Quantifying Gender Bias in Income Prediction: A Comparative Analysis of Regression Models

Kaitlin Day, Nia Hodges and Yuqi Yan

May 7, 2024

Abstract :

Gender disparities in income remain a pressing issue, with significant implications for social justice and economic equality Pew Research Center, 2023. This study focuses on quantifying the impact of gender on income prediction by comparing the predictive quality of popular regression models when gender data is or isn't available. Using a dataset containing demographic, educational, and occupational information, we employ regression analysis to predict income levels. Two sets of models are trained: one with gender data included and one without. The predictive performance of each model is evaluated using metrics such as mean squared error, R-squared, and bias-variance trade-off analysis. Our findings reveal the extent to which gender information influences predictive accuracy and highlight potential biases in income. By systematically comparing the performance of regression models with and without gender data, this study provides insights into the role of gender in income estimation and offers valuable implications for addressing gender disparities in the workforce. Our research contributes to the ongoing dialogue on gender equity and underscores the importance of fair and unbiased predictive modeling practices in tackling gender-based wage inequalities.

Index Terms: gender bias, income prediction, regression analysis, predictive modeling, economic equality

1 Executive Summary

1.1 Decisons to be Impacted

- Research and development strategies
- Policy decisions
- Organizational practices
- Individual career decisions

1.2 Business Value

This project has the potential to enhance organizational performance by enabling more accurate and equitable decision-making processes related to income prediction. By identifying and mitigating gender bias in predictive modeling, businesses can ensure fairer compensation practices, which can lead to higher employee satisfaction, increased productivity, and improved retention rates. Additionally, addressing gender-based wage disparities aligns with corporate social responsibility initiatives and can enhance the organization's reputation as a fair and inclusive employer.

1.3 Data Assets

This project addresses the persistent gender disparities in income levels, aiming to provide empirical evidence and insights to inform efforts to address gender-based wage inequalities and promote gender equity in economic outcomes. By comparing the performance of regression models with and without gender data, the research seeks to elucidate the extent to which gender influences income prediction and identify potential biases in predictive modeling practices.

2 Data Preprocessing

2.1 Data Description

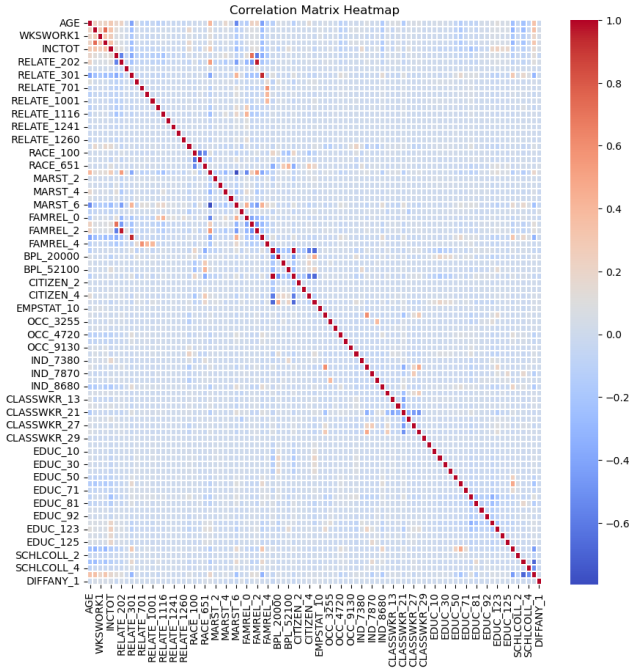
The dataset uses data published by the U.S. Census Bureau on the basic monthly Current Population Survey (CPS). The Current Population Survey (CPS) is a survey of U.S. residents conducted by the U.S. government at regular intervals to understand labor market conditions and changes. The survey covers a variety of data on employment, unemployment, income, and family status, among other things.

This data is important to government organizations, academics, researchers, and businesses due to the insights into the U.S. labor market. It helps inform policy, research, and business decisions.

2.2 Data Cleaning and Outlier Detection

Our original dataset contained 39 columns (features) and 851,273 rows (instances). Through data cleaning and preprocessing, we reduced our dataset to 20 columns and 586,091 rows. The data pre-processing steps are outlined below:

A. Data cleaning: Remove rows containing "NIU" (Not in Universe) values (removed 265,182 rows)
B. Features dropped manually:
    a. **CPSID**, **SERIAL**, **CPSIDV**, and **CPSIDP** are unique identifiers which would mislead our results.
    b. **MONTH** provided no new information due to the census being taken during the same month every year.
    c. **ASECFLAG** and **HFLAG** are related to census sampling methods, not helpful for research.
    d. **ASECWTH** is a family specific weight, and **ASECWT** is a annual social and economic supplement weight, not respondent specific.
    e. **PERNUM** is a person number in sample unit, not helpful for research.
    f. **WKSWORK2** and **WKSUNEM2** are the interval versions of **WKSWORK1** and **WKSUNEM1**, so no new information.
    g. **WKSUNEM1**'s "NIU" is greater than 37%.
    h. **ADJGINC** and **TAXINC** are income variables; we are predicting income using **INCTOT**, so these features would mislead our prediction results.
    i. **NCHILD** and **NCHLT5**'s missing data exceed 50% of the total.



C. One-Hot Encoding: convert categorical variables into numerical by encoding each category into binary 1s (true) and 0s (false) (1172 columns)
D. Dimensionality Reduction: reduce columns from 1172 to 91
    a. **OCC**, **IND**, **BPL**, and **RACE** encoded columns are kept only if there are more than 10,000 of a category.
    b. Delete columns where there are 2 unique values (represented as binary).
    c. Delete column with 1 unique value (**LABFORCE**).

We are using z-score to detect numerical column outliers. After implementing the z-score, 555,865 rows remained in our dataset.

2.3 Correlation Matrix

After processing the data, calculate the correlation between the indicators to generate a correlation matrix. The heat map generated from the correlation matrix shows that red indicates a high correlation and blue indicates a low correlation. From the heat map, we can see that the correlation between the indicators is not strong, which means these indicators can be used for further analysis.

3 Model Updates

Our dataset is high in dimensionality (90 Features) and high in volume (555,865 data). Due to the assumption in the project that gender has an impact on income, we believe that both linear models and neural networks will be the best models to prove or disprove our assumption. Through parallel training, we can do a few things:

- compare model performance when gender is included or excluded as an input feature
- find the optimal weights
- identify the corresponding weights through the gender column
- determine the extent of the impact of gender on income by the positive or negative values

We also know that linear regression is very sensitive to outliers, even if we have removed outliers, we still hope that the model is not so sensitive. Therefore, we have decided to implement ridge regression, lasso regression, and a neural network. At this moment, we have successfully implemented all three models listed.

Lasso Regression: Lasso Regression is a feature selection method that can help identify the most significant features influencing income. Therefore, it can be used to determine the extent of gender's impact on income while excluding other irrelevant features.

Ridge Regression: Ridge Regression is robust against noise and outliers in the data, reducing their influence on model parameter estimates. This can lead to a more stable assessment of gender's impact on income.

Neural Network: Neural networks have powerful nonlinear modeling capabilities to effectively handle complex data patterns and relationships. They can capture higher-order and interaction effects of gender on income.

3.1 Hyperparameters

The dataset is split into a training set and a test set. Various hyperparameters, such as alpha values, are trained and evaluated using cross-validation methods, with 5-fold cross-validation being utilized in this project. The best-performing hyperparameters are then selected to construct the model. Subsequently, the model's performance is assessed using Mean Squared Error (MSE) and Mean Absolute Error (MAE). Adjustments are made when updating model weights to optimize the final regression model for data prediction and analysis.

- Ridge regression
    - alpha = 10
    - fit_intercept = True
    - max_iter = 1000
    - solver = sparse_cg
- Lasso regression
    - alpha = 0.01
    - fit_intercept = False
    - max_iter = 1000
- Simple neural network
    - Input layer size = 89/90 depending on the inclusion or exclusion of the gender feature
    - Number of hidden layers = 1
    - Number of nodes in hidden layer(s) = 10
    - Output layer size = 1
    - Activation function = sigmoid

3.2 Performance Metrics

Mean Square Error (MSE) is a measure of the squared average of the difference between the model's predictions and the true value. It is more sensitive to large errors. Mean Absolute Error (MAE)

measures the average of the absolute values of the differences between the predicted and true values and is more robust to all errors equally.

Version	Model	MSE	MAE
Gender	Lasso	1110197315.4845	11809.8585
	Ridge	**1110187736.0106**	**22809.1819**
	Simple NN	4735728500.236	53243.524
No Gender	Lasso	1138758788.0072	23113.3676
	Ridge	1138749225.9848	23112.573
	Simple NN	4735728500.236	53243.524

Although the performance metrics are not great, error is lower for each model when gender is included as an input feature.

3.3 Feature Weights

After analyzing feature weights for both the Lasso and Ridge regression models, we found that the gender feature was the eleventh most important feature out of ninety input features.

Weights from **Lasso** regression model:

	feature	weight	original_feature
0	Feature 4	56385.356515	UHRSWORKLY
1	Feature 83	48224.021871	EDUC_124
2	Feature 84	45613.719018	EDUC_125
3	Feature 82	27042.281227	EDUC_123
4	Feature 3	26443.523961	WKSWORK1
5	Feature 1	22006.338135	AGE
6	Feature 55	21711.780762	IND_7380
7	Feature 85	20247.286915	SCHLCOLL_1
8	Feature 2	15726.366829	UHRSWORKT
9	Feature 81	12038.198835	EDUC_111
10	Feature 21	11761.036690	SEX_1
11	Feature 65	11606.507760	CLASSWKR_25
12	Feature 40	10340.672992	BPL_52100
13	Feature 47	9338.788656	OCC_430
14	Feature 63	8086.939174	CLASSWKR_14

Weights from **Ridge** regression model:

	feature	weight	original_feature
0	Feature 4	56259.232420	UHRSWORKLY
1	Feature 83	51736.723321	EDUC_124
2	Feature 84	49138.283155	EDUC_125
3	Feature 82	30627.606409	EDUC_123
4	Feature 3	26439.039218	WKSWORK1
5	Feature 1	22005.176979	AGE
6	Feature 55	21702.989243	IND_7380
7	Feature 2	15802.238445	UHRSWORKT
8	Feature 81	15630.554061	EDUC_111
9	Feature 85	13948.058406	SCHLCOLL_1
10	Feature 21	11763.488719	SEX_1
11	Feature 65	11104.904308	CLASSWKR_25
12	Feature 40	10328.441788	BPL_52100
13	Feature 8	9651.429428	RELATE_203
14	Feature 7	9520.622770	RELATE_202

3.4 Machine Learning Morphisms

For Lasso Regression:

$$\mathcal{ML} = \begin{pmatrix} \text{InputSpace} : X = R^{555865\times 90} \\ \text{OutputSpace} : R \\ \text{LearningMorphism} : X \cdot p \\ \text{Parameter} : p \in R^{90} \\ \text{EmpiricalRiskFunction} : \frac{1}{555865}\sum_{i=1}^{555865}(y_i - X_i \cdot p)^2 + \alpha \sum_{j=1}^{90} |p_j| \end{pmatrix}$$

For Ridge Regression:

$$\mathcal{ML} = \begin{pmatrix} \text{InputSpace} : X = R^{555865\times 90} \\ \text{OutputSpace} : R \\ \text{LearningMorphism} : X \cdot p \\ \text{Parameter} : p \in R^{90} \\ \text{EmpiricalRiskFunction} : \frac{1}{555865}\sum_{i=1}^{555865}(y_i - X_i \cdot p)^2 + \alpha \sum_{j=1}^{90} \|p_j\|_2^2 \end{pmatrix}$$

For Neural Network:

$$\mathcal{ML} = \begin{pmatrix} \text{InputSpace} : X = R^{555865*90} \\ \text{OutputSpace} : R \\ \text{LearningMorphism} : \text{F} = F_k(F_{k-1}(F_{k-2}(...F_1(X)))); F_k = W_k X + b_k \\ \text{Parameters} : Weight in R^{90}; W_1 \in R^{10*90}, b_1 \in R^{10}, W2 \in R^{1*10}, b_2 \in R \\ \text{EmpiricalRiskFunction} : MeanSquareError : \frac{1}{555865}\sum_{i=1}^{555865}(y_i - X_i \cdot p)^2 + \alpha \sum_{j=1}^{90} \|p_j\|_2^2 \\ AndCrossEntropy : \frac{-1}{555865}\sum_{i=1}^{555865} y_i log(F(x_i)) - (1-y_i) log(1-F(x_i)) \end{pmatrix}$$

4 Additional Research via Clustering

This article utilizes the HDBSCAN clustering algorithm, which is a density-based clustering method particularly suited for high-dimensional and large datasets. One of its main advantages is

that it does not require pre-specifying the number of clusters to be generated; instead, it determines the number of clusters based on the density distribution of the data, making it highly effective for datasets with complex structures.

Focusing on data from 2023, the study involves 46,707 data points. It includes cross-validation primarily involving systematic adjustments and tests on two parameters: the minimum cluster size (`min_cluster_size`) and the minimum number of samples (`min_samples`). Specific settings are as follows: `min_cluster_size` varied from 5 to 20, in increments of 5, i.e., [5, 10, 15, 20]; `min_samples` ranged from 3 to 9, in increments of 3, i.e., [3, 6, 9]. A total of 12 different parameter combination experiments were conducted to find the optimal clustering setup. After each clustering, the silhouette score was calculated to assess the quality of the clusters, indicating higher similarity within clusters and lower similarity between different clusters, thus ensuring good distinction in the clustering results. The analysis found the optimal settings to be: `min_cluster_size` of 15, `min_samples` of 9, silhouette score of 0.78639, and 670 clusters.

4.1 Occupation Distribution of top 6 clusters

Subsequently, the top six clusters were selected for further analysis based on the number of points in each cluster, with pie charts illustrating the proportion of occupations. The occupations include OCC_2310: Elementary and middle school teachers, OCC_3255: Registered nurses, OCC_4700: First-Line supervisors of retail sales workers, OCC_4720: Cashiers, OCC_9130: Driver/sales workers and truck drivers, OCC_4760: Retail salespersons.

From the graphs, we observe the following:

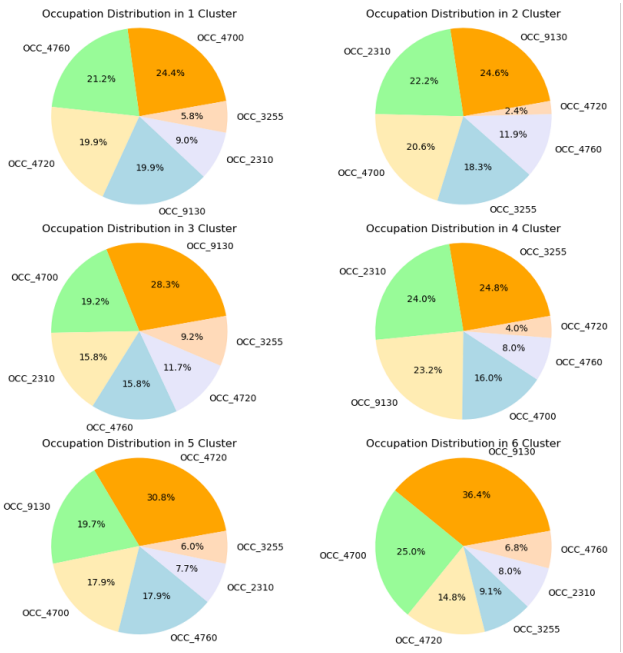
- Cluster 1: First-line supervisors of retail sales workers and Retail salespersons together account for about 50%.
- Cluster 2: Driver/sales workers and truck drivers, and Elementary and middle school teachers together account for about 50%.
- Cluster 3: Driver/sales workers and truck drivers and First-Line supervisors of retail sales workers together account for about 50%.
- Cluster 4: Registered nurses and Elementary and middle school teachers together account for about 50%.
- Cluster 5: Cashiers and Drivers/sales workers and truck drivers together account for about 50%.
- Cluster 6: Driver/sales workers and truck drivers and First-Line supervisors of retail sales workers together account for about 50%.

Next, further analysis is conducted on the positions that have the highest proportions in these six clusters, including the distribution of male and female ratios. In 2023, the position of First-Line supervisors of retail sales workers has a higher proportion of females. In Driver/sales workers and truck drivers, males have a higher proportion. In Registered nurses and Cashiers, females have a higher proportion.

4.2 Income Analysis by Gender in the same Occupation

Subsequently, we analyzed the average income of females and males in these positions:

- In Cluster 1 First-Line supervisors of retail sales workers, the average income for females is $30,002.07, and for males, it is $30,000.75.



- In Cluster 2 Driver/sales workers and truck drivers, the average income for females is $50,000.00, and for males, it is $50,007.00.
- In Cluster 3 Driver/sales workers and truck drivers, the average income for females is $40,006.40, and for males, it is $40,003.62.
- In Cluster 4 Registered nurses, the average income for females is $20,000.56, and for males, it is $20,003.167.
- In Cluster 5 Cashiers, the average income for females is $20,000.56, and for males, it is $20,003.17.
- Driver/sales workers and truck drivers, the average income for females is $35,006.00, and for males, it is $35,001.1.

Combining the data results, in 2023, in the fields of First-Line supervisors of retail sales workers, Driver/sales workers and truck drivers, Registered nurses, and Cashiers, there is little difference in income between men and women.

5 Source Code

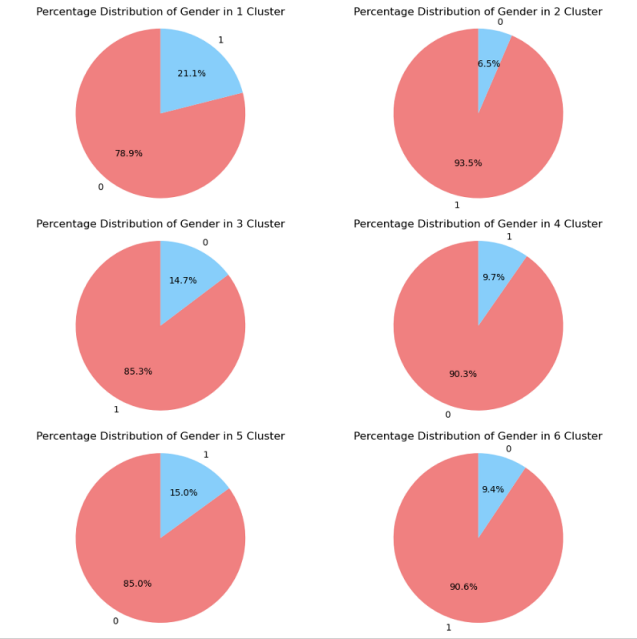
Click here to access our GitHub project repository.

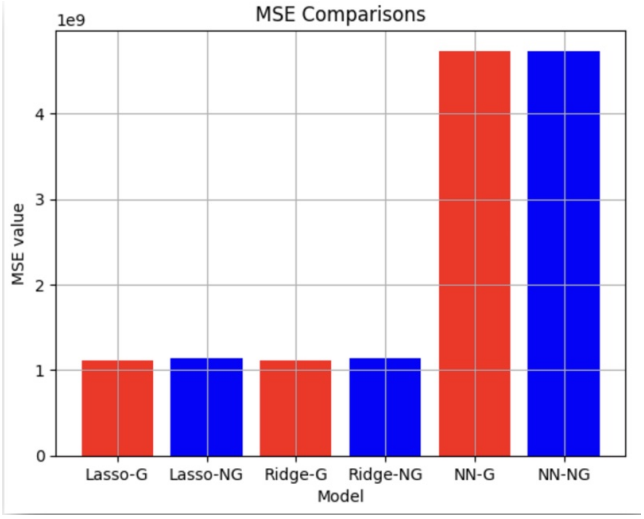
6 Results

In this section, we present and analyze the results obtained from our study, focusing on two primary aspects: predictive model performance and inference results.

6.1 Model Performance

In our investigation, we observed intriguing patterns regarding the predictive performance of various models when gender information was integrated as an input feature. Notably, the Lasso, Ridge, and simple neural network architectures exhibited enhanced performance metrics when gender was included in the feature set. Across different evaluation criteria, such as mean squared error and mean absolute error, these models consistently demon-

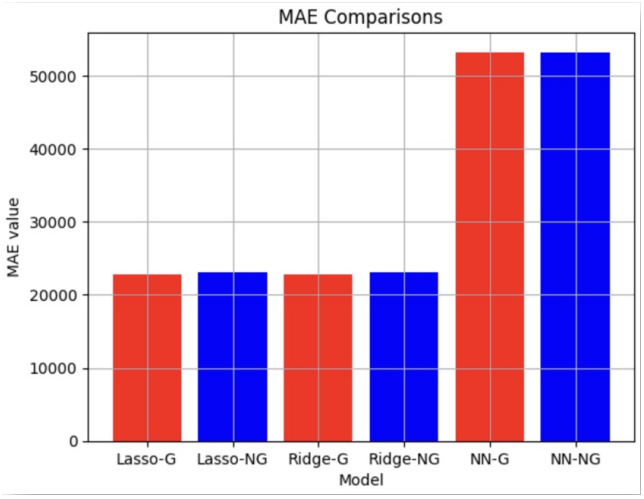


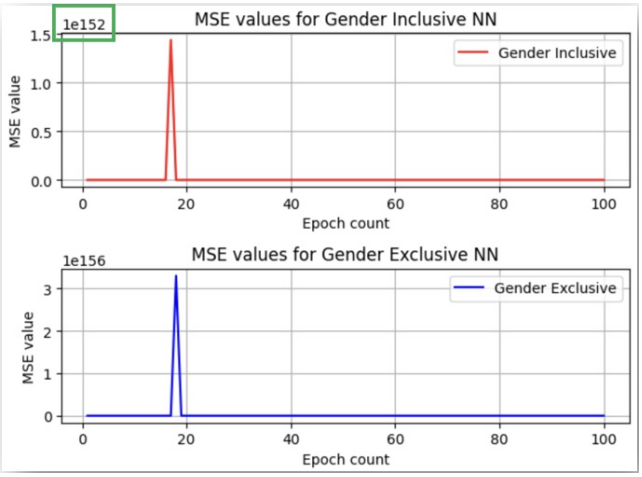


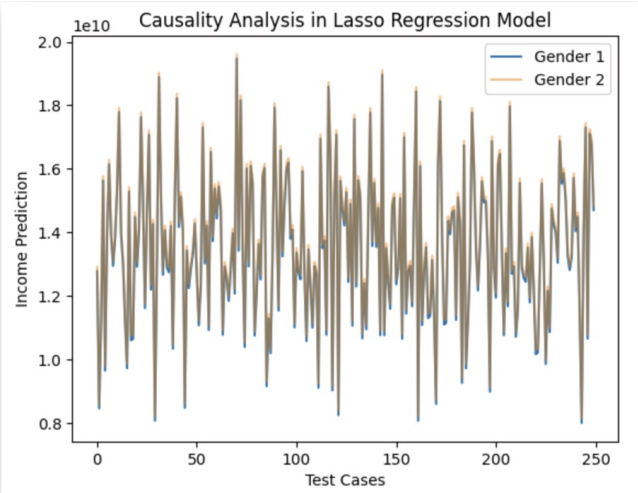
strated lower error rates, indicative of their heightened accuracy in income prediction tasks. This observation underscores the significance of gender as a discernible factor influencing the models' comprehension of the intricate relationships between input features and the resultant income predictions. Our findings shed light on the crucial impact of gender on fine-tuning predictive outcomes, revealing the intricate relationship between demographic attributes and the effectiveness of predictive models. This exploration provides valuable insights into the nuanced dynamics underlying income prediction frameworks.

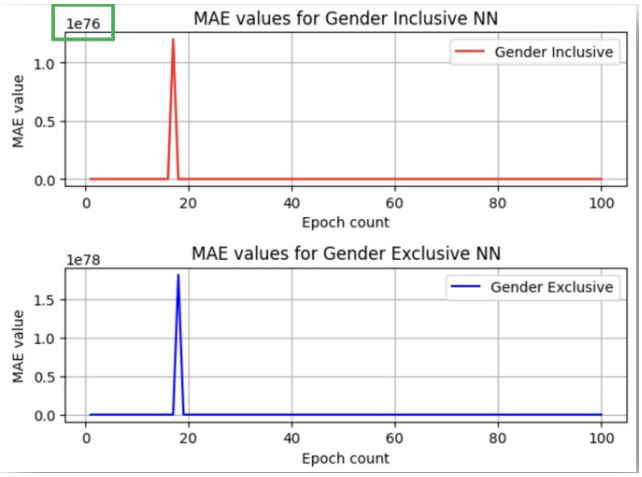
	Version	Model	Performance
0	gender	Lasso-G	22809.8584620818
1	no-gender	Lasso-NG	23113.3675706104
2	gender	Ridge-G	22809.1819256734
3	no-gender	Ridge-NG	23112.5730453263
4	gender	NN-G	53243.5239806428
5	no-gender	NN-NG	53243.5239806428

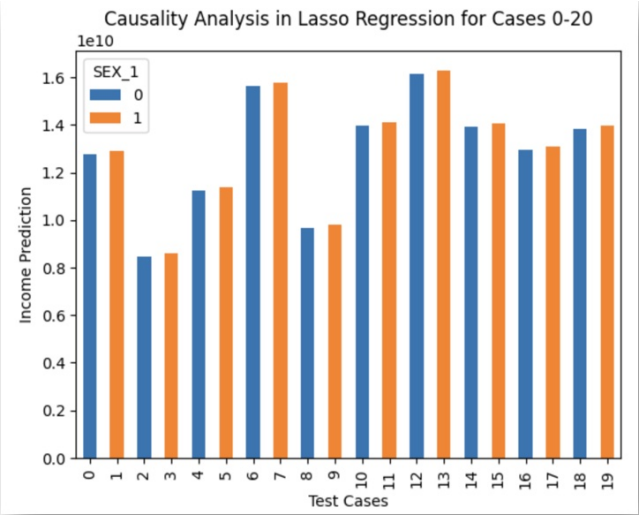
	Version	Model	Performance
0	gender	Lasso-G	1110197315.4845182896
1	no-gender	Lasso-NG	1138758788.0071637630
2	gender	Ridge-G	1110187736.0106124878
3	no-gender	Ridge-NG	1138749225.9848384857
4	gender	NN-G	4735728500.2360105515
5	no-gender	NN-NG	4735728500.2360105515





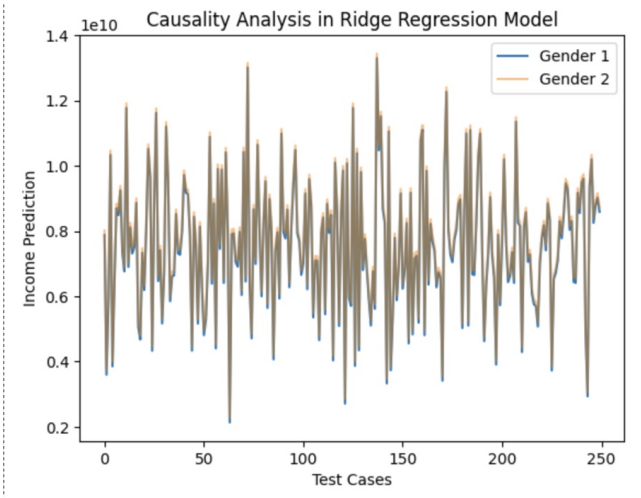


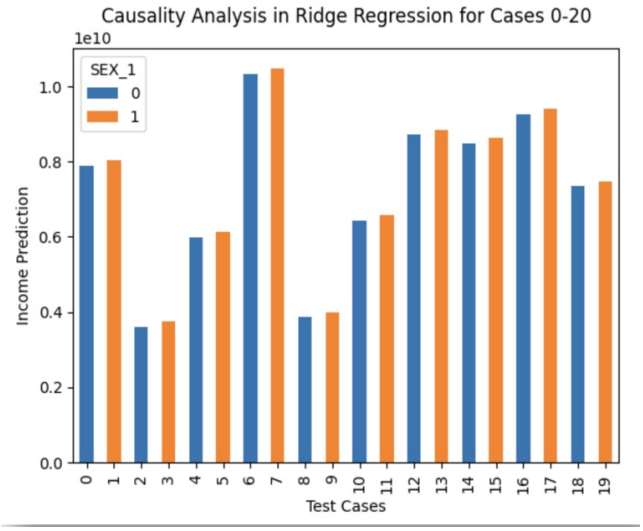




6.2 Inference Performance

During the inference phase, we conducted a thorough causality analysis to examine the impact of gender on income predictions. Employing a rigorous methodology, we generated five hundred test cases, each comprising pairs that differed solely in gender. Subsequently, we utilized our trained models to perform inference, observing distinct variations in predicted income levels. Notably, across all test cases, predictions consistently favored one gender, with income estimates consistently higher for individuals of the same gender within each pair. This phenomenon can be attributed to the inherent biases encoded within the feature weights associated with gender in our models as well as the architecture of regression models. By analyzing these results, our study advances the frontier of quantifying gender bias, providing compelling evidence of consistent disparities in income predictions based solely on gender discrepancies.





7 Insights

Based on the clustering and results in the section above, we have pulled the following insights:

- **Gender is a crucial input feature.** Incorporating gender as an input feature improved the predictive accuracy of our models, underscoring the pivotal role of gender in facilitating income prediction.
- **There is inherent bias in the model predictions.** The consistent tendency for predicted income levels to favor individuals of the same gender within test case pairs highlights the existence of inherent biases embedded within our models. These biases reflect broader systemic disparities within societal structures, emphasizing the imperative for implementing effective mitigation strategies in predictive modeling practices.
- **This study advances gender bias quantification.** Our research adds to the ongoing conversation surrounding the quantification of gender bias by presenting compelling evidence of persistent inequalities in income predictions stemming solely from gender differences. By bringing attention to these disparities, we lay the groundwork for the development of more inclusive and equitable approaches to predictive modeling.
- **Promising Prospects for Gender Equality.** According to the analysis results, in 2023, the income disparity between men and women is minimal in several key sectors, including First-Line supervisors of retail sales workers, Driver/sales workers and truck drivers, Registered nurses, and Cashiers. Despite this encouraging trend towards gender pay equity, the proportion of men and women in these professions varies significantly. This indicates that while income equality is being approached, gender representation within these roles still presents an area for potential improvement.

8 Conclusion

In summary, our research offers a thorough examination of the influence of gender on income prediction and its relevance for mitigating gender disparities in the workplace. Through an analysis of various regression models, both with and without gender data, we have underscored the profound effect of gender on predictive precision. Notably, the inclusion of gender as an input feature consistently bolstered model performance, emphasizing its crucial role in enhancing income estimation accuracy.

Moreover, our examination of causality during inference unveiled intrinsic biases embedded within our models, leading to systematic differences in predicted income levels solely attributable to gender variations. These discoveries emphasize the necessity for robust mitigation measures in predictive modeling to uphold fairness and equity within income prediction frameworks.

Looking ahead, while challenges remain, the observation of smaller gender wage disparities in certain occupational categories in 2023 offers hope. This indicates that gender wage equality is an achievable goal through continuous research and targeted policy interventions. We encourage future research to explore the success factors in these areas and consider how these practices can be extended to a broader range of industries.

Our study adds to the ongoing discussion on gender equity by quantifying gender bias and illuminating its effects on income prediction. Through evidence of gender-based wage disparities, we advocate for the adoption of more inclusive and equitable predictive modeling approaches. Looking ahead, it is essential to confront these discrepancies through ongoing research, policy interventions, and organizational efforts focused on advancing gender equality in the workplace. By deepening our comprehension of the intricate dynamics of income prediction, we can work towards building a fairer and more just society for everyone.

Acknowledgements

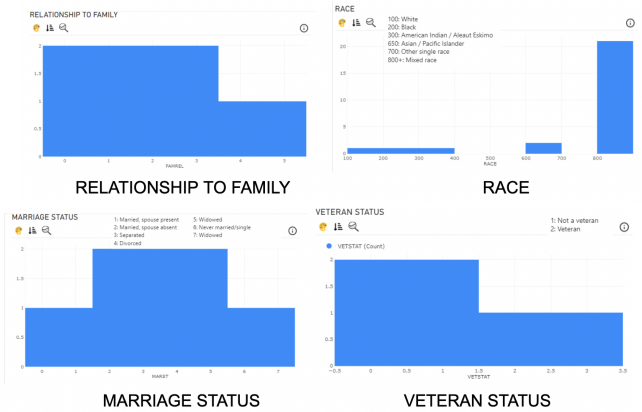
This research received support during the ESE 527 course, instructed by Professor Patricio S. La Rosa, PhD at the School of Washington University in St. Louis.

References

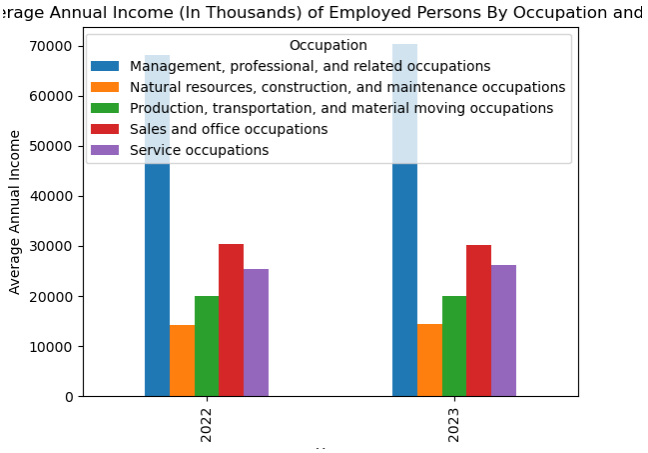
Pew Research Center (2023). *The Enduring Grip of the Gender Pay Gap.* Accessed [Feb. 5, 2024]. URL: https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/#:~:text=The%20gender%20pay%20gap%20%E2%80%93%20the,every%20dollar%20earned%20by%20men.

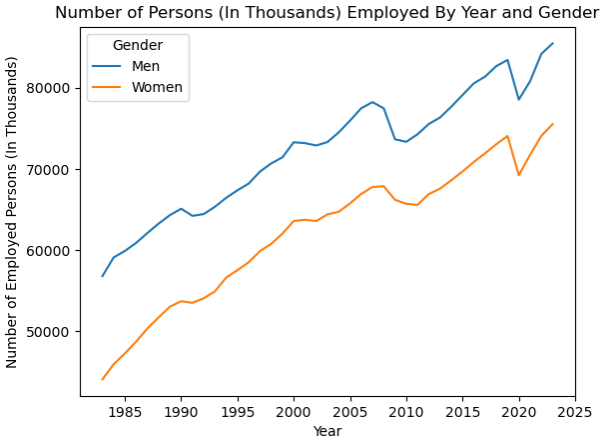
9 Additional Visualizations

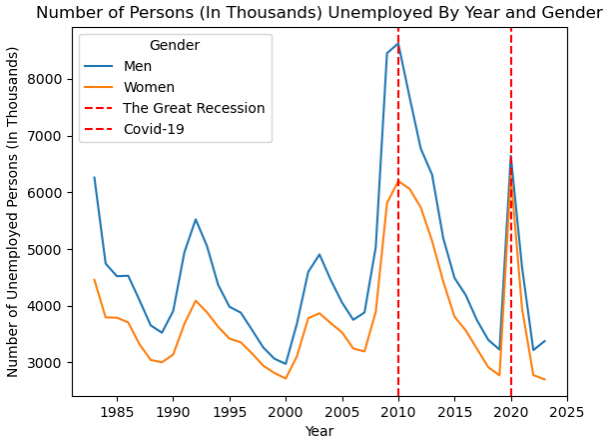
Below are some sample histograms and box plots of our data.




TOTAL INCOME
INDUSTRY
OCCUPATION
CITIZENSHIP STATUS
BIRTHPLACE
EMPLOYMENT STATUS
CLASS OF WORKER
EDUCATION LEVEL
HAS DIFFICULTIES
SCHLCOLL
AGE
UHRSWORKT
UHRSWORKLY
WKSWORK1


erage Annual Income (In Thousands) of Employed Persons By Occupation and
Occupation
Management, professional, and related occupations
Natural resources, construction, and maintenance occupations
Production, transportation, and material moving occupations
Sales and office occupations
Service occupations
Average Annual Income
2022
2023


Number of Persons (In Thousands) Employed By Year and Gender
Gender
Men
Women
Number of Employed Persons (In Thousands)
Year


Number of Persons (In Thousands) Unemployed By Year and Gender
Gender
Men
Women
The Great Recession
Covid-19
Number of Unemployed Persons (In Thousands)
Year