Дипломная работа

На тему: “Анализ продаж британского e-commerce (поиск инсайтов, составление рекомендаций стейкхолдерам, построение предиктивной модели объёмов продаж).”

Студент группы DAU -9:
Гетьман Александр Александрович

2022 год

Содержание

Описание датасета

Как правило, наборы данных электронной коммерции являются проприетарными и, следовательно, их трудно найти среди общедоступных данных. Однако репозиторий машинного обучения UCI создал этот набор данных, содержащий фактические транзакции с 01.12.2010 по 09.12.2011 года. Набор данных поддерживается на их сайте, где его можно найти под названием «Online Retail».

Это транснациональный набор данных, который содержит все транзакции, произошедшие в период с 01.12.2010 по 09.12.2011 для британского и зарегистрированного интернет-магазина. Компания в основном продает уникальные подарки на все случаи жизни. Многие клиенты компании являются оптовиками.

Датасет брался с E-Commerce Data | Kaggle

Наименование столбцов.

Invoice No - Счет-фактура №

Stock Code - Биржевой код

Description - Описание товара

Quantity - Количество купленного

Invoice Date - Дата выставления счета

Unit Price - Цена единицы товара

CustomerID - Идентификатор клиента

Country - Страна покупателя

Описание последовательности действий

Загрузка данных и работа с ключами.

Загружаем данные на юпитер ноутбук предварительно проинпортировав необходимые библиотеки для работы с данными и анализа. Исходный дата фрейм назовём df.

После того как данные успешно загружены посмотрим какого формата у нас стали столбцы через команду info.

После ознакомления с типами данных мы видим что столбец Invoice Date у нас имеет формат object. Так как у нас в данном столбце идёт обозначение времени необходимо сменить формат на datetime64[ns] для того чтоб можно было в дальнейшем создать дополнительные ключи для анализа. При этом видем что столбец CustomerID имеет формат float64 но число с плавающей точкой в данном случае нам не совсем подходит. Перевести в формат int64 в данный момент не можем из-за пропусков.

Сразу создаем новые столбцы Month, Year, Day, Time. Эти ключи нам пригодятся для дальнейшего анализа. После создания наблюдаем что они также создались в формате float64 и преобразовать их в int64 так же мы пока не можем.

Далее столбец Description необходимо привезти к единому регистру для того чтоб не появлялись дубли в уникальных значениях. Так как данные могут быть заполнены не совсем правильно.

Далее проверяем наш дата сет на нулевые значения. Видим что по результатам запроса CustomerID отсутствует примерно в 25% данных.

Чистка данных

Так как 25% от общего количества данных это весьма много то необходимо сохранить данные в отдельный df delete.

После создания нового дата фрейма заполняем столбец CustomerID значением 0.

Видим через команду info что количество пустых значений в столбце CustomerID сократилось до нуля.

Фильтруем датасет чтоб он нам показывал только те значения где CustomerID = 0 и сохраняем полученный результат.

Удаляем пропуски приводя данные к одинаковому количеству данных.

Проверяем выборку на дубли. После проверки видим что они присутствуют. Удаляем.

После того как все данные у нас теперь без пропусков мы можем столбцы CustomerID, Month, Year, Day, Time привести к формату int64 для более корректной работы с ними в дальнейшем.

Создаём ещё несколько столбцов Day month, Month year и Day month year для дальнейшего анализа по кварталам или годам.

Так же добавляем столбец Amount Spent.

Через команду describe смотрим описательную статистику по датасету df new где у нас лежат чистые данные. Видим что столбец Quantity имеет отрицательное значение а Unite Price имеет нулевое значение. Необходимо просмотреть транзакции с отрицательным значением.

После написания кода видим что такие транзакции имеются. Проводим анализ отмененных заказов.

Анализируя первые 5 значений фрейма данных, мы видим, что количество имеет отрицательные значения, нужно проверить верно ли это для всех отмененных заказов. После небольшого анализа наблюдаем что транзакции с отрицательными значениями являются отмененными сделками. Их процент не велик и составляет 1.74% от общего числа данных.

При анализе отмененных заказов мы видим что среди отмененных заказов есть какие-то скидки. Мы видим что их было 77. Необходимо сопоставить значения с исходным df. Данные не рознятся. и теперь мы можем данные по отмененным сделкам удалить.

Проверка транзакции товаров.

Так как 'BANK ' и 'CHARGES' являются служебными то они будут переименованы в 'BANK CHARGES'. Затем spec list будет преобразован в единый список. Теперь можно проверить все транзакции, связанные с Stock Code.

Как мы видим В набор данных включены и другие типы транзакций. Они будут сброшены. Специальные переводы: POST (почтовые расходы), M (ручные), банковские сборы и C2 (перевозка).

Анализ данных.

После того как данные у нас стали чистыми мы можем посмотреть корректное количество заказов по странам и стоимость заказов по странам. Для этого был создан дата фрейм df new.

Строим график количества заказов по странам и видим что большинство заказов было сделано из Соединённого Королевства. Компания базируется в Великобритании, поэтому кажется естественным, что страной с наибольшим количеством продаваемых товаров является Великобритания. Для дальнейшего анализа она будет отброшена. За исключением Великобритании, Германия, Франция и Ирландия тремя странами где клиенты потратили больше всего денег.

Строим график стоимости заказов по странам. Видим что самые большие цены в Нидерландах, Австралии, Японии и Швеции.

Далее строим график которые будут отображать сколько денег потратили клиенты в разных странах на наш продукт. Британию исключить по той же причине. Видим что за исключением Великобритании, больше всего денег на веб-сайте потратили клиенты из Нидерландов, Ирландии, Германии, Франции и Австралии.

Далее посмотрим какие продукты продавались чаще всего. Видим что самыми продаваемыми продуктами являются товары с индексами 23166, 84077, 85099B, 85123A и 22197.

Далее смотрим самые прибыльные продукты. Самыми прибыльными для компании являются товары с индексами DOT, 22423, 85123A, 47566 и 85099B.

Анализ самых продаваемых товаров.

Для данного анализа создадим новый дата фрейм df top prod. В фрейме данных количество представляет собой сумму всех проданных количеств для каждого продукта. Нам требуются все отдельные транзакции связанные с топовыми товарами.

Создаём дата фрейм с топ 50 товаров df top 50.

Выделяем топ 3 самых продаваемых товара где количество транзакции переходит за рубеж в 1000 сделок. В новые фремы попадают товары с индексом 85123A, 85099B, 22423.

Для каждого из 3 самых продаваемых продуктов строим визуализации по ежемесячным продажам. Видим что пики продаж данного продукта присутствуют но графики не показывают четкой закономерности в данных.

Построение модели финансового состояния магазина.

Для начала создаем новый тада фрейм (назовём df new 2). Так как потребуются нам не все столбцы то их можно отсеять.

Создаем столбец с общим доходом компании за месяц (Revenue).

После чего необходимо сгруппировать данные с помощью группировки по двум признакам (Месяц, Год). Так как данные за 12-2011 не полные то нам проще их предсказать удалив существующие.

При построении модели линейной регрессии целевой переменной будут объемы продаж по месяцам. Будем предсказывать на 3 месяца вперед (данные за 12.2011 неполные можно начать прогноз с него строения его до 02.2012). Оцениваем качество модели с помощью RSME.

В нашем случае, то что модель показывает ошибку по RMSE это нормально так как у нас очень маленькая выборка.

Создаём дата фрейм где будут находиться признаки и результаты прогноза (df d). Нам это необходимо для получения прогноза на нужный период.

Получив прогноз на необходимый период склеиваем 2 дата сета в один (df m).

По итогам предсказания и построенной визуализации мы видим резкий спад продаж на 01.2012 и 02.2012.

Пик продаж приходится в период с 10.2011 по 12.2011. Ожидаемо что в преддверии Нового Года данные сервисы пользуются большой популярностью для поиска и покупки подарков. Так как есть трафик из-за рубежа предполагаем что люди закупаются заранее чтоб успеть к празднику.

Выводы и рекомендации.

Делая вывод о ведении БД в данной компании. Качество данных не сильно хорошее.

Данный бизнес неплохо масштабирован по странам но из-за того что там нет указанного города клиента то мы не можем посмотреть регионы где приобретен товар. Еслиб было указано местное время клиента то можно было бы посмотреть на какие часы приходится пик продаж и стимулировать клиента на покупку в не пиковое время различными акциями.

Для данной компании неплохо подошла бы воронка AARRR.

Можно рассчитать время жизни клиента, сколько стоило его завлечь (при имеющихся данных о бюджете компании). Определить средний чек и в зависимости от того сколько он тратит посмотреть стоит ли нам увеличить стоимость товара или дальше работать на кол-во.

По итогам получившегося прогноза рекомендую сделать клиентские акции для увеличения конверсии по продажам и предотвращения возможного спада в продажах.