Laptop Price Predictor

Course : Introduction to Machine Learning DA 515

Done By:
Aswath Santhanakrishnan & Ashwin Santhanakrishnan

Abstract

This study focuses on predicting laptop prices using supervised machine learning models, including Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest. The Random Forest algorithm emerged as the most effective model, achieving an improved R^2 score of approximately 0.88 from an initial score of 0.80. This demonstrates that predictive modeling based on hardware specifications provides significant insights for both consumers and retailers in understanding laptop pricing dynamics.

Introduction

The laptop price predictor functions as a tool to supply cost forecasts for customers through the input of screen size along with touchscreen choices and financial limitations. The device helps retailers form their products to better match what customers want. Through this tool users gain information about accurate price evaluations while receiving insights into different hardware setups to support better decisions.

Dataset Description

Laptop Price Dataset serves as the main dataset for this project and originates from Kaggle. The dataset contains complete specifications along with pricing data from different laptop models enabling efficacious predictive modeling of prices based on components.

Key Characteristics of the Dataset:

- **Total Entries:** 1,303 laptop records
- **Attributes (Columns):** 12
    - **Company:** The research involved study of 19 discrete laptop producers where Dell turned out to be the organization which appeared most often.
    - **TypeName:** Type/category of laptop (categorical, 6 types, most common: Notebook)
    - **Inches:** Screen size measured diagonally (numeric, range: 10.1 to 18.4 inches; mean ~15 inches)

- **ScreenResolution:** Screen resolution details (categorical, 40 unique resolutions, most frequent: Full HD 1920x1080)
- **Cpu:** The study examines processor specifications which include brand information together with clock speed measurement (categorical data contains 118 distinct CPU brands and most common was Intel Core i5 7200U 2.5GHz).
- **Ram:** Amount of RAM installed (categorical, common values range from 2GB to 32GB, most frequent: 8GB)
- **Memory:** Storage type and capacity information (categorical, 39 unique configurations, most common: 256GB SSD)
- **Gpu:** Graphics processing unit specification (categorical, 110 unique GPUs, most frequent: Intel HD Graphics 620)
- **OpSys:** Users choose from 9 possible operating systems for installation with Windows 10 being the most selected.
- **Weight:** Weight of the laptop (categorical, ranging from very light (e.g., 1kg) to heavy, most common weight: 2.2kg)
- **Price:** The selling price of the laptop (numeric, ranging from approximately ₹9,270 to ₹324,954, with an average price around ₹59,870)

Statistical Summary:

- **Price Distribution:**
  - **Minimum Price:** ₹9,270
  - **Average Price:** ₹59,870
  - **Median Price:** ₹52,054
  - **Maximum Price:** ₹324,954
- **Dominant Categories:**
  - **Most Frequent Company:** Dell
  - **Most Frequent Laptop Type:** Notebook
  - **Most Common Screen Resolution:** Full HD (1920x1080 pixels)

The large dataset contains multiple features which help machine learning algorithms properly predict laptop market values. The EDA phase included thorough examinations of each attribute before applying them for laptop price prediction modeling.

Related Work

Researchers and learners commonly use multiple machine learning methods to predict laptop prices through public datasets and online instructional materials. Kaggle and Analytics Vidhya platforms supply resources including statistical instruction and information about laptop pricing in relation to their features which primarily serves educational and modeling needs.

Research & Methodology

Multiple machine learning techniques and analytical procedures were employed during this research to correctly forecast laptop prices. The main analytical approach contained both multiple regression model training and extensive exploratory data analysis (EDA) for feature selection while performing data preprocessing and relationship detection in the laptop dataset.

Machine Learning Models for Price Prediction

Linear Regression:

The statistical method Linear Regression fits mathematical lines to observed data to determine the connection between separate variables and continuous response variables (price). The modeling technique fits an algebraic line that connects observed points of data. Temperature in Celsius maintains a direct linear proportion with temperature in Fahrenheit. The outcome of linear regression offers an estimated connection between variables when their relationship is not strictly linear (for instance, height and weight) and determines their impact on price factors.

Linear Regression predicts the laptop price (y) as a linear function of the input features (x)

$$y=\beta 0 +\beta 1\ X1 +\beta 2\ X2 +\cdots+\beta n\ Xn +\epsilon$$

$\beta 0$ = Intercept

βi = Coefficient for feature Xi

ϵ = Error term

K-Nearest Neighbors:

This non-parametric supervised learner referred to as K-Nearest Neighbors operates in both regression and classification modes. Having received a laptop query the algorithm identifies k nearest instances in the specification-based training data to generate price estimation

through value averaging. KNN exhibits simplicity along with non-linear capability which makes it very dependent on k selection and feature normalization standards.

$$\hat{y} = k1 \sum_{i=1}^{k} y_i$$

Where:

$\hat{y}$ = predicted price

y_i = price of the ith nearest neighbor

Decision Tree:

Decision Tree serves as a tree-structured model that utilizes feature values to split data which then enables predicting target values. Each decision from a feature leads to the creation of a new node until a predicted price emerges at the leaf nodes during regression tasks where the average price serves as estimation. Decision trees are straightforward to understand because they split the data by features for easy classification while handling complex nonlinear patterns in the feature space. When used on its own a single decision tree will yield overfitting because it does not undergo pruning.

$$MSE = n1 \sum_{i=1}^{n} (y_i - \hat{y}_i)2$$

y_i = actual price,
$\hat{y}_i$ = predicted price,
n = number of observations.

Random Forest:

The random forest methodology builds multiple decision trees which it uses to generate combined outcomes. The training of each forest tree depends on random information from data and features. A Random Forest model delivers price predictions through the calculated average from all its constituent decision trees when used for regression tasks. Random Forest enables combined tree outputs to limit the overfitting tendency of single trees and usually leads to better accuracy. The Random Forest model generated the best accuracy rates when forecasting laptop prices.

Random Forest aggregates predictions from multiple decision trees (T) to enhance prediction accuracy by reducing variance:

$$\hat{y}_{RF} = T1 \sum_{t=1}^{T} \hat{y}_t$$

$\hat{y}_t$ = prediction from tree t

Our initial work consisted of both modeling stages as well as performing Exploratory Data Analysis on the laptop data before the training process. EDA examined how data features

distributed and how each parameter related to product price. Identifying anomalies in addition to identifying outliers led us towards performing feature engineering by developing the PPI measurement from resolution and screen size.

Evaluation & Results

Laptop Price Predictor

```
# Lets look at the dataset
df.head()
```

✓ 0.0s

	Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price
0	0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.6832
1	1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	47895.5232
2	2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.0000
3	3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.3360
4	4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.8080

Figure 1: Laptop Data set

Figure 1 provides a snapshot of the laptop dataset. We derived three additional features from the raw data: **PPI** (pixels per inch, calculated from screen resolution and size), and binary indicators for **Touchscreen** and **IPS** display. Selecting the most relevant features was a key challenge due to the diversity of laptop configurations and consumer preferences, which made generalization difficult.

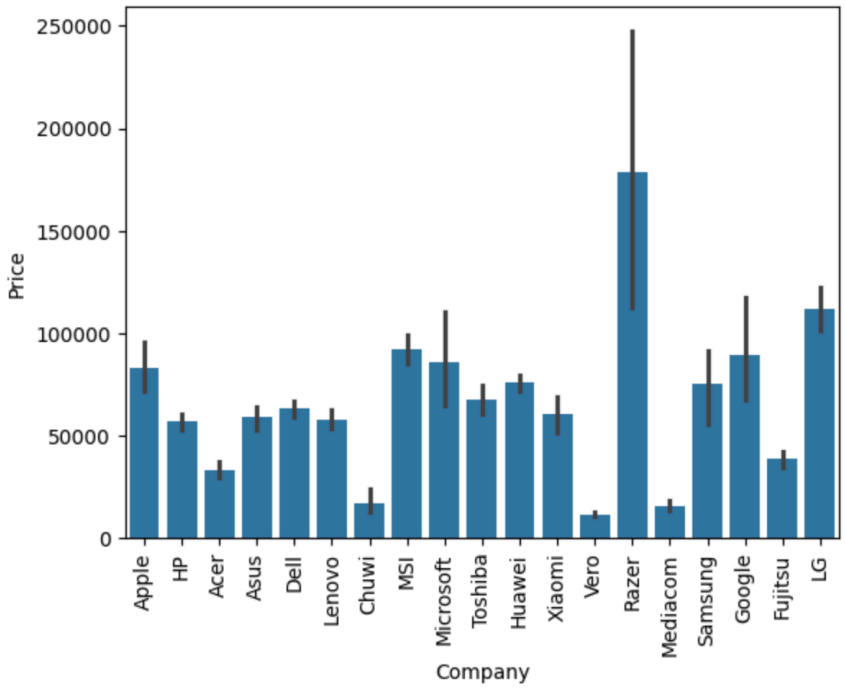


Figure 2 shows the distribution of laptop prices by manufacturer.

We conducted an EDA analysis on the data after successfully processing it. Various plots include Figure 2 among several others that we generated. The presented visual illustration shows how laptop pricing correlates with the manufacturers who produced them. Established and recognized brands within the laptop industry price their laptops at higher levels than competitors do.

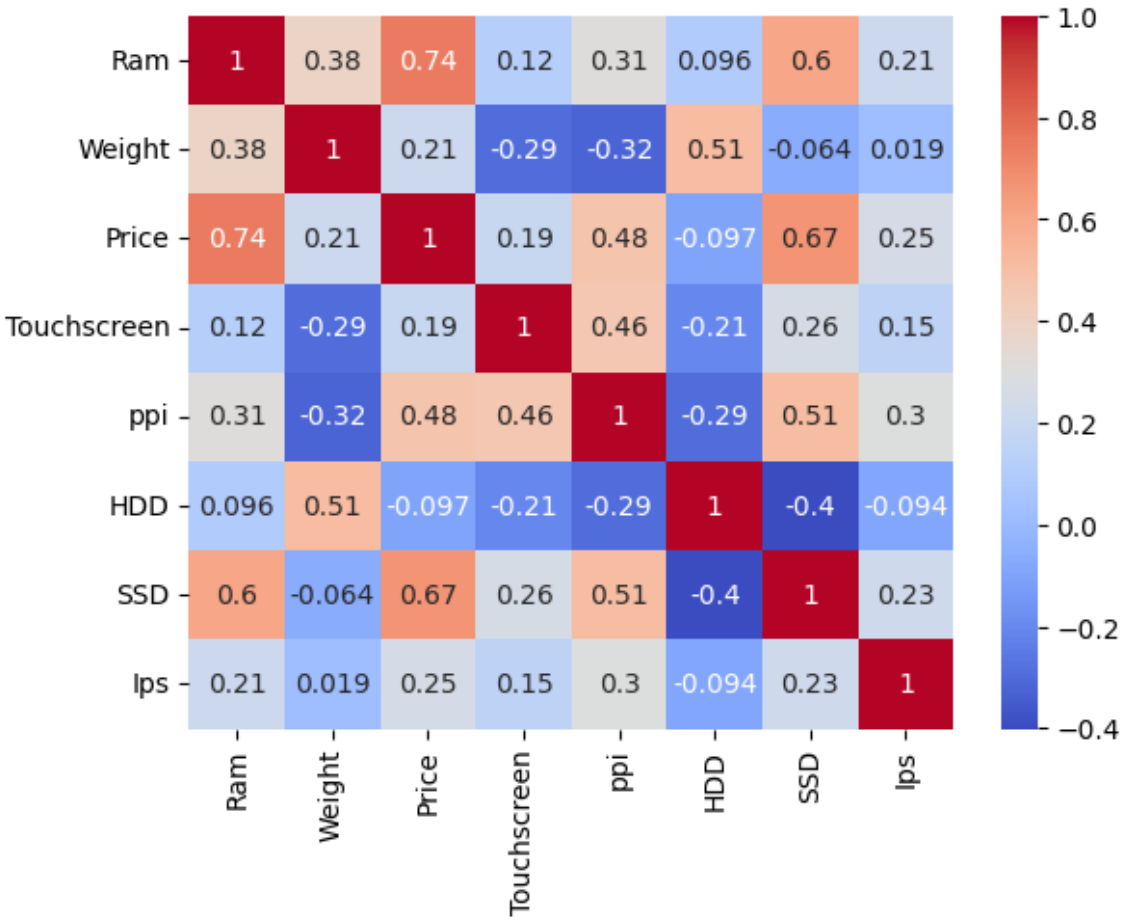


Figure 3: Correlation Matrix here

Figure 3 illustrates a correlation heatmap between all features and the laptop price. We observe that most hardware specifications (e.g., RAM size, SSD capacity) have a positive correlation with price – in other words, laptops with higher specs generally cost more. A few features show weak or negative correlation with price. This correlation analysis helped confirm which features would be important in the prediction models.

S.NO	Algorithm	R2 Score
1	Linear Regression	0.804595
2	KNN	0.79512
3	Decision tree	0.84415
4	Random Forest	0.88564

Figure 4: Model Performance Metrics here

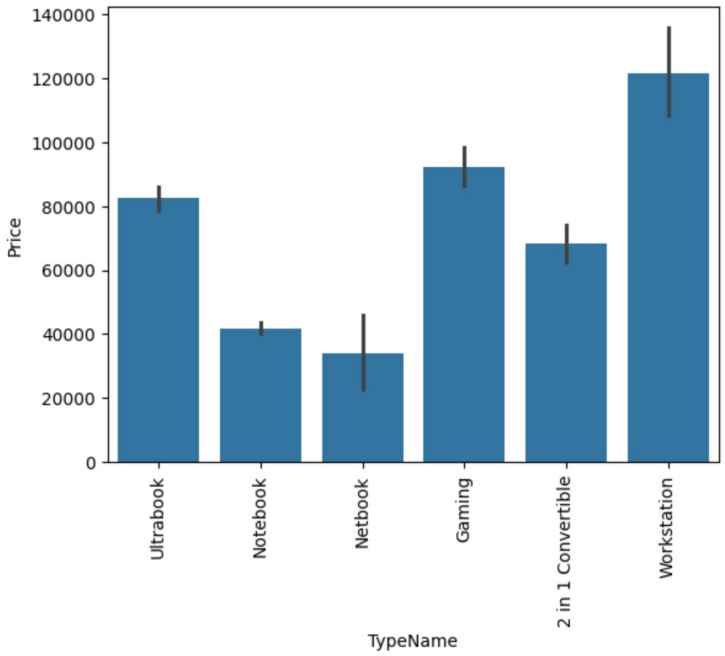
Finally, the figure above showcases the performance of our model. Our model evaluation used R2 scores as the principal measure. The performance measurements of our built models range from 0.8 to 0.9 for all models. Random Forest proved to be the optimal model with its superior performance, which was no mystery. Our model included attributes of discrete form and numerical form since its construction began. NOTE - Our first model came back with an R2 score of 0.80. And we finally managed to get a score of 0.88 in the end, hence improving upon our initial model score by quite a factor!

Model	R^2 Score	Mean Absolute Error (MAE)	Mean Squared Error (MSE)
Linear Regression	0.804	0.2104	0.07475
KNN	0.795	0.1966	0.07837
Decision Tree	0.842	0.1852	0.06077
Random Forest	**0.88**	**0.1607**	**0.04375**

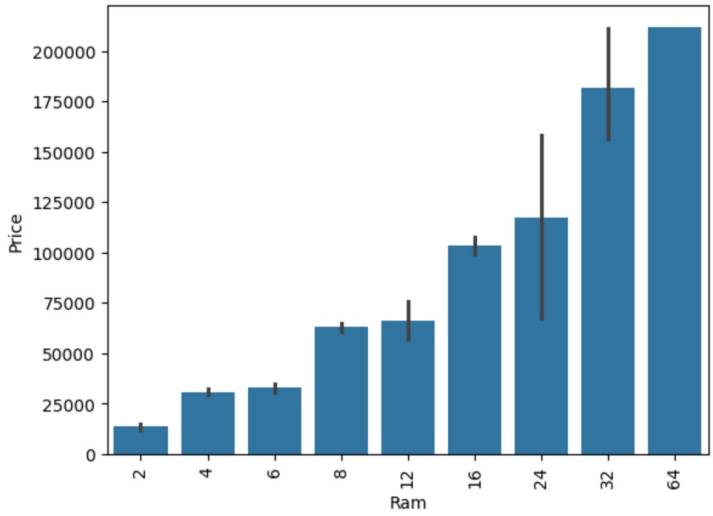
Signal the finest performance with an R ^{2} value of 0.88 as well as the lowest MAE of 0.1607 arrived from Random Forest model, making it the most reliable model for laptop price forecasts. The consistent trend indicates that ensemble methods win over more basic models in case feature sets are highly feature.

Analysis Images:

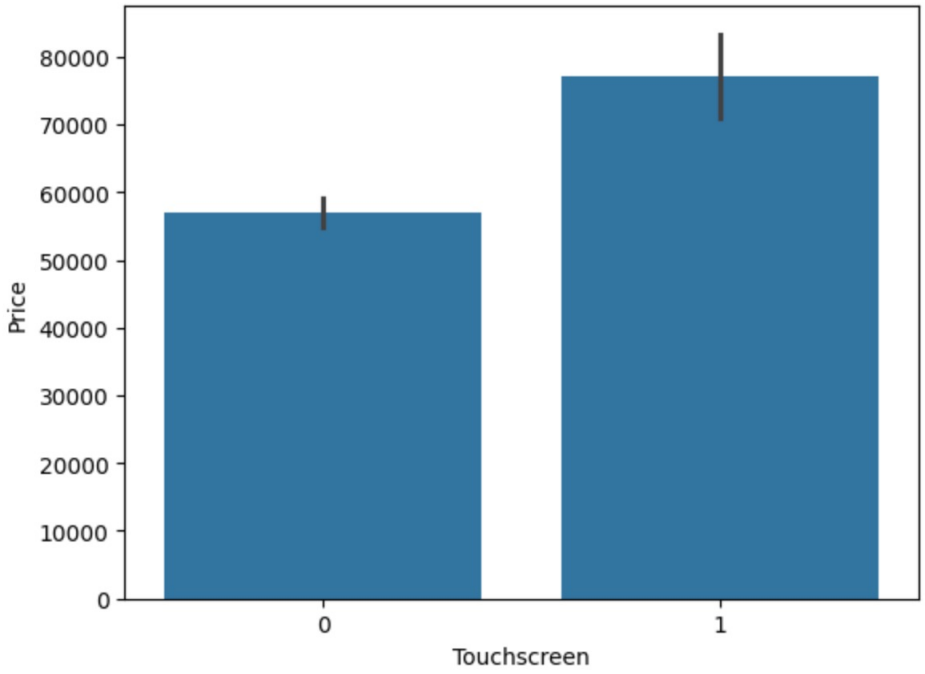
1. **Different types of laptops and there prize range.**



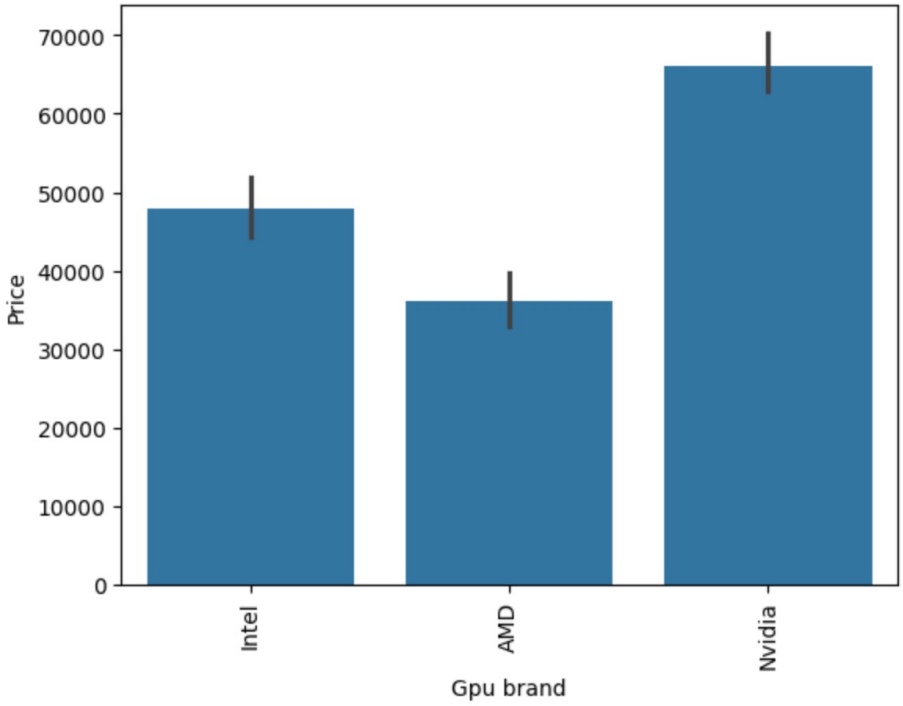
2. **Different types of Ram and there prize range.**



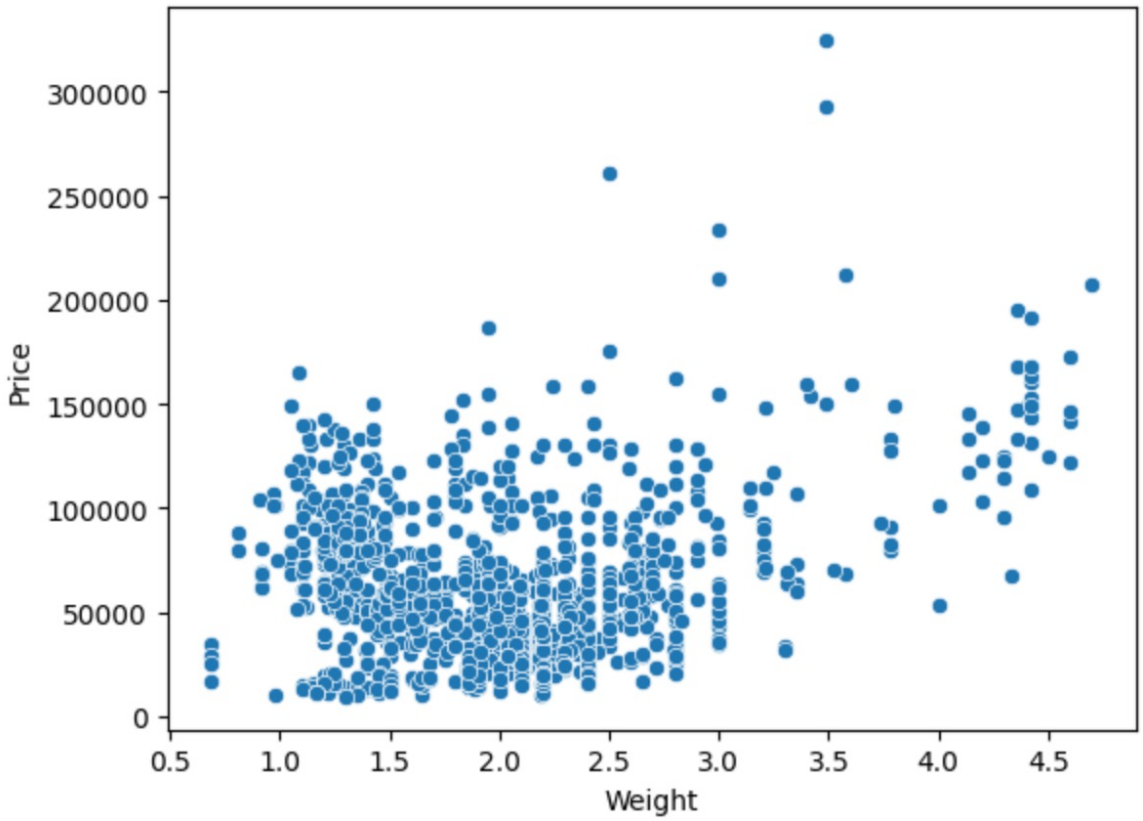
3. Different types of laptops spec in screens and there prize range.



4. Different types of GPU Brand and there prize range.



5. Different types of laptops weights and there prize range.



Final Findings

When applied to laptop price estimation through technical specifications Random Forest ensemble techniques displayed the most effective performance according to the prediction model. Through this technology consumers can find appropriate laptops that fit their budget requirements while retailers use the model's foretelling abilities to develop competitive pricing tactics.

Future Work

Future opportunities include deploying the price prediction model as a web-based application for real-time price guidance. Additionally, further enhancements could involve advanced machine learning techniques such as neural networks or gradient boosting algorithms to increase prediction accuracy.

References

1. Kaggle (Laptop Dataset). "Laptop Price Dataset." Retrieved from: https://www.kaggle.com/datasets/muhammetvarl/laptop-price.

2. IBM Cloud Education. "Linear Regression." IBM Cloud Learn Hub. Retrieved from: https://www.ibm.com/cloud/learn/linear-regression.

3. Scikit-learn Documentation. "Decision Trees." Retrieved from: https://scikit-learn.org/stable/modules/tree.html.

4. Analytics Vidhya (2022). "Laptop Price Prediction – Practical Understanding of Machine Learning Project Lifecycle." Retrieved from: https://www.analyticsvidhya.com/blog.

5. McKinney, Wes (2010). "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56.

6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830. Retrieved from: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

7. Wikipedia contributors. (2024). "Random Forest." *Wikipedia, The Free Encyclopedia*. Retrieved from: https://en.wikipedia.org/wiki/Random_forest.

8. Wikipedia contributors. (2024). "Linear Regression." *Wikipedia, The Free Encyclopedia*. Retrieved from: https://en.wikipedia.org/wiki/Linear_regression.

9. Wikipedia contributors. (2024). "Decision Tree Learning." *Wikipedia, The Free Encyclopedia*. Retrieved from: https://en.wikipedia.org/wiki/Decision_tree_learning