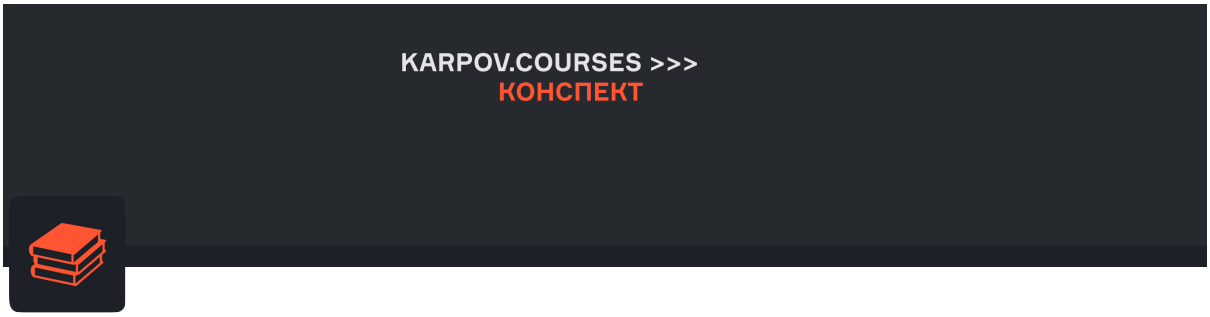


> Конспект > 5 урок > Предсказание диапазонов цен

> Оглавление

> Оглавление
> Мотивация предсказания диапазонов
> Классическая регрессия
> Отличие среднего и медианы
> Медианная регрессия
Пример получения оценки коэффициентов
> Квантильная регрессия
Пример использования
Некоторые библиотеки с квантильной регрессией:
> Резюме

> Мотивация предсказания диапазонов

- Проще выбрать конкретную точку внутри диапазона.
- В случае если модель ошибается в конкретной точке, диапазон может "покрыть" идеальное предсказание.
- Проще верифицировать предсказание у бизнеса.

> Классическая регрессия

В общем случае считаем, что целевая переменная $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$ линейно зависит от регрессоров $\boldsymbol{X} = \{x_1, x_2, ...x_m\}$ с весами $\boldsymbol{\beta} = \{\beta_1, \beta_2, ..., \beta_m\}$ и случайной ошибкой$\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, ..., \epsilon_n\}$. Если тема незнакома, можно почитать больше тут.

Запишем **модель** в векторном виде:

$$\boldsymbol{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Тогда оценка $\hat{\boldsymbol{y}}$ выглядит следующим образом:

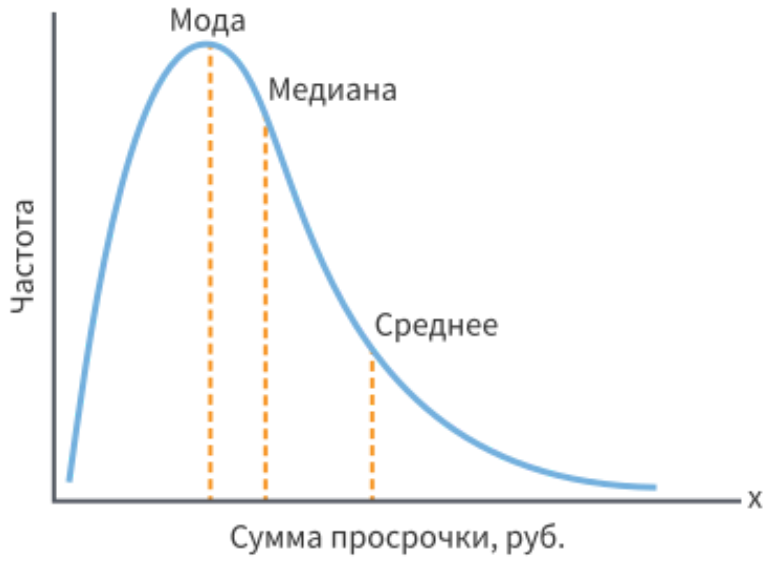
$$\hat{\boldsymbol{y}} = \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}}$$

Предполагаем, что от регрессоров зависит среднее значение $\boldsymbol{y}$.

Функционал качества, который используется при поиске оценки коэффициентов — сумма квадратов разностей наблюдаемого значения и оценки целевой переменной:

$$Q(\hat{\boldsymbol{\beta}}) = \sum_i (y_i - \hat{y}_i)^2 \to \min$$

> Отличие среднего и медианы



Унимодальное распределение

Мода — наиболее частое значение в выборке.

Медиана — квантиль порядка $\frac{1}{2}$, т. е. значение, которое делит распределение на две равновероятные части.

Среднее — среднее арифметическое выборки.

> Медианная регрессия

В отличие от обычной регрессии в медианной предполагается, что от регрессоров зависит медиана $\boldsymbol{y}$:

$$Med(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X} \cdot \boldsymbol{\beta}$$

Уравнение для оценки $\hat{\boldsymbol{y}}$ остаётся таким же, но оценки коэффициентов находятся путём минимизации суммы модулей отклонений:

$$Q(\hat{\boldsymbol{\beta}}) = \sum_i |y_i - \hat{y}_i| \to \min$$

Оценки коэффициентов вычисляются при помощи численных методов.

Пример получения оценки коэффициентов

$Med(y_i|x_i) = x_i \cdot \beta$ — наша модель с одним коэффициентом

$\hat{y}_i = x_i \cdot \hat{\beta}$ — прогнозируемое значение

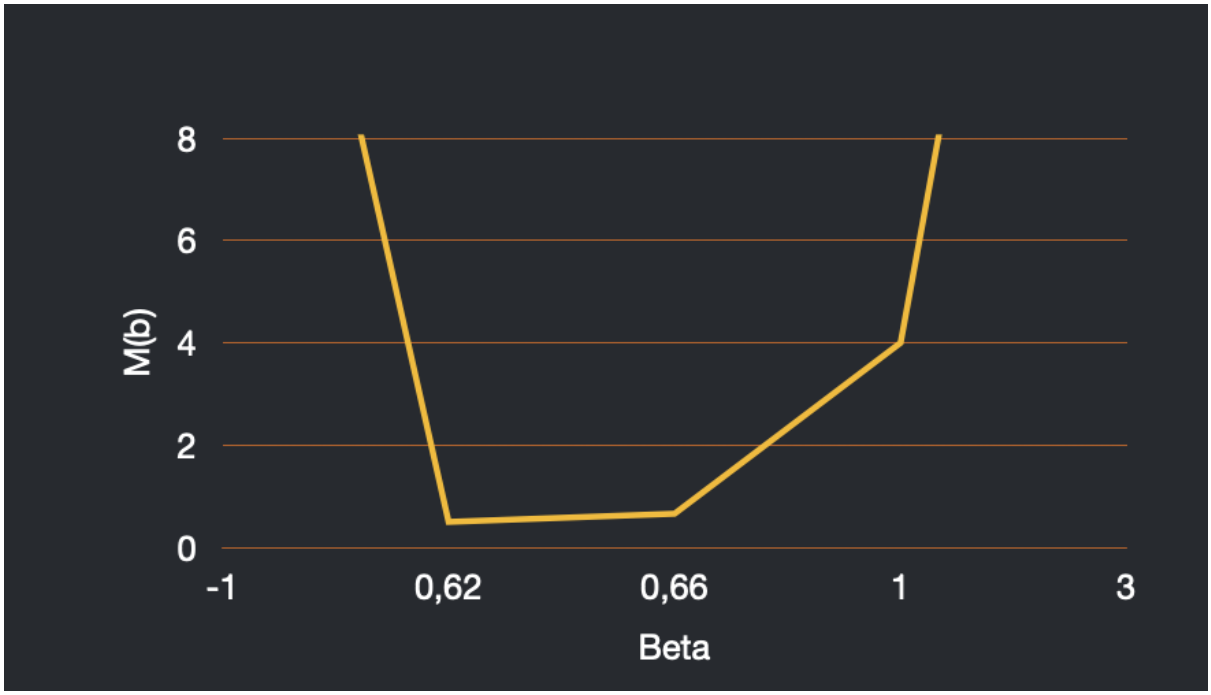
$$Q(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{3} |y_i - \hat{y}_i| \to \min$$

Необходимо найти $\hat{\beta}$.

y	x
1	1
2	3
5	8

$$Q(\hat{\boldsymbol{\beta}}) = |y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + |y_3 - \hat{y}_3| =$$
$$= |1 - 1 \cdot \hat{\beta}| + |2 - 3 \cdot \hat{\beta}| + |5 - 8 \cdot \hat{\beta}| \to \min_{\hat{\beta}}$$

$Q(\hat{\boldsymbol{\beta}})$ не дифференцируемая, кусочно-линейная функция. Можно изобразить её на графике, раскрыв модули.



Очевидно, что оптимальное значение $Q(\hat{\boldsymbol{\beta}})$ принимает при $\hat{\beta} = 0.625$

> Квантильная регрессия

Обобщение случая медианной регрессии. Говорят о квантильной регрессии порядка τ, где τ — **квантиль распределения**.

Медианная регрессия аналогична квантильной регрессии порядка $\tau = 0.5$.

Запишем **модель** квантильной регрессии в предположении о том, что квантиль порядка τ линейно зависит от регрессоров:

$q(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X} \cdot \boldsymbol{\beta}^{\tau}$

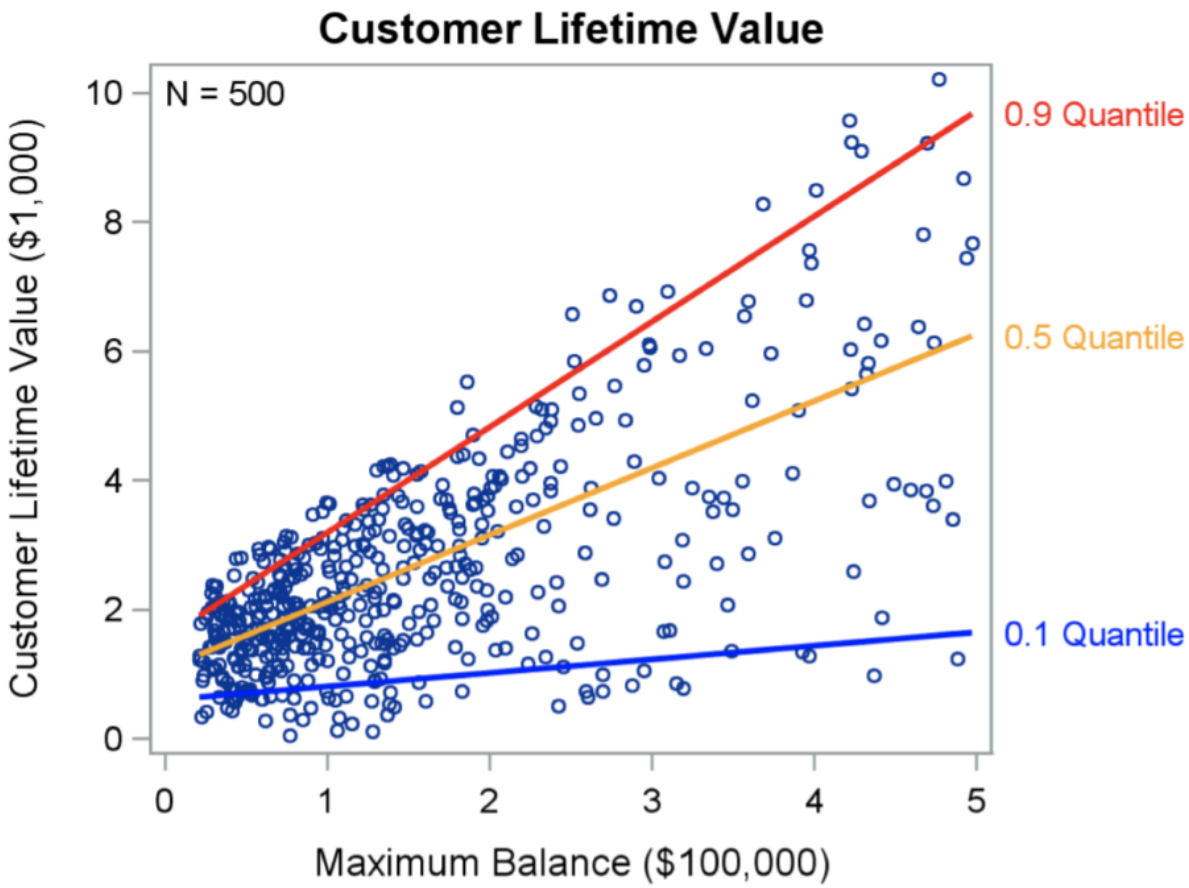
Зависимость для разных квантилей **может быть разной**.

Для поиска оценок коэффициентов используется минимизация асимметричной суммы модулей отклонений:

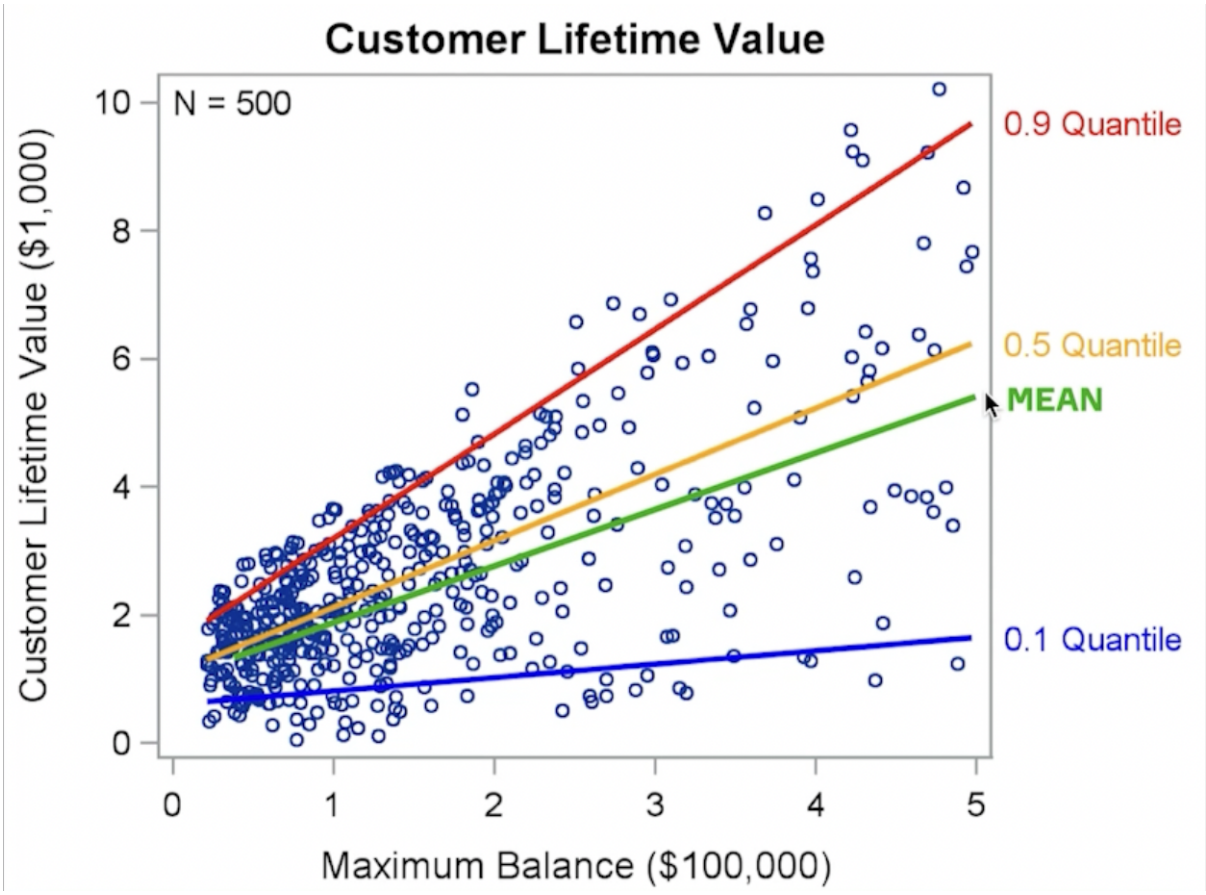
$Q(\hat{\boldsymbol{\beta}}) = \sum_i w_i \cdot |y_i - \hat{y}_i| \to \min$, где веса w_i:

$$w_i = \begin{cases} 1 - \tau, y_i < \hat{y}_i \\ \tau, y_i \geq \hat{y}_i \end{cases}$$

Пример использования



На графике предсказания LTV (lifetime value) по максимальному балансу на счёте изображены линии регрессии для $\tau = \{0.1, 0.5, 0.9\}$. Их наклоны сильно различаются и зависят от квантилей — это будет использоваться в предсказании диапазонов цен.



Нанесём среднее на график

Представим следующую ситуацию: мы предсказываем величину, которая изменяется от -1 до 1, при этом есть лицо, принимающее решение в зависимости от знака предсказания.

В таком случае можно потерять часть хороших предсказаний, которые находятся слева (*в отрицательной части*) около нуля при том, что доверительный интервал покрывает и некоторый отрезок справа от нуля.

Таким образом, можно влиять на принятие решение ответственным лицом: результат может быть разным, когда оперируют конкретными точками и диапазонами.

Некоторые библиотеки с квантильной регрессией:

- Catboost
- LightGBM

> Резюме

1. Вспомнили про регрессию, посмотрели на то, что такое медианная и квантильная регрессии.
2. Выяснили, что с помощью данных моделей можно предсказывать диапазоны цен для товаров.
3. Следует помнить про модели прогнозирования временных рядов.