SCHOLARLY DEBUT
by Scholarly Review

Winter 2025/2026

Predicting Laptop Prices Using Machine Learning and Feature Engineering

By Qianlong He

ABSTRACT

The prices of laptops are highly diverse depending on a blend of the hardware requirements and the brand position, and as such both consumers and retailers find it hard to ascertain their actual worth and the retailer to price the products competitively. This paper uses the machine learning method to estimate laptop prices based on features of the product in detail, using the publicly available Brand Laptops Dataset available on Kaggle, which includes specifications of 991 laptop models. The information contained on the dataset will be processor setup, memory size, storage format, monitor size, operating system and brand name. Following the initial data cleaning, the heavy feature engineering was carried out in the presence of the additional informative variables like the numerical indicators of CPU tier and the operating system type. Unnecessary columns were eliminated, and all the numerical variables were standardized to counterbalance the same contribution as they contribute to the training of the model. This was followed by the development of a linear regression model that was used to estimate price based on these engineered features. An 80/ 2 0 train- test split was used to assess model performance. The model on the training set got a Mean Absolute Error (MAE) of around 17,261, and the validation set got 15,702, which represents an average prediction error of 20 percent of the cost of the laptop. These results indicate that a fairly straightforward regression expression is capable of creating key pricing patterns motivated by fundamental elements of performance. The suggested solution offers a benchmarking tool that could help both consumers and retailers in determining the price fairness and/or the pricing strategy optimization. The future enhancements might consist of more sophisticated machine learning models and other functions like the GPU classification and the build quality metrics.

Keywords: *pricing of laptops, machine learning, regression model, feature engineering, consumer electronics, predictive analytics.*

INTRODUCTION

The diversities in the laptop market are considerable, and pricing differences are very strong based on performance in the hardware, brand perception, and positioning products. Current laptops vary widely in the central processing unit (CPU) architecture, graphics abilities, memory bandwidth, screen technology alongside storage platforms, which support the performance and cost of a product [1]. Meanwhile, strong brand impacts have been determined on the consumer willingness to pay as older manufacturers receive better prices even when similar hardware is offered [2]. A combination of these factors complicates the pricing fairness environments and competitiveness of new products in the market by consumers and the retailers.

Machine learning (ML)[7], has emerged as a more and more effective instrument of consumer electronics price contraction analysis [3]. Previous studies have shown that the models through regression could be used to find out the nonlinear correlations between product specifications and market value, giving price estimations that are more objective and data-driven than the ones achieved by other methods [4]. Nevertheless, most current literature uses high-end datasets or commercial retail platforms, which creates a difference in the available and clear models that may aid consumer decision-making on a bigger scale [5].

The goal of the current research is to come up with a predictive model that approximates the prices of laptops depending on the technical features and brand attributes using a publicly available dataset. Through data preprocessing and feature engineering methods (especially illustrating the usage of CPU tier representation and standardized hardware metrics) the given research intends to identify which of the specifications has the greatest impact on cost. Moreover, the given project gives some information on the comparative relevance of the performance features of RAM and storage versus the factors of branding. The final objective is to make a valuable input regarding a useful baseline model that will help consumers assess the pricing fairness and also help retailers in formulating pricing strategy guided by quantifiable aspects of performance.

METHODS

This analysis utilizes the Brand Laptops Dataset from Kaggle, a comprehensive collection of specifications for 991 different laptops. This dataset was chosen for its detailed features, which provide a robust foundation for analyzing laptop pricing dynamics. It includes 22 attributes such as brand, model, price, processor details, RAM, storage, GPU, and display size. The richness of this data is particularly well-suited for feature engineering, allowing for the extraction of more granular information to improve model accuracy.

Before modeling, the data underwent several preparation steps to ensure quality and create meaningful features for the machine learning algorithms. The process began with data cleaning, where the dataset was checked for missing values. Fortunately, the data was complete, so this step was straightforward. The only initial modification was the removal of a redundant 'index' column.

Significant feature engineering was then performed to convert categorical data into a numerical format suitable for machine learning algorithms [7]. The 'brand' column, for example, was simplified by consolidating fewer common brands into another category, which allowed the model to focus on the market's top three manufacturers: Asus, HP, and Lenovo. The 'processor tier' feature was re-engineered into distinct numerical columns named 'core' and 'ryzen' to represent the CPU series hierarchy, such as Core i3 versus Core i7. This numerical representation better captures the impact of processor performance on price. Similarly, the 'OS' column was converted into a binary feature to indicate whether a laptop runs on Windows.

Several columns were dropped to streamline the model and avoid redundancy. The 'Model' and 'Num threads' columns were removed because processor brand and core count already provided sufficient performance information. The 'secondary_storage_type' and 'secondary_storage_capacity' columns were also excluded, as the vast majority of laptops in the dataset lacked secondary storage. The 'year_of_warranty' was dropped as well due to its limited variability.

Finally, the prepared dataset was split into training and validation sets, with 80% of the data allocated for training the model and the remaining 20% for testing its performance. All features were then scaled using StandardScaler to normalize their range, a crucial step that prevents attributes with larger numerical values from disproportionately influencing the model's predictions.

MODEL TRAINING

With the data prepared, the next step was to train a model to predict laptop prices based on their features. Given the goal of predicting a continuous value (price), this is a regression problem. For this project, a straightforward and interpretable approach was taken by implementing a linear regression model. Linear Regression model from the scikit-learn library used to observe the results.

The linear regression model was trained using the prepared and scaled training dataset. The model's initial performance was evaluated by comparing its price predictions against the actual prices in the training set [8]. A scatter plot of these predicted prices versus the actual prices showed a clear positive trend, indicating that the model was successfully learning the general relationship between a laptop's features and its cost. However, the points were scattered around the ideal diagonal line, which suggested the presence of prediction errors.

RESULTS

The linear regression model indicated that it generally had a high capacity to identify the relationship between the features of the laptop and the pricing patterns [8]. Scatter plotting of the predicted and actual prices of the training data revealed that there is an obvious positive relationship that points to the ability of the model to learn the most important patterns that drive prices. But the diffusion of points about the proposed ideal diagonal line connoted the existence of prediction variance of particular settings.

- Mean Absolute Error (MAE) was used as the model performance metric as it is a generally used price prediction research measure.
- In the training data, the MAE stood at the figure of 17,261, which is about 23 percent of the mean price.

With the validation dataset, the MAE also got better to about 15,702, and this translates to an error of approximately 20%, which shows that there is not much overfitting and that it can be used with new samples with reasonable reliability [3].

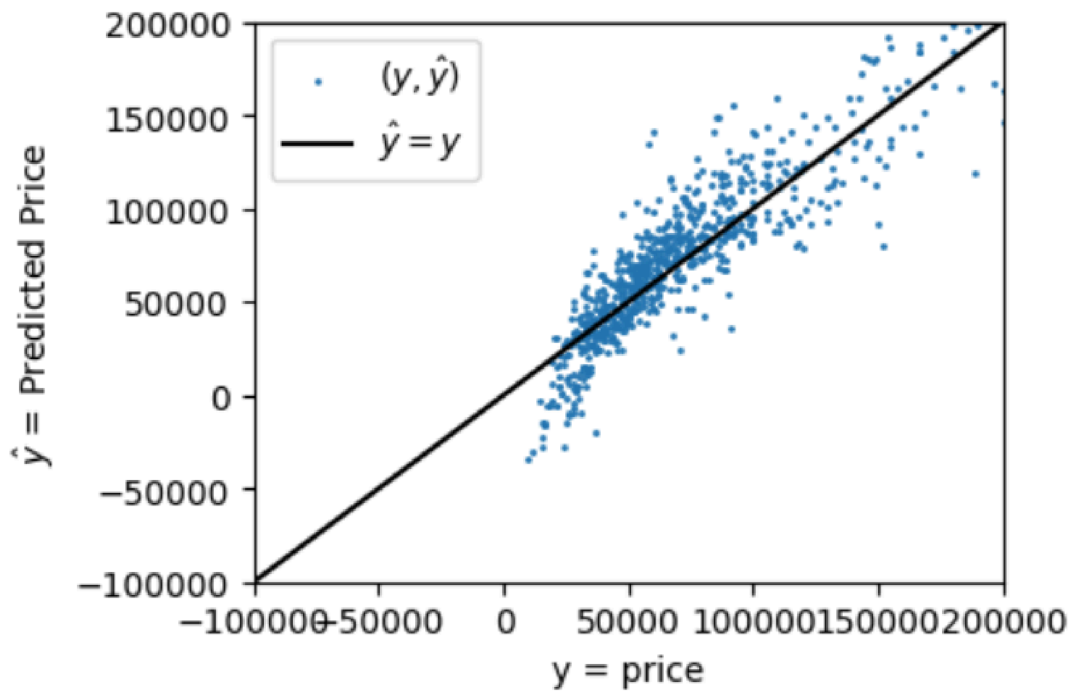

Figure 1: Predicted vs. Actual Laptop Prices on Training Data

The importance of features was evaluated with the help of standardized regression coefficients that were analyzed four times. Components related to core hardware performance, specifically tier of the processors, RAM memory, and primary memory, exhibited the largest coefficients magnitudes, which validates their key position in determining the market price [5]. Overall, such features like the type of operating system and brand category as a group did not add as significant price difference.

These findings, collectively, indicate that a basic linear regression model is capable of creating a foundation price prediction model, with a prediction within a 20 percent error margin and with strong ability to identify performance-based pricing patterns that are often prevalent in consumer electronics markets [6].

DISCUSSION

This paper has elaborated a machine learning-based model in the context of predicting prices of laptops given specifications of products. The results of this paper indicate that the key pricing trends in the consumer electronics market can be represented by the use of linear regression, indicating that it is appropriate as a baseline model when performing a systematic pricing analysis [1].

The most important findings indicate that performance-based factors namely processor tier, RAM and primary storage are the most determining price factors. This corroborates prior findings that the hardware ability is closely related with the consumer-perceived value and retail pricing strategies [2]. In the meantime, secondary characteristics like brand and operating system were of relatively little significance after the performance measures had been taken into consideration. This implies that branding does not have as much influence on consumer choices based on the price model as do quantifiable performance, but at the same time branding does not possess a strong predictive power in price modeling.

The 20% margin of error in the model is indicative of both the potential usefulness of this model and its present constraints. Linear regression is based on the assumption that the features and price are related linearly, yet the consumer electronics pricing can be nonlinear and include temporary market trends, advertising pressure, and nonlinear effects on price that are not included in the dataset [4]. These elements can explain the difference in prediction and imply that more sophisticated models can be used to achieve further accuracy.

Irrespective of these weaknesses, the predictive accuracy is high enough to be applied in practice. Such a model would help consumers to understand price fairness better in their purchasing of laptops and help retailers to maximize pricing strategies in matching products value to market expectations [5]. The next avenue of future research might be nonlinear machine learning models like random forests or gradient boosting that were demonstrated to enhance objectivity in product pricing tasks[6]. Also, one can add the information about GPU performance, display resolution and material build quality, which can improve predictor accuracy [6].

Overall, this paper offers a useful and comprehensible model of predicting the prices of laptops, which proves that the scientific method may be effectively used to facilitate consumer and retailer decision-making.

ACKNOWLEDGEMENT

The author would like to acknowledge his/her research mentor who has helped in the process of developing this project and assisted him/her with constructive comments when writing. The administrators of Kaggle are also given the credit of making the Brand Laptops Dataset publicly available, and thus, open and reproducible machine learning research is made possible.

CONCLUSION

In conclusion, this project successfully developed a machine learning model capable of predicting laptop prices based on their features. The process involved comprehensive data cleaning and feature engineering to transform the raw dataset into a format suitable for modeling. Using a linear regression model, the final prediction accuracy on unseen data was within a 20% margin of error on average. The analysis also confirmed that performance-related specifications, such as the processor, RAM, and storage, are the most significant factors in determining price. This model serves as a valuable tool for consumers seeking to assess fair value and for retailers aiming to create competitive pricing strategies. While the linear model provides a strong baseline, future work could explore more complex algorithms like neural networks to potentially improve prediction accuracy further.

REFERENCES

1. A. M. Chalwadi. Towards a unified model of consumer behaviour when it comes to electronic durable goods: An integrated approach using the SOR framework and the theory of planned behaviour. Master's thesis, Goa University, 2025.
2. B. Rai, P. B. Budhathoki. Factors affecting brand choice behavior of laptop purchases of university students in Nepal. Cogent Arts & Humanities. Vol. 10, pg. 2194126, 2023, https://doi.org/10.1080/23311983.2023.2194126.
3. M. Learning. Artificial intelligence, machine learning, and data science technologies. (No publisher), Year not specified.
4. Brand Laptops Dataset. Kaggle. https://www.kaggle.com/datasets/arnabchaki/laptop-prices, 2024.
5. E. Katya. Exploring feature engineering strategies for improving predictive models in data science. Research Journal of Computer Systems and Engineering. Vol. 4, pg. 201–215, 2023.
6. M. Danish, S. Ali, M. A. Ahmad, H. Zahid. The influencing factors on choice behavior regarding green electronic products: Based on the green perceived value model. Economies. Vol. 7, pg. 99, 2019, https://doi.org/10.3390/economies7040099.
7. N. Safaeian Hamzehkolaei, M. Alizamir. Performance evaluation of machine learning algorithms for seismic retrofit cost estimation using structural parameters. Journal of Soft Computing in Civil Engineering. Vol. 5, pg. 32–57, 2021.
8. H. Yakubu, C. K. Kwong, C. K. Lee. A multigene genetic programming-based fuzzy regression approach for modelling customer satisfaction based on online reviews: H. Yakubu et al. Soft Computing. Vol. 25, pg. 5395–5410, 2021, https://doi.org/10.1007/s00500-020-05461-y.