Generation Z and the New Economic Reality: A Machine Learning Perspective on Financial Challenges

[1*]Abdullah Aljishi, [2]Matin Marjani, [3]Arash Latifi and [4]Lior Shamir

[1, 2, 3]Department of Electrical and Computer Engineering, Kansas State University, Kansas, United States
[4]Department of Computer Science, Kansas State University, Kansas, United States

email: [1*]aljishi@ksu.edu, [2]matinmarjani@ksu.edu, [3]arashlatifi@ksu.edu, [4]lshamir@ksu.edu

**Corresponding author*

Received: 07 February 2025 | Accepted: 25 April 2025 | Early access: 01 May 2025

Abstract - *This study explores the socioeconomic disparities and financial challenges faced by different generational cohorts, with a focus on Generation Z. The research aims to identify patterns in socioeconomic features, such as income distribution and housing affordability, that distinguish generations and impact their financial outcomes. Machine learning models were used, with classification models that predicted generational membership and regression models that estimated the year of birth as a continuous variable. Using mutual information for feature selection, the Explainable Boosting Machine (EBM) achieved the highest classification accuracy of 74.78%, as evaluated using 10-fold cross-validation, while regression analysis demonstrated moderate predictive power (R^2 = 0.6005) with an average absolute error of eight years. The results highlight significant generational differences, with Generation Z experiencing the highest median rent-to-income burden (60.0%) and substantial barriers to homeownership. Despite higher participation in the workforce compared to previous generations at similar life stages, systemic economic challenges, such as rising housing costs and stagnant wages, disproportionately affect Generation Z. These findings underscore the utility of machine learning in identifying generational trends and socioeconomic disparities, offering a framework for further research to refine models and explore additional socioeconomic variables to enhance understanding of generational dynamics. Code and data to reproduce the results are available in GitHub, as detailed in the Dataset Overview subsection.*

Keywords: Generation Z, machine learning, socioeconomic disparities, financial challenges, income inequality.

1 Introduction

For decades, the idea of generational progress has been a cornerstone of societal development: each generation has historically achieved better financial outcomes, improved living standards, and greater opportunities than the one before (Chetty et al., 2017). From the Silent Generation, who endured the hardships of the Great Depression and World War II but benefited from post-war economic growth (Elder, 2018), to the Baby Boomers, who enjoyed access to affordable housing, stable jobs, and rising wages (Patterson, 1996), this upward trajectory seemed inevitable. However, recent discourse suggests that Gen Z may be the first generation to diverge from this trend, facing significant challenges in achieving financial stability and upward mobility relative to their parents and predecessors (Gregory, 2023; Lev, 2021).

Understanding these challenges requires examining the distinct historical, cultural, and socioeconomic contexts that have shaped each generation. Generational divisions, categorized by birth years and corresponding age ranges, are illustrated in Figure 1 (Ipsos, 2023). While Baby Boomers benefited from the economic prosperity of the post-war era, Generation X saw the rise of dual-income households (Bianchi, 2000), increased access to higher education (Bound & Turner, 2007), and advancements in technology that began to reshape industries (Autor et al., 1998). Millennials grew up during the rapid expansion of the internet and digital technologies, which created new opportunities for innovation and connectivity (Palfrey & Gasser, 2011). However, as economic landscapes have evolved, Generation Z is now widely believed to be grappling with skyrocketing housing costs, stagnant wages, mounting student debt, and a job market increasingly dominated by gig work and automation. These economic shifts are thought to have compounded the financial difficulties of Gen Z, potentially leaving them with fewer opportunities for upward mobility compared to prior generations (Twenge, 2023). This paper seeks to investigate whether these claims hold true by examining economic data and generational trends.

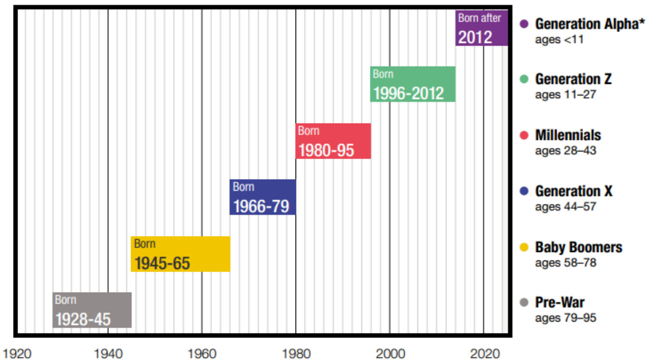


Figure 1: Generational categories by birth year and age in 2023. Adapted from Ipsos (2023)

Previous generations often benefited from pathways to financial security, such as affordable housing and stable income growth. In contrast, Generation Z appears to face significant barriers to wealth-building, particularly through homeownership, which has historically been a key driver of economic mobility (Harvard Joint Center for Housing Studies, 2006). Systemic factors, including growing income inequality, reduced union representation, and living costs consistently outpacing wage growth, are thought to exacerbate these challenges (Mishel & Bivens, 2021; Western & Rosenfeld, 2011).

The issue of income inequality and its evolution across generations has been the subject of extensive research. Prior studies have explored the economic disparities between generational cohorts, such as Baby Boomers, Generation X, Millennials, and Generation Z, focusing on factors like wage stagnation, inflation, and wealth accumulation. For instance, Charles and Hurts (2003) studied how wealth is passed between generations, finding that children's wealth is strongly influenced by their parents' wealth, with lifetime income and asset ownership playing the biggest roles. Gallipoli et al. (2020) examined the joint evolution of cross-sectional inequality in income and consumption across generations.

Research on generational economics has also highlighted the influence of macroeconomic factors, such as recessions and housing market trends, on the financial well-being of different generations. Studies such as Green and Lee (2016) have provided insights into the demand for housing based on age and demographics. However, many existing analyses rely on aggregated data or limited timeframes, which fail to capture the granular differences across individual-level datasets.

In the field of predictive modelling, recent advancements have been made in using machine learning to analyze socioeconomic patterns. For example, Fan et al. (2023) demonstrated the utility of classification and regression models in predicting socioeconomic outcomes based on complex interactions of urban features. While these approaches have proven effective, their application to understanding generational disparities in income remains underexplored.

This study aims to build on this body of work by investigating key factors such as income distribution, housing affordability, and inflation-adjusted wages. By identifying trends and disparities, this research seeks to determine whether Gen Z represents a significant deviation from historical patterns and to understand the systemic changes required to address these challenges.

The remainder of this paper is organized as follows: Section 2 details the dataset and preprocessing procedures, while Section 3 outlines the methodology employed in this study. Section 4 describes the economic metrics used to evaluate generational economic conditions and disparities. Section 5 presents and discusses the findings, and finally, Section 6 concludes the paper and offers recommendations for future research.

2 Data Source and Acquisition

Data for this study was obtained from the IPUMS USA database (Ruggles et al., 2024), a comprehensive resource providing integrated microdata for social and economic research. The dataset comprises anonymized individual-level records with variables covering a range of topics, including demographics, housing characteristics, and employment status. The data span multiple years, enabling longitudinal analysis of generational trends.

2.1 Dataset Overview

The raw dataset obtained from IPUMS spans a comprehensive temporal range, covering the years from 1970 through 2023, which includes individual survey data from the years 1970, 1980, 1990, 2000, and annually from 2001 to 2023. This dataset encapsulates various individual and household attributes across approximately 100,000 respondents, representing a diverse cross-section of the U.S. population.

Key features in the dataset include `YEAR` (year of the survey), `BIRTHYR` (birth year), `RENT` (monthly rent), `VALUEH` (home value), `INCTOT` (total personal income), and categorical variables such as `STATEFIP` (state), `RACE` (race), and `EDUC` (educational attainment). The extensive span and substantial size of the dataset necessitated the implementation of efficient data management techniques, particularly crucial due to the presence of both continuous and categorical variables, as well as coded missing values across several features.

All the code and related files used in the preprocessing pipeline are available in the project's GitHub repository (Aljishi et al., 2025). The repository includes all scripts and resources necessary to replicate the data processing steps, including data splitting, feature engineering, handling missing values, sampling, balancing, and normalization. Additionally, it provides the final clean and balanced dataset used in the analysis, along with detailed instructions on setting up and running the pipeline.

2.2 Data Preparation

Several new features were engineered to facilitate analysis. To ensure comparability of monetary values across the years, features such as `INCTOT` (total personal income) and `VALUEH` (home value) were adjusted for inflation. Using official inflation indices, all monetary values were converted to constant 2023 dollars, providing a standardized economic baseline for longitudinal analysis. A `GENERATION` feature was derived from `BIRTHYR`, assigning individuals to specific generational cohorts: Baby Boomers (1946-1964), Generation X (1965-1980), Millennials (1981-1996), and Generation Z (1997-2012). Individuals under the age of 18 were excluded from the analysis, as the study focuses exclusively on adults. Finally, categorical features such as `STATEFIP`, `SEX`, and `EDUC` were transformed into binary (one-hot encoded) representations to enable their use in machine learning models. This ensured a consistent numerical format across the dataset.

2.3 Handling Missing Values

A systematic approach was implemented to address the missing data. First, explicitly coded missing values (e.g., `9999` for `RENT` or `9999999` for `VALUEH`) were identified and replaced with standard NaN (Not a Number) representations for consistent handling across analyses. Subsequently, rows containing missing values in key features were removed. Finally, features exhibiting a high proportion of missing data, deemed non-essential for the analysis, were excluded entirely.

2.4 Sampling and Balancing

The dataset exhibited imbalances in representation across both years and generations. To mitigate these imbalances and ensure fair comparisons, several sampling strategies were employed. First, to achieve equitable representation across years, the data were sampled to ensure a consistent number of observations per year, accounting for variations in the original yearly dataset sizes. Second, within each year, the generational balance was addressed through a combination of undersampling of overrepresented generations and oversampling of underrepresented generations, particularly Generation Z. Finally, to further refine generational representation and account for age-related biases, an age-based sampling approach was implemented, undersampling older individuals from dominant generations such as Baby Boomers while retaining younger individuals across all generations.

2.5 Final Dataset Structure

The resulting dataset comprises balanced samples across years and generations. Continuous variables, including `INCTOT` and a derived poverty indicator, were standardized using z-score normalization. Categorical variables were fully one-hot encoded, as described previously. Temporal features, specifically `YEAR` and `AGE` features, were retained to facilitate longitudinal and age-based analyses. The detailed descriptions of the features included in the dataset is provided in Table 1.

Table 1: Features Descriptions

FEATURE	DESCRIPTION
GENERATION	Generation of person
YEAR	Census year
ROOMS	Number of rooms
NFAMS	Number of families in household
NCHILD	Number of own children in the household
YNGCH	Age of youngest own child in household
AGE	Age
BIRTHYR	Year of birth
POVERTY	Poverty status
OCCSCORE	Occupational income score
ERSCOR50	Occupational earnings score, 1950 basis
NPBOSS50	Nam-Powers-Boyd score, 1950 basis
STATEFIP	State (FIPS code)
OWNERSHP	Ownership of dwelling (tenure)
KITCHEN	Kitchen or cooking facilities
PLUMBING	Plumbing facilities
UNITSSTR	Units in structure
PHONE	Telephone availability
CBNSUBFAM	Number of subfamilies in household
SEX	Sex
MARST	Marital status
RACE	Race
BPL	Birthplace
SCHOOL	School attendance
EDUC	Educational attainment
EMPSTAT	Employment status
INCTOT_ADJUSTED	Total personal income adjusted for inflation
FTOTINC_ADJUSTED	Total family income
HOUSING_VALUE_ADJ	Monthly contract rent or House value

3 Methodology

The methodology employed in this study integrates statistical analysis and machine learning techniques to examine generational income disparities and predict socioeconomic trends. The analysis incorporates feature selection and correlation analysis to enhance the accuracy and interpretability of the models. It focuses on three main tasks: predicting generational categories and birth years using classification and regression models, analyzing income distributions to assess inequality across generations, and examining how different variables correlate with birth year to identify significant predictors of economic outcomes.

3.1 Feature Selection

For feature selection, we employed mutual information (MI) as a metric to assess the dependency between individual features and the target variable. MI measures both linear and non-linear relationships, making it a robust tool for identifying features with strong predictive potential (Cover & Thomas, 2006). To determine the most relevant features, we applied the filter method, ranking all features based on their MI scores and selecting the top 20 features. This approach ensures that only the most informative features are retained, improving the efficiency and accuracy of the subsequent modeling process (Peng et al., 2005).

In addition to MI, we incorporated the Fisher discriminant score as a supplementary measure. The Fisher score evaluates the class separability of each feature, quantifying how well each feature differentiates between the target classes (Bishop, 2006). Although we did not directly use the Fisher score to select features, it provided valuable insights into the discriminative power of the selected features. By assessing class separability, the Fisher score contributes to a better understanding of which features are most effective in distinguishing between the target classes, thereby enhancing the interpretability of the feature selection process.

To address the potential issue of categorical variables represented by one-hot encoding, we ensured that when a one-hot encoded feature was selected, all related binary features within the one-hot group were also included. This practice prevents the loss of information that could occur if only a single binary feature from the one-hot group were selected. By retaining the full set of related one-hot encoded features, we preserve the consistency of categorical variable representation, ensuring the feature selection process remains aligned with the original structure of the data.

This multi-faceted approach to feature selection—utilizing both MI scores and Fisher discriminant scores—ensures that the chosen features are not only predictive but also contribute meaningfully to the model's ability to differentiate between target classes, while also maintaining consistency in how categorical variables are represented.

3.2 Predictive Models

This study utilized predictive modeling techniques, specifically classification and regression, to analyze and understand socioeconomic trends across generations. Both approaches play a crucial role in uncovering patterns and relationships in the dataset, offering complementary insights into the financial disparities and socioeconomic dynamics that define each generational cohort.

3.2.1 Classification Models

Classification models were employed to predict generational categories (e.g., Baby Boomers, Generation X, Millennials, and Generation Z) based on a range of socioeconomic features. These features included income levels, employment status, homeownership rates, and education attainment. The classification task involved dividing individuals into predefined generational groups based on their birth years and exploring how socioeconomic characteristics differed between these groups.

The classification process began by encoding the generational labels as categorical variables, allowing models to identify patterns within the data. The primary goal of using classification models was to uncover how distinct socioeconomic factors contribute to generational differences. For instance, a classification model might reveal that income levels and housing affordability are significant predictors of generational membership, underscoring their importance in shaping the economic identity of a cohort.

Popular pre-existing classification algorithms—such as Random Forests, Explainable Boosting Machine (EBM), and Logistic Regression—were selected for their balance between predictive accuracy and interpretability. The Explainable Boosting Machine (EBM), in particular, exemplifies this balance—delivering strong classification accuracy and the potential for transparent, feature-level insights. Unlike black-box models such as neural networks, EBM is well-suited for applications where model transparency is valuable, which aligns with our broader research interest in understanding generational patterns. These models were evaluated using accuracy as the primary metric and confusion matrices to provide additional insight into classification performance. Accuracy served as an overall measure of the model's ability to correctly classify individuals into their respective generations, offering a straightforward and interpretable evaluation of performance. The confusion matrices were particularly useful for identifying misclassifications and understanding the model's ability to distinguish between closely related cohorts, such as Millennials and Generation Z, which often share overlapping socioeconomic characteristics. To strengthen the robustness of our evaluation, we employed both an 80/20 train-test split and 10-fold cross-validation. Together, these metrics provided a comprehensive assessment of model effectiveness.

3.2.2 Regression Models

Regression models were utilized to predict the birth year of individuals as a continuous variable, offering a complementary perspective to the classification task. While classification focuses on grouping individuals into discrete generational categories, regression allows for a more granular analysis by identifying how socioeconomic features vary across a continuous timeline. This approach is particularly useful for detecting subtle trends and shifts in financial characteristics over time.

The regression process involved selecting socioeconomic features, such as inflation-adjusted income, education levels, and housing costs, as predictors of birth year. By treating birth year as a continuous outcome, regression models provided insights into how specific economic and demographic factors evolve over time, reflecting broader structural changes in society. For example, a regression analysis might reveal that rising student debt is strongly correlated with more recent birth years, indicating the increasing financial burden on younger generations.

Linear regression was employed in this study to model the relationship between predictors and birth year. We selected linear regression to maintain interpretability and to support the explanatory focus of our analysis. While more complex models—such as ensemble regressors—can offer improved predictive performance, they were not prioritized in this study. Our primary objective was to highlight transparent relationships between socioeconomic factors and birth year, rather than to maximize predictive accuracy. The model was evaluated using several metrics to quantify its performance. The Mean Squared Error (MSE), and Average Absolute Error (AAE) measured the magnitude of deviations between predicted and actual birth years, while the Average Relative Error highlighted the accuracy of predictions relative to the true values. The R^2 Score assessed the proportion of variance in the birth year explained by the model, providing insight into its overall goodness-of-fit. These metrics collectively offered a comprehensive evaluation of the model's performance and its ability to generalize to unseen data.

3.3 Correlation Analysis

Correlation analysis was performed to examine the linear relationships between selected features and the target variable, birth year, using Pearson correlation coefficients. Positive coefficients indicate a direct relationship with more recent birth years, while negative coefficients suggest an inverse relationship with earlier birth years.

To assess statistical significance, p-values were calculated for each coefficient, with values below 0.05 considered significant, highlighting meaningful associations. This analysis aids in identifying relevant features for modeling and complementing feature selection techniques, providing deeper insights into feature importance and interpretability.

4 Economic Metrics

To comprehensively analyze generational economic disparities, this study employed a range of economic metrics to examine income distribution, labor market participation, rent burden, and housing affordability. These metrics provide quantitative insights into the financial realities faced by different generations, highlighting systemic economic shifts and their implications. By evaluating income inequality, employment trends, rent-to-income ratios, and homeownership challenges, this study explores how structural changes have shaped the financial prospects of Baby Boomers, Generation X, Millennials, and Generation Z. Each metric was chosen for its ability to capture specific aspects of generational financial outcomes, enabling a holistic assessment of economic progress and equity. This section details the methods used to calculate these metrics, their significance, and the insights they provide into evolving generational economic dynamics.

4.1 Income Distribution

Income distribution was analyzed using Lorenz Curves, Gini Coefficients (Gini, 1997), income percentile breakdowns, and P90/P10 ratios to capture the breadth of economic inequality across generations. The Lorenz Curve graphically represents cumulative income distribution, offering a visual depiction of inequality. The Gini Coefficient quantifies this inequality, with values ranging from 0 (perfect equality) to 1 (maximum inequality), enabling straightforward comparisons of income disparities between generations. Income percentile breakdowns focused on the bottom 50%, top 10%, and top 1% of earners to illustrate the concentration of wealth within specific groups over time. The P90/P10 ratio, calculated as the income of individuals at the 90th percentile divided by the income of individuals at the 10th percentile, provides a measure of income disparity within a generation. A higher P90/P10 ratio indicates greater gaps between high and low earners, offering additional insight into the extent of inequality beyond the averages. These metrics collectively reveal how income distribution has shifted and whether younger generations experience more pronounced income inequality compared to their predecessors.

4.2 Labor Market Participation

Labor market trends were evaluated by examining employment rates for individuals aged 18–27, as well as across all age groups for each generation. Employment rates, calculated as the proportion of the population engaged in paid work, offer insights into economic activity and engagement. These rates were compared across generations to assess whether younger cohorts, particularly Generation Z, are actively participating in the workforce at levels comparable to or exceeding those of older generations.

4.3 Rent Burden

Rent affordability was assessed using the rent-to-income ratio, calculated as the proportion of income spent on housing costs. This metric highlights the financial strain of meeting basic living expenses, particularly for younger generations such as Millennials and Generation Z. A higher rent-to-income ratio reflects greater economic pressure, inhibiting savings and wealth accumulation.

4.4 Housing Affordability and Homeownership

Housing affordability was evaluated by comparing average inflation-adjusted house values to average inflation-adjusted incomes over time. The maximum affordable mortgage, estimated as three times the household income, was used to assess whether generational earnings align with housing costs. Furthermore, the income increase required to bridge the gap between affordable mortgages and average home values was calculated, highlighting barriers to homeownership for Generation Z. Homeownership is widely recognized as a key pathway to wealth accumulation, particularly for low-income households, making this analysis critical to understanding the long-term financial implications of declining housing accessibility for younger generations.

5 Results

This section presents key findings from both statistical and machine-learning analyses, organized into five subsections. Feature Selection identifies variables most relevant to generational differences. Regression and Classification Results evaluate model performance in predicting birth years and generational cohorts. Correlation Results uncover patterns among socioeconomic variables, while Economic Metrics highlight income inequality and financial disparities across generations. Together, these results provide a comprehensive view of the factors shaping generational differences.

5.1 Feature Selection Results

To understand what factors are most strongly connected to a person's birth year, we used several methods: mutual information (MI), Fisher scores, and p-values. These tools help us measure how much each feature in the data relates to birth year, both in terms of general strength (MI), how well each feature helps to distinguish between people born in different years (Fisher score), and whether the results are statistically meaningful (p-values).

Our results, shown in Table 2, highlight the top 20 features. HOUSING_VALUE_ADJ (monthly contract rent or house value) was the strongest predictor, meaning that housing values often reflect generational differences. Similarly, INCTOT_ADJUSTED (total personal income adjusted for inflation) and NPBOSS50 (Nam-Powers-Boyd score) were also highly important. The NPBOSS50 score reflects the social and economic position of individuals based on their occupations, capturing generational shifts in job types and education levels. These findings suggest that both economic conditions and occupational status play a significant role in distinguishing between generations.

Table 2: Selected Features (Ranked by MI Score)

Feature name	MI Score	Fisher Score	P-Value
HOUSING_VALUE_adjusted	0.7386	9.4550	$< 10^{-5}$
INCTOT_adjusted	0.3768	22.8839	$< 10^{-5}$
NPBOSS50	0.3407	3.1002	$< 10^{-5}$
ERSCOR50	0.3237	3.0257	$< 10^{-5}$
FTOTINC_adjusted	0.1842	13.1118	$< 10^{-5}$
PLUMBING_21	0.1247	1289.6349	$< 10^{-5}$
PLUMBING_22	0.1247	4.3309	$< 10^{-5}$
PLUMBING_12	0.1247	9.4044	$< 10^{-5}$
PLUMBING_20	0.1247	1239.1866	$< 10^{-5}$
PLUMBING_14	0.1247	2.3760	$< 10^{-5}$
YNGCH	0.0990	127.2489	$< 10^{-5}$
OCCSCORE	0.0761	15.7726	$< 10^{-5}$
POVERTY	0.0622	29.5281	$< 10^{-5}$
KITCHEN_3	0.0418	2.2006	$< 10^{-5}$
KITCHEN_5	0.0418	410.0394	$< 10^{-5}$
KITCHEN_4	0.0418	253.6757	$< 10^{-5}$
MARST_6	0.0227	141.5336	$< 10^{-5}$
MARST_5	0.0227	10.4430	$< 10^{-5}$
MARST_4	0.0227	7.9465	$< 10^{-5}$
MARST_3	0.0227	2.3478	$< 10^{-5}$

MARST_2	0.0227	1.9914	$< 10^{-5}$

We also found that household characteristics, such as plumbing and kitchen facilities, were significant predictors. For example, PLUMBING_20 indicates households with complete plumbing, which reflects improved housing standards and modern living conditions often associated with younger generations. Meanwhile, PLUMBING_21 refers to plumbing facilities used only by the household, highlighting private access to utilities, another marker of better housing quality. In contrast, PLUMBING_22 represents plumbing shared with others, a condition more commonly found in older or lower-income housing arrangements, which were typical in past generations.

Kitchen features also proved relevant. KITCHEN_5 refers to households with an exclusive-use kitchen, typically found in more modern homes and reflecting higher living standards. The availability of private kitchen facilities provides further clues about generational shifts in housing quality.

Even features such as YNGCH (age of youngest own child in the household) were helpful for predicting birth year, as family structure and the presence of younger children often vary across generations, providing insights into household demographics and life stages.

By combining these different measures, we selected features that not only improve the accuracy of our predictions but also give us meaningful insights into how social, economic, and household factors vary across generations.

5.2 Correlation Analysis Results

The correlation analysis revealed key relationships between the features and the target variable, birth year. PLUMBING_20 emerged as the most positively correlated feature (Pearson correlation = 0.535291), indicating a strong direct relationship with birth year, followed by MARST_6 (Pearson correlation = 0.410951) and KITCHEN_4 (Pearson correlation = 0.227687). These features demonstrated their relevance in predicting more recent birth years, with high statistical significance ($p < 10^{-5}$).

On the negative side, PLUMBING_21 showed the strongest Pearson inverse correlation with birth year (Pearson correlation = -0.534888, $p < 10^{-5}$), followed by YNGCH (Pearson correlation = -0.307273, $p < 10^{-5}$). These negative correlations suggest that increases in these feature values are associated with older birth years. Other notable negative Pearson correlations include KITCHEN_5 (-0.255492) and POVERTY (-0.143595), reflecting systemic or socioeconomic patterns linked to earlier generations.

Most features demonstrated statistically significant Pearson correlations, with p-values below 10^{-5}, confirming the reliability of the observed relationships. However, a few features, such as MARST_2 (Pearson correlation = 0.006087, p = 0.249559) and KITCHEN_3 (Pearson correlation = -0.008190, p = 0.121364), exhibited weak and statistically insignificant correlations, indicating limited predictive utility.

The correlation analysis results, including Pearson coefficients and p-values for all features, are summarized in Table 3. These findings highlight the varying influence of features on birth year prediction, with both positive and negative correlations critical to the regression model's predictive capacity. These insights will guide feature selection and model optimization.

Table 3: Correlation and P-Values for Selected Features

FEATURE NAME	PEARSON CORRELATION	P-VALUE
PLUMBING_20	0.535291	$< 10^{-5}$
MARST_6	0.410951	$< 10^{-5}$
KITCHEN_4	0.227687	$< 10^{-5}$
MARST_2	0.006087	0.249559
KITCHEN_3	-0.008190	0.121364
ERSCOR50	-0.031943	$< 10^{-5}$
PLUMBING_14	-0.033424	$< 10^{-5}$
NPBOSS50	-0.034276	$< 10^{-5}$
MARST_3	-0.037826	$< 10^{-5}$
PLUMBING_22	-0.041601	$< 10^{-5}$
PLUMBING_12	-0.045062	$< 10^{-5}$
HOUSING_VALUE_ADJUSTED	-0.046916	$< 10^{-5}$
FTOTINC_ADJUSTED	-0.063988	$< 10^{-5}$

OCCSCORE	-0.069503	$< 10^{-5}$
MARST_5	-0.074974	$< 10^{-5}$
MARST_4	-0.094504	$< 10^{-5}$
INCTOT_ADJUSTED	-0.122010	$< 10^{-5}$
POVERTY	-0.143595	$< 10^{-5}$
KITCHEN_5	-0.255492	$< 10^{-5}$
YNGCH	-0.307273	$< 10^{-5}$
PLUMBING_21	-0.534888	$< 10^{-5}$

5.3 Regression Results

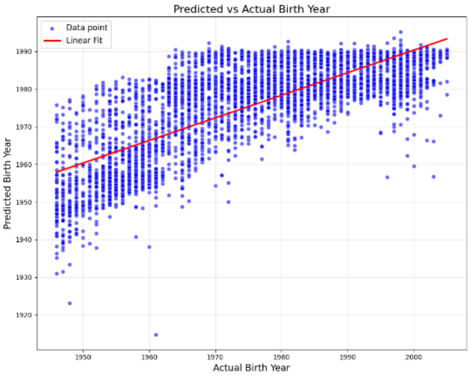


Figure 2: Result of regression illustrates the relationship between predicted and actual birth years, with the red line representing the linear fit. The data points cluster around the diagonal line, indicating that the model generally predicts birth years accurately.

The results of the regression analysis in Figure 2 reveal the model's capability to predict the birth year based on the given socioeconomic features. As shown in the scatter plot of Predicted vs. Actual Birth Year, the linear regression model demonstrates a clear positive trend, with the predicted values generally aligning with the actual birth years. The R^2 score of 0.6005 indicates that approximately 60.05% of the variance in birth year can be explained by the model, showcasing a moderate level of predictive power.

The Mean Squared Error (MSE) of 105.8127 and the Average Absolute Error (AAE) of 8.1880 highlight the magnitude of deviations between the predicted and actual birth years, with the model achieving an average prediction error of just over 8 years. Considering that the median generation length is 16 years, the AAE of 8 years suggests that, on average, the predicted birth year is likely to fall within the same generation as the actual birth year. This reinforces the model's utility in providing generation-level insights, even if exact birth year predictions are not perfect. The Average Relative Error of 0.41% further emphasizes the accuracy of the model in relative terms, suggesting minimal percentage-based deviations from actual values.

The Pearson correlation coefficient between predicted and actual birth years is 0.7749, indicating a strong positive linear relationship. The statistical significance of this correlation is confirmed by a p-value of ($p < 10^{-5}$), which underscores the reliability of the observed relationship. Despite these strengths, the scatter plot reveals some dispersion around the linear fit, particularly for older birth years, indicating areas where the model's predictions deviate more from the true values.

Overall, the results demonstrate that the regression model provides a reasonable and interpretable framework for predicting birth year, capturing meaningful relationships between socioeconomic predictors and the dependent variable. Moreover, the model's performance at the generation level, as indicated by the AAE relative to the median generation length, highlights its practical applicability for generational analysis while leaving room for further refinement to improve precision.

5.4 Classification Results

Table 4: Performance Comparison of Machine Learning Algorithms Based on Accuracy

ALGORITHM	ACCURACY (80/20 SPLIT)	ACCURACY (10-FOLD CV)
ZEROR (BASELINE)	29.72%	29.68%

RANDOM FOREST	70.35%	70.07%
GRADIENT BOOSTING	66.41%	65.61%
LOGISTIC REGRESSION	53.75%	53.51%
DECISION TREE	67.37%	67.72%
EBM	74.62%	74.78%

The ZeroR model achieved an accuracy of 29.72% by always predicting the majority class "Baby Boomers." In cross-validation, ZeroR achieved a similar accuracy of 29.68%, reflecting the baseline nature of the model. This model serves as a baseline for comparison, against which the performance of other models is evaluated. In contrast, all other models achieved significantly higher accuracy, indicating that they can detect meaningful signals in the data that differentiate between generational categories. This high accuracy, observed consistently across both the percentage split and cross-validation, highlights that the dataset contains distinct patterns or features that are characteristic of each generation, validating the relevance of the features used in the models. The classification results for the evaluated models are summarized in Table 4, and the corresponding normalized confusion matrices in Figures 3 provide detailed insights into their performance.

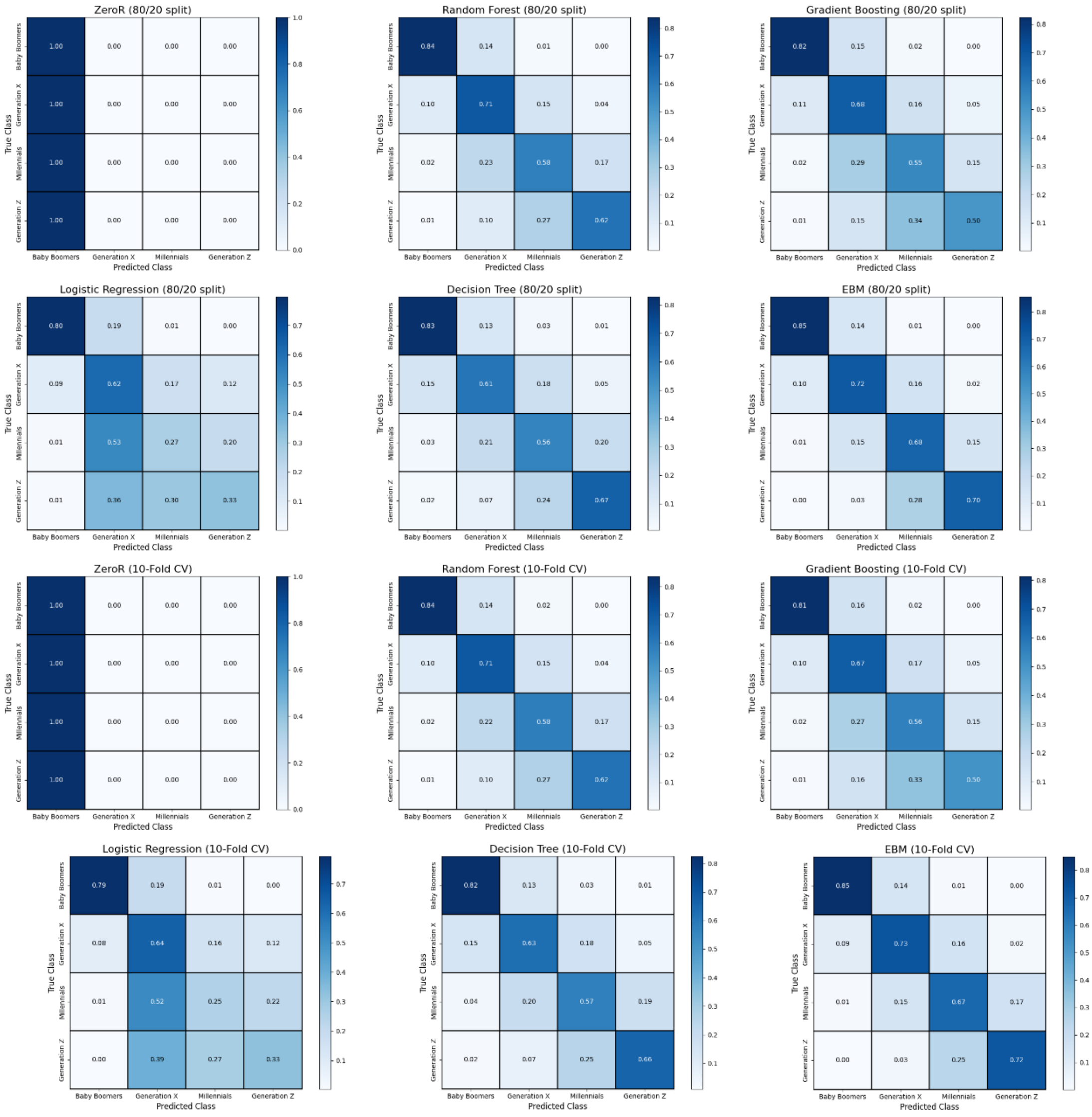
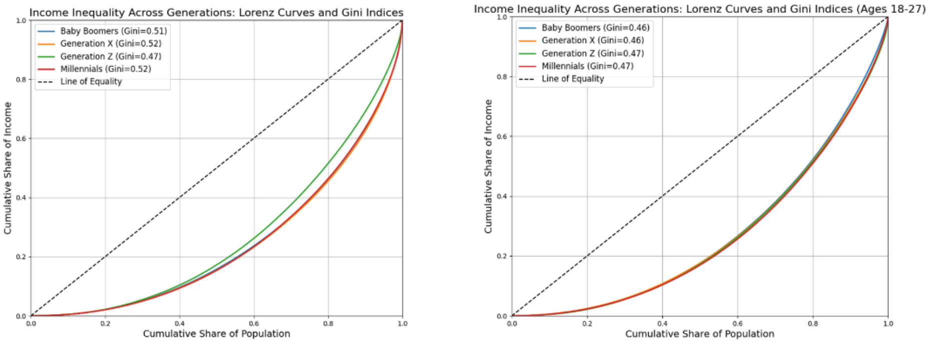


Figure 3: The normalized confusion matrices for the evaluated models, with the top two rows showing results from the 80/20 train-test split and the bottom two rows from 10-fold cross-validation, provide detailed insights into their performance as summarized in Table 4.

Explainable Boosting Machine (EBM) achieved the highest accuracy among all models, with 74.62% in the percentage split and a slightly higher 74.78% in cross-validation, demonstrating its strong capability to handle the dataset while maintaining interpretability. Random Forest followed with accuracies of 70.35% in the percentage split and 70.07% in cross-validation, reflecting solid performance with minimal misclassifications. The Decision Tree model achieved 67.37% in the percentage split and 67.72% in cross-validation, providing interpretable results alongside competitive accuracy. Gradient Boosting, while slightly lower, attained 66.41% in the percentage split and 65.61% in cross-validation, maintaining reasonable predictive power but trailing behind the tree-based models. Logistic Regression delivered lower performance, with accuracies of 53.75% in the percentage split and 53.51% in cross-validation, indicating its limited ability to capture complex patterns within the data.

As shown in the normalized confusion matrices in Figure 3, models such as EBM and Random Forest offer balanced predictions across generational categories, effectively minimizing misclassifications, particularly for "Generation Z" and "Baby Boomers." These results highlight the models' ability to leverage meaningful features that distinguish between generations, reinforcing the validity of the dataset and its potential for deeper generational insights and applications.

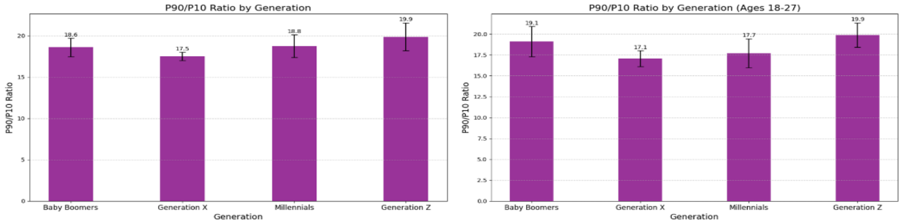
5.5 Economic Metrics Results


Income Inequality Across Generations: Lorenz Curves and Gini Indices

Income Inequality Across Generations: Lorenz Curves and Gini Indices (Ages 18-27)

Figure 4: Lorenz curves and Gini indices illustrating income inequality across generations for all ages (left) and ages 18–27 (right) in the United States.

The results reveal significant trends and disparities across generations in income distribution, housing affordability, and economic opportunities. Generation Z experiences lower levels of overall income inequality compared with individuals of all ages from different generations, as reflected in their Gini index (0.47 versus 0.51–0.52 for other generations), as shown in Figure 4. However, for individuals aged 18–27, the Gini index for Generation Z (0.47) is comparable to Millennials and slightly higher than Baby Boomers and Generation X (both at 0.46), suggesting a narrower gap in income inequality among younger cohorts. Similarly, the P90/P10 ratio for Generation Z is the highest (19.9) across all ages, indicating significant income disparities between the highest and lowest earners within the generation, as depicted in Figure 5. For ages 18–27, the P90/P10 ratio for Generation Z remains the highest, further highlighting these disparities, even among younger individuals.


P90/P10 Ratio by Generation

P90/P10 Ratio by Generation (Ages 18-27)

Figure 5: Bar chart illustrating the P90/P10 ratio, a measure of income inequality, across different generations (All ages vs ages 18-27) in the United States.

This dynamic is further evidenced by the distribution of income shares. Generation Z demonstrates a more equitable income distribution at the lower end compared to other generations across all ages, with 66.7% of total income concentrated within the bottom 50%. However, for individuals aged 18–27, Baby Boomers slightly surpass Generation Z with 67.3% of income held by the bottom 50%, compared to Generation Z's 66.7%, followed by Millennials (66.5%) and Generation X (66.3%). Despite this equitable distribution at the lower end, Generation Z holds the highest income share at the top 1% (7.8%) among individuals aged 18–27, surpassing Millennials

(6.6%), Generation X (7.0%), and Baby Boomers (5.6%). This pattern underscores a widening gap between the highest earners and the rest of Generation Z, particularly when focusing on younger cohorts.

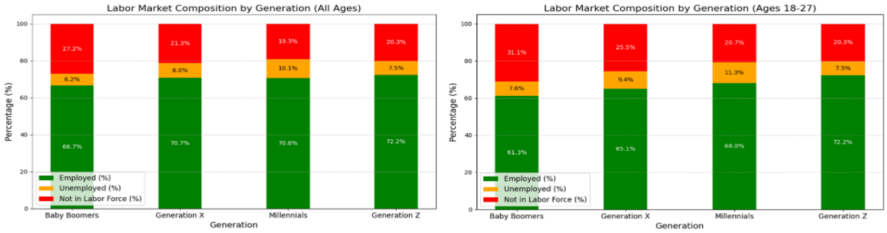


Figure 6: Stacked bar charts illustrating the labor market composition by generation for the entire population (left) and ages 18–27 (right) in the United States.

Moreover, as generations progress, a clear trend emerges: the income share held by the top 10% (excluding the top 1%) shrinks, while the top 1% consolidates an increasingly larger share of total income. For example, the top 10% share decreases from 27.1% for Baby Boomers to 25.6% for Generation Z, while the top 1% share simultaneously increases. This suggests that individuals from lower income groups tend to move into higher income brackets over time, leading to a redistribution of income towards the upper tiers. While this upward mobility may appear positive, it ultimately widens income inequality by concentrating more wealth within the top 1%, creating structural barriers to equitable wealth distribution.

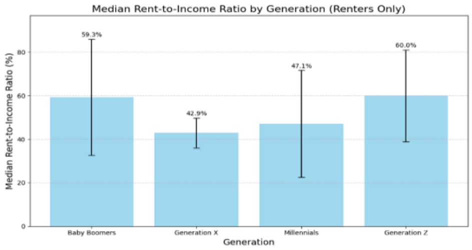


Figure 7: Bar chart illustrating the median rent-to-income ratio for different generations of renters in the United States.

These findings highlight how the concentration of wealth within the top 1% exacerbates economic inequality, disproportionately affecting younger cohorts like Generation Z. This trend compounds their financial challenges and limits their ability to achieve long-term economic stability, underscoring the systemic nature of these disparities.

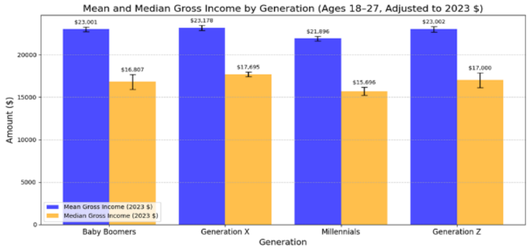


Figure 8: Bar chart showing the mean and median gross income (in 2023 dollars) by generation for ages 18–27 in the United States.

The narrative that members of Generation Z are 'lazy' or 'don't want to work' is contradicted by their higher employment rates compared to those of other generations, as shown in Figure 6, both overall and within the 18-27 age range. At ages 18-27, Generation Z leads with an employment rate of 72.2%, surpassing Millennials (68.0%), Generation X (65.1%), and Baby Boomers (61.3%) at the same life stage, highlighting their active participation in the workforce. Despite their workforce participation, Generation Z continues to face financial challenges and bears the greatest median rent-to-income burden (60.0%), surpassing Millennials (47.1%), Generation X (42.9%), and Baby Boomers (59.3%), as demonstrated in Figure 7, reflecting worsening economic pressures for younger generations. This is particularly concerning given that Generation Z earns a comparable mean income to individuals in the same age range across older generations, as shown in Figure 8. However, despite earning similar incomes, Generation Z suffers from much higher rent-to-income ratios and housing affordability challenges, which significantly impact on their financial stability and ability to build wealth.

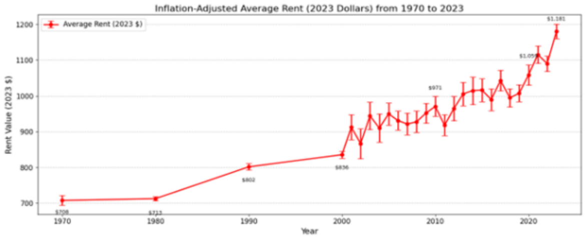


Figure 9: Line chart illustrating inflation-adjusted average rent (in 2023 dollars) from 1970 to 2023 across the United States.

The broader systemic issue of rising housing costs is evident in inflation-adjusted average rent, which has increased steadily from $708 in 1970 to $1,181 in 2023—a 66% rise in real terms, as shown in Figure 9. This sharp rise, particularly after 2000, has compounded the financial strain on younger generations, making it increasingly difficult for them to achieve economic security. These findings underscore the urgent need for policy interventions to address rising rents and ensure equitable access to affordable housing.

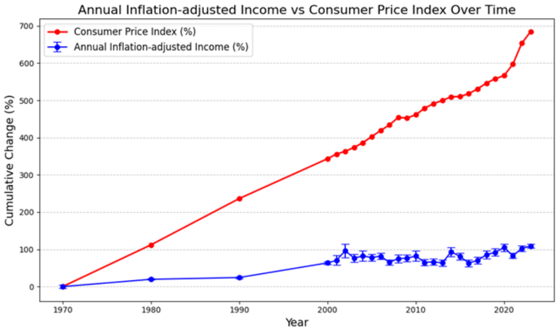


Figure 10: Annual inflation-adjusted income vs. Consumer Price Index (CPI) cumulative change over time (1970–2023). The CPI data were retrieved using the Federal Reserve Economic Data API (U.S. Bureau of Labor Statistics, 2025).

To further understand the economic pressures faced by different generations, it is crucial to examine the drastic disparity between inflation and income growth over time. As shown in Figure 10, the Consumer Price Index (CPI), obtained from the Federal Reserve Economic Data (U.S. Bureau of Labor Statistics, 2025), has surged by 700% over the past five decades, indicating a sharp rise in the cost of living, while inflation-adjusted income has increased by only 100%. This imbalance highlights the shrinking ability of wages to keep pace with rising costs, which disproportionately affects younger generations. To put this into perspective, in 1970, the average cost of a week's worth of groceries for a family was approximately $30 (inflation-adjusted). A worker earning $15 per hour in 2023-equivalent dollars could cover this cost with just 2 hours of work. Today, that same basket of groceries would cost around $210, but a worker’s hourly income would have risen to only $30—requiring 7 hours of work. This stark increase highlights the disproportionate growth of living expenses relative to wages, leaving less financial flexibility for savings, investments, or other essential needs. While this issue disproportionately affects younger generations, particularly Generation Z, who face additional economic barriers such as stagnant wages and skyrocketing housing costs, the trend impacts all generations. If this trajectory continues unchecked, Generation Alpha is likely to face even greater financial challenges than Generation Z, further exacerbating systemic financial insecurity and severely limiting opportunities for upward mobility.

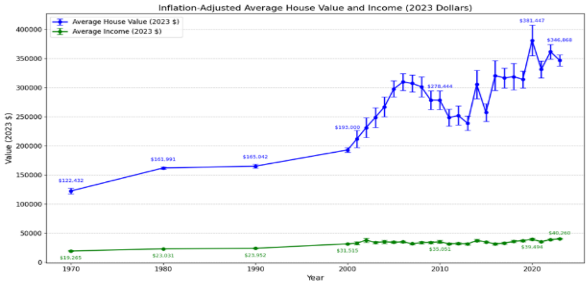


Figure 11: Line chart illustrating inflation-adjusted average house value and income (in 2023 dollars) from 1970 to 2023 in the United States.

Moreover, as highlighted in Figure 11, the disparity between house values and household incomes has widened significantly over time, with house costs increasing at a much faster rate than income. Since 1970, household income has increased by approximately 100%, while the average house value has surged by nearly 180%, exacerbating the affordability gap. In 2023, the average house value was approximately $350,000, while the average household income for two earners was $80,000, making the maximum affordable mortgage $240,000—$110,000 short of the average house price. This stark imbalance underscores the escalating affordability crisis, requiring households to increase their income by approximately 150% to afford a home, an unfeasible goal to achieve on a large scale without systemic changes. This affordability crisis disproportionately impacts Generation Z and severely limits their ability to achieve homeownership. As homeownership is one of the most effective ways for households—especially low-income households—to build wealth over time (Harvard Joint Center for Housing Studies, 2006), the inability of Generation Z to access homeownership not only affects their current financial stability but also undermines their prospects for upward mobility and long-term wealth accumulation, perpetuating systemic inequality across generations.

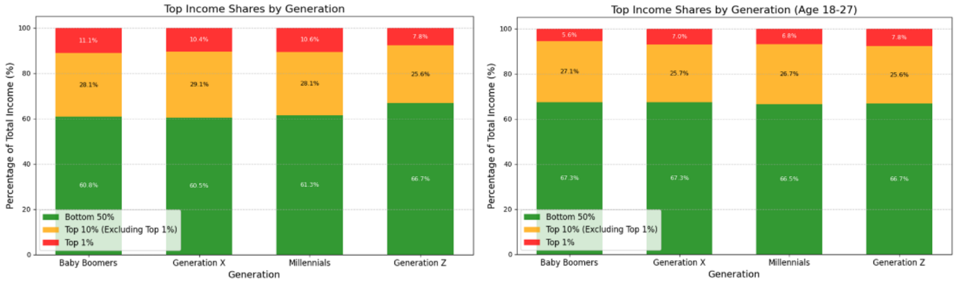


Figure 12: Stacked bar charts illustrating the distribution of income among different generations (all population vs ages 18-27) in the United States.

While these findings highlight the immediate challenges facing Generation Z, an interesting observation emerges when examining how income distribution evolves over time. As generations age, the gap between the top earners and bottom earners widens significantly. For individuals aged 18–27, income distribution is relatively balanced, with the bottom 50% holding 67.3% of total income for Baby Boomers and Generation X. However, this share declines to 60.8% and 60.5% for these same generations across all ages, reflecting a growing disparity. The top 1% share also increases notably, from 5.6% for Baby Boomers (aged 18–27) to 11.1% for all ages (Figure 12).

These trends are further illustrated by key inequality metrics. For example, the P90/P10 ratio for Millennials increases from 17.7 at ages 18–27 to 18.8 across all ages, illustrating how income inequality grows as the cohort ages (Figure 5). Similarly, Gini indices increase from 0.46–0.47 in younger groups to 0.51–0.52 for older ones, reflecting a clear progression of inequality (Figure 4). Together, these measures highlight how wealth and income become increasingly concentrated among top earners as generations progress.

This observation makes sense given that, at younger ages, individuals tend to start from relatively similar socioeconomic positions as they enter adulthood, often with limited income and wealth. Over time, as people make different life choices, pursue varied career paths, and encounter diverse opportunities or challenges, their financial trajectories naturally diverge. Some may benefit from lucrative careers, advantageous networks, or entrepreneurial success, while others might face barriers that limit their upward mobility. These individual trajectories are further compounded by systemic factors—such as unequal access to resources, education, and opportunities—which exacerbate disparities over time. Collectively, these dynamics likely contribute to the persistent and growing inequality observed across generations.

6 Conclusion

This study provides key insights into the socioeconomic disparities faced by different generations, particularly Generation Z, who experience significant financial challenges. Classification models highlighted distinct patterns in socioeconomic features that differentiate generations, with the Explainable Boosting Machine achieving the highest accuracy. Regression analysis further demonstrated the model's ability to capture temporal trends, with predicted birth years aligning closely with actual values, particularly at the generational level.

Economic metrics revealed that Generation Z faces the highest median rent-to-income burden (60.0%) and the greatest barriers to homeownership, with a $110,000 gap between affordable mortgages and average house prices. Despite higher workforce participation (72.2%) compared to previous generations at similar life stages, these efforts are undermined by disproportionate income disparities, as evidenced by their high P90/P10 ratio (19.9). While their income distribution is more equitable at the lower end, systemic issues like unaffordable housing and stagnant wage growth hinder financial stability and wealth accumulation.

The findings highlight the value of machine learning in uncovering generational trends and systemic inequalities. While this study focuses on quantitative analysis, future research would benefit from integrating qualitative insights, such as psychological, behavioral, and cultural perspectives, to enrich the contextual understanding of the data and provide a more holistic view of generational experiences. Future work should also investigate targeted policy interventions to address the economic challenges facing Generation Z and promote equity across generations.

References

Aljishi, A., Marjani, M., Latifi, A., & Shamir, L. (2025). GenZ-Economic-Challenges [Data set and source code]. GitHub. https://github.com/MatinMarjani/GenZ-Economic-Challenges

Autor, D. H., Katz, L. F., & Krueger, A. B. (1998). Computing inequality: Have computers changed the labor market? *The Quarterly Journal of Economics*, 113(4), 1169–1213.

Bianchi, S. M. (2000). Maternal employment and time with children: Dramatic change or surprising continuity? Demography, 37(4), 401–414. https://doi.org/10.1353/dem.2000.0001

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Bound, J., & Turner, S. (2007). Cohort crowding: How resources affect collegiate attainment. *Journal of Public Economics*, 91(5–6), 877–899.

Charles, K. K., & Hurst, E. (2003). The correlation of wealth across generations. *Journal of Political Economy*, 111(6), 1155–1182.

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., & Narang, J. (2017). The fading American dream: Trends in absolute income mobility since 1940. Science, 356(6336), 398–406.

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). Wiley.

Elder, G. H. (2018). Children of the Great Depression: Social change in life experience (25th anniversary ed.). Routledge.

Fan, C., Xu, J., Natarajan, B. Y., & Mostafavi, A. (2023). Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality. *Computer-Aided Civil and Infrastructure Engineering*, 38(14), 2013–2029.

Gallipoli, G., Low, H., & Mitra, A. (2020). Consumption and income inequality across generations. Centre for Economic Policy Research.

Green, R. K., & Lee, H. (2016). Age, demographics, and the demand for housing, revisited. Regional Science and Urban Economics, 61, 86–98.

Gregory, V. (2023). Generational gaps in income and homeownership. Economic Synopses, (15). https://doi.org/10.20955/es.2023.15

Gini, C. (1997). Concentration and dependency ratios. Rivista di Politica Economica, 87, 769–792.

Harvard Joint Center for Housing Studies. (2006). The state of the nation's housing 2006. Harvard Joint Center for Housing Studies.

Ipsos. (2023). We need to talk about generations: Understanding generations. https://www.ipsos.com/en/we-need-talk-about-generations-understanding-generations

Lev, T. A. (2021). Generation Z: Characteristics and challenges to entering the world of work. *Cross-Cultural Management Journal*, 23(1), 107–115. https://doi.org/10.22381/CCMJ2320216

Mishel, L., & Bivens, J. (2021). Identifying the policy levers generating wage suppression and wage inequality. Economic Policy Institute, 13.

Palfrey, J., & Gasser, U. (2011). Born digital: Understanding the first generation of digital natives. Basic Books. https://books.google.com/books?id=wWTI-DbeA7gC

Patterson, J. T. (1996). Grand Expectations: The United States, 1945-1974. Oxford University Press.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

Ruggles, S., Flood, S., Sobek, M., Backman, D., Chen, A., Cooper, G., Richards, S., Rogers, R., & Schouweiler, M. (2024). PUMS USA: Version 15.0 [Data set]. IPUMS. https://usa.ipums.org/usa/

Twenge, J. M. (2023). Generations: the real differences between Gen Z, Millennials, Gen X, Boomers, and Silents—and what they mean for America's future. Simon & Schuster.

U.S. Bureau of Labor Statistics. (2025). Consumer Price Index for all urban consumers: All items in U.S. city average (CPIAUCSL) [Data set]. Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/CPIAUCSL

Western, B., & Rosenfeld, J. (2011). Unions, norms, and the rise in US wage inequality. American Sociological Review, 76(4), 513-537.