FINAL PROJECT WRITE-UP - ALAN CHU

Problem Statement:

In the past, income was relatively a more simple concept and easier to predict. As of now, adult income is determined by many factors, including race, gender, etc. Movements around the world are occurring to equalize income, such as women's suffrage and BLM. For my final project, I aim to develop a predictive model capable of accurately estimating if a person makes over or under 50k dollars by analyzing the trends and patterns within different demographics and working statuses. In general, income is an indicator of someone's success, and the general bar is set at 50k. Fundamentally, this model is useful for researchers and sociologists who want to analyze trends and factors that play into income and what policies should be made to influence that. These predictions will allow them to make informed political decisions that effectively target the root causes of low income, allowing for efficient budget allocation. Historically, much money has been wasted on reducing income inequality such as raising the minimum wage. The successful development of a model will not only enhance predictive capabilities for income scores but will also empower politicians with insights. Additionally, understanding the evolution of income over the years will contribute to reactive and proactive planning to account for trends in the future.

EDA

Data set Description:

This is a dataset pulled from the UCI Machine Repository. Because the website is outdated, I decoded the response content from bytes to a UTF-8 encoded string. After reading through this dataset of Adult Incomes over or under 50k, I found a question to be answered using this data. This dataset is pulled from a clean record census, giving accurate data. Initially, one might naturally assume that factors like being married or having capital gains would make it so that they are likely to make more than 50k, but I am here to put that to the test. Although the dataset includes aspects such as race and workplace, there needs to be further analysis and comparisons to understand the main drivers behind income. Potentially, the dataset can reveal problems within the United States, such as income disparity due to race or ethnicity.

Main Research Question:

How do socioeconomic, demographic, and personal factors influence an individual's likelihood of earning more than $50,000 annually in the United States, and are there underlying factors that make attaining 50,000 annually unfair?

Sub Questions to help guide answers to the main question:

Fundamentally, does any education correlate with having higher income levels, and how is race involved with the furthest education?

Histogram of Income Distribution Vs. Income Level:

At every level of education, there is a significant difference between having education and no education. In general, obtaining a degree or education can significantly boost one's chance of acquiring income over 50k, and this is quite intuitive. Specifically, we can observe that the difference in income occurs most at the HS and some college levels, meaning that basic education like high school is quite essential. It can be argued that income is not only attributed to the education itself; there are also factors such as connections and friends that you make in school, allowing you to place yourself further for a more stable income. Therefore, although high school education is not essential to a general world understanding, the connections and opportunities are quite important. Now, we will delve into other multi-faceted factors that could contribute to a higher income among adults.

Heatmap of Education Vs. Race:

In the heatmap above, there are a few key takeaways:

- Overall, Whites and Asians show a higher attainment of education compared to other races

- In the elementary to high school levels of education, there seems to be varying data ranging across races
- In the Bachelor's degree section, Asians have a disproportionately higher amount
- Although there is not much variability, among people who reached HS Graduation the furthest, blacks and american-indian-eskimo people have the highest concentration.

KEY TAKEAWAY:

We learned through the first question that education indeed does correlate with higher income levels >50k, however, that is unfair considering what race a person is. Overall, there seems to be an education disparity among races, which can be attributed to the economic and social mobility policies within the United States and history. With the existence of historical segregation, there are inherent economic imbalances within the United States, making it harder for minorities to obtain premium education that comes with a high price tag.

Is there evidence of a "peak earning" age range that differs across industries or job types and what does that reveal about certain jobs?

Histogram of Ratio of High Earners (>50K) by Age

Indeed, there is a peak earning age of 50, which is directly in the middle. Many people in this age have either started their ventures or moved to higher positions, so naturally as people age, they make more money in executive positions. However, there are two spikes at the age of 79 and 83, which is intriguing and confusing at the same time. This can be explained in two ways: retirement and small samples. Firstly, every older individual has retirement accounts and savings, which can contribute to a higher income. Furthermore, if there is selection bias or any other sort of bias within a small sample, it is extremely likely that the data could've been skewed. Nevertheless, let's explore what occupations typically associated with higher age, leading to a higher income.

Violin Plot of Education Vs Age

In the violin plot, it is observed that certain jobs have higher age medians. There are 3 positions that have a relatively high median: Executive Manager, Professional Specialty, Private House Service. All 3 of these occupations require specialization and time, which age provides. Naturally, when there are jobs with mobility, age tends to be important and that correlates with high income. Additionally, it seems that many manual labor jobs are done by younger folks, which indicated by the bar plot, tend to make less money.

Key Takeaway

Naturally, seniority plays into higher roles, which makes sense. The graphs and data set checking was mostly just a sanity check to see if the data was relatively correct.

Within working classes, specifically the private sector, is race a large factor in determining one's income? If so, what should be done about it?

Histogram of distribution of income by workclass level

Quite an interesting find. Most people who work in the private sector actually make less than 50,000 a month. Most people would think that working in the private sector provides many benefits such as healthcare and pay, but it seems like that is not the case. Additionally, people who are working in self employed inc have around the same number of people making over or under 50k, meaning that is the best workclass in terms of income. Now, we will be examining the proportion of black people who work in the private sector and their pay relative to other races.

Income distribution of Black Vs All

According to the Pew Research Center, 41% of black people have experienced discrimination in the private workplace, demonstrating 1st degree racism. The comparative bar plot proves this, as the proportion of black people making less than 50k in the private sector is more than other races.

Key Takeaway

Through the three graphs in this question, it is demonstrated that minorities, especially black people, get discriminated against in the workplace. By being treated differently, they are given less income in comparison to other races, and this is observed by looking at the largest workplace by quite a margin: The Private Sector.

What is the difference in income distribution between males and females, and does relationship status matter?

Barplot of proportion of individuals earning over 50k by gender

Naturally, through observing the data presented, Men make more money than Women overall. Although efforts to implement women in the workplace are in effort, there hasn't been a clear balance, as shown in the bars. Observing the calculation, it is shown that most of the men who make over fifty thousand dollars are married, being 89.1%. This reflects a patriarchal norm; the man of the household is supposed to bring home the money. In modern times, we can see that the norm is changing, and this census data may be outdated. With more change towards women in STEM and other male-dominated occupations, women could be making significantly more in just a matter of a few years.

Main findings:

Overall, it is found that there are many factors that play into the possibility of an individual making over 50,000 dollars annually. The gender, race, age, occupation, and many other factors are all important when considering the likelihood of a person being wealthy. Furthermore, after exploring the data, there were many problems unearthed, such as race affecting one's inability to generate a solid income. With the data, there are two outstanding issues. Firstly, race inhibits certain education, and discrimination from employers causes income to be less for minorities. This was found to draw comparisons between race, education, and working classes. Additionally, it was observed that women inherently make less money than men, likely due to a patriarchal belief of the man making the money. This is backed up by the fact that most men who make over 50,000 a year are married.

Modeling

To predict if adults make over or under 50,000 a year, I will be using various classification models because the variable I am trying to predict is qualitative. These models all account for variation and randomness and are generally more accurate and allow me to conduct deeper analysis. For the training and test data, I decided to use 80-20 because that makes it so the model is not over-accurate, but also uses an adequate amount of training data. The goal of modeling is to provide a method of evaluating multiple variables and testing the accuracy of the data set.

Baseline Model

I have decided to predict the accuracy of a baseline model about my target variable, income. It produced an accuracy of 0.759 through using the mode, so we expect to beat this number in more complex models I will be using.

KNN and LGR with a Confusion Matrix

Why?
Using KNN and LGR would be a good start to classifying data, as one takes in close groups and the other is simpler regression. I can see if the income status is based on clusters or similar categories, as I observed before in the EDA. Additionally, a LGR would be important to determine linear relationships, as it is a function to produce a probability value between 0 and 1 for each input. A confusion matrix would be very relevant to evaluating the performance of a predictive model. The confusion matrix is a tool for understanding the accuracy and specific types of errors made by a classification model. In this case, where the goal is to predict whether an individual's income falls into the "<50k" or ">=50k" category, the confusion matrix provides valuable insights. Specifically, the confusion matrices are excellent at identifying false positives and negatives.

Implementation:
After splitting the data into a 80-20 split, we have X variables be other columns and the target column to be income. Then, we encode KNN and LGR pipelines, and we display a confusion matrix where we extract the precision scores from both.

Interpretation and Results:

The score on the KNN matrix was 0.7828751857355126, and the score on the LGR was 0.7393267651888341. The KNN model proved to be more accurate than the baseline, while LGR was not. This exemplifies that if I were to use a neighboring approach through grouping, that would prove to be more accurate. Additionally, by examining the KNN confusion matrix across different demographic groups), I can assess whether the model is exhibiting any biases in its predictions. In general, this proves that the dataset is somewhat coherent and the model is accurate most of the time.

Decision Trees

Why?

To analyze the data further, I decided to use a decision tree classification model to estimate non-linear relationships between data analytics and machine learning because they break down complex data into more manageable parts. Furthermore, Decision Trees are extremely clear, providing a breakdown of the steps it takes to separate data. A potential downside is that decision trees are indeed prone to overfitting, as elements such as entropy play in.

Implementation:

I decided to build a reasonable model with a depth of 3 and with the criterion of entropy. Essentially, I am creating 3 splits with the target to decrease the most entropy, or randomness, in each split. Past that, I plotted the decision tree onto a Decision Tree Boundary Model for visual representation purposes, and the actual process on the right. Finally, I scored accuracy using the .score function.

Interpretation and Results:

The decision tree turned out with a 0.78 accuracy, scoring better than all models before and performing slightly better than KNN's nearest neighbors. Before discussing the accuracy, I wanted to touch upon the expense of decision trees. Because decision trees are more straightforward and require only a splitting process, usually they are more computationally cheaper than KNN, which requires constant comparison and grouping. Fundamentally, the accuracy is likely higher because decision trees can identify key features and create hierarchical rules, while KNN cannot deal with a large data set like this as well.

Random Forests

Why?

Simply put, random forests are likely more accurate than decision trees because they implement more trees, hence the forest name. A downside is that if you were to put a depth of >100000 then that would cause computational lag and it would be costly. I did this following the decision tree because I wanted to initially see if a singular tree could perform better than KNN.

Implementation:

Again, I split the model 80-20, then assigned categorical columns. I used the RandomForestClassifier Pipeline, then assigned depths of 3, 4, 5, 6, 10 and # of trees as 25, 50, 75, and 100. Past that, I used GridSearchCV to cross-validate and fit it. The best parameters were 10 depths and 100 trees, and the score was 0.863.

Interpretation and Results:

Naturally, after determining that the decision tree model is quite accurate, we want to up the ante and incorporate a method of using multiple decision trees to produce a result with the least variability. We end up using the RandomForestsClassifier model, and the depth of 10 and 100 trees are the best parameters we can use. However, as we move past 50 trees, the accuracy does not improve by that much, so to save costs I believe using 50 trees with 10 depth is feasible.

Gradient Boosting

Why?

Gradient boosting is a machine learning algorithm that incorporates weak models and trains them over many rounds to come to the best possible conclusion. I wanted to use this because Gradient Boosting is stepping into a more advanced realm in the sense that there is improvement and constant adjustment of error.

Implementation:

Split the data, then use .get_dummies to encode categorical columns. Instantiate a Gradient Boosting classifier with n_estimators=100,   learning_rate=0.1, max_depth=3, random_state=42. Then, fit the gradient boosting and report predictions.

Interpretation and Results:

With precision, we can determine if the predicted positive results are correct. We have predicted 0.89, under 50k, meaning 89% of the predicted class 0 instances are correct. Additionally, we have predicted

0.80, over 50k, meaning 80% of the predicted class 1 instances are correct. Therefore with a higher precision, Gradient Boosting already outperforms previous models. Recall is mainly involved with predicting true positive cases, and 95% of <50k were predicted. In total, there were 6513 tests, highlighting the rigorous process Gradient Boosting goes through to reduce error. However, there is a clear struggle with identifying >50k than under 50k, as there is likely just less data on people with >= 50k.

Modeling Conclusion

Overall, after comparing all the models for testing, gradient boosting generally had the best outcomes and precision when evaluating my data. This is not to say that we shouldn't use other models, but it seems that gradient boosting would be the best choice.

Next Steps

For the next steps, I would like to do the following:

- Implement Artificial Neural Networks (ANN). It acts similarly to gradient boosting and is more accurate.
- Explore more variables in the table. Although I found many possible causes and factors for income status, I would like to look into the hours per week and native-country categories to determine if immigration and work ethic correlate.
- Improve my model on gradient boosting by creating an optimal threshold, boosting recall for >50k guesses.

This model slightly outperforms the random forest and decision tree because it uses an iterative approach and reduces error with each check.

Final Takeaway

Referring back to the initial research question, the best way we can estimate the different factors into income is by using the gradient boosting model. Given that, some certain dynamics among particular variables can cause a different accuracy, but using all variables, the gradient boosting model stands out.