Predictive Analysis of Laptop Prices Using Machine Learning

Overview

This project focuses on predicting laptop prices using a reliable machine learning model. The dataset consists of 11 columns, each detailing various features of the laptops. The process begins with understanding the data, followed by cleaning, feature engineering, feature selection, and finally, model building and evaluation.

Data Understanding

The dataset includes the following features:

1. Company: There are about 19 unique laptop manufacturers.

2. TypeName: There are 6 unique types of laptops:

- Ultrabook

- Notebook

- Gaming

- 2 in 1 Convertible

- Workstation

- Netbook

3. Inches: The screen size varies from 11 to 17 inches.

4. ScreenResolution: This column includes details about screen resolution, IPS panel display, and touchscreen capability. These details are extracted into separate columns: IPS, Touchscreen, and Resolution.

5. Pixels Per Inch (PPI): Derived from the resolution and inches columns using the formula:

PPI= sqrt{x^2 + y^2} / inches

6. CPU: This column contains detailed information about the processor. Processors are categorized into two main brands: Intel and AMD.

7. RAM: This feature includes unique values of RAM sizes.

8. Memory: This feature details HDD and SSD sizes and indicates whether the laptop has Flash Storage or is Hybrid. These details are extracted into four separate columns.

9. GPU: This column contains detailed information about the graphics processor. GPUs are categorized by brand: Intel, AMD, Nvidia, and ARM.

10. Operating System (OpSys): Details different Windows and MacOS versions, which are grouped by operating system name.

11. Weight: The weight of the laptops ranges from 1 kg to 4 kg.

12. Price: The target column, representing the price of the laptops.

Data Cleaning

Upon reviewing the data, 30 missing values were found and subsequently dropped. Further cleaning was done to remove any unwanted characters.

Feature Engineering

- The ScreenResolution column was split into three columns: IPS, Touchscreen, and Resolution.

- PPI was calculated from the resolution and inches.

- CPU and GPU names were categorized.

- The Memory column was split into four separate columns.

- Unnecessary columns were dropped, and new ones were formed based on the extracted features.

Feature Selection

To identify the least important features for predicting laptop prices, a correlation matrix and RandomForestRegressor were used. It was determined that the price is least dependent on the FlashStorage and Hybrid features, so these columns were dropped.

The refined dataset consists of 12 features and 1 target variable.

Model Building

Given that this is a regression problem, the following models were used:

- Linear Regression

- K-Nearest Neighbors (KNN)

- Support Vector Machine (SVM)

- Decision Tree

- Random Forest

- Gradient Boosting Regressor

Categorical columns were converted to numerical values using OneHotEncoder. The data was split into training and test sets with an 85-15 ratio. Each model was fitted to the data to predict prices, and their performance was evaluated using R2 Score and Mean Absolute Error (MAE).

Among all models, the Gradient Boosting Regressor had the highest R2 score and the lowest MAE.

Hyperparameter Tuning

The Gradient Boosting Regressor model was further optimized using RandomizedSearchCV to find the best estimators.

Deployment

The final model was exported to a pickle file, which was then used to build an application using the Tkinter library.

This project demonstrates a comprehensive approach to predictive analysis, from data understanding to model deployment, ensuring robust and accurate price predictions for laptops.