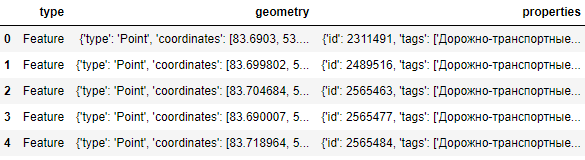
Документация

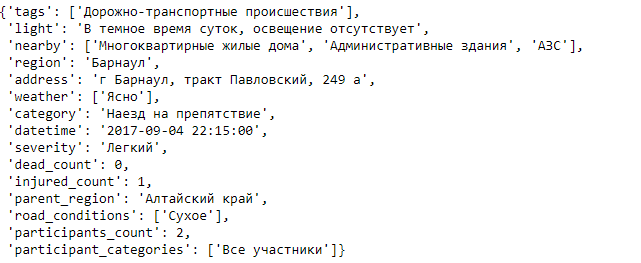
Парсинг и предобработка данных.

Исходные наборы данных в формате geojson находились на сайте <https://dtp-stat.ru/>

После первичной обработки данных и превращения их все в единый датасет:



После дальнейших обработок привел данные к единой структуре:



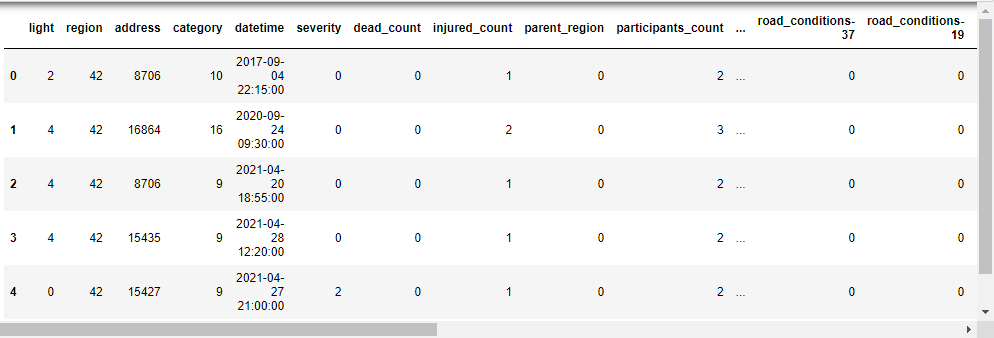


Избавился от пустых значений.

Проверил на уникальность строки датасета и дропнул повторяющиеся.

Заменил все текстовые данные на соответсвующий номер и сохранил эти соответсвия номера записал в датасет на место текста.

Убрал все перечисления сделав из каждой категории отдельный бинарный атрибут где 1 наличие такого параметра у дтп, а 0 отсутсвие.



Экспортировал получившийся датасет для дальнейшей работы с ним

Создание новой фичи.

Новую фичу “индекс” (index\_) формирую по формуле



h\_address – частота дтп по данному адрессу относительно кол-ва дтп в регионе

h\_region - – частота дтп в регионе относительно кол-ва всех дтп

most\_common\_severity\_region – самый часто встречающийся показатель severity в регионе

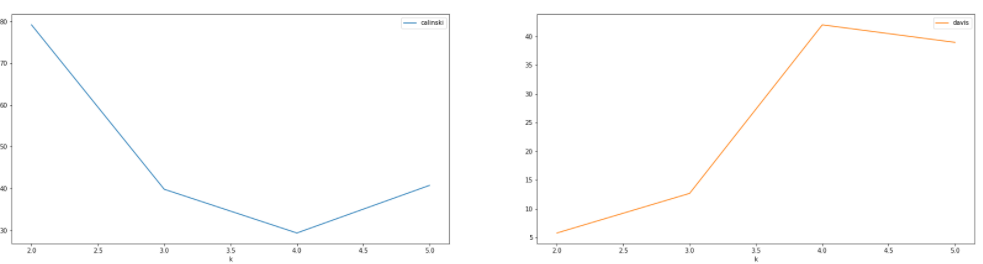
Кластеризация.

Для выбора модели я буду использовать две метрики:

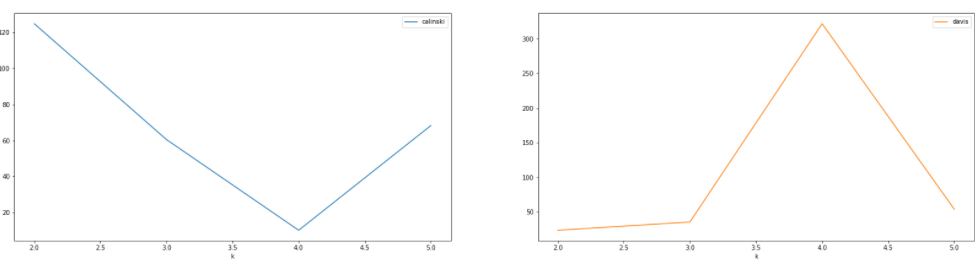
1. Индекс Калинского-Харабаза (синие графики) - чем выше показатель тем лучше

2. Индекс Дэвиса болдина (оранжевые графики) - в идеале должен стремиться к нулю (0 = идельно)

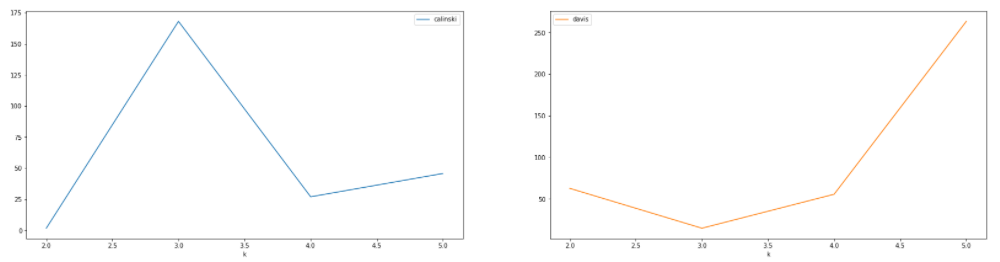
GaussianMixture



KMeans



MiniBatchKMeans



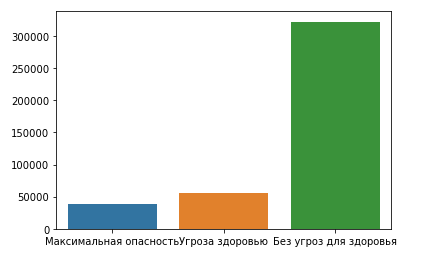
Исходя из этих двух метрик которые я проверил на одинаковом наборе данных для всех моделей (KMeans, GaussianMixture, MiniBatchKMeans) показатели для разного кол-ва кластеров.

В виду того что необходимо четкое кол-во кластеро (три) то по всем показателям метрик лучше всего себя показывает модель кластеризации MiniBatchKMeans.

Я выбираю модель MiniBatchKMeans для распределения ДТП по кластерам.

Даем названия кластерам исходя из распределения данных:

1. Без угроз для здоровья
2. Угроза здоровью
3. Максимальная опасность

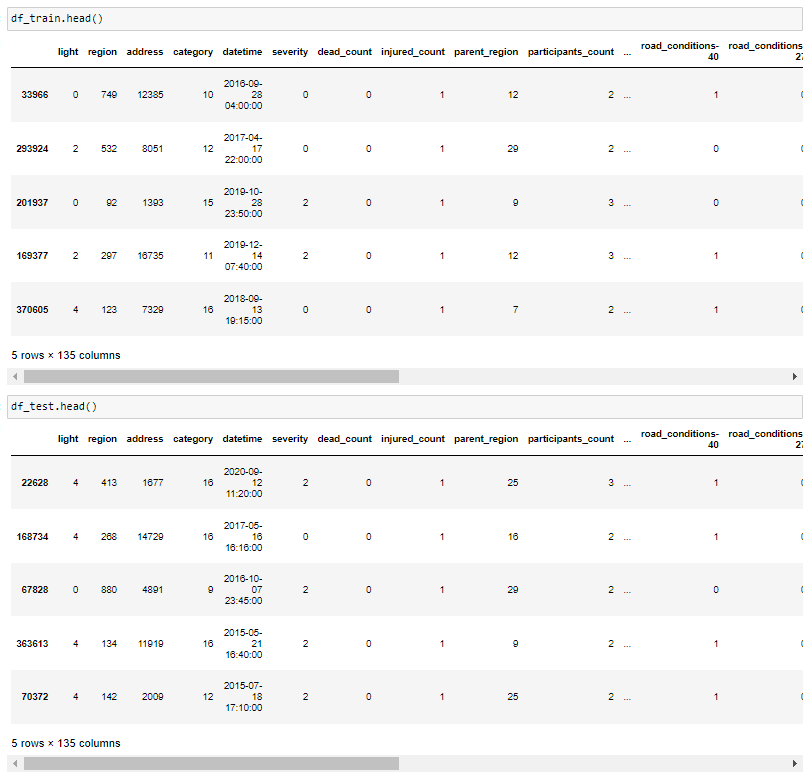


Разбиение датасета на выборки.

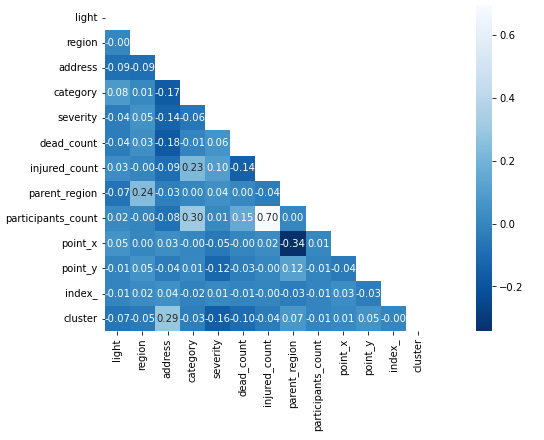
Перемешать датасет чтобы было разнообразнее.

Разделить на целевую переменную и данные.

Разбить датасет на тестовую и обучающую выборки.



Влияние атрибутов на целевую переменную.



Исходя из результатов корреляции Пирсона большее влияние на определение уровня опасности оказывает адресс (атрибут address)

Классификация.

Модели для выбора я подобрал разной сложности:

1. LogisticRegression

2. RandomForestClassifier

3. CatBoostClassifier

Чтобы выбрать модель я буду использовать метрики:

1. F1-score

2. precision

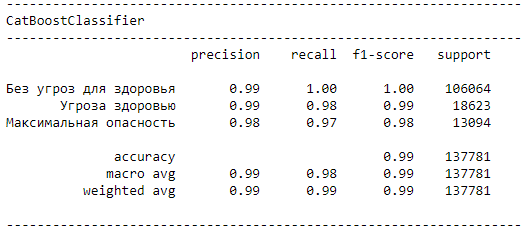
3. recall

4. accurasy

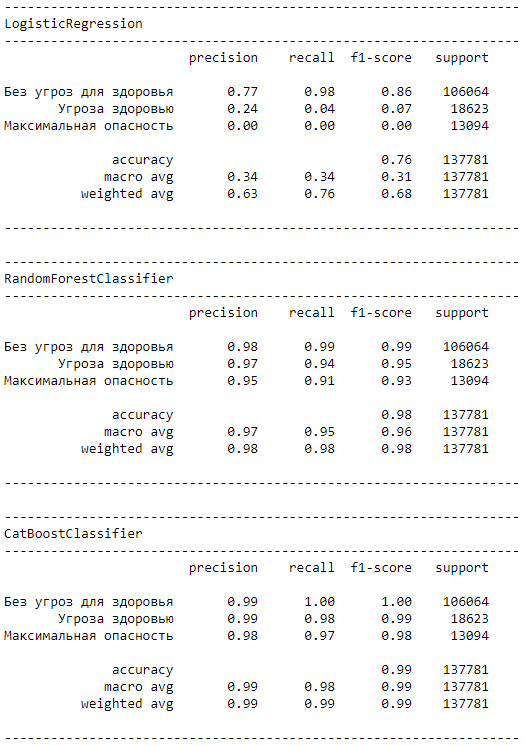
Исходя из выбранных метрик лучше всего себя показала модель CatBoostClassifier для задачи определения уровня опасности дорожной ситуации относительно других моделей.

Описание показателей CatBoostClassifier:

1. F1-score - почти сто процетное определение классов: Угроза здоровью, Максимальная опасность и сто процетное определение класса Без угроз для здоровья.
2. Для метрик precision и recall примерно одинаковое хорошее соотношение угадывания в сторону positive и negative.
3. accurasy - точность ~99%.



Все результаты тестирования:



Визуализация по географическим признакам.



Разработка программного продукта.

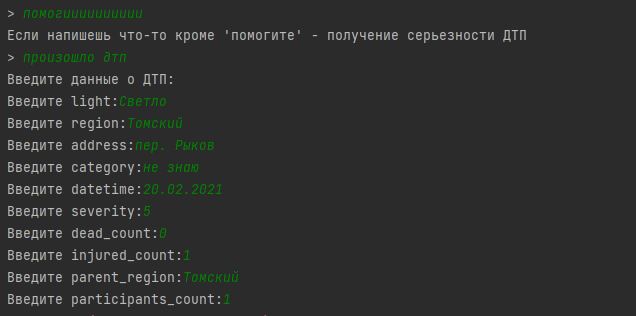
**Бот:**

Консольный бот определяющий команды из естественного языка при помощи модели обученной на ключевых словах команд.

Из функционала:

1. Может объяснить что тебе делать
2. Кинуть данные на сервер с надеждой получить ответ

\* не забудьте включить локальный сервер перед просмотром бота



**API:**

РЕСТУХААААА

Есть только один запрос на получение кластера исходя из данных о дтп которые пришлют ему

