Вариант № 3

Задание для демонстрационного экзамена по комплекту оценочной документации

|  |  |
| --- | --- |
| **Номер компетенции** | F5 |
| **Название компетенции** | Машинное обучение и большие данные |
| **Номер КОД** | 1.1 |

# Описание задания.

# Описание модуля 1:

**Исходные файлы:**

1. train\_first\_part.xlsx (Первая часть обучающего набора данных)
2. train\_second\_part.json (Вторая часть обучающего набора данных)
3. weather.csv (Набор данных о погоде)

**Результаты работы:**

1. Data.zip (Предобработанные данные)
2. Report.html + Report.ipynb (Отчет о проделанной работе)

**ВВЕДЕНИЕ**

Вам предстоит изучить информацию об особенностях транспортных пассажирских перевозках в исследуемой локации для дальнейшего решения задачи прогнозирования времени поездки. Задача является актуальной для оператора компании. Прогнозирование времени поездки позволит оператору оптимизировать свою работу и улучшить финансовые показатели.

С точки зрения учебных целей эта работа также представляет большой интерес. В задаче предстоит работать с различными типами данных, в том числе с геораспределенными. Эти данные позволяют генерировать большое количество новых признаков. Что выделяет одну из особенностей задачи - возможность построения и исследования большого количества новых признаков.

Набор данных представлен в виде csv, xlsx и json-файлов для train и test-выборки. Целевая переменная **Trip duration** - продолжительность поездки в секундах.

В рамках Модуля 1 необходимо выполнить подготовку данных к анализу и построению моделей:

**ЗАДАНИЕ**

|  |
| --- |
| 1.1 Подготовка данных |
| Необходимо выполнить подготовку данных для дальнейшего анализа и построения прогнозных моделей. Следует выполнить загрузку всех необходимых данных по пассажирским перевозкам. Требуется выполнить объединение двух частей набора данных по перевозкам. |

|  |
| --- |
| 1.2 Формирование дополнительных атрибутов |
| Следует выполнить загрузку и преобразование данных по погоде в регионе. Дополнить исходный набор данных по поездкам информацией о погодных условиях в соответствии с временными параметрами. Полученные наборы сохранить для дальнейшего исследования.  Введите дополнительную категориальную переменную «категория по времени поездки»: «1» - очень быстрые (1 квартиль по целевой переменной), «2» - быстрые (2 квартиль по целевой переменной), «3» - средние (3 квартиль по целевой переменной), «4» - медленные (4 квартиль по целевой переменной). |

|  |
| --- |
| 1.3 Описание структуры набора данных |
| Для каждого атрибута подготовьте описание, содержащее текстовое представление (расшифровка, перевод, назначение) и статистику распределения данных (плотность, наличие пустых значений). |

|  |
| --- |
| 1.4 Предобработка данных и выделение значимых атрибутов |
| Данные привести к приемлемому формату и выполнить предобработку – удалить дубликаты строк, удалить техническую информацию, обработать выбросы.  Необходимо определить, какие атрибуты имеют наибольшее влияние на целевую переменную, и оставить только их для последующего обучения. Также необходимо обосновать выбор значимых атрибутов, а также обосновать исключение незначимых. |

|  |
| --- |
| 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. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip |

# Описание модуля 2:

**Исходные файлы:**

1. Data.zip (Предобработанные данные)

**Результаты работы:**

1. Data.zip (Результаты модуля)
2. Report.html + Report.ipynb (Отчет о проделанной работе)

**ВВЕДЕНИЕ**

Вы продолжаете работать с данными, подготовленными в предыдущем Модуле. Предобработанные данные предстоит классифицироватьи визуализировать зависимости данных. Какая-либо работа, обусловленная задачами предыдущего модуля, выполненная в ходе текущего, оцениваться не будет, поэтому проделывайте её только в случае необходимости.

**ЗАДАНИЕ**

|  |
| --- |
| 2.1 Разбиение набора данных |
| Разделите исходный набор данных на обучающую и тестирующую выборки оптимальным образом. Приведите обоснование разбиения |

|  |
| --- |
| 2.2 Визуализация зависимостей данных |
| Используя программные средства, визуализируйте зависимости признаков в наборе данных. Визуализация должна отражать влияние признаков на определение продолжительности поездки (целевую переменную). Приведите интерпретацию полученным результатам. |

|  |
| --- |
| 2.3 Классификация исходных компетенций |
| Проведите классификацию моделей по категории длительности поездки. Рассмотрите не менее трех моделей классификации. Выберите модель классификации данных по продолжительности поездки. Приведите обоснование выбора модели. |

|  |
| --- |
| 2.4 Обучение |
| Проведите обучение выбранной модели на обучающей выборке. Протестируйте работу обученной модели на тестовой выборке. Определите показатели точности работы выбранной модели, сравните с остальными рассматриваемыми моделями. |

|  |
| --- |
| 2.5 Feature Engineering |
| Путём преобразования набора данных, добейтесь более точной работы выбранной модели. Опишите приемы генерации новых данных и результаты, к которым они привели, рассматривая все ранее определенные показатели точности |

|  |
| --- |
| 2.6 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам Модуля 2, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip |