Высшая школа экономики

Факультет экономических наук

Департамент прикладной экономики

Проект

по предмету Эконометрика-2

Применение эконометрического и статистического анализа для оценки влияния различных факторов на ценообразование ноутбуков

Выполнено:

Алгазиновым Александром, студентом исследовательского потока

6 августа 2023 г.

Содержание

1 Абстрактно **2**

2 Мотивация **3**

3 Постановка задачи, описание данных, и плана работы **3**

4 Гипотезы и способы их тестирования **4**

5 Обработка и анализ данных **5**

6 Построение линейных моделей и проверка выполнения предпосылок теоремы Гаусса-Маркова **7**

7 Прогнозируем цену на предложенный нами товар **9**

8 Более сложный анализ **10**

9 Заключение и выводы **11**

References **12**

1 Абстрактно

В этой работе мы провели анализ факторов, определяющих цену ноутбуков. В качестве данных использовался готовый датасет, состоящий из реальных объектов. При этом основная часть работы - написание скриптов по статистическому анализу, эконометрическому моделированию, визуализации данных и результатов экспериментов, а также применению методов машинного обучения. Поэтому этот письменный отчет - лишь приложение к коду с описанием мотивации к исследованию, постановкой задачи, интерпретацией основных выводов и результатов, и так далее. В связи с этим, многие детали, таблицы, и графики здесь будут отсутствовать. При этом все это последовательно и максимально подробно изложено в jupyter блокнотах, на основе которых сделан анализ.

2 Мотивация

В современном мире алгоритмы машинного обучения и нейронные сети глубоко применяются в самых разных сферах - от теоретических точных наук до медицины. Они постоянно развиваются и, особенно в последнее время, демонстрируют прекрасные результаты. При этом у этих инструментов есть ряд недостатков, в частности, низкая интерпретируемость. Если в случае с решающим деревом мы еще можем распечатать его структуру и посмотреть, на основе каких комбинаций признаков агрегируются данные и принимаются решения, то, когда речь идет о графах вычислений - мы чаще всего имеем дело с черными ящиками. Поэтому выделяют целую группу задач и исследований, где необходимо прибегать к другим, более интерпретируемым методам. В таких случаях хорошими вариантами являются статистический и эконометрический анализы. Поэтому в этой работе мы как раз хотим рассмотреть случай, когда нам не столько важно сделать хорошие предсказания, сколько выявить закономерности между признаками и таргетом. Помимо этого, мы также проведем графический анализ, протестируем гипотезы, и сделаем содержательные и интерпретируемые выводы. Обычно как классическая, так и различные модифицированные версии линейной регрессии, неплохо справляются с этим, особенно если мы имеем дело с примерно линейной и дифференцируемой задачей.

Мы будем решать задачу на примере ноутбуков, проанализировав факторы, влияющие на их ценообразование. Кажется, что цена ноутбуков неплохо линейно описывается теми или иными признаками. Так как существует много моделей ноутбуков, отличающихся мощностями, размерами, и так далее, иногда выбор между ними может превратиться в по-настоящему сложный и долгий процесс. Поэтому будет полезно разобраться в примерном алгоритме ценообразования товара, оценив модель. Эта модель поможет понять, во сколько примерно обойдется та или иная техника, а также чем можно принебречь в ноутбуке в цели экономии и получения все еще подходящего под определенные цели товара.

3 Постановка задачи, описание данных, и плана работы

В качестве анализируемой переменной мы возьмем цены на ноутбуки различных брендов и конфигураций. Для этого будем использовать готовый датасет с Kaggle.

Датасет состоит из 1300 наблюдений. Каждое наблюдение - отдельный ноутбук, характеристики которого описаны 11 достаточно стандартными признаками, такими как бренд, оперативная память, процессор, и так далее. То есть мы имеем дело с пространственной выборкой. При этом, несмотря на простоту данных, есть несколько особенностей:
1) Почти все признаки закодированы как категориальные. Это касается даже непрерывных величин (например, вес или память товара), которые в большинстве случаев представлены в виде типа объект с указанием единиц измерения.
2) Некоторые признаки коррелируют друг с другом. Мы знаем, что, в случае использования

линейных моделей, это может быть проблемой, поэтому нужно это учитывать в анализе.
3) В данных присутствуют выбросы. Это касается и таргета, и значений признаков. Линейная регрессия чувствительна к выбросам, поэтому лучше их убирать, или, в крайнем случае, фиксировать их диапазоны.
4) В данных зачастую одно и то же называется по-разному (простой пример - macOS и Mac OS). С одной стороны, это логично, так как мы имеем дело с реальными собранными данными. С другой стороны, это необходимо обрабатывать, чтобы не получить много лишних и неинформативных переменных, которые будут только мешать.

Поэтому датасет перед построением модели требует достаточно внимательной и детальной обработки. При этом процесс приведения датасета к пригодному для построения модели виду зачастую связан с тестированием гипотез и анализом данных. Поэтому план работы может описан в следующих пунктах:
1) Выдвижение гипотез для дальнейшего тестирования.
2) Обработка, анализ данных, тестирование гипотез касательно данных, с которыми мы работаем.
3) Построение различных эконометрических моделей, выбор спецификаций и лучших моделей, тестирование гипотез. Предсказание цены товара.
4) Более глубокий анализ с использованием методов машинного обучения.
5) Выводы по итогам работы.

4 Гипотезы и способы их тестирования

Первая часть гипотез состоит из наших предшествующих убеждений. Мы хотим понять, верны ли наши представления о задаче, с которой мы работаем. Сделаем это путем формулирования соответствующих гипотез, впоследствии их протестировав, и сделав выводы. К этим гипотезам относятся:
1) Линейность и дифференцируемость исходной задачи. Мы проверим это с помощью теста на валидность модели. При этом мы хотим, чтобы нулевая гипотеза о незначимости модели отвергалась на маленьких уровнях значимости. Таким образом мы убедимся, что предсказания линейной регрессии являются хорошими оценками настоящих значений цен товара.
2) Нетрудно заметить, что ноутбуки определенных брендов (например, Apple) в среднем стоят дороже ноутбуков других брендов. Поэтому кажется, что категориальные и качественные признаки важнее непрерывных. То есть субъективные факторы (такие как репутация компании, привлекательность ее бренда в целом) преобладают над объективными (такими как процессор ноутбука и оперативная память) при принятии решений о покупке, что позволяет более известным и доверяемым компаниям отразить это на ценах на продукты. Эту гипотезу можно проверить с помощью графического и численного анализа данных. Помимо этого, можно построить разные модели на урезанном количестве признаков, посмотрев, где лучше качество.
3) Некоторые признаки не являются важными сами по себе (например, вес ноутбука). Мы

думаем, что покупатели игнорируют некоторые объективные факторы ноутбука в пользу других, более важных. Такая гипотеза может быть протестирована несколькими способами: t-test, отсутствие улучшения скорректированного R-квадрата при добавлении этой переменной, зануление веса фактора при регуляризации модели.

Далее мы хотим уже проверить гипотезы, для которых у нас нет первоначальных представлений, но тестирование которых кажется интересным объектом для исследования и получения содержательных выводов. К ним относятся:
1) Комбинации признаков важнее, чем их индивидуальный вклад. То есть если включить попарные комбинации признаков в модель, то качество улучшится (либо с поправкой на увеличенное количество признаков, либо без учета индивидуальных признаков). Эта гипотеза, в случае ее подтверждения, может иметь интересную интерпретацию (источник вдохновления).
2) Некоторые признаки сильно коррелируют друг с другом. Это можно проверить с помощью матрицы корреляций, и общепринятым пониманием того, что есть высокая корреляция, а что нет.
3) Как минимум некоторые предпосылки теоремы Гаусса-Маркова неверны (мы будем последовательно проверять гипотезу о верности каждой из предпосылок, используя Python и релевантные эконометрические тесты и методы).
4) Древовидные модели работают лучше, чем линейная регрессия. Это в какой-то степени связано с гипотезой про комбинации признаков. Дело в том, что такие модели (например, случайный лес) закладывают совершенно другую логику и inductive bias. Это позволяет рассматривать более сложные комбинации признаков на основе предикатов и информационных критериев. Помимо этого, с помощью такого анализа мы можем получить больше информации о значимости категориальных признаков (в классической регрессии их достаточно сложно обобщенно интерпретировать).
5) Более сложные линейные модели работают лучше, чем классическая линейная регрессия. Например, квантильная регрессия робастна к выбросам. Возможно, эта модель отработает лучше.
6) Это не совсем гипотеза, но мы хотим узнать, какая спецификация модели дает наиболее хорошие результаты. Возможно, это тоже позволит нам сделать интересные и неочевидные выводы.

В качестве небольшого примечания хочу обозначить, что все эти аспекты будут затронуты в разных частях работы, и не обязательно будет указано, что конкретно тут мы тестируем конкретную гипотезу. Тем не менее все основные результаты и выводы будут обозначены в отдельных секциях.

5 Обработка и анализ данных

Примечание: так как к работе приложен код с подробным пошаговым описанием и комментариями, я не буду подробно расписывать, что тут происходит, а лишь пройдусь по основным

важным аспектам:
1) Мы перевели численные данные в тип float, убрав единицы измерения.
2) Была сделана достаточно серьезная обработка категориальных признаков (убрали лишние слова, привели к одному формату, назвали одинаково признаки, которые были названы по-разному, но означали одно и то же). Зачастую это делалось достаточно костыльными способами (иногда даже написание условий для замены слов на другие вручную). К счастью, мы не первые, кто работает с этим датасетом, поэтому мы воспользовались уже готовыми **скриптами [5]** .
3) Гипотеза о наличии мультиколлинеарности подтвердилась. При этом проблема не оказалась масштабной, поэтому не пришлось понижать размерность или использовать какие-то сложные техники. Было достаточно убрать пару признаков (при этом мы не сильно потеряли в информативности, так как в большинстве случаев эти признаки очень сильно коррелировали друг с другом).
4) Мы избавились от выбросов, при этом стараясь потерять минимум наблюдений.

Интересные наблюдения, которые мы увидели:
1) Разброс цен на ноутбуки различных операционных систем не обязательно связан с количеством наблюдений, представленных среди них (например, у Mac диапазон шире, чем у Chrome, хотя их почти в 3 раза меньше). То есть как минимум для некоторых операционных систем несложно предсказать цену на ноутбук, так как цены лежат в достаточно узком диапазоне.
2) На основе графического анализа мы сделали вывод, что признаки веса и памяти ноутбука не являются важными и полезными для описания цен. В случае с памятью ноутбука этот результат мне показался достаточно неочевидным, и даже слегка контринтуитивным. В дальнейшем мы будем тестировать гипотезу о незначимости этого фактора.
3) При этом другие числовые признаки (такие как объем оперативной памяти и расширение экрана) неплохо описывают цену ноутбука. Производные цены по этим факторам имеют одинаковый знак, то есть может быть полезно рассмотреть, как взаимодействие этих признаков описывает цену (графический анализ подтвердил, что это стоит сделать в рамках исследования).

На момент завершения этого этапа работы, мы имели два обработанных датасета:
1) Полный датасет: 1184 наблюдения, 9 признаков: компания-производитель, тип ноутбука, процессор, объем оперативной памяти, общая память, GPU, операционная система, вес, и разрешение экрана.
2) Урезанный датасет: то же самое за исключением двух признаков: вес модели, и общая память.

6 Построение линейных моделей и проверка выполнения предпосылок теоремы Гаусса-Маркова

Мы начали с построения первых двух моделей, и проверки гипотез на основе них. Одна модель была построена на полном датасете, другая - на урезанном. На урезанном датасете мы потеряли одну тысячную в качестве как на обычном, так и на скорректированном коэффициентах детерминации. Изменения не кажутся значимыми. При этом на полном датасете фактор веса оказался статистически незначимым на любом общепринятом уровне значимости, а фактор памяти, хоть и значим на 5% уровне значимости, имеет отрицательный вес, что непонятно и противоречит здравому смыслу. Поэтому мы решили по разным причинам принять эти два фактора за ненужные, исключив их из анализа. Таким образом мы получили нашу первую модель, начиная с которой мы строили весь дальнейший эконометрический анализ. Значение скорректированного R^2 примерно 0.73, что делает модель значимой (p-value = 0). В процессе тестирования гипотез на значимость некоторых факторов и модели в целом, мы заметили очень сильную мультиколлинеарность, возникшую в результате кодирования категориальных признаков с помощью one-hot encoding. На этом этапе бороться с ней не стали, так как планировали перейти к проверке выполнения предпосылок теоремы Гаусса-Маркова чуть позднее.

Далее мы протестировали гипотезу о том, что субъективные факторы важнее объективных при ценообразовании. Для этого мы оценили две модели - линейную регрессию чисто на dummy переменных бренда, и регрессию на двух непрерывных признаках: вес и разрешение экрана. Вторая модель по значению критерия качества сильно превзошла первую ($R^2_{adj} = 0.548$ против $R^2_{adj} = 0.206$). Это действительно интересный вывод. Получается, что, либо по этим двум признакам уже примерно понятно, какой бренд произвел этот ноутбук, либо ценообразование определяется больше объективными факторами мощности и качества ноутбука, нежели репутацией и популярностью бренда.

После этого мы перешли к тестированию предпосылок теоремы Гаусса-Маркова:
1) Матожидание остатков равно нулю. Эта предпосылка проверяется посредством взятия среднего арифметического остатков. Предпосылка выполнена, значение крайне близко к нулю.

2) Отсутствие мультиколлинеарности. Мы уже заранее знали, что эта предпосылка нарушилась, и скорее были заинтересованы в выяснении причин и попытке устранения проблемы. Причин оказалось две. Во-первых, в результате кодирования некоторые признаки могли совпадать достаточно сильно, вплоть до 100% (например, операционная система MacOs есть только у бренда Apple). Вторую причину продемонстрируем на примере. Если 75% ноутбуков имеют процессор intel 5, или intel 7, то эти два признака явно будут отрицательно коррелированы, так как у ноутбука только один процессор. Устранить проблему мультиколлинеарности удалось с помощью удаления трех признаков из dummy переменных. Для их обнаружения мы использовали матрицу корреляций и VIF.

3) Гомоскедастичность. Эта предпосылка нарушена, причем на любом уровне значимости

(использовали два теста - тест Уайта и тест Бройша-Пагана, **Источник [3]**). Это объясняется тем, что ноутбуки нетрудно разделить на определенные категории. Ожидаемо, что какие-то категории будут прогнозироваться хуже, а какие-то - лучше (как минимум, у разных брендов разные дисперсии цен и количества выбросов). В приложении в виде кода было описано, почему большинство стандартных способов борьбы с проблемой нам не подойдет. Тем не менее мы попробовали нелинейные преобразования, а в бонусной секции сделали анализ с помощью менее чувствительных к подобным проблемам моделей машинного обучения.

4) Автокорреляция ошибок. Мы использовали тест Дарбина-Уотсона **Источник [6],** предпосылка верна (значение статистики почти равно двум, что и ожидается).

5) Некоррелированность признаков и остатков. **В источнике [2]** описано, как это делается; предпосылка выполнена.

6) Нормальность распределения ошибки. Мы проверяли это графически, нет оснований отвергать нулевую гипотезу о нормальности распределения ошибок.

После этого мы перешли к улучшению нашей модели, пробуя различные спецификации. Так как нас в первую очередь интересует интерпретируемость, вариантов не так много. Напомню, что до этого мы строили самую простую версию модели вида $\hat{y} = w_0 + x^T w$. В связи с проблемой мультиколлинеарности, может быть полезно прологарифмировать непрерывные признаки (которых всего два). Из всевозможных нелинейных преобразований выберем именно логарифмирование, так как в этом случае наша модель все еще имеет неплохую интерпретацию. Также имеет смысл попробовать прологарифмировать таргет. В итоге мы суммарно оценим 4 модели:

1) $\hat{y} = w_0 + x^T w$; Эту модель мы уже оценили.

2) $log(\hat{y}) = w_0 + x^T w$; как 1), только мы предсказываем логарифм таргета.

3) $\hat{y} = w_0 + w_R log(x_R) + w_S log(x_S) + \sum_{i \in cat} w_i x_i$; предсказываем таргет, но логарифмируем непрерывные признаки, не трогая категориальные.

4) $log(\hat{y}) = w_0 + w_R log(x_R) + w_S log(x_S) + \sum_{i \in cat} w_i x_i$; как 3), только предсказываем логарифм таргета.

В качестве метрики оценивания качества мы снова использовали скорректированный R^2. В целом, существенно увеличить качество позволило логарифмирование таргета (например, модель 2) имеет R^2_{adj} выше, чем модель 1) на 0.062). Логарифмирование признаков без таргета привело к переобучению (так как изменение диапазонов значений признаков способствовало присвоению больших весов этим признаков, что является классическим определением переобучения в моделях линейной регрессии), следовательно, ухудшению качества. Если прологарифмировать и таргет, и признаки, результаты будут лучше, чем просто логарифмирование таргета. При этом модель 4) мы выбрали скорее из-за более удобной интерпретации (так как непонятно, что имеется в виду под увеличением расширения экрана на 1 единицу, поэтому процентная интерпретация звучит удобнее). В итоге, наша финальная модель имеет вид:

$$log(\hat{y}) = 2.004 + 0.345log(x_R) + 0.448log(x_S) + 0.619I[Apple] + ... + 0.334I[Windows],$$

где $I[.]$ означает индикатор. Например, $I[Apple] = 1$, если это компания Apple, и 0 в любом противном случае. Так как мы работаем с логарифмами, интерпретация представляет собой эластичности. Если увеличить оперативную память на 1%, то цена ноутбука увеличится на 0.345%. Если увеличить расширение экрана на 1%, то цена ноутбука увеличится на 0.448%. В случае с индикаторами, у нас есть какой-то бренд, выбранный в качестве базового (при этом нас не интересует, какой это бренд конкретно, так как мы заинтересованы в получении обобщенных результатов). Коэффициент при индикаторах означает, что, если ноутбук относится к этому бренду, то цена будет на этот вес процентов выше, чем если бы ноутбук относился к базовому бренду. Например, ноутбук Apple будет стоить на 0.619% дороже, чем аналогичный ноутбук базового бренда.

7 Прогнозируем цену на предложенный нами товар

Предлагаю, с одной стороны, проверить адекватность модели на какой-то реальной конфигурации ноутбука, где мы примерно понимаем его цену, а потом уже пофантазировать и придумать какой-то не очень понятный, но очень интересный товар:

1) Сперва возьмем реальный товар - MacBook, стоимостью 900 евро, после чего увеличим оперативную память на 10%. Предсказанная цена получилась почти 1350 евро, что говорит о том, что мы явно перепредсказали цену. После этого мы посмотрели распределение цен на ноутбуки от компании Apple. Оказалось, мы пытались спрогнозировать самый дешевый товар от бренда с увеличенной оперативной памятью. Наш прогноз слегка превысил медианное значение ноутбуков Apple (1300 евро), и получился меньше среднего (около 1450 евро). Поэтому это значение можно считать скорее выбросным, а прогноз все еще адекватным.

2) Затем мы взяли случайный ноутбук, стоимостью 1500 евро, уменьшили его оперативную память в 5 раз, и увеличили разрешение экрана в 3 раза. Прогноз получился около 1550 евро, что имеет смысл, так как коэффициент при разрешение экрана больше, чем коэффициент при оперативной памяти, то есть модель должна воспринимать ноутбук с высоким показателем расшерения экрана как дорогой товар.

В общем и целом, хоть результаты прогноза не всегда соответствовали ожидаемым результатам, они обоснованы, и могут быть объяснены реальным распределением данных. Эксперимент в лишний раз подтвердил необходимость попробовать построить более робастную к выбросам модель, и сравнить ее с нашей. В таком случае, нулевой гипотезой выступает то, что в модели присутствуют выбросы, которые сложно распознать в явном виде и автоматизированно почистить, не потеряв большого количества данных.

8 Более сложный анализ

Здесь мы пользовались библиотекой sklearn. В ней встроенный класс линейной регрессии использует уже не метод наименьших квадратов, а градиентный спуск для минимизации целевой функции потерь MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2.$$

Это значит, что выбросы (если они есть) будут вносить существенный вклад в оптимизацию, так как большая ошибка на них приведет к сильному увеличению значения функционала потерь. Поэтому мы решили протестировать гипотезу о том, что выбросы негативно влияют на обучение модели. Для этого мы обучили квантильную регрессию с параметром quantile = 0.5, что эквивалентно обучению линейной регрессии на loss-функции MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|.$$

То есть в таком случае сильная ошибка на каком-то объекте уже не является такой критичной, как в случае с MSE. Поэтому модель более робастна к выбросам.

Мы предварительно разделили выборку на обучающую и тестовую. Сравнивать все модели из данной секции мы будем по результатам предсказаний тестовых данных на трех / четырех метриках: $R^2, MAPE, MSE, MAE$.

Квантильная регрессия показала более хорошее качество на всех метриках, что подтвердило нашу гипотезу о негативном влиянии выбросов на качество модели. Несмотря на то, что классическая линейная регрессия оптимизирует MSE, качество на MSE у нее все равно оказалось хуже, чем у квантильной регрессии.

После этого мы обучили случайный лес на дефолтных параметрах. Целью было, с одной стороны, получить дополнительную интерпретацию, и, с другой стороны, проверить гипотезу, что в случайном лесу заложен более правильный inductive bias, что помогает получать более качественные результаты при построении модели. На тот момент лучшее качество давала квантильная регрессия, поэтому сравнивать результаты решили с ней. Показатели значений коэффициента детерминации и средней квадратичной ошибки у случайного леса лучше, в то время как средняя процентная ошибка хуже. Это значит, что лес лучше ловит взаимосвязи и редко ошибается сильно, в то время как в среднем ошибается чуть больше, чем регрессия. Так как значение R^2 у леса значительно выше, чем у квантильной регрессии (0.829 против 0.732) при примерно 2% потерях в MAPE (19.6 против 17.7), наверное, лес является более подходящей моделью в контексте нашего анализа. При этом нужно иметь в виду, что этот вывод нельзя сделать по объективным причинам, тут мы скорее смотрим на соответствие модели нашим ожиданиям. Преимущество дерева в том, что диапазон значений его предсказаний не может выйти за диапазон тех значений таргета, которые он получил на

обучающей выборке. То есть мы не получим неадекватно высокие и неадекватно низкие значения.

С помощью библиотеки shap мы смогли получить топ-20 самых популярных признаков с детализацией, в каких случаях эти признаки являются важными для модели. Самыми важными факторами являются оперативная память, тип ноутбука, разрешение экрана, и процессоры. То есть бренды, которые изначально казались важными, не входят даже в топ-10 (за исключением HP).

Еще одно преимущество леса связано с тем, что мы можем кодировать категориальные признаки с помощью чисел. Обычно из-за этого ухудшается качество, так как LabelEncoder предполагает свойства чисел (например, если Apple закодировано как 1, а HP - как 2, то мы вводим предпосылку, что Apple $<$ HP, что противоречит здравому смыслу). Тем не менее из-за этого мы можем сравнить важность категориальных признаков с важностью непрерывных факторов, использовав feature importance. Вариантов для оценивания важности признаков может быть много. Один из самых распространенных - MDI (**Подробнее - в источнике [9]**). В принципе, результаты построения графика важности признаков сопоставимы с тем, что мы видели до этого. При этом, судя по масштабу, оперативная память сильно важнее, чем даже второй по важности признак, не говоря уже про остальные. Кажется, это связано с тем, что у этого признака не такой большой диапазон значений, и не так много уникальных значений. Поэтому с помощью этого признака можно делить объекты на категории - более дорогие, и более дешевые, соответственно.

9 Заключение и выводы

Подведем итоги на основе гипотез, которые мы выдвинули и в дальнейшем протестировали:

1) Задачу можно аппроксимировать как линейную и дифференцируемую. Несмотря на то, что предпосылки теоремы Гаусса-Маркова выполнились не все, модель линейной регрессии показала достойные результаты, о чем свидетельствуют хорошие метрики качества.

2) Некоторые непрерывные признаки оказались статистически незначимыми, либо продемонстрировали странную зависимость таргета от них (например, при увеличении памяти ноутбука, его цена падает). От таких признаков мы избавились.

3) Мы дважды столкнулись с проблемой мультиколлинеарности. Первый раз - когда проводили первичный анализ. Затем - когда получили dummy переменные. С небольшими потерями в виде информативности датасета, проблему удалось устранить. Более сложно дела обстоят с гетероскедастичностью. Единственный доступный нам способ решения проблемы - логарифмирование данных, то есть изменение интерпретации моделей в сторону эластичностей. Такой подход действительно помог улучшить качество. Так как другие инструменты борьбы с гетероскедастичностью не соответствуют специфике нашего анализа - не имело смысла снова проверять, есть ли у нас гетероскедастичность, или нет.

4) Непрерывные признаки оказались сильно важнее категориальных (несмотря на то,

что их всего два - они вносят гораздо большую информативность, и с помощью них можно оценить статистически значимую модель).

5) Классическая модель линейной регрессии оказалась слишком простой для данных, с которыми мы работаем. Это говорит о том, что мы смогли несколько раз улучшить качество, делая несложные преобразования, и меняя модели на более сложные. Сперва мы прологарифмировали таргет и признаки - это привело к улучшению целевой метрики R^2_{adj}. После этого мы использовали квантильную регрессию - по сути поменяли функцию потерь, чтобы сделать модель менее чувствительной к выбросам. В результате такого незначительного изменения мы смогли улучшить все четыре рассматриваемые метрики. В конце концов, изменение фундамента моделирования - переход от линейного типа моделей к древовидным, тоже помогло улучшить качество. Мы смогли рассматривать более сложные комбинации признаков, что позволило модели выявить более сложные нелинейные связи, и использовать их в прогнозе.

6) Все результаты подтверждены наглядными графиками, таблицами, и цифрами. Мы часто пренебрегали ими в этом отчете, так как целью отчета является краткое представление структуры проекта, основных действий, и содержательных выводов. При этом все это очень подробно представлено в коде, приложением к которому является этот текст.

Список литературы

[1] *Друженелюбная эконометрика,* ссылка на книгу

[2] *Проверка выполнение предпосылок ТГМ,* ссылка на статью

[3] *Гомоскедастичность и как ее проверять,* ссылка на лекцию

[4] *Ноутбуки и их цены,* ссылка на датасет

[5] *Что определяет цену ноутбуков,* ссылка на анализ

[6] *Тест Дарбина-Уотсона,* ссылка на лекцию

[7] *Библиотека shap для интерпретации результатов ML-моделей,* ссылка на статью

[8] *Квантильная регрессия,* ссылка на статью

[9] *Feature Importance,* ссылка на статью