CS 362 / 562 - Assignment 5 Fall, 2025

Government organisations and individuals (perhaps even you) are interested in post-graduation student earnings. This dataset includes survey results of actual earnings for school attendees in California, along with some information about the person responding. The goal of this data challenge (assignment) is to test your ability to write software and models on this real world data. This challenge is designed to mimic those you may receive upon applying to positions as Data Scientist or Machine Learning Engineer. This data challenge is designed to take about two hours, but it is not timed. As always, deliver your results by the due date.

Accept this assignment by clicking on https://classroom.github.com/a/lNrn4Hl9. There are three parts of this assignment: 1. Data Exploration, 2. Model building (with a performance requirement) and 3. Reflection.

Implementation

We provided two CSV files:

- `earnings_train.csv` (~2.7MB)
- `earnings_test_features.csv` (~890KB)

In a Jupyter notebook (local to your machine) or in a Colab notebook (hosted in the cloud) or similar, you are required to do the following:

1. **Data Exploration**
2. **Build a model to predict `WAGE_YEAR4`** — the 4th year salary

Part 1. Data Exploration

To guide your exploration of the data, you will answer the questions below. Submit your answers in README.md in your github repo.

- Data quality: For each feature (column), what is the data type? Is there any missing data?
- Range: What are the unique values for each categorical column? What is the range of values of the numeric columns? Are the numeric column values normally distributed?
- Semantics: What is the meaning of the columns? Are any columns related to other columns? (If so, how?)
- If you are enrolled in CS 562, also provide answers to the following:
    - Which demographic shows the highest `WAGE_YEAR3?` Which demographic shows the lowest `WAGE_YEAR3?`

- Are there any people with negative wage trends? Describe these people by their demographics.
- Are there any people with positive wage trends? Describe these people by their demographics.

You may [read: should] use charts created in matplotlib, seaborn or similar to demonstrate your answers to these questions using the data provided. Include any chart you used in your README.md.

Part 2. Build a model to predict `WAGE_YEAR4`

You will use the file "`earnings_train.csv`" to create a model to predict `WAGE_YEAR4`. You may use any or all of the algorithms discussed in class during this model (i.e. Linear Regression, CART, KNN, etc.). You may not use other algorithms for this model.

You may also use any tools or software we used during practicals. For any other tools or packages, ask the teaching staff for permission. In general, we will grant you use of any tools or software on your computer, or that are freely available on the Internet, as long as the tool works with a Jupyter notebook or Colab and that it is clearly documented. We prefer that you use simpler tools to more complex ones and that you are "lazy" in the sense of using third party APIs and libraries as much as possible. The use of obscure, undocumented "black box" libraries is forbidden.

Once your model is built, you will test it by predicting earnings on data from the file "`earnings_test_features.csv`". Your model will be evaluated against the actual earnings, though you do not have the outcomes (targets) for this. Your prediction must be named "`preds.csv`", a file in CSV format with the one field: WAGE_YEAR4. This file must have one prediction for each facility appearing in the file "`earnings_test_features.csv`", in order. For example:

```
48758.
0.
38139.
```

Part 3. Reflection

Answer the following questions:

- Which features best predict the target outcome (`WAGE_YEAR4`)?
- What does your model say about the people or populations whose data is provided?
- What features, if any, would you like to have had to make a better model?

Submission

1. README.md: answers to Part 1 Data Exploration

2. Jupyter Notebook file or a link to Google Colab: For Part 2 model building and Part 3 reflection, perform all your work in a single Jupyter Notebook (placed in your github repository) or in a single Google Colab (linked from your github repository, shared with anyone with the link). Answer the questions in Part 3 in that notebook / Colab.
3. "`preds.csv`": Export your predictions for `WAGE_YEAR4` to a file in the repository and name it "`preds.csv`".

Grading

There are four (4) parts of this assignment. Each part is assessed on a 4-point scale (Exemplary, Satisfactory, Partial or Unassessable) based on Table 1.

Part	**Assessment**
Data Exploration	Your answers are well supported by your data analysis.
Implementation	Assessment will be based on implementation, decomposition, efficiency and style. ● Implementation: notebook / colab must have fully-executed cells, run without errors, producing the expected output. ● Decomposition: must be structured into cells of reasonable size and scope. ● Style: must use proper names and appropriate comments.
Model Performance	● CS 362: Your model must have an RMSE of 2500~3000 for Satisfactory, less than 2500 Exemplary ● CS 562: Your model must have an RMSE of 2300~2500 for Satisfactory, less than 2300 for Exemplary
Reflection	Assessment will be based on thoughtful discussion of possible uses and ethical considerations.

Table 1: Part Assessment

Overall, we are also interested in your ability to work on a team, which means considering how to package and deliver your results in a way that makes it easy for us to review them. This does NOT mean you are allowed to discuss with others or use their work, including those enrolled in this or similar courses. It does mean that undocumented code and data dumps are virtually useless; commented code and a clear writeup with elegant visuals are ideal. Also consider how asking targeted questions to members of our team may allow you to get more done in less time.

Overall grades for this assignment will also be based on the same 4-point scale according to the following rule set:

- Exemplary: 3 parts Exemplary + 1 part Satisfactory or higher
- Satisfactory: 3 parts Satisfactory + 1 part Partial or higher
- Partial: Any assignment with 2 or more parts submitted
- Unassessable: 2 or more parts not submitted