Программа курса «PYML: ML Практикум: От теории к промышленному использованию»

О курсе: 6-дневный интенсивный практикум предназначен для специалистов, которые хотят не только освоить современные методы машинного обучения, но и научиться внедрять их в реальные бизнес-процессы. Вы получите навыки, которые сразу можно применять в работе: от предобработки данных и построения моделей до их промышленной реализации, и взаимодействия с бизнесом.

Аудитория:

- Data Scientists и аналитики, желающие углубить знания в ML-продакшн.
- Разработчики, планирующие работать с ML-моделями.
- Руководители IT-проектов, которым важно понимать этапы ML-разработки.
- Специалисты смежных областей, стремящиеся перейти в Data Science.

Уровень подготовки:

- Базовый опыт программирования на Python (знание Pandas, NumPy — преимущество).
- Понимание основ математической статистики.

Продолжительность курса: 24 академических часа, 6 дней по 4 ак. часов дистанционно

Содержание программы

1. Задачи машинного обучения и основные подходы к их решению

Теоретическая часть: основные понятия; классификация задач, решаемых с помощью методов машинного обучения; виды данных, понятие датасета.

Практическая часть: первичный анализ датасета, предобработка данных, сквозной пример решения задачи машинного обучения.

2. Базовые методы решения задач классификации и регрессии

Теоретическая часть: определение и примеры задач классификации и регрессии. Линейные методы. Деревья решений. Проблема переобучения. Основные проблемы с данными.

Практическая часть: решение задач классификации и регрессии линейными методами и методами на основе деревьев решений. Подбор гиперпараметров. Примеры переобученных моделей.

3. Продвинутые методы решения задач классификации и регрессии

Теоретическая часть: Стэкинг, бэггинг. Градиентные бустинги на примере LightGBM, xgboost, catboost. GridSearch для подбора гиперпараметров. AutoML.

Практическая часть: подбор гиперпараметров по сетке значений для базовых алгоритмов. Разбор кейсов по обучению моделей градиентного бустинга.

4. Продвинутые методы обработки данных и работа с признаковым пространством

Теоретическая часть: Тематика обучения без учителя - кластеризация, понижение размерности. Визуализация и формирование гипотез о данных. Feature Engineering и Feature extraction.

Практическая часть: примеры решения задач кластеризации и понижения размерности. Feature Engineering для улучшения качества работы ML моделей.

5. Нейронные сети для решения задач машинного обучения

Теоретическая часть: нейронные сети, задачи CV, NLP. Предобученные сети, файнтюнинг. (1,5 часа)
Практическая часть: Примеры решения задач классификации изображений и текстов. Применение полносвязных сетей для решения задач классификации и регрессии. Сравнение с алгоритмами бустинга.

6. Практическое руководство по реализации моделей машинного обучения

Теоретическая часть: сохранение результатов экспериментов и версионирование моделей и датасетов. Воспроизводимость экспериментов. Форматы сериализации моделей и пути продуктивизации: функциональная обертка, REST API сервис. Лучшие практики.
Практическая часть: создание веб-сервиса на фреймворке FastApi. Демонстрация MLFlow и AirFlow.