Analysis of Influential Factors and Discrimination in Gender Pay Gap in the United States

Giulio Caputi 903972988

April 2024

Abstract

The wage gap between men and women is not a novel feature of our society, perpetrating through different times and economic conditions. Studying it is crucial for several reasons. Firstly, understanding the extent and causes of wage disparities between men and women is fundamental to achieve economic equity. Secondly, narrowing the gender wage gap can enhance economic efficiency by ensuring that human resources are used more effectively, regardless of gender. Additionally, reducing wage inequality can help alleviate poverty and improve financial security for families, particularly those headed by women. Finally, addressing pay disparities is also a matter of social justice. In this work, several machine learning and statistical techniques are employed to quantify the extent to which observable characteristics (like education or hours of work) explain the wage differential between men and women, how these extents differ between the two groups, and how much of the income differential is due to unobservable factors, likely discrimination.

Contents

1 Introduction

This paper investigates the gender wage gap in the United States by examining the extent to which variables like schooling, marital status, and occupation influence the probability for individuals to earn more than $50,000 annually. All the models employed here are discussed in depth, prioritizing mathematical and logical soundness and precision. Logistic regression models are implemented to quantify the relevance of each variable for men and women separately, then the same is done using random forests, and lastly a Fairlie decomposition is performed with the aim of quantifying how much of the probability of individuals to earn more than $50,000 per year is due to observable characteristics, and how much is due to discrimination. The hope is to provide insights into the underlying mechanisms that perpetuate economic inequalities. The findings presented here are expected to inform policy recommendations targeted at reducing the gender wage gap and promoting equitable economic opportunities across different segments of the population. The Python code employed for this analysis is available at `https://github.com/CapGiulio/GenderPayGapAnalysis`.

2 Data Description

The dataset used in this analysis, which is available at the aforementioned GitHub repository, was extracted from the United States Census Bureau database by Ronny Kohavi and Barry Becker, and it contains 43,957 observations, each corresponding to a different individual. For each of such individuals, the following variables were collected:

- *age*: the age of the individual
- *workclass*: a categorical variable indicating the work class of the individual, with classes Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
- $fnlwgt$: it stands for "final weight", and it denotes the number of people the census believes the entry represents
- *education*: a categorical variable denoting the highest level of education achieved by the individual, with classes Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool

- $educational - num$: the number of school years passed by the person
- $marital-status$: a categorical variable indicating the marital status of the individual, with classes Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- $occupation$: the general type of occupation of the individual, with classes Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- $relationship$: a variables expressing the relationship status of the individual relative to others, with classes Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- $race$: the individual's ethnicity, with classes White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- sex: the sex of the individual, with classes Female, Male
- $capital - gain$: the capital gains for the individual
- $capital - loss$: the capital loss for the individual
- $hours - per - week$: a variable denoting the number of hours the individual has reported to work per week
- $native - country$: the country of origin of the individual, with several classes, United States being the most popular one

As anticipated, the target variable is binary, and it denotes whether the individual earns more than \$50,000 a year or not. To increase the clarity of the analysis, the names of some columns are changed. Specifically, $educational - num$ becomes $schooling$, $race$ becomes $ethnicity$, $hours - per - week$ becomes $hours$, $native - country$ becomes $immigrant$ (it will be later turned into a binary variable), and $income > 50K$ (the target variable) becomes $target$.

3 Feature Engineering

3.1 Removing Irrelevant Variables

The first step in feature engineering is to remove the following irrelevant predictors:

- *education*, given its collinearity with *schooling*
- $relationship$, since it is largely repetitive, given the presence of $marital - status$
- $capital - gain$, because of the almost-imperceptible impact it has on $target$
- $capital - loss$, for the same reason

Doing so simplifies the analysis carried on below, and increases the focus on more relevant features.

3.2 Dealing with Missing Values

Next, missing values are dealt with. The only columns that contain null values are *workclass*, *occupation*, and *immigrant*. A classic approach when working with incomplete datasets is to build a simple model to predict the missing values from the other variables of the individual. For instance, one might train a shallow neural network with the variables *workclass*, *schooling*, and *gender* to predict *occupation*, and substitute the missing values for *occupation* with the values predicted by such model. In spite of this possibility, the approach followed in this work is to simply discard the individuals with incomplete data. Indeed, the total number of rows with at least one missing value is only 3,230, which corresponds to roughly 7% of the available data, therefore these rows can be removed from the dataset without risking to significantly reduce the predictive power of the models developed.

3.3 Grouping Values

Subsequently, different values of the same categorical variable are grouped together, for the following features:

- *workclass*: all types of government jobs are grouped into a single category, and also all kinds of self-employed jobs are grouped into another category
- $marital - status$: the values are grouped in Married, Separated, Never-Married, and Widowed
- *occupation*: the values of this variable are now restricted to White Collar, Blue Collar, and Service
- *ethnicity*: the values of this variable are now restricted to White, Black, Asian, Eskimo, and Other

- *immigrant*: the individuals originally from the US are given a value of 0 for this variables, while all the others have a value of 1

Also this step aims at simplifying the analysis and the models developed in this work.

3.4 Standardization

Standardization is a preprocessing method that is often performed on data before using such data to train a machine learning model. It involves rescaling continuous features so that each feature has a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean and dividing by the standard deviation for each data point. Therefore, if $X \in \mathbb{R}^{N \times N}$ is our data matrix, the standardization of the i^{th} feature occurs as follows:

$$Z_{:,i} = \frac{X_{:,i} - \mu_i}{\sigma_i^2}$$

Here, Z is the matrix containing the standardized values, $Z_{:,i} \in \mathbb{R}^N$ denotes the i^{th} column of Z, $X_{:,i} \in \mathbb{R}^N$ is the i^{th} column of X, $\mu_i \in \mathbb{R}$ is the mean of $X_{:,i}$, and $\sigma^2 \in \mathbb{R}$ is the standard deviation of $X_{:,i}$. This ensures that the mean of $Z_{:,i}$ is 0 and the standard deviation of $Z_{:,i}$ is 1, for every i corresponding to a continuous feature.
The main reasons why this process is performed are scale invariance and improved convergence. Regarding the first point, in many machine learning algorithms, especially those that use gradient-based optimization methods (like logistic regression, which is implemented in this work), features with larger scales can disproportionately influence the model. Standardization helps in ensuring that the contribution of all features is the same. Regarding the second point, algorithms that use gradient-based optimization algorithms for minimizing the loss tend to converge faster when features are standardized, as the gradient is more balanced. In this work, the standardized features are age, $fnlwgt$, $schooling$, and $hours$.

3.5 Encoding Variables

This process involves converting categorical variables (in this case recorded as strings) to integers. The main reason this is performed is for algorithm compatibility. Indeed, many machine learning algorithms can only handle numerical inputs and do not work with raw categorical data (e.g., strings). Here, two types of variable encoding are performed, namely simple encoding and one-hot encoding.
Simple encoding means turning each class of a categorical variable to a different integer.

It is generally done either when the categories have a natural order (e.g., for educational levels, where it makes sense to record the completion of a doctoral degree with a higher number than a high-school diploma), or when the machine learning model can learn well even in the absence of such natural order (as it is the case for random forests, for example).
Also one-hot encoding is a method employed to convert categorical variables into a form that can be provided to machine learning algorithms to make them work better when making predictions. It is used when the labels of a given categorical variable do not have a natural order, and instead of converting each of them into an integer (as standard encoding involves), one-hot encoding creates a new binary column for each label except one, which is referred to as the base class. This approach removes any ordinal relationship and allows the algorithm to treat each category with equal importance.
For example, the variable *occupation* has three classes, namely Blue Collar, White Collar, and Service. With one-hot encoding, this variable will be substituted by two binary columns, in our case representing whether the individual has a white-collar or a service job, respectively. When both of these columns are 0, the individual has a blue-collar job. In this analysis, the variables which are encoded are $workclass$, $marital-status$, $occupation$, and $ethnicity$. Two data frames are created, one with the simple-encoding representation of these variables, and the other one with their one-hot encoded form. The first dataset is used to train the random forest model, while the latter is used for logistic regression.

4 Logistic Regression

4.1 Theoretical Background on Logistic Regressions

Logistic regression is a method for binary classification, meaning that it is used when the target variable has only two classes, as in the case treated here. Data is denoted as $S = \{x_i, y_i\}_{i=1}^N \in \mathbb{X} \times \{0, 1\}$, which means that each $x_i \in \mathbb{X}$ and each $y_i \in \{0, 1\}$. Logistic regression models the probability for an input x to belong to a certain class, and then uses thresholding (aka quantization) to predict y.
A so-called logistic (also called sigmoid) function $k : \mathbb{R} \to [0, 1]$ is defined as

$$k(x) := \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Two interesting properties of this function are that

$$1 - k(x) = \frac{1+e^x-e^x}{1+e^x} = (1 + e^x)^{-1}$$

and that its first derivative is

$$k'(x) = k(x)(1 - k(x))$$

The posterior probabilities of the two class labels (given input x) are

$$P_{Y|X}(1|x) = k(x^T w + w_0)$$

and

$$P_{Y|X}(0|x) = 1 - k(x^T w + w_0)$$

where w is the vector of weights that the model learns, and w_0 is a kind of intercept called bias term.
It is important to notice that the input of k is always in $\mathbb{R}$ (so it is a number, not a vector). The function k is used to map from $\mathbb{R}$ to $[0, 1]$. The first step is to learn the weights w by maximum likelihood estimation (MLE), which involves finding the w vector that maximizes the likelihood of the observed data (i.e., that makes the observed data more probable). In practice, often the negative log-likelihood (also called cross-entropy) is minimized, as taking the logarithm of the likelihood turns products into sums (with which it is easier to work) and increases numerical stability. The minimizer of the negative log-likelihood is clearly equal to the maximizer of the likelihood, as $argmin_x\{-\log(f(x))\} = argmax_x\{f(x)\}$, provided such logarithm exists. The likelihood function for the logistic model is

$$L(X) = P(X)P(y|X, w)$$

Since $P(X)$ does not depend on w (which is a reasonable assumption), maximizing the likelihood is equivalent to maximizing $P(y|X, w)$, so

$$P(y|X, w) = \prod_{i=1}^{N} k(x_i^T w)^{y_i}(1 - k(x_i^T w))^{1-y_i}$$

As stated above, it is more convenient to work with the negative log-likelihood, which has the following form

$$-\log(P(y|X, w)) = CE(w) = \sum_i (-y_i x_i^T w + \log(1 + e^{x_i^T w}))$$

This formulation is valid only for $y \in \{0, 1\}$; if the classes were different, also the cross-entropy would. Then, w^* (which is the estimated vector of weights) is the vector w that minimizes this quantity. In order to minimize the negative log-likelihood, a standard procedure is setting its gradient with respect to w to 0 and solving for w. As the logarithm is a concave function, this actually maximizes the log-likelihood, and therefore minimizes the negative log-likelihood. The gradient of the cross entropy is

$$\nabla - \log(P(y|X,w)) = \nabla CE(w) = X^T(k(Xw) - y)$$

Because k is not a linear function, inverting the expression above is not possible, thus there is no closed-form solution for w^*. Nevertheless, it is possible to minimize this function efficiently, as it is convex in w. Indeed:

> The function $-y_i x_i^T w$ is convex, as a linear function is both convex and concave. Moreover, $x \approx \log(1 + e^x)$, in fact $e^x \approx 1 + e^x$ for large x, so $\log(1 + e^{x_i^T w})$ is convex (its second derivative is greater than or equal to 0). So, since the cross-entropy is the composition of convex functions, it is itself convex. Additionally, the Hessian of the cross entropy (i.e., the matrix H such that $H[i,j] = \frac{\partial^2}{\partial w_i \partial w_j} CE(w)$) is
>
> $$H = \nabla^2 CE(w) = X^T S X$$
>
> where S is a diagonal matrix whose diagonal elements are $k(x_i^T w)(1-k(x_i^T w))$, which are all ≥ 0. Therefore, S is positive semi-definite, so the Hessian of $CE(w)$ is positive semi-definite, and thus $CE(w)$ is convex in w.

After having learned the weights w, the logistic function uses them to actually make predictions. Employing a standard threshold of 0.5, the function predicts 1 if $P(1|x) \geq 0.5$, and 0 if $P(1|x) < 0.5$.

A very large $|w^T x + w_0|$ corresponds to $P(1|x)$ very close to 1 or to 0, which means the confidence of the predicted label of x is high. Instead, if $|w^T x + w_0|$ is small, then $P(1|x)$ is close to 0.5, so the confidence of predictions is small.

This process effectively learns a hyperplane orthogonal to w. On one side of this hyperplane the prediction is 1, and on the other it is 0, depending on $sign(w^T x + w_0)$. It is interesting to notice how such hyperplane is independent of the scale of w, as changes in extreme values affect predictions only by a little amount (they go "in the right direction", so they are easy points to treat).

The higher the norm of w, the faster the transition between one region and the other. In fact, if the norm of w is ∞, the prediction function is essentially a step function (so that the transition between one label to the other is immediate). If instead the norm of w is small, the transition is smooth. The larger the norm of w is (and so the faster the transition is), the closer probabilities are to 0 and 1, so the higher the confidence is. If the only interest is in predicting the labels (and so quantifying confidence is not relevant), then only the direction of w is important, and its norm is irrelevant.

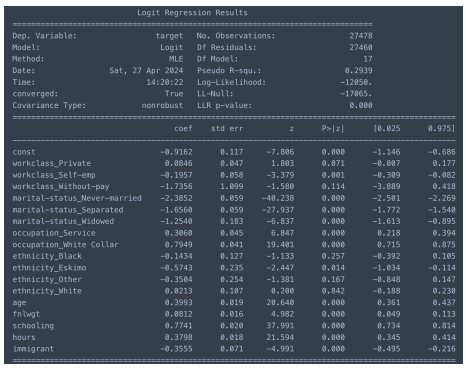
Changing the bias term w_0 shifts the transition surface (so the decision regions) along the w vector, with the transition in prediction happening at vector v such that $v^T w + w_0 = 0$.

Under some conditions, the estimator built in this way is consistent, which means

that, if the data is truly generated by a model within the employed class of models (aka hypothesis space), then $\lim_{N\to\infty} w^* = w_{true}$. Therefore, the maximum likelihood estimator converges to the true vector of parameters, provided such a vector exists.

4.2 Training the Logistic Regression Models

Using the dataset described above, with the categorical variables one-hot encoded and the continuous variables standardized, two logistic regression models are trained, one for men and the other for women. Figure 1 displays the output of the logistic-regression model training for men, and Figure 2 does the same for women. From these outputs, several important insights can be derived about the factors that contribute to an individual earning more than $50,000.

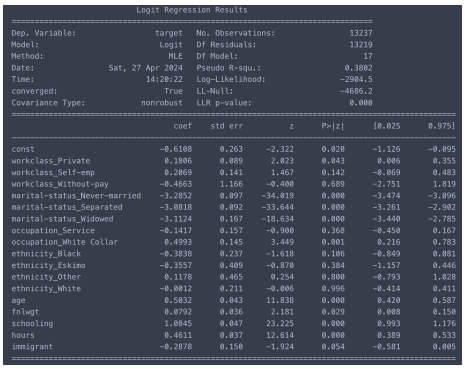


Logit Regression Results

Dep. Variable:	target	No. Observations:	27478
Model:	Logit	Df Residuals:	27460
Method:	MLE	Df Model:	17
Date:	Sat, 27 Apr 2024	Pseudo R-squ.:	0.2939
Time:	14:20:22	Log-Likelihood:	−12850.
converged:	True	LL-Null:	−17065.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	−0.9162	0.117	−7.806	0.000	−1.146	−0.686
workclass_Private	0.0846	0.047	1.803	0.071	−0.007	0.177
workclass_Self-emp	−0.1957	0.058	−3.379	0.001	−0.309	−0.082
workclass_Without-pay	−1.7356	1.099	−1.580	0.114	−3.889	0.418
marital-status_Never-married	−2.3852	0.059	−40.238	0.000	−2.501	−2.269
marital-status_Separated	−1.6560	0.059	−27.937	0.000	−1.772	−1.540
marital-status_Widowed	−1.2540	0.183	−6.837	0.000	−1.613	−0.895
occupation_Service	0.3060	0.045	6.847	0.000	0.218	0.394
occupation_White Collar	0.7949	0.041	19.401	0.000	0.715	0.875
ethnicity_Black	−0.1434	0.127	−1.133	0.257	−0.392	0.105
ethnicity_Eskimo	−0.5743	0.235	−2.447	0.014	−1.034	−0.114
ethnicity_Other	−0.3504	0.254	−1.381	0.167	−0.848	0.147
ethnicity_White	0.0213	0.107	0.200	0.842	−0.188	0.230
age	0.3993	0.019	20.640	0.000	0.361	0.437
fnlwgt	0.0812	0.016	4.982	0.000	0.049	0.113
schooling	0.7741	0.020	37.991	0.000	0.734	0.814
hours	0.3798	0.018	21.594	0.000	0.345	0.414
immigrant	−0.3555	0.071	−4.991	0.000	−0.495	−0.216

Figure 1: Summary of the training of a logistic regression model on the data frame of men



Logit Regression Results

Dep. Variable:	target	No. Observations:	13237
Model:	Logit	Df Residuals:	13219
Method:	MLE	Df Model:	17
Date:	Sat, 27 Apr 2024	Pseudo R-squ.:	0.3802
Time:	14:20:22	Log-Likelihood:	−2904.5
converged:	True	LL-Null:	−4686.2
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	−0.6108	0.263	−2.322	0.020	−1.126	−0.095
workclass_Private	0.1806	0.089	2.023	0.043	0.006	0.355
workclass_Self-emp	0.2069	0.141	1.467	0.142	−0.069	0.483
workclass_Without-pay	−0.4663	1.166	−0.400	0.689	−2.751	1.819
marital-status_Never-married	−3.2852	0.097	−34.019	0.000	−3.474	−3.096
marital-status_Separated	−3.0818	0.092	−33.644	0.000	−3.261	−2.902
marital-status_Widowed	−3.1124	0.167	−18.634	0.000	−3.440	−2.785
occupation_Service	−0.1417	0.157	−0.900	0.368	−0.450	0.167
occupation_White Collar	0.4993	0.145	3.449	0.001	0.216	0.783
ethnicity_Black	−0.3838	0.237	−1.618	0.106	−0.849	0.081
ethnicity_Eskimo	−0.3557	0.409	−0.870	0.384	−1.157	0.446
ethnicity_Other	0.1178	0.465	0.254	0.800	−0.793	1.028
ethnicity_White	−0.0012	0.211	−0.006	0.996	−0.414	0.411
age	0.5032	0.043	11.838	0.000	0.420	0.587
fnlwgt	0.0792	0.036	2.181	0.029	0.008	0.150
schooling	1.0845	0.047	23.225	0.000	0.993	1.176
hours	0.4611	0.037	12.614	0.000	0.389	0.533
immigrant	−0.2878	0.150	−1.924	0.054	−0.581	0.005

Figure 2: Summary of the training of a logistic regression model on the data frame of women

4.3 Driving Insights for Men

For continuous variables (e.g, *age*), the coefficients estimated by the logit models give the expected change in the log-odds of the outcome for a one-unit increase in the predictor variable. For binary predictors from one-hot encoding (e.g., *workclass_Private*), the coefficient represents the log-odds of earning more than $50,000 for individuals in the private work class versus the reference work class. Large coefficients, both positive and negative, indicate variables that have a stronger association with the likelihood of earning more than $50,000.

The $marital - status\_Never - married$, $marital - status\_Separated$, and $marital - status\_Widowed$ variables have large negative coefficients and are highly significant

($p < 0.05$), suggesting that being never married, separated, or widowed is associated with a lower likelihood of earning more than \$50,000 compared to their reference category, which is composed of married individuals.
The $occupation\_WhiteCollar$ coefficient is positive and significant, indicating that individuals in white-collar occupations are more likely to earn over \$50,000 than those in the reference occupation category, which is blue-collar jobs.
The $ethnicity\_Eskimo$ coefficient is significant and negative, suggesting that individuals identified as Eskimo are less likely to earn over \$50,000 compared to the reference ethnic group (which here is composed of Asian men, the ethnicity which generally earns more). The other ethnic groups do not seem to result in statistically significant differences with Asian men.
The variable $workclass\_Private$ has a positive coefficient, but it is not statistically significant ($p > 0.05$), suggesting that working in the private sector does not significantly increase the odds of earning more than \$50,000 compared to other work classes in the dataset.
The variable *age* is a continuous feature with a positive and significant coefficient, indicating that as men get older, they are more likely to earn over \$50,000. This likely reflects the fact that, as individuals age, their stock of human capital grows, and they collect the returns on human-capital investments made previously. Additionally, as workers become more senior, they typically receive more job-specific training, which further increases their productivity, and thus their wage. These views are strongly supported by economic research (Lagakos et al., 2014) (Stein and Yannelis, 2020).
The *schooling* coefficient is positive and highly significant, which is consistent with the notion that higher education levels can lead to higher earnings.
The pseudo R-squared value is relatively low, indicating that while the model is statistically significant, there are likely other factors not included here that influence whether an individual earns more than \$50,000.
The p-value for the likelihood ratio test is less than 0.05, which indicates that the model with predictors fits significantly better than an empty model (intercept only).

4.4 Driving Insights for Women

Regarding the insights for women, the value of the pseudo R-squared is 0.3802, higher than in the previous model for men, indicating that approximately 38.02% of the variability in the target variable is explained by the model, which is a relatively better fit compared to the model for men. Also the log-likelihood value is higher than that of men, which might indicate a better fit. The log-likelihood for the model with no predictors (other

than the intercept) is higher as well, but the overall model is still showing a significant improvement. The LLR p-value is less than 0.05, suggesting that the model provides a significantly better fit than an intercept-only model.
The variables $marital - status\_Never - married$, $marital - status\_Separated$, and $marital - status\_Widowed$ have strong negative coefficients, indicating these are strong predictors for not earning more than \$50,000, as it is the case for men as well.
The $occupation\_WhiteCollar$ variable has again a positive coefficient, suggesting that being in a white-collar job is associated with higher odds of earning more than \$50,000.
The coefficients for the variables indicating ethnicity are not significant, indicating that ethnicity might not be a particularly influential factor for determining the probability of women to earn more than \$50,000.
The coefficient for *age* is positive and highly significant, implying that older women are more likely to earn over \$50,000. This suggests that the acquisition of human capital that likely occurs with age is very important in determining women's wages.
As it is the case for men, also for women there is a strong positive association between the level of schooling and the likelihood of earning more than \$50,000.

5 Random Forest

5.1 Theoretical Background on Decision Trees

A decision tree is a model based on rules that can be employed to predict either continuous value (in the case of regression trees) or a categorical one (in the case of classification trees). Such rules are organized in a binary tree that divides the input space into disjoint regions, each with a constant prediction. The leaves of the tree (so the terminal nodes) contain the actual predictions, while each of the other nodes (the internal nodes and the root) contains a rule (e.g. $x_2 < 6$), and if this rule is satisfied, the algorithm goes left down the tree, otherwise it goes right. Let the total number of regions be T. The algorithm receives as input a vector x_P, and outputs a prediction for it, denoted as $\hat{y_P}$. Formally,

$$\hat{y}_p = \sum_{t=1}^{T} \hat{y}_t \cdot I_{x_p}(R_t)$$

which simply means that $\hat{y_P}$ is the $\hat{y}_t$ for t such that $x_P \in R_t$, with R_t indicating the t^{th} region. This formula is a sum over all terminal nodes (also called leaves). The algorithm computes one decision boundary for each internal node (root node included), and one region for each leaf node.

In order to get T and the regions (so, in order to build the tree) a heuristic algorithm is generally employed. The most popular choice is a recursive binary splitting algorithm, which involves building rules sequentially from the root to the bottom. This is a greedy algorithm, so one that does not optimize the whole tree, but just one little portion of it at the time, and picks the split that is currently the best one. Typically the rules are expressed in the form of "predictor $j <$ something". To determine the first split (at the root), the algorithm looks for two subspaces R_1 and R_2 defined as

$$R_1(j, s) = \{x | x_j < s\}$$

and

$$R_2(j, s) = \{x | x_j \geq s\}$$

where j indicates the splitting variable, and s is the cutpoint. The splitting variable is the variable for which the algorithm checks a condition, and the cutpoint is the actual condition. In the case of regression,

$$\hat{y}_t(j, s) = mean(y_i | x_i \in R_t(j, s))$$

while in the case of classification,

$$\hat{y}_t(j, s) = mode(y_i | x_i \in R_t(j, s))$$

So, the decision tree's prediction for the t^{th} region is the average or the mode of the outputs associated to each input that belongs to the t^{th} region.
In order to quantify the quality of the split (j, s), and to measure how close the predictions are to the actual data, a loss function is necessary. The most popular one for decision trees is the sum of squared errors, that in binary classification is the following function

$$L(j, s) = \sum_{i=1}^{N} (y_i - \hat{y}_1(j, s))^2 I_{x_i}(R_1(\hat{j}, s)) + (y_i - \hat{y}_2(j, s))^2 I_{x_i}(R_2(\hat{j}, s))$$

So, for each node, the algorithm picks the split (j, s) that minimizes L. Once both T and the regions are determined, the decision tree can be used to make predictions.

5.2 Theoretical Background on Random Forests

Random forests are a type of ensemble methods, which themselves are machine learning algorithms that combine instances of weak basic models to create a single, strong model. The set of such base models is called an ensemble. Random forests are ensemble methods that aggregate the predictive power of several decision trees to create a single robust model (hence the name "forest"). The idea is to train each tree slightly differently, and

then use a weighted average (in the case of regression) or majority vote (for classification) to combine their predictions. In the case of random forests, the ensemble set is composed of shallow trees, which are models with high variance and low bias in prediction. When combining them, the aim is to reduce their variance without increasing the bias. Each one of these base models is trained on a different subset of the available training data. The typical approach to select these subsets is bootstrapping, which means creating multiple different training sets of size N starting from the available N data points. This technique involves randomly sampling N elements with replacement from the available data B times, thus obtaining B random and identically distributed sets $S_1, ..., S_B$. Since the sampling is performed with replacement, the resulting sets will (almost surely) contain duplicates. Each so-obtained set is used to train one base model (so the random forest consists of B shallow trees). In the case of regression, the prediction for the overall random forest is the average of those for the base models, while for classification it is their mode.

Aggregating multiple predictions reduces the variance of the model. Indeed, this variance can be expressed as

$$\mathrm{Var}\left[\frac{1}{B}\sum_{i=1}^{B} z_i\right] = \frac{1-\rho}{B}\sigma^2 + \rho\sigma^2$$

where ρ is the average correlation between any two models z_i. The larger B, the smaller the variance of the full model. By increasing B the model does not risk to overfit, even though clearly, since each individual ensemble member can itself overfit, also the whole model can do that, but this effect is not caused nor worsened by increasing the number of bootstrapped sets. Anyway, the increase in B is done in order to reduce the variance of the model, not to obtain a better fit, as this would require getting more data, not sampling with replacement from available one.

If B is large enough, the reduction in variance is only limited by ρ (which is often quite small). To reduce this correlation ρ, random forests introduce additional randomness when creating each tree (hence the name "random" forests). Specifically, during the training, when splitting nodes, not every available p input variable is considered as the possible splitting variable, and instead the model picks a random input subset of size $q < p$ and only considers these q variables as possible splitting variables. At each splitting point, a new random subset of q variables is determined. The random subset selection is done independently for each of the B ensemble members. This ensures that the members of the ensemble are less correlated, so that ρ is smaller. Although the additional randomness has the side effect of increasing σ^2, in practice the reduction in correlation usually dominates over the increase in variance (Breiman, 2001).

5.3 Computing Feature Importance with Random Forests

The reason why random forests are employed in this work is to ultimately compute the importance of each feature in predicting the target variable for both men and women, and then compare such values. As stated above, each decision tree in a random forest is built by recursively splitting the training data into subsets based on the feature values that best separate the classes in the target variable. The choice of which feature and which value to split on is determined by a criterion that measures the uncertainty (or impurity) of the node being split. Typical metrics for computing impurity in the case of classification trees (the one relevant here) are the Gini impurity and the entropy. The former is defined as

$$G = 1 - \sum_{j=1}^{J} p_j^2$$

and the latter is

$$H = -\sum_{j=1}^{J} p_j \log(p_j)$$

In both of these formulae, p_j is the percentage of items of class j in the set.
After the trees are constructed, the importance of a certain feature can be computed based on how effectively splits on this feature reduce impurity across the trees. This process starts by recording, for each tree in the random forest, the reduction in impurity due to a split on a certain feature every time that feature is used. This reduction in impurity is calculated as

$$\Delta I = I_p - (p_l I_l + p_r I_r)$$

where I_p, I_l, and I_r are the impurities of the parent, left child and right child node after the split, respectively, and p_l and p_r are the proportions of samples that go in the left and right child nodes, respectively. Then, the importance of a certain feature is calculated by averaging the decreases in impurity ΔI's over all splits that involve this feature across all trees in the forest and normalizing by the total decrease in impurity for all features. The importance values calculated using individual trees are subsequently joined (usually by averaging) to arrive at a single importance measure for each feature in the random forest model. These values are often normalized so that the sum across all features equals one, making it easier to compare the relative importance of predictors.

5.4 Feature Importance for Men and Women

In this work, two random forest models are trained, one for men and one for women. The data frames used encode categorical variables with the simple-encoding technique,

as one-hot encoding is not necessary for this kind of models. After the training, the models are used to obtain and compare the importance of each feature both for men and women, following the procedure outlined above. Moreover, for each feature, the difference in importance between men and women is computed, along with the average of such two values. The results are shown in Figure 3, where the features are sorted in ascending order based on their average importance. From this output, some factors are worth mentioning.

	Feature	Importance_Men	Importance_Women	Difference	Average
2	fnlwgt	30.259871	24.680234	5.579637	27.470053
0	age	21.641110	19.845440	1.795670	20.743275
4	marital-status	11.757529	21.264358	-9.506829	16.510944
3	schooling	14.110531	12.741629	1.368901	13.426080
7	hours	10.680897	11.250375	-0.569477	10.965636
5	occupation	5.836825	3.538922	2.297903	4.687874
1	workclass	2.959744	3.133205	-0.173461	3.046474
6	ethnicity	1.750908	2.335046	-0.584137	2.042977
8	immigrant	1.002584	1.210791	-0.208207	1.106687

Figure 3: For each feature, this table shows its relative importance for men, relative importance for women, the difference between these two values, and the average importance

First, $fnlwgt$ (final weight) has the highest importance in both models but more so in the male one. The fact that this feature is significant is not surprising. Indeed the final weight, as already said, represents the number of people the census believes that entry represents. The fact that its importance is greater for men than for women might be capturing socioeconomic factors tied to job sectors and demographics that are more varied for men.

The *age* feature is the second most important predictor in both datasets, with slightly more importance for men. As already stated, this is likely to indicate the positive correlation between human capital (derived from experience and specific on-the-job training) and earnings. The slightly higher importance of age for men could suggest a steeper age-earnings profile for men than for women.

The $marital - status$ variable displays a significant difference in importance between men and women. Indeed, its relevance is 12% for men and 21% for women. The fact that this feature is hugely more important for women reflects societal expectations or economic dependencies tied to marital status that differentially affect women. This is a strong indication of discrimination, as it shows that the earning possibilities for women are much more dependant on their marital status than they are for men. Given this result, programs that support single or divorced women might be particularly impactful

for increasing social justice and equity.
The variable *schooling* is quite important for both groups, although slightly more for women. This fact may suggest that educational qualifications could be a key lever for increasing women's earnings potential, and thus that investing in education might yield higher economic returns for women, likely due to a combination of market valuation of skills and educational attainment.
The *occupation* predictor is the forth less relevant variable in both groups, but it is more important for men, which might display occupational segregation where the type of occupation has a stronger link to high earnings for men. Therefore, policies to encourage women in high-paying fields or to mitigate gender stereotypes in career choice could be beneficial.
The other variables, namely the number of weekly hours worked, the work class, the ethnicity, and the country of origin have almost the same importance for both men and women in determining their wage.

6 Fairlie Decomposition Analysis

6.1 Theoretical Background on Fairlie Decomposition Analysis

The Fairlie decomposition is a statistical technique developed by economist Robert Fairlie (Fairlie, 1999), and it is used for understanding differences in an outcome variable between two groups (in this case, men and women). It is an adaptation on the Oaxaca-Blinder decomposition for nonlinear models (e.g., the logit model employed in this work), and for cases in which some predictors are categorical (as it is the case here). The main goal of the Fairlie decomposition is to quantify the extent to which differences in observed characteristics (e.g., age, schooling, or hours of work) between the two groups contribute to the differences in an outcome (high-income attainment in this case). This method has been used extensively in labor economics to dissect the explained and the unexplained difference in binary outcomes between two groups (Fagbamigbe et al., 2021) (Kumar et al., 2021).
The first step to perform the so-called Fairlie Decomposition Analysis (FDA) is to fit a logistic regression model using data for the dominant group, men in this case. Then, it is necessary to shuffle the order of the observations in the disadvataged group (women in this work). This step, which is performed many times (1000 in this work), is crucial, as it removes any order bias and ensures that the results are not dependent on the original order of data. Next, for each different shuffle (so, 1000 times here), the following is done

for each feature j of the d features in the data frame ($d = 17$ in this case, as variables are one-hot encoded):

> The values of the j^{th} column of the men data frame are replaced with the values of the same feature in the women dataset. Then, the probabilities of the target value to be 1 are computed with this modified data frame, and the difference between the probabilities computed with the original men data frame and these new probabilities is calculated. Subsequently, such differences are averaged and stored. This value indicates the impact of feature j on the outcome.

After the d mean differences in probabilities are stored, another average of them is computed (this time it is the sum of the d mean differences over d), and recorded again. After 1000 steps, an average of these values is computed (this time it is the sum of 1000 values over 1000). This final number denotes how much of the difference in outcomes between the groups can be explained by observable characteristics. Specifically, a value of x means that, according to our model and data, the observed characteristics account for x of the difference in the average probability of earning more than \$50,000 between men and women. Any remaining difference in the predicted probabilities is attributed to unexplained factors, which may include discriminatory practices.

6.2 Implementation and Analysis of Results

An FDA of the available data is implemented, and the value obtained is approximately 0.0089828, indicating that the difference, between men and women, in the probability of earning more than \$50,000 a year which is attributable to observed features is only about 0.9%. The total difference in the probabilities is 19.88%, which means that a difference of 18.98% is left unexplained, possibly due to discrimination.

7 Conclusion

This work investigated the impact of personal variables (such as educational attainment or weekly hours worked) on the wage differential between men and women. Rigorous mathematical techniques was employed to evaluate the importance of the predictors in determining whether a certain individual earns more then \$50,000 a year or not. Among the most interesting results is the fact that the marital status is a significantly

stronger predictor of income for women than for men, which displays huge discrimination intrinsic to society, as it reflects societal expectations or economic dependencies tied to marital status that affect women much more than men. Additionally, the difference between men and women in the probability of earning more than $50,000 a year was discerned into an explained and an unexplained part, and it was concluded that the latter significantly dominated the former. While this result simply shows that there are latent relevant variables that explain wages, it is also likely to indicate the possibility of severe discrimination against women. An important fact to consider is that, while these models can show disparities in feature importance, they don't necessarily explain why these differences exist. Further qualitative analysis or more detailed quantitative analysis might be required to delve into underlying causes.

8 References

Lagakos, D., Moll, B., Porzio, T., Qian, N. and Schoellman, T. (2012). *Experience Matters: Human Capital and Development Accounting.* National Bureau of Economic Research Working Paper Series

Stein, L. C. and Yannelis, C. (2020). *Financial inclusion, human capital, and wealth accumulation: Evidence from the Freedman's Savings Bank.* Review of Financial Studies, vo. 33 (11), pp. 5333–5377

Breiman, L. (2001). *Random forests.* Machine learning, vo. 45 (1), pp. 5-32

Fairlie, R. W. (1999). *The absence of the African-American owned business: An analysis of the dynamics of self-employment.* Journal of labor Economics, vo. 17 (1), pp. 80-108

Fagbamigbe, A. F., Oyinlola, F. F., Morakinyo, O. M., Adebowale, A. S., Fagbamigbe, O. S. and Uthman, A. O. (2021). *Mind the gap: what explains the rural-nonrural inequality in diarrhoea among under-five children in low and medium-income countries? A decomposition analysis.* BMC Public Health, vo. 21, pp. 1-15

Kumar, P., Rashmi, R., Muhammad, T. and Srivastava, S. (2021). *Factors contributing to the reduction in childhood stunting in Bangladesh: a pooled data analysis from the Bangladesh demographic and health surveys of 2004 and 2017–18.* BMC Public Health, vo. 21