Proposal: Salary Prediction and Analysis

DATA 450 Capstone

Nicholas Nemkov

February 13, 2025

1 Introduction

A person's salary is determined using various pieces of personal information. Factors such as education level, marital status, and country of birth can lend valuable understanding to current college students on their future careers. These insights help set competitive attitudes, assist in salary negotiations, and create new approaches toward wage equality. Through careful data analysis with the use of machine learning and visualizations, it is possible to predict a person's salary category using their personal information. The visualizations and predictions will show potential trends in people based on their characteristics and profession and help envision a base wage for a person, useful in advertizing potential career choices to students.

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.

3 Data Acquisition and Processing

The data must be prepared before any operations are started, all NA values will either be replaced or removed. Most of the data is taken from the adult.data file. The final column in this file (salary) currently contains two categories (<=50k or >50k) to represent if the person's salary is less than or equal to $50,000 or is greater than that. Using the numerical salaries ot the "Ask A Manager Salary Survey 2021 (Responses) - Form Responses 1.csv" file, the values will be converted to USD. Then, the numerical values will randomly populate the salary column in adult.data, with values being inserted based on the previous categories (ex. person with salary <=50k might get a value of 48056).

4 Research Questions and Methodology

1. Does working longer days result in a high salary? Is a job more demanding based on the level of education? Using a scatterplot, the annual salary in USD 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 3 categories (Primary, Secondary, Higher). Estimated time: ~7 hours
2. Is there a noticable trend between a person's salary and their sex? What about their race or age? Several visuals will be implemented. First, the data will be divided into male and female dataframes. The average salary will be calculated for males and female, this will be plotted on a bar graph. As for race and age, similar steps will be taken to create two more bar graphs (data divided by race and averages calculated). 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. Also, the feature coefficients will be noted for determining which features are most important in increasing the salary. Estimated time: ~9 hours

5 Work plan

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

- import necessary datasets, perform cleaning, and populate salary column with numbers (4 hours)
- Begin work on Question 1 (3 hours).

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

- Finish all necessary visuals for Question 1 (4 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):

- Presentation prep and practice (4 hours)
- Test the model (Question 3) and interpret results (3 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)
- Start writing report (2.5 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)
- Continue writing report (1.5 hours)

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

- Peer feedback (1.5 hours)
- Poster revisions (3 hours)
- Continue writing report (2.5 hours)

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

- Tidy up report and code (7 hours)

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

- write notes for blog post (6.5 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.]