Income modeling using Data Science machine learning techniques

Xin Tang

Bellevue University

INTRODUCTION

Background and Business case:

As an adult who is supporting college age children and does not have any financial support from anyone else, the amount of income is always a top weighted topic for my family. This is also a topic whoever has a steady income and wants to know the growth potential.

Everyone wants to have a higher income. However, besides working hard, what factors may impact your income level and how much the impact is? Does race matter? does education matter? How about the business factor you are in and your age? With the data available out there and various data mining technics, I believe I can use what I am learning from this class to conduct an entry level analysis. Besides, a model could be built to analysis any updated data.

On a broader level, income level from different group and background is a useful information to help decision making on many other fields, like consumer spending, education spending, national GDP growth etc. so any models can successfully predict the income level and discover the most influential factors are valuable business tools, and, will help any business to successfully deal with general population.

Data resource:

After a few rounds of search, I picked a dataset from Kaggle (https://www.kaggle.com/datasets/anaghakp/adult-income-census/data). This dataset is with income information and suitable for data mining/machine learning.

This dataset originates from the 1994 Census Bureau database with information of adult (human being aged >16) income. The income is categorized into 2 groups ( >50K or <50K ). It also contains 11 independent variables like age, work class, education and marital status etc. The only variable that needs to explain is fnlwgt: Final weight. This is an estimation of the number of people each observation in the dataset represents in the population.

The problem I want to address is to build a model to predict income category using available independent variables in the dataset, specifically, to predict if the income will be above or below 50K US dollar. Even though the data is outdated and $50K is now just an average income level for most people, I hope the model created can work on the newer data.

SUMMARY OF MILESTONE 1-3

The analysis and model building follows 3 major steps:

1. EDA, including visualization to show the relationship of the variables.

EDA is an important step to shine the light into the variables inside the dataset: the distribution of the variables; the relationship/correlation between the variables; which variable maybe more important than others; what kind of data are included, is the data clean or make sense?

1. Data cleaning and preparation

After the first step, the data needs to be made clean and ready for model building. Any duplicate or missing need to be exam and treated. The data needs to be evaluated for weight and imbalance. Plus, the data may need to be fitted and transformed. All data also needs to be converted into proper format and type to facilitate the model building.

1. Model building and evaluation.

There are many different models that can be chosen, each has its own pro and cons. So multiple models will be tried out. Some evaluations also will be conducted to disclose strength and performance of each model.

when more information need to be studied, the 3 steps will be iterated and are not in strict of order.

**The complete coding is done at a jupyter notebook named as shown below, the notebook will be submitted as a separate file**. In this summary, I will only explain the critical steps and attach the output chart to show the progress and result.



Last step is conclusion. I will summarize the findings and results of the model building and validation. As well as a high-level summary of the learning from the exercise.

EDA:

* Exam the data after loading up the data:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

The data looks clean with no NA, but later I found out there are some ‘?” or no meaning word which I will need to treat as unknown. It also shown there are both numerical and categorical variables.

A close-up of a computer code

Description automatically generated

* Data visualization:

I used multiple chart types via 2 packages (matplotlib and seaborn). Below are the outputs:

A diagram of a weight distribution

Description automatically generated

This shows the sampling has different population weight, however, late study shows this population weight do not have too much correlation with income and ends up being excluded.

A blue circle with orange triangle and green triangle with black text

Description automatically generated

Looks like the sampling is heavily on white. It is still aligned with race distribution of the population back then, so I do not think it is imbalanced.

A graph with blue and orange bars

Description automatically generated A graph with numbers and a bar chart

Description automatically generated

A graph of a number of people

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

Above few charts show the age, race, marital status and work class will bring a lot of variation to the income. So later some correlation analysis will be conducted to identify which factors are important and if any factors are correlated.

Since age and education length are important factors, check the age and education length distribution. Looks like they are compact enough and resemble a normal distribution with skew, so I do not plan to get them standardized.

A graph of age distribution

Description automatically generated

A graph of a graph of a number of bars

Description automatically generated with medium confidence

Data Preparation:

* There are many “?” in some columns, so need to convert them to unknown.

A close-up of a white background

Description automatically generated

* The education level and education length are almost the same, so drop one of them.

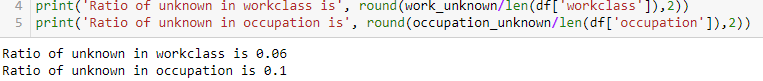


* Check the numerical features and see if they are closely correlated.

A number and numbers on a white background

Description automatically generated

* The ratio of unknown in work class and occupation are low so drop unknowns





* Convert the categorical variables into numerical
  + Income converts to 0 and 1 based on less or above $50K,
  + Sex converts to 0 and 1,
  + Convert rest using pd.get\_dummies



A computer screen shot of a computer code

Description automatically generated

Now the data is clean and ready for data modeling!

Building models:

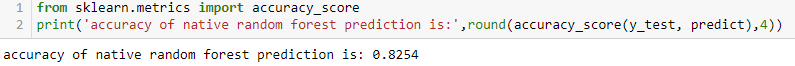
The first question is what model to use. From last few weeks exercises, I used linear regression model, Polynomial model and Random forest model for numerical variable prediction. At later weeks, I used support vector Machine model and logical regression for categorical variable predication.

Personally, I like random-forest since it is a linear regression model but have advantage of overfitting avoid.

I will try random forest regressor first (native one and one with hyperparameter tuning), then I will try Random ForestClassifier model as well.

Please refer to my jupyter notebook for details of my models, I will not attached each single step but just output and charts here.

Model 1. ***RandomForest Regressor without tuning***



A chart of a confusion matrix

Description automatically generated with medium confidence

Model 2. ***RandomForest Regressor ((with hyper parameter optimizer)***

A screenshot of a computer program

Description automatically generated



With hyper parameters tuning, the accuracy improved 0.1 but computing time increased significantly on my 2.4Ghz i5 CPU.

Below is a comparison summary of the 2 models:

|  |  |  |
| --- | --- | --- |
|  | Random Forest simple | Random Forest with hyper parameter tuning |
| Accuracy | 0.8254 | 0.835 |
| R2 | 0.0142 | 0.0683 |
| RSME | 0.4178 | 0.4178 |
| MAE | 0.1746 | 0.1650 |

Classification report:

Random Forest simple:

precision recall f1-score support

0 0.85 0.95 0.89 4420

1 0.70 0.42 0.53 1320

accuracy 0.83 5740

macro avg 0.77 0.68 0.71 5740

weighted avg 0.81 0.83 0.81 5740

Random Forest with hyper parameter tuning:

precision recall f1-score support

0 0.85 0.95 0.90 4420

1 0.74 0.44 0.55 1320

accuracy 0.84 5740

macro avg 0.79 0.70 0.72 5740

weighted avg 0.82 0.84 0.82 5740

From the parameter comparison, for both models, the accuracy is well and comparable. The RSME and MAE are also comparable and acceptable. The R2 seems low. However, since the accuracy is high, I would say the model has limitations but still a good model. The limitation is on minority target class.

Last, one important question is still not answered, which factor is the most important factor to influence the income level. I turn to a RF classifier for help. I want to try out this model since I did not get chance to use it in any prior exercises.

***Model 3. RandomForestClassifier model*** A graph with a bar chart

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated A graph with numbers and a grid

Description automatically generated

looks like age and education length are the 2 most important factors. married with spouse living together can improve the income level as well. occupation matters a little. Native country may not have enough data to tell the difference.

CONCLUSION:

* For this income study, the analysis and model discovered that age and education length are the 2 important factors, which meets common sense range.
* For the 2 prediction models tried, I think both models work well, run within reasonable times, and yield good accuracy. While they predict high income class well, the models has a little difficulty working on minority target class, it may need some other technics to help improving, which I lack knowledge and will need to learn more. I would say for it is ready for deploy for general prediction.
* Among these 2 models, I think they are okay for a quick assessment based on raw data, or, used for general population. On the other hand, they may not predict well on the minority target class. So, better separate those classes out and then run the model.
* There are still many different modeling methods I have not explored for this study. I believe there will be one or multiple model skills that can boost the prediction accuracy to 90% or improve R2.
* This is a simple and clean dataset for education purposes, so the data cleaning is quite minimum. If a freshly collected dataset being presented, the data cleaning may be much more complex and may need a subject expert to aid on variable picking and NA data handling.
* Ethical concern: besides the quality of the dataset, a common challenge is how/where to get the dataset. Often datasets have business values and may have IP restrictions. Besides, the most recent income data may not be suitable to release to any request from any country or groups without checking their intentions. Also, many things in real life (like student financial aid) are depends on income level, so it need to be careful when using the outcome.