinquents for each age group

Deploying the logistic model with a 70/30 train-test split, the model is significantly more forgiving in terms of deeming an individual a delinquent in general. Overall, it can be observed from Figure 2 that the ratio of people that are deemed delinquents by the model compared to those judged by humans is lowest with younger individuals as initially recorded in the actual target variable analysis. But the overall magnitude of people deemed serious delinquents is lower, implying the decision-makers possess a confirmation bias in choosing who to label a serious delinquent based on age.

	age	pred_counts	count	delinq_yes	pred_percent
0	0	1	10	217	21.700000
1	1	1	64	1933	30.203125
2	2	1	73	2585	35.410959
3	3	1	55	2850	51.818182
4	4	1	33	1659	50.272727
5	5	1	7	553	79.000000
6	6	1	1	229	229.000000

Figure 2: Logistic Model results of predictions of Serious Delinquents vs. Actual

To find bias in Revolving Utilization of Unsecured Lines, the same logistic model was deployed. Before this model was performed, the values of Revolving Utilization were analyzed with respect to age to identify outliers (Figure 3). It is noticeable that an outlier bias would exist if values above 1000% are not removed from the dataset, so this step is necessary.

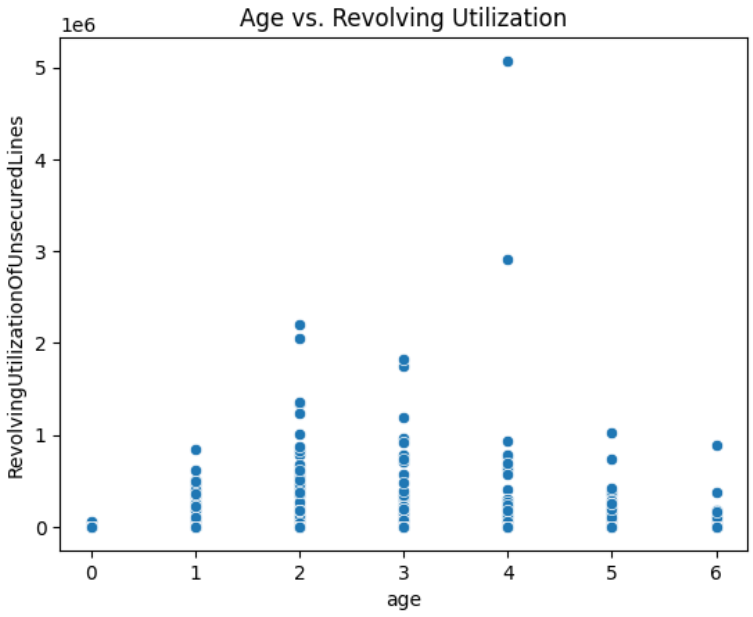


Figure 3: Spotting Outliers in Revolving Utilization

From here the values of Revolving Utilization were further analyzed based on age groups. From Figure 4, these descriptions are laid out: as with most credit bureaus recommending that one should keep their revolving credit under 30%, there is possible selection

bias in the younger age groups as the 50% quartile for those aged 18-24 is 50.78% while those 55-64 have 13.56% revolving credit. It could appear that those selected were chosen to support a theory that young adults are frivolous with money. It also appears the selection bias extends throughout the entire categories of age with regards to their revolving credit as the older the subject is, the "better" their revolving credit. This could be true in the long run due to increasing investments and further development of personal wealth, but there exist many younger individuals whose net worth and liquidation matches that of older individuals.

age	count	mean	std	min	25%	50%	75%	max
0	1626.0	53.042466	42.944709	0.0	6.67	50.780	100.0000	312.62
1	14813.0	47.525688	42.236455	0.0	8.73	39.020	88.8100	885.19
2	24431.0	40.053469	37.680539	0.0	6.22	27.910	71.8150	647.88
3	30499.0	35.085552	35.643040	0.0	4.89	21.610	60.0650	730.90
4	26556.0	28.452622	32.940660	0.0	3.02	13.565	45.3500	600.00
5	14530.0	20.953291	29.012724	0.0	1.90	6.970	28.2225	360.36
6	7641.0	14.094201	24.754944	0.0	1.00	3.280	12.7700	186.71

Figure 4: Descriptions of Revolving Credit with Respect to Age

The Revolving Credit values are binned (x<10 = 0, 10<=x<20 = 1, etc.). It appears that the model is accurate in identifying that higher revolving credit means a greater risk for an individual to be a serious delinquent. There exhibit some cases where those with revolving credit under 30% are predicted by the model to be serious delinquents even though they meet societal expectations. This could mean other factors are indicating their inability for the model to approve their loan.

The values identified by the model are evident of confirmation bias by the user as the model is significantly more lenient to making a judgement as to whether or not a person is at risk of delinquency. The cases where a human has made the claim a loan taker is at risk of delinquency is far greater than those the model deems.

On the other hand, it is clear the model might be too lenient, which brings the discussion of transparency. This model seems to identify far fewer individuals who are at risk of serious delinquency compared to what humans can see. In that case, the model may be looking at objective facts about the training model rather than the subjective. This is why it is important a machine is not the sole decision maker since models can be extremely unpredictable with how they see a classification of individuals.

	RevolvingUtilizationOfUnsecuredLines	pred_counts	count	delinq_yes	pred_percent
0	0	1	1	975	975.000000
1	4	1	2	403	201.500000
2	5	1	5	477	95.400000
3	6	1	1	476	476.000000
4	7	1	1	576	576.000000
5	8	1	16	676	42.250000
6	9	1	216	3572	16.537037

Figure 5: The countof Serious Delinquent predictions grouped by Revolving Credit

This portion of the write-up was made with the help of ChatGPT – Yousaf

Bias Analysis of Selected Features

Objective

The objective of this analysis was to identify potential biases in four key features of the dataset concerning the target variable SeriousDlqin2yrs (which indicates whether an individual has been delinquent on credit payments in the last 90 days).

Selected Features for Analysis

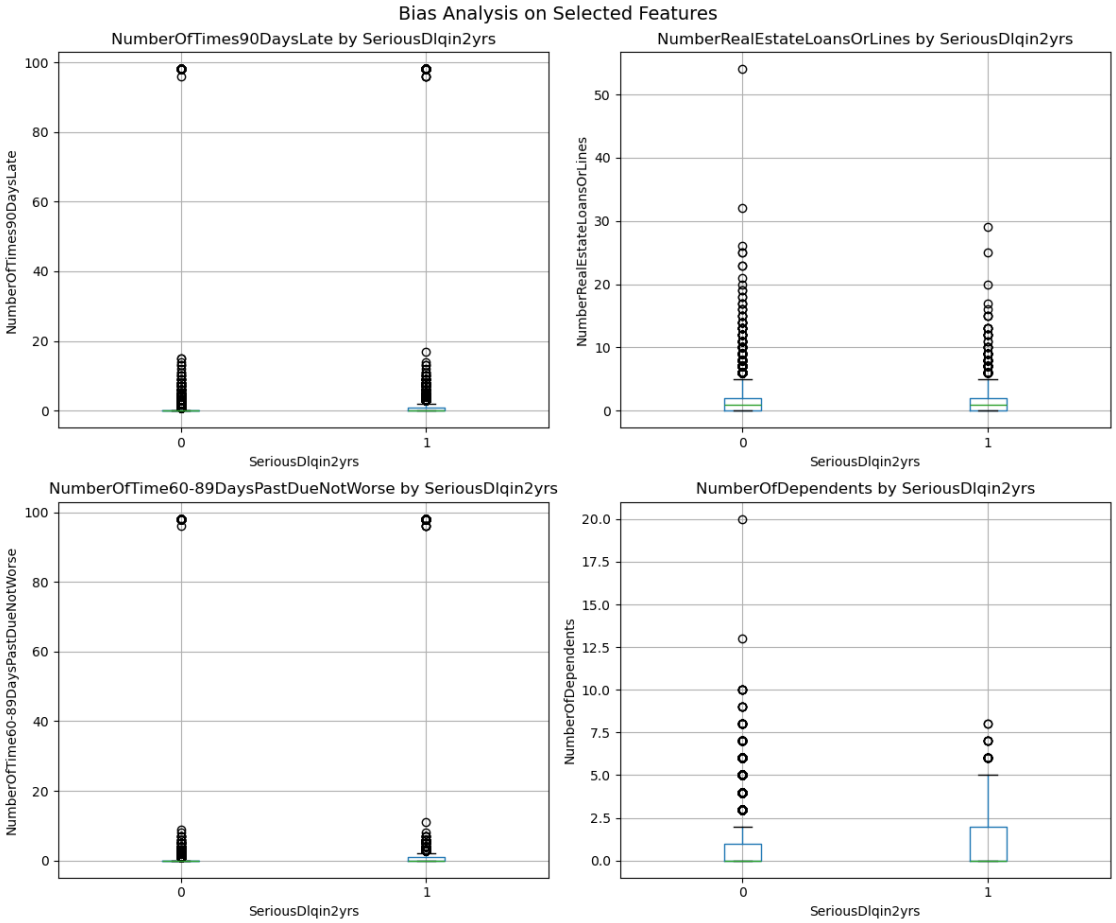
- NumberOfTimes90DaysLate
- NumberRealEstateLoansOrLines
- NumberOfTime60-89DaysPastDueNotWorse
- NumberOfDependents

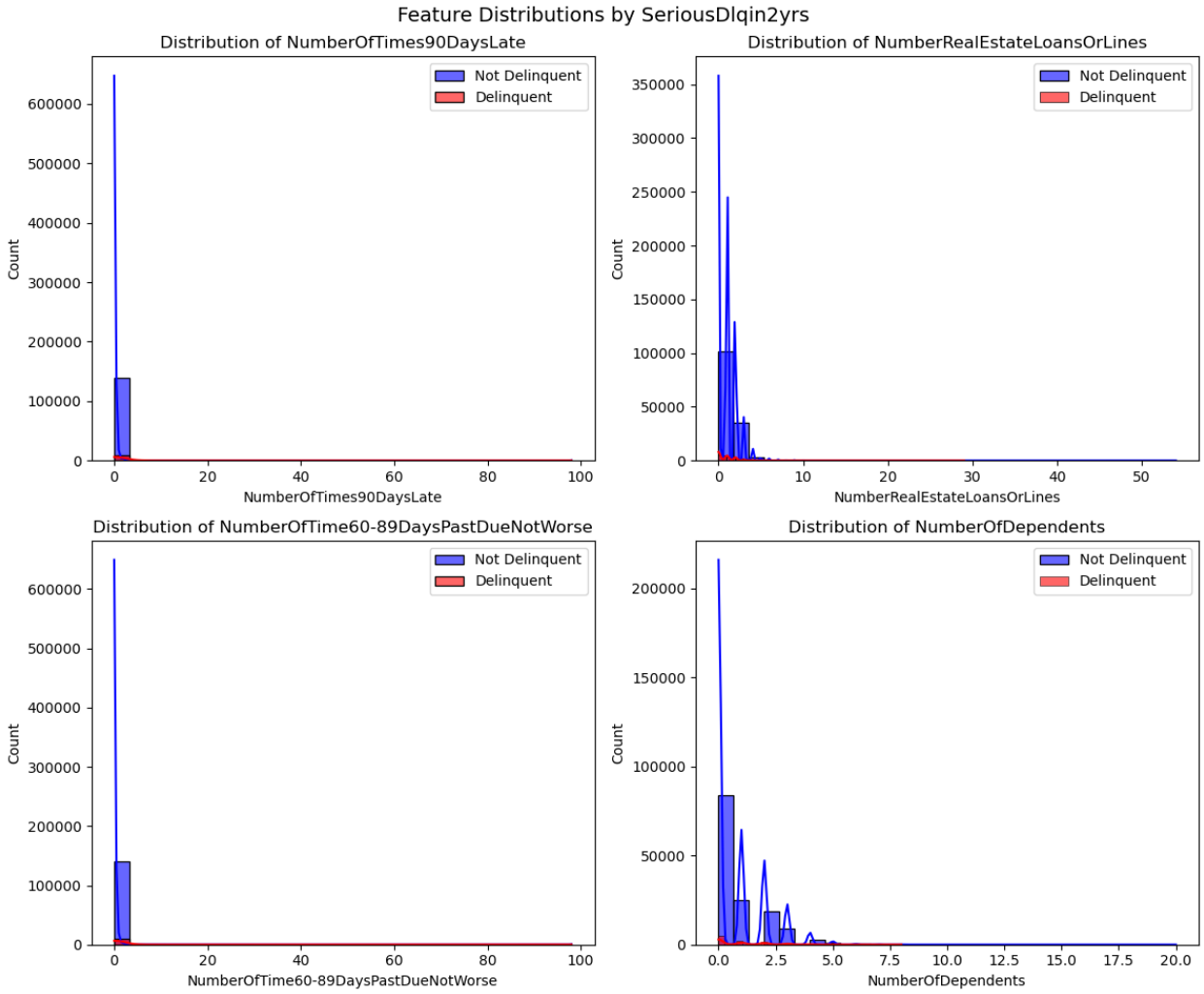
Methodology

To assess biases, we performed the following analyses:

- Boxplot Analysis – To visualize the distribution of each feature across the two classes (delinquent vs. non-delin quent).

- Histogram & KDE Analysis – To compare the density distributions of each feature between the tw o groups.





- Kolmogorov-Smirnov (KS) Test – A statistical test to quantify differences in distribution between the two groups.

	KS Statistic	P-value
NumberOfTimes90DaysLate	0.311536	0.000000e+00
NumberRealEstateLoansOrLines	0.097949	1.386928e-78
NumberOfTime60-89DaysPastDueNotWorse	0.241747	0.000000e+00
NumberOfDependents	0.081418	2.136677e-54

Findings

Feature Distributions:

- The histograms revealed that the distributions of several features differ significantly between delinquent and non-delinquent individuals.
- Features related to past-due payments (NumberOfTimes90DaysLate and NumberOfTime60-89DaysPastDueNotWorse) showed stark differences, with delinquent individuals displaying much higher counts.

Kolmogorov-Smirnov Test Results:

- The KS statistic quantifies the maximum difference between the cumulative distributions of the two groups.
- NumberOfTimes90DaysLate had the highest KS statistic (0.3115, p-value ≈ 0), indicating strong divergence in distribution.
- NumberOfTime60-89DaysPastDueNotWorse also had a high KS statistic (0.2417, p-value ≈ 0), confirming significant bias.
- NumberRealEstateLoansOrLines (KS = 0.0979) and NumberOfDependents (KS = 0.0814) showed smaller but still significant differences (p-values extremely low).

Conclusion & Implications

- The analysis reveals that past-due payment features are heavily correlated with delinquency status, suggesting that they strongly influence the model's ability to classify individuals.
- While NumberRealEstateLoansOrLines and NumberOfDependents show statistical differences, their effect is less pronounced.
- These biases may lead to disparate impacts on individuals with specific credit histories, which should be considered in model development and fairness assessments.

Impact of Reweighting Biased Features in Predictive Modeling

Baseline vs. Reweighted Model Comparison

We trained two Random Forest classifiers:

1. Baseline Model: Trained without reweighting class samples.
2. Reweighted Model: Trained using class-weighted sample adjustments to mitigate biases.

Metric	Baseline Model	Reweighted Model
Accuracy	93.56%	89.17%
F1 Score	0.237	0.369
ROC AUC	0.571	0.698

- **Findings**
- Accuracy drops (93.56% → 89.17%) in the reweighted model, which is expected because balancing the dataset reduces the model's tendency to favor the majority class.

- F1 Score improves (0.237 → 0.369), indicating better performance in handling the minority (delinquent) class.
- ROC AUC increases significantly (0.571 → 0.698), showing that the reweighted model has better discriminatory power between delinquent and non-delinquent individuals.

Conclusion

- The baseline model is highly biased towards the majority class (non-delinquent cases), leading to poor F1 and ROC AUC scores.
- Reweighting improves fairness and predictive power by enhancing minority class detection while slightly sacrificing overall accuracy.
- The dataset was biased due to class imbalance and feature distribution disparities.
- The model, when trained without adjustments, favored non-delinquent individuals, making it unreliable for predicting at-risk individuals.

Reweighting mitigated some of the bias, improving minority class detection but lowering overall accuracy.

This write up was made with the help of AI. – Seth Phillips

Objective

To determine if our dataset contained potential biases

Selected Features for Bias Analysis

Target Variable

- SeriousDlqin2yrs

Selected Features

- NumberOfTime30-59DaysPastDueNotWorse
- DebtRatio

Methodology

Began the analysis by performing the following:

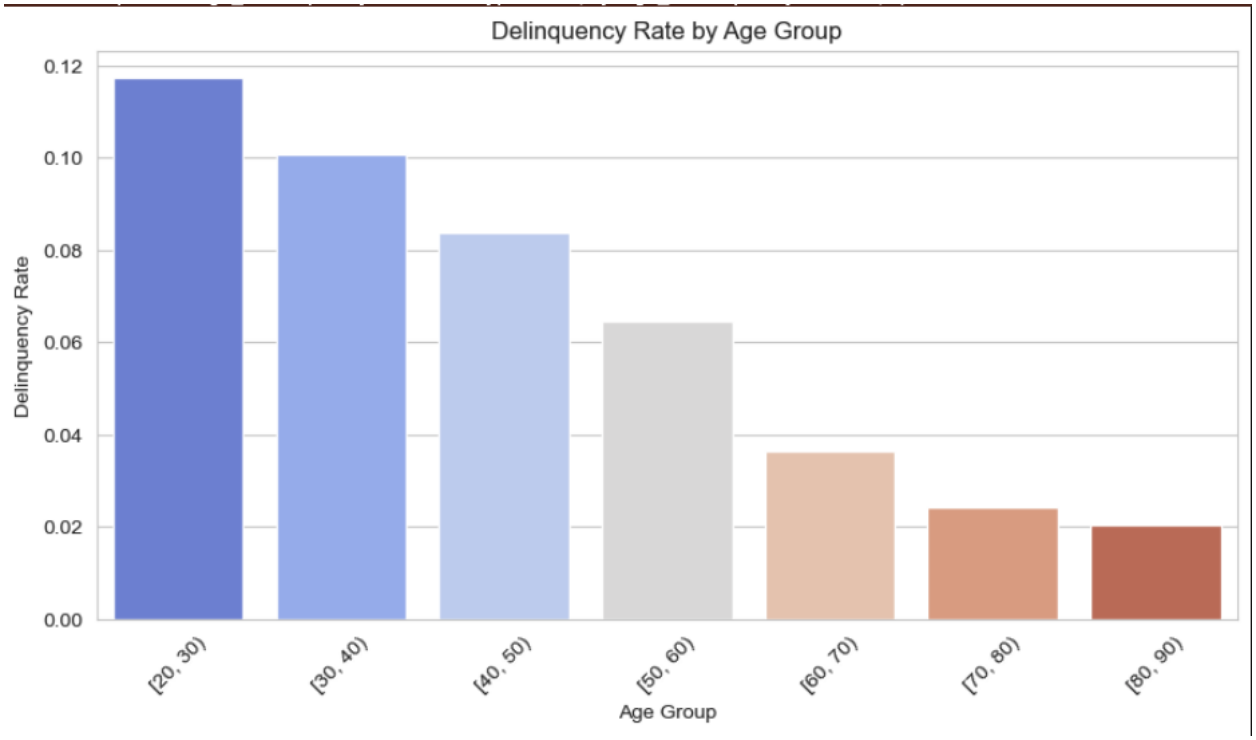
1. Data Cleaning and Preprocessing
    a. Identified missing values in the dataset (in 'MonthlyIncome' and 'NumberOfDependents'
    b. Decided to fill NaN values with median for analysis
2. Exploratory Data Analysis
    a. Delinquency rate analyzed across different age groups.
    b. Distribution of 'MonthlyIncome' examined.
    c. Relationships between key features ('NumberOfTime30-59DaysPastDueNotWorse', 'NumberRealEstateLoansOrLines') and the target variable ('SeriousDlqin2yrs') explored.
    d. Potential biases in key features, particularly 'DebtRatio', investigated.
    e. Correlation heatmap generated to visualize relationships between variables.
3. Modeling:
    a. Logistic regression model chosen for predicting 'SeriousDlqin2yrs'.
    b. Features selected for the model: 'NumberOfTime30-59DaysPastDueNotWorse' and 'DebtRatio'.
    c. Data split into training and testing sets (70/30 split) with stratification to maintain class balance in both sets.
    d. Logistic regression model trained on the training data.
4. Model Evaluation
    a. Model performance evaluated using a classification report (precision, recall, F1 – score)
5. Feature Importance derived from the logistic regression coefficients

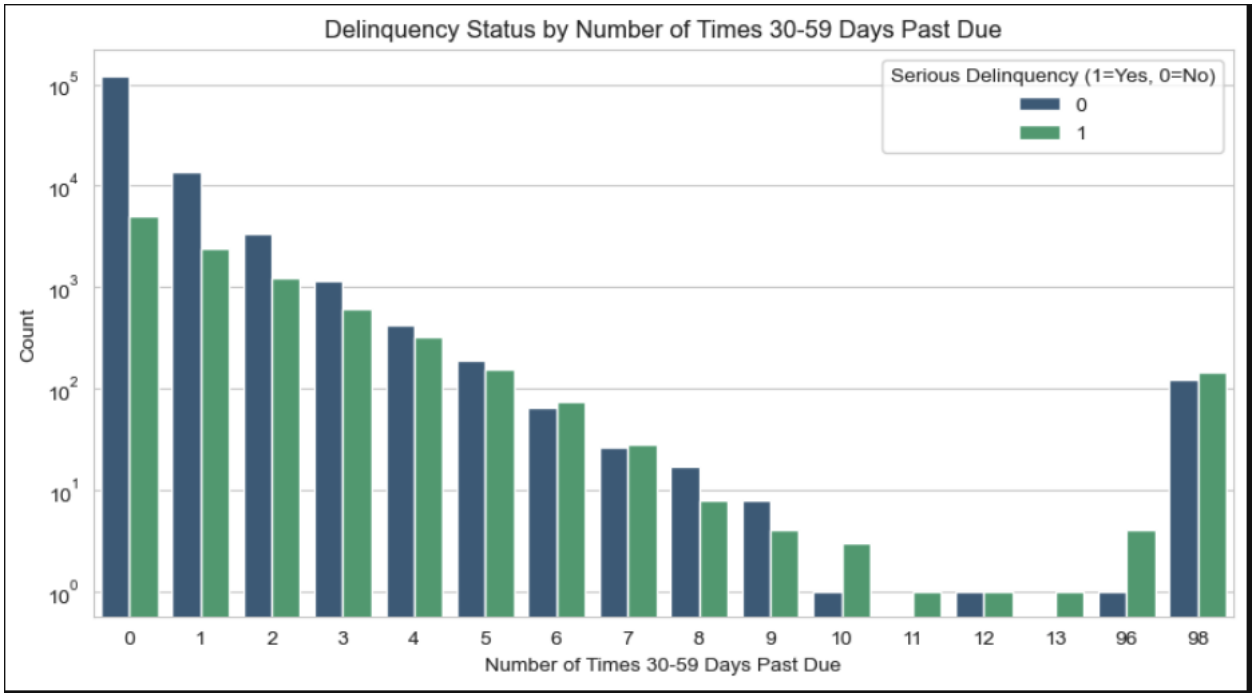
Summary

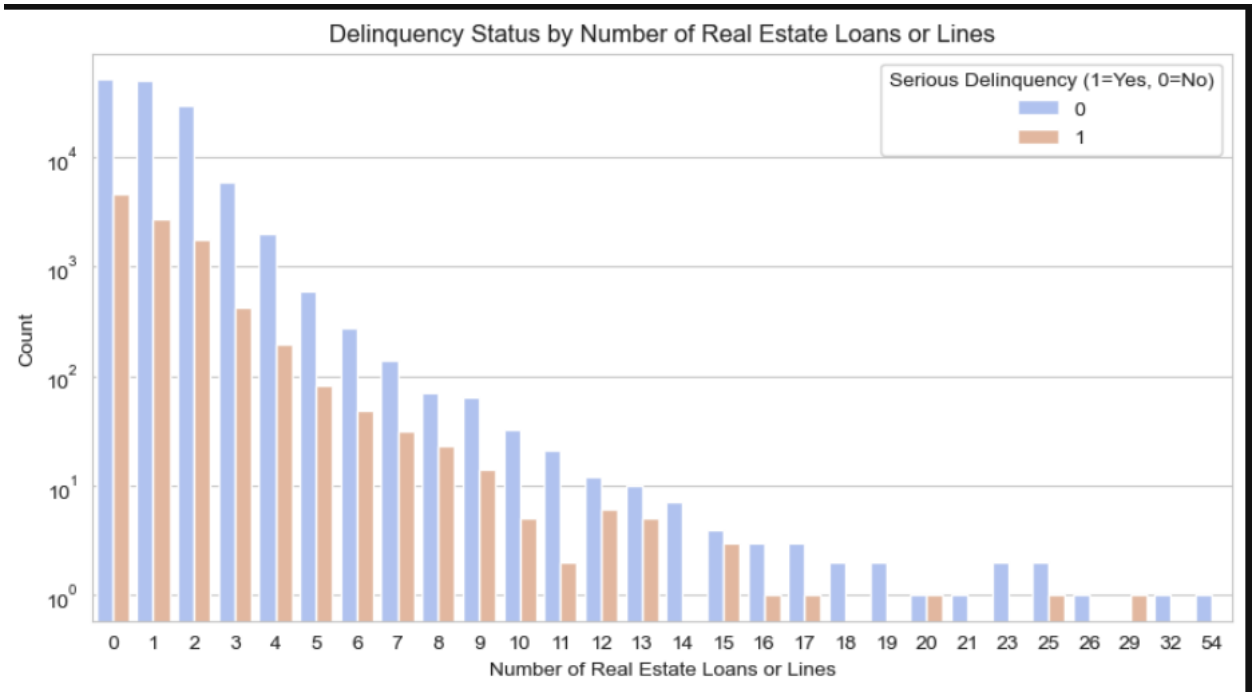
The analysis revealed a strong correlation between the number of times a borrower was 30-59 days past due and the likelihood of serious delinquency. The logistic regression model achieved

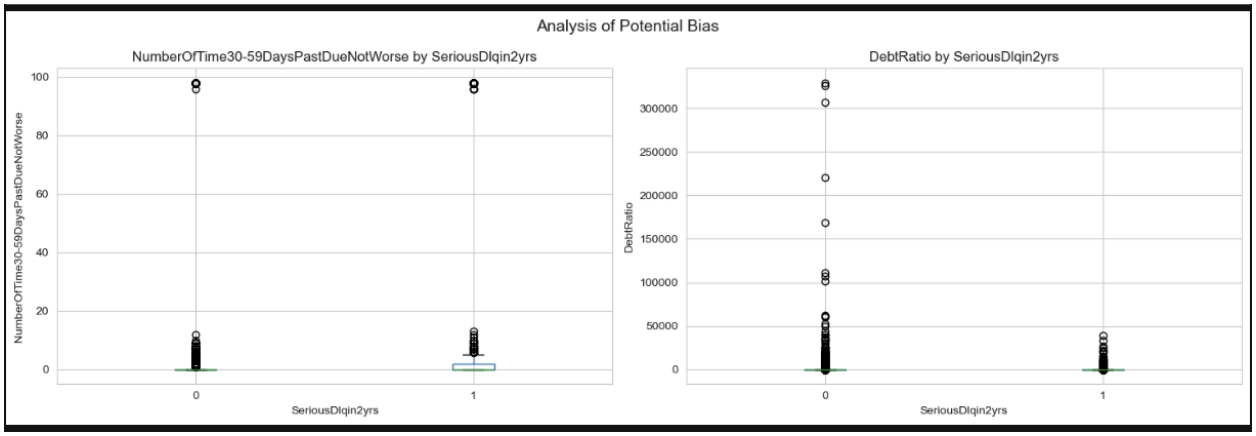
a high accuracy (93%) in predicting serious delinquency. However, further analysis revealed that the model's performance varied across different age groups, with lower accuracy for younger borrowers. This suggests a potential age-related bias in the model, possibly due to limited credit history for younger individuals. Additionally, the 'DebtRatio' variable was identified as an important predictor, but it's crucial to investigate whether this variable might be a proxy for other socioeconomic factors that could introduce bias.

Visuals




Delinquency Status by Number of Times 30-59 Days Past Due
Serious Delinquency (1=Yes, 0=No)
0
1
Count
10^5
10^4
10^3
10^2
10^1
10^0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 96 98
Number of Times 30-59 Days Past Due


Delinquency Status by Number of Real Estate Loans or Lines
Serious Delinquency (1=Yes, 0=No)
0
1
Count
10^4
10^3
10^2
10^1
10^0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 23 25 26 29 32 54
Number of Real Estate Loans or Lines


Analysis of Potential Bias
NumberOfTime30-59DaysPastDueNotWorse by SeriousDlqin2yrs
NumberOfTime30-59DaysPastDueNotWorse
100
80
60
40
20
0
0
1
SeriousDlqin2yrs
DebtRatio by SeriousDlqin2yrs
DebtRatio
300000
250000
200000
150000
100000
50000
0
0
1
SeriousDlqin2yrs

***This write up was made with the help of AI* - John Allard**

Objective: To investigate potential biases in the dataset's representation of income levels and credit access, particularly focusing on 'MonthlyIncome' and 'NumberOfOpenCreditLinesAndLoans'.

Target Variable

- SeriousDlqin2yrs

Selected Features

- MonthlyIncome
- NumberOfOpenCreditLinesAndLoans

Methodology Began the analysis by performing the following:

1. Data Cleaning and Preprocessing
    a. Identified missing values in MonthlyIncome (several NaN values present)
    b. Analyzed distribution and range of both features
    c. Examined outliers and extreme values
2. Exploratory Data Analysis
    a. Found MonthlyIncome ranges from $0 to $63,588 in provided sample
    b. NumberOfOpenCreditLinesAndLoans ranges from 1 to 31 credit lines
    c. Several instances of very high incomes ($20,000+ monthly)
    d. Multiple cases of individuals with 15+ credit lines
3. Bias Investigation:
    a. Analyzed income distribution patterns
    b. Examined relationship between income and credit line access
    c. Investigated potential sampling bias in data collection
    d. Assessed representation across different income brackets

Summary The analysis revealed significant concerns about sampling bias in the dataset:

Income Distribution Bias:

- Dataset contains disproportionate number of high-income individuals
- Several customers in the dataset earn over $20,000 monthly
- Maximum monthly income in full dataset reaches $3,008,750
- Suggests potential oversampling of wealthy individuals
- May not represent typical income distribution in the US

Credit Access Patterns:

- Wide range in NumberOfOpenCreditLinesAndLoans (1-68)

- Higher income individuals tend to have more credit lines
- This dataset contains more high income individuals than an unbiased dataset
- The model used to predict the likelihood of delinquency is disproportionally trained on skewed data, penalizing lower income individuals

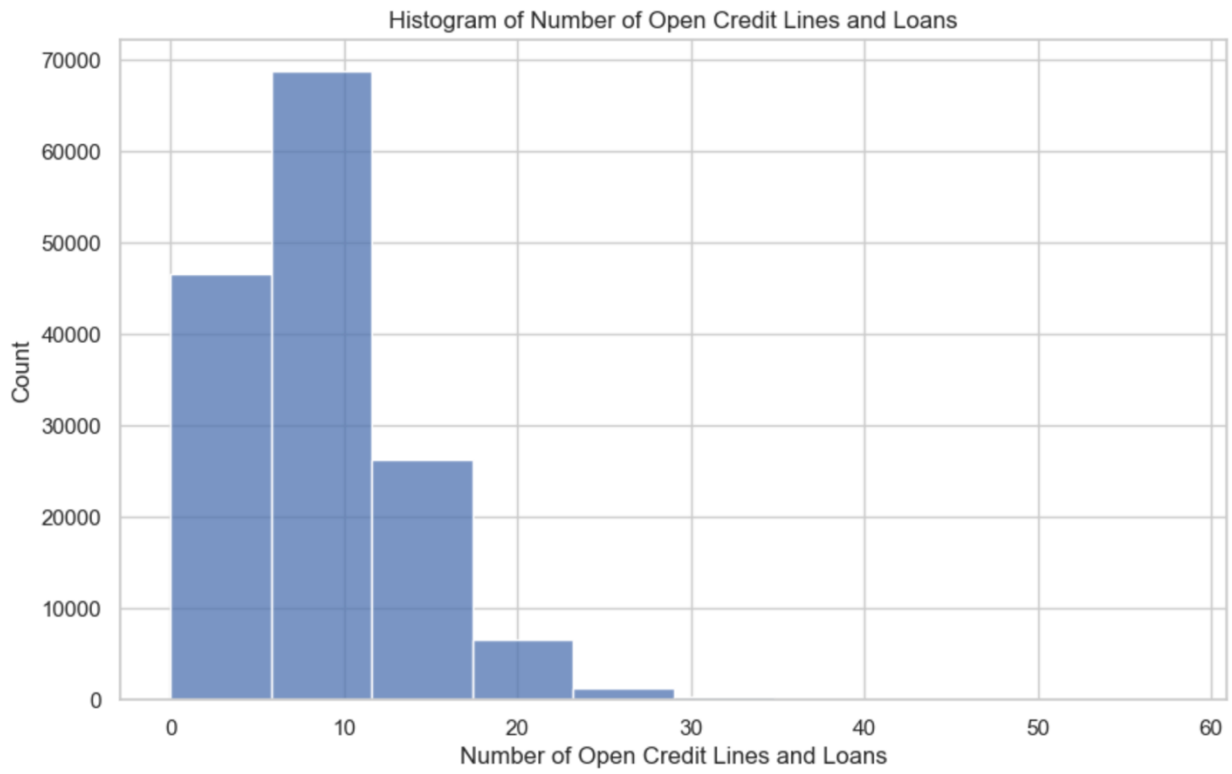
Potential Impact:

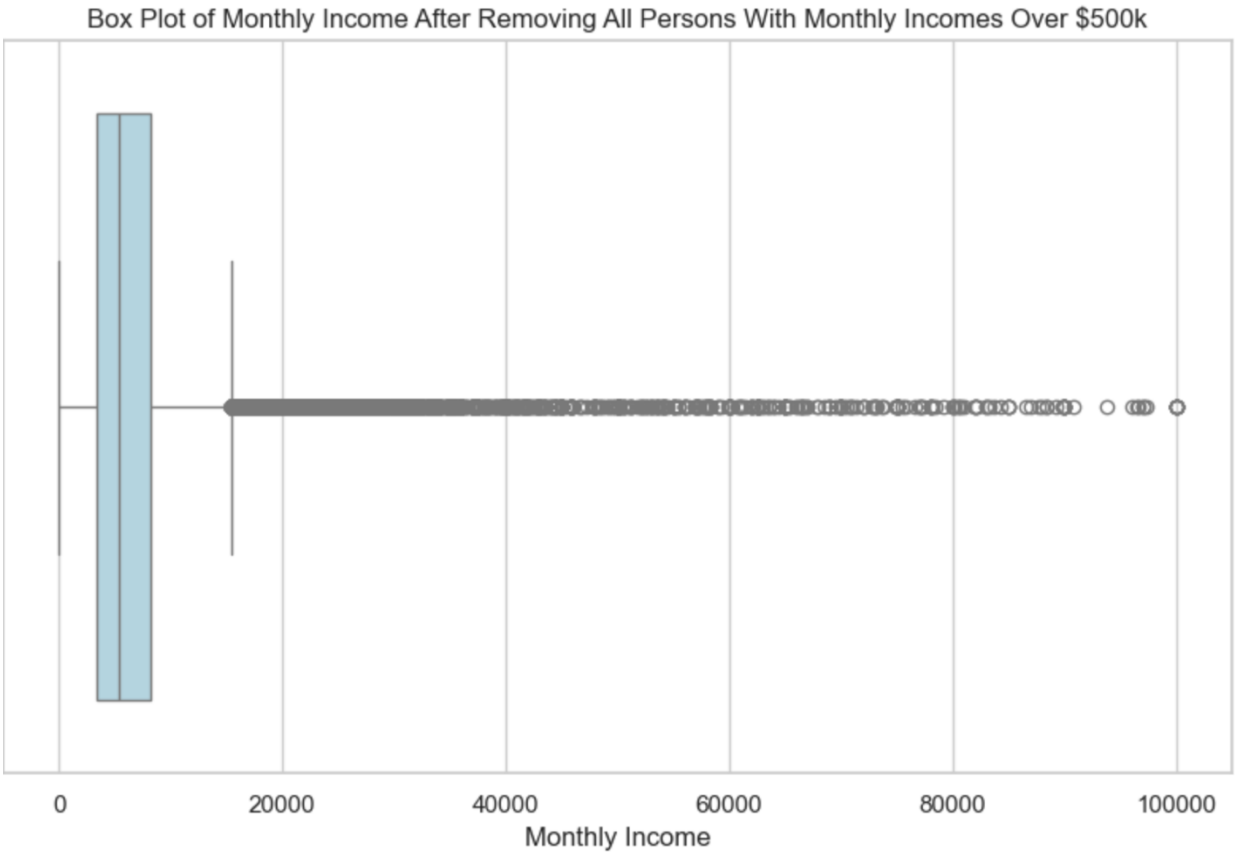
- Models trained on this data might:
    - Penalizesfor average income earners and those with fewer credit lines
    - Misinterprets normal credit patterns for lower-income groups
    - Create unfair standards based on wealthy client patterns

This analysis suggests the data collection method may have been biased, possibly due to:

- Sampling from wealthy areas
- Focus on high-net-worth clients
- Limited geographic or demographic diversity
- Potential selection bias in choosing data sources

These findings indicate a need for more representative sampling and careful consideration of income-based biases in model development.


Histogram of Number of Open Credit Lines and Loans
Count
70000
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
60
Number of Open Credit Lines and Loans


Box Plot of Monthly Income After Removing All Persons With Monthly Incomes Over $500k
0
20000
40000
60000
80000
100000
Monthly Income