Поздравляю вас, теперь вы вооружены знаниями по предобработке данных, и пора применить их на интересном кейсе.

Представлены данные такси, по которым мы должны предсказать возьмёт ли таксист заказ или нет(driver\_response).

Принципы вы знаете, чек-лист аналогичен лабораторной работе.

Чек-лист:

1. Загрузите датасет taxi.csv.

2. Посмотрите на данные. Отобразите общую информацию по признакам (вспомните о describe и info). Напишите в markdown свои наблюдения.

3. Выявите пропуски, а также возможные причины их возникновения. Решите, что следует сделать с ними. Напишите в markdown свои наблюдения.

4. Оцените зависимости переменных между собой. Используйте корреляции. Будет хорошо, если воспользуетесь profile\_report. Напишите в markdown свои наблюдения.

5. Определите стратегию преобразования категориальных признаков (т.е. как их сделать адекватными для моделей).

6. Найдите признаки, которые можно разделить на другие, или преобразовать в другой тип данных. Удалите лишние, при необходимости.

7. Разделите выборку на обучаемую и тестовую.

8. Обучите модель. Напишите в markdown свои наблюдения по полученным результатам. Хорошие результаты дают классификаторы RandomForest и XGBoost

Если возникнут затруднения, то смотрите на материал практических занятий. Данного там должно хватить для выполнения всех пунктов. Желаю успеха!

* *offer\_gk* – unique offer identifier (INT)
* *weekday\_key* – day of week number (Sunday = 0, Monday = 1, etc.) (INT)
* *hour\_key* – hour of day representing an hour part of datetime (value from 0 to 23) (INT)
* *driver\_gk* – unique driver identifier (INT)
* *order\_gk* – unique order identifier (INT). Order may have multiple offers
* *driver\_latitude* – latitude of driver at the time of getting an offer (FLOAT)
* *driver\_longitude* – longitude of driver at the time of receiving an offer (FLOAT)
* *origin\_order\_latitude* – latitude of the order start location at the time of receiving an offer (FLOAT)
* *origin\_order\_longitude* – longitude of the order start location at the moment of receiving an offer (FLOAT)
* *distance\_km* – estimated distance from origin to destination in kilometres (FLOAT). Value -1 means that the destination is not set
* *duration\_min* – estimated duration from origin to destination in minutes (FLOAT). Value -1 means that the destination is not set
* *offer\_class\_group* – class of the order, e.g. Economy, Business, XL (VARCHAR)
* *ride\_type\_desc* – private or business order attribute (VARCHAR)
* *driver\_response* – driver choice of whether to accept the offer or not (VARCHAR)