MIDTERM 1371 ML - Proposal

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

ITAI - 1371 Introduction to Machine Learning

Professor Vishwa Rao

October 27th, 2025.

Introduction

We will use an HR analytics dataset containing 19,158 candidates with demographics, education, work history, city development index, training hours, and a binary target indicating whether a person is seeking a job change. The classes are imbalanced, with roughly three-quarters not seeking a change and one-quarter seeking one. Our goal is to develop a clean, leak-proof preprocessing pipeline and a first supervised model that reliably predicts the target.

Body

We will validate the raw file and assess the quality of the profile data, then standardize categorical text so variants merge into a consistent vocabulary. Ordered fields will be mapped to numbers, including experience to years, company size to midpoints, and education to an ordinal scale. We will impute missing values with the median for numeric features and the mode for categorical features, showing counts before and after imputation. Distributions will be stabilized through a log transform of training hours and standard scaling of numeric features, with outlier capping based on IQR. We will engineer compact features within the pipeline, such as training intensity, job stability, and a city–experience interaction, and add a frequency-encoded view of city for high cardinality. For modeling, we will train a logistic regression baseline and report accuracy, precision, recall, F1, and ROC AUC.

Conclusion

The result will be a notebook with clear before-and-after visuals, a documented feature matrix, and baseline metrics that demonstrate credible separation between classes, ready for future iteration with prediction.

Reference

HR Analytics: Job change of data scientists. (2020, December 7). https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists?resource=download