Xin Tang
August/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 capable to group the people into high- and low-income levels, using indirect information.

The result from this model will also 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.

Using real life data, if this model can predict the income level with sufficient accuracy, I believe credit card companies or mortgage companies will be very interested in adopting and implementing it into their current client qualification process.

Dataset selection

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.

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.

Refer to citation sections for the source of the dataset.

Methods/Results

Methods:

EDA:

My dataset has 11 independent variables and 1 dependent variable. through preliminary analysis, 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.

1. 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.

Visualization:

I mainly used 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).

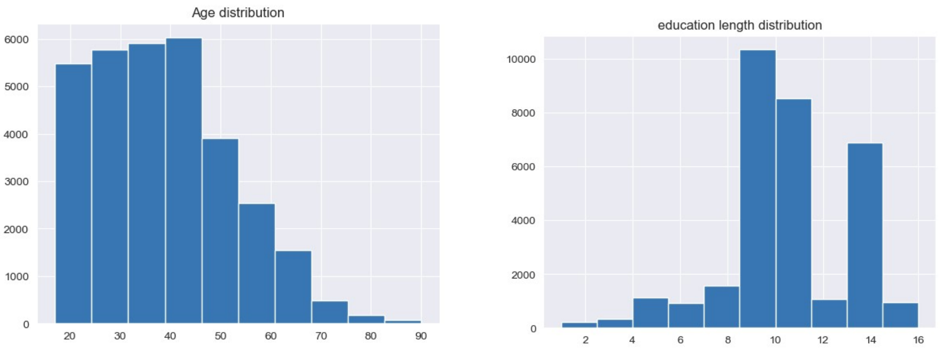


Chart 1. Numerical variables (Age and education length) distribution

- From the charts on categorical variables, one thing noticeable is the data points on low-income people is 3 times as datapoints on high income, which means the data is imbalanced.

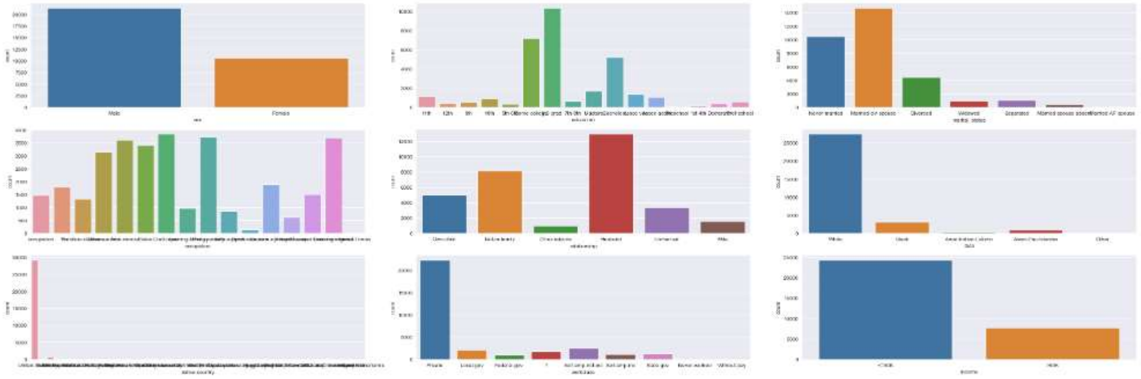


Chart 2: distribution of categorical variables

- 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.

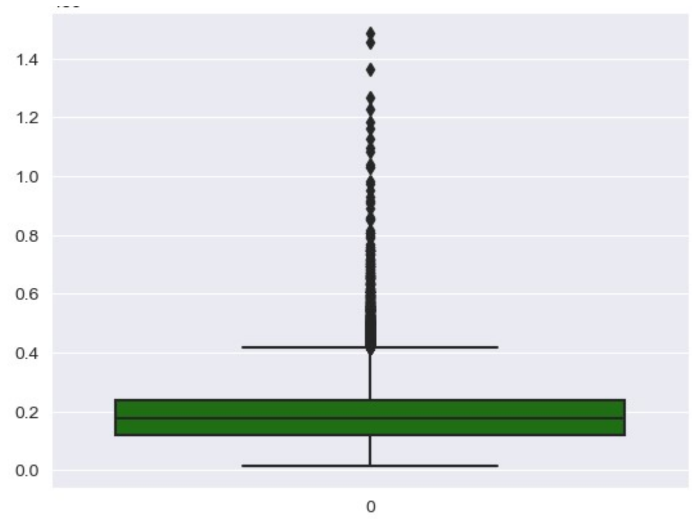


Chart 3. Final weight distribution

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

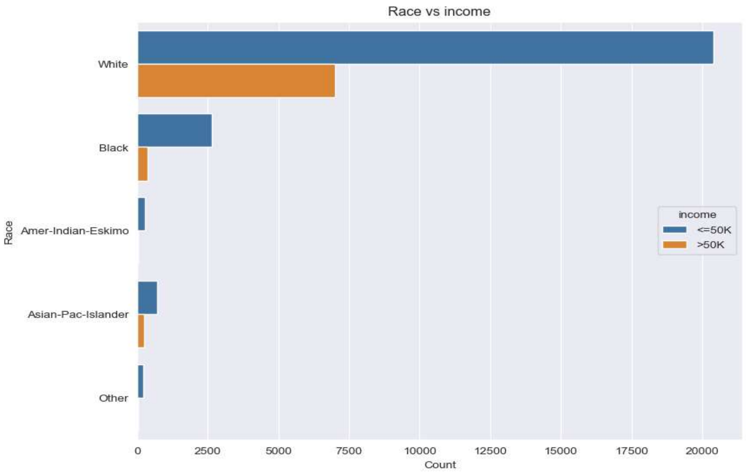


Chart 4. Race vs income (it did not reveal much info, just shows most sample (or the biggest race of population) are white

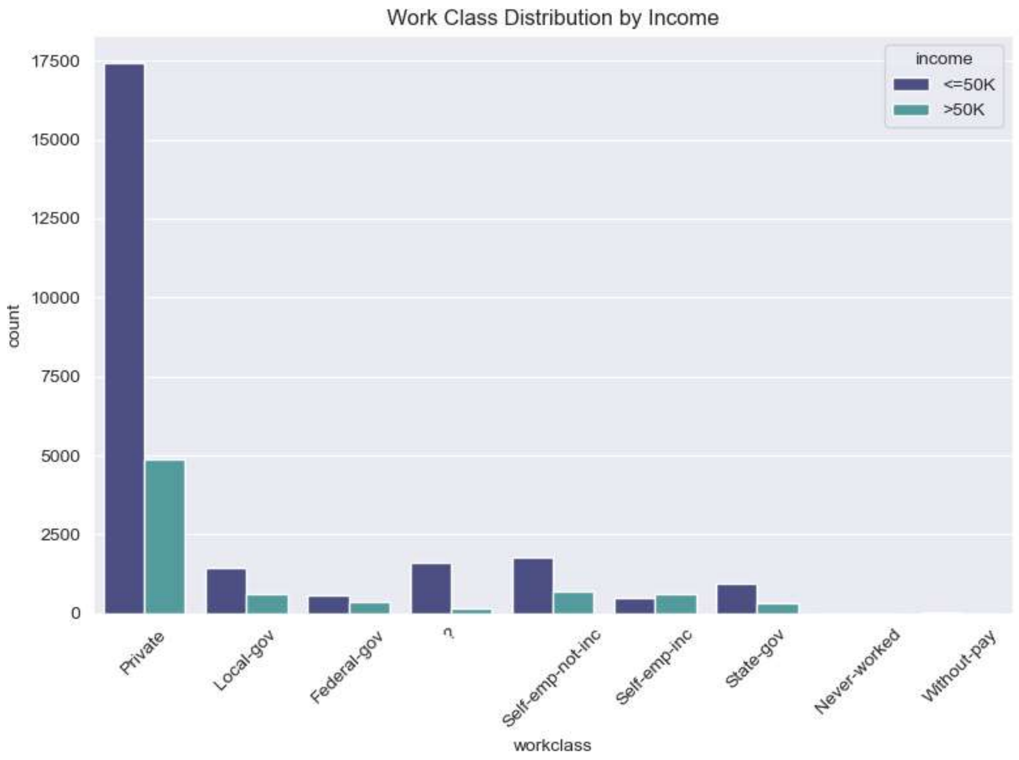


Chart 5. Work class vs income (private sector have biggest income variations)

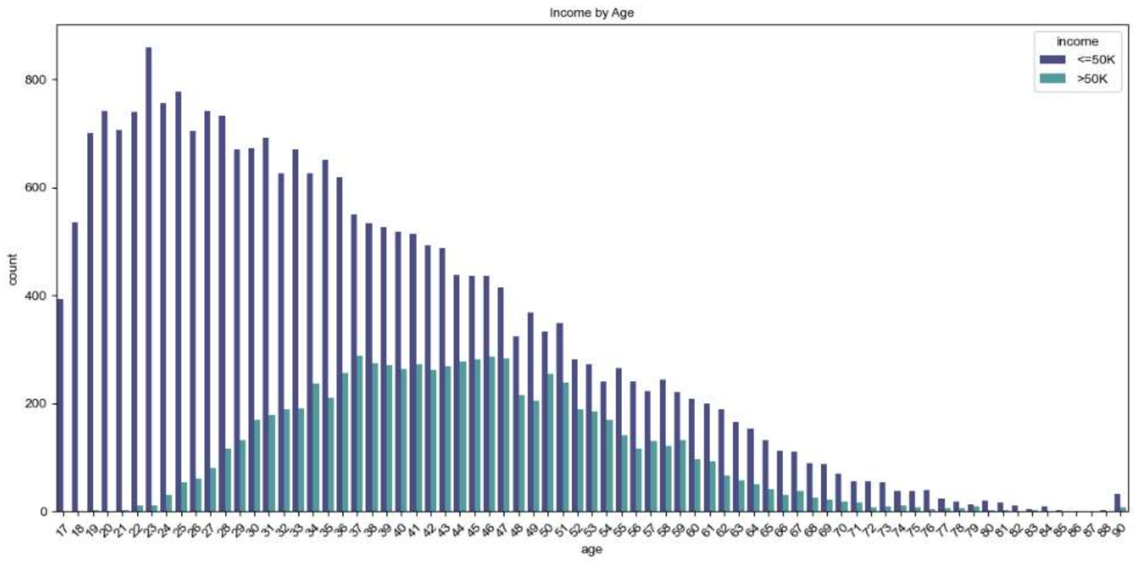


Chart 6. Age vs income (middle age 35-50 years old are most likely have more high income)

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

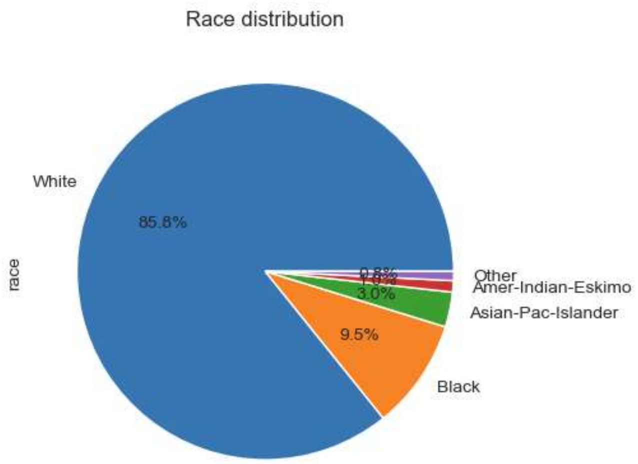


Chart 7. Race distribution (this matches the race distribution of the population back in 1994, white is dominating.)

<u>Prepare the data for modeling</u>:

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 the 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 a 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.

Modeling algorithm selection:

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.

Modeling Evaluation method:

(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 an objective model selection, as well as measuring business impact.

For classification problem, the Wohlwend (Wohlwend, 2023) article has detailed explanation of each term and what the results will indicate.

Below is a quick summary:

- Accuracy: Ratio of correctly predicted observation to total observation. Not work well since data is imbalanced.

- Precision: Proportion of positive predictions which are correct.

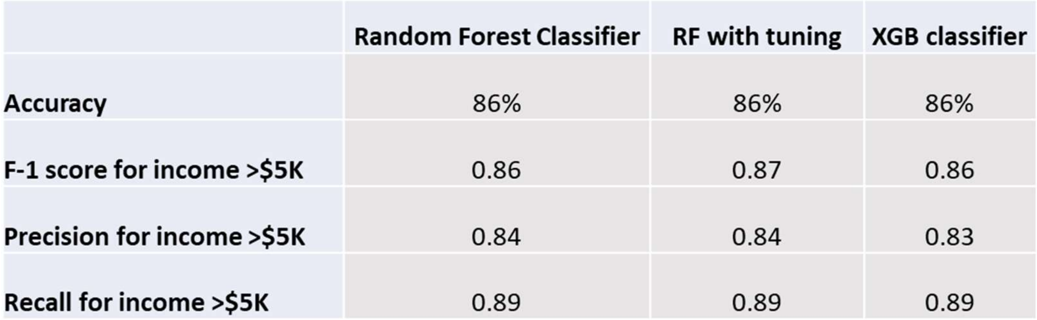
- Recall: Proportion of actual positives that are classified correctly.

- F-1 Score: a harmonic mean of precision and recall.

I will run classification report to get accuracy, precision, recall and F1 score for the 3 models. I will also run confusion matrix to visualize the result, which gives a visual of how much the prediction matches the actual data.

<u>Results</u>:

After model development and evaluation, below is the performance comparison of the 3 models:



	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

Chart 8. Model comparison

Therefore, I would recommend the random forest classifier with tuned parameters.

I also checked the feature importance on features included in the dataset, The result showed that Age, Education and Marriage status are the top 3 factors determining income level. Please refer to the chart below.

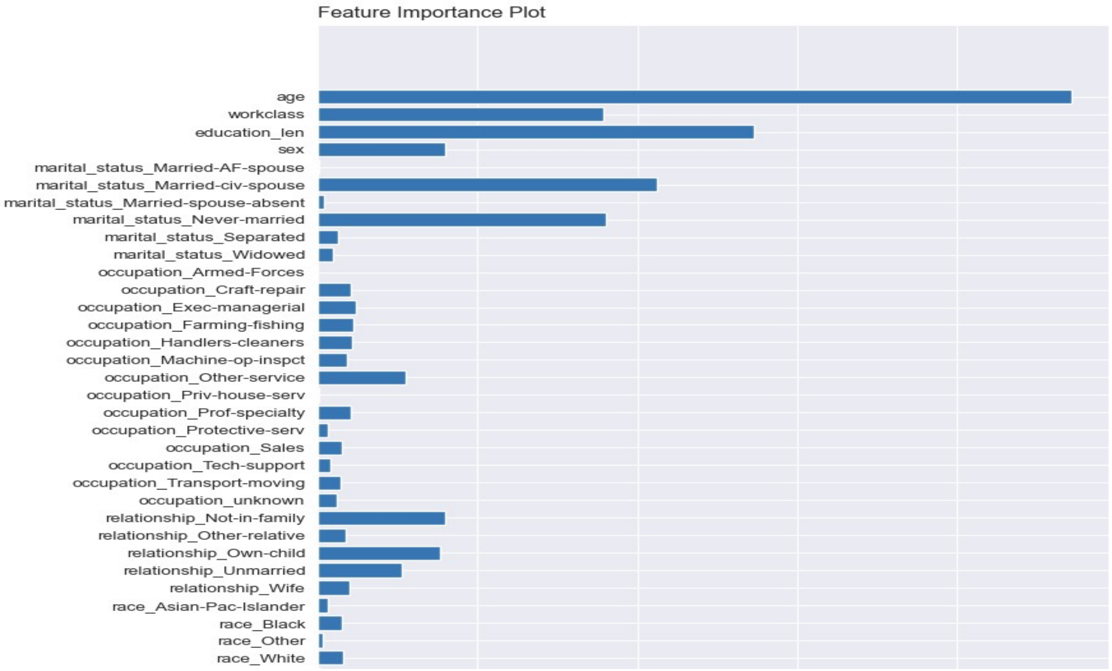


Chart 9. Feature importance

Conclusion:

Learnings:

Developing a model and using it to predict is a systematic project. It involves many steps. Like those outlined in CRISP-DM method. There is no single universal method to follow and needs to be adjusted case by case, e.g.: how to select and reduce feature dimensions, how to handle imbalance data, how to improve the performance of the model. how to assess the evaluation result from end user aspect. Also, the result from evaluation also needs be examined closely and take the end business goal into consideration, for example, does user care more about high-income or low-income group.

Recommendation:

I would recommend the random forest classifier with tuned parameters.

Deployment and next step:

The model development and evaluation so far are purely from a technical aspect. However, since the impact of the result will have a big impact on business decisions, I would present the result to client facing group and management team to do further review, like ROI (return of investment), ease of data collection etc.

Also, to evaluate model performance better, I would suggest getting model tested using much larger real-life data.

Besides, an assessment of hardware requirements is also suggested since it may need more computing resources if deployed.

Another concern is the age of the data. This model is based on 1994 data, it will be no surprising some new factors surfaced recently also have high impacts to the income level. I would recommend working with financial experts to assess if more variables or factors, which are new in today’s situation, are needed to be included in this model.

Ethical concerns and mitigation plan:

Income plays a critical role in determining eligibility of many things, like student loans, social benefits, 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.

Top management cares more on the business impact than how the model was built. To Gain buy in from management team, I need to invite subject experts to do business analysis and prepare the questions may be raised during the presentation.

Before the model goes live, legal and data security teams may also need to be involved to make sure every step, like data collection, the decision-making process etc., meet appliable regulation and law.

Coding

The model was developed using Python, some of the packages used are panda, numpy, sklearn, imblearn, xgboost, the seaborn and matplotlib package are used for visualization.

Refer to attached file for coding details.

DSC630-Xin-Tang-Final-code.pdf

Reference:

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

***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