

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
"Южно-Уральский государственный университет
(национальный исследовательский университет)"
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

Разработка приложения для подбора моделей машинного обучения на основе подхода AutoML

Автор работы:
студент группы КЭ-403
М.А. Щукин

Научный руководитель:
к.ф.-м.н., доцент кафедры СП
С.А. Иванов

Рецензент:
к.ф.-м.н, ст. преподаватель
кафедры ММОМ ФГБОУ ВО
«ЮУрГГПУ»
А.М. Шарафутдинова

Челябинск-2020

Цели и задачи

Цель:

Разработка приложения для подбора моделей машинного обучения на основе подхода AutoML.

Задачи:

- Выполнить обзор научной литературы и существующих решений по данной теме;
- Определить требования к приложению;
- Выполнить проектирование архитектуры приложения;
- Реализовать приложение;
- Выполнить тестирование приложения.

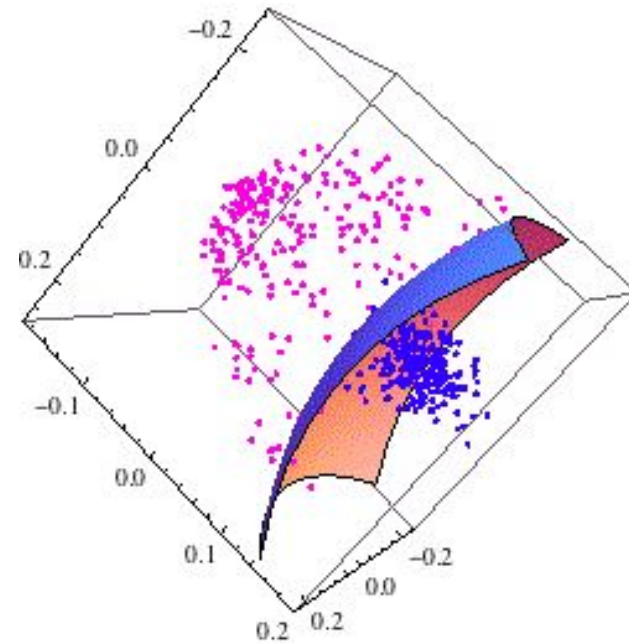
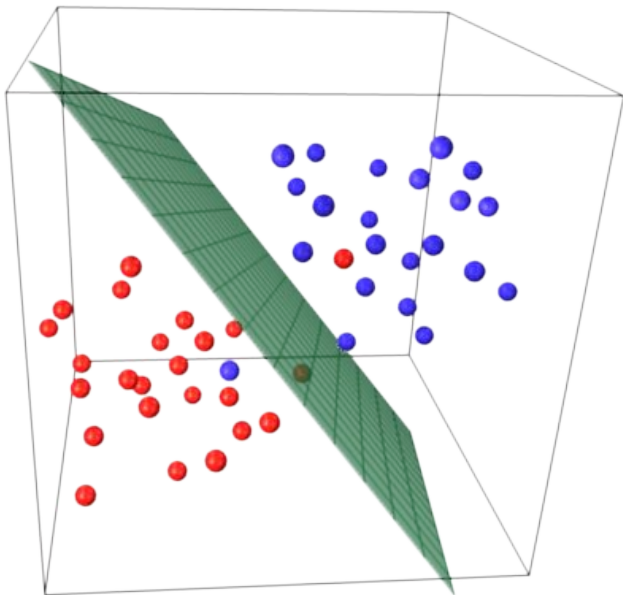
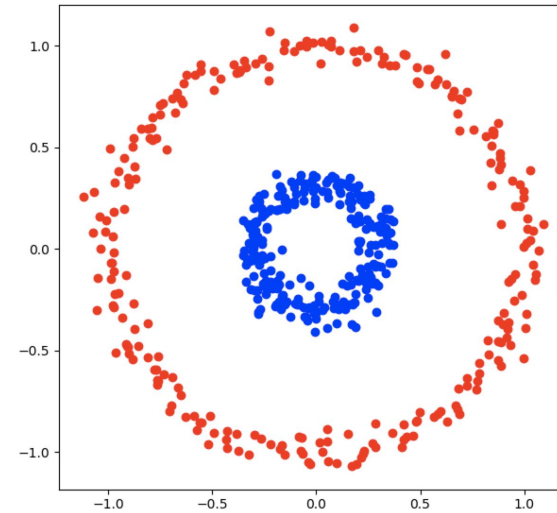
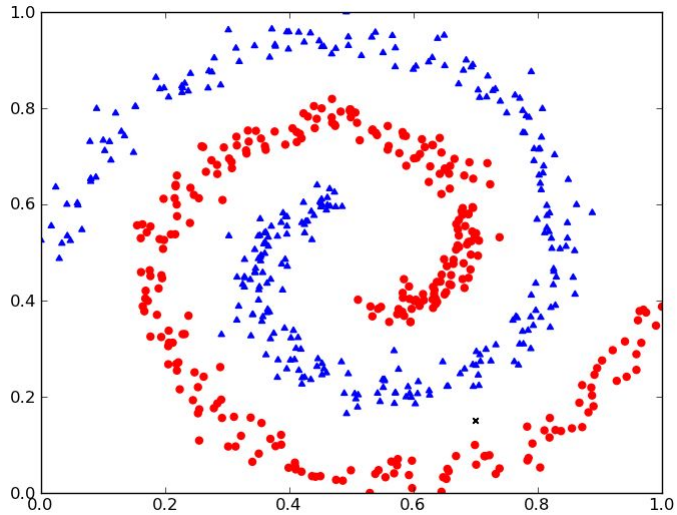
Актуальность

- Активное изучение AutoML
- Демократизация искусственного интеллекта
- Уменьшение потребности в квалифицированных кадрах
- Повышение производительности труда

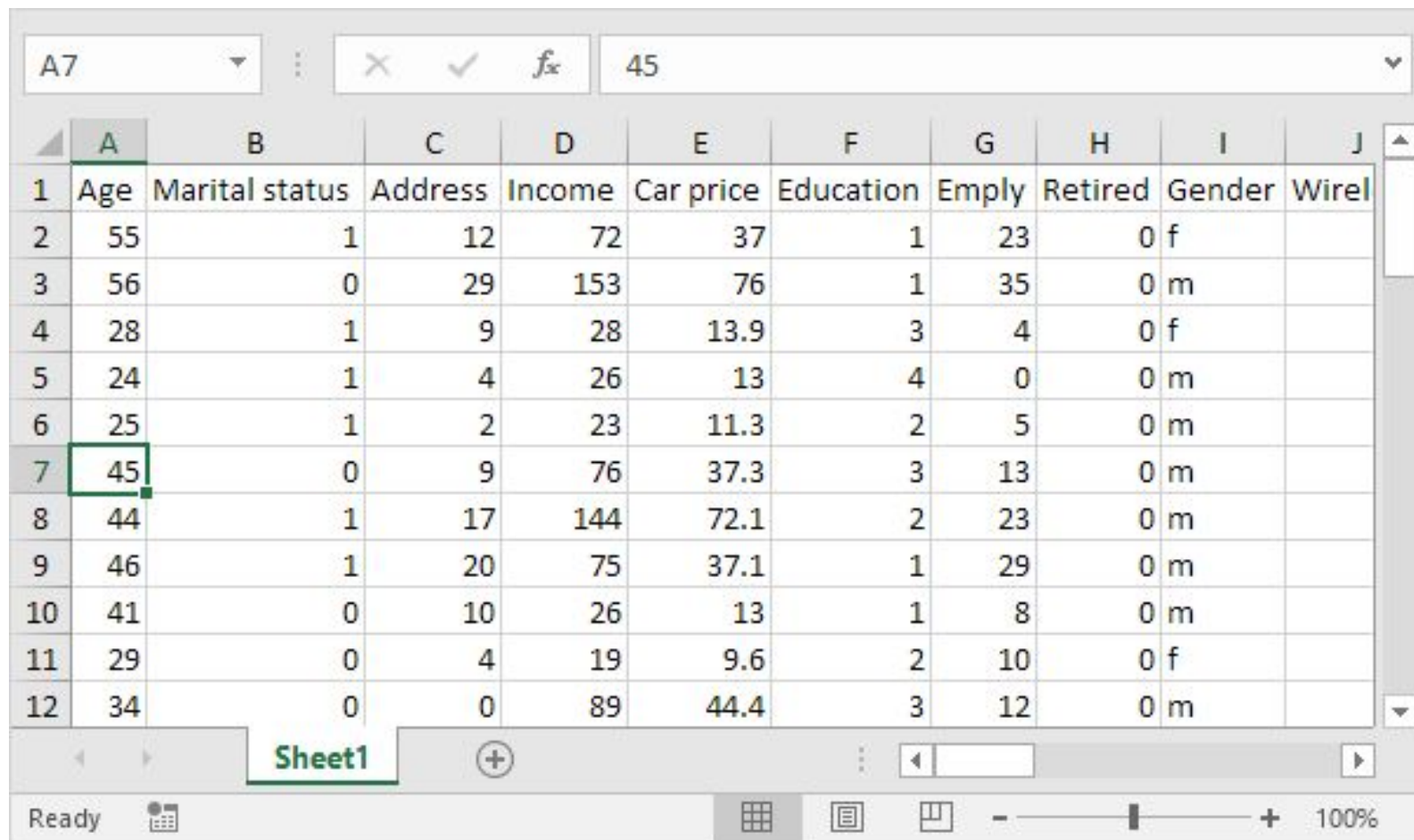
Актуальность (Google Trends)



Бинарная классификация



Формат входных данных



A screenshot of a Microsoft Excel spreadsheet. The formula bar at the top shows the value '45' being entered into cell A7. The spreadsheet contains 12 rows of data with 11 columns. The columns are labeled: Age, Marital status, Address, Income, Car price, Education, Empl, Retired, Gender, and Wirel. The data is as follows:

	A	B	C	D	E	F	G	H	I	J
1	Age	Marital status	Address	Income	Car price	Education	Empl	Retired	Gender	Wirel
2	55	1	12	72	37	1	23	0	f	
3	56	0	29	153	76	1	35	0	m	
4	28	1	9	28	13.9	3	4	0	f	
5	24	1	4	26	13	4	0	0	m	
6	25	1	2	23	11.3	2	5	0	m	
7	45	0	9	76	37.3	3	13	0	m	
8	44	1	17	144	72.1	2	23	0	m	
9	46	1	20	75	37.1	1	29	0	m	
10	41	0	10	26	13	1	8	0	m	
11	29	0	4	19	9.6	2	10	0	f	
12	34	0	0	89	44.4	3	12	0	m	

Что такое AutoML?

AutoML – Automated machine learning.

AutoML – Автоматизированное машинное обучение.

AutoML это:

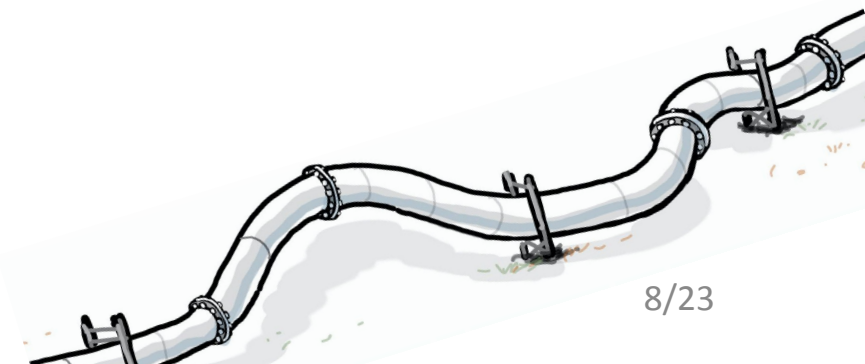
- Автоматизация стандартных ML процессов;
- Минимум вмешательства человека.

Впервые упоминается на **ICML 2014**.

Соревнования по созданию AutoML систем проводятся с 2014 года.

Стандартные ML процессы

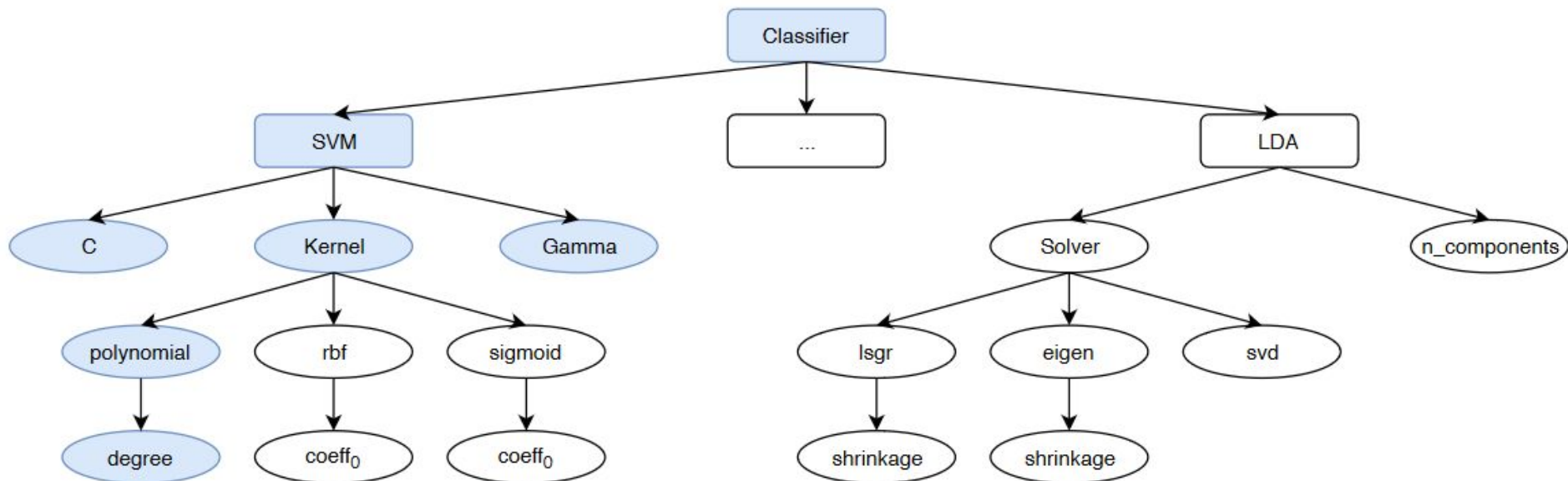
1. Постановка задачи
2. **Обработка отсутствующих значений**
3. **Обработка категориальных значений**
4. Отбор признаков
5. Извлечение признаков
6. **Масштабирование признаков**
7. **Метаобучение**
8. **Выбор метрики**
9. **Выбор способа валидации**
10. **Установка ограничений**
11. **Подбор моделей**
12. **Оптимизация гиперпараметров моделей**
13. Ансамблирование
14. Развертывание системы



Подбор моделей машинного обучения

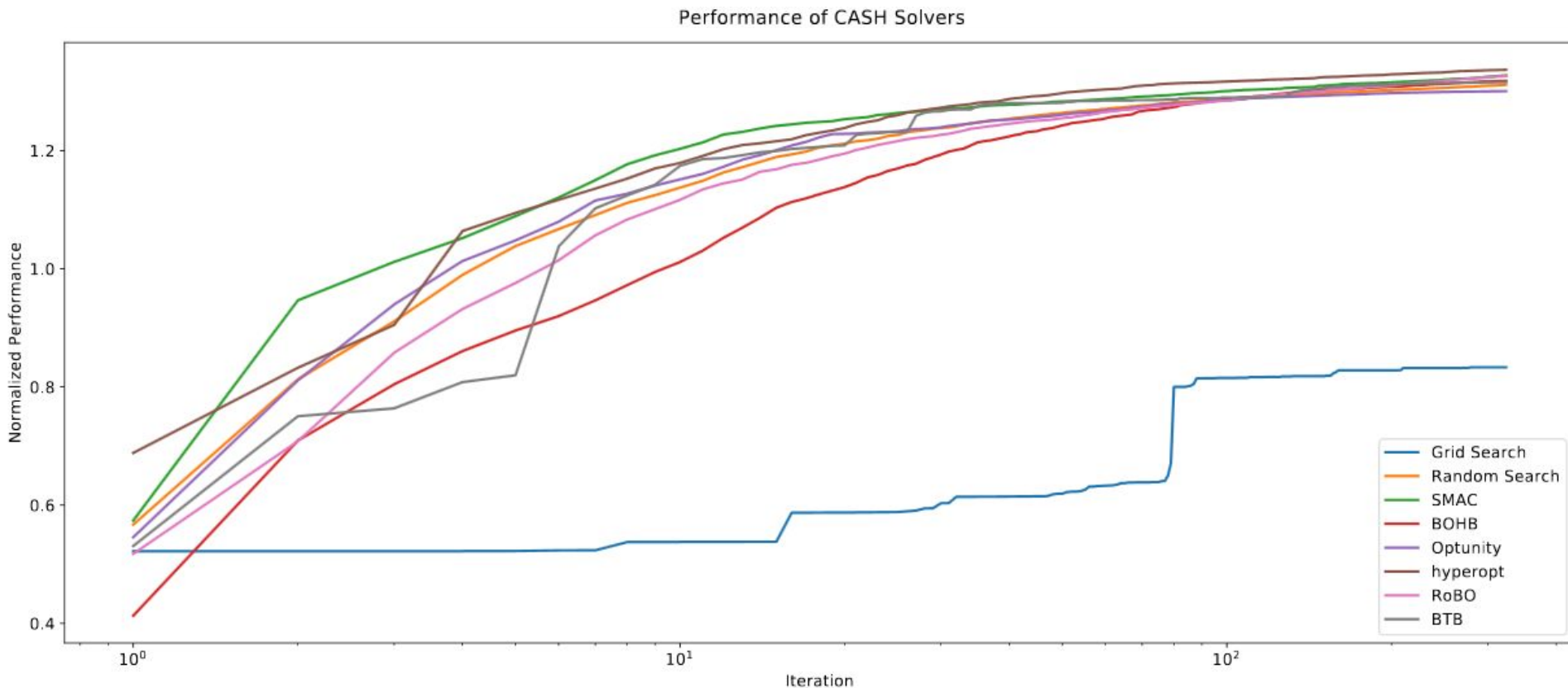
- AdaBoost
- SVM
- random forest
- bagging (SVC)
- extremely randomized trees
- MLP
- histogram gradient boosting
- decision tree
- ridge
- SGD
- k-nearest neighbors
- passive-aggressive
- nearest centroid
- logistic regression
- gaussian process
- LDA
- QDA
- label spreading
- Bernoulli naive Bayes
- Gaussian naive Bayes
- perceptron
- XGBoost
- ELM
- factorization machine
- polynomial network
- deep belief network

Оптимизация гиперпараметров



CASH

CASH – комбинированные подбор алгоритма и оптимизация гиперпараметров.



*Zöller M. A., Huber M. F. Benchmark and Survey of Automated Machine Learning Frameworks. (2020)

Аналоги (Free and Open-source)

Проект	Год	☆ Star	Примечание
TPOT 	2015	7.1k	
H2O AutoML 	2017	4.8k	
MLBox 	2017	1.1k	
Pennai 	2018	130	
AutoKeras 	2017	7.1k	
Hyperopt-Sklearn	2014	1k	
Auto-Sklearn	2015	4.5k	
Auto-Weka 	2013	255	

Аналоги (коммерческие)



Cloud AutoML **2018**



Azure Machine Learning

Automated ML **2018**



Amazon SageMaker

Autopilot **2019**



IBM Watson

AutoAI **2019**

Варианты использования

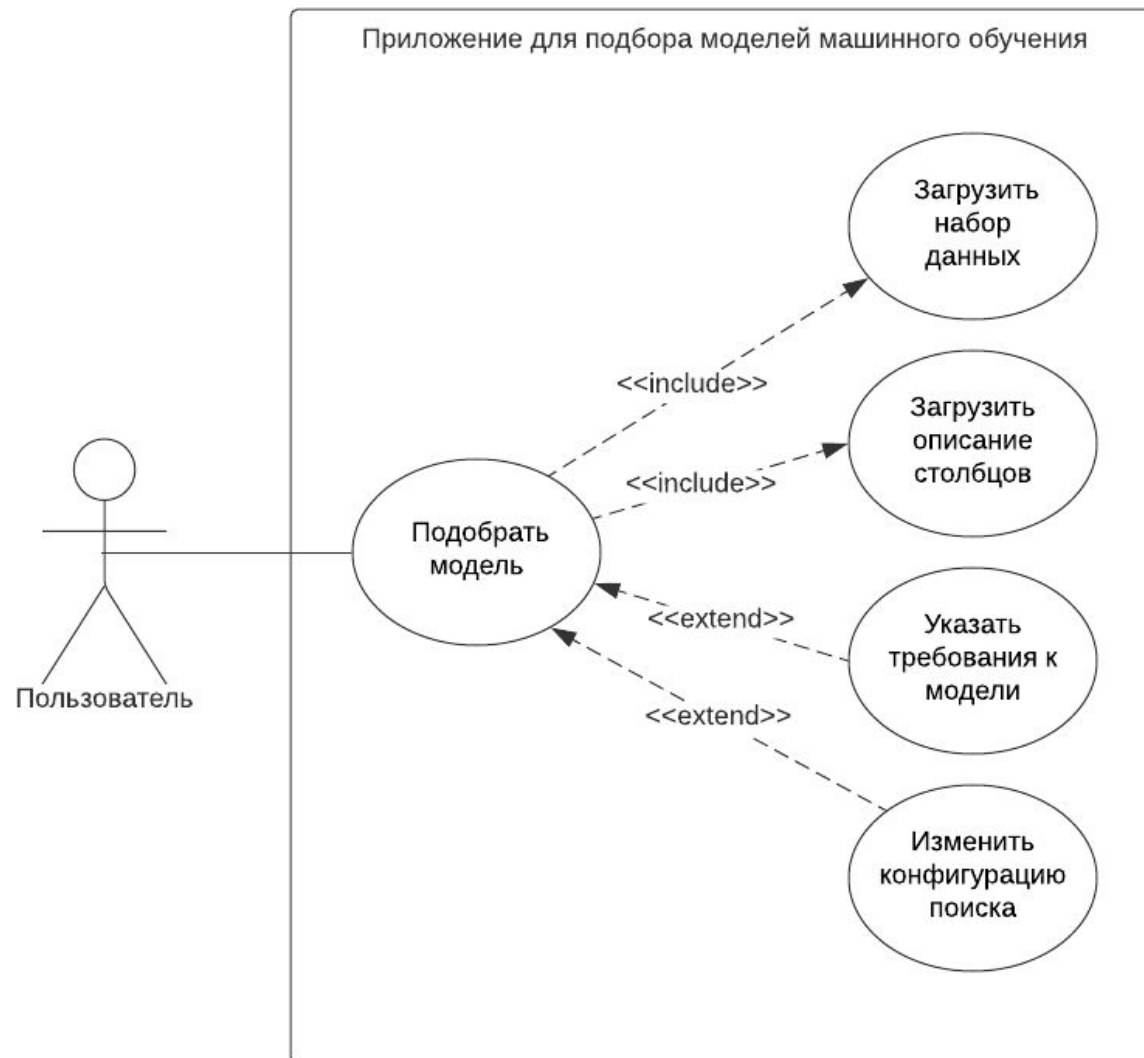
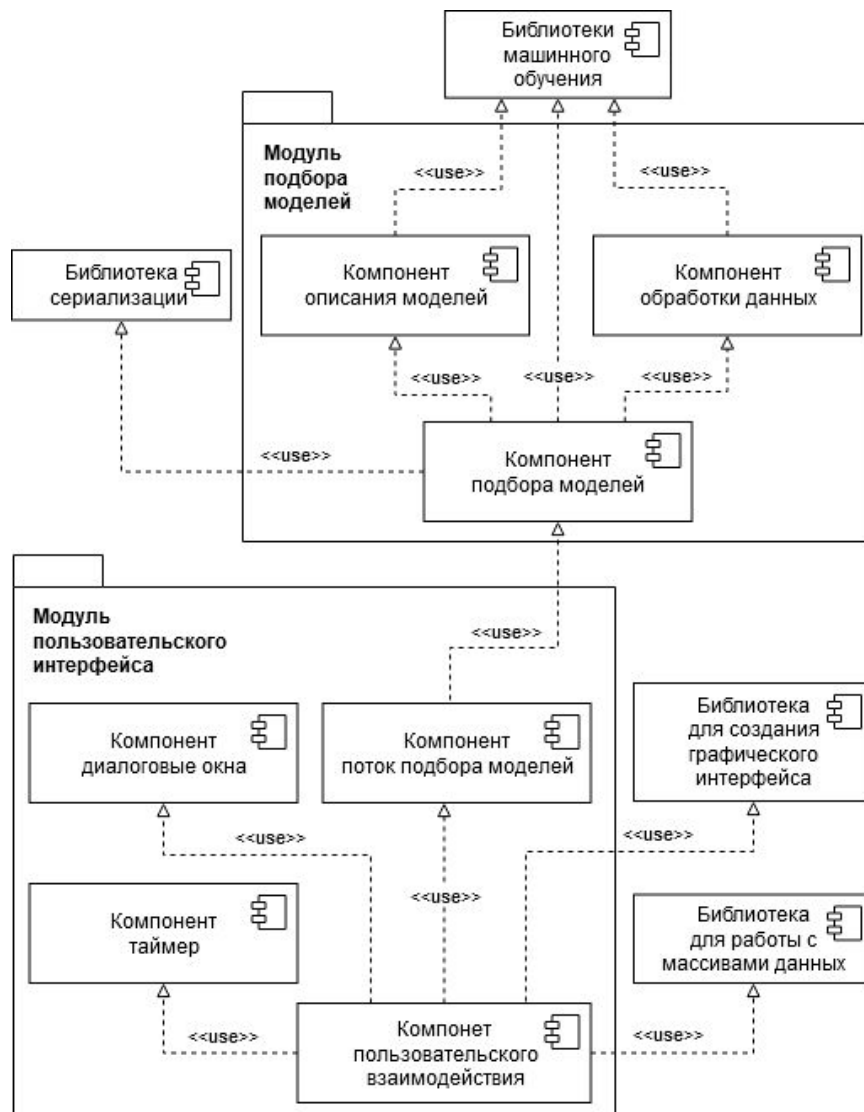


Диаграмма компонентов



Средства разработки



Python 3.6

dmlc
XGBoost



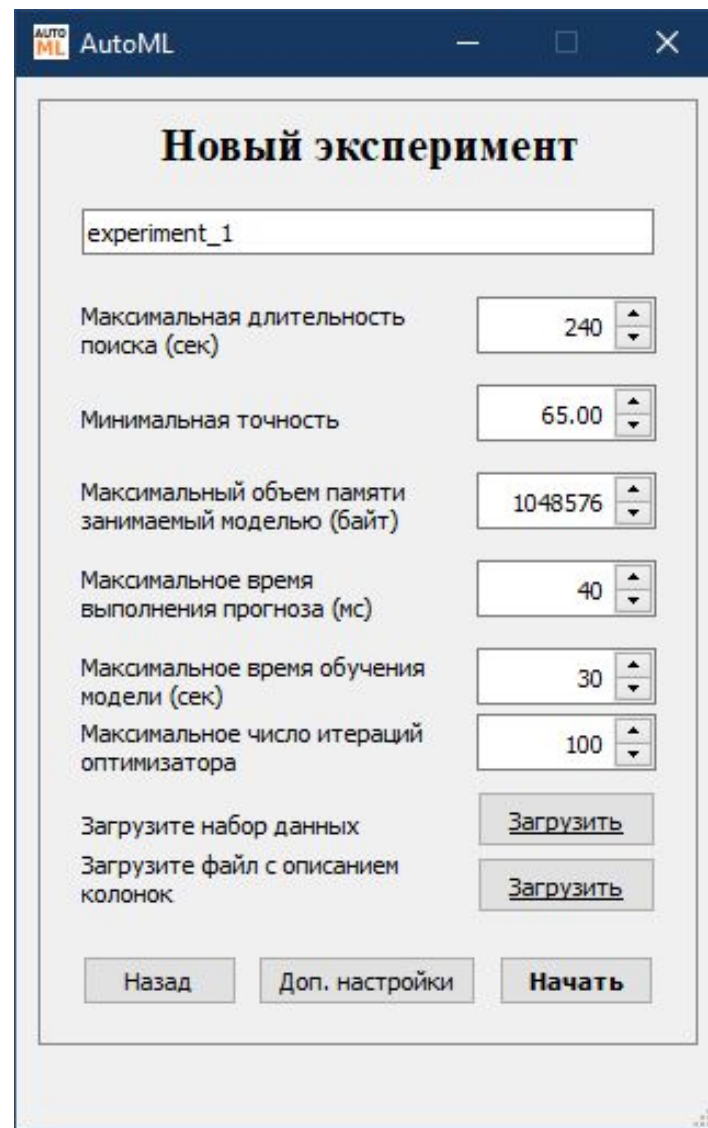
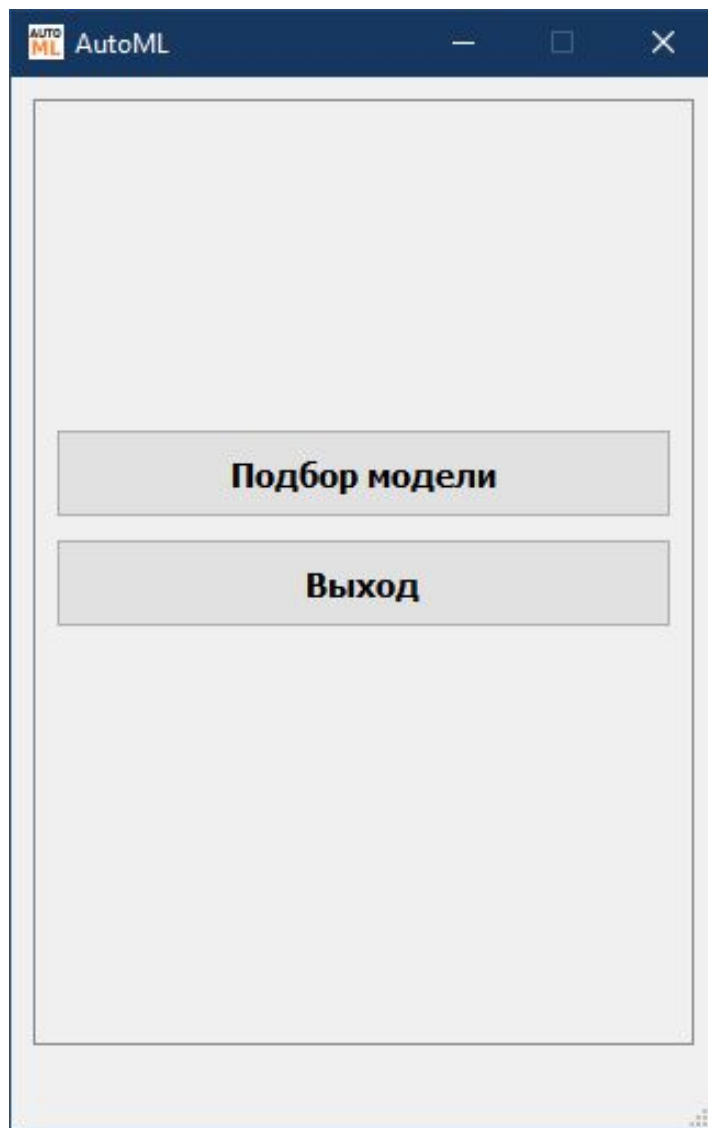
HYPEROPT



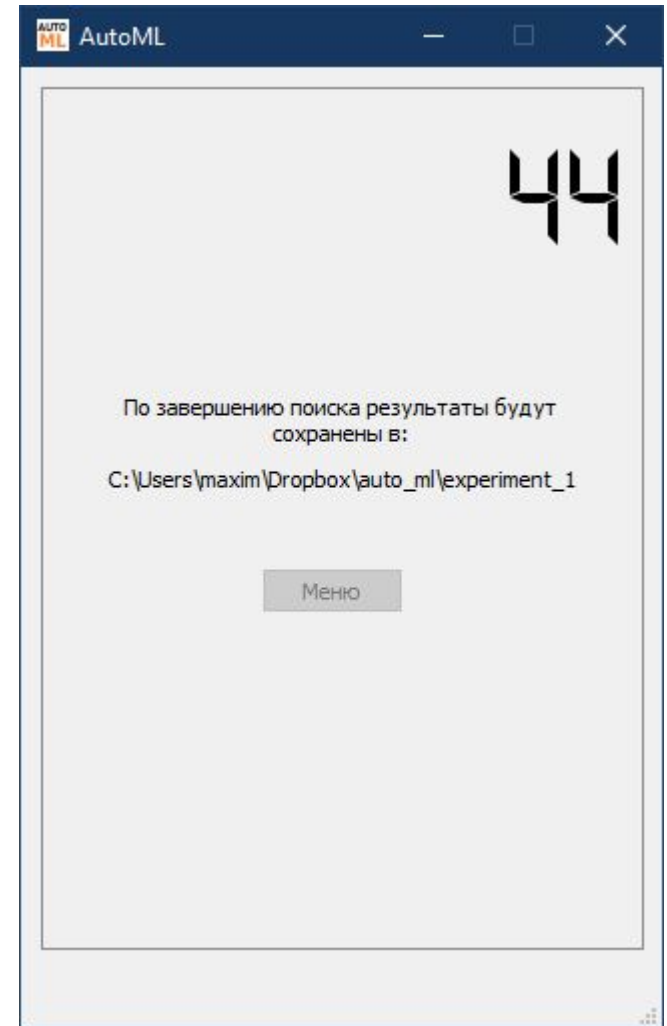
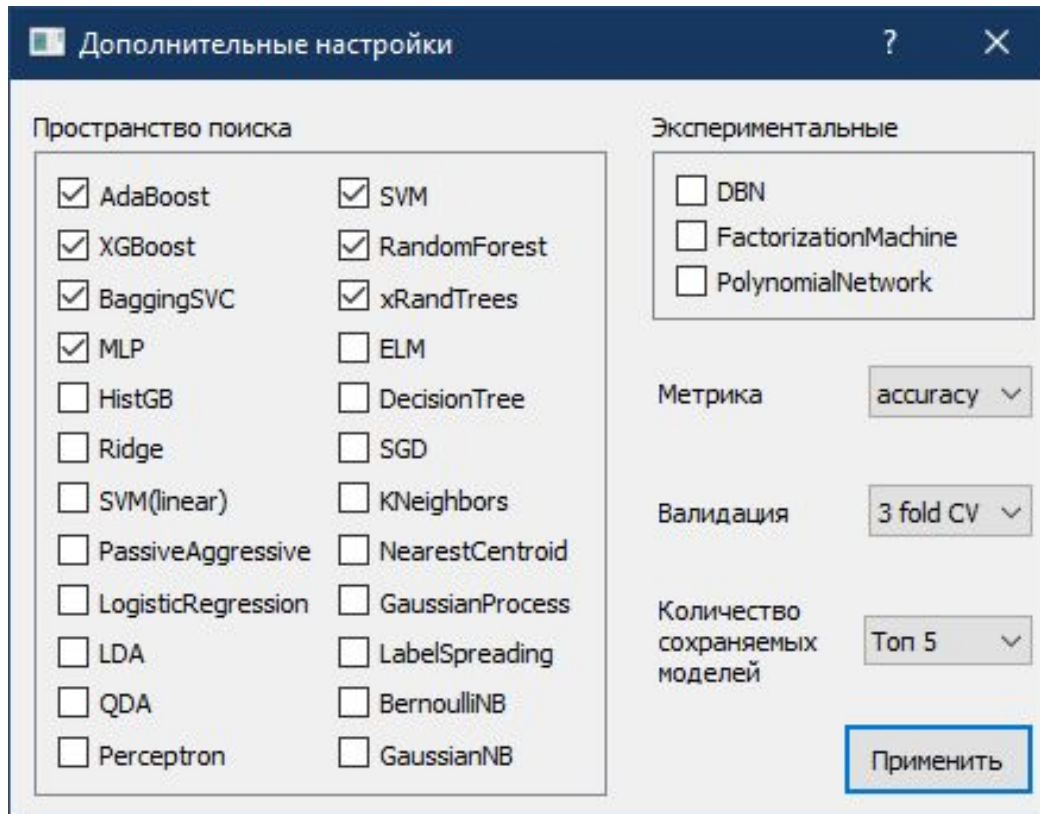
PyQt 5



Внешний вид приложения (1)



Внешний вид приложения (2)



Результаты тестирования (1)

Валидация: 10-fold cross-validation. **Метрика:** predictive accuracy.

dataset	max accuracy OpenML	max accuracy test	time budget (sec)	best models
blood-transfusion-service-center	0.8021	0.7984	300	RandomForest
breast-w	0.9757	0.9713	300	MLP
climate-model-simulation-crashes	0.9296	0.9222	200	RandomForest
credit-g	0.786	0.784	300	XGBoost
Diabetes	0.7878	0.7799	300	LinearSVC
Higgs	0.7333	0.7248	500	HistGB
monks-problems-1	1	1	200	AdaBoost
				HistGB
				RandomForest
monks-problems-2	1	0.9883	200	XGBoost
				AdaBoost
monks-problems-3	0.9892	0.9892	200	RandomForest
				XGBoost
				HistGB
ozone-level-8hr	0.9503	0.9428	500	XGBoost
qsar-biodeg	0.8872	0.8787	300	LinearSVC
Spambase	0.9626	0.9228	500	HistGB
steel-plates-fault	1	0.8906	300	LinearSVC
tic-tac-toe	1	0.9800	500	SVM
Wdbc	0.9842	0.9824	400	Bagging(SVC)

Результаты тестирования (2)

Валидация: 10-fold cross-validation. **Метрика:** predictive accuracy.

XGBoost - 99.48%
HistGB - 99.47%
MLP - 99.31%
SVM - 98.79%

MNIST

HistGB - 97.25%
XGBoost - 97.19%
MLP - 96.47%
SVM - 73.48%

Fashion MNIST

Основные результаты

- Выполнен обзор научной литературы и существующих решений по данной теме
- Выполнено определение требований к приложению
- Выполнено проектирование архитектуры приложения
- Выполнена реализация приложения
- Выполнено тестирование приложения

Исходный код:

<https://github.com/MainTechAI/BachelorThesisAutoML>

Дальнейшее развитие проекта (1)



Star

11.1k

(2017)



CatBoost



Star

5.2k

(2017)

Дальнейшее развитие проекта (2)



Дальнейшее развитие проекта (3)

- Multiclass, Multilabel, Multioutput-multiclass классификация.
- Включение алгоритмов предварительной обработки в задачу CASH.
- Добавление иных алгоритмов классификации.
- Экспериментирование с алгоритмами оптимизации гиперпараметров.
- Neural architecture search (NAS).
- Разработка модуля улучшающего точность за счет применения алгоритмов ансамблирования.
- Изучение методов мета-обучения.
- Поддержка временных рядов.
- Упор на максимальное использование выделенных вычислительных мощностей.
- Сохранение состояния поиска с возможностью в дальнейшем продолжить поиск с этого момента.
- Реализация поиска только по времени или только по максимальному числу итераций или в комбинации (сейчас только в комбинации).
- Сохранение подобранных моделей в формат ONNX.