Final Project - Laptop Pricing

COMP4980: Machine Learning

Timur Rakhimov

T00668753

23.11.2024

Data Description

Dataset i’ve chosen

https://www.kaggle.com/datasets/gyanprakashkushwaha/laptop-price-prediction-cleaned-dataset/data

What is the data about?

The dataset provides detailed information about laptops, including their specifications and prices. It contains 1,273 entries, each representing a different laptop model. The main goal of this data is to understand the relationship between the laptop features (like brand, RAM, and storage type) and the price. This kind of dataset can be used for tasks such as price prediction, understanding market trends, and helping customers make informed buying decisions.

What kind of data is it?

This dataset has a mix of different types of data:

- Numerical Data: Includes values like RAM size (in GB), weight (in kg), and storage capacity (in GB). These are all numbers that can be used directly for analysis.
- Categorical Data: Includes information such as laptop brand (Company), laptop type (TypeName), and the operating system (Os). These are text-based categories that help group the laptops based on their characteristics.

- Binary Data: The dataset also contains yes/no type features, like whether the laptop has a touchscreen (TouchScreen) or an IPS display (Ips). These are represented as 0 (no) or 1 (yes).

Volume: How big is the data?

The dataset has:

- 1,273 rows: Each row represents a different laptop model.
- 13 columns: Each column provides a specific feature of the laptop, such as the brand, RAM, or price.

The total file size is about 129.4 KB, which is small enough to handle easily in Python without any special tools. Even though it’s a small dataset, it provides a good variety of features for building a predictive model.

Variety: What are the variables included? What are their possible values?

The dataset includes the following variables:

1. **Company**: The brand of the laptop (e.g., Apple, Dell, HP).
2. **TypeName**: The type of laptop (e.g., Ultrabook, Notebook, Gaming).
3. **Ram**: The amount of RAM in gigabytes (GB), ranging from 2 GB to 64 GB.
4. **Weight**: The weight of the laptop in kilograms (kg), ranging from 0.69 kg to 4.7 kg.
5. **Price**: The normalized price value (scaled between 9.13 and 12.69).
6. **TouchScreen**: Whether the laptop has a touchscreen (0 = No, 1 = Yes).
7. **Ips**: Whether the laptop has an IPS display (0 = No, 1 = Yes).
8. **Ppi**: The pixels per inch (PPI) value, indicating the screen's sharpness.
9. **Cpu_brand**: The brand of the laptop's processor (e.g., Intel Core i5, AMD Ryzen).
10. **HDD**: The size of the hard disk drive (HDD) in gigabytes (GB).
11. **SSD**: The size of the solid-state drive (SSD) in gigabytes (GB).
12. **Gpu_brand**: The brand of the graphics processing unit (e.g., Nvidia, Intel).
13. **Os**: The operating system (e.g., Windows, Mac, Others).

What format is the database file?

The dataset is in CSV (Comma-Separated Values) format, which is one of the most common formats for data analysis. CSV files are easy to load into Python and other data analysis tools, making this format convenient for further work.

How usable is the data?

The dataset is very user-friendly:

- No Missing Values: Every row and column is fully populated, so there is no need to spend time cleaning up empty or missing data.
- Clean and Consistent: The data types are correct, and there are no obvious errors. For example, numerical values like RAM and weight are all integers or floats, and the categorical values are spelled consistently.
- Ready for Analysis: Since the data is already cleaned, it can be directly used in a Python program for analysis or machine learning tasks without much extra preparation.

Veracity: How reliable is it?

The dataset looks reliable for several reasons:

1. Complete Data: There are no missing entries; all 1,273 rows have values in every column. This shows that the data is well-prepared and does not have gaps.
2. Consistent Values: The data types and values are consistent. For example, the binary columns like TouchScreen and Ips only have values of 0 (no) or 1 (yes), and brand names like Company are spelled correctly without errors.
3. Reasonable Numbers: The numerical columns, such as Weight, Ram, HDD, and SSD, have values that make sense. For instance, laptop weights range from 0.69 kg to 4.7 kg, which fits within normal expectations for laptops.

Velocity: How recent is the data?

The dataset does not include any date or timestamp information, so it’s hard to tell how up-to-date the data is. This could be an issue because laptop specifications and prices change frequently as new models are released. The absence of date information might affect the relevance of any price predictions we make using this data.

Value: How useful is the data?

This dataset is highly useful for several reasons:

- Price Prediction: It can be used to create a machine learning model that predicts laptop prices based on features like RAM, storage, and processor type.
- Market Analysis: It helps identify trends in laptop specifications and how they relate to pricing, which can be valuable for both consumers and retailers.
- Feature Impact: By analyzing the data, we can understand which features (e.g., RAM size, presence of SSD) have the biggest impact on a laptop’s price, providing insights into what consumers value most.

Why is this a good dataset for a machine learning project?

This dataset is a great choice for a machine learning project for several reasons. It has a good mix of different types of data, such as numbers (like RAM size and weight) and categories (like laptop brand and type). This variety makes it possible to try out different machine learning methods, such as predicting prices or finding patterns based on laptop features. The data is very practical because it focuses on laptops, which are common products that many people buy. Predicting laptop prices based on their features can help businesses decide on pricing and help customers make better decisions when shopping for a laptop. It’s a useful way to see how certain features (like having an SSD or more RAM) can affect the overall price.

Another reason this dataset is a good choice is that it is already clean and ready to use. There are no missing entries, so you don’t have to spend a lot of time fixing or organizing the data. This means you can start building models right away, which is great for learning and exploring machine learning techniques quickly.

Although the dataset isn’t very big, it still has a wide range of features that provide valuable information. This makes it a good starting point for learning how to build prediction models and understand which features matter the most.

Data Analysis

General Overview

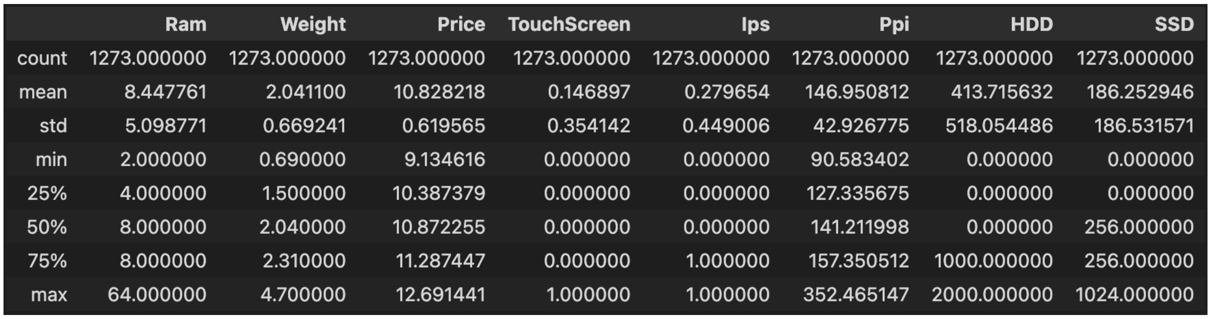
The dataset contains information on 1,273 laptops, with details about their specifications, prices, and features. There are 13 columns in total, including both categorical and numerical variables. Categorical variables such as **Company, TypeName, Cpu_brand, Gpu_brand, and Os** describe qualitative attributes of the laptops. Numerical variables like **Ram, Weight, Price, TouchScreen, Ips, Ppi, HDD, and SSD** provide quantitative details.

All columns have complete data with no missing values, and the dataset's memory usage is 129.4 KB.

Descriptive Statistics

The descriptive statistics reveal some interesting patterns about the laptops. For instance, the average RAM size is **8.45 GB**, which suggests that most laptops in the dataset have 8 GB of RAM, reflecting a standard configuration in the market. However, there are models with as little as 2 GB of RAM, likely budget options, and others with up to 64 GB, which are likely high-end devices aimed at professionals or gamers. The weight of the laptops varies significantly, with lightweight models under 1.5 kg likely representing ultraportable devices, while heavier laptops weighing up to 4.7 kg are likely gaming or workstation models with more powerful hardware.

The price of laptops ranges from **9.13 to 12.69** (scaled values), with most clustered around the mean value of **10.83**. The price distribution is slightly right-skewed, indicating the presence of a few expensive, high-end laptops. This skewness reflects the diversity of products in the market, ranging from affordable laptops to premium devices. Storage options also vary widely, with SSD sizes averaging **186 GB** and some models offering up to 1 TB, reflecting modern trends toward faster, solid-state storage. Meanwhile, traditional HDD storage averages **414 GB**, with some models reaching 2 TB, which may cater to users requiring larger capacities for data storage.

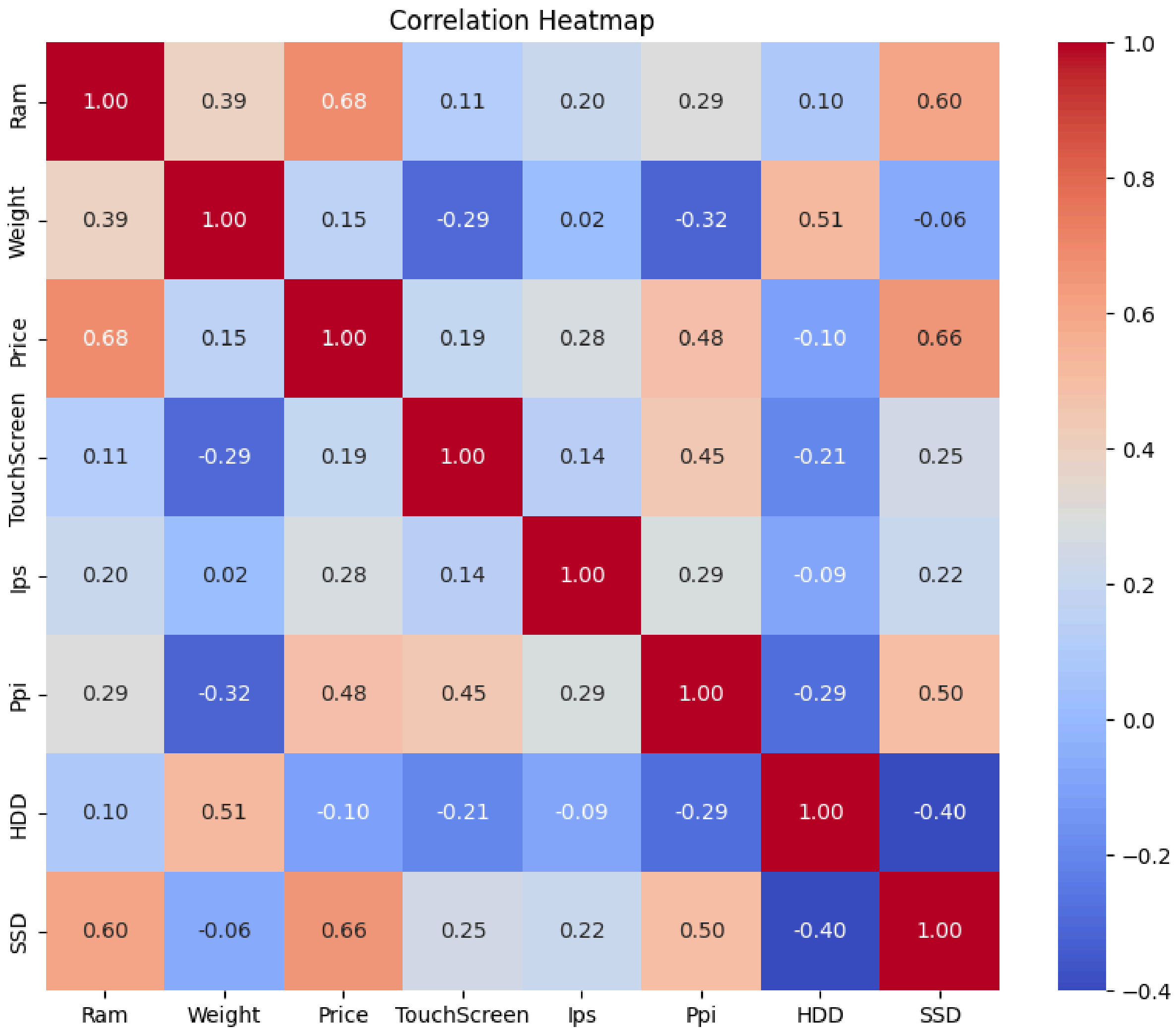


	Ram	Weight	Price	TouchScreen	Ips	Ppi	HDD	SSD
count	1273.000000	1273.000000	1273.000000	1273.000000	1273.000000	1273.000000	1273.000000	1273.000000
mean	8.447761	2.041100	10.828218	0.146897	0.279654	146.950812	413.715632	186.252946
std	5.098771	0.669241	0.619565	0.354142	0.449006	42.926775	518.054486	186.531571
min	2.000000	0.690000	9.134616	0.000000	0.000000	90.583402	0.000000	0.000000
25%	4.000000	1.500000	10.387379	0.000000	0.000000	127.335675	0.000000	0.000000
50%	8.000000	2.040000	10.872255	0.000000	0.000000	141.211998	0.000000	256.000000
75%	8.000000	2.310000	11.287447	0.000000	1.000000	157.350512	1000.000000	256.000000
max	64.000000	4.700000	12.691441	1.000000	1.000000	352.465147	2000.000000	1024.000000

Figure 1: Summary Statistics for Numerical Columns

Correlation Analysis

The correlation analysis highlights some important relationships among the variables. RAM size and price have a **strong positive correlation**, suggesting that higher RAM configurations directly contribute to higher costs. Similarly, SSD size also shows a **strong positive correlation** with price, indicating that laptops with larger SSDs are positioned as premium products. On the other hand, HDD and SSD storage show a **negative correlation**, which may indicate that laptops tend to include either a large SSD or a large HDD, but rarely both, as they cater to different market segments. Interestingly, weight and PPI (pixels per inch) have a **weak negative correlation**, which might suggest that higher-resolution screens are typically found in lighter laptops, likely because high-end ultraportables focus on display quality.



	Ram	Weight	Price	TouchScreen	Ips	Ppi	HDD	SSD
Ram	1.00	0.39	0.68	0.11	0.20	0.29	0.10	0.60
Weight	0.39	1.00	0.15	-0.29	0.02	-0.32	0.51	-0.06
Price	0.68	0.15	1.00	0.19	0.28	0.48	-0.10	0.66
TouchScreen	0.11	-0.29	0.19	1.00	0.14	0.45	-0.21	0.25
Ips	0.20	0.02	0.28	0.14	1.00	0.29	-0.09	0.22
Ppi	0.29	-0.32	0.48	0.45	0.29	1.00	-0.29	0.50
HDD	0.10	0.51	-0.10	-0.21	-0.09	-0.29	1.00	-0.40
SSD	0.60	-0.06	0.66	0.25	0.22	0.50	-0.40	1.00

Figure 2: Correlation Heatmap of Features

Visualisations

Visualizations help to better understand these trends. The price distribution histogram confirms that most laptops fall into a mid-range price category, with a few outliers at the high end. Scatterplots of RAM versus price and SSD versus price clearly show that both higher RAM and larger SSD sizes are linked to increased laptop costs. These relationships are particularly useful for understanding how individual features contribute to overall pricing.

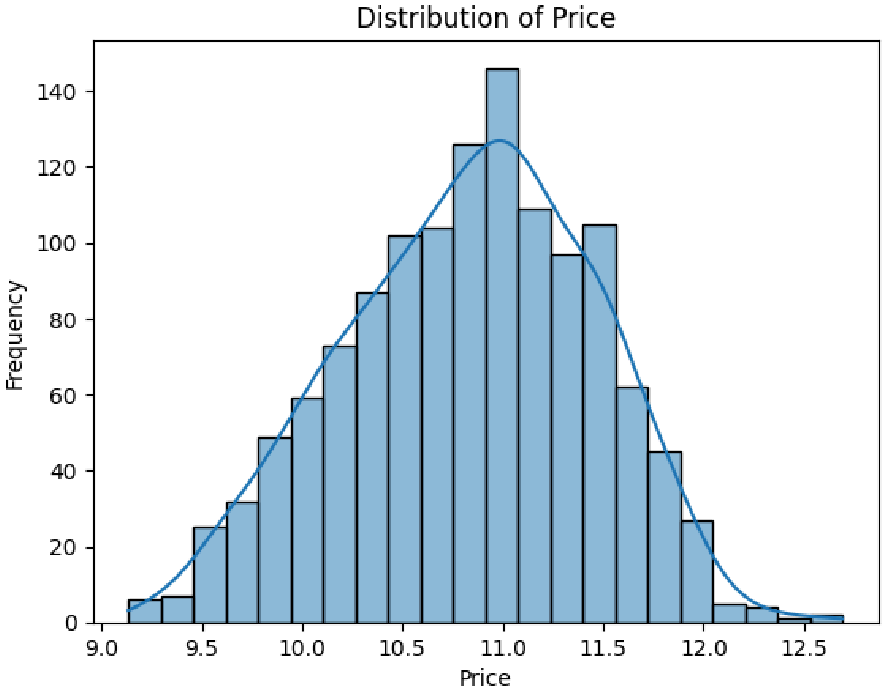


Figure 3: Histogram of Distribution of Price

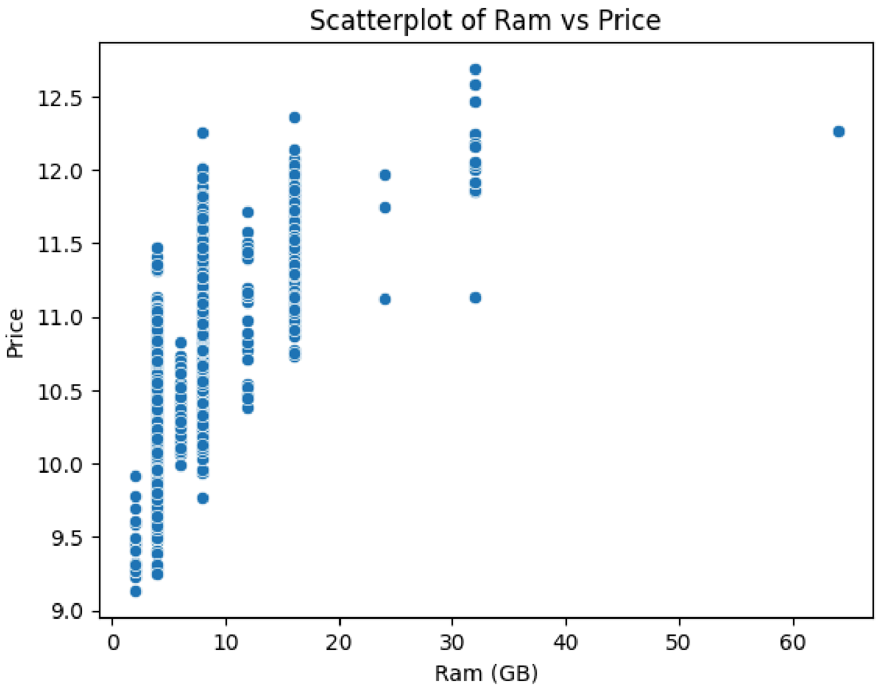


Figure 4: Scatterplot of Relationship between RAM and Price

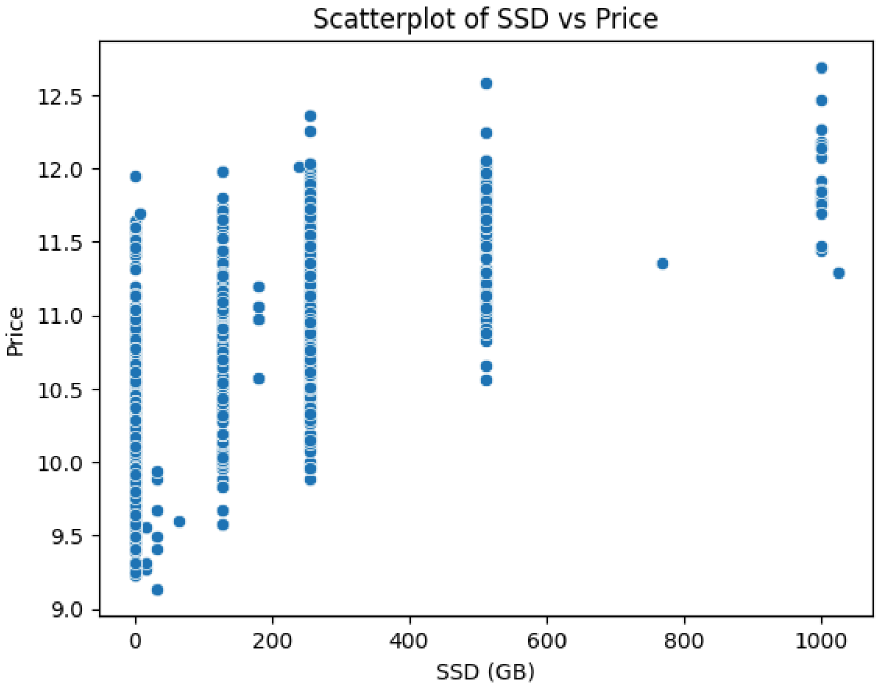


Figure 5: Scatterplot of Relationship between SSD and Price

Data Exploration

Principal Component Analysis (PCA)

PCA was used to reduce the number of features in the dataset while keeping most of the important information. It works by finding patterns in the data and combining related features into new components. The goal was to find out how many components were needed to explain 96% of the data's variability.

The results showed that **7 principal components** are enough to explain 96% of the total variance in the data. The first component alone explains about **40%** of the variability, which means that some features in the dataset have a big impact. The remaining components add smaller amounts of variability. This tells us that not all the original features are equally important, and some of them might be closely related or redundant.

From this, we can understand that features like **RAM, SSD size, and price** likely have the strongest influence on the dataset. PCA helps to simplify the data, making it easier to analyze and work with, especially for building models, while keeping most of the important information.

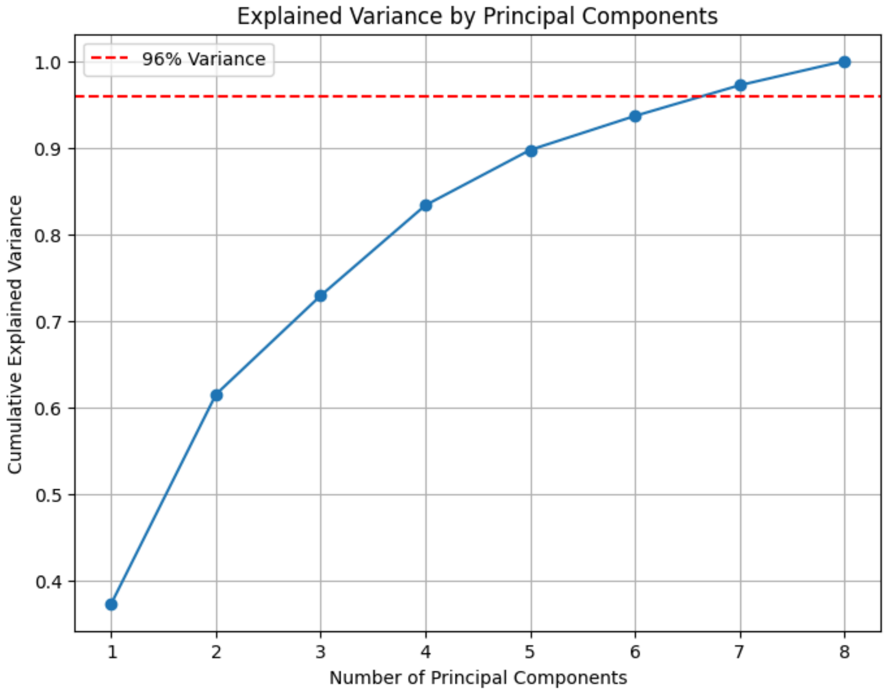


Figure 6: Explained Variance Ratio by Number of Components

Decision Trees

A simple decision tree was created to understand which features are most important for predicting laptop prices. The tree was kept small (maximum depth of 3) so it would be easy to understand. Decision trees work by splitting the data into groups based on the most important features.

The first split in the tree was based on RAM, showing that it is the most important feature for determining price. For laptops with less RAM (e.g., 8 GB or less), the next most important feature was SSD size. Laptops with larger SSDs (e.g., 512 GB or more) tend to have higher prices. For laptops with more RAM (e.g., over 8 GB), PPI (pixels per inch) became important, with higher screen resolutions linked to higher prices.

This tree gives a simple explanation of how laptop features influence prices. For example, laptops with low RAM and small SSDs are usually cheaper, while those with high RAM and better screens are more expensive.

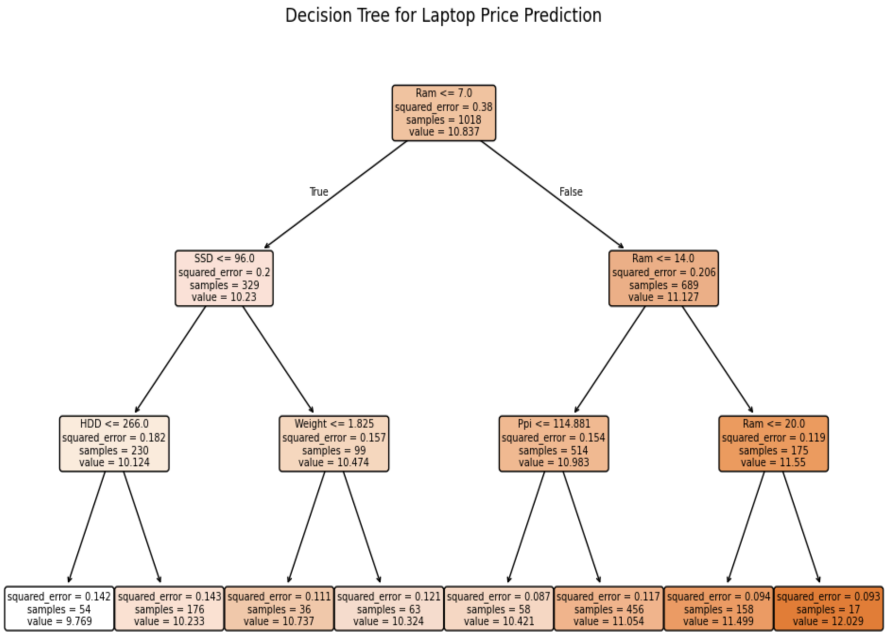


Figure 7: Regression Decision Tree for Laptop Price Prediction

Experimental Method

Hypothesis

I believed that machine learning models like Random Forest, Gradient Boosting, AdaBoost, and Multi-layer Perceptron (MLP) could predict laptop prices based on features such as RAM, SSD size, and screen resolution. I expected that ensemble methods like Random Forest and Gradient Boosting would perform well because they combine multiple decision trees and handle complex patterns. MLP, being a neural network, was expected to need careful adjustments to compete with the ensemble models.

Experiment Details

1. **Feature Preparation:**
    - Numerical features such as Ram, Weight, Ppi, HDD, and SSD, along with binary features like TouchScreen and Ips, were used as inputs.
    - Numerical features were standardized for Gradient Boosting and MLP to improve performance.

2. **Model Training:**
    - Models were initially trained using default hyperparameters to establish baseline performance:
        - Random Forest
        - Gradient Boosting

    - AdaBoost
    - Multi-layer Perceptron (MLP)

3. **Hyperparameter Tuning:**
    - GridSearchCV was applied to optimize key parameters:
        - **Random Forest**: n_estimators, max_depth, min_samples_split, and min_samples_leaf.
        - **Gradient Boosting:** learning_rate, n_estimators, max_depth, min_samples_split, and min_samples_leaf.
        - **AdaBoost**: n_estimators, learning_rate, and estimator.
        - **MLP:** hidden_layer_sizes, activation, learning_rate_init, and max_iter.

4. **Evaluation:**
    - Performance was evaluated using Mean Squared Error (MSE) on the test set and 5-fold cross-validation (CV) to ensure generalization.

Results and Insights

1. Random Forest

- **Performance:**
  - Test MSE: **0.0817**
  - Cross-validation MSE: **0.0802**
- **After Tuning:**
  - Best Test MSE: **0.0782** with n_estimators = 100, max_depth = None, and min_samples_split = 5.
- **Insights:**
  - Random Forest performed consistently well, showing small differences between test and validation results.
  - It worked well because it captures complex patterns in the data without overfitting, thanks to its averaging technique.
- The model showed that **RAM, SSD size, and PPI** were the most important features. Laptops with higher RAM and SSD size tend to have higher prices. Screen resolution (PPI) also mattered, indicating that high-quality displays are a premium feature.

2. Gradient Boosting

- **Performance:**
  - Test MSE: **0.0758**

    - Cross-validation MSE: **0.0807**
- **After Tuning:**
    - Best Test MSE: **0.0725** with learning_rate = 0.05, n_estimators = 200, and max_depth = 5.
- **Insights:**
    - Gradient Boosting outperformed Random Forest by learning patterns in the data more precisely through sequential improvements.
    - It captured subtle relationships between features, such as how **SSD size and PPI together influence price**.
- Gradient Boosting benefited from a smaller learning rate, which allowed the model to improve more gradually and avoid overfitting.

3. AdaBoost

- **Performance:**
    - Test MSE: **0.0987**
    - Cross-validation MSE: **0.1105**
- **After Tuning:**
    - Best Test MSE: **0.0830** with n_estimators = 200, learning_rate = 0.1, and estimator = DecisionTreeRegressor(max_depth = 3).
- **Insights:**
    - AdaBoost was less effective than Gradient Boosting or Random Forest. It struggled to capture the complex relationships in the data.

    - Using shallow trees (max_depth=3) helped the model focus on simpler patterns.
- The performance of AdaBoost shows it works better on simpler problems. While it improved after tuning, it **couldn’t match the accuracy** of Gradient Boosting or Random Forest.

4. Multi-Layer Perceptron (MLP)

- **Performance:**
    - Test MSE: **0.1159**
    - Cross-validation MSE: **0.1180**
- **After Tuning:**
    - Best Test MSE: **0.1050** with hidden_layer_sizes=(100, 100), activation='tanh', and learning_rate_init=0.01.
- **Insights:**
    - MLP struggled to match the performance of ensemble methods. It required careful tuning of layers and learning rates but still had difficulty converging.
    - It is likely less effective for this task because it needs more data or better feature engineering to learn complex relationships.
- While MLP improved with tuning, it remained **less effective** than Random Forest or Gradient Boosting. It’s better suited for tasks with more data or highly complex patterns.

Observations

The analysis showed that some features were very important in predicting laptop prices. **RAM was the strongest factor**, as laptops with more RAM were usually more expensive. Similarly, **the size of the SSD had a big impact** on price. Laptops with larger SSDs cost more because faster and bigger storage is in high demand. **Screen resolution**, measured in pixels per inch (PPI), also **affected the price**. Laptops with higher screen resolutions are typically premium models, so they are priced higher.

Among the models tested, **Gradient Boosting performed the best**, with the lowest test error (MSE: 0.0725) after tuning. **Random Forest was a close second**, with a test error of 0.0782, and it showed stable results across validation tests. AdaBoost did fairly well but wasn't as good at handling the more complex relationships in the data. The MLP model, even after trying different settings, didn't perform as well as the others. It had trouble learning from the data and was sensitive to how the features were scaled.

When comparing the methods, ensemble models like **Random Forest and Gradient Boosting were better** for this dataset. They could handle the relationships between features effectively and didn't need much extra preparation of the data. In contrast, MLP might have performed better with a larger dataset or more advanced changes to the features. Overall, the ensemble models were the most suitable for predicting laptop prices in this case.

Practical Application

The machine learning model can help laptop retailers and manufacturers set the right prices for their products based on features like **RAM, SSD size, and screen resolution**. It can also guide product development by showing which features are most important for higher prices, helping manufacturers design laptops that meet market demands. For online shopping platforms, the model can be used to recommend laptops to customers based on their budget, improving the shopping experience. Additionally, the model can predict trends in laptop prices and demand, helping businesses manage their stock better. These applications make the model useful for both improving business decisions and helping customers make smarter choices.