Итоговое проектное задание по блоку “Машинное обучение”

- Итоговое проектное задание – это работа НЕ на написание кода. Это задание “на подумать”. Этап продумывания структуры проекта самый первый и самый важный в цикле решения задачи анализа данных.
- Задание можно выполнять индивидуально или в небольших группах до 4 человек в группе.
- **До конца дня 27 декабря** запишите свою команду (или просто себя, если работаете индивидуально) в эту таблицу: https://docs.google.com/spreadsheets/d/1whBkwn5m1hJhXSxR2d0tLAcvzKxQNtsGz-hIs3Xf80E/edit?usp=sharing
- На занятии **29 декабря (среда)** мы будем слушать презентации, которые вы подготовите в процессе размышлений над проектом. Презентация ориентировочно должна состоять из 4-7 слайдов (при желании больше) и содержать ответы на вопросы ниже.

1) Подумать и описать некоторую проблему бизнеса, которую можно решить методами ML/AI
   - Если проблема абстрактная (решать её ещё никто не начал), то просто описать на слайде словами максимально подробно.
   - Если над проблемой уже начата работа или планируется начало работы, и уже есть данные, то можно показать кусочек данных на слайде (конечно, закрасить всю персональную информацию, которую нельзя показывать).

2) Формально описать задачу: что является объектом, что – ответом? К какому типу относится задача? (классификация/регрессия/кластеризация/поиск аномалий/что-то другое)?
   - Создать слайд с формальным описанием задачи

3) Какие есть (или могут быть) проблемы в данных, с которыми предполагается работа? Как можно решить эти проблемы, исходя из специфики задачи?
   - Могут ли быть пропущенные значения в данных? Что делать с пропусками? (удалять/заменять на какое-то значение/что-то другое)
   - Могут ли быть выбросы в данных? Из-за чего? Что с ними делать?
   - Нужно ли масштабировать данные?
   - Есть ли ещё какая-то интересная специфика в данных?

4) Какие признаки использовать для решения задачи? (опишите максимально подробно возможные признаки) Где собрать правильные ответы для обучающей выборки?
   - Расскажите, где взять данные для обучения и опишите признаки. Также будет интересно узнать про порядок объема обучающих данных (100 объектов, 1000 объектов, больше?)

5) Какую или какие модели можно использовать в этой задаче? Как думаете, какая модель в задаче предпочтительнее и почему?

6) Какую или какие метрики предпочтительнее использовать для измерения качества в этой задаче? Расскажите почему.

7) Как измерить, что модель даёт экономический эффект?