A06 - Journal of your L06 activities – Let’s be creative

Contributors: Martin Demel, Jiri Musil, Jasmine Campos, Munachimso Benson

Department of Science, Technology, Engineering & Math, Houston Community College

ITAI - 1371 Introduction to Machine Learning

Professor Vishwa Rao

October 12th, 2025.

# Introduction

As a team, we took on the role of data detectives and completed the Module 5 Lab 3 notebook from start to finish. Each of us ran every cell, then we met to compare notes and agree on what the evidence actually revealed. The case file was an employee dataset with a mix of numbers and categories. Our goal was to understand what the corrections meant and how they would benefit a future model.

# Body

A close-up of several graphs

AI-generated content may be incorrect.We began by scanning the scene. The data quality report revealed missing values in performance score, training hours, and education level, along with a small number of outliers and a few duplicate style rows.

This matched a plausible story, as shown in Figure 1. New employees often lack performance reviews, and random entry errors can occur. We confirmed this by examining the missing groups and found that individuals without performance scores had very little experience and were typically at the junior level. This gave us confidence to impute missing values using job level means for performance, the median for training hours due to skewness, and the mode for education. After imputation, nothing was missing, and the distributions remained reasonable.

Figure 1

Next, we encoded the categories. Education and job level had a natural order, so we mapped them to ordered numbers. Department and location did not, so we used one-hot encoding. We then examined the scale. Numerical features ranged across different values, so we experimented with standard scaling and min-max scaling, while leaving salary unchanged as the target. Outliers were handled with simple capping to prevent extreme weeks and project counts from dominating downstream models.

We created metrics such as experience to age ratio, productivity per year, work intensity relative to a normal week, training investment, and an interaction between performance and experience. We also established age and experience bands, as well as a high performer flag.

# Challenges

A screenshot of a computer code

AI-generated content may be incorrect.For Challenge 1, Figure 2, we calculated career progression speed as job level divided by years of experience, with a safety rule that sets the value to zero when experience is zero. It effectively identified fast movers.

Figure 2

For Challenge 2, we added a new work style column with options for Remote, Hybrid, and Office, and updated the pipeline to impute and one-hot encode this field.

For Challenge 3, we compared mean versus median imputation for the performance score and found that median better preserved the distribution in the presence of outliers as it can be seen in figure 3.

A graph of performance and performance

AI-generated content may be incorrect.Conclusion

Figure 3

From the lab, we learned to connect issues to causes and select appropriate fixes. Missing performance scores mainly came from new hires, so we imputed them based on job level, and skewed training hours were replaced with the median. We encoded ordered fields, one-hot encoded the rest, scaled data for comparison, capped outliers, and created simple ratios and an interaction that matched salary rules, confirmed by correlation and a quick random forest. With that foundation, we completed all three challenges, accelerating career progression with a safeguard for zero experience, incorporating work style into the pipeline, and conducting a mean versus median test that favored the median.

# References

Pradère, U. (2025, March 14). Mastering Prompt Engineering with Functional Testing: A Systematic Guide to Reliable LLM Outputs. Towards Data Science. <https://towardsdatascience.com/mastering-prompt-engineering-with-functional-testing-a-systematic-guide-to-reliable-llm-outputs/>

7.1. Pipelines and composite estimators. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/compose.html>

7.3. Preprocessing data. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/preprocessing.html>