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».

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 2

В своей работе я преследовал 2 цели:

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

Слайд 3

С точки зрения бизнес процесса:

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

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

Слайд 4

В машинном обучении моделирование функции спроса представляет собой довольно стандартный подход - обучение с учителем. 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 такая модель не учитывает изменение цены в будущем).

Однако у меня есть гипотеза, что с помощью прогнозирования временных рядов можно учесть влияние быстро устаревающих данных на спрос и, ансамблируя предсказания двух моделей, таким образом улучшить результаты прогнозирования.

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 лучше справляются с задачей предсказания спроса.

Слайд 5

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

Чтобы видеть, куда смещён прогноз, относительно фактических значений, и на сколько процентов я выбрал метрику MPE - средняя процентная ошибка. Если MPE отрицательна - модель завышает прогноз, если положительна - модель занижает прогноз. Стремится к 0.

Слайд 6

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 7

В качестве наивной модели я воспользовался методом прогнозирования спроса - скользящее среднее. Наивная модель показала довольно низкую ошибку, что означало, что можно работать дальше.

Слайд 8

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, что эти записи повлияют на предсказательную способность модели в худшую сторону, но на момент проведения разведывательного анализа я не стал разделять услуги, а оставил это на будущее.

Слайд 9

Наш медицинский центр работает ежедневно, но расписание у врачей не нормированное и это отражает признак weekday. Он показал, что по выходным услуг оказывается меньше, 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.

Слайд 10

Корреляционный анализ показал, что продажи имеют выраженную корреляцию с количеством врачей и медицинской специальностью услуги. В данном случае это ожидаемо и может означать, что чем больше врачей, оказывающих одну и ту же услугу, тем больше продаж по этой услуге, и, также, естественным образом, продажи разнятся по медицинским специальностям.

Слайд 11

На данном слайде представлены три графика зависимости спроса по сезонам, месяцам и дням недели для одной услуги.

Данные графики показывают, что какой-либо ярко выраженной зависимости спроса от времени нет. Спрос не имеет сезонности, трендов, цикличности. Скорее всего, это связано с не стабильным графиком работы врачей данной услуги. К слову, большинство врачей нашего медицинского центра имеют не стабильное расписание, которое, к тому же, часто нарушается из-за разных личных ситуаций. Именно потому что временные промежутки получения спроса неравны друг другу я не рассматриваю данные о спросе как временные ряды.

Слайд 12

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 обстановке 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.

В планах на доработку я планирую подробно проанализировать изменение спроса на каждую услугу в отдельности, чтобы выявить неочевидные признаки, влияющие на спрос, а так же определить эластичность спроса для каждой услуги.

Слайд 13

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

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 во время пандемии на прошлом слайде, эпидемиологическая обстановка является одним из таких факторов.

Слайд 14

На данном 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. Вероятно, поэтому данные услуги и являются уникальными, для нашего города.

Слайд 15

Аномалии в числовых признаках.

Сверху представлены боксплоты признака cost.

Аномальные значения этого признака 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 16

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 порядкого номера делённого на максимальный порядковый номер для всего ряда значений и умноженный на 2\*pi.

Слайд 17

Разделение данных.

Т.к. данные не являются временными рядами, и таргет не имеет зависимости от временных изменений, то для разделения данных я использовал метод train\_test\_split из библиотеки sklearn.

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, эмпирических наблюдений и личной интуиции. Не лучший подход, поэтому в будущем буду подбирать гиперпараметры алгоритмом.

Несмотря на, откровенно говоря, не очень чистые данные, модели показали весьма хорошие результаты.

Слайд 18

На данном слайде представлены графики фактических и спрогнозированных значений. Сверху для CatBoost модели, снизу для LightGBM. Из графиков видно что lightgbm чаще оказывается ближе в предсказаниях к реальным значениям.

Слайд 19

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 20

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 по категориям номенклатурных групп и по медицинским специальностям на отдельные датасеты, и на каждом наборе данных обучать отдельную модель.

Планируется разработка сервиса с возможностью интерактивно вносить изменения в данные и смотреть на изменение спрогнозированного спроса, а так же с возможностью построения интерактивных графиков и подсчета ключевых бизнес метрик.

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

Слайд 21

На этом всё. Благодарю за внимание.