AutoML Report 2

Łukasz Zalewski 329532

January 2024

1 Manual solution

Methodology:

1. First, I did some quick manual tests with knn, random forest, xgb, and decided to proceed with knn as it seemed to work the best.

2. Then executed Sequential Forward Floating Feature Selection (SFFS) from *mlextend* library, using knn with 10 neighbors as evaluation model, for the following number of features: $(20, 22, 24, 26, 28, 30)$. This gave me 6 different sets of most important features.

3. The next step was simple grid search over knn hyperparemeters for each set of features. After finding best hyperparameters for each subset of features, the final evaluation has been conducted using 10-fold cross-validation.

Hyperparameter	**Range**
n_neighbors	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
weights	uniform, distance
p	1, 2

Table 1: Considered knn hyperparameters.

4. The best combination of subset of features and knn hyperparameters was selected. The best achieved 10-fold balanced accuracy was **89.40%** for 24 features, 18 knn neighbors, p=2, and 'distance' weights.

5. Finally, predictions for test set were generated. The final score on the **5%** of testset in the app turned out to be **93.33%**.

The whole hyperparemeter search and evaluation process takes about 10 minutes.

2 AutoML solution

I used *autogluon*.

Parameter	**Value**
presets	'best_quality'
num_stack_levels	3
num_bag_folds	4
time_limit	1200 seconds (20 minutes)

Table 2: AutoGluon fit method parameters.

Model	Validation Score
CatBoost_r9_BAG_L3	0.8670
WeightedEnsemble_L4	0.8670
WeightedEnsemble_L5	0.8670
WeightedEnsemble_L3	0.8665
LightGBM_BAG_L4	0.8665

Table 3: Top5 models from autogluon and their validation scores.

Final balanced accuracy for 5-fold cross validation: **86.7%**
Final balanced accuracy for 5% testset predictions in the app: **96.67%**