Сборный проект

Поздравляем! Вы освоили третий модуль программы. Самое время проверить знания на практике. Выполнять работу вы будете самостоятельно.
Как закончите, отправьте работу на проверку ревьюеру. В течение суток вы получите комментарии. Их нужно учесть: доработать проект и вернуть ревьюеру обновлённый вариант.
Скорее всего, вы будете дорабатывать кейс по комментариям несколько раз. Это нормально.
Проект завершён, когда одобрены все доработки.

Описание проекта

Вы — специалист по Data Sciense в каршеринговой компании. Вам поступил заказ: нужно создать систему, которая могла бы оценить риск ДТП по выбранному маршруту движения. Под риском понимается вероятность ДТП с любым повреждением транспортного средства. Как только водитель забронировал автомобиль, сел за руль и выбрал маршрут, система должна оценить уровень риска. Если уровень риска высок, водитель увидит предупреждение и рекомендации по маршруту.
Идея создания такой системы находится в стадии предварительного обсуждения и проработки. Чёткого алгоритма работы и подобных решений на рынке ещё не существует. Текущая задача — понять, возможно ли предсказывать ДТП, опираясь на исторические данные одного из регионов.
Идея решения задачи от заказчика:

1. Создать модель предсказания ДТП (целевое значение — **at_fault (виновник)** в таблице **parties)**
    - Для модели выбрать тип виновника — только машина (**car**).
    - Выбрать случаи, когда ДТП привело к любым повреждениям транспортного средства, кроме типа SCRATCH (царапина).
    - Для моделирования ограничиться данными за 2012 год — они самые свежие.
    - Обязательное условие — учесть фактор возраста автомобиля.
2. На основе модели исследовать основные факторы ДТП.
3. Понять, помогут ли результаты моделирования и анализ важности факторов ответить на вопросы:
    - Возможно ли создать адекватную системы оценки водительского риска при выдаче авто?
    - Какие ещё факторы нужно учесть?
    - Нужно ли оборудовать автомобиль какими-либо датчиками или камерой?

Заказчик предлагает вам поработать с базой данных по происшествиям и сформировать свои идеи создания такой системы.

Инструкция по выполнению проекта

Шаг 1. Загрузите таблицы sql

Подключитесь к базе данных, используя данные:

```
db_config = {
'user': 'praktikum_student', # имя пользователя,
'pwd': 'Sdf4$2;d-d30pp', # пароль,
'host': 'rc1b-wcoijxj3yxfsf3fs.mdb.yandexcloud.net',
'port': 6432, # порт подключения,
'db': 'data-science-vehicle-db' # название базы данных,
}
```

Шаг 2. Проведите первичное исследование таблиц

- Все ли таблицы имеют набор данных;
- Соответствует ли количество таблиц условию задачи;
- Имеется ли общий ключ для связи таблиц.

Для осмотра таблиц используйте sql-запрос.

Шаг 3. Проведите статистический анализ факторов ДТП

1. Выясните, в какие месяцы происходит наибольшее количество аварий. Проанализируйте весь период наблюдений (таблица **collisions)**.
    - Создайте sql-запрос;
    - Постройте график;
    - Сделайте вывод.
2. Скоро состоится первое совещание вашей рабочей группы. Чтобы обсуждение было конструктивным, каждый сотрудник должен понимать данные. Для этого вы должны создать подходящие аналитические задачи и поручить их решение коллегам. Примеры задач:
    - Проведите анализ серьёзности повреждений транспортного средства, исходя из состояния дороги в момент ДТП (связать **collisions** и **parties**);
    - Найдите самые частые причины ДТП (таблица **parties**).

2.1. Создайте не менее шести задач для коллег. Опирайтесь на примеры и таблицы.

2.2. Пропишите порядок решения для двух задач из списка. Обязательное условие — решение этих задач должно включать связь не менее 2-х таблиц. Пример прописанного порядка:

- Создайте sql-запрос;
- Постройте график;
- Сделайте вывод.

Шаг 4. Создайте модель для оценки водительского риска

1. Подготовьте набор данных на основе первичного предположения заказчика:
    - Выберите тип виновника — только машина (**car**). **
    - Возьмите случаи, когда ДТП привело к любым значимым повреждениям автомобиля любого из участников — все, кроме типа SCRATCH (царапина).
    - Для моделирования возьмите данные только за 2012 год.
    - Подготовка исходной таблицы должна проводиться с помощью sql-запроса.
2. Проведите первичный отбор факторов, необходимых для модели.

Изучите описание факторов. Нужно отобрать те, которые могут влиять на вероятность ДТП. Будет хорошо, если вы аргументируете свой выбор. Пример:

```
columms =['party_type',     # Тип участника происшествия. Таблица parties
          'party_sobriety', # Уровень трезвости виновника (точно может влиять)
Таблица parties
           ......
         ]
```

3. Проведите статистическое исследование отобранных факторов.
    - По результату исследовательского анализа внесите корректировки, если они нужны. Сделайте вывод.
    - Если необходимо, категоризируйте исходные данные, проведите масштабирование.
    - Подготовьте обучающую и тестовую выборки.

Шаг 5. Найдите лучшую модель

1. Смоделируйте не менее 3-х типов моделей с перебором гиперпараметров.
2. 1–2 модели из спринта 2;
3. 1–2 модели из спринта 3.
4. Выберите метрику для оценки модели, исходя из поставленной бизнесом задачи. Обоснуйте свой выбор.
5. Оформите вывод в виде сравнительной таблицы.

Шаг 6. Проверьте лучшую модель в работе

1. Проведите графический анализ «Матрица ошибок». Выведите полноту и точность на график.
2. Проанализируйте важность основных факторов, влияющих на вероятность ДТП.
3. Для одного из выявленных важных факторов проведите дополнительное исследование:
    - Покажите график зависимости фактора и целевой переменной.
    - Предложите, чем можно оборудовать автомобиль, чтобы учесть этот фактор во время посадки водителя.

Пример решения задачи 3:
Выявили, что самый важный фактор ДТП — уровень трезвости виновника **party_sobriety**. Из таблицы исходных данных известно: есть несколько уровней трезвости. Тогда решение по пунктам выглядит так:

- Для графического анализа будем использовать столбчатую диаграмму. В ней отразим зависимость числа ДТП от уровня трезвости. Проанализируем график, сделаем выводы.
- Предложить оборудовать автомобиль анализатором алкогольного опьянения. Измерение состояния при посадке сделать обязательным условием допуска за руль. А чтобы убедиться, что в трубку дышит именно водитель, добавить камеру, направленную на водительское место.

Шаг 7. Сделайте общий вывод по модели

- Кратко опишите лучшую модель.
- Сделайте вывод: насколько возможно создание адекватной системы оценки риска при выдаче авто?
- Какие факторы ещё необходимо собирать, чтобы улучшить модель?

Оформление: Выполните задание в *Jupyter Notebook*. Заполните программный код в ячейках типа *code,* текстовые пояснения — в ячейках типа *markdown*. Примените форматирование и заголовки.

Краткое описание таблиц

- **collisions** — общая информация о ДТП
  Имеет уникальный case_id. Эта таблица описывает общую информацию о ДТП. Например, где оно произошло и когда.
- **parties** — информация об участниках ДТП
  Имеет неуникальный case_id, который сопоставляется с соответствующим ДТП в таблице **collisions**. Каждая строка здесь

описывает одну из сторон, участвующих в ДТП. Если столкнулись две машины, в этой таблице должно быть две строки с совпадением case_id. Если нужен уникальный идентификатор, это case_id and party_number.

- **vehicles** — информация о пострадавших машинах
  Имеет неуникальные case_id и неуникальные party_number, которые сопоставляются с таблицей **collisions** и таблицей **parties**. Если нужен уникальный идентификатор, это case_id and party_number.