Ред. 17.01.23 переделал после замечания Готовцева А. - отделить прогнозирование количества клиентов (чеков) от среднего чека

*«… часто цель ставится из желаний, а не из реальных возможностей … перед тем как поставить цель необходимо сначала сделать прогноз ...»*

https://salers.ru/prognozirovanie-prodazh-tochnyj-raschet-ili-gadanie-na-kofejnoj-gushhe/

**Проект ТЗ дневное прогнозирования продаж**

**Цель:** Создать систему прогнозирования дневных продаж на основании имеющихся данных о продажах за прошлые и настоящие периоды. Дневные продажи по торговой точке описываются двумя усредненными показателями: Количество чеков (трафик) и Средний чека (ассортимент, цены, качество).

**Задача:** На данном этапе требуется создать алгоритм прогноза **Количества чеков** по Торговым Точкам (далее ТТ) на основе некоторых выбранных конкретных факторов. Протестировать алгоритм на прошлых периодах – вывести процент ошибки, выявить аномалии, при необходимости скорректировать количество и качество учитываемых факторов.

**Практическое применение:** Прогнозное **количество чеков** по дням можно суммировать на недельные и месячные прогнозы для использования в качестве основы (базы) для финансового планирования.

**Дальнейшие планы**: Дальнейшей задачей будет прогноз продаж по среднему чеку по такому же алгоритму для использования в качестве основы (базы) для финансового, производственного и маркетингового планирования. В последующем перейти на прогноз по количеству позиций в разрезе категорий продукции итд.

**Гипотеза:** Для того чтобы сделать прогноз мы обычно можем подсчитать усредненное Среднее количество чеков для конкретной ТТ в период который предшествует прогнозу. Соответственно гипотеза звучит так: Существует такой коэффициент **f** который **не** зависит от конкретной даты и конкретной ТТ, а зависит только от комбинации следующих факторов: (Тип ТТ*,* День недели, Время года, Праздник, Низкая температура**),** такой, что:

**Количество чеков** (на прогнозируемый календарный день по ТТ) равно **f** умноженное на **Среднее количество чеков** (далее **расчетная база**) за некоторый период предшествующий прогнозируемой даты по данной ТТ)

Другими словами, идеи такая: мы пытаемся разделить факторы которые отражают особенности конкретной ТТ и конкретное состояние продаж в данное календарное время (за это отвечает **расчетная база**) от факторов которые не зависят от ТТ и конкретного времени года, а зависят от комбинации Типа ТТ*,* Дня недели, Времени года, Праздника, Низкой температуры**.**

Коэффициент **f** оформляется в виде справочной таблицы:

| **Тип ТТ** | **День недели** | **Время года** | **Праздник** | **Низкая температура** | **f** |
| --- | --- | --- | --- | --- | --- |
| 3 вида значения | 7 видов значений | 3 вида значения | 4 вида значения | 3 вида значения | Коэффициент, число от 0 до … |

Размер таблицы будет такой:**3\*7\*3\*4\*3 =756 строк**

Таким образом, предполагается такой алгоритм по прогнозу дневного показателя:

1. На день прогноза, рассчитываем **расчетную базу** - **Среднее количество чеков,** за некоторый период предшествующий дню прогноза по рассматриваемой ТТ
2. По справочной таблице по дню прогноза и типу ТТ определяем **f**
3. Умножаем **f** на расчетную базу - **расчетную базу** и получаем в результате искомый прогноз.

Учитывая соотношение количества данных и количества строк справочной таблицы 11000/756=~15 получается, что данные в этой справочной таблице по основным видам будут усредненными из 15 и более величин, что надеемся, достаточно хорошо сгладит несистемные, случайные факторы, которые мы отбросили вначале. Возможная проблема в этой гипотезе: вероятно будет вводить большую погрешность способ деления ТТ на 3 вида, если будет большой процент ошибок во время тестирования, то придется отказаться от усреднения по типу ТТ и считать f по конкретной ТТ.

**Обоснование выбора используемых для прогноза факторов – параметров:** Факторы, которые будем использовать в модели: (выбор определялся по тому, что факторы являются системными и по ним имеются данные). Они зависят от нашего целевого показателя: **количество чеков**.

| **пп** | **Независимые факторы, параметры влияющие на показатель** | **Пример данных** | **Количество данных** |
| --- | --- | --- | --- |
| *1* | ***Место продажи (ТТ)*** | 1-20 | <20\* |
| *2* | ***Тип ТТ*** | Торговый центр  Домашний  Уличный | 3 |
| *3* | ***Недельные колебания трафика (День недели)*** | Пн-Вс | 7 |
| *4* | ***Календарные предсказуемые события (Праздник)*** | Предпраздничный  Праздничный  Послепраздничный | 3 |
| *5* | ***Время года*** | Лето  Зима  Весна - Осень | 3\*\* |
| *6* | ***Резкие падения температуры в зимний период (Низкая температура)*** | Не было  Было менее 45”С  Резкое падение до 45”С | 3 |

\*<20 можно сократить, оставить только точки, которые работали в период данных

\*\*для минимизации считаем весну и осень идентичными тк текущее состояние учитывается еще в расчетной базе (?)

**Факторы, которые мы игнорируем на данной стадии - не используем в модели**: *Ассортимент, Маркетинговые акции, Качество продавцов, Укомплектованность продавцами, Технические проблемы ТТ, Изменения конкурентной среды, Цены итд,*  тк по ним недоступны данные или данные имели случайный характер или слишком много данных. (Влияние на результат случайных и несистемных возмущений должно уменьшаться при усреднении системных факторов – для этого надо либо увеличить период сбора данных (что невозможно из за ковид) либо попытаться сократить, обобщить количество вариаций используемых системных факторов.)

**Используемые в модели данные:** Период: январь 2021 – январь 2023 - примерно 700 дней

1. **Данные о продажах из IIKO данные вида**

| **Дата** | **ТТ** | **Количество чеков в день, штук** |
| --- | --- | --- |
| примерно 700 дней | <20ТТ | Количество чеков в день, штук |

примерно **11000** данных (если считать 16 точек)

1. **Данные о температуре вида:**

| **Дата** | **Средняя температура за сутки** | **Резкие падения температуры в зимний период (Низкая температура)** |
| --- | --- | --- |
| примерно 700 дней | Значение средней температуры | Не было  Было менее 45”С  день падения до 45”С |

1. **Календарные данные**

| **Дата** | **День недели** | **Время года** | **9** |
| --- | --- | --- | --- |
| примерно 700 дней | Пн-Вс | Лето  Зима  Весна - Осень | Обычный  Предпраздничный  Праздничный  Послепраздничный |

1. **Данные о типе ТТ**

| **ТТ** | **Тип ТТ** |
| --- | --- |
| ~16 - 20 ?ТТ | Торговый центр  Домашний  Уличный |

**Следующая проблема:** Существует трудность в правильном определениирасчетной базы - **Среднее количество чеков** (за некоторый период около (до) рассматриваемого календарного дня данной ТТ) – этот показатель как бы служит базой расчетов для прогнозируемого дня, показывает некоторый средний уровень **Среднего количества чеков** по данной ТТ в данное календарное время. Проблемы в нахождении такой расчетной базы:

1. Усреднение должно быть не по месяцу, а по целым неделям, тк в календарном месяце разное количество дней недели. Разница небольшая, но обнаружено значительное искажение при пренебрежении этим фактом.
2. Для сглаживания влияния случайных и несистемных факторов период усреднения расчетной базы должен быть достаточно большим – желательно 2 полные недели.
3. Прогнозы бывают долгосрочные и краткосрочные, например, для финансового планирования надо спрогнозировать на все дни февраля, а сделать это необходимо примерно 25 января, это значит алгоритм определения расчетной базы для разных прогнозов будет разный.
4. При нахождении расчетной базы необходимо исключить из рассмотрения некоторые «необычные» недели – например недели предновогоднюю и после новогоднюю.
5. Отсюда еще одна проблема – если эта расчетная база будет считаться намного раньше прогнозируемой даты, то будет необходимо делать поправки на изменение тенденции в течении года по ТТ, например будет неправильно планировать на Октябрь по Августу, или на Май по Апрелю.

Для решения данных проблем необходимо:

1. Вести учет даты по номерам недель, а не по месяцам.
2. Создать таблицу недель, которые нужно исключить из расчета среднего, например 1-2 и 51-52 недели года.
3. Создать таблицу корректировок расчетной базы по неделям года.
4. Для долгосрочного планирования создать таблицу корректировок расчетной базы по месяцам или сезонам.
5. Определиться с периодом усреднения – предлагается 2 полные недели

**План:**

1. Экспертным путем определиться с перечнем ТТ, которые исключаем из расчета (по которым было много несистемных возмущений или неполные данные). Экспертным путем определить – стоит ли объединять Весну и Осень.
2. Экспертным путем определить и создать таблицу недель, которые нужно исключить из расчета базы вида:

| Номер недели | Исключить |
| --- | --- |
| 1-52 | да/нет |

1. Создать таблицу корректировок расчетной базы по неделям года вида:

| **ТТ** | **Номер недели** | **Среднее количество чеков за неделю** | **Коэфф -1 исключая необычные** | ***Коэфф -2 исключая необычные*** | **Среднее количество чеков за неделю из расчета за 2 недели с учетом корректировок и исключая необычные** |
| --- | --- | --- | --- | --- | --- |
| 1-20 | 1-52 | шт | Среднее количество чеков за неделю/ Среднее количество чеков за предыдущую неделю | *Среднее количество чеков за неделю/ Среднее количество чеков за неделю -2* | Среднее количество чеков за неделю из расчета за 2 недели с учетом корректировок и исключая необычные |

1. Привести к недельному формату и нормализовать таблицу Данные о продажах из IIKO и получить таблицу вида:

| **Дата** | **Номер недели** | **ТТ** | **f** |
| --- | --- | --- | --- |
| примерно 450дней | 1-52 | <20ТТ | 7\*Среднее количество чеков за день в ТТ тыс руб/ Среднее количество чеков за неделю из расчета за 2 недели с учетом корректировок и исключая необычные |

1. Присоединить остальные данные и провести усреднение, получится справочная таблица вида:

| **Тип ТТ** | **День недели** | **Время года** | **Праздник** | **Низкая температура** | **f** |
| --- | --- | --- | --- | --- | --- |
| 3 вида значения | 7 видов значений | 3 вида значения | 4 вида значения | 3 вида значения | коэффициент **f** |

Размер таблицы будет такой: **3\*7\*3\*4\*3 = 756 строк**

1. Для краткосрочного прогнозирования разработать алгоритм определения расчетной базы (средней выручки за 3 полные недели**)** пользуясь таблицами изпп2 и пп3.
2. Провести тестирование, случайно взять 50 дней из 2022 и 2023 года и по алгоритму рассчитать прогноз, сравнить с фактическим и оценить ошибку.
3. Провести тестирование на реальный завтрашний день, оценить ошибку.
4. Для долгосрочного прогнозирования создать таблицу корректировок расчетной базы по месяцам или сезонам, после этого разработать алгоритм определения расчетной базы на долгосрочный период.
5. Провести тестирование, случайно взять несколько месяцев из 2022 года и по алгоритму рассчитать долгосрочный месячный прогноз, сравнить с фактическим и оценить ошибку