Trường Đại học Công nghệ Thông tin

Lớp DS108.P21

Giảng viên hướng dẫn:
TS.Nguyễn Gia Tuấn Anh – CN.Trần Quốc Khánh

HỆ THỐNG HỎI ĐÁP HỖ TRỢ TƯ VẤN LAPTOP CHO NGƯỜI DÙNG

- Nguyễn Trần Ngọc Ty 23521758
- Đoàn Thị Ngọc Anh 23520042
- Nguyễn Thị Minh Tiến 23521579

NHÓM 22

CONTENT

01 Data Collection

02 Data transformation
Data flow

03 Modelling

04 Deploy

05 Conclusion

01 DATA COLLECTION

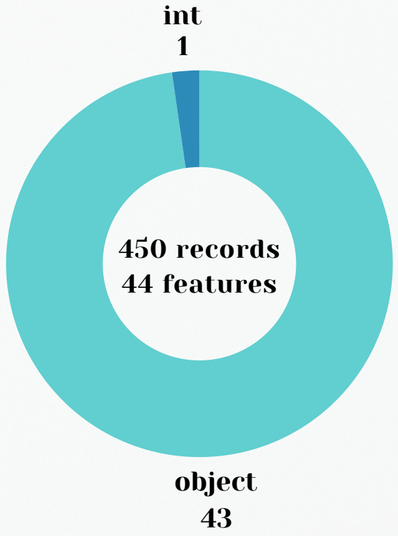
Data sources

Thegioididong.com (Scrapy)
cellphones.com.vn (API + Scrapy)

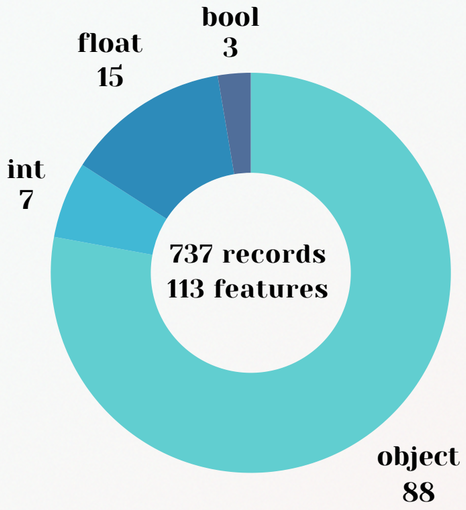
Data extraction

Page source → CSV files

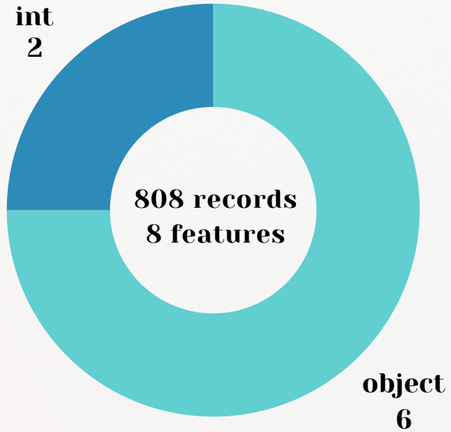
Thegioididong Laptop



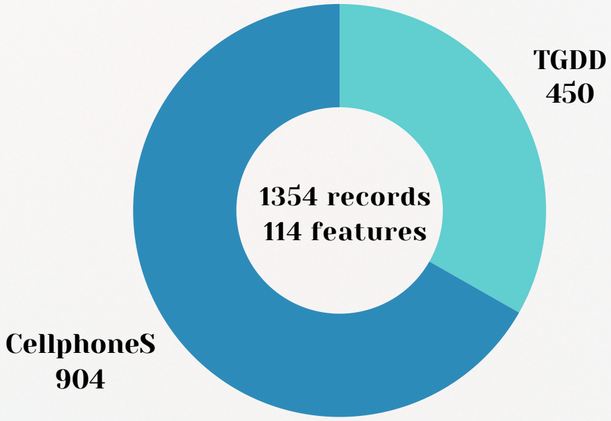
CellphoneS Laptop



CellphoneS Laptop Variant



Total



02 DATA TRANSFORMATION

Data Parsing

Raw, unstructured or semi-structured

↓

Structured data with clear columns and values

Ingestion

Separated files from 2 different sources

↓

Fully combined file

1st EDA

- Understand data
- Detect data issues
- Explore distributions
- Analyze relationships

Data Cleaning

Raw or messy dataset

↓

Cleaned, and consistent dataset

2nd EDA

- Re-examine distributions
- Validate cleaning results
- Feature selection
- Identify predictive variables

Pre-Processing

Cleaned and structured dataset

↓

Machine-learning-ready dataset

2.1

DATA PARSING

114 → 52 features

- Keep
- Delete
- Not sure

Regex

- Simple structure
- Clear patterns

LLMs

Complex, loosely structured text

Rule-based

Hardcoded logic for consistent device models

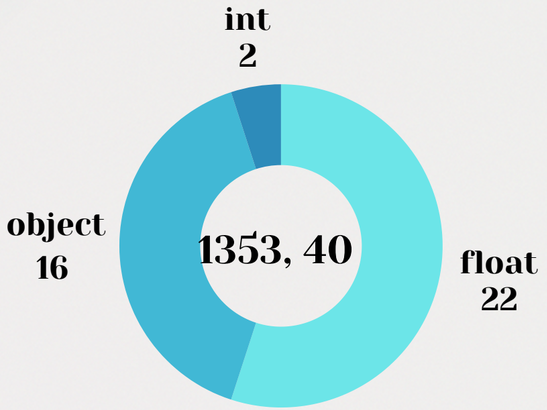
Manual

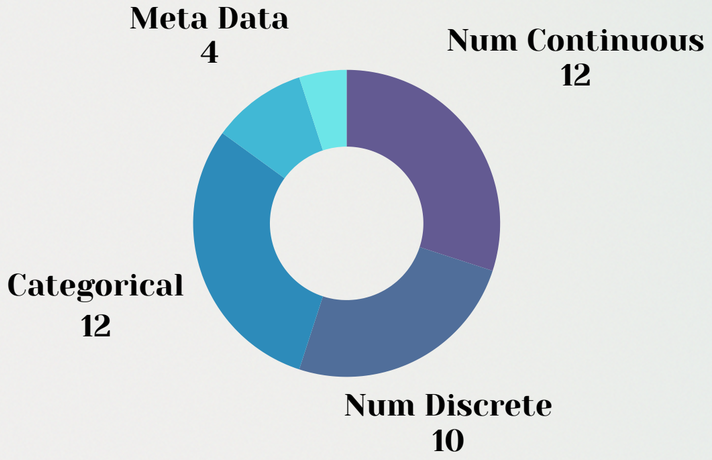
Using domain knowledge for special cases

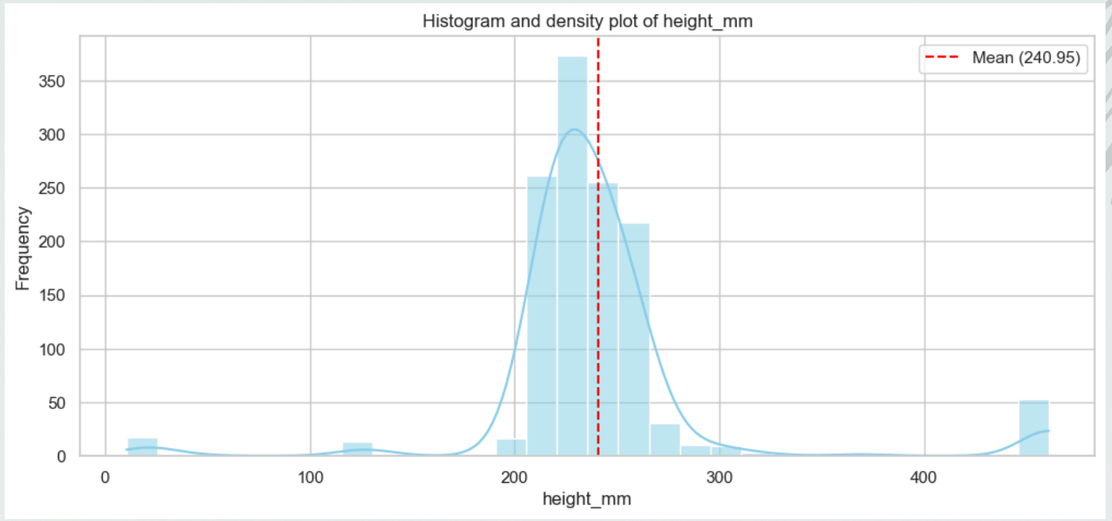
"Dày: 1,13cm - Chiều dài: 30,41cm Chiều rộng: 21,5cm" → Python code → laptop_width, laptop_height, laptop_depth

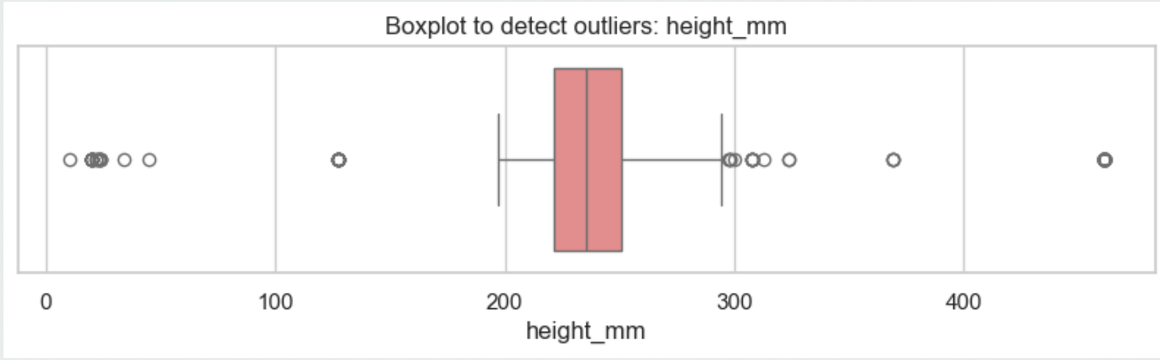
"Intel Core i5-13420H (8 lõi / 12 luồng, up to 4.60 GHz, 12 MB Intel Smart Cache)" → Gemini 2.5 → Review manually → cpu brand, cpu type, cpu cores, cpu threads, cpu_boost clock, cpu_base clock

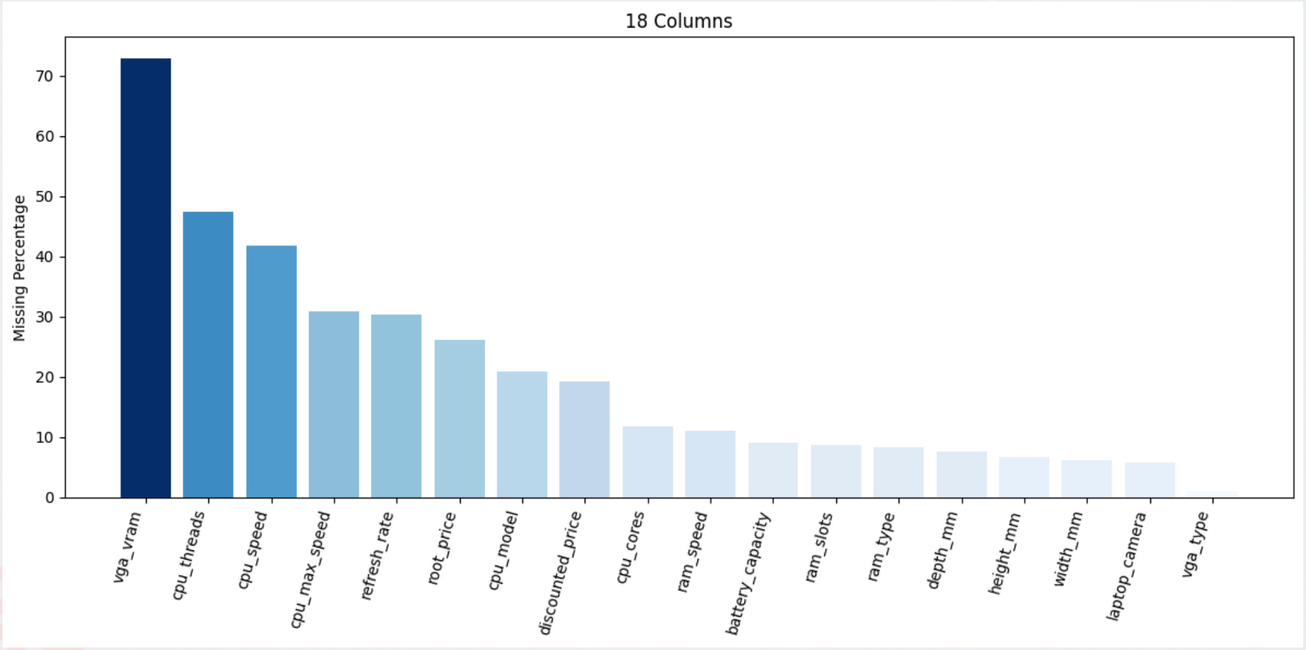
2.2 DATA INGESTION → 1ST EDA

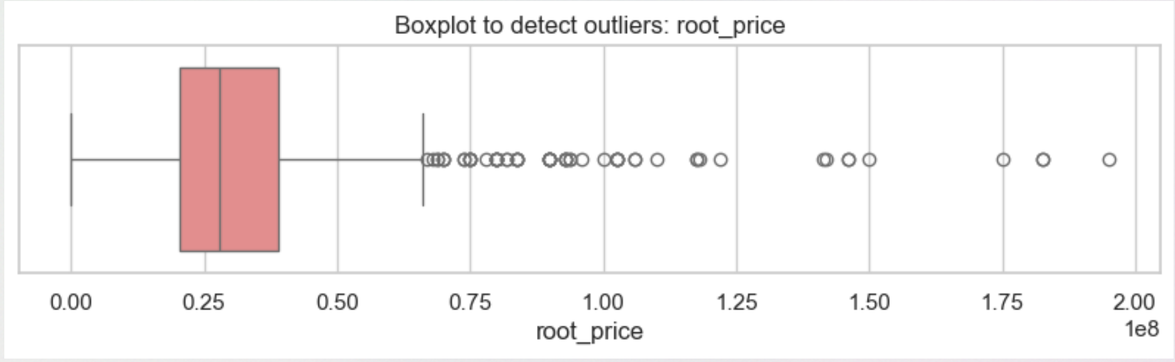












2.3. DATA CLEANING

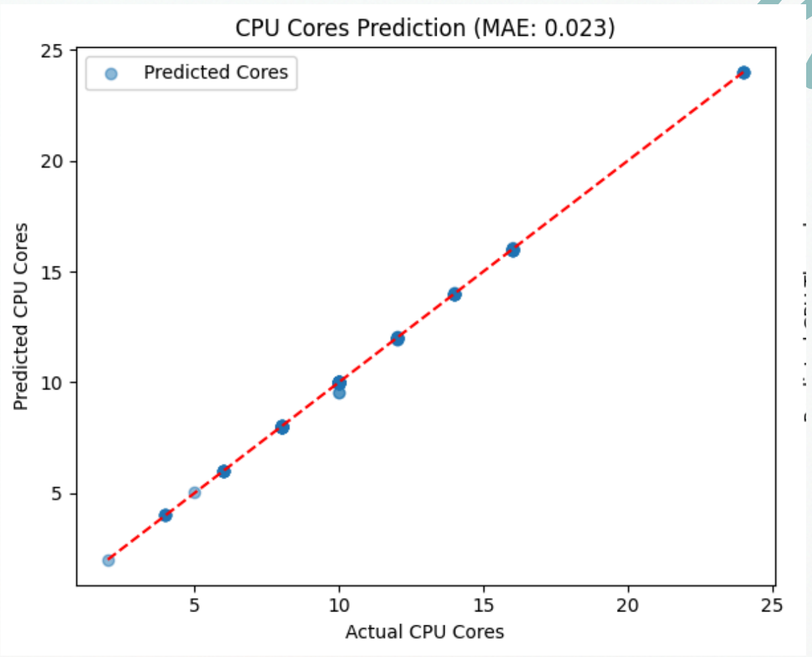
01 Outlier Handling

Manual detection

- Domain-specific correction

02 Rebuilding missing data

- MAR
  - Statistical imputation (mean, mode, model-based - XGBoost và embeddings)
- MCAR
  - LLM-based inference
- MNAR
  - Requires domain knowledge



LLMs	Accuracy	Continuous	Discrete
Gemini 2.5 flash	98,42%	97,76%	99,08%
Gemini 2.0 flash	97,60%	98,68%	96,52%
Llama 3.3	62,90%	60,74%	65,01%

03 Standardize data format

Convert values to consistent and correct types (e.g. number, text)

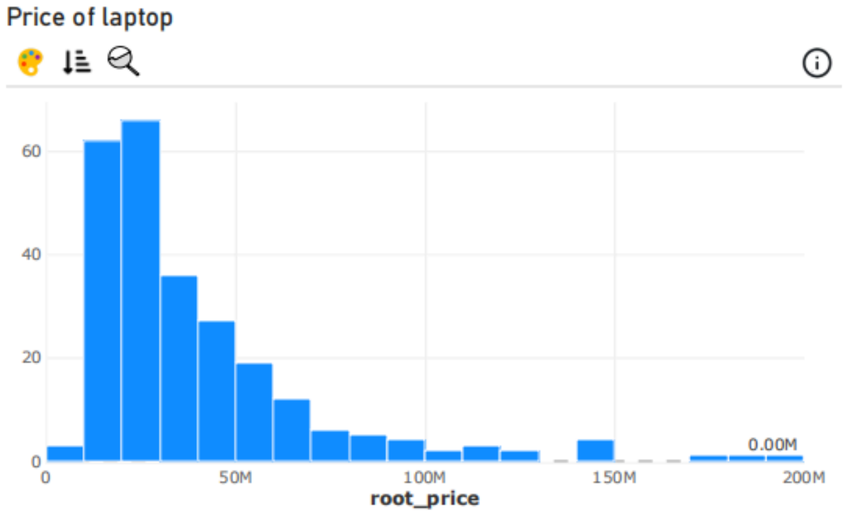
2.4 2ND EDA

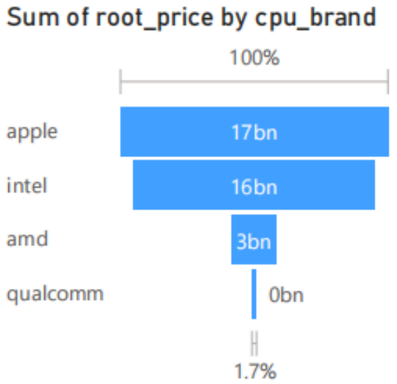
8M
Min of root_price

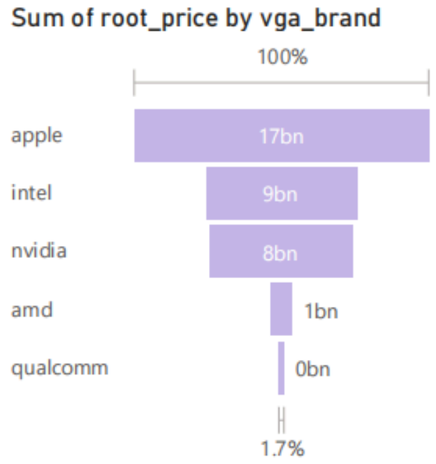
195M
Max of root_price

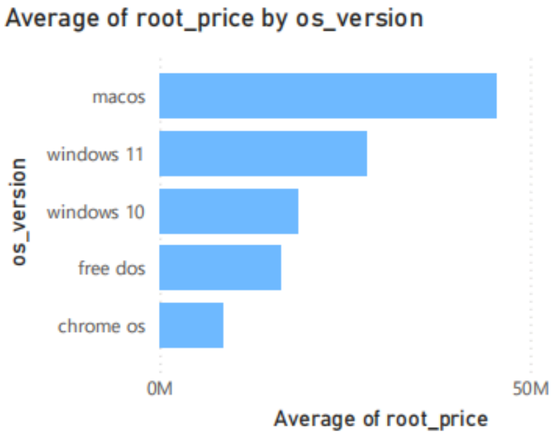
33.90M
Average of root_price

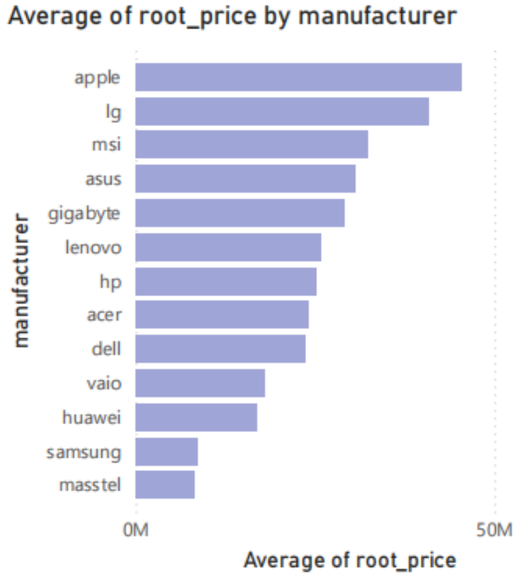
29M
Median of root_price







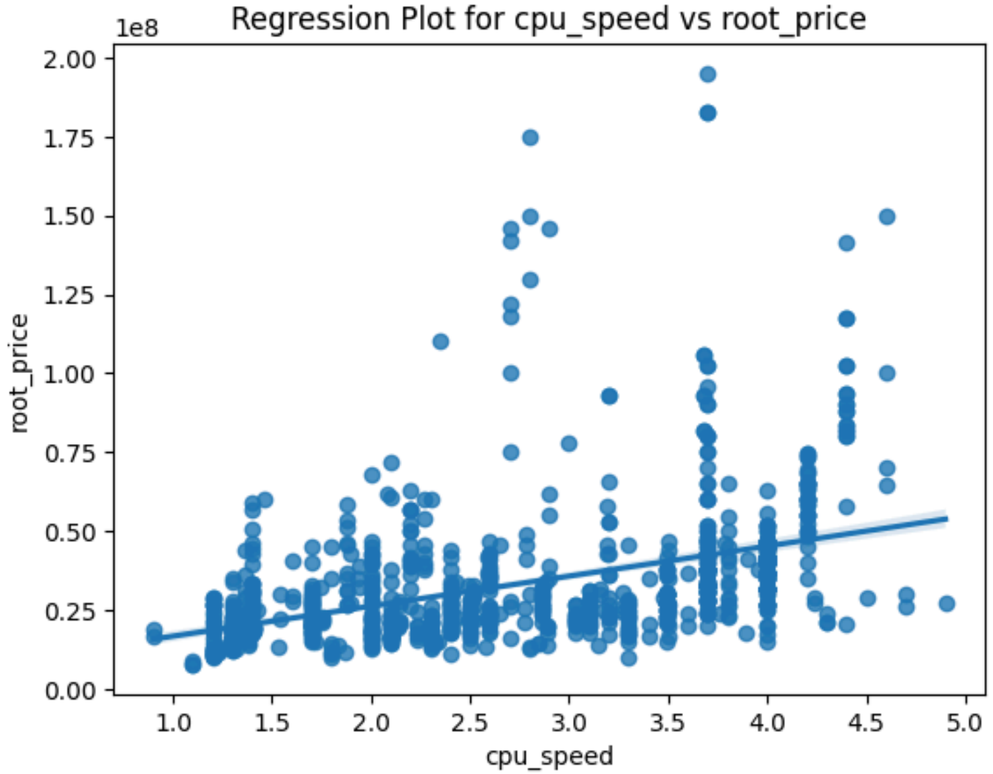




Phân tích đa biến giữa price và các feature số khác:

- Hầu hết các cột đều có tương quan dương với price
- Trừ các biến lên quan đến kích thước thì không có mối quan hệ rõ ràng

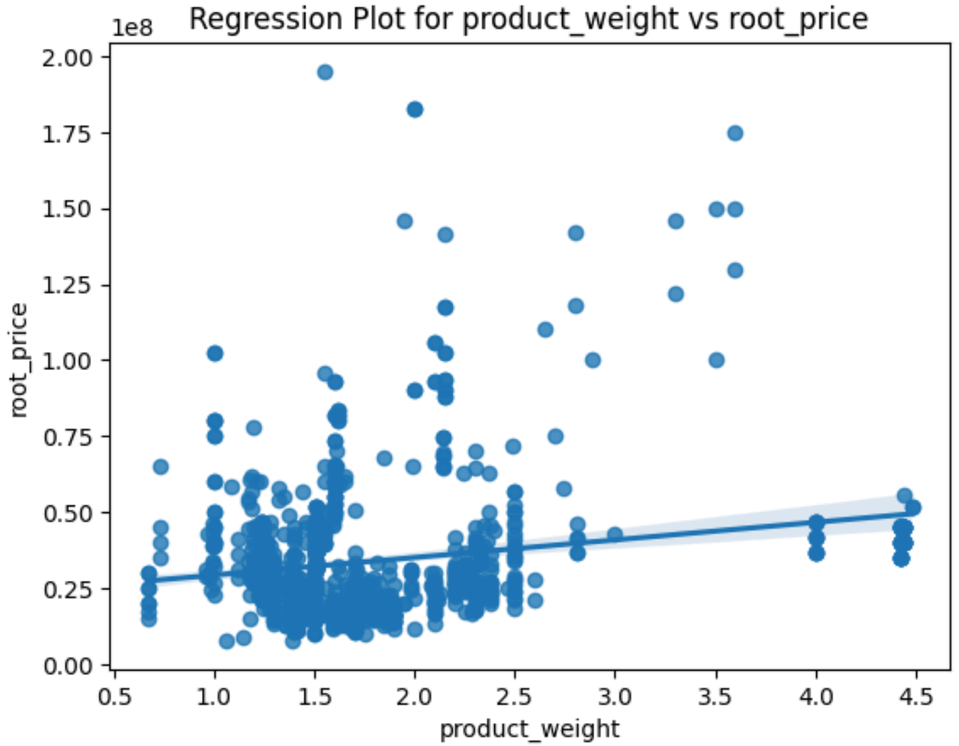
2ND EDA



Có ảnh hưởng

Correlation Coefficient: 0.4508

P-value: 5.4644e-55

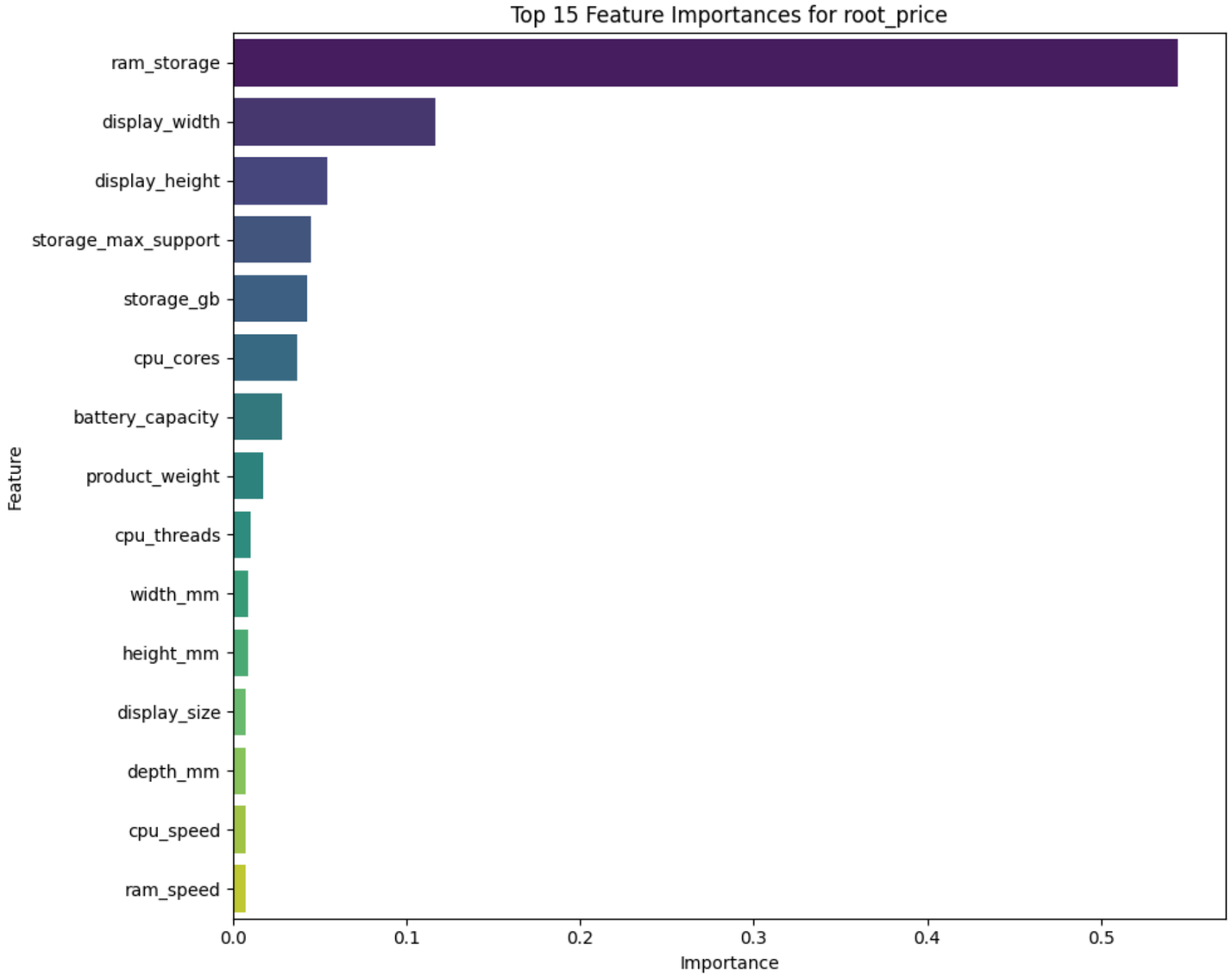


Không ảnh hưởng

Correlation Coefficient: 0.1905

P-value: 2.9614e-10

2ND EDA



Feature importances for root_price

Features	Feature importances
ram_storage	0,544631
display_width	0,116766
display_height	0,054055
storage_max_support	0,045027
storage_gb	0,042426
cpu_cores	0,036955
battery_capacity	0,028198
product_weight	0,017607
cpu_threads	0,01035

2.4 2ND EDA

Thực hiện kiểm định ANOVA với các categorical features với giả thiết với biến target “root_price”:

- **H0: hai biến độc lập**
- **H1: hai biến phụ thuộc nhau**

45 features → 22 features

Index	Features	F-statistic	P-value	Đánh giá
0	material	92,1582	0	Ảnh hưởng
1	manufacturer	19,3462	0	Ảnh hưởng
2	ram_type	46,2786	0	Ảnh hưởng
3	os_version	52,236	0	Ảnh hưởng
4	laptop_color	2,1626	0,018	Ảnh hưởng
5	vga_type	0,017	0,8962	Không ảnh hưởng
6	vga_brand	66,6873	0	Ảnh hưởng
7	lapyop_camera	247,5311	0	Ảnh hưởng
8	cpu_brand	67,3149	0	Ảnh hưởng

2.5

PREPROCESSING

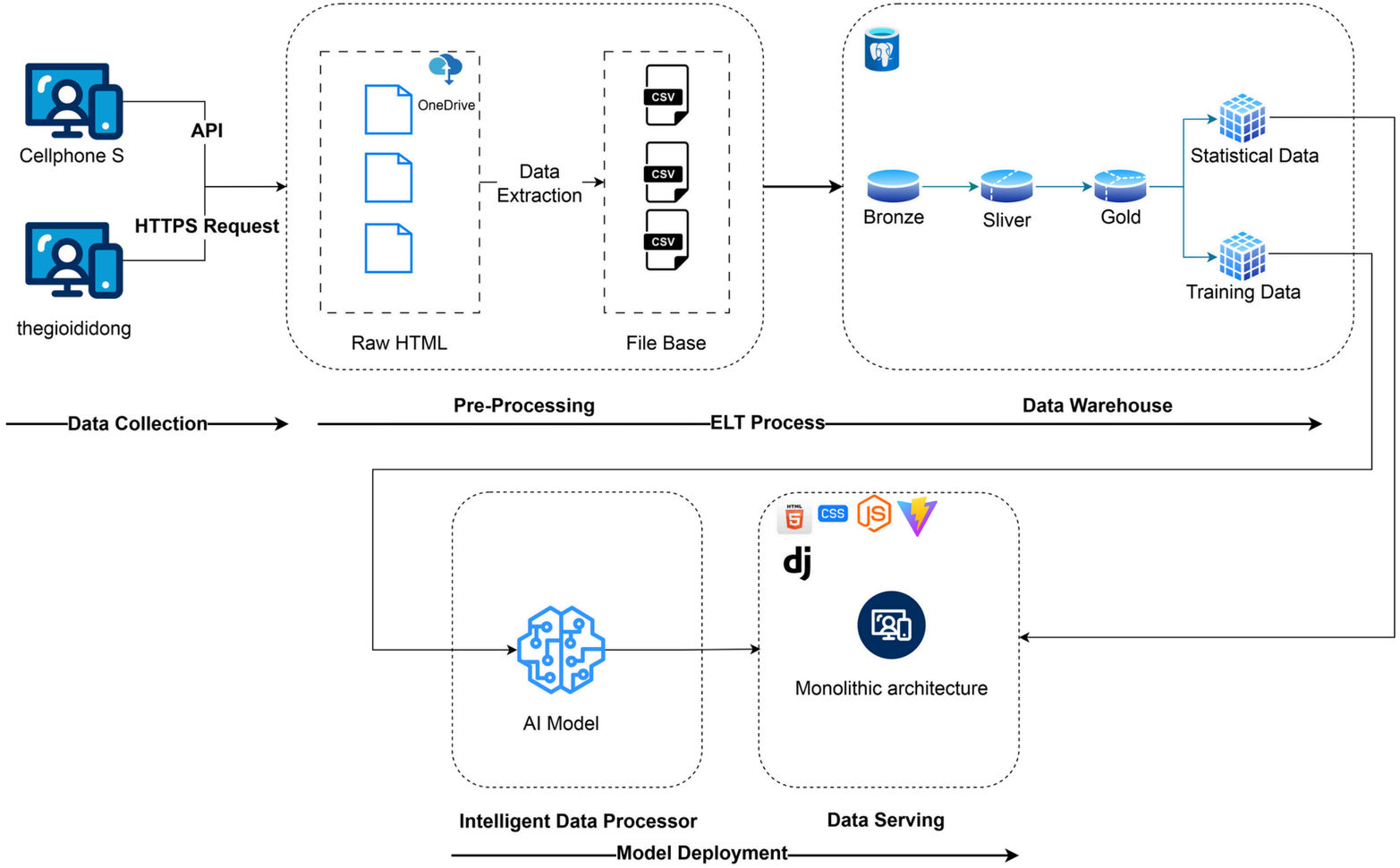
One-hot Encoding

manufacturer', 'cpu_brand',
'material', 'os_version',
'laptop_color', 'vga_brand',
'laptop_camera', 'ram_type',
'laptop_sang_tao_noi_dung',
'do_hoa_ky_thuat',
'cao_cap_sang_trong'

Standard Scaler

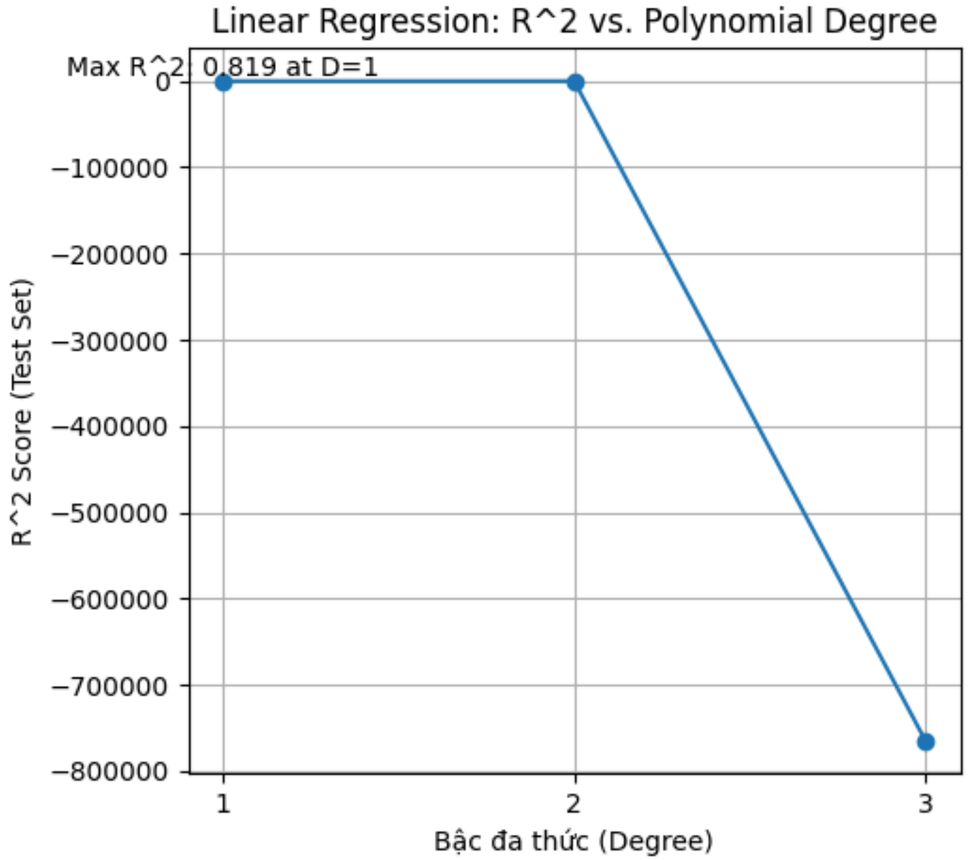
'storage_max_support',
'storage_gb', 'display_width',
'cpu_threads', 'cpu_cores',
'ram_speed', 'cpu_speed',
'ram_storage', 'ram_slots',
'battery_capacity', 'display_height'

2.6 DATA FLOW


Cellphone S
API
HTTPS Request
thegioididong
OneDrive
Raw HTML
Data Extraction
CSV
CSV
CSV
File Base
Bronze
Sliver
Gold
Statistical Data
Training Data
Data Collection
Pre-Processing
ELT Process
Data Warehouse
AI Model
Monolithic architecture
Intelligent Data Processor
Data Serving
Model Deployment

03. MODELLING

Linear Regression với PolynomialFeatures xác định bậc để nắm mối quan hệ phi tuyến



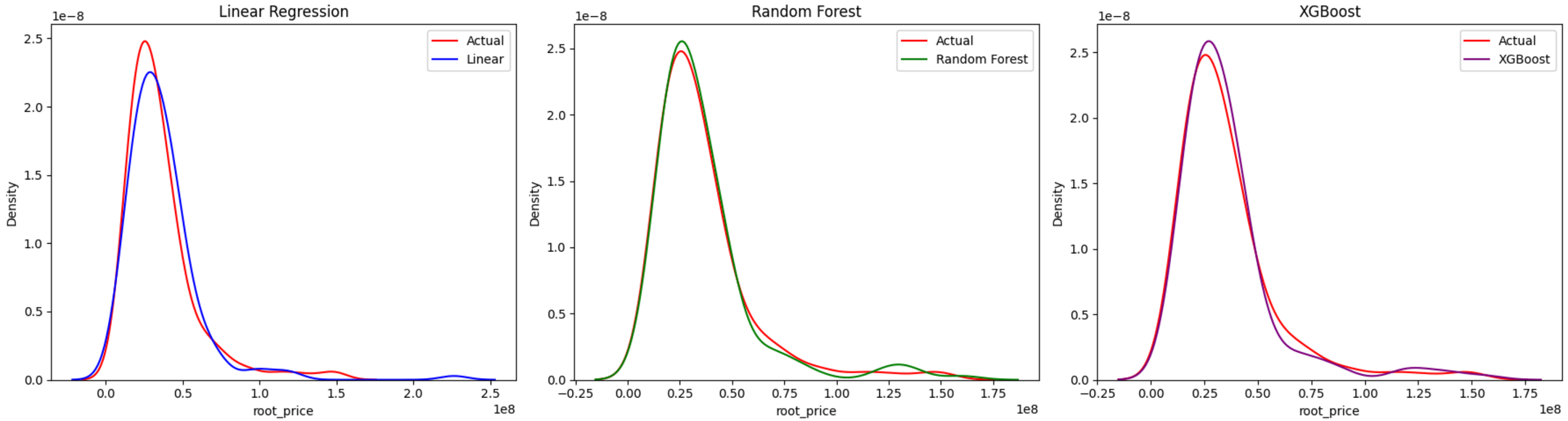
Kết quả: bậc n = 1 cho kết quả dự đoán mô hình ổn định nhất

03. MODELLING

Training và Testing với 3 mô hình với Cross Validation là 5

Mô hình	R^2 (Train)	R^2 (Test)	RMSE (Test)	Cross-validation R^2
Linear Regression	0,891	0,819	10,060,950.43	0,753
Random Forest	0,989	0,916	6,853,476.04	0,840
XGBoost	**0,983**	**0,937**	5,938,349.67	**0,853**

Trực quan hóa

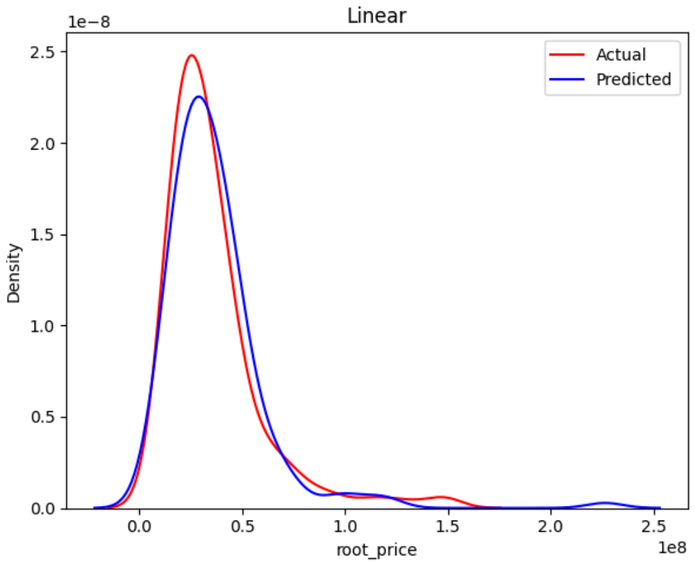


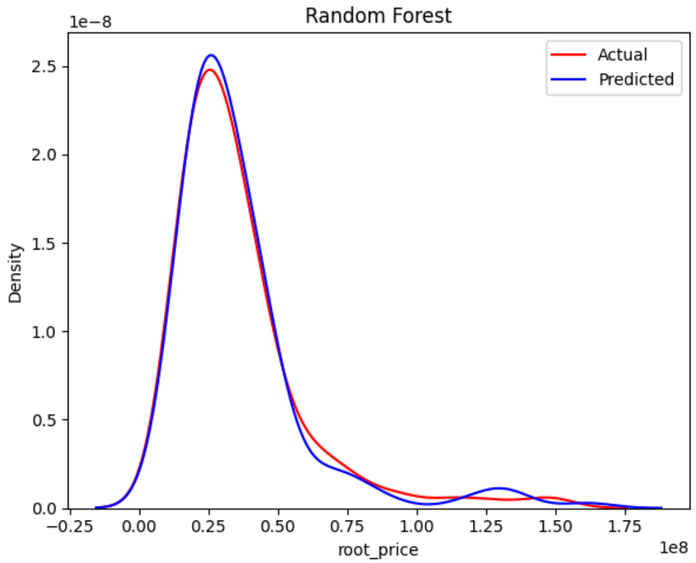
03. MODELLING

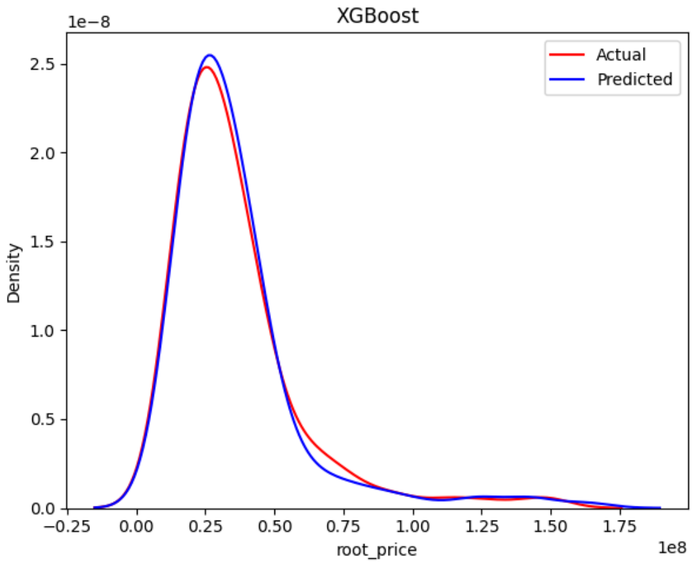
Tunning XGBoost và RandomForestRegressor sử dụng GridSearchCV

Mô hình	Train R^2	Test R^2	RMSE (Test)	CV R^2 (mean)
XGBoost	**0,99080**	**0,94923**	5,320,958	**0,87542**
Random Forest	0,98882	0,91869	6,734,008	0,84166
Linear Regression	0,89092	0,81850	10,060,950	0,75257

Trực quan hóa

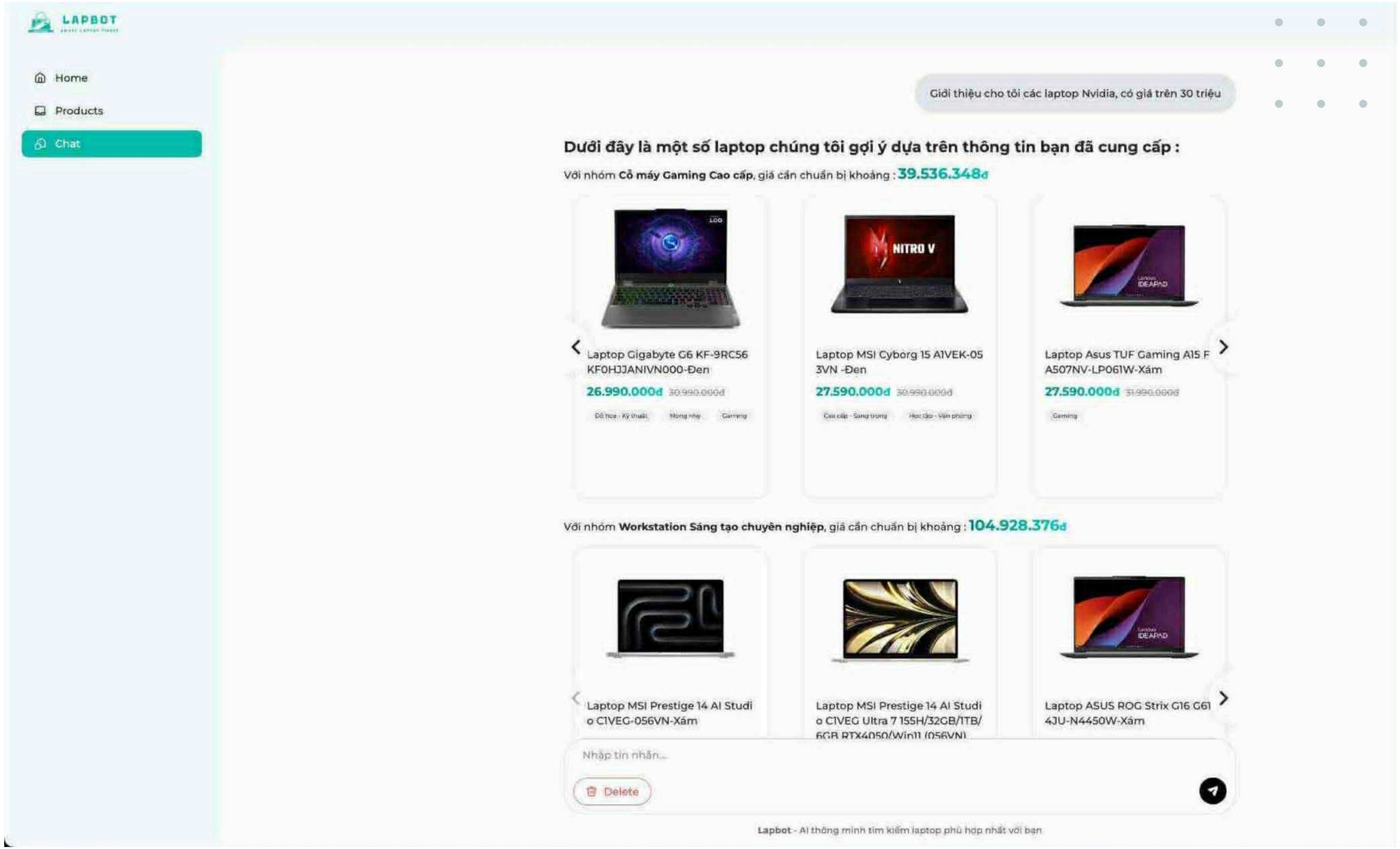




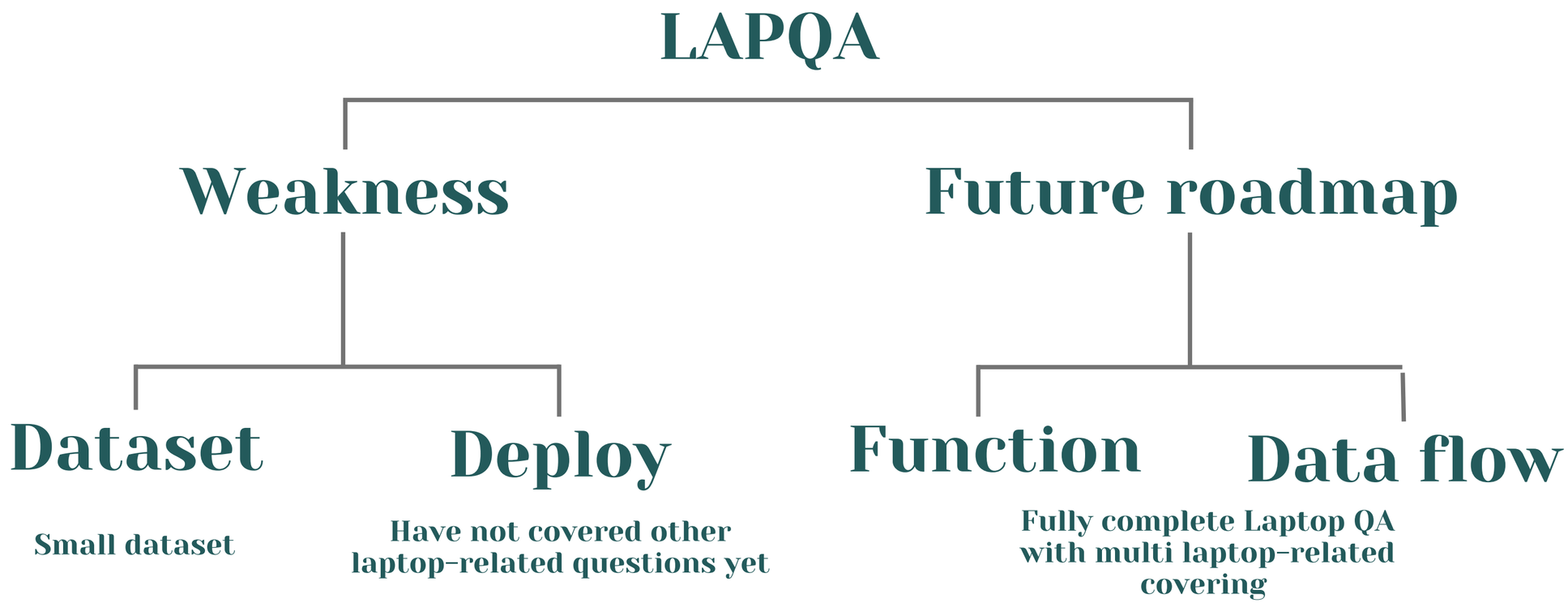


4

DEMO APPLICATION



5. CONCLUSION



Thank you for listening!