Gender Discrimination in the Tech Industry

Hugo David Franco Ávila

A01654856

Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Querétaro

Intelligent System Technologies

Dr. Benjamín Valdés

Abstract

Gender Discrimination is a problem that has affected society for centuries. Women are harassed, paid less, evaluated more harshly, or passed over for promotion because of their gender [1]. This is prevalent over any industry but especially notable in Tech, where salaries have increased steadily over the past decade [2], but the wage gap between men and women stays the same. In this project, by using data from a Glassdoor study on salaries that includes data on gender, position, department, scholarity, base and bonus pay; total compensation (TC), which is calculated as the sum of base and bonus pay for any given year, is predicted by fitting a model to the data using linear regression, and the gap between salaries of men and women can be seen as the difference between the coefficients for both. More data is needed on the subject because the difference in compensation could also be due to religion, race, sexual preference or other factors, but a clear wage gap is visible between men and women.

Keywords: Python, SKLearn, Machine Learning, Regression, Gradient Descent, Discrimination, Gender, Men, Women

Resumen

La discriminación de género es un problema que ha afectado a la sociedad por siglos. Las mujeres son acosadas, pagadas menos, evaluadas de forma más dura, o desestimadas cuando hay posibilidad de una promoción en el trabajo. Esto se ve en todas las industrias pero es especialmente notable en la industria Tecnológica, donde los salarios han incrementado gradualmente durante la última década, pero la brecha salarial entre hombres y mujeres se mantiene igual. En este proyecto, utilizando datos de un estudio de Glassdoor sobre salarios que incluye datos sobre género, posición, departamento, escolaridad, salario base y bono; la compensación total, la cual es calculada sumando el salario base y el bono, se calcula con un modelo de regresión lineal simple que se aplica a los datos, siendo la diferencia entre los coeficientes de hombres y mujeres la brecha que existe entre los salarios. Más datos son necesario ya que la diferencia en compensación puede ser debido a la religión, raza, preferencia sexual u otros factores, pero es claramente visible la brecha salaria entre hombres y mujeres.

Palabras clave: Python. SKLearn, *Machine Learning*, Regresión, Descenso de Gradiente, Discriminación, Género, Hombres, Mujeres

Gender Discrimination in the Tech Industry

Machine Learning can be used to get answers to problems that were previously thought to be unsolvable. One of the biggest problems of the 21st century is the existence of the "wage gap", which is defined as "… the difference in earnings between women and men." [3] While new legislation has passed to be more inclusive of women in the workforce, there still exists some resistance due to the fact that a large portion of the male population doesn't believe the gap is real [4]. To be able to look over this issue more clearly, it is crucial to go through some concepts first.

Machine Learning

According to IBM [5], Machine Learning is a form of Artificial Intelligence (AI) that allows a system to learn from data as opposed to learning from what it was explicitly programmed to. Machine Learning is implemented through different algorithms applied to a set of data, which results in a model. A model is what the system uses to make predictions, either for a value, or a label. That is known as a regression task, or a classification task, respectively.

Linear Regression

Linear regression is a basic and commonly used type of predictive analysis [6]. A linear regression model has an equation of the form $Y = a + bX$, where X is the independent variable and Y is the dependent variable (figure 1). With this equation we can make predictions for input data that exists within the original dataset or make predictions using any other set of values.

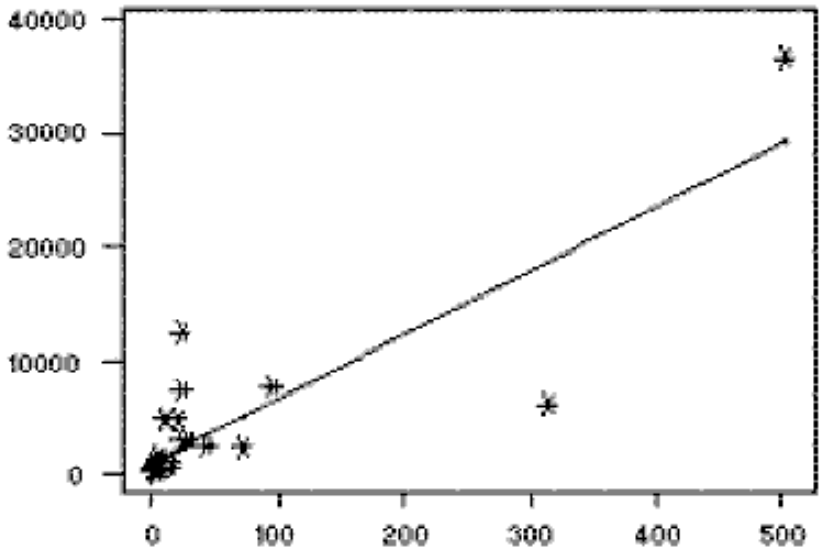

Figure 1. Linear Regression

Gradient Descent

Gradient descent (GD) is an iterative first-order optimization algorithm used to find a local minimum/maximum of a given function. This method is commonly used in machine learning (ML) and deep learning (DL) to minimize a cost/loss function [7]. The "size" of the steps the algorithm takes into the direction of the local minimum are determined by the learning rate, which figures

out how fast or slow we will move towards the optimal weights, or parameters. How the cost is reduced after every iteration is easily observed, as seen in figure 2.

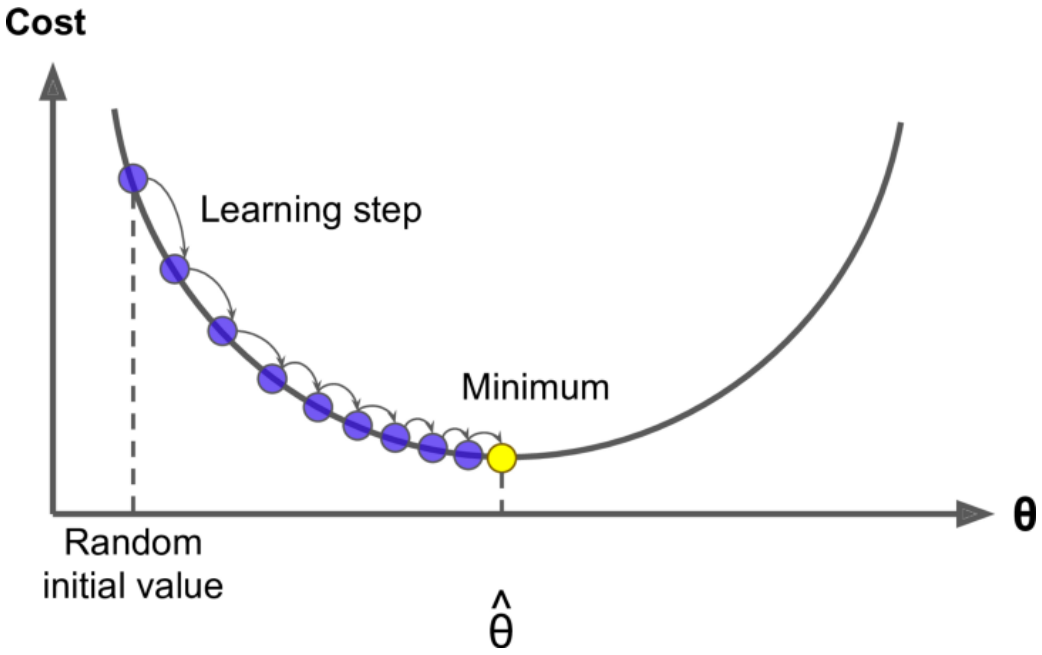

Figure 2. Gradient Descent Optimization

Gender discrimination

Discrimination based on gender (or sex) is a common civil rights violation that takes many forms, including sexual harassment, pregnancy discrimination, and unequal pay for women who do the same jobs as men. [8]. Recent research has shown that although women now enter professional schools in numbers nearly equal to men, they are still substantially less likely to reach the highest echelons of their professions. This lack of success in climbing the professional ladder would seem to explain why the wage gap remains largest for those at the top of the earnings distribution. [9]

Description of the problem

The objective is to be able to predict TC for tech industry professionals, and in the process, observe the difference between the wages of men and women.

Solution

To make the application accessible to multiple users, I used Google Cloud Platform's Cloud Functions, which allows the project to be deployed as a function that can be called many times via the HTTP protocol with different values to get the prediction.

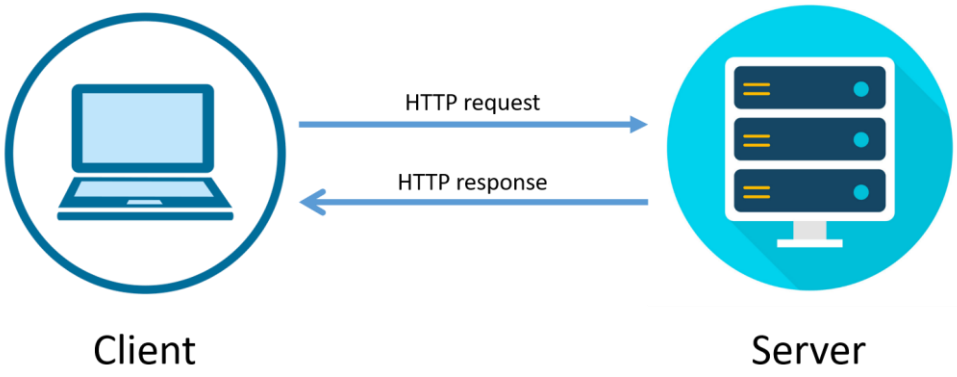

Figure 5. Communication between a client and a server

As for the predictions, I ran Gradient Descent on the dataset from Glassdoor, using the value to be predicted or Y value the TC (base + bonus pay), and for the parameters or X values I used the columns of Job Title, Performance Evaluation, Education, Seniority and Gender. For the columns with categorical value (Job Title, Education and Gender), One-Hot Encoding through Pandas get_dummies function was used.

After running the algorithm for 10000 epochs I obtained a coefficient of determination (R^2) of 0.27, which is a little low, but can be explained given that data isn't perfect and even with a moderately large dataset of 999 instances, more data is needed to be able to overcome bias introduced by larger and smaller values.

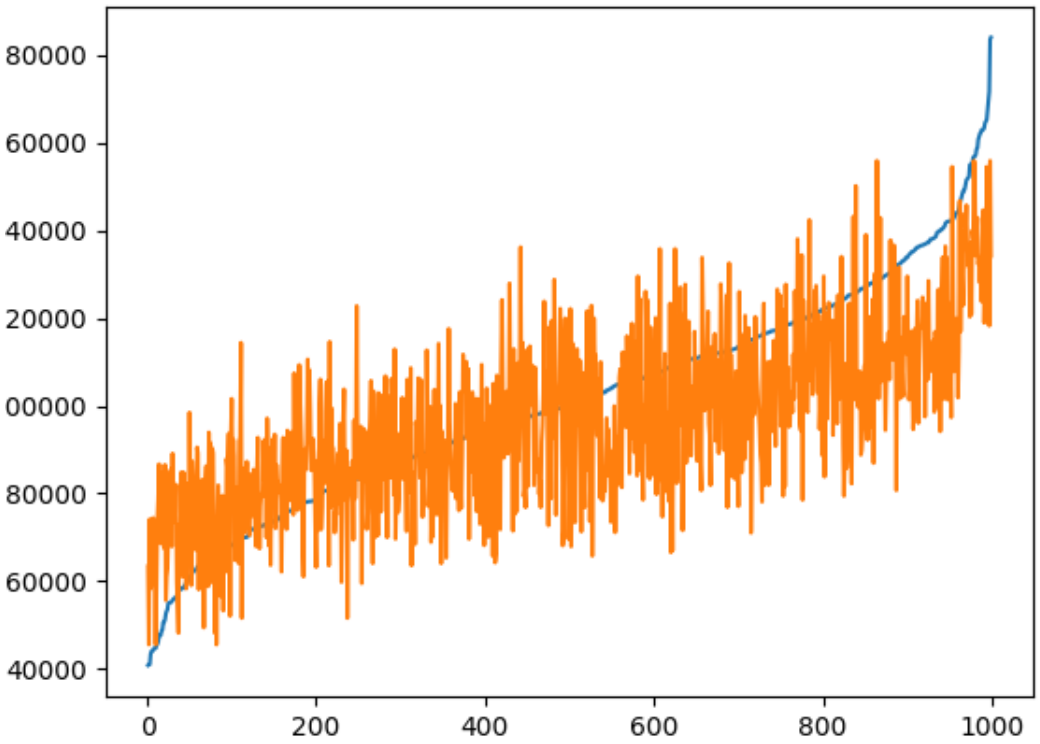

Figure 6. The predictions from Gradient Descent vs the real data

In the next table, we have the column and the coefficient the model assigned to it, which gives us some interesting results:

Parameter	Coefficient
Performance Evaluation	15474.64192
Seniority	40434.82855
Female	41070.8989
Male	40333.32156
College	18848.58522
High School	16208.38989
Masters	22243.33923
PhD	24105.90612
Data Scientist	4164.772927
Driver	1545.66504
Financial Analyst	7713.376001
Graphic Designer	4867.980804
IT	3660.479538
Manager	39414.38519
Marketing Associate	-11723.44287
Sales Associate	7510.605981
Software Engineer	19310.26272
Warehouse Associate	4948.135132

The coefficient for “Female” is slightly higher than the coefficient for “Male”. Also of note the “Marketing Associate” job has a negative coefficient, which would indicate Marketing Associates get impacted negatively in their compensation with respect to other professions.

License

The project was released under the MIT License, which means anyone can use and distribute the software at will. I am however, cleared of any wrongdoing and I do not offer a warranty for the software.

Conclusion

Gender discrimination is a problem that is real and needs to be addressed. While there are many studies that confirm the wage gap is real, in this project with the coefficients obtained we can’t directly prove it, however, it is important to note that a major flaw of our methodology, is using data from different professions, as well as different education levels. If all the women in the study were Managers with PhD’s and the men were Graphic Designers with only their High School degree, we would get a completely unreasonable result, that men are heavily underpaid in relation to women. For future studies, it is recommended to get more data and to try and eliminate variables other than gender.

The first step to solving the issue is recognizing there is a wage gap, and creating new legislation that helps shorten it, eventually making it disappear. Racism and discrimination are problems that are ugly, uncomfortable, but that need to be discussed and addressed.

References

[1] Equal Rights Advocates (n. d.). Discrimination At Work. Equal Rights. Retrieved from https://www.equalrights.org/issue/economic-workplace-equality/discrimination-at-work/

[2] Stone, B. (2022, February 1st). Dice: Tech salaries increased in 2021. TechRepublic. Retrieved from https://www.techrepublic.com/article/dice-tech-salaries-increased-in-2021/#:%7E:text=Despite%20the%20concerns%20with%20returning,employee%20from%202020%20to%202021.

[3] Quick Facts About the Gender Wage Gap. (2022, January 19th). Center for American Progress. Retrieved from https://www.americanprogress.org/article/quick-facts-gender-wage-gap/

[4] Renzulli, K. A. (2019, April 4th). 46% of American men think the gender pay gap is «made up to serve a political purpose». CNBC. Retrieved from https://www.cnbc.com/2019/04/04/46percent-of-american-men-think-the-gender-pay-gap-is-made-up.html

[5] IBM. (2021, September 9th). Machine Learning. IBM. Retrieved from https://www.ibm.com/mx-es/cloud/learn/machine-learning

[6] Statistics Solutions. (2021, August 10th). What is Linear Regression? Retrieved from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

[7] Kwiatkowski, R. (2022, January 6th). Gradient Descent Algorithm — a deep dive - Towards Data Science. Medium. Retrieved from https://towardsdatascience.com/gradient-descent-algorithm-a-deep-dive-cf04e8115f21#:%7E:text=Gradient%20descent%20(GD)%20is%20an,e.g.%20in%20a%20linear%20regression.

[8] FindLaw. (n. d.). Gender Discrimination - FindLaw. Retrieved from https://www.findlaw.com/employment/employment-discrimination/gender-discrimination.html

[9] Yellen, J. L. (2021, January 6th). The history of women's work and wages and how it has created success for us all. Brookings. Retrieved from https://www.brookings.edu/essay/the-history-of-womens-work-and-wages-and-how-it-has-created-success-for-us-all/