Predicting Adult Income That Will Exceed $50,000 per Year

Author: Zhi Ye

Summary of Research Questions and Results

1. How can I accurately predict whether the income of an adult will exceed $50,000/year based on the census data?
    a. I will be performing machine learning to predict whether the income of an adult will exceed $50,000/year based on the provided census data. By doing so, the task will be classification, which will require defining features and labels, creating a model, training the model, making predictions on the data, and assessing its accuracy.
    b. **Result:** In determining whether the income of an adult will exceed $50,000/year, I based my machine learning on specific factors such as education, occupation, race, sex, and income. Doing so, I used the DecisionTreeClassifier and KNeighborsClassifier to see which model would perform better with the given dataset. The result that I found was that the DecisionTreeClassifier produced fairly accurate results, which were better than the KNeighborsClassifier.
2. Is there an income inequality across gender at every race and occupation?
    a. In order to determine whether or not there is an income inequality across gender at every race and occupation, I need to compute the total income for every race and occupation in the dataset for both genders. Additionally, I will graph the results using a bar chart and compare the income between genders for every race and occupation to see if income inequality is present in the dataset.
    b. **Result:** In determining whether or not there was an income inequality across gender at every race and occupation in the dataset, I created two filtered data frames. The first involved race, sex, and income. While the second involved occupation, sex, and income. For both data frames, I graphed percentages for males and females with income >50K or <=50K. By doing so, I discovered that there was an income inequality present in the dataset across gender in terms of race and occupation.
3. How has the education of adults affected income?
    a. To determine if education of adults has an effect on income, I will be computing the average income for each education level. Additionally, I will graph the results using a bar chart to compare the average income across education levels to see how much education has affected income.
    b. **Result:** In determining how has education of adults affect income; I created a filtered data frame involving the columns education and income. Following that, I made computed columns and graphed percentages of adults with income >50K or <=50K by education. By doing so, it was evident that education has a significant impact on income as percentages of adults with >50K income has better correlation with adults that achieved higher education.

Motivation and Background

The background of this Adult Dataset is that it provides data on the income of adult individuals. This data includes relevant information that are potential factors that determines the income of adults such as age, education, gender, race, etc. The motivation behind this project is to use machine learning to accurately predict whether the income of an adult will exceed $50,000 per year based on the census data. By doing so, I can discover if there are any underlying factors that determines whether or not an adult individual will make over $50,000 per year. Additionally, I will push my analysis further by computing income across genders at every race and occupation to discover if there is an income inequality gap present in the dataset. Furthermore, I will analyze how a factor such as education affects the income of adults to see if there is a large disparity of income levels for adult individuals with varying education. Overall, the goal of this project is to determine the underlying factors that affect income for adult individuals. My analysis will be able provide answers to societal questions such as "how does education affect the income of an adult individual?" and "is there an income inequality gap between males and females across race and occupation?"

Dataset

Adult Dataset: http://archive.ics.uci.edu/ml/datasets/Adult

The Adult Dataset from the UCI Machine Learning Repository provides data on the income of adult individuals. In the dataset, there are 48,842 rows and 14 columns. More specifically, the 14 columns included represents age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country.

Methodology

The goal of this project is to predict whether the income of an adult will exceed $50,000/year based on the census data by creating a machine learning algorithm. Then, an analysis on the dataset will be conducted to see if there is an income inequality gap between male and females across race and occupation. Additionally, the analysis will extend to discovering the impact that education has on income. By doing these analyses in addition to the machine learning algorithm, I will be able to gain a better understanding on factors that determines whether an adult individual will have an income that exceeds $50,000 per year.

In the beginning phase of the project, the dataset must first be examined to understand its content. Then, I need to filter the dataset for columns that I am interested in such as education, occupation, race, sex, and income. Next, I will remove any rows that contain missing values. Once the dataset has been cleaned and consolidated, the machine learning algorithm can be conducted. In doing so, the features and labels need to be identified. In this specific case, the label will be income and the features will be the other columns in the cleaned-up dataset such as education, occupation, race, etc. After this step, the dataset needs to be split into 70% training data and 30% testing data, where the DecisionTreeClassifier will be used to train the model. Next, the model needs to be predicted and the accuracy score must be assessed for both training and testing data. Once this is done, the results will need to be analyzed to see where the machine learning algorithm was able to accurately predict whether the income of an adult will exceed $50,000/year based on my filtered census dataset. After

using the DecisionTreeClassifier, the next step is to use nearest neighbor classification using a similar process to determine which classifier predicts the result the best.

Once the machine learning portion of the project is completed, further analysis of the dataset is required to determine if there is an income inequality present between genders and the impact of education on income. To begin the second analysis (research question 2), the necessary columns that will be needed from the dataset that the question calls for will be race, sex, and income. Next, filtered data frames should be made to identify males and females with >50K or <=50K income. Then, computed columns should be made to calculate percentages of males and females with >50K or <=50K income. Then, this process should be repeated to find if there is an income inequality across gender at every occupation, where columns such as occupation, sex, and income will be needed. The task for research question 2 will be to compute the percentage of individuals, who makes >50K or <=50K and identify if they are male or female, their race, and occupation. Then, for the third analysis, perform a similar process that takes all individuals and compare their education level to discover the difference between the percentage of individuals making >50K or <=50K income. Once, these values for each analysis have been completed, plotly should be utilized to plot bar charts to visualize the results. When all three analyses have been completed, explain the methods used and determine if the research questions have been accurately answered.

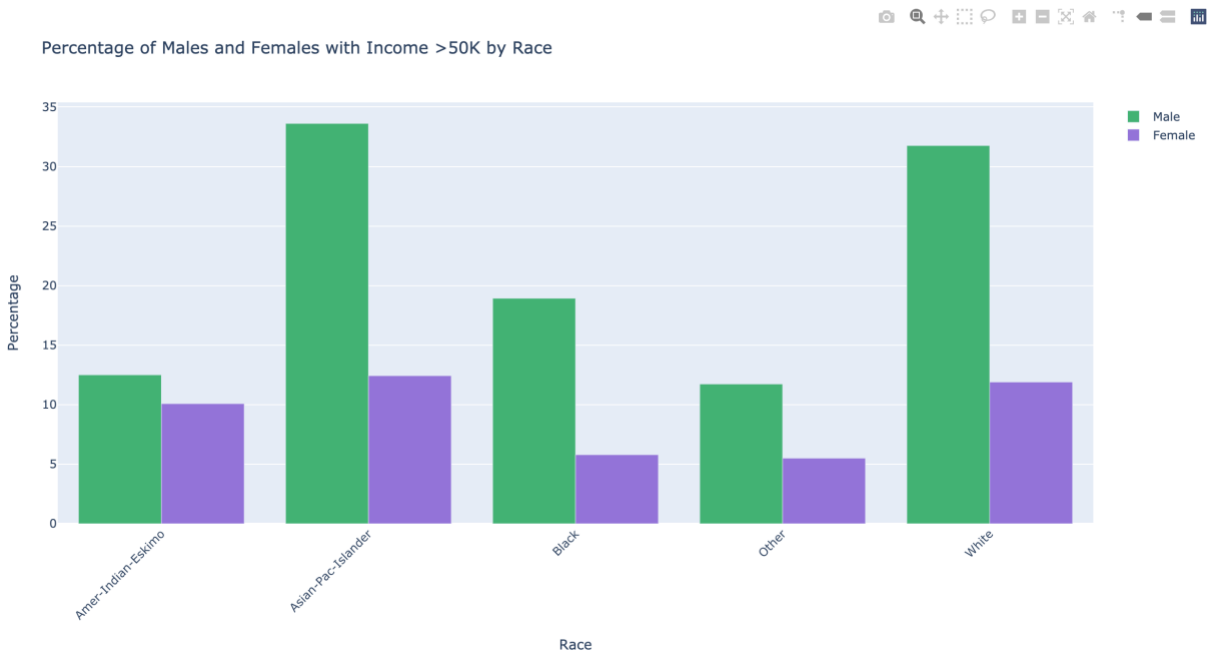
Results

Research Question 1: How can I accurately predict whether the income of an adult will exceed $50,000/year based on the census data?

In performing machine learning to accurately predict whether the income of an adult will exceed $50,000/year based on the census data, I first filtered the adult dataset for columns that I was interested in, which were education, occupation, race, sex, and income. I created my training and testing data and split it into 70% training data and 30% testing data. Then, I applied the DecisionTreeClassifier and KNeighborsClassifier to determine which classifier would produce a better result at predicting. In the case of the DecisionTreeClassifier, it's training accuracy was around 0.7987163984745606 and its testing accuracy was 0.7850477430555556. While KNeighborsClassifier, the f1-score (testing accuracy) for adults with >50K income was around 0.33, which is not very accurate at all. In terms of the classifier that was better at predicting in this case, the DecisionTreeClassifier did a better job. Additionally, for the testing data, 1434 adults exceeded an income of $50,000 while 7782 adults had income of less than or equal to $50,000. Though, the testing accuracy was around 80%, there is an apparent difference between adults with greater than $50,000 income per year and adults with less than or equal to $50,00 income per year when performing machine learning on factors such as education, occupation, race, sex, and income. To view these computations, the image below showcases the performance of each classifier, where "Train Accuracy" to "Total # of Adults" represent the DecisionTreeClassifier and "precision" to "weighted avg" represent KNeighborsClassifier.

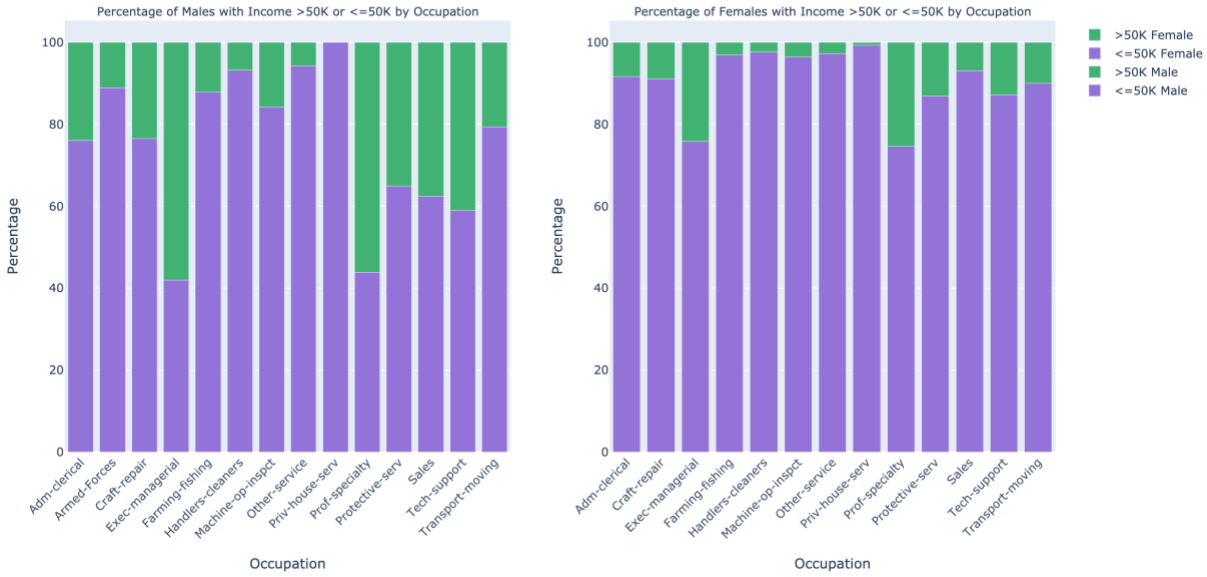
```
Train Accuracy: 0.7987163984745606
Test Accuracy: 0.7850477430555556
# of Adults with income >50K: 1434
# of Adults with income <=50K: 7782
Total # of Adults: 9216
              precision    recall  f1-score   support

       <=50K       0.78      0.94      0.85      6881
        >50K       0.56      0.23      0.33      2335

    accuracy                           0.76      9216
   macro avg       0.67      0.59      0.59      9216
weighted avg       0.73      0.76      0.72      9216
```

Research Question 2: Is there an income inequality across gender at every race?



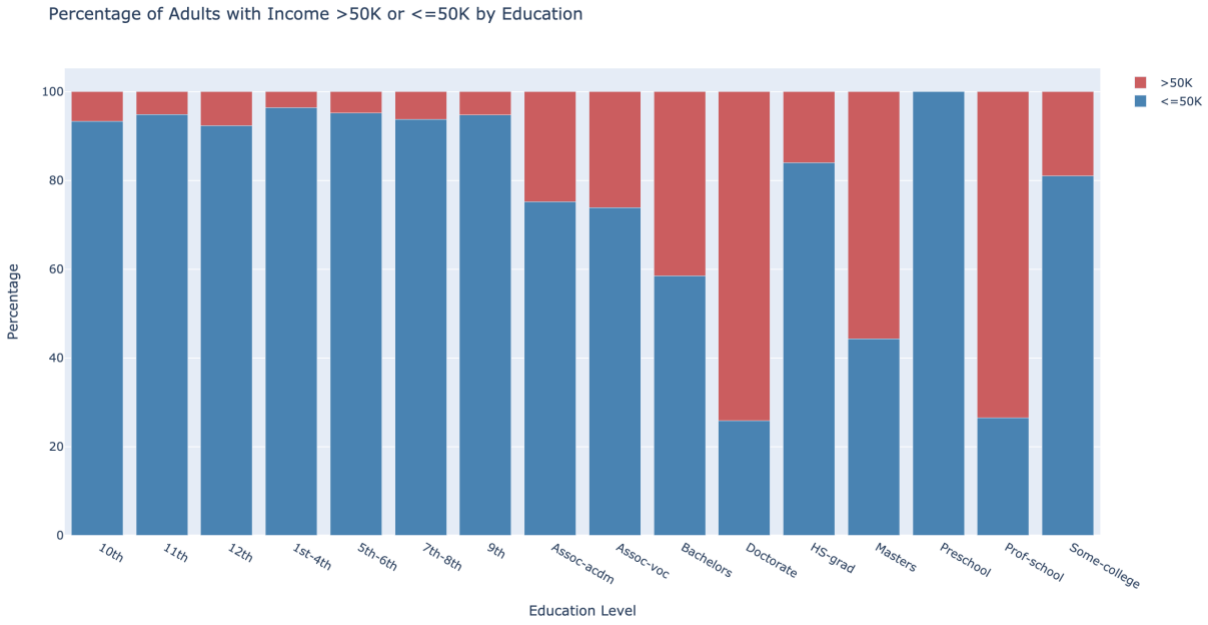
Viewing the visualization above, it showcases percentage of males and females with income >50K by race. Just from a glance, it is apparent that in the adult dataset, there is an income inequality present across gender in every race. In this specific case, the race featured in this dataset are Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, and White. In each racial category, there are higher percentages of men with incomes >50K and if you take the inverse of the visualization above, it's noticeable that there are higher percentages of women with <=50K in each racial category. The standouts in this visualization are for Asian-Pac-Islander and White, where 33.62% of males have an income of >50K while 12.43% of females has an income of >50K. For the racial group, White, 31.76% of males have an income of >50K whereas 11.9% of females have an income of >50K.

Research Question 2: Is there an income inequality across gender at every occupation?



Viewing the visualization above, it showcases percentage of males with income >50K or <=50K by occupation and percentage of females with income >50K or <=50k by occupation. By analyzing each visualization, it is apparent that in the adult dataset, there is an income inequality present across gender in most if not all occupations. In this specific case, there are various occupations featured in this dataset. Some are Exec-managerial, Prof-specialty, and Tech-support. In each of these occupations, there are more males making >50K income than females and more females making <=50K income than males. Looking at Tech-support specifically, there are 41.03% of males making >50K income while females are at 12.93%. for <=50K income in Tech-support, males are at 58.97% whereas females are at 87.07%. Looking at these statistics, it's evident that there's an income inequality across gender throughout the industry.

Research Question 3: How has the education of adults affected income?



Viewing the visualization above, it showcases percentages of adults with income >50K or <=50K by education. By analyzing the visualization, it is apparent in the adult dataset that education has a large impact on income. For example, education levels past High School grad shows that percentages for adults with >50K income are either higher or significantly higher. Looking at a specific example, I will compare Masters to 10th grade. The percentage of adults with >50K income with a master's education level is around 55.66% whereas percentage of adults with >50K income with 10th grade education level is around 6.65%. Moreover, percentage of adults with <=50K income with a master's education level is around 44.34% whereas percentage of adults with <=50K income with 10th grade education level is around 93.35%. Overall, the impact of education on income is huge as there are higher percentage of adults with >50K income when their education level surpasses High School grad.

Challenge Goals

The challenge goals that my project have met are **Machine Learning** and **New Library**. For **Machine Learning**, I predicted whether the income of an adult will exceed $50,000/year based the dataset, *adult.data*. In particular, I created a filtered data frame based on the columns: education, occupation, race, sex, and income. Then, I made the training and testing data to fit my models, which were the DecisionTreeClassifier and KNeighborsClassifier since this was a classification task. By doing so, I compared results between the two classifiers to see which one was able to predict the results the best. I found that in this case, the DecisionTreeClassifier performed better since the testing accuracy was around 0.79 whereas for KNeighborsClassifier, the f1-score (KNN test accuracy) was around 0.45. For the **New Library**, I used plotly to make quality and interactive visualizations to showcase my data and insights from machine learning. More specifically, I created three visualizations to help answer my research questions 2 and 3. The first visualization is a group bar chart that showcases the

percentage of males and females with income >50K by race (research question 2). While the second visualization is a stacked bar chart that utilized plotly's sub-plotting. In this visualization, I made two charts, which are percentage of males with income >50K or <=50K by occupation and percentage of females with income >50K or <=50K by occupation (research question 2). Lastly, for the third visualization, I created a stacked bar chart to showcase percentage of adults with income >50K or <=50K by education (research question 3).

Work Plan Evaluation

1. Analyze and clean dataset **(3 hours)**
    a. Dataset must be examined to understand its content.
    b. Clean the dataset by removing any unnecessary columns such as capital-gain, capital-loss, hours-per-week, educational-num, fnlwgt, and native country.
    c. Remove any rows that contain missing values.
    d. Create a consolidated dataset.
2. Perform machine learning **(12 hours)**
    a. Identify the features and labels (the labels will be >50K and <=50K and the features will be the other columns in the cleaned-up dataset such as age, education, race, etc.)
    b. Split the dataset into 70% training data and 30% testing data, where the DecisionTreeClassifier will be used to train the model.
    c. Next, the model needs to be predicted and the accuracy score must be assessed for both training and testing data.
    d. Analyze results
    e. Once this is done, the results will need to be analyzed to see where the machine learning algorithm was able to accurately predict whether the income of an adult will exceed $50,000/year based on the census data.
    f. After using the DecisionTreeClassifier, perform nearest neighbor classification using a similar process.
    g. Determine which classifier predicts the result the best.
3. Perform computation for research questions 2 and 3 **(6 hours)**
    a. Begin the second analysis (research question 2).
    b. Find the necessary columns such >50K, <=50K, sex, race, and occupation, etc.
    c. Compute the difference of individuals, who makes >50K or <=50K and identify if they are male or female, their race, and occupation.
    d. Begin the third analysis (research question 3).
    e. Find the necessary columns such as education, >50K, <=50K, etc.
    f. Perform computations that takes all individuals and compare their education level to discover the difference income.
4. Plot results for research questions 2 and 3 **(3 hours)**
    a. Once the values for each analysis have been completed, use plotly to plot a bar chart to visualize the results.
5. Analyze results and determine if research questions have been accurately answered **(4 hours)**

a. Explain the methods used and determine if the research questions have been accurately answered.

In my work plan, my high-level tasks followed by its estimate of time in hours were analyze and clean dataset (3 hours), perform machine learning (12 hours), perform computation for research questions 2 and 3 (6 hours), plot results for research questions 2 and 3 (3 hours), and analyze results and determine if research questions have been accurately answered (4 hours). These estimates I found were somewhat accurate; however, some sections took longer than expected while others were not as long as estimated. Starting with analyzing and cleaning the dataset, I would say this did took around 3 hours as I took my time to examine the dataset to understand its content. Additionally, I took time to consider which columns in the dataset were important for my research and by doing so, I filtered the dataset according to those columns to create a consolidated dataset for each section of my project.

The next high-level task was performing machine learning. This task surprisingly did not take as long as I estimated. The majority of time for this task was learning about the KNeighborsClassifier and going back into the lessons to strengthen my understanding of the DecisionTreeClassifier. Once I gained a better understanding of both these classification algorithms, I experimented with the machine learning portion of my project in Jupyter Notebooks. After I found that I was producing the correct results, I moved the development in Visual Studio Code. This process in total took around 8 hours, which was 4 hours faster than I expected.

Once machine learning was completed, I moved into performing computations for research questions 2 and 3. Initially believing this would take 6 hours, I was wrong. This task in reality took around 12 hours since it required various filtering, data frame editing, merging, and computing. Research question 2 took up most of this time as I had to filter and create a data frame for males with >50K income, filter and create a data frame for males with <=50K income, merge the two data frames together, make computed columns for percentages, and then repeat a similar process for females in the dataset. Additionally, in the case for occupation in research question 2 and education in research question 3, I had to fill in a missing occupation or education level in my filtered data frame, which took some time to figure out. Once I finished making computations for research question 2, I moved into research question 3. This section required less time since it followed a similar process to research question 2 while having less parts to it.

When all the computations were done, I moved into the next high-level task, which was plotting results for research questions 2 and 3. The initial estimation for this task (3 hours) was far off than the actual hours required (12 hours). I found that I spent a lot of time in researching and experimenting with plotly. I had to figure out how to create bar charts, what type of bar charts to use, how to create labels, customize colors, save the plot, and how to create subplots. I thought plotly would be similar to other libraries I used in the past; however, that wasn't the case causing more time spent on understanding the new library to use it correctly.

Lastly, the final high-level task was analyzing my results. My initial estimation of 4 hours was accurate since there were a good number of parts to analyze in my project. Each research question was fairly equal in terms of time that it took for the analysis. Overall, the total time in

terms of hours for my work plan was estimated around 28 hours, while the total time required was 39 hours.

Testing

Initially for testing, it was difficult to identify what parts to test my project since it deals with machine learning and plotting. However, after much consideration, I decided to split some of my functions up into separate functions to return values that I can test on. For example, I made the filtered_df_train_test function to return X_train, X_test, y_train, y_test. Then, I created a helper function called helper_length_train_test_data in my test file, income_projec_test.py that takes the train_test_data as a parameter to help test the filtered_df_train_test function by returning the length of the train_test_data. Next, I defined a function called test_filtered_df_train_test that takes in the length of the train_test_data and used assert_equals from cse163_utils to determine if the expected value matches the received value. If they do not match, assert_equals will crash the program and tell me what was wrong. The way that the testing looks like is shown in this image:

```
def test_filtered_df_train_test(len_train_test_data):
    """
    Takes len_train_test_data as a parameter to test the
    filtered_df_train_test function using my own test file
    to ensure that it is working properly using the imported
    assert_equals function from cse163_utils. By doing so, I can
    confirm if the expected value matches the received value. If they
    do not match, assert_equals will crash the program and tell me what
    was wrong.
    """
    assert_equals(14, len_train_test_data[0])
    assert_equals(6, len_train_test_data[1])
    assert_equals(14, len_train_test_data[2])
    assert_equals(6, len_train_test_data[3])
```

This testing strategy was then repeated throughout my machine learning functions to test that my functions were returning the correct lengths for my data. For functions such as compare_gender_race, compare_gender_occupation, and compare_education_income, I returned merged data frames and tested with assert_equals to see if the correct values were in certain columns and rows. How this looks like is shown in the image below:

```
def test_compare_education_income(merged_edu_data):
    """
    Takes merged_edu_data as a parameter to test the
    compare_education_income function using my own test file
    to ensure that it is working properly using the imported
    assert_equals function from cse163_utils. By doing so, I can
    confirm if the expected value matches the received value.
    If they do not match, assert_equals will crash the program and
    tell me what was wrong.
    """
    assert_equals(' Bachelors', merged_edu_data.loc[0, 'education'])
    assert_equals(3, merged_edu_data.loc[1, 'Count (<=50K)'])
    assert_equals(' Masters', merged_edu_data.loc[2, 'education'])
    assert_equals(33.33, merged_edu_data.loc[2, 'Percent (<=50K)'])
```

Lastly, an important note when testing my project, I created my own custom test file called adult.test that features only a small portion of data from the main dataset, adult.data to help test in a more controlled environment.

Collaboration

I did not collaborate with anyone, I worked on this project individually. In doing so, I experimented with Jupyter Notebook in the beginning phase of my project before moving the development into Visual Studio Code.