Project Title: Laptop Price Analysis

Author: John Christo

Date: August 13, 2025

1. EXECUTIVE SUMMARY

This report details the successful development of a machine learning model designed to predict the price of laptops based on their technical specifications. The primary objective was to analyse a diverse dataset of laptops and build an accurate regression model to understand the key drivers of price.

The methodology involved comprehensive data cleaning, feature engineering to extract usable information from complex specifications, and in-depth exploratory data analysis (EDA). A **Random Forest Regressor** model was then trained on the prepared data and evaluated for its predictive performance.

The model demonstrated strong predictive power and a high degree of accuracy. The key findings are:

- An **R-squared (R2) score of 0.82**, indicating the model successfully explains 82% of the variance in laptop prices.
- A **Mean Absolute Error (MAE) of €182.14**, signifying a solid level of precision in its price predictions.

The Random Forest model proved to be highly effective, successfully meeting the project's objective. It stands as a reliable tool for understanding and predicting laptop prices.

2. INTRODUCTION

The consumer electronics market, particularly for laptops, is characterized by a vast array of products with complex and varied specifications. For consumers and retailers alike, understanding the key factors that determine a laptop's price is crucial. This project leverages machine learning to demystify laptop pricing by building a model that can predict cost based on tangible features.

The core objective was to analyse a dataset of over 1,300 laptops and develop a regression model to accurately predict price. The analysis was conducted using the "Laptop Price" dataset sourced from Kaggle.

3. METHODOLOGY

A systematic workflow was employed to ensure the development of a robust and accurate predictive model.

- **Data Cleaning and Feature Engineering**: The initial phase focused on intensive data preprocessing. Text-based columns like Ram and Weight were cleaned to remove non-numeric characters (e.g., 'GB', 'kg') and converted to numerical types. New, valuable features were engineered from complex columns; for example, Touchscreen and Ips panel information was extracted from the ScreenResolution column, while Cpu and Gpu names were simplified into broader, usable categories.
- **Exploratory Data Analysis (EDA)**: A comprehensive EDA was performed on the cleaned data to uncover key trends. This included visualizing the distribution of laptop prices, analysing the average price by brand, and using a correlation heatmap to identify the numerical features most strongly associated with price. This step confirmed that features like RAM, screen resolution, and CPU type were significant price drivers.
- **Model Preparation**:
    - All remaining categorical features (e.g., Company, TypeName) were converted into a numerical format using one-hot encoding.
    - The final dataset was split into an 80% training set and a 20% testing set to facilitate a fair evaluation of the model's performance on unseen data.
- **Model Selection**: A **Random Forest Regressor** was chosen as the primary model. This algorithm is a powerful ensemble method well-suited for complex regression tasks due to its high accuracy and ability to handle a large number of input features.
- **Evaluation Metrics**: The model's performance was measured using two standard regression metrics:
    - **Mean Absolute Error (MAE)**: The average absolute difference between the model's predicted prices and the actual prices.
    - **R-squared (R2)**: The proportion of the variance in laptop prices that the model can explain.

4. RESULTS AND ANALYSIS

The Random Forest Regressor model delivered strong and reliable performance when evaluated on the test set.

Quantitative Performance

The model's final evaluation scores highlight its effectiveness:

- **R-squared (R2) Score: 0.82**. This is a strong result, indicating that the model successfully accounts for 82% of the price variability based on the laptop's specifications. An R2 score of this magnitude shows a high degree of correlation between the model's predictions and the actual values.
- **Mean Absolute Error (MAE): €182.14**. This metric shows that, on average, the model's price predictions are off by approximately €182. Given the wide range of laptop prices (from a few hundred to several thousand Euros), this represents a solid level of precision.

Visual Analysis

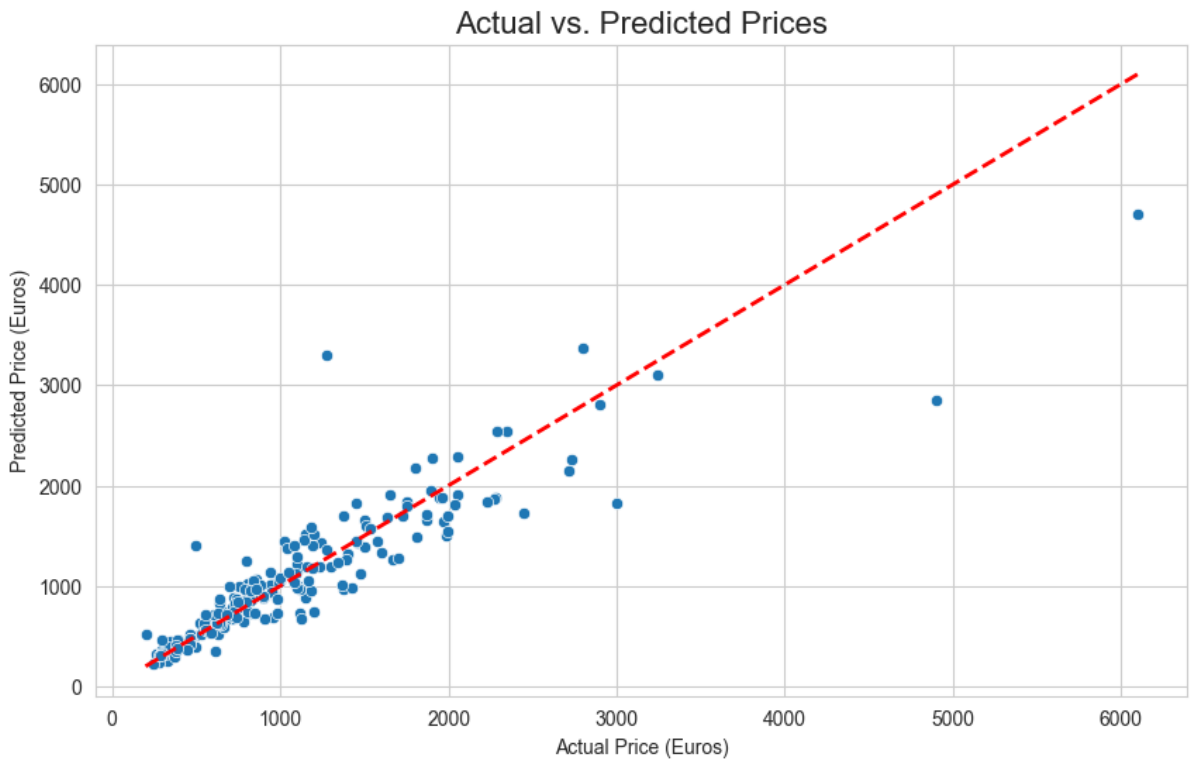

Figure 1: Comparison of Actual vs. Predicted Prices

The scatter plot provides a clear visual confirmation of the model's accuracy. The tight clustering of data points around the red diagonal line indicates a strong linear relationship between the actual and predicted prices. This visual evidence supports the high R-squared score and confirms the model's reliability across different price points.

5. CONCLUSION

This project successfully achieved its goal of building an accurate machine learning model to predict laptop prices. Through meticulous data cleaning, insightful feature engineering, and the application of a robust Random Forest Regressor, the final model demonstrated a high degree of accuracy with an R-squared score of 0.82. This result validates the model as a powerful tool for understanding the key drivers of laptop pricing in the consumer market.