CINIMEX
Вместе быть лучшими!

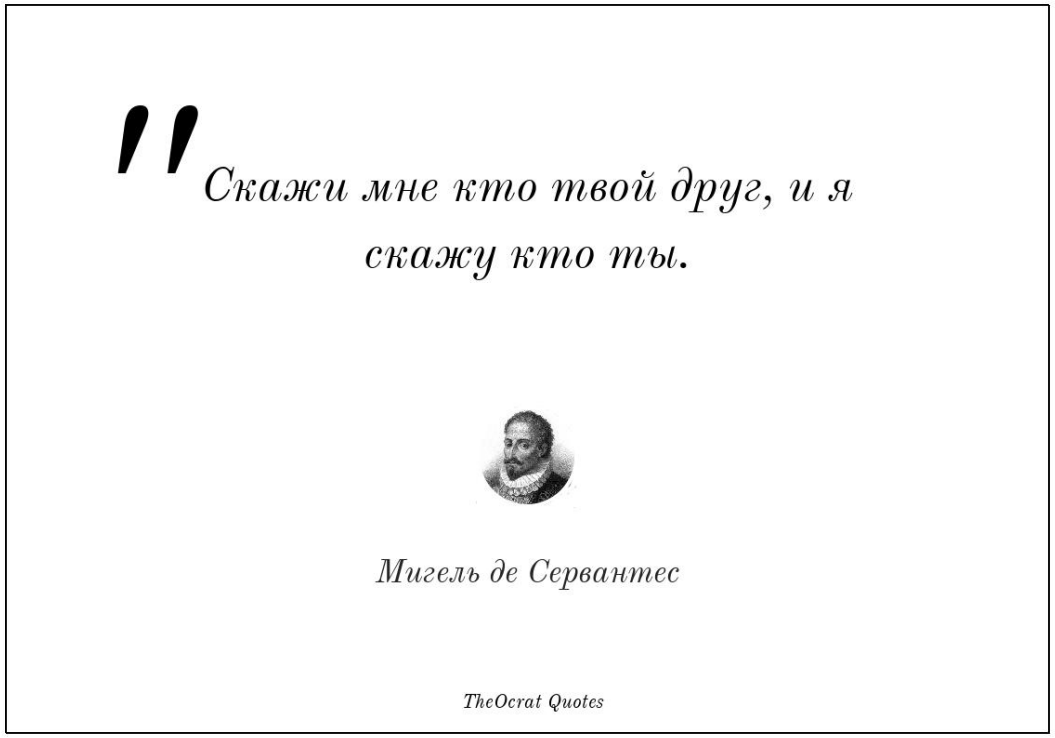
Практика применения ML моделей в оценке рисков

Март 2023 г.

www.cinimex.ru

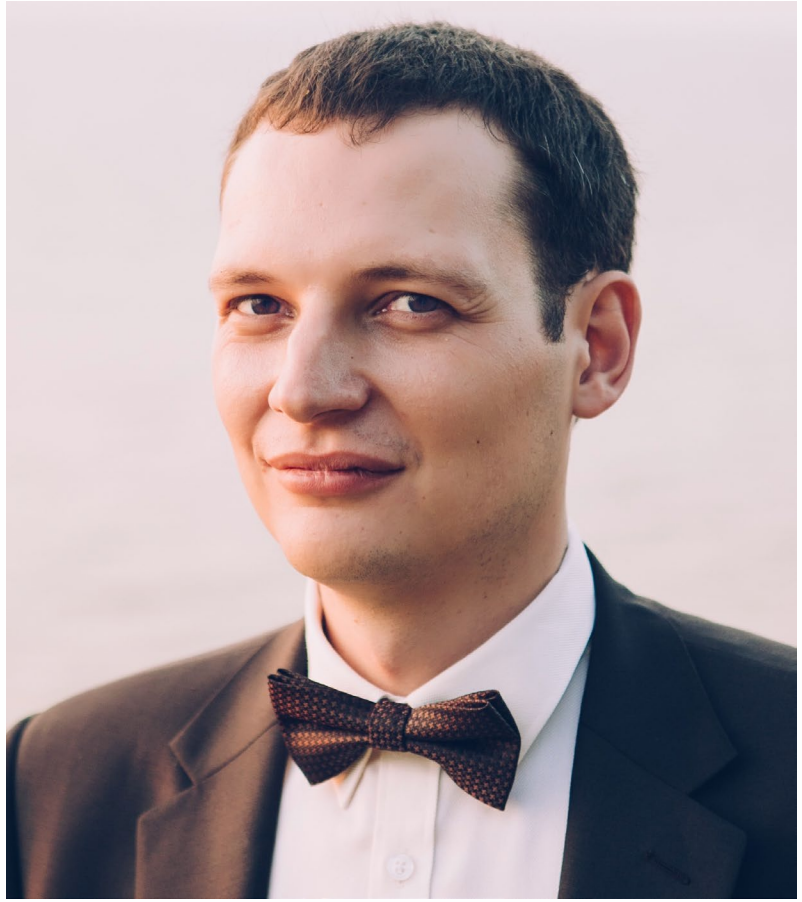
Ключевые тезисы

VS



Давайте знакомиться

Мартынов Родион
Старший руководитель проектов дирекции лаборатории данных, компания «Синимекс»
25 лет в IT, 5 лет в DS & DE & AI



Шихалиев Фрэнк
Консультант по развитию машинного обучения, Росгосстрах
15 лет в страховании и анализе данных

Компания «Синимекс» – разработчик ИТ–систем для бизнеса

Мы команда высококвалифицированных ИТ–специалистов, обладающих уникальной экспертизой и обширными компетенциями в области создания и внедрения бизнес–ориентированных и инфраструктурных решений в крупнейших организациях.

Наши преимущества

- **с 1997 года** на рынке
- **3000+** реализованных уникальных проектов
- **500+** сотрудников
- **отмечены** ведущими РА (CNews, RAEX, TAdviser)
- **партнерство** с ведущими мировыми поставщиками решений

Наши рейтинги

8 МЕСТО

TADVISER

Крупнейшие поставщики сторонних ИТ–решений из реестра отечественного ПО

10 МЕСТО

RAEX

Рэнкинг крупнейших ИТ–компаний и групп в области разработки программного обеспечения (2022 год)

13 МЕСТО

CNews Analytics

Разработчики и поставщики российских ИКТ–решений 2021

22 МЕСТО

CNews Analytics

Крупнейшие поставщики ИТ для финансового сектора

23 МЕСТО

RAEX

Рэнкинг крупнейших российских групп и компаний в области информационных и коммуникационных технологий (2022 год)

29 МЕСТО

CNews Analytics

Крупнейшие ИТ–разработчики России 2021

Наша география

Головной офис компании находится в **Москве**

Дистанционные сотрудники работают по **всей России**

Открыто **5 офисов** на территории России

К открытию готовятся еще **3 офиса**

Наши решения

- Заказная разработка
- Консалтинг (аналитика, архитектура, дизайн, MVP, исследования)
- Стаффинг команд и специалистов
- Внедрение Enterprise платформ
- FinTech и R&D
- DevOps (консалтинг, аудит и выстраивание процессов ЖЦ разработки ПО)

Интеграционные решения

Автоматизация бизнес–процессов

Внедрение стороннего ПО

Предоставление API (открытых интерфейсов)

Решения на микросервисной платформе

Машинное обучение и анализ данных

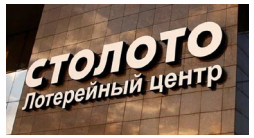
Машинное обучение и анализ данных

В наши услуги входит диагностика, проверка текущих гипотез по автоматизации, разработка моделей машинного обучения.

- Рассчитываем эффективность внедряемого решения
- Помогаем в систематизации текущих данных
- Проводим поддержку и актуализацию внедренных решений
- Всесторонне и объективно оцениваем новые технологии
- Имеем большой штат экспертов и специалистов в разных сегментах бизнеса

Ключевые проекты

Pyxis — Международный проект по оптимизации маркетинговых кампаний в интернете



Прогнозирование спроса и распределение лотерейных билетов по всей России

Разработка чат–бота и базы знаний на основе искусственного интеллекта

Сегментирование клиентов по территориальным факторам, определение высокорисковых клиентов

Определение SKU с помощью компьютерного зрения

python, R, pandas, scikit-learn, dmlc XGBoost, CatBoost, ANACONDA, TensorFlow, PyTorch, Python, Numba, matplotlib, seaborn, NumPy, Apache Spark, HORTONWORKS, jupyter, kafka, Flask, GRPC, kubernetes, docker, PostgreSQL, ClickHouse, ORACLE, cassandra, Microsoft SQL Server, Power BI

Итак!

Сегментация 2.0. Intro по проекту

Публикация о проекте

ML–проект Сегментация 2.0 компании Cinimex и Росгосстрах стал победителем в конкурсе Global CIO «Проект года 2022» в номинации «Лучшая модель машинного обучения для страховой компании».

Сегментация 2.0. Дорожная карта проекта

09/20

Старт проекта

Знакомство с источниками данных
Проработка гипотез
Анализ best practice

05/21

Геосегментация

Подтверждена гипотеза
Подтвержден экономический эффект
Визуализация рисков и потенциала продаж для различных страховых продуктов
Разработан и интегрирован в тарифный модуль РГС микросервис с вызовом модели

12/21

Графовая модель

Подтверждена гипотеза
Подтвержден экономический эффект
Построены инструменты инкрементного обновления графа
Разработан и интегрирован в тарифный модуль РГС микросервис с вызовом модели

06/22

Агенты и партнеры

Проработаны гипотезы
Построены витрины с инкрементным обновлением
Управленческий дашборд
Проект полностью сдан
Передача в поддержку и на развитие

Сегментация 2.0. Общие вводные

Геосегментация

Геосегментация. Гипотеза

Проверка влияния локального окружения территории на риск аварийности

Места ДТП

Данные договоров

Картографические данные

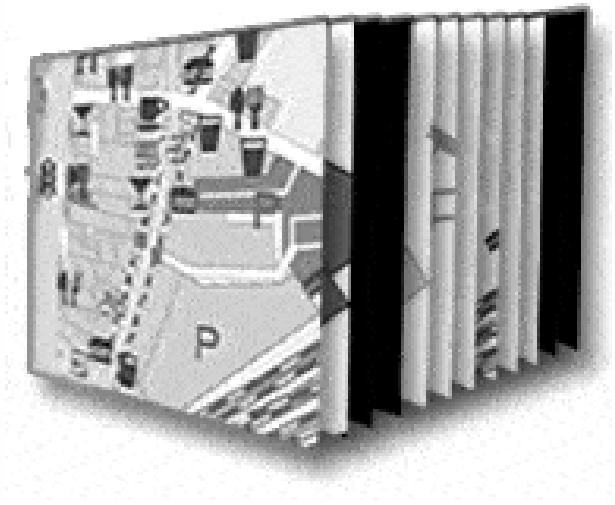
Местность

Дороги

...

Здания в окрестности

Дополнительные геометки



Набор карт для каждого убытка

Геосегментация. Подходы Geo2Vec

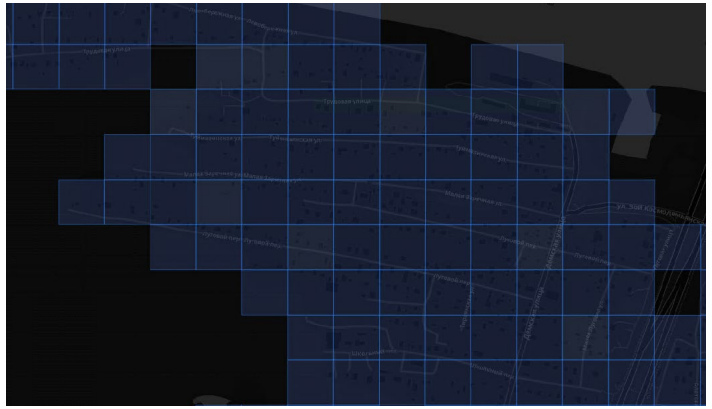
Исследовано около 10 релевантных публикаций и best practice, из которых наиболее подходящим был выбран метод **Loc2Vec** (Learning location embeddings with triplet–loss networks).

Модель Loc2Vec учитывает локальные географические свойства местности, стараясь построить такое векторное пространство, в котором расстояния между «похожими» векторами минимально, а между «непохожими» максимально.

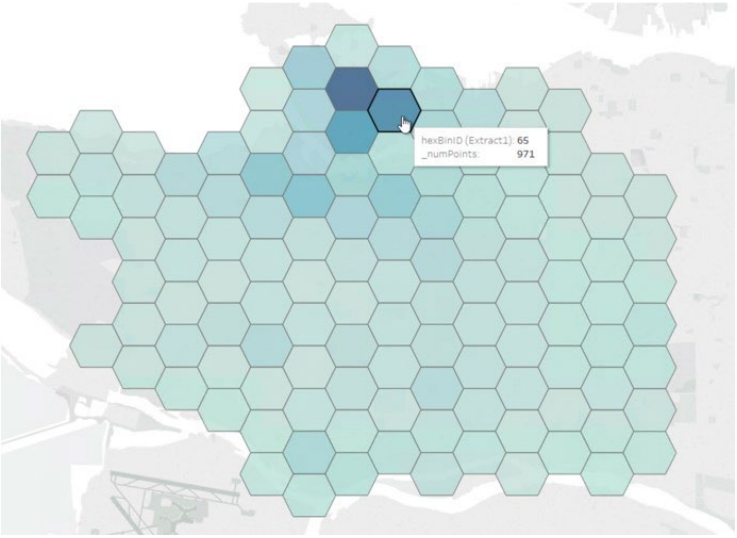
На вход модели подаются геоточки, в которых происходили ДТП / убытки, таким образом, модель учится «находить» места, похожие на те, где в прошлом уже реализовывались риски.

Геосегментация. Построение сетки полигонов

На основе административного деления



На основе квадратных тайлов



Гексагональная сетка

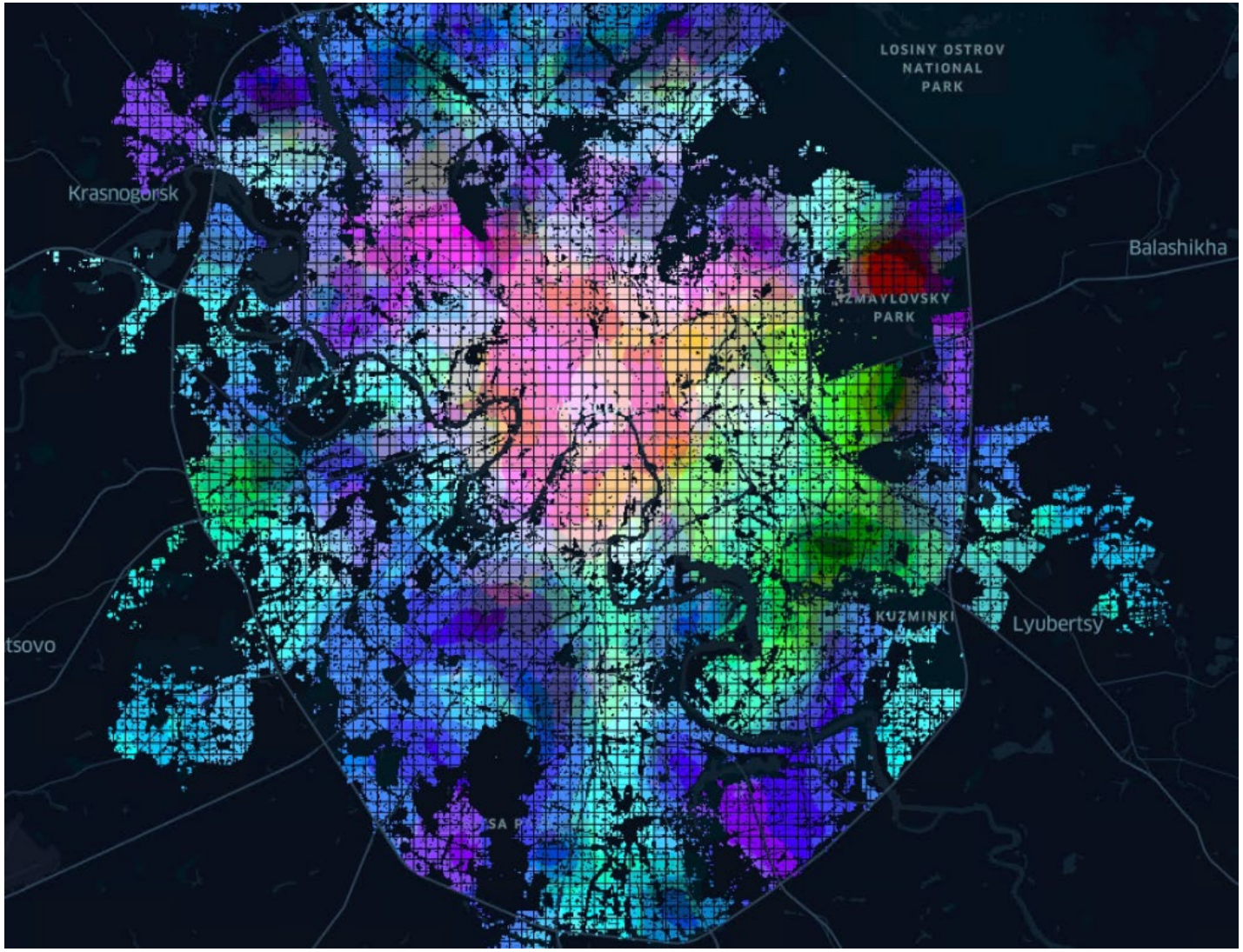
Геосегментация. Пример визуализации

Один из примеров визуализации полученных векторов.

Он не содержит оценку, и не говорит какой район более «рисковый».

Модель на данном этапе показывает только различия.

Модель говорит, что вот тут более непохоже, чем вот там.



Геосегментация. А как же скоринг?

В дальнейшем полученный из модели Loc2Vec вектор используется как дополнительный набор признаков моделью следующего уровня.

Векторное представление геолокации + дополнительные геопризнаки территории

Модель второго уровня

Оценка риска

Также модель второго уровня на вход принимает «классические» численные геопризнаки, описывающие оцениваемую территорию внешними для нее факторами.

Геосегментация. Направления

Область построения сетки полигонов
Размер и форма полигона

Эксперименты с нейросетью. Добавление «голов»

Список слоев OpenStreetMap
Размерность вектора на выходе модели Loc2Vec

Список фичей: демографические, инфраструктурные, объекты социальной инфраструктуры

Интеграция со скоринговыми моделями и модулем тарификации

Разработка инструментов визуализации

Эксперименты с loss функцией. Сегментирование по продукту, риску и размеру убытка

Геосегментация. Результаты

Оценка рисков

- Проведена большая аналитическая работа
- Целевая метрика на отложенной выборке улучшена

Визуализация

- Удалось повысить детализацию имеющихся решений геосегментирования
- Разработаны инструменты оценки развития и потенциала продаж в регионах

Графовая модель

Графовая модель. Intro

Цель: повышение точности текущих скоринговых моделей РГС при помощи добавления к оценке объекта анализа его окружения, построенного с применением графовых методов.

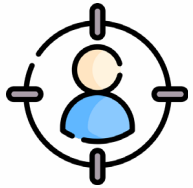
Базовая гипотеза: влияние окружения клиента при оценке риска клиента на этапе заключения договора страхования.

Понимание бизнес процесса

Задача – разработать модель, выявляющую высокорисковых клиентов на стадии заключения договора, которая:

Определяет окружение клиента

Анализирует степень рисковости окружения



Дает заключение по рисковости самого клиента

Новые признаки и скоры модели используются в качестве дополнения к существующей скоринговой модели РГС.

Котировка

Транспортное средство

Категория ТС: Легковые ТС (В)

Мощность двигателя: 200,00 л.с.

Тип ТС: Легковой

Цель использования: Личная

☑ ТС зарегистрировано в иностранном государстве

☐ ТС используется с прицепом

Марка: Daewoo

Год выпуска: 2000

Модель: Matiz

☐ Не нашел свою марку или модель

Название модели в полисе: Daewoo Matiz

Номер VIN автомобиля: 11111110111111111

Гос. номер автомоб: Ч222ЧЧ22

☐ VIN отсутствует

☐ Автомобиль не п

Серия и номер СТС: 2222 111111

Заключению договора предшествует «котировка», содержащая информацию о страхователе, застрахованных объектах и т.д.

При обработке котировки дается совокупная оценка риска, влияющая на стоимость полиса.

Водитель №1

☐ Водительское удостоверение иностранного государства

Фамилия	Имя	Отчество	Дата рождения
Иванов	Иван	Иванович	12.12.1917

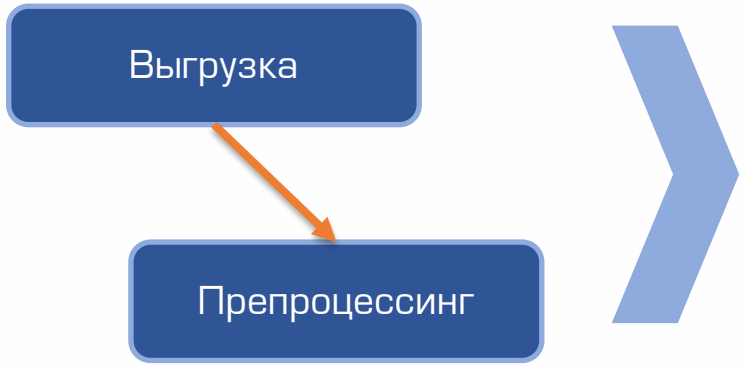
Водительское удостоверение		Начало стажа	
0101	010101	01.08.2022	Стаж вождения 0 лет

☐ Данные водителя изменились с прошлой покупки ОСАГО

Работа с данными

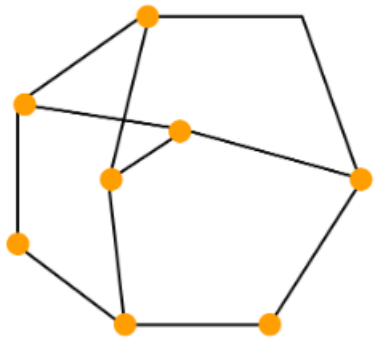
Сложности:

- Богатый атрибутный состав журнала котировок (600+ полей);
- Эффект масштаба, кол–во котировок измеряется сотнями миллионов в год;
- Недостаточно жесткие правила валидации при вводе исходных данных;
- **Фактор времени** при образовании связи в графе.

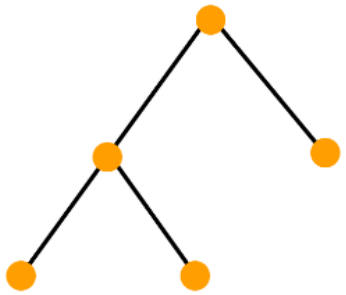


1. Получение данных из DWH;
2. Препроцессинг строковых значений;
3. Получение унифицированных колонок в соответствии с типами узлов;
4. Фильтрация выбросов;
5. Очистка и дедупликация;
6. ...

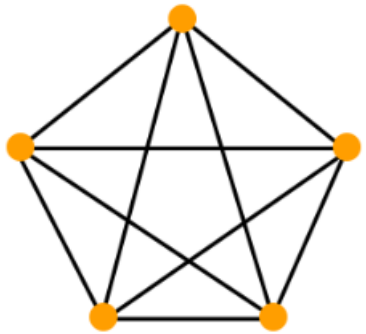
Виды графов



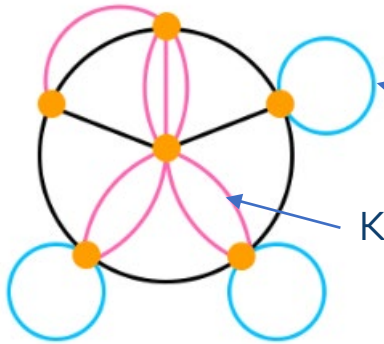
Обыкновенный



Дерево

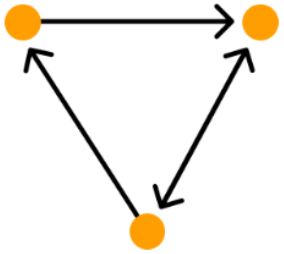


Полный

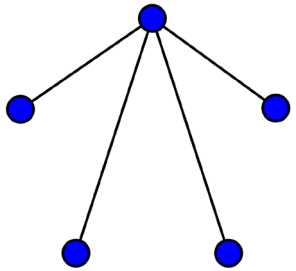


ПсевдоМультиГраф

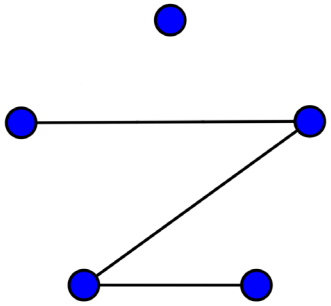
Метро	Коллеги по работе	Гиперссылки	



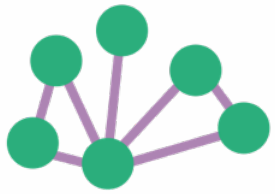
Ориентированный



Связный

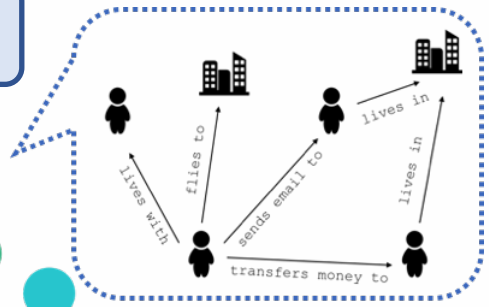


Несвязный



homogeneous

heterogeneous



Дорожное сообщение	Молекула	Дома и дороги	Пользователи платформы	Пользователи+ товары

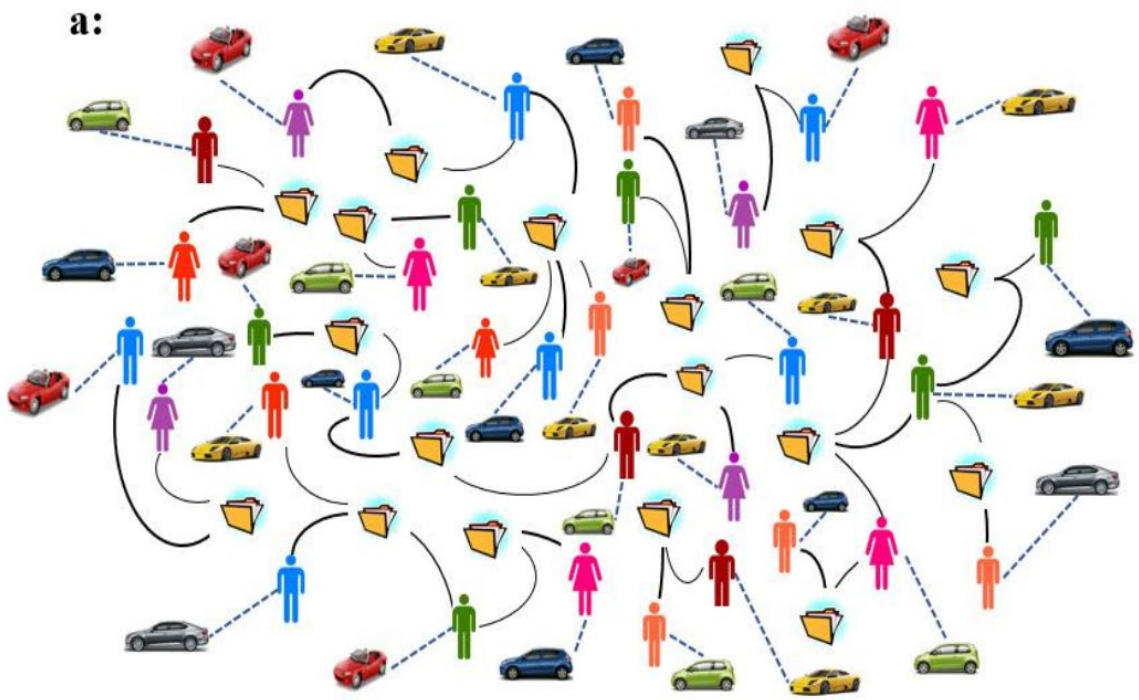
Большая подзадача – выбор целевого графа

Определение вида графа является первостепенным, т.к. от него зависит все остальное.

Для построения графа нужно определиться, какая структура будет оптимальной для хранения, удовлетворения требованиям и получения признаков.

Потенциальные структуры:

1. Однородный граф котировок;
2. Однородный граф узлов связи при сохранении истории их появления в котировках;
3. Гетерогенный граф узлов связи и котировок;
4. ...
5. Вариантов великое множество, см. пред. слайд.



Извлечение графовых признаков

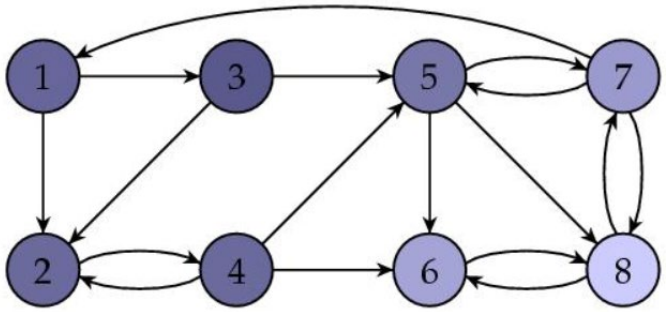
- Математические – оцениваются по структуре графа.
- Бизнес–признаки – рассчитываются для решения конкретных задач по бизнес–логике

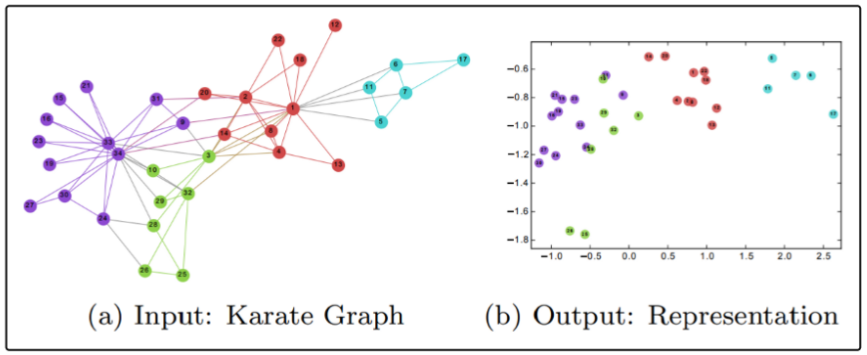
По узлам

- Node degree
- Eigenvector centrality
- Betweenness centrality
- Pagerank
- DeepWalk
- ...

По графу

- Adjacency matrix
- Laplacian matrix
- Density
- Efficiency
- Агрегирование признаков узлов
- Наличие циклов, клик
- ...

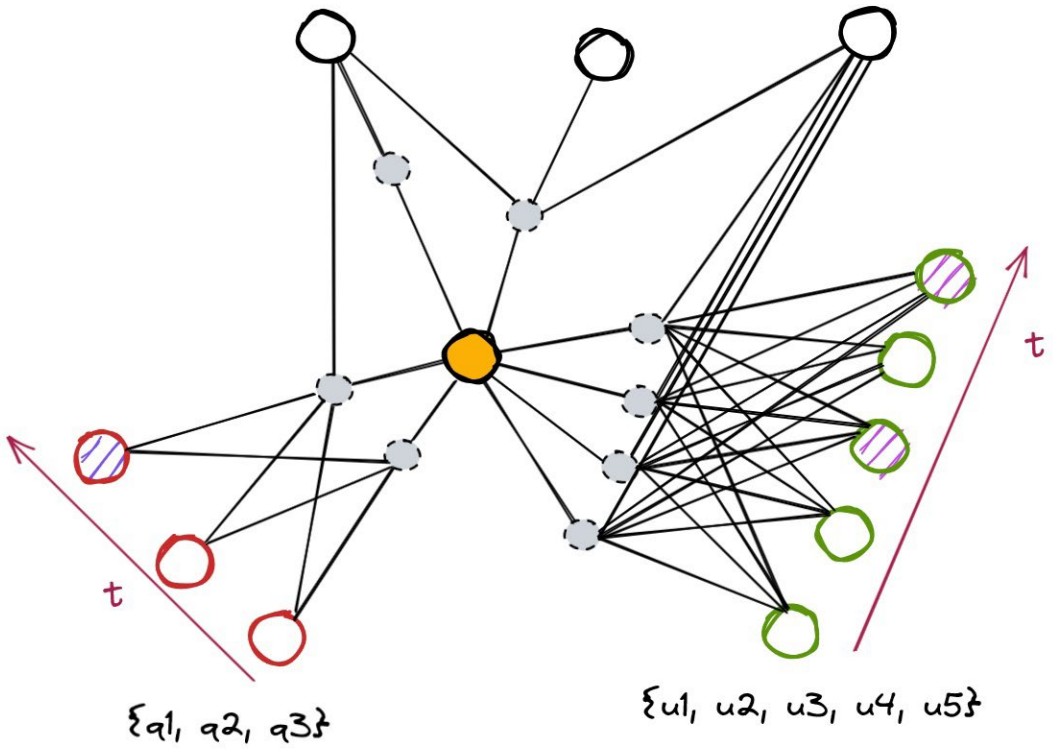



(a) Input: Karate Graph
(b) Output: Representation

Дедупликация

Проблематика:
Котировки могут повторяться, например автоматический расчет пролонгации и т.д.

Решение:
Чтобы «не надувать» граф и не искажать признаковое пространство графа необходимо выбрать **стратегию дедупликации** узлов и ребер графа.



R&D. Будни data science

1 Шаг — Давайте перестроим граф

2 Шаг — Извлечение признаков из нового графа

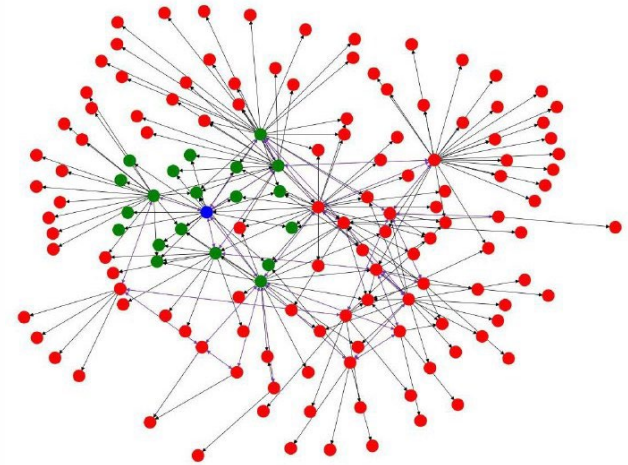
3 Шаг — Перепроверка всех имеющихся гипотез с новым набором данных

4 Шаг — Повторить с п.1)

NetworkX – чрезмерное потребление памяти и низкая скорость обхода (даже при наших кастомных патчах)
Graph-tool – значительная оптимизация потребления памяти и скорости обхода

Как подтверждали гипотезу

- Для графовой модели совместно с командой РГС был сформулирован критерий «высокорисковости»;
- Набор данных был размечен в соответствии с данным критерием;
- Во избежание лик–фактора все признаки считаются по прошедшей истории каждой котировки, помним про «фактор времени образования связи».



Признаки, извлеченные из графа для целевой котировки

→

Модель второго уровня

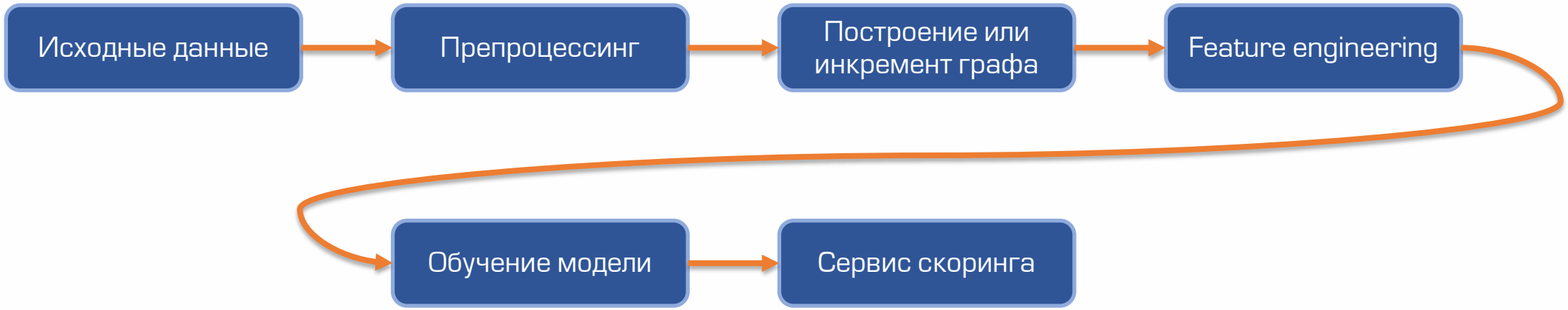
→

$$P(X|\Theta)$$

Оценка риска

Пайплайн модели

Пайплайн можно схематично представить:



Как работает модель в режиме оценки

Обработка котировки моделью на этапе inference:

- На вход получаем информацию по необходимым атрибутам котировки;
- Котировка "приклеивается" к большому графу, все узлы в нем носят характер доступной истории;
- Для котировки строится окружение;
- Модель возвращает графовые признаки и признаки котировок;
- Формируется score для целевой котировки;
- Score и значения признаков передаются в модуль тарификации.

Решение позволяет построить интерактивную визуализацию графа с полной информацией о его окружении необходимого порядка.

Стек технологий

pandas

NumPy

LightGBM

graph-tool

jupyter

DASK

Numba

Apache Airflow

SQLAlchemy

Flask

mlflow

OPTUNA

PostGIS
Spatial PostgreSQL

Вернемся снова к заглавному тезису

Make Graph Great Again!

Окружение имеет значение!

А вы что думаете?

Вопросы?

СПАСИБО ЗА ВНИМАНИЕ!

Лаборатория Данных

Узнайте больше –>>