Этические и юридические аспекты MLOps

Ирина Степановна Трубчик

Лекция 12

https://t.me/+PsC-JDrwrvsxNmVi

Responsible AI (Ответственный ИИ): три столпа

⚖️ Этика

справедливость, прозрачность, подотчётность

💼 Право

GDPR, CCPA, EU AI Act, соответствие законам

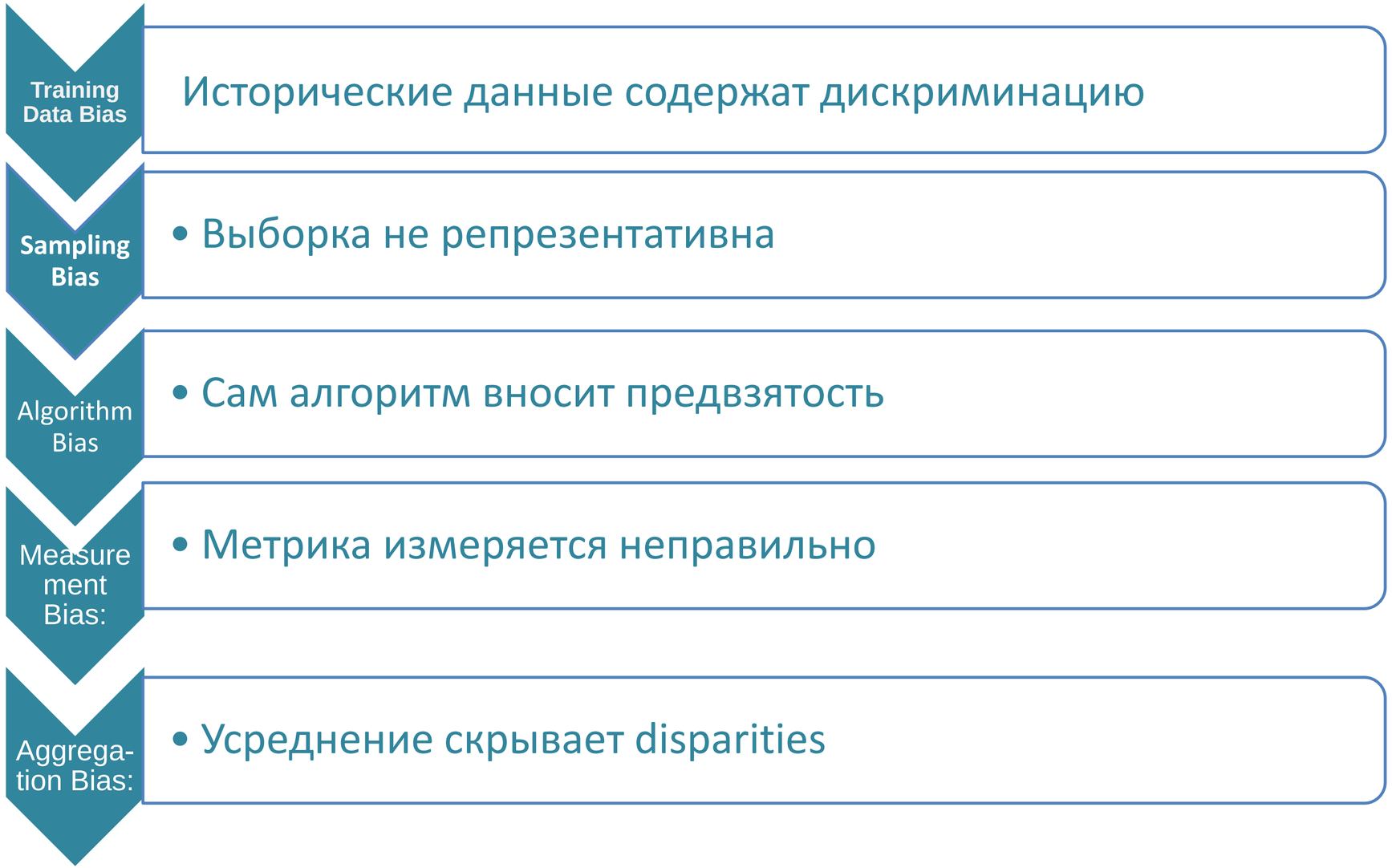
Управление

Процессы, структуры, ответственность

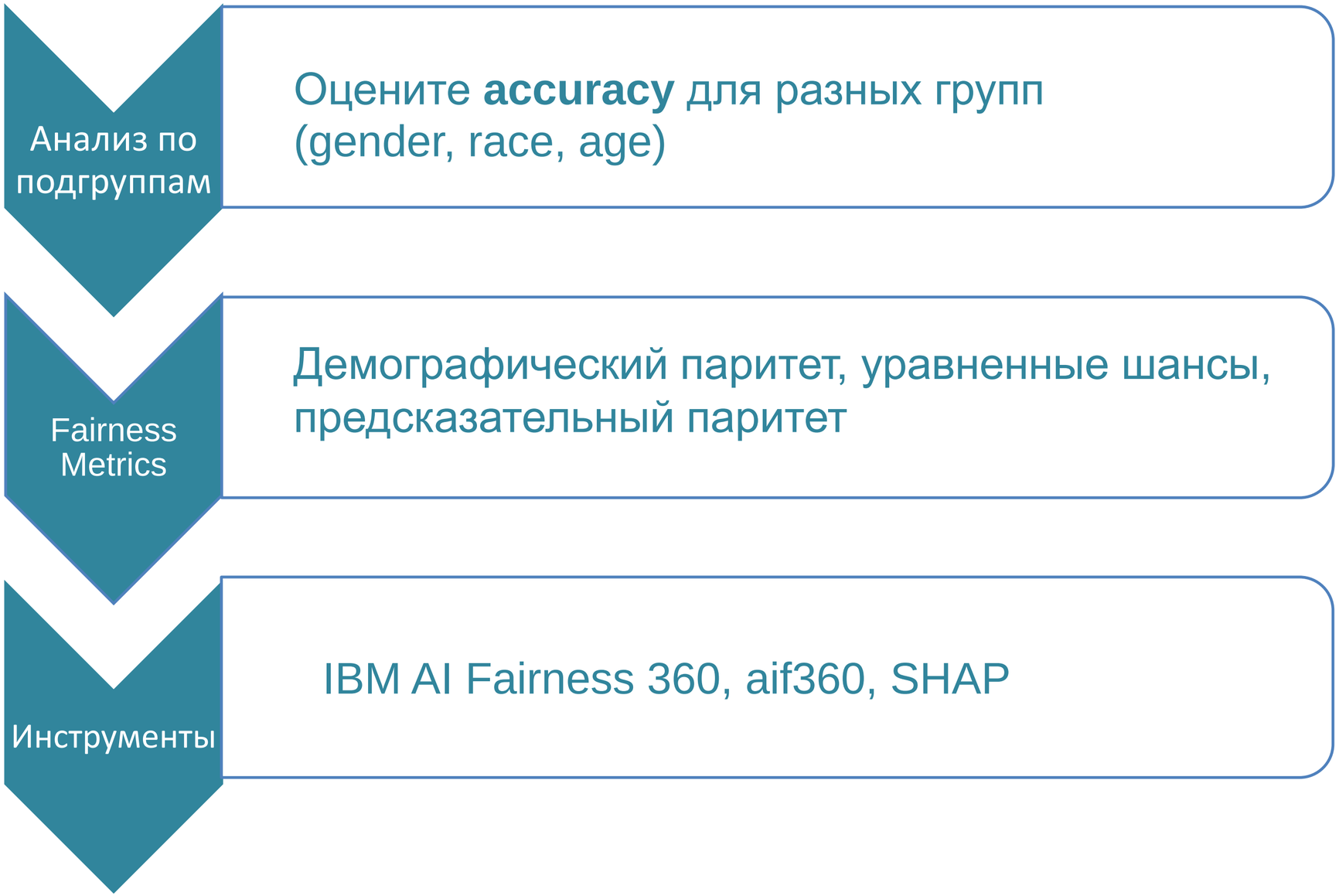
Высокопрофильные инциденты: мировые и российские

🚨 Amazon (2018): ML отклоняла женщин — дисбаланс в исторических данных
🚨 Google Photos (2015): Классифицировала чёрных людей как "gorillas"
🚨 СберБанк (2023+): Комплексная работа с bias в credit scoring — 110M+ клиентов
🚨 Яндекс.Метрика: GDPR compliance issues при обработке данных EU пользователей
🚨 X5 Retail: Оптимизация loss функции для справедливого ценообразования

Типы Bias в ML моделях



Как обнаружить Bias



1. Demographic Parity (Демографический паритет)

Формула: $P(Y=1|A=0)=P(Y=1|A=1)$

Смысл: Вероятность положительного решения (Y=1) должна быть одинаковой для обеих групп, независимо от защищённого атрибута (A).

2. Equalized Odds (Уравнённые шансы)

Смысл: True Positive Rate (TPR) и False Positive Rate (FPR) должны быть одинаковыми для всех групп.

3. Predictive Parity (Предсказательный паритет)

Смысл: Precision (точность положительных предсказаний) должна быть одинаковой для всех групп.

На практике: правило 80%

```
Accuracy для женщин = 0.85
Accuracy для мужчин = 0.92

Ratio = 0.85 / 0.92 = 0.924 = 92.4%

92.4% > 80% ✅ (но близко к порогу, требует внимания)
```

Fairness Metrics служат инструментом контроля качества:

✅ Обнаружение bias при разработке модели
✅ Документирование справедливости в Model Card
✅ Мониторинг performance по подгруппам в production
✅ Принятие обоснованных решений о deployment

Главный вывод: Мы не можем управлять тем, что не измеряем.

Законодательные акты

🌐 GDPR (EU)

- ✓ Lawfulness & transparency
- ✓ Data minimization
- ✓ Right to be forgotten
- ✓ Data portability
- ✓ Privacy by design

152-ФЗ (РФ)

- ✓ Согласие на обработку
- ✓ Ограничение хранения
- ✓ Безопасность ПДн
- ✓ Уведомление об инцидентах
- ✓ Документирование процессов

✓ Lawfulness & transparency

Данные обрабатываются только на законных основаниях (согласие, договор, закон и т.д.), а пользователю понятно, что именно с его данными делают. В ML-контексте: у пользователя есть понятное описание, какие данные собираются, зачем, как обучается модель и как используются предсказания.

✓ Data minimization

Собирать и хранить только те данные, которые действительно нужны для задачи, а не «на всякий случай». Для ML-системы это означает: лишние поля (ФИО, точный адрес, паспорт и т.п.) выкидываются, если они не критичны для качества модели.

✓ Right to be forgotten

Пользователь может потребовать удалить свои персональные данные, и компания обязана это сделать. Для ML: нужно уметь удалять записи из хранилищ и, при необходимости, переобучать или обновлять модели так, чтобы «забыть» вклад этих данных.

✓ Data portability

Пользователь имеет право получить свои данные в структурированном, машиночитаемом виде и перенести их к другому поставщику. В ML-контексте: должна быть возможность выгрузить профиль, историю операций, события трекинга и т.п. в стандартном формате.

✓ Privacy by design

Приватность закладывается в систему с первого дня проектирования, а не прикручивается «сверху» потом. Для ML это означает: продумывать анонимизацию, агрегацию, шифрование, сроки хранения, доступы, логирование и работу с согласием ещё до того, как начата сборка датасета и обучение модели.

Privacy by Design: Практика

1. **Collect minimum:** Собрать только необходимые данные
2. **Aggregate:** Анонимизировать, группировать
3. **Delete:** Удалить после 90 дней
4. **Encrypt:** Шифровать в transit и at rest
5. **Log:** Логировать все доступы

Почему это нужно

Healthcare
Врачи должны знать логику

Legal
Судьи должны понимать

Finance
Клиенты должны знать причину

GDPR требует: "Right to an explanation" (статья 22)

Методы объяснения моделей

SHAP: SHapley Additive exPlanations — вклад каждого признака

LIME: Local Interpretable Model-agnostic Explanations

Decision Trees: Понятные if-then rules

Attention: Для neural networks

Counterfactuals: "Что если" сценарии

Model Card: Документирование

📝 **Model Details:** Версия, дата, тип, input/output

🎯 **Intended Use:** Где применять, где нельзя

📊 **Performance:** Accuracy по подгруппам

⚠️ **Limitations:** Что модель не может делать

✅ **Bias & Fairness:** Тесты на bias, результаты

🔍 **Explainability:** Как объяснять решения

📚 **References:** Ссылки на код, данные

ML Governance: 4-этапный процесс

Development: Разработка, тестирование, Model Card

Ethics Review: ML Ethics Committee проверяет

Staging: Тестирование в production-like окружении

Production: Canary rollout, monitoring, audits

EU AI Act: Классификация риска

⊘ **PROHIBITED: Social scoring, real-time biometric ID**

● **HIGH-RISK: Employment, criminal justice, credit, autonomous vehicles**

LIMITED-RISK: Chatbots, recommenders, deepfakes

MINIMAL-RISK: Spam filters, game AI

High-Risk AI: 6 требований

1 **Risk Assessment:** Идентифицировать потенциальные уязвимости

2 **Data Governance:** Документировать данные, качество

3 **Human Oversight:** Люди могут переопределить решения

4 **Transparency:** Model Card, тестирование, логирование

5 **Monitoring:** Непрерывная проверка performance

6 **Quality & Robustness:** Тестирование на edge cases

Case Study: Bias в найме и финансах

🌐 Amazon (2018)

❌ Bias против женщин в hiring

✅ Урок: Bias-test BEFORE deploy

RU СберБанк (2023+)

⚠️ Credit scoring: 100% ML decisions, 110M+ клиентов

✅ Решение: Fairness audits, explainability

Case Study: Privacy & данные

🌐 Clearview AI (2021)

❌ 20B фото без согласия

Штраф: $100M+

RU Российское ПЗ (152-ФЗ)

🛍️ До 4% оборота за нарушения

✅ Privacy by design критична

Best Practice #1: Ethics Committee

- **Кто:** Технологи, этики, юристы, domain experts
- **Функции:** Проверять high-risk модели ДО deployment
- **Требования:** Model Card, bias audit, privacy review
- **Timeline:** 1-2 недели на проверку
- **Решения:** Одобрить / Отклонить / Исправить

Best Practice #2: Monitoring в production

Daily: Accuracy, precision, recall

Weekly: Bias metrics по подгруппам

Monthly: Полный compliance audit

Alerts: Автоматические оповещения при аномалиях

Compliance Checklist

- ☑ Model Card documentation
- ☑ Bias audit report
- ☑ Privacy Impact Assessment
- ☑ Explainability documentation
- ☑ Audit trail & logging system
- ☑ Incident response plan
- ☑ Ethics Committee approval

Российские best practices

Кодекс этики AI в России (2023)

Подписали: СберБанк, Яндекс, МТС, VK, Gazprom Neft, РФПИ. 180+ организаций

СберБанк: ответственное AI в финансах

100% ML решения в credit scoring + fairness audits для 110M+ клиентов

Яндекс & X5 Retail: оптимизация с учётом bias

Loss functions, которые учитывают справедливость, а не только accuracy

Кто что делает: RACI Matrix

Task / Процесс	Data Scientist	ML Engineer	Committee	Product
Detect bias	R	C	A	I
Deploy to prod	C	R	C	A
Monitor drift	C	R	I	C
Final decision по модели	C	C	A	C

R=Responsible, A=Accountable, C=Consulted, I=Informed

R – делает, A – несёт итоговую ответственность, C – консультирует, I – просто в курсе.

Ключевые выводы

Ethics is Technical — Требует инженерных решений

Regulation is Real — GDPR/CCPA/AI Act have real consequences

Documentation Saves — Model Cards критичны

Test Early & Often — От дня 1, не после deployment

Accountability Matters — Чёткое распределение ответственности

Домашнее задание

📝 **Обязательное:** Создать Model Card для своей модели

🔍 **Обязательное:** Провести bias audit по гендеру

📊 **Рекомендуется:** Реализовать SHAP explainability

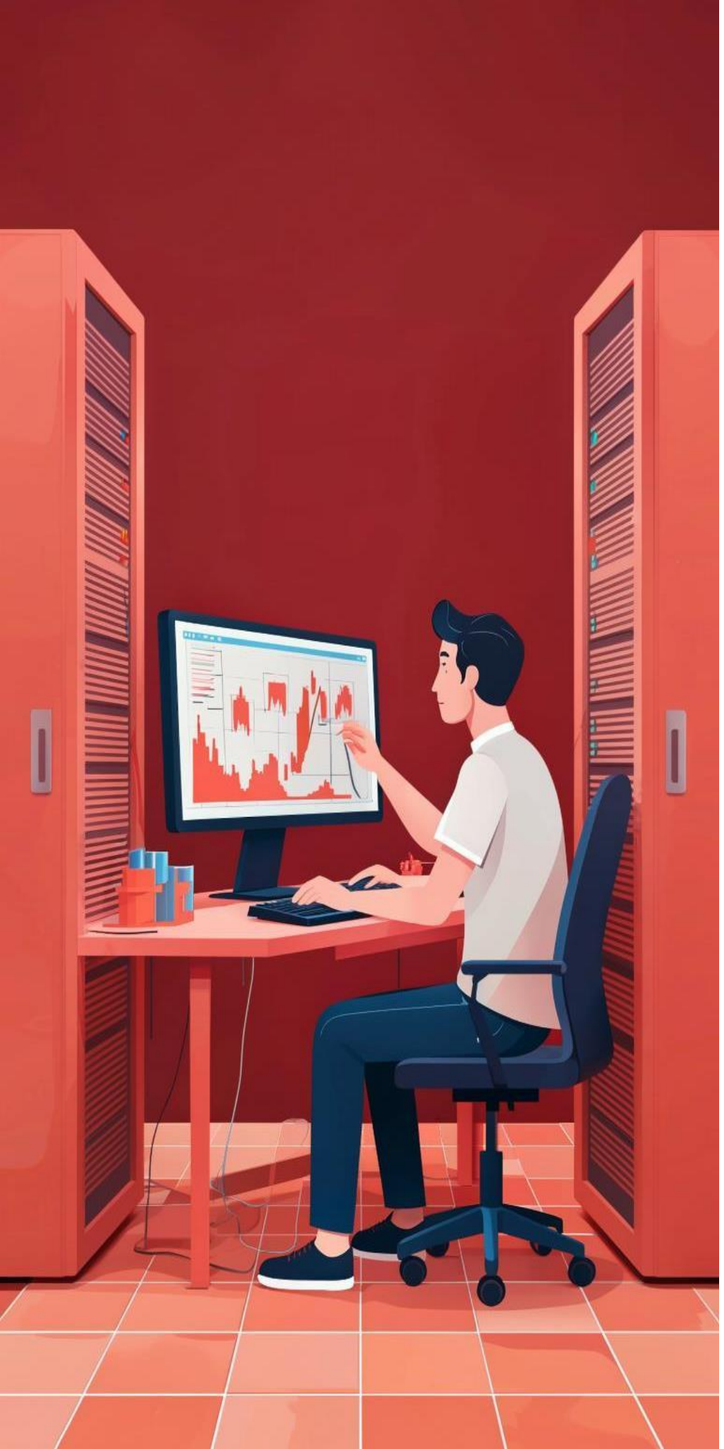
🎯 **Опционально:** Полный fairness audit с метриками

Домашнее задание

Создайте краткий Model Card (0.5-1 page) для модели из вашего проекта ЛР.

Включите:

1. **Model Details**: - Название, версия, дата создания; - Автор(ы); - Тип модели
2. **Intended Use**: - Для чего используется; - Кто пользователи; - Что НЕ использовать
3. **Performance:** - Основные метрики; - Performance по подгруппам (gender, age, etc. if applicable)
4. **Limitations**: - Когда модель может ошибаться; - Edge cases
5. **Ethical Considerations:** - Есть ли bias?; - Какие защиты реализованы?



Ключевые вопросы для самопроверки:

Вопрос 1: Назовите 5 типов bias, которые могут присутствовать в ML модели. Для каждого типа:

1. Дайте определение
2. Приведите конкретный пример
3. Объясните, как его обнаружить

Вопрос 2: Вы разрабатываете ML систему для прогноза дефолта клиентов банка. Эти данные включают:- Доход, возраст, семейный статус, кредитная история- Место жительства, род деятельности. Система будет использоваться в EU и California.

Ответьте:

1. Какие GDPR принципы нужно соблюдать?
2. Какие данные можно / нельзя использовать?
3. Какие права должны иметь пользователи?
4. Какие штрафы возможны за нарушение?

Вопрос 3: Объясните концепцию SHAP (SHapley Additive exPlanations) и почему она важна.

Ответьте:1. Что такое SHAP value для отдельного признака?
2. Как интерпретировать SHAP plot?
3. Почему SHAP лучше чем просто feature importance?
4. Приведите пример SHAP объяснения для вашей модели

Вопросы

Телеграм https://t.me/+PsC-JDrwrvsxNmVi

СКИФ (https://do.skif.donstu.ru/course/view.php?id=7508)