Data mining для долгосрочного прогноза продаж в ритейл сети

PolyAnalyst

ПОСТАНОВКА ЗАДАЧИ

АНАЛИЗИРУЕМЫЕ ДАННЫЕ

База данных по объему ежедневных продаж по 4 уровням номенклатуры товарной иерархии во всех магазинах сети. Основные характеристики базы:

- Данные включают 6-летний период
- База данных охватывает 688 магазинов сети. Важно, что среди них есть как старые так и только что открывшиеся магазины.
- Общее число записей – 133 000 000.

База данных по характеристикам магазинов. Примерный состав данных:

- Регион
- Дата открытия
- Площадь, расположение и другие параметры магазина
- Демографическая ситуация в окрестности магазина

РЕШАЕМАЯ ЗАДАЧА

Основная задача - построение предиктивной модели на основе предоставленных данных, которая позволит предсказывать объем ежедневных продаж в каждом из магазинов сети по каждой из 10 товарных групп на глубину 1.5 года. Вследствие большого разброса объемов продаж по разным магазинам и товарным группам целью является минимизация относительного (в %) отклонения прогноза от факта.

СПЕЦИФИКА ЗАДАЧИ

СИЛЬНЫЕ СЕЗОННЫЕ ИЗМЕНЕНИЯ ПРОДАЖ

ИМЕЕТСЯ КАТЕГОРИЯ ТОВАРОВ («НОВОГОДНИЕ ТОВАРЫ»), СЕЗОННОСТЬ ПРОДАЖ КОТОРЫХ РЕЗКО ВЫРАЖЕНА И СИЛЬНО ОТЛИЧАЕТСЯ

РЕЗКО ВЫРАЖЕНА НЕДЕЛЬНАЯ ПЕРИОДИЧНОСТЬ ПРОДАЖ. ОНА СУЩЕСТВЕННО МЕНЯЕТСЯ В ЗАВИСИМОСТИ ОТ ВРЕМЕНИ ГОДА, А ТАКЖЕ ОТ МАГАЗИНА К МАГАЗИНУ

АНОМАЛЬНЫЕ ПРОФИЛИ ПРОДАЖ В НЕКОТОРЫЕ ПРАЗДНИЧНЫЕ И ПРЕДПРАЗДНИЧНЫЕ ДНИ

МЕТОДОЛОГИЯ И ЭТАПЫ АНАЛИЗА

ЭТАПЫ АНАЛИЗА ДАННЫХ

СИЛЬНЫЕ СЕЗОННЫЕ ИЗМЕНЕНИЯ ПРОДАЖ

Очистка данных

Сегментация данных

Вычисление временных производных для трансформации задачи анализа временных рядов в классическую задачу data mining

Разбиение данных на обучающую и тестовую выборки

Построение компонент модели

- Так как целью является минимизация относительной, а не абсолютной продажи, для всех параметров, характеризующих объемы продаж, переходим к их логарифмам

- Так как важной характеристикой магазина является его структура продаж, вычисляем характеризующие ее производные: самую продаваемую подкатегорию товара в каждой категории, разнообразие (энтропию) продаж, доли покупок со скидками и т.д.

МЕТОДОЛОГИЯ И ЭТАПЫ АНАЛИЗА

ЭТАПЫ АНАЛИЗА ДАННЫХ

Сильные сезонные изменения

ОЧИСТКА ДАННЫХ

Сегментация данных

Вычисление временных производных для трансформации задачи анализа временных рядов в классическую задачу data mining

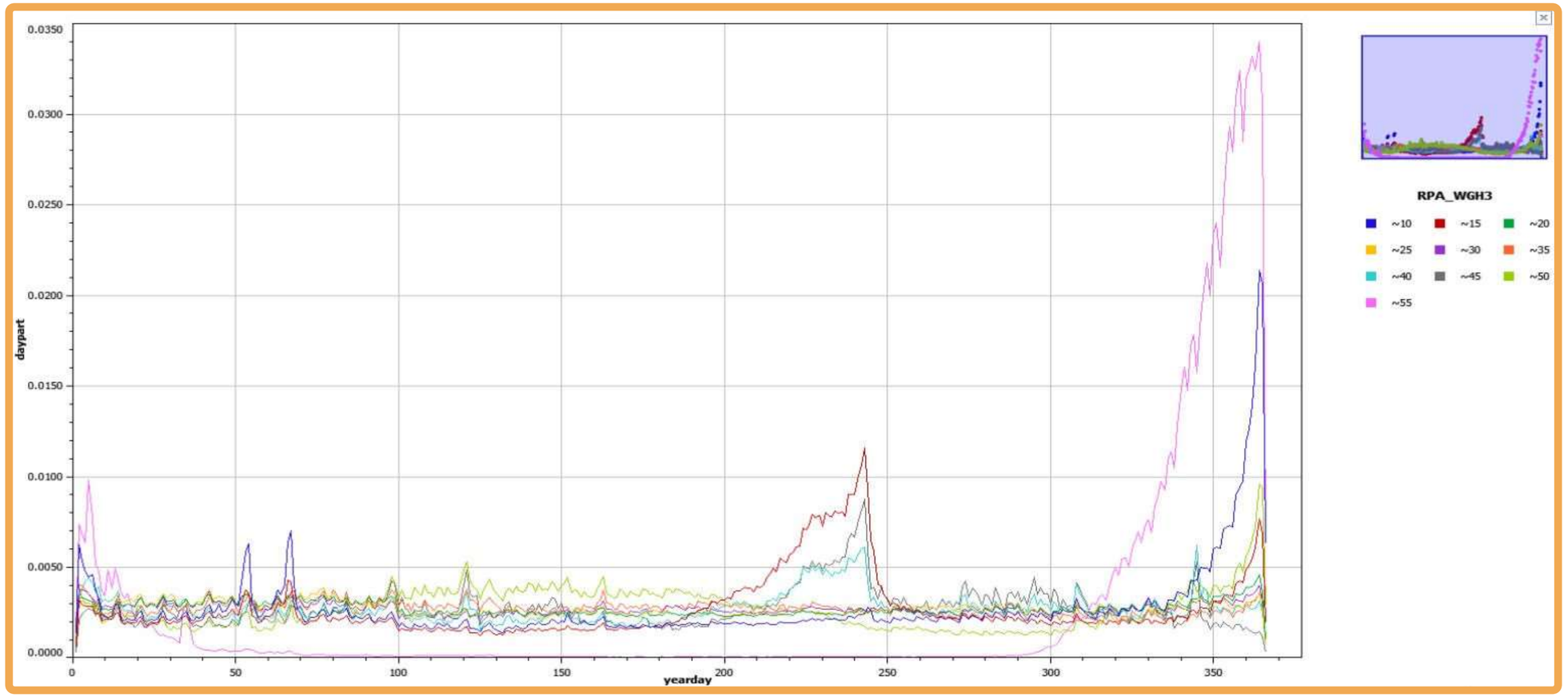
Разбиение данных на обучающую и тестовую выборки

Построение компонент модели

- Проводится фильтрация отладочных и некорректных данных, данных по ныне не существующим магазинам и т.д.

ПРОДАЖИ ПО ДНЯМ ГОДА

Общее распределение продаж разных товарных групп по дням года.



Есть товарная группа (55 – новогодние товары), сильно отличающаяся от продаж других товарных групп. Средства визуализации и манипулирования данными в PolyAnalyst позволяют быстро находить типичные и нетипичные случаи.

МЕТОДОЛОГИЯ И ЭТАПЫ АНАЛИЗА

ЭТАПЫ АНАЛИЗА ДАННЫХ

Сильные сезонные изменения

Очистка данных

СЕГМЕНТАЦИЯ ДАННЫХ

Вычисление временных производных для трансформации задачи анализа временных рядов в классическую задачу data mining

Разбиение данных на обучающую и тестовую выборки

Построение компонент модели

- Некоторые товары требуют подхода к моделированию продаж, значительно отличающегося от других.

- Очевидно, подходы к прогнозу продаж для давно существующих и только что открывшихся магазинов должны быть совершенно разными

КАК ОБРАБАТЫВАТЬ ДАННЫЕ?

ДАННЫЕ ПРЕДСТАВЛЯЮТ СОБОЙ ВРЕМЕННЫЕ РЯДЫ.

PolyAnalyst имеет инструментарий для работы с временными рядами, но надо ли его здесь применять?

Тип зависимости продаж складывается из заранее известных компонент:

- рост продаж («раскрутка») новых магазинов
- годичная сезонность
- недельная сезонность
- аномалии в праздничные и предпраздничные дни
-эффект от уменьшения продаж при открытии неподалеку другого магазина сети («каннибализация»)

Надо выяснить параметры этих зависимостей, и тогда мы превратим задачу анализа временных рядов в классическую задачу data mining.

ОБЩАЯ СТРУКТУРА МОДЕЛИ

ln(ОБЪЕМ ПРОДАЖ В ДАННЫЙ ДЕНЬ)

=

ln(СРЕДНИЙ ДНЕВНОЙ ОБЪЕМ ПРОДАЖ ДЛЯ ДАННОГО ГОДА)

+

ln(ПОПРАВКА ДЛЯ ДАННОГО ДНЯ)

МЕТОДОЛОГИЯ И ЭТАПЫ АНАЛИЗА

ЭТАПЫ АНАЛИЗА ДАННЫХ

Сильные сезонные изменения

Очистка данных

Сегментация данных

ВЫЧИСЛЕНИЕ ВРЕМЕННЫХ ПРОИЗВОДНЫХ ДЛЯ ТРАНСФОРМАЦИИ ЗАДАЧИ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ В КЛАССИЧЕСКУЮ ЗАДАЧУ DATA MINING

- День недели
- День года, фаза года (sin и cos «угла года от 1 января»)
- Праздничный или предпраздничный день

Разбиение данных на обучающую и тестовую выборки

Построение компонент модели

МЕТОДОЛОГИЯ И ЭТАПЫ АНАЛИЗА

ЭТАПЫ АНАЛИЗА ДАННЫХ

Сильные сезонные изменения

Очистка данных

Сегментация данных

Вычисление временных производных для трансформации задачи анализа временных рядов в классическую задачу data mining

РАЗБИЕНИЕ ДАННЫХ НА ОБУЧАЮЩУЮ И ТЕСТОВУЮ ВЫБОРКИ

Построение компонент модели

- Последний год в данных будет использован для оценки точности создаваемых моделей.

МЕТОДОЛОГИЯ И ЭТАПЫ АНАЛИЗА

ЭТАПЫ АНАЛИЗА ДАННЫХ

Сильные сезонные изменения

Очистка данных

Сегментация данных

Вычисление временных производных для трансформации задачи анализа временных рядов в классическую задачу data mining

Разбиение данных на обучающую и тестовую выборки

ПОСТРОЕНИЕ КОМПОНЕНТ МОДЕЛИ

- Модель прогноза среднего объема дневных продаж;
- Модель отклонения объема продаж в данный день от среднего.

ОБЪЕМ ДНЕВНЫХ ПРОДАЖ

МОДЕЛЬ ПРОГНОЗА СРЕДНЕГО ОБЪЕМА ДНЕВНЫХ ПРОДАЖ

БЛИЖНИЙ ПРОГНОЗ (на следующий год)

На основе модели изменения среднедневных продаж по сравнению с текущим годом

ДАЛЬНИЙ ПРОГНОЗ (на год после следующего)

На основе глобального тренда развития продаж данного магазина

ПРОГНОЗ ДЛЯ НОВЫХ МАГАЗИНОВ

На основе построения нормы среднедневных продаж для магазина с данными характеристиками, умноженный на линейный временной тренд (эффект «раскрутки»)

ИЗМЕНЕНИЕ ДНЕВНЫХ ПРОДАЖ

МОДЕЛЬ ИЗМЕНЕНИЯ СРЕДНЕДНЕВНЫХ ПРОДАЖ ПО СРАВНЕНИЮ С ТЕКУЩИМ ГОДОМ НА ОСНОВЕ СВОЙСТВ МАГАЗИНА И ПАРАМЕТРОВ, ХАРАКТЕРИЗУЮЩИХ ПРОДАЖИ В ТЕКУЩЕМ ГОДУ

Модель вряд ли будет очень точной, мало примеров, при большом количестве предикторов → Опасность переобучения → Следует использовать простой устойчивый метод

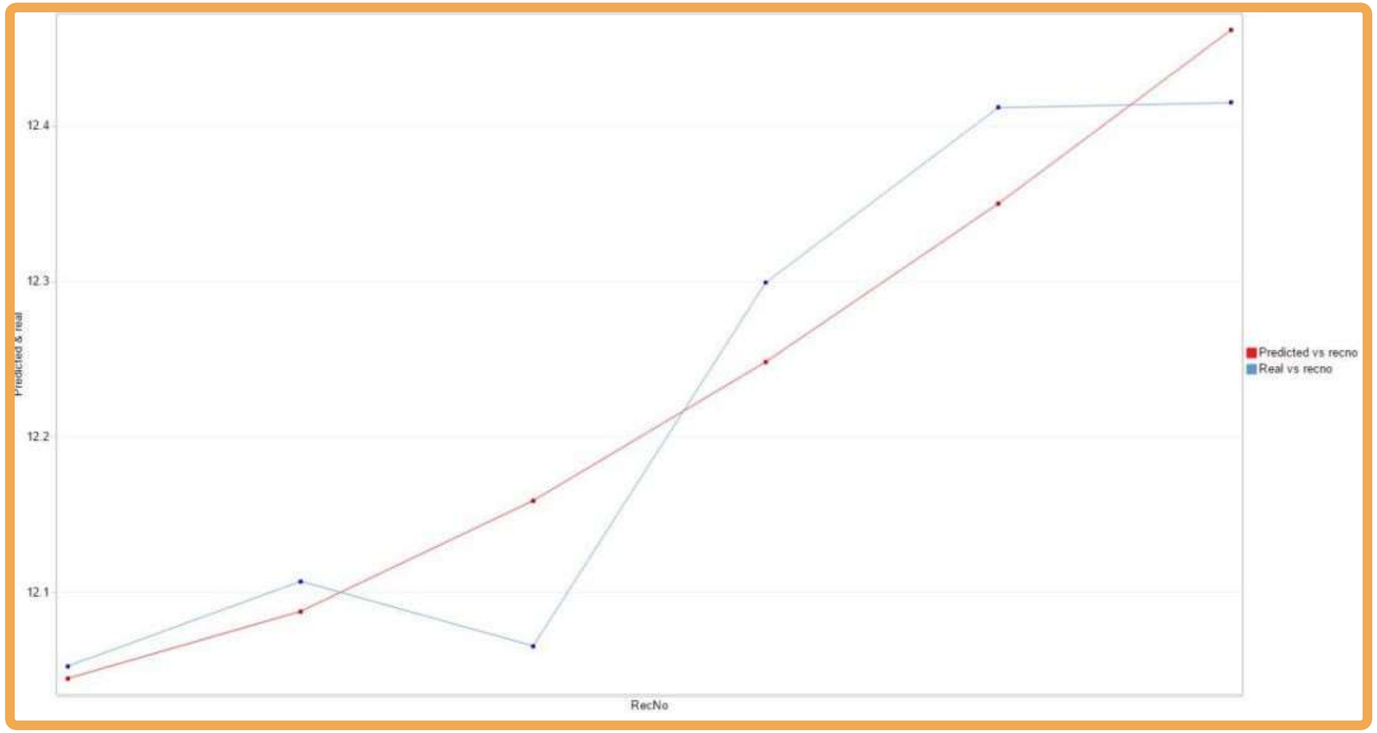
Например, для игрушек простую но достаточно точную модель дает дерево решений

Для совсем новых магазинов -рост 26%
Для недавно открытых (менее полутора лет) – рост 14%
Для старых магазинов на юге – рост 11% (любопытный факт!)
Для остальных старых магазинов – рост 3%

ГЛОБАЛЬНЫЙ ВРЕМЕННОЙ ТРЕНД

ГЛОБАЛЬНЫЙ ВРЕМЕННОЙ ТРЕНД РАЗВИТИЯ ПРОДАЖ ДАННОГО МАГАЗИНА

- Мало данных (6 точек), задача экстраполяции , поэтому требуется устойчивость => мало слагаемых, невысокий порядок.
- Используется модель ln (sales) = a0 + a1xt + a2x√t, где √t = год – 2010



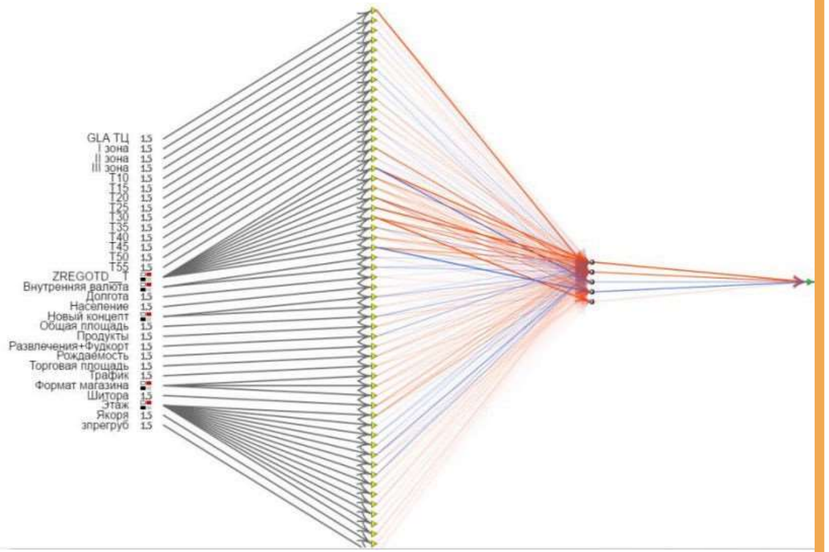
Например, для магазина в Москве на улице Вавилова:

ln(sales) = +0.170629*(+1.00356+1.30073*t-3.29720*sqrt(t)+71.6480)

НОРМЫ СРЕДНЕДНЕВНЫХ ПРОДАЖ

ПОСТРОЕНИЕ НОРМЫ СРЕДНЕДНЕВНЫХ ПРОДАЖ ДЛЯ МАГАЗИНА С ДАННЫМИ

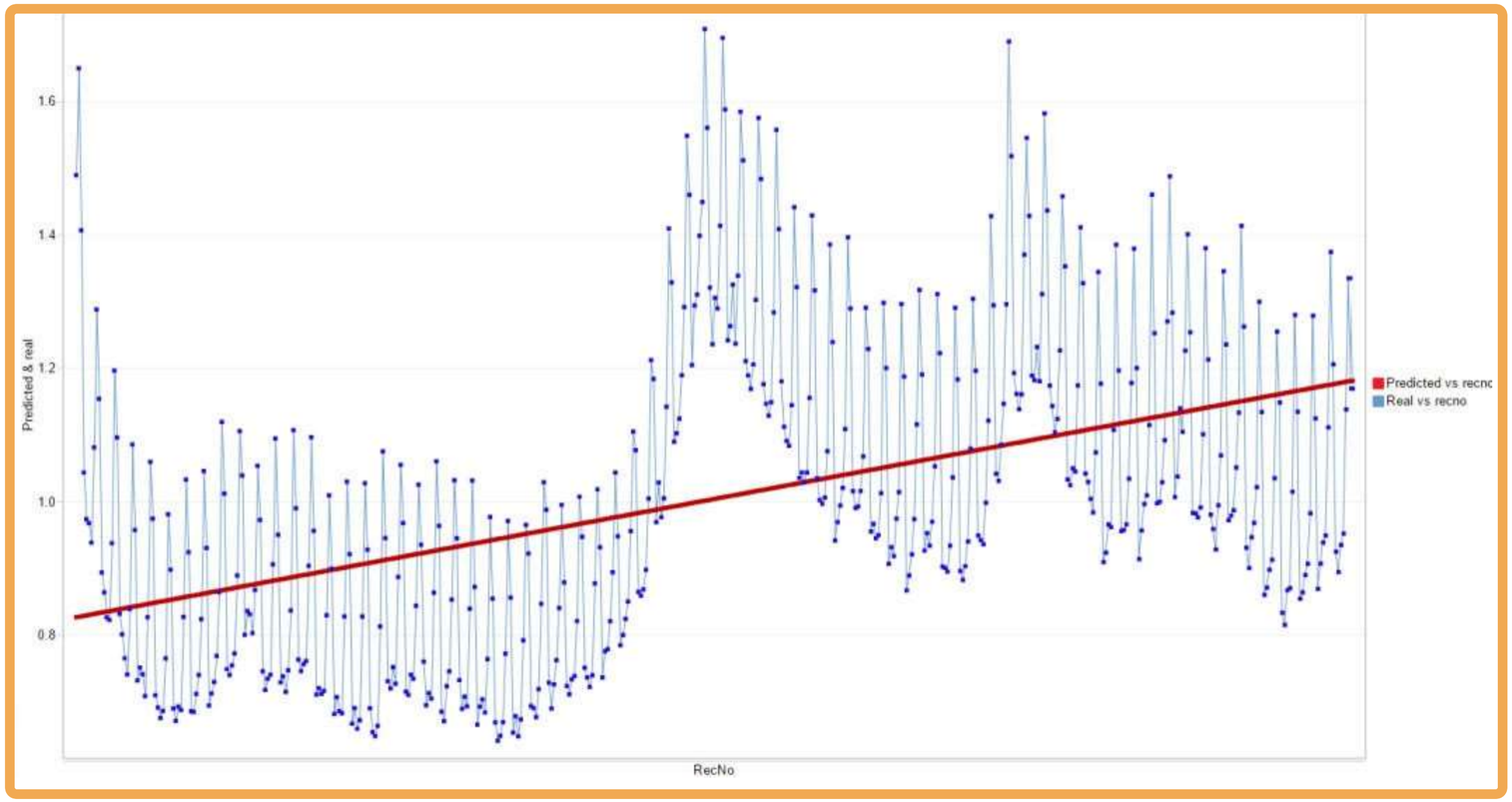
- Модель может быть достаточно точной, мало примеров, при большом количестве предикторов => опасность переобучения => следует использовать простой устойчивый метод
- Лучшей оказалась простая нейронная сеть (1 скрытый слой с 5 нейронами)



Самые важные факторы:
регион, этаж, на котором находится магазин, формат магазина.

ЭФФЕКТ «РАСКРУТКИ» НОВЫХ ТОЧЕК

Было обнаружено, что рост логарифма дневных продаж для новых (существующих меньше полутора лет) магазинов можно грубо описать линейной зависимостью от времени, но коэффициент этого роста для всех товаров разный. Быстрее всего растут продажи подгузников, медленнее всего – товаров для спорта и активного отдыха.



Пример этой аппроксимации для канцелярии, книг, мультимедиа.

ОТКЛОНЕНИЕ ДНЕВНЫХ ПРОДАЖ

МОДЕЛЬ ОТКЛОНЕНИЯ ОБЪЕМА ПРОДАЖ В ДАННЫЙ ДЕНЬ ОТ СРЕДНЕГО

- Модель может быть достаточно точной, много примеров => опасность переобучения умеренная
- Желательна интерпретируемость модели
- Лучшим оказалось дерево решений
- Достигнуто R2 = 0.459

НЕИЗВЕСТНЫЕ ФАКТОРЫ

УЧЕТ НЕИЗВЕСТНЫХ ДОЛГОВРЕМЕННЫХ ФАКТОРОВ

ПРОБЛЕМА

Далеко не все факторы, влияющие на продажи, отражены в данных (деятельность конкурентов, макроэкономические факторы, влияющие на покупательскую способность, долговременные погодные аномалии, открытие или закрытие больших предприятий поблизости и т.д.)

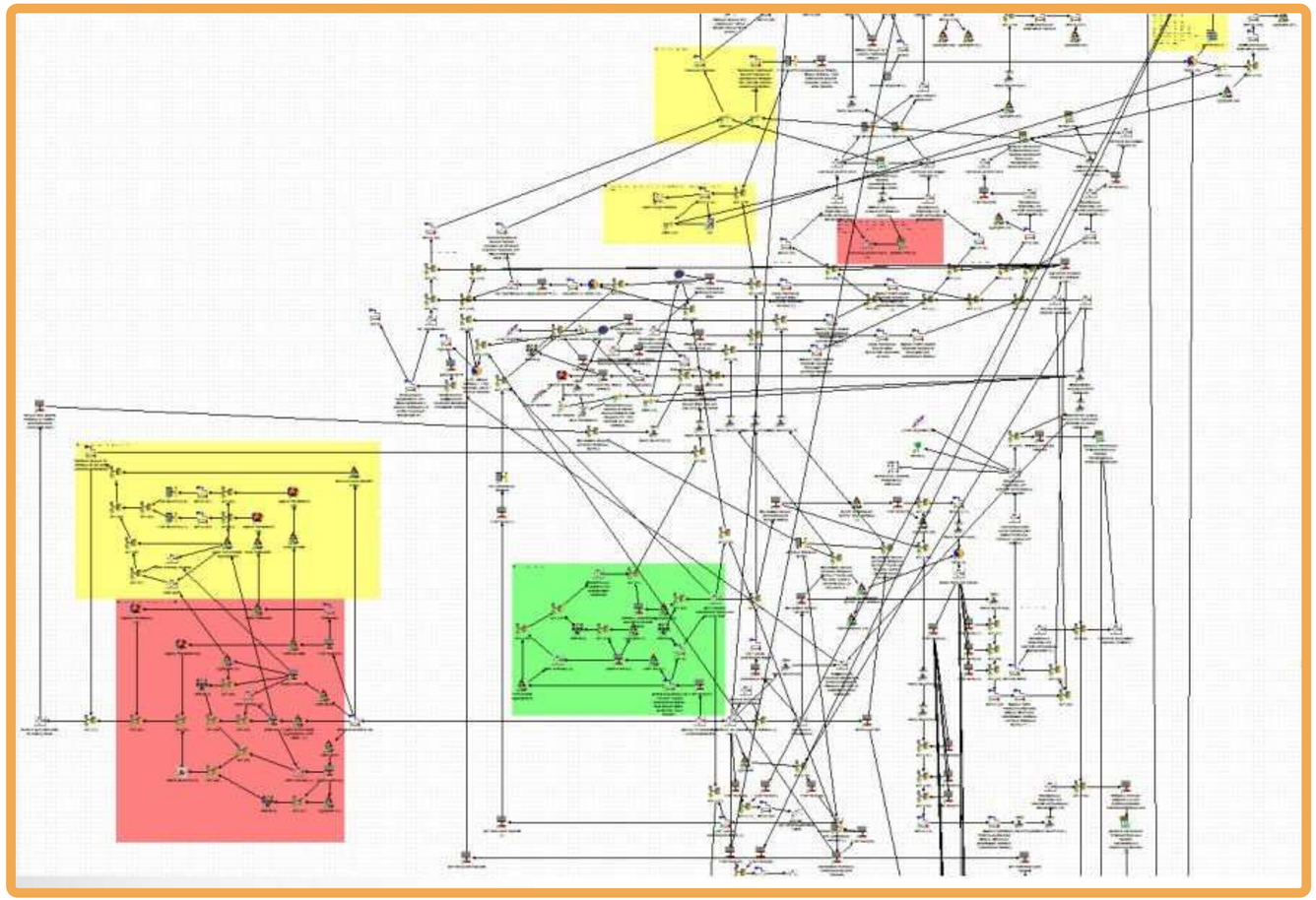
РЕШЕНИЕ

Динамическая коррекция модели. Если мы видим, что прогнозная модель систематически ошибается, мы понимаем, что начал действовать некий неизвестный нам фактор, и требуется корректировать прогноз.

Оптимизация схемы этой коррекции делается на основе обучения на имеющейся истории.

ПРОДУКЦИОННЫЙ СКРИПТ

ФРАГМЕНТ ПРОДУКЦИОННОГО ПРОГНОСТИЧЕСКОГО СКРИПТА



Были перечислены основные факторы, влияющие на прогноз, но жизнь большой торговой сети включает много нюансов. Поэтому продукционный скрипт для этой задачи включает несколько сотен узлов. Тем не менее, удобные средства многопользовательской разработки в PolyAnalyst делают поддержку и развитие этого скрипта аналитическим отделом ритейл сети относительно несложной задачей.

РЕЗУЛЬТАТЫ

ВЫГОДЫ ПРИОБРЕТАЕМЫЕ РИТЕЙЛ СЕТЬЮ С ВНЕДРЕНИЕМ СИСТЕМЫ POLYANALYST

Возможность прогнозировать объем продаж на каждый день в каждом магазине по каждой товарной группе на полтора года вперед со средней точностью 46%.

Для старых магазинов эта точность повышается до 39.2%

Чем больше объемы продаж (по магазину или по товарной группе), тем точность прогноза выше. Например, точность прогноза по игрушкам (на старых магазинах) составляет 10.19%, а общая точность прогноза по самым крупным магазинам доходит до 27%.

О КОМПАНИИ МЕГАПЬЮТЕР ИНТЕЛЛИДЖЕНС

PolyAnalyst 6.5

МЕГАПЬЮТЕР ИНТЕЛЛИДЖЕНС

Извлекаем и структурируем факты из текстовых документов

Оцифровываем и роботизируем бизнес-процессы

Строим модели на основе аналитики и Искусственного Интеллекта

107 разработчиков, 16 лингвистов и аналитиков, 9 кандидатов наук

Предоставляем кластерную платформу для анализа Больших Данных

Поддерживаем четверть компаний из списка Fortune 100 и еще более 100 клиентов

Член Ассоциации Разработчиков Программных Продуктов «Отечественный софт»

Платформа PolyAnalyst включена в реестр Российского ПО. Свидетельство №4414

КОМАНДА МЕГАПЬЮТЕР

Давид Сазонов

Руководитель направления текстового анализа

sazonov@megaputer.ru

Алексей Русских

Генеральный директор

+7 (915) 424-23-45

russkikh@megaputer.ru

Сергей Ананян

Исполнительный директор

sananyan@megaputer.ru

Гольцов Дмитрий

Заместитель ген. директора
Коммерческое направление

+7 (916) 111-95-49

goltsov@megaputer.ru

Мегапьютер Интеллидженс

Москва, ул. Бауманская 6, офис 723
+7 (499) 753-01-29
www.megaputer.ru