ДЗ «Урок 6. Семинар. Создание проекта машинного обучения» (теория)

Примерами задач, которые решаются благодаря Data Science, являются:

1. Прогноз и сокращение оттока клиентов;

2. Создание персонализированных предложений для покупателей;

3. Оптимизация закупок на производстве;

4. Поиск целевой аудитории для продукта;

5. Автоматизация прогноза цены на товары и услуги в зависимости от сезона;

6. Анализ загруженности на автодорогах.

В бизнесе берется какая-то конкретная задача в этой области (рис. 21). Например, нужно провести оптимизацию цены закупки с целью повышения прибыли организации. Необходимы проведение анализа данных, их визуализация и принятие обоснованных решений на основе полученных результатов. Это является задачами Data Science, где машинное обучение может быть одним из инструментов для оптимизации цены закупки, а не самостоятельной областью рассмотрения.

Необходимо решить данную задачу для определения метода снижения средней закупочной цены на 2%. В результате этого прибыль организации должна увеличиться на 4%. В масштабах большой компании эти 4% приведут к добавочной прибыли, измеряемой миллионами и миллиардами денежных единиц.

Рис. 21. Применение DaS в бизнесе

В современном бизнесе собираются тысячи метрик [6, 9]. Задача аналитиков – собрать эти данные, проанализировать и представить на дашборде (от англ. dashboard) – инструменте, на котором

визуализированы проанализированные данные[1]. Чтобы структурировать данные и затем проводить анализ, аналитик использует различные инструменты, например:

• Яндекс.Метрика – сервис для просмотра, анализа и визуализации статистики веб-сайтов, их оценки посещаемости и анализа поведения пользователей;

• Power BI – набор программного обеспечения от Microsoft для структурирования, анализа и визуализации данных. На основе этой системы аналитики придумывают, в каких схемах будут храниться данные, описывают структуры таблиц;

• Tableau – сервис для визуализации и анализа данных. С его помощью аналитик может оформлять отчеты в виде визуальных элементов, которые облегчают восприятие сложной статистической информации;

Продолжим рассмотрение проблемы из рис. 21. Работа автоматизированного блока по ценообразованию позволяет обновлять цены каждый день в нескольких тысячах магазинов. Остановка этого процесса на ночь недопустима. Отсутствие актуальных цен и продажа по старым ценам неизбежно влечет штрафы и санкции. Работа строится следующим образом. Выделяется проектная группа, которая изучает проблему, на чём она строится, как с этим работать. Они изучают данные: как формируется ценообразование, как складываются цепочки закупок, куда эта информация поступает, кто принимает решение и т.д. – то есть идет сбор информации. Дальше, когда все эти данные собраны, они подвергаются чистке: могут быть ошибки, могут быть товары, которые уже не используются, которые вывели из продуктовой матрицы и т.д. [18].

«Очистили» эти данные, дальше делается их анализ: сравнивается, что значимое, что не значимое, что оказывает наибольшее влияние на цену. Таким образом, формируется некая бизнес-модель с набором параметров, которые влияют на оптимальное ценообразование. Примеры параметров:

1. Удаленность поставщика от складов;

2. Объем партии, которую мы закупаем;

3. Сезонность, спрос;

4. Количество и местоположение складов и др.

Здесь стоит остановиться и упомянуть такое понятие, как «Многомерный статистический анализ». Рассмотрим две переменные: площадь и цена квартиры. У нас есть 8 парных значений, мы можем нанести их на график синими точками. Каждая точка имеет 2 координаты (площадь, цена). Площадь здесь будет независимой переменной $\boldsymbol{x}$ (признак, причина), а цена квартиры – зависимой переменной $\boldsymbol{y}$ (результирующая, следствие) – рис. 22.

[1] Визуализация данных позволяет не только упростить исследование, но и представить сложные данные в наглядной форме для их презентации, помогая аналитикам объяснить свои выводы заказчикам и другим заинтересованным лицам.

Apartment area	Apartment price
27	1,2
37	1,6
42	1,8
48	1,8
56	2,6
57	2,5
77	3
80	3,3

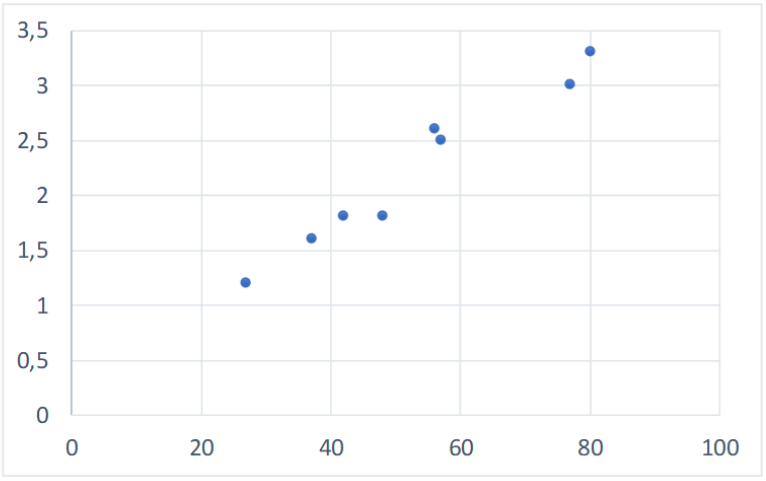

Рис. 22. Линейная регрессия. Двумерные данные

Но мы могли бы иметь несколько признаков, как в примере с продуктовой матрицей, которые влияют на результирующий фактор ***y***. Тогда каждая точка на графике задавалась бы не только тремя координатами, но также изменялись бы её цвет, размер, форма, непрозрачность и градиенты. На рис. 23, построенном при помощи языка программирования Python, отображены данные об автомобилях. Визуализируем 6 измерений для 205 машин. Такие данные уже будут называться многомерными. Кроме очевидных количества лошадиных сил, снаряжённой массы (совокупной массы автомобиля с водителем) и стоимости на графике эмулированы ещё три измерения:

• пробег в городских условиях, который уменьшается с более светлым оттенком маркера. Можно заметить, что пробег меньше у машин с большей ценой, мощностью мотора и массой;

• размер двигателя прямо пропорционален размеру маркера. Чем больше двигатель, тем выше цена и меньше пробег;

• форма маркера позволяет отобразить до десяти характеристик, но здесь квадрат отображает 4 двери, круг – 2 двери.

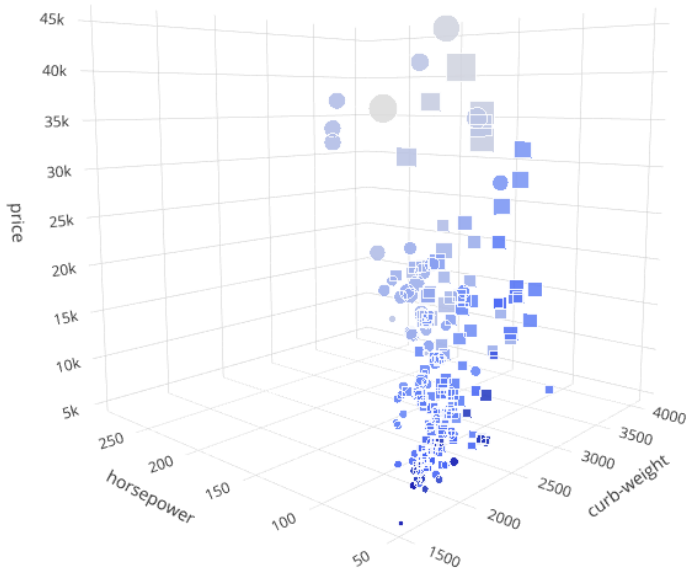

Рис. 23. Анализ многомерных данных на примере рынка автомобилей

Занимается изучением этих данных многомерный статистический анализ – раздел математической статистики, который посвящен исследованиям экспериментов с многомерными наблюдениями.

Вернемся к бизнес-модели с продуктовыми складами. В этой модели мы находим оптимальные параметры и составляем её так, чтобы суметь выбрать оптимальных поставщиков на оптимальные группы товаров с оптимальными размерами закупок. Получившаяся модель детально описывается и интегрируется в систему для ежедневного определения оптимального ассортимента закупок товаров в соответствии с их объемами для различных поставщиков. Это способствует автоматизации закупочных процессов и повышению их эффективности в рамках ежедневного функционирования системы.

После завершения разработки проводится внедрение, последующее тестирование и доступные усовершенствования, что позволяет бизнесу покупать товары по сниженным ценам, а компании – обеспечить себе дополнительные возможности для повышения прибыли. Однако, в реальной жизни намного сложнее. Для того, чтобы это сделать: собрать параметры, проанализировать, построить модель, автоматизировать и запустить, чтобы это в итоге давало доход бизнесу – требуется от нескольких недель до нескольких месяцев [6, 9], в зависимости от размера компании, сложности задачи, от её параметров, а также опыта и квалификации IT-аналитиков.

При анализе производственного процесса необходимо применить оптимизацию, учитывая значительное количество перемещений, складов, сроков исполнения и согласования всех этапов, а также другие аспекты (см. рис. 24).

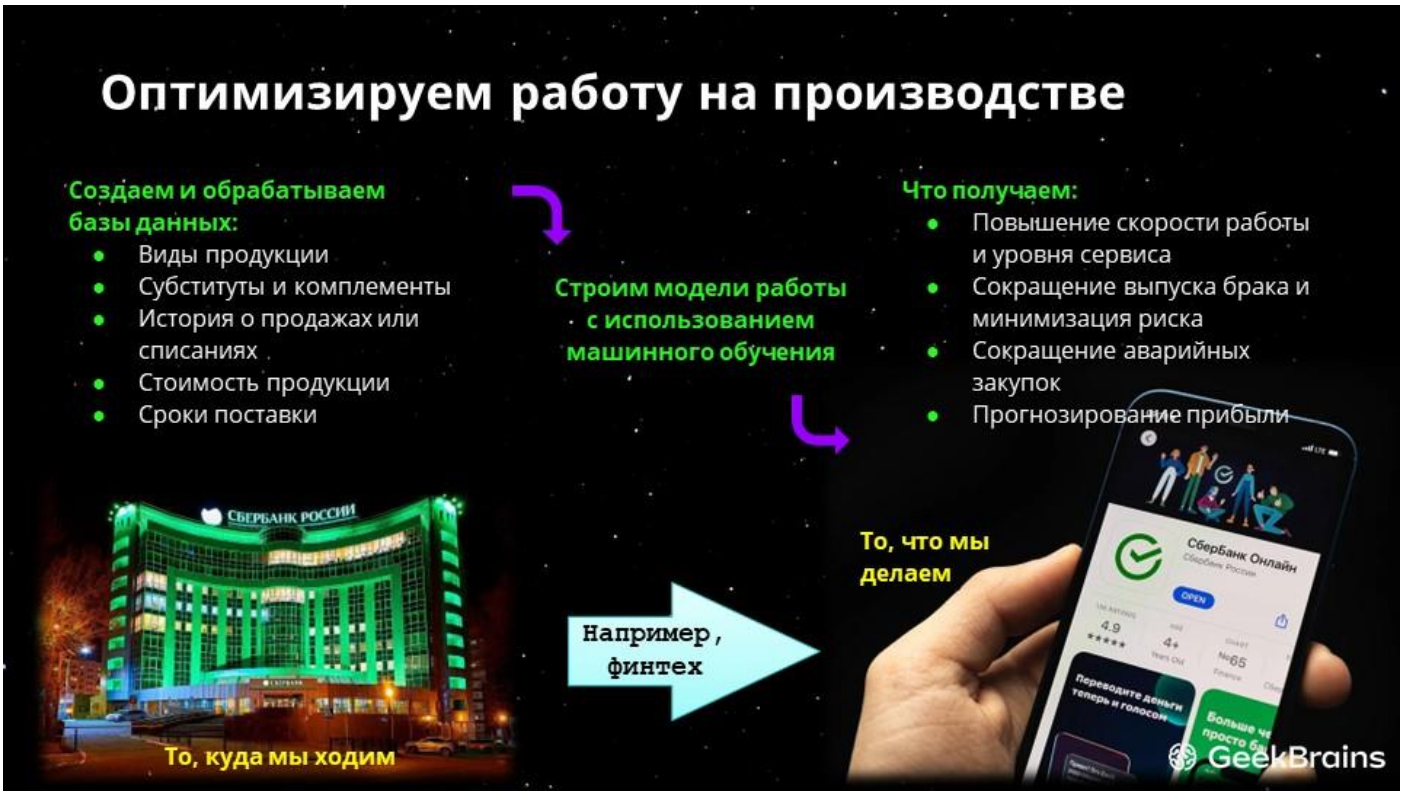

Рис. 24. Пример оптимизации работы на производстве [6]

Здесь тоже строится многофакторная модель, определяются некие оптимальные маршруты движения деталей или собираемого товара, на выходе при производстве это работает быстрее, сокращается количество операций, которые необходимо выполнять, количество перемещений, идет

меньше бракованного товара и т.д. Поэтому на производстве наука о данных и машинное обучение также активно используются для оптимизации [19].

Говоря о моделях более конкретно, можно сказать, что для решения задачи оптимизации цен закупки и повышения прибыли организации может быть использовано несколько типов моделей машинного обучения в зависимости от доступных данных и конкретных характеристик проблемы. Вот несколько возможных вариантов:

1. Линейная регрессия – этот метод может быть полезен для анализа данных о ценах закупки и прибыли организации. Он позволяет выявить линейные зависимости между различными переменными и предсказать прирост прибыли при изменении цен закупки.

2. Деревья решений и случайный лес – эти методы могут использоваться для выявления сложных зависимостей между ценами закупки и прибылью организации. Они могут учитывать нелинейные взаимосвязи и включать в анализ большое количество переменных.

3. Градиентный бустинг – этот метод также хорошо подходит для решения задачи оптимизации прибыли. Он может использоваться для построения ансамблей моделей, которые учитывают различные аспекты данных и позволяют добиться лучших результатов.

4. Кластеризация – если имеются данные о различных группах товаров или поставщиков, кластеризация может помочь выделить оптимальные группы для закупок и определить наиболее эффективные стратегии ценообразования.

Выбор конкретного типа модели зависит от специфики данных, доступных ресурсов и требований к точности предсказаний. Важно также учитывать необходимость регулярного обновления модели и её интеграции в процессы компании для эффективного использования в реальном времени.