Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

КУРСОВАЯ РАБОТА

Программный проект на тему

"Построение модели расчёта каннибализации"

Выполнила студентка группы 186, 3 курса,
Наумова Анастасия Константиновна

Руководитель КР:
Наумов Данила Константинович

Соруководитель КР:
Глазкова Екатерина Васильевна

Москва 2021

Содержание

Аннотация на русском языке

В современном ретейле существует множество параметров, которые влияют на конечный спрос продукта. При проведении промо-акций эффект обычно оценивают, как прирост промо-продаж относительно регулярного спроса. При этом очень часто товары, стоящие в промо, снижают спрос на похожие товары. Этот эффект называется «каннибализацией».

Игнорирование такого эффекта приводит к некорректной оценке промо-эффективности и снижает коммерческую выгоду для ретейлера.

В рамках данного проекта проведено изучение соответствующей экономической модели, исследование, реализация и сравнение различных подходов к построению модели для оценки продаж товаров с учётом эффекта каннибализации на данных по продажам компании «Утконос».

Аннотация на английском языке

In modern retail there are many characteristics that impact on product's final demand. The effect of promotion compaigns is evaluated as sales growth during promotion relatively to regular sales while product in promotion can reduce sales for similar products. This effect is called "product cannibalization".

Ignoring this effect leads to an incorrect assessment of promotional efficiency and reduces the commercial benefit for the retailer.

Under the project a corresponding economic model was studied, research on various approaches to building a model was made and the model for assessing sales of goods, taking into account the effect of cannibalization on Utkonos sales data was implemented.

Ключевые слова

Product cannibalization, promotion compaigns, promo strategy, Machine Learning, effectiveness.

1 Введение

Ретейлеры часто используют промо-акции, так как они являются одними из наиболее эффективных методов привлечения клиентов. Иногда промо-акции являются вынужденной мерой в борьбе с конкурентами, но наиболее весомой причиной использования промо-акций является получение прямой выгоды за счёт эластичности спроса.

Продажи товара наиболее сильно зависят от цены на него самого. Для подсчёта изменений продаж при изменении цены используют эластичность спроса по цене. Она показывает, на сколько процентов изменится величина спроса при изменении цены. Однако продажи также зависят от цен на другие товары.

Эффект от проведения промо-акции можно декомпозировать по элементам. Часть из них связаны с поведением товаров-заменителей и товаров, которые покупают вместе. При подсчёте данных элементов используется перекрёстная эластичность или кросс-эластичность — показатель процентного изменения в количестве купленного товара или услуги в ответ на изменение в цене другого товара или услуги. В экономике выделяется два основных кросс-эффекта: Хало (или Гало) и каннибализация. Хало эффект возникает при продаже товаров комплементов, то есть «дополнителей». Например, при скидках на кроссовки могут вырасти продажи носков.

Аналогично, эффект каннибализации возникает на товары субституты, «заменители». Существует множество примеров, в которых с экономической точки зрения доказывается существование каннибализации. Например, продажи iPhone могут уменьшить продажи iPad в определённый период времени (On product cannibalization. A new Lotka-Volterra model for asymmetric competition. (2016).), или большие скидки на шоколад Lindt могут забирать продажи не только у своего аналога (элитного шоколада), но и у более дешёвых товаров данного сегмена(McKinsey Digital. Эффективное промо: разобраться и перенастроить. (2019).). Учёт этого эффект является важным при

планировании промо-акции, так как снижение продаж каждого отдельного товара может быть не столь велико, однако, суммарное снижение продаж может оказаться большим.

Ключевая сложность задачи оценки эффекта каннибализации состоит не в том, чтобы вычислить, насколько снизились продажи товаров — на этот вопрос можно ответить методами базовой аналитики. Необходимо определить, какие промо-акции привели к снижению продаж каких товаров.

Таким образом, исходная задача работы — построить модель, обученную вычислять коэффициенты каннибализации и находить по нему средний объём, на который уменьшатся продажи товара, из-за другого, стоящего в промо-акции. Во время работы над проектом задача свелась к построению модели, позволяющей предсказывать продажи товаров с учётом эффекта каннибализации. Исследование и реализация методов построения модели, а также измерение результатов, их сравнение и описание было проделано в данной работе.

Данная работа имеет следующую структуру. В главе 2 проводится обзор существующих решений. В 3 рассматривается экономическая модель и приводится идея решения задачи с точки зрения машинного обучения. Глава 4 рассказывает о проведённой работе с данными. Описание первого и второго этапа экспериментов приведено в главах 5 и 6 соответственно. Дальнейшие направления для исследования обозначены в главе 7. Выводы из данной работы описаны в главе 8.

1.1 Актуальность

Данная работа позволит улучшить точность регулярного и промо прогноза продаж, уменьшить количество неэффективных промо-акций. Повышение точности прогноза продаж может привести к увеличению прибыли, что является наиболее важной задачей ретейлера. Этим обусловлена актуальность курсовой работы по реализации модели прогноза продаж с учётом эффекта

каннибализации.

1.2 Цели и задачи

Цель данной работы состоит в изучении эффекта каннибализации и построении модели, позволяющей проводить его оценку. Исходя из этого, были поставлены следующие задачи:

1 Изучить теорию и литературу по данной теме

2 Провести исследование различных подходов к подготовке данных

3 Провести исследование различных подходов к измерению каннибализации

4 Построить и сравнить различные модели для выбранных подходов к измерению каннибализации, подобрать наилучшие параметры регуляризации

5 Провести интерпретацию моделей

6 Привести практическое доказательство существования каннибализации

2 Обзор существующих решений

Существует несколько экономических статей (Product Cannibalization and the Role of Prices. (2001), On product cannibalization. A new Lotka-Volterra model for asymmetric competition. (2016), New model introductions, cannibalization and market stealing (2014).), исследующих явление каннибализации, показывающих последствия этого эффекта, приводящих различные примеры. Такие статьи не показались насыщенными для моей работы, так как их основной задачей было нахождение и доказательство существования эффекта каннибализации, а не его подсчёт и реализация модели. Безусловно, необходимо провести хотя бы краткое практическое доказательство существования эффекта каннибализации в данных, с которыми проводится работа, но основным вопросом является построение модели и подсчёт эффекта каннибализации. В этом оказались более полезны менее научные, но более современные источники, такие как презентации компании SAS (SAS Institute Inc. Математика цен. Эластичность и ценовые эффекты. (2017), SAS Institute Inc. Анализ промоакций. (2015)) и пост McKinsey Digital. Эффективное промо: разобраться и перенастроить. (2019). В них говорится о важных вопросах и деталях, на которые необходимо обратить внимание при решении описываемой задачи. Стала очевидной необходимость сужения сегмента рассматриваемых товаров на узкие категории и более серьёзная очистка данных.

Кроме того, эти источники раскрывают декомпозицию спроса и помогают понять, как применяется подсчёт каннибализации в современном ретейле. Они помогли составить более полное видение задачи и сформировать план работы.

Подавляющее большинство источников говорит о том, что для подсчёта каннибализации подходят модели экспоненциальной регрессии. Они являются довольно простыми и легко интерпретируемыми, что обусловливает отсутствие ложных зависимостей в предсказаниях, которые возникают, так как спрос обладает высокой волатильностью. Логарифмирование данных о

продажах помогает стабилизовать дисперсию. К тому же модель экспоненциальной регрессии достаточно точна, так как каннибализация является эффектом кросс-эластичности, которая имеет экспоненциальный вид.

Однако в статье On product cannibalization. A new Lotka-Volterra model for asymmetric competition. (2016) рассказывается о нелинейной модели вида Lotka-Volterra. В данной работе не будет рассматриваться эта модель, так как она редко используется в экономике и более подходит для биологии. Модель Lotka-Volterra используется в этой области для исследований изменений поголовья хищников и их травоядных жертв в зависимости от различных условий. Для задачи ретейла такие условия трудно и неточно интерпретируются.

Открытого практического решения задачи нельзя найти, так как выполнение данной работы представляет из себя бизнес-задачу, решаемую каждой отдельной компанией, которой необходимо как можно более точно подсчитывать объем продаж.

2.1 Выводы

Экономические статьи приводят научное доказательство существования эффекта каннибализации. Менее научные источники (посты и презентации) позволяют понять декомпозицию спроса и применение подсчёта каннибализации, что формирует видение задачи. Источники советуют использование модели экспоненциальной регрессии. Открытого программного решения данной бизнес-задачи не существует.

3 Постановка задачи

3.1 Экономическая модель

В рамках данной работы для упрощения понимания товар А будет называться "каннибалом", если в ходе промо-акции он отнял продажи (или если

предполагается, что это может произойти), "отканнибализировал"их, у товара Б — "жертвы".

Эластичность спроса – динамика реагирования рынка на изменение предложения (вида продукта, его цены, комплекса услуг и т.д.). При проведении промо-акции для эластичных товаров при снижении цены растёт спрос. Но это не единственный эффект от промо-акции. Их можно декомпозировать по элементам, как показано на Рисунке 3.1

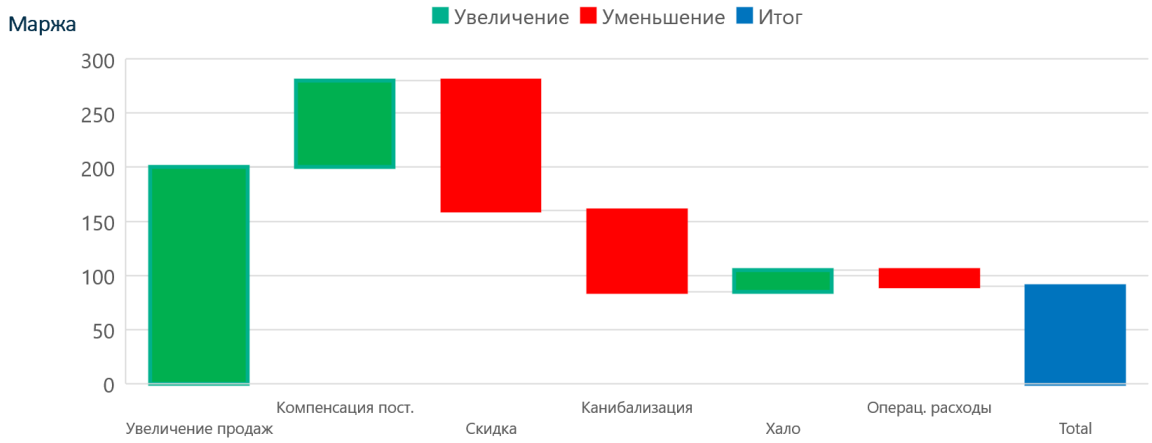

Рис. 3.1: Эффект от промо-акции, декомпозированный по элементам

Эффект каннибализации возникает между товарами-субститутами, для них можно посчитать коэффициенты кросс-эластичности по следующей формуле:

$$\mathcal{E}_c = \frac{\%q}{\%p_{\text{др.товар}}}$$

где $\%q$ — процентное изменение объёма продаж товара, $\%p_{\text{др.товар}}$ — процентное изменение цены на другой товар (товар-каннибал).

Таким образом, между товарами возникает каннибализация, если коэффициент кросс-эластичности положителен, так как и $\%q$, и $\%p_{\text{др.товар}}$ отрицательны (в связи с промо снизится цена каннибала и из-за этого снизятся продажи жертвы). Однако в предсказании продаж жертвы участвует большое число других факторов, так как спрос на товары сильно волатилен, поэтому для предсказания продаж необходимо использовать методы машинного обучения.

3.2 Модель машинного обучения

Исходя из экономической теории было принято решение не вычислять коэффициенты напрямую и находить по ним средний объём, на который уменьшатся продажи товара, а построить модель, позволяющую предсказывать продажи товаров с учётом эффекта каннибализации.

В данной работе были построены модели, предсказывающие продажи товара-жертвы. При выставлении промо-цены товара-каннибала, для товаров, которые могут являться его жертвами (определяются с помощью разбиения всех товаров на группы), будут производиться предсказания продаж с учётом новой цены каннибала. Если эти предсказанные продажи будут меньше, чем продажи предсказанные регулярным прогнозом продаж, будет считаться, что объём, на который предсказание модели с учётом каннибализации меньше регулярного прогноза, был отканнибализирован.В таком случае будет выдаваться рекомендация ставить товар-каннибал в промо одновременно с товаром жертвой.

В приведённых далее экспериментаж жертвой является фиксированный товар.

Данный метод не масштабируется: для каждого товара-жертвы нужна отдельная модель, так как для всех товаров подбираются различные каннибалы. В секции 7 будет описана идея модели, которая не вошла в рамки данной курсовой работы, но будет рассмотрена, как хорошо масштабируемая модель без экономической основы.

3.3 Выводы

Эксперименты описанные далее будут проводиться для предсказания продаж выбранного товара-жертвы. В основе идей для моделей лежит экономическая формула кросс-эластичности.

4 Работа с данными

4.1 Подготовка данных

С начала исследования было ясно, что необходимо строить отдельные модели для разных групп товаров так, чтобы внутри одной группы находились только товары субституты, так как эффект каннибализации основан на поведении товаров-заменителей. Однако было необходимо выбрать, каким образом их подбирать: есть несколько различных уровней категорий товаров, которые постепенно сужают их число. Например, на высоком уровне могут быть категории "Масла"и "Свежие продукты", а на наиболее глубоком может быть только "Ультрапастеризованное коровье молоко"с определённой жирностью и объёмом. Изначально для проведения экспериментов внутри одной высокой группы выбирались товары с наибольшими продажами. Но, как уже упоминалось ранее большие скидки на элитные товары могут забирать продажи не только у своего аналога, но и у более дешёвых товаров данного сегмена(McKinsey Digital. Эффективное промо: разобраться и перенастроить. (2019).). А товары с большой разницей в цене обычно имеют значительный разрыв в объёме продаж. Для дальшейших исследований был выбран один из наиболее глубоких уровней категорий, так как именно на нём находятся товары-аналоги. Эксперименты проводились на различных группах товаров: 5 товаров из одной категории молока, 10 товаров из категории лимонадов, 17 из категории шоколада. В данной работе представлены эксперименты на группе молока (5 товаров), так как на ней удобнее приводить результаты и отслеживать их изменение. Для всех групп различных размеров наилучший результат был получен в одинаковом эксперименте (6.2).

Для дата-сета берутся данные с сентября 2018 года по начало мая 2021 (за последние два года и восемь месяцев), так как за этот период они являются полными и достоверными. Из этого периода исключаются данные, когда товары находились в статусе новинок или были выведены из ассортимента. Также исключаются периоды с начала марта до конца апреля 2020 года и

промежуток 28 декабря - 4 января каждого года (в этот период наблюдается аномально резкий прирост продаж). Учитываются только не отменённые заказы.

Данные о продажах необходимо было очистить от аномалий. Ими считались заказы с большим объёмом какого-либо товара (а не большой объём продаж продукта в день, так как данные будут очищены от тренда и сезонности). Для каждого продукта был подсчитан 98 квантиль его объёма в заказах. Все данные, превышающие это значение, были заменены на него.

Для товаров известно их наличие на складе. Оно измеряется от 0 до 1, как доля времени суток, когда товар находился на складе. Принято считать, что, продажи товара являются не достоверными, если товар не был в наличие больше 2.5 часов в сутки (меньше 0.9 в долях). Для товаров существует подсчитанный восстановленный спрос — спрос, очищенный от "плохого"наличия и аномально крупных заказов. В дни, когда наличие товара на складе было недостаточным, продажи товара были заменены на востановленные. Данные были очищены от тренда и сезонности, чтобы модель могла обучаться на периодах в несколько лет и делать предсказания для любого последующего промежутка времени, проще говоря для стационарности временного ряда.

4.2 Разведочный анализ данных

Для примеров экспериментов выбрана группа с пятью товарами. Суммарное число дней товаров из группы в промо за весь период : 801 Суммарное число различных(уникальных) дней : 898 Длина датасета (суммарное число дней для всех товаров) : 3555. Больше информации о датасете можно увидель из графиков 4.1, 4.2, 4.3 и таблицы 4.1.

Таблица 4.1: Разведочный анализ данных

Good	Number of days in sale	Number of days with zero sales	Number of days in promo
Victim	727	16	276
Good 1	609	6	139
Good 2	877	33	87
Good 3	888	52	159
Good 4	454	108	140

Number of goods in group	Total number of days (Length of dataset)	Number of days in promo	Number of different days
5	3555	801	898

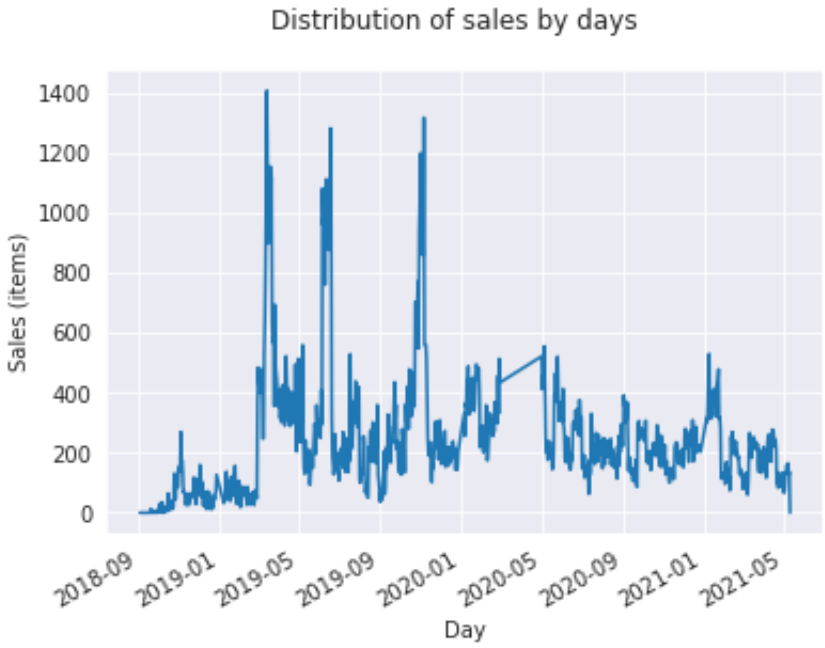

Рис. 4.1: Распределение продаж в группе по дням

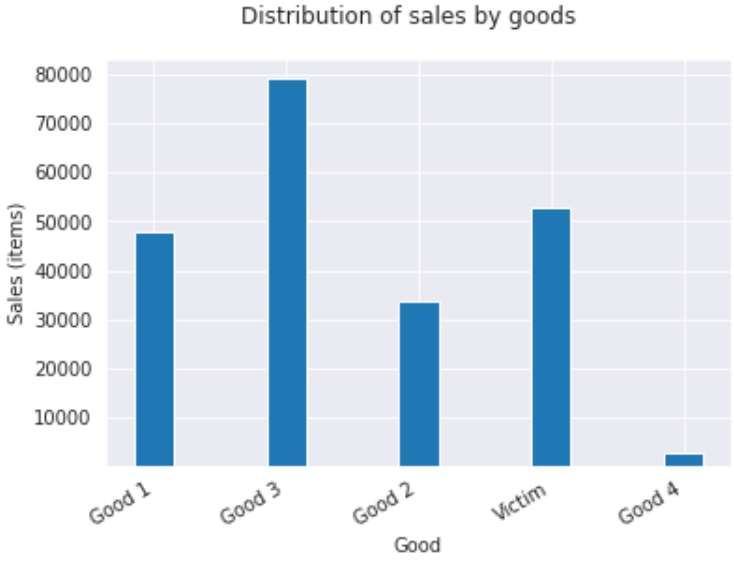

Рис. 4.2: Распределение суммарных продаж по товарам

Рис. 4.3: Распределение среднего цен по товарам

5 Описание первого этапа экспериментов

5.1 Первый эксперимент

Для начала были построены несколько простых моделей для пары товаров: модель предсказывала продажи жертвы в зависимости от цены самой жертвы и каннибала (то есть модель имела всего два признака). Идея заключается в том, что такие модели делают предсказание на основе цен товаров и предобученного коэффициента кросс-эластичности, как и предполагает экономическая модель. Для таких данных были выбраны простые и интерпретируемые модели: Curve_fit, линейная регрессия, (нелинейная) обобщенная аддитивная модель (GAM). И для сравнения был обучен градиентный бустинг.

Из 2 был сделан вывод о необходимости использования экспоненциальной регрессии, в связи с этим целевая переменная логарифмируется перед обучением модели. Обобщенная аддитивная модель была выбрана за свою простоту и интерпретируемость. Понять, что такое GAM, проще отталкиваясь от того, что все линейные модели входят в "класс" обобщённых аддитивных моделей. Линейная модель обучается давать признаку некоторый константный вес. Её

формулу можно представить в следующем виде:

$$a(x) = \sum_{i=1}^{l} f_i(x_i) + bias, \quad f_i(x_i) = c_i x_i$$

где $a(x)$ — предсказание модели, l — число признаков объекта, x_i — значение i-го признака объекта, c_i — вес i-го признака, $bias$ - свободный коэффициент

А предсказание обобщённых аддитивных моделей можно представить по формуле:

$$a(x) = \sum_{i=1}^{l} f_i(x_i) + bias, \quad f_i(x_i) = \sum_{j=1}^{n\_splines} g_j(x_i)$$

где $a(x)$ — предсказание модели, l — число признаков объекта, x_i — значение i-го признака объекта, $bias$ - свободный коэффициент, g_j — сплайн с обученными коэффициентами.

Таким образом, вес каждого признака описывается суммой значений функций от значения признака. Эти функции называются сплайнами. Для GAM сплайнами являются функции, отличные от нуля на небольшом отрезке. На примере из pyGAM, balancing predictive power and interpretability with generalized additive models (2018)

На рисунке 5.1 полупрозрачными цветами показаны сплайны для некоторого признака, а чёрным — их сумма, то есть функция веса признака в зависимости от его значения.

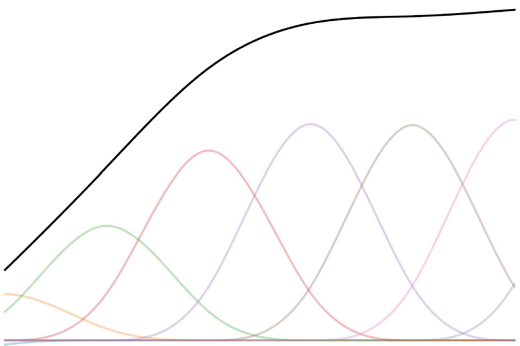

Рис. 5.1: Обученная функция веса для некоторого признака GAM

Curve_fit является моделью в библиотеке scipy.optimize, которая позволя-

ет обучать коэффициенты в заранее заданной функции. Для задачи предсказания продаж с учётом каннибализации использовалась следующая функция, исходя из экономической модели:

$$Q_a = \exp(\gamma + \alpha p_a + \beta p_b)$$

где α, β, γ — обучаемые коэффициенты, p_a — цена жертвы, p_b — цена каннибала, Q_a — продажи жертвы.

Для задачи были выбраны следующие метрики и ошибки: средняя квадратическая ошибка (MSE), коэффициент детерминации (R^2), взвешенная абсолютная процентная ошибка (WAPE).

Ошибка WAPE имеет следующую формулу:

$$WAPE = \frac{\sum_{i=1}^{n} |a(x_i) - y_i|}{\sum_{i=1}^{n} y_i}$$

где y_i - фактическое значение объёма продаж (целевое значение для объекта x_i), $a(x_i)$ - предсказанное значение, x_i - объект, для которого делается предсказание (в данном эксперименте цена жертвы и каннибала).

Эта ошибка часто используется для прогноза продаж, так как она является симметричной и наименее чувствительна к искажениям числового ряда.

Молоко является сильно эластичным товаром, поэтому его продажи резко возрастают во время промо-акции, что можно видеть на рисунке 5.2. Выбранные модели умеют хорошо отлавливать такое изменение в продажах. Результаты предсказаний имеют ступенчатый вид, поскольку модель зависит всего от двух признаков, которые редко изменяются. Если посмотреть на веса признаков линейной регрессии и curve_fit из таблицы 5.1 можно заметить, что даже в хорошем случае подбора каннибала и жертвы (каким данный пример и является), вес цены каннибала очень мал по сравнению с ценой жертвы. Аналогичное можно увидеть из графика 5.3 изменения веса признака в зависимости от его значения для GAM.

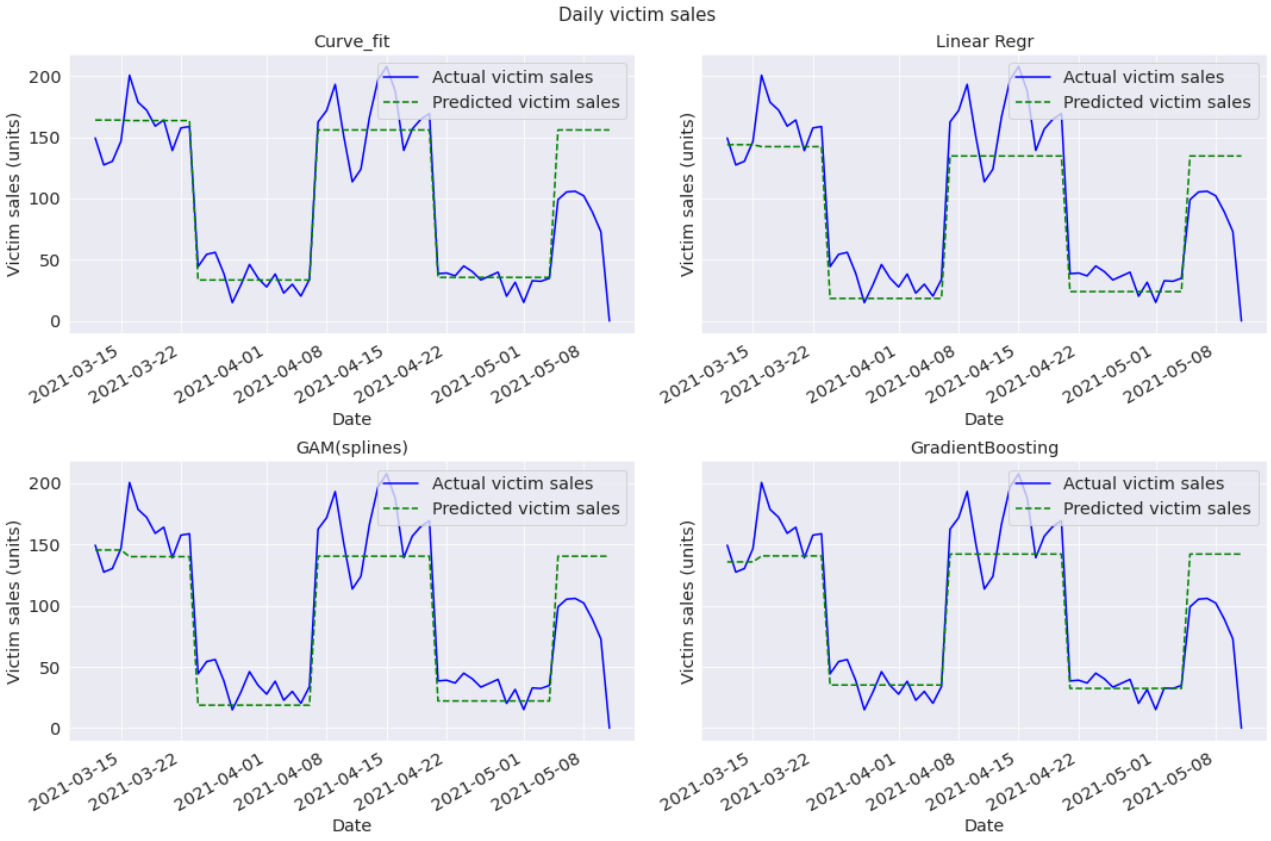

Рис. 5.2: Предсказания ежедневных продаж для данных с двумя признаками

Таблица 5.1: Веса признаков для моделей ежедневных продаж

Model	Victim price weight	Cannibal price weight	Bias
Linear Regression	-0.1098	0.0108	10.7469
Curve_fit	-9.5906	2.6756	10.7910

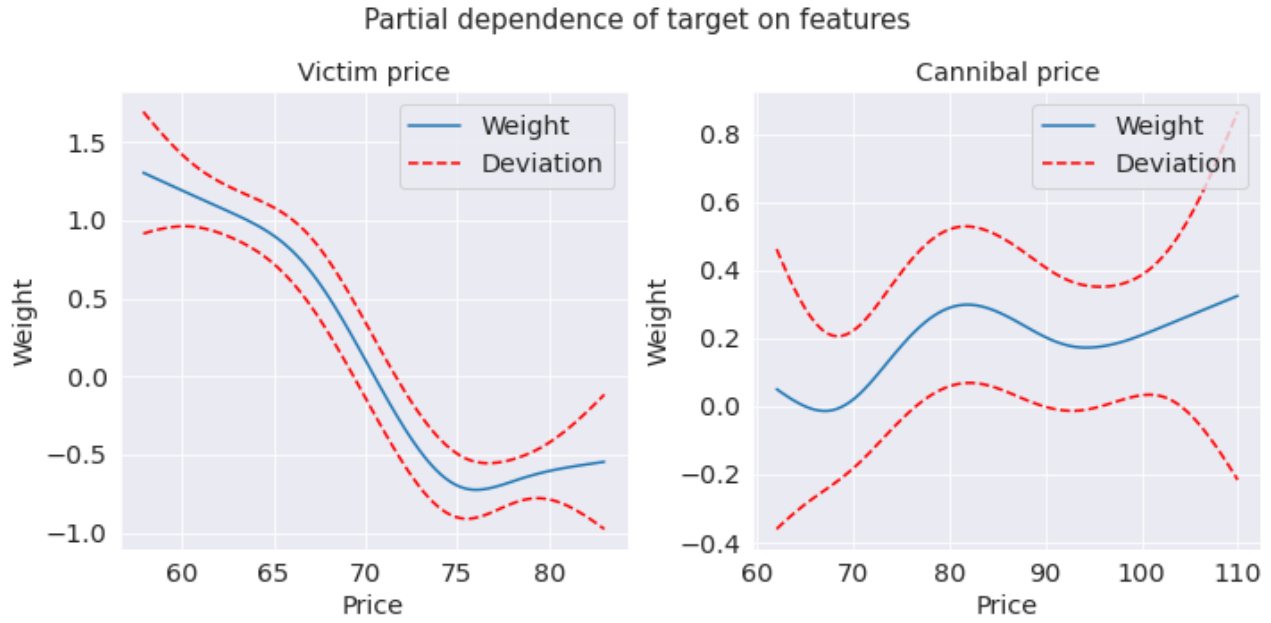

Рис. 5.3: Зависимость веса признака от его значения (GAM)

Также можно посмотреть на предсказания на данных, в которых из каждой недели был оставлен только один медианный день (рисунок 5.4). Такой подход сглаживает целевую переменную и позволяет делать более точное предсказание, однако, из-за сокращения данных модели могут недообучаться. Можно посмотреть таблицу результатов 5.2: ошибки при втором подходе заметно уменьшаются. В обоих случаях лучшей моделью оказался градиентный бустинг, худшей — GAM.

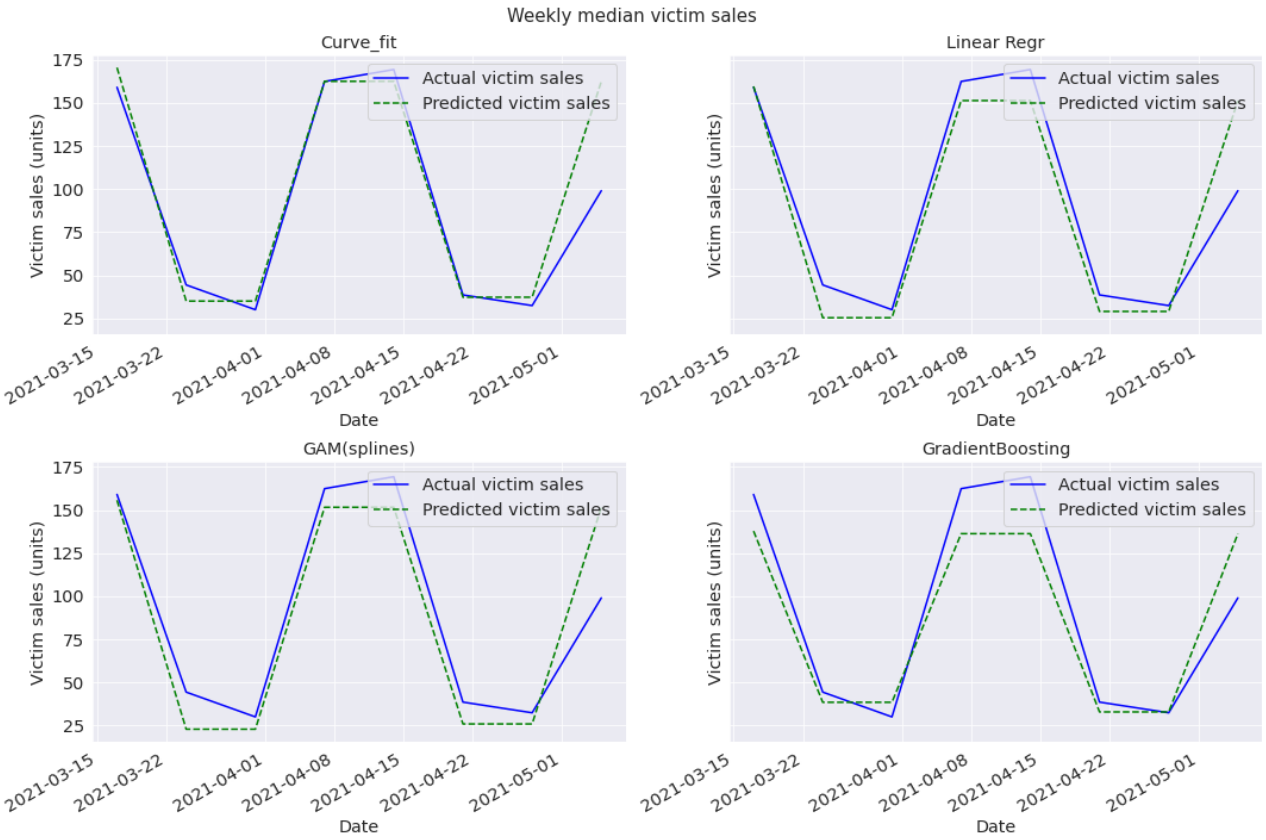

Рис. 5.4: Предсказания медианных продаж за неделю для данных с двумя признаками

Таблица 5.2: Ошибки моделей с двумя признаками

Model	MSE	R^2	WAPE	Sales type
Linear Regression	1073.9703	0.7300	0.2607	daily
Curve_fit	1074.5068	0.7299	0.2171	daily
GAM(splines)	1097.8482	0.7240	0.2628	daily
GradientBoosting	1003.0871	0.7478	0.2265	daily
Linear Regression	447.1252	0.8699	0.1828	weekly median
Curve_fit	416.2168	0.8789	0.1338	weekly median
GAM(splines)	449.3272	0.8693	0.1728	weekly median
GradientBoosting	388.0007	0.8871	0.1454	weekly median

5.2 Второй эксперимент

Следующим шагом стало добавление к имеющимся двум признакам цены остальных товаров из групп. Это обусловлено тем, что у одного товара может быть несколько каннибалов, что может позволить усложнить известную на теорию и найти зависимости, делающие прогнооз полее точным. Градиентный бустинг был заменён на XGBoost, так как он показал более хорошие результаты. В рамках данной задачи линейную регрессию можно сравнить с curve_fit, так как они отличаются тем, что в регрессии целевой переменной являются логарифмированные продажи, а в curve_fit формула регрессии стоит в показателе экспоненты. На данных с ценами других товаров из группы в качестве признаков линейная регрессия оказалась лучше, чем curve_fit, поэтому от второй модели отказались. Нелинейная обобщённая аддитивная модель в дальнейшем тоже не рассматривается. Были добавлены модели линейной регрессии с регуляризацией, причина этого поясняется результатами обучения моделей.

Однако появилась проблема, связанная с данными: поскольку не для всех товаров есть данные за три года (например, товар ещё не ввели в ассортимент), в случае, когда все цены товаров являются признаками, остаются данные о продажах только за общие известные дни, что может в некоторых случаях сократить объём данных. С такими случаями тоже необходимо работать, так как итоговую модель необходимо будет применять ко всем группам товаров.

Результаты предсказаний можно увидеть на графиках 5.5 и 5.6. Из ошибок моделей в таблице 5.3 видно, что для ежедневных предсказаний линейной регрессии качество ухудшилось с добавлением многих цен каннибалов. Это происходит из-за того, что продажи жертвы не сильно от них зависят, модели ищут ложные зависимости и переобучаются в смысле ошибок Train-Test выборок (роста весов моделей регрессии не наблюдается). Для борьбы с этим была добавлена регуляризация, такие модели показали несколько более хо-

рошие результаты, по сравнению с линейной регрессией в этом эксперименте, но примерно в три раза более плохие (по MSE), чем регрессия в предыдущем эксперименте.

Если посмотреть на веса моделей из таблицы 5.4, можно увидеть, что они и так очень малы, поэтому l2-регуляризация не дала сильного улучшения — модель всё ещё находит ложные зависимости, в отличие от l1-регуляризации, которая обратила в ноль некоторые веса. Тем не менее, модели показали плохие результаты. Если посмотреть на медианные подажи в неделю, можно заметить, что линейная регрессия улучшила свой результат даже по сравнению с предыдущим экспериментом, а модели с регуляризацией сравнимы по ошибкам с ошибками линейной регресси в предыдущем эксперименте.

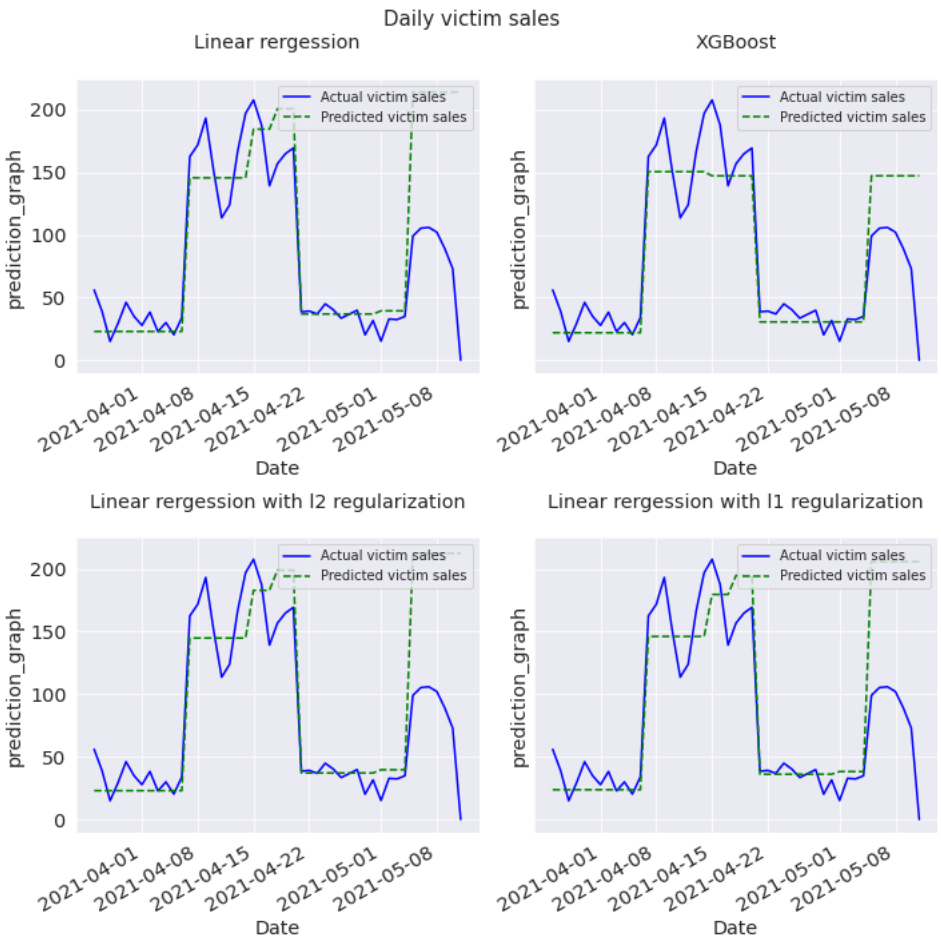

Рис. 5.5: Ежедневные предсказания для данных с признаками многих цен

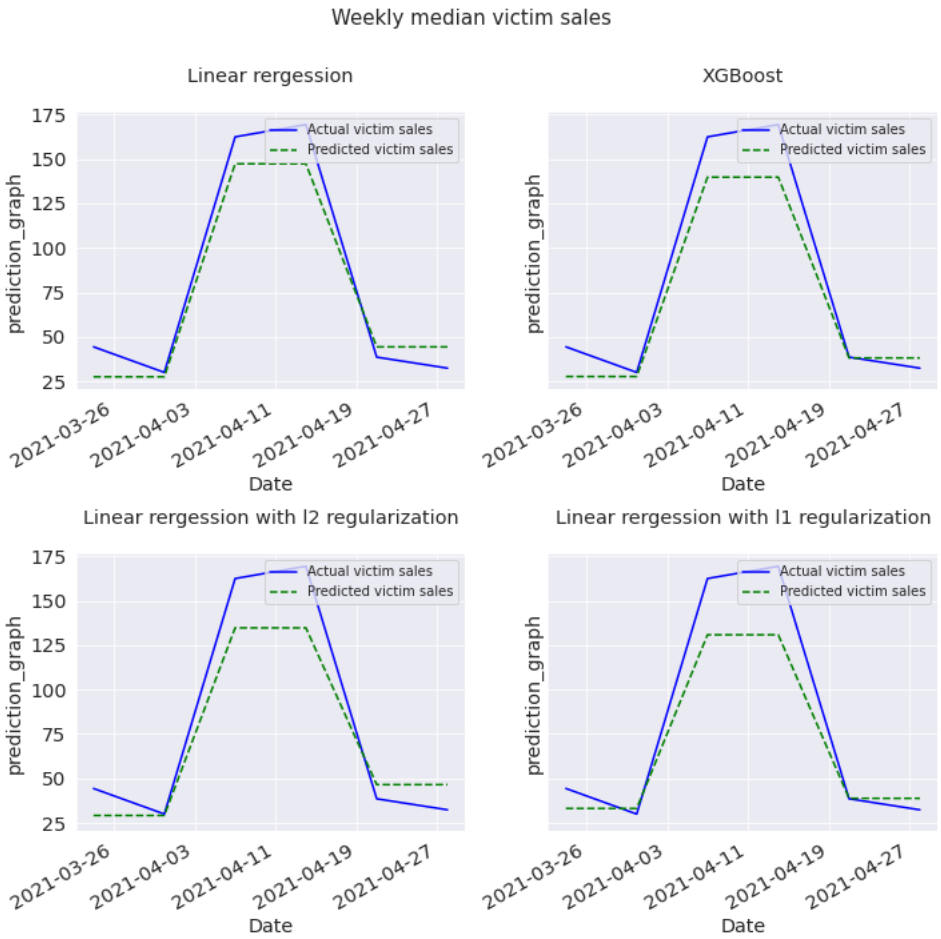

Рис. 5.6: Еженедельные предсказания для данных с признаками многих цен

5.3 Выводы

Эксперименты помогли понять, что не все товары даже внутри узкой группы являются одновременно каннибалами друг друга, основной вес приходится на пару товаров. Было выдвинуто следующее предположение: с помощью линейной регрессии с l1-регуляризацией находить возможных каннибалов на валидационной выборке (которая не участвует в обучении моде-

Таблица 5.3: Ошибки моделей с признаками многих цен

Model	MSE	R^2	WAPE	Sales type
Linear Regression	3164.6141	0.1591	0.4145	daily
LR l1 regularization	2815.2717	0.2520	0.393	daily
LR l2 regularization	3099.0021	0.1766	0.4116	daily
XGBoost	1176.4135	0.6874	0.2841	daily
Linear Regression	196.7889	0.9476	0.1555	weekly median
LR l1 regularization	443.6541	0.8818	0.1912	weekly median
LR l2 regularization	410.8279	0.8905	0.2106	weekly median
XGBoost	283.7340	0.9244	0.1621	weekly median

Таблица 5.4: Веса признаков для моделей ежедневных продаж

	Feature weight					
Model	Victim price	Good 1	Good 2	Good 3	Good 4	Bias
LR	-0.1085	0.0072	0.0372	0.0092	-0.0036	7.5933
LR l1	-0.1078	0.0051	0.0325	0.0078	-0.	7.9627
LR l2	-0.1075	0.0074	0.0368	0.0092	-0.0035	7.5133

лей) и обучать итоговую модель, где ценовыми признаками будут только эти товары-каннибалы.

6 Описание второго этапа экспериментов

6.1 Изменения в витрине данных и описание модли

На втором этапе экспериментов к данным добавили много признаков, была построена полноценная модель. Итоговым стал следующий список признаков:

- Промо цена жертвы (числовой признак)
- Регулярная цена жертвы (числовой признак)
- Доля скидки (числовой признак)
- Номер дня в неделе (категориальный признак)
- Номер недели в году (категориальный признак)
- Номер дня промо-акции жертвы (поскольку модель будет предсказывать её продажи). В случае отстутствия промо признак имеет значение 0 (категориальный признак)
- Суммарное число заказов в компании день. Существует модель, обученная предсказывать это значение, оно и используется, как признак (числовой признак)
- Цены других товаров в группе (числовой признак)

Можно заметить, что в списке нет признаков, описывающих характеристики товаров. Признаки выбраны так, потому что модель будет строиться отдельно для каждой группы, внутри которой товары только незначительно отличаются, следовательно, такие признаки были бы для них одинаковыми и константными.

Для подготовки данных применялось One Hot Encoding (для категориальных признаков) и масштабирование (для числовых).

6.2 Третий эксперимент

На данном этапе экспериментов рассматриваются следующие модели: линейная регрессия (LR), метод опорных векторов для задачи регрессии (SVM), линейная регрессия с l1-регуляризацией, линейная регрессия с l2-регуляризацией, случайный лес (Random Forest), XGBoost.

Для модели SVM верно, что, если инвертировать цель модели классификации, можно получить решение задачи регрессии. SVM для бинарной классификации находит наибольшую возможную полосу, которая разделяет два класса и ограничивает нарушителей. Метод опорных векторов в задаче регрессии старается уместить наибольшее возможное количество объектов на полосе с ограничением нарушителей. По нахождению объекта на построенной полосе делает предсказание о нём.

Результаты обучения показаны на графике 6.1. Выбранные модели снова переобучаются в смысле ошибок Train-Test выборок (без увеличения весов моделей линейной регрессии), ошибки в таблице 6.1.

В случае если модели обучить без признаков цен каннибалов, результат незначительно улучшается. Если же для выбранной жертвы подобрать каннибала по валидационной выборке, результат станет заметно лучше 6.2: модели меньше переобучаются, результат на тестовой выборке улучшился для нескольких моделей. Наилучшей оказался XGBoost. Если рассматривать модели с медианными продажами в неделю, результаты и на обучающей вы-

борке плохи (и для случая с одним и четырьмя каннибалами) из-за того, что модели остаются недообученными.

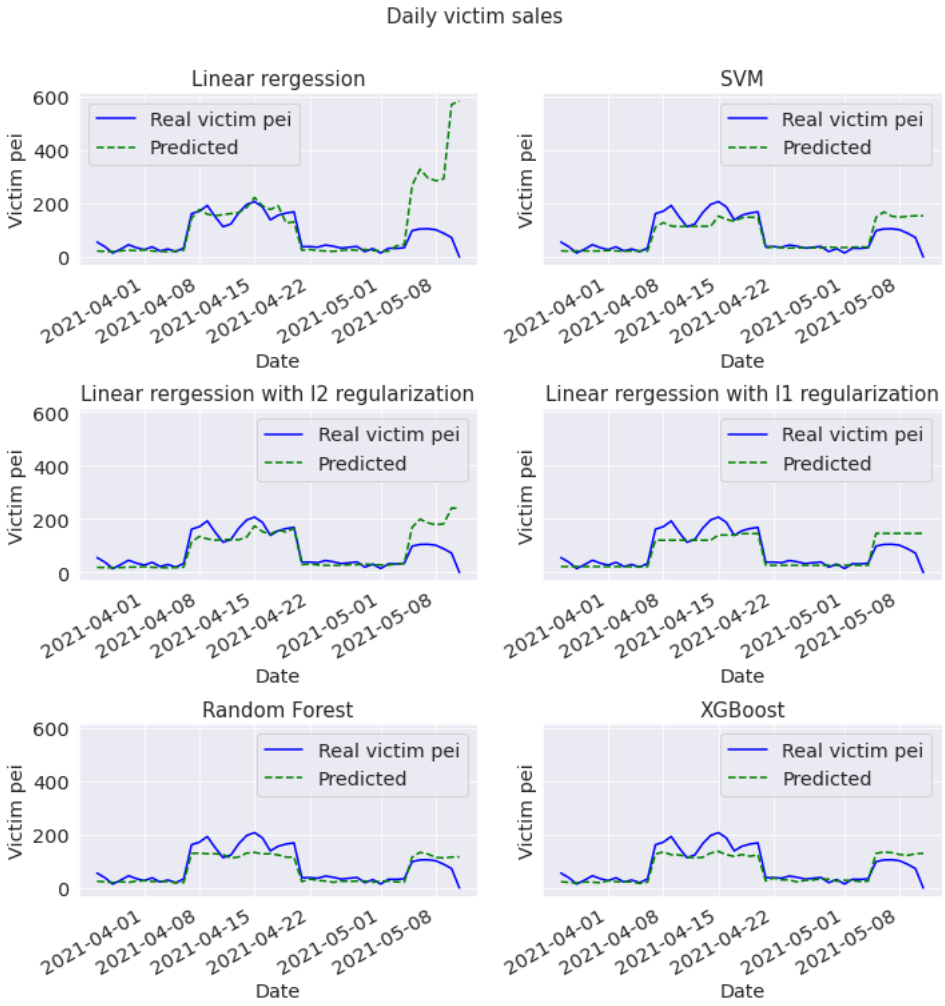

Рис. 6.1: Ежедневные предсказания для данных со многими признаками

6.3 Интерпретация моделей

Для интерпретации моделей в данной работе использовалась библиотека SHAP и explainerdashboard. Все приведённые в этой секции графики строились для модели XGBoost, так как она получила наилучшие результаты. Также рассматриваются данные с 4 каннибалами, чтобы можно было проследить разницу в их влиянии на предсказания продаж жертвы.

На графике 6.2 показано влияние значения отдельных признаков в сравнении с их "базовыми"значениями на конкретный объект из тестовой выборки для модели XGBoost с 4 каннибалами. По координатной оси отложена экспонента предсказания модели. Она равна 52.48 (результат выделен жирным

Таблица 6.1: Ошибки моделей со многими признаками и 4 каннибалами

	Train			Test			
Model	MSE	R^2	WAPE	MSE	R^2	WAPE	Sales type
Linear Regression	629.5178	0.8273	0.2338	16921.4739	-3.4961	0.7136	daily
SVM	1322.2935	0.6486	0.2932	1618.6506	0.5699	0.3298	daily
LR l1	1301.9925	0.6540	0.3074	1446.5270	0.6156	0.3260	daily
LR l2	742.3750	0.8027	0.2435	3049.7576	0.1896	0.3982	daily
Random Forest	563.6302	0.8502	0.2121	1104.6321	0.7064	0.2986	daily
XGBoost	117.0568	0.9688	0.0962	998.2195	0.7347	0.2611	daily
Linear Regression	0.9589	1.00	8.99e-16	1.83e+02	-16.1025	4.09e-01	weekly median
SVM	1.60e+04	-0.1258	7.03e-01	5.32e+02	-94	6.97e-01	weekly median
LR l1	4.56e+03	0.6792	3.52e-01	4.78e+03	-10	2.08e+00	weekly median
LR l2	8.81e-01	0.9999	4.76e-03	1.80e+02	-15.7	4.06e-01	weekly median
Random Forest	1.68e+04	-0.1834	7.89e-01	1.94e+03	-271	1.33e+00	weekly median
XGBoost	6.22e+03	0.5626	3.59e-01	4.73e+02	-88	6.57e-01	weekly median

Таблица 6.2: Ошибки моделей со многими признаками и одним каннибалом

	Train			Test			
Model	MSE	R^2	WAPE	MSE	R^2	WAPE	Sales type
Linear Regression	1184.5885	0.8200	0.3109	3434.7575	0.1222	0.4895	daily
SVM	1712.5749	0.7399	0.3286	1232.6494	0.6849	0.2791	daily
LR l1	1934.1278	0.7062	0.3628	1326.6924	0.6609	0.3047	daily
LR l2	1490.5406	0.7736	0.3420	1977.4176	0.4946	0.3747	daily
Random Forest	762.8071	0.8841	0.2284	1040.5788	0.7340	0.2628	daily
XGBoost	545.4042	0.9171	0.2071	948.6566	0.7619	0.2502	daily

шрифтом). Некоторое "базовое"предсказание (которое сделано, когда признаки принимают свои средние или наиболее частые значения) подписано серым над координатной осью - *base value*. Значения под красными и синими отрезками являются признаками в формате "{Название признака} = {Значение, которое оно принимает}". Признаки на красном и синем отрезке увеличивают и уменьшают значение предсказания соответственно. На сколько предсказание изменяется из-за конкретного признака можно увидеть из длины отрезка, отвечающего за выбранный признак.

Слева направо можно увидеть следующие признаки:

- $week\_num\_18 = 1.0$ - номер недели. С помощью использования OHE признак стал бинарным для всех возможных значений (всего 52)
- $Good1 = 0.584$ - отмасштабированная цена первого каннибала
- $PROMO = -1.39$ - отмасштабированная промо-цена жертвы (в случае отсутствия промо равно регулярной)
- $discount\_rate = 1.285$ -отмасштабированное значение доли скидки (до масштабирования было меньше единицы)
- $promo\_day\_0 = 0.0$ - нулевой номер дня акции (в этот день была промо-акция)
- $orders = -0.741$ - отмасштабированное количество заказов в конкретный день. Заказов было меньше, чем обычно.
- $Good2 = 0.269$ - отмасштабированная цена второго каннибала.

Остальные признаки внесли куда более маленький вклад.

Можно сделать вывод, что сильнее всего на предсказание влияет бинарный признак "Продаётся ли товар по регулярной цене". В данный день товар стоял в промо, из-за этого его продажи увеличились по значению данного признака. Интересно также то, что на предсказание повлиял номер недели, в это время были майские праздники.

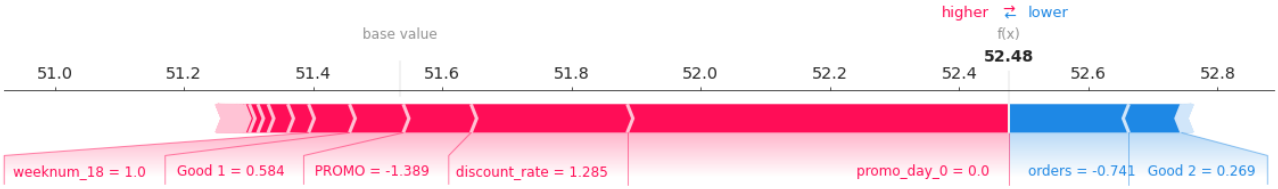

Рис. 6.2: Влияние признаков на предсказание одного объекта

График 6.3 показывает, как повлияли значения отдельных признаков в сравнении с их "базовыми"значениями на всей тестовой выборке. По оси X показан номер объекта выборки, y - экспонента значения предсказания (то есть объём продаж в штуках). На html-странице при выборе курсором объекта становятся видны значения основных синих и красных признаков. Так же как и на предыдущем гафике наибольшее значение имеет признак "Продавался ли товар по регулярной цене". Если просмотреть несколько объектов, можно увидеть, что для разных дней наибольший вклад вносили разные признаки. Для сраврения представлен график 6.4.

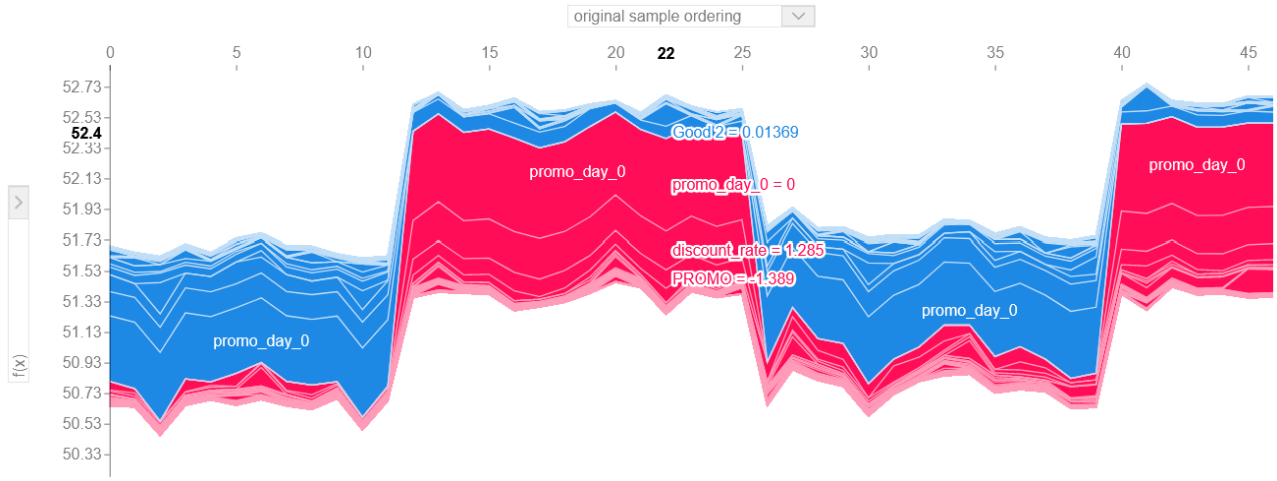

Рис. 6.3: Влияние признаков на предсказания всей тестовой выборки

На графике 6.5 можно увидеть, как сильно признаки влияют на предсказание (в долях). При этом значения бинарных признаков, принадлежащих одному категориальному, объединяются. Сильнее всего влияют признаки:

- *promo_day* - номер дня в промо-акции. Признак так много значит, скорее всего, из-за того, что при one hot encoding появляется бинарный

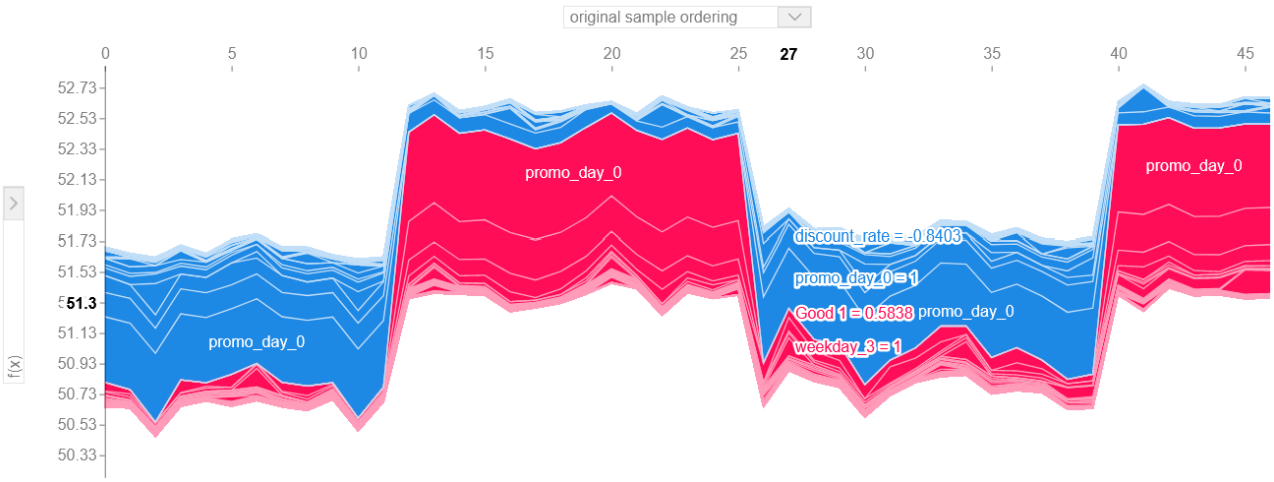

Рис. 6.4: Влияние признаков на предсказания всей тестовой выборки

признак "Продаётся ли товар по регулярной цене", так как дням, когда акции отсутствуют, присвоен отдельный номер (0).

- *discount_rate* - доля скидки (судя по всему, модель обучается различным предсказаниям при различных скидках - маленьких и больших,нулевых)
- *Good* 2 - цена второго каннибала. Вероятно, этот товар действительно способен каннибализировать продажи у жертвы.
- *PROMO* - собственная цена жертвы.

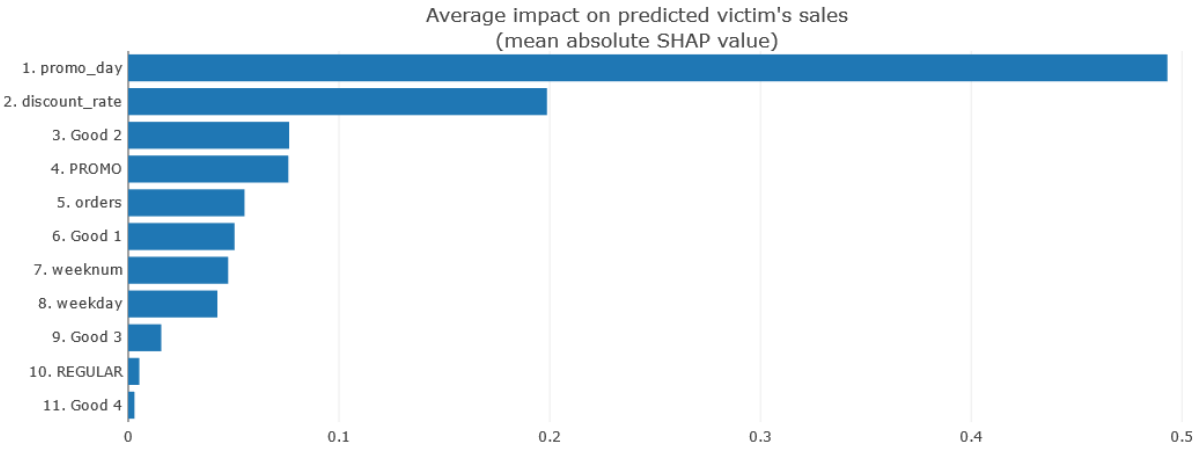

Рис. 6.5: Влияние признаков на предсказание в долях

График 6.6 даёт более детальные объяснения, как и какие значения признаков влияют на прогноз. Точки на графике представляют собой, какие зна-

чения принимает признак. Они имеют три характеристики:

1 значение по оси y - показывает, к какому признаку точки относятся

2 значение по оси x - показывает, как признак повлиял на модель (на сколько он уменьшил или увеличил предсказание)

3 цвет - значение признака, голубой цвет означает, что признак принял наименьшее значение, розовый - наибольшее

Из этого графика можно вынести, что

- Если модель имеет нулевой промо-день, предсказание будет уменьшено (из-за отсутствия промо-акции), и наоборот
- Скидка (которая принимала примерно одинаковое значение для выбранного товара) увеличивает продажи
- Низкая цена второго товара уменьшает продажи, а высокая увеличивает. Подтверждается предположение о том, что второй товар является каннибалом.
- На 18 и 19 неделе в годах (майские праздники) растут продажи
- Продажи жертвы растут во вторник и среду (можно предположить, из-за того, что в эти дни обычно начинаются промо-акции)

Второй товар является потенциальным каннибалом. Для проверки существования каннибализации на практике были выделенны некоторые условия, при одновременном выполнении которых, подтверждается существования каннибализации (условия выдвигались и изменялись во время работы над данной задачей):

1 Медианные продажи жертвы в течение промо ниже медианы продаж за 2 недели до промо

2 Товар-каннибал продаётся по промо-цене

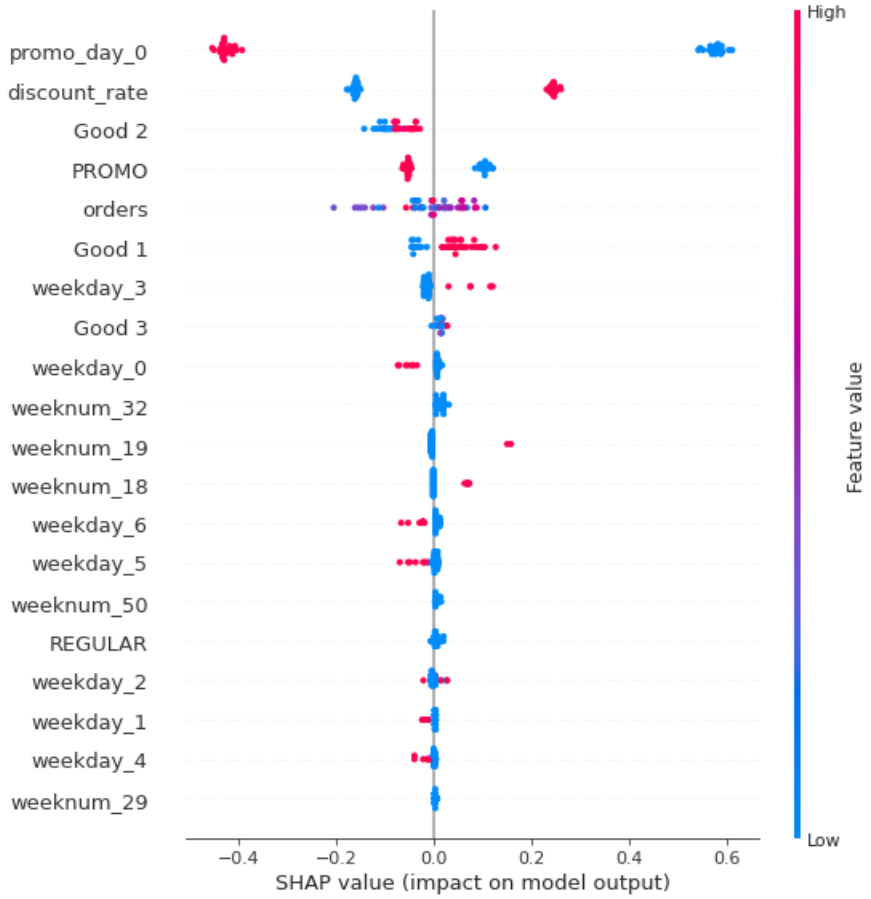

Рис. 6.6: Влияние признаков в зависимости от их значения

3 Цена жертвы регулярна и константа в период промо-акции каннибала и за неделю до неё

И действительно, для данной пары товаров существуют такие периоды: рисунок 6.7. На низнем графике представлено изменение цен жертвы и каннибала. На верхнем показаны ежедневные и медианные за неделю продажи каннибала и жертвы.

6.4 Выводы

По данной работе необходимо проводить дальнейшие исследования, так как модели подвержены переобучению. Для товара-жертвы из приведённых экспериментов найден как минимум один каннибал.

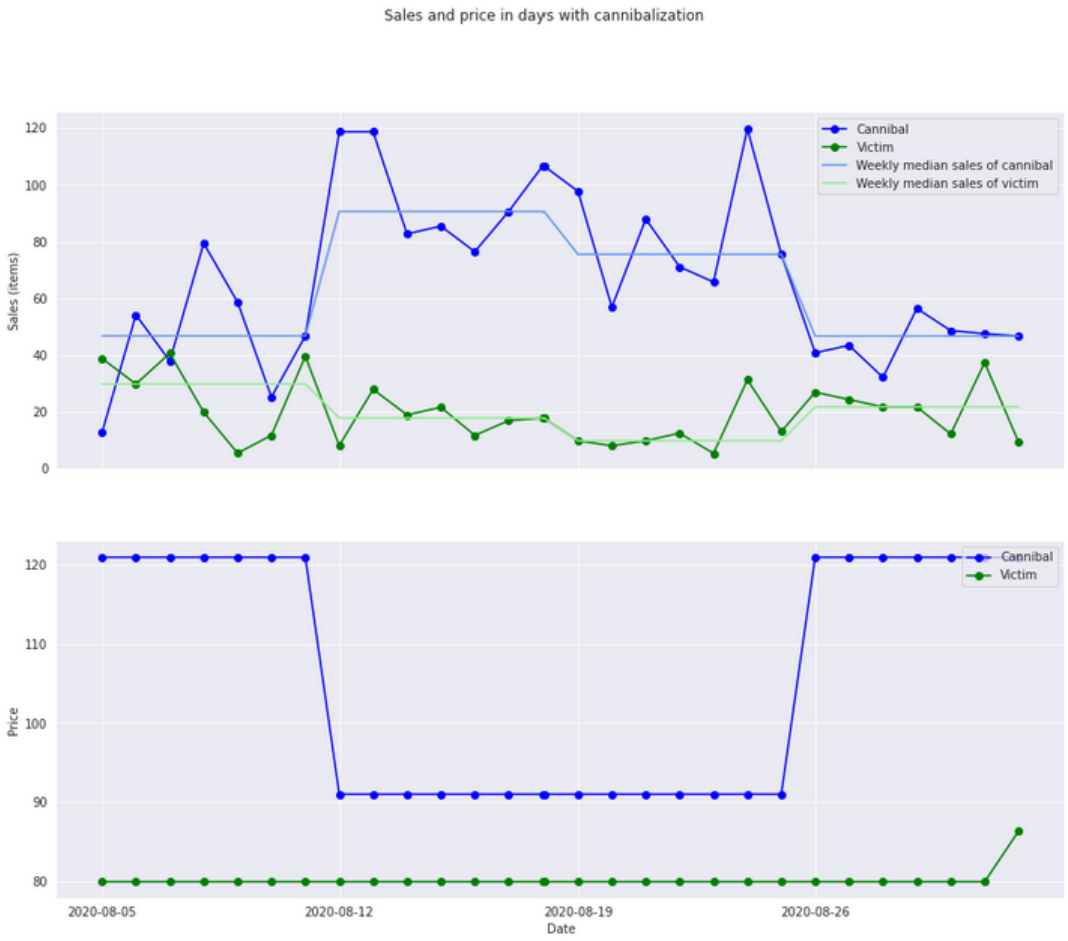

Рис. 6.7: Пример эффекта каннибализации

7 Дальнейшие исследования

Так как данная модель не масштабируется и показывает недостаточно хорошие результаты была предложена следующая идея модели.

Каннибалом будет считаться любой товар, стоящий в промо-акции. Жертвой - любой товар, продающийся по регулярной цене. Таким образом для каждого каннибала жертвами по очереди будут разные товары (аналогично можно смотреть на ситуацию со стороны жертвы). Для каждой такой пары в каждый день, удовлетворяющий условию (каннибал в промо, жертва - нет), данные будут занесены в дата-сет в следующем виде:

- Цена каннибала
- Цена жертвы
- Общие характеристики (например, день недели, неделя года)
- Различне характеристики (например, день промо-акции (для каннибала не равен нулю, для жертвы равен), срок хранения товара)

Таким образом, количество данных сильно увеличится, у модели будет гораздо больше примеров для находжения каннибализации. Такая модель может хорошо масштабироваться: можно обучить одну модель для всех товаров. Однако в таком случае, могут возникнуть проблемы с описанием различных характеристик товаров из разных групп (например, для товаров из категории фруктов нет данных по значению характеристики жирность). Для решения этой проблемы можно несколько сузить категорию и обучить больше моделей.

8 Выводы

В данной работе были исследованы различные подходы построения модели машинного обучения для подсчёта объёма продаж с учётом эффекта

каннибализации, в том числе модели, основанные на экономической теории каннибализации. Результаты были измерены и проанализированы. Была проведена интерпретация моделей, обозначены дальнейшие направления для исследования. В данной работе было найдено практическое подтверждение эффекта каннибализации на данных компании "Утконос".

Построенные модели оказались сильно подвержены переобучению, но последний эксперимент показал наиболее хороший результат и наименьшую степень переобучения в моделей. Были найдены неочевидные и значимие зависимости продаж товаров от многих признаков.

Список литературы (или источников)

1. Meredith L., Maki D. Product cannibalization and the role of prices //Applied Economics. – 2001. – Т. 33. – №. 14. – С. 1785-1793.

2. Guidolin M., Guseo R. On product cannibalization. A new Lotka-Volterra model for asymmetric competition in the ICTs. – 2016.

3. Haynes M., Thompson S., Wright P. W. New model introductions, cannibalization and market stealing: evidence from shopbot data //The Manchester School. – 2014. – Т. 82. – №. 4. – С. 385-408.

4. Бодунов, Дмитрий. Математика цен. Эластичность и ценовые эффекты. [Электронный ресурс] / SAS Institute Inc. – Электрон. текстовые дан. – Москва: [б.и.], 2017. – Режим доступа: https://docplayer.ru/61265761-Matematika-cen-elastichnost-i-cenovye-effekty-keysy-bodunov-dmitriy-copyright-sas-institute-inc-all-rights-reserved.html, свободный.

5. McKinsey Digital. Эффективное промо: разобраться и перенастроить. [Электронный ресурс] / McKinsey Digital. – Электрон. текстовые дан. – Москва: [б.и.], 2019. – Режим доступа: https://vc.ru/mckinsey/69835-effektivnoe-promo-razobratsya-i-perenastroit, свободный.

6. Власова, Валентина. Анализ промо-акций. [Электронный ресурс] / SAS Institute Inc. – Электрон. текстовые дан. – Москва: [б.и.], 2015. – Режим доступа: https://docplayer.ru/38861332-Analiz-promo-akciy-valentina-vlasova-vedushchiy-konsultant-v-oblasti-resheniy-dlya-roznichnogo-biznesa.html, свободный.

7. Understanding forecast. Accuracy: MAPE, WAPE, WMAPE [Электронный ресурс] / – Электрон. текстовые дан. – 2021. – Режим доступа: https://www.baeldung.com/cs/mape-vs-wape-vs-wmape, свободный.

8. Daniel Servén Marín. pyGAM balancing predictive power and interpretability with generalized additive models [Электронный ресурс] / PyData – Электрон. репозиторий – Berlin: [б.и.], 2018. – Режим доступа: https://github.com/dswah/PyData-Berlin-2018-pyGAM, свободный.

9. Dan Becker. Machine Learning Explainability [Электронный ресурс] / kaggle – Электрон. курс – [б.и.] – Режим доступа: https://www.kaggle.com/learn/machine-learning-explainability, свободный.

10. Dipanjan (DJ) Sarkar. Hands-on Machine Learning Model Interpretation. [Электронный ресурс] / Medium. – Электрон. текстовые дан. – [б.и.], 2018. – Режим доступа: https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608, свободный.