Introduction to Machine Learning

Challenge 2

Due Date: Nov 30, 2025 (11:55 PM)

Kaggle Competition Link:
https://www.kaggle.com/t/b078d923aae5480aa5774a84f0c5a0fd

Task 1: Linear Models & Regularization

1. Load the provided train.csv from the competition.
2. Create your own train-validation split:
    - Use 70% of the data for training, 30% for validation
    - Set random_state = YOUR_ERP_ID (e.g., random_state=210345)
    - Do **not** use the competition's test.csv for tuning
3. Implement and systematically compare the following **only**:
    - Ordinary Linear Regression (baseline)
    - Polynomial Regression (degree 2 and 3), with and without interaction terms
    - Ridge Regression (L2)
    - Lasso Regression (L1)
    - Elastic Net
4. Build a proper preprocessing pipeline that handles missing values, one-hot/ordinal encoding, scaling, feature engineering, etc.
5. Use GridSearchCV or RandomizedSearchCV to tune:
    - Regularization strength α
    - Polynomial degree and interaction on/off
6. Track and report RMSE on your validation set:
    - Baseline (plain linear)
    - Best polynomial configuration
    - Best Ridge / Lasso / Elastic Net
    - Final best linear-model RMSE
7. Using the exact same train/validation split, train ONLY one gradient boosting model of your choice (CatBoost, LightGBM, XGBoost)
8. Deliverable: A short report (Report 1) showing your pipeline, key results, and insights. How did the best model of Step 6 performed in comparison to the boosting model (in terms of training and inference time and performance). Also provide an organized Python code/notebook

Reminder: Your train-validation split (Step 2 above) for Task 1 must use your ERP ID as the random seed to ensure that everyone works on a slightly different dataset and discourages copying.

Task 2: Kaggle Submission

1. Now load the full provided train.csv from the competition (do not use your Task 1 split).
2. Train different models on the entire training set (100% of train.csv). You may use any model from standard ML libraries. But make sure that the models we have covered in the course must be evaluated, including but not limited to:
   a. Regression Tree
   b. Linear Regression Variants (your best model from Task 1 now trained on full dataset)
   c. GradientBoosting
   d. LGBM
   e. CatBoost
   f. XGBoost
3. Appropriately handle missing values, one-hot/ordinal encoding, scaling, feature engineering, etc.
4. Generate predictions on the competition's test.csv and submit to Kaggle.
5. Deliverable: A detailed report (Report 2) that include:
   a. A comparison table of MSE (on your own validation set) for all models tried
   b. Your final Kaggle public leaderboard score
   c. Analysis: Which model performed best?
   d. Also provide an organized Python code/notebook.