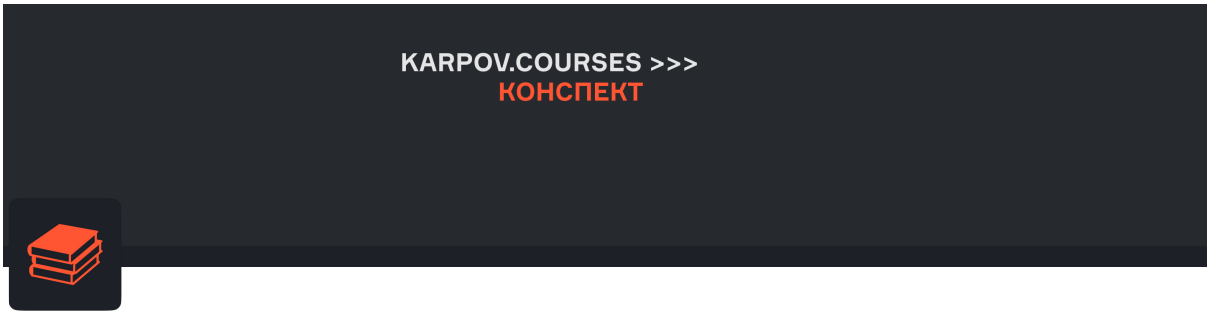



# > Конспект > 6 урок > Скрещивание динамического и классического ценообразования

## > Оглавление



Ноутбук и данные для занятия лежат тут: скачать

Ноутбук в формате `html`: скачать

## > Кластеризация временных рядов

Кластеризация имеет важное практическое значение в ценообразовании: она помогает обратить внимание на отдельные специфические группы, для которых можно предсказать отдельную цену и выиграть в обороте/прибыли. С помощью неё можно выделить, например, такие группы, как:

1. "Генераторы трафика" — товары с наибольшими продажами.

2. "Генераторы прибыли" — товары с низкой эластичностью, невысокой ценой и средними продажами.
3. "Промо-товары" — товары, которые продаются только в промо.
4. "Усилители прибыли" — товары с низкой эластичностью и эпизодическим характером продаж.

Очень часто выгодно кластеризовать товары не только по каким-то статистическим признаками, сгенерированным по временному ряду, но и по самой динамике изменения цены/продаж. На эту тему Станислав провёл отдельный вебинар, который мы настоятельно рекомендуем посмотреть. Ноутбук с вебинара доступен по ссылке. Его мы также рекомендуем изучить, чтобы разобраться, как всё работает. Однако всё же приведём основные техники:

1. Time Series K-Means из пакета tslearn;
2. Dynamic Time Warping (DTW) — с помощью этой техники можно рассчитывать расстояние между временными рядами и потом кластеризовать их по этому показателю;
3. Variational Recurrent AutoEncoder (VRAE) — позволяет получить embedding, кодирующий всю временную последовательность. Соответственно, временные ряды можно кластеризовать с использованием полученного вектора;
4. Deep Temporal Clustering;
5. Time2Vec — техника работает аналогично VRAE.

# > Quantile loss

Постановка бизнес-задачи: нужно найти такое значение цены (а именно ниже рыночной), при которой мы увеличиваем позицию на рынке, чтобы покупатели пришли именно к нам.


Нужно уметь предсказывать цену ниже рыночной или на уровне рынка. При этом модель должна сильнее штрафовать за перепрогноз (предсказали больше), чем за недопрогноз (предсказали цену меньше). Под описание подходит лосс для $10$ квантиля.

Для $50$ квантиля одинаковый штраф слева и справа.

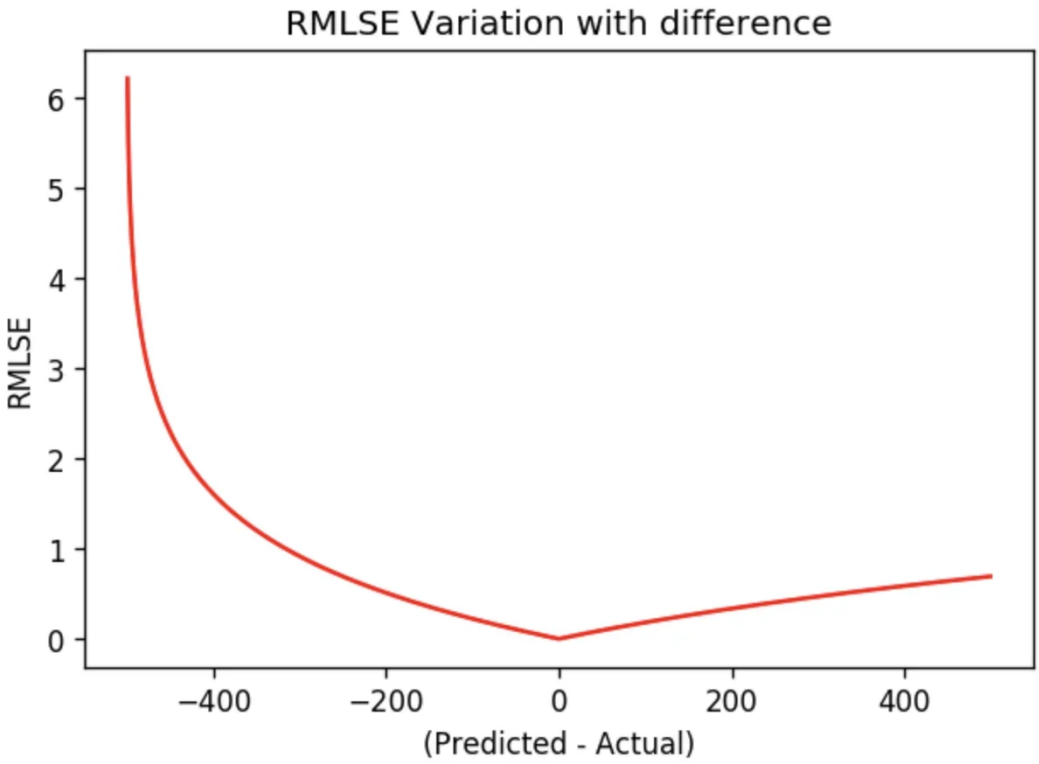
Для $90$ квантиля больше штрафуем за недопрогноз: например, хотим получить максимальную маржинальность продажи, поэтому лучше ставить цену выше.

# > Root Mean Squared Log Error (RMSLE)

RMSLE (среднеквадратичная логарифмическая ошибка) — одна из многочисленных ошибок, пригодная для задач регрессии. Она рассчитывается по следующей формуле:

$$RMSLE\left(y,\hat{y}\right)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log\left(y_i+1\right)-\log\left(\hat{y}_i+1\right)\right)^2}$$

Графически её рост с отклонением $y$ от $\hat{y}$ выглядит следующим образом:




Она обладает следующими свойствами:

1. Очень большая стабильность по отношению к выбросам в данных.
2. Из-за свойств логарифма она может быть трактована как относительная ошибка (разность логарифмов = логарифм частного), из-за чего масштаб величины перестаёт так сильно влиять (опять возвращаемся к выбросам).
3. Более высокие значения за недооценку, в сравнении с наказанием за переоценку на одно и то же значение (это наглядно видно из графика).

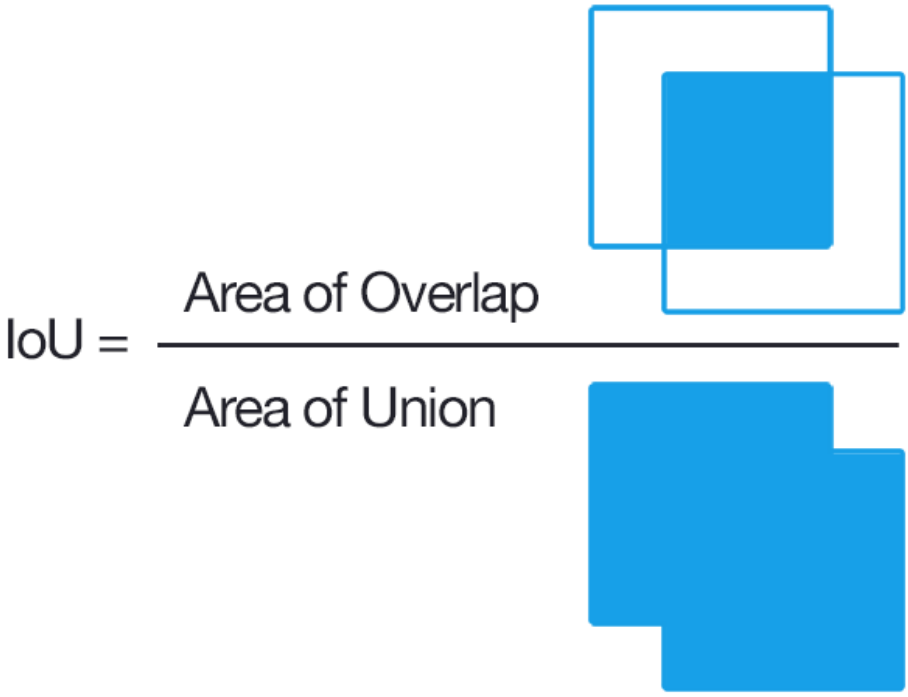
# > Метрики качества прогноза диапазонов

Одной из наиболее популярных метрик для оценки диапазонов является IoU (Intersection over Union) или индекс Джаккара. Эта метрика, как и похожая на нее Dice, пришла из компьютерного зрения: там она используется в задачах сегментации и обнаружения объектов.

Классическое определение выглядит так:

$$IoU = \frac{A \cap B}{A \cup B}$$

где $A$ и $B$ — множества пикселей, принадлежащих 2-м объектам (обычно одно из них — это предсказанная моделью маска, а второе — маска разметки).




Поняв смысл этой метрики, её легко переделать под работу с временными рядами. Для каждого момента времени рассчитаем интервалы предсказанного диапазона и интервалы реального диапазона (разметки). Соответственно, ширина (мера) пересечения уходит в числитель, а ширина объединения — в знаменатель. Если диапазоны не пересекаются, метрика будет равна $0$, а если полностью совпадают, то $1$. Финальная метрика по временному ряду получается усреднением метрик по каждой временной точке.

# > Дополнительные материалы

1. **Quantile regression**

   https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf

2. **XGBOOST + Quantile**

   https://towardsdatascience.com/regression-prediction-intervals-with-xgboost-428e0a018bhttps://colab.research.google.com/drive/1KlRkrLi7JmVpprL94vN96lZU-HyFNkTq?usp=sharing#scrollTo=wPm3w9CneeuQ