University of Amsterdam

AML 2025: Unified Tabular Learning Group 9

Datasets

- Higgs - particle physics dataset for signal vs background classification. Consists of 28 feature variables (21 kinematic + 7 derived). 175,000 training samples.
- Heloc - credit risk dataset for Good vs Bad repayment prediction. Consists of 23 feature variables (credit history metrics). 9 000 training samples.
- CoverType - forest cover type dataset for 7-class classification. Consists of 54 feature variables (10 continuous + 44 binary). 58 000 training samples.

- covtype_test_submission.csv
- covtype_train.csv
- heloc_test.csv
- heloc_test_submission.csv
- heloc_train.csv
- higgs_test.csv

Baseline model

TabSTAR

TabSTAR: A Tabular Foundation Model for Tabular Data with Text Field:

Alan Arazi, Eilam Shapira, Roi Reichart

- Parameter-free model
- Pretrained on multiple tabular datasets in a multitask setup (classification + regression)

Research Question

Which modeling approach achieves **the best overall performance** across three tabular classification datasets, and does it **outperform** the baseline model?

Exploration

Classification problem with up to 11 classes. Variants explored:

- 9-class variant (binary outcomes for HIGGS and Heloc unified)
- 11-class variants

Problems encountered:

- Target coding across 11 classes of 3 datasets → **artificial coding and handling irrelevant columns**
- Class and data imbalance → **smoothed class reweighting**
- Potential proxy features (e.g. HIGGS "weight") → **removed after sanity checks**

Explored feature importance

- Mutual Information + PCA

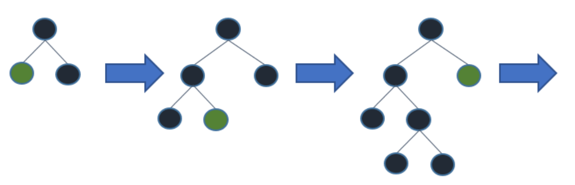
Models explored

Logistic Regression, Random Forest, Gradient Boosting, LightGBM)

	dataset	setting	logreg	rf	gb	lgbm
0	heloc	RAW	0.701009	0.705258	0.701540	0.693043
1	heloc	FS+PCA	0.707913	0.708444	0.707913	0.699416
2	covtype	RAW	0.725497	0.884778	0.771620	0.882196
3	covtype	FS+PCA	0.709922	0.847690	0.745547	0.839687
4	higgs	RAW	0.751400	0.839971	0.832914	0.842971
5	higgs	FS+PCA	0.727000	0.829429	0.819371	0.831857

Selected model = RAW LGBM

Light Gradient-Boosting Model Architecture



(lightgbm.readthedocs.io)

Based on decision trees → Builds multiple trees in sequence

- Initialize with one tree
- Train a new tree with the gradients
- Reiterate until maximum trees or no more loss

Why LGBM?

- LGBM minimizes the multi-class loss → Accurate probability estimates
- Leaf-wise tree growth → faster computation
- Uses both gradients and Hessians to compute leaf updates → Efficient learning

Our model parameters →

```
params = {
    "objective": "multiclass",
    "num_class": 11,
    "metric": "multi_error",
    "learning_rate": 0.05,
    "num_leaves": 63,
    "min_data_in_leaf": 50,
    "feature_fraction": 0.8,
    "bagging_fraction": 0.8,
    "bagging_freq": 1,
    "lambda_l2": 1.0,
    "verbosity": -1,
    "seed": 42
```

How an instance moves through the model

1. Take row to the shared feature space
2. Initialize 11 per-class scores
3. For one tree: follow feature splits to a leaf.
4. Add leaf values to class scores (boost corrections)
5. **Repeat steps 3–4 across all trees (accumulate scores).**
6. *Softmax final* scores => get class probabilies
7. Pick max probability via *argmax*
8. Assign class prediction to the row

Error and results analysis

TabSTAR

(scores on validation from train-test split)

	precision	recall	f1-score
bad/background	0.96	0.97	0.97
good/signal	0.94	0.93	0.94
Spruce/Fir	0.65	0.78	0.71
Lodgepole Pine	0.78	0.68	0.73
Ponderosa Pine	0.61	0.90	0.72
Cottonwood/Willow	0.00	0.00	0.00
Aspen	0.00	0.00	0.00
Douglas-fir	0.47	0.17	0.25
Krummholz	0.66	0.64	0.65
accuracy	0.89	0.89	0.89

Accuracy (Kaggle)

0.73190

LGBM

(scores on validation from train-test split)

	precision	recall	f1-score
bad	0.73	0.76	0.74
good	0.71	0.68	0.70
background	0.88	0.87	0.88
signal	0.76	0.78	0.77
Spruce/Fir	0.86	0.85	0.85
Lodgepole Pine	0.87	0.88	0.88
Ponderosa Pine	0.89	0.91	0.90
Cottonwood/Willow	0.87	0.85	0.86
Aspen	0.76	0.64	0.70
Douglas-fir	0.78	0.78	0.78
Krummholz	0.89	0.90	0.89
accuracy	0.84	0.84	0.84

Accuracy (Kaggle)

0.94152

Comparison with baseline model

- Prediction works best with **11 classes** (LGBM) vs 9 classes (TabSTAR).
- **Balancing** applied through weights helps predicton.
- Comparable on train data, but LGBM is much better in test set = **better generalizability**
- TabSTAR does very well on *heloc* + *higgs*, but poorly on *covtype*, which hurts overall generalizability.
- **LightGBM wins overall**

Potential improvements

- Constraining the model slightly could lead to better generalization
- Stronger validation set → Stratify also using the target variable
- Use NaNs or other imputed values instead of imputing 0, as this value is taken into consideration by the model and can affect the findings

Conclusion

A single unified LightGBM model trained on an 11-class dataset with feature reweighting **can learn effectively across HELOC, HIGGS, and Covertype datasets, outperforming** an unbalanced but pre-trained baseline TabSTAR model.

Contributors: Annabelle Donato. Jeremi Wasilewski, Henri Siltaniemi

Group number: 9

Github repository: https://github.com/bLacKr0sE17/AML2025_unified_model_group9