Project Report: Laptop Price Prediction

Author: [Saba Akram]
Date: September 19, 2025

Abstract

This report outlines the process of developing a machine learning model to predict the price of laptops based on their technical specifications. The project utilizes a dataset containing various laptop attributes, including Company, CPU, GPU, RAM, Memory, and screen details. The process involved comprehensive data cleaning, feature engineering, and exploratory data analysis to uncover insights. Several regression models were trained and evaluated, with the **Random Forest Regressor** demonstrating the highest accuracy, achieving an R-squared score of approximately 89%. This report details the methodology, model performance, and conclusions drawn from the project.

1. Introduction

The price of a laptop is determined by a complex interplay of its hardware and software components. For consumers and manufacturers, understanding the value of these components is crucial. This project aims to build a robust regression model that can accurately predict the price of a laptop given its specifications. By analyzing features such as RAM, storage type (SSD/HDD), processor brand, and screen quality, the model provides an estimated market price, serving as a useful tool for price benchmarking and analysis.

Dataset: The primary dataset used for this project contains information on over 1300 laptops. The key attributes include:

- Company, TypeName, OpSys
- Inches, ScreenResolution
- Cpu, Gpu
- Ram, Memory, Weight
- Price (Target Variable)

2. Data Cleaning and Feature Engineering

The initial dataset required significant preprocessing to be suitable for modeling. The following steps were undertaken:

- **Data Type Conversion:**
  - The Ram column was converted from a string format (e.g., '8GB') to a numerical integer format (e.g., 8).
  - The Weight column was converted from a string format (e.g., '1.37kg') to a numerical float format (e.g., 1.37).
- **Feature Engineering from Existing Columns:**
  - **Touchscreen:** A binary feature (Touchscreen) was created by parsing the

ScreenResolution column to identify laptops with touch capabilities.
    - **IPS Display:** A binary feature (IPS) was created to indicate whether a laptop has an IPS panel, also derived from the ScreenResolution column.
    - **Screen Resolution (X & Y):** The ScreenResolution was split into two new numerical columns, X_res and Y_res, representing the horizontal and vertical pixel counts.
    - **Pixels Per Inch (PPI):** A ppi feature was engineered using the screen size (Inches) and resolution (X_res, Y_res) to create a more informative display quality metric.
    - **CPU Brand:** The Cpu column was simplified to extract the brand (e.g., Intel Core i5, Intel Core i7, AMD Ryzen, etc.), creating a new Cpu_brand feature.
    - **Memory Type:** The Memory column was transformed to create four new binary features representing the presence and capacity of different storage types: HDD, SSD, Hybrid, and Flash_Storage.

After cleaning and feature engineering, irrelevant columns like ScreenResolution, Inches, Cpu, and Memory were dropped to avoid data redundancy.

3. Exploratory Data Analysis (EDA)

EDA was performed to understand the data distribution and the relationship between various features and the laptop's price.

- **Price Distribution:** The target variable, Price, was found to be right-skewed. A logarithmic transformation (np.log(df['Price'])) was applied to normalize its distribution, which helps improve the performance of linear-based regression models.
- **Categorical Features:** Bar charts were used to visualize the relationship between categorical features (like Company, TypeName, OpSys, Cpu_brand) and the average price. It was observed that companies like Razer and Apple have a significantly higher average price, while certain operating systems like Mac OS and Windows 10 Pro are associated with more expensive devices.
- **Numerical Features:** Scatter plots and correlation analysis revealed strong relationships between numerical features and price.
    - A strong positive correlation was observed between Price and Ram, SSD capacity, and ppi.
    - Weight also showed a positive correlation with price.
    - The correlation heatmap confirmed these relationships and highlighted the low correlation between HDD, Flash_Storage, and Price.

4. Model Development

The final phase of the project focused on building and evaluating various regression models.

- **Data Preparation:** The dataset was split into training and testing sets with an 80:20 ratio.
- **Categorical Encoding:** The categorical columns (Company, TypeName, OpSys, Cpu_brand, Gpu_brand) were transformed using OneHotEncoder to convert them into a numerical format that the models can process.

- **Modeling Pipeline:** A ColumnTransformer and Pipeline from Scikit-learn were used to streamline the process of encoding categorical features and training the regression models.

The following regression algorithms were implemented and compared:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Decision Tree Regressor
5. Support Vector Regressor (SVR)
6. Random Forest Regressor
7. AdaBoost Regressor
8. Gradient Boosting Regressor
9. XGBoost Regressor
10. Voting Regressor (ensemble of Random Forest, Gradient Boosting, and XGBoost)

5. Model Evaluation and Comparison

The models were evaluated based on two key metrics: **R-squared (R2) Score** and **Mean Absolute Error (MAE)**. The R2 score indicates the proportion of the variance in the price that is predictable from the features, while MAE measures the average magnitude of the errors in a set of predictions.

The performance of each model is summarized below:

Model	R2 Score	Mean Absolute Error
Random Forest	**0.89**	**0.16**
Voting Regressor	0.88	0.16
Gradient Boosting	0.87	0.17
XGBoost	0.87	0.17
Decision Tree	0.83	0.20
AdaBoost	0.76	0.24
Linear Regression	0.69	0.29
Ridge Regression	0.69	0.29
Lasso Regression	0.66	0.30

Support Vector Regressor	0.57	0.32

The **Random Forest Regressor** emerged as the best-performing model with the highest R2 score of 0.89 and one of the lowest MAE values (0.16). This indicates that the model can explain about 89% of the variability in laptop prices, making it a reliable and accurate predictor.

6. Conclusion

This project successfully developed a machine learning model for predicting laptop prices. Through meticulous data cleaning, insightful feature engineering, and a comparative analysis of multiple regression algorithms, the Random Forest model was identified as the most effective solution.

The final model and the preprocessing pipeline were saved using pickle for future deployment, allowing for easy prediction on new, unseen laptop data.

Future Work:

- **Hyperparameter Tuning:** Further improvements could be achieved by performing hyperparameter tuning on the top-performing models (e.g., Random Forest, XGBoost) using techniques like GridSearchCV.
- **Advanced Models:** Exploring more complex deep learning models could potentially capture more intricate patterns in the data.
- **Data Expansion:** Incorporating a larger and more current dataset could enhance the model's accuracy and generalizability.