E-commerce

Проект. Построение модели ценообразования для кроссовок

Бахишев Никита, Блохин Павел

БЭАД222

June 4, 2025

Вступление

Индустрия кроссовок — одна из самых динамично растущих ниш глобального fashion-рынка. По данным *Statista*, совокупный годовой объём продаж спортивной обуви превысил около 140 млрд USD в 2024 г. и продолжает расти темпами 5–7 % в год, а сегмент коллекционных и «ресейл»-кроссовок демонстрирует двузначные темпы прироста. Лимитированные коллаборации и нишевые модели формируют высокий и разнонаправленный спрос. Для бизнеса это означает серьёзную неопределённость в вопросах ценообразования: ошибочно установленная цена ведёт либо к потере маржи, либо к заморозке капитала на складе.

Научная и практическая значимость проекта заключается в развитии количественных подходов к оценке стоимости кроссовок. В отличие от большинства существующих работ, где анализ ограничен агрегированными показателями или простыми регрессионными моделями, в данном исследовании предложены:

1. **Богатый первичный датасет**, полученный парсингом платформы GOAT — крупнейшего международного маркетплейса для продажи кроссовок. Это обеспечивает репрезентативность ($\approx$ 1000 моделей) и высокую вариативность признаков (бренд, материал, технология подошвы, гендер, уникальность, время с релиза и т. д.).
2. **Многоуровневая методология анализа**:
   - лог-преобразование цены с учётом лог-нормального распределения;
   - робастная OLS-оценка с поправкой Уайта на гетероскедастичность;
   - тесты Бокса–Кокса и Рамсея для выбора функциональной формы и проверки спецификации;
   - квантильная регрессия для выявления неоднородных эффектов на разных уровнях цен.
3. **Акцент на интерпретируемости**: коэффициенты переведены в эластичности и процентные изменения, что облегчает использование результатов.

Наша модель позволяет:

- прогнозировать рыночную цену новой или вторичной пары с 95 % доверительным интервалом;
- количественно оценивать вклад конкретных характеристик (бренд, технология *Air*, редкость расцветки, гендер и т. д.) в премию или дисконт;
- оптимизировать стратегию закупок и ценообразования для маркетплейсов, дистрибьюторов и ресейлеров;
- служить основой для автоматизированных решений (API-оценщик, динамический прайсинг).

Товар и данные

Для нашего проекта мы выбрали строить модель ценообразования для такого предмета обуви, как кроссовки. Данные по кроссовкам взяли с сайта имностранной ритейл платформы `goat.com`. В таблице 1 ниже представлены характеристики, которые у нас получилось достать, с их описаниями.

Характеристика	**Описание**
Бренд	Nike, Adidas, Puma и прочее
Модель	Текстовое название модели
Дата выхода	Дата первого выпуска кроссовок в формате yyyy-mm-dd
Количество цветов	Числовой признак ширины доступной цветовой линейки
Основной цвет	Категориальный признак, основной цвет кроссовок
Материал верха	Материал, из которого изготовлен верх кроссовок (кожа, нубук и прочее)
Технология	Технология изготовления кроссовок
Количество размеров	Ширина доступной размерной сетки
Уникальность	Флаг уникальности/редкости пары кроссовок
Цена	Цена пары кроссовок в евро

Table 1: Характеристики кроссовок, спаршенные с сайта

С помощью даты выхода пары кроссовок мы создали новый признак – количество дней, прошедших с даты релиза (на момент 3 мая 2025). Так мы получили дополнительный численный признак.

Описательные статистики числовых признаков (и флага уникальности) представлены в таблице 2. Сразу видно, что в цене есть много экстремально больших значений. Также из интересного – около 7% кроссовок в выборке являются уникальными.

Изначально мы спарсили только данные из одного раздела, в результате чего получили, что из 683 пар только 35 были женскими. Мы посчитали такой перекос нерепрезентативным, так как хотелось оценить разницу в ценах между женскими и мужскими кроссовками. Поэтому мы решили отдельно спарсить еще только женские кроссовки, в результате получив 645 мужских пар и 384 женские пары. Более наглядно распределение представлено на рисунке 1.

	len_colorway	len_size	unique	price	days_since_release
count	1029	1029	1029	1029	1029
mean	4.14	34.5	0.07	1646.88	4401.12
std	1.63	7.07	0.26	11816.96	3210.68
min	0	8	0	35	2
25%	3	26	0	135	1729
50%	4	39	0	316	3845
75%	5	39	0	613	6697
max	11	41	1	300589	14732

Table 2: Описательные статистики

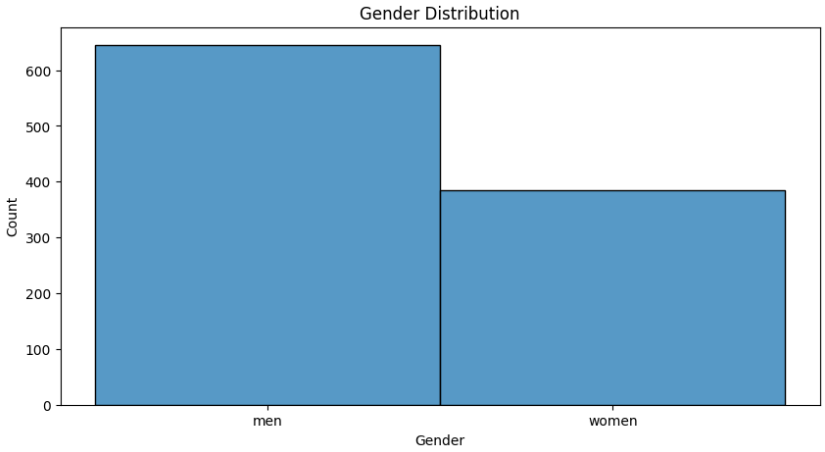

Figure 1: Распределение пар кроссовок по полу

Самая важная характеристика для нас – это цена кроссовок. Как и многие денежные факторы, цена пары кроссовок распределена скорее лог-нормально, поэтому сразу было принято решение логарифмировать ее. Распределение логарифмированной цены представлено на рисунке 2. Как можно заметить, в выборке присутствуют наблюдения, являющиеся выбросами по цене. Однако мы решили не выкидывать данные наблюдения, так как они необязательно будут выбросами с точки зрения самой модели.

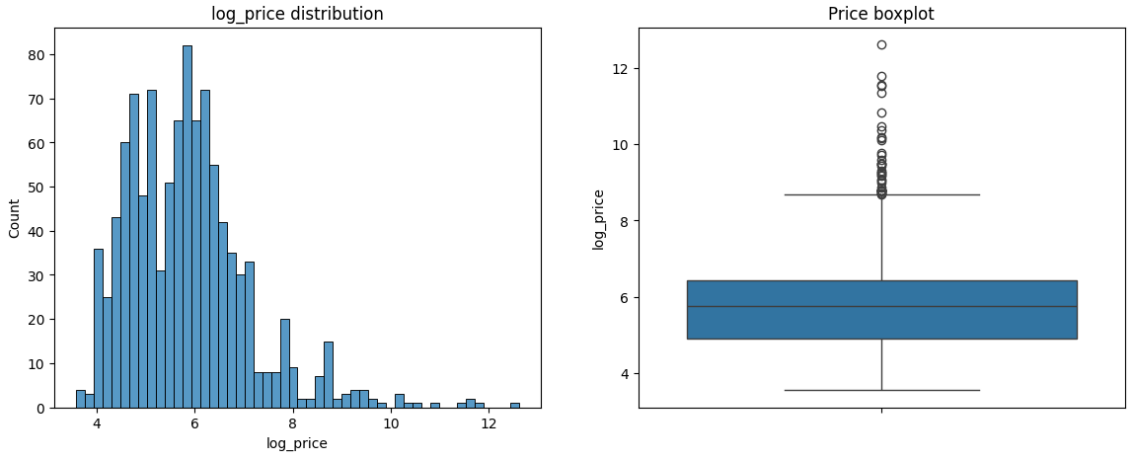

Figure 2: Распределение логарифмированной цены

Для выявления потенциальных связей между признаками и ценой была построена корреляционная тепловая карта (рисунок 3) для числовых признанков (у нас их всего три, кроме цены).

Несложно заметить, что логарифмированная цена достаточно сильно коррелирует с количеством дней с релиза и шириной размерной сетки. Эти признаки могут стать достаточно хорошими регрессорами.

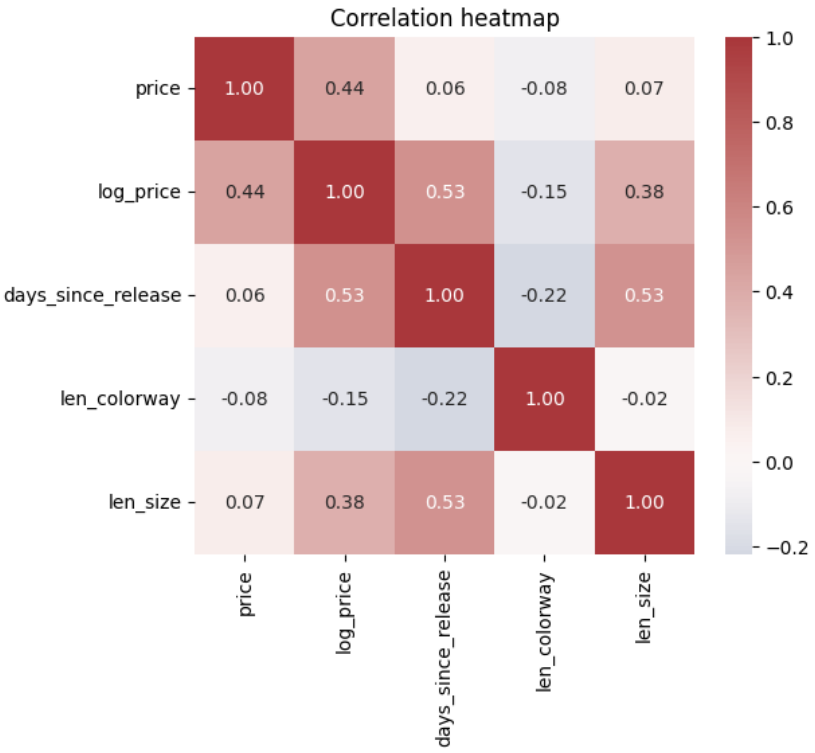

Figure 3: Корреляционная карта

В наших данных есть несколько категориальных признаков с достаточно большим количеством категорий. Кодировать каждый – не лучшая практика, поэтому мы решили посмотреть на распределение пар кроссовок по каждой из категорий, чтобы сгруппировать какие-то из них или выделить наиболее важные. Распределения брендов, цветов, материалов и технологий представлены на рисунках 4, 5, 6 и 7 соответственно.

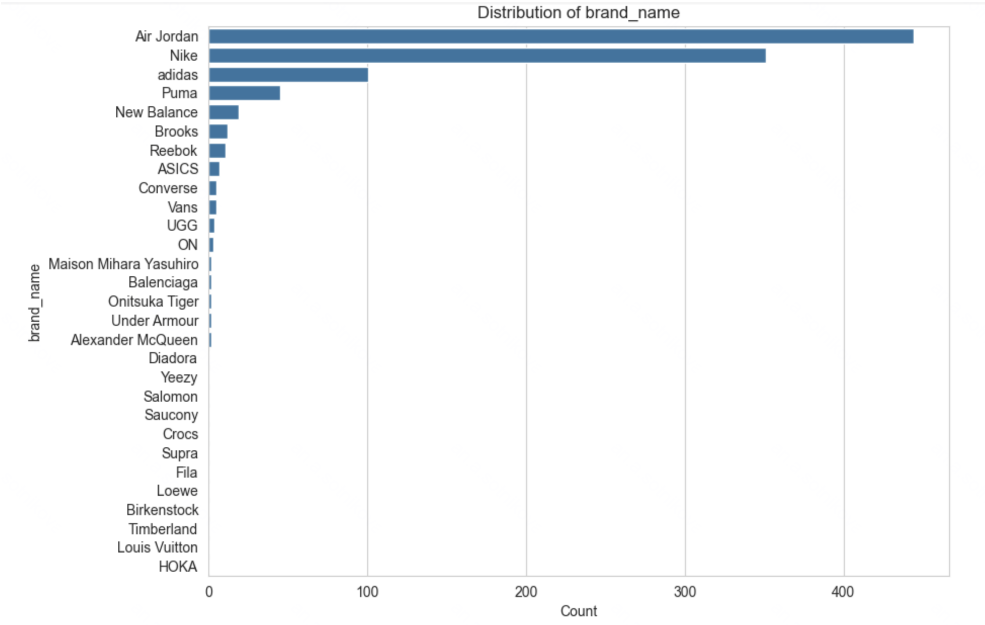

Figure 4: Распределение брендов кроссовок

Несколько брендов сильно выделяются по количеству наблюдений, в качестве регрессоров

можно использоавть дамми-переменные на самые популярные бренды кроссовок.

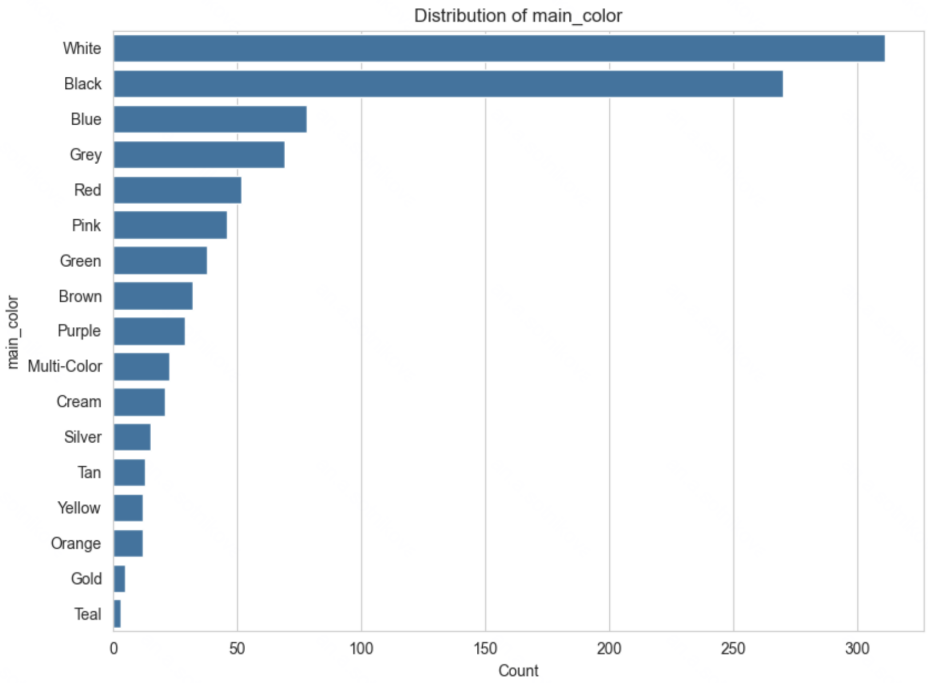

Figure 5: Распределение основных цветов кроссовок

Белый и черный – самые частые цвета кроссовок, что не удивительно. Эти цвета можно выделить в отдельный регрессор. Также можно сделать дамми-перменную на яркие цвета кроссовок (розовый, фиолетовый и прочее), так как, например, более яркие кроссовки могут быть более редкими/особенными, что напрямую влияет на их цену.

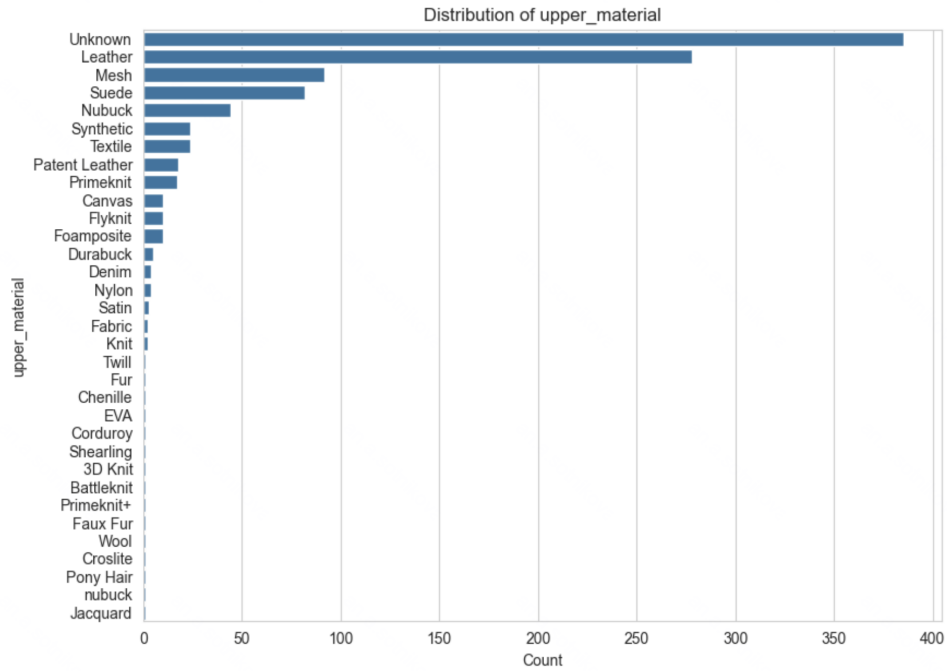

Figure 6: Распределение материалов верха кроссовок

Сложно выделить что-то особенное в плане материалов, можно сконцентрироваться на влиянии использования более редких материалов.

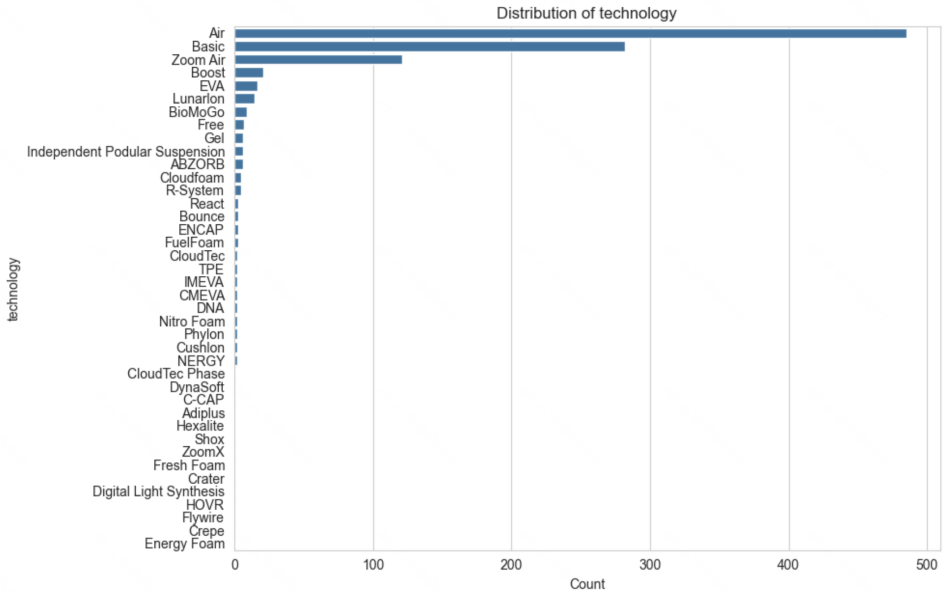

Figure 7: Распределение технологий изготовления кроссовок

В плане технологий преобладают `Air`-технологии. Можно сделать дамми-переменные на этот тип технологий или на те технологии, которые используются реже всего.

Чтобы предварительно оценить, какие регрессоры могут быть важными при включении их в модель мы построили графики ящики с усами для распределения логарифмированной цены в зависимости от наших бинарных переменных (рисунок 8). Заметим, что в зависимости от бренда цена на кроссовки может быть как выше (Nike), так и ниже (остальные бренды). Также женские кроссовки стоят заметно меньше мужских, а редкость технологии изготовления скорее удешевляет кроссовки. А вот наличие Air-технологии – наоборот делает кроссовки более дорогими. Кажется, что все эти переменные будут неплохими регрессорами для модели.

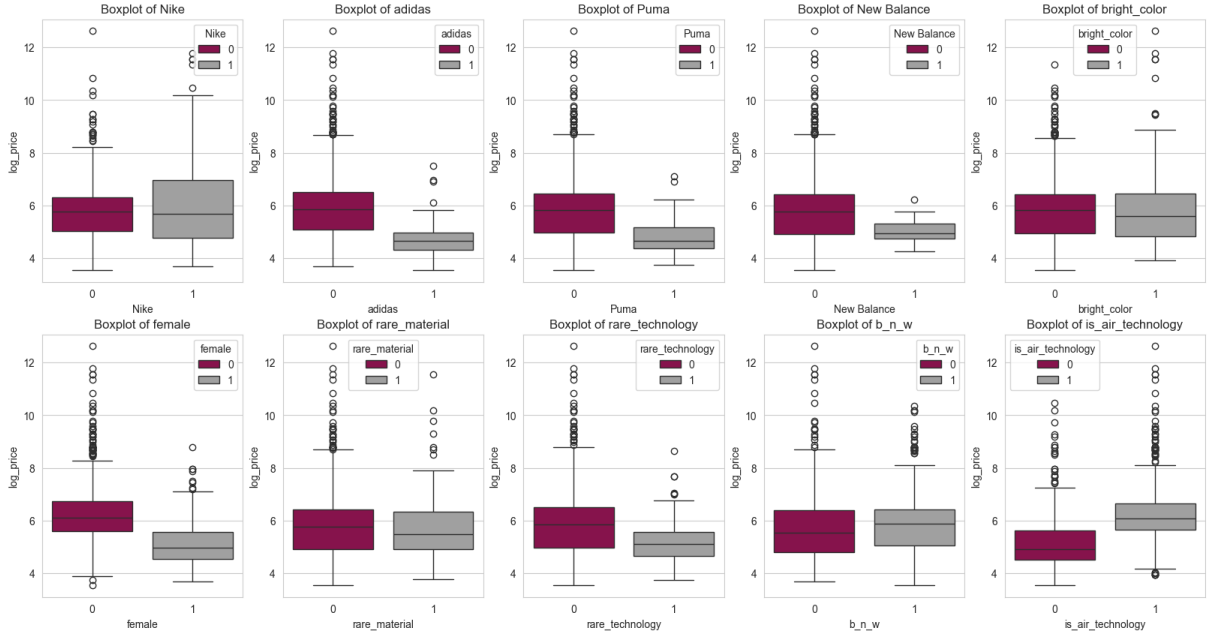

Figure 8: Распределение лог-цены в зависимости от дамми-переменных

Базовая модель ценообразования

Функциональная форма

Мы решили проверить несколько фукнциональных форм для нашей модели. В каждой из моделей есть общие регрессоры, это: константа, ширина цветовой палитры (len_colorway), ширина размерной сетки (len_size), уникальность пары (unique), количество дней с релиза (days_since_release), дамми-переменные на топ-5 брендов (Air Jordan, Nike и проч.), дамми переменная на женские кроссовки (female). Зависимой переменной выслупает логарифмированная цена.

Поскольку таргетная переменная $log(price)$ то при изменении X_j на , цена кроссовка изменится на 1, $e^{\beta_j \Delta X_j} - 1 = e^{\beta_j} - 1$

В первой модели мы решили добавить такие регрессоры, как дамми-переменные на яркий основной цвет кроссовок (bright_color) и на редкость используемой технологии (rare_technology). Результаты оценки такой регрессии представлены ниже.

Количество наблюдений	1 029
Рассчитан R^2	0.384
Скорр. R^2	0.377
F-статистика	55.82 ($p < 0.001$)
Лог-правдоподобие	−1449.7
AIC / BIC	2927 / 2997
Тип ковариационной матрицы	HC0 (robust)

Модель в целом значима, о чем нам говорит значение F-статистики и ее p-value. Однако коэффициенты перед добавленными нами регрессорами оказались незначимыми.

Существенные предикторы при $\alpha = 0.05$ (и их интерпретация):

- `unique` — увеличение лог-цены на 0.563 (+56%).
- `days_since_release` — +0.0002 ($\approx$ 0.02 % за +день).
- `Nike` — +0.365 (+36% к базовому бренду).
- `female` — −0.674 (−49%).

Уравнение регрессии

$$\begin{aligned}\widehat{\ln(\text{price})} &= 5.4636 - 0.0162\,\text{len\_colorway} - 0.0073\,\text{len\_size} \\ &\quad + 0.5627\,\text{unique} + 0.0002\,\text{days\_since\_release} \\ &\quad + 0.1631\,\text{Air Jordan} + 0.3648\,\text{Nike} - 0.1337\,\text{Puma} \\ &\quad + 0.0155\,\text{New Balance} - 0.2796\,\text{adidas} + 0.1440\,\text{bright\_color} \\ &\quad - 0.6741\,\text{female} + 0.2024\,\text{rare\_material} - 0.0891\,\text{rare\_technology}\end{aligned}$$

В следующей функциональной форме мы решили изменить переменные, связанеые с цветом кроссовок и их технологией. Вместо предыдущих регрессоров мы добавили дамми-переменные на черный/белый основной цвет (так как это одни из самых базовых цветов) и на Air-технологию (так как она является достаточно популярной). Результаты оценки представлены ниже, модель в целом значима (p-value около 0), а также значимы коэффициенты перед новыми признаками.

Количество наблюдений	1 029
Рассчитан R^2	0.392
Скорр. R^2	0.384
F-статистика	61.01 ($p < 0.001$)
Лог-правдоподобие	−1443.7
AIC / BIC	2913 / 2978
Тип ковариационной матрицы	HC0 (robust)

Существенные предикторы при $\alpha = 0.05$:

- `unique` — увеличение лог-цены на 0.522 (+68%).
- `days_since_release` — +0.0002 за день (≈0.02%).
- `b_n_w` (чёрно-белая расц.) — −0.200 (−18%).
- `female` — −0.618 (−46%).
- `is_air_technology` — +0.353 (+42%).

Уравнение регрессии

$$\begin{aligned}\widehat{\ln(\text{price})} &= 5.5958 - 0.0283\,\text{len\_colorway} - 0.0064\,\text{len\_size} \\ &\quad + 0.5217\,\text{unique} + 0.0002\,\text{days\_since\_release} \\ &\quad - 0.0873\,\text{Air Jordan} + 0.2294\,\text{Nike} - 0.1673\,\text{Puma} \\ &\quad - 0.1060\,\text{New Balance} - 0.2355\,\text{adidas} - 0.2004\,\text{b\_n\_w} \\ &\quad - 0.6175\,\text{female} + 0.3529\,\text{is\_air\_technology}\end{aligned}$$

Данная функциональная форма дала нам более высокий скорректированный R^2 (0.384 против 0.377).

Во время изучения данных мы заметили, что средняя лог-цена может имметь квадратичную зависимость от len_colorway, поэтому в слеудющей функциональной форме решили добавить к предыдущим регрессорам еще квадрат от len_colorway. Ниже представлены результаты оценки модели. Модель в целом значима.

Количество наблюдений	1 029
Рассчитан R^2	0.392
Скорр. R^2	0.384
F-статистика	56.50 ($p < 0.001$)
Лог-правдоподобие	−1443.4
AIC / BIC	2915 / 2984
Тип ковариационной матрицы	HC0 (robust)

Существенные предикторы при $\alpha = 0.05$:

- `unique` — $\beta = 0.522$ $\Rightarrow$ цена выше на $\approx 69\%$.
- `days_since_release` — $\beta = 0.0002$ $\Rightarrow$ +0.02 % за день.
- `b_n_w` (чёрно-белая расц.) — $\beta = -0.199$ $\Rightarrow$ цена ниже на $\approx 18\%$.
- `female` — $\beta = -0.616$ $\Rightarrow$ цена ниже на $\approx 46\%$.
- `is_air_technology` — $\beta = 0.354$ $\Rightarrow$ цена выше на $\approx 42\%$.

Уравнение регрессии

$$\begin{aligned}\widehat{\ln(\text{price})} &= 5.6871 - 0.0818\,\text{len\_colorway} + 0.0059\,\text{len\_colorway2} - 0.0063\,\text{len\_size} \\ &\quad + 0.5221\,\text{unique} + 0.0002\,\text{days\_since\_release} \\ &\quad - 0.0777\,\text{Air Jordan} + 0.2337\,\text{Nike} - 0.1650\,\text{Puma} \\ &\quad - 0.1029\,\text{New Balance} - 0.2242\,\text{adidas} - 0.1985\,\text{b\_n\_w} \\ &\quad - 0.6160\,\text{female} + 0.3535\,\text{is\_air\_technology}\end{aligned}$$

Заметим, что скорректиованный R^2 не улучшился, а коэффициенты перед len_colorway и len_colorway^2 оказались незначимыми. В таком случае мы решили остановиться на наборе регрессоров из второй функциональной формы.

Для зависимой переменной мы сразу выбрали лог-цену, так как это соответствует здравому смыслу, однако считаем нужным проверить, точно ли модель с логарифмированной зависимой переменной лучше. Для этого мы провели тест Бокса-Кокса с преобразованием Зарембки. Тестовая статистика оказалась равна 3631, что сильно больше критического значения для распределения хи-квадрат и 5% уровня значимости. Значение RSS сильно меньше у модели с лог-таргетом, поэтому со спокойной душой выбираем ее.

Table 3: Экономический смысл коэффициентов в лучшей модели

Фактор	β	e^{β}	Изм. цены	**Комментарий**
unique	0.522	1.68	+68 %	Лимитированная/уникальная пара редка на рынке, поэтому за неё платят почти на две-трети больше.
is_air_technology	0.353	1.42	+42 %	Наличие фирменной «воздушной» подошвы повышает воспринимаемую ценность, добавляя 40 % к цене.
b_n_w	–0.200	0.82	–18 %	Минималистичная ч/б расцветка считается менее «хайповой», поэтому её переплата ниже на 18 %.
female	–0.618	0.54	–46 %	Спрос на женские размеры меньше, что почти вдвое снижает ожидаемую цену на вторичном рынке.
days_since_release	0.0002	1.0002	+0.02 %	Каждый день после релиза пара в среднем слегка дорожает (скудное предложение, растущий «ностальгический» спрос).

Проверка предпосылок

Перед проверкой предпосылок мы избавиилсь от незначимых коэффициентов и переоценили модель, чтобы работать с более короткой и интерпретируемой моделью.

Первая предпосылка ТГМ, которую мы решили првоерить, – предпосылка о гомоскедастичности ошибок (т.е. $\sigma_i^2 = \sigma_j^2 = \sigma^2 \quad \forall i, j$). Был построен график предсказанные значения-остатки (рисунок 9), по которому можно предположить о наличии гетероскедастичности в наших данных. Для получения подтверждения был проведен тест Бройша-Пагана ($H_0 \colon \sigma_i^2 = \sigma_j^2 = \sigma^2 \quad \forall i, j, \quad H_1 \colon \exists i \colon \sigma_i^2 \neq \sigma^2$), который подтвердил наши опасения на 5% уровне значимости (тестовая статистика = 30.26, p-value = 0).

Предпосылка о гомоскедастичности не выполнена, то есть дисперсии ошибок отличаются, что ведет к тому, что оценки стандартных ошибок коэффициентов регрессии смещены, оценки метода наименьших квадратов неэффективны, t-статистики коэффициентов неадекватны, следовательно мы неправильно делаем вывод о значимости коэффициентов.

Для решения этой проблемы мы заранее при оценке моделей использовали робастные стандартные ошибки в форме Уайта HC0. Эта поправка влияет только на способ расчёта матрицы ковариаций оценок коэффициентов (и, следовательно, их стандартных ошибок, t-статистик и p-value), но не на сами коэффициенты и не на остатки. Поэтому выводы о значимости коэффициентов были сделаны без возможных заблуждений из-за гетероскедастичности.

Следующая предпосылка – отсутствие мультиколлинеарности. В целом эта проблема не должна нас затронуть, так как мы имеем достаточно много наблюдений, однако мы все равно проверили VIF для наших регрессоров (таблица 4). Все значения VIF меньше 10, так что предпосылка об отсутствии мультиколлинеарности выполнена.

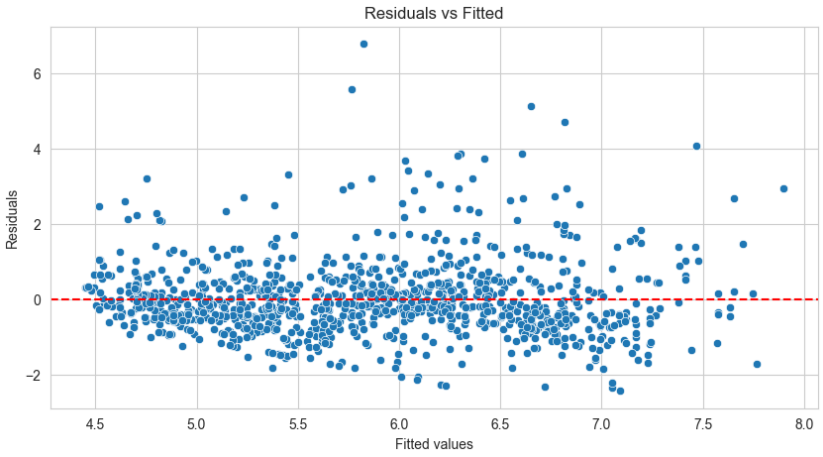

Figure 9: График предсказанные значения-остатки

Регрессор	**VIF**
unique	1.01
days_since_release	1.28
b_n_w	1.06
female	1.53
is_air_technology	1.58

Table 4: Значения VIF для регрессоров

Третья предпосылка – корректность спецификации модели. Что касается формы зависимой переменной – это было проверено выше, модель с логарифмированной ценой значительно лучше. Для проверки наличия пропущенных регрессоров воспользуемся тестом Рамсея с одним вспомогательным регрессором. Для этого оценивается вспомогательная регрессия: $y = X\beta + \alpha\hat{y}^2 + \eta$. Гипотезы: $H_0 : \alpha = 0, H_1 : \alpha \neq 0$.

Тестовая статистика теста оказалась равна 0.16, что соответствует p-value = 0.69, исходя из чего мы не отвергаем нулевую гипотезу об отсутствии пропущенных регрессоров на уровне значимости 5%. Соответственно, спецификация нашей модели корректна.

Далее проверим остатки нашей модели на нормальность распределения. Для этого воспользуемся тестом Харке-Бера. Гипотезы: $H_0 : S = 0, K = 3, H_1 : S \neq 0, K \neq 3$. Тестовая статистика оказалась равна 1721, что соответствует p-value = 0, откуда можем сделать вывод, что остатки распределены не нормально. Это может быть связано с наличием выбросов в нашей модели. Для определения выбросов мы воспользовались DFFITS. Исключив 81 влиятельное наблюдение, мы переоценили модель. В результате мы получили $R^2_{adj} = 0.53$, что является сильным улучшением относительно предыдущих моделей. Далее мы заново провели тест Харке-Бера: тестовая статистика стала меньше: теперь она равна 26, однако нулевая гипотеза о нормальности распределения остатков все еще отвергается. Скорее всего дело в коэффициенте ассиметрии (он сильно больше 0), так как коэффициент эксцесса оказался достаточно близким к 3.

В целом нормальность остатков является важной для проверки различных гипотез, но не для корректной оценки модели. В случае ненормальности распределения остатков все еще можно проверять гипотезы, используя бутстрап.

Модель квантильной регрессии

В процессе изучения данных и оценки базовой модели множественной регрессии у нас появилось несколько гипотез, для проверки которых может быть полезна модель квантильной регрессии:

1. В финальной модели коэффициент при уникальности пары кроссовок оказался значимым. Наша гипотеза заключается в том, что уникальность кроссовок сильнее влияет на цену более дорогих кроссовок.

2. Также значимым и отрицательным оказался коэффициент при дамми-переменной на белый/черный основной цвет кроссовок. Гипотеза следующая: белый и черный цвет – самые базовые для кроссовок, поэтому этот фактор почти не влияет на более дешевые кроссовки, но оказывает сильное влияние на более дорогие пары (так как более дорогие кроссовки как правило более необычные).

Мы оценили квантильную регрессию для квантилей 0.1, 0.25, 0.5, 0.75 и 0.9. С помощью бустрап-процедуры были получены оценки дисперсий коэффициентов. На рисунке 10 представлены графики зависимости значений коэффициентов от квантилей.

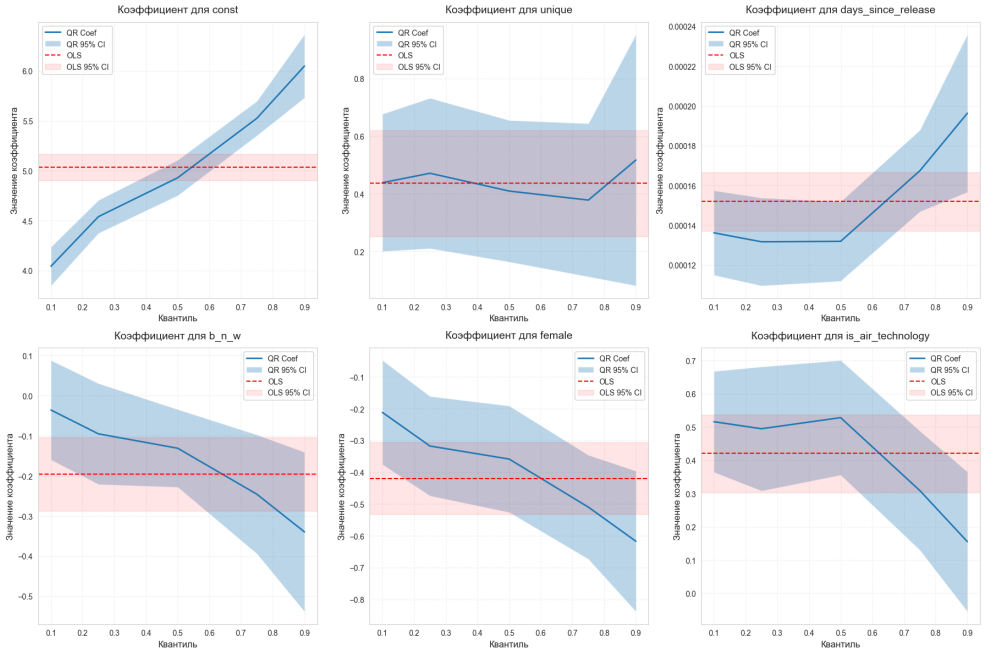

Figure 10: Графики значений коэффициентов в зависимости от квантилей

На глаз уже можно сказать, что скорее всего первая наша гипотеза не подтвердится (коэффициент при unique находится почти на одном уровне), а вторая не будет отвергнута (виден ниспадающий тренд). Однако проверим это формально, используя следующую статистику, имеющую асимптотически нормальное стандартное распределение:

$$t = \frac{coef_{q_1} - coef_{q_2}}{\sqrt{se(coef_{q_1} - coef_{q_2})}} \sim (asy)N(0,1)$$

Для каждого из случаев гипотезы выглядят так (будем сравнивать именно квантили 0.1 и 0.9):

$$H_0 : coef_{0.9} - coef_{0.1} = 0$$

$$H_1 : coef_{0.9} - coef_{0.1} \neq 0$$

1. Для первой гипотезы тестовая статистика равна -0.3099, p-value = 0.7567. Соответственно, на уровне значимости 5% мы не отвегаем нулевую гипотезу о равенстве оценок коэффициентов для квантилей 0.1 и 0.9. Следовательно, фактор уникальности примерно одинаково влияент на цену как дорогих, так и дешевых кроссовок.

2. Для второй гипотезы тествая статистика равна 2.5417, p-value: 0.0110. Соответственно, на уровне значимости 5% мы не отвегаем альтернативную гипотезу о наличии статистически значимых различий между оценками коэффициентов при b_n_w. Получается, что белый/черный цвет кроссовок действительно сильнее удешевляет более дорогие кроссовки. Также коэффициент при b_n_w оказался незначимым для квантиля 0.1, то есть также подтверждается гипотеза и том, что цвет в целом не влияет на более дешевые кроссовки.

Прогноз по модели

Теперь придумаем свой товар и спрогнозируем цену на него. Пусть мы хотим узнать, сколько стоили бы мужские кроссовки, выпущенные ровно год назад (365 дней), в белом цвете, не являющиеся уникальными и произведенные без Air-технологии.

Так как наша модель не прошла тест на нормальность остатков, то для получения корректных доверительных интервалов мы применили бутстрап. В итоге получили следующий прогноз и 95%-ый доверительный интервал для него:

$$\widehat{price} = 133.7 \text{ euro}$$

$$price \in (117.43, 152.71)$$

Заключение:

В процессе работы мы спарсили данные, исследовали их на наличие закономерностей и потенциальных факторов для нашей модели (провели feature-engineering). Проанализировали поведение таргет-переменной и привели ее в логарифмированный вид для работы с лог-нормальным распределением, которому часто подчиняются цены.

Далее мы оценили несколько моделей гедонистического ценообразования кроссовок и выбрали наиболее лучшую по высоте скорректированного R^2_{adj}, проверили выполнения предпосылок теоремы Гаусса-Маркова, чтобы понять с каким эконометрическими ограничениями мы столкнемся в процессе работы. Получили итоговую спецификацию и на основе нее сделали предсказание для произвольной пары обуви, которое оказалось достаточно адекватной и соответствующей реальным ценам на рынке.

Наконец, мы построили модель квантильной регрессии, для которой проверили гипотезы о важности уникальности кроссовок и их основного цвета. Мы построили графики зависимостей значений коэффициентов от квантилей и формально проверили наши гипотезы.

По результатам анализа мы получили интерпретируемые и согласующиеся со здравым смыслом результаты, например, женские кроссовки стоят в среднем на 46 % дешевле, уникальность пары делает ее дороже в среднем на 68 %, а базовые цвета (черный и белый) наоборот удешевляют на 18 %.

Данное исследование можно развивать дальше, взяв более сложные факторы, обработав данные с других популярных сайтов (Poison, StockX), чтобы найти более интересные закономерности, которые могли бы помочь для анализа текущего рынка кроссовок бизнесу. В целом, данный рынок очень богат интересными паттернами для изучения и развивается с каждым годом все больше и больше, поэтому является очень перспективным для эконометрических исследований и, конечно, для методов машинного и глубинного обучения.