Анализ ценообразования ноутбуков

Александр Алгазинов

Высшая школа экономики

15 мая 2023

Все, что нужно знать

Постановка задачи

Мы хотим понять, что определяет цену ноутбуков, и насколько точно это можно смоделировать. Для этого мы будем использовать эконометрический анализ и инструменты машинного обучения

Описание данных

Возьмем готовый датасет с Kaggle. Содержит 1300 наблюдений, которые описываются 11 признаками. Особенности: пространственная выборка, содержит выбросы, требует тщательную предобработку

План работы

Выдвижение гипотез; обработка и анализ данных; тестирование гипотез и построение линейных моделей; проведение более глубокого анализа; подведение итогов исследования

Гипотезы

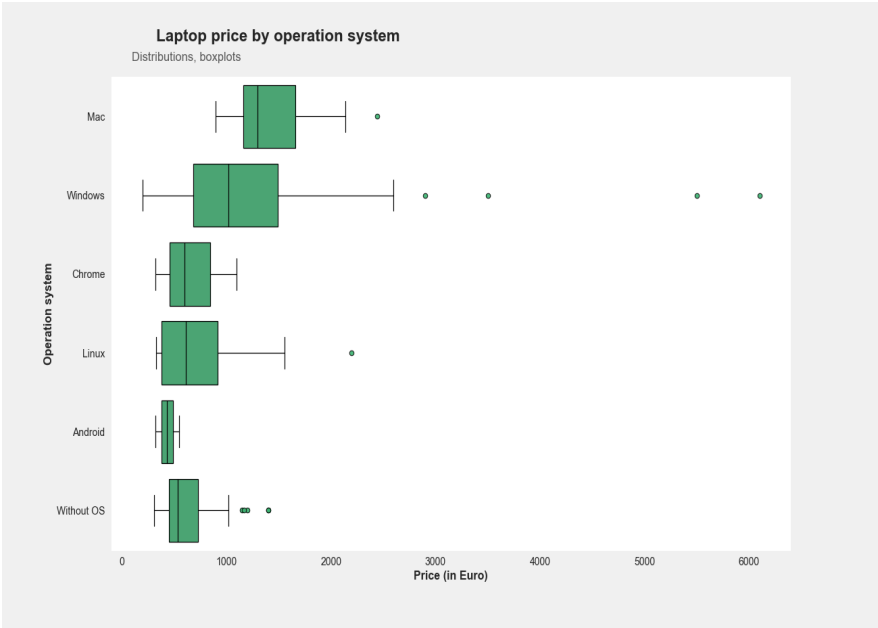
То, во что мы верили изначально

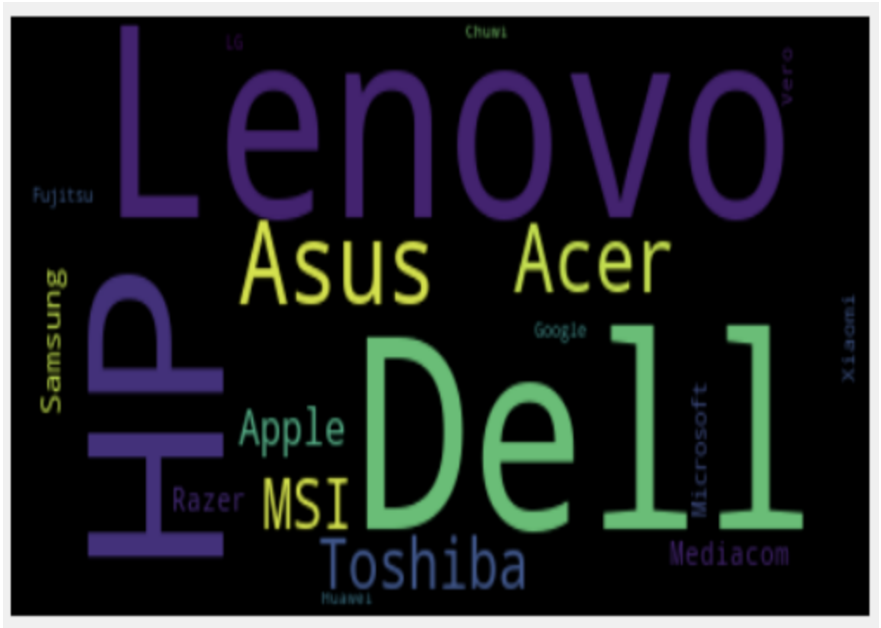
- 1. Линейность и дифференцируемость исходной задачи
- 2. Категориальные и качественные признаки важнее непрерывных
- 3. Некоторые признаки не являются значимыми

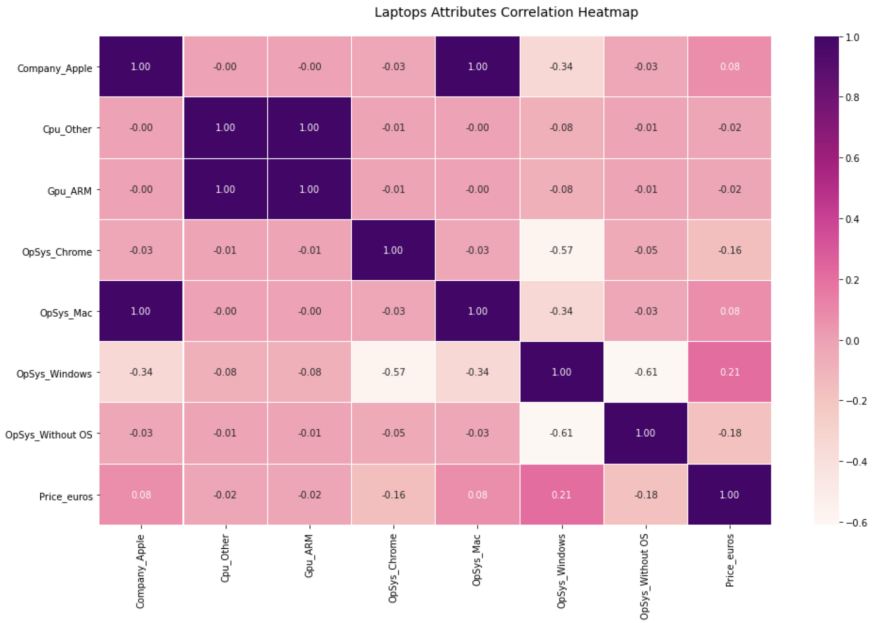
Более абстрактные, но все еще интересные идеи

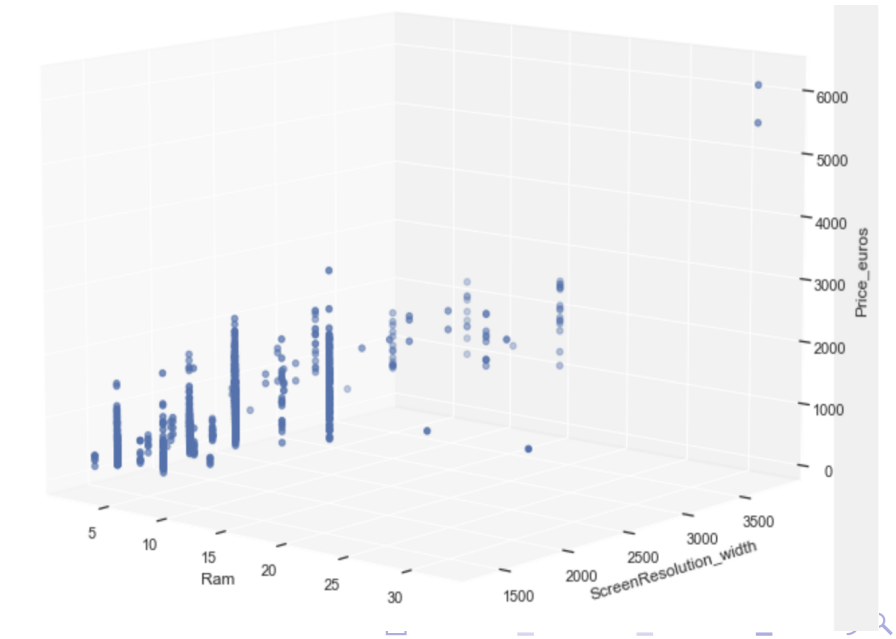
- 1. Комбинации признаков улучшают модель
- 2. Неверность некоторых предпосылок ТГМ
- 3. Inductive bias линейных моделей не очень удачный
- 4. Более робастные линейные модели работают лучше

Анализ


Laptop price by operation system
Distributions, boxplots
Mac
Windows
Chrome
Linux
Android
Without OS
Operation system
0
1000
2000
3000
4000
5000
6000
Price (in Euro)


Lenovo
HP
Dell
Asus
Acer
Apple
MSI
Razer
Toshiba
Samsung
Fujitsu
LG
Chuwi
Vero
Google
Microsoft
Xiaomi
Mediacom
Huawei


Laptops Attributes Correlation Heatmap
Company_Apple
Cpu_Other
Gpu_ARM
OpSys_Chrome
OpSys_Mac
OpSys_Windows
OpSys_Without OS
Price_euros
1.00 -0.00 -0.00 -0.03 1.00 -0.34 -0.03 0.08
-0.00 1.00 1.00 -0.01 -0.00 -0.08 -0.01 -0.02
-0.00 1.00 1.00 -0.01 -0.00 -0.08 -0.01 -0.02
-0.03 -0.01 -0.01 1.00 -0.03 -0.57 -0.05 -0.16
1.00 -0.00 -0.00 -0.03 1.00 -0.34 -0.03 0.08
-0.34 -0.08 -0.08 -0.57 -0.34 1.00 -0.61 0.21
-0.03 -0.01 -0.01 -0.05 -0.03 -0.61 1.00 -0.18
0.08 -0.02 -0.02 -0.16 0.08 0.21 -0.18 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Price_euros
6000
5000
4000
3000
2000
1000
0
Ram
5
10
15
20
25
30
ScreenResolution_width
1500
2000
2500
3000
3500

Результаты тестирования гипотез

Предпосылки ТГМ

- 1. ✗Отсутствие мультиколлинеарности (избавились удалением признаков, критерий - VIF > 10 + матрица корреляций)
- 2. ✗Гомоскедастичность (группы объектов, тесты Уайта и Бройша-Пагана, p-value $= 0$)
- 3. ✓Остальные предпосылки выполнены

Другие наблюдения

- 1. Модель значима (p-value $= 0$, $R^2_{adj} = 0.727$)
- 2. Статистически незначимые признаки: Weight (p-value $= 0.486$), Memory (маленький и отрицательный коэффициент)
- 3. Непрерывные признаки важнее категориальных ($R^2_{adj} = 0.548 > R^2_{adj} = 0.206$)
- 4. Комбинации непрерывных признаков не важны (переменная Ram x Screen не улучшила результаты)

Моделирование (МНК)

Мы оценили четыре модели:

- 1. $\hat{y} = w_0 + x^T w; R^2_{adj} = 0.728$
- 2. $log(\hat{y}) = w_0 + x^T w; R^2_{adj} = 0.790$
- 3. $\hat{y} = w_0 + w_R log(x_R) + w_S log(x_S) + \sum_{i \in cat} w_i x_i; R^2_{adj} = 0.710$
- 4. $log(\hat{y}) = w_0 + w_R log(x_R) + w_S log(x_S) + \sum_{i \in cat} w_i x_i; R^2_{adj} = 0.796$

Основные выводы:

- 1. Лучше логарифмировать таргет
- 2. Оценивание модели 3. привело к переобучению
- 3. Выбираем модель 4. из-за удобства интерпретации

$$log(\hat{y}) = 2.004 + 0.345 log(x_R) + 0.448 log(x_S) + 0.619 \mathbb{1}[A] + ... + 0.334 \mathbb{1}[W]$$

Содержательная интерпретация приведена в отчете

Более сложный анализ

Мы разделили выборку на обучающую и тестовую, после чего обучили три модели, измерив качество на тесте

Модель	**Функция потерь**	**MSE**	R^2	**MAPE**
LinReg	MSE	102239	0.732	19.298
QuantReg ($\tau = 0.5$)	MAE	81723	0.786	17.715
Random Forest	MSE	65115	0.829	19.597

Таблица: Сравнение моделей на цифрах

Нас интересует способность модели понимать, что происходит в целом, поэтому Random Forest $\succ$ QuantReg $\succ$ LinReg. Оценка МНК оказалась слишком простой, но все еще хорошей для моделирования задачи

Выводы по итогам работы

- Задача линейна и дифференцируема, несмотря на частичное невыполнение предпосылок ТГМ
- Данные имеют сложную структуру (мультиколлинеарность, разделение объектов на группы)
- Результаты чувствительны к изменениям априорного представления о задаче
- Не все признаки статистически значимы. Более того, некоторые даже мешают оценивать модель
- Непрерывные признаки существенно важнее категориальных

Спасибо за внимание!