Proposal: Salary Prediction and Analysis

DATA 450 Capstone

Nicholas Nemkov

February 6, 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 an potential employer. These insights help set competitive wages, assist in salary negotiations, and create fair laws toward wage equality. Through careful data analysis with the use of machine learning and visualizations, it is possible to predict a person's salary using their personal information. The visualizations and predictions will show potential trends in people based on their characteristics and help envision a base wage for a future employee, creating an efficient part of the hiring process.

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.

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 Ask A Manager Salary Survey 2021 (Responses) - Form Responses 1.csv

"Ask a Manager Salary Survey." TidyTuesday, 2021, https://github.com/rfordatascience/tidytuesday/blob/main
05-18/readme.md.

The data consists of responses given by many people on various questions concerning their jobs. In particular, one column will be used for its salary numbers.

- "What is your annual salary?": Represents a numberical values of yearly salary. Currency varies for people, this will all be converted to USD in the data processing.

3 Data Acquisition and Processing

[In this section, if applicable, describe how you will obtain the data (if it's anything more complicated than a simple download). Discuss what data processing steps will be needed, such as recoding variables, data cleaning, data tidying, imputing missing values, etc. See sections 1c, 1d, 1e in the "Good Enough Practices" paper.]

The data must be prepared before any operations are started, all NA values will either be replace or removes. 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

[In this section, list each of the questions you will explore. Following each question, provide a detailed and specific plan for how you plan to answer the question. Include the specific steps you will take, what form the answer will take (a number? table? visualization? model? Give all the specifics), and estimate how many hours each question will take to complete.]

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 high salary, and with what accuracy can a salary be predicted for a person? A linear regression model will be created, data will be 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 hours)
- Poster revisions (1.5 hours)

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

- Peer feedback (1.5 hours)
- Poster revisions (2 hours)

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

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

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.]