**Модуль А**

**Исходные файлы:**

1. Модуль А.pdf (Инструкция к модулю А)

**Результаты работы:**

1. Data.zip (Результаты модуля А)
2. Report.html + Report.ipynb (Отчет о проделанной работе)
3. Readme.txt (Дополнительные комментарии)

**ВВЕДЕНИЕ**

На этом чемпионате вам предстоит провести исследование по выявлению динамики изменения параметров пассажирских перевозок сервисами такси за 2019-2021 годы. Результаты исследования помогут в будущем оптимизировать работу предприятий, осуществляющих пассажирские перевозки, и позволит спрогнозировать распределение соответствующих ресурсов.

Данные по статистике перевозок такси доступны для загрузки на страницах <https://data.cityofnewyork.us/Transportation/2021-Yellow-Taxi-Trip-Data/m6nq-qud6>, <https://data.cityofnewyork.us/Transportation/2020-Yellow-Taxi-Trip-Data/kxp8-n2sj>, https://data.cityofnewyork.us/Transportation/2019-Yellow-Taxi-Trip-Data/2upf-qytp и представлены в формате .xlsx за 2019-2021 годы. Каждый файл имеет следующую структуру:

**Columns in this Dataset**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Type** |
| VendorID |  | Number |
| tpep\_pickup\_datetime |  | Date & Time |
| tpep\_dropoff\_datetime |  | Date & Time |
| passenger\_count |  | Number |
| trip\_distance |  | Number |
| RatecodeID |  | Number |
| store\_and\_fwd\_flag |  | Plain Text |
| PULocationID |  | Number |
| DOLocationID |  | Number |
| payment\_type |  | Number |
| fare\_amount |  | Number |
| extra |  | Number |
| mta\_tax |  | Number |
| tip\_amount |  | Number |
| tolls\_amount |  | Number |
| improvement\_surcharge |  | Number |
| total\_amount |  | Number |
| congestion\_surcharge |  | Plain Text |

В рамках всего конкурсного задания вам потребуется предобработать данные, выполнить анализ данных и выявить ключевые зависимости, построить необходимые модели машинного обучения, разработать API и пользовательское приложение, позволяющее выполнять прогнозы на несколько лет вперёд.

В этом модуле необходимо подготовить наборы данных и произвести их предобработку для дальнейшего исследования и построения моделей обучения.

**ЗАДАНИЕ**

|  |
| --- |
| 1.1 Загрузка данных |
| Создать базу данных Taxi, содержащую структуру, согласно исходным данным. Загрузить данные с ресурсов https://data.cityofnewyork.us/Transportation/2021-Yellow-Taxi-Trip-Data/m6nq-qud6, https://data.cityofnewyork.us/Transportation/2020-Yellow-Taxi-Trip-Data/kxp8-n2sj, https://data.cityofnewyork.us/Transportation/2019-Yellow-Taxi-Trip-Data/2upf-qytp. Необходимо оставить не менее 40% от исходного количества записей. Загрузить содержимое наборов в базу данных, предусмотреть обработку исключений при загрузке. В наборах данных для последующего обучения должны быть включены записи по каждому месяцу, присутствующему в исходных наборах, в пропорциональном количестве. Базу данных необходимо поддерживать в актуальном состоянии на протяжении всей работы. |

|  |
| --- |
| 1.2 Предобработка данных и выделение значимых атрибутов |
| Для каждого набора данных (по годам) предполагается применить методы кластерного анализа для создания групп сходных объектов. Такие кластеры в будущем помогут выявить динамику зависимости различных параметров для оптимизации ресурсов предприятия пассажирских перевозок.  Необходимо определить, какие атрибуты имеют наибольшее влияние на определение таких групп объектов (кластеров), и оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора. |

|  |
| --- |
| 1.3 Описание структуры набора данных |
| Для каждого атрибута подготовьте описание, содержащее текстовое представление (расшифровка, перевод, назначение). Выполнить проверку нормальности распределения значений в атрибутах. Для определения отклонений от нормы провести графическую и статистические оценки. Обосновать выбор методов оценки. На основе полученных результатов сделать вывод о типе распределения и мере скошенности. |

|  |
| --- |
| 1.4 Определение численных характеристик наборов данных |
| Определите характеристики в каждом наборе (по годам), удовлетворяющие следующим требованиям:   1. Средний по районам процент чаевых, оставленных пассажирами в поездках, длительность которых превышала 20% и не превышала 80% от максимальной длительности, от оплаты проезда; 2. Количество поездок, расстоянием не менее 1 км, при которых стоимость на километр не превышала наиболее вероятную стоимость на километр поездок по такому же тарифу для каждого способа оплаты; 3. Пять самых популярных районов, в которых высаживались люди, совершавшие поездки вместе в количестве меньше среднего количества человек на поездку, расположенных в порядке убывания средней стоимости поездки.   Сравните полученные численные значения и выявите динамику изменений по годам. |

|  |
| --- |
| 1.5 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам сессии в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip |