Personal Computer (PC) Value Estimator

Alexander Kim
STAT 418
University of California, Los Angeles

Abstract

This paper presents a machine learning-based web application designed to estimate the fair market value of CPUs and GPUs using performance benchmarks and observed retail prices. The tool addresses the growing challenge consumers face in determining whether they are overpaying for hardware and in comparing similar components based on value. By leveraging data science tools to scrape benchmark data, implement CatBoost regression models, and deploy a containerized architecture, the system enables real-time price estimation and component comparison. With a marginal error of approximately $42 in estimating prices for both CPUs and GPUs, the application provides users with a data-driven and reliable method to assess fair market value.

I. Introduction

Due to the COVID-19 pandemic in 2020, a significant portion of the global workforce shifted to remote work to reduce the spread of the virus. This sudden transition created a surge in demand for personal computing devices, particularly custom-built personal computers (PCs), as individuals sought to upgrade or build systems suitable for work-from-home and remote learning environments. Consequently, the market for PC components became highly competitive, with noticeable strain on the supply of CPUs and GPUs. According to a study published in *Patterns*, global PC shipments saw a substantial increase during this period, underscoring how the pandemic catalyzed growth in the consumer hardware market [1].

With the vast amount of new hardware released, consumers often face difficulty identifying the most cost-effective components within a specific budget. Furthermore, price fluctuations driven by supply chain disruptions, market speculation, and vendor markups have made it increasingly challenging to assess whether a component is fairly priced [2]. To address this problem, this project introduces a machine learning-based web application that estimates the fair market value of CPUs and GPUs using publicly available benchmark scores and retail price data. By combining web scraping, regression modeling, and an interactive user interface, the tool empowers users to make more informed purchasing decisions when building or upgrading a PC.

II. Data Collection and Preprocessing

Data was collected using the requests library and BeautifulSoup in Python, scraping performance benchmarks and product pricing from PassMark Software. Extracted CPU and GPU specifications included component names, PassMark benchmark scores, ranks, value scores, and associated prices. A total of 5,140 CPUs and 2,695 GPUs were retrieved.

Entries with incomplete features were omitted to minimize bias in the training dataset, and extreme outliers were removed to improve model generalization. Figures 1 and 2 showcase the before and after of outlier removal to illustrate the effect of data cleaning on the price distributions for both CPU and GPU components. The cleaned distributions exhibit reduced skewness and more realistic pricing ranges, aligning better with the target use case of enthusiast-level PC builds.

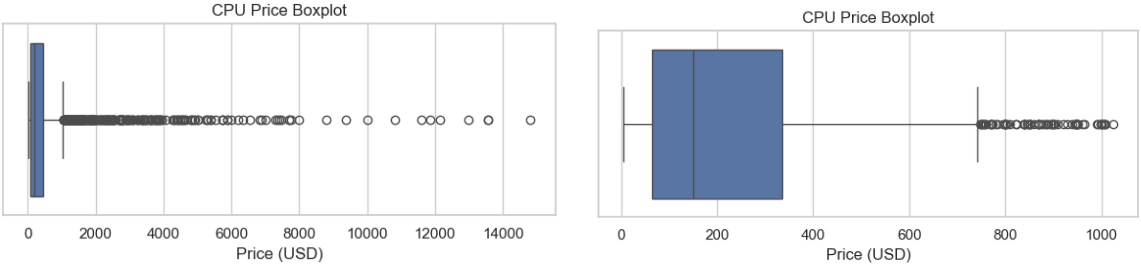

Figure 1: CPU Price Distribution Before and After Cleaning. Left: Raw CPU price distribution with extreme outliers exceeding $10,000. Right: Cleaned distribution excluding CPUs priced above $1,000 to reflect realistic consumer builds.

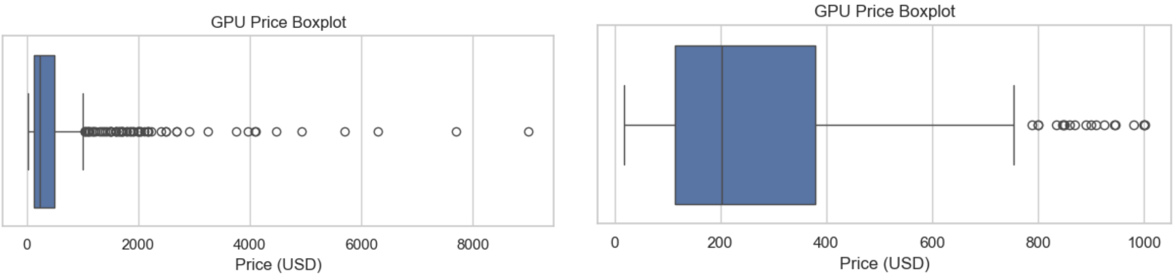

Figure 2: GPU Price Distribution Before and After Cleaning. Left: Raw GPU price distribution with prices extending beyond $8,000. Right: Filtered GPU prices capped at $1,000 to target the mainstream and enthusiast consumer segments.

To reflect typical enthusiast-level PC builds, the datasets were further filtered to include only components priced below approximately $1,000. Although some CPUs and GPUs exceed this threshold, the majority of observations are concentrated below $2,000, and meaningful performance trends are captured within this range. Higher-priced components were assumed to represent niche use cases and were excluded as outliers for standard consumer PC components. Figures 3 and 4 present the price distributions before and after this subsetting process for CPUs and GPUs, respectively.

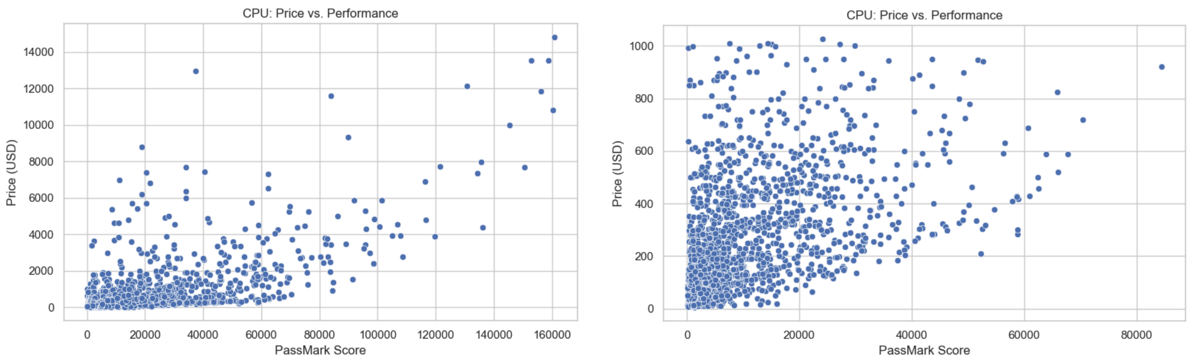

Figure 3: CPU price vs. PassMark performance before (left) and after (right) data cleaning and processing. The initial data contains extreme outliers and pricing inconsistencies, whereas the cleaned dataset on the right shows a more interpretable relationship between performance and price, improving suitability for regression modeling.

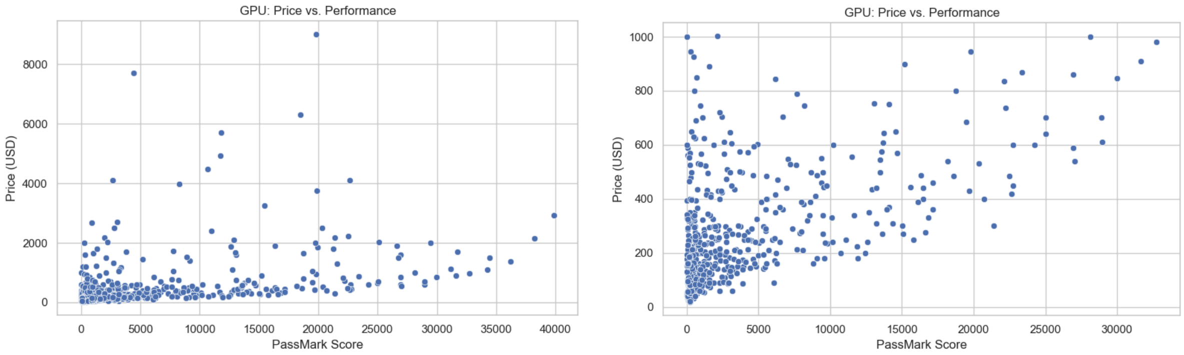

Figure 4: GPU price vs. PassMark performance before (left) and after (right) data cleaning and processing. The left plot shows numerous extreme outliers and missing structure, while the right plot—after removing overpriced entries and normalizing pricing—reveals a clearer trend between performance and price.

To ensure a more robust and representative modeling process, price distributions for both CPUs and GPUs were visually inspected before and after data cleaning. As illustrated in Figures 5 and 6, the raw datasets exhibited extreme right-skewness due to a minority of ultra-high-end components priced above $2,000. These outliers were removed to better align the dataset with typical consumer-level builds. After filtering and subsetting, prices were standardized using a log transformation to reduce skew and stabilize variance. The resulting distributions reflect a more realistic market segment, centered around sub-$1,000 components, and are more conducive to regression modeling. This preprocessing step was essential in minimizing the influence of extreme values.

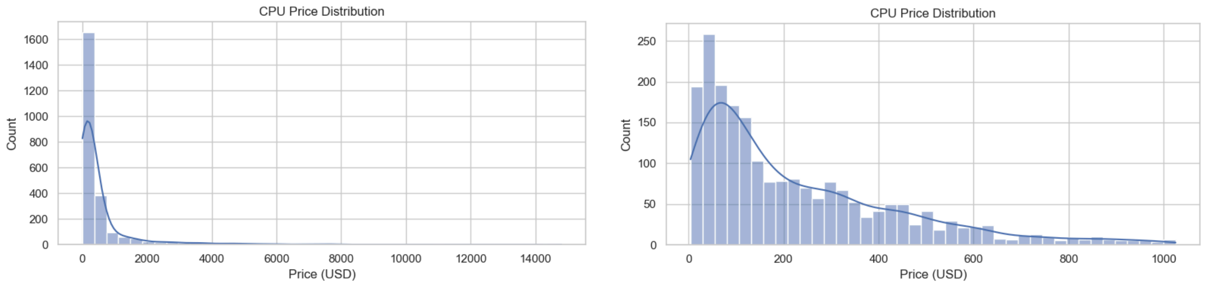

Figure 5: CPU price distributions before (left) and after (right) applying outlier removal and log transformation.

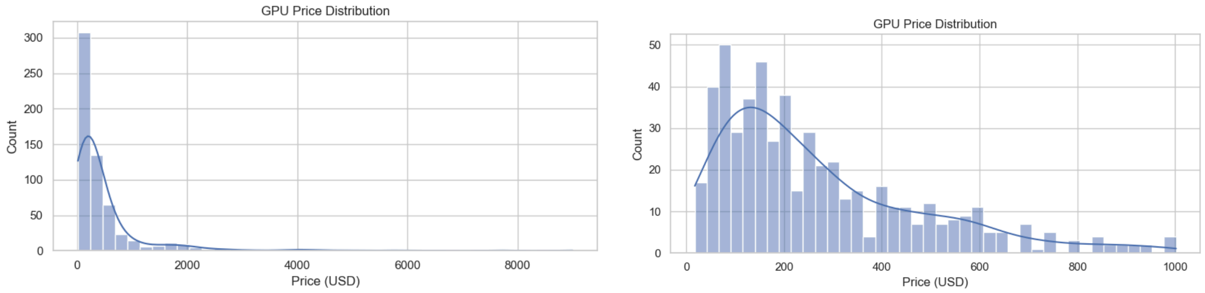

Figure 6: GPU price distributions before (left) and after (right) cleaning and standardization. Post-cleaning distributions reflect prices under $1,000 to target realistic consumer builds.

III. Modeling Approach

CatBoost regression models were trained independently for CPU and GPU datasets due to their differing price-performance dynamics. CatBoost was selected for its strong performance on tabular data, native handling of categorical variables, and robustness with minimal hyperparameter tuning. Unlike other gradient boosting algorithms, CatBoost processes categorical features natively using techniques like ordered boosting and target statistics, which help prevent overfitting and reduce the need for extensive preprocessing [3]. Model performance was assessed through visual diagnostics, including residual histograms, Q-Q plots, and actual-versus-predicted scatter plots, to ensure no systematic patterns of error across the feature space.

IV. Results

The trained CatBoost models yielded reliable performance on both CPU and GPU datasets, as reflected by their root mean square error (RMSE) values. The models effectively captured the relationship between benchmark performance metrics and price, incorporating features such as PassMark scores, rankings, and brand information.

The residual plots for CPUs and GPUs in Figure 7 revealed centering around zero and no evident heteroscedasticity as well as symmetric error distributions, indicating well-calibrated predictions. Although some outliers are present—particularly at the higher end of the predicted price range—they are not systematically skewed, implying that model performance is stable across the price spectrum.

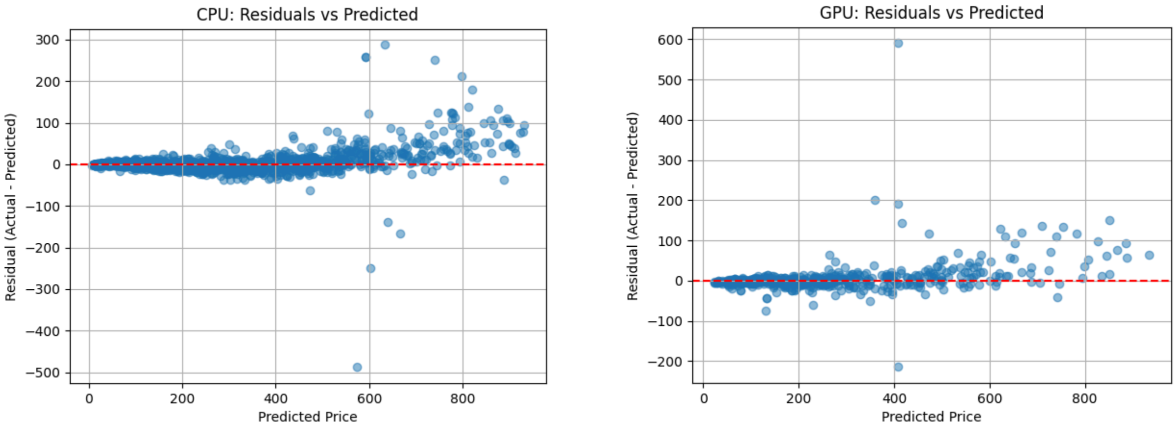

Figure 7: Residuals vs. Predicted Price for CPUs (left) and GPUs (right). Residuals are centered around zero with no clear pattern or heteroscedasticity, indicating well-calibrated models. A few high-end outliers are present but do not suggest systematic error.

From Figure 8, the actual versus predicted price plots show a strong alignment along the identity line, indicating minimal systematic bias in the model predictions. Quantitatively, the RMSE for CPU price estimation was approximately $44.10, while the GPU model achieved an RMSE of $39.61. These errors are relatively modest in the context of retail pricing and represent differences that, in practical terms, often fall within typical sales tax margins or regional pricing fluctuations. As such, these values demonstrate the models' ability to generalize effectively across a wide range of hardware within the sub-$1000 enthusiast segment. The majority of predictions fall close to the actual prices, reflecting low bias, good calibration, and reliable predictive accuracy.

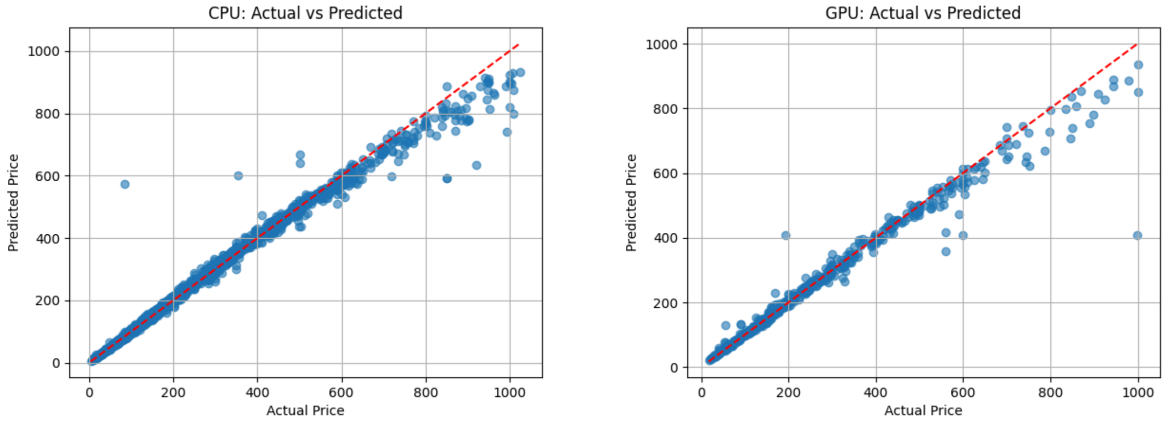

Figure 8: Actual vs. Predicted Prices for CPUs (left) and GPUs (right). Most data points closely align with the 45-degree line, indicating that the models are effectively capturing the underlying price-performance relationship. The strong linear agreement and low RMSE values (CPU: $41.32, GPU: $37.05) confirm accurate predictions and minimal systematic bias.

Lastly, the Q-Q plots in Figure 9 exhibit reasonably linear alignment with the theoretical normal distribution, particularly within the central quantiles. This indicates that the residuals approximate normality and that the error variance remains well-behaved across the dataset. Taken together, these diagnostic results support the conclusion that the CatBoost models generalize effectively and provide reliable price estimates within the target domain.

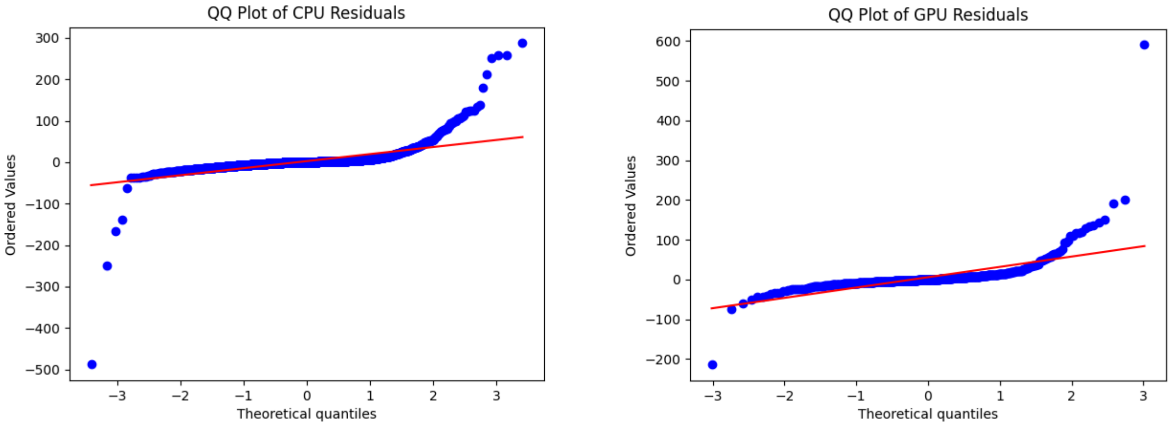

Figure 9: Q-Q plots for CPU (left) and GPU (right) residuals. Residuals demonstrate approximate normality, especially around the central quantiles. Deviations in the tails are expected but limited, indicating that model errors are well-behaved and the CatBoost models generalize effectively within the target domain.

V. Conclusion

The PC Value Estimator demonstrates the effectiveness of combining benchmark performance metrics with statistical learning to evaluate component pricing in a transparent, data-driven manner. By leveraging structured benchmark and pricing data, the system produces consistent and accurate price estimates, with RMSE values of approximately $44 for CPUs and $40 for GPUs. Users can interact with a dynamic tool that reflects current performance-to-price dynamics, moving beyond static review sites and anecdotal comparisons. Residual diagnostics confirm minimal bias and stable error variance, supporting the reliability of the CatBoost models. This work provides a scalable framework with practical applications, from aiding consumers in purchase decisions to enabling more intelligent market analysis. Future directions include incorporating time-based price tracking, extending support to additional hardware categories, and enabling full system valuation based on user-defined configurations.

Sources

[1] G. Wiederhold, "COVID-19 and the Acceleration of Digital Health," *Patterns*, vol. 1, no. 6, pp. 100103, Oct. 2020, doi: 10.1016/j.patter.2020.100103.

[2] M. Chowdhury, M. Khan, A. R. Uddin, and M. A. U. Mazumder, “COVID-19 pandemic-related supply chain studies: A systematic review,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 148, p. 102271, 2021. doi: 10.1016/j.tre.2021.102271

[3] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf