DRIVING TOWARDS WAGE EQUITY WITH MACHINE LEARNING:

Predicting Annual Wages by Job Title

CRAIG T. CHOSNEY

A capstone project submitted in partial fulfillment for the degree of

MASTER OF SCIENCE in QUANTITATIVE METHODS IN THE SOCIAL SCIENCES

CUNY GRADUATE CENTER

NEW YORK, NEW YORK

FALL 2024

I. RESEARCH PURPOSE

The genesis of this project was a conversation I had with an HR manager about designing a compelling HRIS and data science adjacent project to address some challenges and needs within their department and the human resources field. They told me that a practical utility for them would be a standardized system to assist with their organization's annual compensation increases that could accommodate inflation, market rate changes, and individual tenure. With this conversation in mind, I conceptualized this project as a way to use data science and analytics techniques to model future compensation for a workplace organization and create an annual deliverable that a human resources department may use as guidance for such compensation increases.

Thus, the purpose of this research is to explore the differences in annual compensation between the internal data of a mid-size NYC cultural institution and national external market job valuation data and to construct a machine learning model that models future compensation per job title for the mid-size cultural institution accounting for inflation, market rate wage changes, and tenure. Ultimately, it is my hope that such an exploration of local versus national wage parity may promote a more equitable wage dispersion within the cultural institution and society at large.

II. LITERATURE REVIEW

As previously stated, this project will explore the gap in wage compensation between the internal data of a mid-size NYC cultural institution and external market job valuation data and construct a machine learning model utilizing this wage gap to predict future compensation per job title for the institution. This modeling is predicated on work that many social scientists have undertaken to understand and explain the theoretics of wage equity and inequality. This literature review attempts to connect this gap in wage compensation to its modeling through an examination of the conception of wage inequality as a function of both human capital and cumulative advantage and disadvantage.

A. WAGE EQUITY, DISPERSION, AND INEQUALITY

The sociologist's understanding of wage as a unit of both equity and inequity can be traced back to the writings of Karl Marx, who in *Wage Labor and Capital* writes simply "that wages are the amount of money which the capitalist pays for a certain period of work or for a certain amount of work" (22). Marx elucidates that within our economic system of capitalism the individual must exchange their labor-power for commodities including "wage," Marx's primary commodification of this labor-power: "*Wages* therefore are only a special name for the price of labor-power, and are usually called the price of work; it is the special name for the price of this peculiar commodity, which has no other repository than human flesh and blood" (23). Wages are important to the individual within the framework of capitalism simply because the wage worker must sell their labor-power to the capitalist in order to live (24). Within Marx's capitalism the laborer is removed from the "means of subsistence." This removal of the individual laborer from their means of subsistence may have seemed quite novel in 1849, the time of Marx's writing of *Wage-Labor and Capital* and of the shift from agrarian living to industrialization. Today, however, in an age of post-industrialization these facts of this removal seem quite banal. For most of us, if we are hungry and want to eat a salad, we cannot simply get one from the land. We have to walk to Sweetgreen or the supermarket and buy one, most likely with the money that we earned by exchanging our labor-power for wages. It is in this exchange of “wages” for the “means of subsistence” that the power of wage as a unit of both equity and inequity is produced.

In contemporary convention, pay equity is the requirement of “equal pay” for “equal work,” meaning that people across all categories of gender, race, age, sexual orientation, disability, or any other marginalizing classification should get the same pay for doing identical jobs. Pay equity can be further distinguished into the principles of wage parity and comparable worth. Wage parity refers to equal wages for all peoples doing the same work, and comparable worth refers to equal wages for doing different work with similar skills, qualifications, working conditions, and levels of responsibility (Gupta, Singh, and Balcom 2022: 126).

Equitable pay has many benefits that contribute to the intrinsic motivations of a worker. Intrinsically, pay equity is linked psychologically to increased worker morale, productivity and group cohesiveness. There is a reciprocal nature between employee and employer; “thus, employees will put forth more effort when they feel they are receiving a fair wage” (Leete 2000: 426). Furthermore, wage compensation constitutes an effective reward system for employees. Failla, et al. write that “reward systems are closely associated with perceptions of equity, which are often driven by social comparisons among colleagues. Rewards that are perceived as inequitable can result in declining job satisfaction, reduced effort, counterproductive work behavior, and… employees quitting their jobs” (2024: 1).

Despite the clear benefits of wage equity to both employees and employers, inequity can and often exists within an organization's dispersion of wages. Wage dispersion within an organization exists as either horizontal or vertical. Horizontal wage dispersion describes dispersion within an organization's hierarchical levels, and vertical wage dispersion describes dispersion across an organization's hierarchical levels. Simply, vertical dispersion would describe differing wages between a manager and subordinate. Horizontal dispersion describes wages amongst members of an organization's single hierarchical level differing due to the conflation of easily observable characteristics such as tenure and difficult to observe factors such as individual productivity (Failla, et al. 2024: 2). Inequity can be present in either the natural vertical or horizontal dispersion of an organization's wages. Take a gender pay gap in a healthcare organization for example. Inequity could present horizontally as female doctors earning less than their male counterparts. Vertical inequity could present itself as nurses making less money than doctors, with a greater proportion of nurses being female and doctors being male. This gender pay inequity is not necessarily a product of de facto pay discrimination, but rather a function of systemic societal devaluation of certain occupations. Gupta, et al. (2022) write, "that female-sterotypic internalized values (oriented towards caring for others) tend to be less valued in society compared to male-stereotypic values (focusing on status, competition, and upward mobility)" (126-7). Within our example of horizontal gender pay inequity between doctors, this could present as pediatricians, a medical speciality oriented towards caring for others, earning less than cosmetic plastic surgeons, a medical specialty focusing on status, competition, and upward mobility.

A final way that inequity can present in wage capital is longitudinally through the life course process of stratification. Cheng (2021) presents evidence of a cohort approach to inequality that sees different birth cohorts experience differing life course wage patterns of wage inequality. Cheng writes, “the specific historical and social contexts that shape individuals’ career paths and life experiences may affect how their life course trajectories unfold, and this in turn translates into the cohort shifts in the patterns of aggregate inequality” (2). In Cheng’s example, earlier cohorts that experienced the labor market before “the breaking down of internal labor markets along with the growing organizational and institutional flexibility in employment from the late 1970s” experience a lesser extent of wage dispersion than later cohorts entering the labor market following the onset of these trends (5). This cohort approach adds a needed temporal aspect to our conceptualization of wage inequality.

B. HUMAN CAPITAL AND THE HUMAN CAPITAL EARNINGS FUNCTION

Much of the analysis of earnings and their distributions owes a debt to the contributions of Jacob Mincer and the development of the “human capital earnings function.” Prior to Mincer’s work, economics literature was interested in the inequality and shape of of the distribution of income or earnings, but “the literature in this area considered the consequences of combining separate distributions of ability or considered the effects of purely random or stochastic events as a determinant of the distribution of income” (Chiswick 2003: 345). Mincer is credited with pioneering a model of the effect of labor market experience or on-the-job training on the determination and distribution of earnings based on rational economic behavior by individuals in the labor market in his “path breaking” doctoral dissertation and the seminal 1958 article, “Investment in Human Capital and Personal Income Distribution” (345). Mincer’s work primarily concerns the relationships between education, on-the-job training, and earnings. Mincer found that “overall school and experience investments are positively correlated across individuals,” demonstrating an empirical positive relationship between these two variables (346).

Gary Becker and Barry R. Chiswick build upon Mincer’s theoretical and empirical work on the specific human capital of schooling and on-the-job-training to codify the concept of “human capital” as a tool to define the factors that explain wage variance in a population. In *Education and the Distribution of Earnings*, they put forth a theory of the distribution of earnings in which, “each person is assumed in effect to maximize his economic welfare by investing an appropriate amount in human capital, and the distribution of of earnings is determined by the distribution of investments and their rates of return” (1966: 368). In this theory the determinants of distribution are “inheritance of property income, equality of opportunity, distribution of abilities, subsidies to education, and other human capital, etc.” (368). Put differently by Hasl, et al., “individuals’ human capital involves their educational achievement, cognitive and socioemotional skills, and work experience, among other things” (2022: 505).

In perhaps Mincer’s most lasting contribution to wage modeling, he advances his work and its extension by Becker and Chiswick to develop the “human capital earnings function,” put forth in his 1974 study *Schooling, Experience, and Earnings: “*In this study, Mincer shows that “the inclusion in the earnings function of even crude measures of ‘post school investments’ in addition to schooling lends a great deal of scope to the analysis of distribution," and he coins the term "the human capital earnings function" for this expanded relationship" (Chiswick 351). Mincer presents the human capital earnings function as the formula,

$$\ln(w) = \alpha + \beta S + \gamma_1 X + \gamma_2 X^2 + \epsilon,$$

where the logarithm of gross earnings can be expressed as a quadratic function of years of labor and market experience. In the above function, the natural logarithm of wages ($\ln(w)$) represents the dependent variable, which corresponds to an individual's earnings. The independent variables include the number of years of schooling (S), which captures the education level, and the years of labor market experience (X). The model also includes the square of labor market experience (X^2) to account for diminishing returns to experience over time. The intercept term is denoted by α, while β is the coefficient that measures the return on each additional year of schooling. The coefficients γ_1 and γ_2 measure the effect of experience and its squared term on wages, respectively. Lastly, the error term (ϵ) accounts for unobserved factors that may influence an individual's wages (353). Chiswick notes that the power of this function lies in its "high explanatory power for earnings in spite of the simple measures of investment in human capital" (354) and that its lasting impact is "that the structure of the equation has become a standard feature of research in labor economics, the economics of education, and in labor market analyses by sociologists" (356).

While developed by Mincer as an economic model, Tomaskovic-Devey, Thomas, and Johnson develop an explicitly sociological variant on human capital theory that can be further utilized in the modeling of earnings inequalities. By the time of their study in 2005, they show how mainstream Mincer's work has become, writing "The human capital model is so well known that it requires little introduction. The basic idea is that individuals make investments in their own set of productive skills through education, general experience in the labor market, and specific experiences with current employers" (60). However, their sociological critique continues that, "human capital investment is often not a voluntary and almost never an individual choice. Human capital acquisition is a social process" (61). They expand Mincer's human capital function to include race and ethnicity, showing how new terms can be added to interact with age and education (66):

$$\ln(wage_{it}) = \alpha_i + \alpha_t + \beta_1 age_{it} + \beta_2 education_{it} + \beta_3(age_{it} \times education_{it})$$
$$+\beta_4(age_{it} \times black_{it}) + \beta_5(age_{it} \times Latino_{it}) + \beta_{6-N} control_{it} + \epsilon_{it}.$$

The power of Tomaskovic-Devey, Thomas, and Johnson's approach is that it allows the incorporation of the social mechanisms that produce inequalities into the human capital earnings function.

C. CUMULATIVE ADVANTAGE

The concept of cumulative advantage (CA) was developed as an extension of the mechanism by which different human capital produces inequality across any temporal process. Cumulative advantage was originally developed by R.K. Merton to explain advancement in science careers under the label "The Matthew effect," but since its development it has been employed to take on multiple meanings going beyond its original application to scientific careers (Crystal, Shea, and Reyes 2017: 911). The term "cumulative advantage" has become a shorthand for "cumulative advantage and disadvantage," a description of "processes by which the effects of early economic, educational, and other advantages can cumulate over the life course" and "a theoretical model that described and examined the process of disparate life outcomes as one of iterative interaction of initial advantages and societal institutions over the life course" (911).

Diprete and Erich expand on the mechanism by which cumulative advantage produces inequality. They write:

> The central descriptive idea in the CA literature is that the advantage of one individual or group over another grows (i.e., accumulates) over time, which is often taken to mean that the inequality of this advantage grows over time. The advantage in question is typically a key resource or reward in the stratification process, for example, cognitive development, career position, income, wealth, or health. (2006: 272)

In this description, the advantages and disadvantages that drive the mechanisms of inequality are human capital such as "cognitive development, career position, income, wealth, or health"; thus, it is shown how human capital drives the mechanism of cumulative advantage. Interestingly, DiPrete and Eirich note that the CA process can occur at the level of populations as well as at the level of individuals, cautioning social researchers that "the failure to recognize the possibility that CA is at work can lead to incorrect models and incorrect predictions about the evolution of population distributions over time" (284).

D. WAGE MODELING

Now that the theoretical assumptions and mechanism for wage modeling have been established, contemporary wage modeling methods utilizing the framework of the human capital earning function and cumulative advantage can be discussed. There are many methods of wage modeling, but the following three methods were selected due to their utilization in recent wage prediction studies.

The first method of discussion is logistic prediction. A logistic prediction model fits wage data to a logistic function, allowing the capture of wage trends with an initial slow growth phase followed by rapid increases, eventually stabilizing. The general form of the logistic model is as the following equation:

$$y = \frac{K}{1 + ae^{-bt}},$$

wjhere K is the saturation level, a is the coefficient, b is the relative growth rate, e is the exponential constant, and t is time. Y in the equation is the model's output and its value represents the average salary every year (Wei, Jiang, and Zhao 2014: 616). An example of this model's use is predicting long-term wage trends based on historical wage data and economic growth assumptions, such as for workers in Sichuan Province (618).

Quantile regression is the second approach for discussion. Quantile regression extends traditional linear models by estimating the impact of variables across different points of the wage distribution (e.g., median, upper, and lower quantiles). This approach extends the standard capital human wage function to incorporate risk and skewness:

$$ln(w_i) = \alpha_0 + \alpha_1 X_i + \theta_1 R_{ij} + \theta_2 S_{ij} + v_i,$$

where R_{ij} is wage variance, S_{ij} is skewness, and w_i represents wages (Vieira, Constancia, and Teixeira 2020: 195). The quantile regression method is useful for capturing heterogeneous impacts of wage variation on different wage groups, such as evaluating how education and risk effect wage outcomes at different points in the wage distribution (197).

Finally, Dynamic Structural Equation Modeling (DSEM) integrates cumulative advantage, human capital theory, and autoregressive processes to study wage inequality and dynamics over time. This model utilizes two equations. The first,

$$Y_{it} = \alpha_i + \phi_i Y_{i,t-1} + \varepsilon_{it},$$

represents the within-person level model, with an autoregressive structure that captures the dynamics of individual wage growth over time (Hasl, et al. 2022: 508). A second set of equations,

$$\alpha_i = \beta_{00} + \beta_{01}\text{IQ}_i + \beta_{02}\text{GPA}_i + \beta_{03}\text{pSES}_i + \beta_{04}\text{Edu}_i + e_{0i}$$

$$\phi_i = \beta_{10} + \beta_{11}\text{IQ}_i + \beta_{12}\text{GPA}_i + \beta_{13}\text{pSES}_i + \beta_1 4\text{Edu}_i + e_{1i}$$

represents a between-person model that captures individual differences in initial wages (α) and growth rates (ϕ), linking age dynamics to education and socioeconomic factors (in this instance adolescent IQ, GPA, parental socioeconomic status, and the highest level of education in adulthood). Importantly, the between person equations, "highlight the idea that even if individuals (on average) do not experience strict CA processes, a pattern of growing inequality can arise from individual differences in these between-person variables" (509). Dynamic Structural Equation Models are useful for modeling wage dynamics and understanding cumulative advantage, focusing on individual growth trajectories over time.

In summary, these models offer complementary insights, with logistic models focusing on trends, quantile regression capturing distributional differences, and DSEM examining dynamic wage growth and inequality across the lifespan.

E. COMPENSATION MANAGEMENT

The ultimate goal of tracing a line from the concepts of wage equity and dispersion through human capital and cumulative advantage to the eventual modeling of wage inequality is to better inform a more equitable strategy of compensation management. Effective compensation management is related both to the internal fairness of the enterprise of human resource management as well as to the impact on the competitiveness of a company in its industry (Li 2021: 1). Compensation planning involves the collection of historical data from third party sources (compensation surveys, best practices reports, and financial statements) to be studied and analyzed. Furthermore, in an essay, "Future Compensation: Principle-Based Analytics and Perceived Value," James Sillery notes that,

> …the past is not always a good predictor of what will be needed in the future. In the future, companies will use tools similar to Monte Carlo simulations. This will involve playing out multiple scenarios to identify a path to the desired state, and then modeling that state back to the present to identify the resources needed along the way. (Risher 2014: 72)

Sillery understands that data-driven predictive modeling is a key to future wage compensation studies.

Indeed, technology and the implementation of HRIS (Human Resource Information Systems) has greatly impacted how organizations structure wage compensation. In addition to producing the required information for making compensation decisions faster and at a lower cost, “the most fundamental value of [HRIS] technology is its ability to encourage new thinking that removes the need for layers of administration” (Kovach and Cathcart 1999: 276). Emphasis is placed by Wang (2024) on HRIS and data analytics’ contributions to strategic planning and decision support: “Throughout statistical analysis, such as correlation coefficients and regression models, organizations can develop a mathematical model that predicts the success of HRIS implementation based on variables like organizational readiness, system compatibility, and user engagement levels” (224). Li (2021) advocates for the use of advanced analytical tools such as SQL and Python to improve HRIS compensation management practice in the era of big data and ultimately drive productive business decisions (2). Li’s primary benefits for using such advanced tools are increased scale of processing data and increased complexity of the analysis operation. In addition to paying attention to just total salary and benefits budgets, advanced compensation analysis allows for the inclusion of “other factors’ effects, such as the number of employees of the company, the company’s past experience, and the external comparisons in the same industry” (3). This advanced analysis of salary data is comprised of descriptive analysis, predictive analysis, and guidance analysis. Li (2021) delineates these analysis types as such:

> Descriptive analysis mainly refers "what is happening now" and relates to average median, frequency, etc., which are achieved through Excel; predictive analysis refers "what will happen" and "why this happened," realizing through Python by using machine learning tools (such as KNN, Naive Bases, and decision trees, etc.); guidance analysis is the highest level and mainly aimed at "what should we do" through the decision model, providing customized solutions for future decisions and strategies. (4)

Li then suggests that quantitative analysis in the salary field be carried out from the six aspects of labor cost analysis, internal fairness analysis, external competitiveness analysis, personal reward, salary structure analysis, and salary satisfaction analysis. Of particular interest for its relation to equity, human capital, and competitive advantage is internal fairness analysis, the aspect concerning "the relationship between employees' education background, skills, department, position, rank, position and internal equity" (4). It is these predictive and descriptive quantitative analyses of salary wage data that this project will employ to explore the gap in wage compensation between the internal data of a mid-size NYC cultural institution and external market job valuation data.

III. RESEARCH QUESTIONS

The primary research question this project seeks to answer is as follows:

Can machine learning be used to model future annual wage compensation per job title for a mid-size NYC cultural institution to account for inflation, market rate wage changes, and individual tenure?

Secondary research questions are:

What are the features/variables that contribute to pay disparity between external job market valuation and internal wage compensation? How do these features/variables relate to pay equity?

IV. DATA SOURCES

There are two sources of data for this project. The first data source is the Employee Census of our mid-size NYC cultural institution. This Employee Census was provided by the institution's human resources department upon promise of confidentiality for the institution and its staff. The institution is operated by a nonprofit organization, and it is dedicated to hosting exhibits, events, and education programs related to its mission. It is one of many well known major cultural and educational institutions in New York City. The institution has approximately 250 employees, and has an annual operation budget of around $20 million USD.

The second data source for this project is Occupational Employment and Wage Statistics (OEWS) data from the U.S. Bureau of Labor Statistics. The OEWS survey is a semiannual survey measuring occupational employment and wage rates for wage and salary workers in non-farm establishments in the United States, and it produces employment and wage estimates annually for approximately 830 occupations. The OEWS survey draws its sample from state unemployment insurance files. This project utilizes OEWS survey data from the years 2012 through 2023 that was collected from the U.S. Bureau of Labor Statistics website (bls.gov). Survey data is available from 1997 to the present. Data from 2012 to 2023 was selected for this project to provide a robust enough data sample for model training and because the OEWS survey data prior to 2012 does not contain a field for NAICS code. NAICS refers to the North American Industry Classification System. The OEWS data sample for this project is selected for NAICS industry code 712000, referring to Museums, Historical Sites, and Similar Institutions. The OEWS Survey data was selected because it is the most comprehensive wage survey data available that is most comparable to the data collected in our other data source, the Employee Census.

V. PRIMARY VARIABLES

The variables found in our Employee Census are as follows:

- GENDER
- JOB TITLE/CODE
- BENEFIT ELIGIBILITY DESCRIPTION/CODE
- HIRE DATE
- ANNUAL SALARY
- EEOC JOB CLASSIFICATION CODE/DESCRIPTION

Additionally, I have coded a variable OCC_CODE, which is the six digit Standard Occupational Classification (SOC) code used by federal agencies to classify workers into occupational categories for the purpose of collecting, calculating, or disseminating data. I have used my personal judgment to assign the best fitting SOC code for each of the institution’s job titles.

The variables selected from the OEWS Survey Data to correspond with the variables found in our Employee Census are as follows:

- NAIC - *North American Industry Classification System (NAICS) code for the given industry*
- NAICS_TITLE - *North American Industry Classification System (NAICS) title for the given industry*
- OCC_CODE - *The 6-digit Standard Occupational Classification (SOC) code or OEWS-specific code for the occupation*
- OCC_TITLE - *SOC title or OEWS-specific title for the occupation*
- TOT_EMP - *Estimated total employment rounded to the nearest 10 (excludes self-employed)*
- EMP_PRSE - *Percent relative standard error (PRSE) for the employment estimate.*
- PCT_TOTAL - *Percent of industry employment in the given occupation*
- A_MEAN - *Mean annual wage*
- MEAN_PRSE - *Percent relative standard error (PRSE) for the mean wage estimate*
- A_PCT10 - *Annual 10th percentile wage*
- A_PCT25 - *Annual 25th percentile wage*
- A_MEDIAN - *Annual median wage (or the 50th percentile)*
- A_PCT75 - *Annual 75th percentile wage*
- A_PCT90 - *Annual 90th percentile wage*

Additionally, because the annual OEWS datasets have been bound into one dataset for the Museums, Historical Sites, and Similar Institutions industries, I have added a YEAR variable referring to the survey year, 2012-2023.

VI. ANALYTIC STRATEGY

The primary analytic strategy for this project is to train a machine learning model on the combined OEWS survey and Employee Census data, utilizing a constructed wage gap feature (annual employee salary [ANNUAL_SALARY] from the Employee Census minus mean annual wage [A_MEAN] from the OEWS survey) as our label to be predicted by the other features or variables of the dataset. The trained model is then used to process the Employee Census dataset to generate wage gap predictions for each job title present in the census that are more aligned with the broader Museums, Historical Sites, and Similar Institutions industry salary data. These wage gap predictions will be used to construct a final adjusted salary per job title for our NYC cultural institution that has greater parity with the industry salary data and adjusts compensation for inflation, market rate wage changes, and individual tenure. The effective implementation of this machine learning model will allow us to answer our primary research question regarding the efficacy of using such a model for wage prediction.

A feature analysis of the model will determine which variables are the most relevant for predicting the wage gap between our NYC cultural institution and the broader Museums, Historical Sites, and Similar Institutions industry for the job titles in our Employee Census. This feature analyses will serve to answer our secondary research question regarding what features/variables contribute to pay disparity within the Museums, Historical Sites, and Similar Institutions industry.

VII. METHODS

To begin, 2012 to 2023 OEWS data from the Bureau of Labor Statistics was read into the RStudio IDE. This data was bound into one dataset after column names were standardized across all years, and the imported data was filtered by NAICS industry code 712000 for “Museums, Historical Sites, and Similar Institutions”. The data was then filtered by a list of all OCC_CODES that were coded by job title into the Employee Census. Next, the OEWS data and Employee census data were joined on OCC_CODE to create a combined dataset. The WAGE_GAP feature was engineered by subtracting A_MEAN from ANNUAL_SALARY. Because we only have ANNUAL_SALARY data from 2023, but A_MEAN data from 2012-2023, we must find a way to adjust our wage gap data for the years 2012 to 2022. To engineer this adjustment, I subset the 2023 wage gap data from the combined dataset. I then constructed a “percent wage gap” column by dividing WAGE_GAP by A_MEAN just for the 2023 data. This PCT_GAP column was joined back to the combined dataset, and the WAGE_GAP feature was reconstructed by multiplying PCT_GAP by A_MEAN. There are limitations to this approach, mainly that it assumes that the wage gap is the same percentage of annual salary for each year in the dataset, which would not be true if we were to analyze actual annual salary data from our NYC cultural institution for the years 2012 to 2022. We now have our final combined dataset for machine learning.

To train our machine learning model, the combined dataset was output from RStudio as a .csv file, and uploaded into the Colab Python IDE. The columns 'HIRE_DATE', 'HIRE_YEARS', 'JOB_TITLE', 'BENEFIT_ELIGIBILITY', ‘PCT_GAP’, 'EEOC.JOB.DESCRIPTION', 'NAICS_TITLE', 'OCC_TITLE', and ‘ANNUAL_SALARY' were dropped from the model for reasons of redundancy and model skewing. A Random Forest Regressor model was first tuned and and then trained on our data, giving the following performance metrics:

Metric	**Value**
Mean Absolute Error (MAE)	2,900.10
Mean Squared Error (MSE)	34,935,411.14
Root Mean Squared Error (RMSE)	5,910.62
R^2 Score	0.9034

Table 1: Random Forest Metrics

Next, a Gradient Boosting Regressor model was tuned and trained on our combined dataset, giving the following performance metrics:

Metric	**Value**
Mean Absolute Error (MAE)	2,490.02
Mean Squared Error (MSE)	29,680,849.85
Root Mean Squared Error (RMSE)	5,448.01
R^2 Score	0.9180

Table 2: GBR Metrics

The Gradient Boosting Regressor model was chosen as our final model due to its superior metrics. The GBR’s R^2 Score of 0.9180 indicates that about 91.8% of the variability in the target variable is explained by the model and indicates a strong fit to the data; furthermore, the errors MAE and RMSE are relatively low, indicating reliable performance.

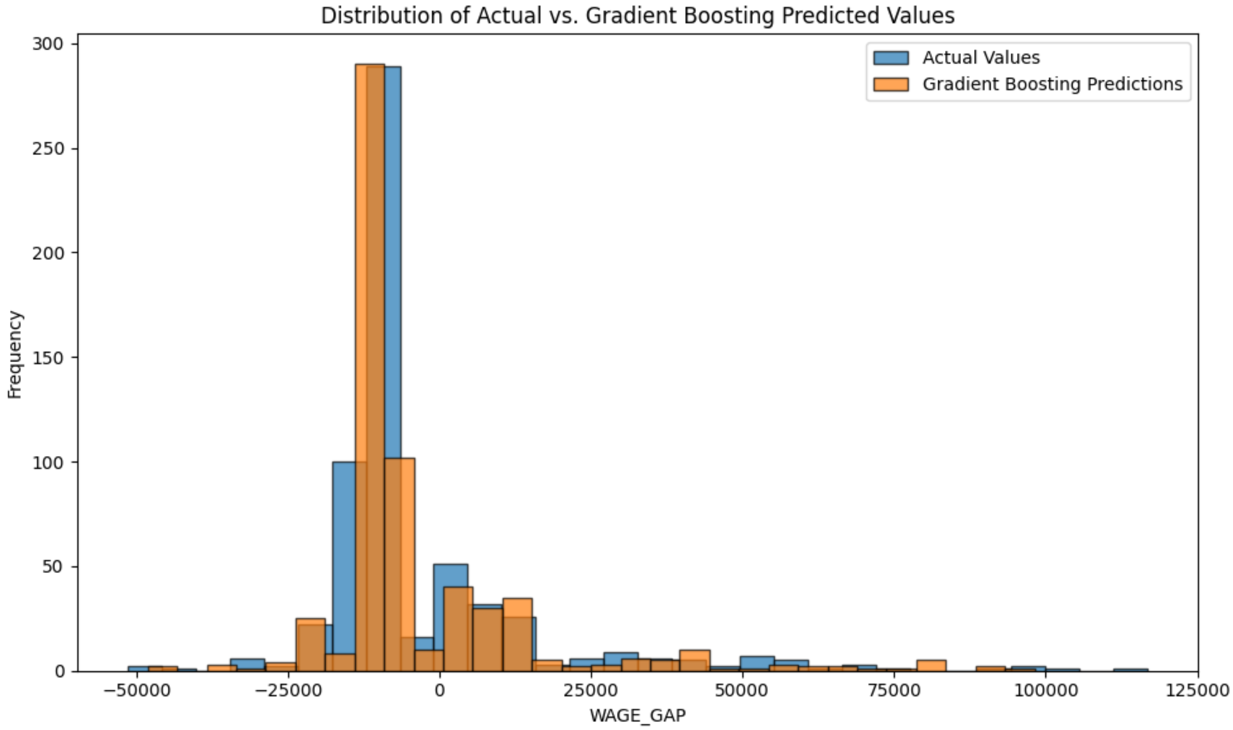

Figure 1: GBR Model Value Distribution

Further visual analysis of the GBR model's predicted value distribution versus actual values in Figure 1 indicates a good central fit for this model, and contributed to its selection.

Following the training of the selected GBR model, the Employee Census dataset was processed for predictions by aligning its structure with that of the training data. The GBR model was then used to predict on the processed Employee Census data, and these predictions were saved. The distribution of these wage gap predictions are shown below in Figure 2. We see the distribution of these saved predictions aligns with the GBR model value predictions distribution shown in Figure 1, indicating prediction validity. The saved predictions were grouped by JOB_TITLE and merged back into the

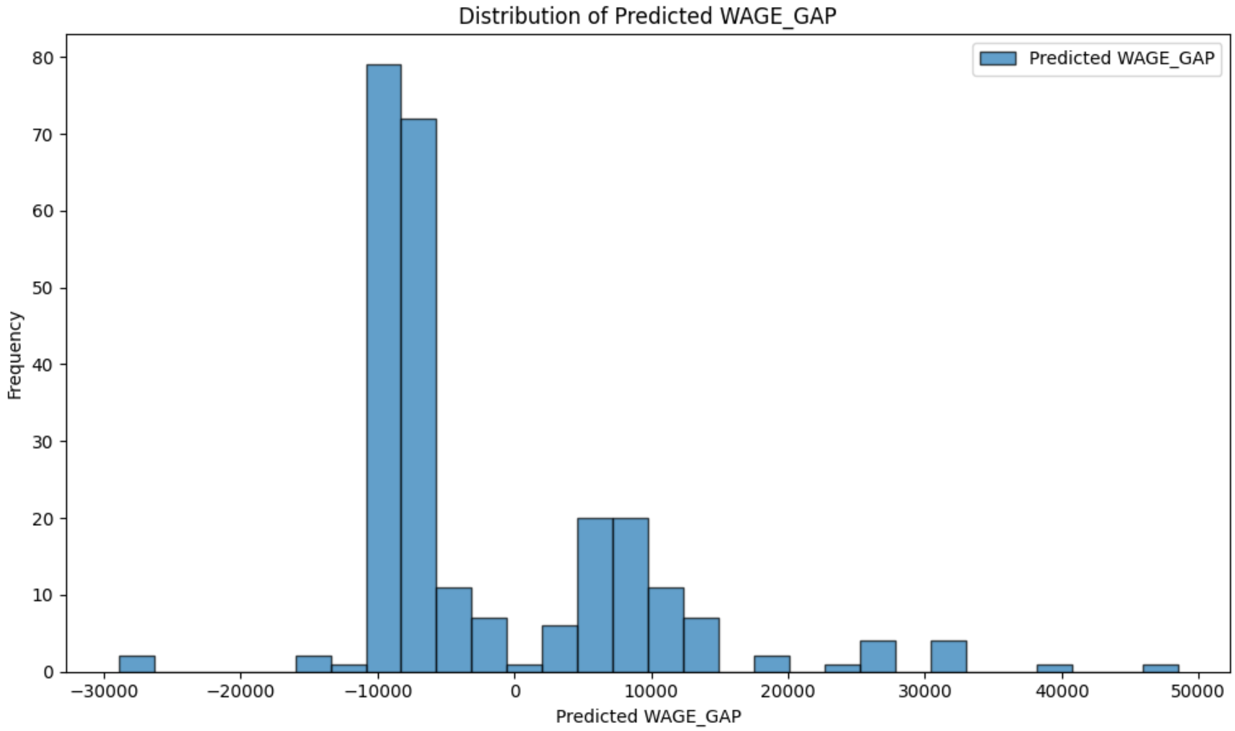

Figure 2: Predicted Wage Gap Distribution

Employee Census. This new predicted WAGE_GAP was then added to ANNUAL_SALARY to compute a new predicted annual salary per job title. This new predicted annual salary was then adjusted for inflation, market rate wage changes, and individual tenure using the following framework:

Adjustment Type	Description	Formula
Inflation Adjustment	Adjust salaries for inflation using historical CPI data or an annual inflation rate to bring past salaries to current-year dollars	Inflation Adjusted Salary = Predicted Salary × (1 + Inflation Rate)^n Where n = Number of years to adjust
Market Rate Changes	Account for changes in market conditions by applying an estimated market growth rate to the inflation-adjusted salary	Market Adjusted Salary = Inflation Adjusted Salary × (1 + Market Rate Change)

Tenure Adjustment	Reflect tenure-based increases using the number of years since hire (Tenure Years) and applying an annual growth rate	Tenure Adjusted Salary = Market Adjusted Salary × (1 + Tenure Rate)^(Tenure Years)

Table 3: Annual Salary Adjustment Framework

For the adjustments I selected an inflation rate of 0.025 representing 2.5% annual inflation, a market rate change of 0.03, representing 3% annual market rate change, and a tenure rate of 0.005, representing a .05% increase in salary per year of tenure. The formulas in our adjustment framework were applied to the new predicted annual salary per job title to compute an Inflation Adjusted Mean Salary, Market Adjusted Mean Salary, and Tenure Adjusted Mean Salary each per job title. Finally, these three adjusted salaries were integrated to create a Final Adjusted Salary utilizing the following formula:

$$\begin{aligned}\text{Final Adjusted Salary} = {} & \alpha(\text{Inflation Adjusted Salary}) \\ & +\beta(\text{Market Adjusted Salary}) \\ & +\gamma(\text{Tenure Adjusted Salary})\end{aligned}$$

where α, β, and γ are weights assigned to each adjustment type. I selected weights of $\alpha = 2/5$, $\beta = 2/5$, and $\gamma = 1/5$ for our model's Final Adjusted Salary. The rates and weights for these adjustments were selected by me for their comparability to standard job market adjustment rates and weights, but they can be easily modified by any organization implementing the model to reflect the standards and needs of such organization. Lastly, the Final Adjusted Salary is merged back to the Employee Census, creating a deliverable for the organization with machine learning model predicted annual salaries per job title.

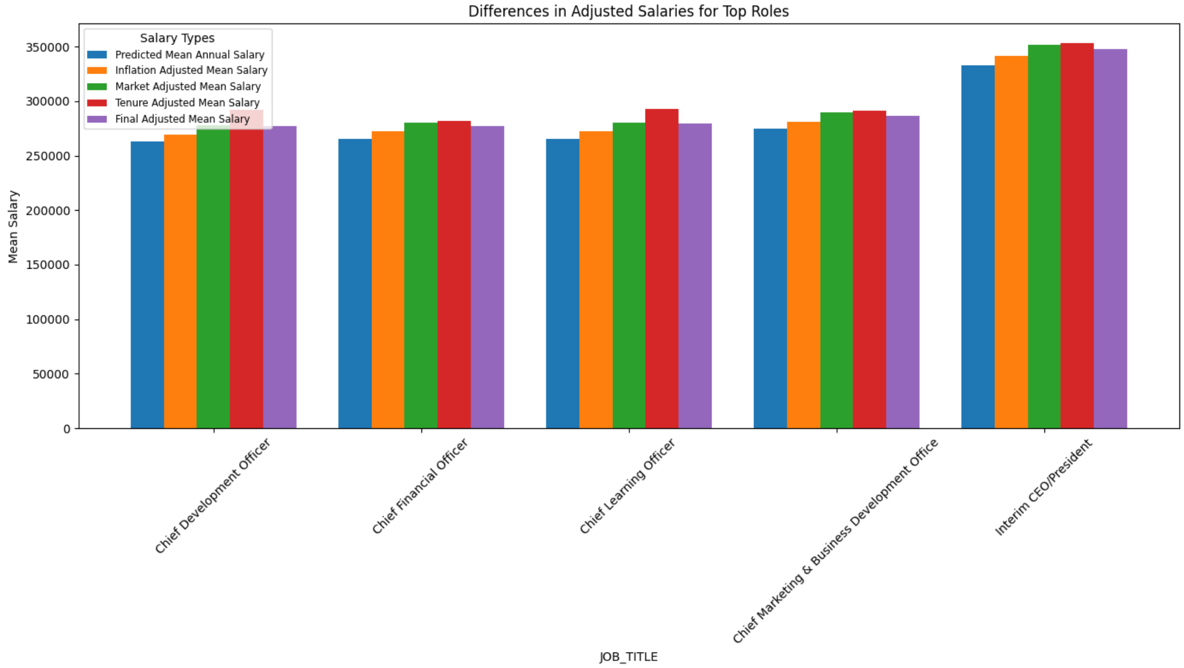

Figure 3: Model Adjusted Salary Visualization

Figure 3 visualizes the differences between the original annual salary and the four adjusted annual salaries for the five most highly compensated job titles in the NYC cultural institution. This figure is shown to give a visual representation of how each adjusted salary differs from the original, and how the Final Adjusted Mean Salary integrates the three other adjusted salaries.

VIII. RESULTS

Our final deliverable for the NYC cultural institution is an updated Employee Census for the organization with a wage gap in USD that represents the difference in salary between the actual annual compensation for a job title within the organization and the RGB model's predicted annual compensation for that job title based upon mean annual salaries from the U.S. Bureau of Labor Statistics designated occupation code for that job title within the NAICS designated industry of Museums, Historical Sites, and Similar Institutions. When the outliers from this wage gap variable are removed using the IQR method, we are presented with the following distribution:

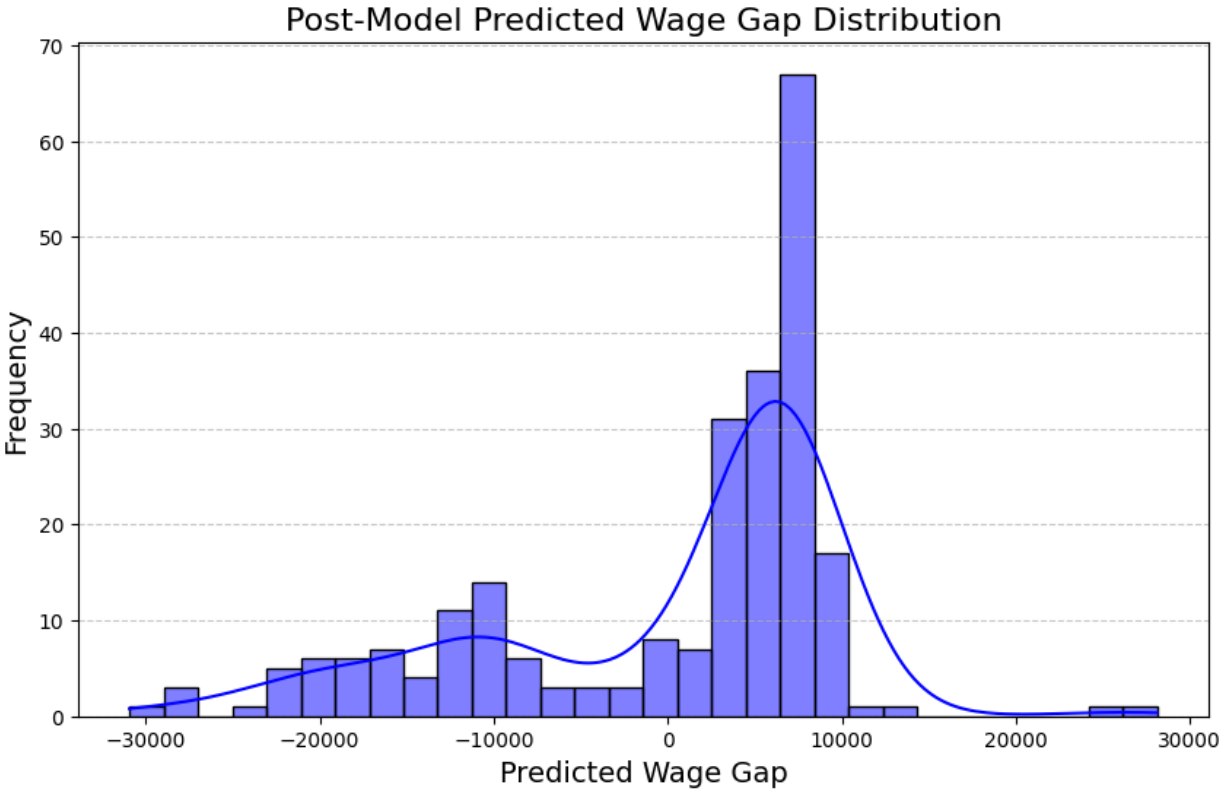

Figure 4: Final Wage Gap Distribution

The mean wage gap is 108.59, indicating that on average actual organizational salaries are slightly above the industry mean. The highest mode of wage gaps in our distribution are between 2,000 and 10,000 indicating that overall the majority of job titles in the organization are compensated above industry mean. The final deliverable also contains a Model Predicted Salary, calculated by adding our wage gap to the actual salary, and a Final Adjusted Salary, calculated by applying the formulas from our adjustment framework. When we compare the actual salaries per job title in the final deliverable to the predicted and final adjusted salaries per job title as in Figure 5 we

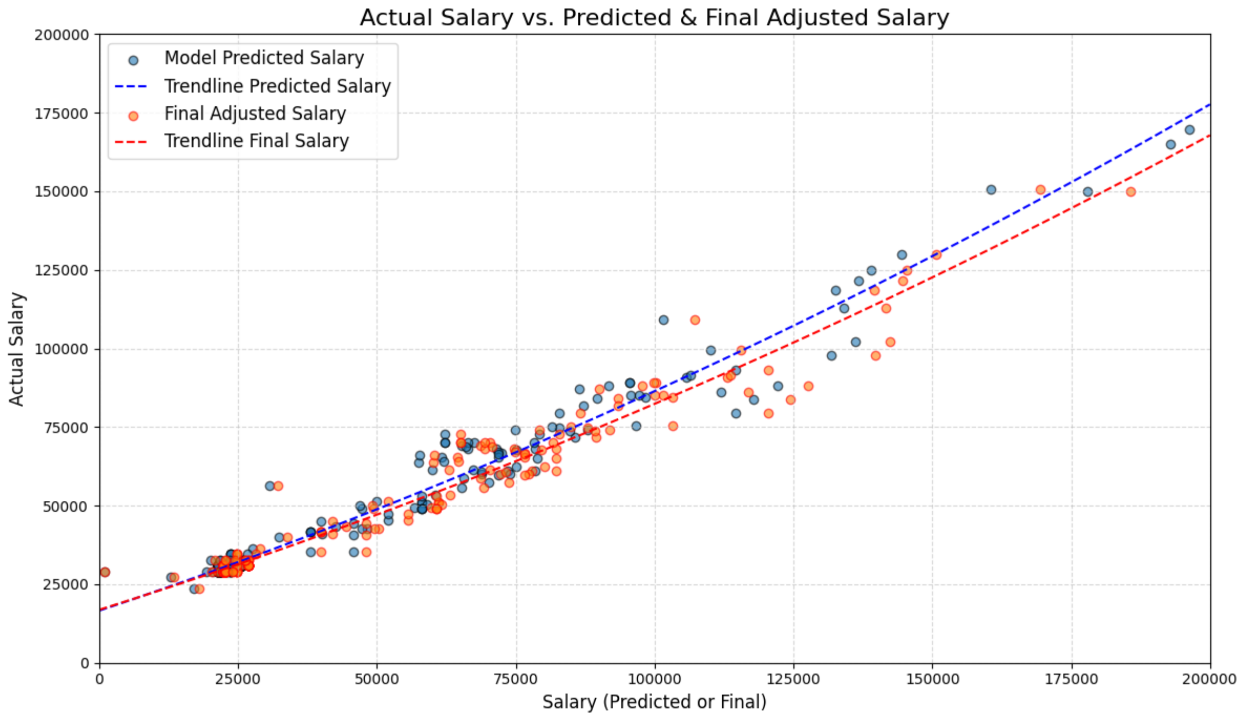

Figure 5: Final Deliverable Salary Comparisons

see that the data points generally align along a 45-degree line, indicating that the predicted and final adjusted salaries strongly correlate with the actual annual salaries. This is not surprising, as our model predicted salaries are based upon our wage data; however, the trend lines' close approximations of a 45-degree line suggest good overall accuracy for the model. Our deliverable and subsequent analysis of its results proves that machine learning can be utilized to accurately fit a model to industry wage compensation data to predict future annual wage compensation per job title. These predictions can then be augmented with calculations to adjust them for inflation, market rate wage changes, and individual tenure.

Interestingly, the predominance of outlying job titles with large negative wage gaps in our final deliverable are at the levels of upper management within the organization. We see in Figure 6 that the five least compensated job categories

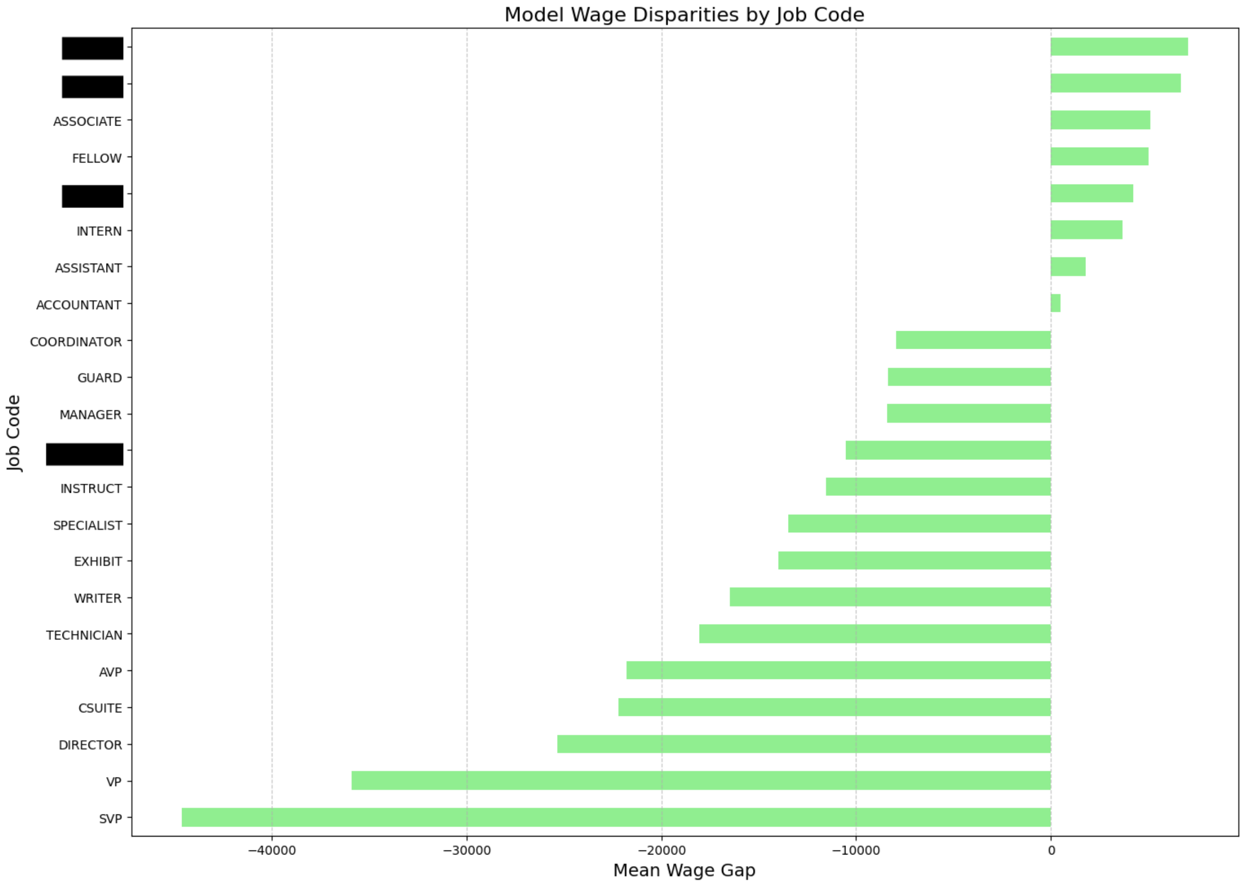

Figure 6: Wage Disparities by Job Code

compared to the industry mean within the organization are “SVP,” ‘VP,” “DIRECTOR,” “CSUITE,” and “AVP,” all upper management categories. The large negative wage gaps for these upper management categories indicate that either upper management roles within the organization are under-compensated or that these upper management roles are overcompensated industry-wide, representing an area within which the organization may wish to look in an effort to increase wage parity with the overall industry.

Feature analysis of our model shows that the most predictive features for wage gap are the features relating to compensation and job code/occupation. It is evident in Figure 7 that the most predictive feature is A_MEAN, followed by job code and

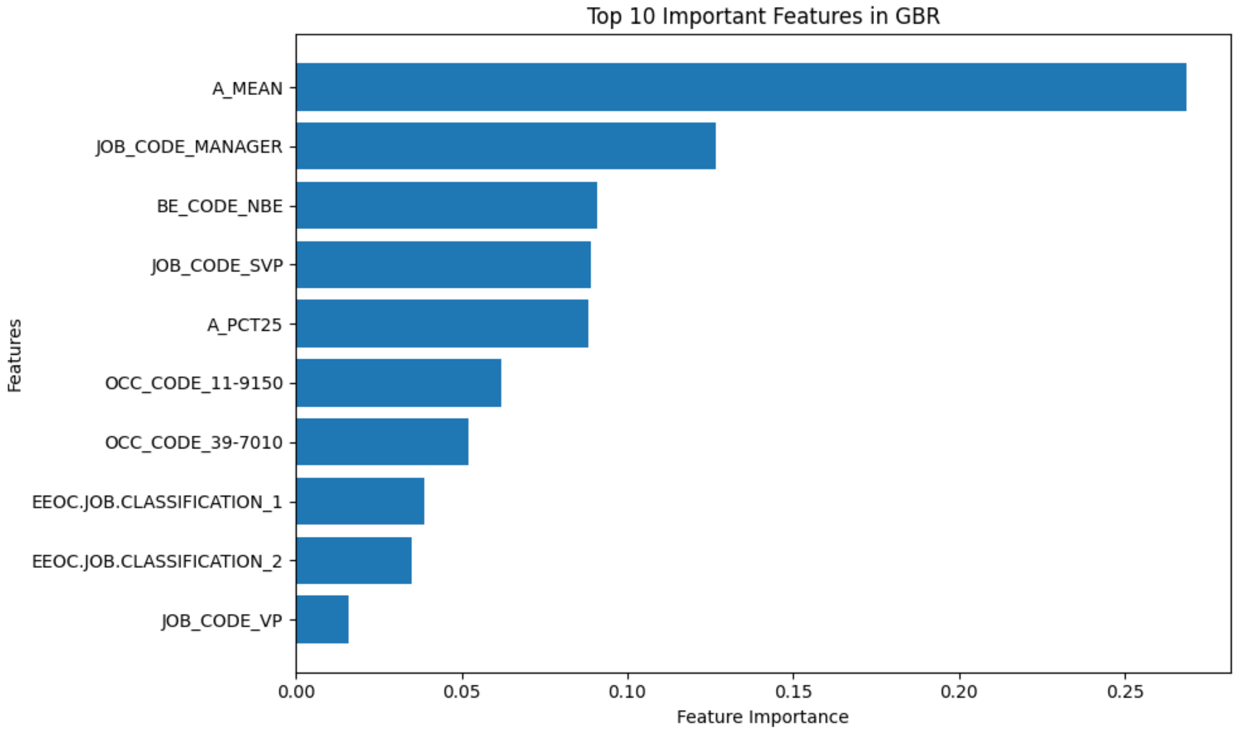

Figure 7: GBR Model Feature Analysis

occupation code features. It is no surprise that the best predictor of wage compensation is itself wage compensation. The OEWS data does not contain demographics like gender, race, and age, so we cannot determine if these categories effect pay disparity for our study. However, we can conclude that the feature/variable that most strongly contributes to pay disparity between external job market valuation and the internal wage compensation of our organization is managerial status. It is unclear how managerial status relates to pay equity within our organization without further demographic analyses, but within the broader context of national industry wages, we can consider that burgeoning wealth inequality has resulted in larger pay gaps between more highly paid upper management roles and lower paid non-managerial roles.

IX. CONCLUSION

The purpose of this capstone project was to explore the differences in annual compensation between the internal data of a mid-size NYC cultural institution and national external market job valuation data and to construct a machine learning model to predict future compensation per job title for the institution accounting for inflation, market rate wage changes, and tenure, thus creating a deliverable for potential use by such an organization to promote a more equitable wage dispersion within the organization and society at large. The data analytic tools of RStudio and Python were successfully used to achieve this purpose, and our final deliverable and subsequent analysis of its results proved that machine learning can be utilized to accurately fit a model to industry wage compensation data in order to predict future annual wage compensation per job title and that these predictions can then be augmented with calculations to adjust them for inflation, market rate wage changes, and individual tenure. Finally, we concluded that the feature/variable that most strongly contributes to pay disparity between external job market valuation and the internal wage compensation of our organization is managerial status but that determining its relation to wage equity is beyond the scope of this project and requires further analyses.

Although it has been effective at answering our research questions and creating our deliverable, there are some limitations to the approach utilized by this project. The first limitation lies in the coding done by myself of the job titles from the Employee Census to SOC code. As this coding was done using my judgment, it is subject to a degree of bias inherent in any qualitative assignment. Furthermore, the coding from job title to SOC code represents a jump from the specific to the general. In coding a specific job title as a more general occupation code we lose the specificity unique to a job title and are at risk of flattening the nuance of these roles to a broad salary band. A second limitation of this project is that our wage gap for the years 2012 to 2023 was engineered based on actual salary data only from the year 2023. It was my attempt with this project to implement a deliverable solution based upon the data provided, but for more accurate compensation predictions I would recommend utilizing an organization's historical compensation records to more thoroughly and fully depict wage gaps over the time series of the dataset. Finally, the project is limited by restricting itself to the OEWS survey data for NAICS industry code 712000, referring to Museums, Historical Sites, and Similar Institutions. By restricting to one NAICS industry code, we assume that the mean annual wages collected by the OEWS survey for this industry code are the most comparable wages for our organization; however, our organization may benefit from a broader survey of multiple industry codes or a more specific subset of one industry code in order to achieve the external job market wage parity which that organization desires.

This discussion of organizational preference regarding coding, data selection, and industry classification raises the idea that the notion of "wage parity" is subjective, and that the parity an organization seeks must be conceptualized and driven towards by that organization's leadership. Despite the project's limitations, it is my hope that this work may be utilized by organizational leadership to drive such wage parity and contribute to the minimization of wage inequity within both workplace organizations and society at large.

X. REFERENCES

Becker, G. S., & Chiswick, B. R. (1966). “Education and the Distribution of Earnings.” *The American Economic Review*, 56(1/2), 358–369.

Cheng, S. (2021). “The Shifting Life Course Patterns of Wage Inequality.” *Social Forces*, 100(1), 1–28.

Chiswick, B. R. (2003). “Jacob Mincer, experience and the distribution of earnings.” *Review of Economics of the Household*, 1(4), 343–361.

Crystal, S., Shea, D. G., & Reyes, A. M. (2017). “Cumulative Advantage, Cumulative Disadvantage, and Evolving Patterns of Late-Life Inequality.” *The Gerontologist*, 57(5), 910–920.

DiPrete, T. A., & Eirich, G. M. (2006). “Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments.” *Annual Review of Sociology*, 32(1), 271–297.

Failla, V., Foss, N. J., Melillo, F., & Reichstein, T. (2024). “When inequality means equity: horizontal wage dispersion and the propensity to leave current employment across different organizational settings.” *Industrial and Corporate Change*, 00, 1–26.

Gupta, N., Singh, P., & Balcom, S. A. (2022). “When Pay Equity Policy Is not Enough: Persistence of the Gender Wage Gap Among Health, Education, and STEM Professionals in Canada, 2006-2016.” *Canadian Studies in Population*, 49(3–4), 123–148.

Hasl, A., Voelkle, M., Kretschmann, J., Richter, D., & Brunner, M. (2023). “A Dynamic Structural Equation Approach to Modeling Wage Dynamics and Cumulative Advantage across the Lifespan.” *Multivariate Behavioral Research*, 58(3), 504–525.

Kovach, K. A., & Cathcart, C. E. (1999). “Human Resource Information Systems (HRIS): Providing Business with Rapid Data Access, Information Exchange and Strategic Advantage.” *Public Personnel Management*, 28(2), 275–282.

Leete, L. (2000). “Wage equity and employee motivation in nonprofit and for-profit organizations.” *Journal of Economic Behavior & Organization*, 43(4), 423–446.

Li, B. (2021). “Quantitative Analysis of Salary Data in the Big Data Era.” *Journal of Physics. Conference Series*, 1881(3), 032022.

Marx, K. (1902). *Wage-labor and Capital.* New York: Labor News Co.

Risher, H. (2014). “Expert Views on the Future of Salary Management.” *Compensation and Benefits Review*, 46(2), 66–73.

Tomaskovic-Devey D., Thomas, M., & Johnson, K. (2005). Race and the accumulation of human capital across the career : A theoretical model and fixed-effects application. *The American Journal of Sociology*, 111(1), 58–89.

Vieira, J., Constância, C., & Teixeira, J. (2020). “Education and risk compensation in wages: a quantile regression approach.” *Applied Economics Letters*, *27*(3), 194–198.

Wang, A. (2024). “Enhancing HR management through HRIS and data analytics.” *Applied and Computational Engineering*, 64, 223–229.

Wei, Z. X., Liang, J. M., & Zhao, P. (2014). “Average Wage Prediction Based on Logistic Model.” *Applied Mechanics and Materials*, 483, 615–618.