MTH 4330: Intermediate Report

Predicting Wage Inequality Across New York State Industries and Regions

Group 7: Fnu Tenzin Dhonyo, Thomas Lee, Trisha Wu, Jordan Yim

November 19, 2025

1 Introduction

This project investigates wage inequality across industries and regions in New York State using machine learning techniques applied to the Quarterly Census of Employment and Wages (QCEW) dataset. Our task is to predict average annual wages based on multiple factors including industry classification (NAICS code), geographic location (county/region), employment size, ownership type (private vs. government), and temporal trends from 2015 to 2024.

Since our project proposal, we have made progress in implementing a complete data pipeline and training baseline models. This study also considers how New York's shifting industry climate—marked by emerging high-growth sectors and long-run wage trends—shapes which fields are most promising to pursue. The 2015–2024 window captures major structural forces, including rapid technological expansion, the COVID-19 shock, and the surge in at-home and care-based services.

Progress Summary: We have successfully loaded and cleaned the QCEW dataset, performed exploratory data analysis, engineered relevant features, and trained two models: a baseline linear regression and an Extreme Gradient Boosting (XGBoost) model. Initial results are strong, with clear room for improvement through tuning and added features.

2 Background

Prior research on wage inequality primarily uses regression-based decomposition methods and supply–demand frameworks to explain historical wage structures. These approaches, while valuable for understanding past trends, often fail to capture complex, high-dimensional interactions across technology adoption, policy changes, and demographic shifts (Autor & Katz, 1999).

The QCEW dataset provides near-census coverage of employment and wages in New York State, covering approximately 97% of nonfarm employment. However, the data presents several practical challenges including confidentiality suppression for small establishments, NAICS industry reclassification over time, and place-of-work counting that may double-count individuals with multiple jobs (New York State Department of Labor, 2024).

Our contribution is a predictive machine learning approach that leverages historical QCEW data to produce transparent, forward-looking wage estimates. Unlike traditional econometric models, our ensemble methods can capture nonlinear relationships between predictors and identify complex interactions between industry, geography, and time.

3 Dataset

Dataset Description

We use the New York State QCEW dataset (231,777 rows). After cleaning, our analysis focuses on:

- **Average Employment**: mean employee count per NAICS region–industry
- **Total Wage**: total annual wages paid
- **Year**: 2015–2024
- **Ownership**: numeric encoding of categories (Private, Local Gov., Total)
- **Annual Average Salary**: Total Wage / Average Employment

Additional fields such as **NAICS code**, **NAICS title**, and **region/area** support filtering and exploratory analysis but are excluded from the baseline models. These features are still critical because they reveal structural patterns across industries and geography that guide deeper modeling and interpretation.

Train/Dev/Test Split

Given the dataset size, we use a standard random split: 70% training (162,244), 15% development (34,767), 15% test (34,767). We also apply 5-fold cross-validation on the training set to obtain stable performance estimates and guide model selection.

Evaluation Metrics

With a continuous target (annual average salary), we evaluate models using standard regression metrics:

- **Mean Squared Error (MSE)**: average squared prediction error, minimized by the OLS model and sensitive to large errors.
- **Root Mean Squared Error (RMSE)**: square root of MSE, expressed in salary units (dollars), providing an interpretable measure of typical prediction error.
- **$\mathbf{R^2}$ (Coefficient of Determination)**: proportion of variance in salaries explained by the model; useful for comparing overall model fit.

Together, these metrics quantify both accuracy (RMSE), error sensitivity (MSE), and explanatory power (R^2) across our baseline and ML models.

4 Methods

4.1 Baseline Method: Ordinary Least Squares (OLS)

Feature Construction: All numerical features (Employment, Total Wage, Year) are cleaned and standardized. The categorical Ownership variable is encoded into numerical labels. The response variable, Annual Average Salary, is computed as Total Wage divided by Average Employment.

Input-to-Output Mapping: The model predicts annual average salary using the linear function:

$$\hat{y} = \beta_0 + \beta_1(\text{Employment}) + \beta_2(\text{Total Wage}) \\ + \beta_3(\text{Year}) + \beta_4(\text{Ownership}) + \epsilon. \qquad (1)$$

Mapping Steps:

- First, continuous inputs (Employment, Total Wage, Year) are multiplied by their corresponding coefficients.
- Second, the encoded Ownership value contributes an additive shift determined by its coefficient.
- Third, the intercept β_0 and all weighted feature contributions are summed to produce the prediction $\hat{y}$.

4.2 Machine Learning Method: XGBoost Regression

Feature Construction: XGBoost uses the same cleaned and encoded features as OLS. No standardization is required.

Input-to-Output Mapping: XGBoost builds a sequence of small regression trees, where each new tree learns to correct the errors of the previous model. The final prediction is the sum of contributions from all trees, allowing the model to capture nonlinear patterns more effectively than OLS.

Key Hyperparameters:

- **nrounds**: number of boosting iterations (trees)
- **eta**: learning rate controlling how much each tree contributes
- **max_depth**: complexity of individual trees
- **subsample**: introduce randomness and prevent overfitting

Hyperparameters are selected using XGBoost's built-in 5-fold cross-validation, and we choose the iteration with the lowest validation RMSE via early stopping.

Summary: Together, the OLS and XGBoost models show how linear and nonlinear approaches differ in their ability to predict wage outcomes.

5 Experiments

We evaluate OLS and XGBoost using 5-fold cross-validation and RMSE on the development and test sets. Table 1 shows that XGBoost achieves lower error across all splits, demonstrating a clear improvement in predictive accuracy.

The predicted-vs-actual scatterplots further highlight this difference. The OLS model displays a broad, dispersed pattern of points that drift noticeably from the 45-degree line, showing that the linear specification cannot capture essential wage dynamics, particularly at the extremes of the wage distribution. In contrast, XGBoost predictions stay much closer to the diagonal, showing that the model captures patterns that OLS misses.

Table 1: Comparison of OLS and XGBoost Model Performance (RMSE)

Model	CV_RMSE	Dev_RMSE	Test_RMSE
Baseline OLS	36230.45	35111.03	35831.65
XGBoost	2486.14	2089.04	2212.52

Overall, both the quantitative metrics and the visual diagnostics indicate that XGBoost provides a more accurate and flexible approach for modeling the QCEW wage data.

Predicted vs. Actual Plots

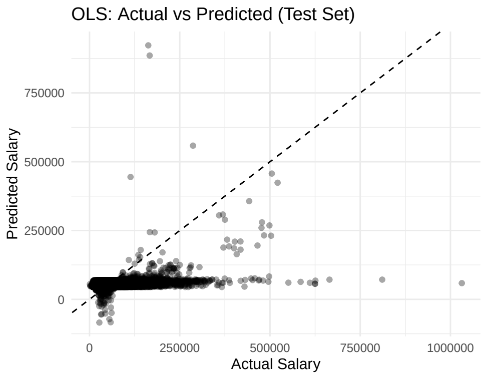

Figure 1: OLS: Actual vs Predicted (Test Set)

The OLS predictions spread widely around the diagonal, showing that the linear model has difficulty capturing wage differences across industries. By contrast, XGBoost predictions stay close to the diagonal, indicating that the model captures wage patterns much more accurately.

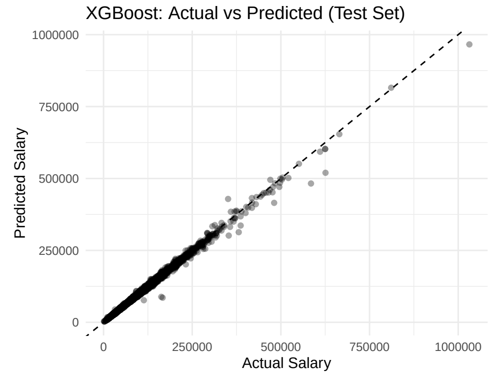

Figure 2: XGBoost: Actual vs Predicted (Test Set)

6 Conclusion

Our error analysis on the development set shows clear differences between the two models. OLS exhibits larger systematic errors, especially for industries with very high or very low wage levels. Its residuals indicate consistent under-prediction for high-wage sectors and over-prediction for low-wage sectors, reflecting the limits of a linear specification.

In contrast, XGBoost errors are smaller and more evenly distributed, although the model still struggles with extreme wage values and sharp year-to-year fluctuations. These poorly predicted cases generally correspond to industries with volatile employment or irregular wage spikes.

Based on these findings, we propose two actionable improvements:

1. Add nonlinear or interaction features (e.g., log transformations, spline terms, or Year × Industry effects) to better capture variation in underlying industry structures.

2. Incorporate additional predictors such as region indicators, NAICS subgroup dummies, or lagged wage features to provide more signal for volatile industries. Adding geographic variables—especially those that distinguish NYC—would further refine our predictions.

These refinements should reduce error variance and improve overall predictive performance for the final report.

References

Autor, D. H., & Katz, L. F. (1999). Changes in the wage structure and earnings inequality. *Handbook of Labor Economics*, *3*, 1463–1555. https://doi.org/10.1016/S1573-4463(99)03007-2

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

New York State Department of Labor. (n.d.). *Quarterly Census of Employment and Wages (QCEW) Annual Data* [Dataset]. New York Open Data. https://data.ny.gov/Economic-Development/Quarterly-Census-of-Employment-and-Wages-Annual-D
shc7-xcbw