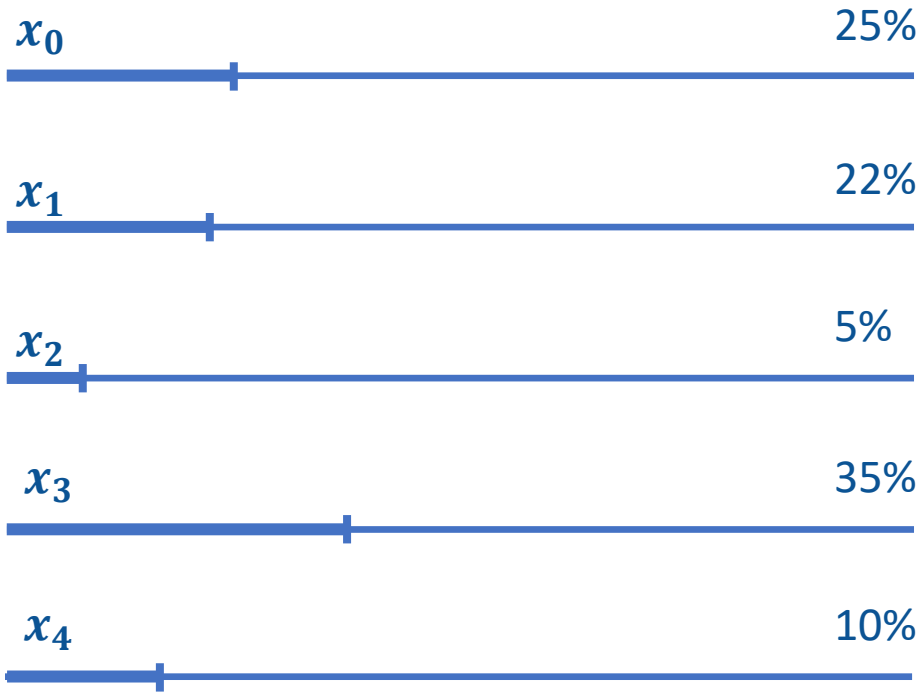
Отбор измеряемых величин

Для построения эффективной модели важно подавать данные по тем величинам, которые действительно влияют на target. Использование в прогнозировании только значимых величин позволяет лучше прогнозировать поведение target переменной и сократить затраты на лишние измерения и хранения информации.



Статистический анализ

Признаки между собой слабо связаны, но не все из них информативны для построения прогноза по целевой переменной. Статистически наиболее значимы оказались переменные с индексами 0,1,3 и 4.

Модель

Зависимость целевой переменной от отобранных признаков имеет нелинейный характер. Для восстановления этой зависимости лучше себя показала модель случайного леса (параметр значимости слева переменных).

Выводы

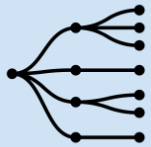
Для построения прогноза достаточно измерять величины с индексами 1-4. Это не наносит ущерба в прогнозах и позволяет экономить ресурсы сбор на хранении данных.

Кластеризация товаров

Кластеризация продуктовых товаров по группам может помочь решить задачи оптимизации поставок товаров в магазины, рекомендовать похожие товары и отфильтровывать список товаров в интернет магазине, оптимизировать хранение товаров на складе, выбор субкатегорий товаров, определение порядка выставления на полку.

Метрика дальности

Необходимо определить как измерять расстояния между объектами. Сравнивалось два: редакторское расстояние Левенштейна и косинусное расстояние.



Число кластеров

С помощью дендрограмм было установлено, что минимальное число кластеров равно 5.

Метод кластеризации

Для кластеризации использовался метод k ближайших соседей, в котором число кластеров задавалось не меньше 5.