The Minority Report

Christopher Yi, Jim Banya, Shafiullah Hashimi, Rohit Barua

CS 484-002, GMU Spring 2023

1 Project Motivation

Our selected problem is to study how to detect and correct against discriminatory hiring and recruiting practices. As a cohort of senior CS students from various minority backgrounds, this is of particular relevance to us. As we are all on the cusp of entering a very competitive job market, the "what-ifs" surrounding the eventual outcomes of our search for employment looms over us. In a socioeconomic climate where racial issues in particular have taken center stage in the public consciousness, we feel it would be beneficial to examine discriminatory practices on a broader scale and see if data mining techniques can help examine these lingering issues. To do so, we endeavored to examine existing data sets with various demographic information and to examine any relationships between them, be it whether job performance metrics are a good predictor for a listed racial group, whether non-subjective metrics can be a good predictor for compensation, and so on.

2 Contributions

Christopher compiled the results of the entire group's individual efforts into this report, as well as adding the cursory historical contextual information to the Jupyter notebook. He also examined a simulated HR data set using KNN classification, as well as as much larger data set of salaries across the largest U.S. tech companies through multiple linear regression.

3 Method

Christopher chose to examine two separate data sets through two lenses: classification and regression. The first data set was examined using the K-nearest neighbors algorithm, or KNN:

$$\hat{y} = f(x) = y_{nn}(x)$$

$$nn(x) = \arg\min_{n=1,\ldots,N} ||x - x_n||_2^2 = \arg\min_{n=1,\ldots,N} \sum_{d=1}^{D} (x_d - x_{nd})^2$$

The theory is that any chosen features will prove to be a poor predictor for an output categorical feature (in this case race or sex). The general algorithm would be:

1. Preprocess data and choose relevant input features.
2. Separate features and categorical labels into two matrices.
3. Further separate these features and labels into test and validation submatrices.
4. Calculate Euclidean distance row by row between validation sample and training feature matrix.
5. Select arbitrary K lowest distances. Their indices (arg min in above formula) correspond to the closest labels.
6. Evaluate results.

The second data set was examined through the lens of a multiple linear regression model.

$$f(x) = w_0 + \sum_{d=1}^{D} w_d x_d = w^T x$$

Each of the weights beyond the base term constant w_0 correspond to one of the input features used to predict the output feature y. This function will be evaluated with the goal of minimizing the difference between the predicted labels (the projected linear equation) and actual labels. These differences, or residuals, can be squared and summed to become a residual sum of squares, or RSS. The equivalent expression using matrices is:

$$RSS(w) = (y - Xw)^T (y - Xw)$$

The list of weights w can be evaluated and selected by minimizing this loss function:

$$w^* = \arg\min_w RSS(w)$$

Further, the ordinary least squares (OLS) method can be used to estimate the actual weights themselves. The expansion and derivation (with respect to w) of the RSS function yields this expression: $w = (X^T X)^{-1} X^T y$. All told, the general algorithm using all of this information is:

1. Preprocess data and choose relevant input features.
2. Separate features and categorical labels into two matrices.
3. Further separate these features and labels into 5 submatrices for 5-fold cross validation.
4. Calculate list of weights w for each fold.

5. Calculate RSS for each list of weights.
6. Choose minimum RSS and its corresponding list of weights w.
7. Calculate predicted labels by dot product of weights and augmented feature matrix.
8. Evaluate results.

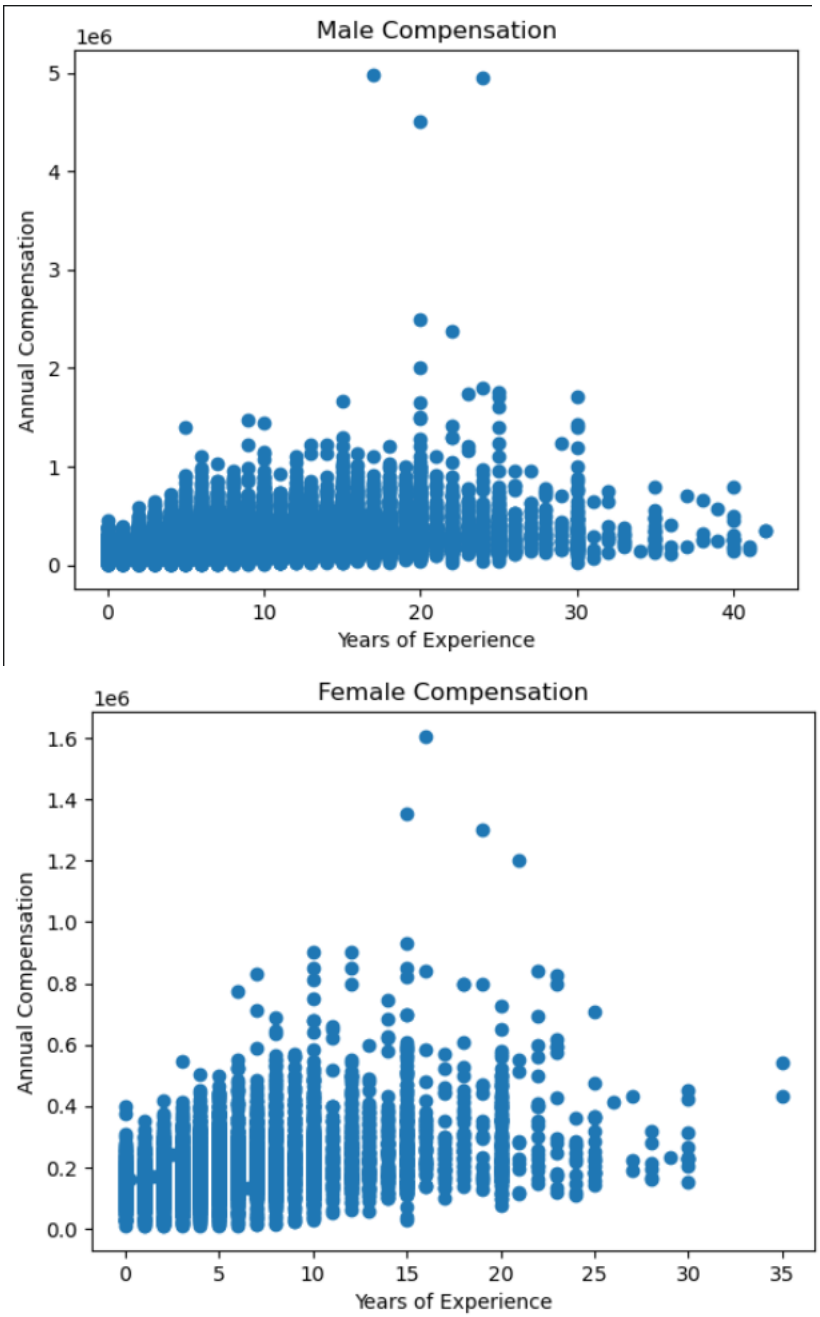
4 Experiment (Chris)

Christopher's first experiment was on a simulated HR data set with multiple columns with little significance, so a number of post-processing steps were necessary. First, only purely quantitative features were saved. In order to preserve discriminatory identifiers, gender and race were preserved by converting their string identifiers to integers. This left the employee's performance scores, engagement, race, and gender. Race was chosen to be the categorical ground truth label, and served as the buckets to which the closest neighbors in the KNN algorithm would be classified as. Then, the KNN algorithm described in "Methods" was applied to the remaining features after consolidating them into a 1:4 split into validation and training sets.

The metrics used to analyze the neighbor classifications were accuracy and micro/macro F1. An accuracy/micro F1 of 38.9% and a macro F1 of 19.2% seemed to indicate that this data set demonstrated little to no relationship between the features and race. While this had little explanatory power, particularly given it was a simulated data set, it was more a proof of concept / test run for the real data to follow.

Christopher's second experiment used a real data set: tech sector salaries of the largest U.S. companies collected from 2017 to 2021. This time, a linear regression modeling technique was used. The preprocessing of the data was more involved since there were so many more samples, and many of the responses included null responses. Culling the data meant losing a significant portion of the data, but unfortunately there was little that could be done given many of the unknown responses involved identifying features, i.e. race and gender. In addition, some non-tech occupations were culled from the data as well.

Before proceeding with the linear regression modeling, scatter plots were created to see if any relationship between years of experience and compensation could be seen visually. After creating a separate data frame centered around binary gender identifiers, the following plots were created:

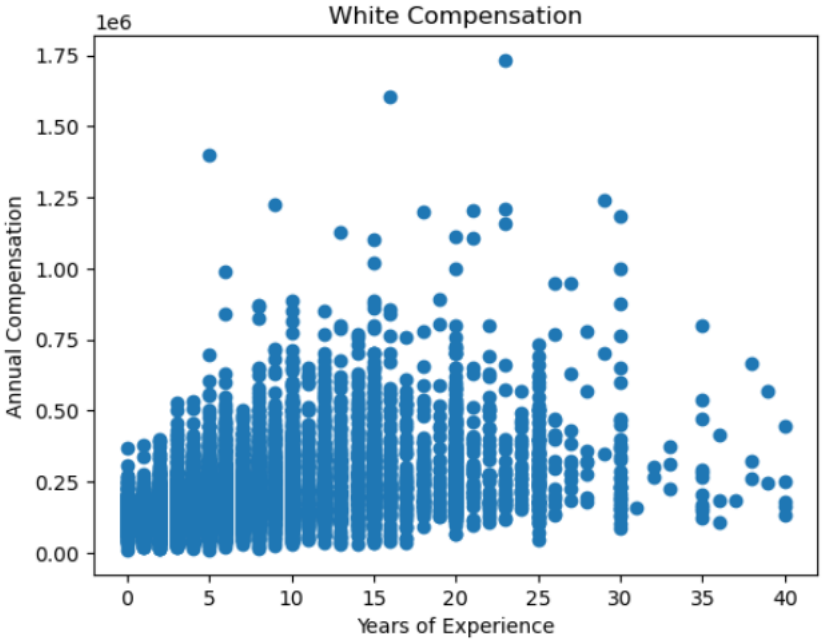


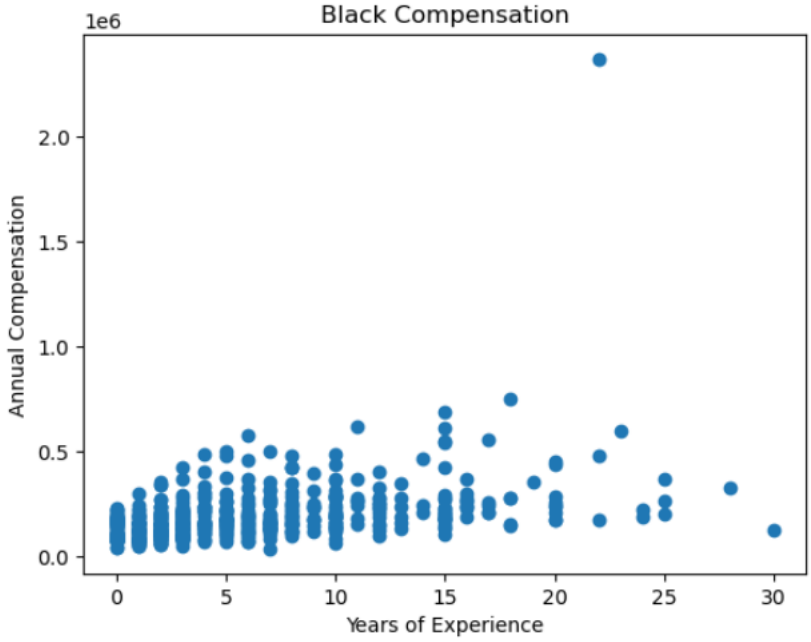
At first glance, there seems to be little to no correlation between experience and compensation across both genders. And if anything, female employees seem to have slightly higher salaries overall. However, given that female employees are very under-sampled compared to male employees (6,252 to 34,749), it is

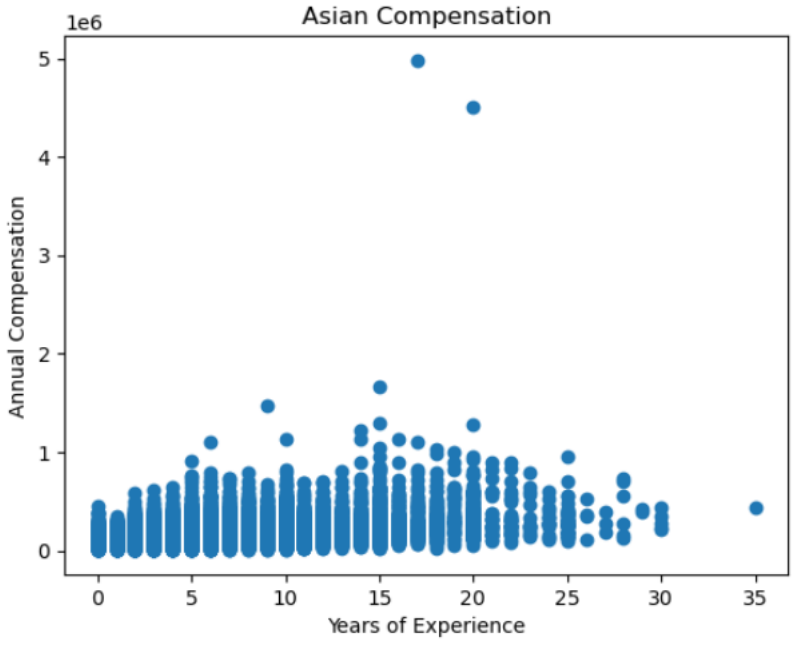
possible that the fewer female employees are comparatively more adept.

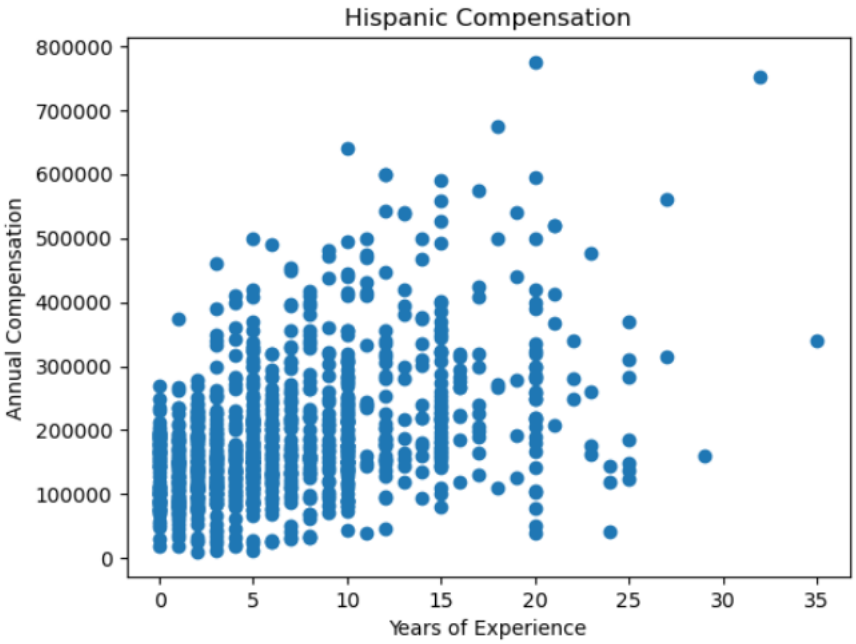
This gender-based data frame was then split, using compensation as labels and a list of male/female identifiers, years of experience, and years at the current company as features. The weights yielded by the fold with the lowest RSS were 57425.51618, 50944.33253, 10002.95321, and -1561.86543. It seems that years at the company are slightly negatively correlated, while the gender identifiers are weighted significantly higher than years of experience, the only other objective metric in the chosen set of features.

A second run with linear regression was attempted, this time on an additional data frame created around race identifiers instead. Again, scatter plots were created to model the relationship between experience and compensation separately across single-race identifiers:




Black Compensation
1e6
2.0
1.5
1.0
0.5
0.0
Annual Compensation
0
5
10
15
20
25
30
Years of Experience


Asian Compensation
1e6
5
4
3
2
1
0
Annual Compensation
0
5
10
15
20
25
30
35
Years of Experience



Again, there seemed to be little to no correlation that can be obviously discerned. After again applying 5-fold OLS, the weights yielded with the minimum RSS were 28983.65744, 32572.50561, 35057.71418, 17732.38652, 9038.66236, and -3784.53103. Once again, years at their company is slightly negatively correlated. The racial identifiers