Proposal: Salary Prediction and Analysis

DATA 450 Capstone

Nicholas Nemkov

February 17, 2025

1 Introduction

A person's wage is influenced by various personal characteristics and broader state-level economic factors. Attributes such as education level, marital status, and work experience can play a crucial role in determining individual earnings, while state-level wages provide insight into regional economic trends and job market conditions. Understanding these factors can help individuals make informed career decisions, negotiate salaries effectively, and identify potential wage disparities across different states. Through data analysis, machine learning, and visualizations, we can explore how personal characteristics impact wages and compare these trends across states. These insights can highlight common patterns in earnings, reveal disparities, and provide valuable guidance for job seekers and policymakers.

2 Dataset

There are multiple datasets used for this project:

2.1 adult.data

Becker, Barry and Ronny Kohavi. "Adult." UCI Machine Learning Repository, 1996, https://doi.org/10.24432/C5XW20.

This dataset, extracted from the 1994 Census database, contains a 15 columns and 32561 observations. Each row represents a person, with 14 of their personal characteristics (age, sex, marital status, etc.) and a categorical target column stating whether their salary is less than or equal to $50k or greater than $50k. The majority of the employees are born in the United States, with others being from all around the world.

adult.data contains the following variables:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: Continuous, numerical representation of education column.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: Country of birth.
- salary: This is the target variable, represents categories (<=50k or >50k) in USD.

2.2 state_M2023_dl.xlsx

U.S. Bureau of Labor Statistics. (2023). Occupational Employment and Wage Statistics (OEWS): New York. Retrieved from https://www.bls.gov/oes/current/oes_ny.htm

This dataset contains the 2023 occupational employment and wage estimates, calculated with data collected from employers in all industry sectors in metropolitan and nonmetropolitan areas in the US (including US territories).

This dataset contains the following columns:

- area: U.S. (99), state FIPS code, Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA) code, or OEWS-specific nonmetropolitan area code
- area_title: Area name
- area_type: 1= U.S.; 2= State; 3= U.S. Territory; 4= Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA); 6= Nonmetropolitan Area
- prim_state: The primary state for the given area. "US" is used for the national estimates.
- naics: North American Industry Classification System (NAICS) code for the given industry.

- naics_title: North American Industry Classification System (NAICS) title for the given industry.
- i_group: Industry level. Indicates cross-industry or NAICS sector, 3-digit, 4-digit, 5-digit, or 6-digit industry. For industries that OEWS no longer publishes at the 4-digit NAICS level, the "4-digit" designation indicates the most detailed industry breakdown available: either a standard NAICS 3-digit industry or an OEWS-specific combination of 4-digit industries. Industries that OEWS has aggregated to the 3-digit NAICS level (for example, NAICS 327000) will appear twice, once with the "3-digit" and once with the "4-digit" designation.
- own_code: Ownership type: 1= Federal Government; 2= State Government; 3= Local Government; 123= Federal, State, and Local Government; 235=Private, State, and Local Government; 35 = Private and Local Government; 5= Private; 57=Private, Local Government Gambling Establishments (Sector 71), and Local Government Casino Hotels (Sector 72); 58= Private plus State and Local Government Hospitals; 59= Private and Postal Service; 1235= Federal, State, and Local Government and Private Sector
- occ_code: The 6-digit Standard Occupational Classification (SOC) code or OEWS-specific code for the occupation
- occ_title: SOC title or OEWS-specific title for the occupation
- o_group: SOC occupation level. For most occupations, this field indicates the standard SOC major, minor, broad, and detailed levels, in addition to all-occupations totals. For occupations that OEWS no longer publishes at the SOC detailed level, the "detailed" designation indicates the most detailed data available: either a standard SOC broad occupation or an OEWS-specific combination of detailed occupations. Occupations that OEWS has aggregated to the SOC broad occupation level will appear in the file twice, once with the "broad" and once with the "detailed" designation.
- tot_emp: Estimated total employment rounded to the nearest 10 (excludes self-employed).
- emp_prse: Percent relative standard error (PRSE) for the employment estimate. PRSE is a measure of sampling error, expressed as a percentage of the corresponding estimate. Sampling error occurs when values for a population are estimated from a sample survey of the population, rather than calculated from data for all members of the population. Estimates with lower PRSEs are typically more precise in the presence of sampling error.
- jobs_1000: The number of jobs (employment) in the given occupation per 1000 jobs in the given area. Only available for the state and MSA estimates; otherwise, this column is blank.
- loc quotient: The location quotient represents the ratio of an occupation's share of employment in a given area to that occupation's share of employment in the U.S. as a whole. For example, an occupation that makes up 10 percent of employment in a specific

metropolitan area compared with 2 percent of U.S. employment would have a location quotient of 5 for the area in question. Only available for the state, metropolitan area, and nonmetropolitan area estimates; otherwise, this column is blank.

- pct_total: Percent of industry employment in the given occupation. Percents may not sum to 100 because the totals may include data for occupations that could not be published separately. Only available for the national industry estimates; otherwise, this column is blank.
- pct_rpt: Percent of establishments reporting the given occupation for the cell. Only available for the national industry estimates; otherwise, this column is blank.
- h_mean: Mean hourly wage
- a_mean: Mean annual wage
- mean_prse: Percent relative standard error (PRSE) for the mean wage estimate. PRSE is a measure of sampling error, expressed as a percentage of the corresponding estimate. Sampling error occurs when values for a population are estimated from a sample survey of the population, rather than calculated from data for all members of the population. Estimates with lower PRSEs are typically more precise in the presence of sampling error.
- h_pct10: Hourly 10th percentile wage
- h_pct25: Hourly 25th percentile wage
- h_median: Hourly median wage (or the 50th percentile)
- h_pct75: Hourly 75th percentile wage
- h_pct90: Hourly 90th percentile wage
- a_pct10: Annual 10th percentile wage
- a_pct25: Annual 25th percentile wage
- a_median: Annual median wage (or the 50th percentile)
- a_pct75: Annual 75th percentile wage
- a_pct90: Annual 90th percentile wage
- annual: Contains "TRUE" if only annual wages are released. The OEWS program releases only annual wages for some occupations that typically work fewer than 2,080 hours per year, but are paid on an annual basis, such as teachers, pilots, and athletes.
- hourly: Contains "TRUE" if only hourly wages are released. The OEWS program releases only hourly wages for some occupations that typically work fewer than 2,080 hours per year and are paid on an hourly basis, such as actors, dancers, and musicians and singers.

2.3 MST-EST2023-ALLDATA.csv

U.S. Census Bureau. (2023). Annual State Population Totals: 2020-2023 Dataset. Retrieved from https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#v2023

This U.S. Cansus Bureau dataset collects the estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, District of Columbia, and Puerto Rico from 2020 to 2023.

I will be using a single column, for use in Question 5:

- POPESTIMATE2023: 2023 resident total population estimate

3 Data Acquisition and Processing

Both datasets must be prepared before any operations are started, all NA values will either be replaced or removed. Each dataset will be explored and all unnecessary columns will be removed. Datasets will be used together for certain research questions.

4 Research Questions and Methodology

1. Is a job more demanding based on the higher level of education and the age of the person? Does the income level of a job determine how hard a person has to work? This is a task for the adult.data file. For the first question, using a scatterplot, the age can be plotted against the weekly hours worked. In turn the points are colored by the level of education. To create a more presentable visual, education level will be grouped into 4 categories (Preschool, Primary, Secondary, Higher). For the second part, using the salary categories (e.g., <=50K or >50K), a boxplot is created to show how income level relates to weekly hours worked across education categories. The visual will be as such, weekly hours worked will be plotted against the salary, colored by the 4 education categories. Estimated time: ~6 hours
2. Is there a noticable trend between a person's salary and their sex? What about their race? Multiple visuals will be implemented using the adult.data dataset. The total occurence of each sex will be plotted using a filled bar chart, with the total percentage being divided by the salary categories (<=50k and >50k). As for race, there will be a stacked bar plot showing the different races, each column representing the total number of race occurences and being stacked by the salary categories. Estimated time: ~5 hours

3. What characteristics makes a person have a salary greater than $50k or less than $50k, and with what accuracy can it be predicted for a person? A logistic regression model will be created using the adult.data file, the salary column will be the target variable and the rest of the column the features. The data will be standardized and split into training and testing sets, and the accuracy of the model will be recorded using precision and recall. Also, the feature coefficients will be noted for determining which features are most important in increasing the salary. Estimated time: ~9 hours

4. Which US state has the greatest average annual wage and hourly wage? Using the state_M2023_dl.xlsx dataset, the data will be filtered to only contain US states. It will then be further cleaned to only contain a occ_title of 'all occupations'. Then, two bar plots will be made to visualize all 50 states with their corresponding mean annual wages and mean hourly wages. In addition, to visualize the locations of the US states, two choropleth maps will be created to show a gradient of both annual and hourly average wages throughout the US. Estimated time: ~6 hours

5. Do high employment levels indicate high average salaries (both hourly and annual)? This question uses the filtered state_M2023_dl.xlsx dataset (only 'All Occupations' for OCC_TITLE). A set of bubble charts can visualize annual or hourly mean wage against the TOT_EMP (total employment) in a state, the size of the bubbles representing the median salary (annual or hourly). In order for the data to be compared fairly, the TOT_EMP of each state has to be normalized by being divided by the 2023 total state estimated population (POPESTIMATE2023) from the NST-EST2023-ALLDATA.csv table. Estimated Time: ~6 hours

6. IT Job Analysis: Which IT profession is most common across the US states? To prepare the data, the state_M2023_dl.xlsx dataset will be filtered to include only IT professions and only US states. A ridgeline plot will then be constructed to visualize the distribution of IT occupations across states based on the number of jobs per 1,000 total jobs, colored by the specific IT job. This will allow us to observe which IT professions are widely distributed across all states.

Additionally, the difference between A_PCT90 and A_MEDIAN (90th percentile and median annual IT wages) will be calculated for each IT occupation. This will help assess income disparity within IT professions, revealing whether the wage distribution is relatively even or if there is a significant gap between mid-level and top earners. Estimated Time: ~6 hours

7. Case Study: Does a rural state like Nebraska share similar professions to New York and New Jersey? This specific question is asked to see what jobs are common in an relatively urban area (NY and NJ), and a rural area (Nebraska). The state_M2023_dl.xlsx dataset will be filtered to contain only professions in NY, NJ, and NE. There will be two visual: a ridgeline plot and a horizontal bar plot (both using the JOBS_1000 column for finding the most common job). Ridgeline will be used for the distribution of all jobs across the three states. As for the bar plot, I wish to filter the top 5 job for each state and visualize which specific job is most common in all three states. Estimated Time: ~4 hours

5 Work plan

Week 4 (2/10 - 2/16):

- import necessary datasets, perform cleaning (4 hours)
- Begin work on Question 1 (3 hours).

Week 5 (2/17 - 2/23):

- Finish all necessary visuals for Question 1 (3 hours)
- Begin coding for Question 2 (3 hours)

Week 6 (2/24 - 3/2):

- Finish code for Question 2 (2 hours)
- Prepare data for regression model for Question 3 (5 hours)

Week 7 (3/3 - 3/9):

- Test the model (Question 3) and interpret results (3 hours)
- Begin Question 4 (2 hours)

Week 8 (3/10 - 3/16): *Presentations given on Wed-Thu 3/12-3/13.*

- Poster prep (4 hours)
- Presentation peer review (1.5 hours)
- Finish Question 4 (4 hours)

Week 9 (3/24 - 3/30): *Poster Draft 1 due Monday morning 3/24 at 9am. Poster Draft 2 due Sunday night 3/30.*

- Peer feedback (2.5 hours)
- Poster revisions (2.5 hours)
- Begin Question 5 (2 hours)

Week 10 (3/31 - 4/6): *Final Poster due Sunday 4/6.*

- Peer feedback (1.5 hours)
- Poster revisions (3 hours)
- Finish Question 5 (4 hours)

Week 11 (4/7 - 4/13):

- Do Question 6 (6 hours)

Week 12 (4/14 - 4/20):

- Do Question 7 and tidy up rest of code (5 hours)
- Begin final report (2 hours)

Week 13 (4/21 - 4/27): *Blog post draft 1 due Sunday night 4/28.* [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).

Week 14 (4/28 - 5/4):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/5 - 5/8): *Final blog post due Tues 5/7. Blog post read-throughs during final exam slot, Thursday May 8th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)
- [Do not schedule any other tasks for this week.]