Курсовая работа

Интеллектуальный анализ статистических данных поездок такси

Задача заключается в том, чтобы считать статистические данные, предобработать их, найти в них новые признаки и подготовить для машинного обучения с целью предсказания стоимости поездки.

Шаг 1. Чтение данных

Данные представляют собой набор csv-файлов.

1.1. Считать данные и проанализировать структуру таблицы.

1.2. **Проанализировать общую информацию по каждому столбцу**. В MATLAB для этого есть функция **summary**. Изучите ее работу.

Далее примените ее к данным. Изучите полученную информацию по каждому столбцу. "Запомните" те столбцы, где есть "странные" данные.

Шаг 2. Предобработка данных

2.1. **Удалить лишние столбцы**. Подумайте, от каких столбцов можно избавиться, не испортив данные для решения главной задачи.

Обоснуйте Ваше решение.

2.2. **Оптимизировать типы данных**. В таблице есть данные **RatecodeID** и **payment_type**. Какого они типа? Преобразуйте их к этому типу.

2.3. Стандартизировать пропущенные значения в признаках **fare_amount** и **trip_distance** (функция **standardizeMissing** в MATLAB).

2.4. Удалить пропуски из таблицы

Функция **rmmissing** в MATLAB

2.5. Удалить 1% выбросов

rmoutliers(T1, 'percentiles', [0.5 99.5], 'DataVariables', ["trip_distance","fare_amount"]) в MATLAB

2.6. Проанализировать общую информацию по каждому столбцу (функция **summary** в MATLAB).

Шаг 3. Получение новых признаков

3.1. Для каждой поездки найти час, когда человек сел в машину.

3.2.Для каждой поездки найти ее длительность в минутах

3.3. Очистить данные от выбросов по длительности поездки (функция rmoutliers в MATLAB).

Шаг 4. Анализ данных

4.1. Визуализировать стоимость поездки от её длительности (функция scatterhistogram в MATLAB).

4.2. Добавить в визуализацию цветовое разделение типов поездки (по признаку RatecodeID).

4.3. Проанализировать, как время суток влияет на стоимость и количество поездок.

Определить, как время поездки влияет на стоимость. Построить гистограмму, чтобы определить, когда люди чаще совершают поездки. Добавим к гистограмме график зависимости стоимости от типа поездки.

4.4. Найти медианные значения стоимости, дальности и длительности поездки ("fare_amount","trip_distance","trip_minutes") в зависимости от ее типа ("RatecodeID"). Визуализировать полученные характеристики в пространстве.

4.5. Отфильтровать неадекватные данные по географической широте (оставить pickup_latitude > 30 & pickup_latitude < 45).

4.6. Отобразить поездки на карте, разделив разные типы поездок цветом (функция geoscatter в MATLAB).

Шаг 5. Выводы

5.1. Сделать выводы по результатам исследования данных.