Predicting Laptop Prices

In today's rapidly evolving technology landscape, consumers are faced with a complex challenge, determining the fair market value of laptops based on their specifications and manufacturers. This paper delves into the significance of developing a predictive model that estimates laptop prices. The project's goal is to assist consumers in making informed decisions when purchasing laptops, ensuring they neither overpay nor compromise on quality. To achieve this, the study employs advanced machine learning techniques on a dataset obtained from Kaggle, exploring the relationships between laptop prices, specifications, and manufacturers.

Solving the problem of predicting laptop prices has far-reaching implications that resonate with both individual consumers and the broader market. For consumers, having access to accurate price predictions empowers them to make informed decisions, ensuring that they neither overspend nor compromise on the quality of their purchase. This, in turn, fosters consumer satisfaction and trust, vital factors for brand loyalty and positive word-of-mouth.

On a larger scale, stakeholders in the laptop industry stand to gain significantly from the development of such a predictive model. Manufacturers can better understand the price sensitivities of consumers, allowing them to optimize their pricing strategies and product offerings. Retailers can tailor their promotions and discounts more effectively, enhancing customer engagement and driving sales. Moreover, a reliable laptop price prediction model can contribute to a more transparent and competitive market, fostering healthy competition and innovation.

The dataset utilized for this project originates from Kaggle, a well-known platform for sharing and exploring datasets. It contains comprehensive information about various laptop models, including specifications, prices, and manufacturers. This dataset forms the foundation for all subsequent analyses, including exploratory data analysis (EDA), data preparation, and model building.

Following data selection, the project was structured into three milestones: EDA (Exploratory Data Analysis), Data Preparation, and Model Building. In Milestone One, the initial step involved importing the data from a CSV (Comma Separated Value) file into a data frame. Afterwards the data set was investigated for missing values, but none were found. Multiple visualizations were generated to enhance the comprehension of the data. The initial collection of graphs consisted of histograms depicting the distribution of all variables. This endeavor aimed to provide deeper insights into the data's distribution patterns. A significant number of variables, including the dependent variable, were not normally distributed, as can be seen in figure 1. This information was valuable when selecting model types in milestone 3.

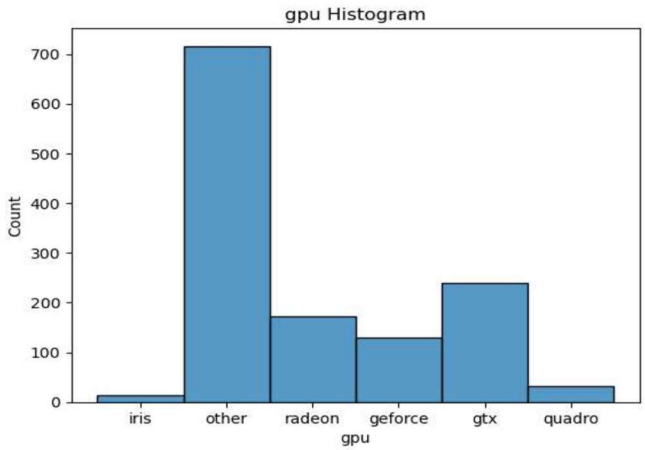

Figure 1: GPU Histogram

Afterwards, scatterplots were generated to visualize the relationships between variables and laptop prices. This proved beneficial in revealing correlations and price overlaps, even as specific hardware specifications were increased. This underscored the necessity for all variables in the model, as individual variables alone didn't determine the price. The correlation and price overlap are depicted in Figure 2.

The final visualization employed a bar graph to illustrate the average laptop prices across different manufacturers. This approach aimed to deepen the understanding of how brand names

influence laptop prices. While interpretation remains open, especially considering brands producing high-specification laptops, a distinct trend emerges for broader consumer-oriented companies offering similar equipment. The bar graph in Figure 3 exemplifies how certain brands incorporate a premium for their status or reputation, as evident in the price variances among Apple, HP, and Acer.

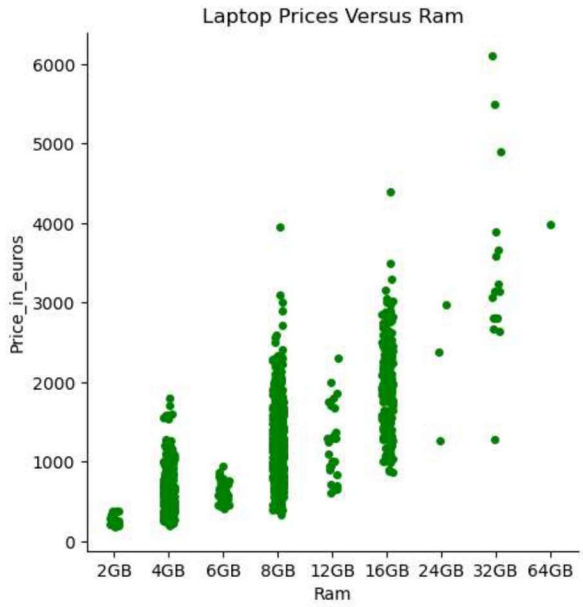

Figure 2: Scatterplot – Laptop Prices Versus Ram

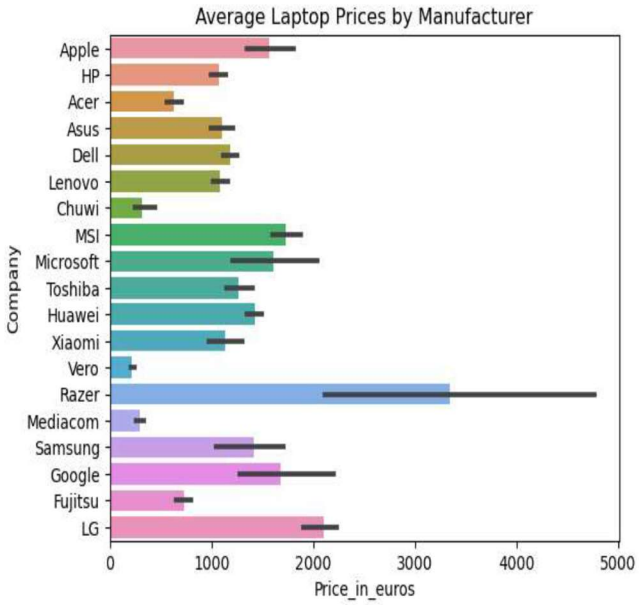

Figure 3: Bar Graph – Price versus manufacture

In milestone two, the initial focus was on preparing the data for effective model building and evaluation. The process began with the standardization of header information and text strings. This encompassed converting all characters to lowercase and removing any white spaces. Subsequently, extraneous columns like 'laptop_id' and 'product' were eliminated. Additionally, efforts were directed towards achieving uniformity in metric units and international currencies, which were converted to USD.

A substantial transformation involved the conversion of text strings into both categorical and numeric features. Specifically, the features 'ram', 'screenresolution', 'cpu', 'gpu', and 'memory' underwent tailored transformation approaches. While each feature demanded distinct treatment, several consistent methods were employed. The 'split' function was recurrently utilized to segment strings into meaningful text and numerical components. Following this, the 'extract' and 'replace'

functions played a pivotal role in modifying or eliminating irrelevant details from the strings. With the extraction of categorical labels and quantifiable numbers into separate features, data types were then adjusted, and steps were taken to reduce abundant categorical features.

At the end of this milestone, the data was configured to facilitate the subsequent model construction in the upcoming milestone. This entailed dividing the data frame into two distinct sets: one dedicated to the dependent variable and the other to independent variables. These sets were further partitioned into four subsets, encompassing two for training and two for testing purposes. Notably, the training data underwent one-hot encoding to convert categorical attributes into numeric representations.

The last milestone was to build a model that would predict laptop prices based on the data provided. The models chosen were all decision tree style models, due to the lack of normally distributed variables. XGBoost, CatBoost, and LightGBM were chosen to be tested and evaluated. These models are renowned for their exceptional performance in handling extensive datasets, capturing intricate nonlinear relationships, and offering inherent feature importance metrics. Their utilization of the gradient boosting framework, which combines weak learners to form a robust predictive model, was another contributing factor to their selection.

The experimentation involved a systematic grid search approach to optimize hyperparameters, focusing on n_estimators, max_depth, and learning_rate. Each model underwent separate grid searches to expedite the process and facilitate quick adjustments. Evaluation metrics were thoughtfully chosen to comprehensively assess the regression model's effectiveness: MAE for robustness against outliers, RMSE for balanced precision evaluation with outlier sensitivity, and R-squared to gauge the model's variance explanation. Both training and testing datasets were evaluated using these metrics, enabling a comprehensive understanding of potential overfitting across the different models.

The analysis and model building provided valuable insights into the relationship between laptop specifications and prices. The XGBoost model demonstrated the best predictive performance, yet its overfitting issue warrants further attention. While the model's R-squared value of 0.852 signifies a strong explanatory capability, the presence of overfitting suggests caution in immediate deployment.

To enhance the model's accuracy, the incorporation of additional features, such as laptop build materials, backlit keyboards, and build quality, is recommended. Collecting data over multiple time points would also enable tracking price fluctuations, providing a more accurate representation of market dynamics. Addressing these limitations would result in a more robust and deployable model.

The project has illuminated several challenges, including the absence of quality of life features and potential fluctuations in prices due to data sets web scraping methodology. Addressing these challenges is essential to improving the model's accuracy. Additionally, exploring ensemble methods and regularization techniques could further enhance the model's generalization capabilities. The project's success opens doors to broader applications, such as other consumer electronics.

In conclusion, the journey through exploratory data analysis, data preparation, and model building has shed light on the intricate relationship between laptop specifications, manufacturers, and prices. While the XGBoost model shows promise, further refinement and feature enrichment are required before deployment. This project underscores the importance of data-driven decision-making for both consumers and stakeholders in the ever-evolving technology market.