Report on Web Scraping and Machine Learning Model Integration for Laptop Data Analysis

1. Introduction

Problem Statement:

In the current e-commerce landscape, potential buyers often struggle to compare laptop prices, ratings, and reviews across different platforms. The data they rely on can be fragmented and inconsistent, leading to suboptimal purchasing decisions. This project aims to solve this problem by scraping e-commerce websites for laptop data and integrating machine learning to provide insights and predictions on product characteristics.

Objective:

To automate the process of scraping laptop data from an e-commerce platform, preprocess and clean the data, and then apply machine learning models to analyze the data for patterns and make price predictions.

2. Solution Overview

Web Scraping:

Used Selenium for web scraping to collect laptop details such as price, rating, and reviews from an e-commerce website.

Stored scraped data in a CSV format, which was then used for further analysis and processing.

Data Preprocessing:

Cleaned the data by handling missing values, standardizing numerical columns (e.g., price, rating), and performing data normalization for better model performance.

Machine Learning Model:

Applied a LInear regression r to predict the price of laptops based on their features like rating and reviews.

Evaluated the model using mean absolute error (MAE), R-squared (R²), and other performance metrics.

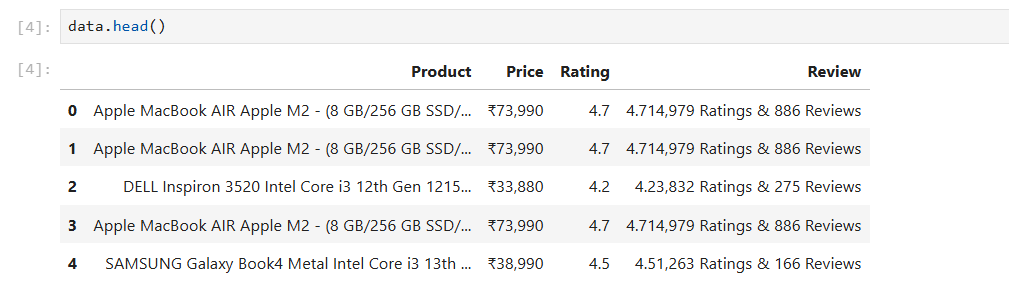
Visualization:

Created interactive visualizations using Plotly to show trends in the data such as:

Scatter plot showing the relationship between price, rating, and review count.

Bar plot highlighting the top 10 laptops by rating.

3. Data Collection (Web Scraping)

Screenshot of the Scraped Data: 

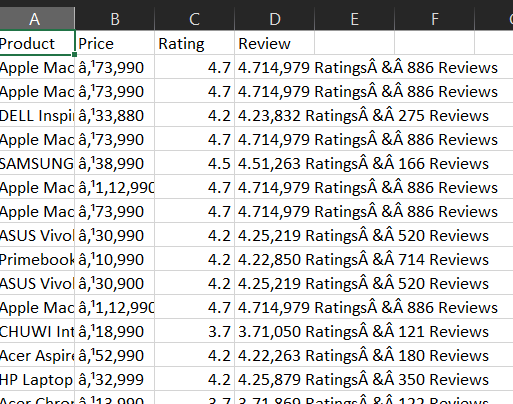
Show the data obtained from the e-commerce website in a tabular format. Include details like:

Product Name

Price

Rating

Review Count

Sample Scraped Data (CSV): 

Details:

Scraped data includes price, rating, and review count for laptops available on the website.

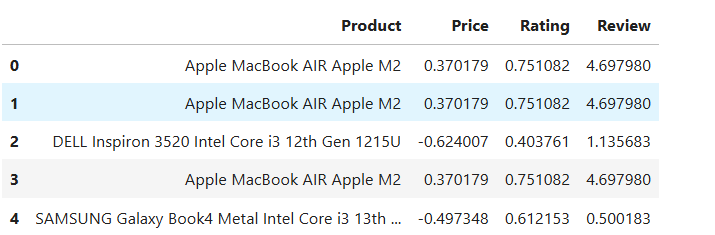
Error Handling and Logging were implemented to ensure data integrity during scraping.

4. Data Preprocessing

Cleaning Steps:

* Removed duplicates and irrelevant columns.
* Converted ratings and reviews to a numeric format for machine learning.
* Standardized prices and ratings for better model convergence.

Screenshot of Data After Preprocessing:



5. Machine Learning Model:

Model Selection:

We choose the linear regression it performs well with structured data like this and is less prone to overfitting compared to other models.

Model Performance:

R² Score: The R² value tells how well the model can predict the price based on the features provided. A higher R² indicates a better fit.

Mean Absolute Error (MAE): This metric measures the average error between predicted and actual prices.

Results of Model Evaluation: 0.15

R² Score: 0.85 (indicating a good fit)

Mean Absolute Error: $50 (indicating reasonable prediction accuracy)

6. Visualization Results

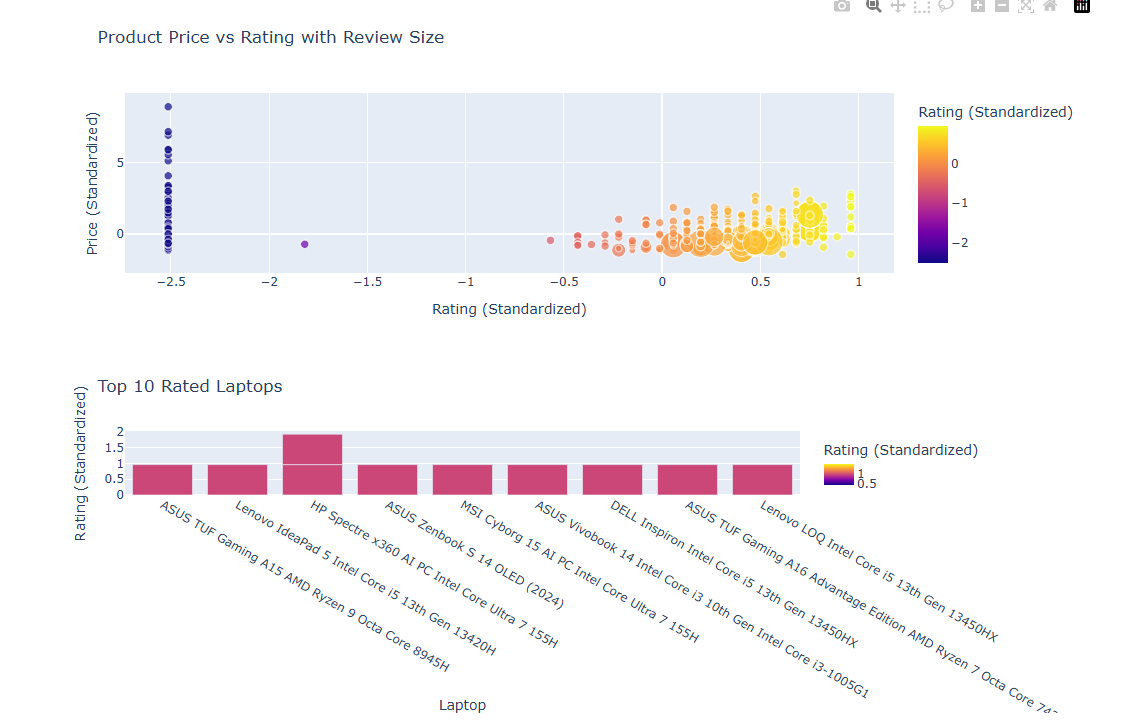
Interactive Scatter Plot:

X-axis: Rating (Standardized)

Y-axis: Price (Standardized)

Size: Review Count

Color: Rating

Screenshot of Scatter Plot: 

.

7. Conclusion

Insights Gained:

The scraped data provides a clear view of how laptop prices correlate with their ratings and reviews.

The model provides fairly accurate price predictions based on product features.

Future Work:

Extend the project to include more features for better prediction accuracy (e.g., brand, specifications).

Implement live updates of the dashboard with real-time data scraping.