ML System Design Doc

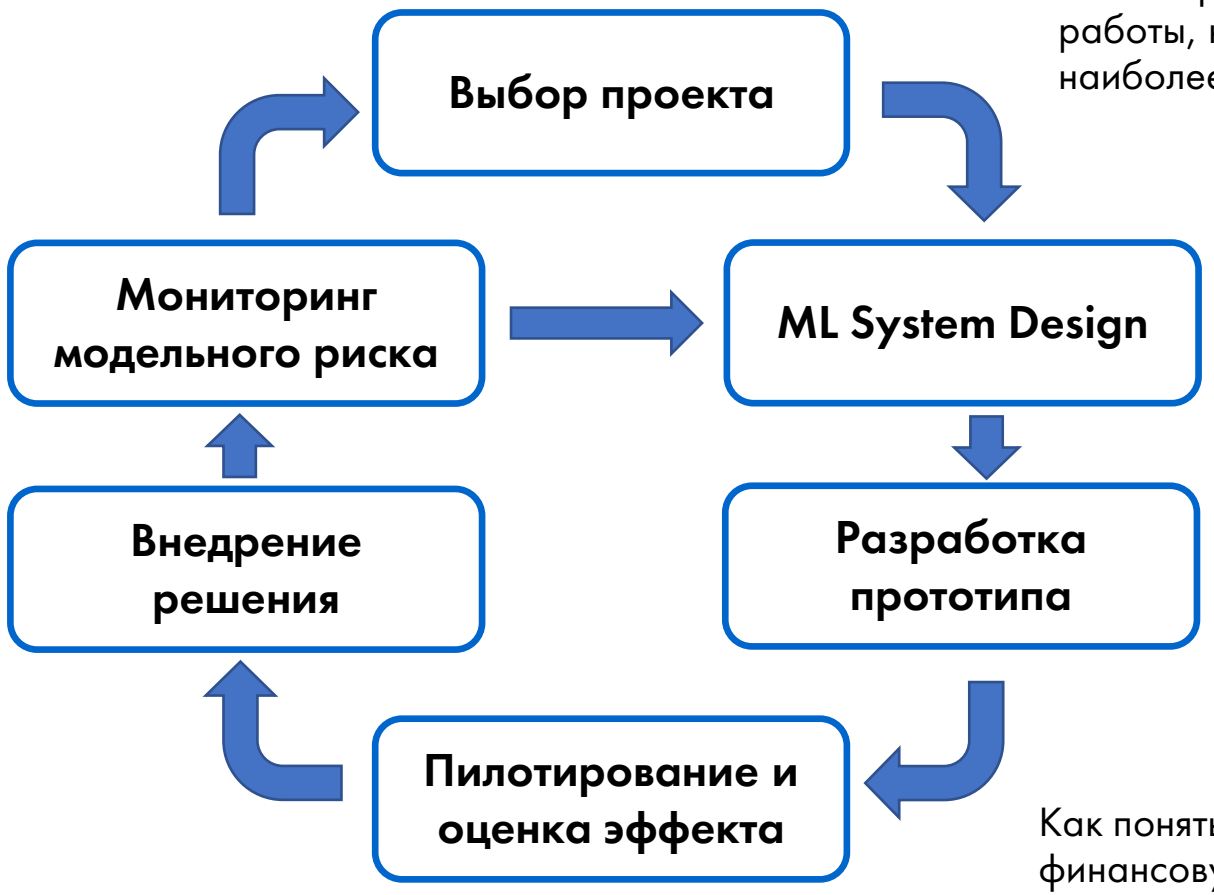
ODS ML System Design Course 2022

Лекция-бонус от Reliable ML

Ирина Голощапова

Reliable ML

Фреймворк по внедрению и развитию продвинутой аналитики



Как выбрать направление работы, которое будет наиболее полезно компании

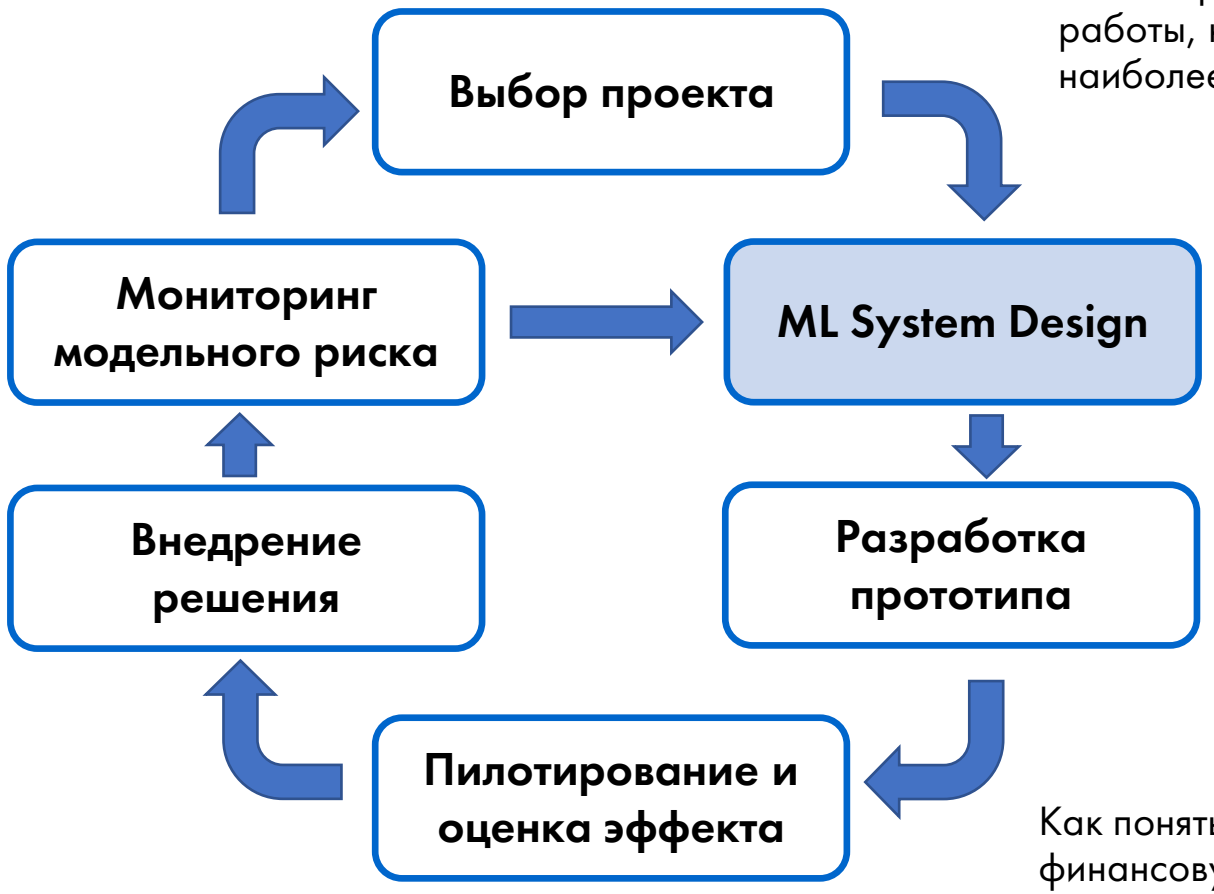
Как создать решение, чтобы:
- оно могло встроиться в бизнес-процесс
- можно было понять, хорошо или плохо оно работает

Как понять, приносит ли финансовую прибыль решение в боевых условиях

Как сделать из прототипа стабильное боевое решение и определить, когда его нужно дорабатывать

Reliable ML

Фреймворк по внедрению и развитию продвинутой аналитики: ML System Design



Как выбрать направление работы, которое будет наиболее полезно компании

Как создать решение, чтобы:
- оно могло встроиться в бизнес-процесс
- можно было понять, хорошо или плохо оно работает

Как понять, приносит ли финансовую прибыль решение в боевых условиях

Как сделать из прототипа стабильное боевое решение и определить, когда его нужно дорабатывать

А что делать то? Писать ML System Design Doc

Один документ, чтоб править всеми

- **ML System Design Document для крупных проектов помогает Data Science подразделению:**
  - ✓ Структурировать собственные мысли: БТ, архитектура решения, результат, применение
  - ✓ Задать все критические вопросы бизнесу, уточнить бизнес-требования
  - ✓ Понять бизнес-процесс и нюансы применения ML-системы
  - ✓ Понять, что реализация проекта возможна и какие ожидают трудности
  - ✓ Синхронизировать ожидания технической и бизнес-команд
  - ✓ Установить стандарты работы
- **ML System Design Doc активно набирает обороты по применению в DS процессах:**
  - ✓ Шаблон Reliable ML для ML System Design Doc – GitHub Repo
  - ✓ Международные шаблоны и материалы тут

Когда нужно писать, а когда нет

ML System Design Doc

- **Всегда – при разработке продукта:**
  - ✓ Разработка продукта разбивается на итерации, после которых проходят пилоты
  - ✓ Для каждой итерации пишем Design Doc
- **Рекомендуется – при длительном проекте (> 3 мес.):**
  - ✓ Пишем Design Doc для всего проекта, если нужно – тоже разбиваем по итерациям
- **Иногда – при кратком проекте (< 3 мес.):**
  - ✓ Пишем, если скоуп проекта сложный как технически (много моделей, много этапов вычислений), так и бизнесово (много бизнес-правил, много сценариев использования результата)
  - ✓ Пишем, если ожидания бизнеса туманны (видимость < 2 метров) и нужна синхронизация
  - ✓ Пишем, если в ДС команде нет согласья (четкого понимания), как именно реализовать вычисления
  - ✓ Когда много стейкхолдеров
  - ✓ По решению техлида команды и Product Owner Design Doc может упрощаться для малых проектов

Workflow

ML System Design Doc

- **Заполнение документа происходит итерационно:**
  - ✓ **Старт работы и заполнение вводных –** Product Owner
  - ✓ **Техника –** Data Scientist
  - ✓ **Ревью –** Data Scientist & Product Owner
  - ✓ **Встреча с бизнесом –** Data Scientist & Product Owner
  - ✓ **Финализация –** Data Scientist & Product Owner
  - ✓ **RoadMap –** Data Scientist & Product Owner
- **Результатом работы над дизайн-доком является** реалистичная и структурная **дорожная карта** работы над ML решением
- **Документация к ML решению != ML System Design Doc**, это отдельная задача, в которой дизайн-док хорошо помогает

ML System Design Doc: общие принципы составления

Что держим в голове при заполнении и проверке документа

Шаблонно = плохо

- ✓ Если один и тот же дизайн док можно применить к 2м и более проектам, то это плохой дизайн док
- ✓ Дизайн док должен показывать схему решения для конкретной задачи, поставленной в части 1

Не держите в голове - записывайте

- ✓ Детальная фиксация ключевых моментов в документе – благо
- ✓ Максимально точно формулируем ключевые параметры, влияющие на моделирование и его результаты – согласовываем с заказчиком
- ✓ Обдумываем и записываем риски ко всем этапам – что может пойти не так и насколько это критично – согласовываем с заказчиком

EDA в помощь

- ✓ Как правило, выполнить принципы выше очень помогает проведение EDA в процессе написания дизайн-дока

ML System Design Doc: Вводные

Заполняет Product Owner, по технике помогает Data Scientist

1. Цели и предпосылки

1.1. Зачем идем в разработку продукта?

1.2. Бизнес-требования и ограничения

1.3. Что входит в скоуп проекта/итерации, что не входит

1.4. Предпосылки решения

1. Цели и предпосылки

1.1. Зачем идем в разработку продукта?

- Бизнес-цель `Product Owner`
- Почему станет лучше, чем сейчас, от использования ML `Product Owner` & `Data Scientist`
- Что будем считать успехом итерации с точки зрения бизнеса `Product Owner`

1.2. Бизнес-требования и ограничения

- Краткое описание БТ и ссылки на детальные документы с бизнес-требованиями `Product Owner`
- Бизнес-ограничения `Product Owner`
- Что мы ожидаем от конкретной итерации `Product Owner`.
- Описание бизнес-процесса пилота, насколько это возможно - как именно мы будем использовать модель в существующем бизнес-процессе? `Product Owner`
- Что считаем успешным пилотом? Критерии успеха и возможные пути развития проекта `Product Owner`

1.3. Что входит в скоуп проекта/итерации, что не входит

- На закрытие каких БТ подписываемся в данной итерации `Data Scientist`
- Что не будет закрыто `Data Scientist`
- Описание результата с точки зрения качества кода и воспроизводимости решения `Data Scientist`
- Описание планируемого технического долга (что оставляем для дальнейшей продуктивизации) `Data Scientist`

1.4. Предпосылки решения

- Описание всех общих предпосылок решения, используемых в системе – с обоснованием от запроса бизнеса: какие блоки данных используем, горизонт прогноза, гранулярность модели, и др. `Data Scientist`

ML System Design Doc: Методология (1/4)

Постановка и блок-схема: заполняет Data Scientist

2.1. Постановка задачи

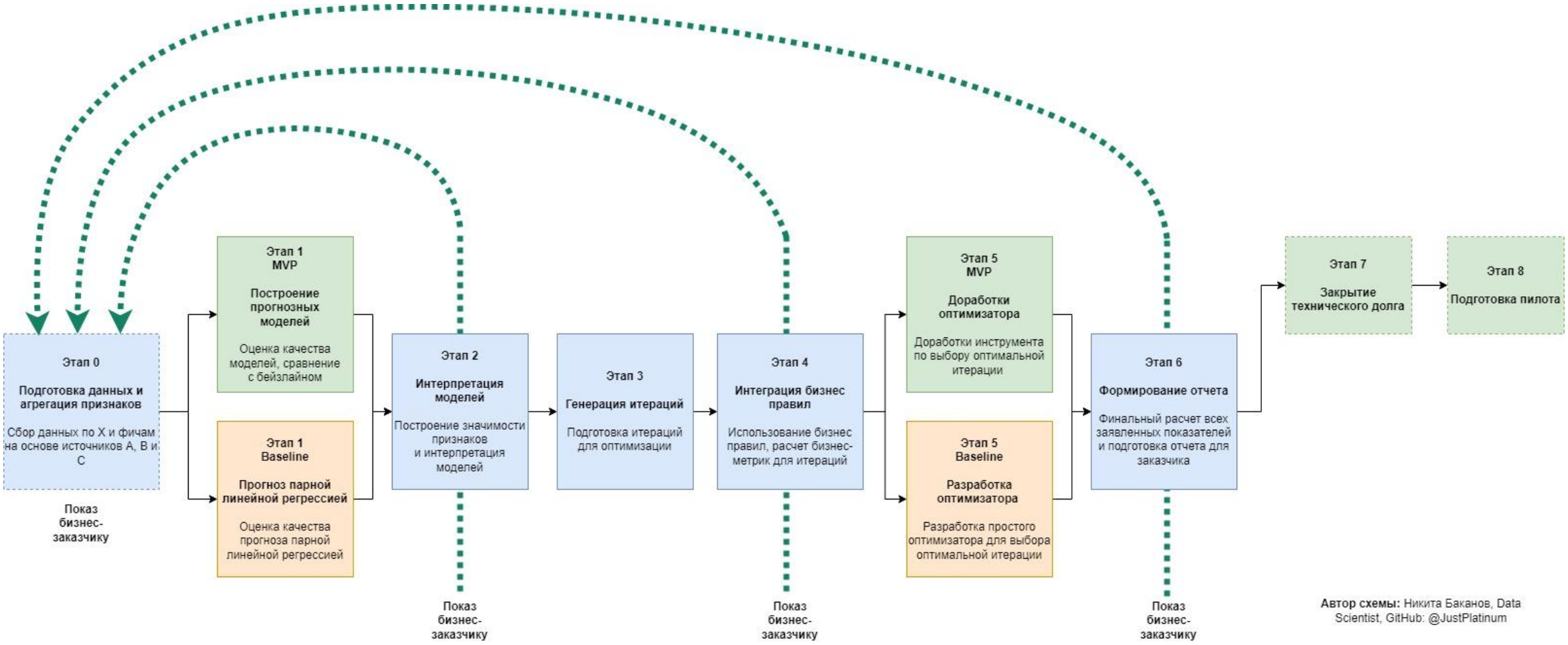
- ✓ Что делаем с технической точки зрения: рекомендательная система, поиск аномалий, прогноз, оптимизация, и др.

2.2. Блок-схема решения

- ✓ Блок-схема для бейзлайна и основного MVP с ключевыми этапами решения задачи: подготовка данных, построение прогнозных моделей, оптимизация, тестирование, закрытие технического долга, подготовка пилота, другое.

ML System Design Doc: Методология (2/4)

Пример блок-схемы



ML System Design Doc: Методология (3/4)

2.3. Этапы решения задачи: заполняет Data Scientist

Этап 1 - это обычно, подготовка данных.

В этом этапе должно быть следующее:

- ✓ Данные и сущности для обучения ML модели. Отдельная таблица для целевой переменной (либо целевых переменных разных этапов), отдельная таблица – для признаков.

Название данных	Есть ли данные в компании (если да, название источника/витрин)	Требуемый ресурс для получения данных (какие роли нужны)	Проверено ли качество данных (да, нет)
Продажи	DATAMARTS_SALES_PER_DAY	DE/DS	+
...	...	...	...

- ✓ Краткое описание результата этапа - что должно быть на выходе: витрины данных, потоки данных, др.

! Чаще всего заполнение раздела невозможно без EDA. Как минимум, команда ML/DS должна удостовериться в адекватности коммитмента на ожидаемые бизнесом метрики успешности пилота, способы расчета целевых переменных и фичей

ML System Design Doc: Методология (4/4)

2.3. Этапы решения задачи: заполняет Data Scientist

Этапы 2 и далее, помимо подготовки данных.

Описание техники для каждого этапа **отдельно для MVP** и **отдельно для бейзлайна**:

- ✓ Выборка для обучения, теста и валидации. Выбор репрезентативных данных для экспериментов, обучения и подготовки пилота
- ✓ Горизонт, гранулярность, частота необходимого пересчета прогнозных моделей
- ✓ Определение целевой переменной, согласованное с бизнесом
- ✓ Какие метрики качества используем и почему они связаны с бизнес-результатом, обозначенным Product Owner в разделах 1 и 3
- ✓ Необходимый результат этапа
- ✓ Какие могут быть риски и что планируем с этим делать
- ✓ Верхнеуровневые принципы и обоснования для: feature engineering, подбора алгоритма решения, техники кросс-валидации, интерпретации результата (если применимо).
- ✓ Предусмотрена ли бизнес-проверка результата этапа и как будет проводиться

ML System Design Doc: Подготовка пилота

Заполняют Data Scientist, Product Owner, AB Team

3.1. Способ оценки пилота

- ✓ Краткое описание предполагаемого дизайна и способа оценки пилота

3.2. Что считаем успешным пилотом

- ✓ Формализованные в пилоте метрики оценки успешности

3.3. Подготовка пилота

- ✓ Что можем позволить себе, исходя из ожидаемых затрат и времени на расчеты

ML System Design Doc: Внедрение (1/2)

Заполняет Data Scientist

4.1. Архитектура

- ✓ Блок схема и пояснения: сервисы, назначения, методы API

4.2. Описание инфраструктуры и масштабируемости

- ✓ Какая инфраструктура выбрана и почему
- ✓ Плюсы и минусы выбора
- ✓ Почему финальный выбор лучше других альтернатив

4.3. Требования к работе системы

- ✓ SLA, пропускная способность и задержка

4.4. Безопасность системы

- ✓ Потенциальная уязвимость системы

ML System Design Doc: Внедрение (2/2)

Заполняет Data Scientist

4.5. Безопасность данных

- ✓ Нет ли нарушений GDPR и других законов

4.6. Издержки

- ✓ Расчетные издержки на работу системы в месяц

4.7. Integration points

- ✓ Описание взаимодействия между сервисами (методы API и др.)

4.8. Риски

- ✓ Описание рисков и неопределенностей, которые стоит предусмотреть

Материалы для дополнительного изучения

Welcome расширять подборку!

- ✓ Шаблон ML System Design Doc [EN] от AWS и статья с объяснением каждого раздела
- ✓ Верхнеуровневый шаблон ML System Design Doc от Google и описание общих принципов его заполнения
- ✓ ML Design Template от ML Engineering Interviews
- ✓ Статья Design Documents for ML Models на Medium. Верхнеуровневые рекомендации по содержанию дизайн-документа и объяснение, зачем он вообще нужен
- ✓ Краткий Canvas для ML-проекта от Made with ML. Подходит для верхнеуровневого описания идеи, чтобы понять, имеет ли смысл идти дальше.

ML System Design Doc: Что дальше?

Шаблон Reliable ML и его развитие

Развитие шаблона – обязательно!

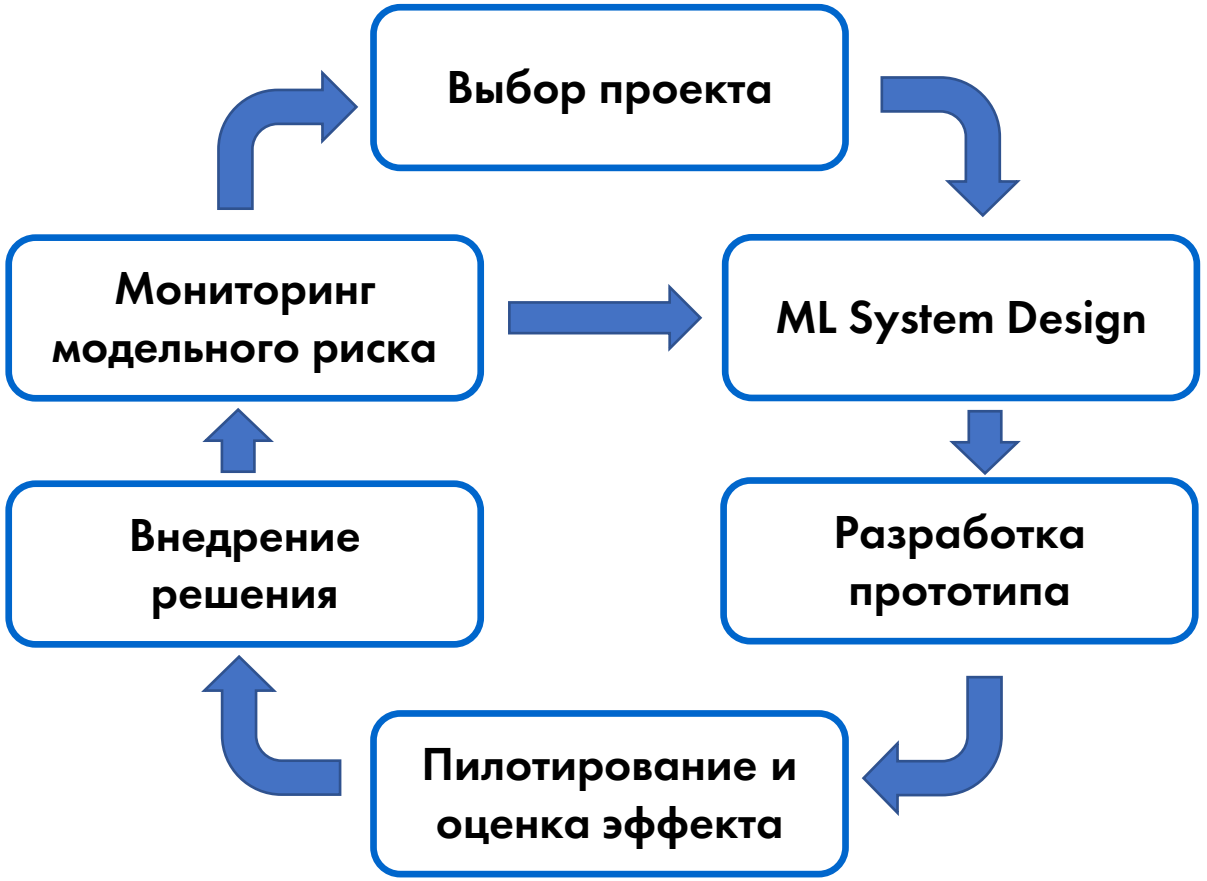
- ✓ Pull-requests с дополнениями/расширениями/комментариями – приветствуются!

Добавление примеров – мы готовы помочь!

- ✓ Если вы готовы разместить в открытом доступе заполненный шаблон ML System Design Doc для вашего проекта, пишите в тг @irina_goloshchapova
- ✓ Разберем и отшлифуем ваш пример в парадигме Reliable ML
- ✓ Разместим ссылку на финальную версию документа в вашем репозитории от вас как автора – в коллекции примеров в репозитории ML System Design Doc
- ✓ В 1 кв. 2023 г. готовы взять 5 примеров

Telegram-канал Reliable ML

Что делать, чтобы результат работы Big Data был применим в бизнес-процессах и приносил финансовую пользу



i.o.goloshchapova@gmail.com

@irina_goloshchapova