A Machine Learning Approach to Understanding the Determining Factors of the Gender Wage Gap

Sophia Guan

Gender inequality is a complex subject consisting of a variety of issues and nuances. In this project, we choose to study gender income inequality—a prevalent issue in current society. Among the many factors that play a role in the gender wage gap, we focus on the affects of marital status, race, geographical location (by state), age, and years of education. By using these variables to create a model able to predict the hourly wage gap between a woman and their equivalent male counterpart, we can analyze the impact of each variable to better understand the role they play in the income gap. Utilizing income data from the Current Population Survey, we train and test five models—a Linear Regression, Decision Tree Regressor, Random Forest Regressor, KNeighbors Regressor, and MLP Regressor. Our Linear Regression model found that there is a correlation between being a never married worker and a smaller gender wage gap, as well as being a married worker with an absent spouse and a greater gender wage gap. In general, though, our models found little correlation between the variables provided and the predicted hourly age gap.

1. Introduction

Gender income inequality is not as simple as "equal pay for equal work"—a variety of personal and societal factors contribute to the gender wage gap. By studying the affect of different attributes on the gender wage gap, we can better understand both the scale of this issue and its possible solutions. So, we explore the question, how does a worker's marital status, along with other variables, impact the gap in hourly wage between male and female workers? We seek to create a model able to predict the gender wage gap given a set of variables—age, years of education, race, state, and marital status. Through our model, we can study the weight of each variable in deciding how big this gap will be. This model can be used to determine public policy for childcare. For example, if an area has a bigger predicted gap, policies can be implemented in that area for free childcare or to expand afterschool program options.

2. Related Works

Similar studies have been done from an economics perspective. Particularly, a paper published in the Journal of Economic Literature 2017 analyzes the decline in importance of human-capital factors and the increased significance of psychological characteristics and non-cognitive skills in regards to the gender wage gap (Blau & Kahn, 2017, p. 1).

From a machine learning perspective, there has been research done on the significance of different variables on the predicted income of an individual. In a report by Junda Chen, Chen (2021) analyzes the importance of different features using a Logistics Regression and a Random Forest model. In this research, however, gender is merely one of the factors in predicting income (Chen, 2021, p. 1).

3. Data

The data used in this project comes from the Gender Pay Gap Dataset on Kaggle, which provides information on 234 variables including hourly wage and gender (Blau & Kahn, 2017). Of these 234 variables, we utilize seven that were most relevant to our model—sex, age, years of education, race, state, hourly wage and marital status. By including these other variables, we can explore beyond just marital status to understand how this gap may differ across the country, or at different ages. To our benefit, this data is already numerical. For example, a "1" in the sex variable represents a male, while a "2" represents a female.

3.1 Dataset Overview

The variables we utilize for our model are:

- sex (2 integer options): The sex of the individual, either male or female
- age (continuous positive integer): The age of

the individual

- race (4 integer options): The race of the individual, either White non Hispanic, Black non Hispanic, Hispanic, or Other non Hispanic
- statefip (50 integer options): The state the individual lives in
- sch (20 integer options): The years of education the individual has attained, from grade school to an advanced degree
- marst (6 integer options): The marital status of the individual, either married (spouse present), married (spouse absent), separated, divorced, widowed, or never married
- realhrwage (continuous positive dollar amount): The hourly wage of the individual.

3.2 Limitations

Though this dataset provides thorough information on our topic, it does include some limitations. In the "race" variable, the dataset lacks specification on Census-recognized groups, such as Asian, Pacific Islander or Native American, including them under the "Other non Hispanic" section (Jensen et al., 2021).

In addition, of the many years of data included, we focus on data from 2013, which is the most recent year provided. Using slightly older data to train and test our model may result in a slightly dated model. Gender inequality is a constantly changing topic in society, and data from 2013 might not fully reflect the current situation.

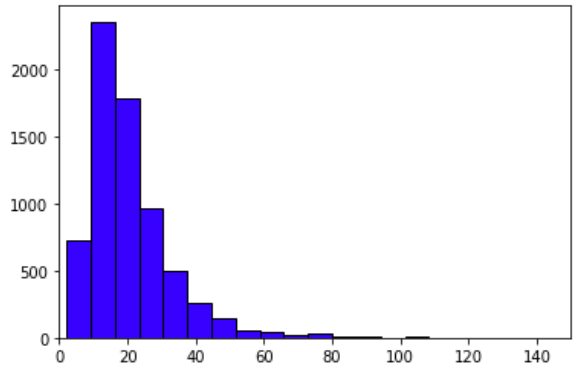
Moreover, the dataset doesn't include information on the amount of hours a worker has worked. This variable could be important in deciding a worker's hourly wage.

Finally, the dataset only provides information on American workers, so it does not accurately reflect the entire world.

4. Methodology

4.1 Data Preprocessing

In order to begin training our models, we need to take a few steps to pre-process our data. First, our data for race, marital status and state is numerical. For example, a value of "1" in race represents White non Hispanic, a "2" represents Black non Hispanic, a "3" represents Hispanic, and a "4" represents Other non Hispanic. However, this representation of race does not make sense within the context of a machine learning model; one race does not carry a greater value than another. Similar logic applies to states—California does not have a greater value than Florida. So, in order to alleviate this issue, we use one hot encoding. Essentially, for each worker, we split the race variable into four separate variables (White non Hispanic, Black non Hispanic, etc.) Then, for the variable that matches the worker's race, we assign the value 1. For the other variables, we assign the value 0. So, if a worker is Hispanic, the value stored in the White non Hispanic, Black non Hispanic, and Other non Hispanic variables would all be 0, while the value of the Hispanic variable would be 1. This way, no race is deemed to have a greater value than another within our model. We apply the same one hot encoding technique for the "state" and "marst" variable. In our model, we include a the hourly wage of the male worker in order to study the change in wage gap as the income of a worker rises. For example, we may find that the wage gap between male and female workers increases as their hourly wages increase. However, if one of the inputs of our model is the exact wage of the male worker, our model will end up relying heavily on that variable to predict the gap. So, we categorize the male worker's hourly wage.



Number of Workers vs Hourly Wage
[Figure 1]

Given the distribution shown above, we categorize the male worker's wages in increments of 10. So, an hourly wage between 0 and 10 is represented by a 1, an hourly wage between a 10 and 20 is represented by a 2, and so on. In order to take into account the fact that there are technically two

wages included in our data (male and female), we add a "realhrwage_squared" variable which is the square of the categorized wage variable.

4.2 Constructing the Gap

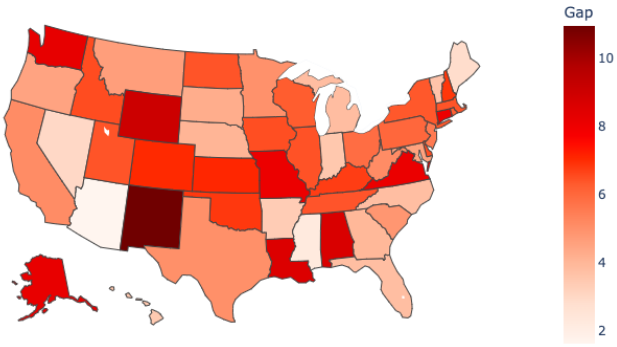
In order to train a model that can predict the gender hourly wage gap given a set of variables, our data must include the gap between a male worker and an identical female worker (in terms of our variables). However, the Current Population Survey does not provide this, so we must calculate the gap. In order to find the hourly wage gap, we need to compare a female and male worker who each share the same traits (in age, years of education, race, state, and marital status). In some cases, multiple females or multiple males in the dataset share the same traits. In order to mitigate this occurrence, we group our data by the age, years of education, race, state, and marital status and take the mean of their hourly wages. Essentially, we average the wages of identical female workers and identical male workers.

In order to match female and male workers, we first split the dataset into two on the basis of sex—one for female workers, and one for male workers. For each worker, we create an "id" variable that corresponds to the age, years of education, race, marital status, and state variables of that worker. Then, we merge the two datasets based on the "id" variable. For each pair of workers in this new dataset, we subtract the female worker's hourly wage from the male worker's hourly wage and label this new variable the "gap." Our reconstructed dataset essentially contains 6,965 pairs of identical workers—one male, and on female. Each row includes the age, years of education, state, race, marital status, male wage, and wage gap of the identical pair. We no longer need the "sex" variable, since we know each row contains one male and one female.

In some cases, a female worker's hourly wage will be more than a male worker's hourly wage, resulting in a large negative "gap" variable. These values can be difficult for machine learning models to interpret, so we normalize the value between -1 and 1. We use this "gap_normalized" variable to train and test our models.

4.3 Data Exploration

In order to understand our models, we must first explore the new dataset we have constructed. We found that the mean gap across the whole dataset was $6.778 per hour. The figures below provide more information on the data:



Median Gap by State
[Figure 2]

Race	Median Gap (Dollars)
White Non Hispanic	20.417
Black Non Hispanic	16.136
Hispanic	13.831
Other Non Hispanic	19.363

[Figure 3]

5. Multiple Models

Before training and testing, we must select which models to use. Our data is numerical, so the models we use will be Regressors (rather than a Classifiers). So, as a baseline model, we will use a Linear Regression. Then, we will test slightly more complex models—a Decision Tree Regressor, a Random Forest Regressor, and K-Nearest Neighbors model. Finally, we will test a neural network— the MLP Regressor.

Before beginning to train our models, we must define the input and output variables. The goal of our model is to be able to predict the income gap based on the age, race, state, years of education, hourly wage, and marital status. So, we set our output "y"

variable as our normalized gap variable and set our input "x" variable as everything else in our dataset. Then, we must split our dataset into two sections— a training section, and a testing section. We use 67% of our data for training, and 33% for testing. Then, we train our model, and test it. In order to analyze our model, we must compare the predicted gap values to the actual gap values. However, since we used the normalized gap to train the model, the predicted gap values will also be normalized. So, we must first denormalize these values, then evaluate the accuracy by analyzing the root mean squared error, mean absolute error, and r^2 score. We go through this process for our five chosen models.

Linear Regression. Before training this model, we remove the "state" variables from our input "x" as it causes our coefficients to be extremely large. In this case, we let the model automatically choose its own hyperparameters:

- 'copy_X': True
- 'fit_intercept': True
- 'n_job's: none
- 'normalize': 'deprecated'
- 'positive': False

Decision Tree Regressor. For this model, we include the following hyperparameters:

- 'random_state': 0
- 'min_samples_split': 200
- 'max_depth': 5

Random Forest Regressor. For this model, we include the following hyperparameters:

- 'random_state': 0
- 'min_samples_split': 200
- 'max_depth': 5
- 'n_estimators': 20

KNeighbors Regressor. For this model, we include the following hyperparameter:

- 'n_neighbors': 4

MLP Regressor. For this model, we include the following hyperparameters:

- 'random_state': 0
- 'max_iter': 1000
- 'hidden_layer_sizes': (5,10)

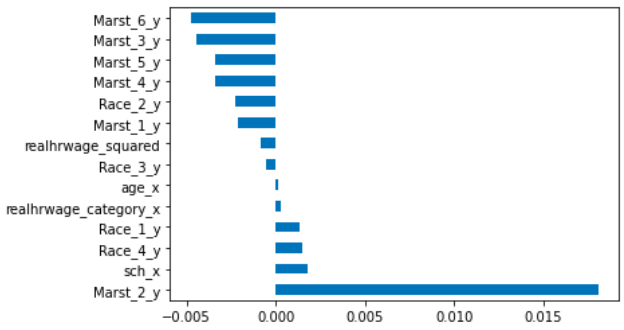
6. Results

	Root Mean Squared Error	***Mean Absolute Error***	***R^2 Score***
Linear Regression	19.832	11.578	0.353
Decision Tree Regressor	20.150	11.798	0.333
Random Forest Regressor	20.123	11.530	0.335
KNeighbors Regressor	22.602	12.557	0.161
MLP Regressor	124.369	75.307	-24.410

[Figure 4]

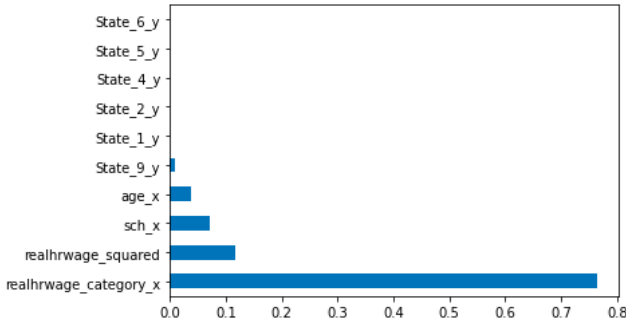
Based on the error values and R^2 score shown above, these models lack predictive power. This will be addressed in Section 7.2. However, we will choose the three most predictive models to analyze.

Linear Regression. A Linear Regression model includes a coefficient for each variable, which helps us understand the weight of each variable in determining the gap. The graph below includes the the coefficients of each variable:



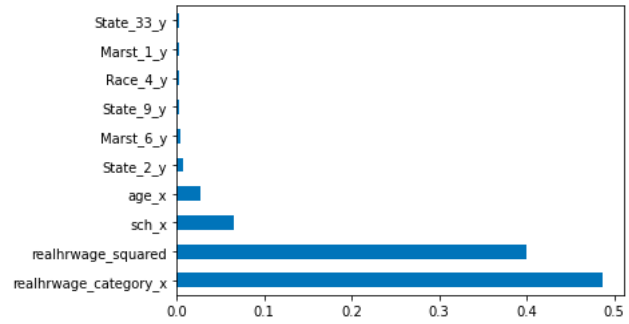
[Figure 5]

Decision Tree Regressor. A Decision Tree Regressor includes a "feature importance" attribute which we use to evaluate the importance of each variable. The graph below includes the 10 variables with the greatest importances:



[Figure 6]

Random Forest Regressor. A Random Forest Regressor includes a "feature importance" attribute which we use to evaluate the importance of each variable. The graph below includes the 10 variables with the greatest importances:



[Figure 7]

7. Discussion

7.1 Interpreting Model Results

When analyzing the importance and impact of variables among different models, it may seem necessary to focus on the values for coefficients and importances provided by the models. Though these values can be insightful, it is often more important to analyze the relationship between different variables, which we will be exploring in this section. In particular, the Linear Regression model is able to give us the most information on each variable—we can see which variables have a positive or negative affect on the wage gap, rather than just the importance.

Linear Regression. We can judge the importance, or weight, of a variable in a Linear Regression by looking at each variable's coefficient. Coefficients of greater magnitude have a greater impact on the predicted gap.

From the graph we produced in Figure 5, we can see that the "Marst_2_y" variable, which represents a married person with an absent spouse, has the greatest positive coefficient. Essentially, according to our model, if a given pair of workers are a married man with an absent spouse and a married woman with an absent spouse, their hourly wage gap is likely to be greater.

On the opposite end of the graph, the "Marst_6_y" variable, which represents a never married worker, has the greatest magnitude negative among variables. According to our model, if a given pair of workers are a never married man and a never married woman, their hourly wage gap is likely to be smaller. In the context of the real world, this seems logical as never married women may find it necessary to be financially independent, so they would be more likely to earn wages closer to that of their male counterparts.

Then, in order from greatest to least magnitude, the coefficients of separated, widowed, divorced, and married workers were all negative. Essentially, each variable would have made the predicted hourly wage gap smaller, but by decreasing amounts.

For the race variables, the coefficient of 'Race_4_y' had the greatest positive value, followed by 'Race_1_y'. Essentially, our model predicts that a pair of workers who are both not White, Black, or Hispanic are likely to have a greater hourly wage gap. Next, a pair of workers who are White non Hispanic are also likely to have a slightly greater hourly wage gap.

In contrast, the 'Race_2_y' and 'Race_3_y' variables had negative coefficients. Our model predicts that a pair of workers who are Black non Hispanic, or a pair of workers who are Hispanic are likely to have a smaller hourly wage gap.

The 'sch_x' variable, representing years of education, had a positive coefficient, indicating that the hourly wage gap would grow as the amount of years of education for a pair of workers increased. For example, a woman and man with an advanced degree are predicted to have a greater wage gap than

a women and man with just a high school education.

The 'age_x' variable, representing age, seems to have little affect on the hourly wage gap.

In comparison to the race, age, and years of education variables, the variables for marital status are generally greater in magnitude, so they have a greater impact on the final predicted gap.

Decision Tree Regressor. We use the feature importance attribute to analyze the importance of different variables in a Decision Tree Regressor.

In Figure 6, we can see that the 'realhrwage_category_x' variable has the greatest importance, followed by the 'realhrwage_squared.' Essentially, the model deems the categorical wage of the pair of workers as most important in predicting the hourly wage gap. This variable is important to include our model as it can help us understand the way the wage gap changes as income changes. According to the results from the Decision Tree Regressor, the categorical wage of a worker is extremely important in deciding the income gap between a male and female worker.

Following in magnitude are the 'sch_x' and 'age_x' variable, which are third and fourth in importance. The rest of the variables, including marital status, seem to have almost no importance in this model.

Random Forest Regressor. We use the feature importance attribute to analyze the importance of different variables in a Random Forest Regressor.

According to Figure 7, the relationship between the importance values of the variables of the Random Forest Regressor is very similar to that of the Decision Tree Regressor. However, the Random Forest Regressor does include the "Marst_6_y" variable, which represents never married, and "Marst_1_y" variable, which represents married (spouse present), in its list of top 10 importances. For the Random Forest Regressor, these two variables are important, but not as important as the categorical wage, years of education, and age variables.

7.2 Alternate Approach

Given the poor accuracies and R^2 score in Figure 4, we want to understand what is causing this lack of predictive power. Since the hyperparameters used were already optimized, we take an alternate approach. In this approach, instead of predicting the wage gap between a male worker and female worker, we seek to predict the *hourly wage* of a worker using 'sex' in addition to the same set of variables. This allows us to see the direct relationship between our variables and the hourly wage of a worker. After training and testing these models, we found that the accuracies for these models were even worse. The best model was the Linear Regression, which had a Root Mean Squared Error of 22.355, a Mean Absolute Error of 8.715, and a R^2 score of 0.367. Given these error values, it is possible that the variables we use aren't enough to explain a worker's income well, so our models aren't able to make accurate predictions.

8. Conclusion

Though these results may seem ambiguous, there are still a few conclusions we can draw. First, our models, particularly our Linear Regression model, provide insight on the affects of different marital status and race on the gender wage gap. This information can better help us understand the topic of gender inequality as a whole, and can also be implemented in public policy in the future. We can also conclude that the variables we included aren't enough to explain the gender wage gap well. This invites further research—which variables are important in explaining the gender wage gap? Will adding more variables to our current set improve our models? In addition, further research can be done using data on other countries, or the world. Perhaps on a greater geographical scale, these variables play a more important role in predicting the gender wage gap. It may also be insightful to study the gender wage gap among specific professions—this could give us insight on the areas or industries that this issue is most prevalent in. Gender income inequality is an evolving issue in society and machine learning can be used as an important tool to better understand and combat it.

Acknowledgments and References

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, *55*(3), 789-865. https://doi.org/10.1257/jel.20160995

Chen, J. (2021, September 1). *Feature Significance Analysis of the US Adult Income Dataset*. https://minds.wisconsin.edu/bitstream/handle/1793/82299/TR1869%20Junda%20Chen%203.pdf?sequence=1&isAllowed=y

Gender Pay Gap Dataset. (n.d.). Kaggle. Retrieved September 3, 2022, from https://www.kaggle.com/datasets/fedesoriano/gender-pay-gap-dataset

Jensen, E., Jones, N., Orozco, K., Medina, L., Perry, M., Bolendor, B., & Battle, K. (2021, August 4). *Measuring Racial and Ethnic Diversity for the 2020 Census*. United States Census Bureau. Retrieved September 3, 2022, from https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html

Code: https://github.com/sophiaguan/genderwagegapanalysis

Special thanks to Mrs. Ana Sofia Muñoz Valadez from the University of Chicago for her mentorship throughout this project.