Morris Patin

Professor Ayyagari

Intro to Machine Learning

24 October 2025

Lab 5 Report

In this lab, I focused on the importance of data preprocessing and feature engineering, which are critical steps in building reliable machine learning models. The challenges in this assignment helped me go beyond just cleaning data—I learned how to prepare and transform raw information into meaningful features that actually improve how well a model performs. Each challenge added a new layer of understanding to the data pipeline process, from creating calculated features to properly encoding and imputing data.

The first challenge involved designing a new feature called *career progression speed*. This feature measured how quickly an employee's job level advanced compared to their total years of experience. Creating this required careful handling of potential division errors, especially when years of experience equaled zero. I handled this by setting those cases to zero, preventing any mathematical issues while still keeping the dataset consistent. This step taught me how feature engineering can help reveal patterns that might not be obvious at first glance. It also showed that creativity and logic both play a role when designing features that can make models smarter.

In the second challenge, I learned how to handle a new categorical variable called `work_style`, which represented whether employees worked remotely, hybrid, or in the office.

To integrate this into the model, I updated the preprocessing pipeline to include one-hot encoding for the new column. This process converted each category into numerical values that a machine learning model could understand. It reminded me that as new data sources or variables are introduced, the preprocessing steps must be flexible enough to adapt. It also reinforced how categorical data, when handled properly, can bring valuable context to models that might otherwise overlook key behavioral differences between groups.

The final challenge compared two imputation methods—mean and median—for handling missing values in the `performance_score` column. By visually comparing the distributions after each imputation, I noticed that the median approach preserved the original shape of the data more effectively. The mean imputation pulled the data toward the center, slightly distorting the overall distribution, while the median kept the results closer to their original pattern. This exercise emphasized how even small decisions in data preprocessing can affect the outcome of analysis and model training. Choosing the right imputation method can make a major difference, especially when dealing with skewed or non-normal data.

Overall, this lab deepened my understanding of how crucial data preprocessing and feature engineering are in the overall machine learning workflow. It showed that the goal is not just to "fill in blanks" or "encode labels," but to prepare data in a way that maintains accuracy, fairness, and interpretability. I also realized that building strong data pipelines takes both technical skills and intuition—knowing when to use certain techniques and why they matter. Each challenge connected back to a real-world scenario where thoughtful preprocessing can improve insights and help models make better predictions. In the end, this lab helped me see data not just as

numbers in a table, but as meaningful information that can tell a more complete story when prepared carefully and intelligently.