Advanced Machine Learning
Project 2: Cost-Aware Customer Selection

Bruno Tobiasz, Mikołaj Trębski

Warsaw University of Technology

June 2025

Outline

Problem overview

- **Goal:** Identify electricity customers likely to exceed usage threshold
- **Business case:** Target 1,000 households for energy-saving offers
- **Challenge:** Balance accuracy vs. feature acquisition costs

Dataset

- 5,000 training samples
- 500 anonymized features
- Binary target (usage above/below threshold)

Cost-aware evaluation

Score function

$$\text{Score} = 10 \times \text{True Positives} - 200 \times \text{Number of Features} \qquad (1)$$

Example 1:

- 850 correct predictions
- 12 features used
- Score: €6,100

Example 2:

- 300 correct predictions
- 2 features used
- Score: €2,600

Key insight: Feature cost dramatically impacts profitability!

Feature selection strategy

Methods used

1. `SelectKBest()` - Top k features by statistical score
2. `SelectFromModel()` - Random Forest feature importance

Key finding

Only 13 out of 500 features had importance > 0.003

Observation: Performance degraded with >25 features due to:

- Lower accuracy
- Dramatically increased costs

Machine Learning algorithms

5 algorithms were tested:

1. Logistic Regression
2. Random Forest Classifier
3. AdaBoost
4. Gradient Boosting Classifier
5. Bagging Classifier

Hyperparameters tuned:

- Regularization parameters (C, penalty)
- Number of estimators
- Maximum tree depth

Model performance comparison

Algorithm	Score	Accuracy	Features
Logistic Regression	4265	69.2%	2
Random Forest	4380	72.4%	4
AdaBoost	3860	70.6%	2
Gradient Boosting	3342	71.7%	4
Bagging	3055	70.7%	6

Winner

Random Forest achieved the highest cost-adjusted score of **€4,380**

Final model selection

Chosen model: Random Forest

Configuration:

- 400 estimators
- Maximum depth: 5
- 4 features (SelectKBest)

Performance

- **Score:** €4,380
- **Accuracy:** 72.4%
- **Features used:** 4 out of 500

Why random forest worked best

Advantages for this problem:

- **Ensemble method** - Reduces overfitting
- **Built-in feature selection** - Natural importance ranking
- **Robust to noise** - Important with 500 features
- **Good bias-variance tradeoff** - Especially with limited features

Key insight

Ensemble methods can achieve high performance while maintaining parsimony in feature selection

Key findings

1. **Feature costs matter:** Cost-aware selection is crucial
2. **Less is more:** 4 features outperformed larger sets
3. **Ensemble advantage:** Random Forest provided best cost-benefit
4. **Effective evaluation:** Score function balanced accuracy vs. cost

Thank you for your attention!

Questions?