Journal of High School Science

The factors influencing financial success: A Machine Learning approach

Zhou M[1], Ramezani R[2]

Submitted: May 26, 2024, Revised: version 1, June 26, 2024
Accepted: July 5, 2024

Abstract

This paper explored the various socioeconomic factors that contribute to individual financial success using machine learning algorithms and approaches. Financial success, a critical aspect of an individual's well-being, is a complex concept influenced by various factors. The objective of this study was to understand the determinants of financial success. The National Longitudinal Survey of Youth 1997, by the Bureau of Labor Statistics was examined, consisting of longitudinal data of a sample of 8984 individuals. The dataset comprised income variables and a large set of socioeconomic variables. The findings highlighted the significant influence of the highest education degree obtained, occupation and gender as the top three determinants of individual income among the socioeconomic factors examined. Yearly working hours, age and work tenure followed as the three secondary influencing factors. All the other factors including parental household income, industry, parents' highest education grade and others were identified as tertiary factors. These insights will allow researchers to better understand the complex nature of financial success, and advance broader societal well-being by providing insights for policymakers during decision-making processes.

Keywords

Machine learning, Financial success, Socioeconomic factors, Longitudinal data, Individual income, NLSY1997, Income analysis, Random Forest algorithm, SHapley Additive exPlanations, Receptor Operator Curve

[1]Corresponding author, Michael Zhou, Interlake High School, Bellevue, WA, 98008, USA. michaelzhou2025@gmail.com

[2]Ramin Ramezani, Adjunct Associate Professor, Computer Science, Samueli School of Engineering, University of California at Los Angeles, California, 90095, USA. raminr@ucla.edu

Introduction

In the contemporary world, financial success and achieving financial success has gradually gained more weight in the minds of all individuals. As a result of the capitalism pervading most of the world's economies, the notion of financial success, which once was a personal ambition, has become a worldly phenomenon with many far-reaching implications. Recognizing and navigating the intricacies of financial success is one of the most paramount endeavors that researchers must take into account in order to grasp the forces that shapes aspirations, decision-making, and the broader socio-economic fabric of society.

To better understand the complexities associated with financial success and income analysis, we will start off by defining some terms. "Financial success" in the context of this study encompasses the monetary gains of individuals by the same individuals in a one year time-frame. "Income analysis" refers to the systematic evaluation of patterns between those who earn a higher income and those who earn a lower income. Analyzing these key concepts will help dissect the complex nature of financial success and provide nuanced insights into the factors shaping income dynamics.

Objectives

In this big data era, advanced data science has been increasingly used in many industries, including economic research. There is a growing interest in utilizing machine learning methods in economic research, surpassing traditional statistical models. References (2-5) are some examples using machine learning methods in economic research. One prominent effort in this field of study is to predict individual income based on a range of socioeconomic factors. These factors typically include education, employment, demographic information, socioeconomic status, and other relevant variables. References (6-13) provide illustrative examples in this research area.

In this study, we used the data from National Longitudinal Survey of Youth 1997 (hereinafter abbreviated as NLSY97) to conduct a multi-class classification task to predict individual income levels with machine learning algorithms. The primary purpose was to identify dominant socioeconomic factors that influence an individual's income. Various socioeconomic factors, including demographic, educational, and occupational factors, were examined for their influence on individual income levels. This list is not exhaustive but serves as a starting point for further exploration. Each factor targeted a different aspect of an individual's life as it relates to financial success. When combined, these factors provided a more nuanced view than any single factor alone. Through the analysis of these factors, this study sought to provide a holistic view of their influence on individual economic well-being and financial success, both independently and collectively. Furthermore, the longitudinal data in the survey data was explored to enhance prediction accuracy. In contrast to existing research on individual income prediction, which primarily focuses on exploring population characteristics in the dataset, this study also explored the trajectories of individuals over time in the longitudinal data to improve the accuracy of prediction tasks.

Literature review

Yeh et al. (2) used a satellite-based deep learning approach to predict economic well-being in Africa with satisfactory results. Sheetal et al. (3) demonstrated that machine learning methods can be used to predict people's attitudes toward income inequality. They found that that a greater emphasis on hard work was associated with a greater justification of income inequality. Kannan et al. (4) showed that the machine learning approach can provide a robust model to predict a household's financial vigilance. Aiken et al. (5) used traditional survey data to train machine-learning algorithms to recognize patterns of poverty in mobile phone data and reduced errors of exclusion by 21%, highlighting the potential for targeting humanitarian assistance. Previous research also explored various socioeconomic factors related to financial success using machine learning methodologies. Kamdjou et al. (6) estimated the income returns to education using a machine learning approach and concluded that each additional year of education, on average, provided a rate of return of 10.4%. Gomez-Cravioto et al. (7) employed supervised machine learning predictive analytics to predict alumni income. Their study centered on survey data collected in 2018 by Tecnologico de Monterrey and yielded statistically significant results, with particular accuracy in predicting the alum's first income after graduation. Lazar (8) employed principal component analysis and support vector machine methods to generate and evaluate income prediction data based on the Current Population Survey provided by the U.S. Census Bureau and achieved accuracy values as high as 84% for binary classification tasks. Matz et al. (9) applied an established machine learning method to predict individual income from the digital footprints people left on Facebook with an accuracy up to 0.49. Khongchai et al. (10) proposed a salary prediction system based on a decision tree technique with seven features to increase students' motivation for studying. It yielded an overall 41.39% accuracy and demonstrated positive satisfaction with 50 student samples. Rana et al. (11) investigated income prediction performance of eleven machine learning models on the UCI Adult Dataset and showed that demographic characteristics helped some models predict income levels with greater accuracy. Zaid et al. (12) developed a decision tree algorithm with an accuracy of 84.3 on predicting income class for an "adult income dataset" acquired on the open source Kaggle platform. Chakrabarty et al. (13) developed a Gradient Boosting Classifier model on the UCI Adult Dataset to predict individuals' yearly income in the US and achieved a prediction accuracy of 88.2%.

Methods

In this section, one of the predictive tasks was to classify provided input into predefined income categories. Compared to the covariate analysis conducted in the exploratory analysis, a multivariate analysis was performed to fully explore the collective influence of all independent variables on the dependent variable. Weka, an open-source machine learning software toolkit developed by the University of Waikato in New Zealand, was utilized. Weka offers a diverse array of machine learning algorithms, encompassing classification and regression, among others.

Dataset

The dataset used in this study originated from the National Longitudinal Surveys with many cohorts tracking the life trajectories of several groups of American men and women. The NLS is a long-withstanding resource used by many economists and sociologists and provided credible and reliable data for this research.

NLSY97 Data

The NLSY97 cohort was purposely chosen for this investigation. "The NLSY97 consists of a nationally representative sample of 8,984 men and women born during the years 1980 through 1984 and living in the United States at the time of the initial survey in 1997." "The NLSY97 collects extensive information on respondents' labor market behavior and educational experiences. The survey also includes data on the youths' family and community backgrounds to help researchers assess the impact of schooling and other environmental factors on these labor market entrants." Following the initial survey, the cohort included follow-up interviews and continuations.

NLSY97 has a total of 87227 variables as of this study date, from which a small subset of variables pertaining to education, employment, demographic information, socioeconomic status were purposely chosen for this research, some of which with repeated measurements or data across selected years. Table 1 lists the variables considered, where the income variable is the dependent variable to be predicted and all the other variables are independent variables used to predict the dependent variable income. In this paper, independent variables and features are used interchangeably.

To utilize the temporal information within the longitudinal data, we included repeated measures of individuals from several different time points, spanning the years 2021, 2019, 2017, and 2015, consisting of a total of $8984 \times 4 = 35936$ entries. The Appendix provides details of the NLSY97 variables used in this study, including their NLSY97 variable title, definition, possible values and how they are encoded in this study when applicable.

Data preprocessing

Data preprocessing played an important role in preparing data for machine learning prediction tasks. The objective was to ensure that the input data was structured appropriately for machine learning. In this study, data preprocessing served several purposes. First, it extracted longitudinal information from the original dataset, which was essential for understanding trends and patterns over time for longitudinal analysis. This process ensured that the temporal aspect of the data or the trajectories of individuals over time was preserved and could be effectively utilized in subsequent prediction tasks. Second, it filtered out invalid entries from the dataset, thus enhancing the quality and reliability of the data used in prediction tasks. Lastly, it encoded the original data types into a more suitable format when necessary to improve prediction accuracy. This included, for example, converting categorical variables into nominal variables with one-hot encoding, scaling numerical variables to a common range, and properly labeling missing values.

Table 1. Variables chosen from NLSY97 for this study (all of numeric type)

Variables	NLSY97 variable names	Measures in years
sex	KEY!SEX, RS SEX (SYMBOL)	1997
race	KEY!RACE, RACE OF R (SYMBOL)	1997
degree	HIGHEST DEGREE RECEIVED	XRND (Cross Round)
biological father's highest grade	BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED	1997
biological mothers highest grade	BIOLOGICAL MOTHERS HIGHEST GRADE COMPLETED	1997
residential fathers highest grade	RESIDENTIAL FATHERS HIGHEST GRADE COMPLETED	1997
residential mothers highest grade	RESIDENTIAL MOTHERS HIGHEST GRADE COMPLETED	1997
parental household income	GROSS HH INCOME IN PAST YEAR	2003
highest grade	RS HIGHEST GRADE COMPLETED	2021
age	RS AGE IN MONTHS AS OF INTERVIEW DATE	2015, 2017, 2019, 2021
industry	YEMP, TYPE OF BUS OR INDUSTRY CODE (2002 CENSUS 4-DIGIT) 01 (ROS ITEM)	2015, 2017, 2019, 2021
occupation	YEMP, OCCUPATION/JOB CODE (2002 CENSUS 4-DIGIT) 01 (ROS ITEM)	2015, 2017, 2019, 2021
total work weeks	# WEEKS EVER WORKED AT JOB 01, 02, 03, 04, 05, 06, 07, 08	2015, 2017, 2019, 2021
yearly work hours	YYYY EMPLOYMENT: TOTAL HOURS WORKED WEEK WW (YYYY is year and WW ranges from 01 to 52)	2015, 2017, 2019, 2021
income (yearly)	TOTAL INCOME FROM WAGES AND SALARY IN PAST YEAR	2015, 2017, 2019, 2021

Extract and represent longitudinal information

Longitudinal data and analysis have been extensively explored in classic statistical models (14) (15). In this study, we explored the utilization of longitudinal data information in machine learning model settings.

To utilize the trajectory information of individuals over time in the longitudinal data, repeated measures of the same individual across different years were compiled into separate entries. In other words, the longitudinal information in the dataset was extracted by encoding one entry for each individual with measures from N years into N entries for this individual, each for one year's measure of the individual, as shown in Table 2 (original entry with repeated measures over years) and in Table 3 (encoded entries with one for each year). For new data in Table 3, each entry represents a unique observation, i.e. one individual at a specific time point. With this

data transformation, the time-varying information was captured in the by-individual by-time data structure, making it robust against missing values and irregularly spaced measures.

Table 2. Original dataset entry with repeated measures over the years

Variables	Entry 1
individual id	2
sex	1
race	5
degree	2
biological father's highest grade	17
biological mothers highest grade	15
residential fathers highest grade	14
residential mothers highest grade	15
parental household income	75000
highest grade	14
age 2019	448
industry 2019	7680
occupation 2019	3820
total work weeks 2019	646
yearly work hours 2019	2700
income (yearly) 2019	128400
age 2021	472
industry 2021	7680
occupation 2021	3820
total work weeks 2021	749
yearly work hours 2021	3220
income (yearly) 2021	115000

Filtering out invalid entries

Out of 35936 entries, 15245 of them did not have valid income values (with negative values from NLSY97 data). These were hence removed from the study, leaving a total of 20691 entries. In order not to lose too much data, the entries with missing values for independent variables (or non-income variables) were retained. All missing values for nominal and numeric attributes in a dataset were replaced with the modes and means from the training data.

Data type conversion and feature engineering

Table 1 lists the variables of NLSY97 raw data used in this study, all being of numerical type

with negative values -1 through -5 representing various not-available cases. Entries with invalid income values were filtered out in the previous step, while all negative values for other independent variables were marked as missing values for machine learning prediction tasks. The following data type conversion and feature engineering were performed on the raw data, to improve the performance of machine learning models. The dependent variable income was categorized into three nominal values throughout this study as shown in Table 4, thereby converting the prediction tasks in this study into multi-class classification tasks.

Table 3. Encoded entries with one for each year

Variables	Entry 1	Entry 2
individual id	2	2
sex	1	1
race	5	5
degree	2	2
biological father's highest grade	17	17
biological mothers highest grade	15	15
residential fathers highest grade	14	14
residential mothers highest grade	15	15
parental household income	75000	75000
highest grade	14	14
age	448	472
industry	7680	7680
occupation	3820	3820
total work weeks	646	749
yearly work hours	2700	3220
income (yearly)	128400	115000

Table 4. Data type conversion and feature engineering

Variables	Raw data type and values	New data type and values
sex	numeric: 1, 2	nominal: 1, 2
race	numeric: 1, 2, 3, 4, 5	nominal: 1, 2, 3, 4, 5
degree	numeric: 0, 1, 2, 3, 4, 5, 6, 7	nominal: 0, 1, 2, 3, 4, 5, 6, 7
biological father's highest grade	numeric: 1-20, 95 (ungraded)	numeric: 0-20, 95
biological mothers highest grade	numeric: 1-20, 95 (ungraded)	numeric: 0-20, 95
residential fathers highest grade	numeric: 1-20, 95 (ungraded)	numeric: 0-20, 95
residential mothers highest grade	numeric: 1-20, 95 (ungraded)	numeric: 0-20, 95
parental household income	numeric	numeric
highest grade	numeric: 1-20, 95 (ungraded)	numeric: 0-20, 95
age	numeric	numeric
industry	numeric: 170-9990	nominal: 1-18
occupation	numeric: 10-9990	nominal: 1-33
total work weeks	numeric	numeric
yearly work hours	numeric	numeric
income (yearly)	numeric: 0-380288	nominal: - 1 (Less than 50K), - 2 (between 50K and 100K), - 3 (More than 100K)

Exploratory analysis

After the data processing on the raw NLSY97 data, the new dataset contained 20691 entries and 15 numeric and nominal variables, 9 being numeric and 6 being nominal. The objective of the exploratory data analysis was to identify potential patterns and variables that could be useful in prediction tasks. For this purpose, a covariate analysis was conducted to examine the correlation between each independent variable and the dependent variable (income). The covariate analysis was performed on the survey data of 20691 entries. Subsequently, multivariate analysis to examine the collective influence of independent variables on the dependent variable was conducted with machine learning algorithms, as described below.

We used the Spearman correlation coefficients between all pairs of variables to identify independent variables that showed a high correlation with the dependent variable. Then, we narrowed down the set of independent variables by removing repeated highly correlated independent variables. We calculated the Spearman correlation coefficient for a pair of variables with Excel, first ranking the entries with the RANK.AVG function, then calculating the correlation coefficient with the CORREL function on the ranking numbers.

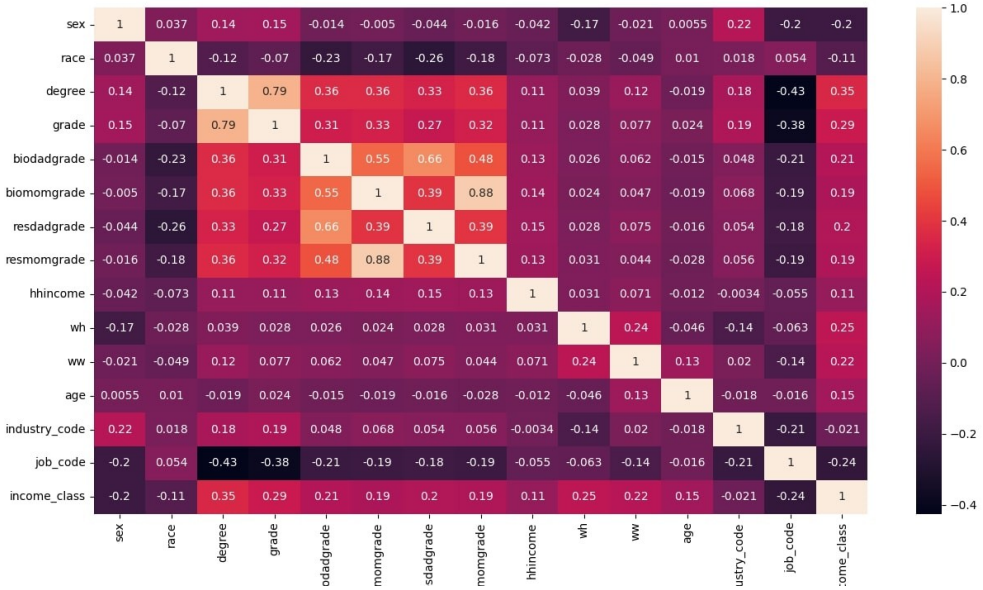

Figure 1. Heatmap of Spearman correlation coefficients for all variable pairs listed in Table 1

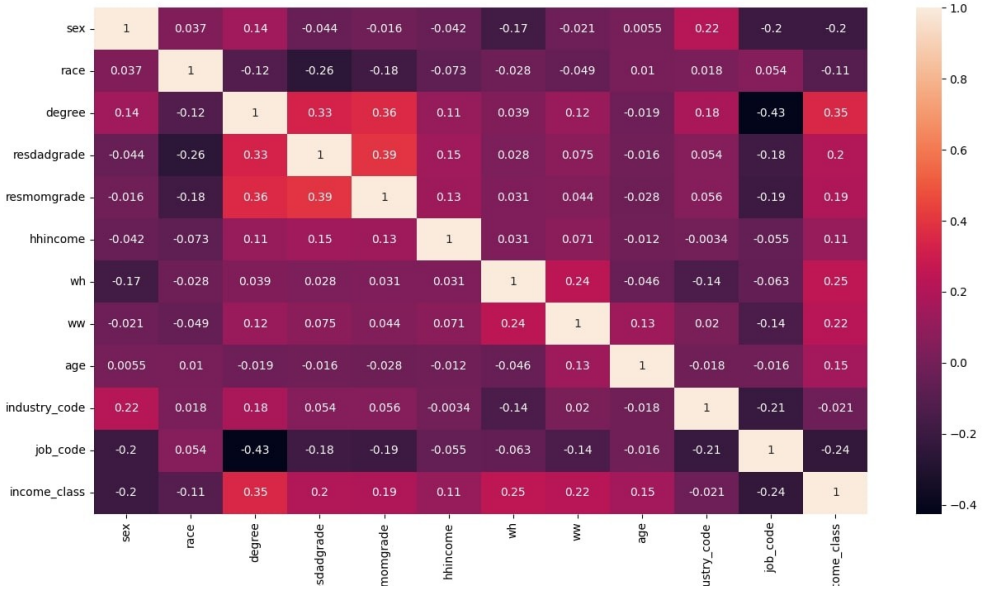

Figure 2. Updated heatmap of Spearman correlation coefficients for remaining 12 variables

Figure 1 displays the heatmap of the Spearman correlation coefficients for all variable pairs listed in Table 1. There was high correlation between pairs (highest degree, highest grade),

(biological father's highest grade, residential fathers highest grade) and (biological mothers highest grade, residential mothers highest grade), hence we decided to keep one variable from each pair, *viz.* highest degree, residential fathers highest grade, and residential mothers highest grade, leaving us with 12 variables for subsequent study and analysis as shown in Table 5. Figure 2 displays the updated heatmap of Spearman correlation coefficients for the remaining 12 variables.

Table 5. The twelve variables and their data type used in this study

Variables	Data type and values
sex	nominal: 1, 2
race	nominal: 1, 2, 3, 4, 5
degree	nominal: 0, 1, 2, 3, 4, 5, 6, 7
residential fathers highest grade	numeric: 0-20, 95
residential mothers highest grade	numeric: 0-20, 95
parental household income	numeric
age	numeric
industry	nominal: 1-18
occupation	nominal: 1-33
total work weeks	numeric
yearly work hours	numeric
income (yearly)	nominal: - 1 (Less than 50K), - 2 (Between 50K and 100K), - 3 (More than 100K)

Models

A baseline model was first established to set a benchmark for further analysis. Then, the effectiveness of several popular machine learning algorithms in weka was assessed for income classification. This included the multilayer perceptron neural network (16), sequential minimal optimization (17) for Support Vector Machines (18) and Random Forest (19) - three powerful and representative machine learning models. Each model has proven effective in conducting prediction tasks in prior research. The top-performing algorithm was then selected to conduct additional prediction tasks, to facilitate further in-depth analysis.

Baselining the data

In exploratory analysis, the income classes presented with a distribution of 57.6%, 31.3% and 11.1% respectively for the three income classes (<50K, >=50K and <100K, and >=100K). The baseline model is to predict the majority class <50K for all items, which resulted in the baseline in Table 6.

Table 6. Benchmark by baseline model

Models	Correctly Classified Instances	ROC Area
vote majority	57.6%	0.500

Multilayer Perceptron Neural Network

A multilayer perceptron (MLP) is a type of artificial neural network consisting of multiple layers of neurons. The neurons in the MLP typically use nonlinear activation functions, allowing the network to learn nonlinear relationships in data.

Sequential Minimal Optimization for Support Vector Machine

Sequential Minimal Optimization (SMO) is a widely used algorithm for training Support Vector Machines (SVMs). SVMs are powerful supervised learning models for solving classification and regression tasks by finding the hyperplane that best separates the data points into predefined classes while maximizing the margin between the classes. SMO, developed by John Platt, can efficiently solve the quadratic programming problem for training SVMs.

Random Forest

The Random Forest algorithm consists of a group of decision trees, each being trained independently on a random subset of the training data and variables. For predictions, the final result is the average (for regression tasks) or majority vote (for classification tasks) over the independent predictions from all the trees in the forest. The Random Forest algorithm has many advantages in prediction tasks, including high predictive accuracy, resistance to overfitting, robustness to noise and outliers, adeptness at handling missing values, and scalability.

Analysis

Model analysis

To ensure a fair and consistent model performance analysis, we adopted a uniform test option - percentage split - in the classification tasks in this study. This involved allocating 80% of the data for model training and the remaining 20% for model testing. This train-test split helps detect and prevent model overfitting. We also conducted the analysis using 10-fold cross-validation.

The evaluation metrics used to assess model performance included the prediction accuracy in Equation (1) (measured by Correctly Classified Instances metrics in weka) and Area Under the Curve (AUC - a weighted average of ROC Area in weka) for classification tasks.

$$\text{accuracy} = \frac{\text{number of correctly classified instances}}{\text{total number of instances}} \times 100\% \quad (1)$$
$$= \frac{\text{number of true positive instances} + \text{number of true negative instances}}{\text{total number of instances}} \times 100\%$$

The Receiver Operating Characteristic (ROC) curve is a graphic representation of the performance of a classification model across different threshold values. The AUC measures the area under the ROC curve. A perfect classifier would have an AUC of 1 and a random classifier would have an AUC of 0.5, which is no better than chance.

Longitudinal data analysis

To evaluate the influence of longitudinal data on prediction performance, two classification tasks were designed: one using NLSY97 data solely from one year and the other using NLSY97 data from multiple years including 2021, 2019, 2017 and 2015. Then we assessed the model performance metrics for both tasks to determine whether integrating longitudinal data improved the model prediction performance.

Factor significance analysis

Compared to traditional statistical methods, machine learning models usually lack a clear interpretation of significance levels of individual factors in prediction tasks. SHapley Additive exPlanations (SHAP) values (20) were used for factor significance analysis in machine learning models in this study. SHAP is a game theory approach to explain the output of machine learning models. It assigns each feature an important value - a SHAP value. SHAP values quantify the contribution of individual features to model predictions, enhancing the interpretability of complex models and helping identify the most influential features in the model's decision-making process. In this study, the Tree SHAP algorithm was used in calculating SHAP values, which is tailored for tree-based algorithms like Random Forest and; importantly; does not rely on a feature-independence assumption.

Results and Discussion

In this section, the analysis results for model selection, longitudinal data contribution and factor significance are compiled. Weka was used in classification tasks with different machine learning models.

Model selection

To assess the effectiveness of the three machine learning algorithms for the classification tasks in this study, NLSY97 data from the year 2021, 2019, 2017 and 2015 was fed into each model. Table 7 shows the prediction accuracy and the area under the ROC for the three algorithms in this prediction task. The Random Forest algorithm presented with both, a greater prediction accuracy and a larger ROC area. Hence, this model was chosen for all the subsequent analyses.

Table 7. Model performance comparison

Models	Correctly Classified Instances	ROC Area
MLP	68.0 %	0.807
SVM	67.9 %	0.732
Random Forest	72.6 %	0.849

For the testing dataset consisting of twenty percent of the input data (20% of 20691, i.e., approximately 4138 instances), the three machine learning models produced the confusion matrixes presented in Table 8 through Table 10.

Table 8. Confusion matrix for the MLP model

Actual class ↓ Predicted class →	Less than 50K	Between 50K and 100K	More than 100K
Less than 50K	1888	460	34
Between 50K and 100K	435	762	100
More than 100K	74	222	163

Table 9. Confusion matrix for the SVM model

Actual class ↓ Predicted class →	Less than 50K	Between 50K and 100K	More than 100K
Less than 50K	2036	338	8
Between 50K and 100K	538	739	20
More than 100K	81	342	36

Table 10. Confusion matrix for the Random Forest model

Actual class ↓ Predicted class →	Less than 50K	Between 50K and 100K	More than 100K
Less than 50K	2064	299	19
Between 50K and 100K	473	749	75
More than 100K	87	180	192

Longitudinal data contribution

The following experiments were designed to examine the contribution of longitudinal data to the model performance in predicting individual income. Five random forest models were trained as presented in Table 11. Four models were trained with data from individual years, while the fifth model was trained with data points from multiple years, to enable utilization of longitudinal data.

Table 11. Procedure for evaluating the advantage of incorporating longitudinal data

Models	Training data
Model 2015	4000 random data points from year 2015
Model 2017	4000 random data points from year 2017
Model 2019	4000 random data points from year 2019
Model 2021	4000 random data points from year 2021
Model 2015-2021	4000 random data points from year 2015-2021, 1000 data points from each year

Four sets of test data from the year 2015 through 2021 respectively were used to evaluate the model performance. Each set consisted of 1000 test data points randomly drawn from each year (excluding those used in model training). Table 12 shows the prediction accuracy and ROC area of the random forest machine learning models with and without longitudinal data on the four sets of test data. The fifth Model 2015-2021 trained with longitudinal data outperformed other models trained without longitudinal data, in terms of both correctly classified instances and ROC areas.

Table 12. Comparison prediction performance with and without longitudinal data

Tasks	Model performance without longitudinal data	Model performance with longitudinal data
1000 test data points from 2015	Model: Model 2015 Correctly Classified Instances: 70.3% ROC Area: 0.768	Model: Model 2015-2021 Correctly Classified Instances: 72.2% ROC Area: 0.801
1000 test data points from 2017	Model: Model 2017 Correctly Classified Instances: 65.1% ROC Area: 0.772	Model: Model 2015-2021 Correctly Classified Instances: 67.6% ROC Area: 0.810
1000 test data points from 2019	Model: Model 2019 Correctly Classified Instances: 65.3% ROC Area: 0.795	Model: Model 2015-2021 Correctly Classified Instances: 68.4% ROC Area: 0.819
1000 test data points from 2021	Model: Model 2021 Correctly Classified Instances: 64% ROC Area: 0.793	Model: Model 2015-2021 Correctly Classified Instances: 65.9% ROC Area: 0.813

The result demonstrated a significant improvement in model prediction performance with the use of longitudinal data, confirming our initial expectation that incorporating longitudinal data can enhance machine learning models for individual income prediction.

Factor significance

Figure 3 and 4 show the Random Forest feature importance with SHAP values. These graphs show how each factor impacted the model output.

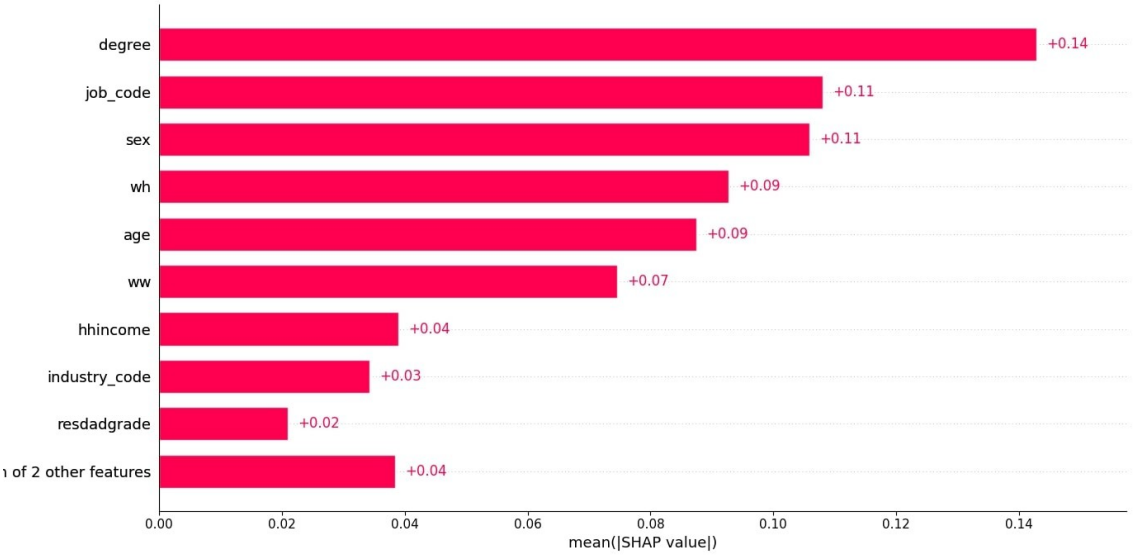

Figure 3. SHAP value (impact on model output) - mean absolute

In Figure 3, the most influential variables for predicting individual income are an individual's highest education degree, occupation and sex, in that order. Following closely are the variables of yearly working hours, age and work tenure, with all other variables considered less influential.

In Figure 4, it is not surprising to observe that variables such as the highest degree attained, yearly working hours, age, work tenure, and parental household income exhibit a positive correlation with individual income. Additionally, specific occupations are associated with higher incomes, and males tend to earn higher incomes compared to females.

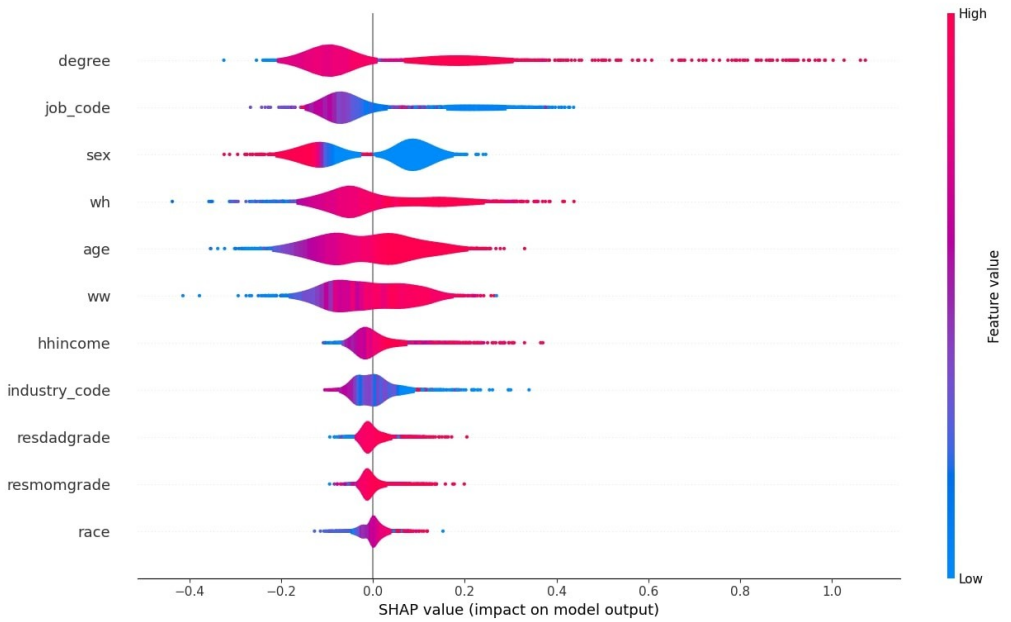

Figure 4. SHAP value (impact on model output)

Table 13. Prediction performance with different feature sets

Experiments	Correctly classified instances	ROC Area
baseline: with all 11 features	72.6 %	0.849
without degree	69.5 %	0.825
without occupation	70.1 %	0.818
without sex	71.6 %	0.840
without yearly working hours	71.8 %	0.844
without age	71.4 %	0.844
without work tenure	72.2 %	0.841
without parental household income	71.3 %	0.839
without industry code	70.2 %	0.831
without residential father's grade	72.3 %	0.844
without residential mother's grade	71.9 %	0.843
without race	71.0 %	0.839

In addition to examining SHAP values for individual features, the following experiments were performed to directly assess the quantitative impact of each individual feature on income prediction performance. We established a baseline using all 11 features, then systematically dropped one feature at a time to produce 11 additional experiments. The prediction performance of the Random Forest machine learning model for all 12 experiments is presented in Table 13. As seen, the Area under the ROC and the correctly classified instances are consistent with previous results.

This study reinforces the predictive power of machine learning models observed in prior economic research on income prediction. Compared to previous research in this area, this study makes several distinctive contributions. Firstly, it leveraged both individual longitudinal (intra-individual) characteristics and population (inter-individual) characteristics in national longitudinal survey data, yielding increased accuracy and ROC area in predicting individual income levels. Secondly, aside from achieving satisfactory accuracy at predicting income levels, this study explored the significance levels of various socioeconomic factors in the prediction task, leading to valuable insights. With these unique aspects, this work sheds new light on the potential of employing longitudinal characteristics in research to increase the accuracy of individual income prediction, and providing valuable practical insights on the relationship between socioeconomic factors and individual income levels, for both, individuals striving for financial success, and policymakers seeking to promote broader societal well-being.

Conclusion and perspective

This study demonstrated that with careful data preprocessing, feature engineering, proper model selection, and effectively leveraging longitudinal data, machine learning methodologies can predict individual income with reasonable accuracy. It also provided valuable insights on how various socioeconomic factors influence individuals' income levels, aiding individuals striving for financial success in maintaining focus, and assisting policymakers in making informed decisions for the broader socioeconomic well being of society.

There were several findings unique to this study. Firstly, making proper use of longitudinal information in input dataset yielded increased prediction accuracy in individual income prediction tasks. Secondly, the analysis made use of SHAP values and auxiliary approaches, leading to valuable insights into the significance levels of various socioeconomic factors on individual incomes, with potentially significant impact in practice. The top three influential variables for predictding individual incomes were identified as an individual's highest education degree, occupation and sex. These were followed by yearly working hours, age and work tenure. The actionable variables, such as highest degree, occupation, and yearly working hours, hold particular significance for individuals aiming for higher incomes.

Interesting future directions for this research were identified. This study focuses on the NLSY97 data, while similar methodologies employed herein could be extended to other datasets. Additional socioeconomic factors,

beyond those listed in Table 1, such as health-related and geographic factors, could be included to discover more patterns and influential factors affecting individual incomes. When removing the sex/gender variable from the machine learning model (as listed in Table 5), a lower prediction accuracy (71.5% *versus* 72.6%) and smaller ROC area (0.84 *versus* 0.85) were observed. These results highlight the significant issue of sex/gender bias as influencing individual income. This study exploited longitudinal information in the input dataset by disaggregating repeated measures on an individual into multiple entries, each corresponding to a specific time point for the individual, before input into machine learning models. There may be alternative methods for leveraging longitudinal data. One limitation with the use of longitudinal data in income prediction is that due to the inflation of individual incomes over time, employing longitudinal data over an extensive time frame might lead to a deterioration in the accuracy of prediction tasks. There may be methods to mitigate the impact of income inflation in input data, allowing for a more effective use of longitudinal data.

Acknowledgement

I would like to thank Professor Ramin Ramezani from UCLA for the informative and engaging meetings on AI and Machine Learning. I also want to express my gratitude to my TA, Joanna Gilberti from New York University, for her invaluable assistance in editing and maintaining progress with this project.

References

1. Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1997 cohort, 1997-2021 (rounds 1-20). Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. Columbus, OH: 2024. https://www.nlsinfo.org/investigator/

2. C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, **11**, Article number: 2583, 2020. https://www.nature.com/articles/s41467-020-16185-w

3. A. Sheetal, S. H. Chaudhury, K. Savani. A deep learning model identifies emphasis on hard work as an important predictor of income inequality. *Scientific Reports*, **12**, Article number: 9845, 2022. https://www.nature.com/articles/s41598-022-13902-x

4. R. Kannan, K. W. Shing, K. Ramakrishnan, H. B. Ong, A. Alamsyah. Machine learning models for predicting financially vigilant low-income households. *IEEE Access*, vol. **10**, pp. 70418-70427, 2022. https://ieeexplore.ieee.org/document/9810947

5. E. Aiken, S. Bellue, D. Karlan, C. Udry, J. E. Blumenstock. Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, **603**, 864–870, 2022. https://www.nature.com/articles/s41586-022-04484-9

6. H. Teguim. Estimating the returns to education using a machine learning approach – evidence for different regions. *Open Education Studies*, **5**(1), 2023. https://doi.org/10.1515/edu-2022-0201

7. D. A. Gomez-Cravioto, R. E. Diaz-Ramos, N. Hernandez-Gress, J. L. Preciado, H. G. Ceballos. Supervised machine learning predictive analytics for alumni income. *Journal of Big Data*, **9**, Article number: 11, 2022. https://doi.org/10.1186/s40537-022-00559-6

8. A. Lazar, Income prediction via support vector machine, *2004 International Conference on Machine Learning and Applications*, 2004. Proceedings., Louisville, KY, USA, pp. 143-149, 2004. https://ieeexplore.ieee.org/document/1383506

9. S. C. Matz, J. I. Menges, D. J. Stillwell, H. A. Schwartz. Predicting individual-level income from Facebook profiles. *PLoS ONE* **14**(3): e0214369, 2019. https://doi.org/10.1371/journal.pone.0214369

10. P. Khongchai, P. Songmuang. Improving students' motivation to study using salary prediction system. *2016 13th International Joint Conference on Computer Science and Software Engineering* (JCSSE), Khon Kaen, Thailand, pp. 1-6, 2016. https://ieeexplore.ieee.org/abstract/document/7748896

11. M. A. Islam, A. Nag, N. Roy, A. R. Dey, S. F. A. Fahim, A. Ghosh. An investigation into the prediction of annual income levels through the utilization of demographic features employing the modified UCI adult dataset. *2023 International Conference on Computing, Communication, and Intelligent Systems* (ICCCIS), 1080-1086, 2023. https://ieeexplore.ieee.org/document/10425394

12. M. Zaid, T. Rajendran. Higher classification accuracy of income class using decision tree algorithm over naive bayes algorithm. *Advances in Parallel Computing Algorithms, Tools and Paradigms*, 2022. https://ebooks.iospress.nl/doi/10.3233/APC220079

13. N. Chakrabarty, S. Biswas. A statistical approach to adult census income level prediction. *2018 International Conference on Advances in Computing, Communication Control and Networking* (ICACCCN), Greater Noida, India, pp. 207-212, 2018. https://ieeexplore.ieee.org/document/8748528

14. G. M. Fitzmaurice, N. M. Laird, J. H. Ware. *Applied Longitudinal Analysis* (2nd ed.). Wiley, 2011. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119513469

15. H. Donald, G. Robert D. *Longitudinal Data Analysis* (1st ed.). Wiley-Interscience, 2006. https://onlinelibrary.wiley.com/doi/book/10.1002/0470036486

16. D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning representations by back-propagating errors. *Nature*, **323**, 533–536, 1986. https://www.nature.com/articles/323533a0

17. J. Platt. Sequential Minimal Optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning*. 208, 1998. https://www.researchgate.net/publication/2624239

18. N. Cristianini, J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000. https://doi.org/10.1017/CBO9780511801389

19. L. Breiman. Random Forests. *Machine Learning* **45**, 5–32, 2001. https://doi.org/10.1023/A:1010933404324

20. L. Shapley. A value for n-person games. In: *Contributions to the Theory of Games*, vol. **II** Princeton University Press, 307–317, 1953. https://doi.org/10.1515/9781400829156-012

Appendix

Explanation of Variables for NLSY97 Data Used in This Study

All the following information is available in NLS Investigator NLSY97 codebook, https://www.nlsinfo.org/investigator/pages/home

All variables can take on their specific valid input values and the following invalid/skip values:

- -1: Refusal
- -2: Don't Know
- -3: Invalid Skip
- -4: VALID SKIP
- -5: NON-INTERVIEW

Variables, their NLSY97 variable title and their valid input values:

- Sex
    - NLSY97 variable title: KEY!SEX, RS SEX (SYMBOL)
    - Definition: Sex of Youth
    - Possible values of the variable:

        - 1 Male
        - 2 Female
        - 0 No information
- Race
    - NLSY97 variable title: KEY!RACE, RACE OF R (SYMBOL)
    - Definition: Race
    - Possible values of the variable:
        - 1 White
        - 2 Black or African American
        - 3 American Indian, Eskimo, or Aleut
        - 4 Asian or Pacific Islander
        - 5 Something else? (SPECIFY)
        - 0 No information
- Degree
    - NLSY97 variable title: HIGHEST DEGREE RECEIVED
    - Definition: The highest degree received
    - Possible values of the variable:
        - 0 None
        - 1 GED
        - 2 High school diploma (Regular 12 year program)
        - 3 Associate/Junior college (AA)
        - 4 Bachelor's degree (BA, BS)
        - 5 Master's degree (MA, MS)
        - 6 PhD
        - 7 Professional degree (DDS, JD, MD)
- Biological father’s highest grade
    - NLSY97 variable title: BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED
    - Definition: Highest grade completed by respondent's biological father
    - Possible values of the variable:
        - 0 NONE
        - 1-12 Nth GRADE
        - 13 1ST YEAR COLLEGE
        - 14 2ND YEAR COLLEGE
        - 15 3ND YEAR COLLEGE
        - 16 4ND YEAR COLLEGE
        - 17 5ND YEAR COLLEGE
        - 18 6ND YEAR COLLEGE
        - 19 7ND YEAR COLLEGE
        - 20 8TH YEAR COLLEGE OR MORE

        - 95 UNGRADED
- Biological mothers highest grade
    - NLSY97 variable title: BIOLOGICAL MOTHERS HIGHEST GRADE COMPLETED
    - Definition: Highest grade completed by respondent's biological mother
    - Possible values of the variable:
        - Same as above for Biological father's highest grade
- Residential fathers highest grade
    - NLSY97 variable title: RESIDENTIAL FATHERS HIGHEST GRADE COMPLETED
    - Definition: Highest grade completed by respondent's residential father
    - Possible values of the variable:
        - Same as above for Biological father's highest grade
- Residential mothers highest grade
    - NLSY97 variable title: RESIDENTIAL MOTHERS HIGHEST GRADE COMPLETED
    - Definition: Highest grade completed by respondent's residential mother
    - Possible values of the variable:
        - Same as above for Biological father's highest grade
- Parental household income
    - NLSY97 variable title: GROSS HH INCOME IN PAST YEAR
    - Definition: Gross household income in the previous year of 2003
    - Possible values of the variable:
        - -999999 to 999999
- Highest grade
    - NLSY97 variable title: RS HIGHEST GRADE COMPLETED
    - Definition: The highest grade completed as of the survey date
    - Possible values of the variable:
        - Same as above for Biological father's highest grade
- Age
    - NLSY97 variable title: RS AGE IN MONTHS AS OF INTERVIEW DATE
    - Definition: Age in months as of the interview date
    - Possible values of the variable:
        - 0 to 520
- Industry
    - NLSY97 variable title: YEMP, TYPE OF BUS OR INDUSTRY CODE (2002 CENSUS 4-DIGIT) 01 (ROS ITEM)
    - Definition: These are 2002 Census Industry Codes
    - Possible values of the variable:
        - 170 TO 290: AGRICULTURE, FORESTRY AND FISHERIES

        - 370 TO 490: MINING
        - 570 TO 690: UTILITIES
        - 770: CONSTRUCTION
        - 1070 TO 3990: MANUFACTURING
        - 4070 TO 4590: WHOLESALE TRADE
        - 4670 TO 5790: RETAIL TRADE
        - 5890: ARTS, ENTERTAINMENT AND RECREATION SERVICES
        - 6070 TO 6390: TRANSPORTATION AND WAREHOUSING
        - 6470 TO 6780: INFORMATION AND COMMUNICATION
        - 6870 TO 7190: FINANCE, INSURANCE, AND REAL ESTATE
        - 7270 TO 7790: PROFESSIONAL AND RELATED SERVICES
        - 7860 TO 8470: EDUCATIONAL, HEALTH, AND SOCIAL SERVICES
        - 8560 TO 8690: ENTERTAINMENT, ACCOMODATIONS, AND FOOD SERVICES
        - 8770 TO 9290: OTHER SERVICES
        - 9370 TO 9590: PUBLIC ADMINISTRATION
        - 9670 TO 9890: ACTIVE DUTY MILITARY
        - 9950 TO 9990: ACS SPECIAL CODES
    - These values are categorized and encoded into 1 to 18 in the above order in this study
- Occupation
    - NLSY97 variable title: YEMP, OCCUPATION/JOB CODE (2002 CENSUS 4-DIGIT) 01 (ROS ITEM)
    - Definition: These are 2002 Census Occupation Codes
    - Possible values of the variable:
        - 10 TO 430: EXECUTIVE, ADMINISTRATIVE AND MANAGERIAL
        - 500 TO 950: MANAGEMENT RELATED
        - 1000 TO 1240: MATHEMATICAL AND COMPUTER SCIENTISTS
        - 1300 TO 1530: ENGINEERS, ARCHITECTS, AND SURVEYORS
        - 1540 TO 1560: ENGINEERING AND RELATED TECHNICIANS
        - 1600 TO 1760: PHYSICAL SCIENTISTS
        - 1800 TO 1860: SOCIAL SCIENTISTS AND RELATED WORKERS
        - 1900 TO 1960: LIFE, PHYSICAL, AND SOCIAL SCIENCE TECHNICIANS
        - 2000 TO 2060: COUNSELORS, SOCIAL, AND RELIGIOUS WORKERS
        - 2100 TO 2150: LAWYERS, JUDGES, AND LEGAL SUPPORT WORKERS
        - 2200 TO 2340: TEACHERS
        - 2400 TO 2550: EDUCATION, TRAINING, AND LIBRARY WORKERS

        - 2600 TO 2760: ENTERTAINERS AND PERFORMERS, SPORTS AND RELATED WORKERS
        - 2800 TO 2960: MEDIA AND COMMUNICATION WORKERS
        - 3000 TO 3260: HEALTH DIAGNOSIS AND TREATING PRACTITIONERS
        - 3300 TO 3650: HEALTH CARE TECHNICAL AND SUPPORT
        - 3700 TO 3950: PROTECTIVE SERVICE
        - 4000 TO 4160: FOOD PREPARATIONS AND SERVING RELATED
        - 4200 TO 4250: CLEANING AND BUILDING SERVICE
        - 4300 TO 4430: ENTERTAINMENT ATTENDANTS AND RELATED WORKERS
        - 4460: FUNERAL RELATED OCCUPATIONS
        - 4500 TO 4650: PERSONAL CARE AND SERVICE WORKERS
        - 4700 TO 4960: SALES AND RELATED WORKERS
        - 5000 TO 5930: OFFICE AND ADMINISTRATIVE SUPPORT WORKERS
        - 6000 TO 6130: FARMING, FISHING, AND FORESTRY
        - 6200 TO 6940: CONSTRUCTION TRADES AND EXTRACTION WORKERS
        - 7000 TO 7620: INSTALLATION, MAINTENANCE, AND REPAIR WORKERS
        - 7700 TO 7750: PRODUCTION AND OPERATING WORKERS
        - 7800 TO 7850: FOOD PREPARATION
        - 7900 TO 8960: SETTER, OPERATORS, AND TENDERS
        - 9000 TO 9750: TRANSPORTATION AND MATERIAL MOVING WORKERS
        - 9800 TO 9840: MILITARY SPECIFIC OCCUPATIONS
        - 9950 TO 9990: ACS SPECIAL CODES
    - These values are categorized and encoded into 1 to 33 in the above order in this study
- Total work weeks
    - NLSY97 variable title: # WEEKS EVER WORKED AT JOB 01 - 10
    - Definition: Number of weeks of total tenure at any job reported in the employer roster
    - (YEMP) to the survey date
    - The sum of 10 variables for the 10 jobs is used in this study
    - Possible values of the variable:
        - 0 to 2000
- Yearly work hours

    - NLSY97 variable title: YYYY EMPLOYMENT: TOTAL HOURS WORKED WEEK 01 to 52
    - Definition: Total number of hours worked in this week at a civilian job in YYYY
    - The sum of 52 variables for the 52 weeks of the year is used in this study
    - Possible values of the variable:
        - 0 to 500
- Income (yearly)
    - NLSY97 variable title: TOTAL INCOME FROM WAGES AND SALARY IN PAST YEAR
    - Definition: During 2020, how much income did you receive from wages, salary, commissions, or tips from all jobs, before deductions for taxes or for anything else?
    - Possible values of the variable:
        - 0 to 99999999