Dean&Francis
ISSN 2959-6130

# Income Group Classification Case Study

**Kunyun Han**

**Abstract:**

Energy is the significant resources for global economic growth and development and for human daily life. In recent years, the increasing scarcity of energy and the continuous rise in energy prices have posed more severe challenges to many countries in the current international environment. Residents' energy consumption, which can be categorized into indirect consumption from production driven by consumption and direct consumption for daily living, is intricately linked to their consumption behavior. And as we know, income levels play a decisive role in shaping this consumption behavior. Different income brackets often lead to distinct energy consumption patterns; for instance, higher - income individuals might consume more energy through luxury goods production and larger living spaces, while lower - income groups focus on basic energy needs for survival and daily activities.

Given this crucial connection, understanding the factors that influence income becomes paramount. This paper uses a sample size of 65,062 and applies two methods, univariate analysis and bivariate analysis, to comprehensively analyze the impacts of micro - factors such as an individual's years of education and macro - factors such as industry characteristics on personal income. A logistic regression model is constructed to predict the level of personal income based on different income - influencing factors, aiming to identify citizens with a wage level of less than $50,000 and assist the government in formulating relevant policies.

The study finds that: The coefficients of some levels of education, workclass, and native country are positive. An increase in these factors will lead to an increase in the probability of a person having a salary of <= 50,000. The coefficients of age, fnlwgt, marital_status, working_hours_per_week, some levels of education, workclass, and native country are negative. An increase in these factors will lead to a decrease in the probability of a person having a salary of <= 50,000. Based on the above research results, this paper puts forward corresponding policy recommendations for enhancing the income level of vulnerable groups and improving their quality of life.

**Keywords:** Personal Income, Logistic Regression, Wage Prediction, Influencing Factors

# 1. Introduction

## 1.1 Research background

Energy stands as a cornerstone for global economic growth, development, and human daily life. However, in recent years, the escalating scarcity of energy resources, coupled with the continuous surge in energy prices, has presented formidable challenges to numerous countries within the current international landscape. These issues not only impede economic progress but also disrupt the stability of daily life, highlighting the urgency of addressing energy - related concerns.Consumption behavior, which is intricately intertwined with energy usage, is predominantly shaped by income levels. Higher - income individuals typically exhibit consumption patterns that lead to greater energy consumption, such as the acquisition of luxury goods, which require substantial energy inputs during production, and the occupation of larger living spaces that demand more energy for heating, cooling, and lighting. Conversely, lower - income groups tend to prioritize basic energy requirements for survival and essential daily activities. This disparity in energy consumption based on income brackets underscores the significance of understanding the factors that influence income.

From the perspective of individual characteristics, factors such as gender, age, and education level have a profound impact on income distribution. Gender differences have led to the widespread existence of unequal pay for men and women in the workplace. There are obvious differences in the income-earning ability among people of different age groups. Younger employees are often restricted in their income due to factors such as lack of experience, while older groups may also face problems such as career development bottlenecks that affect their income growth. Education level is directly related to an individual's career choice and salary level. People with higher education usually have stronger competitiveness in the labor market and can obtain higher income. In terms of occupation, the income gap between different occupations has been continuously widening. In addition, the relationship between working hours and income is not a simple linear one. For some positions with high intensity and long working hours, the income return does not match the effort.

This situation of income inequality has attracted widespread attention around the world. It not only concerns the economic well-being of individuals but also poses a potential threat to social stability and harmonious development. The continuous widening of the gap between the rich and the poor may lead to the solidification of social classes, limit social mobility, and hinder the sustainable growth of the economy. Against this backdrop, the government is actively collaborating with the World Health Organization to jointly formulate policies to improve the living conditions of vulnerable groups and mitigate the negative impacts brought about by income inequality.

## 1.2 Purpose of research

This study holds great significance both at the theoretical and practical levels.

Theoretically, by applying univariate analysis and bivariate analysis methods, it deeply analyzes the action mechanisms of micro factors such as an individual's years of education and macro factors such as industry characteristics on personal income, thus enriching the research content of the income distribution theory. Constructing a logistic regression model to predict personal income provides a new perspective and practical case for the research methods in this field. It helps to further improve and expand the theoretical system of income distribution and offers theoretical references and methodological examples for subsequent related studies.

Practically, this study aims to accurately identify the group of citizens whose wage level is lower than $50,000, providing crucial data support and a basis for decision-making for the government to formulate targeted policies. Based on the research results, the government can formulate targeted education support policies to improve the education level of vulnerable groups and enhance their employability and income-earning ability. In terms of industrial policies, it can guide resources to incline towards industries that can absorb more low-income groups for employment and increase their income levels, promoting the optimization of the employment structure and fairness in income distribution. Moreover, through policy tools such as taxation and social security, the government can adjust income distribution and narrow the gap between the rich and the poor. This has important practical guiding significance for improving the quality of life of vulnerable groups, promoting social fairness and justice, and driving the healthy and stable development of the economy.

## 1.3 Literature review

In the article "Real and Nominal Interstate Income inequality in the United States: Further Evidence", the author Ram points out that the individual income inequality in the United States has become a hot topic in contemporary society, and the high level of income inequality among states is likely to reduce the annual economic growth rate of the United States (Ram et al., 2015)[1]. Given the urgency and importance of improving the quality of life of vulnerable groups, over the past 20 years, scholars have used regression models and combined with different variable factors to conduct predictive analyses of personal income. For example, Jacob Mincer used the Ordinary Least Squares (OLS) regression method to construct the Mincer equation. By quantifying the impacts of work

ISSN 2959-6130

experience and education level on personal income, he predicted the level of individual income, providing data support for policy-making. Another example is that Wang Shiting from Renmin University of China applied multiple regression analysis and added variables such as gender and family background, further improving the accuracy of predicting personal income. These studies share similarities with this study in terms of methodology, all aiming to construct effective predictive models by exploring key influencing factors[2]. However, this study has a much larger sample size (with a sample size of 65,062). Moreover, it comprehensively applies two methods, namely univariate analysis and bivariate analysis, to conduct a detailed analysis of the influencing factors. It also continuously optimizes the logistic regression model and eliminates multicollinearity, showing certain characteristics in terms of the comprehensiveness of model construction and the analysis of influencing factors.

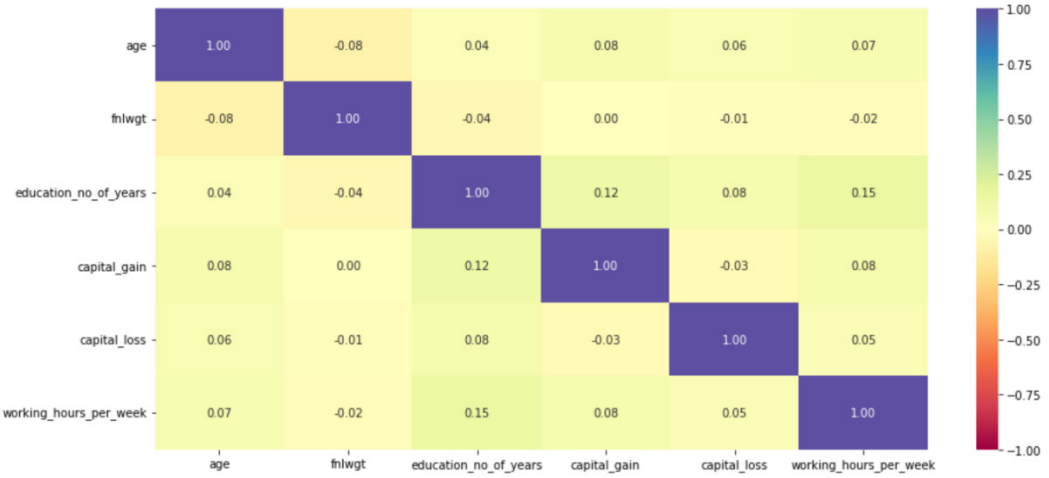
2.EDA

## 2.1 Univariate analysis

In the univariate analysis, the average age in this dataset is approximately 38 years, with the age range spanning from 17 to 90 years. The average years of education are around 10 years, and the average working hours per week are 40, with the majority of individuals coming from private enterprises.And ~24% of the observations have a salary above 50K and while ~76% have a salary below 50K

## 2.2 Bivariate analysis

In the multivariate analysis, we first conducted a correlation test. As can be seen from the graph, there is no significant correlation among various variables. After clarifying the correlations between variables, we can more rationally select the variables to be included in the model. Variables that have a strong correlation with the dependent variable and a weak correlation with each other should be given priority, which can improve the explanatory power and prediction accuracy of the model.




### 2.2.1 The influence of gender on income

Gender has a positive impact on income level, that is, there exists a gender-based salary gap in society. Among them, approximately 30% of men have a salary exceeding $50,000, while only about 10% of women have a salary exceeding $50,000.

### 2.2.2 The influence of education on income

The number of years of education (educational level) has a positive impact on income level, that is, as the number of years of education (educational level) increases, income also increases. Among them, about 70% of people with a doctoral degree have a salary exceeding $50,000.

### 2.2.3 The influence of occupation on income

Occupation also has a significant impact on income level. Among people working as administrative managers and professional professors, about 50% of them have a salary exceeding $50,000. People engaged in occupations such as private domestic service, cleaners, and agriculture and fishery are more likely to have a salary lower than $50,000.

### 2.2.4 The influence of workclass on income

About 50% of self-employed people have a salary exceeding $50,000. Next, about 40% of federal government employees have a salary exceeding $50,000. Approximately 20% of private sector employees have an income exceeding $50,000.

### 2.2.5 The influence of age on income

There is a positive correlation between age and income, that is, the older the age, the higher the income level. It can be seen from the data that people with a salary exceeding $50,000 are usually older, with an average age of about 48 years old. The average age of people with a salary lower than $50,000 is about 36 years old.

#### 2.2.6 The influence of working hours on income

There is no absolute linear relationship between working hours and income. Most people with a salary exceeding $50,000 work about 40 hours per week. Compared with those with a salary lower than or equal to $50,000, people with a salary exceeding $50,000 have a wider range of working hours, but there are also outliers.

3.Model building

### 3.1 Build a logistic regression model

In the process of data analysis and modeling, data preprocessing is a crucial step to ensure the accuracy and reliability of the research. In this study, we carried out the following data preprocessing operations: Through preliminary exploratory analysis, it was found that many variables had outliers. For example, in the age variable, although the age range theoretically covers from 17 to 90 years old, some extreme values may deviate from the age distribution characteristics of the actual research objects. Variables such as the number of years of education and the number of working hours per week also have similar situations. There are significant differences between the minimum value and a specific percentile (such as the 25th percentile), and between the maximum value and the 75th percentile, indicating that there may be interference from outliers. If these outliers are not dealt with, they may have an adverse impact on subsequent statistical analysis and model fitting. For instance, they can distort statistics such as the mean and standard deviation, and reduce the robustness of the model. Therefore, we adopted the boxplot method to deal with the variables with outliers, so as to purify the dataset and make it better reflect the true distribution characteristics of the variables.

In this study, the logistic regression model was mainly selected for predicting the wage level. This model can effectively handle binary classification problems. In this study, the classification result of whether an individual's income is higher than $50,000 (yes/no) is the core focus. The logistic regression model can not only clearly show the relationship between various influencing factors and personal income, but also make a probability prediction of the individual's income level based on these factors, meeting the dual needs of this study to explore the influence mechanism of variables and achieve wage prediction.

We set whether an individual's income is higher than $50,000 as the dependent variable (Y). If an individual's income is higher than $50,000, the value of Y is 0; if an individual's income is lower than or equal to $50,000, the value of Y is 1. At the same time, taking into account both theoretical research and practical situations, we selected a series of independent variables such as age, gender, educational attainment, and working hours and included them in the model. These independent variables cover various factors that may affect personal income, providing data support for a comprehensive analysis of the income influencing mechanism.

The preprocessed data was divided into a training set and a test set at a ratio of 7:3. The training set is used for estimating the model parameters, while the test set is used for evaluating the performance of the model. Based on the selected independent variables and the dependent variable, a logistic regression model was constructed. The basic expression of the model is:

$$log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n$$

Among them, P(Y = 1)represents the probability that an individual's income is less than or equal to50,000. $\beta_0$ is the intercept term, $\beta_1\beta_2...\beta_n$ is the regression coefficient corresponding to each independent variable, and $X_1X_2...X_n$ represents different independent variables respectively. After that, the model parameters are estimated to enable the model to fit the data as accurately as possible on the training set, and to determine the direction and degree of the influence of each variable on personal income.

### 3.2 Model evaluation and optimization

After constructing the model, the test set was used to evaluate the performance of the model. Indicators such as accuracy, precision, recall, and the F1 score were adopted to measure the accuracy and reliability of the model's predictions. It can be seen from the following data that the model performs relatively well.

| Accuracy | Recall | Precision | F1 |
|---|---|---|---|
| 0.835624 | 0.923759 | 0.867897 | 0.894957 |

#### 3.2.1 Detecting and Dealing with Multicollinearity

After constructing the logistic regression model, diagnosing multicollinearity among independent variables is a crucial step to ensure the model's effectiveness. We used the Variance Inflation Factor (VIF) to detect the degree of multicollinearity among independent variables. Generally, when the VIF value is greater than 10, it indicates a serious multicollinearity problem among independent variables.

Upon testing, we found that some independent variables had relatively high VIF values, which is a common multicollinearity phenomenon in multiple linear regression. For example, there was a high degree of multicollinearity among "education_no_of_years", "education", and certain

ISSN 2959-6130

classification levels of “workclass”, “native_country”, and “race”, suggesting a strong linear correlation among them. This multicollinearity can lead to unstable estimates of regression coefficients and an increase in standard errors, thereby reducing the model’s prediction accuracy and reliability.

To address this issue, we took the following measures: First, we conducted a correlation analysis on the independent variables with multicollinearity to clarify the specific degree of association between variables. Then, based on the analysis results, we considered removing variables with excessively high correlations or integrating variables. For instance, we removed the variable “occupation_Unknown” because we could obtain the same information from “workclass_Unknown”.

After removing several variables, we found that removing “education_no_of_years” and “race” did not significantly improve the model’s performance. This means that in the current model, these two variables have weak explanatory power for the dependent variable (such as predicting personal income levels), and removing them would not reduce key performance indicators such as the model’s prediction accuracy and goodness - of - fit.

We then tested the VIF values again to ensure that the VIF values of all variables were within a reasonable range, effectively mitigating the negative impact of multicollinearity on the model.

#### 3.2.2 Removing high p-value variables

To further ensure the accuracy of the model, we also carried out an operation to remove variables with high p - values. In the initially constructed logistic regression model, the p - values of some independent variables failed to reach the ideal significance level. For example, variables such as “workclass_Never - worked” and “workclass_Without - pay” had p - values greater than the common significance threshold (e.g., 0.05), indicating that statistically, their influence on the dependent variable (whether an individual’s income is higher than $50,000) was not significant. This means that in the model, these variables failed to fully demonstrate their explanatory power for personal income, which might affect the overall prediction accuracy of the model and the accurate analysis of the income - influencing mechanism.

To improve this situation, we took the following measures. Since only the p - values of some classification levels were high, we chose an iterative approach to remove these variables. First, we constructed a model and checked the p - values of the variables. Then, we removed the variable column with the highest p - value. Based on the model after removing this variable, we created a new model, checked the p - values of the variables again, and then removed the variable column with the highest p - value at that time. This process was repeated until there were no variable columns with p - values greater than 0.05.

After the above series of optimization operations, we re - estimated the model parameters. The results showed that the p - values of the variables that originally had high p - values were significantly reduced and successfully fell into the acceptable significance range (less than 0.05). This indicates that the optimization measures effectively improved the significance of these variables in the model, enabling them to more accurately reflect their influence on personal income.

## 4. Results

After multiple adjustments and optimizations, the logistic regression model constructed in this study demonstrated certain performance on the test set. The accuracy of the model is 0.826%, the precision is 0.886%, and the recall is 0.885%. The F1 - score, which integrates precision and recall, reaches 0.885. It is a comprehensive measure of the model’s performance, indicating the model’s effectiveness in balancing accurate prediction and comprehensive capture of high Income group.

In terms of coefficient interpretation, we have identified that factors such as age, working hours, marital status, and educational attainment have a significant impact on individual wage income, providing a basis for understanding the income - influencing mechanism. For example, when keeping other characteristics constant, a one - unit increase in age will reduce the odds of an individual’s salary being less than or equal to $50,000 to approximately 0.97 times the original, which is equivalent to a decrease of about 3%. This indicates that as age increases, the likelihood of an individual’s salary being at a lower level (≤$50,000) gradually decreases. This may be because the increase in age is accompanied by the accumulation of work experience and skills, thereby enhancing the income level.

## 5. Discussion

### 5.1 Interpretation and Significance of Results

We have successfully constructed a predictive model with an F1 - score of 0.89 on the training set. This model can be utilized by the government to identify citizens with a salary of less than $50,000, enabling them to formulate corresponding policies. Moreover, all logistic regression models have demonstrated certain generalization performance on both the training and test sets, indicating their reliability and practicality, which provides an effective tool for subsequent policy - making.

Regarding the impact of variable coefficients, the coefficients of some levels of education, workclass, and native country are positive. This means that an increase in these

variables will lead to an increase in the probability of an individual having a salary ≤ $50,000. Conversely, the coefficients of age, fnlwgt, marital_status, working_hours_per_week, and some levels of education, workclass, and native country are negative. An increase in these variables will lead to a decrease in the probability of an individual having a salary ≤ $50,000. These findings reveal the complex mechanisms by which different variables affect personal salary levels and provide key insights for understanding income - influencing factors.

### 5.2 Comparison with Previous Studies

Previous research has also explored the determinants of personal income. Similar to our findings, studies by Jacob Mincer and Wang Shiting have shown that education and work experience (related to age in our study) are significant factors in income determination. However, our study differs in several aspects. With a larger sample size of 65,062, our results may be more generalizable. Additionally, our comprehensive use of univariate and bivariate analysis, along with the meticulous handling of multicollinearity and high - p - value variables in the logistic regression model, provides a more in - depth and accurate analysis of the influencing factors. For example, in some previous studies, the issue of multicollinearity may not have been addressed as comprehensively. Ignoring multicollinearity can lead to inaccurate coefficient estimates and unreliable model predictions. Our approach of using the Variance Inflation Factor (VIF) to detect and address multicollinearity, as well as the iterative removal of high - p - value variables, contributes to a more robust and accurate model.

### 5.3 Limitations of the Study

Despite the valuable insights gained, this study has several limitations. Firstly, the sample, although relatively large, may not be fully representative of the entire population. It may be subject to selection bias, as the data collection method or source might have excluded certain groups. For example, individuals in remote areas or specific industries that are difficult to access may not be adequately represented in the sample, which could limit the generalizability of the results.

Secondly, the study only considered a limited number of variables. There are many other potential factors that could influence personal income, such as personal abilities, family inheritance, regional economic development, and social networks. Omitting these variables may lead to an incomplete understanding of the income - influencing mechanism. For instance, in some regions with booming economies, individuals may have more opportunities to earn higher incomes regardless of their education or work experience.

Finally, the logistic regression model used in this study assumes a linear relationship between the log - odds of income and the independent variables. In reality, the relationship may be more complex, and non - linear relationships may exist. This simplification may not fully capture the true nature of the relationship between variables and income, potentially affecting the accuracy of the predictions.

## 6. Conclusion

In conclusion, this study has delved into the factors influencing personal income through the construction of logistic regression models. Our analysis has identified a range of variables, including age, education, workclass, marital status, and working hours, that significantly impact the likelihood of an individual earning a salary of ≤ $50,000. The model we developed, which demonstrates good generalization performance with an F1 - score of 0.89 on the training set, provides a practical tool for predicting citizens' income levels. This can assist the government in formulating targeted policies.

Given these findings, it is imperative that the government takes proactive steps. Reforms should be implemented to ensure private - sector employees are fairly compensated for their work. This not only addresses potential salary disparities within the private sector but also promotes overall income fairness. Additionally, government - formulated policies must prioritize equal pay and actively work to reduce the existing pay gap in society. By doing so, the government can contribute to creating a more equitable economic environment, enhancing social stability, and promoting the well - being of all citizens. Overall, this research not only enriches our understanding of personal income determinants but also offers actionable recommendations for policy - making in pursuit of a more just and prosperous society.

## 7. Acknowledgement

Here, I would like to express my sincere gratitude to my professor, for the meticulous guidance and support during the research and writing process. I also thank all those who participated in this study.

## 8. References

[1] Ram R. Real and nominal interstate income inequality in the United St ates: Further evidence[J]. International Advances in Economic Research, 2015, 21: 131-132.

[2] 王诗婷 . 居民收入的影响因素研究 —— 基于 2015 年 CGSS 调查数据的分析 [J]. 新经济 ,2020,(07):76-80.(in Chinese)