Прогнозирование дефолта компании с использованием графов юридических лиц

СР2

Основная информация

- GitHub: https://github.com/Sanjar-Javodov/sna-2025-company-default-risk
- Telegram канал: https://t.me/+CP_szPWFMUg1NjM0
- Участники проекта(роли): Джаводов Санджар (ML Ops, Data Scientist), Кахоров Пайрав (Data Scientist, Data Analyst)

Ключевая идея и описание (из CP1)

- Используйте структурные шаблоны в графах финансовых транзакций, чтобы идентифицировать компании с высоким риском дефолта.

- Используйте функции, основанные на графах, и машинное обучение для выявления компаний, допустивших дефолт (косвенный признак: поведение, связанное с отмыванием денег).

Цель и задачи

Цель: Предсказать финансовый дефолт, используя график транзакционных взаимосвязей.

Шаги:

- Сбор данных (с минимальными затратами на борьбу с отмыванием денег)
- Построение ориентированного графа: счета в виде узлов, транзакции в виде ребер
- Назначение меток из флагов отмывания денег (по умолчанию = 1)
- Извлекать графические характеристики (центральность, степень, рейтинг страницы)
- Обучение (случайный лес + SMOTE)
- Визуализировать и интерпретировать результаты

Описание исследовательского набора данных

- Набор данных: IBM AMLSim synthetic dataset (HI-Small)
- Источник:https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml
- Тип: Полностью синтетический, имитирующий взаимодействие компаний и частных лиц посредством финансовых транзакций
- Маркировка: Транзакции, помеченные как отмывочные, используются для определения рисков компании (по умолчанию = 1)

Почему этот набор данных подходит

- Общедоступный и этически безопасный в использовании (синтетический, без реальных персональных данных)
- Содержит реалистичные схемы финансового взаимодействия (банковские переводы, покупки)
- Отражает поведение, связанное с отмыванием денег, которое является надежным показателем рискованного финансового поведения
- Позволяет строить крупномасштабные графики (>500 тыс. узлов), типичные для реальных систем

Проектирование и создание сети

- Узлы = Счета (предполагаемые компании)
- Направленные ребра = Финансовые транзакции от отправителя к получателю
- Вес ребра = Сумма транзакции
- Метка узла = Причастен к отмыванию денег → дефолтирован = 1

Описание сети

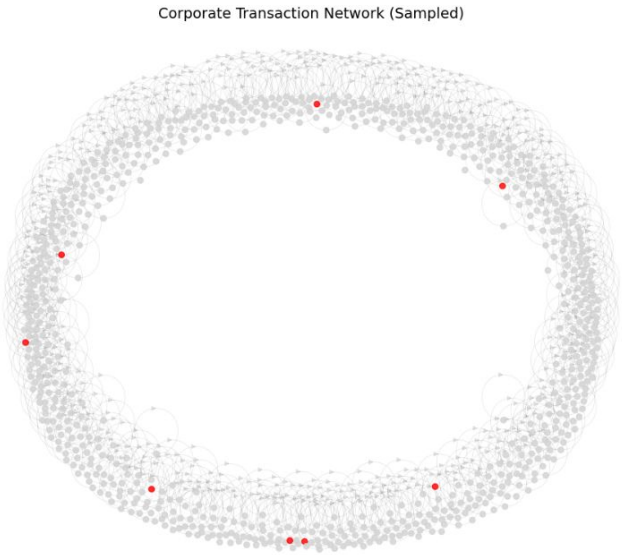
- Количество узлов: 514 210
- Число ребер: 989 036
- Плотность графа: 0,00000374 (очень разреженный)
- Средняя степень: 3,85
- Компоненты связи (слабые): 121 480
- Наибольший размер компонента: 360 653 узла (~70%)
- Диаметр: Слишком велик для эффективного вычисления

Визуализация графа

- Выбранный подграф с 800 узлами, визуализируемыми с помощью spring-макета
- Узлы окрашены в соответствии со статусом по умолчанию:

Красный = Значение по умолчанию (метка = 1)

Серый = Обычный (метка = 0)



EDA анализ

- Все объекты (in_degree, out_degree, pagerank, degree_centrality) сильно смещены вправо
- Узлы, используемые по умолчанию, нечетко разделены структурными элементами
- Класс 1 (по умолчанию) показывает лишь незначительные отклонения (в некоторых случаях немного выше рейтинг страницы / степень защиты)
- Это означает, что поведение по умолчанию встроено в структуру, близкую к нормальной, и ее нелегко отделить

Моделирование

Базовая модель: Случайный лес с class_weight="сбалансированный"

- Очень низкий уровень запоминания в классе 1 (3,4%), несмотря на высокую точность (98%)

Эксперимент SMOTE

- SMOTE использовался для балансировки тренировочного набора
- Результат: Количество отзывов увеличилось с 3,4% до 50,5%
- Но: количество ложных срабатываний увеличилось → точность снизилась до 2,1%, погрешность снизилась до 75%

Необходимо проанализировать

- SMOTE повышает чувствительность модели к редким событиям
- Это эффективный компромисс между возможностью запоминания и точностью
- Более продвинутые методы (временные, основанные на графиках, обнаружение аномалий), необходимые для CP3