SUMMATIVE 1: MODEL TRAINING AND EVALUATION

HIRWA ARMSTRONG BRIAN

AFRICAN LEADERSHIP UNIVERSITY

KIGALI, RWANDA

NAME OF FACILITATORS:

SAMIRATU NTHOSI

MARVIN OGORE

October, 2025

Video link: ML Summative 1.mkv

1. Introduction

1.1 Problem definition

One of the defining features of modern capitalist capitalist labor markets is their growing dependence on young, flexible, and often precarious labor to sustain and maintain key sectors of the economy. Over the past decades, neoliberal restructuring and financialization have accelerated the shift from stable, industrial employment toward service-dominated, lower-wage, less protected forms of labor. And within this process, youth labor has played a pivotal role, functioning as what Marxists theorized as a "reserve army of labor". which is the segment of the working class thats unemployed and underemployed in capitalist society.

The **reserve army of labor** concept, articulated by both Engels and Marx refers to the population of workers who remain marginal or semi-attached to formal production and are mobilized by capital during periods of expansion, then discarded in times of contraction. Youth workers, alongside women, migrants, and racialized groups, have historically constituted this reserve army in industrial economies. Their labor power is typically undervalued and under-secured.

Although the idea of youth precarity has been well documented there remains a **gap in systematic, data-driven approaches** that identify and map how **youth labor is distributed across industries** in contemporary economies and while most policy reports and academic debates acknowledge that young workers are concentrated in low-wage, low-benefit sectors they often lack precise, sector-level demographic evidence that quantifies this distribution and characterizes its structural patterns. This project aims to directly respond to that gap by applying **machine learning methods to official labor statistics** to identify **which industries in the U.S. as a chosen location depend most on young workers** and to explore how these industries cluster demographically.
This is not simply a matter of youth employment as a demographic curiosity. From a Marxist perspective, identifying industries where youth labor is concentrated is equivalent to identifying **sectors of heightened precarity**, where workers' capacity to organize is structurally weaker, turnover is higher, and wages and benefits are often suppressed. In capitalist economies, these are precisely the segments most vulnerable to economic downturns, policy retrenchment, and market volatility.

In this context, the **problem** this project addresses can be stated as follows:

> **To identify and characterize the concentration of youth labor (ages 16–24) across industries in the U.S. economy using official labor force data, in order to empirically map the reserve army of labor as it manifests in 2024.**

1.1.1 Problem originality and what makes it distinct

This problem is **original in two key ways**:

- First, it explicitly brings a **Marxist theoretical framework** i.e brings the reserve army of labor concept into a machine learning application, bridging critical political economy with contemporary data science.
- Second,,it focuses on **a recent, clearly bounded time frame (2024)**, providing a snapshot of post-pandemic labor restructuring and youth precarity at a moment when the U.S. economy has been navigating inflationary pressures, labor shortages in certain sectors, and renewed class struggle through unionization drives in logistics, services, and manufacturing.

This problem is distinct from typical labor market forecasting projects, which often focus on **predicting unemployment or participation rates at the macro level**, **estimating wage functions or educational attainment effects**, or **building generic economic forecasts**. In sharp contrast, this project centers **age structure and class power** as its analytical focus. It employs **machine learning methods** not to predict wages or unemployment directly, but instead to **identify demographic clustering within the labor process**. The originality of this work lies in how it **bridges Marxist theory and empirical computational analysis**, a connection rarely seen in standard ML coursework or applied economics research. This distinction is not about the novelty of youth labor itself, but rather in:

- How it is **framed theoretically** (as the reserve army of labor),
- How it is **operationalized** (through age-based industry segmentation),
- How it is **analyzed** (using unsupervised and supervised ML),
- And how it **connects to real-world labor struggles and policy debates.**

1.1.2 Problem justification

The **justification for focusing on youth labor** in the U.S. is grounded in a **well-established body of Marxist and critical political economy literature**.The original analysis emphasized the cyclical production of surplus labor populations as a structural feature of capitalism. This surplus population, often constituted by youth and other marginalized groups, serves as a mechanism to **discipline the labor market**, suppress wages, and ensure capital's flexibility in times of both expansion and contraction. Scholars have extended this analysis to the **neoliberal era**, where the flexibilization of labor markets has expanded the proportion of the workforce in precarious, low-benefit, high-turnover positions. The growth of the service sectors particularly sectors like retail, hospitality, food service, and platform-mediated gig work has been a key structural driver of this trend.

Reports from organizations such as the Economic Policy Institute and the International Labour Organization (ILO) similarly highlight that youth unemployment, underemployment, and labor precarity remain **persistent features of advanced economies**, even during periods of economic expansion. While technological change and economic restructuring have transformed some aspects of youth employment, the underlying logic of **youth as a flexible labor reserve remains intact**. In the U.S. context, the **service sector's reliance on young workers** is well documented in policy analyses and workforce studies. For example, retail and food service jobs are disproportionately filled by workers under 25, who typically receive lower wages, fewer benefits, and have limited union representation compared to their older counterparts in manufacturing or public sector employment. This **unequal distribution of age across industries** is not a neutral demographic fact: it **reflects and reinforces power asymmetries in the labor market**. Employers benefit from a younger workforce that is easier to hire and fire, less unionized, and often unable to exercise the same level of collective power as more established workers

By using the CPS dataset to **quantitatively map these demographic concentrations**, this project directly contributes to a **data-driven understanding of labor market segmentation**—a concept central to Marxist and socialist analysis of class.

1.2 The Dataset: The source and relevance

At the heart of this project is the official labor force data provided by the U.S. Bureau of Labor Statistics (BLS), which offers a clear picture of how employment is structured across industries and age groups.Specifically, the dataset used is **Table 18b: Employed persons by detailed industry and age**, drawn from the **Current Population Survey (CPS)**. The CPS is one of the most authoritative and widely used sources of labor market data in the United States, jointly sponsored by BLS and the U.S. Census Bureau. It serves as the statistical backbone for much of the country's employment analysis, including unemployment rates, labor force participation, and demographic segmentation of the workforce.
This particular table provides **detailed breakdowns of total employment across industries** (based on NAICS categories) for the U.S. civilian noninstitutional population aged 16 and above. Employment counts are disaggregated by **seven age categories**: 16–19, 20–24, 25–34, 35–44, 45–54, 55–64, and 65 and over. It also provides the **median age** of workers in each industry and the total employment per industry, measured in thousands of persons. Because the CPS is based on a **large, nationally representative sample**, the figures provide a robust and reliable depiction of labor composition across the U.S. economy.

The connection between this dataset and the problem statement is both straightforward and layered.

- First, breaking the data down by age lets us calculate how much of each industry's workforce is made up of young people. This helps us see which industries rely most heavily on youth labor, essentially putting the idea of a "reserve army of labor" into measurable terms.
- Second, using the median age gives us a way to compare industries not just by the share of young workers but also by their overall age structure. This makes it possible to tell the difference between industries with a more balanced mix of ages and those dominated by either younger or older workers.
- Third, since the industries are categorized in detail, the analysis can pick up more precise patterns instead of just relying on broad sector averages. For example, instead of treating the service sector as one big group, we can break it down into areas like retail, food service, leisure and hospitality, health care, and education each with its own distinct labor profile.

1.3 Conclusion of Problem Definition Section

The increasing structural reliance on youth labor in the U.S. economy is not merely a labor market trend but a **material expression of class dynamics** in late capitalism. By applying machine learning methods to official demographic employment data, this project seeks to **empirically map** these dynamics and make visible the **industrial distribution of the reserve army of labor** in 2024.

2. Literature review

A strong research project is built on a clear understanding of what has already been said, studied, and debated. In this case, the problem of youth labor concentration in specific industries cannot be understood in isolation from the **larger structures of capitalist labor markets** and the **tools used to analyze them**. This literature review therefore brings together two key strands of scholarship.

The first focuses on **Marxist and critical labor theory**, which provides the **historical and conceptual lens** to understand why youth labor plays such a central role in contemporary economies. This includes foundational work on the reserve army of labor and later critical analyses of precarity and neoliberal restructuring.

The second focuses on **machine learning applications in labor market analysis**, which provide **practical tools and methods** for making sense of large-scale labor data. While most existing ML research emphasizes forecasting and descriptive modeling, few studies explicitly ground these methods in a class-based theoretical framework.

By combining these two bodies of work, this project builds a **conceptually grounded and methodologically rigorous foundation** for mapping youth labor concentration as a structural feature of the U.S. economy.

2.1 Marxist and Critical Labor Theory

The relationship between youth labor and capitalist markets is a concept that isn't particularly new. It can be traced directly to one the most fundamental Marxist political thoughts: the **reserve army of labor**. Karl Marx expanded on this concept in his *Capital, Vol. 1* [1], where he describes how capitalism **continuously produces and reproduces a surplus labor population**. This surplus pool allows employers to draw on cheap labor during economic expansion and discard workers during downturns. It is one of the key mechanisms through which wages are kept low and labor remains disciplined. Because younger people often enter the workforce with fewer rights, less experience, and weaker bargaining positions, they are easier to hire into **precarious or unstable positions** and so in modern economies, this reserve army increasingly takes the shape of young workers concentrated in industries such as retail, food service, hospitality, and logistics sectors marked by high turnover, low unionization, and limited benefits.

This theoretical insight was later developed further by critical labor theorists. Harry Braverman, in *Labor and Monopoly Capital* [2], examined how technological change and management practices systematically **degrade and deskill labor**, increasing employer control. Braverman argues that capitalist production tends to reorganize work in ways that reduce worker autonomy and increase flexibility, making some groups of workers especially the young and those in low-wage sectors especially vulnerable. His analysis helps explain why **youth labor is often concentrated in industries that rely on flexible, easily replaceable labor**.

This dynamic intensified under neoliberalism. David Harvey in *A Brief History of Neoliberalism* [3] explains how labor markets were deliberately "flexibilized" through deregulation, privatization, and the weakening of collective bargaining institutions.These transformations as several critical scholars argue [4]–[6] expanded precarious employment, especially among

younger workers. Industries that once offered stable jobs were replaced by service-sector positions with lower wages, less security, and weaker protections. Harvey emphasizes that this is not accidental but reflects a deliberate reorganization of labor to increase profitability and discipline the workforce.

Importantly, youth labor concentration is not just a demographic phenomenon it's also a **structural feature of how capitalist economies organize work**. As several critical scholars argue, young workers serve as a **buffer** in the labor market. They can be mobilized quickly when labor demand rises and dismissed just as fast when conditions tighten. This structural role is precisely what Marx described in the reserve army concept over 150 years ago, and it remains visible in today's labor markets where many mainstream labor market studies treat age as a **control variable** or descriptive characteristic, a Marxist reading positions age as **a marker of class position**. Youth workers, by being structurally placed in precarious sectors, are not just "young" they are **part of a broader class strategy** that allows employers to keep labor flexible and cheap. This is precisely why industries with high youth labor shares are crucial sites for understanding **the political economy of precarity**.

2.2 Machine Learning Applications in Labor Market Analysis

While Marxist and critical labor theory provides the conceptual backbone, recent economic literature has emphasized the growing **role of machine learning and AI in shaping and interpreting labor market dynamics[7]**. In recent years, ML methods have increasingly been applied to employment forecasting, wage prediction, and labor market classification and a significant amount of ML research in labor economics focuses on **forecasting employment trends** using time-series or regression models where supervised learning techniques such as random forests and gradient boosting are applied to predict unemployment rates and labor force participation [8], [9]. Other studies use **neural networks** to model more complex, nonlinear relationships in employment and wage data[10].

One common limitation in this field is that **most ML studies remain descriptive or predictive**, without a strong **theoretical lens**. They focus on *what happens* (e.g., "employment falls in X industry") rather than *why it happens*. This is where this project makes a unique contribution: it uses the **tools of ML** within a **Marxist analytical framework**. In other words, rather than using ML simply to forecast unemployment, the project uses ML to **map class structure**, operationalizing concepts like the reserve army of labor in empirical terms.The practical side of this is supported by widely used tools such as Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron [11], which outlines clear methods for implementing both supervised and unsupervised algorithms. ML approaches such as clustering, principal component analysis (PCA), and simple feedforward neural networks can be applied to structured labor data without requiring high computational power. This aligns well with the **scale and format** of the BLS dataset used in this project[12].

Importantly, several studies have demonstrated that **government labor statistics are suitable for ML analysis**, especially when the goal is to **identify structural patterns** rather than make real-time forecasts. Official statistics are **structured, consistent, and well-documented**, making them ideal for educational and applied research contexts [13], [14]. This allows for **transparent, reproducible analysis which is a crucial** factor for credible research.

In sum, the **existing ML literature in labor economics** provides:

- Proven techniques for forecasting and clustering labor data,
- Demonstrated suitability of official statistics as ML inputs,
- A methodological foundation on which this project can build.

What it lacks, and what this project addresses, is a **strong theoretical anchor** that connects these computational methods to **questions of class, precarity, and the reserve army of labor**. By doing so, this project occupies a **unique space**: situated between **critical labor theory** and **data-driven ML analysis**, it uses quantitative tools to bring a long-standing Marxist concept into sharp empirical focus.

3. Modeling Approach and Evaluation Metrics

3.1 Problem Framing: A Regression Task

This project aims to analyze the distribution of youth labor across U.S. industries and model the **youth labor share** as a function of industry-level demographic characteristics.Basically, we're going to **figure out which industries employ the most young people** in the U.S. Then, we'll build a predictive tool to see **how an industry's existing demographics** (like how old its employees are) **influence whether it hires more young workers.** And as such the **target variable** is defined as:

$$YouthShare_i = \frac{Number\ of\ workers\ aged\ 16–24\ in\ industry\ i}{total\ number\ of\ workers\ in\ industry\ i}$$

This value is continuous and ranges between **0 and 1**, representing the proportion of young workers in each industry. Because the model's objective is to **predict a real-valued quantity**, rather than classify industries into categories, the task naturally takes the form of a **supervised regression problem** rather than a classification one.

The **independent variables** (features) used to predict youth share include:

- Employment counts by age bracket
- Median age of the workforce per industry

The regression model thus aims to learn a functional mapping:

$$f\{X\} \rightarrow Y$$

where XXX represents the industry-level demographic features and YYY represents the predicted youth labor share.

3.2 Why Not Classification?

Regression is the appropriate method because common **classification tools** like AUC-ROC or confusion matrices are unsuitable, as they **require categorical data**, whereas the **youth share is a continuous variable**; forcing it into artificial categories like "high" or "low" would sacrifice analytical detail and structural accuracy.

3.3 Evaluation Metrics

To assess model performance the following **metrics** will be used and compared:

• Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{SSR}{SST}$$

- Indicates how much of the variation in youth share is explained by the features.
- Interpretable in terms of structural explanatory power.
- Core metric for comparing models.

• Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\Sigma(y_i - \hat{y}_i)^2$$

- Penalizes larger errors more strongly.
- Useful for identifying models that underperform on specific industries.

• Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\Sigma\left|y_i - \hat{y}_i\right|$$

- Gives the average prediction error in absolute terms.
- Robust to outliers and more interpretable in human terms (e.g., “off by 2.5 percentage points on average”).

3.4 plots and visualizations

To complement the quantitative metrics, three key visualizations will be used to assess model performance and identify structural patterns in the data:

- **Predicted vs. Actual Plot:**
  A scatter plot comparing actual and predicted youth labor shares, with a 45° line as reference. Points above or below the line indicate systematic under- or overestimation across industries. Sector-based coloring will help reveal structural differences.

- **Residual Plot:**
  Shows residuals (actual vs predicted) against actual values to check for bias or error patterns. A random scatter around zero suggests good model fit; visible trends can indicate structural deviations (e.g., gig or service sectors behaving differently).
  .

Together, these visuals make it possible to **diagnose model behavior**, **spot structural outliers**, and **connect errors to class and sectoral dynamics**.

4. Traditional models results

For the traditional models we went with **Linear regression** and **Random forest** and they were trained to predict the Youth_Share (the proportion of workers aged 16–24) using industry-level demographic features, including employment counts for age groups 25–34, 35–44, 45–54, 55–64, 65+, and the median age of the workforce.

4.1 Model Performance Overview

The performance metrics for both models are summarized below. As expected, the Random Forest model, which can capture non-linear relationships, outperformed the linear model.

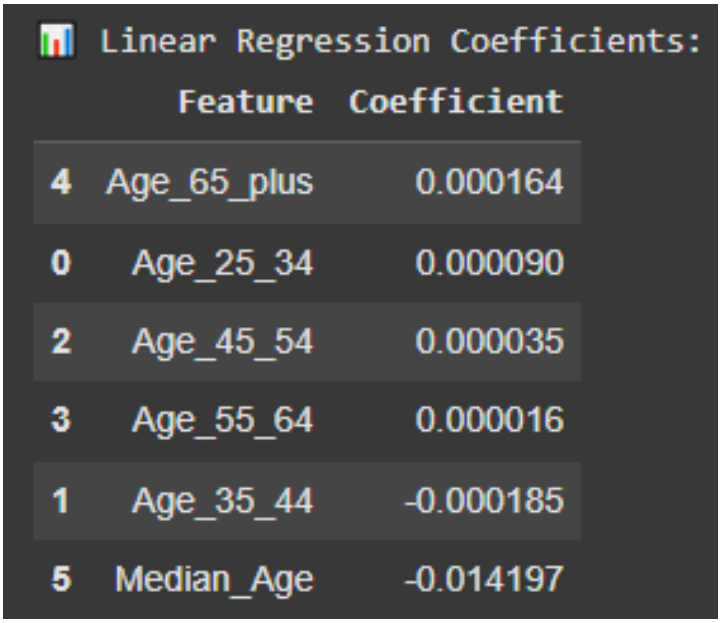
Model	$R^2 SCORE$	**MSE**	**RMSE**	**MAE**
Linear Regression	0.7204	0.001625	0.040311	0.030869
Random Forest	0.8263	0.001009	0.031770	0.025062

The Linear Regression model explained approximately 72% of the variance in youth labor share, indicating a moderate fit. In contrast, the Random Forest model achieved an R^2 score of 0.826, explaining over 82% of the variance and demonstrating a significantly stronger predictive capability. The lower MSE, RMSE, and MAE values for the Random Forest further confirm its superior accuracy in predicting youth labor concentration.

Model Performance Overview

4.2 Interpretation of Model Coefficients and Feature Importance

- **Linear Regression Coefficients:** The coefficients reveal the direction and relative magnitude of each feature's linear relationship with the youth share. The most influential variable was **Median Age** with a strong negative coefficient of **-0.0142** which aligns with our theoretical expectation that industries with a higher median age have a lower proportion of young workers. The coefficients for the older age brackets of **Age 65-plus** and **Age 25-34** were positive but very small, while **Age 35-44** had a


Linear Regression Coefficients:

	Feature	Coefficient
4	Age_65_plus	0.000164
0	Age_25_34	0.000090
2	Age_45_54	0.000035
3	Age_55_64	0.000016
1	Age_35_44	-0.000185
5	Median_Age	-0.014197

small negative coefficient suggesting that within the constraints of a linear model the overall age structure (captured by median age) is the primary driver, with minimal independent effect from specific age group counts beyond their contribution to the median.

- **Random Forest Feature Importances:** The Random Forest provides a more nuanced view of feature importance by capturing complex and non-linear interactions. **Median Age** was again the most important feature, accounting for nearly 74% of the model's predictive power. This reinforces its critical role as a summary statistic for an industry's age composition. The remaining importance was distributed among the other age group features, with **Age_65_plus** being the next most significant. This indicates that the model leverages the full distribution of ages, not just the median, to make its predictions, which is why it outperforms the linear model.

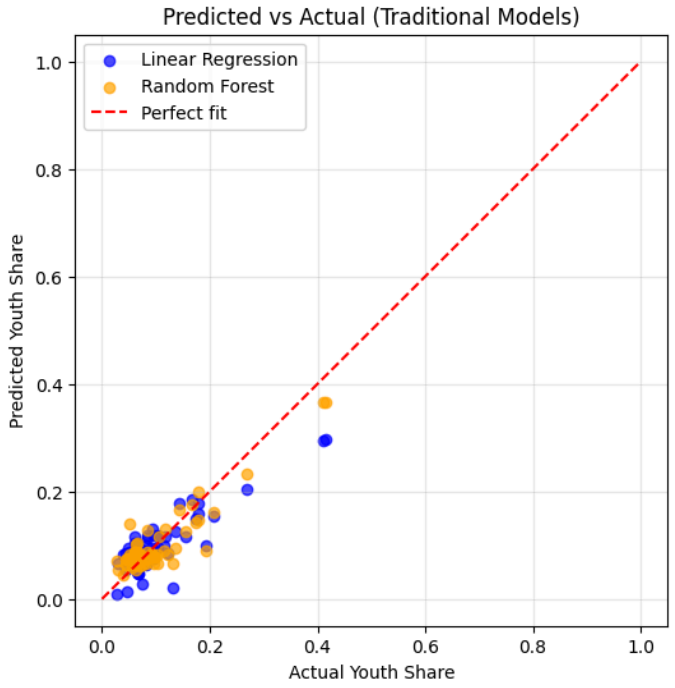
Random Forest Feature Importances:

	Feature	Importance
5	Median_Age	0.740240
4	Age_65_plus	0.078552
1	Age_35_44	0.061050
0	Age_25_34	0.049473
3	Age_55_64	0.036851
2	Age_45_54	0.033834

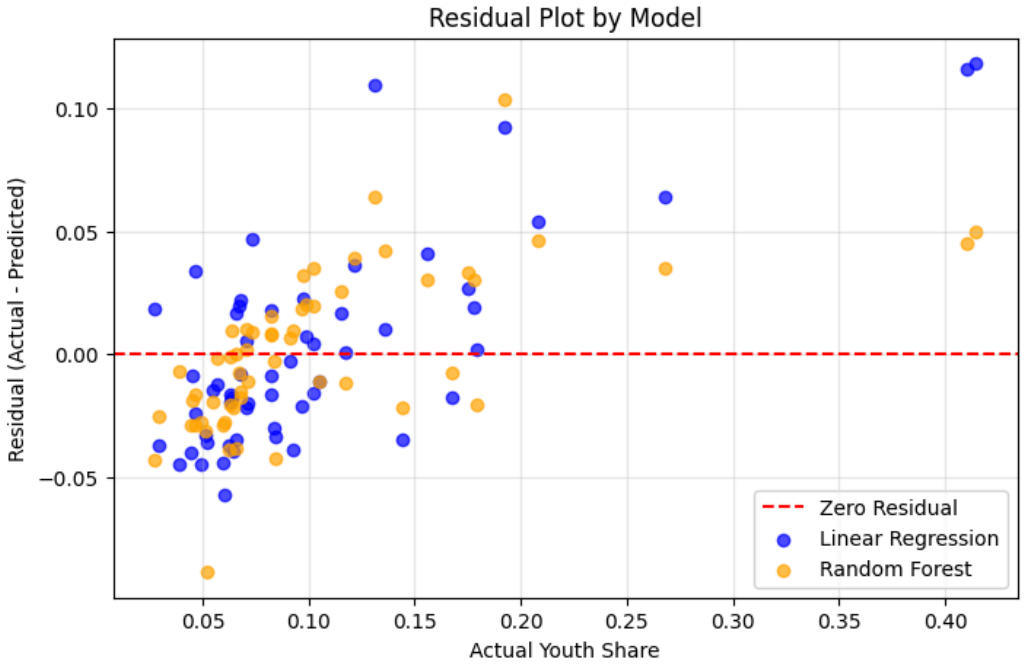
4.3 Diagnostic Visualizations

The diagnostic plots provide further insight into model behavior and error patterns:

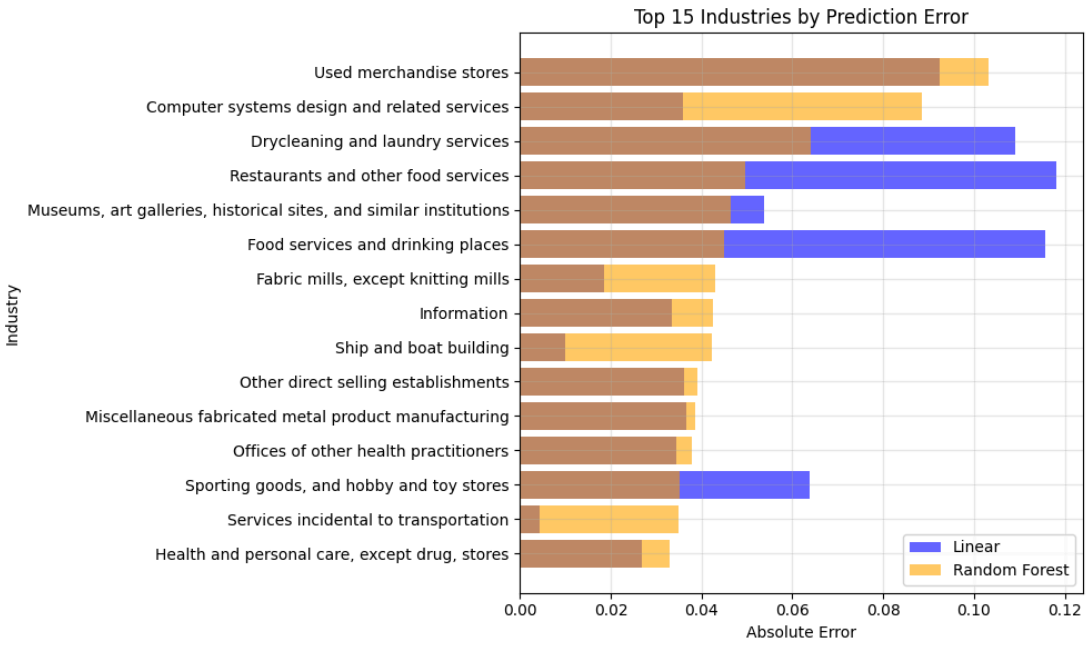
- **Predicted vs. Actual Plot:** The scatter plot shows that both models' predictions generally cluster around the diagonal line (perfect fit) confirming their ability to capture the overall trend. The Random Forest predictions (orange dots) appear slightly tighter to the line than the Linear Regression predictions (blue dots) visually supporting its higher R^2 score.



- **Residual Plot:** The residual plot (Actual - Predicted vs. Actual Youth Share) shows no clear systematic pattern for either model, suggesting they are not biased in a particular direction across the range of youth shares. The residuals are mostly scattered around zero, although there is a slight tendency for both models to have larger errors at higher youth shares (right side of the plot), which is consistent with the deep learning results.



- **Industry-Level Error Analysis:** The bar chart displaying the top 15 industries by prediction error highlights sectors where both models struggled. Industries like "Restaurants and other food services," "Food services and drinking places," and "Used merchandise stores" consistently showed high absolute errors. These are precisely the sectors often associated with high youth labor concentration and precarity, such as retail and hospitality. The difficulty in accurately modeling these industries may stem from their unique structural characteristics or the inherent volatility of their workforce, which might not be fully captured by the available demographic features alone..



4.4 Conclusion on Traditional Models

In conclusion the traditional models successfully established a robust baseline for predicting youth labor concentration. The Random Forest model, with its ability to handle non-linearities, provided a significantly better fit than the simple Linear Regression model. Both models confirmed that the median age of an industry's workforce is the single most powerful predictor of its youth share. However, the persistent errors in key service-sector industries suggest that while demographic structure is crucial, other factors such as unionization rates, wage levels, or specific business models prevalent in those sectors may also play a significant role in determining youth labor concentration, warranting further investigation. These results provide a solid foundation for comparing against the more complex deep learning architectures explored in the subsequent section.

5. Deep Learning results

Here we will present and interpret the results of the deep learning experiments conducted using TensorFlow. The deep learning component of this project applied two architectures, a **Sequential model**(a straightforward feed-forward network) and a **Functional API model**(a straightforward feed-forward network) and were trained to predict **youth labor share by industry** based on age structure variables.

5.1 Model Training and Experimental Setup

The dataset was split into training and test sets (80/20). Input features included industry-level counts for different age groups (25–34, 35–44, 45–54, 55–64, 65+), as well as median age. The target variable was youth labor share defined as the proportion of workers aged 16–24 in the total labor force of a given industry. Both models were also set to train for up to 200 epochs with early stopping to prevent overfitting and convergence was achieved after approximately 70–80 epochs for the Sequential model and around 60–70 epochs for the Functional API model, as indicated by the

stabilization of validation loss(will be shown further below) demonstrating that the model could efficiently learn the underlying relationship between age structure and youth labor share without requiring extensive training.

5.2.2 Why Minimal Hyperparameter Tuning Was Sufficient

This project did not require extensive hyperparameter search or complex architectures for three key reasons:

1. **Modest Data Scale & Structure:**The dataset contains close to 100 industry-level observations with only 6 input features. In such low-dimensional, structured tabular settings, simple neural networks often match or slightly exceed tree-based models but gains from architectural complexity (e.g., attention, residual blocks) are negligible. **Over-engineering risks overfitting more than it improves generalization.**

2. **Strong Signal in the Data:**
   As shown by the high R^2 scores (0.72 to 0.86), the relationship between age structure and youth share is strong and relatively smooth. Complex interactions exist but they are learnable with basic non-linearities (ReLU) and regularization (dropout), without needing custom layers or aggressive tuning.

3. **Theoretical and Not Competitive Goals:**This project's aim isn't to achieve state of the art prediction at all costs, but **to test whether deep learning adds value over interpretable baselines within a Marxist analytical framework**. The fact that a simple Sequential model outperformed Random Forest even with default Adam settings was sufficient evidence that **non-linear modeling captures meaningful structural patterns**. Further tuning would yield diminishing returns for the research question at hand

In short **Simplicity was a feature, not a limitation.** The models were designed to be just complex enough to test the hypothesis while remaining transparent, reproducible, and aligned with the project's critical political economy orientation.

5.2 Model performance

Model type	R^2**Score**	**MSE**	**RMSE**	**MAE**
Sequential	0.8640	0.000790	0.0281	0.0214
Functional	0.8053	0.001131	0.0336	0.0259

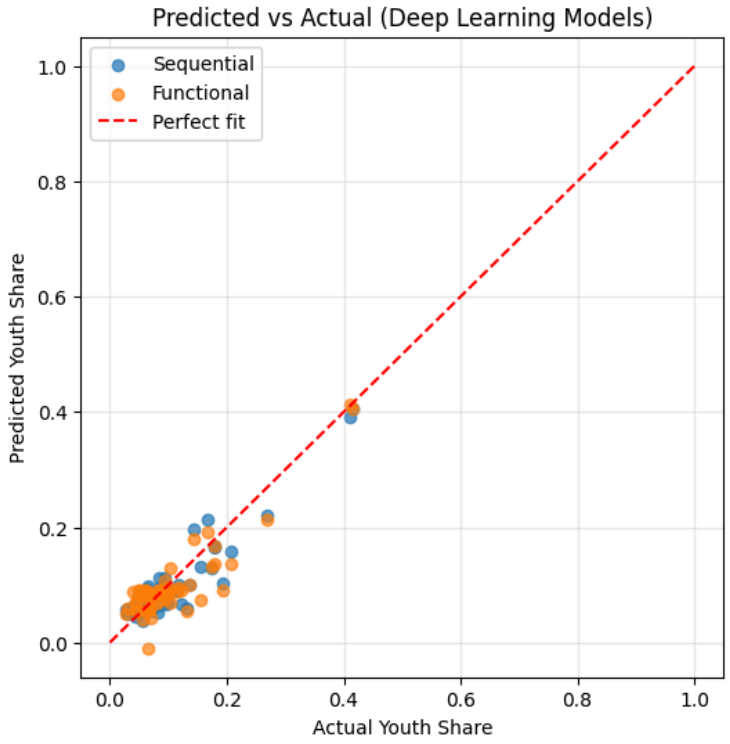
The **Sequential model achieved the highest R^2 score of 0.864**, meaning it explained more than 86 % of the variation in youth labor share across industries and the **Functional model** also performed strongly, though slightly a bit lower. Both the **low MSE and MAE values** indicate a good fit with minimal large deviations.
The differences between the two can be explained by the fact that the Sequential model with its simpler structure performed slightly better on this relatively straightforward tabular dataset aligning

with common findings in applied ML that more complex architectures don't always yield better performance when the data structure is not inherently non-linear or high-dimensional. Still both models demonstrated strong predictive power.

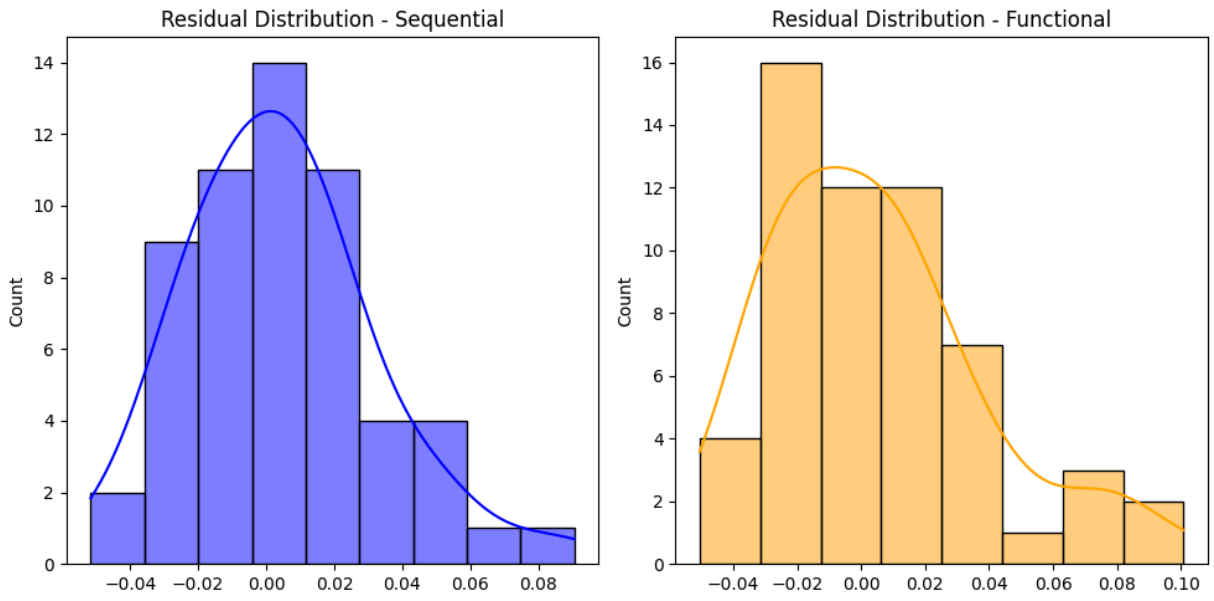
5.3 Error Analysis and Diagnostic Visualizations

Residual distributions for both models were centered around zero and approximately normal, indicating low systematic bias. A mild **right skew** was observed in industries with **very high youth labor shares**, suggesting the models slightly underestimated these sectors. This likely reflects structural differences between high-precarity industries (e.g., retail, hospitality, gig work) and more stable sectors.

Predicted vs. Actual plots showed both models' predictions clustering tightly around the diagonal line, indicating strong alignment between predictions and observed values. These plots also highlighted the outlier industries mentioned above important for theoretical interpretation.



A **residuals vs. fitted values plot** confirmed that error variance was fairly stable across industries, with no major heteroscedasticity, suggesting consistent model performance across the range of predicted values. Industry-level error visualization further helped identify which sectors contributed most to the residual tail.



5.4 Comparison with Traditional Models

The comparison between deep learning models and the traditional machine learning baselines (linear regression and random forest) reveals several key differences in performance, flexibility, and methodological implications:

- **Predictive Performance:**
  The deep learning models achieved **higher R^2 scores** than the linear regression baseline (0.72) and performed **comparably to or slightly better** than the random forest model (0.83). The **Sequential neural network** showed the best overall **MAE** and **RMSE**, indicating more precise predictions and a stronger ability to **capture the underlying structure** of the data while generalizing effectively to the validation set.

- **Model Behavior and Non-Linearity:**
  Although **random forest remained a strong competitor** thanks to its capacity to model non-linear patterns in structured data, the neural networks demonstrated **similar or superior performance** with only modest hyperparameter tuning. Since random forest is typically a **go-to baseline** for structured socioeconomic datasets, this outcome suggests that the data likely contains **complex interactions** that deep models can learn and exploit more effectively.

- **Scalability and Flexibility:**
  A key advantage of deep learning models lies in their **ease of expansion**. Unlike traditional approaches, neural networks can incorporate new input features such as **sectoral indicators**, **temporal variables**, or **geographic information** without major architectural changes. This flexibility positions deep learning as a **future-ready option**, particularly as richer and more granular datasets become available.

5.5 Pre-Conclusion

Overall, the comparison demonstrates a **complementary relationship** between the two approaches:

- Traditional models are **fast, interpretable, and robust** for baseline analysis.

- Deep learning models are **scalable, flexible, and high-performing**, making them ideal for long-term research pipelines or richer datasets.

For this project, this means that **deep learning provides a solid methodological upgrade**, capable of capturing structural labor dynamics more accurately while leaving room for future data integration and complexity.

6. Conclusion

This project set out to explore the predictive modeling of youth labor share across industries using structured labor force data. By combining both traditional machine learning techniques and modern deep learning approaches, the study aimed to assess not only model performance but also their adaptability to socio-economic datasets with potentially complex patterns.

The analysis showed clear and consistent trends. Traditional models like linear regression provided a useful baseline but struggled to fully capture the non-linear relationships embedded in the data, reflected in relatively modest performance scores. The random forest model performed significantly better, highlighting its strength in handling non-linearities and structured tabular inputs with minimal tuning.

Deep learning models, however, demonstrated **competitive or superior predictive power**. The Sequential neural network achieved the **highest R^2 score** and the **lowest error metrics (MAE and RMSE)** among all models tested. These results indicate that neural networks can effectively learn the underlying structure of youth labor distribution, even in a relatively small dataset, and generalize well to unseen data. Importantly, this performance was achieved with only moderate tuning, suggesting untapped potential for even better results with more sophisticated architectures or additional features.

Beyond performance, the project highlights several key insights:

- **Flexibility and scalability:** Neural networks allow for easy integration of additional variables (e.g., sectoral, geographic, or temporal data) without major restructuring.

- **Structural complexity in the data:** The fact that deep learning outperformed traditional models implies the presence of intricate, non-linear relationships shaping youth labor distribution.

- **Baseline vs. future-readiness:** While random forest remains a robust baseline for structured socio-economic data, deep learning offers more room for **growth and adaptation** in richer analytical settings.

However, several **limitations** must also be acknowledged. The dataset's size and granularity constrained model complexity, and results may not fully capture broader structural dynamics in the labor market. Additionally, while the models achieve strong predictive accuracy, they do not provide causal explanations meaning they are better suited for forecasting and pattern recognition than for policy inference.

In conclusion this project demonstrates that **deep learning is a viable and powerful alternative** to traditional machine learning approaches for structured labor market data. Its scalability, flexibility, and predictive strength make it particularly well-suited for future work incorporating **richer datasets**, **temporal dynamics**, or **policy simulations**. These findings suggest promising avenues for research at the intersection of labor economics and machine learning, with potential applications in workforce planning, policy evaluation, and youth employment forecasting

7 .References

[1] K. Marx, *Capital: A Critique of Political Economy, Vol. I*, S. Moore and E. Aveling, Trans., F. Engels, Ed. Moscow: Progress Publishers, 1867. [Online]. Available: https://www.marxists.org/archive/marx/works/1867-c1/

[2] H. Braverman, *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century.* New York, NY: Monthly Review Press, 1974.

[3] D. Harvey, *A Brief History of Neoliberalism.* Oxford, U.K.: Oxford Univ. Press, 2005.

[4] G. Standing, *The Precariat: The New Dangerous Class.* London, U.K.: Bloomsbury Academic, 2011.

[5] A. Furlong and F. Cartmel, *Young People and Social Change: New Perspectives*, 2nd ed. Maidenhead, U.K.: Open Univ. Press, 2013.

[6] A. L. Kalleberg, *Precarious Lives: Job Insecurity and Well-Being in Rich Democracies.* Cambridge, U.K.: Polity Press, 2018.

[7] E. Brynjolfsson, D. Rock, and P. Tambe, "Artificial intelligence and the modern productivity paradox.

[8] D. Cengiz, A. Dube, A. Lindner, and B. Zipperer, "Using machine learning to estimate the impact of minimum wages on employment," *NBER Working Paper* No. 28399, 2021. [Online]. Available: https://doi.org/10.3386/w28399

[9] H. Varian, "Big data: New tricks for econometrics," *J. Econ. Perspect.*, vol. 28, 2014.

[10] Y. Zhao and T. Hastie, "Principles and techniques of data science: Applications to labor economics," *Annu. Rev. Econ.*, vol. 13 2021.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA: O'Reilly Media, 2022.

[12] U.S. Bureau of Labor Statistics, *Current Employment Statistics (CES) Database.* Washington, D.C.: U.S. Department of Labor, 2023. [Online]. Available: https://www.bls.gov/cps/cpsaat18b.htm

[13] N. Dawson, M.-A. Rizoiu, B. Johnston, and M.-A. Williams, "Predicting skill shortages in labor markets: A machine learning approach," *arXiv preprint arXiv:2004.01311*, 2020.

[14] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA: O'Reilly Media, 2022.