Xin Tang
6/14/2024
Predictive Analytics
DSC630-T302 (2247-1)

Income level prediction

for credit card/loan application invitation

Project Introduction:

In United States, Income provides economic resources that shape choices about housing, education, childcare, food, medical care, and more. For a person or a family, it plays a crucial role in meeting basic needs such as education, shelter, water and sanitation. From enterprises' angles, It is also imperative for companies to know this information to target the most suitable groups and mitigate the financial risk in situations like credit card or loan application, even housing aid application.

For various reasons, consumers are often unable or unwilling to provide their true income information, which makes it difficult to trust the self-reported income data.

However, in the big data era, there are other ways to reveal the truth through other more indirect objective information, like education level and marriage status. With this objective in mind, I would like to develop a predictive model to group the people into high- and low-income levels.

The result from this model will provide a cross-check point on self-reported data, as well as an indication of what consumer background will yield high- or low-income level. From there, the interested companies will target correct consumer groups and ensure a better yield of investment.

Dataset selection and approach:

Dataset:

The data resource for this project is from Kaggle, which is an open-source data science community with many data suitable for data analysis. This dataset has 11 independent variables like age, work class, education and marital status etc., which are all objective information. It contains almost 32 thousand individual's records, so is large and complex enough to create a predictive model.

As a backup, I picked up a world population dataset, which is also interesting and impacts us on a great scale. As a second backup, I have an insurance dataset I am interested in exploring, if the first 2 do not work out.

For the income dataset, I would like to:

1. Build a model can predict the level of income (higher or lower than >$50K, which is the criteria used in dataset).
2. Find out which factor is the most important to influence the income level.

All datasets can be found in Kaggle. Refer to citation sections for details:

Approach:

What types of model or models do you plan to use and why?

I peeked into the dataset, they have mixed numerical and categorical data. But not a text-based dataset, so At least I do not plan to do sentiment analysis or tf-idf matrix.

This is a classification problem; the outcome will be with a binary classification label. For a classification problem, First I like to try random forest classifier, (IBM) which is an effective tool for estimating missing values as it maintains accuracy when a portion of the data is missing, also will help to avoid overfitting. I will also try to do hyper parameter tuning to see how much it could improve the performance. Next, I will try XGboost. (Nvidia)XGBoost is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

How do you plan to evaluate your results?

(Wohlwend, 2023), Evaluation plays an indispensable role in the process of building robust and effective classification models. It will help to understand model performance, make a objective model selection, as well as measuring business impact.

I will run classification report to get accuracy, precision, recall and F1 score for the 2 models. I will also run confusion matrix to visualize the result.

What do you hope to learn?

Developing a model and using it to predict is a systematic project. It involves many steps. Like those outlined in CRISP-DM method. I am interested in learning how to improve the performance of the model. How to assess the evaluation result from end user aspect, since sometime the results from one index may not equally important, or not even relevant.

Assess any risks or ethical concerns with your proposal.

Risk1: imbalanced data. → in income data, it is common to have majority of data belongs to low income, high income is a relatively small portion. This can be addressed using proper packages specific to handle this issue.

Risk2: The factors impact income level changes over time. For example, A college graduate majored in art will have dramatic income difference from one majored in computer science. Most census data do not provide college major details.

Risk 3: Ethical concerns: income will play a critical role in determining eligibility of many things, like student loan, social benefit, tax, credit limit for consumption and many more. An unvalidated data resource or incorrect model will bring serious consequences. To develop a model ready for deployment, the dataset it depends on needs to be carefully validated. The final model selected also needs to be tested heavily.

Identify a contingency plan if your original project plan does not work out.

To avoid the last-minute start over and working on an alternative dataset, I would spend a lot of time on EDA, visualize the data distribution, check correlations, decide weight etc.

To avoid running into technical roadblock, I plan to use experience from others, like professor and some classmate who working on the data science field already.

Last, Since I will travel a lot and not everywhere I can access database in US. I will keep a local copy of everything on my computer, if needed, I will email professor some evidence of work and ask for an extension.

Include anything else you believe is important.

Following the timeline is the key to the on time finish the project. Preparing in advance for answers to the non-technical questions (like model's application, ROI etc) will also help to show the real value of the project.

Milestone 3 addition

Will I be able to answer the questions I want to answer with the data I have?

My dataset has 11 independent variables and 1 dependent variable. through preliminary EDA. I found out:

1. The data has no empty NA fields. However, some variables have "?", or no meaning answers, which are equivalent to NA and need to be handled. At the same time, the ratio of NA or meaningless data are low, so data is still good to use.
2. The data has mixed numerical and categorical variables, which covers many important aspects of personal information, like work, education, marital status etc.

Based on this, I believe the dataset is clean and has enough information to build the model, as well as discover the impact of each variable to the dependent variable, which is income level.

What visualizations are especially useful for explaining my data?

I mainly use bar charts, box plot and bivariate plot for my visualizations.

- A bar chart is powerful to reveal the distribution of the variables, I used it to have an overall view of every variable.

- From the chart on numerical variables, looks like the dataset has a good representation of the population. People aged between 20-50 are the main force of earning income, also most working people have some college and post college education (education length 10 -14).

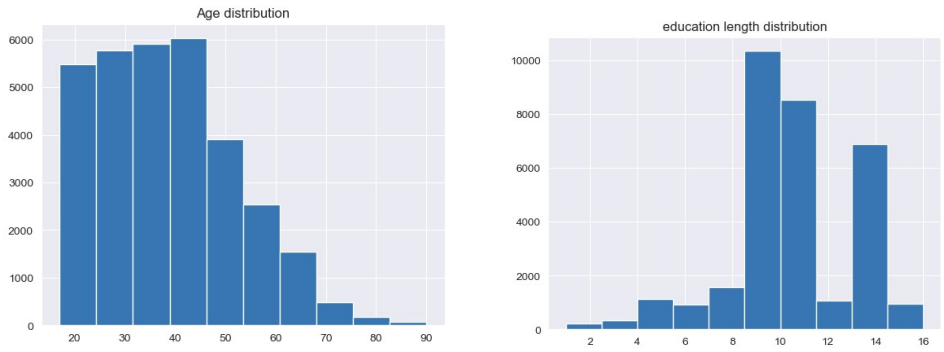

Chart 1. Numerical variables (Age and education length) distribution

- From the charts on categorical variables, one thing noticeable is low income people is 3 times of high income, which is imbalanced.

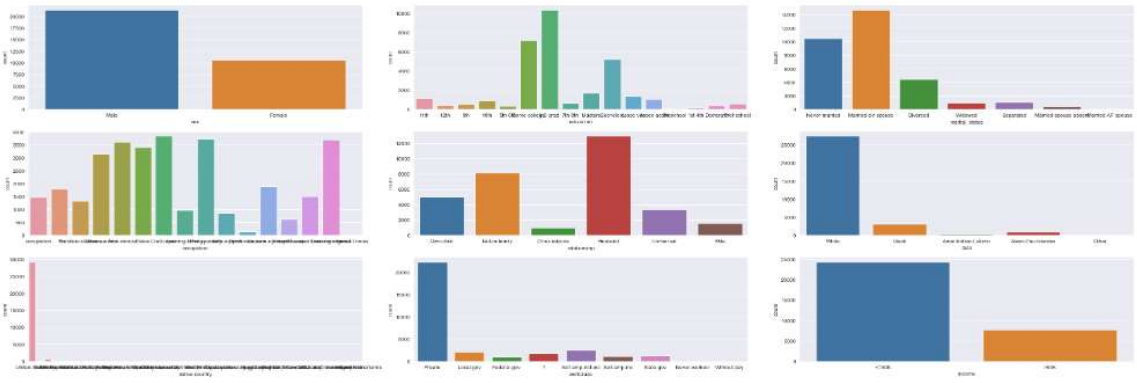

Chart 2: distribution of categorical variables

- A box plot will discover any outliner for numerical variables. I will try on one variable: “fnlwgt”. However, this variable seems not relevant to income, so the outliner is not important. Similarly, most variables are not in normal distribution

and outliner may not mean abnormal condition, so I did not use box plot too much.

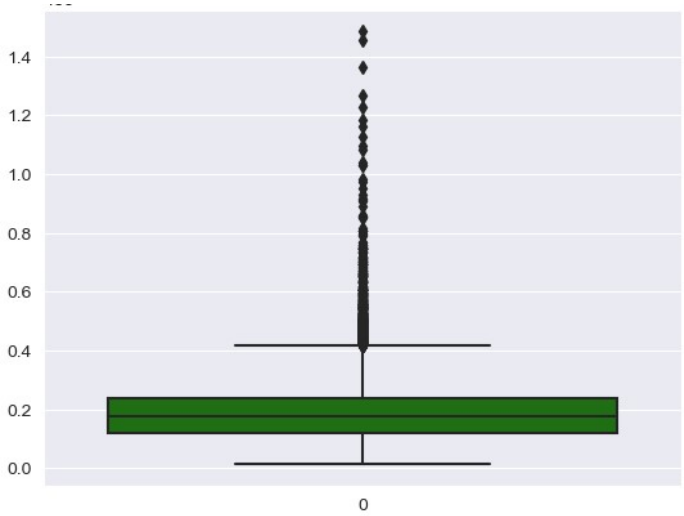

Chart 3. Final weight distribution

- A bivariate plot will quickly expose the relationship between independent variables and dependent variables. Some interesting information could be revealed.

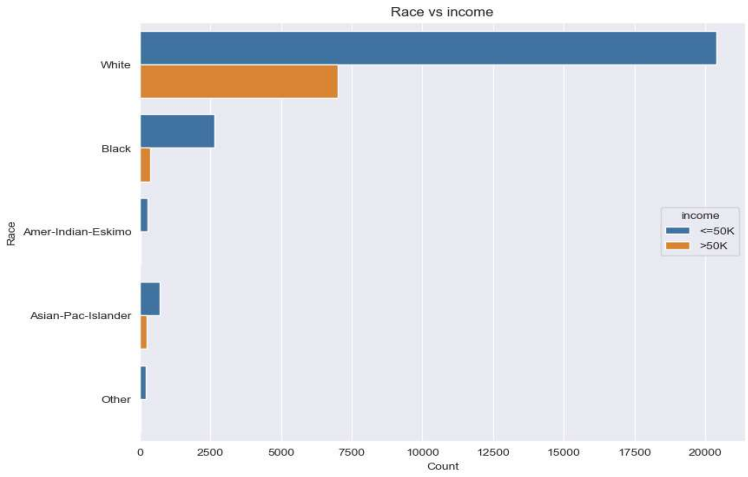

Chart 4. Race vs income (it did not reveal much info, just shows most sample (or the biggest race of population) are white

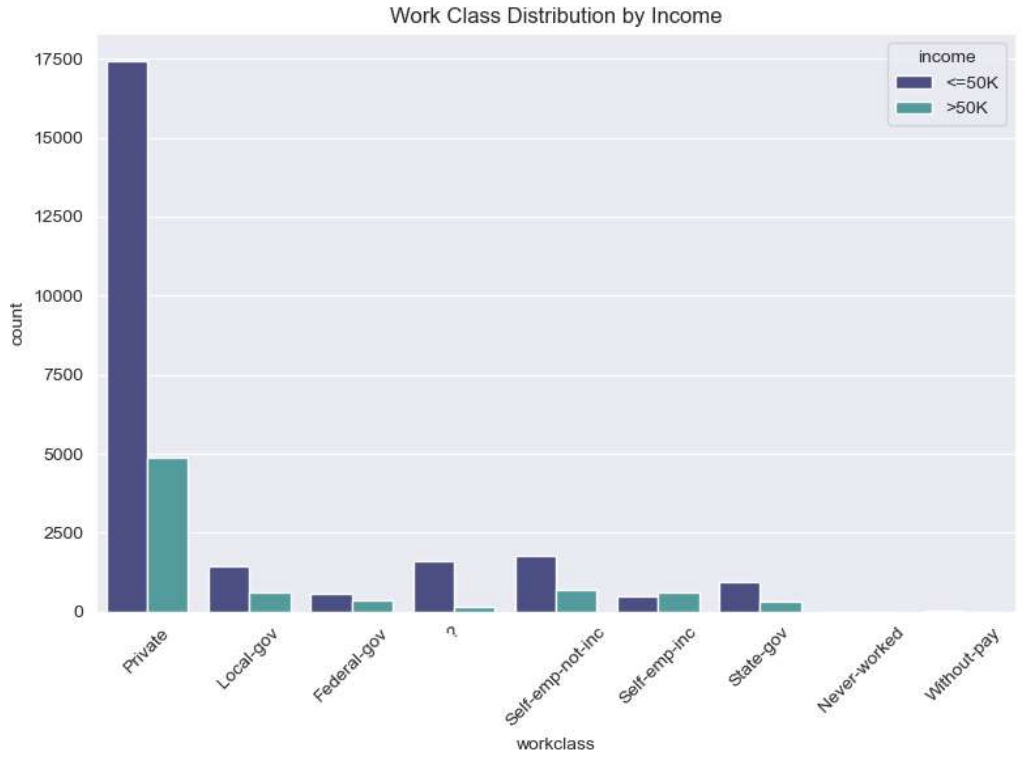

Chart 5. Work class vs income (private sector have biggest income variations)

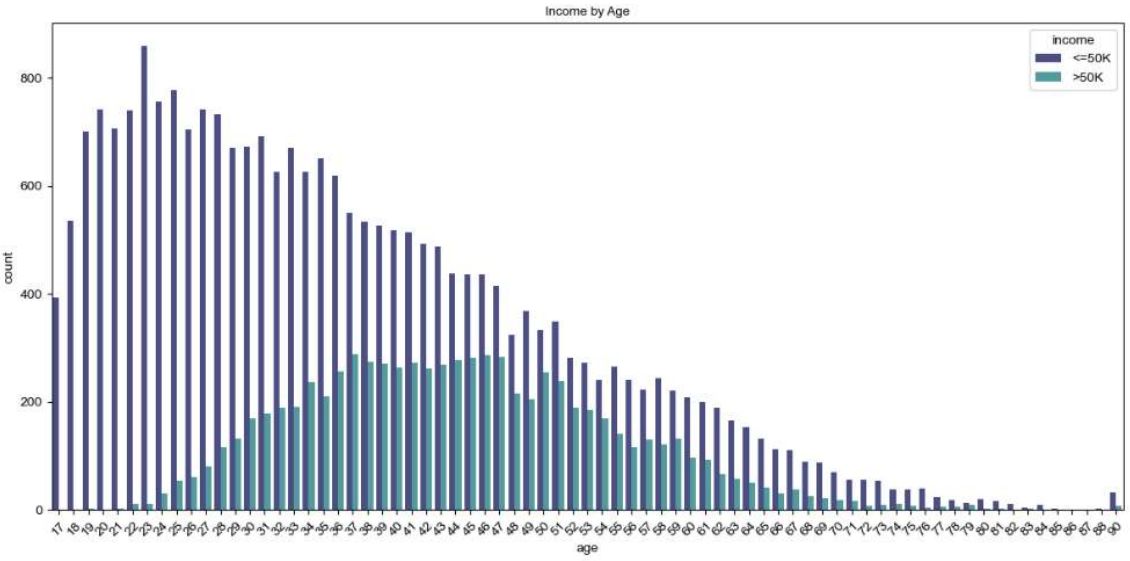

Chart 6. Age vs income (middle age 35-50 years old are most likely have more high income)

- A pie chart could be a nice alternative to a bar chart. I used it on variable: ‘race’.

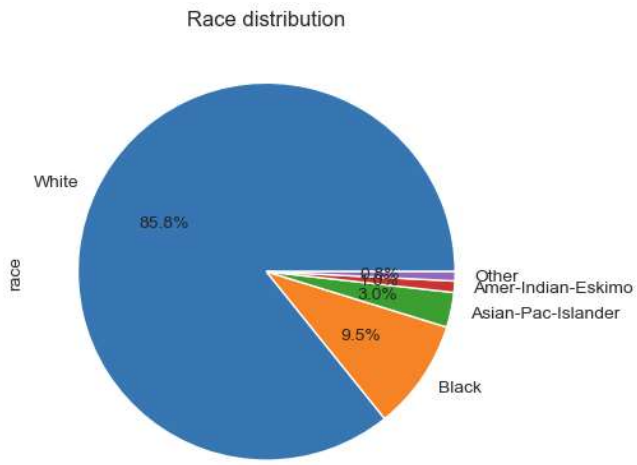

Chart 7. Race distribution (this matches the race distribution of the population back in 1994, white is dominating.)

Do I need to adjust the data and/or driving questions?

I believe some adjustments are needed to prepare data for model build.

- Even there is no direct NA in the dataset, data has NA equivalent data point (like ‘?’ or occupation for occupation), so need to get them handled. I will check their percentage in whole data and decide if they can be removed.

- One variable, ‘fngwgt’ does not have too much influence on income. After visualization and understanding its meaning, I will decide what to do with it.

- The variable, ‘education level’ pretty much means the same thing as another variable, ‘education number (length of education year)’. More education length in

years always means the higher education level. So, I plan to remove the education level and keep education length only.

- I also plan to encode all categorical variables into some sort of numerical variables to develop the models.
- The income groups are not balanced, there are always way more low-income count than high-income count. This will cause some trouble for the model, so I will plan need to do some data balancing to ensure similar count of samples pulled from each income group be used to train the model.

Do I need to adjust my model/evaluation choices?

I originally proposed to use random forest and XGBooster model. So far, I think they are still good choices for classification problem I am trying to resolve.

Random forest normally needs to get parameters tuned to get better performance possible. I will use hyper parameter tuning to see how much I can improve the performance.

Since I am developing more than one model, the results need to be compared to pick the best choice. There are several ways to evaluate classification models (Wohlwend,2023). I will check the model accuracy, the precision, recall etc. I will pay attention to how much computing resource each model will take.

Last, to help answer the question that which variables are more important to influence the income level, I plan to use feature importances function from Random Forest Classifier to get the answer.

Are my original expectations still reasonable?

I would think my original expectations (develop a model, find which factors are important to income level) are still in correct path to get the business case success. So, I plan to stick to the original plan and try to focus on this dataset only.

Milestone 4 addition

Explain your process for prepping the data:

- The first step is clean the data. I converted all meaningless answers to NA. Then I calculated the proportion of NA to the whole data. Since it is less than 1%, I removed the NA.
- Second, I removed the variable: 'fngwgt', since it does not have any relationship to income.
- Third, I encoded all categorical variables into dummy numerical variables to develop the models. I used mean encoding on work class. For the rest, I just turned them into dummy variables.
- Last, before splitting the data into train/test set, I sued SMOTE to keep sampling from each income level balanced.

Build and evaluate the model:

- First, I tried the random forest classifier. Then I used hyper parameter tuning to get the optimized parameters.

- Second round, I tried the XGBooster, I am curious to see if it will get better performance.

Below is a summary of the evaluation:

	Random Forest Classifier	RF with tuning	XGB classifier
Accuracy	86%	86%	86%
F-1 score for income >$5K	0.86	0.87	0.86
Precision for income >$5K	0.84	0.84	0.83
Recall for income >$5K	0.89	0.89	0.89

From the above chart. I concluded the 3 models have similar performance, out of the 3, RF model with optimized parameters performs best.

Results and Recommendations:

Using Predictive modeling technics, I built a Random Forest classification model, which will be able to take some personal information, like age, education length, work class etc., then predict if the individual's income is above or below $5k (equivalent to $10K now). The model has 86% accuracy, meanwhile, 89% of the predicted high-income individuals are truly high-income individuals.

This could be used for financial institutes, like banks, credit card issuers and mortgage companies to predict customers' income, or cross check their self-reported income.

I would recommend getting more data from financial institutes, or any potential user, to evaluate the model.

Also, since this model is based on 1994 data, I would recommend working with financial experts to assess if more variables or factors, which are new in today's situation, need to be included in this model.

I also recommend doing a legal and ethical review to ensure the modelling and its application complies with all legal regulation and properly using customer's personal information.

Resources:

Income dataset:

ANAGHA K, P (Oct. 2023). Adult Income Census, Predict whether income exceeds $50K/yr based on census data, Retrieved June 10, 2024 from https://www.kaggle.com/datasets/anaghakp/adult-income-census/data

World population dataset:

SOURAV BANERJEE (June. 2022). World Population Dataset, Global Headcount: World Population Dataset by Country/Territory, Retrieved June 10, from

https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset

Insurance dataset:

KUSH SHAH (June. 2020). The Insurance Company (TIC) Benchmark. Retrieved June 10, from https://www.kaggle.com/datasets/kushshah95/the-insurance-company-tic-benchmark

Nvidia *(date unknown)*: *XGBoost – What Is It and Why Does It Matter?*

https://www.nvidia.com/en-us/glossary/xgboost/

IBM (*date unknown*): *What is random forest?*

https://www.ibm.com/topics/random-forest#:~:text=Key%20Benefits&text=Feature%20bagging%20also%20makes%20the,or%20contribution%2C%20to%20the%20model.

Brandon Wohlwend *(Ju16,2023): Evaluating classification models*

https://medium.com/@brandon93.w/machine-learning-evaluating-classification-models-18713af3d764