Barry Becker Classification Model

Bao Dinh, Wesley Smith, Dakoda Fisher, Mohammed Qurneh, Andreas Cedron

School of Data Science, University of North Carolina at Charlotte

DTSC 2302: Modeling and Society

Dr. Iván Flores Martínez and Ted Carmichael

April 29, 2024

1 INTRODUCTION

Based on data scraped by Barry Becker's 1994 Census Database "Census Income Dataset", the data used for our study focuses on whether income exceeds $50K/yr based on census data. Our purpose with this study is to examine and use certain factors such as education, demographics, field of work and experience to predict income, getting people to realize their maximum earning potential and make better decisions on their industrial careers. As people get into this predictive model, they could start benchmarking against peers as they map out and analyze any choices they could make for their futures in certain areas or a possible change in industries. The stakeholders here include those considering moving to the US, those who are interested in working in other fields, students considering continuing their studies or not, those who are merely curious about their peers' performance, and even those who want to make strategic career plans.

For this study, we had to go to various data analytical techniques (such as heatmaps or the centroid clustering method) that provide better insights towards it (Statswork, 2020). Taking a step back, we also had to take a deeper look at the various factors that impact income such as education and how it's, on average, an effective tool on reducing income inequality (Abdullah et al., 2013) and also how the gender pay gap has had a gripping effect on women and their income is affected worse by other factors such as age, ethnicities, lifestyles or "family-friendly policies" (Kochhar, 2023).

We predict people with certain factors like more work experience or a higher level of education are more likely to earn a $50k+ salary as opposed to others who lack in such areas.

2 DATA EXPLORATION AND PREPROCESSING

2.1 Dataset

The dataset consists of 48,842 entries over 15 variables, of which six are integer values and nine are categorical. Our approach was split into three main steps: clean invalid values, condense unnecessary variables, and convert categorical variables for modeling. We knew the data held many invalid data points, either as null values or empty strings. First, we found invalid values in the *workclass*, *occupation*, and *native-country* variables. We did this by searching all unique values for each variable. There were the standard null and Nan values, as well as values such as '?' which we ruled to be invalid. These invalid entries were standardized into null and Nan values. In total the three variables held over 3,000 entries with invalid data. Removing all invalid entries left us with 45,222 entries.

2.2 Features and Processing

Next, we sought to remove multicollinearity by condensing or removing columns that share similar data points. *Education* and *education-num* represent the same information, a person's highest level of education achieved, one in integer format and the other with categories. We chose to keep the integer format as it could be more easily used in ordinal calculations and removed the *education* variable. *Capital-gain* and *capital-loss* both represented the same idea, but were mutually-exclusive variables. Since an entry could not contain capital gain as well as capital loss, we chose to combine these two variables into one integer capable of holding negative values whenever capital loss was tracked. During our analysis of invalid entries, we found that the binary value of *income* held four possible values due to misinput. Correcting this gave us two possible values '¡=50K' and '¿50K'. At the same time, we discovered that the *capital-gain* and *hours-per-week* variables held odd maximum values at 99,999 and 99 respectively. These values stood out and led to us determining that these were the maximum allowed values. The 'hours-per-week' variable appeared normal, with unique values steadily reaching the "maximum". The 'capital-gain' variable, on the other hand, had a large gap between the "maximu" and next highest values. While we considered culling these entries as outliers, the number of entries with this max value were too great to cut without impacting the data significantly. Lastly, we found inconsistent meaning in the *fnlwgt* (meaning "final weight") variable. The data suggest that the final weight represented an estimate of how many people fit each entry. After using summation, we found that the final weight of all entries equaled over eight billion. We acknowledged that this number could not fit the data accurately, especially since the values were estimations, so before modeling we removed the variable.

Finally, we used the dummy encoding technique to convert all of our categorical variables into binary values that would better fit the logistic regression model. After these alterations, we ended with a dataset of 45,222 entries spanning 80 variables. A majority of the expanded variables came from the *native-country* and *occupation* variables, which held the greatest variation in valid data.

2.3 Visualization

Before putting the model through training, we observed a few trends of the data set. One of the strongest predictions was that *hours-per-week* would play a large part in determining whether someone would make more than $50,000.

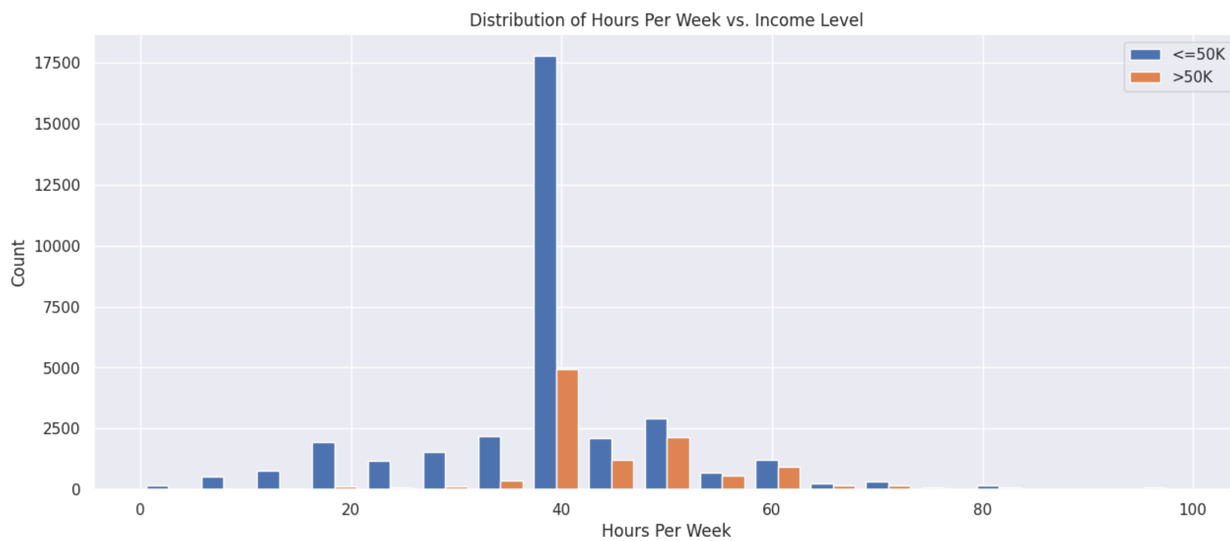

Figure 1: Hours Per Week versus Income

This visual shows that the data held an abnormally high number of entries around 40 hours, though this was to be expected as 40 hours falls right on the 9am to 5pm schedule that was commonplace. It also shows that there were very few entries able to earn over 50 thousand without working 40 hours or more per week.

We also found that over 80% of the entries were submitted by individuals who identified as White. As shown in the diagram below, White individuals make up nearly 40,000 data entries, with roughly a third of them earning greater than 50 thousand. Meanwhile, despite the lower count, we can observe that less than a fifth of the Black identifying group earn more than 50 thousand. Despite this graph showing that all groups are more likely to make less than 50 thousand, we can see an inequality in the percentages of each group.

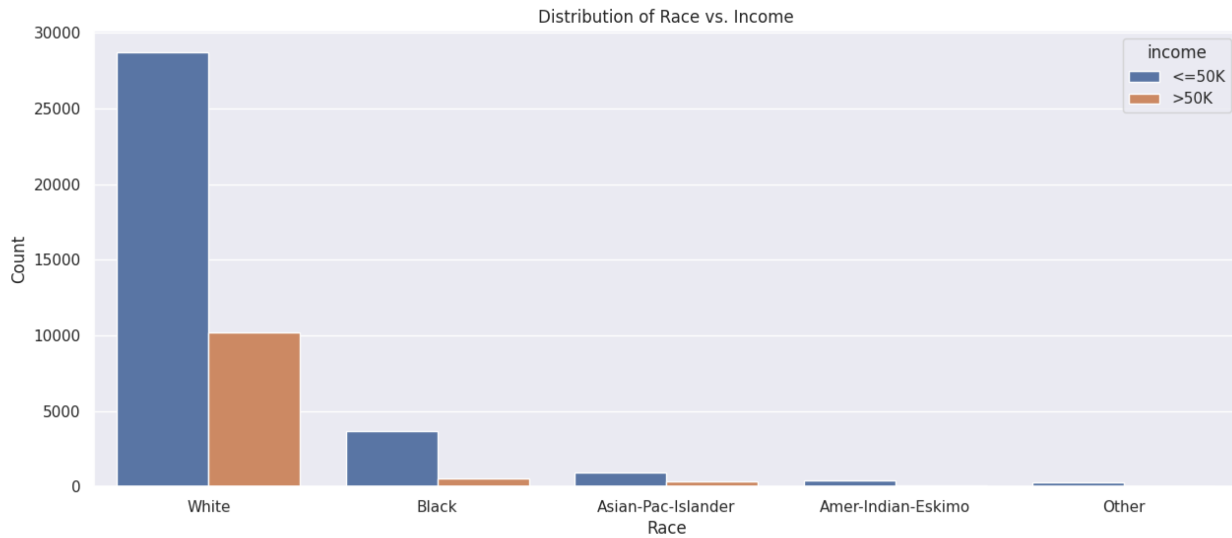

Figure 2: Race versus Income

Lastly, with the common knowledge that 30 years ago, when this data was collected, women were not equally paid in their employment, we wanted to visualize the difference between male and female earnings. We predicted that, just like with individuals identifying as White, being male would not guarantee higher earnings, but would clearly impact the likelihood of earning more.

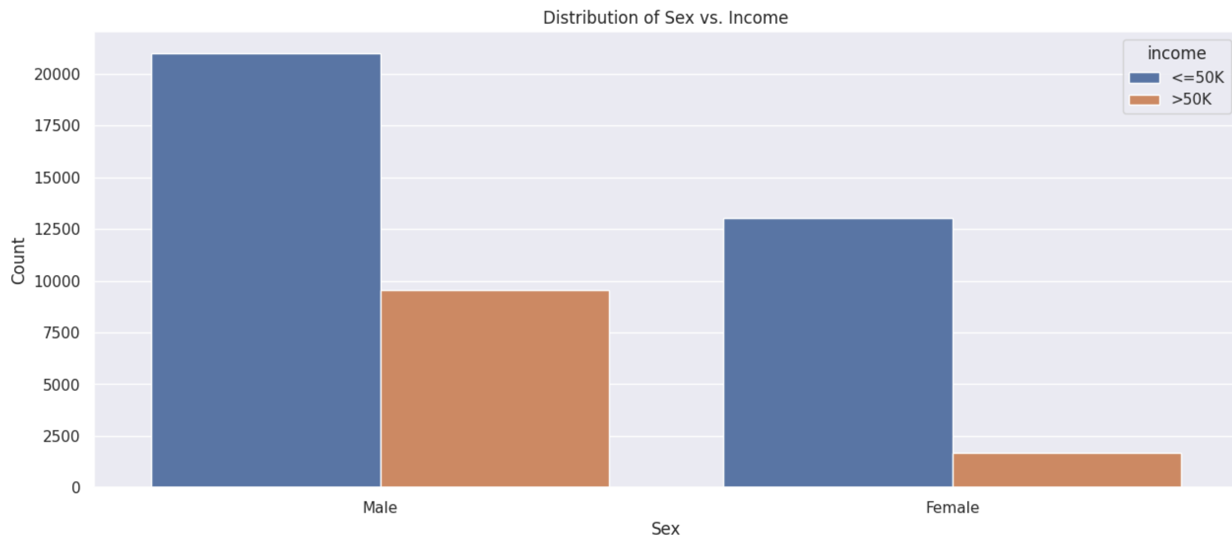

Figure 3: Sex versus Income

Above we can see that the majority of data entries are male. This alone does not confirm that women made less, but when we look at the percentage of each group to earn more than 50 thousand, we see that the model reflects what was predicted. We observe that roughly 30% of males earned more than $50,000, while roughly 15% of females earned the same amount.

3 MODEL SELECTION

For this project we decided to use logistic regression and decision tree models. We have a variety of reasons to choose these models, one being that keeping it to two models helps us get effective results within our time frame. The logistic regression model is the best way for us to answer our research question because it shows the relationship between our variables and income. The logistic regression model works really well for our project because we want to compare two variables and get a relationship which is what a logistic regression does. Using the logistic regression model we are able to accurately predict if certain variables will lead to making over or under 50,000. The main issue with using a logistic regression model is that it can be hard to understand especially if one is not used to working with them. The decision tree will be used to see what variables have a significant impact on the income. One of the biggest strengths of a decision tree is that it is easier to understand from a human perspective. This will allow us to understand our data better and to be able to accurately predict based on our models but the decision tree does have the downside of being less reliable. Both of our models have various weaknesses and strengths. The logistic regression model is reliable and in general will have better and more consistent results. The decision tree on the other hand can be used to show others our results as it is much easier to understand and it is easier to make predictions from it. When we use both of these models we can have them make up for the others weakness and achieve the most accurate predictions that we can.

Once we chose our models we had to make sure that they would actually give accurate results and fine tune them to make sure we have the best possible models. To do this we used the grid search method. We did this by setting up our hyperparameters to be within our desired scope of the project and then created models within the parameter until the best model was found. When we did this we got the f1 scores from each of the tuned models and compared them to the f1 scores of the original models. We found that the original logistic regression model has an f1 score of 0.679 and the tuned model has an f1 score of 0.686. For the original decision tree model it has an f1 score of 0.625 and the tuned model has an f1 score of 0.675. For both the models we decided to use the tuned models because even though there was not a significant increase both models did improve with tuning. Now that we have the models that work best for our project and they are fine tuned we can start to analyze the results and make predictions.

4 RESULTS

4.1 Model Fine-Tuning

After tuning the hyperparameters of the logistic regression model, the impacts to the model's performance are:

Table 1. Initial Model vs. Tuned Model

Metric	Initial Model	Tuned Model
Accuracy	0.7975	0.8283
Precision	0.5623	0.6286
Recall	0.8382	0.7561
F1	0.6731	0.6865
ROC-AUC	0.9016	0.9016

We improved the model's accuracy from 79.75% to 82.83%; in other words, our tuned model correctly predicted 3% more outcomes of the test set than our initial model. Furthermore, we improved the model's precision from 56.23% to 62.86%; when our tuned model predicts an individual makes an income of greater than $50,000, it is correct 6% of the time more than our initial model. However, we sacrificed our model's recall which decreased from 83.82% to 75.61%; in other words, our tuned model correctly identified 7% of all individuals less than our initial model. Overall, our model's accuracy, in terms of F1-score, improved from 67.31% to 68.65% — an improvement of approximately 1%. Lastly, there was no change in the model's ROC-AUC which indicates there was no improvement to the model's discriminative ability. Despite the minimal improvement, the hyperparmeter-tuning puts the logistic regression ahead of the decision-tree in terms of F1-score; thus, we'll focus on the results of the logistic regression model for our discussion.

4.2 Model Coefficients

Instead of analyzing all 79 predictors, we'll explore some notable ones; a predictor's notability will be based on its odds ratio. Predictors will be considered "notable" if their odds ratio is above 8; we chose 8 as the threshold because most of the predictors (75 out of 79) have an odds ratio of less than 8. Moreover, predictors' odds ratio were calculated using `sklearn` whereas their statistical significance used `statsmodel`; we used different libraries because `scikit-learn` doesn't support statistical significance. After exponentiating our logistic regression's coefficients, the notable predictors we found were:

Table 2. Notable Variables

Variable	Odds Ratio	p-Value
Age	8.3878	0.000
Hours Worked Per Week	21.2043	0.000
Education	78.0407	0.000
Capital Profit	1092140487.4064	0.000

Age

Age had an odds ratio of 8.3878 which means that for each additional year of age, the odds of making an income of greater than $50,000 increases by approximately 8 times. Moreover, *age* had a p-value of 0.000 which means that it is a statistically-significant predictor. Overall, there is a strong, positive, statistically-significant relationship between *age* and making an income of greater than $50,000.

Hours Worked Per Week

Hours worked per week had an odds ratio of 21.2043 which means that for each additional hour worked per week, the odds of making an income of greater than $50,000 increases by approximately 21 times. Moreover, *hours worked per week* had a p-value of 0.000 which means that it is a statistically-significant predictor. Overall, there is a strong, positive, statistically-significant relationship between *hours worked per week* and making an income of greater than $50,000.

Education

Education had an odds ratio of 78.0407 which means that for each additional level of education, the odds of making an income of greater than $50,000 increases by approximately 78 times. Moreover, *education* had a p-value of 0.000 which means that it is a statistically-significant predictor. Overall, there is a strong, positive, statistically-significant relationship between *education* and making an income of greater than $50,000.

Capital Profit

Capital profit had an odds ratio of 1,092,140,487.4064 which means that for each additional dollar of capital profit, the odds of making an income of greater than $50,000 increases by approximately 1,000,000,000 times. Moreover, *capital profit* had a p-value of 0.000 which means that it is a statistically-significant predictor. Overall, there is a strong, positive, statistically-significant relationship between *capital profit* and making an income of greater than $50,000.

5 DISCUSSION

5.1 Limitations and Assumptions

According to Google for Developer's article on imbalanced data, our dataset has a mild class imbalance. Class imbalances can lead to our model being biased towards the majority class which negatively impacts the model's ability to predict the minority class. In the case of our dataset, the class imbalance undermines our model's ability to predict individuals making an income of over $50,000. To address the class imbalanced, we used class weighting to upweight our minority class; however, upweighting doesn't completely avoid the class imbalance. In the future, we would like to gather a dataset with a balanced class distribution.

Furthermore, our dataset is from 1994 which negatively impacts our model's applicability to the current job market because the job market has changed drastically since 1994 (2016, Pew Research Center). The predictors that had a strong relationship with making an income of greater than $50,000 may not be as strong in today's job market. To address the dated nature of our model, we would like to gather a dataset with dataset from the last 5 years in order to be more representative of the current job market.

On the topic of unexpected findings, the odds ratio of *capital profit* was abnormally high. We suspect the odds ratio is abnormally high because people tend to use their disposable income to invest in capital assets.

5.2 Practical Implications and Applications

Our predictive model offers significant potential for application in real-world scenarios and informs decision-making processes across various domains. For instance, government policymakers can use the model's insights to craft targeted interventions that address income disparities. By identifying demographic segments with higher probabilities of earning above $50K/year, policymakers can design more effective taxation policies, allocate resources for welfare programs, and strategize social initiatives to promote economic equity and social mobility. Just like how companies can use these predictions to fine-tune their marketing strategies and customize product offerings for their target audience's income levels, financial institutions can also gain by creating personalized financial products and services specifically designed for various income groups. This way, they enhance customer satisfaction and loyalty in the process.

Both opportunities and potential consequences arise from implementing the model's recommendations and predictions. Leveraging the insights can lead to more informed decision-making and targeted resource allocation, but it is necessary to mitigate potential risks such as preserving biases or making worse prevalent inequalities. For instance, if not carefully calibrated, the model's recommendations could inadvertently reinforce existing gender or racial disparities in income distribution.

The model's insights are useful for multiple industries, organizations and involved parties. Government agencies like departments of social services, labor or economic development can use these predictions to distribute their resources better and deal with poverty alleviation more correctly. Marketing and advertising companies might find the capacity of this model to help them customize their campaigns towards certain income groups very beneficial; therefore, they can improve the return on investment for their customers. In the same way, the financial services field can apply these forecasts to tailor financial products and services for various income categories, improving their competitiveness and market significance.

Nonetheless, it's the everyday people who are most likely to gain a lot from the insights that come with our prediction model. People thinking about their careers, whether they want to further their education, or even those considering moving can use the model to get an idea of how much they might earn and make better decisions. For example, a person considering more education can evaluate the possible return on investment by understanding how education levels link with income. Likewise, people who are looking into changing careers or finding new jobs can utilize the model for comparing themselves with others and spotting areas they must improve their skills in order to increase how much money they earn. In essence, the model provides people with a stronger sense of control and understanding as they move through their work-related and financial paths.