MIDTERM 1371 ML

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

ITAI - 1371 Introduction to Machine Learning

Professor Vishwa Rao

October 27th, 2025.

Introduction

We aimed to create a clean, leak-proof preprocessing pipeline for the HR attrition dataset and to train an initial supervised model that predicts whether a candidate is looking for a job change. We used a structured, notebook-based approach that combined thorough data quality checks, text normalization, ordinal conversions, imputation, feature engineering within a scikit-learn pipeline, and a clear view of the transformed features before and after. The dataset included 19,158 rows and fourteen columns, with an imbalanced target: about 75% labeled as not seeking a job change and 25% as seeking one.

Body

We started by verifying that no lines were skipped during load, printing the exact row count and schema, then examining the class distribution. We assessed data quality by counting missing values per column, summarizing the cardinalities of all categorical fields, and plotting a heatmap of missingness along with histograms for two numeric reference points, as seen in Figure 1.

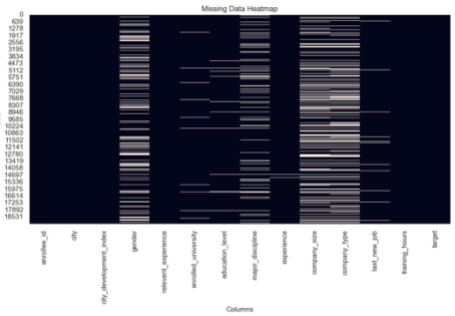
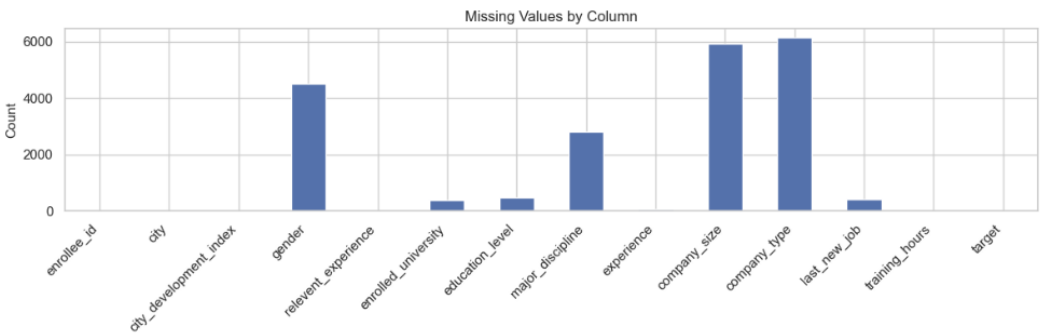


Figure 1



This initial review confirmed that company type, company size, gender, and major discipline had the most missing values. We standardized text across all categorical columns by converting to lowercase, trimming whitespace, and aligning common variants such as company size ranges and university enrollment labels. This standardization alone corrected 86,824 string values, making downstream encoding stable and reproducible.

We converted ordered categories into numeric form to maintain their natural order. Experience strings were mapped to approximate years, while the last new job interval was scaled to a small integer range with a cap for long gaps. Company size was estimated using midpoints, and education level was ordinalized. These conversions created four easy-to-interpret numeric fields for analysis and modeling. Missing values were handled next. We used median imputation for numeric features and mode imputation for categorical ones as a demonstration on a test copy, leaving the main modeling framework unchanged. The analysis showed that 20,733 missing values would be fixed by the imputation process, and the training-only pipeline would use the same strategies without leaking information from the test split. We next addressed scale and distribution. Training hours were highly right-skewed, so we applied a log1p normalization within the numeric branch of the pipeline, followed by standard scaling as seen in Figure 2.

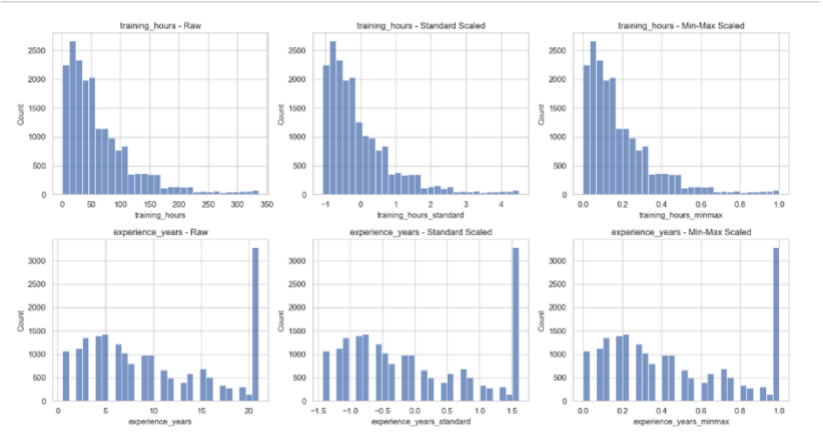


Figure 2

Because extreme values can dominate linear models, we implemented an IQR-based cap for training hours and experience years, fitted only on the training fold. The outlier illustration showed that 984 training-hour values exceeded the IQR upper bound in the full dataset, which motivated the capping step within the train-fit pipeline.

We integrated feature engineering into the pipeline after splitting the data. A custom DomainFeatureEngineer transformer generated three simple, interpretable features from existing columns. Training intensity was calculated by normalizing training hours with career length; job stability compared experience to the time since the last job change; and a city–experience interaction captured how development context might influence the effect of tenure. Besides these engineered features, we kept a dedicated branch that frequency-encoded the city and then scaled it, which helped manage a high-cardinality identifier without significantly increasing dimensionality. The column mapping combined the numeric branch with log-normalization and capping, the categorical branch with one-hot encoding and an infrequent bucket, and the city–frequency branch. The final processed matrix included 133 model features, with fifteen raw inputs feeding into sixteen total branch inputs before expansion. A side-by-side preview confirmed the one-hot outputs and the engineered columns.

Results

We trained two logistic regression baselines on the processed matrix: a standard model and a class-weighted variant to handle imbalance. The standard model achieved an accuracy of 0.777 and a ROC AUC of 0.775. The class-weighted model increased recall to 0.686 while maintaining a ROC AUC of 0.777, representing the best operating point for reliably detecting job-change seekers. Precision was 0.488 for the balanced model, which is expected given the class imbalance, and the ROC curve demonstrated stable separability on the held-out set, as seen in Figure 3. The overall outcome is summarized in Figure 4.

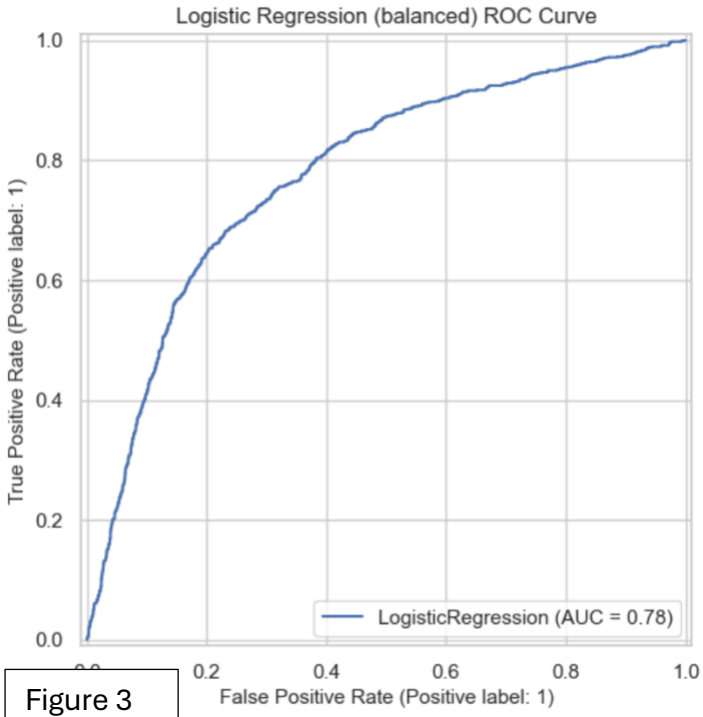


Figure 3

	accuracy	precision	recall	f1	roc_auc
LogisticRegression	0.777	0.585	0.363	0.448	0.775
LogisticRegression (balanced)	0.743	0.488	0.686	0.571	0.777

Figure 4

Conclusion

We developed a comprehensive preprocessing workflow that is mathematically sound and prevents data leakage, transforming messy tabular data into a consistent feature matrix for supervised learning. The method combines text cleaning, ordinal conversions, imputation, distribution-aware scaling, outlier detection, and domain-specific features within a scikit-learn pipeline. The result is a straightforward end-to-end process from raw data to 133 processed features, with clear before-and-after views and baseline metrics that are easy to interpret. This provides a solid foundation for future model improvements without needing a complete overhaul of the data layer.

Team contribution

The work was planned and executed collaboratively. The entire team reviewed dataset selection decisions and discussed modeling options to ensure the project aligned with mid-term assignment goals. Martin Demel and Jiri Musil led the technical design and implementation, covering data checks, feature engineering, pipeline development, and evaluation.