Predictive Analysis Report

Machine Learning Final Project

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.

Data sources

The core data consists of the datasets provided as part of the assignment, these include a labeled learning dataset, an unlabeled prediction dataset, and several auxiliary tables which describe employment status, retirement history, sport participation, and geographic identifiers. All tables were merged using a unique individual identifier.

In addition to the original data, external geographic information was incorporated. This includes administrative classifications at the department and regional levels, city typology indicators, and population-based measures. We merged these variables by using official municipality identifiers.

Data preprocessing

All the preprocessing steps we implemented were set up inside model pipelines.

Numerical variables were imputed using the median. Categorical variables were imputed using the most frequent category and encoded using one-hot encoding. High-cardinality categorical variables, like detailed job codes and municipality identifiers, were simplified or excluded in favor of aggregated representations (for example, occupational groupings, department, and region). In addition, several indicators were created to try to capture the presence or absence of employment, retirement, pension income, and sport participation, as well as aggregated counts such as the number of sport clubs per individual.

Model selection and hyperparameter tuning

Model selection was performed through hyperparameter optimization using 5-fold cross-validation with grid search. The same cross-validation folds were used for all candidate models. Model selection was based on the cross-validated root mean squared error and its variability across folds. When performance differences were within statistical noise, the simpler model was preferred.

Two ensemble methods were evaluated: 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 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 optimal configuration consisted of n_estimators=2000, learning_rate=0.04, max_depth=8, min_samples_split=100, and subsample=0.8, for a superior 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 predictive performance of the final selected model was estimated using the held-out test set. The expected performances of the final model on new data are as follows: the root mean squared error is TO FILL , the mean absolute error is TO FILL , and the coefficient of determination is TO FILL .

Assessment of prediction quality

Figure diagnostics_performance.png presents the relationship between true and predicted values, the distribution of residuals, and residuals as a function of predicted values.

Variable importance

Variable importance was assessed using permutation-based methods. Figure diagnostics_importance.png displays the top fifteen most influential variables, measured by the increase in prediction error when each variable is permuted. The most important predictors include TO FILL .

Distribution comparison

To verify that the test set is representative of the training data, feature distributions were compared between the two samples. Figure diagnostics_distributions.png shows the distributions of selected key variables in the training and test sets.