22 июня 2018

Как правильно заказывать проекты с Machine Learning

Data Scientist департамента BI ГК «КОРУС Консалтинг» Павел Матюсов рассказал изданию Roem.ru об основных сложностях, которые встречаются в проектах по Machine Learning.

Несколько лет работы с проектами в области машинного обучения, участие в соревнованиях Kaggle, работы как с платформенными платными и бесплатными решениями, так и решение задач с помощью языков RPython дают мне возможность сделать небольшой обзор того, как обычно начинается проект по Machine Learning, какие сложности возникают и как их преодолеть.

Часть 1. Взгляд изнутри


Кому Matiusov Pavel

Привет. Нужно сделать КП для компании Ромашки. Компания занимается выращиванием и сбытом ромашек. Нужно прогнозировать отток постоянных покупателей по каждому сорту, текущий результат неудовлетворительный, хотят точности больше 90%. Напиши примерные трудозатраты.

--
С уважением,

«Как можно верно оценить такую задачу?, — подумал я. — Еще и критерий точности >90%?. Интересно, как они его считают и эта фраза «по каждому сорту», сколько их там? Ладно, можно готовить очередной опросник».

Я немного утрирую ситуацию, которая происходит достаточно часто, но давайте пройдемся по пунктам:

- Заказчик обычно не знает возможностей машинного обучения и специфики его успешного применения. Например, предсказать спрос по каждому сорту (допустим, этих сортов около 1000) будет в разы сложнее, чем предсказать общий спрос на всю продукцию. А для некоторых сортов и попросту невозможно. Добиться 90% точности для каждого — невозможно почти на 100%. Хотя бы потому, что 10 сортов будут продавать по 1–3 штуке в абсолютно случайном месяце без каких-либо зависимостей, а еще 10 зависят только от желания зарубежных оптовых заказчиков и абсолютно не коррелируют ни с одним показателем из возможных.

Из этого можно вытекает один вывод: некоторые задачи, полностью или почти полностью нереализуемые еще на стадии формулировки. Я подчеркну, речь не о сроках, данных, а именно о формулировке задачи, и наиболее популярные камни преткновения тут — это точность и горизонт предсказания. Невозможно на два года вперед сказать, сколько будет продано молока с ошибкой в 3% максимум, как невозможно для абсолютно всей продукции вашего магазина сделать точный прогноз. Казалось бы, это логично, но на практике «космических» запросов на реализацию хватает.

- Почти всегда встречаются понятия: точность, спрос, отток и другие популярные формулировки. Но мало кто задумывается, как сильно можно и даже нужно углубляться в формализацию этих понятий. Например, разберем понятия точность и отток.

Точность можно измерять разными способами, в интернете легко найти минимум 10 разных метрик. Возьмем одну их простых: модуль разности (прогноз — факт) и поделим результат на факт. Допустим, по факту было продано 100 машин, предсказали 110 машин. Считаем, 110−100 = 10, затем этот результат делим на факт — 10/100 = 0.1. Ошибка составила 0.1 * 100% = 10%. Значит, точность будет равна 1 — ошибка = 90%, что, согласитесь, звучит неплохо.

А теперь представьте, что наша задача предсказать продажи дорогих и редких машин, их продают по 4−10 штук в месяц, подойдет ли нам такая метрика? Допустим, по факту было продано 7 машин, предсказали 9 машин. Считаем, 9−7 = 2, затем делим этот результат на факт — 2/7 = 0.28. Ошибка составила 0.28 * 100% = 28%. Значит, точность 1 — ошибка = 72%. Очень далеко

до 90% верно? А ошиблись-то всего на 2 машины...

Очень важно до старта проекта убедиться и обговорить метрику точности решаемой задачи, особенно если от нее зависит его успешная реализация.

Или, например, возьмем понятие отток. Надо понимать, что, если отток воспринимать как «покупатель наших машин в следующем месяце ничего у нас не купит», то это один тип и сложность задачи. А если отток — это «покупатель наших машин еще три месяца покупает наши машины, затем месяц покупает только запчасти для них, затем ничего не покупает», сложность точного прогнозирования этой задачи становится в разы сложнее.

Тут у нас появляется следующий вывод, который несомненно поможет как заказчикам, так и исполнителям: максимально формализуйте и уточняйте требования до старта проекта. Не только требования, но и сами термины. Это, безусловно, важно во всех типах проектах, но в машинном обучении требуется максимально четкой детализации. Несколько примеров, как нечеткая формулировка заставляла пересматривать сроки проекта, а порой и способы его реализации:

- В популярной задаче создания рекомендательных систем был использован алгоритм коллаборативной фильтрации, давайте называть его просто алгоритм. В нем необходимо были данные по оценке пользователем неких объектов. Например, объект Б, оценен на 3 балла из 5. После первичного анализа, появилось слишком много девиантов — людей, чье мнение постоянно не совпадало с мнением большинства. Но как оказалось в дальнейшем, это была оценка пользователей, которые не смотрели фильм или сделали вывод по трейлеру. Пришлось их убирать, а это

существенное уменьшение обучающей выборки и некоторые другие сложности. Вот так определили новое понятие — оценка.

- В задаче предсказания некого объекта было определено, что объект един и неделим. И заказчик будет оценивать именно предсказание этого объекта. Оказалось, что объект на определенном горизонте распадается. Т.е. на три месяца вперед, предсказав 100 объектов, мы будем очень точны. Но проблема в том, что на 4, эти 100 уже будут 200-ми, и мы сильно потеряем в качестве. Пришлось усложнять проект, вводив дополнительную модель предсказания деления объекта на два новых объекта. Сроки существенно увеличились.

Заканчивая эту часть хочется упомянуть об обучающей выборке. В тех случаях, когда мы используем алгоритмы, которые требуют обучения на исторических данных, важно, чтобы история была, верно? И здесь есть несколько важных моментов:

- Нам еще непонятно, какие данные придется выгружать и какие признаки для решения задачи формулировать.
- Достаточно часто разговор скатывается на уровень «дайте нам, все что есть», «нет, вы скажите, что вам нужно, мы выгрузим».

Например, чтобы спрогнозировать отток постоянных покупателей, нам понадобятся характеристики сортов ромашек, данные по продажам, цены, средние значения и так далее. И здесь начинается самое интересное. Потому что, например, использование API Google.Maps или Яндекс.Карты может показать, где расположена ближайшая цветочная точка конкурентов

и подскажет, где много ли сколько вокруг жилых домов, и есть ли рядом кладбище. Можем ли мы это использовать? Ведь если рядом через месяц построят оптовую цветочную базу, спрос упадет, а мы этот фактор не внесли в обучение.

А вдруг у заказчика уже есть готовое мобильное приложение, которое позволяет ставить лайки понравившимся сортам ромашек по торговым точкам? И если количество лайков падает, можем ли мы это использовать? В каком виде это хранится?

И логичная ситуация, что по нескольким сортам история продаж — три недели. А прогноз нужен еще на три. Это тоже совсем нерадостная ситуация

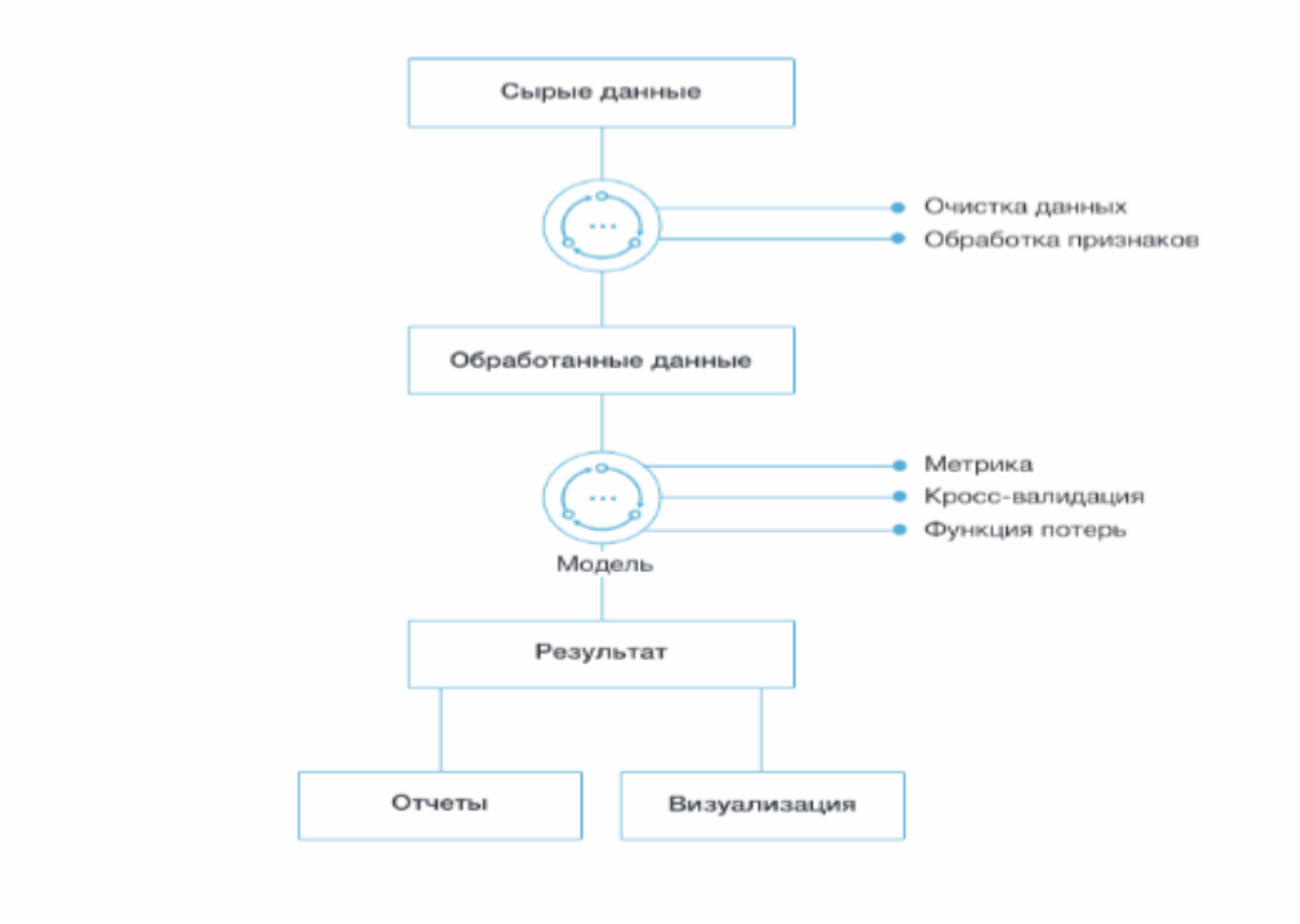
Тут нет готового рецепта, на этой стадии нужно очень активно сотрудничать с бизнес-заказчиком задачи. А лучше непосредственно с тем, кто занимается вплотную этой задачей. От этих людей можно узнать множество дополнительной информации. Продавец может вам сказать, что сорт ромашек номер 22 можно предсказывать по-другому, он всегда идёт по 1 штуке на букет других ромашек.

Делаем вполне очевидный четвертый вывод: продумывайте заранее, какие данные могут потребоваться, в каких объемах, и оценивайте качество и полноту обучающей выборки. Чем больше данных, тем лучше, но некоторые могут быть очень трудозатратными ещё на стадии их получения. И как можно чаще и больше общайтесь и узнавайте, какие факторы или особенности есть у всех процессов. Добавить что-то в конце часто бывает проблематичным.

Часть 2. Процесс

— Добрый день, через две недели сможете показать первые результаты? Как ждать три месяца???

В этой части хочется разобрать сам процесс реализации ML-проектов — если очень кратко, то выглядит он примерно так:



Перед самим процессом нужно понять две очень важные вещи, про которых частенько забывают или просто не знают. Сначала поймем, что умеет в общем машинное обучение.

Машинное обучение умеет:

- Получив объекты (человек, изображение, сим-карта, цветок, пара людей), извлечь признаковое описание объектов (рост, цвет волос, размер одежды, количество детей, образование, наличие смартфона);
- Посмотрев на объекты, научиться:
- Классифицировать (мужчина/женщина)
- Прогнозировать значения для объектов (возраст, доход, рост)
- Группировать (школьники, бизнесмены, политики, любители чая).

Вспомнив всё то, что мы узнали из первой части, приходим к очень необычному выводу: машинное обучение может решать почти все задачи, которые мы можем придумать в рамках тех требований и ограничений, о которых я писал выше.

- Предсказать, сколько погибнет ромашек, а сколько вырастет? Можно.
- Предсказать, какой новый цвет ромашек будет в тренде в следующем сезоне? Легко.
- Классифицировать для нас работников, которые уволятся через год? Запросто.

И т.д.

Что нам понадобится для этого?

Для первого пункта: данные о температурах, грунте на грядках, чем обрабатывали цветы, их сорта, кол-во осадков, состав воды поливы и т. д.

Для ответа на второй вопрос подключаемся к лидерам мнений (в чем была Бузова последний раз в Instagram?). Распознаем цвет и фасон платьев ковровой дорожки Каннского кинофестиваля, собираем цветовую гамму модных домов Парижа.

Для прогноза об увольнениях собираем данные по опозданиям, рабочим места, количеству сотрудников, близости к метро, логов веб-браузеров и т. д.

И первая важная вещь, которую мы понимаем: при наличии обучающей выборки и исторических данных для реалистичного горизонта в будущем, мы можем предсказать почти все. Вопрос здесь только в точности и целесообразности.

Но надо четко понимать, что чем сложнее и нетривиальнее задача, тем трудозатратнее и с бóльшей долей неизвестности качества она будет решена. Примерный список задач, которые можно решать на текущий момент, которые не находятся на стадии «исследования» и существуют достаточное количество успешных практик и вариантов решения:

- Обработка естественного языка (Natural Language Processing) — машинный перевод, анализ отзывов, выделение названий, логические выводы.
- Анализ социальных сетей (Social Network Analysis) — рекомендация друзей, поиск сообществ, выделение лидеров мнения.
- Анализ изображений и видео (Computer Vision) — выделение лиц на изображениях, извлечение номеров, названий с камер, стиля, трендов в одежде.

- Анализ аудио сигналов (Signal Processing) — распознавание речи, классификация музыки, рекомендация плейлиста.
- Рекомендательные системы (Recommended Systems) — рекомендация товаров, друзей, прогнозирование оценок к фильмам.
- Поиск ассоциативных правил (Association Rule Learning) — построение логических правил, анализ чеков.
- Поиск спама (Spam Detection) — Gmail, Mail.ru, Яндекс.Почта, ...
- Сегментация потребителей (Customer Segmentations) — Facebook, Google, Яндекс, Вымпелком...
- Выявление фрода (Fraud Detection) — Google, Facebook, Вымпелком, Сбербанк...
- Прогнозирование оттока (Churn Prediction) — Amazon, Netflix, Вымпелком, МТС, Мегафон, МГТС...
- Распознавание речи (Speech Understanding) — Apple (Siri)...
- Классификация изображений (Image Understanding) — Facebook, Google, Instagram, Яндекс, Mail.ru...

Такие задачи уже имеют свой алгоритм решения, и их можно почти без рисков реализовывать и внедрять.

Вторая важная вещь, которую нужно понимать перед началом реализации проекта, это требования к ресурсам и результаты пилота. Специфика реализации всего вышеперечисленного такова, что трудозатраты пилота

и не-пилота часто бывают идентичны, а мощности для реализации часто на время пилота требуются такие же или даже выше, того, что будет в продуктивном решении.

Разберем эти моменты поподробнее на нашем случае предсказания оттока покупателей каждого сорта ромашки. Давайте подумаем, как тут можно ограничить пилотный проект? Например, так:

Пилот	Полный проект
Предсказываем 1–2 сорта	Предсказываем все сорта
Предсказываем все сорта, но для одной точки	Предсказываем для всех точек
Предсказываем три разные по типу точки и трем максимально разным сортам	Предсказываем сорта и точки

А теперь по трудозатратам:

Что делаем для пилота	Что делаем для полного проекта
Собираем все признаки	Собираем все признаки
Делаем преобразование и подготовку данных	Делаем преобразование и подготовку данных

Перебираем различные алгоритмы, ищем лучший результат	Уже знаем лучшие алгоритмы, настраиваем их и получаем результат

В этой схеме время на продуктивное решение может получится даже меньше, чем время в пилотном проекте. Конечно, бывают и иные ситуации, но и эта совсем не редка. Более того, некоторые проекты, в которых я участвовал, в принципе не могли пилотироваться, потому что для требуемой точности нужны были все возможные данные и ресурсы.

После того, как мы поняли эти две важные вещи, перейдем к популярному списку проблем, которые могут возникнуть при реализации:

- Новые форматы данных или их виды. Если для обучения мы использовали температуры в градусах, а теперь берем данные по ромашкам на грядках, то, как только для предсказания поступят данные в других измерениях, результат будет плачевный.
- Данные, которые мы используем для обучения, совсем не похожи на данные, на которых мы будем проверять успешность проекта.
- Требования показать промежуточный результат или пилотный результат за очень короткий срок.
- Нетривиальные задачи, которые можно реализовать, но для них невозможно предоставить референс, уникальные задачи.
- Многочисленные ошибки в исходных данных. Один департамент считает готовую продажу, это когда есть запись в базе, а другой, только когда

данные придут на расчетный счет.

- Невозможность соединить данные. В одной среде это ID покупателя, а в другой ID операции. Связи между ними нет. Ее можно сделать, но понадобиться время.
- Использование будущих данных. Например, полные продажи, которые учитывают еще и законтрактованные оптовые продажи за два месяца.

Это далеко не весь список, есть еще много технических моментов, но они скорее относятся к технической части, нежели чем к идейной, которую я раскрываю в статье.

В заключении хочется сказать, что понимание вышеописанных особенностей задач сферы машинного обучения существенно упростят реализацию проектов, а иногда помогут вообще избежать заранее рисковых проектов. Они помогут правильно подходить к общей оценке проектов и понимать достаточный и необходимый набор данных и условий.

Источник: Roem.ru