Machine Learning Final Project: Predictive Analysis Report

Yesmine Hachana and Yufeng Jiang

Objective

The objective of our project is to build a predictive model for the target variable provided in the learning dataset and to generate predictions for unseen individuals.

The link to our github repository is the following:

https://github.com/jemappelleyesmine/Machine_Learning_Final_Project

Data sources

All the data used within our data preparation script is located in a subdirectory titled "project-56-files".

The final model was trained on 40,036 observations (80% of the labeled dataset) with 43 predictor variables after preprocessing. The labeled dataset of 50,046 individuals was split into a training set (40,036 observations) and a test set (10,010 observations, 20%).

The predictor variables include 14 numeric features (like age, working hours, remuneration, community size, and sport club count) and 29 categorical features (including sex, family type, qualification, occupation codes, employment characteristics, and geographic indicators). We dropped the original postal codes (insee_code, 13,702 unique values) after extracting geographic features (department, region, city type, and community size), and also replaced the detailed job description codes (job_desc and job_desc_retired, 408 unique codes each) with their simplified N2-level equivalents (approximately 50 categories).

Additionally, we created seven indicators as binary features to distinguish "real" missing values from absence of information (for example, distinguishing individuals without job data because they are unemployed versus those with incomplete employee records). This reduced the dataset from an initial 48 variables to 43 predictors.

No external data from sources beyond the project bundle were added. All data files used in this analysis were provided as part of the project dataset. However, we integrated the supplementary datasets included in the project bundle that were not part of the core person description files. Specifically, we incorporated job-related information (employment type and detailed job descriptions for both active workers and retirees), retirement pension data, sport club membership records, and geographic information.

The geographic data integration involved merging five separate files: city administrative boundaries (city_adm.csv), GPS coordinates (city_loc.csv), population statistics (city_pop.csv), department mappings (departments.csv), and region mappings (regions.csv). Additionally, we used the job description mapping (code_job_desc_map.csv) to transform the PCS-ESE 2017 codes (N3 level, 429 unique codes) into broader categories (N2 level, approximately 50 categories).

Data preprocessing

All preprocessing was set up within scikit-learn pipelines.

For numeric features (14 variables including age, working hours, remuneration, and community size), we applied median imputation. For categorical features (29 variables including sex, family type, occupation codes, and employment characteristics), we applied most-frequent imputation followed by one-hot encoding. Seven numeric-coded categorical variables (such as employee count and employer type) were manually reassigned from the numeric feature list to the categorical feature list. The same preprocessing was done for all models.

Model selection and hyperparameter tuning

Model selection was performed through hyperparameter optimization using 5-fold cross-validation with grid search. For each hyperparameter configuration, the model was trained on four folds (approximately 32,000 observations) and evaluated on the held-out fifth fold (approximately 8,000 observations), and the process was repeated five times. The same cross-validation folds were used for all candidate models. Gradient Boosting was chosen if its cross-validation RMSE plus a threshold (maximum of 0.005 absolute tolerance or the larger model's standard deviation) was lower than Random Forest's RMSE, otherwise the simpler Random Forest model would be preferred.

We evaluated three regression models, including two ensemble methods: a baseline Decision Tree with fixed maximum depth of 5, Random Forest, and Gradient Boosting. We produced several iterations during the hyperparameter tuning process in order to progressively refine parameter ranges based on edge detection to try to identify optimal configurations.

For Random Forest, the initial exploration tested n_estimators in the range 200-400, max_depth values of 15, 20, and None, min_samples_split from 2 to 10, and min_samples_leaf from 1 to 4 across 81 combinations. Future iterations revealed that performance plateaued with minimal gains. For Random Forest, the final parameters were n_estimators=2500, max_depth=None, min_samples_split=7, and min_samples_leaf=1, for a cross-validated RMSE of 0.3028.

For Gradient Boosting, the initial grid tested n_estimators from 200-400, learning_rate from 0.05-0.15, max_depth from 5-9, and min_samples_split from 2-10. During our fourth try, we introduced the subsample parameter and tested values of 0.8 and 1.0. During round five, we looked at the tradeoff between slow learning and many trees by testing n_estimators up to 2500 and learning_rate as low as 0.04. For Gradient Boosting, the optimal configuration consisted of n_estimators=2000, learning_rate=0.04, max_depth=8, min_samples_split=100, and subsample=0.8, for a cross-validated RMSE of 0.2749.

The Gradient Boosting model was a 34.7% improvement over the decision tree baseline (RMSE 0.4213) and seemed to outperformed Random Forest by 0.0294 RMSE points, beyond the established threshold of 0.0061 based on cross-validation variance. The iterative tuning process we did gave diminishing returns, with the final iteration producing an improvement of 0.0021 RMSE, so less than the model's variance.

Therefore, Gradient Boosting was selected as the final model.

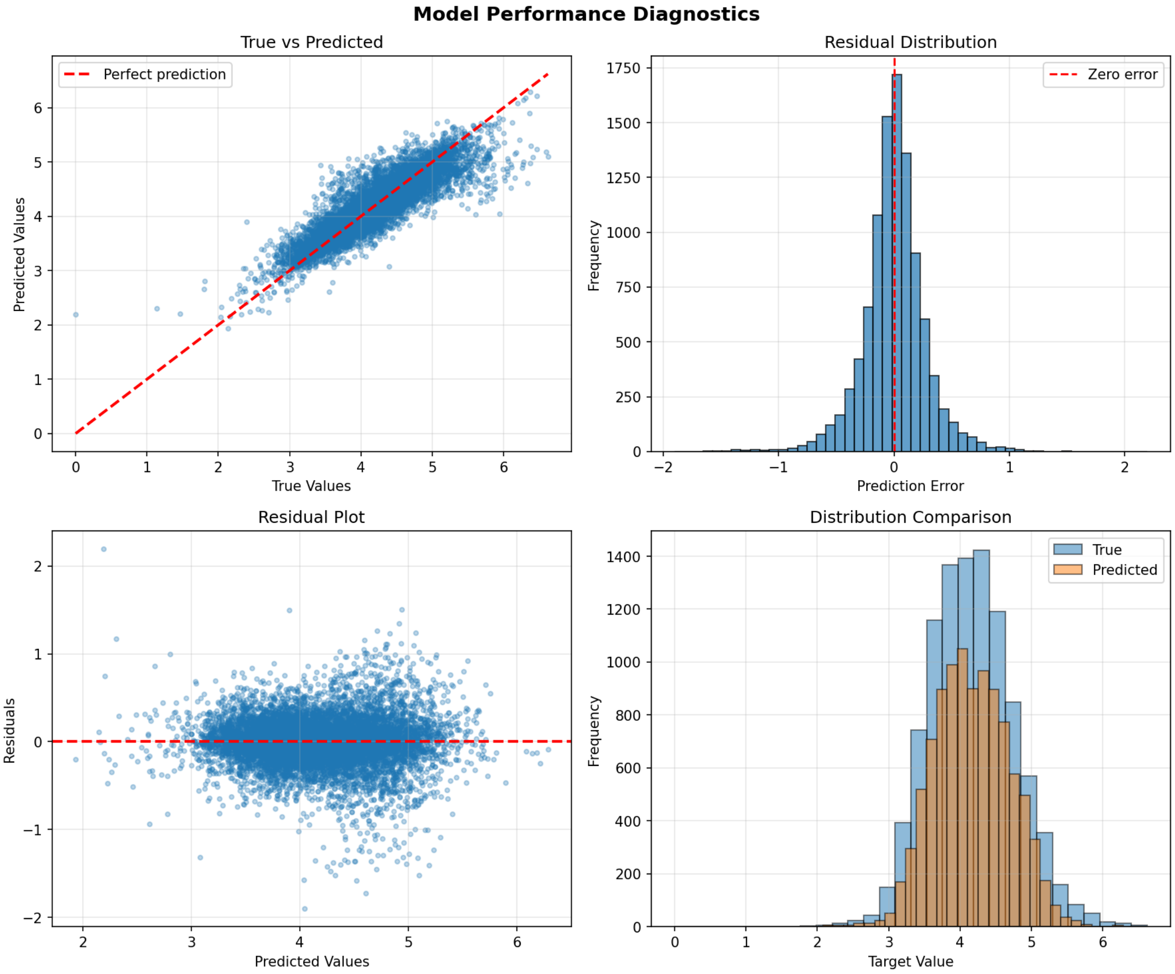
Expected predictive performance on new data

The expected performances of our final model on new data are a root mean squared error of 0.2740, a mean absolute error of 0.1932, and an R squared of 0.7901. We obtained these estimates by evaluating the chosen Gradient Boosting model on a held-out test set of 10,010 observations (20% of the data) that was never used during model training or hyperparameter selection.

The test set RMSE (0.2740) aligns quite well with the cross-validation RMSE from model selection (0.2749).

Assessment of the predictions of the model

Performance diagnostics



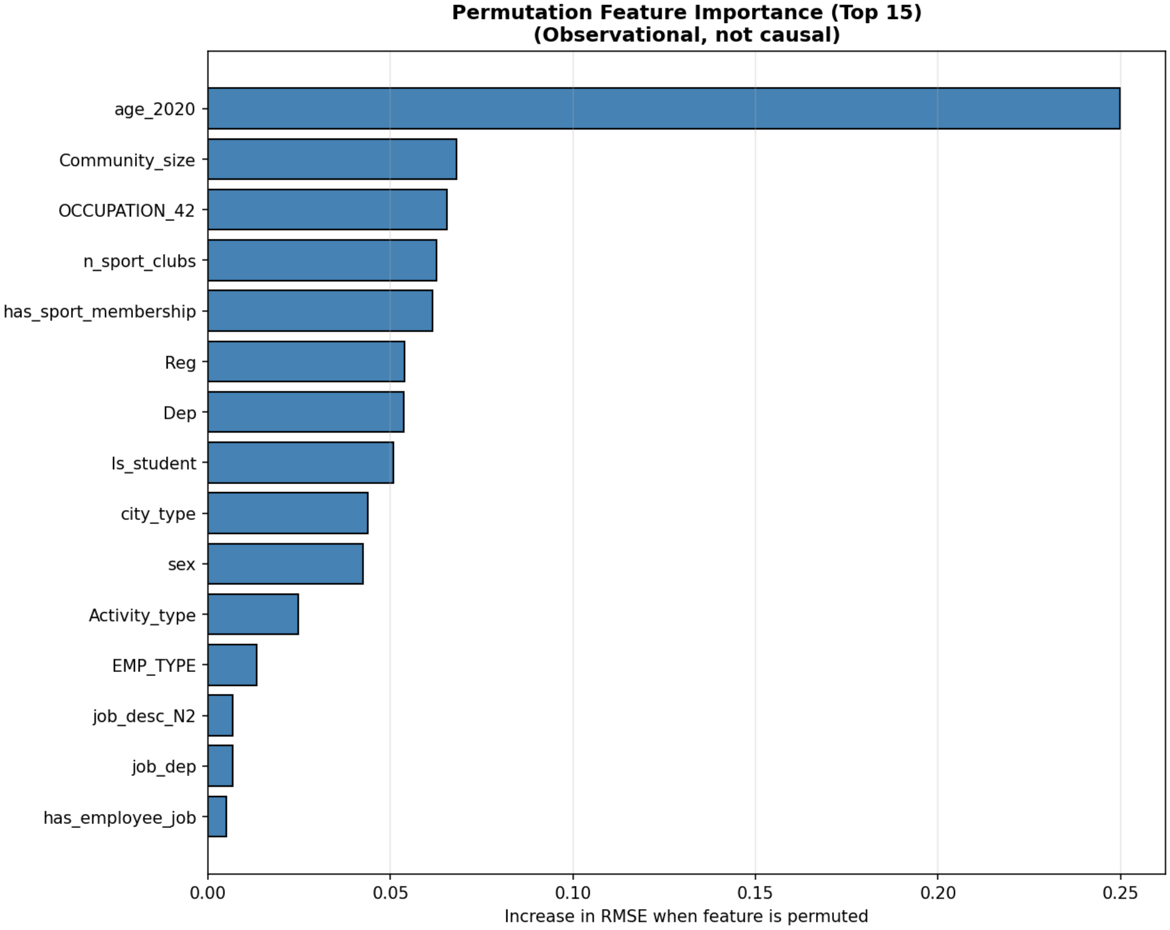
The model shows predictive performance with small bias. The True vs Predicted scatter plot (upper left panel) shows predictions clustering quite tightly around the perfect prediction line across the full range of target values. The residual distribution (upper right panel) illustrates an approximately normal distribution centered at zero (mean error = 0.0022). The residual plot (lower left panel) shows errors with no discernible patterns as a function of predicted values. The distribution comparison (lower right panel) illustrates overlap between true and predicted target distributions, though predicted values show reduced variance compared to true values.

Error analysis

The model achieves a median absolute error of 0.1354. Ninety-five percent of prediction errors fall within the interval [-0.5681, 0.5648].

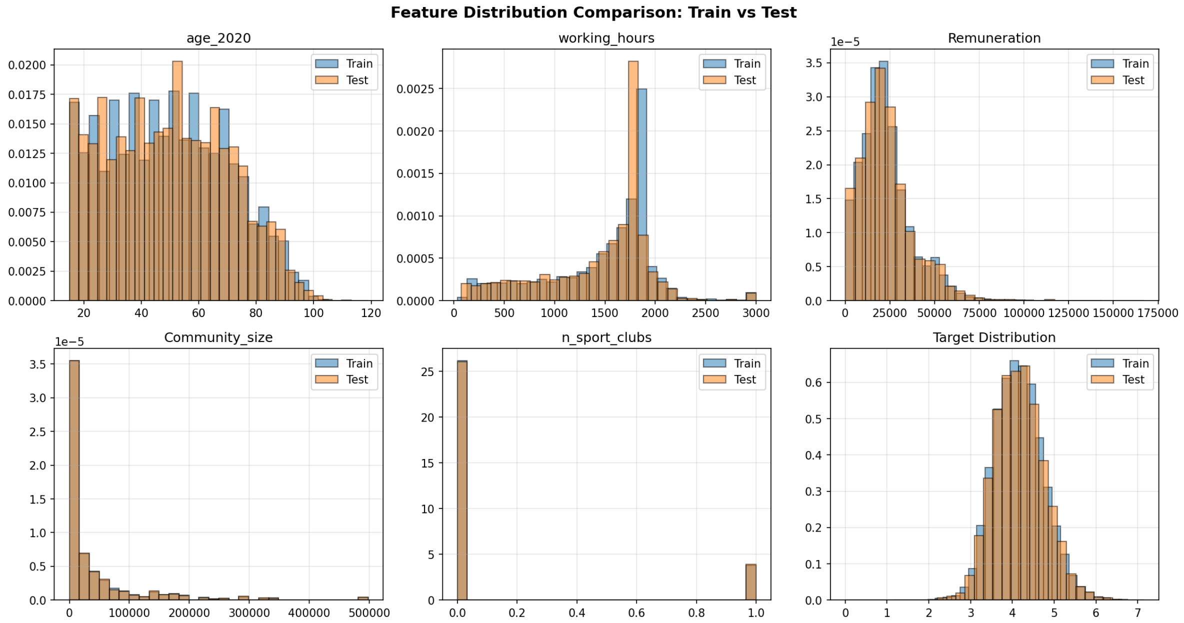
Analysis of the ten worst predictions reveals absolute errors ranging from 1.47 to 2.19, with the most severe case involving a true value of 0.0000 predicted as 2.1925. We examined the characteristics of these extreme errors. From our understanding, worst predictions predominantly occur for individuals from smaller communities (mean community size of 15,783 versus 40,998 overall), with lower current remuneration (18,637 versus 22,884) but much higher retirement pay (35,158 versus 18,774). Six of the ten worst cases are male, four belong to family type "typem4-1," and three share the same socio-professional category (csp_8_4). This pattern could imply that the model has difficulty with individuals that exhibit unusual combinations of employment and retirement characteristics.

Feature importance



We conducted permutation-based importance analysis on the full test set. It identifies age as the overwhelmingly dominant predictor, with ΔRMSE = 0.2499, nearly ten times the importance of the second-ranked feature. When age information is randomly permuted, model performance degrades from RMSE = 0.2740 to approximately 0.5239, and R^2 decreases by 0.5575 (from 0.7901 to 0.2326). Community size, socio-professional category (OCCUPATION_42), and sport club participation (both count and binary membership indicator) form the second tier of importance with ΔRMSE values ranging from 0.062 to 0.068. Geographic features (region and department codes) rank sixth and seventh with ΔRMSE values around 0.054. Employment-related features like job type, economic sector, and remuneration have surprisingly low importance (ΔRMSE $<$ 0.01). The simplified job description codes (job_desc_N2) also show minimal importance (ΔRMSE = 0.0068).

Distribution validation



We compared the training and test set distributions. The target variable shows nearly identical distributions between training (mean = 4.1676, std = 0.5940) and test (mean = 4.1669, std = 0.5981) sets, with differences of less than 0.2% of the mean and 0.7% of the standard deviation. Age shows overlapping distributions centered around 40-50 years, working hours peak near 1600-2000 annual hours for both sets, remuneration distributions are nearly identical with concentration between 20,000-40,000, community size shows a long-tailed distribution with most individuals in small to medium cities, sport club membership shows most individuals having zero clubs in both sets, and the target distributions are quite aligned in both location and spread.