

Медведев Давид Валерьевич

Data Scientist / ML-Инженер

После инженерного проекта по автоматическому расчёту инсоляции помещений (степень освещенности помещений солнечным светом), основанного на методах линейной алгебры, я всерьёз увлекся математикой, что в конечном итоге привело меня в МФТИ и к сфере анализа данных и машинного обучения. Хочу развиваться Data Science и ML-инжиниринга, сосредотачиваясь на прикладных задачах. Также рассматриваю возможность поступления в аспирантуру и участия в исследовательских проектах.

КОНТАКТЫ:

Telegram: @david_medvedev

GitHub: <https://github.com/SalLangg>

Интерактивное резюме: <https://sallangg.github.io>

ОБРАЗОВАНИЕ:

2024 — по настоящее время

МФТИ, ФПМИ — Магистратура

Факультет: Прикладная математика и информатика

Направление: Современная комбинаторика

2018 — 2022

Тюменский индустриальный университет — Бакалавриат

Направление: Расчеты строительных конструкций

ОБУЧЕНИЕ И КУРСЫ:

- **Центр "Пуск" МФТИ:** Продвинутые методы машинного обучения.
- **Кафедра интеллектуальных систем:** Введение в машинное обучение (Константин Воронцов) и программирование на Python (Мурат Апишев)

ПРОЕКТЫ:

BuildBIMClassify — ML-система классификации BIM-объектов по сметным работам.

Разработал и внедрил пилотную MVP модель для автоматического сопоставления элементов BIM-модели (Revit) с позициями сметного справочника - классическая задача, выполняемая по большей мере вручную.

Цель: сократить до минимума ручное сопоставления элементов по работам. Когда сбор справочника будет завершен, модель будет полностью обучена и внедрена в боевые проекты, что позволит значительно сократить ручной труд и повысить точность сметных расчётов.

Технологии: Scikit-learn, LightGBM

Описание:

Использован LightGBM для multi-label классификации с фокусом на Precision (92%).

При обучении нарочно были введены пропуски в ключевых переменных, что бы повысить устойчивость модели к человеческому фактору.

Подготовил пайплайн для дообучения модели при расширении справочника.

Сейчас работаю над повышением качества классификации к уровню 95-98%. После этого заменю категориальные признаки на их эмбединги, чтобы нивелировать грамматические ошибки при заполнении параметров.

MorseNet — Декодер аудио файлов с кодом Морзе

Цель: построить модель декодирования сигналов морзе, используя технологии, похожие на обработку естественного языка.

Технологии: PyTorch, FastAPI, MLflow, Docker

Описание:

- Построена нейросеть *CNN → LSTM* с *CTC Loss* для декодирования из аудиофайлов
- Для извлечения признаков использовались *Mel-спектрограммы* и *аугментации (time/freq masking)*
- Логирование метрик обучения происходит через *MLflow*
- Сохранение моделей на сервере
- Качество: **0.433 no Levenshtein distance** на *Kaggle* (15 место, лидер — 0.24)
- Реализован *FastAPI*-сервер с возможностью дообучения модели, независимо от инференса
- Решение упаковано в *Docker*

GitHub: https://github.com/SalLangg/Morse-Decoder_V2

Классификация изображений

Цель: построить модель для классификации 42 персонажей по JPEG-изображениям

Технологии: PyTorch, torchvision, seaborn

Описание:

- Разработана *CNN*-модель с 3 сверточными блоками
- Использованы техники *аугментации* и расширения тестовой выборки: случайные повороты, изменение яркости/контраста, горизонтальное отражение
- **96.56% accuracy** на тренировочной выборке для самой базовой модели.

GitHub: https://github.com/SalLangg/Personality_Prediction

Предсказание личности

Цель: Цель: разработать модель предсказания личности человека на основе данных

Технологии: Pandas, Numpy, Matplotlib, Seaborn, StratifiedKFold, CatBoostClassifier, scikit-learn

Описание:

- Проанализирована степень важности пропусков в данных
- Созданы новые признаки для расширения выборки
- Протестированы различные модели - *CatBoost*, *XGBoost*, *RandomForest*, а также их стейкинг

GitHub: <https://github.com/SalLangg/Image-classification>

RAG-LMM помощник инженера (В разработке)

Цель: разработать систему умного поиска по внутренней базе знаний компании с выводом найденной информации в качестве контекста LLM.

Технологии: модели компьютерного зрения, ORC, LLM, RAG

Задачи, которые предстоит решить:

- Решение проблем с неструктурированной документацией с помощью *ORC*
 - Интеграция примечаний с изображениями в систему
 - Проверка актуальности норм
 - Фильтрация галлюцинаций LLM
-

ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:

BIM-Менеджер / ООО «Партнер.Проект»

02/2023 — по настоящее время

- Внедрение и сопровождение информационных технологий в проектной организации
- Разработка систем автоматизации проектирования для поддержки и ускорения рабочих процессов
- Сбор данных с информационной модели по запросу аналитиков, с последующей выгрузкой в PostgreSQL
- Координация работы всех смежных отделов (около 40 человек)
- Обучение сотрудников, создание обучающих материалов и техническая поддержка
- Сбор аналитики с информационных моделей и плагинов

Разработал **пилотную ML-модель** для классификации объектов из информационной модели (BIM) по справочнику сметных работ.

Цель: сократить количество ручного сопоставления элементов работам и человеческий фактор.

- Модель построена на **стейкинге CatBoost, Random Forest и Logistic Regression**
- Основная целевая метрика — **Precision 97.3%**. Нужно минимизировать количество неверно классифицируемых объектов
- Подготовил пайплайн для сбора новых данных и полноценного обучения, когда справочник работ будет полностью готов и введен в работу

НАВЫКИ:

Языки программирования: Python, C# (базово)

ML: PyTorch, sklearn

Прочее: SQL, Docker, FastAPI, MLflow,