# Порядок выполнения проекта

1. Постановка задачи и выяснение общей картины

* Определите цель в бизнес-понятиях
* Как будет использоваться ваше решение?
* Как вы должны измерять производительность?

1. Получение данных

* Представьте данные в формате, который позволяет легко манипулировать данными (не изменяя сами данные).
* Выясните размер и тип данных (временной ряд, выборка, географические данные и т.д.).
* Произведите выборку испытательного набора, отложите его в сторону и никогда не смотрите в него (никакого подглядывания за данными!).

1. Исследование данных

* Создайте копию данных для исследования (при необходимости производя выборку, чтобы получить поддающийся управлению размер).
* Создайте тетрадь Jupyter для сохранения записи об исследовании данных.
* Изучите каждый атрибут и его характеристики:
* имя;
* тип (категориальный, целочисленный/с плавающей точкой, ограниченный/неограниченный, текстовый, структурированный и т.д.);
* процент отсутствующих значений;
* возможная польза для задачи;
* Для задач обучения с учителем идентифицируйте целевой атрибут (атрибуты).
* Визуализируйте данные.
* Исследуйте взаимосвязи между атрибутами
* Идентифицируйте перспективные трансформации, которые возможно захотите применить.
* Документируйте все, что вы узнали.

1. Подготовка данных

* Работайте с копиями данных (сохраните исходный набор данных незатронутым).
* Напишите функции для всех применяемых трансформаций данных по следующим пяти причинам
* Так вы сможете легко подготовить данные в следующий раз, когда получите свежие данные
* Так вы сможете применить эти трансформации в будущих проектах.
* Чтобы очистить и подготовить испытательный набор.
* Чтобы очистить и подготовить новые образцы данных после того, как решение начнет существовать.
* Чтобы облегчить трактовку подготовительных вариантов как гиперпараметров.
* Очистите данные.
* Исправьте или удалите выбросы (необязательно).
* Заполните отсутствующие значения (например, нулевым, средним, медианным или каким-то другим значением) или отбросьте их строки (либо столбцы).
* Выберите признаки (необязательно).
* Отбросьте атрибуты, которые не несут в себе никакой полезной информации для задачи.
* Масштабируйте признаки: стандартизируйте или нормализуйте их.

1. Составление окончательного списка перспективных моделей

* Обучите множество созданных на скорую руку моделей из разных категорий (например, линейную, наивную байесовскую, SVM, случайный лес, нейронную сеть и т.д.), используя стандартные параметры
* Измерьте и сравните их производительность.
* К каждой модели примените перекрестную проверку с контролем по N блокам и вычислите среднее значение и стандартное отклонение меры производительности для N блоков.
* Составьте окончательный список из первых трех-пяти самых перспективных моделей.

1. Точная настройка системы

* Проведите точную настройку гиперпараметров с применением перекрестной проверки.
* Испытайте ансамблевые методы. Комбинация лучших моделей часто будет работать более эффективно, чем индивидуальные модели
* После обретения уверенности в отношении финальной модели измерьте ее производительность на испытательном наборе, чтобы оценить ошибку обобщения