Using a Machine Learning (ML) model to predict individual's income

Ben Hughes

1 Introduction

Income prediction models are used to optimise decision making across the world. Being able to predict an individual's income can give you a great insight into their financial stability and larger socioeconomic trends, enabling financial institutions and policymakers to make better decisions.

Applications for financial institutions:

Income prediction can be useful for banking and financial institutions as they asses a individual's credit rating, which affects their ability to repay loans and in turn informs businesses decisions. Many current credit models have been shown to reinforce systematic inequalities against certain groups with black and hispanic people having a disproportionate amount of 'credit invisibile' individuals in the US[1], reinforcing the lower black home-ownership rate of 44% compared to the 74% for non-Latinx white people[2] which acts as a barrier for wealth building within these communities. A more accurate model that uses income prediction could be used to address the inequalities, provided it is able to address biases within its training data, a problem with some ML models[3] which I have tried to address

Applications for policy-making

An income prediction tool could be used for policy-making so policy-makers can enact targeted social programs to address inequalities within populations and to improve policy around taxing which can disproportionately affect certain groups.

1.1 Methods

In order to create a an income prediction model I have used 2 machine learning methods (Linear and Logistic Regression) to predict a given individual's annual income. I have modeled the problem as a binary classification problem, with the categories being whether the income is below or above $50K. My models take 9 features (age, workclass, education, marital-status, occupation, living-situation, sex, hours-per-week and capital).

1.2 Results

After tuning i managed to a achieve an F1 score of 0.670 for both logistic and linear regression with the linear having a higher Cohen Kappa score 0.549 than logistic 0.537. To improve the models a more balanced dataset could be used or undersampling the ¡=50K could be used, but this would also have issues with underrepresenting lower wage groups in the data and could introduce new biases.

Figure 1 a flowchart to show the system for developing a Machine Learning model to predict income

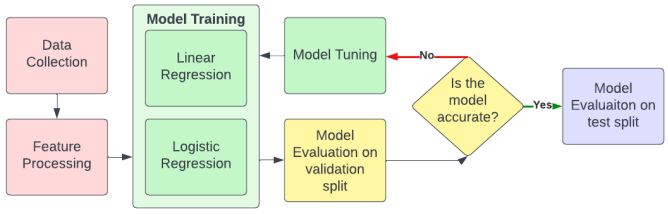

Fig. 1. figure for how the ML system will work

2 Data

2.1 Data Overview

The dataset I have selected for this model is the 1994 United States Census dataset compiled by Ronny Kohavi and Barry Becker in 1996[4]. It is a binary classification dataset for income above or below $50K, this classification is represented in the data by two labels: `>50K` and `<=50K`. It contains 48842 instances each with 14 features. The split between the 2 labels is shown in Fig 1, there is a disproportionate number of entries 23.93% being `<=50K` which will need to be addressed when training and evaluating the model to ensure it is not biased toward `>50K`

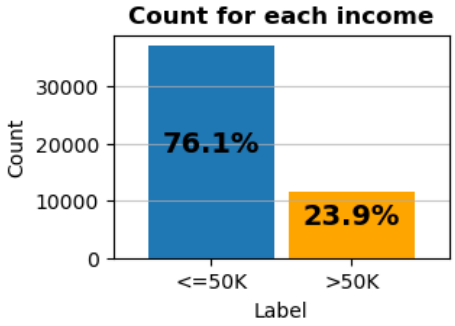

Fig. 2. Income Distribution in the Dataset

The data was already split into a test train section, I have combined both sets together in order to transform the data

consistently and then select my own test train validation split.

2.2 Data construction and transformation

For many features I have grouped together similar values and merged smaller categories in order to simplify the model. For some categories I have had to take a more in depth approach which are outlined below

2.2.1 Label

For the Label I corrected inconsistencies in the labels and encoded the labels as with `<=50K` as 0 and `>50K` as 1

2.2.2 age

The age feature has been grouped into student `<22` and retirement over `>61` categories where income is low and into categories for other ages between according to their income distribution as seen in figure 2. Grouping ages together by income merges smaller categories and reduces noise.

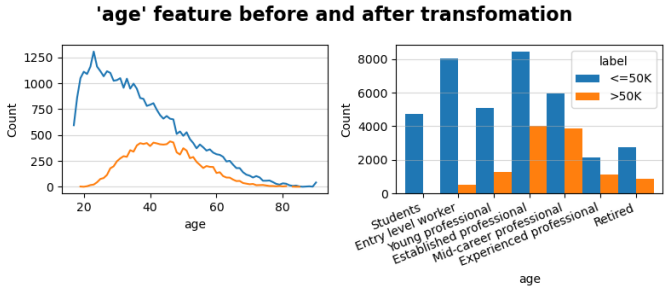

Fig. 3. Income Distribution in the Feature 'age'

2.2.3 workclass

For the workclass column I first removed the never worked, which was mostly students and without pay and retirees as these were small categories where the label was almost entirely `>50K`, both groups with known incomes - pension or student loans. I also grouped together all the government employees, as they have a similar pay and are similar organisations and merged the self-employeed categories together as they were small.

2.2.4 fnlwgt

Fnlwgt is a number generated by the dataset authors that represents the number of people the entry represents within the whole census. After binning the data into bins of width 50,000 there is a uniform distribution of `>$50K` labels, This feature will not provide any value and will increase noise and so I have removed it.

2.2.5 education

education and education-num are both features representing the same data, with education-num already being integer encoded. In order to better understand the data i removed education-num and just used education. The different categories were unbalanced and so the first step I took was to group them together by education level to make the categories more even, and then encoded them according to the level of education.

2.2.6 marital-status

I have grouped together marital-status into 3 categories that align with income: Married, Unmarried and Separated. The categories that made these up all had very similar income and were infrequent. Merging them would make the model generalise better to new, less detailed data.

2.2.7 occupation

There are 2 infrequent occupations I have grouped into other categories due to their low count. I have binned together Armed-Forces into Protective-serv (protective services) and Priv-house-serv into Handlers-cleaners as they both are similar occupations and both have similar incomes to that group.

2.2.8 relationship

I found from my data exploration, as shown in figure 4, relationship was unreliable. There were cases where people's gender differed from their gender listed in their relationship, and there were categories that were unclear as to their meaning. I used the data to bin them together into whether

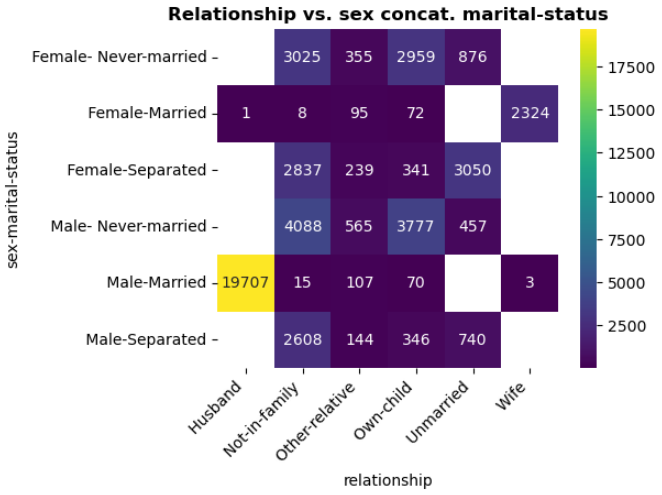

Fig. 4. Relationship vs sex and marital status

they lived alone (Not-in-family, Other-relative, Own child, Unmarried) or they were cohabiting (Husband, Wife), as I thought people who lived alone would have higher income, however the inverse pattern was seen in the data, likely caused by married people being older and further in their career.

2.2.9 race

I have removed the race column in order to try and avoid introducing a racial bias in my model. Predictive models that use race as a feature can often times lead to discriminatory results[4] that can reinforce inequalities against certain groups during model deployment, as was the case. On top of that using race as a metric in decision making can create legal issues in many parts of the world and thus removing it ensures the model complies with all legal frameworks. As you can see in Table 1 the demographics distribution is inaccurate with data from 2020 with white people being disproportionately represented in the dataset.

TABLE 1
Demographic Distribution

Category	White	Black	Asian	Native	Other
Percentage from data	85.5%	9.6%	3.1%	1.0%	0.8%
Percentage today[5]	75.5%	13.6%	9.3%	1.3%	0.3%

2.2.10 *sex*

The dataset has sex in 2 categories, male and female as shown in table 3 the data is unbalanced towards male. Despite this imbalance I have kept the data in as sex is a strong predictor of income due to a pay gap in gender in the US that was 68%[6] in 1996, when the census was taken that the dataset was compiled from. This disparity is also present in the dataset as shown in table 2

TABLE 2
Income Distribution by Sex

Sex	Male	Female
% making >$50K	30.4%	10.9%

2.2.11 *capital-gain and capital-loss*

I have replaced capital-gain and capital-loss with a new category of capital which is defined as any entry with a capital-gain or capital-loss value that is not 0. This category is supposed to represent people who have some investment in some sort of capital 0. I have assumed that if someone is able to invest in capital then they will have a larger disposable income and therefore a higher income, this assumption is reflected in the data.

2.2.12 *hours-per-week*

I have grouped hours per week into categories of 40 and above or below, as the continuous data was very spread out and may produce noise in the model.

2.2.13 *native-country*

For the native-country the data was extremely skewed towards the United States with it taking up 89.7% with the next highest being mexico with only 1.9% I tried to combat this by binning countries into their continent however it was still to skewed. Due to the data already have a large imbalance in the label I thought it would be best to remove this category

2.3 Dataset split

I have chosen to use a split shown in tablex

TABLE 3
Data Split for Training, Validation, and Test Sets

	Train	Validation	Test
Percentage	80%	10%	10%

The benefit of using a validation and test split is that I can tune the hyperparameters around the validation set and then use the test set to evaluate the final model to avoid overfitting hyperparameters to the test data.

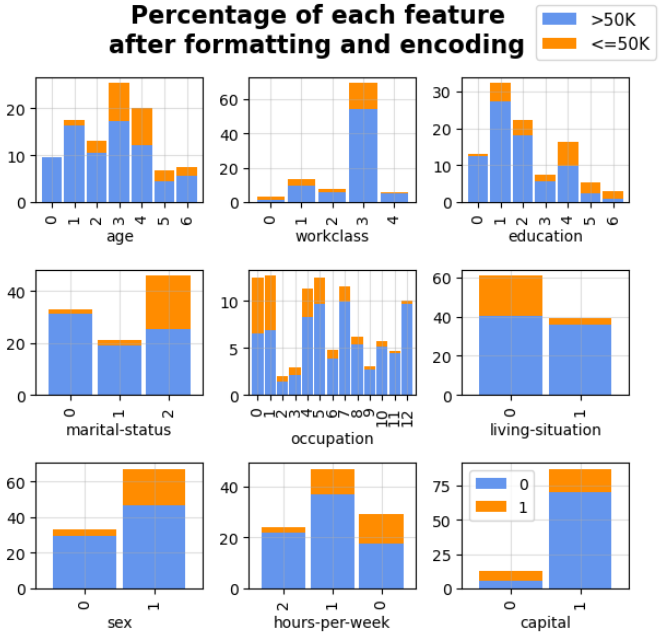

Fig. 5. Income Distribution across all Features in the Data

3 Model Evaluation

The models I have selected are Linear Regression with a threshold to determine the model's prediction, as a baseline and a logistic regression as an improved model.

3.1 Evaluation Metrics

Due to the unbalanced count of labels the model will favour predicting the `>$50K` which will result in the accuracy score being high despite the low precision of results. This is an example of the accuracy paradox and is the reason why accuracy is not always the best metric for binary classification tasks. I have chosen to use the F1 score, confusion matrix, and Cohen's kappa, a measure of agreement beyond chance for the models prediction to evaluate the quality of my model as they will all account for the imbalance of labels in the dataset.

3.2 Model Selection

When encoding the categories I chose to encode them according to either their natural order (age and education being ordered by their natural scale) or their order of percentage `>$50K`. By ordering them this way it will enhance performance of regression models on this data, and so I have chosen linear regression and logistic regression for my model. I believe logistic regression will perform better as it is able to fit curves to the data and can handle categorical data well, unlike linear regression Linear regression will also be sensitive to outliers which there may be lots of due to the data imbalance.

3.3 Before Tuning

I first trained both models on data from the validation set so that I could make sure I was not overfitting the hyperparameters to the test set. The results are shown in

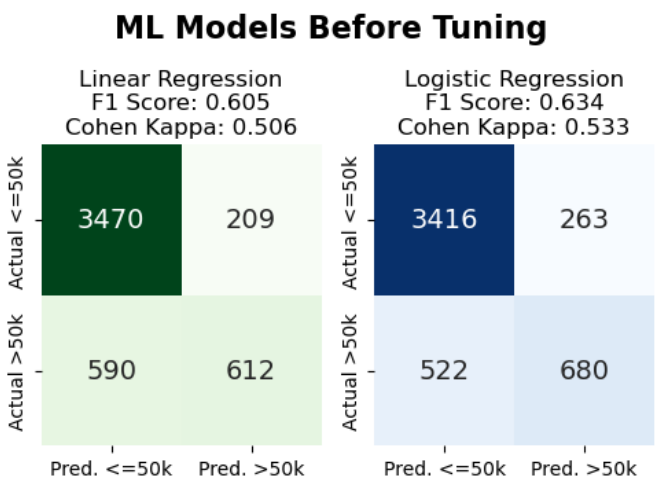

Fig. 6. Income Distribution in the Dataset

figure 6. You can see from the confusion matrix that both models are very biased towards predicting `<=$50K` but they both very rarely predict `>$50K` incorrectly, this is due to the large bias in the data. When predicting `>$50K` values the linear regression had 0.509 sensitivity, quite low whereas logistic was better slightly at predicting `>50K` with `0.566` sensitivity. Both models have low F1 and Kappa, with the Logistic regression still doing better than linear.

3.4 Tuning Process

The first step I took in tuning both models was to combat the imbalance in the target data by introducing a bias in both models. For logistic regression I set the class weight to balanced and for linear regression I changed the threshold from 0.5 to 0.35, a value I found to give the best F1 score. This bias would make them both more sensitive towards the `>$50K` prediction. I also changed the logistic model to use the liblinear solver as I had best results with it and decreased the C value to 0.01.

3.5 After Tuning

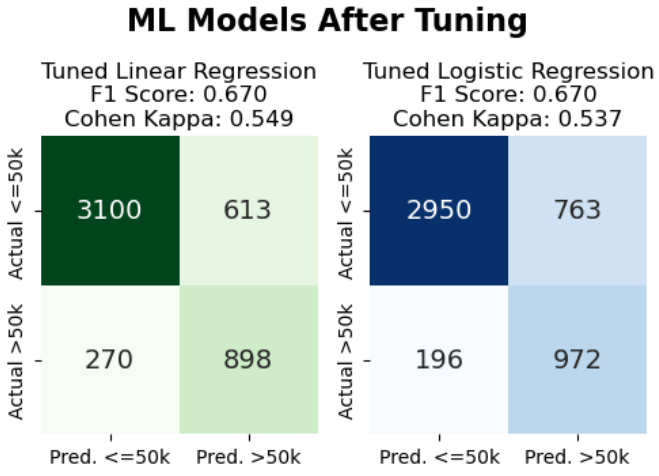

Fig. 7. Income Distribution in the Dataset

the confusion matrix shown in figure shows that after tuning the F1 score is much better and for predicting `>$50K`, the sensitivity scores are also much better for both models. The cohen's Kappa has had a small increase for both, with a greater increase for the linear regression, due to the balance between false positives and false negatives that logistic does not have, however it is still what would be classed as a moderate agreement between the expected and observed predictions. Both models now have much higher positive predictive value for `>$50K`, which will be caused by the bias making them much more likely to predict `>$50K`. The logistic regression has not improved as much as linear regression after the tuning, I believe this is because the logistic regression has become too biased towards the `>$50K`.

4 Reflection

The Machine Learning submodule is the module I have found most interesting so far this year. Going in to the module I was familiar with basic linear regression models and did not realise the breadth of different algorithms there are for machine learning, and their applications for different types of problems. I found instance based models especially K-nearest-neighbours to be most interesting in the way that it compares new data to the values it has already seen, I tried to use it for my model but found it did not work well due to the imbalance in data. This coursework has taught me a lot about hyper-parameter tuning and how it can be difficult to find the optimal hyperparameters that increase performance across different metrics. If I were to do this coursework again I would try and address the imbalance in the data, perhaps by under or over sampling different labels and I would experiment with more complex models. I like to think that the capital feature that i created is quite unique as i have read other papers using this dataset and many remove the capital gain and loss features.

References

[1] Kenneth P. Brevoort, Philipp Grimm, Michelle Kambara. *Data Point: Credit Invisibles*. May 2015. Consumer Financial Protection Bureau. https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

[2] U.S. Census Bureau. *Quarterly Residential Vacancies and Homeownership, Third Quarter 2023*. Release Number: CB23-173. October 31, 2023. https://www.census.gov/housing/hvs/files/currenthvspress.pdf.

[3] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner. *Machine Bias*. May 23, 2016. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Frederik Zuiderveen Borgesius. *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. Council of Europe, 2018. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

[5] *QuickFacts: United States*. Census Bureau, United States. https://www.census.gov/quickfacts/fact/table/US/PST045222.

[6] The Council of Economic Advisers. *Explaining Trends in the Gender Wage Gap*. June 1998. https://clintonwhitehouse4.archives.gov/WH/EOP/CEA/html/gendergap.html#:~:text=The%20data%20that%20permit%20disaggregation,to%2074%20percent%20in%201996.