AI-Powered Lifecycle Financial Planning – Predictive Modeling in Wage Income

October 2023

1 Introduction

1.1 Background

The goal in this project is to predict people's wage income in the future and help them make wiser decisions on their financial plans. For our project specifically, it is not realistic to require our users to fill in all the information needed, especially about some private and personal information, to predict their income. Considering this, we need to clean the data set we get at the very beginning, and do feature selection to find out the most important information and also improve the accuracy of predictions.

In the data cleaning part, our cleaning steps are mainly based on simple statistics. For the following feature selection section, we use some machine learning model to help us find out which predictors are important. We mainly focus on two different types of model, linear model and decision tree model. While selecting predictors in various models, we can also compare the performance of different models. The model selection and variables selection are not independent, but intersecting. In this project, we got decent results in Elastic-Net Method, Light Gradient Boost (LightGBM) and Neural Network model. Random Forest and Extra tree models are stopped because of the long time of tuning parameters. Additionally, we aim to find out whether linear model can be used because it is simpler and can be trained easily. Therefore we try Generalized linear model (GLM) in the last section. Generalized linear model with Gaussian distribution and backward elimination has been done, and more distribution family should be considered in the future.

1.2 Dataset Description

The data set we use is Public Use Microdata Sample (PUMS) data set, which comes from the sample of the responses to the American Community Survey (ACS) created by U.S. Census Bureau. PUMS files for an individual year contain data on approximately one percent of the United States population.

There are two types of PUMS files, one for Person records and one for Housing Unit records. We use the person records data, and each record in the Person file represents a single person.

In Table 1, it shows some samples and several columns:

PWGTP	AGEP	PUMA	SERIALNO	DEAR
13	85	800	2021GQ0000026	2
51	67	800	2021GQ0000031	1
17	74	1200	2021GQ0000063	2
61	16	1700	2021GQ0000067	2
15	83	500	2021GQ0000100	1

Table 1: Sample of PUMS data

PUMS data has 287 variables in total. We remove the last variables which is irrelevant to personal information:

- 131-206 column Indicates whether the variable mentioned before has been modified.
- 207-287 column is about person record-replicate weights.

Then, in order to facilitate the analysis, we group rest variables by their meanings in Table 2:

Income	Location	Migration	Race	Ancestry
Language	Marriage	Employment	Vehicle	Disability
Children	Insurance	Military	Education	Others

Table 2: Groups of variables in PUMS

The "Others" group includes the rest several variables, like age, sex and quarter of birth.

In the "Income" group, we have many types of income, and we show some in the following:

- **INT**: Interest, dividends, and net rental income past 12 months.
- **SEMP**: Self-employment income past 12 months.
- **WAGP**: Wages or salary income past 12 months.
- **PERNP**: Total person's earnings. (= WAGP + SEMP)
- **PINCP**: Total person's income.

Considering that we only want to predict the wage income in our group and other income like investment and pension will be calculated by other group, we choose total person's earnings ("PERNP") to be our response. Some people are self-employed and we need to include their income ("SEMP") in their future.

1.3 Data Preprocessing

During our preprocessing of the data, we use two techniques:

- **Preliminary screening based on the missing rates and the level of detail of variables.**

  First, we remove the variables with high missing rates and too detailed variables.

  For example, the missing rate of CITWP (which means the year of naturalization write-in) is over 0.93. It is not possible to fill the missing values, and even if we do that, too many guess values make this variable meaningless.

  For RAC2P (which is recoded detailed race code), there are 68 races which is not necessary, and we have other brief variable about race. So, we can drop it.

- **Group the levels of some variables.**

  Other than variables like RAC2P, there are also some too detailed variables which is logically important and cannot be replaced by other variables. But users are not willing to find their choice through too many options.

  For example, we group the levels of INDP (which means industry recode). There are over 500 types of industry. We only keep the field of industry based on the first 3 letters to reduce the number of levels.

  In Table 3, some levels of INDP are shown.

Code	Meaning of code
0270	AGR-Logging
0280	AGR-Fishing, Hunting And Trapping
0290	AGR-Support Activities For Agriculture And Forestry
0370	EXT-Oil And Gas Extraction
0380	EXT-Coal Mining
0390	EXT-Metal Ore Mining
0470	EXT-Nonmetallic Mineral Mining And Quarrying
0490	EXT-Support Activities For Mining
0570	UTL-Electric Power Generation, Transmission And Distribution
0580	UTL-Natural Gas Distribution
0590	UTL-Electric And Gas, And Other Combinations

Table 3: Some levels of INDP

- **Similar variables are removed.**

  In the same group of variables in Table 2, we have variables with similar meanings. In the following example, we keep the second one and remove the first. For example:

  POWPUMA: Place of work PUMA based on 2010 Census definition.

  POWSP: Place of work - State or foreign country recode.

- **Deal with missing data.**

  Most missing values have specific meaning. Take "ENG" as an example. The meanings of "ENG" and its levels are shown in Table 4. The missing value of "ENG" means the person is less than 5 years old or he speaks only English. So, we can use "b" to represent this part of people and fill in the missing value.

ENG	Ability to speak English
N/A	less than 5 years old/speaks only English
1	Very well
2	Well
3	Not well
4	Not at all

Table 4: Meanings of "ENG" and its levels

- **Use one hot encoding to deal with category variables.**

  One hot encoding are used to change category variables into dummy variables. There are only 4 numeric variables: AGEP, JWMNP, WKHP, PERNP. All the others should be transformed.

  An example of "FER" (which indicates whether the person gave birth to child within the past 12 months) are shown in Table 5.

Original FER		**After one hot encoding**		
b	less than 15 years/greater than 50 years/ male	FER b	FER 1	FER 2
1	Yes	1	0	0
2	No	0	1	0
		0	0	1

Table 5: : One hot encoding for "FER"

2 EDA

2.1 Target Variable

For PUMS Data, our target value is **PERNP** (Total person's earnings.) Here is a distribution of the target variable.

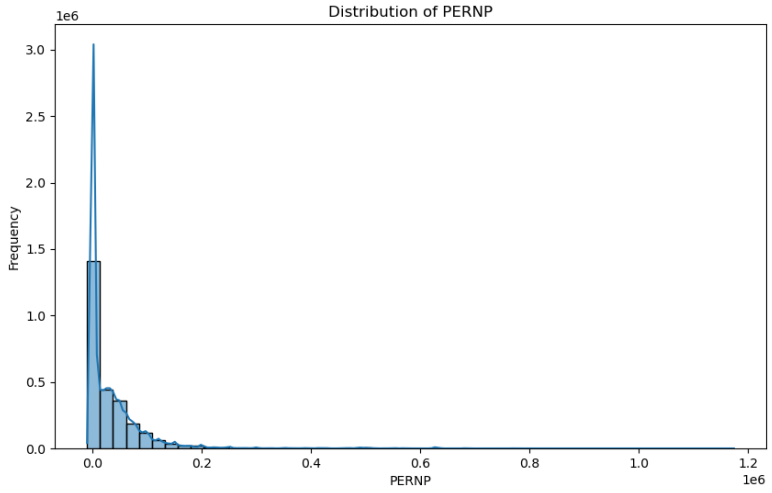


Figure 1: Distribution of PERNP

The plot has an extreme long right tail with very big value. Therefore, we consider to use log transform to fix this. As there are many 0 values, we use $\log(x+1)$ here.

The new distribution plot is shown as below.

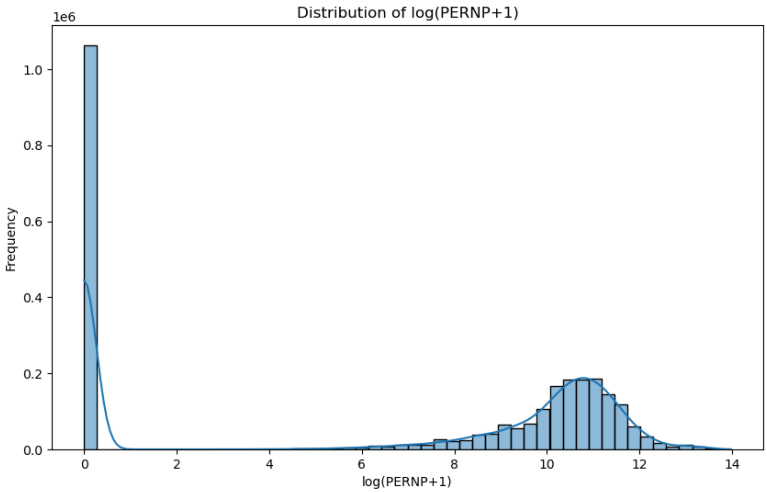


Figure 2: Distribution of log(PERNP+1)

There are still a large amount of 0 values as we are doing $\log(0+1)$, but the rest data distribution seems normal like.

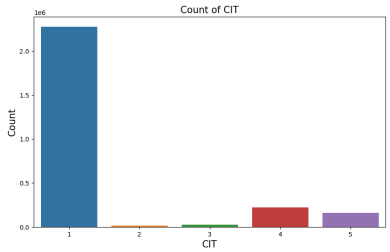
2.2 Feature Variable

In order to have a clear look about the relationship between target variable and feature variables, we pick several examples from each group and split them between categorical and numerical variables.

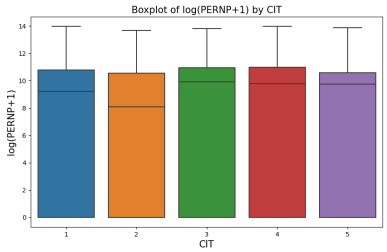
2.2.1 Categorical Variable

CIT

We get the distribution of "CIT" variable and the boxplot of log(PERNP+1) by "CIT".



(a) Count of CIT

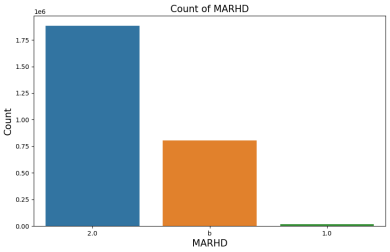


(b) Boxplot of log(PERNP+1) by CIT

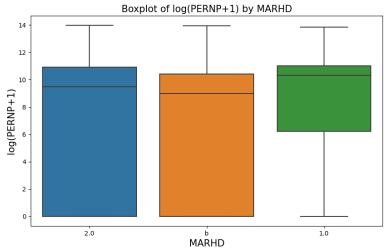
Figure 3: CIT plots

MARHD

We get the distribution of "MARHD" variable and the boxplot of log(PERNP+1) by "MARHD".



(a) Count of MARHD

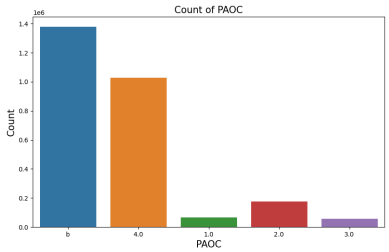


(b) Boxplot of log(PERNP+1) by MARHD

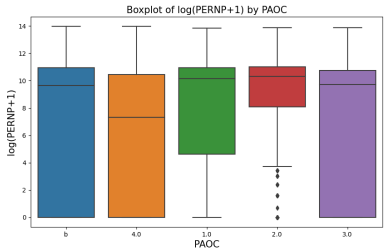
Figure 4: MARHD plots

PAOC

We get the distribution of "PAOC" variable and the boxplot of log(PERNP+1) by "PAOC".



(a) Count of PAOC



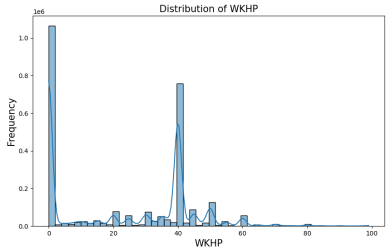
(b) Boxplot of log(PERNP+1) by PAOC

Figure 5: PAOC plots

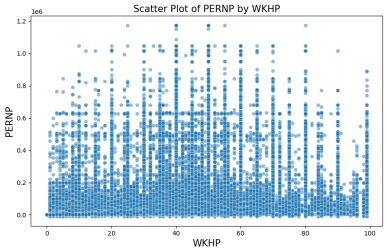
2.2.2 Numerical Variable

WKHP

We get the distribution of "WKHP" variable and the scatter Plot of "PERNP" by "WKHP".



(a) Distribution of WKHP

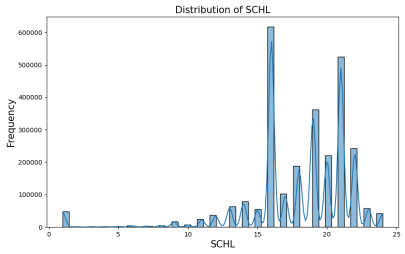


(b) Scatter Plot of PERNP by WKHP

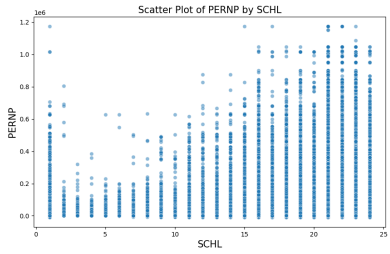
Figure 6: WKHP plots

SCHL

We get the distribution of "SCHL" variable and the scatter Plot of "PERNP" by "SCHL".



(a) Distribution of SCHL

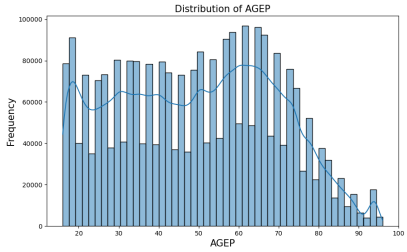


(b) Scatter Plot of PERNP by SCHL

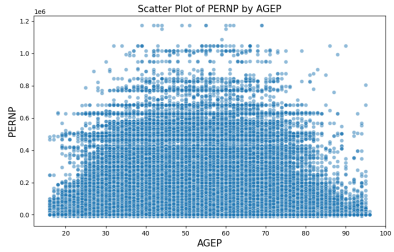
Figure 7: SCHL plots

AGEP

We get the distribution of "AGEP" variable and the scatter Plot of "PERNP" by "AGEP".



(a) Distribution of AGEP



(b) Scatter Plot of PERNP by AGEP

Figure 8: AGEP plots

2.3 Heat Map

We also create a heat map to check the correlation between each numerical variables.

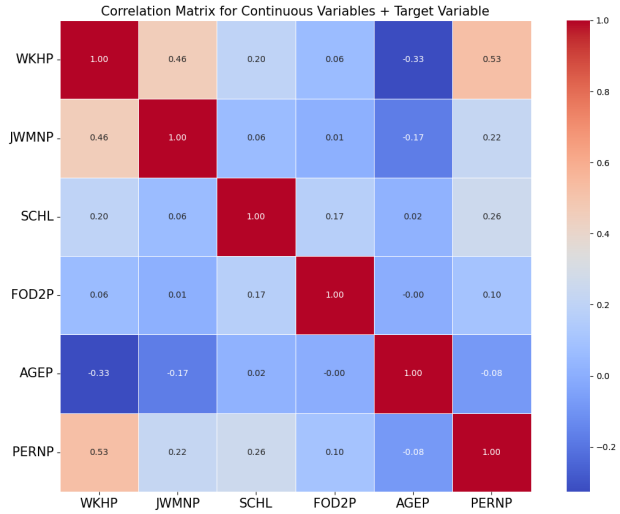


Figure 9: Heat map for Numerical Variables