Формирование витрины данных по датасету “Поездки такси в Нью-Йорке”

Итоговая аттестация

Яковлев Андрей

Цели проекта

Бизнес-задача:

получить инструмент анализа данных по поездкам такси в Нью-йорке

Требования:

инструмент должен быть масштабируемый (минимальное количество строк в датасете 6 млн.);

инструмент должен быть устойчивый к “выбросам и шуму”;

на основании полученной витрины данных можно будет строить графики и дашборды;

План реализации

1. загрузка данных;
2. анализ данных;
3. очистка данных
4. выбор инструментария для построения витрины;
5. построение витрины;
6. проведение анализа зависимости чаевых от дальности поездок и количества пассажиров;

Используемые технологии

Инструмент:

PySpark

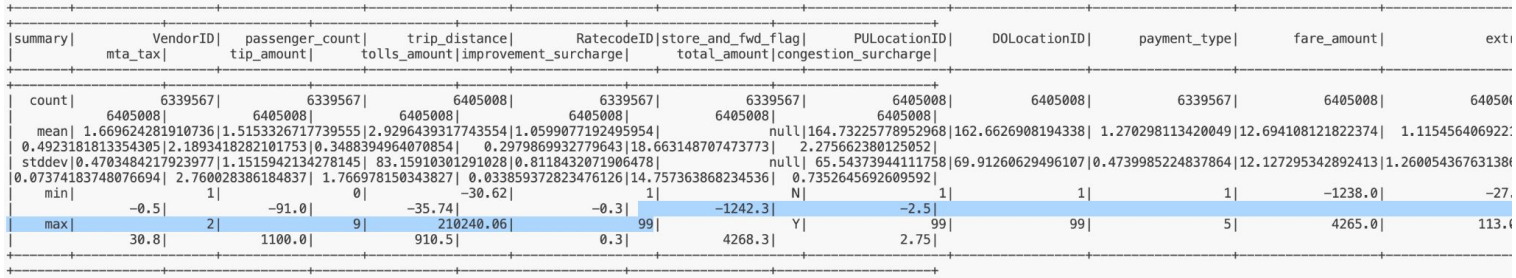
Обоснование:

- масштабируемый инструмент, позволяющий вести вычисление на кластере;
- имеет собственные библиотеки для машинного обучения, что позволит построить регрессионную модель зависимости чаевых от дальности поездок и количества пассажиров;
- позволяет использовать прочие фреймворки и библиотеки ML/AI (например, CatBoost, PyTorch и т.д.) для более глубокой аналитики и построения моделей;

Общая информация о датасете

Общее количество строк: 6 405 008

Количество строк с пустыми значениями по ключевым полям: 65 441



```
+-------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+
|summary|          VendorID|   passenger_count|     trip_distance|        RatecodeID|store_and_fwd_flag|      PULocationID|      DOLocationID|      payment_type|       fare_amount|             extr
|       |          mta_tax|        tip_amount|      tolls_amount|improvement_surcharge|      total_amount|congestion_surcharge|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+
|  count|           6339567|           6339567|           6405008|           6339567|           6339567|           6405008|           6405008|           6339567|           6405008|           640500
|       |           6405008|           6405008|           6405008|           6405008|           6405008|           6405008|
|   mean| 1.669624281910736|1.5153326717739555|2.9296439317743554|1.0599077192495954|              null|164.73225778952968|162.66269081943381| 1.270298113420049|12.694108121822374|  1.115456406922
| 0.4923181813354305|2.1893418282101753|0.3488394964070854|   0.2979869932779643|18.663148707473773|   2.275662380125052|
| stddev|0.4703484217923977|1.1515942134278145| 83.15910301291028|0.8118432071906478|              null| 65.54373944111758|69.91260629496107|0.4739985224837864|12.127295342892413|1.260054367631380
|0.07374183748076694| 2.760028386184837| 1.766978150343827| 0.033859372823476126|14.757363868234536|  0.7352645692609592|
|    min|                 1|                 0|             -30.62|                 1|                 N|                 1|                 1|                 1|           -1238.0|             -27.
|              -0.5|             -91.0|             -35.74|              -0.3|           -1242.3|              -2.5|
|    max|                 2|                 9|         210240.06|                99|                 Y|                99|                99|                 5|            4265.0|             113.0
|              30.8|            1100.0|             910.5|               0.3|            4268.3|              2.75|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+
```

```
root
 |-- VendorID: string (nullable = true)
 |-- tpep_pickup_datetime: timestamp (nullable = true)
 |-- tpep_dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- trip_distance: float (nullable = true)
 |-- RatecodeID: string (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- PULocationID: string (nullable = true)
 |-- DOLocationID: string (nullable = true)
 |-- payment_type: string (nullable = true)
 |-- fare_amount: float (nullable = true)
 |-- extra: float (nullable = true)
 |-- mta_tax: float (nullable = true)
 |-- tip_amount: float (nullable = true)
 |-- tolls_amount: float (nullable = true)
 |-- improvement_surcharge: float (nullable = true)
 |-- total_amount: float (nullable = true)
 |-- congestion_surcharge: float (nullable = true)
```

Анализ зависимости чаевых от расстояния поездки и количества пассажиров (часть 1)

Что сделано: построена регрессионная модель

результат:

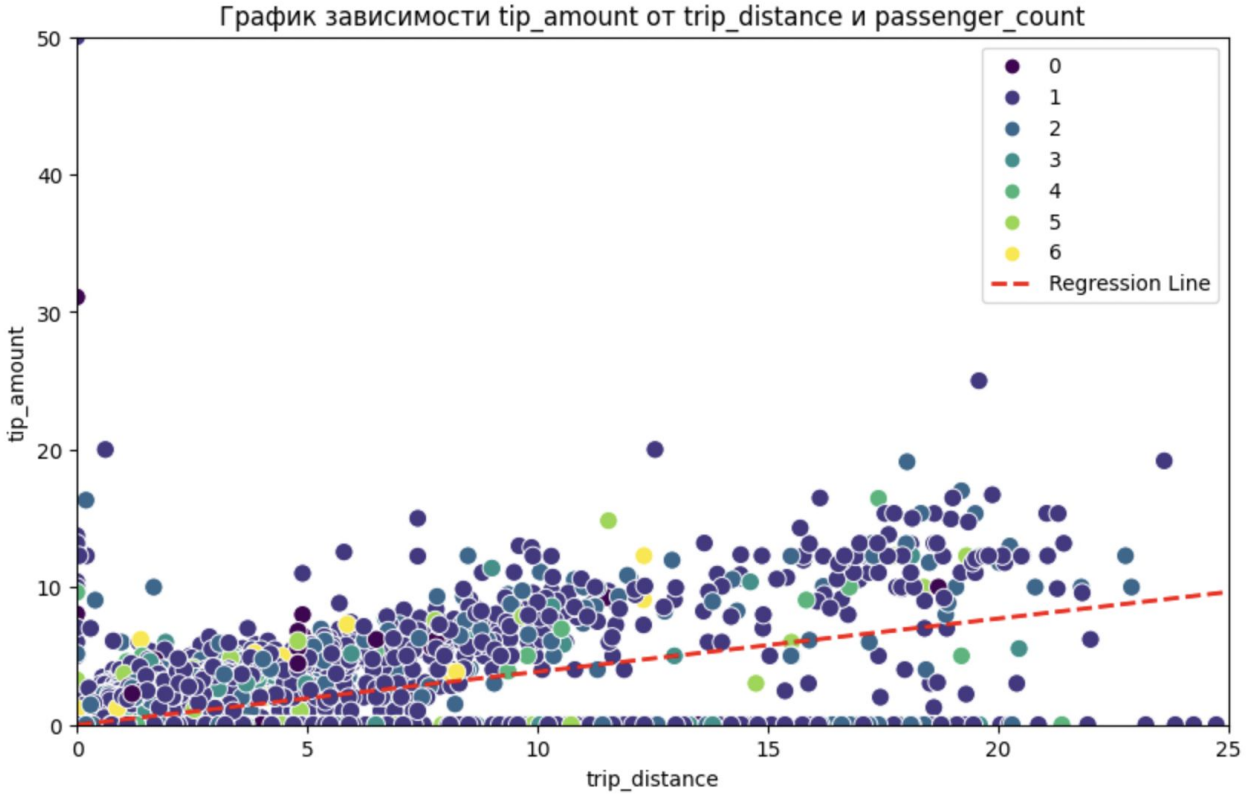
коэффициент для расстояния поездки: 0.3865701661367091

коэффициент для количества пассажиров: -0.01520730217113689

коэффициент (Intercept, т.е. значение при нулевых значениях): 1.1342261078907034

Качество модели (на тестовых данных): r^2 = 0.278741

Анализ зависимости чаевых от расстояния поездки и количества пассажиров (часть 2)



выбранные фичи не описывают полностью значение целевой переменной !! Требуется дополнительный анализ

Выводы

- в рамках проекта реализован масштабируемый инструмент анализа датасета по поездкам такси;

- проведен анализ зависимости чаевых от дальности поездок и количества пассажиров;