Поздравляем! Вы освоили четвёртый 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 Science 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)**
   1. Для модели выбрать тип виновника — только машина (**car**).
   2. Выбрать случаи, когда ДТП привело к любым повреждениям транспортного средства, кроме типа SCRATCH (царапина).
   3. Для моделирования ограничиться данными за 2012 год — они самые свежие.
   4. 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:
   1. Возможно ли создать адекватную системы оценки водительского риска при выдаче авто?
   2. Какие ещё факторы нужно учесть?
   3. 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)**.
   1. Создайте sql-запрос;
   2. Постройте график;
   3. 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:
   1. Проведите анализ серьёзности повреждений транспортного средства, исходя из состояния дороги в момент ДТП (связать **collisions** и **parties**);
   2. 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. Реализуйте его. 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**

Подготовьте набор данных на основе первичного предположения заказчика:

* 1. Выберите тип виновника — только машина (**car**). *\*\**
  2. Возьмите случаи, когда ДТП привело к любым значимым повреждениям автомобиля любого из участников — все, кроме типа SCRATCH (царапина).
  3. Для моделирования возьмите данные только за 2012 год.
  4. Подготовка исходной таблицы должна проводиться с помощью sql-запроса.

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*

*......*

*]*

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

**Шаг 5. Найдите лучшую модель**

1. Смоделируйте не менее 3-х типов моделей с перебором гиперпараметров.
2. Выберите метрику для оценки модели, исходя из поставленной бизнесом задачи. Обоснуйте свой выбор.
3. 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:
   1. Покажите график зависимости фактора и целевой переменной.
   2. 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.

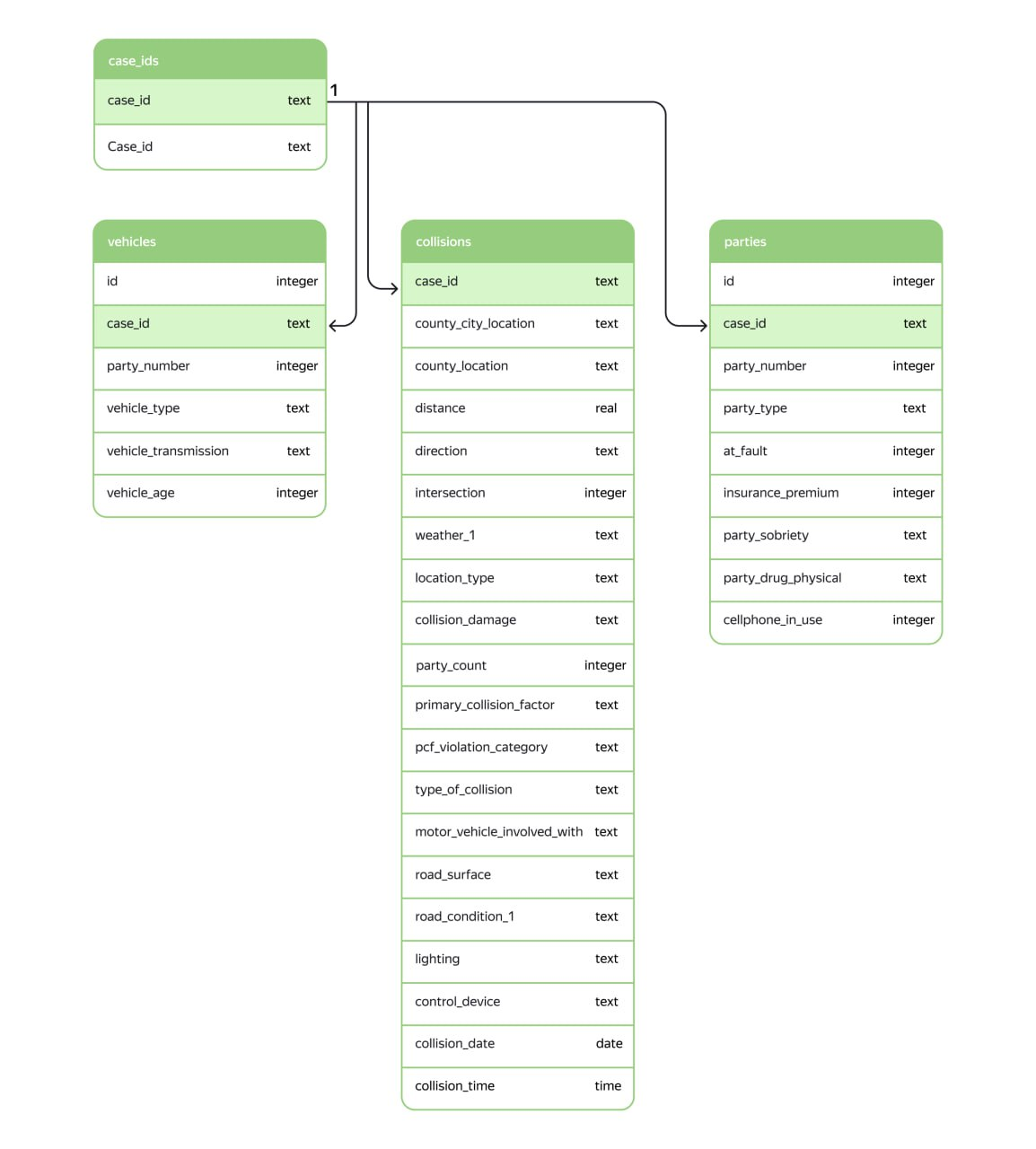
**Шаг 7. Сделайте общий вывод по модели**

* 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.



**collisions (информация о происшествиях)**

|  |  |  |
| --- | --- | --- |
| **Описание** | **Обозначение в таблице** | **Подробнее** |
| Идентификационный  Номер в базе данных | CASE\_ID | Уникальный номер для зарегистрированного происшествия в таблице происшествий. |
| Дата происшествия | COLLISION\_DATE | Формат год/месяц/день |
| Время происшествия | COLLISION\_TIME | Формат: 24-часовой |
| Является ли место происшествие перекрёстком | INTERSECTION | Y — Intersection (перекрёсток) |
|  |  | N — Not Intersection (не перекрёсток) |
|  |  | -- — Not stated (Не указано) |
| Погода | WEATHER\_1 | A — Clear (Ясно) |
|  |  | B — Cloudy (Облачно) |
|  |  | C — Raining (Дождь) |
|  |  | D — Snowing (Снегопад) |
|  |  | E — Fog (Туман) |
|  |  | F — Other (Другое) |
|  |  | G — Wind (Ветер) |
|  |  | - — Not Stated (Не указано) |
| Серьёзность происшествия | COLLISION\_DAMAGE | 1 — FATAL ТС (Не подлежит восстановлению) |
|  |  | 2 — SEVERE DAMAGE (Серьёзный ремонт, большая часть под замену/Серьёзное повреждение капитального строения) |
|  |  | 3 — MIDDLE DAMAGE (Средний ремонт, машина в целом на ходу/Строение в целом устояло) |
|  |  | 4 — SMALL DAMAGE (Отдельный элемент кузова под замену/покраску) |
|  |  | 0 – SCRATCH (Царапина) |
| Основной фактор аварии | PRIMARY\_COLL\_FACTOR | A — Code Violation (Нарушение правил ПДД) |
|  |  | B — Other Improper Driving (Другое неправильное вождение) |
|  |  | C — Other Than Driver (Кроме водителя) |
|  |  | D — Unknown (Неизвестно) |
|  |  | E — Fell Asleep (Заснул) |
|  |  | - — Not Stated (Не указано) |
| Состояние дороги | ROAD\_SURFACE | A — Dry (Сухая) |
|  |  | B — Wet (Мокрая) |
|  |  | C — Snowy or Icy (Заснеженная или обледенелая) |
|  |  | D — Slippery (Muddy, Oily, etc.) (Скользкая, грязная, маслянистая и т. д.) |
|  |  | - — Not Stated (Не указано) |
| Освещение | LIGHTING | A — Daylight (Дневной свет) |
|  |  | B — Dusk-Dawn (Сумерки-Рассвет) |
|  |  | C — Dark-Street Lights (Темно-Уличные фонари) |
|  |  | D — Dark-No Street Lights (Темно-Нет уличных фонарей) |
|  |  | E — Dark-Street Lights Not Functioning (Темно-Уличные фонари не работают) |
|  |  | - — Not Stated (Не указано) |
| Номер географических районов, где произошло ДТП | COUNTY\_CITY\_LOCATION | число |
| Названия географических районов, где произошло ДТП | COUNTY\_LOCATION | список разных названий, категориальный тип данных |
| Направление движения | DIRECTION | N — North (Север) |
|  |  | E — East (Восток) |
|  |  | S — South (Юг) |
|  |  | W — West (Запад) |
|  |  | - or blank — Not State (Не указано) |
|  |  | на перекрёстке |
| Расстояние от главной дороги (метры) | DISTANCE | число |
| Тип дороги | LOCATION\_TYPE | H — Highway (Шоссе) |
|  |  | I — Intersection (Перекрёсток) |
|  |  | R — Ramp (or Collector) (Рампа) |
|  |  | - or blank — Not State Highway (Не указано) |
| Количество участников | PARTY\_COUNT | число |
| Категория нарушения | PCF\_VIOLATION\_CATEGORY | 01 — Driving or Bicycling Under the Influence of Alcohol or Drug (Вождение или езда на велосипеде в состоянии алкогольного или наркотического опьянения) |
|  |  | 02 — Impeding Traffic (Препятствие движению транспорта) |
|  |  | 03 — Unsafe Speed (Превышение скорости) |
|  |  | 04 — Following Too Closely (Опасное сближение) |
|  |  | 05 — Wrong Side of Road (Неправильная сторона дороги) |
|  |  | 06 — Improper Passing (Неправильное движение) |
|  |  | 07 — Unsafe Lane Change (Небезопасная смена полосы движения) |
|  |  | 08 — Improper Turning (Неправильный поворот) |
|  |  | 09 — Automobile Right of Way (Автомобильное право проезда) |
|  |  | 10 — Pedestrian Right of Way (Пешеходное право проезда) |
|  |  | 11 — Pedestrian Violation (Нарушение пешеходами) |
|  |  | 12 — Traffic Signals and Signs (Дорожные сигналы и знаки) |
|  |  | 13 — Hazardous Parking (Неправильная парковка) |
|  |  | 14 — Lights (Освещение) |
|  |  | 15 — Brakes (Тормоза) |
|  |  | 16 — Other Equipment (Другое оборудование) |
|  |  | 17 — Other Hazardous Violation (Другие нарушения) |
|  |  | 18 — Other Than Driver (or Pedestrian) (Кроме водителя или пешехода) |
|  |  | 19 — Speeding (Скорость) |
|  |  | 20 — Pedestrian dui (Нарушение пешехода) |
|  |  | 21 — Unsafe Starting or Backing (Опасный старт) |
|  |  | 22 — Other Improper Driving (Другое неправильное вождение) |
|  |  | 23 — Pedestrian or “Other” Under the Influence of Alcohol or Drug (Пешеход или «Другой» в состоянии алкогольного или наркотического опьянения) |
|  |  | 24 — Fell Asleep (Заснул) |
|  |  | 00 — Unknown (Неизвестно) |
|  |  | - — Not Stated (Не указано) |
| Тип аварии | TYPE\_OF\_COLLISION | A — Head-On (Лоб в лоб) |
|  |  | B — Sideswipe (Сторона) |
|  |  | C — Rear End (Столкновение задней частью) |
|  |  | D — Broadside (Боковой удар) |
|  |  | E — Hit Object (Удар объекта) |
|  |  | F — Overturned (Опрокинутый) |
|  |  | G — Vehicle (транспортное средство/ Пешеход) |
|  |  | H — Other (Другое) |
|  |  | - — Not Stated (Не указано) |
| Дополнительные участники ДТП | MOTOR\_VEHICLE\_INVOLVED\_WITH | Other motor vehicle (Другой автомобиль) |
|  |  | Fixed object (Неподвижный объект) |
|  |  | Parked motor vehicle (Припаркованный автомобиль) |
|  |  | Pedestrian (Пешеход) |
|  |  | Bicycle (Велосипедист) |
|  |  | Non-collision (Не столкновение) |
|  |  | Other object (Другой объект) |
|  |  | Motor vehicle on other roadway (Автомобиль на другой проезжей) |
|  |  | Animal (Животное) |
|  |  | Train (Поезд) |
| Дорожное состояние | ROAD\_CONDITION\_1 | A — Holes, Deep Ruts (Ямы, глубокая колея) |
|  |  | B — Loose Material on Roadway (Сыпучий материал на проезжей части) |
|  |  | C — Obstruction on Roadway (Препятствие на проезжей части) |
|  |  | D — Construction or Repair Zone (Зона строительства или ремонта) |
|  |  | E — Reduced Roadway Width (Уменьшенная ширина проезжей части) |
|  |  | F — Flooded (Затоплено) |
|  |  | G — Other (Другое) |
|  |  | H — No Unusual Condition (Нет ничего необычного) |
|  |  | - — Not Stated (Не указано) |
| Устройство управления | CONTROL\_CONDITION\_1 | A — Functioning (Функционирует) |
|  |  | B — Not Functioning (Не функционирует) |
|  |  | C — Obscured (Затемнённый) |
|  |  | D — None (Нет) |
|  |  | - — Not Stated (Не указано) |

**Parties (описание участников происшествия)**

|  |  |  |
| --- | --- | --- |
| **Описание** | **Обозначение в таблице** | **Подробнее** |
| Идентификационный  номер в базе данных | CASE\_ID | Уникальный номер для зарегистрированного происшествия в таблице происшествий. |
| Номер участника происшествия | PARTY\_NUMBER | От 1 до N — по числу участников происшествия |
| Тип участника происшествия | PARTY\_TYPE | 1 — Car (Авто) |
|  |  | 2 — Road bumper (Дорожные знаки) |
|  |  | 3 — Building (Строения) |
|  |  | 4 — Road signs (Отбойник) |
|  |  | 5 — Other (Другое) |
|  |  | 6 — Operator (Оператор) |
|  |  | - — Not Stated (Не указано) |
| Виновность участника | AT\_FAULT | 0/1 |
| Сумма страховки (тыс. $) | INSURANCE\_PREMIUM | число |
| Состояние участника: физическое или с учётом принятых лекарств | PARTY\_DRUG\_PHYSICAL | E — Under Drug Influence (Под воздействием лекарств) |
|  |  | F — Impairment — Physical (Ухудшение состояния) |
|  |  | G — Impairment Unknown (Не известно) |
|  |  | H — Not Applicable (Не оценивался) |
|  |  | I — Sleepy/Fatigued (Сонный/Усталый) |
|  |  | - — Not Stated (Не указано) |
| Трезвость участника | PARTY\_SOBRIETY | A — Had Not Been Drinking (Не пил) |
|  |  | B — Had Been Drinking, Under Influence (Был пьян, под влиянием) |
|  |  | C — Had Been Drinking, Not Under Influence (Был пьян, не под влиянием) |
|  |  | D — Had Been Drinking, Impairment Unknown (Был пьян, ухудшение неизвестно) |
|  |  | G — Impairment Unknown (Неизвестно ухудшение) |
|  |  | H — Not Applicable (Не оценивался) |
|  |  | - — Not Stated (Не указано) |
| Наличие телефона в автомобиле (возможности разговаривать по громкой связи) | CELLPHONE\_IN\_USE | 0/1 |

**Vehicles (Описание автомобиля)**

|  |  |  |
| --- | --- | --- |
| **Описание** | **Обозначение в таблице** | **Подробнее** |
| Индекс текущей таблицы | ID | Номер в таблице |
| Идентификационный номер в базе данных | CASE\_ID | Уникальный номер для зарегистрированного происшествия в таблице происшествий. |
| Тип кузова | VEHICLE\_TYPE | MINIVAN |
|  |  | COUPE |
|  |  | SEDAN |
|  |  | HATCHBACK |
|  |  | OTHER |
| Тип КПП | VEHICLE\_TRANSMISSION | auto (Автоматическая) |
|  |  | manual (Ручная) |
|  |  | - — Not Stated (Не указано) |
| Возраст автомобиля (в годах) | VEHICLE\_AGE | число |