Personal Total Income of Individual in the United States of America in 2023*

A Predictive Model Using the American Community Survey Data

Justin (Jiazhou) Bi and Weiyang Li

October 25, 2024

This project aims to build a predictive model for personal total income using various demographic attributes, such as marital status, occupation, and residing state. We explored three popular machine learning algorithms: Random Forest, Linear Regression, and Extreme Gradient Boosting. Ultimately, Extreme Gradient Boosting was selected due to its superior predictive performance. While the model achieved a high R-squared value, it also exhibited a high Mean Squared Error, indicating challenges in accurately predicting all data points. Future steps for improving the model may involve incorporating additional demographic features to enhance model accuracy and reduce errors. Additionally, fine-tuning hyperparameters, experimenting with feature engineering, and handling outliers could further improve performance.

1 Introduction

Income is a fundamental concept as it affects nearly every member of society, influencing access to resources, quality of life, and overall economic stability. At the same time, it is also a highly complex concept to define, as it can be interpreted differently depending on various perspectives, such as economic, social, legal, and tax purposes. Therefore, we must examine various demographical factors that may impact income to understand its concept better. Our project defines income as "all forms of net financial resources generated or lost from work, investments, or other activities." This project aims to predict personal income using data from the 2023 American Community Survey (ACS) (Ruggles et al. (2024)). The ACS provides detailed demographic information, including age, education, employment, marital

*Inspired and instructed by: https://github.com/RohanAlexander/marriage. All the project related files can be found at: https://github.com/Jiazhou-Bi/ACS-Income-Model/tree/main.

status, mortgage status, veteran status, etc… As a result, it provides a robust dataset for studying potential factors that may influence income.

Due to the timeframe limitations, the findings of this project are specifically applicable to the 2023 dataset. As such, the results may vary across different periods, and their generalizability has not been assessed within this project. The raw data used in this analysis was sourced from IPUMS using the IPUMS API. (IPUMS (2024)) We utilized pandas (McKinney (2010)) for data processing, and data visualization was carried out using Seaborn. (Waskom (2021)) Additionally, all tables presented in this report were generated with Plotly. (Inc. (2015))

Income inequality is a serious problem and can lead to significant consequences and it may affect social structure, economic stability, as well as policy-making processes. Researchers found that demographic factors, such as education level and occupation type, can significantly impact an individual's financial standing. (Autor (2013)) Similarly, social relationships, such as marital status and household characteristics, affect income dynamics. (Waite and Gallagher (2000)) Therefore, applying machine learning models to analyze the ACS dataset and using the information mentioned to predict personal income is reasonable.

We have chosen 2023 ACS dataseet because it offers many demographical details, as well as self-reported total income of the individuals. The sample is also representative and covers all geographical locations in the USA. Furthermore, it provides easy access to their database through API. We will employ advanced predictive models, including random forest, lineaer regression, and extreme gradient boosting machine to estimate personal income based on demographic factors to provide insights into which potential factors have the most significant impact on personal income. We hope the findings generated from our project can provide insights to policy-making process so that we can, hopefully, reduce income inequality and support further economic development in the future.

If we can successfully predict personal income using demographic information from the survey, this analysis will be able to contribute to a nuanced understanding of the American economy in 2023. It can help to reveal how demographic changess, educational attainment, and employment status/type can impact personal income. On the other hand, if the model cannot predict personal income with a high degree of confidence, this would suggest that the factors included in the analysiscannot not fully capture the intricacies of economic behavior and individual circumstances. As a result, additional variables, such as detailed education background like school of graduation, languages spoken, health status, etc, might be necessary in accurately predicting income.

2 Data

The dataset used here is downloaded from the Toronto Open Data website via (**TorontoOpenData?**). This dataset contains all the reported crimes that happened in Toronto from 2014 to 2023. This dataset is grouped by the year of the reported crime, its category and belonging subtype,

Table 1: Example of Cleaned Data

	STATEICP	GQ	OWNERSHP	MORTGAGE	SEX	AGE	MARST	EDUC	SCHLTYPE	OCC
1	41	1	1	3	2	51	6	10	1	800
2	41	1	1	3	1	61	1	8	1	4810
3	41	1	1	1	1	63	1	7	1	8030
4	41	1	1	1	2	36	4	7	1	120
5	41	1	1	1	1	17	6	5	3	4000

and the count of the subtype being reported and cleared for that year for each division. Because I am examining the crime pattern in the city, I have dropped the division information and aggregated the existing data according to their subtype and the year of the crime being reported. In the following subsections, I will review all the variables used in this report and provide some basic descriptive statistics. The first five rows of the cleaned data used for analysis are attached (Table 1).

2.1 Report Year

The report year variable is the number of crimes being reported. In this dataset, the data spans from 2014 to 2023, encompassing ten years. No month or date information was given; thus, there are only ten different values for this variable in chronological order.

2.2 Category

The category includes information about the nature of the crime. There are six crime categories: Crimes Against Property, Crimes Against the Person, Other Federal Statute Violations, Other Criminal Code Violations, Controlled Drugs and Substances Act, and Criminal Code Traffic. They are listed in the table below (**?@tbl-table2**).

2.3 Subtype

There exist multiple subtypes under each crime category. The following is an exhaustive table (**?@tbl-table3**) of all crimes' subtypes and their respective category.

2.4 Count

In the original table, this value is grouped by the subtype of the crime, the division, and the year when the crime was reported. The original count indicates the number of a specific subtype of crime reported within a particular division for the year. However, as mentioned before, because I am only interested in all the crimes in the City of Toronto, I have dropped the division information and aggregated the count from all the divisions to a single value. Therefore, for each subtype of the crimes, a total count of that subtype is reported in a single year.

2.5 Count_Cleared

These are the counts of crimes identified as cleared. In plain words, these are crimes that are dealt/solved. I have taken the same approach for this column as the previous one. After cleaning the data,f, or each subtype of the crimes, there is a total count of that subtype being reported that is also cleared in a single year.

2.6 Case_Clearing_Rate

This column was not included in the raw dataset but was created by dividing the cleared crimes by total crimes. A higher case-clearing rate for a particular subtype of crime usually suggests a higher effectiveness of law enforcement in dealing with this subtype of crime. The value is ranged from 0 to 100%.

3 Model

Our modeling strategy aims to explore the relationship between occupation, age, gender, education, and income. To achieve this, we implemented three models: Linear Regression with Interaction Terms, Random Forest, and XGBoost, each selected based on their specific strengths for predictive accuracy and interpretability.

3.1 Linear Regression Model with Interaction Term

3.2 Random Forest

3.3 XGBoost

4 Results

5 Discussion

5.1 Weaknesses and next steps

6 Appendix

6.1 Data cleaning

6.2 Model Details

References

Autor, David H. 2013. "The "Task Approach" to Labor Markets: An Overview." *Journal for Labor Market Research* 46 (3): 185–99. https://doi.org/10.1007/s12651-013-0125-7.

Inc., Plotly Technologies. 2015. "Collaborative Data Science." Montreal, QC: Plotly Technologies Inc. 2015. https://plot.ly.

IPUMS. 2024. "IPUMS API." IPUMS; Data retrieved via IPUMS API.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/ 10.25080/Majora-92bf1922-00a .

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0 [dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

Waite, Linda J., and Maggie Gallagher. 2000. *The Case for Marriage: Why Married People Are Happier, Healthier, and Better Off Financially*. New York, NY: Broadway Books.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.