Прогнозирование в разрезе SKU

Новые возможности повышения адекватности прогнозов

Алексей Субботин
Loginom Company (ex. BaseGroup Labs)

Трудности прогноза спроса

- Большое количество товарных позиций
- Особенности каждой SKU: сезонность, характерный масштаб объемов
- Наличие товаров-заменителей
- Изменение ассортимента: исключение старых и появление новых SKU
- SKU с редкими продажами

Путь 1: попозиционный прогноз

Достоинства:

1. Возможность учета индивидуальных особенностей (сезонность, редкие продажи, …)

Недостатки:

1. Трудоемкость поддержки и актуализации одновременно множества моделей
2. Длительность обсчетов
3. Потеря информации об отношениях между позициями (заменители, сопутствующие)
4. Игнорирование изменений в номенклатуре
5. Сложности построения композитных моделей из-за переменного числа SKU

Попозиционный прогноз

Пример расчета времени:

- Количество SKU – 100 000
- Количество моделей на 1 SKU – 10
- Время расчета 1 модели - 100 мсек

Итого:

- 100 000 SKU * 10 моделей* 100 мсек = 27.8 часов

Ошиблись? Нужны правки? Переделываем…

Путь 2: прогноз по группам

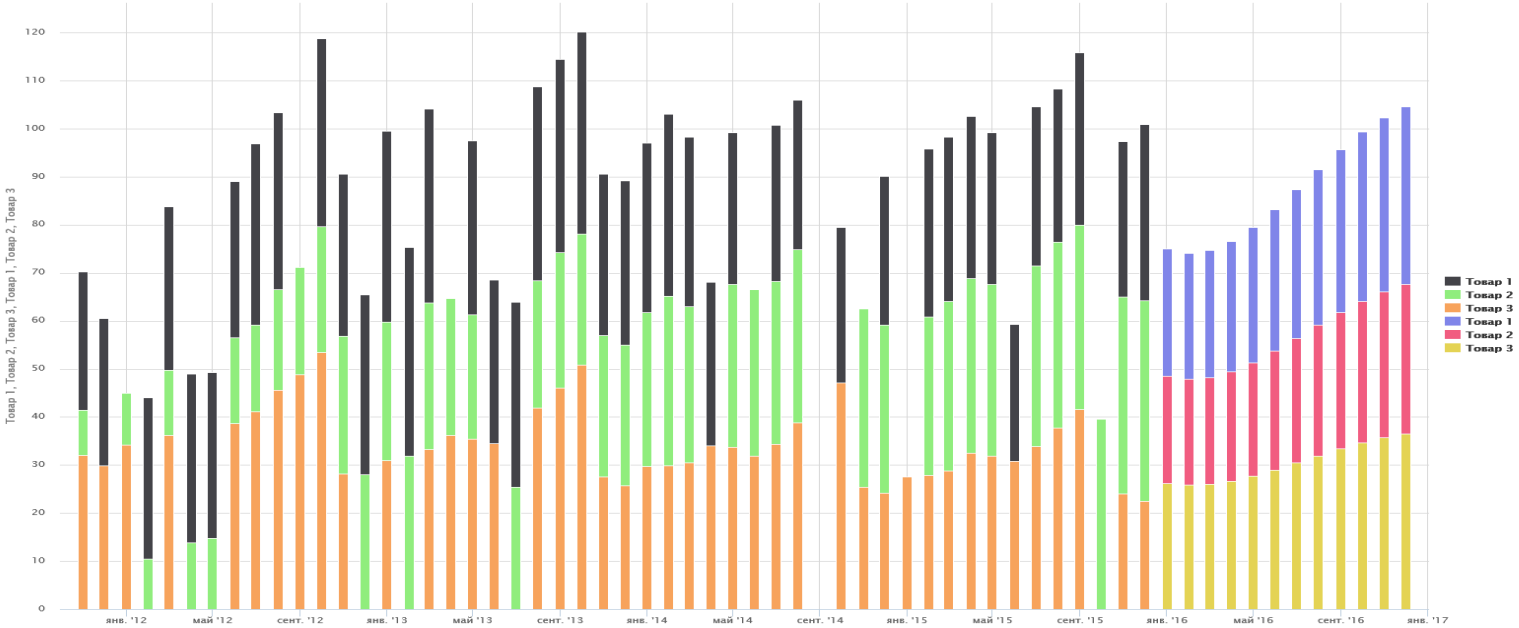
Достоинства:

1. Более стабильные прогнозы
2. Небольшое количество групп
3. Редкие изменения групп
4. Возможность обработки «сверху вниз»

Недостатки:

1. Нет информации по конкретному SKU, а это – конечная цель
2. Сложность получения из прогноза группы прогноза по SKU
3. Сложность балансировки попозиционного прогноза

Путь 2 - прогноз по группам

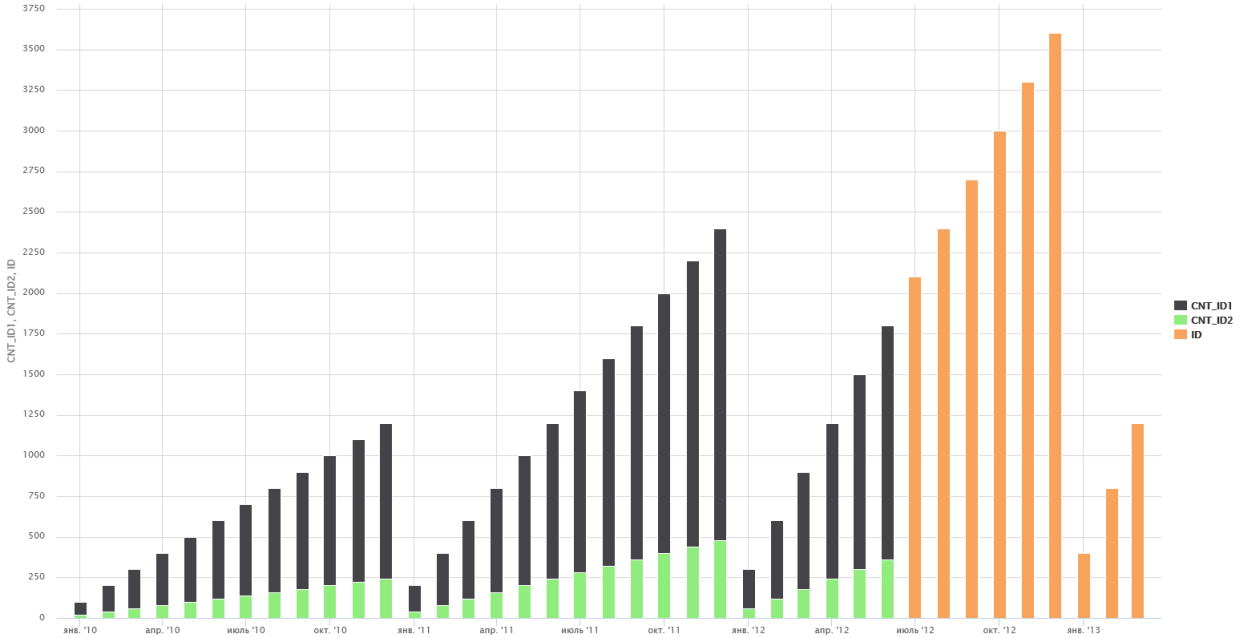


Адекватность – необходимое (но не достаточное!) условие точности.

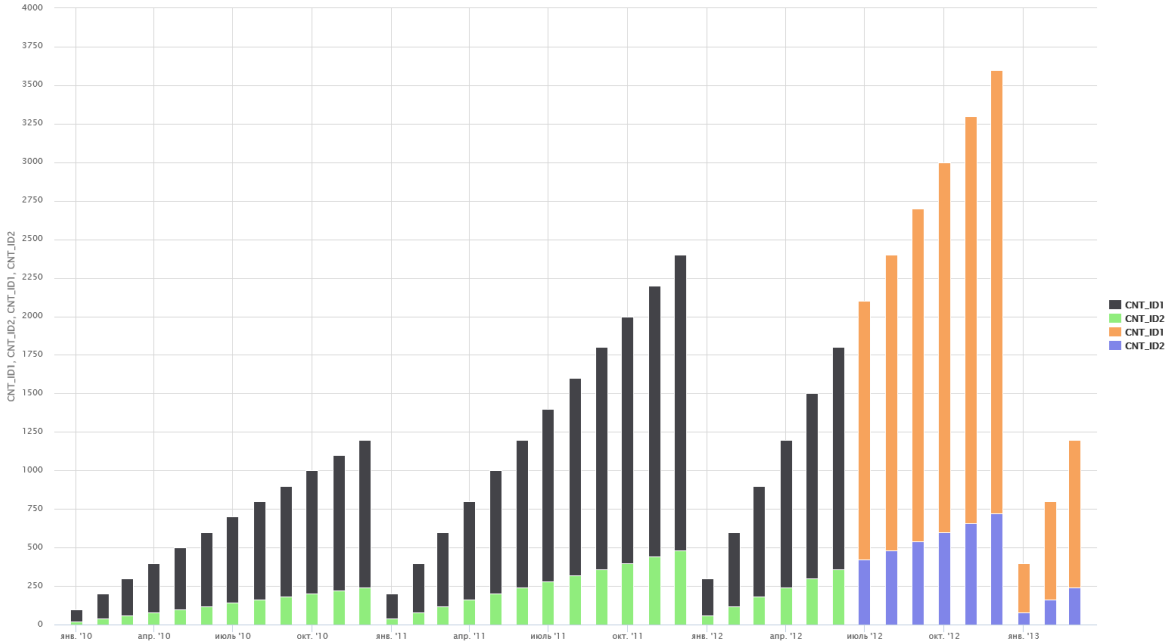
Неадекватный прогноз спроса:

- Некачественное планирование
- Избытки/недостатки на складах
- Ухудшение отношений с поставщиками
- Потеря денег

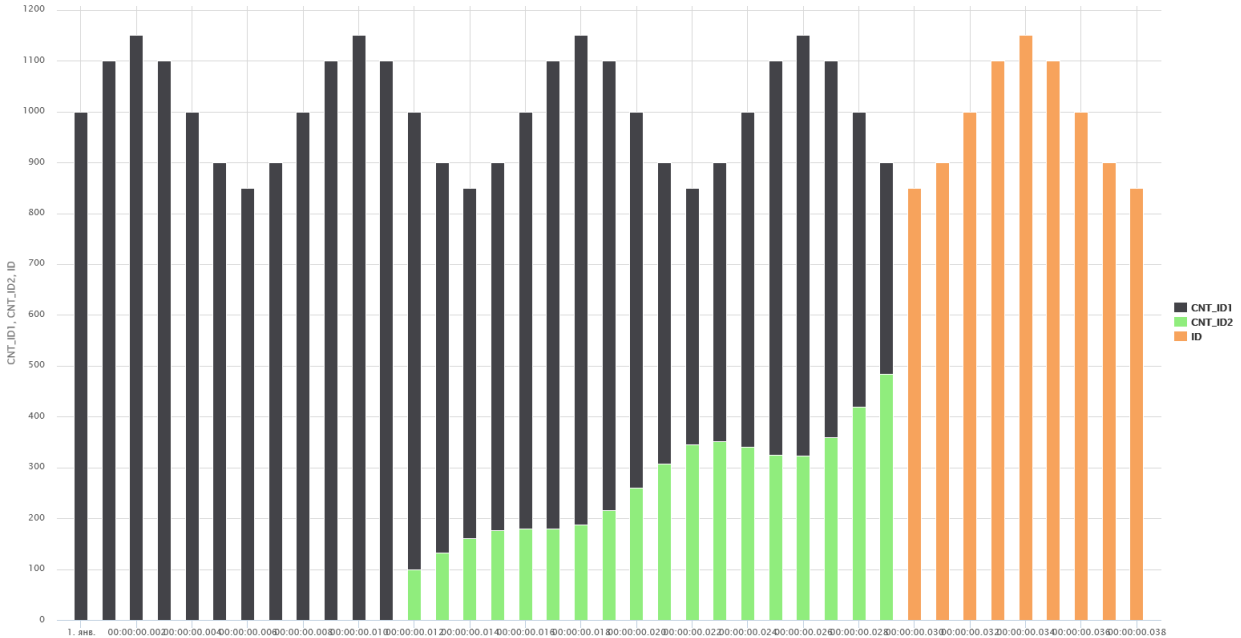
Адекватность это…



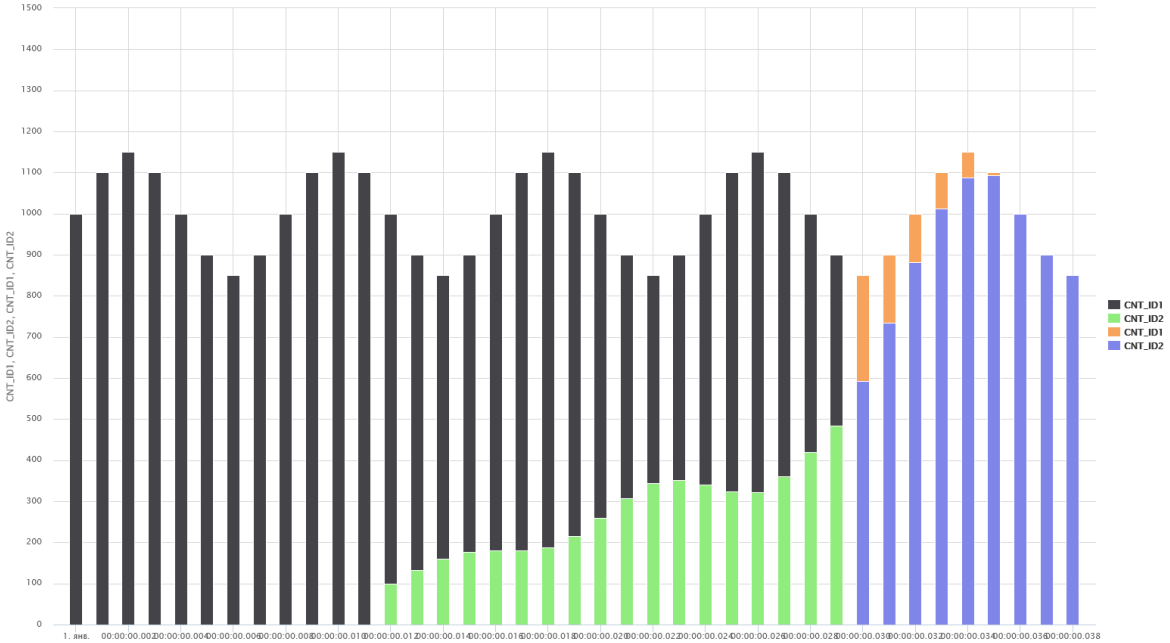
Интуитивная понятность



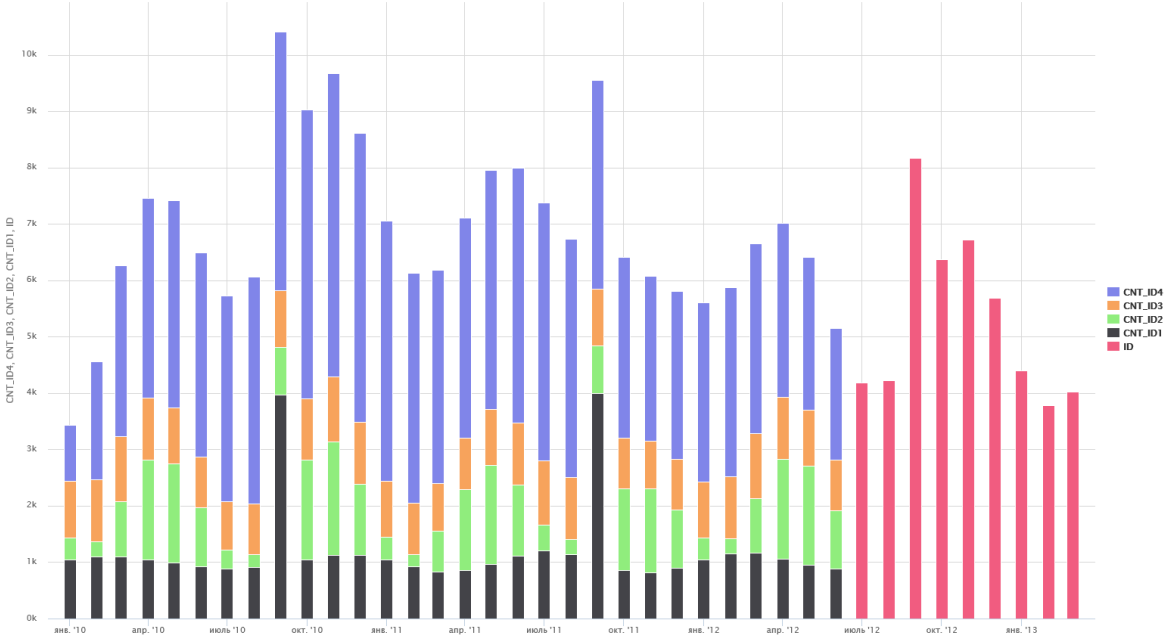
Адекватность это…



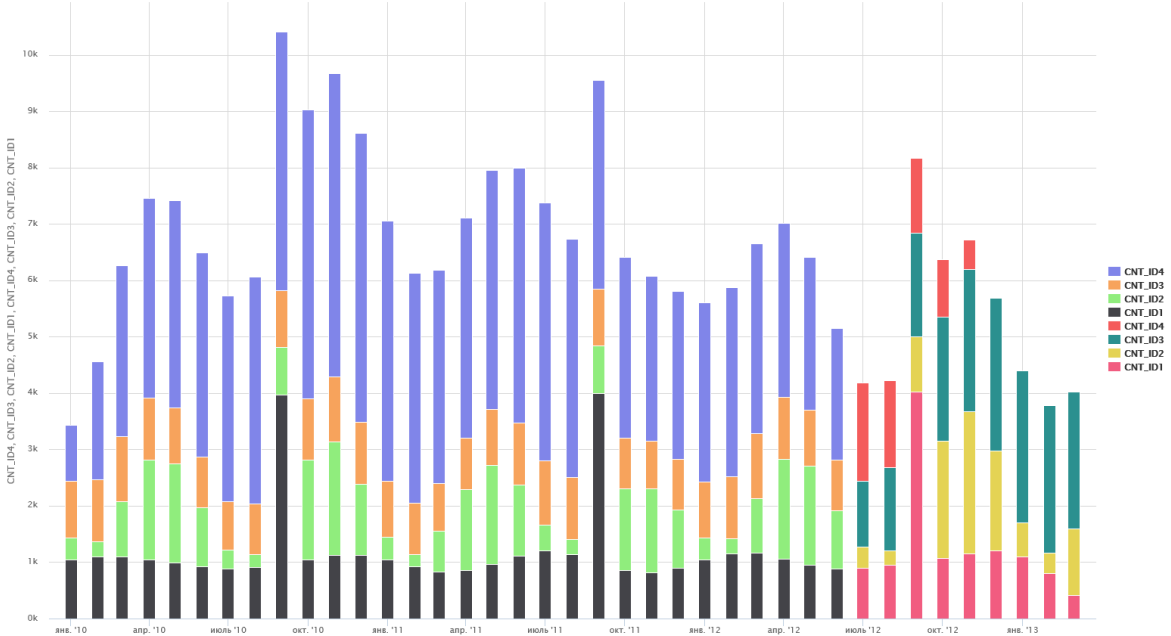
Учёт трендов, масштабов и отношений внутри групп



Адекватность это…



Учёт индивидуальных особенностей SKU



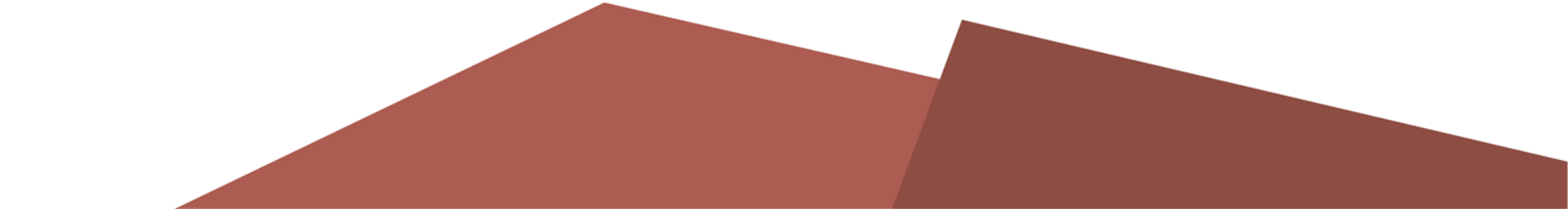
Адекватность это…

Универсальный алгоритм, одинаково работающий с любыми данными, не приспосабливаясь специально к «модельным» ситуациям

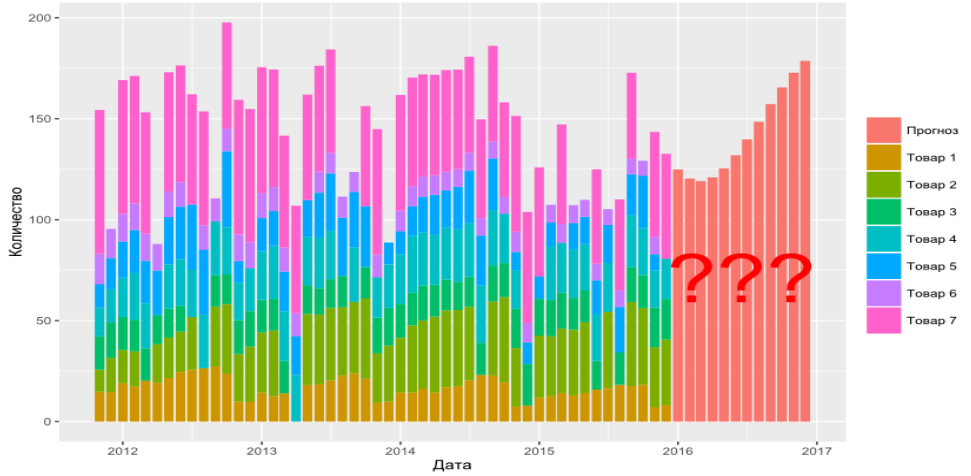
Решение

1. Объединить преимущества двух подходов
2. Стараться максимально учесть априорные данные

Алгоритм



Выбор базовой модели



$$\sum_{k=1}^{N} \hat{y}_{k,N-1+i} = \hat{Y}_{N-1+i}\,,$$

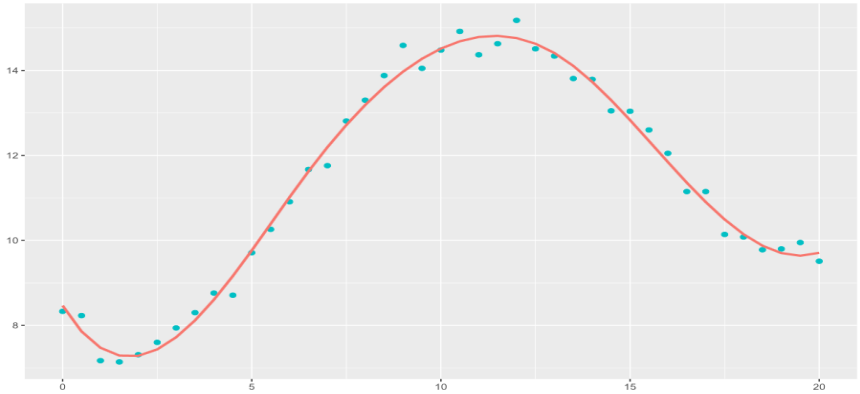
$$i = 1..P$$

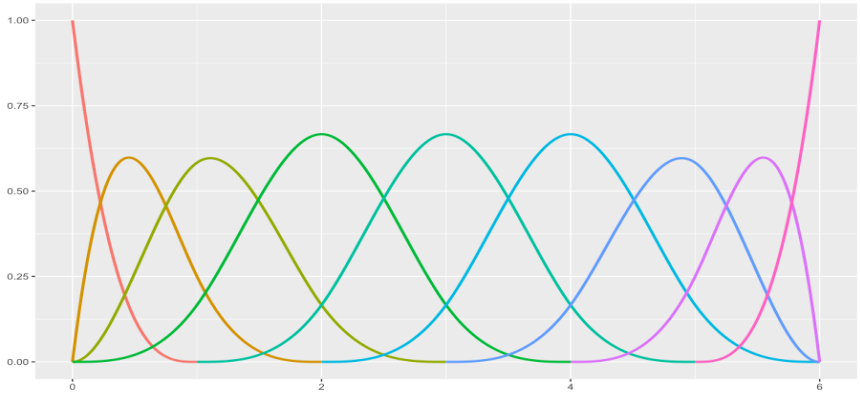
Проблемы функциональных трендов (линейный, полиномиальный…):

- Малое количество и глобальность степеней свободы
- Проблемы с точностью для моделей высокого порядка

Сплайны

Индивидуальный тренд





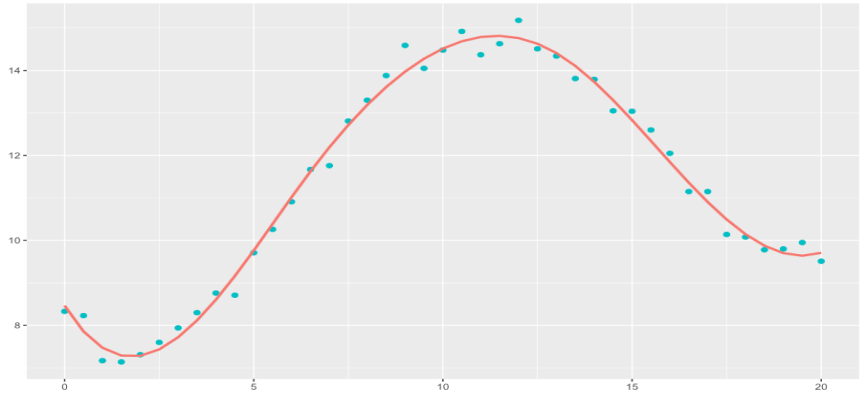
Сглаживание

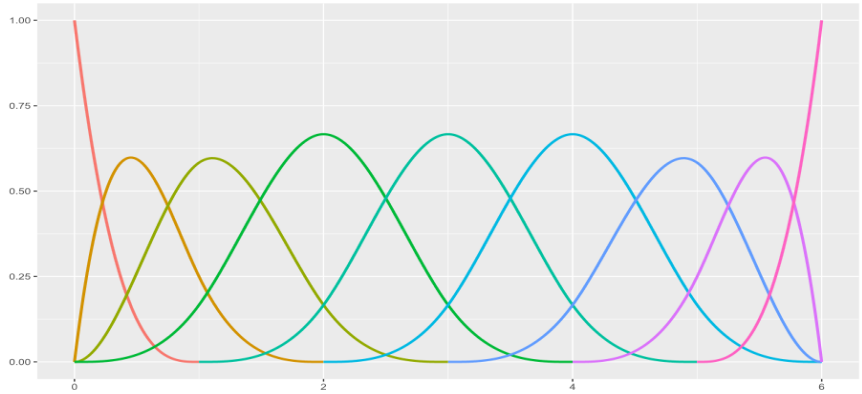
- $\hat{s}(t) = \arg\min_{\xi(t)}\{p \cdot \|\boldsymbol{\xi} - \boldsymbol{s}\|_2^2 + (1-p) \cdot \|\xi''(t)\|_2^2\}$

Метод конечных элементов

- $\hat{s}(t) = \sum_{j=1}^{K} \beta_j B_j(t)$
- $\hat{s}(t_i) = \sum_{j:B_j(t_i)\neq 0} \beta_j B_j(t_i)$

Индивидуальный тренд

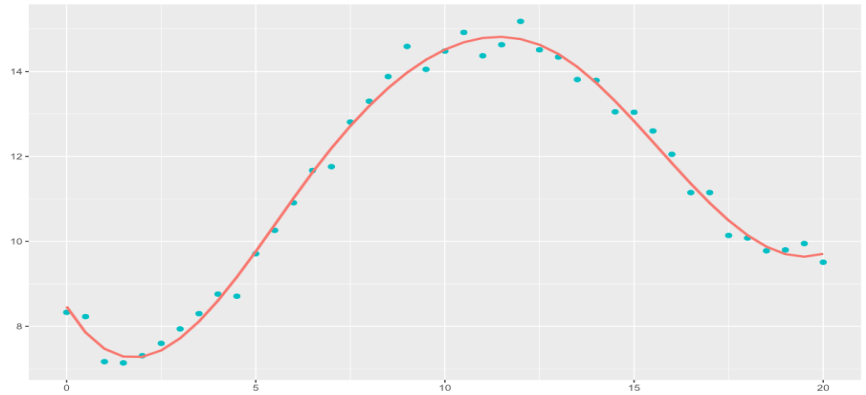


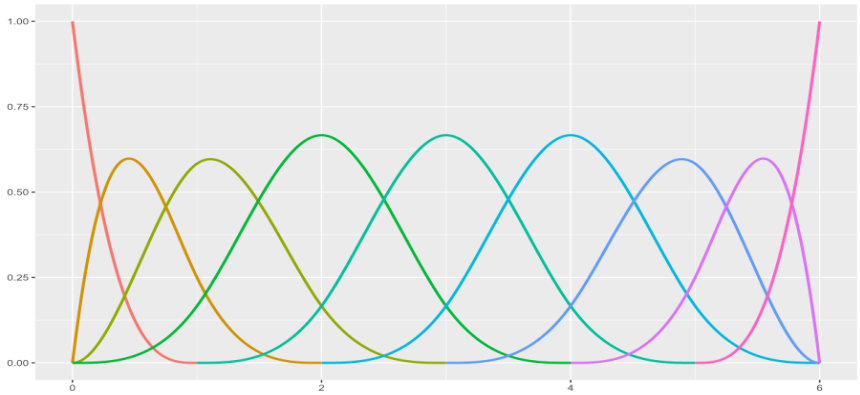


$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}\left\{p\cdot\sum_{i=1}^{N}[\hat{s}(t_i)-s(t_i)]^2+(1-p)\cdot\left(\int_{t_1}^{t_N}\hat{s}''(t)dt\right)^2\right\}=$$

$$=\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}\left\{p\cdot\sum_{i=1}^{N}\left[\sum_{j:B_j\neq 0}\beta_j B_j(t_i)-s(t_i)\right]^2+(1-p)\cdot\left(\sum_j\beta_j\cdot\int_{B_j(t)\neq 0}B_j''(t)dt\right)^2\right\}=0$$

Индивидуальный тренд

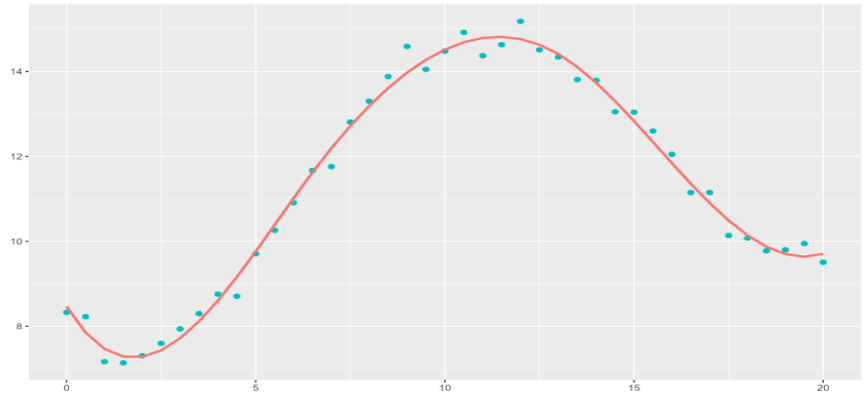


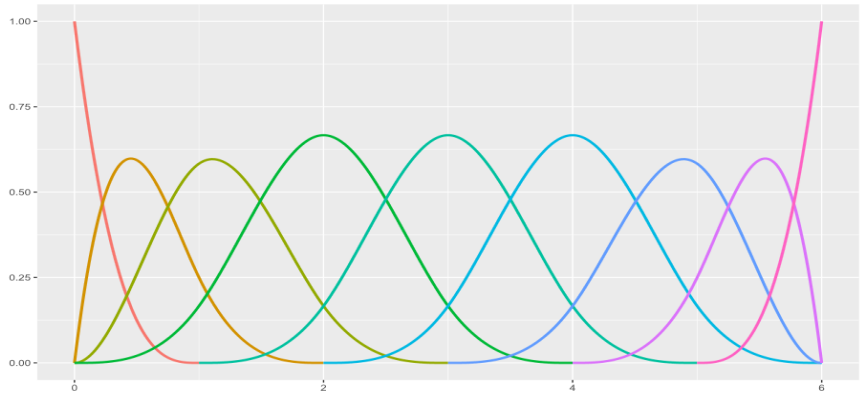


$$\{p \cdot \mathbf{B^T B} + (1-p) \cdot \mathbf{Q}\} \cdot \boldsymbol{\beta} = p \cdot \mathbf{B^T} \cdot \mathbf{s}$$

$$\Rightarrow \boldsymbol{\beta}$$

Базовая модель тренда





Преимущества регрессионных сплайнов:

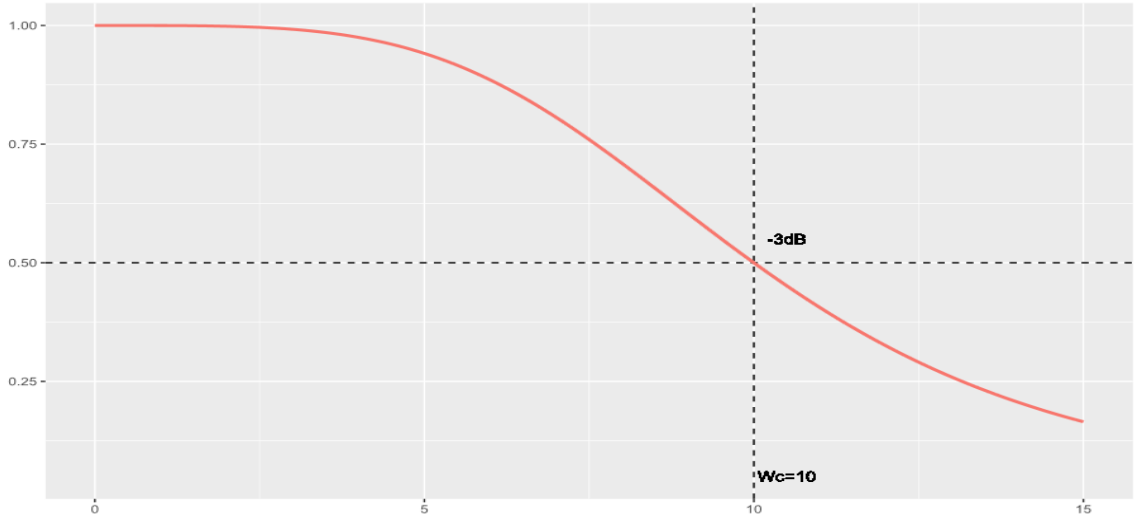
- Степеней свободы много, их число легко контролируется, и они локализованы
- Параметры можно подбирать грубо
- Нет проблем с устойчивостью вычислений
- Уравнения сводятся к системам с разреженными матрицами – время $O(N)$ вместо $O(N^3)$

Учет сезонности

$$\{p \cdot \mathbf{B}^{\mathbf{T}}\mathbf{B} + (1-p) \cdot \mathbf{Q}\} \cdot \boldsymbol{\beta} = p \cdot \mathbf{B}^{\mathbf{T}} \cdot \mathbf{s}$$

$$\{p \cdot \mathbf{B}^{\mathbf{T}}\mathbf{B} + (1-p) \cdot \mathbf{Q}\} \cdot \boldsymbol{\beta} = p \cdot \mathbf{B}^{\mathbf{T}} \cdot \mathrm{diag}(\boldsymbol{\eta})^{-1} \cdot \mathbf{s}$$

Подбор степени сглаживания



$$\frac{1}{1+\left(\frac{\omega}{\omega_c}\right)^4},$$

$$\omega_c = \left(\frac{p}{1-p}\right)^{\frac{1}{4}}$$

Редкие SKU без сезонности

Исключаем то, что невозможно прогнозировать. Необходимо использовать другие подходы, например, перенести моделирование на более крупный временной масштаб.

Объединение трендов в общую модель

$$\Phi(\hat{s}_1, \dots, \hat{s}_M) = \sum_{k=1}^{M} \alpha_k \cdot \Phi_k(\hat{s}_k) =$$

$$= \sum_{k=1}^{M} \alpha_k \cdot \{p_k \cdot \left\| \hat{\boldsymbol{s}}_{\boldsymbol{k}} - \mathbf{diag}(\boldsymbol{\eta}_{\boldsymbol{k}})^{-\mathbf{1}} \boldsymbol{s}_{\boldsymbol{k}} \right\|_2^2 + (1 - p_k) \cdot \|\hat{s}_k''(t)\|_2^2\}$$

$$\rightarrow min$$

$$\sum_{k=1}^{N} \left[\eta_{k,N-1+i} \hat{s}_{k,N-1+i}\right]_+ = \hat{Y}_{N-1+i}\,, \qquad i = 1..P$$

$$[x]_+ = \begin{cases} x, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Балансировка функционала – «изюминка» модели

$$\alpha_k \sim \left[Var\left(p_k \cdot \mathbf{B_k^T} \cdot \mathrm{diag}(\boldsymbol{\eta_k})^{-1} \cdot \mathbf{s_k} \right) \right]^{-\frac{1}{2}}$$

Решение

- Выпуклая оптимизация с нелинейными ограничениями
- Метод множителей Лангранжа

$$\frac{\partial}{\partial(\boldsymbol{\beta}, \boldsymbol{\lambda})} \Lambda(\hat{s}_1, \dots, \hat{s}_M) =$$

$$= \frac{\partial}{\partial(\boldsymbol{\beta}, \boldsymbol{\lambda})} \{\Phi(\hat{s}_1, \dots, \hat{s}_M) + \sum_{k=1}^{N} \lambda_k \cdot g_k(\hat{s}_1, \dots, \hat{s}_M)\} = 0$$

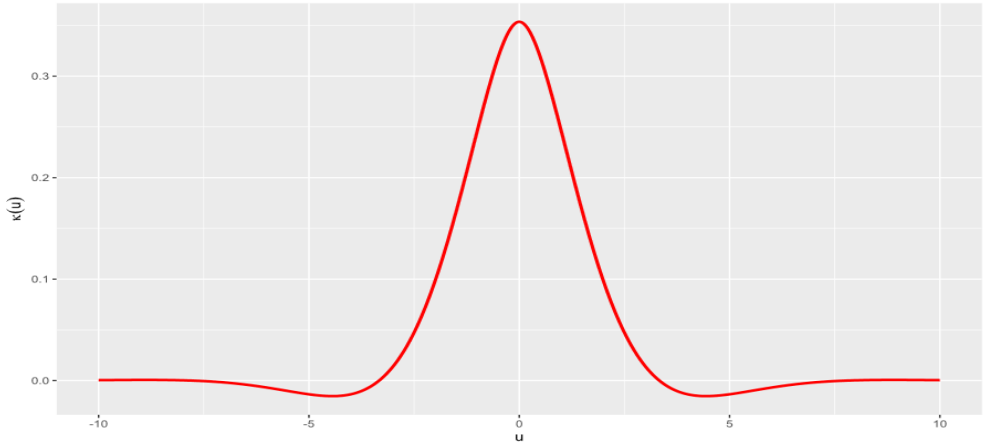
Метод активных множеств

$$\Phi(\hat{s}_1, \dots, \hat{s}_M) = \sum_{k=1}^{M} \alpha_k \cdot \Phi_k(\hat{s}_k) = \to min$$

$$\sum_{k=1}^{N} \left[\eta_{k,N-1+i} \hat{s}_{k,N-1+i}\right]_+ = \hat{Y}_{N-1+i}, \qquad i = 1..P$$

$$\mathbf{F} = \begin{pmatrix} \mathbf{H} & \mathbf{A^T} \\ \mathbf{A} & \mathbf{0} \end{pmatrix}$$

Алгоритм: оптимизация



Эквивалентное ядро

$$\kappa(u) = \frac{1}{2} \cdot e^{-\frac{|u|}{\sqrt{2}}} \cdot \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right)$$

$$d = l \cdot \frac{2\pi}{\omega_c} = l \cdot 2\pi \cdot \left(\frac{1-p}{p}\right)^{\frac{1}{4}}$$

1. Снижаем количество данных, требуемых для прогноза
2. Увеличиваем скорость работы.

Округление значений

- Округленная сумма не равна сумме округленных слагаемых
- Разные способы распределения дефекта: пропорционально, поровну, «всё одному»

Выводы

1. Быстрый метод с низкими требованиями по трудоемкости
2. Автоматическая балансировка и учет масштабов рядов
3. Адекватность на модельных данных, доверие к результату

Вопросы?

Алексей Субботин
aleksey.subbotin@loginom.ru

Loginom Company
(ex. BaseGroup Labs)
loginom.ru