AutoML research

Gakubu

Mikołaj **Ga**łkowski
Kacper **Ku**rowski
Hubert **Bu**jakowski

Faculty of Mathematics and Information Science
Warsaw University of Technology

June 9, 2022

Overview

1. Motivation
2. Autogluon
3. Our framework
4. Benchmark – OpenML
5. Summary

Motivation for AutoML

- rapid growth of machine learning
- complexity of machine learning algorithms is often beyond the reach of non-experts
- accelerating the research done by researchers
- filling the gap between “supply” and “demand” in Data Science market

Autogluon

- AutoML for Text, Image, and Tabular Data
- Easy-to-use and easy-to-extend
- Intended for both ML beginners and experts
- Extensive documentation: `https://auto.gluon.ai/stable/index.html#`

Autogluon

- AutoML for Text, Image, and Tabular Data
- Easy-to-use and easy-to-extend
- Intended for both ML beginners and experts
- Extensive documentation: `https://auto.gluon.ai/stable/index.html#`

```
from autogluon.tabular import TabularDataset, TabularPredictor

train_data = TabularDataset('train.csv')
test_data = TabularDataset('test.csv')
predictor = TabularPredictor(label='class').fit(train_data=train_data)
predictions = predictor.predict(test_data)
```

Our framework

- solving binary classification problems

Our framework

- solving binary classification problems
- preprocessing:
  - removing outliers
  - removing identifier columns
  - imputing NaNs in numeric columns with mean
  - KNNImputer in categorical columns
  - scaling numeric columns
  - one-hot encoding categorical columns

Our framework

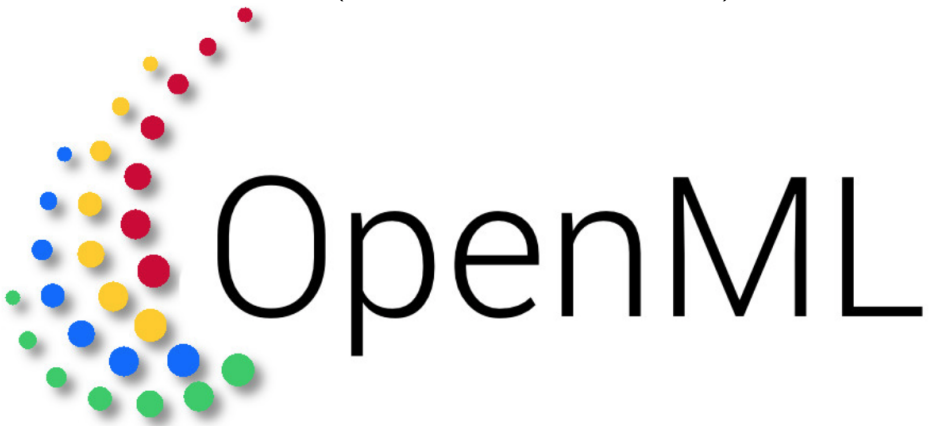
- solving binary classification problems
- preprocessing:
  - removing outliers
  - removing identifier columns
  - imputing NaNs in numeric columns with mean
  - KNNImputer in categorical columns
  - scaling numeric columns
  - one-hot encoding categorical columns
- models inside:
  - Logistic Regression, Random Forest Classifier, XGBClassifier, SVC, Voting Classifier (based on previous models)

Our framework

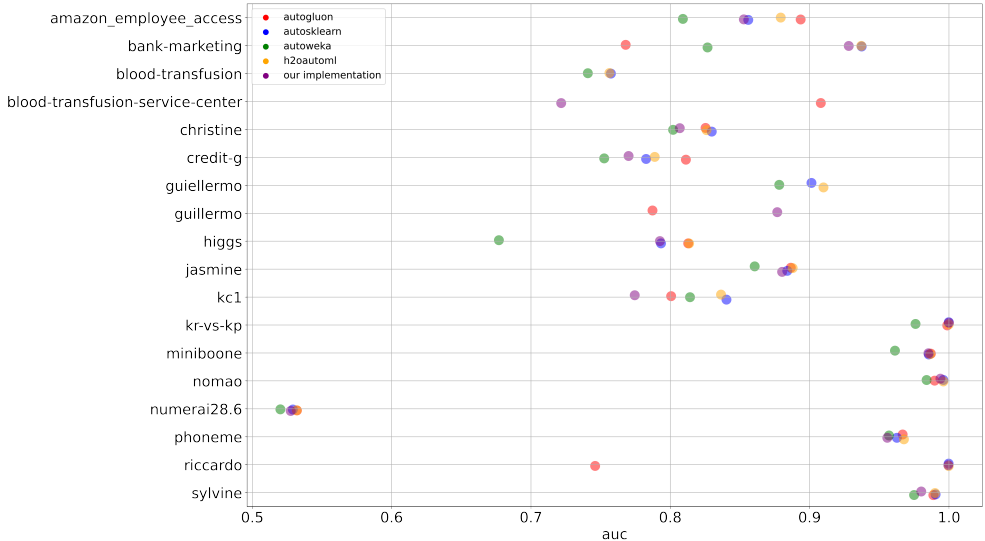
- solving binary classification problems
- preprocessing:
  - removing outliers
  - removing identifier columns
  - imputing NaNs in numeric columns with mean
  - KNNImputer in categorical columns
  - scaling numeric columns
  - one-hot encoding categorical columns
- models inside:
  - Logistic Regression, Random Forest Classifier, XGBClassifier, SVC, Voting Classifier (based on previous models)
- metrics:
  - roc auc score, f1 score, recall, precision, accuracy

OpenML benchmark

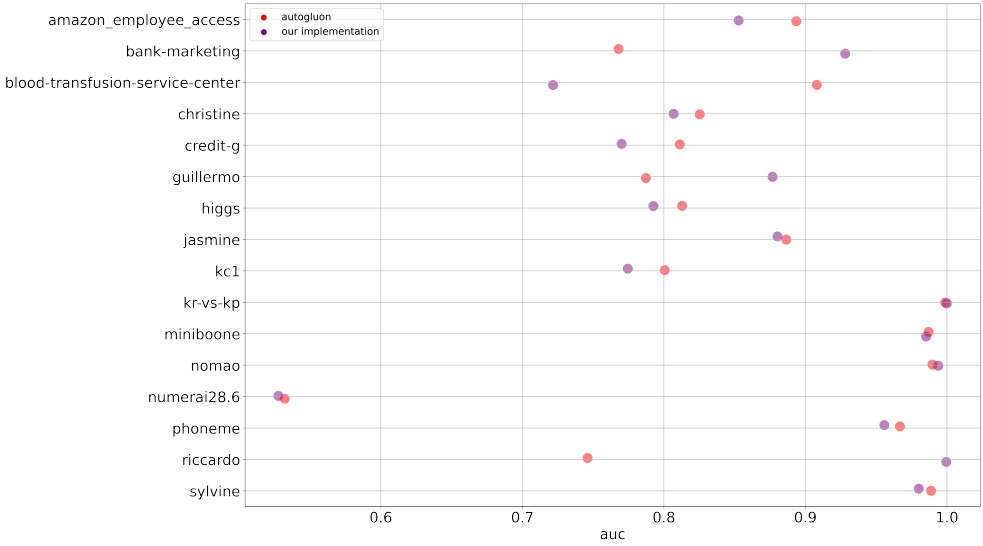
- based on 39 datasets, 20 of them are binary classification problems
- open-source and extensible (you can add your own datasets)



Benchmark results – OpenML



Benchmark results – AutoGluon vs our framework



Summary

- Rapid growth of AutoML framework
- Once implemented, AutoML frameworks are very easy to use
- Our implementation isn't much worse than AutoGluon – sometimes it even is better!
- However, humans are (still) not replaceable

The end!