

Задание на курсовую работу:

Название: Трекинг ML-экспериментов в ClearML

1. Цель курсовой работы

Цель курсовой работы – реализовать полное сопровождение экспериментов по построению модели машинного обучения с использованием ClearML, включая:

- отдельные эксперименты для подготовки выборок и обучения модели;
- версионирование датасетов;
- сохранение артефактов (модели, данных);
- логирование гиперпараметров и метрик качества;
- анализ результатов через веб-интерфейс ClearML.

2. Исходные данные и технологии

Нужно использовать:

- кейс «Моделирование предрасположенности клиентов» из дисциплины «Машинное обучение»;
- библиотеки:
 - для анализа данных и ML: `pandas`, `numpy`, `scikit-learn` (при необходимости — `xgboost`, `lightgbm`, `catboost` и др.);
 - для трекинга экспериментов: `clearml`;
- среду выполнения: локальную среду с поддержкой ноутбуков и AI-ассистента.

Настройка подключения к ClearML может выполняться через задание переменных окружения в коде (демонстрационный вариант). В пояснительной записке необходимо явно указать, что хранение ключей в коде недопустимо в реальных проектах и должно заменяться конфигурацией/переменными окружения вне кода.

3. Требования к постановке экспериментов в ClearML (вариант «на 5»)

3.1. Структура экспериментов

Обязательно:

1. **Отдельный эксперимент (Task)** для подготовки выборок из raw-датасета:
 - загружается «сырой» датасет;
 - выполняется предобработка (расчеты и т.п.);

- формируются обучающая/тестовая выборка или предобработанный датасет;
- результат подготовки (файлы с предобработанными данными) сохраняется как артефакт и/или как версия Dataset.

2. Минимум два независимых ML-эксперимента (Task) для построения модели:

- каждый эксперимент:
 - загружает данные через Dataset (определённая версия);
 - обучает модель;
 - логирует гиперпараметры;
 - сохраняет модель как pickle-артефакт;
 - логирует метрики качества.
- эксперименты должны отличаться:
 - набором гиперпараметров;
 - и/или набором признаков;
 - и/или используемым алгоритмом ML. Можно использовать тривиальную модель (на 1 признаке).

Итого: минимум **один Task подготовки данных + два Task обучения моделей**.

3.2. Использование нескольких датасетов с версионированием (Dataset)

Обязательно использовать объект Dataset не менее чем для **двух версий датасета**:

1. Создать и финализировать **базовую версию датасета**, например:
 - исходный или минимально очищенный датасет;
2. Создать **как минимум ещё одну версию** датасета:
 - Обучающие и тестовые выборки.

Новая версия также должна быть загружена и финализирована.

3. В каждом ML-эксперименте:
 - явно загрузить нужную версию датасета через Dataset.get(...).get_local_copy() или аналогичный метод;

3.3. Сохранение ML-модели в виде *.pickle-объекта

Обязательно:

- в каждом ML-эксперименте обученная модель должна быть:
 - сериализована в файл формата *.pickle (через joblib или pickle);

- загружена в соответствующий Task как артефакт ClearML.

3.4. Сохранение гиперпараметров ML-алгоритма

Обязательно:

- в каждом ML-эксперименте гиперпараметры модели должны быть доступны в интерфейсе ClearML (во вкладке Configuration или аналогичном разделе);
- допустимые варианты реализации:
 - использование `Task.connect` / `Task.connect_configuration` для объекта с параметрами;
 - использование конфигурационных файлов, автоматически подхватываемых ClearML;
 - явное логирование словаря гиперпараметров.

В пояснительной записке необходимо:

- перечислить, какие гиперпараметры варьировались;
- указать, как именно они были зафиксированы в ClearML.

3.5. Сохранение метрик качества модели

Обязательно:

- в каждом ML-эксперименте должна быть залогирована **как минимум одна итоговая метрика качества** (ROC-AUC и т.п.), отображающаяся в разделе RESULTS → SCALARS;

В пояснительной записке необходимо:

- провести сравнительный анализ результатов минимум двух ML-экспериментов.

4. Требования к структуре и качеству кода

Вы можете использовать за основу код кейса вашей практической работы. Допустимо разбить код на несколько ноутбуков.

5. Пояснительная записка (15-25 страниц)

Объём: **15–25 страниц** (без учёта титульного листа, содержания, списка литературы и приложений).

Рекомендуемая структура:

5.1. Введение

- актуальность трекинга экспериментов в ML;

- цель и задачи курсовой работы;
- объект и предмет исследования;
- краткое описание выбранного датасета и постановки задачи (моделирование предрасположенности клиентов).

5.2. Теоретическая часть

Рекомендуемое содержание:

1. **Основы трекинга экспериментов в машинном обучении:**
 - проблемы воспроизводимости;
 - понятия: эксперимент, артефакт, конфигурация, метрика.
2. **Системы управления ML-экспериментами (MLOps / experiment tracking):**
 - типовые задачи таких систем;
 - краткое сравнение подходов (в т.ч. место ClearML среди других решений).
3. **Архитектура и компоненты ClearML:**
 - Task (эксперимент);
 - Logger (метрики, графики, артефакты);
 - Dataset / Hyper-Datasets (версионирование данных);
 - артефакты и конфигурации.
4. **Версионирование данных в ML-проектах:**
 - зачем фиксировать версии датасетов;
 - как это реализовано в ClearML.

Теоретическая часть должна опираться на 5–10 источников (книги, статьи, документацию).

5.3. Практическая часть

Рекомендуемые разделы:

1. **Описание исходных данных и постановка задачи:**
 - источник и структура датасета;
 - целевая переменная и признаки;
 - формулировка задачи (классификация/регрессия и т.п.).
2. **Подготовка данных и эксперимент подготовки выборок:**
 - этапы предобработки;
 - логика разделения на выборки;
 - описание Task в ClearML;
 - описание созданных версий Dataset (для варианта «на 5»).

3. Описание ML-экспериментов:

- инициализация Task для обучения;
- используемые алгоритмы и гиперпараметры;
- использованные версии датасетов;
- логируемые артефакты (модели, данные).

4. Логирование гиперпараметров, метрик и анализ результатов:

- список зафиксированных гиперпараметров;
- перечень метрик;
- сравнительный анализ минимум двух ML-экспериментов (для «4»/«5»).

5. Обеспечение воспроизводимости:

- способы фиксации версий данных;
- как воспроизвести эксперименты по ноутбуку и данным в ClearML.

5.4. Заключение

- выводы по работе;
- оценка достигнутых результатов;
- направления дальнейшего развития.

5.5. Список литературы

- оформление по требованиям ГОСТ (с этим сейчас хорошо справляются нейросети!);
- включить источники по ML, MLOps и ClearML.

5.6. Приложения

- скриншоты из ClearML.

6. Скриншоты из UI ClearML (минимальный набор)

В пояснительную записку необходимо включить скриншоты с подписями.

Обязательно:

1. Список задач (Task Table):

- показаны:
 - Task подготовки данных (для варианта «на 5»);
 - минимум два Task обучения моделей (для «4»/«5»);
- пояснение, какая задача что делает.

2. Карточка Task подготовки данных (для «на 5»):

- обзор / детали;
- вкладка Artifacts (файлы данных или ссылки на Dataset).

3. Карточки двух ML-экспериментов:

- вкладка Configuration / Hyperparameters (гиперпараметры);
- вкладка Artifacts (модели в pickle);
- вкладка Scalars (метрики).

4. Сравнение экспериментов:

- скриншот с сопоставлением метрик/графиков разных ML-экспериментов (для «4»/«5»).

Желательно:

- скриншот раздела Datasets с версиями датасетов (для варианта «на 5»).

7. Итоговый комплект материалов

По завершении курсовой работы должен быть представлен следующий комплект:

1. **Пояснительная записка** объёмом 15-25 страниц в текстовом формате (DOCX) по структуре, близкой к описанной.
2. **Jupyter Notebook(и):**
 - с кодом:
 - подготовки данных;
 - создания/использования Dataset (для варианта «на 5»);
 - инициализации Task в ClearML;
 - логирования моделей, гиперпараметров и метрик;
 - с необходимыми комментариями.
3. **Комплект скриншотов** из UI ClearML:

8. Вариативные требования по ClearML для оценок «5», «4» и «3»

Ниже обобщаются требования именно к части по ClearML, которые определяют уровень оценки.

8.1. Требования для оценки «5» (полный вариант, описан в разделах 3-7)

Критерии:

- Есть **отдельный Task** для подготовки данных из raw-датасета.
- Есть **минимум два ML-эксперимента (Task)** для обучения моделей, отличающихся гиперпараметрами / признаками / алгоритмами.

- Используется объект Dataset с **минимум двумя версиями** датасета:
 - каждая версия создаётся и финализируется;
 - в каждом ML-эксперименте явно загружается нужная версия.
- В каждом ML-эксперименте:
 - модель сохраняется в виде *.pkl-файла и логируется как артефакт;
 - гиперпараметры зафиксированы и доступны в UI ClearML;
 - залогированы метрики качества, причём хотя бы один эксперимент содержит динамику метрики (например, loss по эпохам).
- Пояснительная записка содержит:
 - теоретическую часть по трекингу экспериментов и ClearML;
 - подробное описание реализованных Task, Dataset и артефактов;
 - сравнительный анализ экспериментов с опорой на данные ClearML;
 - нужные скриншоты.

8.2. Требования для оценки «4» (облегчённый вариант)

Часть по ClearML считается выполненной на уровень «4», если:

- выполнены **не менее двух ML-экспериментов (Task)** для обучения модели;
- эксперименты различаются по гиперпараметрам / признакам / алгоритмам;
- **Task для подготовки данных может отсутствовать** (подготовка допускается в тех же Task, что и обучение);
- использование объекта Dataset **не является обязательным**;
- в ClearML:
 - хотя бы один вариант данных (исходный или предобработанный) сохранён как артефакт;
 - **минимум для одного ML-эксперимента** модель сохранена как *.pkl и залогирована как артефакт (желательно – для всех);
 - для каждого ML-эксперимента зафиксированы гиперпараметры (видны в UI ClearML);
 - для каждого ML-эксперимента залогирована хотя бы одна итоговая метрика качества;
- пояснительная записка:
 - описывает структуру экспериментов;
 - показывает, как логируются модели, данные, гиперпараметры и метрики;
 - содержит скриншоты хотя бы по двум ML-экспериментам.

8.3. Требования для оценки «3» (минимальный вариант)

Часть по ClearML считается выполненной на уровень «3», если:

- реализован **как минимум один эксперимент (один Task)** в ClearML, включающий:
 - загрузку и предобработку данных;
 - обучение модели;
 - расчёт метрик качества;
- Task для подготовки данных отдельно **не требуется**;
- объект Dataset **не используется**;
- сохранение гиперпараметров в ClearML **не является обязательным**:
 - можно использовать модель с параметрами по умолчанию;
 - явное логирование гиперпараметров не требуется;
- желательно:
 - модель сохранена в виде *.pkl и прикреплена к Task как артефакт (это рекомендуемый минимум для устойчивой «3»);
- обязательно:
 - в ClearML залогирована как минимум одна итоговая метрика качества, отображающаяся в RESULT → SCALARS;
- в пояснительной записке:
 - кратко описан один эксперимент;
 - приведён скриншот Task и метрики из ClearML.