Мониторинг скоринговых моделей в продакшене

Слайд. Заголовок: Мониторинг скоринговых моделей в продакшене.

Текст: DevFest - 2025. Докладчик: Туров Алексей Data science expert в Beeline. Речь: Меня зовут Туров Алексей я являюсь data science expert в компании Beeline Kyrgyzstan на текущий момент. До этого я был senior data science specialist в банке Компаньоне. Где мы с нуля строили строили и построили за последние полгода новую инфраструктуру онлайн кредитования.

Слайд. Заголовок: Путь скоринговых моделей в бизнесе

Текст: Ситуация. Бизнес приходит и говорит что хочет ML -> Провели исследование в тетрадки -> получили крутую модель -> go go go to production А ЧТО ДАЛЬШЕ?????

Слайд 2.   
Вспомним теорию что такое показатели AuC roc.   
**AUC «на пальцах»: два быстрых способа**

## **Способ 1: путь по решётке (ROC-путь)**

1. Отсортируй по скору **по убыванию** и возьми только метки: у нас это **[1, 1, 0, 0, 1]**.
2. Есть **m=3** единицы и **n=2** нуля → решётка **3×2**.
3. Старт снизу-слева. Идём **вверх** за каждую 1, **вправо** за каждый 0.  
   * шаги: ↑, ↑, →, →, ↑
4. **AUC = доля клеток под маршрутом** = 4 клетки из 6 возможных пар = **4/6 = 0.67**.

Это интуитивно: вправо — «ошибки по нулям», вверх — «попали в единицы», площадь под кривой — сколько пар «1 выше 0» мы «собрали».

-> выдали(потеряли) деньги плохим заемщикам -> кредиты не вернули -> все пропало -> нафиг эти модели -> выдаем по бизнес правилам -> бизнес лучше знает.

«Почему мониторинг важен» — 2-3 тезиса, 1 картинка/мем.

Речь.Все начинается с цели для бизнеса Что?. В крупных компаниях только малая часть из моделей которые были на исследования попадает в прод, почему? Есть целые отделы валидации, Зачем? В прод выводят модели только если есть вся инфраструктура и правильный мониторинг. Когда?

3 Слайд.Заголовок: Зачем нужен мониторинг моделей в продакшене?

Текст:

3.1 Это делает бизнес вовлеченным в процесс. Наглядность(superset, tablue, power BI). Возможность управление процессом кредитования. Нет черным ящикам - полностью управляемый процесс. Не бизнес говорит нам что у нас проблемы а мы бизнесу

3.2 Вовлечение в процесс develops и разработка с ними процесса мониторинга сервиса (Grafana, prometheus, logging и т.д.). Живой сервис или нет, подключение к БД, кол-во заявок, время заявок, кто приходит и т.д. Разделение ответственности

3.3 Для команды data science это возможность отслеживать процессы, вовремя замечать деградацию модели или ее неправильную работу. Предугадывать проблемы а не ждать их. Проводить дальнейшие исследования и изыскания.

4. Наглядность

Проблема: Я провел исследование а дальше модель живет сама

Решение: Настроить мониторинг или купить готовое решение(databrics и другие). Для закрытых систем только свое решение и оно более гибкое

4.1 вы можете посадить аналитика и сказать чтоб он сделал так как вам надо, но без знаний в data science процесс скорее всего растянется до того, как аналитик не станет дата сайнс специалистом. Долго, дорого

4.2 Сделать самим. Стек power BI, sql, dwh, grafana, prometeus.

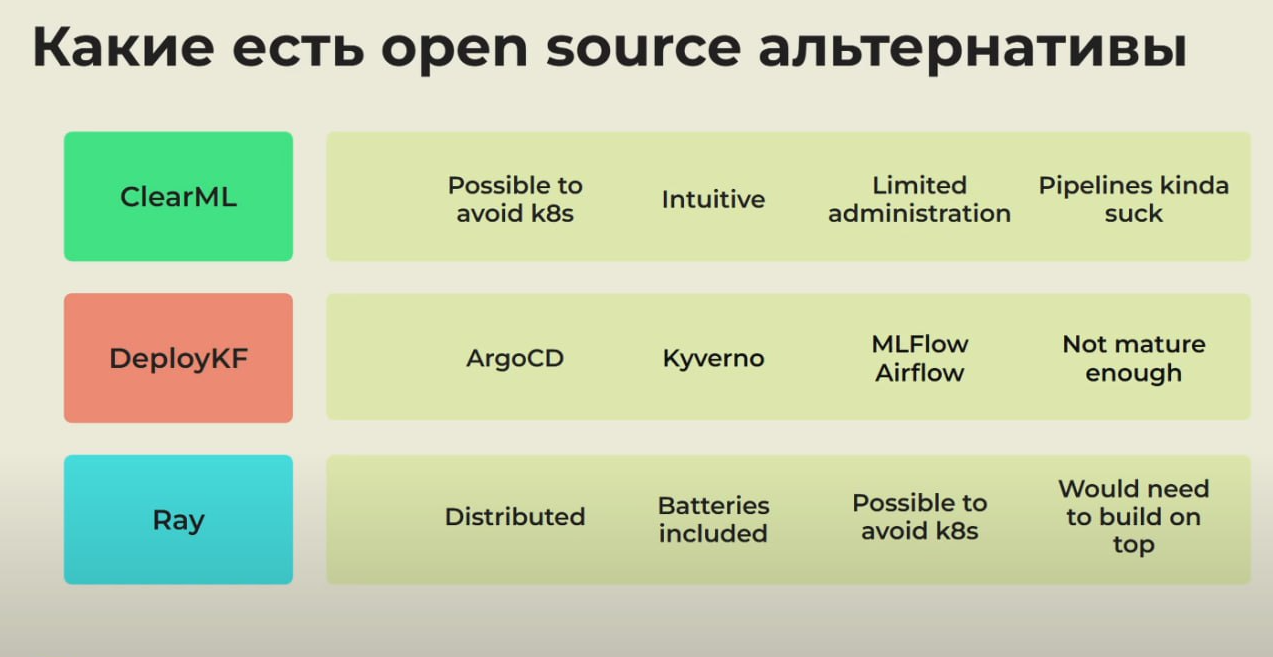
Просим у бизнес тех задание что они хотят видеть (FPD, SPD, par30-90, выдачи, социальный портрет и другие метрики - не имеющие отношение к модели только если в фичах). Может сделать и аналитик

Добавляем свои метрики модели(Mlflow, Airflow, Kafca, Featcher stories, CI/CD, Docker)

5. Что это за метрики. Что нужно дата сайентисты отслеживать в модели.

Что происходит когда дата сайнтист заканчиает исследования и отправляет модель в продакшен? Если у вас есть команда дата валидаторов то они еше до оправки проверяют модель на тестовых данных и контрольно валидируют и могут завернуть модель. Если нет такой команды то и валидация в лкчшем случае ложится на коллег п обычно на том же дата сайнтисте.

Как этого избежать есть много открытых программ для контрольирования экспериметров ML flow, Clran Ml. И они помогаю настроить весь процесс обучения развертывания в один pipline. Ml flow например позвволяет настроить выгрузки в продашен и имеет несколько стадий. test-dev-prod. Можно хранить артифакты и так далее. Почему это актуально, потому что некотрые хранят модели в gitlab и не понимают почему так нельзя.

(Здесь начинается MLops, оперируем не кодом а экспериментами, версии всего моделей кода датасетов, Kube Flows  


У нас есть эксперименты и теперь мы можем на валидации делать такие фишки как проверка data drift detected, drift concept()Короткое определение + чек-лист сигналов (χ², PSI) есть небольшой тестовый код), колмогорова смирнова хи квадрат Что такое дата дрифт это когда поменялись входящие данные. Очень актуально для скорингов так там огромное кол-во правил как внутренних так и регулятора.

И бывает почти всегда так что обучались на одних данных а модель предсказывает совсем иные.

что делаем дальше. Если у нас нет дрифта то не обязательно что он не появится и его надо отслеживать тут испльзуем airflow grafama mlflow

если вдруг появился дрифт модели - это плохо или хорошо ???? Смотрим что сдвинулось дрифт бывает разным и это отдельная тема для осуждения. Нашли дрифт Пришли моловые люди в сентябре - студенты и подали много заявок… А что делать?? Можем посмотрить исторические данные, и если данная группа склонна к риску то бизнем должен повесить ухудшающие коэффиценты. Так же можно использовать

6. CBPE: как оценить Gini без таргета  
Проблем азадержки таргета в скоригах. У нас есть вероятности(не приедикт). Но таргет срееет месяц два три шесть. А качество моджели нужно сейчас и тут на помощь временнные ряды CBPE  
Он помогает предсказать качество текущей модели и там есть также его диапозоны и алерты. Можно запустить простой код и показать наглядно

**Заголовок:** Что мониторим ежедневно: бизнес-прокси, пороги и сегменты

**Бизнес-прокси (Tableau):**

* Выдачи: шт/сумма (LCY)
* Заявки (входящий поток)
* Approve rate (AR), Issue rate (IR)
* FPD/SPD/TPD, PAR30/60/90
* Bad rate, EL/PL (если есть)
* Score-дистрибуции (корзины/пороги)

**Порог и политика:**

* Влияние threshold на AR/IR
* Доля «серой зоны» и hit-rate ручной проверки
* Смещение score у порога, исключения/лимиты по сегментам

**Кто подаёт заявки (сегменты):**

* Соц-демо: пол, возраст (баки/группы)
* Доход, тип занятости, стаж
* Регион / канал / продукт
* Новый vs повторный клиент
* История кредитов (лимиты, DPD)
* Устройство/платёжный метод/время суток
* Риск-корзины по скорбаллу

**Алерты (триггеры разбора):**

* AR/IR вышли за исторические коридоры
* Всплеск отказов/5xx/таймаутов
* Сдвиг score-дистрибуции у порога
* Рост FPD/SPD/TPD, PAR выше цели

Если хочешь — перенесу это прямо в твой PPTX-шаблон как готовый слайд и выровняю типографику под шаблон DevFest.

# **Слайд: Data Science monitoring — стек и практики**

**Заголовок:** Data Science Monitoring: стек и что именно отслеживать

**Пункты (для слайда):**

1. **MLflow / ClearML / Kubeflow** — трекинг экспериментов и **Model Registry** *параметры, метрики, артефакты; стадии: Staging → Production*
2. **Airflow** — оркестрация и автозапуски  
    *ETL/фичи/скoring/ретрейн; SLA/ретраи/backfill*
3. **Feature Store** (offline/online) — **point-in-time** корректность  
    *ключи сущностей, версионирование фич, backfill на истории, online-выдача на inference*
4. **CI/CD + Docker** — поставка моделей  
    *юнит/данные/контракт-тесты; образы; blue-green/canary; теги model\_version*
5. **Data & Score Drift** — стабильность входа/выхода  
    *числовые:* ***KS/PSI****, категориальные:* ***χ²****; score-дистрибуции, approve-rate*
6. **Статпорогa** (быстрые ориентиры)  
    *PSI: <0.1 ok, 0.1–0.25 предупреждение, >0.25 — дрейф; p-value χ²/KS < 0.01 — тревога*
7. **CBPE** — качество **без таргета сейчас** *онлайн-оценка AUC/Gini по уверенности модели и историческим калибровкам; коридоры и алёрты*

## **Что сказать к слайду (30–45 сек)**

«Трекаем эксперименты и версии в MLflow/Model Registry, оркестрируем пайплайны в Airflow. Фичи живут в Feature Store: offline для обучения и backfill’ов, online — для быстрых предсказаний с point-in-time гарантией. Поставляем модели через CI/CD и Docker с метками model\_version. В проде мониторим и вход (data drift), и выход (score drift): для числовых берём KS/PSI, для категорий — χ², плюс следим за approve-rate. Когда нет таргета — используем CBPE, чтобы понимать, куда движется AUC/Gini уже сегодня.»