RESEARCH ARTICLE OPEN ACCESS

Web Scraping for Laptop Data from Amazon Using Machine Learning

Sandhyarani Kamble*, Snehalata Jadhav *, Dr. Urmila Pol**, Dr. Tejashree T. Moharekar***, Dr. Parashuram S. Vadar***

*(PG Student, Computer Science Department, Shivaji University, Kolhapur, India)

** (Associate Professor, Computer Science Department, Shivaji University, Kolhapur, India Email: urp_csd@unishivaji.ac.in)

***(Assistant Professor, Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur, India
Email : ttm.50649@unishivaji.ac.in)

Abstract:

The goal of this project is to use web scraping techniques to extract laptop data from Amazon and then use a variety of machine learning algorithms to analyze the data. Product details, costs, ratings, and reviews are among the information gathered. Machine learning models are used to analyze market trends, categorize laptops according to user preferences, and forecast pricing trends. In this project, we utilize various machine learning algorithms to analyze the scraped data. Logistic Regression is applied to classify products based on customer sentiment, while Random Forest is used for both price prediction and classification tasks, ensuring high accuracy in identifying patterns. K-Means Clustering helps segment laptops into different categories based on features such as price and specifications, enabling market analysis. Linear Regression is employed for predicting continuous outcomes like price estimation, and Support Vector Machine (SVM) is used for classifying laptops into distinct categories based on user feedback and other product attributes. These algorithms together enable comprehensive analysis of the data.

Keywords — **Web Scraping, Machine Learning, Amazon laptop data, Price Prediction, Logistic Regression, Random Forest, K-Means, Linear Regression, Support Vector Machine (SVM).**

I. INTRODUCTION

In this research project, we will investigate web scraping methods to gather laptop information from Amazon for machine learning analysis. By gathering information like product details, prices, and customer reviews, we can derive useful insights into consumer behavior and market trends. The data will be analyzed and processed using machine learning algorithms to determine patterns and forecast future trends. This project will also discuss the issue of web scraping, such as data accuracy and ethical issues. Finally, our aim is to create a system that can aid businesses and customers in making an informed decision. Our research will be part of the emerging concept of data-driven decision-making for e-commerce.

II. LITERATURE REVIEW

This paper aims to leveraging the powerful capabilities of selenium webdrivers, we were able to automate the extraction of valuable data from various online shopping websites. In the web scraping, using selenium project, evaluation and validation are essential for access the quality of the scraped data including completeness, consistency and relevance [1]. Researchers have been using web scraping to collect data from websites for analysis, but it can be tricky. This paper reviews existing methods and finds that BeautifulSoup, a Python library, is a popular and powerful tool for web scraping. It's fast, easy to use, and can handle complex websites. The paper highlights how BeautifulSoup has been used in various studies to collect data on things like product prices, social media posts, and job listings. However, it also notes some challenges like handling anti-scraping measures and dealing with large amounts of data. Overall, the paper concludes that BeautifulSoup is a great tool for web scraping and can help researchers and businesses extract valuable insights from online data [2]. The outcome of the "Web scraping for e-commerce website" project is a powerful and efficient utility that enables companies with actionable e-commerce insights. By extracting product information systematically, price information, and consumer feedback, the project allows companies to make smart choices regarding pricing strategies, inventory management, and marketing efforts [3]. This research paper is the process of particularly important in field such as business intelligence in the modern age this paper looks at the what web scraping technologies, stages, AI, data science, big data, cyber security. This paper data extract from sites using the HTTP protocol used by web browser. The web scraping process is fetching stage, extracting stage, transformation. In web scraping code reuse and maintenance are especially critical. Python is for easy implementation [5]. This research paper shows how machine learning can be used to analyze the sentiment of online reviews and ratings scraped from websites. The authors used web scraping to collect data from ecommerce sites and then applied machine learning algorithms to determine the sentiment of the reviews. The results show that this method can accurately identify positive, negative, and neutral sentiments. This can help businesses understand their customers' opinions and make informed decisions. The paper concludes that machine learning-based sentiment analysis is a powerful tool for web scraping and can benefit various applications such as product recommendation, reputation management, and market research [7]. In this research paper they describe, Web scraping’s are sets of techniques created to spontaneously retrieve data via a web rather than copy/duplicating data manually. The purpose of web scraping tools are to find a particular kind of data, get it, as well as combine the data to a newer page (webpage). Particularly scraping tools focus on converting unstructured set of information followed up by storing it in clean databases. This article focuses specifically on techniques for extracting content from web pages. Especially in the field of web advertising, we use scraping technology [9] This project, named as Price comparison website using web scrapping is the place where shoppers could find the great deals on the products. The best deals will be clearly highlighted. To obtain best deals from Price comparison websites web scrapping techniques are used to fetch detailed information. This way, paper aims to provide solution for online customers to buy products at good deal and save their valuable time, effort, and money [11]. AI-based methods and tools used for adjust themselves to scraping the data. Web scraping use machine learning and AI technologies. Data are not in structured formats and Difficulty of extracting relevant data from web pages. It highlights Scrapy as a powerful web scraping tool, offering speed, extensibility, and efficient data extraction capabilities. Develop a more efficient and accurate way to extract and classify web content using AI and machine learning algorithms techniques [12]. Social media is a significant catalyst for acquiring and disseminating information in several domains such as entertainment, commerce, science, politics, and crisis management. It enables users to publish and share a diverse range of media formats, including text, videos, pictures, and audio. By conducting data analysis on social media, a person can access a wide range of information, including trends, concerns, and key individuals. Sentiment Analysis (SA) aims to ascertain the emotional response of individuals towards a specific service, business, or product[13].

III. RESEARCH METHODOLOGY

1. Web Scraping: The first step involves scraping laptop data from Amazon using web scraping techniques. Tools like BeautifulSoup or Scrapy are used to extract product details such as price, specifications, ratings, and reviews from Amazon's product pages. This data is stored in a structured format, such as a CSV or database, for further analysis.

2. Data Preprocessing: After the data is scraped, it undergoes preprocessing to clean and transform it into a usable format. This includes handling missing values, removing duplicates, and standardizing textual data (such as reviews). Numerical data, like price and ratings, is normalized to ensure consistency across different data points.

3. Feature Engineering: Relevant features are selected and extracted from the raw data. Features such as product category, brand, price, user reviews, and ratings are used to create meaningful input variables for machine learning models. Textual data (reviews) is transformed into numerical form using techniques like TF-IDF or sentiment analysis.

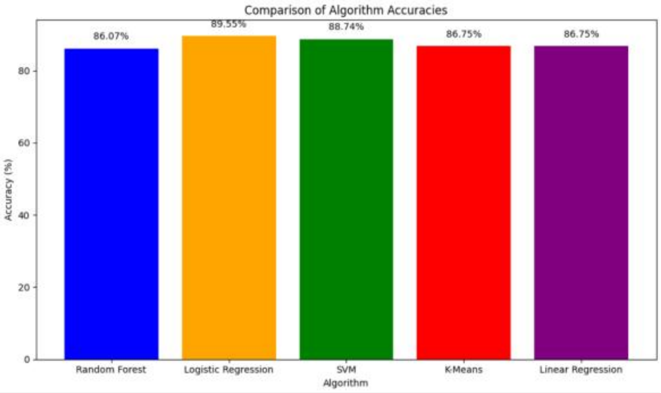
4. Model Selection and Training: Various machine learning algorithms are used to model and analyze the data: o Logistic Regression is applied to classify products based on binary sentiment outcomes (positive or negative feedback). o Random Forest is used for regression and classification tasks to predict laptop prices and classify products based on features. o K-Means Clustering groups similar laptops together based on product features to segment the market. o Linear Regression is used for price prediction by modeling the relationship between numerical features and the laptop price. o Support Vector Machine (SVM) is used to classify laptops into categories (e.g., budget vs. high-end) based on customer reviews and product attributes.

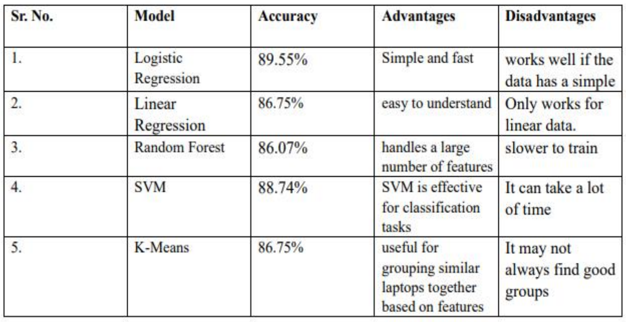
5. Model Evaluation: The performance of each model is evaluated using various metrics, such as accuracy, precision, recall, F1-score for classification tasks, and Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) for regression tasks. Cross-validation techniques are employed to ensure the robustness of the models.

6. Analysis and Insights: Once the models are trained, their outputs are analyzed to gain insights into laptop pricing trends, consumer preferences, and product segmentation. The results can be used for predictions about future trends, helping businesses and consumers make informed decisions.

7. Deployment and Visualization: Finally, the models can be deployed into a dashboard or web application for real-time predictions and data visualization. Interactive graphs and reports can display insights such as price trends, market segmentation, and product classifications.

IV. COMPARING MODEL PERFORMANCE





Sr. No.	Model	Accuracy	Advantages	Disadvantages
1.	Logistic Regression	89.55%	Simple and fast	works well if the data has a simple
2.	Linear Regression	86.75%	easy to understand	Only works for linear data.
3.	Random Forest	86.07%	handles a large number of features	slower to train
4.	SVM	88.74%	SVM is effective for classification tasks	It can take a lot of time
5.	K-Means	86.75%	useful for grouping similar laptops together based on features	It may not always find good groups

V. CONCLUSIONS

In this study, we were able to implement a number of machine learning algorithms such as Random Forest Regression, SVM, Linear Regression, Logistic Regression, and K-Means clustering to classify laptop data web-scraped from Amazon. The data comprising features such as processor type, RAM, storage, and customer reviews was cleaned and preprocessed for model training. Logistic Regression yielded the best predictions of laptop prices, while SVM and Random Forest Regression could classify laptops efficiently based on specifications. K-Means clustering assisted in segmenting similar laptops for better analysis of product types. The outcomes show the strength of the use of web scraping and machine learning for e-commerce data analysis. This method can be applied to real-time price, product suggestions, and customer analysis. Future research can involve improving model performance by adding more sophisticated features and external data.

REFERENCES

[1] Lakkakula. Sai Lakshmi, D. Lakshmi Sumithra, K. Jhansi, J. Vandana(2024) "Web Scraping And Data Analysis For Online Shopping With Selenium."

[2] Abodayeh, A., Hejazi, R., Najjar, W., Shihadeh, L., & Latif, R. (2023, March). "Web Scraping for Data Analytics: A BeautifulSoup Implementation." In 2023 Sixth International Conference of Women in Data Science at Prince Sultan University

[3] Sri G. Shridhar, Shadula Sathwika, Kethiri Sreehitha Reddy(2024)."Web Scraping for ECommerce website."

[4] 4Woodall, R., Kline, D., Modaresnezhad, M., & Vetter, R. (2021). "A cloud-based system for scraping data from amazon product reviews at scale." In Proceedings of the Conference on Information Systems Applied Research, Washington DC, USA.

[5] Khder, M. A. (2021). "Web scraping or web crawling: State of art, techniques, approaches and application." International Journal of Advances in Soft Computing & Its Applications.

[6] Srividhya, V., & Megala, P. (2019). "Scraping and Visualization of Product Data from E-commerce Websites." Int. J. Comput. Sci. Eng.

[7] Sahu, S., Divya, K., Rastogi, N., Yadav, P. K., &Perwej, Y. (2022). "Sentimental Analysis on Web Scraping Using Machine Learning Method." Journal of Information and Computational Science.

[8] Zhao, B. (2022). "Web scraping." In Encyclopedia of big data (pp. 951-953). Cham: Springer International Publishing.

[9] Niranjan Krishna, Anvith Nayak, Sana Badagan, Chethan Jetty, Dr. Sandhya N(2022)"A study on Web Scraping."ISSN (Online) 2394-2320

[10] SCM de S Sirisuriya. (2023)"Importance of Web Scraping as a Data Source for Machine Learning Algorithms – Review."Faculty of Computing, General Sir John Kotelawala Defence University Sri Lanka.

[11] Arman Shaikh, Raihan Khan, Komal Panokher Mritunjay, Kr Ranjan Vaibhav Sonaje.(2023)"E-commerce Price Comparison Website UsingWeb Scraping"

[12] Prof. K. N. Aaglave, Shivanjali Santosh Jadhav, Amaan Firoj Khatib, Rohini Laxman Khurangale(2023). "A Survey on the Web Scraping : In the Search of Data"

[13] Galiveeti Poornima, Meenakshi, Malik Jawarneh, A Shobana, KP Yuvaraj, Urmila R Pol, Tejashree Tejpal Moharekar (2025), "Machine Learning for Sentiment Analysis Using Social Media Scrapped Data" Natural Language Processing for Software Engineering, 143-154, John Wiley & Sons, Inc

[14] Aswad shaikh, Aniket sonmali, sohamwakade (2023)." product comparison website using web scraping and machine learning"

[15] Gandhe Vineeth Kumar, Hema M S, Aishwarya R, K R Mamatha (2013). "Web Scraping for E-Commerce Websites"

[16] Kasereka Henrys(2021). "Importance of web scraping in e-commerce and emarketing"

[17] Aliya thaseen, A.Vijay Kumar , Dr.Anusha Ampavathi , Dr A. Obulesh, Dr Mohd Nazeer, Dr Abdul Ahad. "CUSTOMER-TARGETED E-COMMERCE WEBSITE USING WEB SCRAPING"

[18] Priya Matta, Nikita Sharma, Devayani Sharma, Bhasker Pant, SachinSharma(2020). "webscraping: applications and scraping tools.".