“Автоматизация бизнеса методами машинного обучения”

Automating Business Processes using Machine Learning

by

Павел Велихов https://ru.linkedin.com/in/velikhov

Леонид Жуков http://hse.ru/staff/lzhukov

Аннотация

Машинное обучение стремительно меняет мир, каждая успешная современная организация сегодня рассматривает машинное обучение и Data Science как один из основных инструментов повышения бизнес показателей, будь то доходность, размер аудитории и объем продаж. Вы уже разобрались как работают алгоритмы машинного обучения и хотите применить их для решения задач бизнеса? Тогда этот курс для вас!

При применении алгоритмов машинного обучения для оптимизации бизнес процессов компаний и организаций возникает множество сложных практических задач. В этом курсе мы рассмотрим всю цепочку шагов от сбора, интеграции и чистки данных, построения моделей, выбор функций потерь, сравнение, тестирование и мониторинг внедренных моделей с помощью метода нахождения аномалий.

В этом курсе почти не будут рассматриваться сами алгоритмы машинного обучения. Основная цель курса - познакомить слушателей с полной картиной применения машинного обучения, с различными задачами и процессами, которые при этом возникают. После окончания курса вы будете теоретически и практически подкованы для решения задач Data Science на практике и будете знакомы с самыми перспективными подходами к решению всего цикла задач Data Science. Все задачи будут приложены к примеру реального бизнеса кредитных организаций.

Пререквизиты

- Базы данных
- Машинное обучение

- Программирование на Python

Система оценки:

- 80% складывается из теоретических и практических заданий в курсе
    - 3 задания на программирование
    - 4 теоретических заданий
- 20% экзамен

Экзамен засчитывается автоматом, если качественно выполнены все задания.

Состав учебного курса

Лекция 1: Введение

Почему нельзя просто натренировать модель машинного обучения, быстро внедрить и заняться другим проектом? Какие возникают дополнительные задачи, какие риски во время эксплуатации модели, как понять какую пользу приносит модель. Общий обзор системы принятия решений на основе Data Science. Обзор основного кейса для курса - пример онлайн кредитной организации, модель бизнеса, основные показатели эффективности, ограничения на возможные решения, фундаментальные проблемы.

Семинар 1: Детальный обзор бизнеса кредитной организации и как это отражается на системе автоматизации принятия решений на основе data science.

Лекция 2: Данные, разные типы данных, модели данных

Данные в современном бизнесе бывают очень разных видов, от табличных или реляционных данных, до текстов, картинок, видео и списков гетерогенных событий. В этой лекции мы рассмотрим как моделировать такие данные и какие алгоритмы и методы лучше всего работают с этими типами данных.

Семинар 2: Обзор данных, которые накапливает и к которым имеет доступ кредитная организация. Обзор первого задания на программирование.

Лекция 3: Событийная модель как основная модель данных бизнеса. Kappa и Lambda архитектуры. Работа с логом событий (Process mining).

Бизнес является совокупностью бизнес процессов, а событийная модель данных бизнеса фиксирует все важнейшие точки бизнес процессов. Лог таких событий является основой современных архитектур информационных систем компаний, таких как Kappa и Lambda. Восстановление процессов из логов событий, тестирование гипотез на логе событий. Создание удобных для машинного обучения витрин данных из лога событий.

Семинар 3: Разбор событийной модели кредитной организации. Построение витрин для удобства построения запросов к данным. Ad-hoc запросы к данным. Process Mining и типичные запросы в этой парадигме анализа. Когортный анализ. Запросы к логу событий на языке XQuery.

Лекция 4: Интеграция и чистка данных.

Типичные проблемы с качеством данных, классы ошибок в данных, методы повышения качества данных. Задача интеграции данных. Понятие семантической интеграции, интеграция на уровне представления данных. Виртуальная интеграция данных и создание хранилищ данных.

Семинар 4: Детальный разбор проблем интеграции и качества данных в кредитной организации. Оброз второго задания на программирование.

Лекция 5: Версионирование данных

Проблемы при работе с большим количеством источников и версий. Системы версионирования данных. Воспроизводимость результатов. Проблемы версионирования, возникающие при командной работе.

Семинар 5: Детальный разбор модели контроля версий git.

Лекция 6: Построение моделей и системы принятия решений.

Создание признаков для моделей. Понятие витрины признаков и пайплайнов. Workflow для создания признаков. Версионирование признаков. Выбор функции

потерь. Выбор функции выгоды. Различные проблемы оптимизации экономики компании.

Семинар 6: Детальный разбор проблем оптимизации экономики кредитной организации. Инструменты созданий пайплайнов и workflow.

Лекция 7: Коллективное машинное обучение

Примеры коллективного машинного обучения. Задача детекции мошенничества в кредитной организации.

Семинар 7: Основные отличия коллективного машинного обучения от традиционного.

Лекция 8: Сравнение и тестирование моделей.

Сравнение моделей в терминах эффекта для бизнеса. Сравнение моделей для разных сегментов клиентов бизнеса. Комбинирование моделей. Тестирование моделей перед эксплуатацией.

Семинар 8: Разбор конкретных кейсов сравнения моделей для кредитной организации. Как более слабую модель можно выдать за равноценную более сильной, и как избегать таких сравнений.

Лекция 9: Эксплуатация моделей, отслеживание аномалий

Мониторинг системы принятия решений. Понятие аномалии, отслеживание аномалий. Аномалии в много-размерном пространстве признаков. Современные модели детекции аномалий. Машины Больцмана и вариационные авто-кодеры.

Семинар 9: Аномалии для объектов со сложной внутренней структурой на примере данных о кредитной истории и истории транзакций клиентов.

Лекция 10: Перспективные направления в Data Science, подведение итогов курса.

Организации только сравнительно недавно начали массово проектировать и внедрять системы принятия решений, основанные на data science. Эти системы часто кардинально меняют устоявшиеся подходы ведения бизнеса и построения остальных

IT систем организации. Мы рассмотрим несколько перспективных направлений в data science, включая интересные старт-апы в этой области.

Семинар 10: Технологические старт-апы в Data Science: Anodot, Tamr, Trifacta.

Задания на программирование

В этом курсе предусмотрены три задания на программирования для закрепления материала курса и приобретения практического опыта работы в области data science. Все три задания строятся на синтетических данных гипотетической кредитной организации, но максимально приближены к реальным данным настоящих компаний.

Задание 1: Готовим данные для моделей

В этом задании нам даны основные исторические данные кредитной организации по своим клиентам в формате лога событий. Также есть данные о поведении клиента на сайте компании. Кроме этого, надо проинтегрировать следующие источники данных:

- данные по преступности в регионах
- данные об экономическом благосостоянии граждан
- данные о кредитных историях из 2-х разных бюро
- геолокационные GPS данные некоторых пользователей

После интеграции мы построим профили клиентов. Мы применим два различных метода, и в следующем задании сравним результаты.

Задание 2: Строим новую модель скоринга, доказываем менеджменту, что она лучше существующей

В этом задании мы подготовим признаки для машинного обучения с помощью пайплайнов. После этого мы обучим новую модель для скоринга клиентов и сравним ее с существующей моделью. Наша задача будет доказать менеджменту, что новая модель работает намного лучше чем существующая, поэтому надо будет найти показатели, которые в наилучшем виде характеризуют нашу новую модель. Так же мы оценим эффект от двух методов построения профилей клиентов.

Задание 3: Ловим фрод через коллективное обучение и аномалии

Наш бизнес на основе data science заработал, но он подвергся атаке фродстеров. Хитрые хакеры написали роботов, которые притворяются людьми, берут кредиты на настоящие и поддельные документы и не возвращают деньги. Мы научимся обнаруживать такие атаки с помощью аномалий и автоматически фильтровать всплески активности фродстеров с помощью коллективного машинного обучения.

Теоретические задания

Задание 1: Разновидности данных

В этом задании мы научимся выбирать правильные модели для данных, выражать и тестировать гипотезы с помощью этих данных. Также поймем в каких случаях надо использовать когортный анализ.

Задание 2: Интеграция и версионирование данных

В этом задании мы придумаем и оценим алгоритмы чистки и интеграции данных, а также проанализируем разные схемы версионирования и их свойства.

Задание 3: Функции потерь и выгоды моделей

В этом задании мы проанализируем разные функции выгоды моделей, научимся подбирать подходящие функции для разных типов задач, и научимся выбирать функции потерь для машинного обучения, которые лучшим образом соответствуют заданной функции качества.

Задание 4: Оценка экономической эффективности

В этом задании мы научимся оценивать модели с точки зрения экономической эффективности. Конечно, создать полную экономическую модель организации и использовать ее для оценки на практике - нерешаемая задача, поэтому мы будем использовать различные упрощенные модели.