Report of CS340 Assignment 2: Bias Mitigation

Yilai Chen 陈驿来 12013025

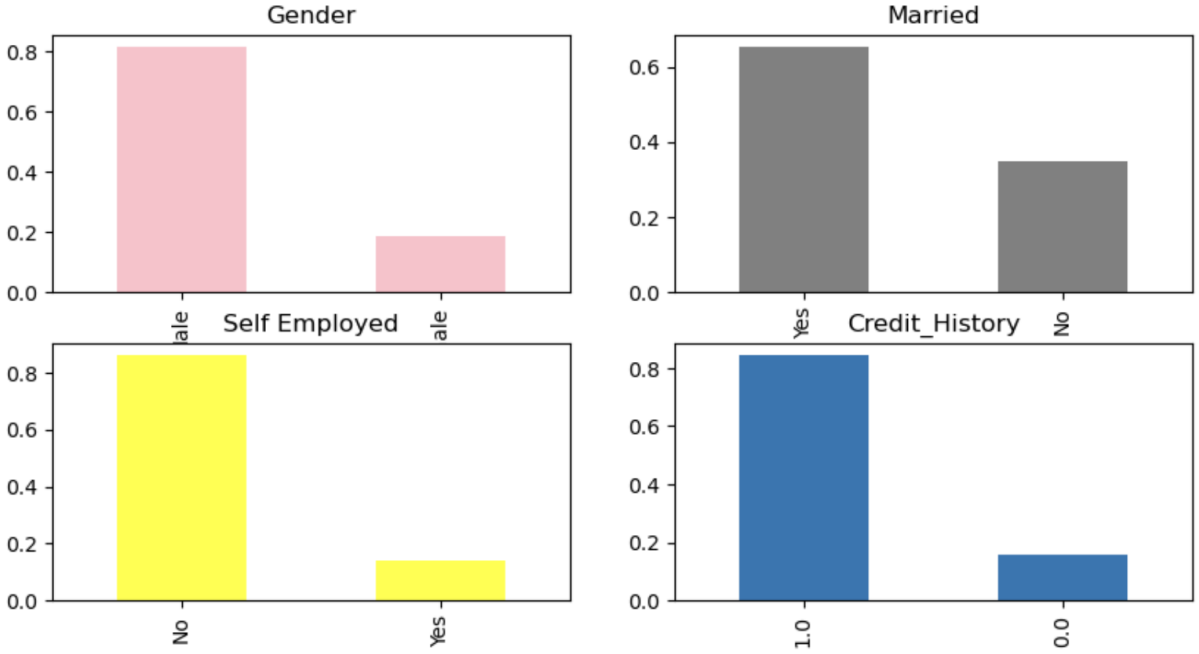
Overview

In this assignment, I built a series of predictive models to predict user loan status based on a provided dataset. Due to severe bias in the data set, if use it directly to train a model, the prediction results may be very different. Therefore, I first assess the degree of bias of the dataset and then train and predict on the dataset. There are two data sets. The training set provides some information about loan applicants and the outcome of their loan application (allowed or denied). The test set provides some information about loan applicants, but not the results of their loan applications. It is necessary to train a classification model on this data and make predictions on the test set data. And compare the bias of the model before and after mitigate.

Statistics and Visualization of Bias in Data Sets

First let's analyze the dataset, understood its biases, and considered how these biases could impact our predictive model.

Visualizing Categorical Variables



- **Gender bias:**

  80% of the data set are men, which shows that the data set has obvious gender bias, and the number of male samples far exceeds the number of female samples.
  This gender bias may cause the model to perform poorly when approving loans for women, as the model may be more inclined to make decisions based on the characteristics of men.

- **Marital status bias:**

  Around 65% of the applicants in the data set were married, indicating a higher number of married individuals.
  This bias could cause the model to favor loan approvals for married applicants and lower loan approval rates for unmarried applicants.

- **Self-employment bias:**

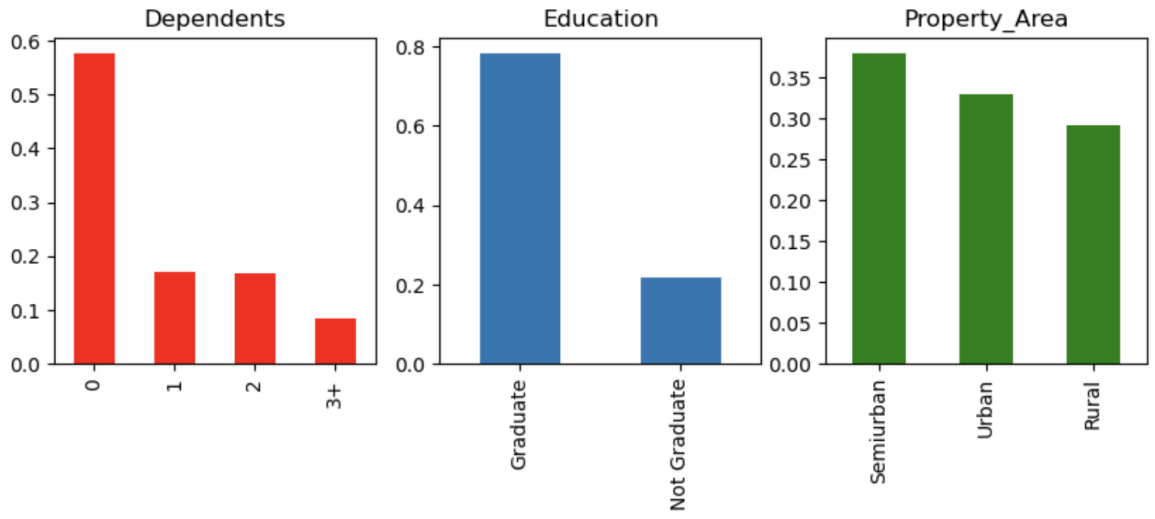
  Around 85% of applicants in the data set are not self-employed, indicating the dominance of the non-self-employed population.
  This bias may cause the model to favor loan approvals for non-self-employed applicants and lower loan approval rates for self-employed applicants.

- **Credit history bias:**

  Around 85% of the applicants in the data set have loan history, which indicates that most applicants have previous loan experience.
  This bias may cause the model to favor loan approvals for applicants with loan history and lower loan approval rates for applicants without loan history.

By analyzing these biases, we can realize that the model may be biased in loan approval for different groups, and measures need to be taken to mitigate the impact of these biases to ensure the fairness and accuracy of the model.

Visualizing Ordinal Variable



- **Number of dependents:**

  Around 58% of applicants in the data set had zero dependents, while 17% had one dependent, 17% had two dependents, and 8% had three or more dependents. The distribution of dependents shows a relatively even picture, although there is a slight bias towards applicants with zero dependents.

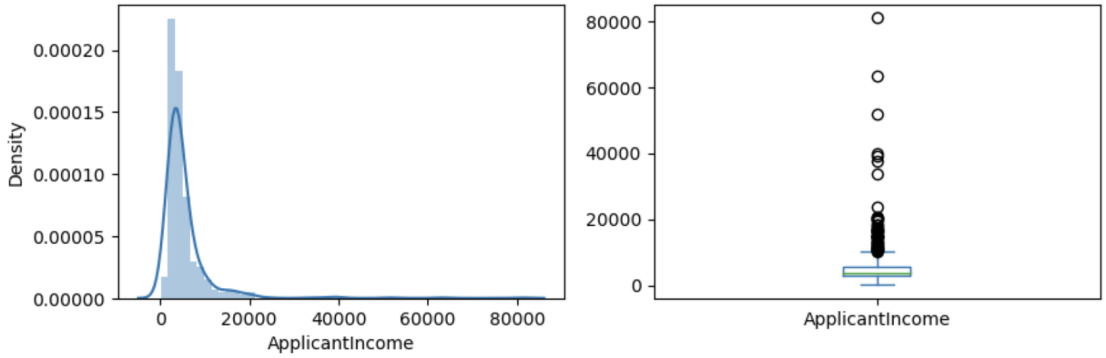
- **Education:**

  Around 78% of the applicants in the data set were educated, while 22% were not educated. Most applicants are educated, but a significant number are uneducated.

- **Property Area:**

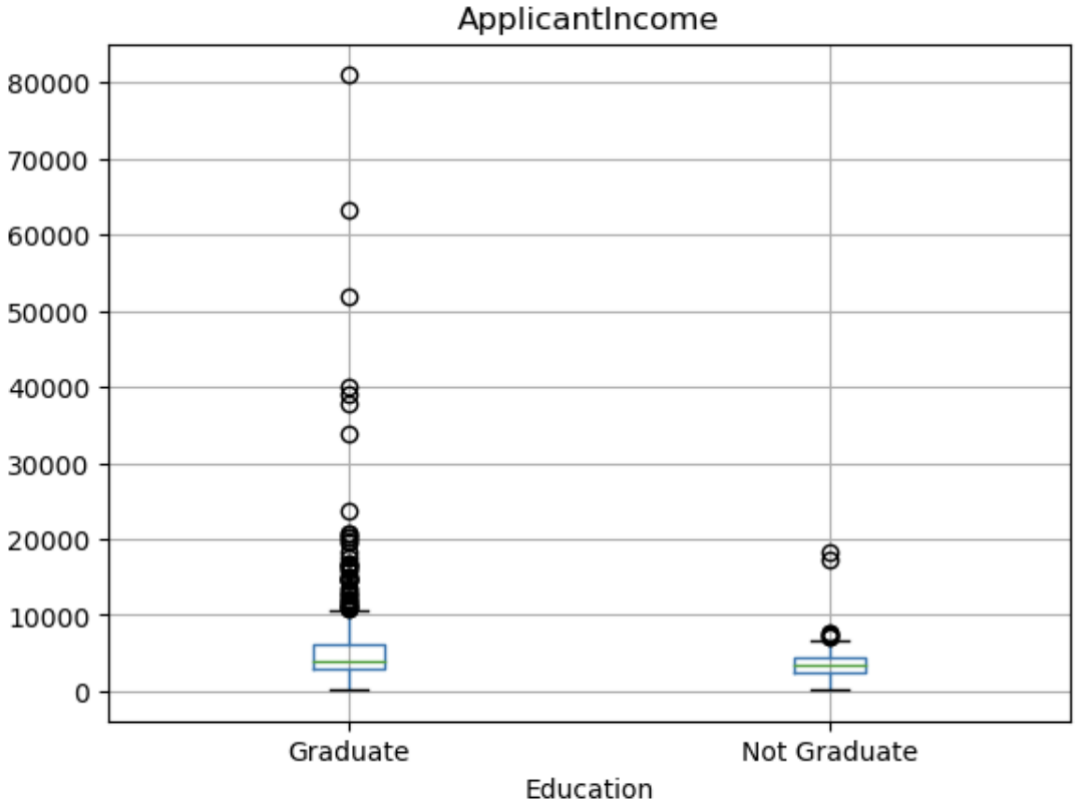
  The largest property area in the data set is semi-urban, accounting for 39%; followed by urban, accounting for 33%; and finally rural, accounting for 28%. The distribution of property areas shows some diversity, but semi-urban areas have the largest number of applicants.

Visualizing Numeric Variables and Checking for Outliers



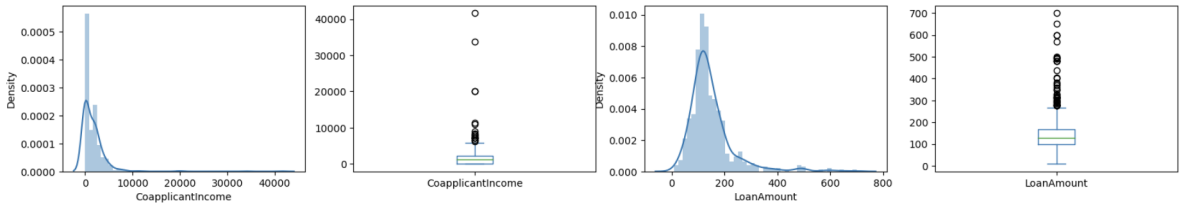
- **Applicant Income:**

  Due to the high degree of left skew, that is, most applicants have lower incomes, but there are also a small number of people with higher incomes, which may have an impact on the predictions of the model. Higher-income applicants may tend to be approved for a loan, while lower-income applicants may face a higher risk of rejection. Therefore, special attention needs to be paid to possible outliers and appropriate measures are taken to deal with them to mitigate their impact on the model.



- Most outliers in applicant income occur among educated people, while fewer outliers occur among noneducated people. This may mean that educated applicants generally have higher or more unstable incomes, leading to more anomalies. This situation may affect the model's prediction results, as high or unstable income may affect the loan approval decision. Therefore, during the modeling process, special attention needs to be paid to educated

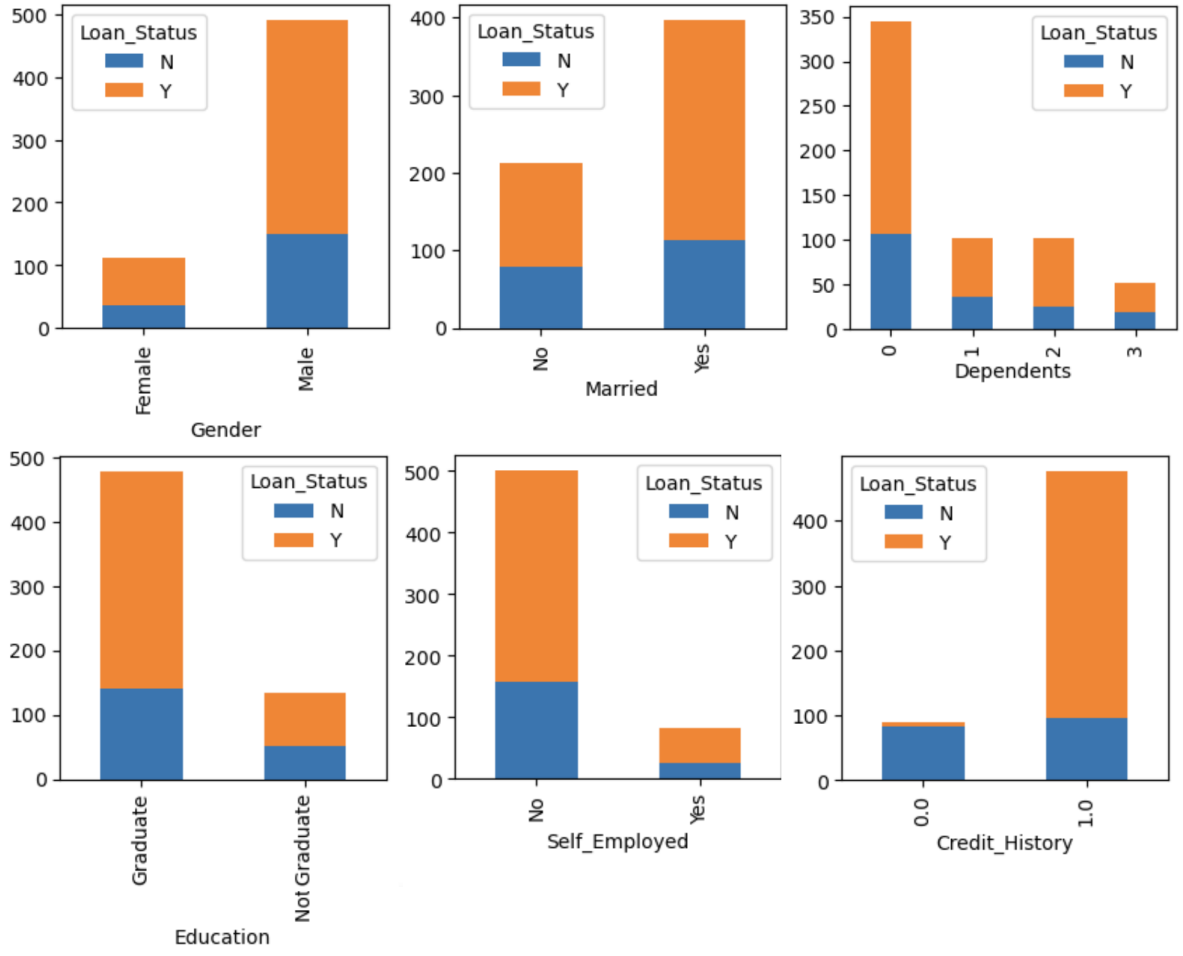
applicants and appropriate measures to handle possible anomalies to ensure the accuracy and fairness of the model.



- Coapplicant Income: Again, a high degree of skewness indicates that most coapplicants have lower incomes, but there is also a small minority of coapplicants with higher incomes. This may result in unfair judgment on the outcome of the loan application, as the income of the co-applicant may affect the loan approval decision. Possible outliers need to be identified and dealt with to ensure the robustness and fairness of the model.
- Loan Amount: A slight skew may indicate a relatively even distribution of loan amounts, but there are still possible outliers to be aware of. Outliers in loan amounts may influence loan approval decisions, as larger loan amounts may increase the risk of default. Therefore, appropriate outlier handling is required to ensure that the model can accurately predict loan approval outcomes.

In summary, for these skewed features, outlier processing is required to mitigate their impact on model predictions, ensure that the model can accurately predict loan approval results, and avoid bias.

Bivariate Analysis



- **The impact of Gender on Loan Eligibility:**

  A similar proportion of 500 men and 100 women qualified for a loan, indicating that there is no clear bias in terms of gender. This means that the model may not be significantly biased with respect to gender.

- **The impact of marital status on loan qualifications:**

  The proportion of 400 married people who are eligible for loans is significantly smaller than the proportion of 200 unmarried people who are eligible for loans. This may imply that the model has a certain bias in terms of marital status and may be more inclined to provide loans to unmarried people.

- **The impact of the number of dependents on loan qualifications:**

  The number of dependents has little impact on whether you are eligible for a loan. This suggests that the model may not be too biased with respect to the number of dependents.

- **The impact of education level on loan eligibility:**

  The proportion of people who qualified for loans was similar among the 480 educated and 120 uneducated people. This suggests that the model may not be too biased in terms of educational level.

- **The impact of employment status on loan eligibility:**

  The proportion of people who qualify for a loan is about the same among the 500 non-self-employed and the 100 self-employed people. This suggests that the model may not be too biased in terms of employment status.

- **The impact of credit history on loan qualifications:**

  The most obvious thing is that 400 of the 510 people with credit history were qualified for loans, but almost none of those without credit history were qualified for loans. This suggests that the model may be biased towards providing loans to people with a credit history and may be reluctant to provide loans to people without a credit history.

From the above visualization its clear that mostly graduates, self-employed and applicants with 0 dependents have high chances of loan approval Applicants whose credit history is 0 have very low chances of getting loan.

Correlation of Dataset

In this part I use heat-map to analyze the correlation between each two features:

- **Co-Applicant Income and Race (0.62):**

  The high correlation (0.62) may mean that the sensitive attribute Race may be related to Co-Applicant Income during model training, which may lead to an increase in the model's sensitivity to Race when making loan decisions.

  The model may be biased towards giving loans to people in certain Race categories and biased against people in other Race categories.

  The model may need to be adjusted to ensure it is fair and accurate in terms of Race.

- **Loan Amount and Race(0.5):**

  The correlation is 0.5, which may mean that Race has some influence on the loan amount.

  The model may tend to size loan amounts based on Race categories, which may lead to bias in the loan approval process.

  Models need to be noted and adjusted to ensure that they are not affected by Race when deciding loan amounts.

- **Loan Amount and Applicant Income(0.57):**

  The correlation is 0.57, indicating that applicant income may have a larger impact on loan amounts.

  If the model relies primarily on the applicant's income during training to determine loan amounts, it may introduce bias in loan approvals, especially when income is correlated with Race.

  There is a need to ensure that models take into account other factors in addition to an applicant’s income when deciding loan amounts to reduce potential bias.

- **Loan Status and Credit History(0.56):**

  The correlation is 0.56, indicating that credit history may have a larger impact on loan status.

  If the model relies heavily on credit history to predict loan status during training, it may lead to bias against applicants with no credit history or poor credit history.

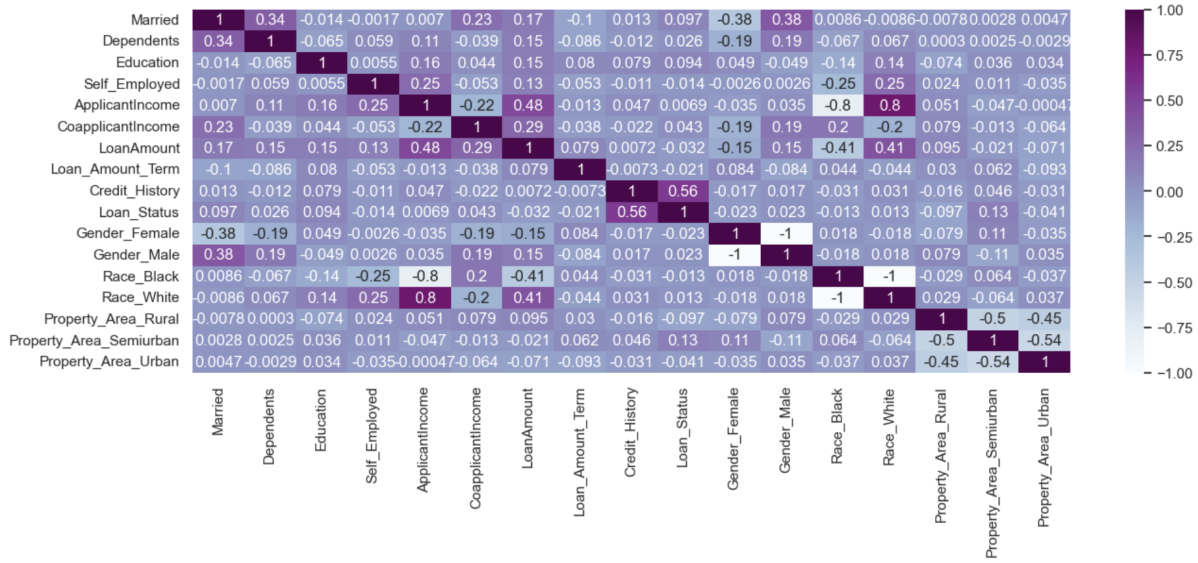
  There is a need to ensure that models consider a variety of factors when determining loan status to reduce possible bias, particularly with regard to credit history.

Train a Predictive Model

Preprocessing

Before officially starting to train the model, some additional missing data must be added.
First, for the five data types of Applicant Income, Co-applicant Income, LoanAmount, Loan_Amount_Term, and Credit_History, we take the mode of the existing corresponding data to fill in the missing part. We also do the same operation for objects such as Loan_ID, Gender, Race, Married, Dependents, Education, Self_Employed, Property_Area, and Loan_Status.
And, we Use get dummies for the remaining object columns for which mapping or encoder cant be

used.



By the way, after performing one-hot encoding, I found that Gender and Race have a very important impact on the judgment results. This will cause a very serious bias phenomenon, and measures need to be taken to eliminate it in the future.

Using Logistic Regression Supervised ML Classification Model

We then evaluate the model's performance using three pieces of data that provide different perspectives on the model's performance:

- Accuracy: refers to the ratio of the number of samples predicted correctly by the model to the total number of samples. It is one of the most intuitive evaluation indicators. In this model, the accuracy is 0.80, which means that the model's prediction accuracy on the test set is 80%.
- Confusion Matrix: Provides a detailed comparison between the model's predicted results and actual results for each category. Among them, the diagonal line of the matrix represents the number of samples predicted correctly by the model, while the elements on the off-diagonal represent the number of samples predicted incorrectly by the model. In this example, you can see the prediction effect of the model on category 0 and category 1, including True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
21	25	4	98

- Classification Report: Provides precision, recall and F1-score for each category, as well as weighted average and macro average indicators. The precision rate represents the proportion of predicted positive cases that are actually positive cases, the recall rate represents the proportion of actual positive cases that are correctly predicted as positive cases, and the F1 value comprehensively considers the precision rate and recall rate.

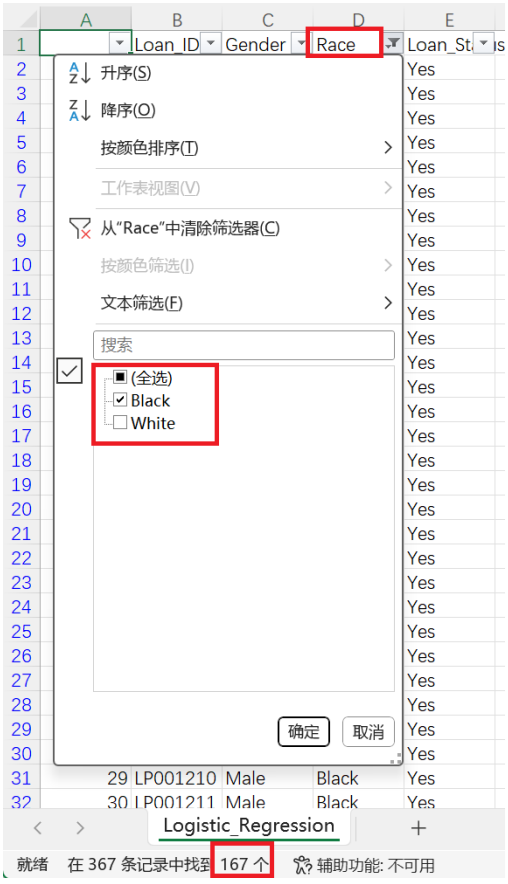
	precision	recall	f1-score	support
0	0.84	0.46	0.59	46
1	0.80	0.96	0.87	102

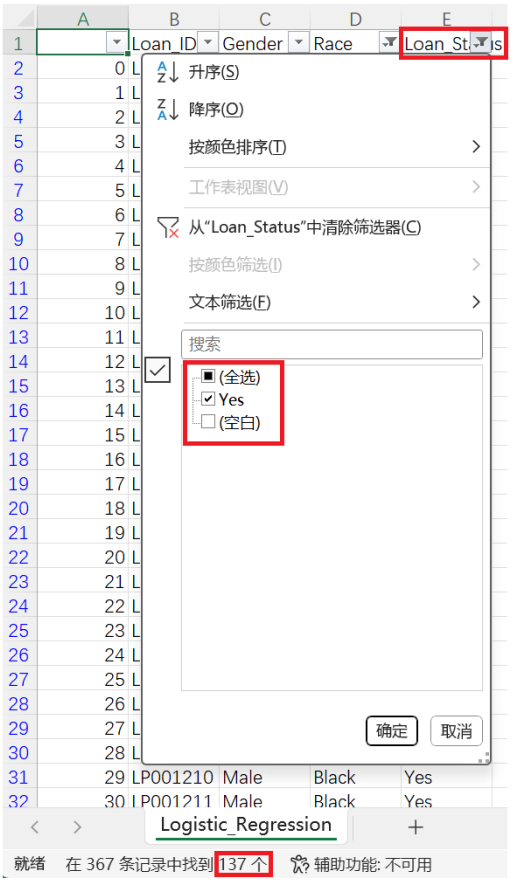
	precision	recall	f1-score	support
accuracy			0.80	148
macro avg	0.82	0.71	0.73	148
weighted avg	0.81	0.80	0.78	148

We then evaluate model bias after the model has processed on the test file using demographic parity.

Race

- Black: Pr(Y'=1 | A=0) = 137 / 167 = 0.8204
- White: Pr(Y'=1 | A=1) = 143 / 200 = 0.7150





Synthesis to a Ratio	Value
Difference	0.1054
Ratio	0.8715

Gender

- Male: Pr(Y'=1 | A=2) = 219 / 286 = 0.7657
- Female: Pr(Y'=1 | A=3) = 51 / 70 = 0.7286

Synthesis to a Ratio	Value
Difference	0.0371
Ratio	0.9516

Based on the data provided, we can calculate the bias (bias) of the linear regression model's predicted results with respect to race and gender. Here we use different sensitive characteristics (race and gender) to calculate the proportion of each group that is predicted to be in the positive category (to qualify for a loan).
It can be observed from these proportions that the proportion predicted to be in the positive category differs across racial and gender groups, suggesting that the model's predictions are biased along race and gender lines. For example, in the black group, the proportion predicted to be in the positive category is higher, while in the white group, the proportion predicted to be in the positive category is lower.

Comprehensive analysis shows that the linear regression model has prediction bias for different racial and gender groups, that is, the model is more inclined to make unfair predictions for certain racial and gender groups. This bias may affect the fairness and accuracy of the model, so corresponding measures need to be taken to reduce the bias and improve the fairness and reliability of the model.

Mitigating model bias by expanding data based on the original data set.

The specific method is to randomly select half of the original data, and for these pieces of data, generate three other pieces of data that are the same, with only gender and race being different combinations.

10	LP001018	Male	Black	Yes	2	Graduate	No	4006	1526	168	360	1 Urban	Y
11	LP001018	Female	Black	Yes	2	Graduate	No	4006	1526	168	360	1 Urban	Y
12	LP001018	Male	White	Yes	2	Graduate	No	4006	1526	168	360	1 Urban	Y
13	LP001018	Female	White	Yes	2	Graduate	No	4006	1526	168	360	1 Urban	Y

Also,

- A new feature in the test set, Total-Income, was calculated, representing the applicant's total income, including the applicant's own income and the income of co-applicants.
- A new feature in the test set, EMI, was calculated, representing the monthly loan amount that needs to be repaid, i.e., the loan amount divided by the loan term.
- A new feature in the test set, Balance-Income, was calculated, representing the applicant's remaining monthly income, which is the total income minus the monthly loan amount that needs to be repaid.

Then, the original features Applicant-Income, Co-applicant-Income, Loan_Amount_Term and LoanAmount were removed from the training and test sets since new features have been calculated to replace them. Doing so simplifies the dataset and helps the model understand the data better.

We then re-train and re-evaluate model bias after the Bias Mitigation of data

Race

- Black: Pr(Y'=1 | A=0) = 138 / 167 = 0.826
- White: Pr(Y'=1 | A=1) = 141 / 200 = 0.705

Synthesis to a Ratio	Value
Difference	0.121
Ratio	0.854

Gender

- Male: Pr(Y'=1 | A=2) = 218 / 286 = 0.762
- Female: Pr(Y'=1 | A=3) = 51 / 70 = 0.739

Synthesis to a Ratio	Value
Difference	0.0409
Ratio	0.963

After bias mitigation, this could result in less disparity if demographic characteristics of gender and race (such as the number of males and females, the number of blacks and whites) were more balanced across categories. Bias mitigation methods may make models treat different demographic characteristics more fairly, thereby reducing variability in predictions. In addition, bias mitigation methods may also enhance the model's ability to generalize to the overall data, thereby reducing the variability between different classes. But the effect is not as good as expected, so I will try other models next.

Using Decision Tree Classifier Model

We then evaluate model bias after the model has processed on the test file using demographic parity.

Race

- Black: Pr(Y'=1 | A=0) = 138 / 167 = 0.826
- White: Pr(Y'=1 | A=1) = 141 / 200 = 0.705

Synthesis to a Ratio	Value
Difference	0.121
Ratio	0.854

Gender

- Male: Pr(Y'=1 | A=2) = 107 / 152 = 0.704
- Female: Pr(Y'=1 | A=3) = 27 / 40 = 0.675

Synthesis to a Ratio	Value
Difference	0.029
Ratio	0.959

The demographic parity results of Decision Tree Classifier before Bias Mitigation are better than those of Logistic Regression, mainly in terms of gender. The proportion of men and women is closer. This may be due to the following reasons:

- Decision trees are a nonlinear model that may be better able to capture complex relationships in the data than Logistic Regression.
- Decision trees can automatically learn the interaction and combination between features during the training process, while Logistic Regression requires manual selection of features and assumes a linear relationship between them.

- Decision trees have relatively little impact on outliers during training because they do not directly affect the partitioning process of the decision tree. In contrast, Logistic Regression is more sensitive to outliers and may be biased due to the presence of outliers.
- Decision trees can output feature importance scores to help us understand how the model makes predictions.

We then re-train and re-evaluate model bias after the Bias Mitigation of data

Race

- Black: Pr(Y'=1 | A=0) = 138 / 167 = 0.826
- White: Pr(Y'=1 | A=1) = 143 / 200 = 0.715

Synthesis to a Ratio	Value
Difference	0.111
Ratio	0.866

Gender

- Male: Pr(Y'=1 | A=2) = 215 / 286 = 0.752
- Female: Pr(Y'=1 | A=3) = 54 / 70 = 0.771

Synthesis to a Ratio	Value
Difference	0.019
Ratio	0.975

In the Decision Tree Classifier model, the demographic parity results of Race and Gender before Bias Mitigation are greatly improved, which means that during the model training process, the decision tree model learns the patterns and features in the data more effectively, and in the process of analyzing these When dividing features, different Race and Gender are treated more fairly. This shows that our above-mentioned Bias Mitigation measures have a certain effect on Decision Tree Classifier.

Using Random Forest Classifier Model

We then evaluate model bias after the model has processed on the test file using demographic parity.

Race

- Black: Pr(Y'=1 | A=0) = 139 / 167 = 0.832
- White: Pr(Y'=1 | A=1) = 138 / 200 = 0.690

Synthesis to a Ratio	Value
Difference	0.242
Ratio	0.829

Gender

- Male: Pr(Y'=1 | A=2) = 218 / 286 = 0.762
- Female: Pr(Y'=1 | A=3) = 50 / 70 = 0.714

Synthesis to a Ratio	Value
Difference	0.048
Ratio	0.937

The demographic parity results of the Random Forest Classifier model before Bias Mitigation are worse than those of the Decision Tree Classifier Model and Logistic Regression. There may be the following reasons:

- Random Forest is an ensemble learning algorithm that contains multiple decision tree models and trains each decision tree by randomly selecting features and samples.
- Random Forest is prone to over-fitting when dealing with high-dimensional data, especially when there is a lot of noise or redundant features in the training set.
- Some hyper-parameters in Random Forest (such as the number of trees, depth of trees, minimum sample split per tree) have a great impact on the performance of the model. If these hyper-parameters are not properly tuned and optimized, the performance of the model may degrade, resulting in poor demographic parity results.
- Random Forest is less sensitive to class-imbalanced data, but if the data is extremely imbalanced, it may affect the performance of the model. If there are far fewer classes in a sample than other classes, the model may tend to predict classes that occur more frequently, leading to biased results.

We then re-train and re-evaluate model bias after the Bias Mitigation of data

Race

- Black: Pr(Y'=1 | A=0) = 139 / 167 = 0.832
- White: Pr(Y'=1 | A=1) = 147 / 200 = 0.735

Synthesis to a Ratio	Value
Difference	0.097
Ratio	0.883

Gender

- Male: Pr(Y'=1 | A=2) = 224 / 286 = 0.783
- Female: Pr(Y'=1 | A=3) = 52 / 70 = 0.742

Synthesis to a Ratio	Value
Difference	0.041
Ratio	0.948

In the Random Forest Classifier model, if the demographic parity results of Race and Gender have improved to a certain extent before Bias Mitigation, it may be due to the influence of feature selection and sampling: The Random Forest model will be used during the training process of each decision tree. Features and samples are randomly selected for training. This randomness helps reduce the model's dependence on specific features, thereby mitigating potential bias. And the streamlining and refining of features in Bias Mitigation also plays a big role.

Using XGB Classifier Model

We then evaluate model bias after the model has processed on the test file using demographic parity.

Race

- Black: Pr(Y'=1 | A=0) = 125 / 167 = 0.749
- White: Pr(Y'=1 | A=1) = 129 / 200 = 0.645

Synthesis to a Ratio	Value
Difference	0.104
Ratio	0.861

Gender

- Male: Pr(Y'=1 | A=2) = 198 / 286 = 0.602
- Female: Pr(Y'=1 | A=3) = 48 / 70 = 0.686

Synthesis to a Ratio	Value
Difference	0.084
Ratio	0.878

In the XGB Classifier Model's demographic parity results before Bias Mitigation, Race is better than both Decision Tree Classifier Model and Logistic Regression, but the gender data is worse for the following reasons:

- XGB is an ensemble learning algorithm that is generally more complex than Decision Tree and Logistic Regression. Because XGBoost is able to capture complex relationships between features, XGBoost may perform better when dealing with features like Race that have a greater impact on model predictions.
- If Race's data distribution is relatively balanced, then XGBoost may be more accurate in predictions. However, if the data distribution of gender is unbalanced, for example, the number of male samples is much larger than the number of female samples, then the model may be biased in predicting female-related outcomes.
- There are many parameters that need to be adjusted in XGBoost, such as learning rate, tree depth, leaf node weights, etc. If these parameters are not adjusted appropriately, it may cause the model to perform poorly on certain features.
- XGBoost can provide an importance score for each feature, thereby helping us understand how much the model pays attention to different features. If the Race feature has a higher importance score in the XGBoost model, the model's prediction of Race may be more accurate.

We then re-train and re-evaluate model bias after the Bias Mitigation of data

Race

- Black: Pr(Y'=1 | A=0) = 126 / 167 = 0.755
- White: Pr(Y'=1 | A=1) = 129 / 200 = 0.645

Synthesis to a Ratio	Value
Difference	0.110
Ratio	0.854

Gender

- Male: Pr(Y'=1 | A=2) = 202 / 286 = 0.706
- Female: Pr(Y'=1 | A=3) = 45 / 70 = 0.643

Synthesis to a Ratio	Value
Difference	0.063
Ratio	0.911

In the XGB Classifier model, if the Gender data has improved to a certain extent but the Race has not improved much before Bias Mitigation, it may be due to the difference in feature importance: In the XGB Classifier model, due to the gradient boosting method, the model will Partition the data based on the importance of features. It may be that the Gender feature has a higher importance in the model, so the model pays more attention to this feature, thereby achieving improvements in Gender data. And this may be because XGB Classifier is not suitable as a training model for this data set. In comparison, you can try the previous decision tree and random forest models, which have achieved better results under Bias Mitigation.