SberMarket & 1535

Весна 2021

1 Вступление

Привет, Аналитик! Наша компания SberMarket выросла в 120 раз за последние 2 года — достичь нам этого позволили ребрендинг, инвестиции Сбера, привлечение лучших специалистов на рынке и расширение географии. Не последнюю роль в нашем сверхбыстром росте сыграла и пандемия.

Внешние факторы очень помогли нам, но без настройки внутренних процессов в условиях жесткой конкуренции не выжить. Пора оптимизировать денежные затраты, так как большая часть денег уходит на маркетинг. Условно каналы маркетинга можно разделить на 2 вида:

- можно напрямую отслеживать, через какие источники пользователи узнают о нас и начинают пользоваться нашим продуктом. Например, мы точно знаем, сколько людей пришли через интернет-рекламы, так как они кликнули на баннер.
- нельзя напрямую отслеживать, но можно оценить влияние: ТВ, SMM (продвижение в соц сетях - Facebook, VK, Instagram, ...).

Тебе нужно будет оценить эффект ТВ-рекламы на количество совершаемых заказов в неделю.

2 Задание

Для оценки эффекта ТВ на кол-во заказов нужно понять, какие факторы влияют на количество заказов — основную переменную в нашей задаче, которая очень важна для компании. Будем с этого момента называть эту переменную зависимой, или y. Она зависит от различных факторов: зимой больше заказов из-за праздников и холодов, зависимость может быть от СМС, количества показов рекламы на ТВ и т.д. Попытаемся построить простую линейную зависимость между всеми этими параметрами X — в анализе данных они называются признаками — и целевой переменной y:

$$y = \sum_{i=1}^{n} w_i \cdot x_i + b \tag{1}$$

где w_i — вес i-того признака, x_i — значение i-того признака, для данного объекта выборки, а b — $bias$, или сдвиг. Конечно, такая простая модель не сможет идеально предсказывать количество покупов товаров, однако если окажется, что она делает это с приемлемой погрешностью, то ей можно будет пользоваться. Такая модель называется линейной регрессией. К тому же, у линейной регрессии есть свои плюсы — ей легко пользоваться, вычислительно она очень проста, а также ее можно явно интерпретировать.

Процесс подбора (в случае линейной регрессии их можно вычислять напрямую) оптимальных весов w_i назвается обучением модели. Во время обучения подбираются такие веса, при которых модель позволяет лучше всего предсказывать значения целевой переменной на имеющихся данных, которые еще называются обучающей выборкой. Имея конкретные значения весов можно вычислить предсказанное значение Y для каждой точки в обучающей выборке и измерить среднюю ошибку по всем точкам. Эту среднюю ошибку и необходимо минимизировать. Ее можно вычислить по формуле

$$MSE = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_{true} - y_{predicted})^2 \tag{2}$$

где N — размер обучающей выборки, y_{true} — реальное значения целевой переменной для i-того объекта (точки) выборки, $y_{predicted}$ — предсказанное моделью значения целевой переменной для него. MSE расшифровывается как Mean Squared Error. О том, почему берется именно квадрат ошибки, а не сама ошибка, мы здесь говорить не будем, хотя про это можно погуглить или спросить. При обучении линейной регрессии производится минимизация именно этой ошибки MSE на обучающей выборке.

Конечно, подбор/вычисление оптимальных весов не нужно производить вручную. Для этого можно воспользоваться уже готовым алгоритмом из библиотеки scikit-learn.
Он называется **sklearn.linear_model.LinearRegression**. Можно почитать о использовании линейной регрессии на странице с ее документацией. Обучение модели производится в одну строчку кода (при условии, что данные для обучения уже подготовлены)!

После обучения модели значения получившихся весов можно интерпретировать. В случае линейной регрессии важность признака соответствует просто модулю от его веса — чем больше вес (по модулю), тем важнее признак.

2.1 Критерии

Тебе надо построить модель линейной регрессии, чтобы получить из весов модели значимость ТВ. Критерием качества модели является значение метрики MAPE на отложенной тестовой выборке и хорошо продуманная интерпретация. Для этого нужно пройти шаги по построению модели и понять, почему каждый из шагов важен.

Для оценки качества модели создано соревнование на сайте Kaggle. Туда можно будет загрузить предсказания для тестовой выборки и получить оценку качества модели.

3 Шаги

1. Что такое линейная регрессия? Построй линейную регрессию количества заказов на основе признаков X, которые, по твоему мнению, имеет смысл использовать.
2. Что такое BLUE оценка? Расшифруй каждую букву и дай вербальное объяснение - почему оценка должна обладать этими свойствами.
3. Проверь гипотезу о значимости коэффициентов при X. Что такое значимость, почему важно проверять значимость?
4. Проверь скоррелированность X и компонента ошибки. Построй графики с координатными осями y=ошибка модели и x=признак из модели для каждого из признаков. Если есть корреляция, то это проблема эндогенности. К чему это ведет? Какие причины? Какая буква из BLUE затрагиваются и как влияет на оценку? Дай интерпретацию существующей корреляции ошибки с переменной.
5. Проверь, как ведет себя ошибка — построй график ошибки по времени, по y. Какие выводы можно сделать? Можешь ли ты назвать момент, когда поведение этой завсисимости резко меняется и причину, которая за этим скрывается?
6. Объясни на словах, что такое гетероскедастичность и автокорреляция. Какая буква из BLUE страдает и почему? К чему ведет?
7. Какие методы по решению проблемы авторкорр и/или гетероскедастичности можно применить? Примени эти методы.

4 Входные данные

Тебе даны следующие признаки, агрегированные поо неделям:, за период с 2019.01.01 по 2020.12.31:

1. Охват постов в ВКонтакте
2. Охват постов в Facebook
3. Охват ТВ рекламы
4. Кол-во выходов рекламы по радио
5. Охват по СМС
6. Охват по push-уведомлениям
7. Охват email рассылок
8. Охват PR активностей
9. Индекс потребительских цен по РФ

10. Индекс цен компании Сбермаркет
11. Затраты конкурентов на маркетинговые активности
12. Сезонность
13. Кол-во поисковых запросов про Пандемию

Вся выборка разбита на обучащую и тестовую. Для обучающей выборки даны значения целевой переменной, а для тестовой их необходимо предсказать.

5 Выходные данные

На выходе необходимо получить:

1. Предказания количества покупок в неделю для тестовой выборки
2. Оценка влияния охвата ТВ рекламы на количества покупок в неделю