Dean&Francis
ISSN 2959-6157

# Application of Logistic Regression in Life

**Yukai Kong**

**Abstract:**

This research employs logistic regression analysis to discover fundamental determinants of income among vulnerable groups to help government bodies create effective social intervention policies. The study processed a complete dataset of 32,531 records from governmental sources gathered by the World Health Organization with systematic data cleaning followed by exploratory data analysis and feature engineering to develop logistic regression models and then subject these models to strict evaluation and refinement. The choice of Python for this analysis stemmed from its superior data management features alongside its sophisticated analytical library support. The study's analysis revealed that income levels above $50,000 are significantly linked to higher education attainment and particular employment categories. Researchers addressed multicollinearity by conducting Variance Inflation Factor (VIF) assessments and refined model variables through iterative evaluation of statistical significance (p-values). The completed logistic regression model demonstrated strong predictive power through an impressive F1 score of 0.89. The findings suggest we need to expand access to education while establishing fair pay systems and better working environments as well as developing specific policies to reduce economic inequality. Upcoming advancements will likely include extended demographic data collection and experimentation with superior machine learning methods along with the implementation of interpretability tools to enhance policy understanding.

**Keywords:** Logistic Regression, Income Prediction, Data Analysis, Socioeconomic Factors, Multicollinearity, Feature Engineering, Policy Recommendations

## Introduction

This comprehensive report outlines a detailed data science project aimed at leveraging logistic regression to identify factors influencing income levels among vulnerable populations. This initiative directly supports governmental efforts in formulating targeted social policies by accurately predicting individuals who are most likely to fall within the low-income bracket. Utilizing a robust dataset provided by governmental sources, initially gathered by the World Health Organization (WHO), the project emphasizes the significance of precise data analysis in enhancing the effectiveness and precision of existing social support frameworks.

## Business Problem Overview and Solution Approach

### Problem Definition

The primary challenge addressed by this project is identifying key determinants of income among vulnerable groups. Understanding these determinants is crucial for the government to develop and implement social policies that effectively target and allocate resources to those most in need, consequently improving their economic conditions and overall quality of life.

### Solution Approach

Our methodology comprised several critical phases, each employing Python for rigorous data analysis. Python was specifically selected due to its demonstrated effectiveness, flexibility, and extensive analytical libraries, positioning it as superior to conventional statistical methods for this particular application. The methodological phases implemented in our approach included: firstly, Data Cleaning, wherein Python's Pandas library was utilized for its capability to efficiently manage large datasets, ensuring data integrity and quality, which are fundamental to achieving accurate modeling outcomes. This phase encompassed systematically removing inconsistencies, addressing missing values, and verifying the dataset's reliability. (McKinney,2010) Secondly, we conducted Exploratory Data Analysis (EDA), leveraging Python's advanced visualization libraries, Matplotlib and Seaborn, which offer extensive graphical functionalities. These tools facilitated a thorough examination of data distributions and inter-variable relationships, enabling the identification of relevant trends, outliers, and underlying patterns essential for subsequent modeling steps. Thirdly, the Feature Engineering stage capitalized on Python's flexibility in generating intricate and meaningful derived variables. Informed directly by insights obtained from the EDA phase, this step effectively captured complex nonlinear relationships within the data, significantly enhancing the predictive capabilities of our model. Fourthly, the Model Development phase involved implementing logistic regression, selected specifically for its suitability in addressing binary classification tasks. Logistic regression provides clear probabilistic predictions and easily interpretable results, granting it distinct advantages over more complex machine learning techniques such as neural networks. Finally, the Model Evaluation and Refinement phase employed Python's robust statistical metrics, including the Variance Inflation Factor (VIF) and the F1 score, to systematically evaluate and enhance the logistic regression model. This process addressed critical modeling issues such as multicollinearity among predictors and potential model overfitting, thereby ensuring reliability and generalizability of the final analytical results.

## Data Overview

The dataset analyzed comprises 32,531 records, each with 14 attributes, including demographic and socioeconomic factors such as age, education level, capital gain, and working hours per week. Crucially, the dataset exhibited no missing or duplicate values, ensuring high-quality data suitable for thorough analysis.
Exploratory Data Analysis Results

### Univariate Analysis

Work Hours: The majority of individuals indicated an average workweek of approximately 40 hours, representing a standard workweek.
Work Class: Analysis revealed approximately 70% of the population were employed in the private sector.
Native Country: A substantial majority (94%) were North American natives, with a minor proportion (2.1%) identifying as Asian natives.

### Bivariate Analysis

In this section, we will compare and analyze the relationship between salary and variables such as Education, Work Class, Work Hour, and Age
Salary Relationships: Strong positive correlations were identified between higher education levels and elevated income levels. Approximately 70% of individuals possessing doctoral degrees earned above $50K annually. Similar income trends appeared among self-employed and federal government employees, highlighting educational attainment and job class as significant factors.

ISSN 2959-6157

### Data Preprocessing

Data preprocessing was crucial to ensure meaningful results. Notable steps included:
Capital Gain and Loss: These variables were removed due to their high proportion of zero values, not contributing meaningful variance to the analysis.
Outlier Treatment: Extreme values in variables such as age, "fnlwgt" (final weight), education years, and weekly working hours were systematically adjusted using box plot methodologies to maintain data consistency and reliability.

### Statistical Measures in Logistic Regression Analysis

The analysis emphasized several critical statistical measures vital for interpreting logistic regression results:
Standard Deviation (SD): This measure indicated variability in data, providing insights into the sensitivity of the model to changes in predictors.
Z-value: Representing the coefficient divided by its standard error, the Z-value tested the null hypothesis, with higher absolute values indicating more significant predictors.
P-value: Low values (<0.05) indicated statistical significance, affirming predictors' contributions to the model.
Coefficient: Quantified changes in log odds for the dependent variable per unit change in predictors, directly indicating predictor influence.(Hosmer et al.,2013)
Model Building and Performance

### Logistic Regression Model

The logistic regression model categorized income levels, with outcomes coded as "less than $50K" (1) or "$50K or more" (0). Critical attention was given to minimizing false negatives and positives, strategically optimizing the F1 score, which encapsulates a balance between precision and recall.

### Addressing Multicollinearity and Model Refinement

Utilizing the Variance Inflation Factor (VIF), we systematically identified and corrected multicollinearity among predictor variables. Iterative removal and assessment of problematic variables ensured reliability and interpretability of the logistic regression coefficients.(DeLange et al.,2018)
In the previous analysis, the data for the variables of working class and occupation were unknown wherever they were located. The VIF value also emphasizes the high correlation between these variables. When obtaining the same information from occupation and working class, they will be removed in this calculation. At the same time, Education-no_of_years and race_ White has high VIF values We will be dropping and one at a time, building separate models, and checking their performances to see which variable has a significant impact on the model's performance. After testing, discarding education_no_of_years and race_ White will not have a significant impact on the model's performance.
After filtering the data, we will still remove and evaluate variables with issues. We will build a model, check the p-values of the variables, and drop the column with the highest p-value.
Then, creating a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
Finally repeating the above two steps till there are no columns with p-value > 0.05, to ensure the reliability and interpretability of logistic regression coefficients.

### Model Performance Summary

Our logistic regression model achieved an impressive F1 score of 0.89, indicating strong predictive accuracy. Precision-recall curves and Receiver Operating Characteristic (ROC) analyses identified the optimal classification threshold. We selected three thresholds of 0.5, 0.58, and 0.76 respectively. After comparing, the default threshold (0.5) provided a superior balance, maximizing the F1 score and ensuring robust classification performance. (Humphrey et al.,2022)

## Conclusion and Recommendations

This study has been able to build a predictive model that can be used by the government to find the citizens having less than 50K salary with an f1_score of 0.89 on the training set and formulate policies accordingly.
Coefficient of some levels of education, workclass, and native country are positive an increase in these will lead to increase in chances of a person having <=50K salary.
Coefficient of age, fnlwgt, marital_status, working_hours_per_week,some levels of education, workclass, and native country are negative increase in these will lead to decrease in chances of a person having <=50K salary.
The government should promote education among citizens, they should make policies to make education accessible to all, as we say in our analysis that people who have higher education are more likely to have a salary above 50,000 Dollars.
Working hours is one of the significant predictors of salary, The government should implement laws to ensure that people are paid fairly for their work and are not overworked for the increase in salaries. This would improve

work-life balance.

Reforms should be made for private-sector employees so that they are paid fairly for their work.

Policy formulated by the government should be considerate of equal pay and counter the pay gap that exists in society.

Reflections and Improvements

*Lessons Learned*

This project underscored the importance of structured data analysis pipelines and logistic regression's robust applicability to social science classification challenges. Comprehensive EDA and thoughtful feature engineering substantially influence predictive model performance.

*Areas for Improvement*

Future improvements identified include:

Data Collection: Expanding dataset demographics to include broader and diversified social-economic variables, enabling more nuanced insights.

Model Complexity: Experimenting with ensemble techniques or more sophisticated machine learning models could potentially enhance accuracy and sensitivity to social-economic nuances.(Pedregosa et al. ,2011)

Post-Model Analysis: Advanced interpretability tools (e.g., SHAP, LIME) could significantly improve understanding of feature-specific influences on predictions, offering deeper insights for policy implications.(Lundberg & Lee, 2017)

## References

Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.
https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.Nick, T. G., & Campbell, K. M. (2007). Logistic regression. Topics in biostatistics, 273-301.

Long, J. T., Neogi, S., Caldwell, C. M., & DeLange, M. P. (2018). Variation inflation factor-based regression modeling of anthropometric measures and temporal-spatial performance: Modeling approach and implications for clinical utility. Clinical Biomechanics, 51, 51-57.

Humphrey, A., Kuberski, W., Bialek, J., Perrakis, N., Cools, W., Nuyttens, N., ... & Cunha, P. A. C. (2022). Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth. Monthly Notices of the Royal Astronomical Society: Letters, 517(1), L116-L120.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Lundberg, S.M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30.