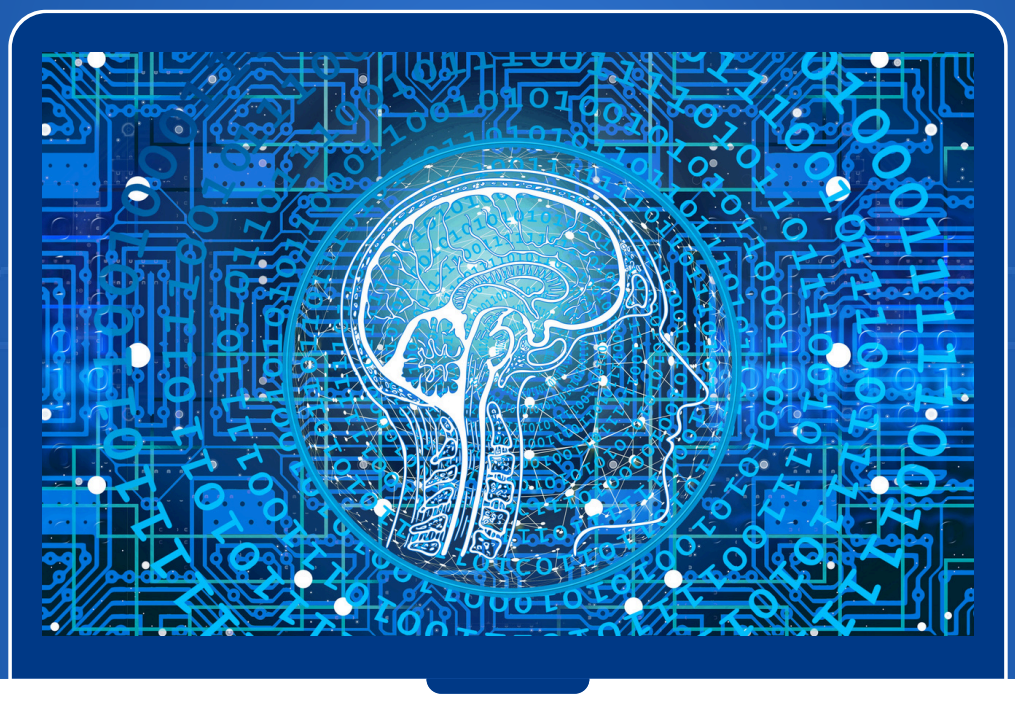
การทำนายราคาของแล็ปท็อป

จากข้อมูลของชิ้นส่วนในแล็ปท็อป



LAPTOP PRICE

PROJECT PROGRESS

ปัญหา : จะสามารถทำนายราคาของแล็ปท็อปจากข้อมูลของชิ้นส่วนภายในได้หรือไม่ และหากทำได้ ข้อมูลชิ้นส่วนใดที่มีผลต่อราคามากที่สุด

ประโยชน์สำหรับผู้บริโภค

1.การประเมินราคาอย่างแม่นยำ

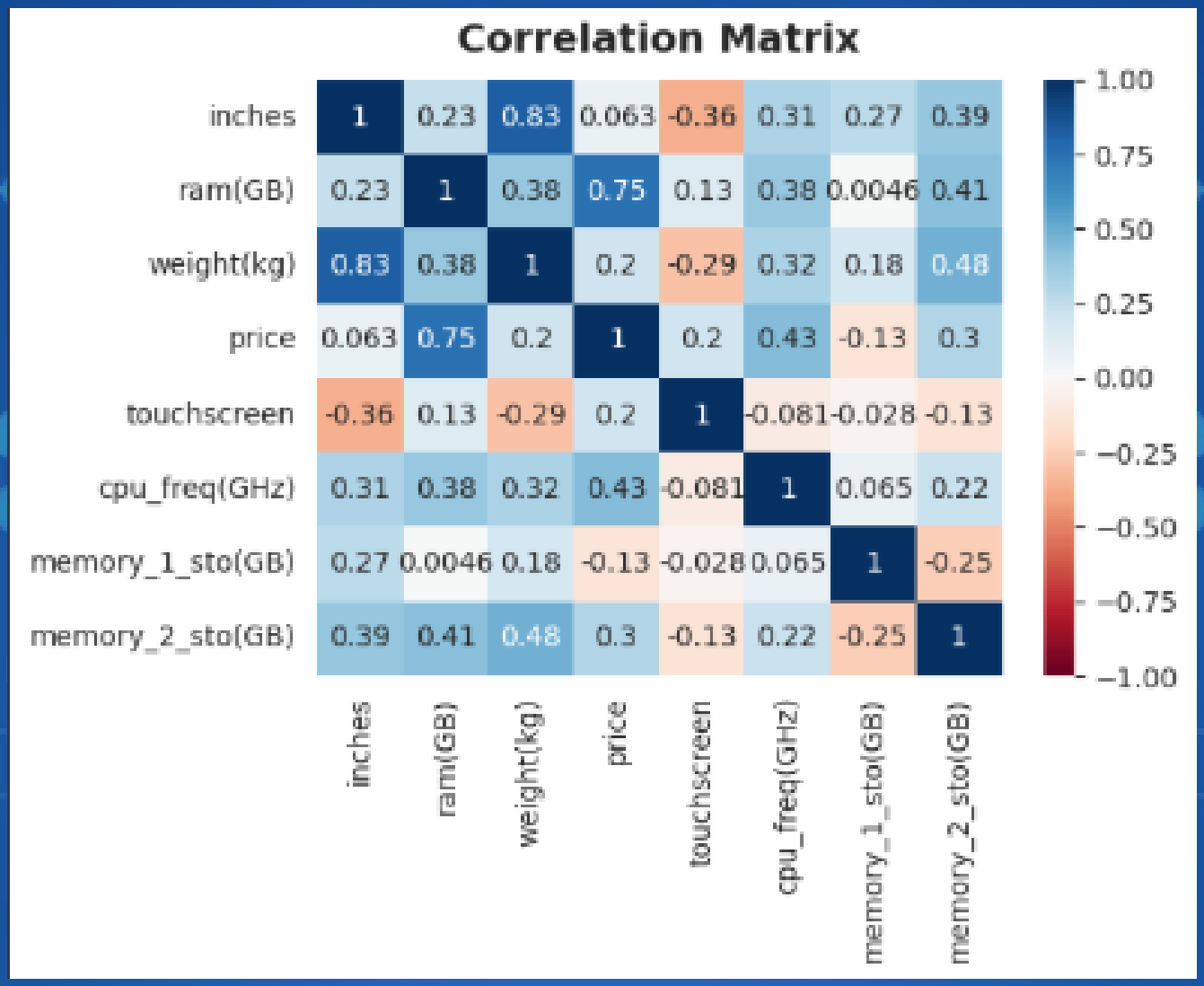
2.การเปรียบเทียบที่ง่ายขึ้น

3.ความมั่นใจในการซื้อ

Data Acquisition

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1302 entries, 0 to 1301
Data columns (total 20 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   company            1302 non-null   object 
 1   product            1302 non-null   object 
 2   typename           1302 non-null   object 
 3   inches             1302 non-null   float64
 4   cpu                1302 non-null   object 
 5   ram(GB)            1302 non-null   int64  
 6   gpu                1302 non-null   object 
 7   opsys              1302 non-null   object 
 8   weight(kg)         1302 non-null   float64
 9   price              1302 non-null   float64
 10  resolution         1302 non-null   object 
 11  screentype         364 non-null    object 
 12  touchscreen        1302 non-null   float64
 13  cpu_freq(GHz)      1302 non-null   float64
 14  memory_1_sto(GB)   1302 non-null   float64
 15  memory_1_type      1302 non-null   object 
 16  memory_2_sto(GB)   1302 non-null   float64
 17  memory_2_type      208 non-null    object 
 18  cpu_brand          1302 non-null   object 
 19  gpu_brand          1302 non-null   object 
dtypes: float64(7), int64(1), object(12)
memory usage: 203.6+ KB
```

Correlation Matrix



- RAM มีความสัมพันธ์เชิงบวกสูงกับราคา (+0.75) โน้ตบุ๊กที่มีราคาแพงกว่ามักจะมี RAM ที่สูงกว่า
- CPU Frequency มีความสัมพันธ์เชิงบวกปานกลางกับราคา (+0.45)
- Inches และ Weight มีความสัมพันธ์เชิงบวกสูง (+0.82) เนื่องจากโน้ตบุ๊กที่มีหน้าจอขนาดใหญ่กว่ามักจะมีน้ำหนักมากกว่า

MODELที่ใช้ในการแก้ปัญหา

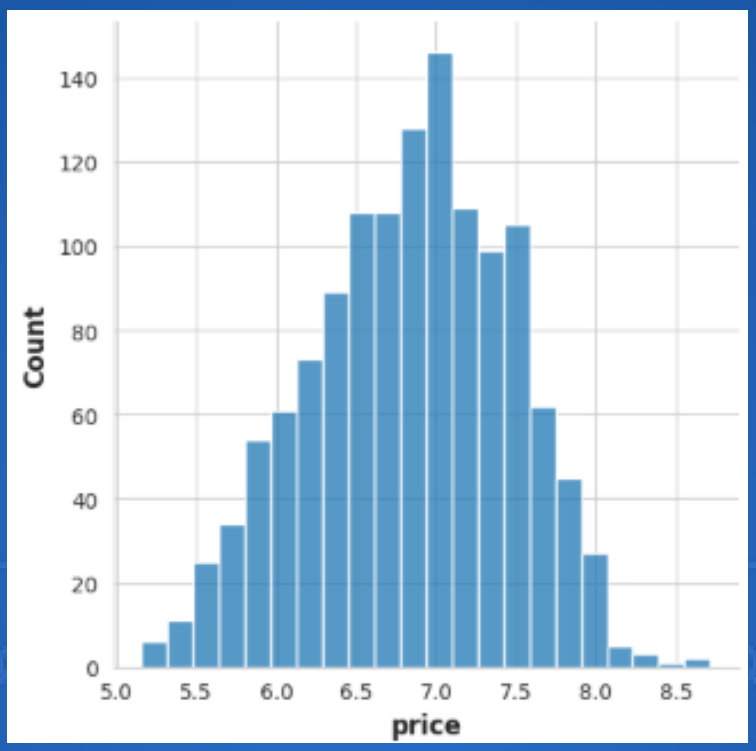
โมเดลหลักสองแบบในการทำนายราคาของแลปทอป ได้แก่ Random Forest และ Linear Regression โดยทั้งสองโมเดลเป็น Supervised Learning ที่ใช้ข้อมูลที่มี labeled data ในการฝึกฝน เพื่อให้โมเดลสามารถทำนายค่าเป้าหมาย ซึ่งในกรณีนี้คือราคา ของแลปทอปที่เป็น continuous data ได้อย่างแม่นยำ ขั้นตอนการทำงานมีดังนี้

1. การเตรียมข้อมูล (DATA PREPROCESSING)

- เราจะเริ่มต้นด้วยการแปลงข้อมูลราคาของแลปทอปที่มีการกระจายตัวแบบขวา โดยใช้ Log Transformation เพื่อปรับให้ข้อมูลมีการกระจายที่เหมาะสมมากขึ้นสำหรับการค่าโมเดล
- สำหรับข้อมูลเชิงหมวดหมู่ (categorical features) เช่น 'company, product, 'cpu' และอื่นๆ เราจะทำ Label Encoding เพื่อแปลงข้อมูลเหล่านี้เป็นค่าตัวเลขที่โมเดลสามารถใช้งานได้
- และมีการทำ scaling ให้ linear regression

DATA PRE-PROCESSING

CATEGORICAL FEATURES ENCODING




```
[ ]  1 catCols =  ['company','product','typename','cpu','gpu','opsys','resolution','screentype','resolution','memory_1_type','memory_2_type','gpu_brand','cpu_brand']
```

```
[ ]  1 #Label encoding
     2 en = LabelEncoder()
     3 for cols in catCols:
     4     df1[cols] = en.fit_transform(df1[cols])
     5 print('Dataframe encoded by Label encoding dimension : ', df1.shape)
```

```
Dataframe encoded by Label encoding dimension :  (1301, 20)
```

```
1 df1.head()
```

	company	product	typename	inches	cpu	ram(GB)	gpu	opsys	weight(kg)	price	resolution	screentype	touchscreen	cpu_freq(GHz)	memory_1_sto(GB)	memory_1_type	me
0	1	299	4	13.3	51	8	57	8	1.37	7.200194	10	1	0.0	2.3	128.0	3	
1	1	300	4	13.3	51	8	50	8	1.34	6.801216	1	2	0.0	1.8	128.0	0	
2	7	50	3	15.6	57	8	52	4	1.86	6.354370	3	2	0.0	2.5	256.0	3	
3	1	299	4	15.4	65	16	9	8	1.83	7.838915	12	1	0.0	2.7	512.0	3	
4	1	299	4	13.3	51	8	58	8	1.37	7.497540	10	1	0.0	3.1	256.0	3	

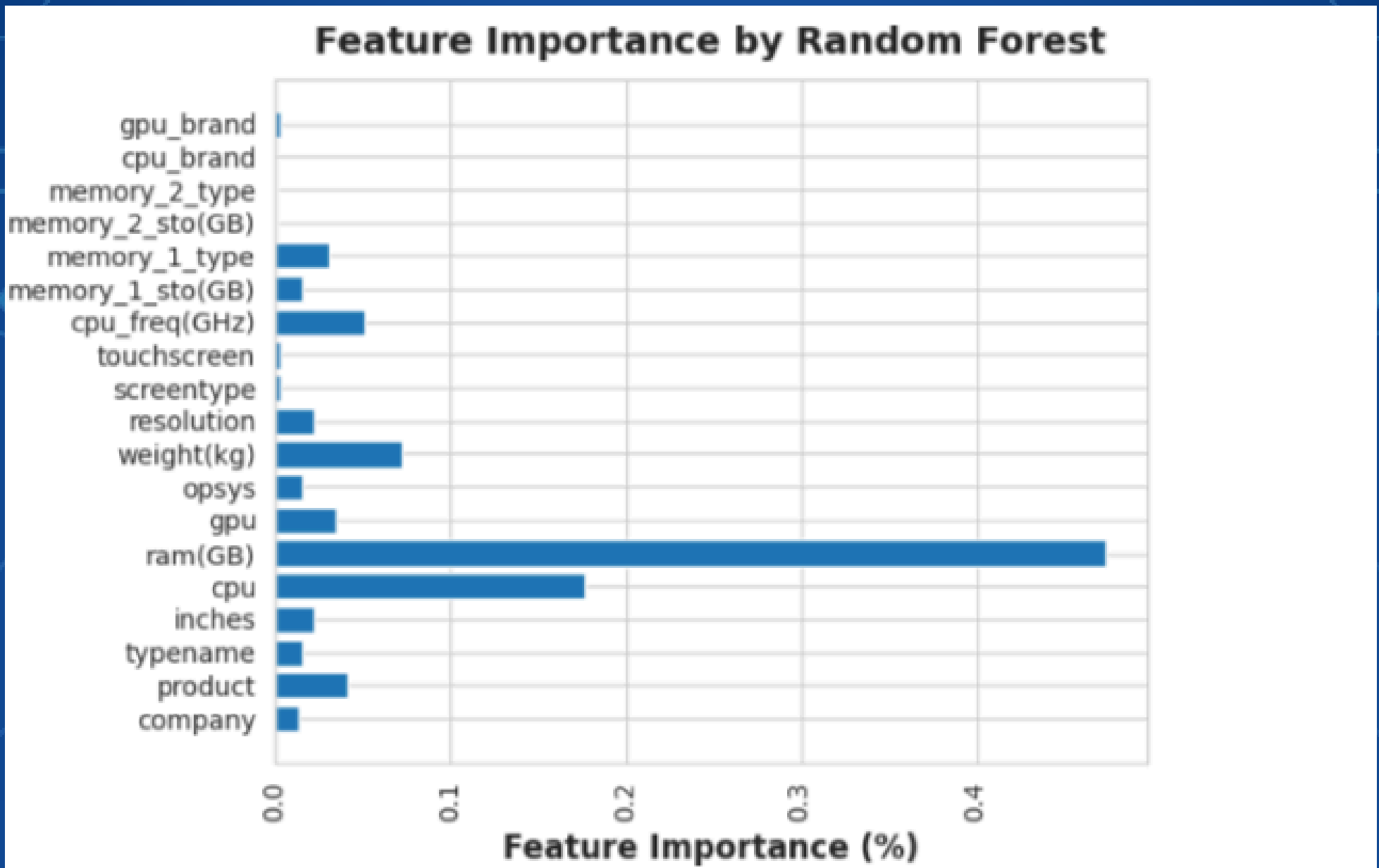
2. การแบ่งข้อมูล (DATA SPLITTING)

- เราจะแบ่งข้อมูลออกเป็น Train Set และ Test Set โดยใช้สัดส่วน 70% สำหรับการฝึกโมเดลและ 30% สำหรับการทดสอบโมเดล " นอกจากนี้เรายังทำการแบ่งข้อมูลจาก Train Set เป็น Training Set และ Validation Set เพื่อใช้ในการตรวจสอบโมเดล

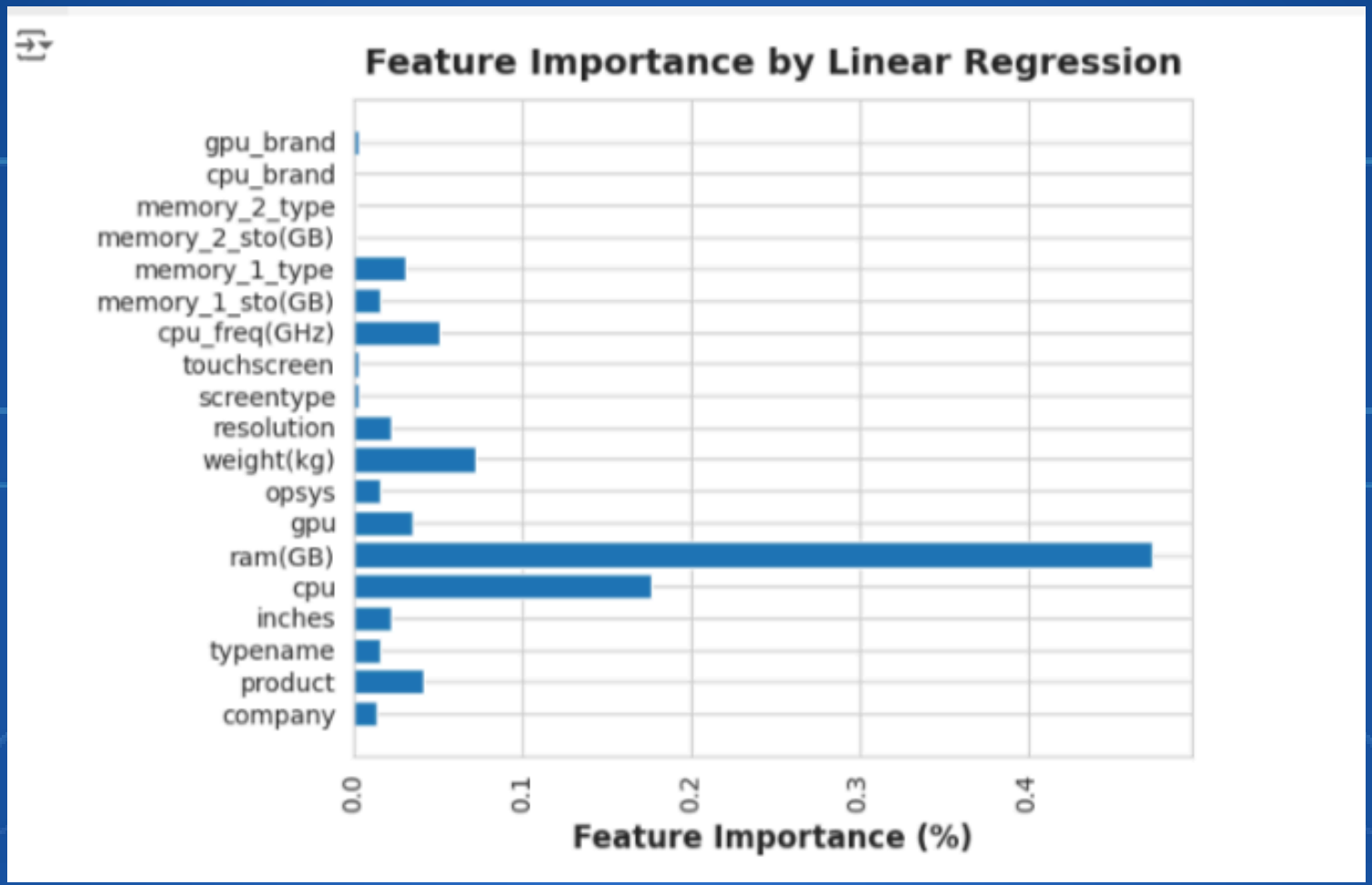
3. การประเมินประสิทธิภาพของโมเดล (MODEL EVALUATION)

- เราจะใช้ R-squared ในการวัดประสิทธิภาพของโมเดลในการทำนายราคา นอกจากนี้ยังมีการใช้ Root Mean Squared Error (RMSE) เพื่อวัดค่าผิดพลาดของการทำนาย

RANDOM FOREST



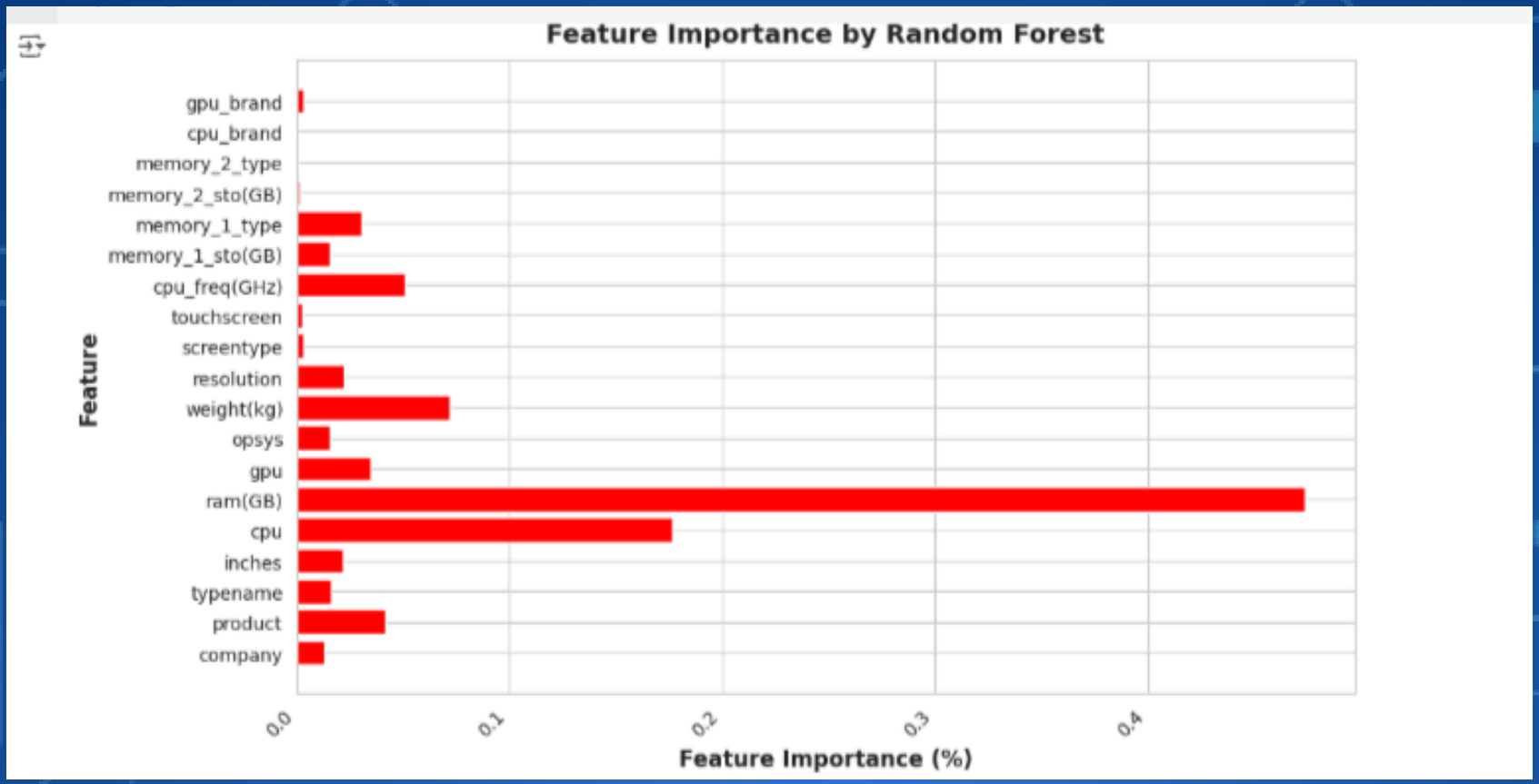
LINEAR REGRESSION

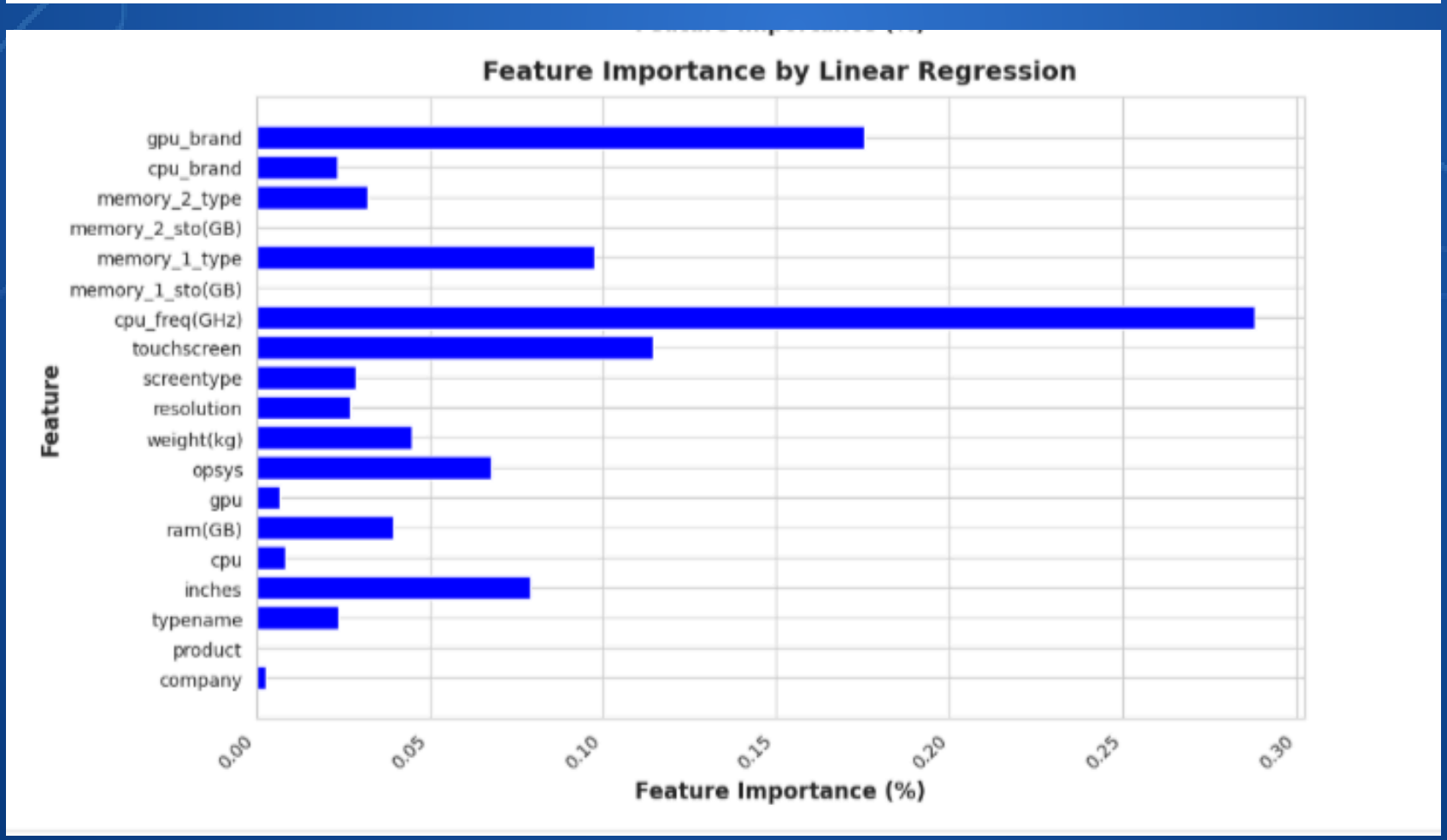


SUMMARY


Actual vs Predicted Price: Linear Regression vs Random Forest
RF, R2 91.11 %
LR, R2 78.23 %
Predicted Price
Actual Price
0
1000
2000
3000
4000
5000
6000
7000

SUMMARY


Feature Importance by Random Forest
gpu_brand
cpu_brand
memory_2_type
memory_2_sto(GB)
memory_1_type
memory_1_sto(GB)
cpu_freq(GHz)
touchscreen
screentype
resolution
weight(kg)
opsys
gpu
ram(GB)
cpu
inches
typename
product
company
Feature
0.0
0.1
0.2
0.3
0.4
Feature Importance (%)


Feature Importance by Linear Regression
gpu_brand
cpu_brand
memory_2_type
memory_2_sto(GB)
memory_1_type
memory_1_sto(GB)
cpu_freq(GHz)
touchscreen
screentype
resolution
weight(kg)
opsys
gpu
ram(GB)
cpu
inches
typename
product
company
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Feature Importance (%)