Описание работы модели

Подготовка данных

1. Кластеризация

Изначально данные обучающей выборки кластеризуются по признакам широты и долготы с помощью алгоритмы «DBSCAN» с параметром eps = 0.012, что эквивалентно эллипсу с главными хордами примерно 1.2 и 0.8 км.

Далее плохо разделённые кластеры снова кластеризуются на более мелкие с помощью «Affinity Propagation». Каждому кластеру даётся номер в дальнейшем называемый сектором.

2. Добавление признаков

После кластеризации добавляются признаки, связанные с временем начала сбоя т. е. год, месяц, день недели, день месяца, час начала сбоя. Предполагается, что эти признаки помогут выявить сезонные закономерности. Далее генерируются признаки, которые должны показывать некоторые параметры статистики за последний месяц до предсказания по каждому сектору . В частности это среднее значение длительности сбоев, минимальное значение, максимальное значение и мода по сектору.

Обучение

Наконец, когда данные готовы, они даются на вход линейной регрессии. Далее за сериализацию как модели так и некоторых вышеописанных параметров отвечает класс «Regressor». Он также отвечает за преобразование тестового датасета к формату, пригодному для получения прогноза на обученной модели линейной регрессии.

«Regressor»

Метод «processing» отвечает за классификацию нового объекта, то есть отнесение его н какому-либо сектору и сопоставление этому объекту статистик (параметров) прошлого месяца. Метод «create\_time\_feats» отвечает за формирование признаков, связанных с временем начала сбоя. Метод «predicting» формирует прогноз по подготовленной тестовой выборке.