Текст защиты

Слайд 1

Здравствуйте. Меня зовут Скикевич Игорь. Я работаю программистом в частном медицинском центре.

Тема моего финального проекта «Интеллектуальная система оптимизации ценообразования услуг частного медицинского центра».

Слайд 2

Данная тема была выбрана не случайно. Наша команда разрабатывает медицинскую информационную систему. В ней разрабатывается модуль статистики и финансовой аналитики медицинской организации, а выбранная тема является одной из частей этого модуля.

Слайд 3

В своей работе я преследовал 2 цели:

1. Изучить задачу оптимизации цены с точки зрения бизнеса.
2. Разработать прототип модели прогнозирования для будущего сервиса.

Слайд 4

С точки зрения бизнес процесса:

Система ценовой оптимизации строится на алгоритме, который каким-либо образом моделирует функцию спроса.

Процесс системы цикличен: сначала вводятся данные для прогноза, затем прогнозируется спрос, далее подсчитываются ключевые бизнес метрики (объём выручки, прибыль, маржа – в зависимости от поставленной бизнес цели). Следом процесс повторяется, каждый раз для новой цены, до тех пор, пока не выявится оптимальная цена.

Слайд 5

В машинном обучении моделирование функции спроса представляет собой довольно стандартный подход - обучение с учителем. Конкретно для оптимизации цены решается задача регрессии.

Важно отметить, что прогнозировать спрос можно с помощью прогнозирования временных рядов, но такой подход практикуется для оптимизации запасов в розничной торговле, и для оптимизации цены он не подходит, (потому что для такой модели необходимо, чтобы данные собирались через равные промежутки времени).

На основе информации о методах решения данной задачи средствами машинного обучения мною было принято решение взять, в качестве прототипа, модель, основанную на градиентном бустинге, потому что он отлично справляется с задачей предсказания спроса.

Слайд 6

Т.к. предстоит решать задачу регрессии, для оценки предсказательной способности моделей, я выбрал метрику RMSE - корень из среднеквадратической ошибки.

А для того чтобы видеть куда смещён прогноз относительно фактических значений и на сколько процентов я выбрал метрику MPE - средняя процентная ошибка. Если MPE отрицательна – значит модель завышает прогноз, если положительна – наоборот, модель занижает прогноз. Сама метрика стремится к 0.

Слайд 7

В обучающие данные вошли как внутренние, так и внешние факторы. Внутренние данные представляют собой характеристики услуг, оказанных с конца января 2017 года по начало августа 2022 года. Внешние данные связаны с датой оказания услуг.

Слайд 8

В качестве наивной модели я решил взять простую линейную регрессию от числовых признаков. Величина отклонения прогноза довольна небольшая и модель справляется с задачей лучше чем, если бы я прогнозировал продажи случайным числом, но модель явно есть куда улучшать.

Слайд 9

Для ускорения анализа данных я воспользовался библиотекой pandas\_profiling.

Оказалось что в данных слишком много записей из номенклатурной группы «Анализы». Примерно 50% всех записей. Так же примерно такой же объём записей относится к медицинской специальности «Клиническая лабораторная диагностика». Это почти все записи, которые относятся к номенклатурной группе «Анализы».

В этот момент у меня появилась гипотеза о том, что такое смешение данных внесёт сильный разброс в распределение спроса, что сильно ухудшит предсказательную способность модели, но на момент проведения разведывательного анализа я не стал разделять услуги на группы, потому что это потребовало бы сбора большего числа данных и, следовательно больших временных затрат. Это я оставил на будущее.

Слайд 10

Наш медицинский центр работает ежедневно, но расписание у врачей меняется от месяца к месяцу. Признак weekday показал, что по выходным оказывается меньше услу, чем в будние дни, при этом записи об оказании услуг по будням распределены равномерно.

Признак holiday показал, что в праздничные дни и дни с сокращённым рабочим днём, установленные производственными календарями, присутствует спрос на услуги медицинского центра. Следовательно, данный признак можно оставить.

Слайд 11

Корреляционный анализ по Пирсону показал, что продажи имеют выраженную положительную корреляцию с количеством врачей. В данном случае это может свидетельствовать о том, что чем больше врачей, оказывающих одну и ту же услугу, тем больше продаж по этой услуге. Можно заметить что признак pandemic положительно коррелирует с продажами, и с праздничными днями.

Корреляция Крамера показывает, что медицинские специальности практически полностью скоррелированы с номенклатурной группой услуги. Этот факт подтверждает, что данные нужно разделять согласно гипотезе, озвученной ранее. Месяцы и Сезоны, а так же дни недели и выходные полностью попарно скоррелированы. Это ожидаемо – месяцы определяют время года, а выходные являются днями недели.

Слайд 12

На этом и следующих четырёх слайдах представлены графики для одной услуги, случайно взятой.

Из графика по сезонам видно что, спрос не имеет сезонности, трендов или цикличности. Вероятно, это связано с графиком работы врачей и характером медицинских проблем данной услуги. Как упоминалось ранее, расписание работы врачей меняется от месяца к месяцу, и к тому же, часто нарушается из-за разных личных ситуаций.

Слайд 13

На слайде представлен график изменения спроса по месяцам

Слайд 14

На этом графике показано как спрос распределялся по дням недели в каждый год.

Какие-то значимые выводы сделать трудно, потому что мне не достаёт данных о расписании работы врачей за представленный период данных.

Слайд 15

Спрос во время пандемии.

Сверху график спроса на выбранную услугу во время пандемии. Снизу – график спроса на эту же услугу до и после пандемии.

Как видно из графиков, спрос на данную услугу во время пандемии увеличился, относительно спроса до изменения эпидемиологической обстановки в стране. Интересно, что выбранная услуга относится к специальности акушерство и гинекология, и поэтому для меня остаётся загадкой по какой причине спрос на данную услугу возрос.

Слайд 16

На данном слайде представлен график распределения спроса на всё ту же услугу и изменение цены этой услуги на протяжении всего временного промежутка.

Судя по графику, вероятнее всего спрос на данную услугу неэластичен, а увеличение спроса при увеличении цены является следствием каких-то других факторов. Как показал график спроса во время пандемии на прошлом слайде, эпидемиологическая обстановка является одним из таких факторов.

Слайд 17

На этом графике представлен суммарный спрос на уникальные, для нашего города, услуги. По-видимому, уникальность услуг не прибавляет им популярности, хотя, возможно, цены на эти услуги слишком завышены, или же данные услуги не пользуются популярностью в связи с характером медицинских проблем, связанных с этими услугами.

Слайд 18

Аномалии в числовых признаках.

Сверху представлены боксплоты признака cost.

Аномальные значения цен обусловлены отсутствием разбиения данных на группы по медицинской специальности и номенклатурным группам. На данном этапе я решил логарифмировать признак.

Снизу боксплоты признака amount\_dostors.

На мой взгляд, в данном признаке можно оставить всё как есть. Если в будущем я смогу получить точную информацию о расписании работы врачей, а так же эпизодических появлениях приезжих врачей, тогда можно будет изучить данный признак детальнее, вкупе с дополнительными данными, и понять какие значения являются выбросами и как их лучше сгладить.

Слайд 19

Циклические признаки.

Такие признаки как время года, месяц, день недели, являются циклическими и для того чтобы алгоритм «понимал» что за зимой следует весна, за декабрём – январь, а после воскресенья наступает понедельник, я нашёл интересный способ кодирования циклических признаков. На графике представлен пример признака month после кодирования. Вместо оригинального значения признака берутся синус и косинус от частного порядкого номера значения признака и максимального порядкового номера для всего ряда значений, уноженного на 2\*pi.

Слайд 20

Модели.

Я решил использовать CatBoost и LightGBM из-за их хороших показателей при высокой скорости работы даже без использования графического ускорителя. Ко всему прочему, данные модели обладают простым в использовании API, по сравнению с XGBoost, по крайней мере, мне так показалось. Так же предпочтение этим моделям было отдано по причине их умения самостоятельно обрабатывать категориальные признаки.

Все гиперпараметры подбирались вручную исходя из распространённых примеров и эмпирических наблюдений. Не лучший подход, поэтому в будущем буду подбирать гиперпараметры алгоритмом.

Несмотря на то, что данные, не очень чистые, модели показали весьма хорошие результаты.

Слайд 21

Сверху представлены результаты обучения модели CatBoost, хоть в среднем ошибка и небольшая, с выбросами модель не справляется.

Снизу представлены результаты обучения модели LightGBM. Результаты чуть лучше, чем у CatBoost, но модель так же не справляется с прогнозированием больших значений.

Слайд 22

Выводы.

В ходе работы были обучены 2 модели, на базе алгоритмов градиентного бустинга.

Обе модели показали неплохие результаты даже на таких разрозненных данных.

Хоть модели и показали неплохие результаты, прототип требует доработки, как по части данных, так и самих моделей.

Одна из обученных моделей будет являться основой для построения сервиса оптимизации цен и выявления, эластичных по цене, услуг частного медицинского центра.

Слайд 23

Итоги и планы доработок

Я выяснил, что для поставленной задачи собранные мною данные слишком разрозненны. В данных содержатся услуги разных медицинских специальностей и разных номенклатурных групп, что вносит высокий уровень случайности в таргет, что в свою очередь усложняет прогнозирование. Для улучшения качества прогнозирования нужно разбить данные на отдельные датасеты, и на каждом наборе данных обучать отдельную модель.

Планируется разработка сервиса с возможностью интерактивного изменения данных и построения интерактивных графиков, а также подсчета ключевых бизнес метрик.

Планируется разработка системы автоматизированного сбора и обработки данных.

Так же планируется проведение экспериментов с моделями на основе глубокого обучения и обучения с подкреплением.

И после введения в эксплуатацию первой версии сервиса планируется выяснение периода для переобучения модели на новых данных.

Слайд 24

На этом всё. Благодарю за внимание.