Предсказание гедонистической ценовой функции смартфонов на российском рынке

Ссылка на гитхаб: https://github.com/amtevs/phone-price-prediction

Тема

Данная работа посвящена построению гедонистической ценовой функции для смартфонов, продающихся на российском рынке. В качестве данных были использованы характеристики и цены смартфонов на сайте `https://www.mvideo.ru`.

Сбор данных

Для сбора данных был разработан парсер на языке Python, осуществляющий запросы к внутреннему API сайта `https://www.mvideo.ru`.

Парсер последовательно обходит список идентификаторов товаров (`productId`) и для каждого из них отправляет GET-запрос по адресу `/bff/product-details`. Для получения цены товара отдельно отправляется GET-запрос по адресу `bff/products/prices` Запросы сопровождаются необходимыми HTTP-заголовками и cookies, чтобы имитировать поведение браузера.

Из полученного ответа извлекаются следующие сведения:

- название товара (`name`),
- бренд (`brandName`),
- полный набор характеристик (`properties.all`),
- цена товара со скидкой и без скидки.

Каждая запись сохраняется в формате JSON в файл. Между запросами реализованы случайные паузы, чтобы избежать блокировок со стороны сервера. Всего была получена информация о 1000 смартфонах.

После парсинга получается перекрестная (пространственная) выборка объектов с характеристиками:

- **name** - наименование товара.
- **brand** - компания производитель товара.
- **basePrice** - цена товара в рублях.
- **salePrice** - цена товара со скидкой или цена товара, по которой в текущий момент продается модель телефона в рублях.
- **Гарантия** - количество лет, в течение которых на устройство предоставляется гарантийное обслуживание.

- **Страна** - страна производитель товара, возможны расхождения с фактическим местом сборки из-за наличия производственных линий в разных странах.
- **Год релиза** - год выпуска модели телефона в продажу
- **Состояние** - новый или восстановленный, восстановленным является б/у устройство, которое было отремонтировано и выставлено на продажу.
- **Экран (Дюймы)** - диагональ экрана.
- **Безрамочный** - бинарная переменная, принимает значение 1, если экран занимает почти всю лицевую поверхность устройства, и 0 — если присутствуют заметные рамки.
- **Разрешение экрана** - количество пикселей по горизонтали и вертикали.
- **Технология экрана** - Технология получения изображения, на которой построен экран девайса. На данный момент выделяют TFT-, OLED-, AMOLED- и IPS-дисплеи, которые различаются по яркости, качеству цветопередачи, углам обзора и энергопотреблению.
- **Частота обновления экрана (Гц)** - это количество раз в секунду, с которым экран обновляет изображение (герцы).
- **Яркость (кд/кв.м)** - указывает на то, какой световой поток способен обеспечить дисплей.
- **Тип процессора**, **Тип графического ускорителя** - наименования установленных центрального и графического процессоров.
- **Количество ядер (шт)** - количество ядер процессора. переменная влияет на производительность процессора и возможность одновременной обработки нескольких потоков данных.
- **Память (ГБ)** - основная память телефона, которую можно использовать на приложения, фотографии, видео и т.д.
- **Количество основных камер (шт), Количество фронтальных камер (шт)** - количество камер, расположенных на задней/передней части смартфона.
- **Основная камера (МПикс), Фронтальная камера (Мпикс)** - разрешение в мегапикселях основной и фронтальной камер соответственно,
- **Разрешение видеосъемки (Пикс)** - максимальное разрешение видеосъёмки, поддерживаемое основной камерой.
- **Цифровой зум (х), Оптический зум на увеличение (х)** - во сколько раз изображение можно увеличить с помощью цифровой обработки или за счёт физической конструкции линз соответственно.
- **Съемка видео в портретном режиме, Оптическая стабилизация** - бинарные переменные, принимающие значение 1, если соответствующая технология присутсвует в смартфоне и 0 иначе.
- **Поддержка симкарт** - допустимое количество одновременно используемых SIM-карт и их формат (например, nano-SIM).
- **Поддержка стандартов** - информация о поколениях мобильных сетей (например, 2G/3G/4G), с которыми совместимо устройство.

- **Поддержка Wi-Fi, Технология NFC, Сенсор распознавания лица** - бинарные переменные, принимающие значение 1, если соответствующая технология присутсвует в смартфоне и 0 иначе.
- **Сканер отпечатка пальца** - при наличии указывается его расположение (например, на экране, сбоку, сзади).
- **Материал корпуса** - основные материалы, использованные в конструкции корпуса устройства (например, пластик, стекло, металл).
- **Степень защиты (IPXY)** - обозначение международного стандарта защиты корпуса от пыли (X) и влаги (Y).
- **Блок питания, Кабель, Чехол** - бинарные переменные, принимающие значение 1, если предмет указан в комплектации смартфона и 0 иначе.
- **Мощность блока питания (Вт)** - мощность, с которой зарядное устройство передаёт энергию в смартфон.
- **Емкость аккумулятора (мАч)** - параметр, указывающий на возможность аккумулятора прибора поддерживать его автономную работу, чем выше ёмкость, тем дольше смартфон работает автономно, однако увеличивается и время полной зарядки.
- **Габаритные размеры (В*Ш*Т/В*Ш*Г мм)** - высота, ширина и толщина (или глубина) устройства, измеренные в миллиметрах.
- **Вес (г)** - вес устройства в граммах.

Некоторые характеристики, такие как поддержка Apple Pay или наличие сканера LiDAR (характерные для устройств Apple), являются уникальными для отдельных моделей или производителей. Поскольку эти параметры либо присутствуют только у ограниченного числа устройств, либо часто вовсе не указываются в описании, они были исключены из анализа из-за значительного количества пропущенных значений (NaN) и невозможности получить полную информацию по всему набору данных.

Анализ описательных статистик и графический анализ переменных

Для анализа были доступны две переменные: basePrice (заявленная цена без скидки) и salePrice (фактическая цена продажи). На рис.1 видно, что распределения цен визуально похожи, но salePrice имеет медиану меньше, что согласуется с логикой. В качестве зависимой переменной для модели была выбрана переменная salePrice, поскольку она отражает реальную стоимость устройства на момент анализа, с учетом всех действующих скидок, акций и особенностей текущего рыночного спроса. В то время как basePrice представляет собой рекомендованную или стартовую цену производителя.

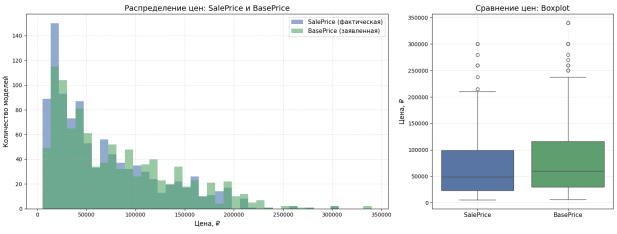


Рис. 1: Сравнение распределений цен

В Таблице 1 представлены описательные статистики зависимой переменной. Можно заметить, что средняя цена значительно выше медианы, что может свидетельствовать о наличии небольшой группы смартфонов, которая дороже большинства. Большинство товаров находятся в более низком ценовом сегменте, с ценами ниже среднего. Диапазон цен на телефоны достаточно широкий, что подтверждается довольно высоким значением стандартного отклонения. Так как минимальная и максимальная цена отличаются в несколько порядков друг от друга, чтобы снизить влияние выбросов, уменьшить стандартное отклонение и сделать распределение ближе к нормальному имеет смысл рассматривать регрессию логарифма цены.

Статистика	Значение
Количество	929
Количество NaN	0
Среднее	62275.987
Стандартное отклонение	51447.065
Минимум	5499.0
25%	21999.0
Медиана (50%)	44999.0
75%	89990.0
Максимум	299999.0

Таблица 1: Описательные статистики для переменной salePrice

Компания-производитель может являться важным фактором, определяющим цену телефона, потому что между компаниями могут отличаться технологии производства, страны производства, средний ценовой сегмент продукции, лояльность и популярность среди людей. Графически на рис. 2 видны существенные различия между средними ценами на телефоны для 7 наиболее часто встречающихся брендов в выборке. Основными, понятными многим, техническими характеристиками телефона являются разрешение экрана, память, количество основных и фронтальных камер, разрешение видеосъемки и фотографий, цифровой и оптический зум, емкость аккамулятора, степень защиты от воды и пыли, поддержка стандартов.

Статистика	Значение
Количество	929
Количество NaN	0
Среднее	10.690
Стандартное отклонение	0.872
Минимум	8.612
25%	9.999
Медиана (50%)	10.714
75%	11.407
Максимум	12.612

Таблица 2: Описательные статистики для логарифма salePrice

Наиболее часто встречающее количество фронтальных камер у телефонов в выборке - 1, остальные значения имеют несущественно малую долю. Также частота встречаемости поддержки любых стандартов, кроме 5G примерно одинаковая, в выборке большинство телефонов поддерживают стандарты 2G, 3G, 4G, LTE. Можно также проверить, что распредление цены телефона не зависит от того, поддерживается ли каждый из этих стандартов. Значимое различие в распределении есть только в поддержке стандарта 5G.

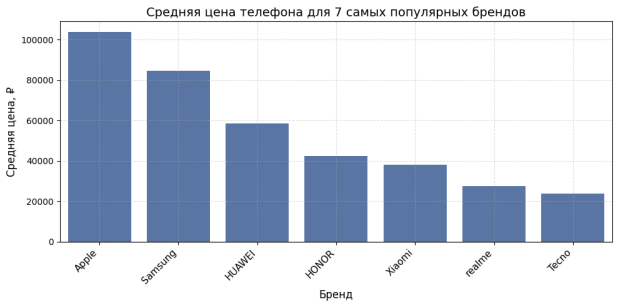


Рис. 2: Различия в средних ценах по брендам

Визуальными характеристиками смартфона являются такие факторы, как тип экрана (безрамочный или нет), материал корпуса, габаритные размеры. Остальные факторы, которые могут влиять на цену телефона: вес, состояние, год релиза. По распределениям остальных характеристик можно понять, что большинство значений для признака ”глубина” смартфона идентичные, есть несколько телефонов с большей глубиной, но их значительно меньше. Также распределение цены в зависимости от того, новым является смартфон или восстановленным, указывает на то, что чаще всего восстанавливают телефоны среднего/дорогого сегмента, так как визуально

медианы двух распределений очень близки друг к другу.

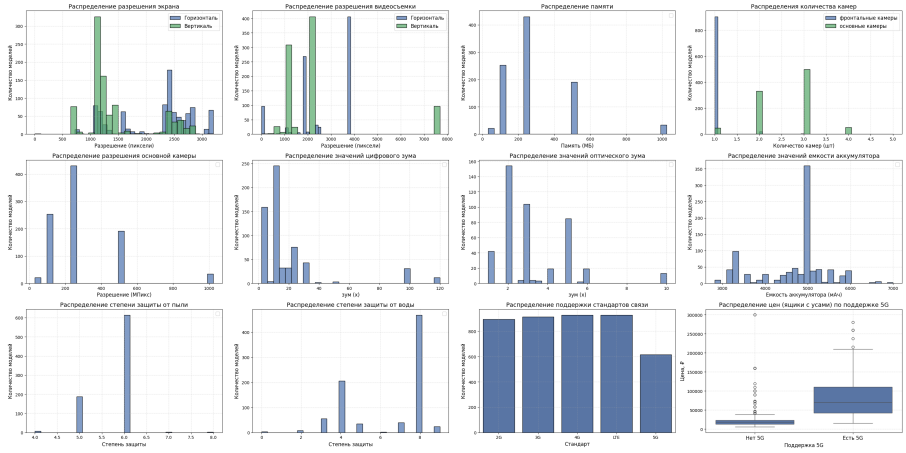


Рис. 3: Распределения технических характеристик

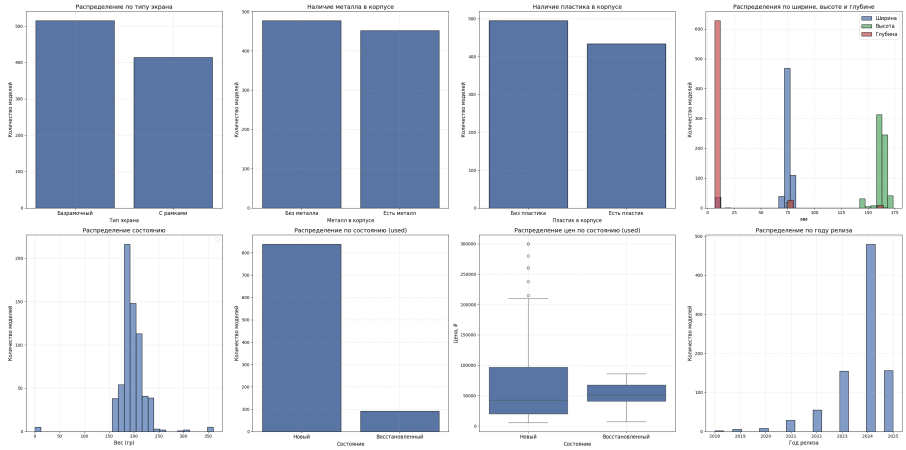


Рис. 4: Распределения остальных характеристик

Обработка пропущенных значений

Пропущенные значения в дамми переменных, таких как безрамочный, состояние, is_5G говорят о том, что особыми характеристиками объект не обладает (онии не указаны на сайте), поэтому они были заполнены по принципу большинства. Пропущенные значения в остальных переменных были заполнены средним значением.

Рис. 5: График остатки-предсказания

Построение основной модели

Построение модели началось с признаков: ’battery’, ’main_cams’, ’memory’, ’height’, ’width’, ’depth’, ’video_resolution_v’, ’video_resolution_h’, ’screen_resolution_h’, ’screen_resolution_v’, ’screen_diag’, ’dust_IP’, ’water_IP’, ’zoom’, ’weight’, ’frame_has_metal’,

’frame_has_plastic’, ’is_new’, ’is_5G’, ’frameless’, ’year’, ’opt_zoom’, ’front_camera_mp_total’, ’main_camera_mp_total’, ’brand_Apple’, ’brand_Samsung’, ’brand_Xiaomi’, ’brand_Tecno’, ’brand_HONOR’, ’brand_HUAWEI’, ’brand_Infinix’.

Несмотря на то, что выборка достаточно большая 929 наблюдений, я все равно обращала внимание на мультиколлинеарность признаков и старалась максимально уменьшить эту проблему.Чтобы проанализировать мультиколлинеарность, использовались два признака: корреляции признаков друг с другом и VIF-фактор. Так, высокая корреляция друг с другом наблюдалась у признаков frame_has_metal, frame_has_plastic ≈ -0.85. Поэтому было принято решение объединить их в одну переменную, отвечающую за прочность: ’strong_frame’ = ’frame_has_metal’ - ’frame_has_plastic’, принимающую значения -1, 0, 1, где -1 означает, что в составе корпуса указан только пластик, а 1 - только металл.

Также у некоторых дамми-переменных на бренды наблюдалась сильная корреляция с battery (Емкость батареи), что говорит о том, что, у большинства моделей телефонов, например, компании Apple, большинство устройств имеют емкость батареи примерно в одном диапазоне. Из-за этого VIF-фактор brand_Apple был больше 10. Производительность устройства в большей степени зависит не от емкости батареи, а он того, как оно тратит энергию при работе, поэтому признак battery было принято решение удалить.

После этого была построена базовая модель, имеющая вид: $y = \beta_0 + X\beta + e$

Чтобы проверить модель на гетероскедастичность, был построен график ’Остатки-Предсказания’. На рис.5 видно, что график освидетельствует о неправильно выбранной функциональной форме. Ранее предполагалось, что логарифмирование зависимой переменной поможет лучше описать данные. Так как нужно сравнить модели с линейной и логарифмированной зависимлй переменной, чтобы определить лучшую, воспользуемся тестом Бокса-Кокса с преобразованием Зарембки.

Рис. 6: График остатки-предсказания

H_0 : Линейная модель лучше описывает данные

H_1 : Полулогарифмическая модель лучше описывает данные

Тестовая статистика получилась равна: 20850.042, p_{value} : 0.0000. Поэтому на любом разумном уровне значимости можно отвергуть гипотезу H_0 в пользу H_1. После логарифмирования зависимой переменной график ’Остатки-Предсказания’ стал похож на облако точек, в котором не наблюдалась зависимости остатков от предсказаний.

В качестве теста на гетероскедастичность применим тест Уайта.

H_0 : Разброс остатков не зависит от прогнозных значений (гомоскедастичность)

H_1 : Разрброс остатков зависит от прогнозных значений (гетероскедастичность)

Тестовая статистика получилась равна 655.9175, p_{value} : 0.0000. Поэтому на любом разумном уровне значимости гипотеза H_0 отвергается в пользу гипотезы H_1 : ошибки гетероскедастичны. Получается, что в модели оценки стандартных ошибок коэффициентов регрессии смещены, оценки МНК коэффициентов регрессии неэффективны, t-статистики коэффициентов регресии неадекватны. Чтобы избавиться от проблем с гетероскедастичностью далее оценим регрессию с поправкой на гетероскедастичность: $HC3$.

Было решено объединить несколько коррелирующих признаков, которые по своей сути похожи друг на друга. Так, video_resolution_v, video_resolution_h коррелируют между собой и один

из них оказывается незначим в итоговой регресии. При этом кажется, что они оба несут в себе важную информацию для ценообразования. Поэтому применим к ним PCA и разложим на две главные компонентны с процентами объясненной дисперсии соответственно: 0.695, 0.305. Аналогично признаки: height, width, depth. При анализе признаков было выявлено, что у большинства объектов значение depth (Глубина) одинаковое. Поэтому эту переменную было принято решение удалить. А height, width по отдельности заменить на полупериметр: P = height + width. Аналогично dust_IP, water_IP - защита от воды и пыли. В зависимости от параметров, один из этих признаков становился незначимым в итоговой модели, при этом между ними есть заметная корреляция и для телефонов они указываются вместе. Поэтому было принято решение аналогично объединить признаки в сумму (IP). Таким образом, VIF некоторых факторов стал еще ниже.

Далее я протестировала спецификацию и функциональную форму модели тестом Рамсея на вторую, третью, четвертую степени признаков.

H_0 : Модель корректно специфицирована (добавление степеней предсказанных значений не улучшает мо

H_1 : Модель некорректно специфицирована (добавление степеней предсказанных значений улучшает мод

Тест Рамсея показал, что модель специфицирована неверно, поэтому изменили функциональные формы некоторых признаков, опираясь на логику и R^2 регресии зависимой переменной на этот признак. Новые добавленные связи: $IP^2, P^2, screen\_resolution\_h^2, log(screen\_resolution\_v)$, $year^2, log(zoom), weight^2, frameless\_year = frameless{\cdot}year, log(memory), brand\_Apple\_memory = brand\_Apple \cdot memory$,
$brand\_Samsung\_front\_camera\_mp\_total = brand\_Samsung \cdot front\_camera\_mp\_total$, $brand\_HUAWEI\_front\_camera\_mp\_total, brand\_Samsung\_main\_camera\_mp\_total$, $brand\_HUAWEI\_main\_camera\_mp\_total$

После этого на любом разумном уровне значимости можно было принять гипотезу о том, что модель корректно специфицирована по тесту Рамсея для каждой из вышеперечисленных степеней.

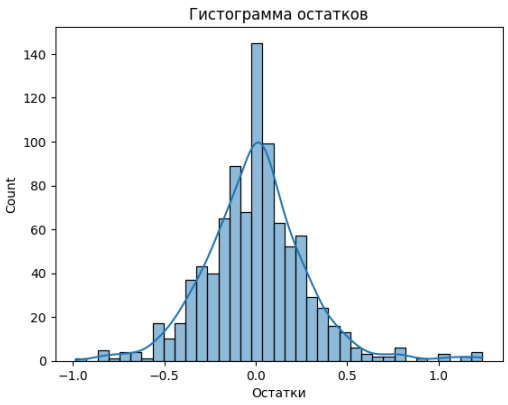
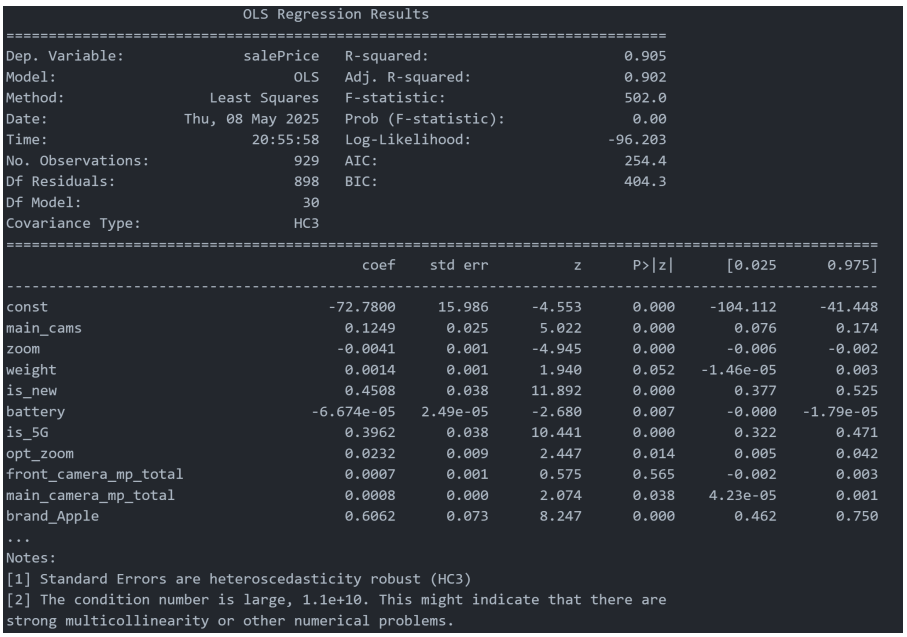


Рис. 7: График остатки-предсказания

Затем были удалены незначимые переменные и протестирована нормальность остатков. По тестам на равенство распределений получили, что на любом разумном уровне значимости гипотеза о нормальности остатков отвергается. При этом визуально распределение остатов близко к нормальному: остатки распределены приблизительно симметрично вокруг нуля, есть легкая остроконечность (высокий пик в центре) и возможные лёгкие ”тяжёлые хвосты” по краям. Предполагаем, что у нас достаточно большая выборка, что нивелирует проблемы связанные с распределением остатков.

Для этой модели предполагаем, что эндогенности нет, потому что данные взяти из надежного источника, поэтому ошибки измерения быть не может, пропущенной переменной также не может быть, потому что изначально взято очень много переменных, признаки, описанные выше односторонне влияют на цену, нет такого, что цена телефона влияла бы на его характеристики. Таккже была попытка взять несколько переменных в качестве инструментов и эндогенность не была обнаружена.

Итоговая модель, в которой не удалены незначимые признаки имеет вид:



OLS Regression Results

Dep. Variable:	salePrice	R-squared:	0.905
Model:	OLS	Adj. R-squared:	0.902
Method:	Least Squares	F-statistic:	502.0
Date:	Thu, 08 May 2025	Prob (F-statistic):	0.00
Time:	20:55:58	Log-Likelihood:	-96.203
No. Observations:	929	AIC:	254.4
Df Residuals:	898	BIC:	404.3
Df Model:	30		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-72.7800	15.986	-4.553	0.000	-104.112	-41.448
main_cams	0.1249	0.025	5.022	0.000	0.076	0.174
zoom	-0.0041	0.001	-4.945	0.000	-0.006	-0.002
weight	0.0014	0.001	1.940	0.052	-1.46e-05	0.003
is_new	0.4508	0.038	11.892	0.000	0.377	0.525
battery	-6.674e-05	2.49e-05	-2.680	0.007	-0.000	-1.79e-05
is_5G	0.3962	0.038	10.441	0.000	0.322	0.471
opt_zoom	0.0232	0.009	2.447	0.014	0.005	0.042
front_camera_mp_total	0.0007	0.001	0.575	0.565	-0.002	0.003
main_camera_mp_total	0.0008	0.000	2.074	0.038	4.23e-05	0.001
brand_Apple	0.6062	0.073	8.247	0.000	0.462	0.750

...

Notes:

[1] Standard Errors are heteroscedasticity robust (HC3)

[2] The condition number is large, 1.1e+10. This might indicate that there are strong multicollinearity or other numerical problems.

Рис. 8: Модель с незначимыми признаками

$F-$статистика = 502.0, $p_{value} = 0$, что говорит о значимости модели в целом. На уровне значимости 5% незначимыми оказываются такие признаки, как main_cams, battery, front_camera_mp_total, main_camera_mp_total, brand_Xiaomi, brand_Tecno, brand_HONOR, log_screen_resolution_v, log_zoom, weight_sq, frameless_year,brand_Samsung_front_camera_mp_total,brand_Samsung_main_camera_mp_total, brand_HUAWEI_main_camera_mp_total.

Модель проанализирована на выбросы с помощью трех тестов.

1. Стьюдентизированные остатки, 2. Точки Левериджа, 3. DFFITS

Выбросом я считала наблюдение, которое обозначается таким в каждом из трех тестов. После удаления 5 выбросов и всех незначимых переменных итоговая модель имеет вид:

$ln(price) = -55.96 + 0.40 \cdot ln(memory) + 0.7 \cdot brand\_Apple + 0.47 \cdot is\_new + 0.292 \cdot strong\_frame + 0.42 \cdot is\_5G + 0.2 \cdot main\_cams + 0.099 \cdot ln(zoom) + 0.000015 \cdot year^2 - 0.0004 \cdot brand\_Apple \cdot memory + 0.00005 \cdot frameless \cdot year + 0.08 \cdot dim\_PC2 - 0.000005 \cdot P^2 + 0.001 \cdot IP^2 - 0.004 \cdot brand\_HUAWEI \cdot front\_camera\_mp\_total + error$

Интерпретация переменных:

1) Увеличение памяти на 1% связано с ростом цены на 0.4%.

2) Устройства Apple дороже на $e^{0.7} - 1 \approx 101\%$ (в два раза) по сравнению с другими брендами при одинаковых характеристиках.

3) Новые устройства дороже восстановленных на $e^{0.47} - 1 \approx 60\%$

4) Увеличение прочности корпуса на 1 единицу (то есть добавление металла в пластиковый корпус, переход с пластикового корпуса на корпус, полностью состоящий из металлов) увеличивает цену в среднем на 0.34%

5) Устройства, поддерживающие 5G в среднем дороже на $\approx 52\%$

6) Влияние года выпуска на цену нелинейное и усиливается с течением времени. с каждым годом прирост цены ускоряется на $0.000015(2year + 1)$

7) Для устройств Apple дополнительный гигабайт памяти увеличивает цену менее значительно, чем для других брендов - на 0.39996%

8) С каждым годом безрамочные устройства (выпущенные в этом году) становятся дороже на 0.005% дополнительно.

9) Рост размера телефона в ширину или высоту в среднем снижает цену, но эффект не очень большой.

10) Увеличение степени защиты телефона от воды или от пыли ведет к росту цены.

11) Увеличение цифрового зума смартфона на 1% приводит к росту цены примерно на 0.099%.

По постороенной модели, примерная цена нового телефона имеющего 512 Гигабайт памяти, неизвестного бренда с металлическим корпусом, поддержкой 5G, защитой от воды и пыли IP68, цифровым зумом в 25 единиц, выпущенным в 2025 году, 65мм в ширину и 150мм в длину, dim_PC2 = 1, 2 основные камеры составит около 47.200 рублей.

Квантильная регрессия

Классическая линейная регрессия предполагает, что коэффициенты при переменных одинаковы для всех наблюдений. Однако в экономике и маркетинге это редко выполняется: покупатели дорогих смартфонов руководствуются другими критериями, чем покупатели бюджетных моделей. Квантильная регрессия позволяет зафиксировать это поведение: оценка коэффициентов проводится отдельно для каждого квантиля, что даёт гораздо более гибкую и точную картину. Так же квантильная регрессия адаптируется к асимметрии распределения цен, устойчива к гетероскедастичности и выбросам. Для построения модели использовались те же переменные, что и в итоговой модели линейной регрессии.

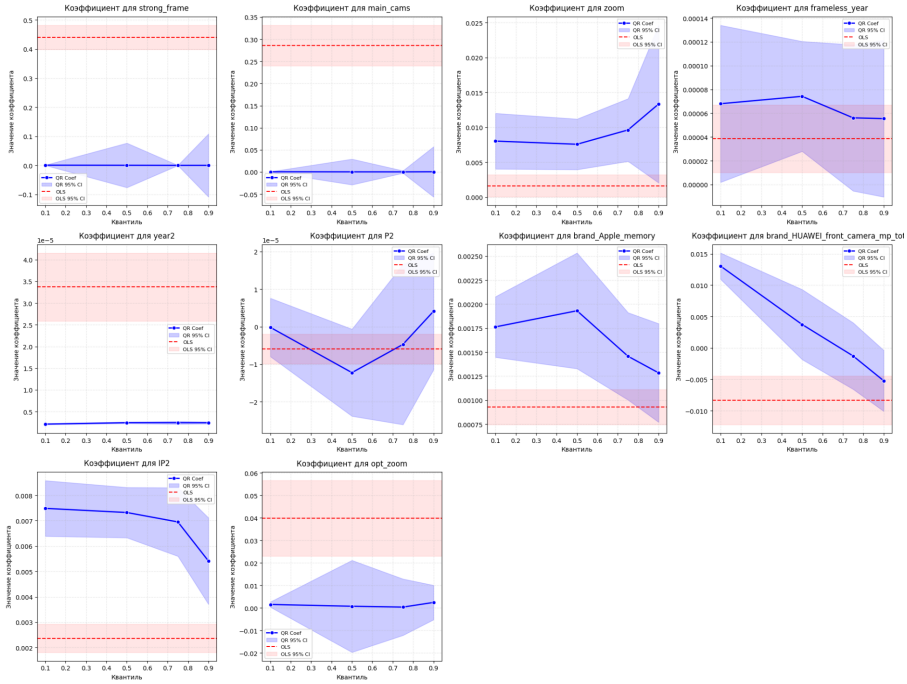
Незначимые переменными оказались ‘brand_Apple‘, ‘brand_Samsung‘, ‘is_5G‘, ‘wireless_charging‘, ‘memory_log‘, ‘is_new‘, ‘strong_frame‘, ‘log_zoom‘, ‘dim_PC2‘

Для бюджетных телефонов (q=0.1) критичными оказались

- Качество фронтальной камеры (+0.0131, p=0.026)
- Защита IP2 (+0.0075)
- Оптический зум (+0.0016)

Для премиальных (q=0.9) ключевыми оказались

- Зум (+0.0132)
- Год выпуска (year2)
- Толщина рамки (frameless_year)



‘zoom‘ монотонно растёт от 0.1 к 0.9 квантилю: влияние зума становится сильнее в дорогих смартфонах.

‘frameless_year‘ умеренно положительный, незначительно уменьшается к высокому квантилю.

‘year2‘ стабилен и почти не изменяются между квантилями.

‘P2‘ показывает чёткую смену знака: отрицательное влияние в низких квантилях и положительное — в верхних.

‘brand_Apple_memory‘ уменьшается с ростом квантиля, но остаётся положительным.

‘brand_HUAWEI_front_camera_mp_total‘ показывает явное уменьшение влияния, в верхних квантилях коэффициент близок к 0.

‘IP2‘ показывает постепенное снижение коэффициента от 0.1 к 0.9.

‘opt_zoom‘ все значения близки к нулю, широкие интервалы.

‘main_cams‘, ‘strong_frame‘ коэффициенты близки к 0 и малозначимы по всем квантилям.

Так же был проведен тест Вальда на равенство коэффициентов между квантилями ’zoom’, ’frameless_year’, ’year2’, ’brand_HUAWEI_front_camera_mp_total’, ’brand_Apple_memory’ и коэффициенты различаются только у ’brand_HUAWEI_front_camera_mp_total’, ’year2’

Рассмотрим еще телефон с

’is_new’ = 1

’brand_Apple’ = 0

’is_5G’ = 1

’memory_log’ = 7.600902459542082

’strong_frame’ = -1

’main_cams’ = 2

’log_zoom’ = 3.433987

’zoom’ = 30.000000

’frameless_year’ = 2024

’year$^{2\prime} = 4096576$

’dim_PC2’= 0.691285

’const’ = 1

’P$^{2\prime} = 56786.8900$

’brand_Apple_memory’ = 0.0

’brand_HUAWEI_front_camera_mp_total’ = 0.0

’IP$^{2\prime} = 196.0$

’opt_zoom’ = 3.0

10-й перцентиль (q=0.1): Телефон будет стоить 35,900

Медиана (q=0.5): Телефон будет стоить 65,200

75-й перцентиль (q=0.75): Телефон будет стоить 95,400

90-й перцентиль (q=0.9): Телефон будет стоить 123,500