НИУ МАИ

Автоматический поиск аномалий в логистической сети

Мочалова Юлия Дмитриевна,
группа 8О-203М

Ревизников Дмитрий Леонидович,
д.ф.-м.н., профессор каф. 810Б МАИ

Цель и задачи

Цель: реализация автоматического поиска аномалий в объектах почтовой связи

- Анализ структуры данных
- Обзор существующих методов, алгоритмов и метрик
- Выбор технологии обработки данных с учетом требований к большому объему и скорости поступления данных
- Реализация прототипа модели определения аномалий
- Тестирование работы прототипа

В связи с увеличением количества получаемых данных, анализ и исследование становятся затратными с точки зрения человеческих ресурсов.

Термины

- РПО - Регистрируемое почтовое отправление
- ОПС - Объект почтовой связи
- Трассировка - Маршрут следования регистрируемого почтового отправления в логистической сети

Предметная область и исходные данные

- 50 тыс. объектов почтовой связи
- ~ 14 млн. действующих в момент времени регистрируемых почтовых отправлений (РПО)
- над каждым РПО совершается 3-10 операций в день

`merge_base.matreshka_merge`

РПО штрихкод	время операции	тип операции	атрибут операции	индекс ОПС	индекс следующего ОПС	тип операнда
123456789012	2020-02-20 20:20	1	2	125080	665825	1
234567890123	2020-02-21 21:21	8	2	125080	665825	2
345678901234	2020-02-22 22:22	8	6	665825	125080	2
456789012345	2020-02-23 23:23	1	2	665825	125080	1
567890123456	2020-02-24 00:24	1	1	665825	125080	2
...	...	...	...	...	...	...

Структура данных

Структура данных по ОПС

Временной ряд с интервалом в 4-12 часов, с признаками:

anomaly.train_table

индекс ОПС	время операции	количество принятых РПО	количество РПО в пути	количество обработанных РПО	количество РПО, покинувших ОПС	остатки	день недели	часть дня
125080	2020-02-20 00:00	100	200	100	665825	1000	7	1
125080	2020-02-20 04:00	300	400	200	665825	1200	7	1
125080	2020-02-20 08:00	500	600	300	125080	1300	7	1
125080	2020-02-20 12:00	700	800	400	125080	1400	7	1
125080	2020-02-20 16:00	900	1000	500	125080	1300	7	2
...	...	...	...	...	...	...	...	...

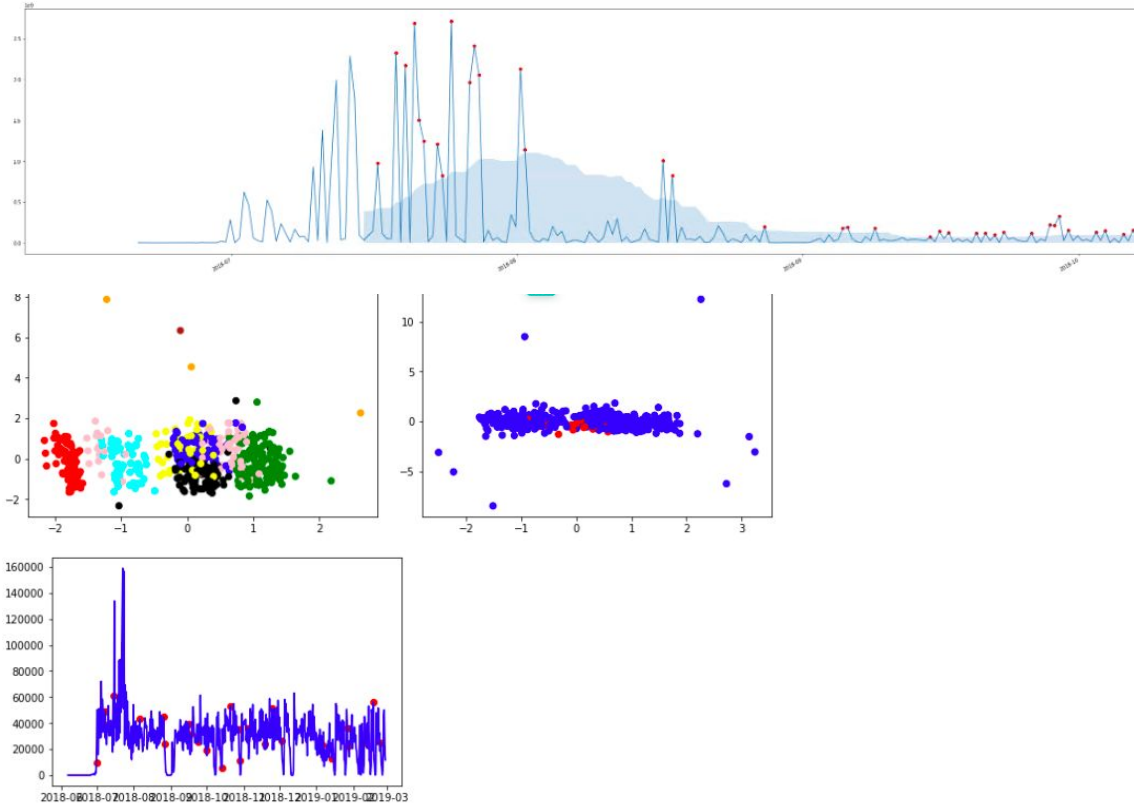
Методы, используемые в логистике

Наиболее распространенные методы базируются на следующих подходах:

- Математические подходы
- Методы исследования операций
- Кибернетические методы
- Прогностические методы

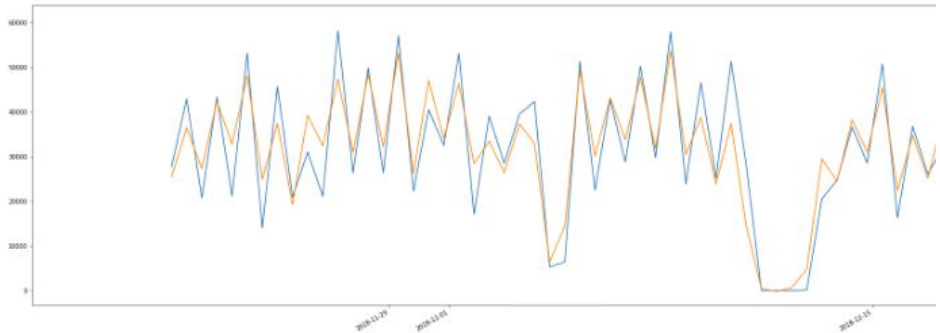
Что было опробовано?

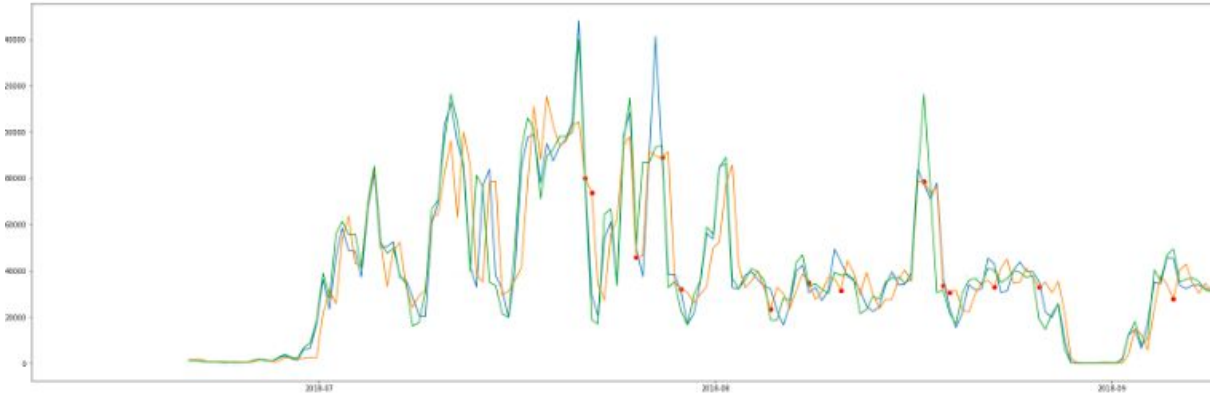
- Статистический анализ
- Кластеризация
  - k-means
  - DBScan



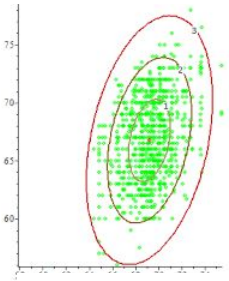
Что было опробовано?

- Регрессия
  - Линейная регрессия с регуляризацией
  - RandomForestRegressor
  - GradientBoostingRegressor
- Нейронные сети
  - RNN
  - LSTM

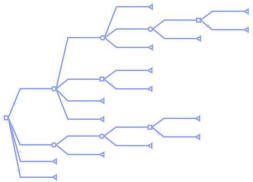




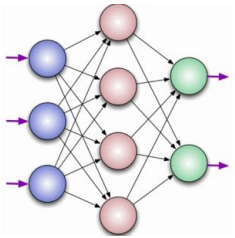
Выбранные модели и метрики



Расстояние Махаланобиса - мера между векторами случайных величин, учитывающая распределения выборки (рассчитывается от центра масс до вектора)



Градиентный бустинг. Легко интерпретируемая модель



Нейронная сеть LSTM с пятью скрытыми слоями

Расстояние Махаланобиса от многомерного вектора $x = (x_1, x_2, x_3, \ldots, x_N)^T$

до множества со средним значением

$$\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$$

и матрицей ковариации определяется следующим образом

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}.$$

Как работает?

индекс ОПС	время операции	количество принятых РПО	количество РПО в пути	количество обработанных РПО	количество РПО, покинувших ОПС	остатки	день недели	часть дня
125080	2020-02-20 00:00	100	200	100	665825	1000	7	1
125080	2020-02-20 04:00	300	400	200	665825	1200	7	1
125080	2020-02-20 08:00	500	600	300	125080	1300	7	1
125080	2020-02-20 12:00	700	800	400	125080	1400	7	1
125080	2020-02-20 16:00	900	1000	500	125080	1300	7	2
...	...	...	...	...	...	...	...	...

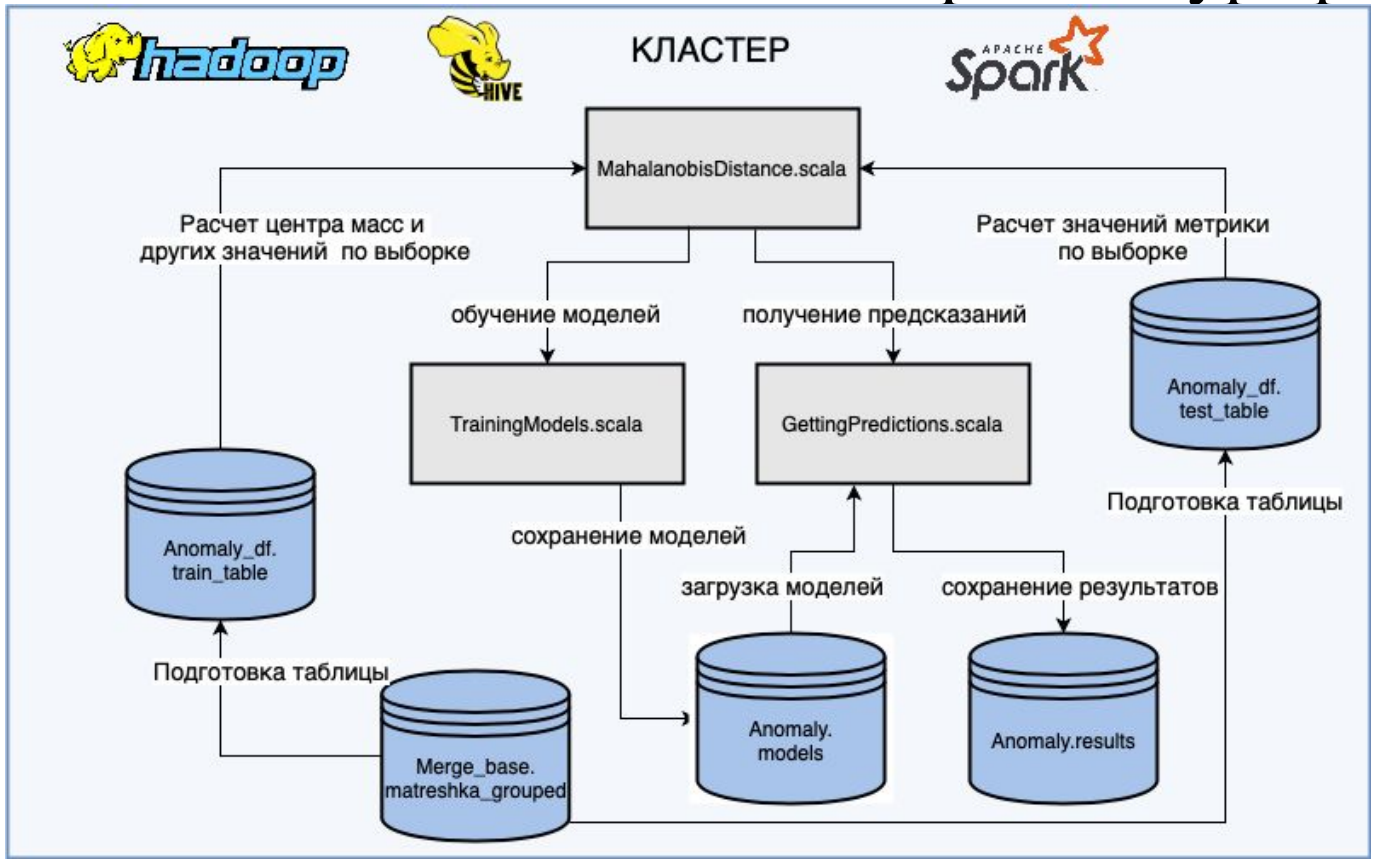
Mahalanobis Distance

МЕТКА
2
6
3
12
3
...

Библиотеки

	Плюсы	Минусы
Spark.ML / **Spark.mlib**	алгоритмы машинного обучения	отсутствие нейронных сетей
Sparkling Water H2O	поддерживает работу с нейронными сетями, размер библиотеки	новая библиотека, много недоработок и несовместимостей версий
deepLearning4Java (dl4j)	поддерживает работу с нейронными сетями	размер библиотеки

Архитектура решения

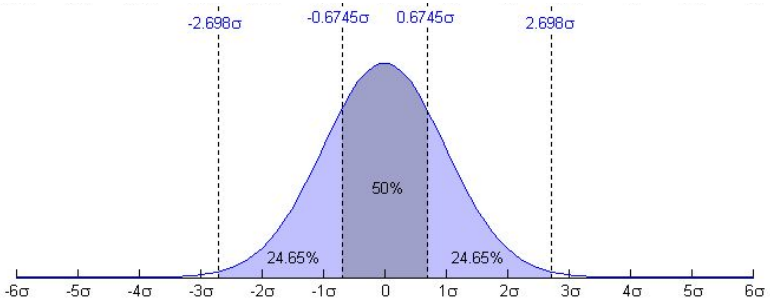


Валидация

Пространство: События (объекты) - как точки в пространстве с центром масс

Модель: Объединение результатов трех моделей

Порог - квантиль или задается экспертно



$$\rho(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$

S – ковариационная матрица.

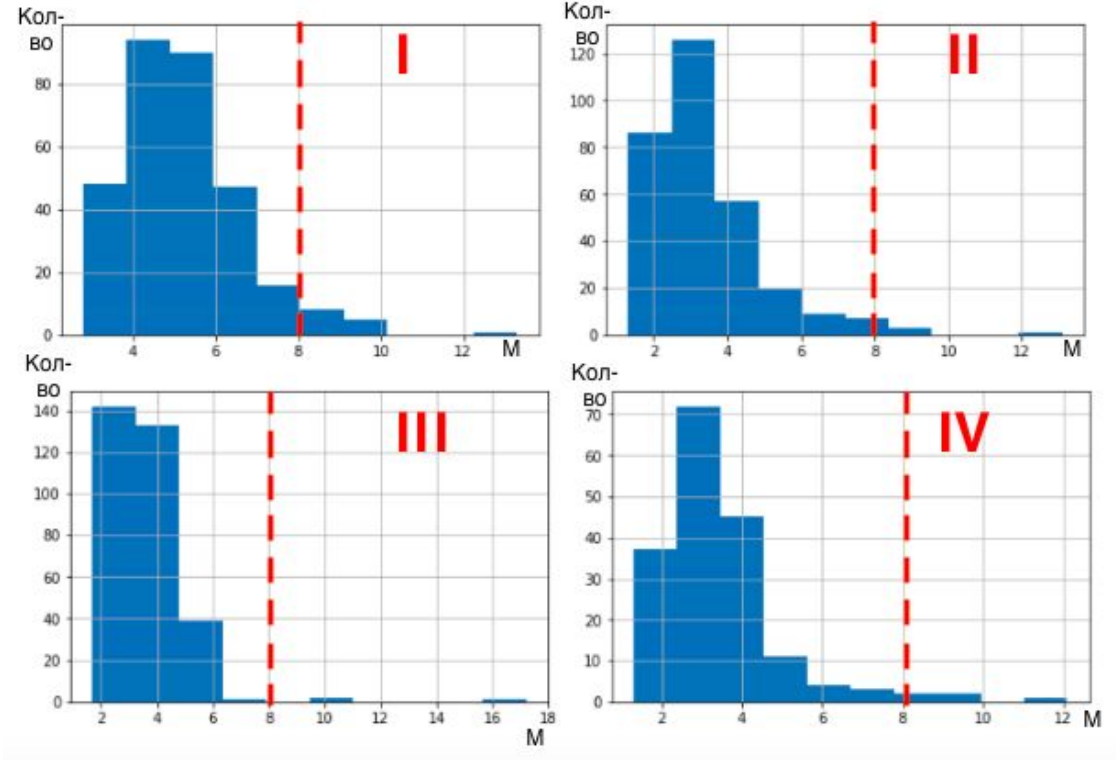
Пример

Визуализация результатов работы на четырех ОПС

1 ОПС: превышение остатков в пути, перегруз мощности ОПС, не справлялись с объемом

2, 4 ОПС: замедление сортировки, связанное с производственными причинами

3 ОПС: объемы поступлений были больше среднего, но в пределах нормы



*ОПС - объект почтовой связи

Выводы

При выполнении работы над реализацией поиска аномалий в логистической сети были выполнены следующие задачи:

- Проанализирована структура данных
- Проведен обзор существующих методов, алгоритмов и метрик
- Выбраны технологии обработки данных
- Разработаны модели определения аномалий
- Реализованы программные модули
- Протестирована работы прототипа
- Модель внедрена в производственный процесс

Таким образом в результате работы был реализован программный продукт, осуществляющий автоматический поиск аномалий для различных ОПС логистической сети

НИУ МАИ

Публикации

Публикация тезисов:

- Мочалова Ю.Д., “Автоматический поиск аномалий в динамической сети” - Международной научно-практической конференции на Digital UAV Forum, 2019 г.
- Мочалова Ю.Д., “Автоматический поиск аномалий в логистический сети” - XLV Международная молодёжная научная конференция «Гагаринские чтения – 2019»

Выступление на конференции:

- Международная научно-практическая конференция на Digital UAV Forum – 2019 г., Армения, Ереван (диплом за 3 место)