EEE G513: Machine Learning for Electronics Engineers

T2 Evaluative Component : Salary Prediction Competition

Team:

Ayushman Muduli (2025H1400065G)
Akshit Sharma (2025H1400069G)

Our solution predicts salaries by modelling log1p(salary_average) and minimizing RMSPE, which focuses on percentage error. The key problems are high cardinality categories, strong geographic effects, and unseen city–role combinations. The method tackles these with hierarchical encodings, city grouped cross validation, and an optimized ensemble of gradient boosted models.

Data preparation:
Train and test are merged with the cost of living table. COL features are imputed by country medians, then global medians. Categorical fields (country, state, city, role) are label encoded. The target is transformed to log space.

Feature engineering:
The core features are hierarchical target encodings built only from training data: country role, state role, country, state, and role level means, medians, standard deviations, min–max ranges, and sample counts. Each row receives features through strict fallbacks so the model remains stable for rare or unseen combinations.

These encodings are expanded with interaction signals: ratios (for example country_mean divided by role_mean), deviations (country_role_mean minus country_mean), confidence weights (log1p of sample counts), variance features, and COL adjusted versions of the salary statistics.

Cost of living is summarized through mean, median, standard deviation, range, percentiles, interquartile range, coefficient of variation, and grouped category proxies (housing, services, goods). Additional features capture purchasing power and salary per COL unit. All infinities and missing values are replaced with zero.

Cross validation:
A five fold GroupKFold on the city prevents leakage and mirrors the real test setting. For each fold, target encodings are recomputed on the fold's training cities and features are built independently for train, validation, and test.

Models and ensemble:
The stack includes two LightGBM variants, one XGBoost model, and a GradientBoostingRegressor. All models predict log salary. Out of fold predictions are collected for every model.

Final predictions come from a weighted ensemble. The weights are not hand tuned; they are optimized with SciPy's SLSQP solver to directly minimize RMSPE on the out-of-fold predictions under the constraint that weights sum to one. These optimized weights are applied to the models' averaged test predictions and exponentiated back to salary space.