Surrogate Optimization in Generative Design Problems for Composite Machine Learning Models

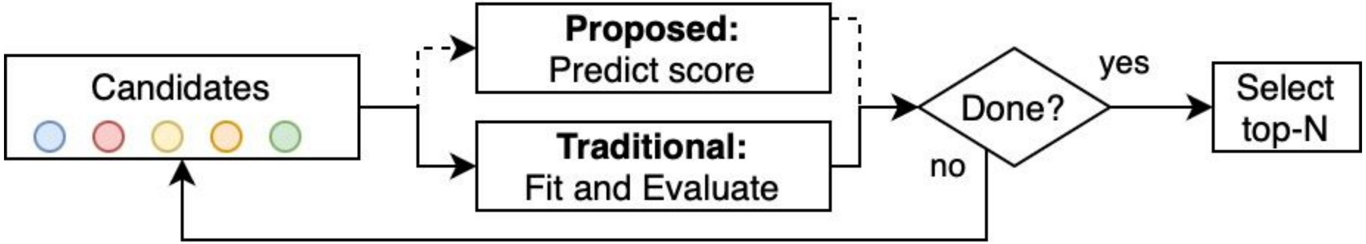
N. Nikitin, K. Cherniak, R. Mardyshkin, E. Shikov, P. Shevchenko, I. Iov, M. Pinchuk, G. Kirgizov, A. Stebenkov, A. Kalyuzhnaya
NSS Lab
ITMO University

Sochi
2025

Contents

Introduction

AutoML is the process of automatically generating and optimizing the various components and steps involved in a machine learning pipeline



ML-pipeline: chained machine learning models where the previous model output serves as the next model input

FEDOT

N. Nikitin, P. Vychuzhanin, M. Sarafanov, I. Polonskaia, I. Revin, I. Barabanova, G. Maximov, A. Kalyuzhnaya, A. Boukhanovsky, Automated evolutionary approach for the design of composite machine learning pipelines, Future Generation Computer Systems, 2021

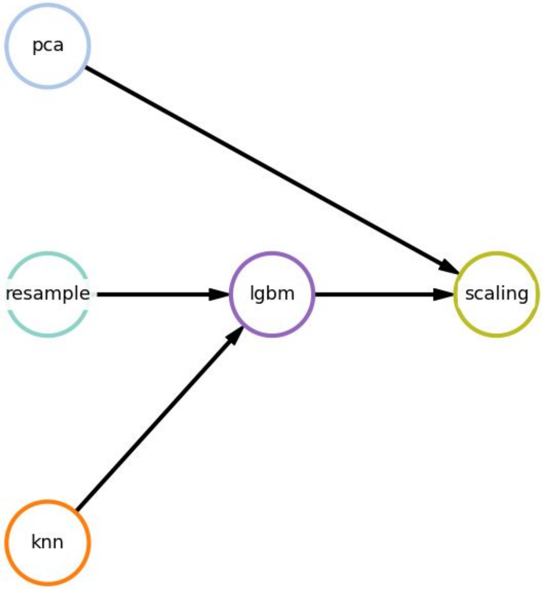
FEDOT Framework

Pipeline:

- Directed acyclic graph of arbitrary size and arbitrary rank.
- Nodes: ML-operations
- Edges: data flow.
- Operations' hyperparameters have different semantics, thus pipeline graph is heterogeneous

Pipeline optimization:

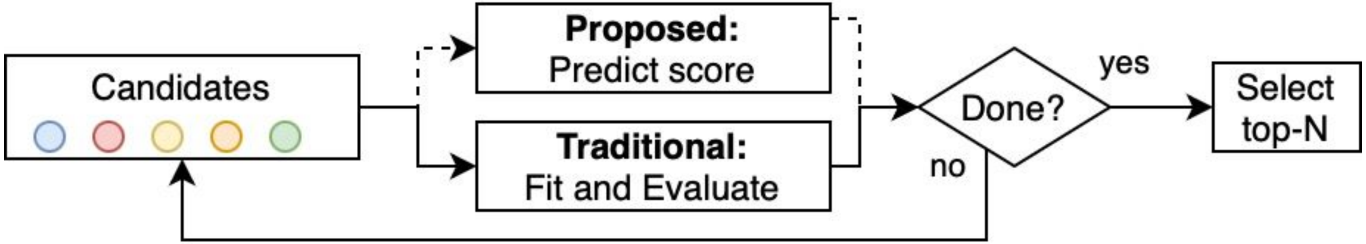
- Initial population of solutions forms
- Each pipeline is trained and assessed
- Selection: top pipelines are selected
- Crossover: some pipeline pairs exchange nodes
- Mutation: some nodes and links replaced randomly
- Process repeats with enhanced population



Problem Statement

- Pipeline optimization takes a significant amount of time and compute
- Tabular data has a lot of matching patterns
- Previously obtained knowledge is not used to enhance the initial guess

Could composite ML-pipelines be ranked based on own features and dataset description query?



Existing solutions: RankGNN

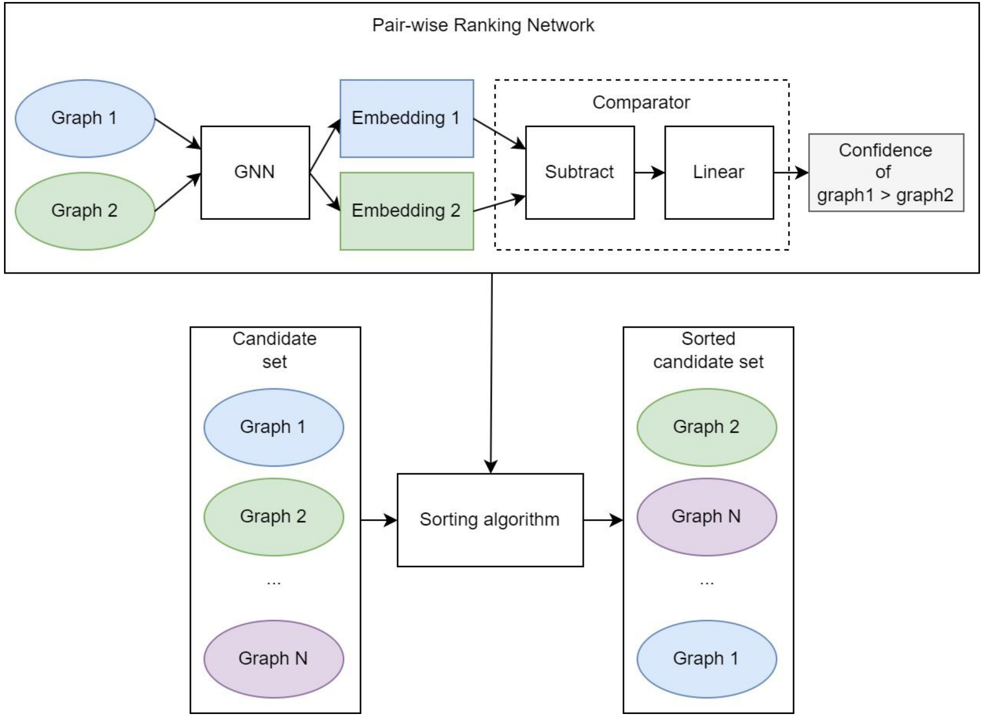
Main idea:

Use the surrogate model as comparator in a graph sorting algorithm.

Weakness:

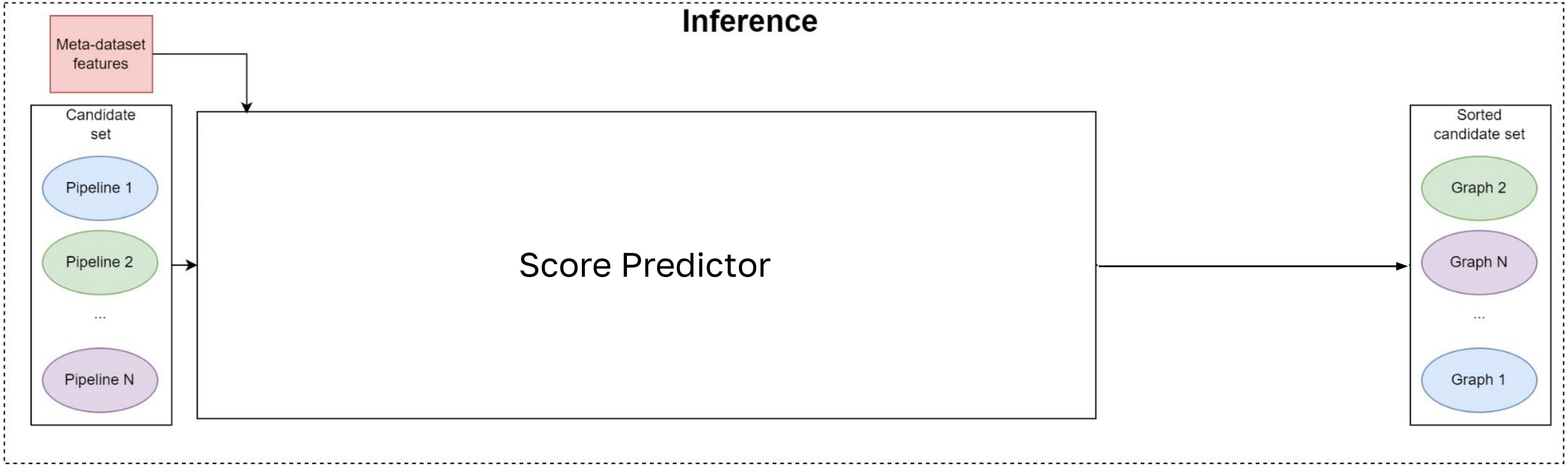
1) No query to rank candidates is accepted

2) Pair-wise ranking complexity is up to $O(N^2)$



[1] Damke, Clemens, and Eyke Hüllermeier. "Ranking structured objects with graph neural networks." *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*. Springer International Publishing, 2021.

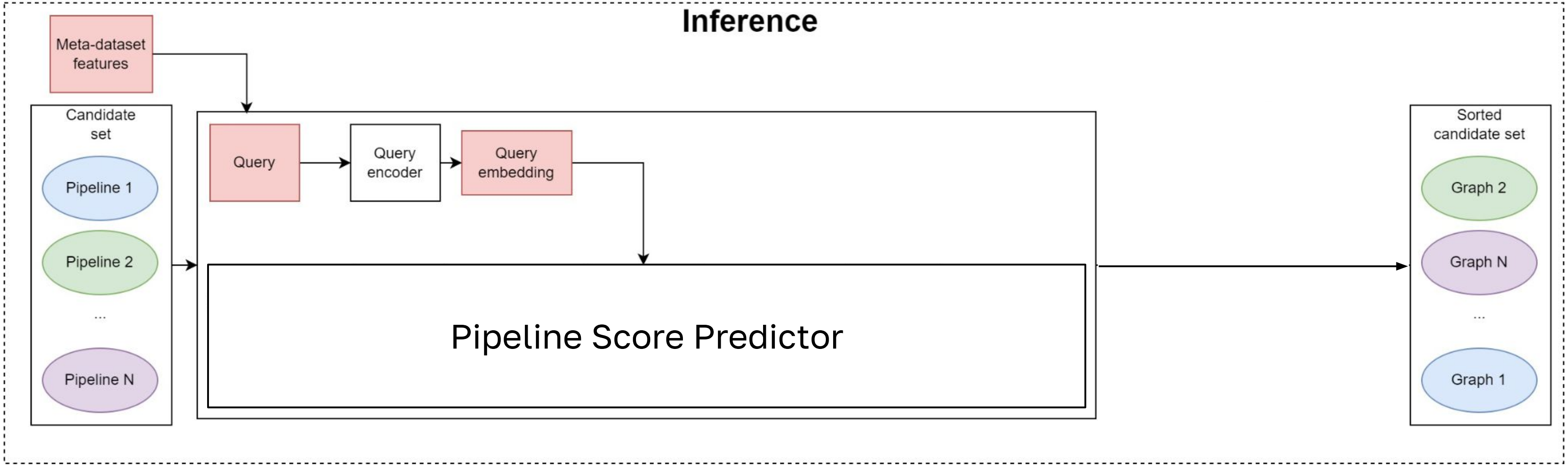
Proposed solution



Main idea:

1) Train a model to assign a score for a candidate with respect to a query
2) Introduce a query to ranking algorithm in early-fusion fashion

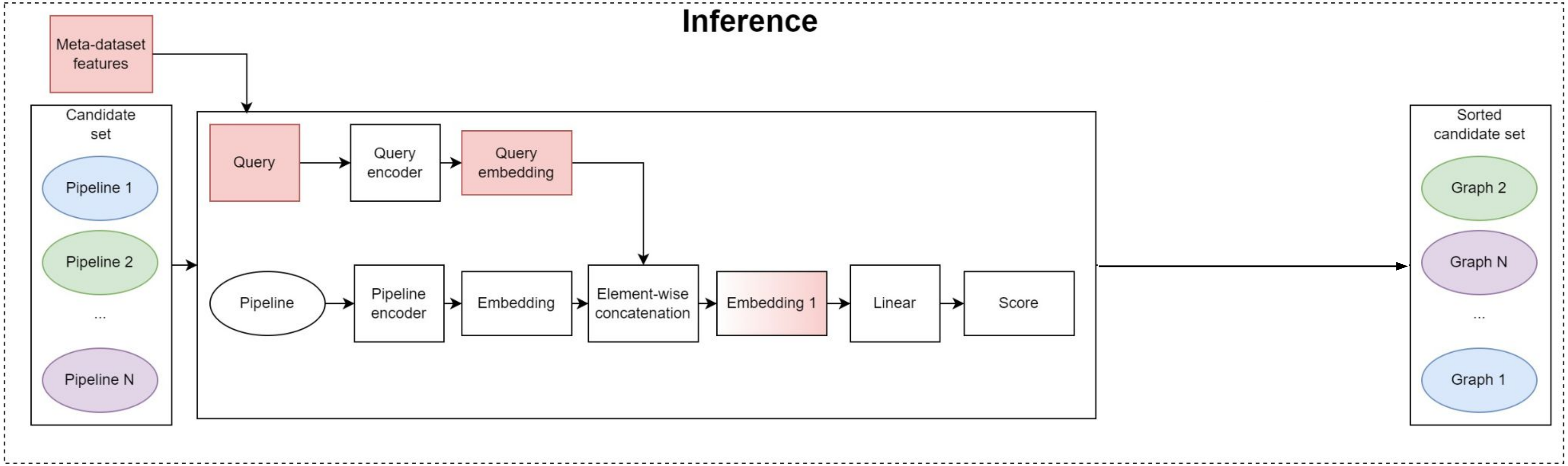
Proposed solution



Main idea:

1) Train a model to assign a score for a candidate with respect to a query
2) Introduce a query to ranking algorithm in early-fusion fashion

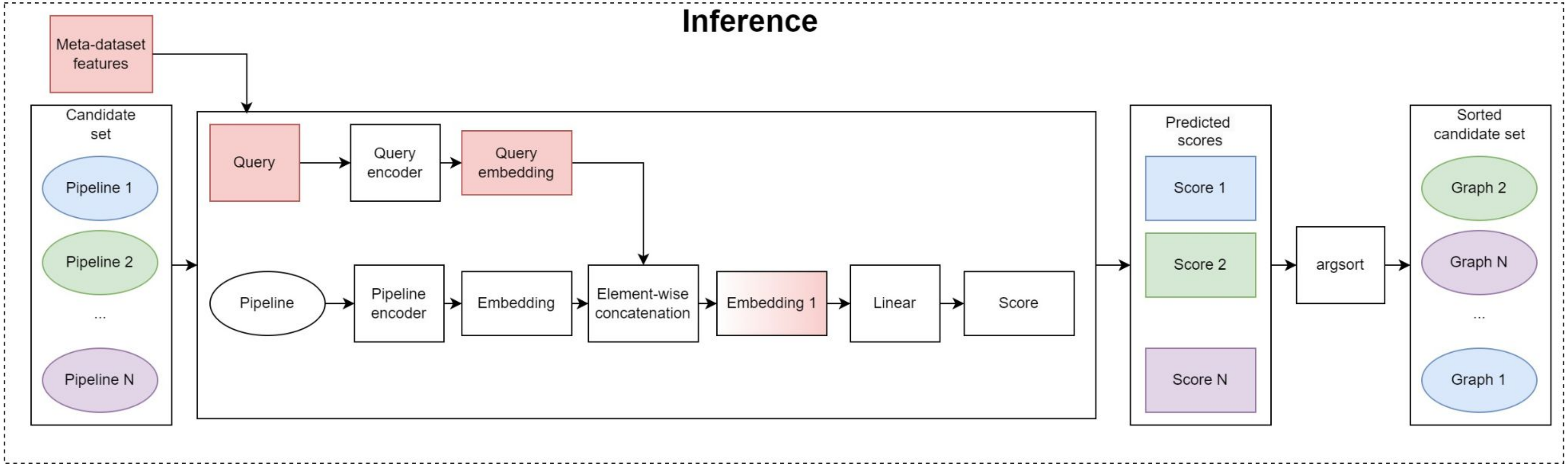
Proposed solution



Main idea:

1) Train a model to assign a score for a candidate with respect to a query
2) Introduce a query to ranking algorithm in early-fusion fashion

Proposed solution

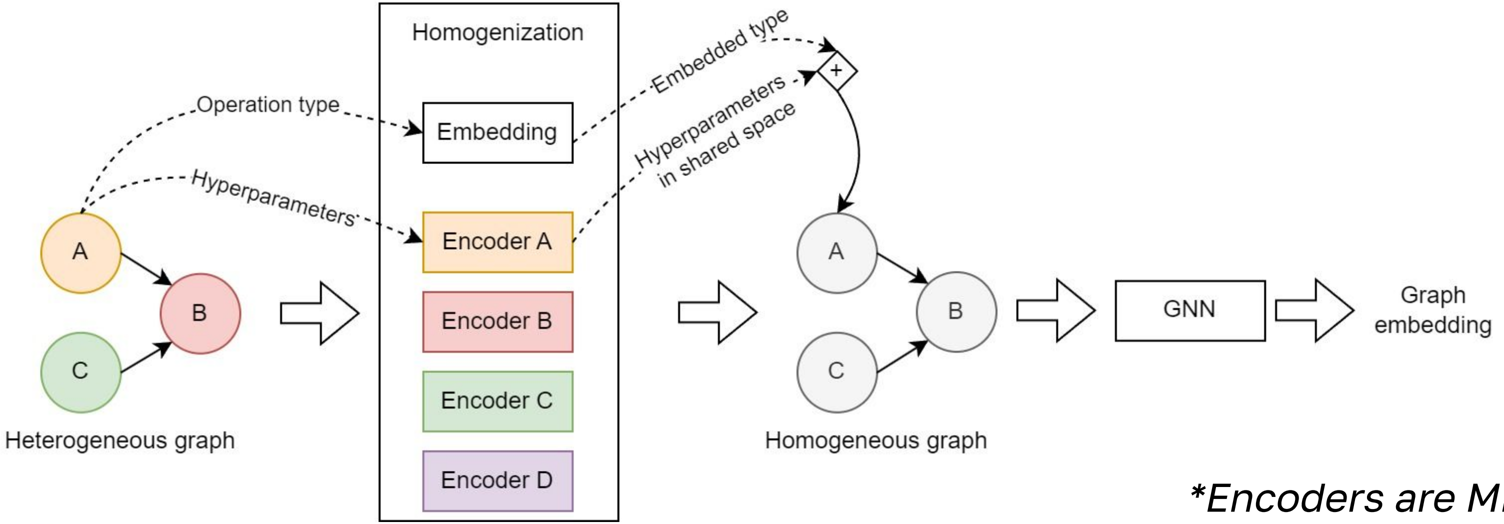


Main idea:

1) Train a model to assign a score for a candidate with respect to a query
2) Introduce a query to ranking algorithm in early-fusion fashion

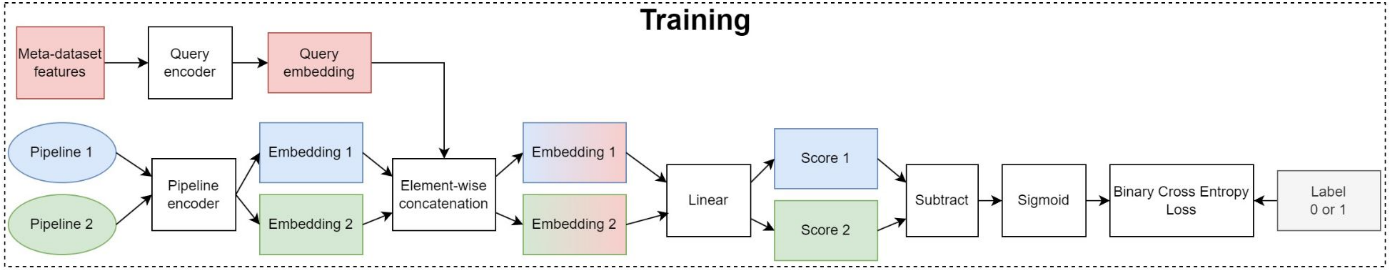
Heterogeneous graph encoding

- Graph heterogeneity arises from the diverse nature of operation types
- Each type is accompanied by its unique set of hyperparameters, differing in size and semantics
- The general type of each node is consistent — they all represent operations



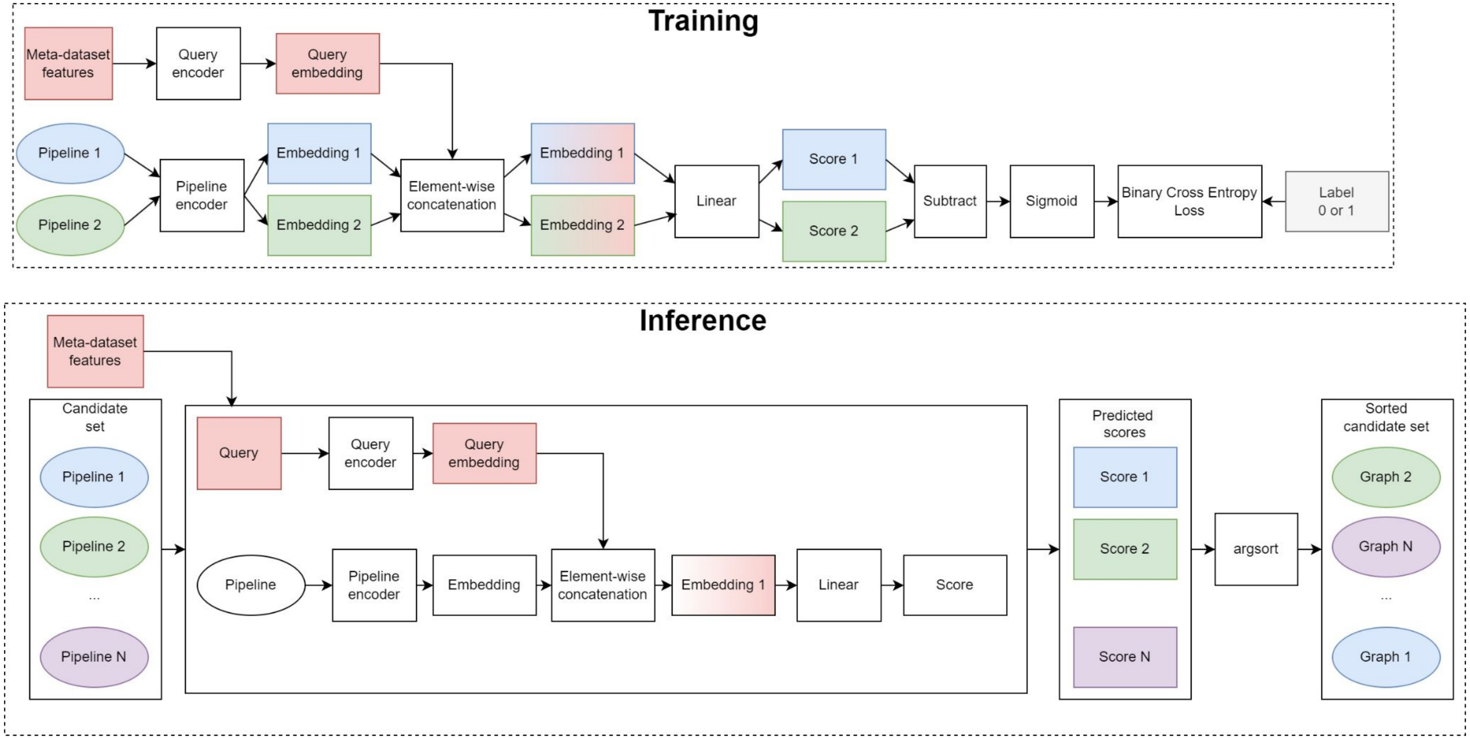
**Encoders are MLPs*

RankNet (point-wise inference)



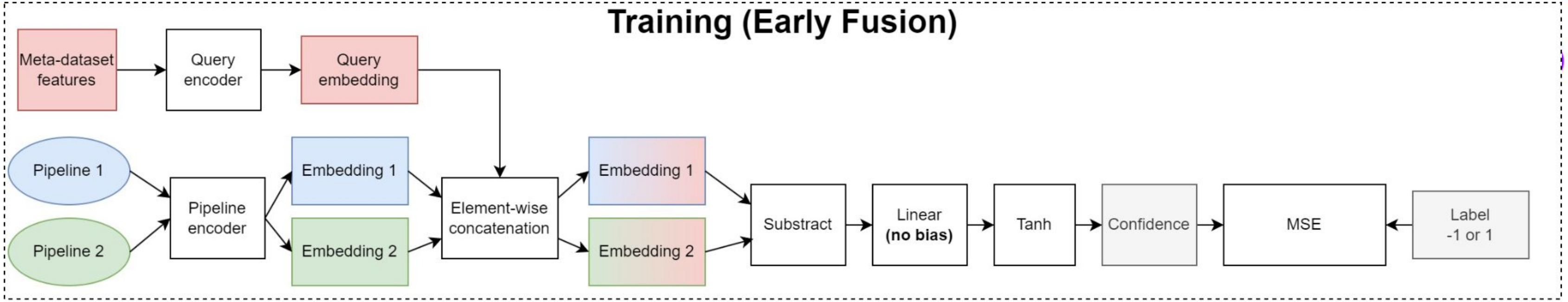
[1] Burges, Chris, et al. "Learning to rank using gradient descent." Proceedings of the 22nd international conference on Machine learning. 2005.

RankNet (point-wise inference)



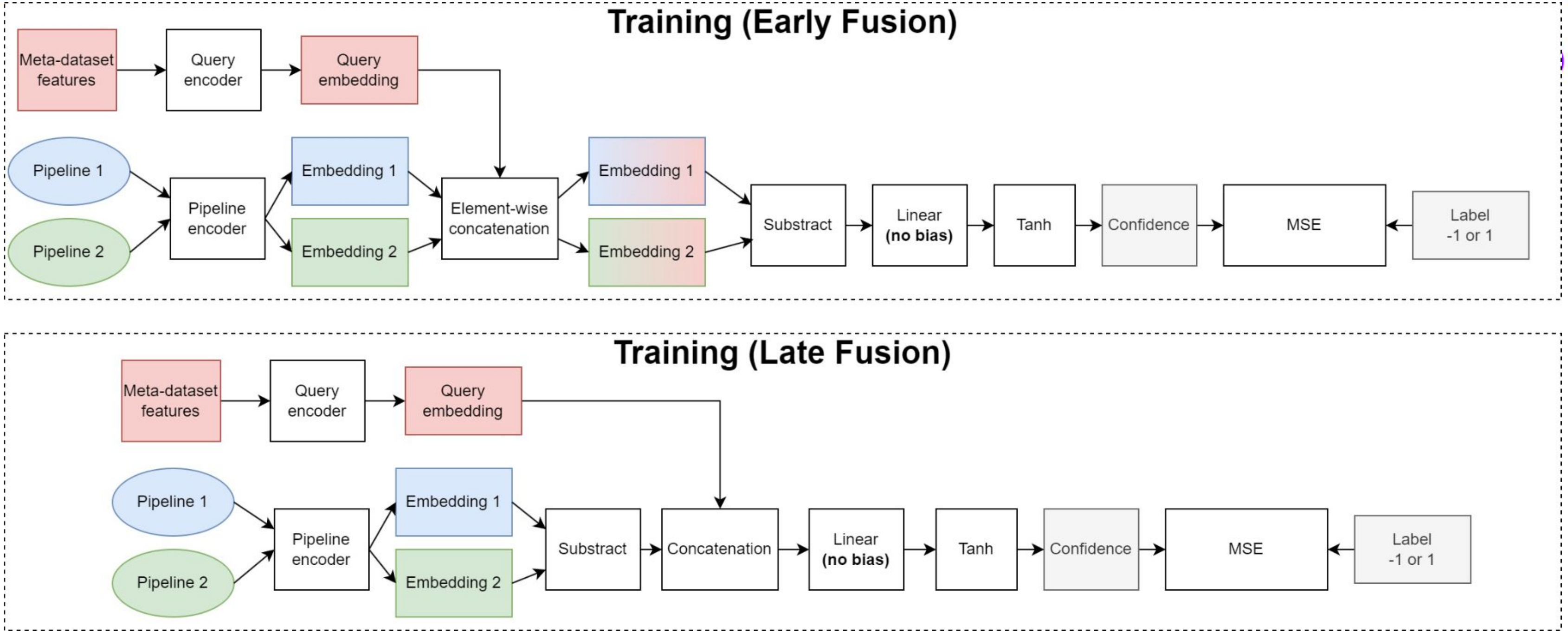
[1] Burges, Chris, et al. "Learning to rank using gradient descent." Proceedings of the 22nd international conference on Machine learning. 2005.

DirectRanker (pair-wise inference)



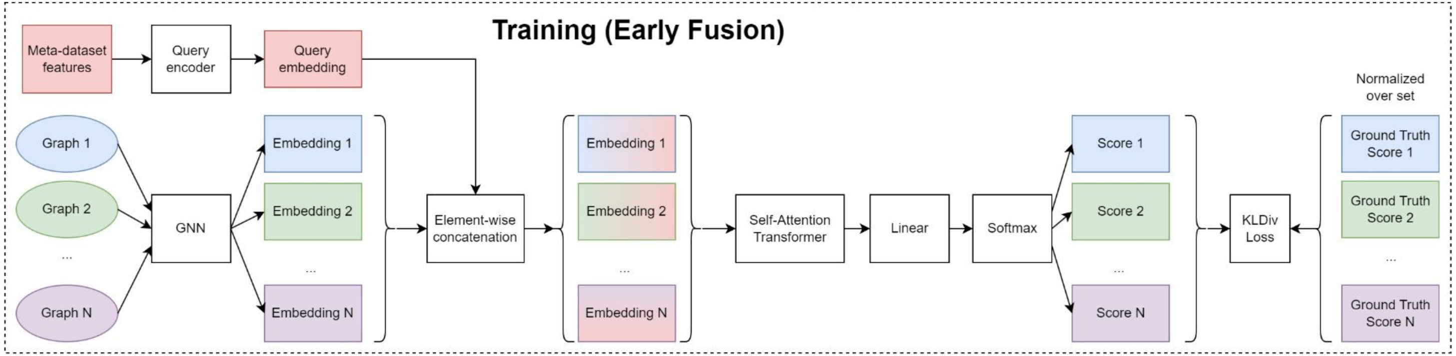
[1] Köppel, Marius, et al. "Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III. Springer International Publishing, 2020.

DirectRanker (pair-wise inference)



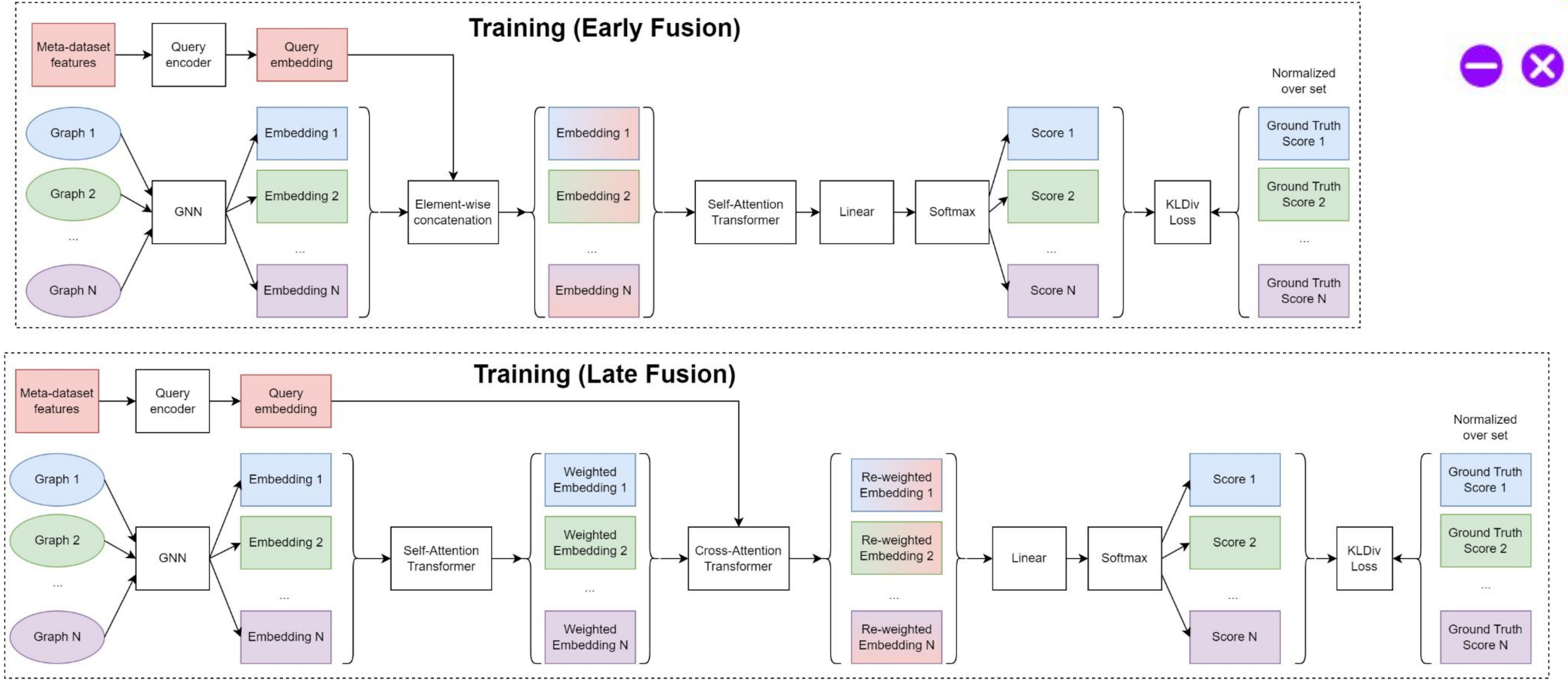
[1] Köppel, Marius, et al. "Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III. Springer International Publishing, 2020.

SetRank (list-wise inference)



[1] Pang, Liang, et al. "Setrank: Learning a permutation-invariant ranking model for information retrieval." Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. 2020.

SetRank (list-wise inference)



[1] Pang, Liang, et al. "Setrank: Learning a permutation-invariant ranking model for information retrieval." Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. 2020.

Experiment Setup

Data:

- 10 meta-datasets, 10 folds per meta-dataset, 100 total
- Unique pipeline architectures per meta-dataset: ~189-1869
- Variations per architecture: 50
- Total pipelines: 2.5 million

Time to collect dataset ~ 1 month (16 core CPU)
Train/test split: 80%/20% pipelines

Candidates set per meta-dataset is randomly formed from pipelines of different scores.

Ranking approach selection

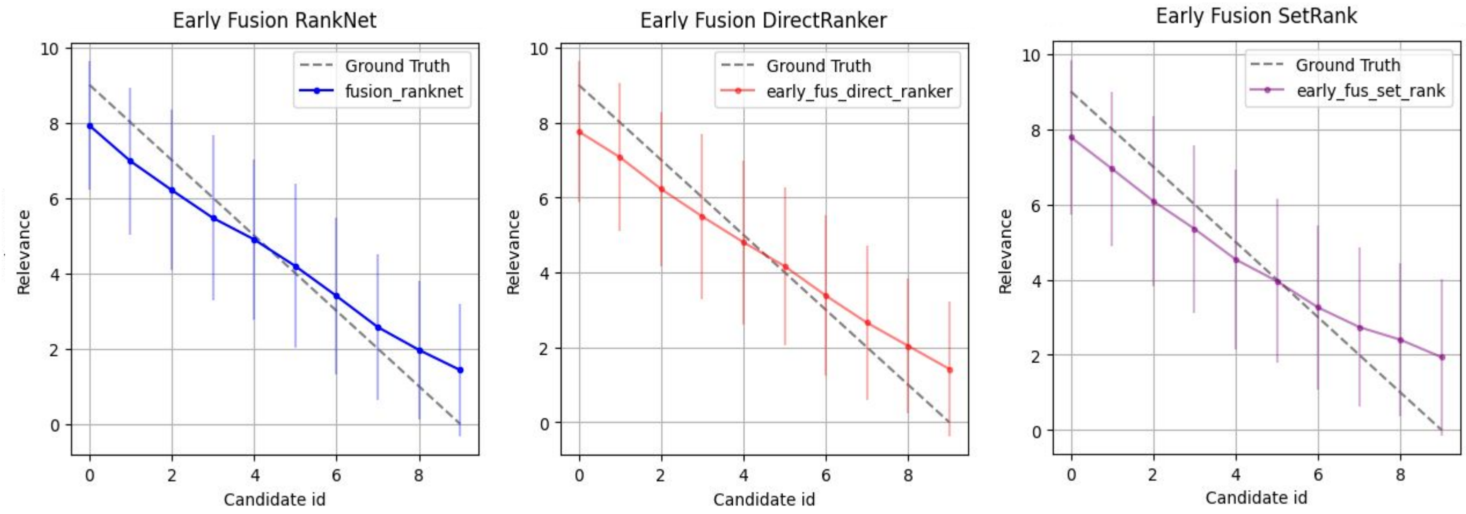
**Ranking of 10 candidates*

		OpenML meta-features		PyMFE meta-features	
Surrogate type	Fusion	Kendall-Tau	Precision	Kendall-Tau	Precision
Random	-	0.0	0.1	0.0	0.1
RankNet	Early	**0.52**	**0.24**	**0.53**	**0.25**
DirectRanker	Early	0.51	0.24	0.52	0.24
	Late	0.42	0.19	0.41	0.19
SetRank	Early	0.44	0.21	0.44	0.21
	Late	0.41	0.2	0.36	0.18

Features importance

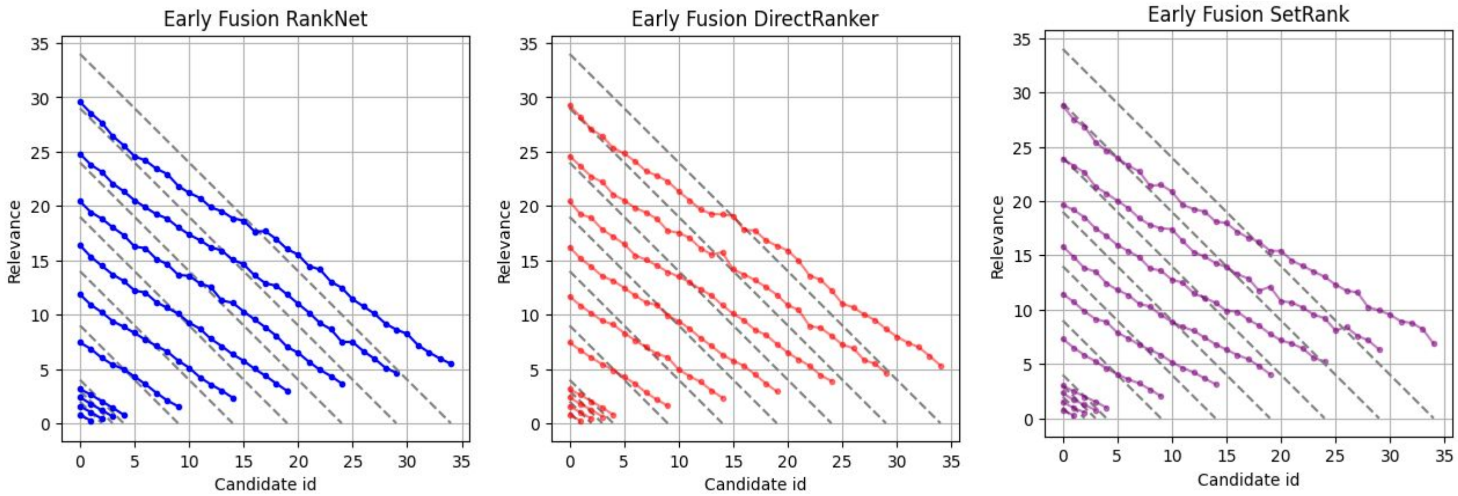
		Change of Kendall-Tau correlation coefficient			
Surrogate	Fusion	Operation type, %	Operation hparams, %	Edges between operation, %	Dataset meta-features, %
RankNet	Early	**-77**	-40	-62	-38
Direct Ranker	Early	**-80**	-41	-63	-37
	Late	-50	-46	**-68**	-24
SetRank	Early	**-66**	-36	-57	-36
	Late	-58	-42	**-58**	-19

Mean and std of sorting



Predicted relevancy over true relevance for pipelines trained on ROC AUC score

Sorting sets of different size



Predicted relevancy over true relevance for candidate sets of different size

Performance as AutoML component

The surrogate model was introduced to FEDOT framework as a pipeline evaluator

During the experiment a pipeline was designed for an unseen dataset

	w/ surrogate	w/o surrogate
Pipeline ROC AUC for the test subset	0,98	1.0
Process duration, sec	**20,4**	246,7

Conclusion

- GNN-features were shown to be applicable for ranking composite ML-pipelines
- Meta-dataset features can be utilized for ranking composite ML-pipelines
- Best ranking is achieved with point-wise learning-to-rank head

Avenues for future research

1) **Evaluation Scope**
- Augment the surrogate to accommodate other types of machine learning tasks beside classification
- Extend the evaluation dataset

2) **Training Scope**
- Explore methods for obtaining of embeddings of new operations without retraining
- Multiobjective optimization with regard to the compute requirements
- Multimodal features support, LLM usage for auxiliary inputs

Thank you for attention

FEDOT Framework

ITs MOre than a UNIVERSITY

ITMO OpenSource

Metafeatures

PyMFE		OpenML
Entire meta-dataset features	Averaged over meta-dataset columns features	Entire meta-dataset features
attr_to_inst	attr_ent	MajorityClassSIze
class_ent	eigenvalues	MinorityClassSize
eq_num_attr	freq_class	NumberOfClasses
gravity	joint_ent	NumberOfFeatures
inst_to_attr	kurtosis	NumberOfInstances
nr_attr	max	NumberOfNumericFeatures
nr_class	mean	NumberOfSymbolicFeatures
nr_cor_attr	min	
nr_inst	mut_inf	
nr_num	range	
ns_ration	skewness	
	var	

Surrogate Performance

	AutoGluon	Our	Our w/ surrogate	Baseline	Best Baseline	Best Pipeline
	RocAuc					
kddcup09_appetency	0,836	0,833	0,696	0,820	**0,842**	–
guillermo	0,915	–	0,882	0,897	**0,917**	–
albert	**0,770**	0,726	0,728	0,756	0,766	0,676
christine	0,824	–	0,804	0,817	**0,829**	–
numerai28_6	0,522	0,525	0,519	0,524	**0,532**	0,510
amazon_employee_access	0,865	–	0,705	0,843	**0,874**	–
airlines	**0,731**	0,713	0,674	0,709	0,728	0,650
sf-police-incidents	–	–	**0,678**	0,645	0,676	–
	Time, sec					
kddcup09_appetency	**537**	15734	1120	0	0	0
guillermo	**16367**	–	32072	0	0	0
albert	7416	**1334**	1491	0	0	0
christine	**1851**	–	3283	0	0	0
numerai28_6	5622	14443	**1863**	0	0	0
amazon_employee_access	**422**	–	2127	0	0	0
airlines	8058	16598	**2081**	0	0	0
sf-police-incidents	–	–	**1855**	0	0	0

Goal & objectives

Goal:

Development of a method for ranking composite ML-pipelines according to given data description

Objectives:

- Analytical review of relevant solutions, choice of quality metrics
- Collecting dataset
- Adapting relevant solutions in accordance to utilized input data
- Testing method and choosing final solution