Гео-платформа и гео-эмбединги

Гео-платформа

Единая гео-сетка позволяет в обезличенной форме объединить данные по более чем 120 гео-слоям, что позволяет решать задачи прогнозирования на качественно новом уровне.

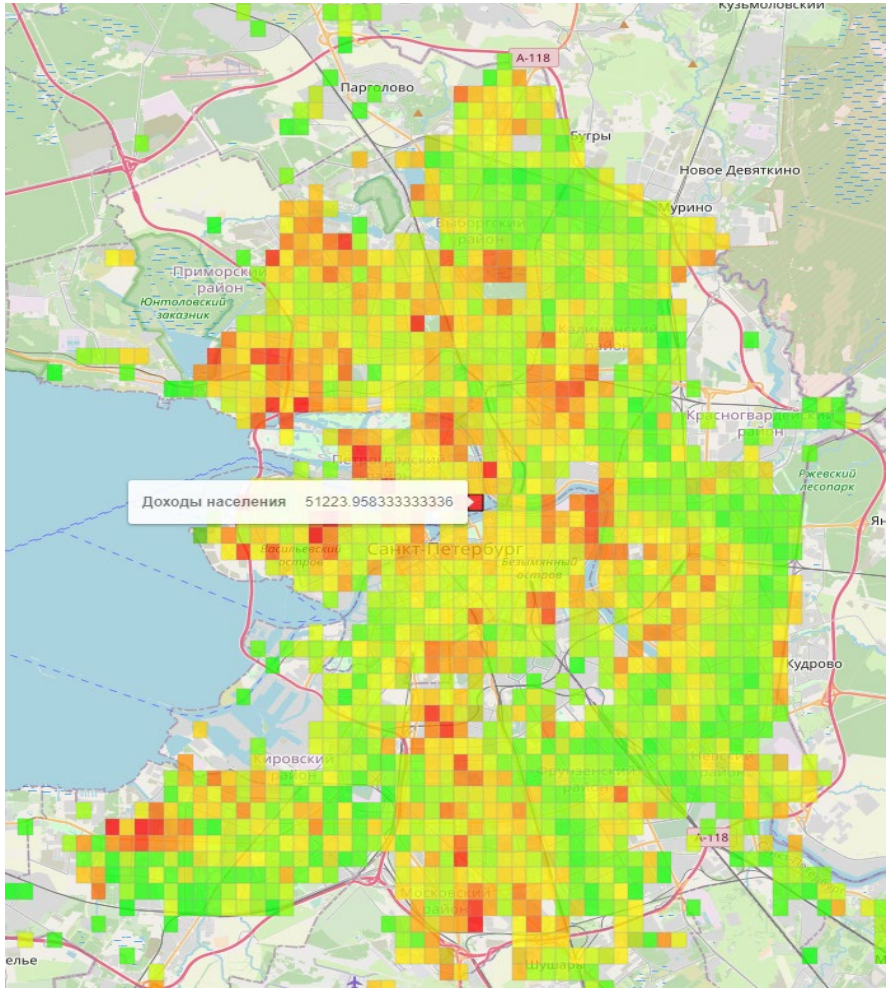
Задача гео-платформы: интегрировать максимально широкий периметр гео-данных в «стандартный» процесс анализа данных и применения моделей машинного обучения.

Сотовые операторы – параметры населения
- Численность работающих/проживающих
- Доходы, возраст, пол
- Перемещения и маршруты

Циан – параметры застройки
- Стоимость, возраст жилой недвижимости
- Арендные ставки коммерческой недвижимости

ОФД – торговая активность
- Средний чек
- Количество транзакций

Соц.сети – интересы населения
- Интерес к банковским продуктам
- Интерес к покупке авто, недвижимости



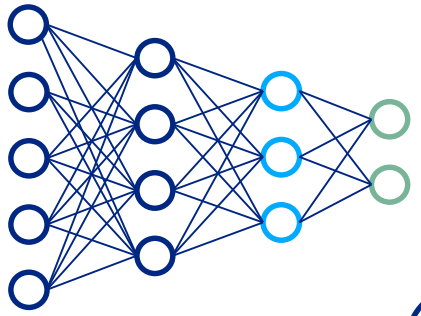
Возможные направления сотрудничества

Платформа изначально ориентирована на применение ИИ для задач гео-аналитики, что обуславливает уникально широкий периметр ее применения:

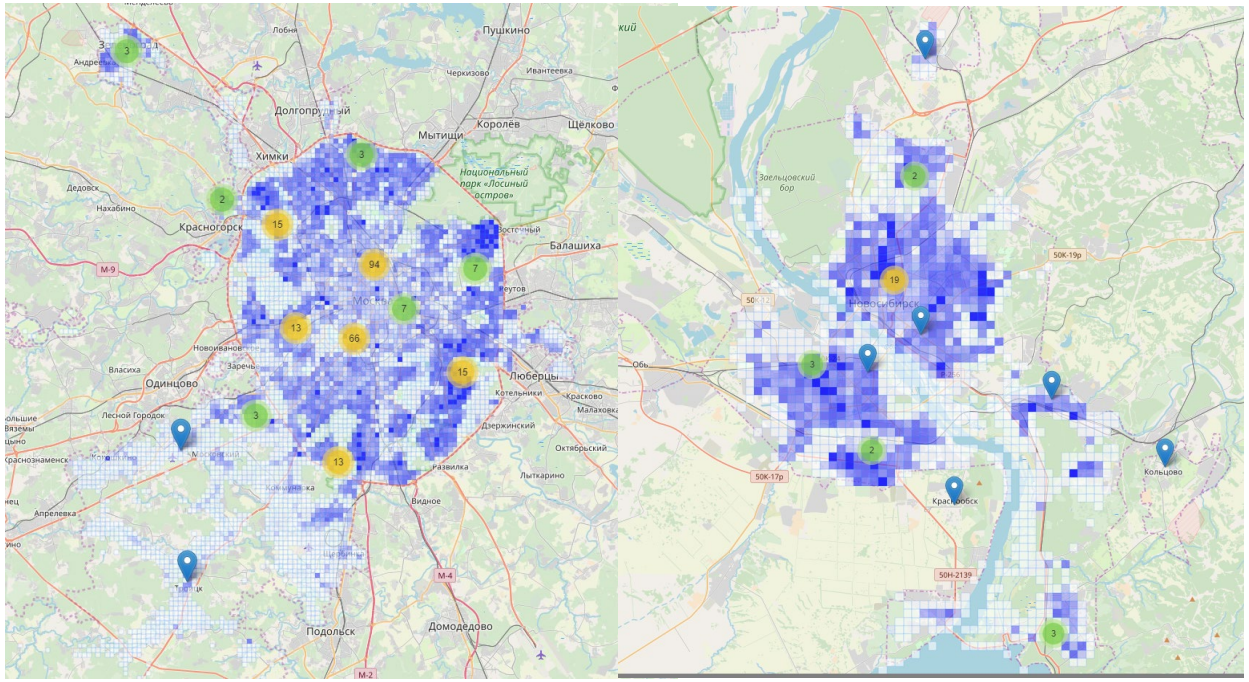
- Рекомендательные системы оптимального размещения объектов инфраструктуры и точек продаж, оценки гео-потенциала и умных KPI
- Оптимальные решения по распределению клиенто-потока по точкам продаж с учетом емкости и специализации точек продаж
- Оценки справедливой оценки стоимости объектов недвижимости, ставок аренды и инвестиционной привлекательности
- Уникальные инструменты аналитики и взаимодействия в части развития существующих моделей закачиков.

Размещение инфраструктуры и точек продаж

Уникальные периметр данных и алгоритмов машинного обучения строить прогнозы на любом уровне детализации: клиенто-поток, средний чек, спрос на заданную услугу или продукт.



- **Поиск оптимальных локаций для новых точек продаж**
- **Оценка потенциала новых точек продаж**
- **Установления умных KPI в ранее открытых точках**
- **Стратегический анализ на уровне городов / укрупненных районов**



- Успешно реализован ряд проектов в Банковской сфере
- Успешно реализован проект по оптимизации размещения наружной рекламы для компании застройщика
- Ведется разработка пилотных рекомендательных сервисов для СМБ компаний по оптимальному размещению точек продаж

4

«Полуфабрикаты» данных

Каким образом можно объединить данные для достижение синергетического эффекта?

Объединить первичные данные

- **В ряде случаев невозможно из-за законодательных ограничений.**
- **Требует компетенций по работе с ними со стороны всех участников процесса.**

Построить две независимые модели, объединить скоры.

- **Склейка на уровне результатов прогнозирования ведет к потере информации о внутренних зависимостях в данных.**
- **Отсутствует полноценный синергетический эффект от объединения данных и компетенций.**

Data Fusion подход: «полуфабрикаты» данных

- **Позволяют соблюдать законодательные ограничения и privacy клиентов**
- **Сохраняют максимальную ценность для задач моделирования.**
- **Эффективно «распределяют» компетенции команд**

Полуфабрикаты гео-данных: geo2vec

Гео-эмбединги – новый Data Fusion продукт гео-платформы

Многие компании имеют опыт прогнозирования продаж и активно развивают Data Science компетенции.

Развитие компетенций в гео-аналитике требует значимых инвестиций:

- **Закупка первичных «универсальных» гео-данных**
- **Наличие команды со специализированными навыками.**

Заказ альтернативных моделей не дает синергетического эффекта, компетенции остаются полностью на стороне подрядчика.

Эмбединги – возможность управлять трансформацией информации для формирования представлений данных с заданными свойствами

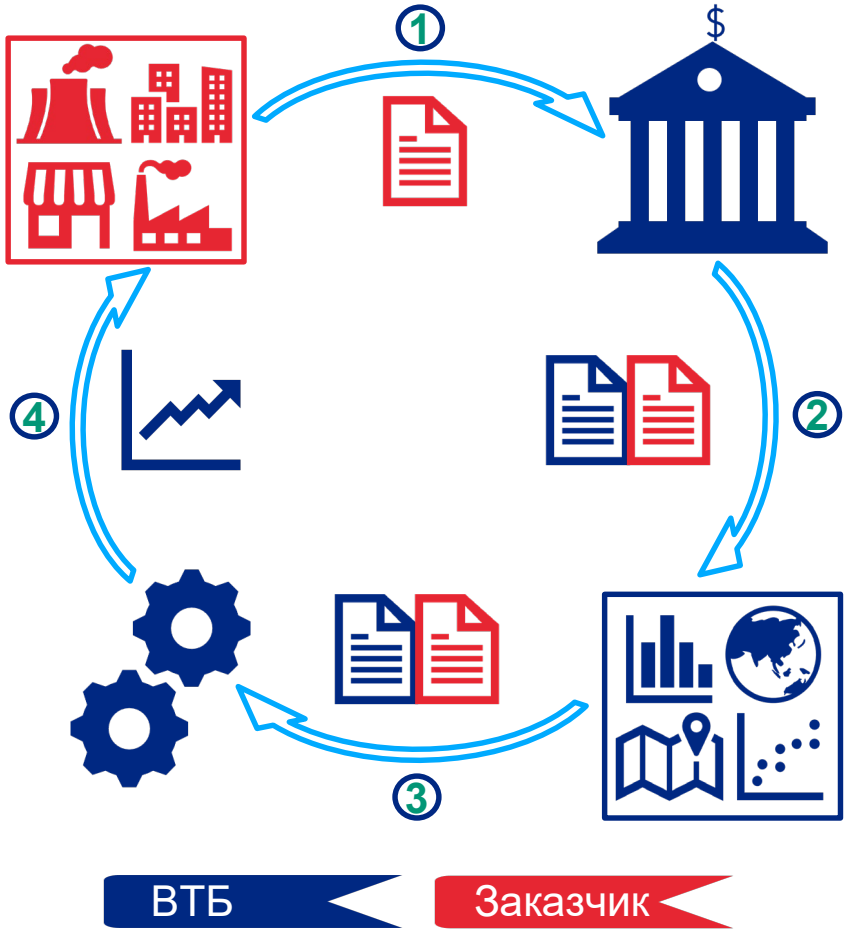
- ✓ **Содержат всю релевантную информацию из данных гео-платформы для решения задачи заказчика.**
- ✓ **Не требуют специализированных компетенций для применения в DS задачах, в том числе, обогащения ранее разработанных моделей**

Гео-аналитическая платформа

Задача платформы: **консолидация гео-данных и их интеграция в процессы анализа данных и разработки моделей машинного обучения**

1. ➢Передача заказчиком целевых переменных и гео-позиций объектов в гео-аналитическую платформу
2. ➢Обогащение переданных данных гео-признаками
3. ➢Обучение моделей на обогащённом датасете
4. ➢Передача результатов работы моделей (прогнозы целевых переменных, эмбеддинги и др.) по действующим и новым объектам обратно заказчику

Новый продукт в рамках платформы: **ориентирован на компании, которые хотят развивать направление Data Science**



Опыт решения задач гео-аналитики

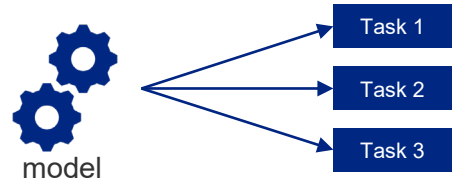
Особенности задач гео-аналитики:

- Набор размеченных данных сильно ограничен
- Имеется несколько целевых переменных
- Признаковое пространство характеризуется очень высокой размерностью
- Возникает проблема отбора признаков
- Имеется огромное количество неразмеченных данных, которые не используются
- Необходимость передачи данных внешним контрагентам без раскрытия детальной информации

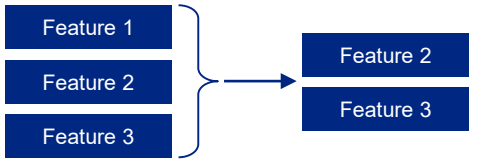
Применение гео-аналитической платформы для прогнозирования чека и трафика розничных магазинов

Методы

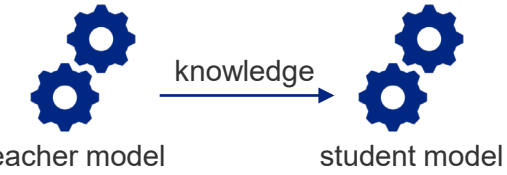
Multi-task learning



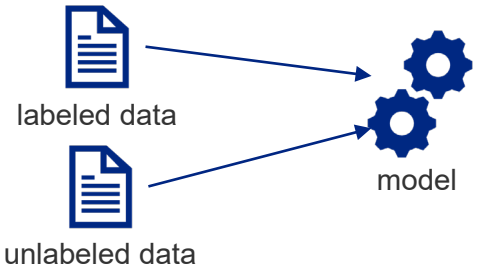
Feature Selection



Teacher-student technique



Semi-supervised learning



Neural network embeddings

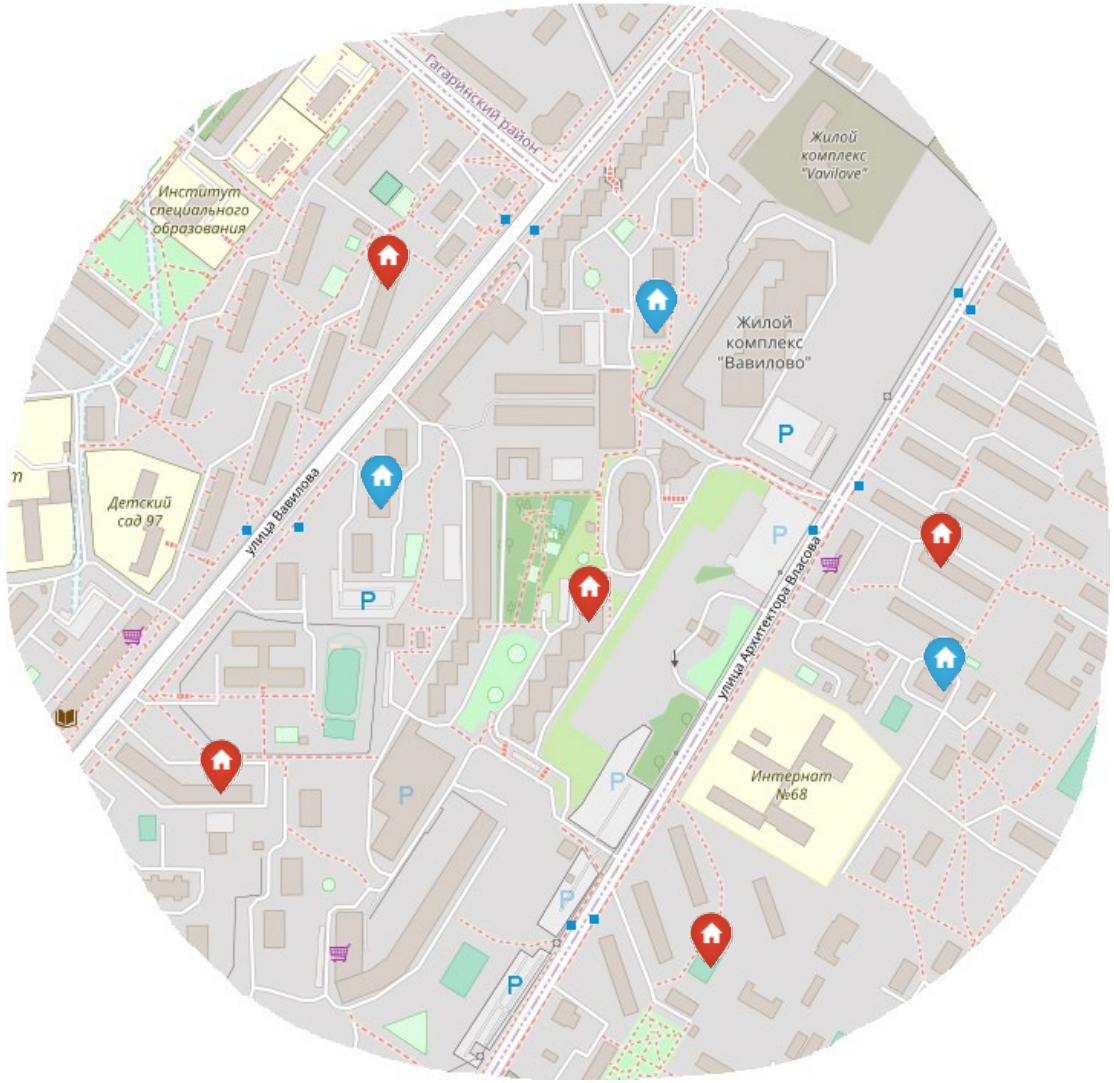
Описание бизнес-задачи

- Торговая сеть состоит множества магазинов, расположенных в различных городах
- Рассматриваются магазины определенного типа
- Известно географическое расположение действующих магазинов
- Каждый магазин характеризуется его чеком и трафиком
- Известны гео-позиции точек, где можно открыть новые магазины

Обозначения:

– действующие магазины

– точки, где можно открыть магазин



Перед тем как принять решение о выборе места для открытия нового магазина, хотелось бы оценить его характеристики: чек и трафик

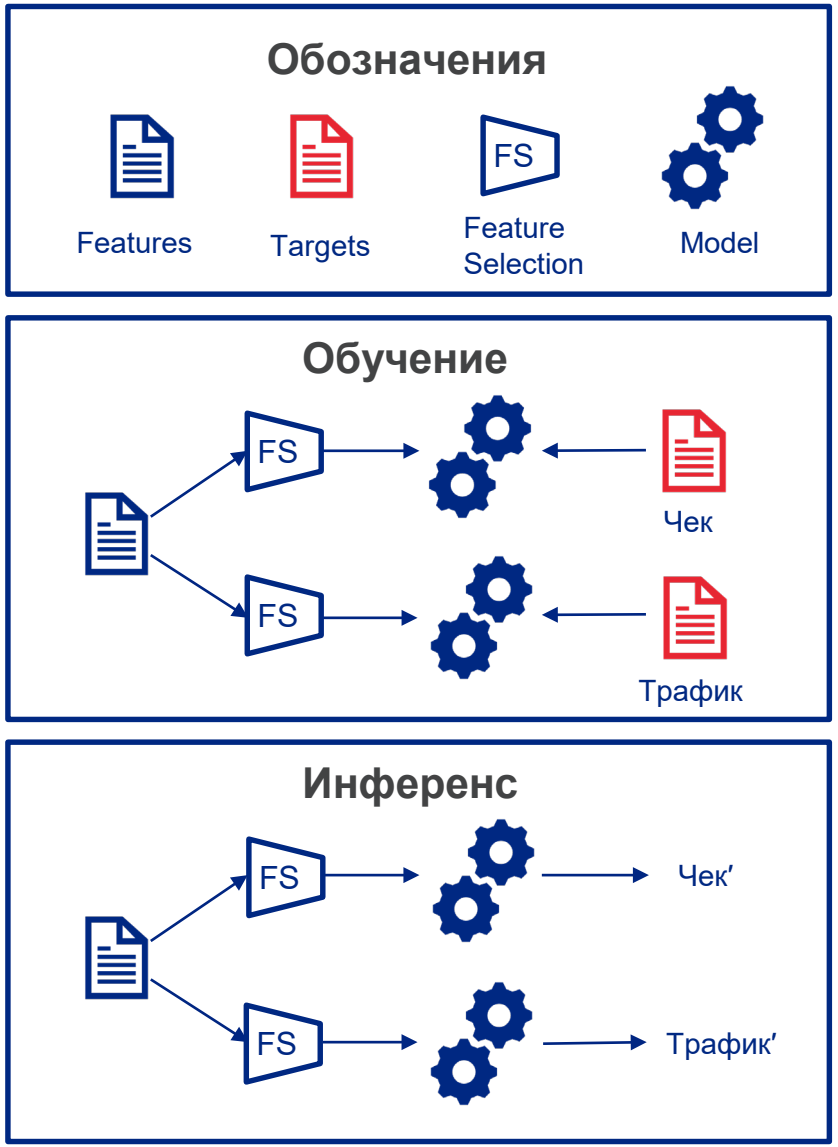
Модель: Gradient Boosting

Цель: обучить модель, которая прогнозирует чек и трафик на основе гео-признаков

Проблема: количество признаков превышает объем выборки

Решение: Feature Selection (FS) методом Feature-Wise Kernelized Lasso[1]

- Для чека и трафика сформированы два соответствующих подмножества признаков
- Метод Feature-Wise Kernelized Lasso выбрал признаки которые привели к более устойчивому поведению моделей в смысле выбранной метрики качества
- Обучены две модели lightgbm для чека и трафика соответственно



[1] https://arxiv.org/pdf/1202.0515.pdf

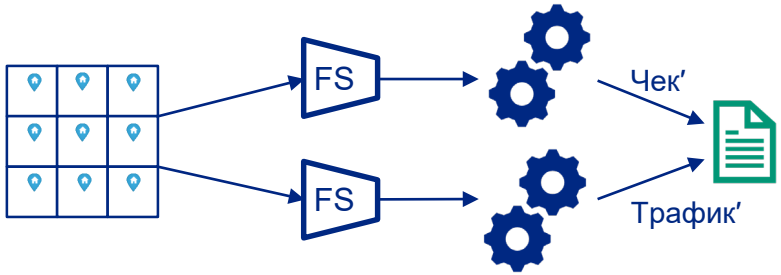
Модель: Dense Neural Network

Цель: **получение эмбеддингов гео-признаков для обогащения существующих моделей заказчика**

Проблема: **объём выборки слишком мал**

Решение: **teacher-student techniques**[2] [3] [4]

Шаг 0: рассматривая центр каждой зоны в сетке как место для открытия нового магазина, сделаем с помощью предыдущей модели прогноз чека и трафика по всем зонам

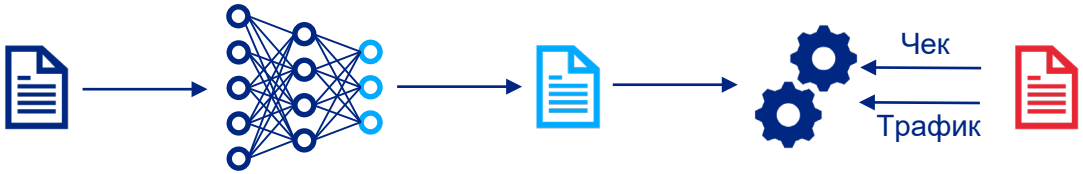


Объем обучающей выборки 28657
train / valid / test = 70 / 15 / 15

Шаг 1: предобучим на датасете с предыдущего шага Dense Neural Network, получая в результате инициализацию весов сети лучше, чем случайная

Шаг 2: сделаем fine-tuning предобученной сети на исходных данных

Шаг 3: протестируем эмбеддинги, обучив на них бустинг и сравнив результаты с изначальной моделью

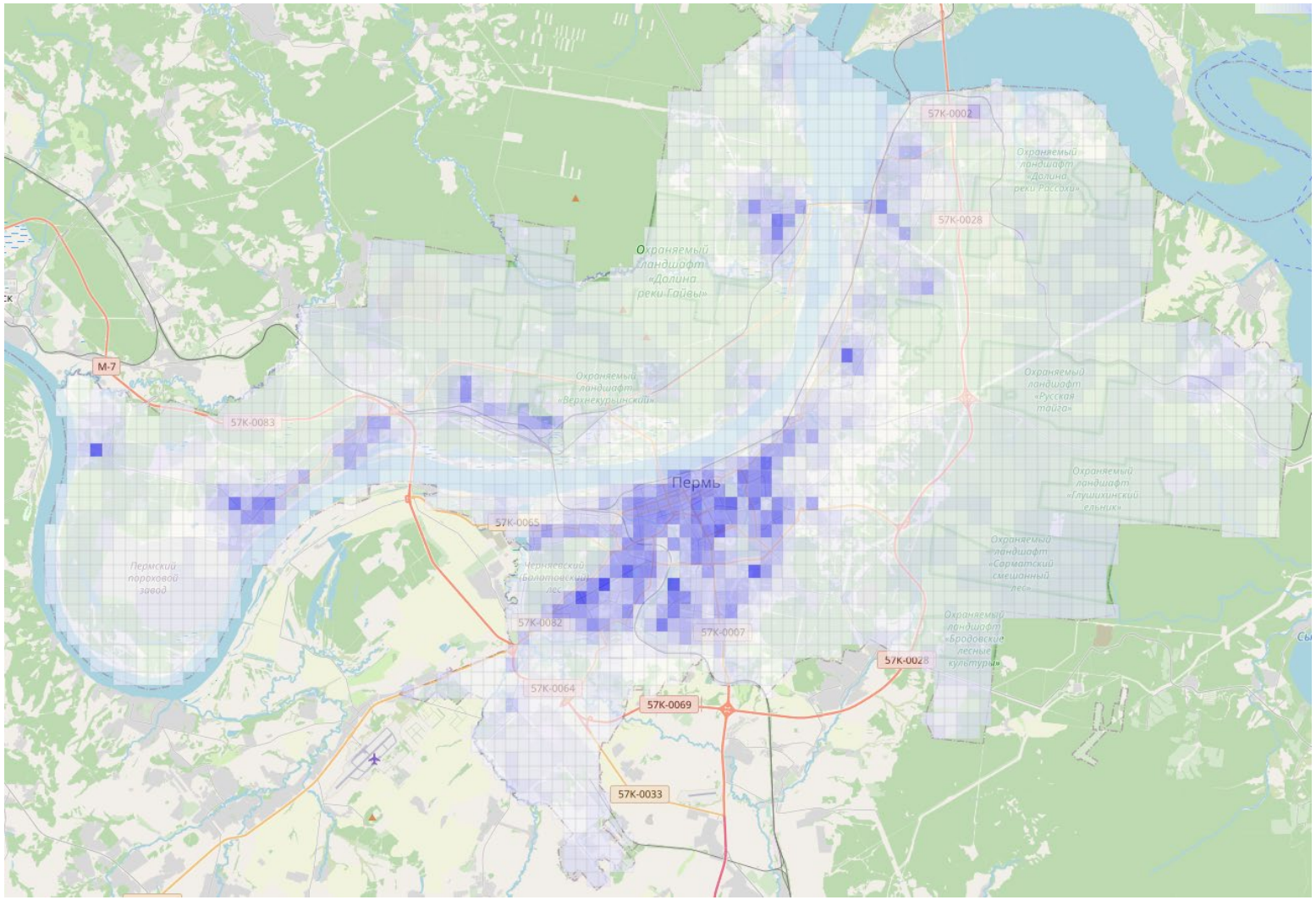


$$Loss = (y_{check} - \hat{y}_{check})^2 + 2(y_{traffic} - \hat{y}_{traffic})^2$$

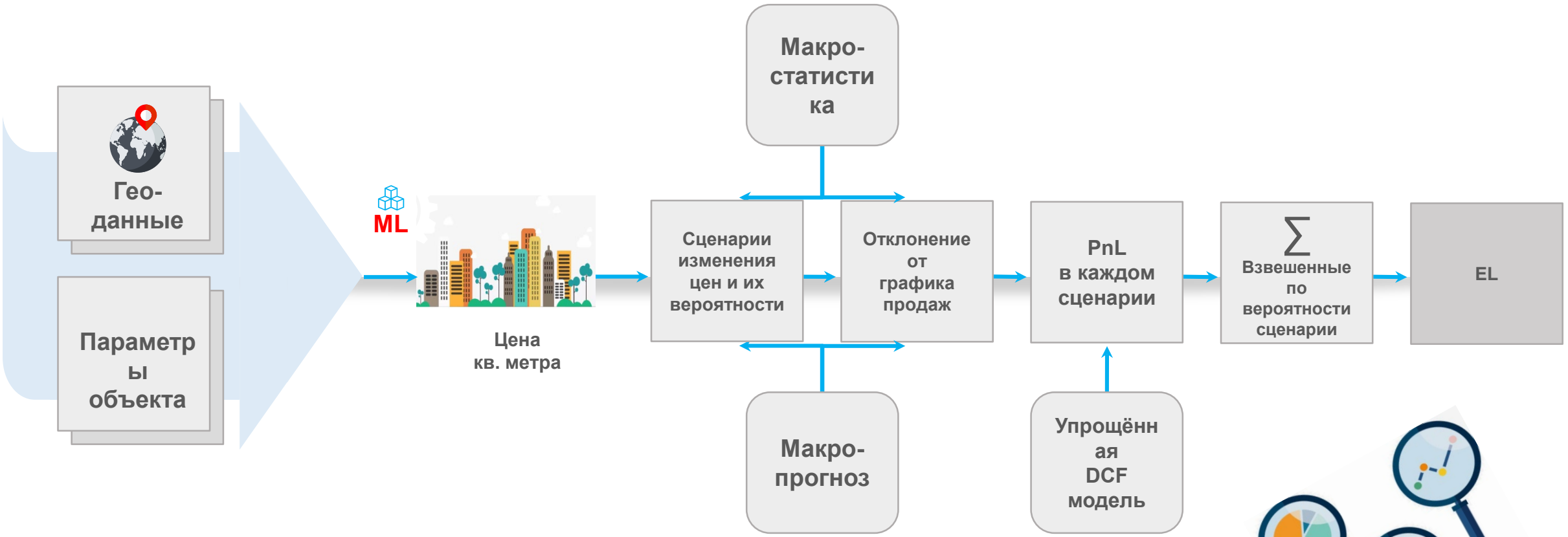
[2] https://arxiv.org/abs/2004.03281
[3] https://arxiv.org/pdf/1603.01670.pdf
[4] https://arxiv.org/pdf/1511.05641.pdf

Результаты

Возможность передачи сжатого представления гео-признаков за пределы Банка: 16 признаков вместо 1800



Возможная схема моделирования рисков в рамках 214-ФЗ



Риск гео-эмбединги

Рынок: существующие на рынке решения ориентированы исключительно на показатель оценки текущей стоимости.

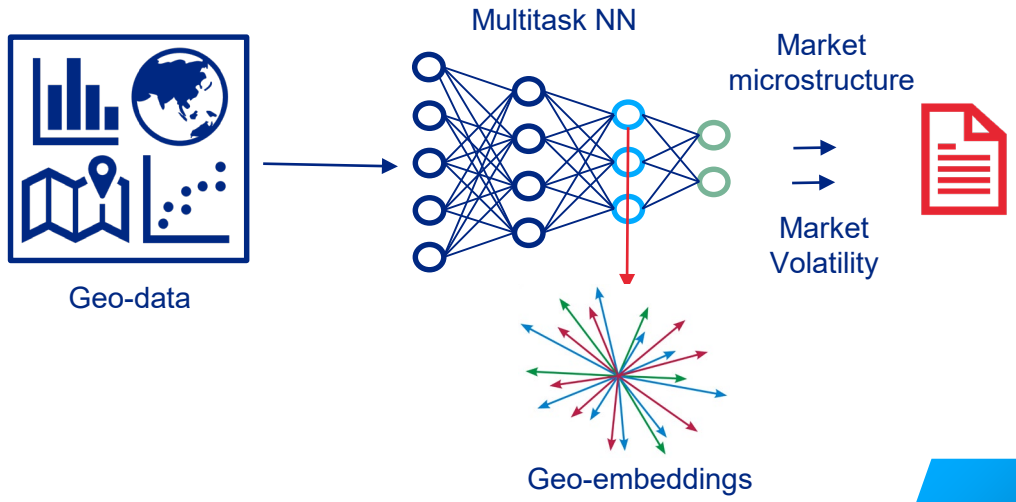
Для оценки рисков необходимо:

- Учитывать микроструктуру рынка: разброс цен на аналогичные объекты в текущий момент времени.
- Учитывать волатильность цен во времени.

Технология гео-эмбедингов: позволяет инкапсулировать информацию о любых целевых переменных в компактное векторное представление.

Для риск параметров, гео-эмбединги могут применяться:

- в моделях оценки LGD по ипотечным кредитам.
- в моделях проектного финансирования.



По задачам ИИ и продвинутой аналитики наша команда оказывает услуги внешним компаниям в рамках Платформы Больших Данных. Мы рады сотрудничеству.

Суржко Денис / dasurzhko@vtb.ru

СПАСИБО ЗА ВНИМАНИЕ!

Platforma
БОЛЬШИХ ДАННЫХ