Правительство Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет «Высшая школа экономики»
Факультет компьютерных наук
Образовательная программа бакалавриата 09.03.04 «Программная инженерия»

ОТЧЕТ

по учебной (технологической) практике

в VK Education Practice

Выполнил студент
группы БПИ235
А. В. Васюков

_____
(подпись)

Проверил:

Руководитель практики от факультета компьютерных наук

к.э.н., доцент Департамента программной инженерии С. А. Лебедев

«___» _____ 2025 г. _____ _____
(оценка) (подпись)

Москва – 2025

АННОТАЦИЯ

Учебная практика проходила в компании VK в рамках программы VK Education Practice. Практика была направлена на изучение и применение современных методов машинного обучения и разработки программных решений для обработки и анализа данных. В ходе работы были освоены технологии Python, NumPy, Pandas, Catboost и другие инструменты для построения моделей, а также реализован проект по предсказанию пола пользователей по их данным.

СОДЕРЖАНИЕ

1. ЦЕЛЬ И ЗАДАЧИ ПРАКТИКИ

1.1. Цель

Целью практики является освоение и закрепление знаний по классическим алгоритмам машинного обучения (линейные и метрические модели, деревья решений, ансамблевые методы, рекомендательные системы), а также разработка модели машинного обучения для предсказания пола пользователей социальных сетей с последующим применением результатов в задачах персонализированной рекламы.

1.2. Задачи

1. Изучить теоретический материал и лекции по классическим алгоритмам машинного обучения: линейные и метрические модели, деревья решений, ансамблевые методы, рекомендательные системы.
2. Выполнить практические задания к лекционным материалам для закрепления знаний.
3. Провести анализ исходных данных: исследование структуры, выявление особенностей и подготовка признаков.
4. Выполнить предобработку и очистку данных для корректного обучения моделей.
5. Разработать и обучить модель машинного обучения для предсказания пола пользователей социальных сетей.
6. Оценить качество модели с использованием соответствующих метрик.
7. Подготовить отчет о проделанной работе, включающий описание этапов разработки, результаты экспериментов и выводы.

2. ОПИСАНИЕ МЕСТА ПРОХОЖДЕНИЯ ПРАКТИКИ

Практика проходила в компании VK [1] в рамках программы VK Education Practice [2]. VK – одна из крупнейших российских IT-компаний, объединяющая социальные сети, сервисы коммуникации, развлечений, образования и облачные платформы. Компания активно развивает направления, связанные с машинным обучением, анализом данных и созданием высоконагруженных систем.

Программа VK Education Practice представляет собой масштабную онлайн-практику, разработанную совместно с ведущими университетами. Ее цель – дать студентам опыт работы с современными технологиями в условиях, приближенных к реальным индустриальным задачам.

Структура программы включает:

- лекции по машинному обучению;
- практические задания на основе реальных данных;
- проектную работу по разработке и обучению моделей;
- консультации и обратную связь от экспертов VK.

3. ОБЗОР ИЗУЧЕННЫХ МАТЕРИАЛОВ, ТЕХНОЛОГИЙ И МЕТОДОВ

В процессе практики был проведен обзор теоретических основ и практических методов машинного обучения, необходимых для решения задачи предсказания пола пользователей социальных сетей. Особое внимание уделялось не только изучению алгоритмов и моделей, но и инструментам анализа данных, их предобработки и визуализации, а также современным библиотекам и подходам, применяемым в индустрии.

3.1. Классические методы машинного обучения

- Задачи классификации и регрессии.
- Алгоритмы: метод k-ближайших соседей, наивный байесовский классификатор, решающие деревья, логистическая регрессия.
- Методы оптимизации: градиентный спуск.
- Метрики качества: точность, полнота, F1-мера и другие показатели оценки моделей.
- Методы работы с признаками: извлечение и преобразование данных, обработка пропусков, отбор признаков.

3.2. Продвинутые методы и ансамбли моделей

- Ансамблирование: бэггинг, бустинг, стекинг.
- Случайный лес, градиентный бустинг (CatBoost, XGBoost).
- Методы кластеризации: k-means, иерархическая и спектральная кластеризация, оценка качества кластеризации.
- Понижение размерности, выбор оптимальных гиперпараметров.
- Поиск ассоциативных правил.

3.3. Обработка текстов (NLP)

- Специфика текстовых данных, методы предобработки и ручного извлечения признаков.
- Использование эмбеддингов слов, векторное представление текстов.
- Применение NLP в реальных задачах (классификация и анализ текстовой информации).

3.4. Геоаналитика

- Основы работы с пространственными данными и их проекциями.
- Использование GeoPandas, OSM, OSMnx, QGIS для анализа геоданных.
- Методы пространственного объединения (spatial joins).

3.5. Рекомендательные системы

- Основные подходы: item-to-item, методы на основе совстречаемости, неперсонализированные и персонализированные рекомендации.
- Коллаборативная фильтрация и матричная факторизация.
- Метрики оценки качества рекомендательных систем.

3.6. Методы анализа поведения пользователей

- CRM и CVM-модели.
- Методы расчета CLTV (Customer Lifetime Value).
- Look-alike модели, response-моделирование, uplift-модели.
- Подходы к контролю качества данных и мониторингу работы моделей.

3.7. Технологии и инструменты

В процессе практики активно использовались библиотеки Python [3] для анализа и визуализации данных: pandas [4], NumPy [5], Matplotlib [6], Seaborn [7]. Для обработки данных применялись json и datetime. Для машинного обучения использовались scikit-learn [8], CatBoost [9], XGBoost [10]. Разработка велась в Jupyter Notebook с использованием системы контроля версий Git.

4. ОПИСАНИЕ МЕТОДОВ, ТЕХНОЛОГИЙ И СРЕДСТВ РАЗРАБОТКИ, ИСПОЛЬЗОВАННЫХ ДЛЯ РЕШЕНИЯ ПОСТАВЛЕННЫХ ЗАДАЧ

Для реализации проекта по предсказанию пола пользователей социальных сетей был применен комплекс методов машинного обучения и инструментов анализа данных, обеспечивший решение задачи в соответствии с целями практики. Работа строилась поэтапно и включала исследование данных, их подготовку, разработку и обучение моделей, а также оценку качества полученных решений.

4.1. Методы машинного обучения

- Ключевыми методами стали алгоритмы классификации, что было обусловлено бинарным характером целевой переменной (мужской/женский пол).
- Логистическая регрессия была выбрана как базовая модель, позволяющая получить интерпретируемое решение и выступающая в качестве контрольной точки для последующего сравнения.
- Решающие деревья и случайный лес использовались для выявления нелинейных закономерностей и учета взаимодействий между признаками.
- Ансамблевые методы (бэггинг, градиентный бустинг) применялись с целью повышения устойчивости и точности предсказаний за счет объединения нескольких слабых моделей в сильную.
- Для настройки качества применялся подбор гиперпараметров (GridSearchCV, RandomizedSearchCV), что обеспечивало выбор оптимальных параметров моделей и предотвращало переобучение.

Таким образом, выбор конкретных алгоритмов был продиктован необходимостью как исследовать базовые интерпретируемые модели, так и протестировать более сложные ансамблевые подходы, широко используемые в индустрии.

4.2. Методы обработки данных

Для корректной работы алгоритмов была проведена всесторонняя предобработка:

- Разведочный анализ данных (EDA) позволил выявить распределения признаков, наличие выбросов и пропусков.
- Обработка пропусков выполнялась с учетом природы признаков (заполнение модой, медианой или специальными категориями).
- Кодирование категориальных переменных (One-Hot Encoding, Label Encoding) обеспечивало их преобразование в числовую форму.

- Масштабирование признаков применялось для корректной работы алгоритмов, чувствительных к масштабу (например, логистическая регрессия, k-NN).
- Инженерия признаков включала извлечение новых характеристик из исходных данных, что повышало информативность признакового пространства.

4.3. Технологии и инструменты разработки

- Основным языком разработки выступал Python, предоставляющий широкий спектр библиотек для анализа и моделирования данных.
- Для анализа и обработки данных использовались pandas и NumPy, которые являются основными и самыми популярными для данной работы.
- Для визуализации результатов применялись Matplotlib и Seaborn, что позволяло наглядно интерпретировать распределения, корреляции и метрики качества.
- Библиотека scikit-learn обеспечила реализацию базовых алгоритмов машинного обучения и метрик оценки (accuracy, precision, recall, F1).
- Для ансамблей применялись специализированные библиотеки CatBoost и XGBoost, предоставляющие эффективные реализации градиентного бустинга.
- Эксперименты проводились в среде Jupyter Notebook, что способствовало документированию кода и промежуточных результатов.
- Для контроля версий и организации работы над проектом использовались Git и GitHub.

4.4. Организация процесса работы

Разработка велась итеративно:

1. Первичный анализ данных и выявление особенностей датасета.
2. Предобработка и формирование признакового пространства.
3. Обучение базовых моделей (логистическая регрессия, решающее дерево) и получение контрольных результатов.
4. Построение ансамблевых моделей и их сравнение с базовыми по метрикам качества.
5. Подбор гиперпараметров и оптимизация финальной модели.
6. Систематизация результатов и формирование выводов о применимости методов.

Таким образом, использованные методы и технологии обеспечили комплексное решение задачи классификации, продемонстрировав практическую значимость изученных алгоритмов и инструментов.

5. ОПИСАНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

В рамках практики была разработана и обучена модель машинного обучения для предсказания пола пользователей социальных сетей.

5.1. Предобработка данных

Был создан единый датасет пользователей, включающий геолокацию, страну, регион, часовой пояс, рекламные запросы и параметры устройств.

Из признаков были извлечены дополнительные характеристики:

- domain – домен сайта из запроса referer;
- hour – час запроса из request_ts;
- browser, os – тип браузера и операционной системы;

Пропуски в region_id заполнялись значением country_id, остальные пропуски – значением none.

Данные агрегировались по пользователям с использованием статистик частоты и среднего значения для признаков.

5.2. Обучение модели

Данные были разделены на тренировочную, валидационную и тестовую выборку в пропорции 8:1:1.

Для числовых признаков применялся StandardScaler, категориальные признаки обрабатывались с помощью Pool библиотеки CatBoost.

Настройка CatBoostClassifier:

iterations=1500,
early_stopping_rounds=15,
auto_class_weights=’Balanced’,
random_seed=42,
verbose=100.

Также проводились эксперименты с логистической регрессией и XGBoost, однако наилучший результат показал CatBoost.

5.3. Результаты оценки модели

Data	AUC-ROC	Accuracy	F1-score
Train	0.9030	0.9033	0.8985
Test	0.8760	0.8762	0.8704

На тестовых данных модель достигла Accuracy = 87.6%, ROC-AUC = 87.6%, F1-score = 87%, что свидетельствует о высокой точности и сбалансированности предсказаний.

5.4. Дополнительные эксперименты:

Были протестированы новые признаки: синусно-косинусное преобразование времени, день недели запроса, OneHotEncoding для browser и os.

Эксперименты с альтернативными моделями (логистическая регрессия, XGBoost) не привели к улучшению качества.

5.5. Выводы

Наиболее эффективной моделью для задачи классификации пола пользователей является CatBoost с настроенными параметрами.

Предобработка данных и инженерия признаков оказались критически важными для повышения качества модели.

6. ЗАКЛЮЧЕНИЕ

В ходе прохождения практики в компании VK успешно были выполнены все поставленные задачи и достигнуты цели.

Был освоен теоретический и практический опыт работы с алгоритмами машинного обучения, включая логистическую регрессию, решающие деревья и градиентный бустинг, а также с современными инструментами анализа данных и визуализации.

В рамках проекта была разработана и обучена модель для предсказания пола пользователей социальных сетей, которая показала высокие показатели точности на тестовой выборке в 87%.

Практика позволила закрепить теоретические знания и приобрести навыки работы с реальными данными, что может быть использовано в дальнейшем для курсовых и дипломных работ, а также в профессиональной деятельности в области машинного обучения и анализа данных.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Официальный сайт VK. Электронный ресурс. URL: https://vk.com/ (дата обращения 23.08.25)
2. Официальный сайт VK Education Practice. Электронный ресурс. URL: https://education.vk.company/program/vk_education_practice (дата обращения 23.08.25)
3. Официальная документация Python. Электронный ресурс. URL: https://docs.python.org/3/ (дата обращения 23.08.25)
4. Официальная документация pandas. Электронный ресурс. URL: https://pandas.pydata.org/docs/ (дата обращения 23.08.25)
5. Официальная документация NumPy. Электронный ресурс. URL: https://numpy.org/doc/ (дата обращения 23.08.25)
6. Официальная документация Matplotlib. Электронный ресурс. URL: https://matplotlib.org/stable/index.html (дата обращения 23.08.25)
7. Официальная документация Seaborn. Электронный ресурс. URL: https://seaborn.pydata.org/ (дата обращения 23.08.25)
8. Официальная документация scikit-learn. Электронный ресурс. URL: https://scikit-learn.org/stable/ (дата обращения 23.08.25)
9. Официальная документация Catboost. Электронный ресурс. URL: https://catboost.ai/docs/en/ (дата обращения 23.08.25)
10. Официальная документация XGBoost. Электронный ресурс. URL: https://xgboost.readthedocs.io/en/stable/ (дата обращения 23.08.25)

ПРИЛОЖЕНИЕ

Рабочий план-график прохождения практики с отметками о выполнении:

№ п/п	Сроки проведения	Планируемые работы	Отметка о выполнении
1	01.08.25	Ознакомление с VK Education Practice	+
2	01.08.25	Выбор теоретического курса и проектной задачи	+
3	01.08.25 – 13.08.25	Просмотр лекций	+
4	01.08.25 – 13.08.25	Изучение теоретического материала	+
5	08.08.25 – 13.08.25	Выполнение практической задачи	+
6	14.08.25	Ознакомление с обратной связи от экспертов по проделанной работе	+