Project Report

on

Product price prediction by machine learning method through web scrapping

Submitted in partial fulfillment of the requirement for the award of the degree of

B.Tech in Computer Science and Engineering

Under The Supervision of

Name of Supervisor: Mr. Rajeev Kumar.

Designation: Assistant Professor

Submitted By

Name of Students: Shalini Sinha, 19SCSE1010390

Tanishq Pundir, 19SCSE1010519

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

INDIA

MAY, 2023

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled "**Product price prediction by machine learning method through web scrapping** " in partial fulfillment of the requirements for the award of B.Tech submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of 1 February 2023 to 15 May 2023, under the supervision of **Mr Rajeev Kumar** Assistant Professor Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Shalini Sinha

Tanishq Pundir

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Rajeev Kumar

Assistant Professor

CERTIFICATE

The Project Viva-Voce examination of **Shalini Sinha, 19SCSE1010390, Tanishq Pundir**, **19SCSE1010519,** entitled "**Product price prediction by machine learning method through web scrapping** " has been held on 15 May, 2023 and their work is recommended for the award of B.Tech.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Program Chair **Signature of Dean**

Date: 15 May, 2023

Place: Greater Noida

ABSTRACT

In this emerging world of the internet, there is lots of data present and retrieving this data becomes very complicated. As a result, web scraping is one of the important method of data gathering. Web scraping is a technique of extracting data from various websites and depending on the tool end-users can access the data in several formats such as spreadsheet, csv, json, xml and database. Web scraping is used in many fields like e-commerce, market research, brand monitoring and etc. Our system proposes a method of fetching product data from e-commerce websites and comparing them. For extracting data different tools are used such as Scrapy, BeautifulSoup, Selenium, etc. Our system uses Selenium for extracting data. After extraction data is stored into MySQL database. This data is then displayed in a comparable format on our webapp. Visiting websites one by one and comparing product details is time consuming, so to overcome this our system will display all the product details from various websites, which will help the end-user to compare the products. Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Prediction model is an information output generated by an ML algorithm trained on historical input data. A machine learning prediction is simply a model's output when provided with an input. Reliable ML predictions offer valuable insights leading to more confident and guided decisions by businesses. e.g., Business sales forecast for the next quarter, Likelihood of customer churn for a specific brand, etc.

Keywords - Web scraping, E-commerce, Data extraction, Web crawler, Python

Table of Contents

CHAPTER-1

Introduction

INTRODUCTION

Websites are an endless source of information that are available to everyone. The most recent advancement in technology compelled us to change the way we do business. The new venue for doing business is online. Understanding how to use the Internet and the different opportunities it may give is one of the keys to success in e-marketing and e-commerce.

Data crawling or web scraping or data harvesting has been into the existence for as long as the web itself. It is always associated with web content extraction, at the begin it wasn't always served this purpose. Web scraping can be considered as a method of retrieving or extraction content from a website for Web scraping is the method of retrieving or extracting content from a website for the purpose of using it for purposes beyond the control of the website owner.

Web scraping was originally used to construct connections with test frameworks. Companies such as IP-Label have built technologies that allow web developers and webmasters to monitor website performance on a regular basis using tools such as Selenium (xbyte, 2021). Previously, extracting online data required physically transferring the text accessible on a web page to a local file; this method was exceedingly inefficient and could not be utilised for commercial applications. Spreadsheet tools such as Microsoft Excel and Google Sheets offer some rudimentary web scraping features and were mostly used to retrieve HTML tables from webpages.

Every company aspires to exceed itself in a world of fierce competition. The question that the majority of other business owners are surely concerned about. Competitor research is not an option but a need in an industry where everything revolves around the consumer. Access to so much data can provide you a competitive advantage in your sector.

Web scraping is generally regarded as an effective and strong strategy for gathering big data since a large volume of heterogeneous data is continually created on the web. Most data scientists are familiar with the notion of web scraping, which is growing more popular as a result of the massive amount of data available on the internet and new companies who do not want to spend time gathering data that can be accessed fast on the internet.

OVERVIEW OF CONCEPTS

What is E-commerce?

Ecommerce, often known as electronic commerce or online commerce, is the purchasing and selling of products or services through the internet, as well as the transmission of money and data to complete these transactions. E-commerce, according to Wikipedia The action of purchasing or selling things electronically using online services or the Internet is known as electronic commerce.

Type of e-commerce:

- Business to business
- Business to consumer
- Consumer to consumer
- Consumer to business

What is E-marketing?

E-marketing (or electronic marketing) is marketing done over the Internet. Internet Marketing and Online Marketing are two synonyms for E-Marketing that are commonly used interchangeably. Electronic marketing is the method of promoting a brand (company, product, or service) over the Internet via computers and mobile devices.

What is Web Scraping?

Web scraping is a means of automatically obtaining vast volumes of data from websites. The majority of this data is unstructured HTML data that is turned into structured data in a spreadsheet or database before being used in various applications. To gather data from websites, web scraping may be done in a variety of methods. These include employing internet services, specific APIs, or even writing your own web scraping programmes from scratch. Many huge websites, such as Google, Twitter, Facebook, StackOverflow, and others, provide APIs that allow you to access structured data. This is the greatest alternative, however there are other sites that do not allow users to access big volumes of data in an organised format or that need users to register or they are simply not that technologically advanced. In that situation, it's best to use Web Scraping to scrape the website for data.

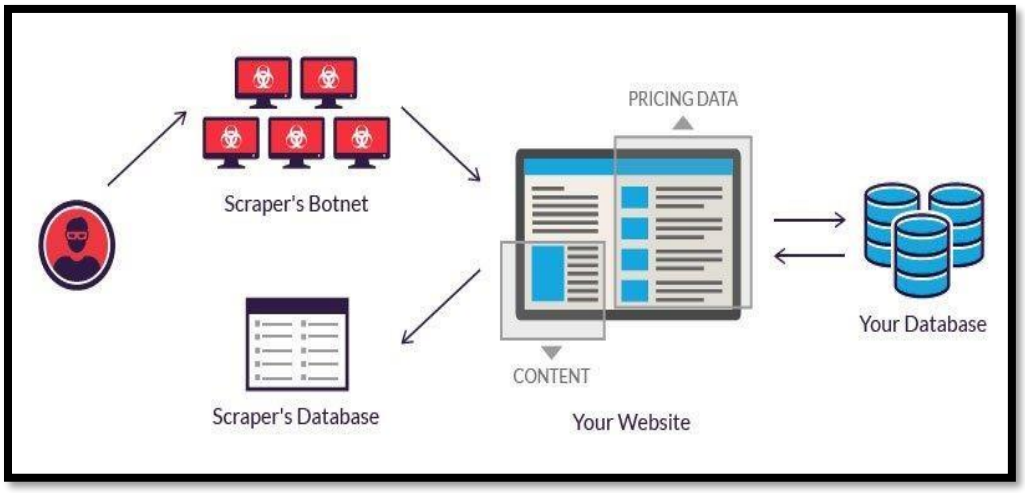


Fig 1.1 Web Scrapping

Benefits of Web Scrapping

Web scraping has emerged as an important strategy for e-commerce businesses, especially in delivering rich data-based insights.

- Price-monitoring and Product Research
- Online price comparison
- Better Customer analysis
- Market Analysis
- Better advertisements
- Influences Marketing and Sales Strategy
- Brand monitoring
- Extract business details from business directory
- Helps in Future analysis

Challenges of web scraping

- It is not always easy to know which site to scrape. Sometimes choose between scraping the data owner or an aggregator site.
- The internet is dynamic. Each web site has a particular structure, which may be subject to changes anytime.
- Data is volatile.
- Legal issues and informing web site owners.

How Web Scrapping Work?

In general, to extract data from websites, Web Scraping methods follow three steps as mentioned below -

Make an HTTP Request to The Website URLs - In the first step, you need to provide a list of URLs to the Web Scrapers from which you want to extract the data and it will make HTTP requests to these URLs.

Load and Parse the Websites Code - Once the HTTP request is successfully executed to the URLs mentioned in the first step, Web Scrapers can load the HTML or XML code of the websites. Some Web Scrapers can also load CSS or JavaScript elements as well. Further, Web Scraper will parse the HTML code to identify and extract relevant data which is predefined by you.

Save Relevant Data Locally - Once all relevant data is extracted, Web Scrapers can store the data in a structured format, usually as an Excel Spreadsheet or CSV file. These scrapers can also save the data in other formats, such as JSON, etc.

Different Types of Web Scrapers

Web Scrapers can be categorized based on various factors, including Self-built or Pre-built Web Scraper, Browser Extension or Software Web Scrapers, and Local or Cloud Web Scrapers.

Self-Built Web Scrapers require advanced programming knowledge to build and develop. If you want more features, it will require a more advanced understanding of programming languages. Pre-Built Web Scrapers can be downloaded and used directly. Some Scrapers also provide options to customize their features based on your requirements.

Browser Extension Web Scrapers can be added to your browsers. These Scrapers generally have a limited set of features as these are dependent on compatibility with your browser. Software Web Scrapers does not suffer from these limitations as these can be downloaded and installed directly on your computer.

Cloud Web Scrapers run on the cloud and will not use your local computer resources such as RAM, CPU, etc. Local Web Scrapers run on your computer and will consume its resources to perform Web Scraping.

How to Scrape The Web (Step-by-Step)

Identify the URLs

In this first step, identify all URLs you want to scrape and extract relevant data for further analysis. For example, you could scrape Amazon's Website to collect product information and customer reviews.

Inspect the Webpage

Before performing Web Scraping, you need to inspect the Webpage's HTML content. You can do it by right-clicking on the Webpage and selecting Inspect or View Page Source.

Identify the Data You Want to Scrape

In this step, you need to analyze the Webpage's HTML code to identify the unique HTML tags which contain the relevant data. For example, if you want to scrape the reviews from Amazon, you need to find tags that contain customer reviews, as shown in the below figure.

identify-data-to-scrape

Write Code for Web Scraping

Once you have inspected the HTML content of the Webpage and identified all appropriate HTML tags, you can use Python libraries to write your own Web Scraping function. You need to specify what information you want to scrape and parse from the Website.

Code Execution

In this step, you execute your written code. Then, it will make an HTTP request to the URLs, scrape the data and parse it as mentioned in the previous steps.

Store Final Data

Once everything required is extracted, the next step is to store this data in a structured format for further analysis. It could be stored locally or in a database in any format, such as CSV, Excel, JSON, etc.

CHAPTER-2

Literature Survey

LITERATURE SURVEY

A variety of the concerning articles were be discussed, in order to identify some takeaways from traditional prediction model and how different web scrapping model work. There are developments in artificial intelligence technologies owing to the hardware strengthening of computers. The amount of data required for artificial intelligence is increasing day by day. The Internet is the largest resource used to access this data. It is necessary to obtain this data quickly, structurally and systematically. However, it is a very costly process to handle these works with manpower. For this reason, web scraping techniques are used to overcome these problems [1]. There are some techniques for retrieving data from online source such as online retailers in prediction system. Machine learning has gained popularity in many application fields because it can process large data sets with many variables [2]. The applications of machine learning range from creating better recommendation systems on Netflix to facial recognition in pictures to cancer prediction and prognosis.

Information retrieval, news collecting, website monitoring, competitive marketing, and other topics are covered by web scraping. Web scraping makes it quick and straightforward to access the large quantity of information available online. Compared to manually pulling data from websites, it is far faster and easier [3]. These days, web scraping is more and more common.

A lot of data collection and information extraction may be done quickly and easily using an online data scraping software. Compared to manually pulling data from websites, it is far faster and easier. These days, web scraping is more and more common [4]. A lot of data collection and information extraction may be done quickly and easily using an online data scraping software. However, when individuals use the term "web scrapers," they often refer to computer programs. extract meaningful data.

These bots can quickly retrieve enormous volumes of data by automating this process. In the digital era, when big data plays such a significant role and is continuously updating and changing, this has obvious advantages [5]. Web scraping has several uses, particularly in the area of data analytics. Web crawling and web scraping technique are commonly used for retrieving data from website

A software agent that simulates the human experience of Web surfing interaction is used in the web scraping process, which includes systematic content extraction and combining from the Internet [6]. It might also be viewed as an extraction procedure that converts Internet material that is unstructured into a structured format that is simple to read and utilize for various studies.

The web scraping is basically obtaining information autonomously using software from the internet [7]. The libraries are used for web scraping such as Storm Crawler, Jauntium, Jaunt, Scrapy, Norconex, Apify, Colly, Selenium, Beautiful Soup and Grablab. Selenium and Beautiful Soup libraries were used in this work.

Although the Selenium library is slower, it is very useful on websites have a static URL. On the other hand, although the Beautiful Soup library is quite fast, it is inadequate in websites with embedded page structures.

Selenium is a library that can perform automatic web browsing through the web driver. The HTML and CSS codes of the page are accessed as a result of the mechanical processes performed. Web scraping takes place as a result of manipulating them [8]. Then, the data obtained is cleaned and made structural. The Selenium library is also used for testing operations. All elements on the page can be manipulated with the functions to be written in these languages.

Another library can be used in web scrapping is the Beautiful Soup. This library is generally used to analyze HTML and XML structures. Unlike the Selenium library, it does not need any web driver structure. It reads the source codes of the target site and parses those source

codes [9]. Thus, it provides data to be obtained very quickly compared to the Selenium library.

The web scraper uses a request to get access to the website, and then uses the HTML code to discover specified components, extract them, and store them into a structured manner using a data frame. This method of removing unstructured material from websites can be used in a variety of contexts. It may be used, for instance, to compare product pricing across several websites, to remove advertisements from connected pages, to address issues with inadequate statistics data, or to supplement official datasets with additional data [10]. Web scraping may also be utilized in the human resources field to locate open positions on various websites and categorize those using Nave Bayes algorithms.

A major task in almost all of discussed literature is the extraction of features from textual data. Textual data is known to be unstructured and hard to interpret without having an idea about the context. This particular challenge has evolved itself into the field of Natural Language Processing (NLP) [11]. The research field has developed multiple methods to turn the unstructured text data into a more structured format that is easier to digest by statistical models.

To capture the relationships between scraped features and the prices as accurately as possible, it will be of valid importance which type of prediction algorithm is used. Prediction algorithms use different mathematical optimization techniques to approximate the effect of its features on the outcome variable. Where some algorithms are known to be good in capturing linear effects, others are found to be better in capturing non-linear effects. As data shapes and relationships are different per research subject, there is no one-model-fits-all solution [11]. To find out which algorithm fits best to the prediction problem in this study, it can already help to review earlier studies with comparable prediction problems.

Logistic regression is useful tool to predict dependent variable that contain binary outcome. Literature has looked at logistic regression technique to develop customer satisfaction model basing from the fact that both dependent and independent variables are categorical.

Response for dependent variable was Yes or No while response for all independent variables was categorical.

In study, different prediction models such as Support Vector Machine (SVM), Random Forest and Linear Regression, XGBoost, Decision Tree were used, and 83% score was obtained with ADA Boost. proposed a model using the yelp dataset to predict the success and rating of the new model. They performed the Chi-square test and stochastic gradient descent to classify the model features as having the greatest weight. They used classification algorithms such as SVM, Random Forest, Logistic Regression and Multilayer Neural Networks. Random Forest achieved 56% and Multilayer Neural Networks obtained 60% accuracy [12]. While using clustering algorithms, accuracy increased to 85%. proposed a viable technique on the prediction model on available data from Kaggle, where 75 features were extracted for supervised machine learning.

A better result was obtained with the linear regression value of 53.13% to solve the sales forecasting problem in small data. The study examined a wide variety of ML techniques based on real datasets. LSTM and GRU neural networks were developed to assist with the gradient problem. A 10-year analysis period was chosen [13]. Logistic Regression gave better results in studies with multiple discriminant analysis.

A daily sales forecast was required for products with a short shelf life, and a weekly sales forecast was required for products with a long shelf life. Lagged variables as input variables were the main mechanism by which propositional learning algorithms capture the relationship between the past and present values of a series [14]. It was emphasized that sequential lagging variables could be averaged over a single field to reduce the number of input variables, as a large number of input variables might have a negative effect on some learning algorithms.

One of the methodologies that scores best is VADER, which seems very well able to classify sentences into the three sentiment buckets positive, negative and neutral. Besides, this rule-based algorithm is able to label more than 80% of the sentences served to the

algorithm, which was more than most of the investigated algorithms. This is a useful property, especially when the textual data is not abundantly available through time. The VADER algorithm is readily available as a Python module [14]. This made VADER the most promising sentiment scoring algorithm to use.

To investigate the possibility of non-linear dependencies between sentiments from text document and financial market movements, several studies investigate the use of non-linear models to develop their prediction models. hen a Self-Organizing Fuzzy Neural Network model was used, a significant directional prediction accuracy 86.7%. Additionally, when the same algorithm was used over a different time period, similar prediction results (75.56%) where obtained. These findings support the assumption that sentiment scores maintain a non-linear relationship.

In all the mentioned literatures, no detail implementation has been reported on the web scraping technique. In [15] the researchers present the taxonomy of web extraction tools but rather focused on the theoretical than demonstrating the implementation codes.

Thus, in this research we will propose an online price prediction system. For evaluating the usefulness and acceptance of the proposed system, we will measure the acceptance by user's perspective using Technology Acceptance Model (TAM) after implementing prototype system. We will also find the factors that have positive influence on acceptance of the system from user.

Chapter-3

Project Diagram

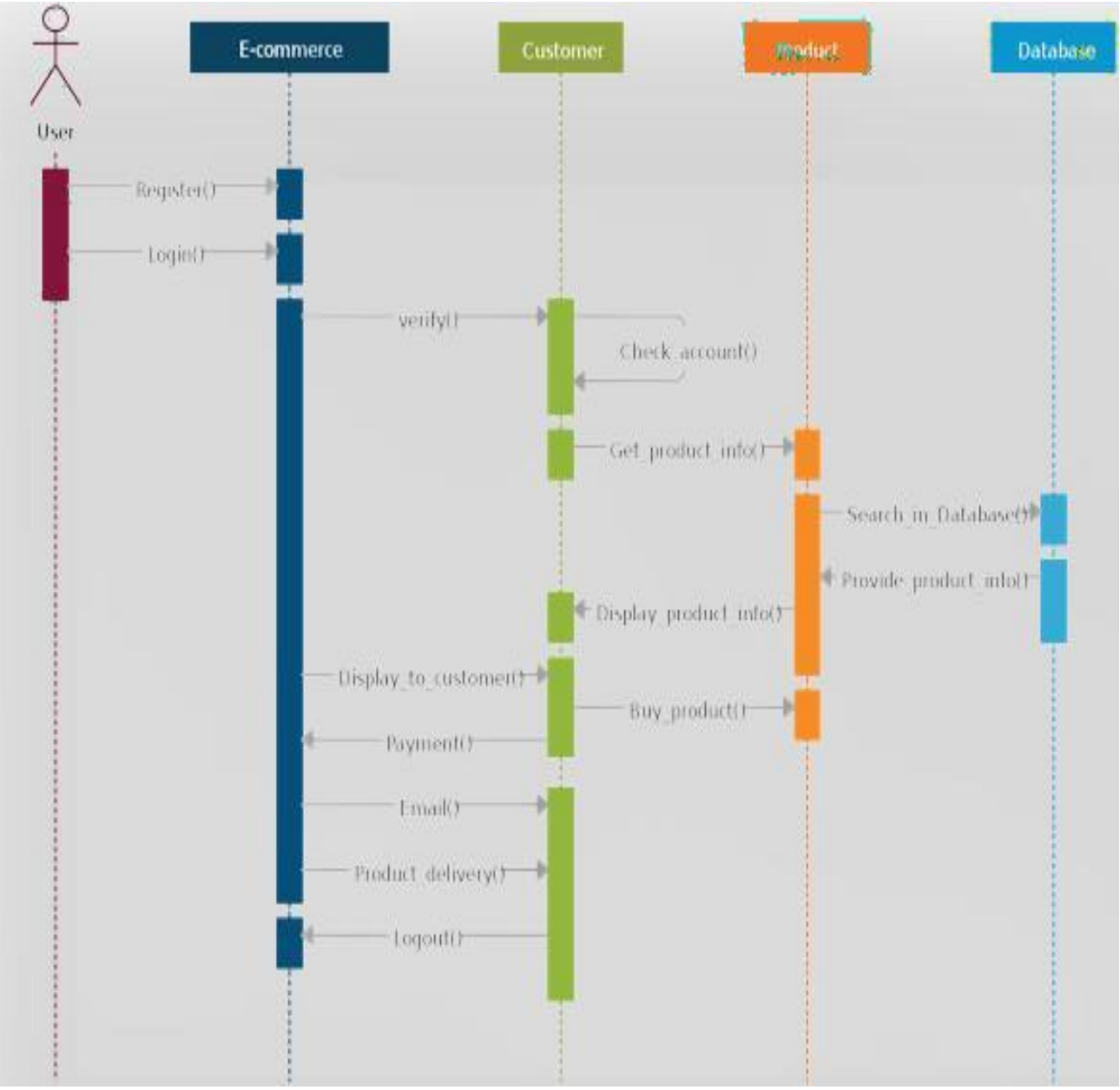


Fig 3.1 Web Scrapping Sequence Diagram

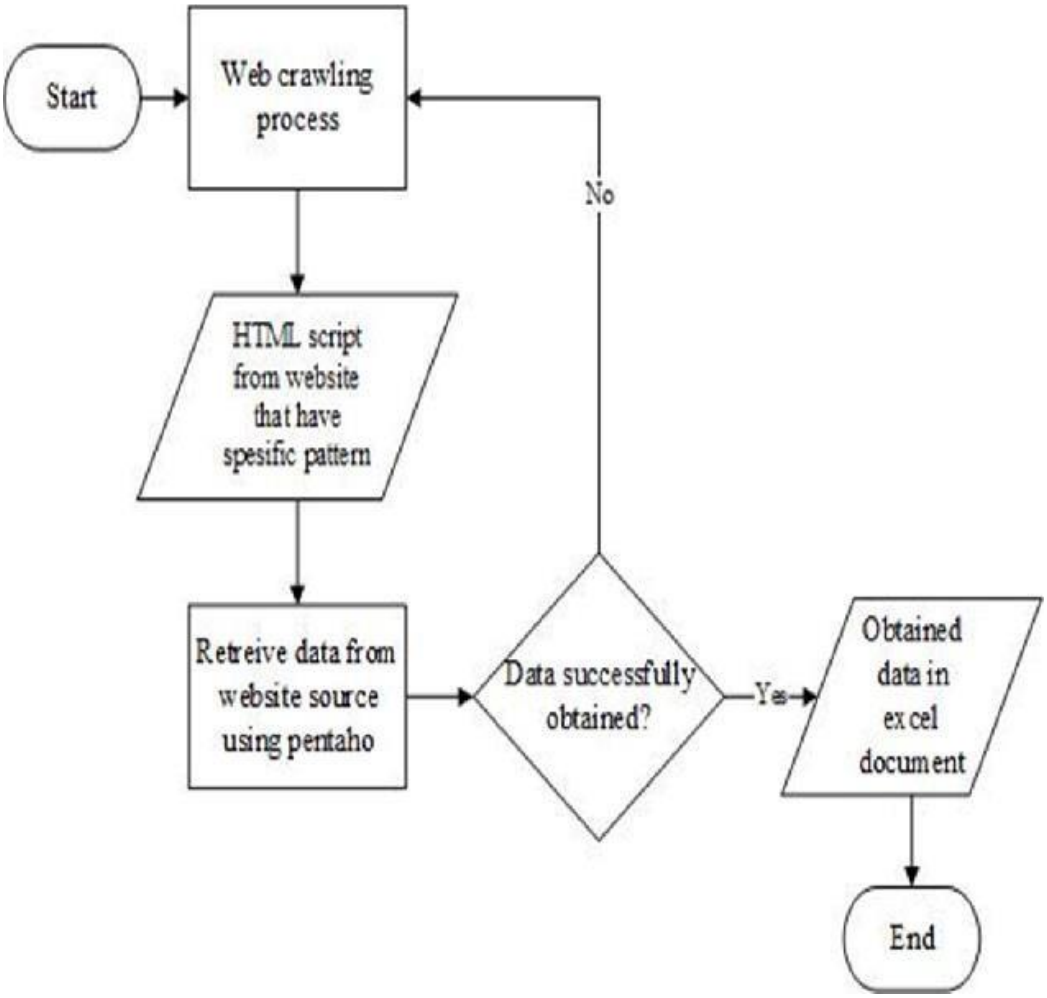


Fig 3.2 DFD diagram

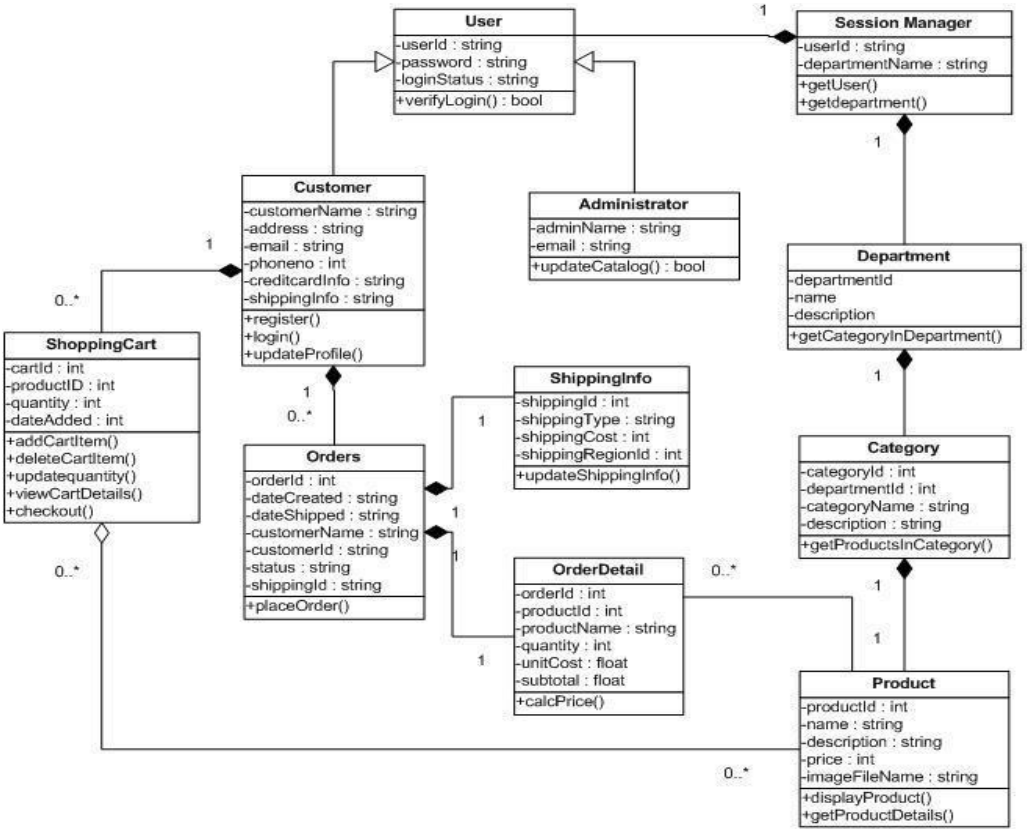


Fig 3.3 Class Diagram

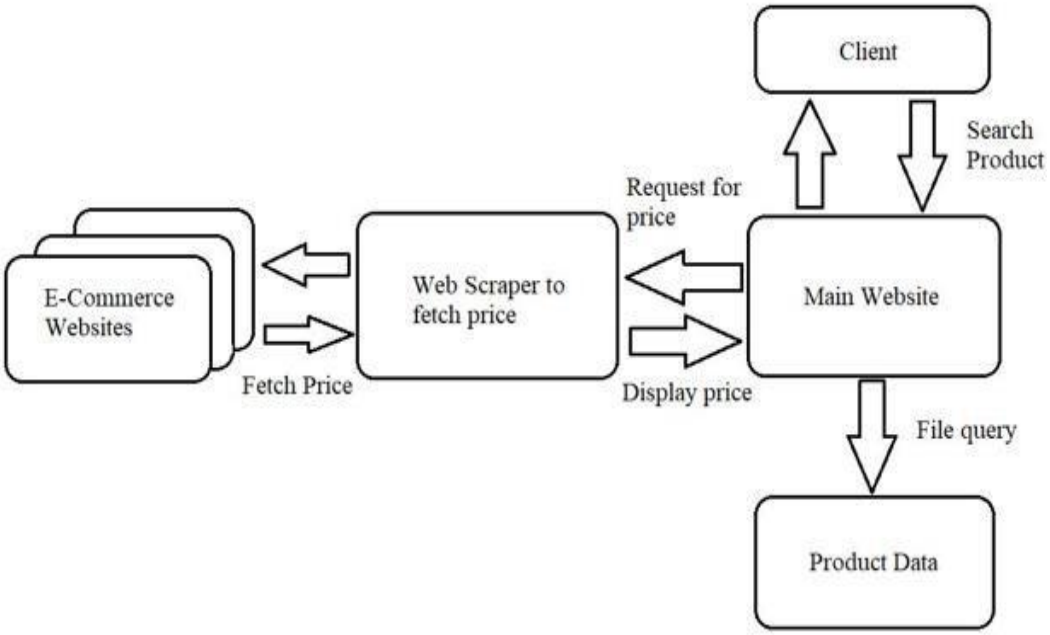


Fig 3.4 Block Diagram

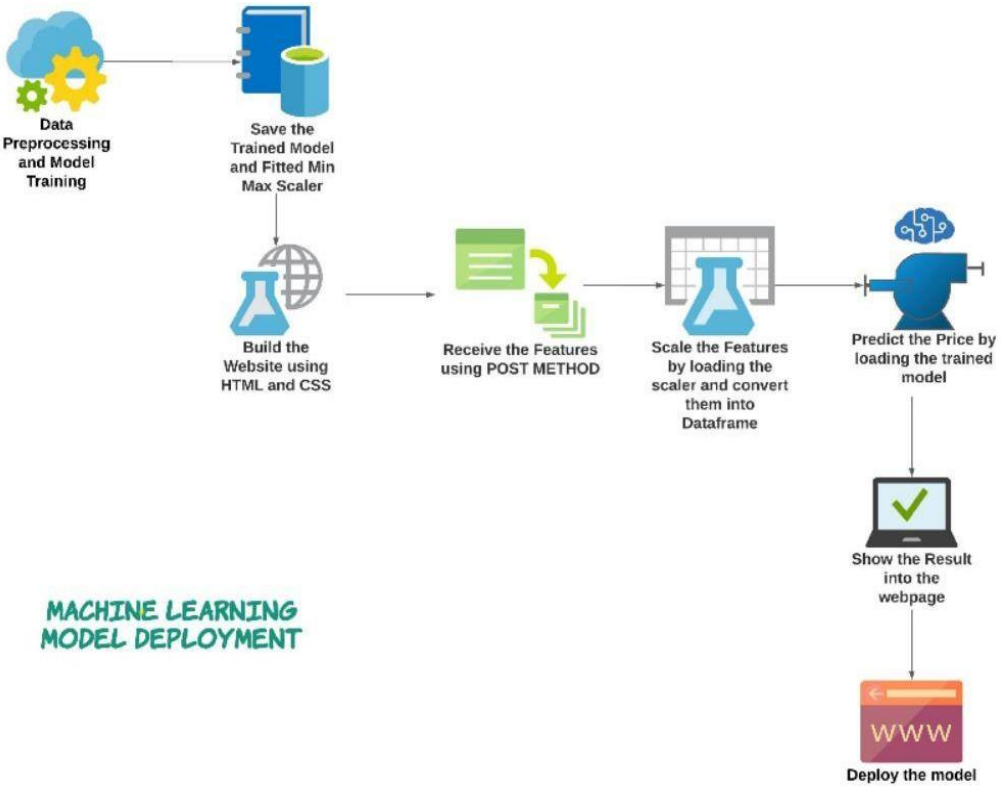


Fig 3.5 ML Model

Chapter-4
Module Description

HOW DO WEB SCRAPERS WORK?

- First step, robots.txt: One of the most important and overlooked step is to check the robots.txt file to ensure that we have the permission to access the web page without violating any terms or conditions.
- Secondly, the web scraper will be given one or more URLs to load before scraping. The scraper then loads the entire HTML code for the page in question. More advanced scrapers will render the entire website, including CSS and JavaScript elements.
- Then the scraper will either extract all the data on the page or specific data selected by the user before the project is run.
- Ideally, the user will go through the process of selecting the specific data they want from the page. For example, you might want to scrape an Amazon product page for prices and models but are not necessarily interested in product reviews.
- Lastly, the web scraper will output all the data that has been collected into a format that is more useful to the user.

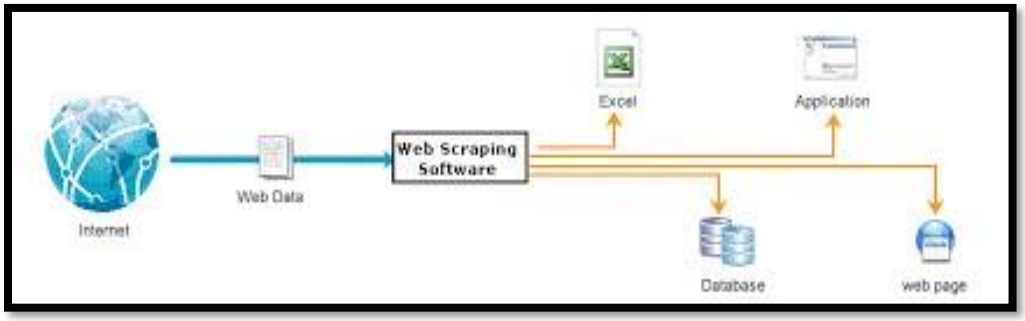


Fig 4.1 Architecture of web scrapping

Required Tools

Hardware Requirements:

- RAM: 4GB and Higher
- Processor: Intel i3 and above
- Hard Disk: 500GB

Software Requirements

- OS: Windows or Linux
- Python IDE: python 2.7.x and above
- Language: Python, MySQL
- WebScrapper
- Supervised Learning
- Library: Selenium

Data for price prediction is obtained from the web. The diagram of the overall architecture is indicated in Figures 1 and 2. Firstly, data is extracted by scraping technology. Then data is sent to a csv file to make analysis. Bagging Algorithms, ANN, SVM, XGBoost, KNN and RF are applied to data. As a result, the model generates the predicted outputs.

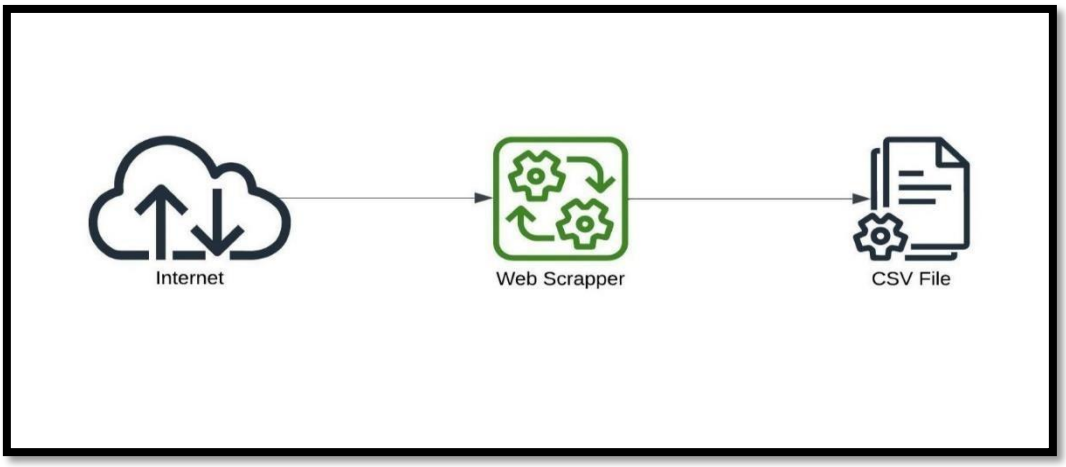


Fig 4.2 Working of Web Scrapper

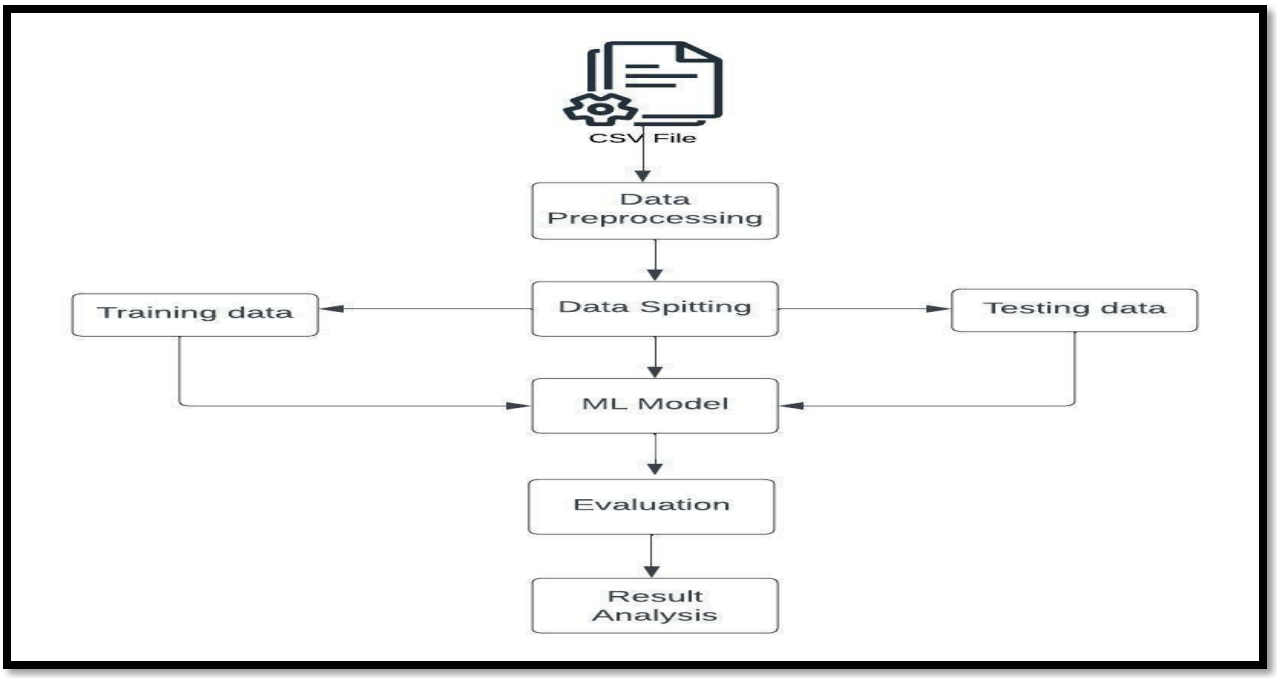


Fig 4.3 ML model

Python can run on many different platforms and has a simple syntax similar to the English language, therefore, it is easy to code. First of all, the goal of web scraping is to collect web data that would be in HTML format or Json. Python provides a library called Requests, which is a simple HTML library that allows you to integrate your Python programs with web services. Once you find the relevant data for your project on the webpage, you can download it for valuable insight. To do this, Python provides another library called

BeautifulSoup, which helps you retrieve particular content from a web page, remove HTML tags, and save the information. The last step in web scraping is to save the collected data in a structured form. With the help of Python Pandas Library, you can store the data in the desired format.

The success rate of information retrieval depends on the information required and what percentage of overhead the user receives . The created system can display information according to the user's wishes with a precision rate 93.9% and recall rate of 100%. Precision rate is calculated using

Precision= (no of relevant data retrieved)/(total no of data retrieved)

The number of relevant data retrieved is the data corresponding to the needs of the system. A total number of data retrieved is all data stored in the database, which is the result of scraping. Precision data that does not reach 100% due to data that is not as expected. Differences in the categorization of goods on each website to be one of the alleged occurrences of it.

Recall rate is calculated using:

Recall = (no of relevant data retrieved)/(total no of relevant data)

To maintain the present information. Scraping will be done periodically every week. However, in certain circumstances scraping can be done at any time. Changes in scraping schedules will result in deletion of old data automatically and the storage of the latest data scraping results. The next scraping period will be updated automatically by the system.

After the web scraping process, the data pre-processing phase is performed. Firstly, it is checked whether there is a deficiency in the attributes of the data at this stage. In our work, these missing data are removed from the system since the identified deficiency is very low compared to the total data. We did a number of preprocessing procedures on the data as a

whole to make it simpler to process the review texts. To assess our model for the best accuracy, we use 80% training and 20% testing. In addition, machine learning models are not deterministic models. Many algorithms use randomization to split variables and evaluate the outcome. A by product of this trait is that the predictive models will vary slightly each time the algorithm is implemented. Even with these slight variations, machine learning models are highly effective and used in many predictive applications.

Although all the data sets are known in this case study, the test set will be put through the predictive model, and accuracy will be determined by comparing actuals obsolescence statuses and obsolescence date with the one predicted by the model. This practice is known as validation and is a best practice for model creation and evaluation because the data used to create a prediction model are never used to validate its accuracy. Currently, the majority of the obsolescence forecasting models in the literature estimate model accuracy by using the same data used to create the model.

The next step in the case study was to run the training data set through a machine-learning algorithm to create a predictive model. Machine learning has many algorithms and infinitely more if counting all the slight variations that can be done to increase accuracy. Three machine-learning algorithms, artificial neural networks (ANNs), support vector machines (SVMs), XG Boost , Multi Linear Regression (MLR) and random forest (RF) will be applied to this case study.

The random forest algorithm is a supervised learning algorithm used for classification and regression problems. As we can predict on the name, the random forest algorithm consists of decision trees trained with the bagging technique. In the bagging method, base learners are randomly trained with subsets in the training set. As the number of trees in the random forest algorithm increases, the algorithm gives more precise results. The biggest advantage of the random forest algorithm is to perform well in datasets with missing data. At this point, random forest algorithms are frequently used for both large datasets and small

datasets. Another advantage of the random forest algorithm is that it provides a deeper exploration of the dataset by establishing various models on the dataset.

Artificial neural networks are an algorithm inspired by the working principles and functions of the human brain. The working principle of artificial neural networks is the same as the working of neural networks in the brain. Learning in biological systems is provided by synaptic connections between neurons. Information from people's sense organs updates synaptic connections. In artificial neural networks, on the other hand, samples represent information coming from sense organs. Learning occurs as a result of using examples and associating them with results. Training, on the other hand, refers to the process that continues until the determination of connection weights using examples and obtaining the best results.

K-Nearest Neighbor algorithm is used in classification and regression modeling and is considered the easiest supervised learning algorithm compared to other algorithms. This algorithm first emerged for the solution of classification problems, then it began to be used for solving regression problems. K Nearest Neighbors algorithm is one of the other supervised learning algorithms. Unlike, it does not have a training stage. In the approach of this algorithm, training and testing are considered the same operation. The absence of a training set makes this algorithm very easy to implement, but it is not recommended to be used on large data sets because it has low performance compared to other algorithms. In the K Nearest Neighbor algorithm, predictions are made based on observation similarity. It is a non-parametric supervised learning algorithm compared to other algorithms.

XGBoost algorithm is a decision tree-based machine learning algorithm. This algorithm is known as the best among decision tree-based algorithms. The reason for this is that it has been developed with various optimization and software techniques in order to make better predictions with less resource usage.

Multiple linear regression is a method used to see how independent features are related to the feature affected by other features. In order to determine the order of importance of the independent variables, MLP determines the effect of the independent variables in the regression equation on the dependent variable. Simple linear regression looks at whether only one feature explains the dependent feature, while multiple linear regression looks at the status of more than one feature.

MSE is an evaluation method that measures the performance of the estimator in regression models. In this evaluation metric, the estimated values are subtracted from the real values and squared. RMSE is a metric that calculates the distance between predicted and actual values and measures the magnitude of the error. In simpler words, it is the standard deviation of the predicted values. RMSE is calculated by squaring the value resulting from the operations. The RMSE value can range from 0 to ∞. Algorithms with a lower value on this metric are considered to perform better. MAE is an error metric created by summing the absolute error values. In this metric, the first estimated values are subtracted from the actual values. These values are then summed up in their absolute value and divided by the number of observations. MAE value is always positive and can take an infinite value. Algorithms with lower values in this evaluation metric are considered to perform better.

In the final step, once the algorithm constructs a predictive model, each part or element from the unknown data set is run through the model and receives a predicted label.

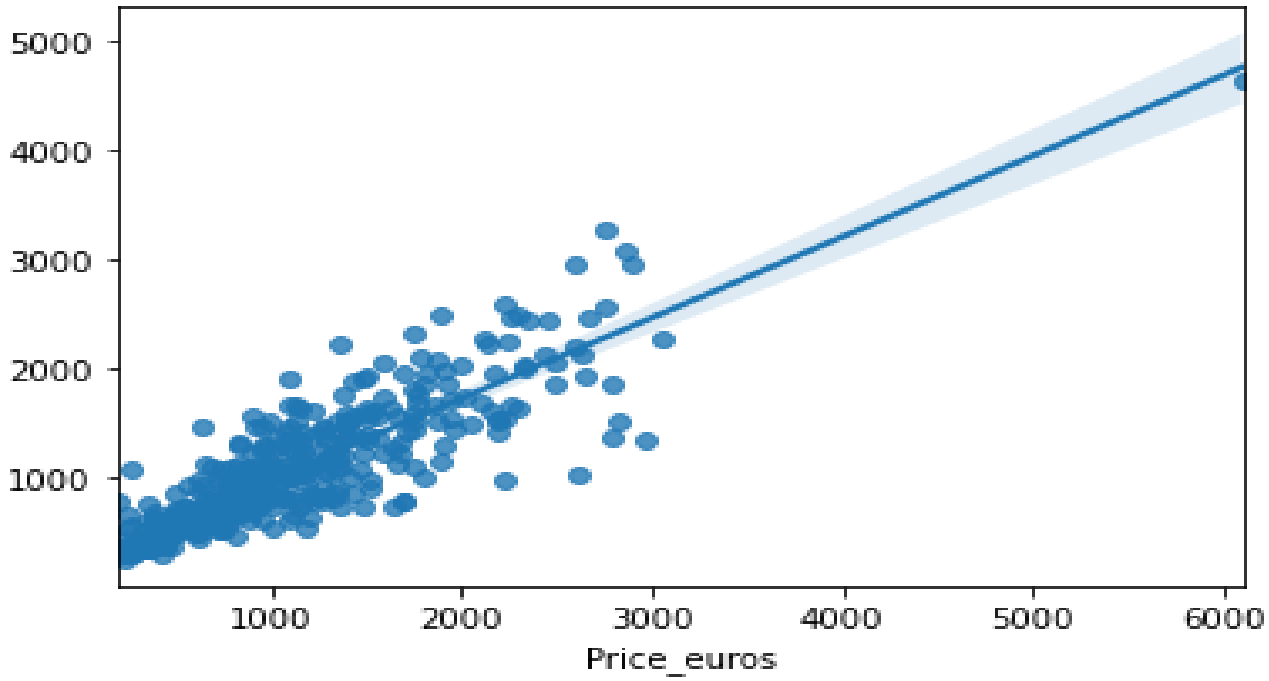
CHAPTER-5

Result

The databases are rather small: they only contains a little more than 300 observations. We stacked observations coming from both websites to get a more robust model. In production, it would also imply an easier estimation, as it would avoid having several different models (this would also avoid multiple reestimations every year for each type of product). We try to fit a linear regression on a small subset of 16 variables simultaneously present in both databases. We first modelled untransformed prices, which lead to disappointing results. The coefficients are highly volatile on the three databases. The accuracies are low. The accuracy of the linear model on all data is around 81%. A lasso regression discard four variables : weight, backlit keyboard and screen size. The accuracy of the linear regres sion drop to 78%. The coefficients are sometimes very different from one database to the other. The variables may be filled differently. The pricing policies could be different between two websites. The number of variables is quite low. Some key features might have been omitted. The model seems to be very much influenced by high prices. It could be a good idea to remove some highly influencial laptops. Only some of the characteristics are common to both bases. Some laptops may be present on both website. It would be possible to compare their collected features.

We introduced all of the variables selected in one of the random forests in a linear regressions to estimate price changes. We tried to predict the transformed and untransformed prices. To take the logarithm of prices has one main advantage : it always produces positive estimates. A few estimates were negative with the regression in levels. We had to drop them from the calculation. The accuracy of the linear models is lower than the one of the random forests. Random forests are non linear estimators. They are fitted to the complex characteristics set of laptops. But they can’t select the best subset of features to use in linear regression. The average accuracy on the 100 randomly drawn samples is

lower. It is below 86 %. We used a subset obtained by non linear method to estimate a linear model. That's why there is such a drop in accuracy.



Web Implementation testing Scraping for Data Retrieval on Marketplace Sites

what has been done is as follows:

1. Based on the results of black box testing the system can run according to the functionality derived from the analysis need.

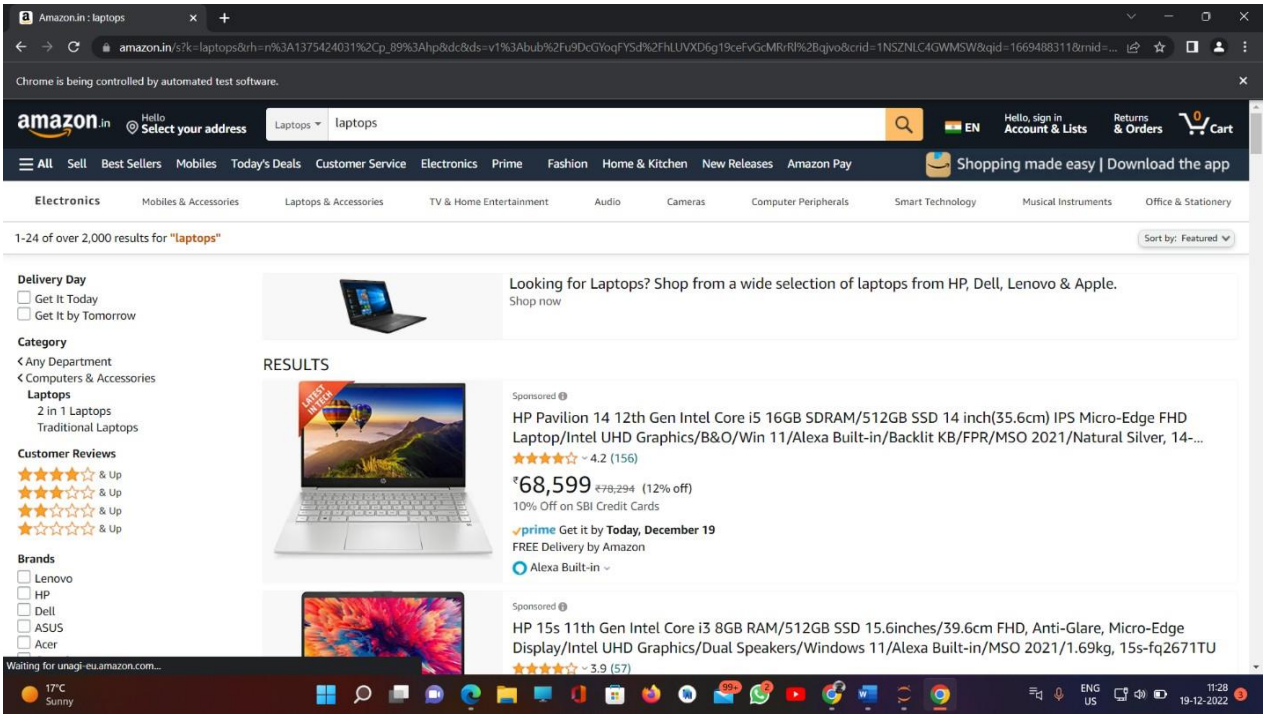
2. Based on the results of the system white box testing successfully run with the condition of the data retrieved according to the script that is run, where every one series of process statements in the program has executed at least once during testing and all logical conditions have been tested and worked.

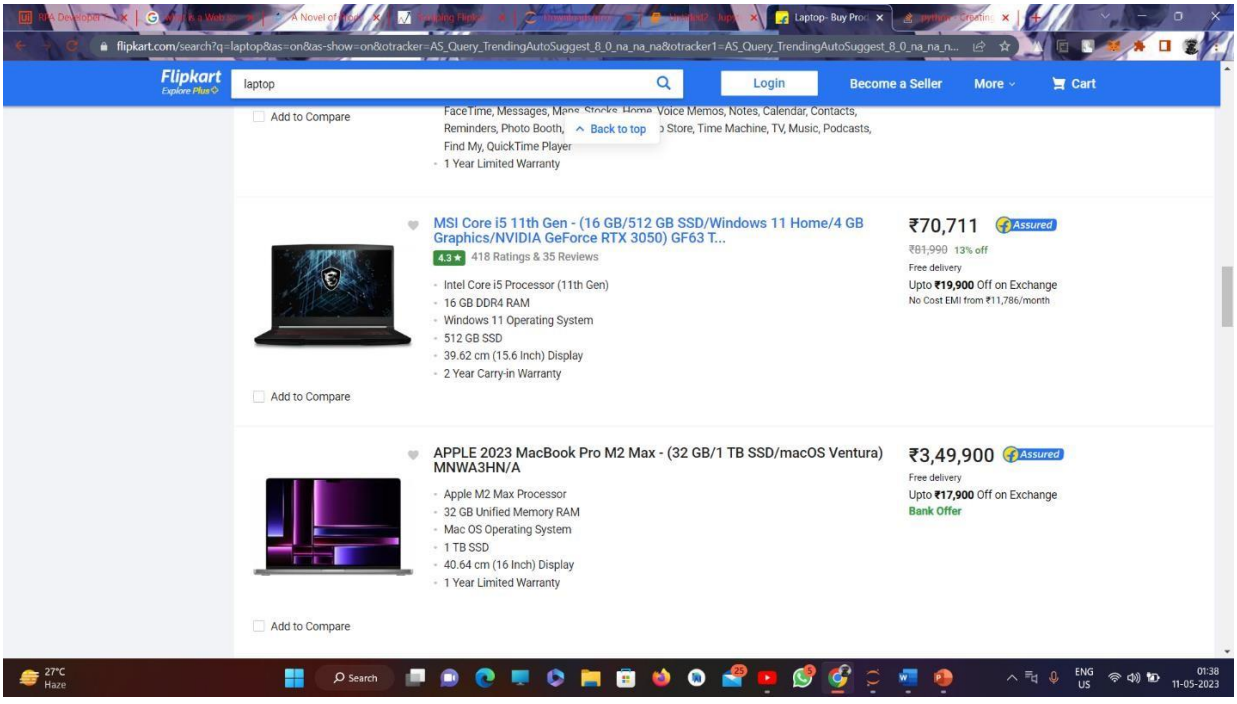
Source Code

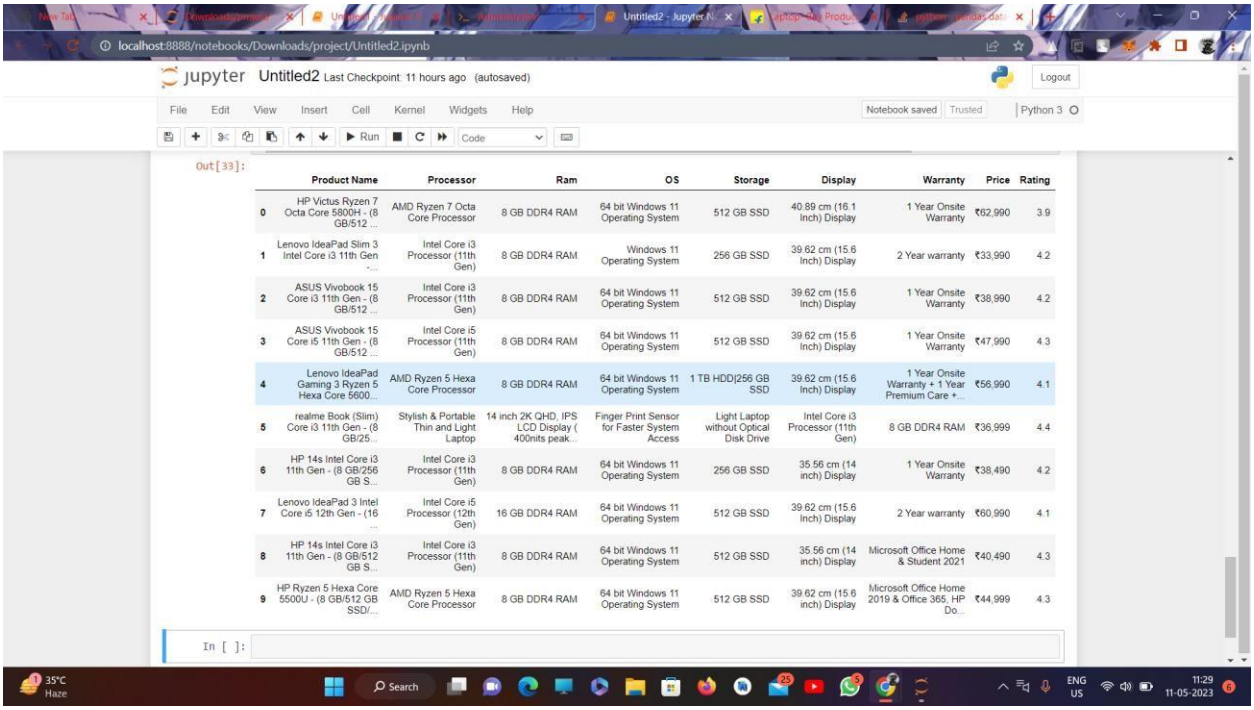
```
import csv

from bs4 import BeautifulSoup

from selenium import webdriver

def get_url(search_term):


t="https://www.amazon.in/s?k=laptops&rh=n%3A1375424031%2Cp_89%3A{}&dc&ds=v1%3Abub%2Fu9DcGYoqFYSd%2FhLUVXD6g19ceFvGcMRrRl%2Bqjvo&crid=1NSZNLC4GWMSW&qid=1669488311&rnid=3837712031&sprefix=laptops%2Caps%2C560&ref=sr_nr_p_89_3"

    search_term=search_term.replace(' ','.')

    url=t.format(search_term)

    url+='&page{}'

    return url

def extract_record(item):

    atag=item.h2.a

    d=atag.text.strip()

    url='https://www.amazon.in'+atag.get('href')

    try:

        price=item.find('span','a-price')

        price=price.find('span','a-offscreen').text

    except AttributeError:

        return

    try:

        rating=item.i.text

    except AttributeError:
```

```
            rating=''


    result=(d,price,rating,url)
    return result
def main(search_term):
    driver=webdriver.Chrome()
    url=get_url(search_term)
    record=[]
    for page in range(1,21):
        driver.get(url.format(page))
        soup=BeautifulSoup(driver.page_source,'html.parser')
        result=soup.find_all('div',{'data-component-type':"s-search-result"})
        for item in result:
                r=extract_record(item)
                if r:
                    record.append(r)
    driver.close()
    with open('result.csv','w',newline='',encoding='utf-8') as f:
        writer=csv.writer(f)
        writer.writerow(['Description',"price",'Rating','Url'])
        writer.writerows(record)
main("hp")
```

Output


Amazon.in : laptops
amazon.in/s?k=laptops&rh=n%3A1375424031%2Cp_89%3Ahp&dc&ds=v1%3Abub%2Fu9DcGYoqFYSd%2FhLUVXD6g19ceFvGcMRrRl%2Bqjvo&crid=1NSZNLC4GWMSW&qid=1669488311&rnid=...
Chrome is being controlled by automated test software.
amazon.in
Hello Select your address
Laptops
laptops
EN
Hello, sign in Account & Lists
Returns & Orders
Cart
All Sell Best Sellers Mobiles Today's Deals Customer Service Electronics Prime Fashion Home & Kitchen New Releases Amazon Pay
Shopping made easy | Download the app
Electronics Mobiles & Accessories Laptops & Accessories TV & Home Entertainment Audio Cameras Computer Peripherals Smart Technology Musical Instruments Office & Stationery
1-24 of over 2,000 results for "laptops"
Sort by: Featured
Delivery Day
Get It Today
Get It by Tomorrow
Category
Any Department
Computers & Accessories
Laptops
2 in 1 Laptops
Traditional Laptops
Customer Reviews
& Up
Brands
Lenovo
HP
Dell
ASUS
Acer
Looking for Laptops? Shop from a wide selection of laptops from HP, Dell, Lenovo & Apple.
Shop now
RESULTS
Sponsored
HP Pavilion 14 12th Gen Intel Core i5 16GB SDRAM/512GB SSD 14 inch(35.6cm) IPS Micro-Edge FHD Laptop/Intel UHD Graphics/B&O/Win 11/Alexa Built-in/Backlit KB/FPR/MSO 2021/Natural Silver, 14-...
4.2 (156)
₹68,599 ₹78,294 (12% off)
10% Off on SBI Credit Cards
prime Get it by Today, December 19
FREE Delivery by Amazon
Alexa Built-in
Sponsored
HP 15s 11th Gen Intel Core i3 8GB RAM/512GB SSD 15.6inches/39.6cm FHD, Anti-Glare, Micro-Edge Display/Intel UHD Graphics/Dual Speakers/Windows 11/Alexa Built-in/MSO 2021/1.69kg, 15s-fq2671TU
3.9 (57)
Waiting for unagi-eu.amazon.com...
17°C Sunny
ENG US
11:28 19-12-2022


flipkart.com/search?q=laptop&as=on&as-show=on&otracker=AS_Query_TrendingAutoSuggest_8_0_na_na_na&otracker1=AS_Query_TrendingAutoSuggest_8_0_na_na_n...
Flipkart
laptop
Login
Become a Seller
More
Cart
Add to Compare
FaceTime, Messages, Maps, Stocks, Home, Voice Memos, Notes, Calendar, Contacts, Reminders, Photo Booth, Back to top, Store, Time Machine, TV, Music, Podcasts, Find My, QuickTime Player
1 Year Limited Warranty
MSI Core i5 11th Gen - (16 GB/512 GB SSD/Windows 11 Home/4 GB Graphics/NVIDIA GeForce RTX 3050) GF63 T...
4.3★ 418 Ratings & 35 Reviews
Intel Core i5 Processor (11th Gen)
16 GB DDR4 RAM
Windows 11 Operating System
512 GB SSD
39.62 cm (15.6 Inch) Display
2 Year Carry-in Warranty
₹70,711 Assured
₹81,990 13% off
Free delivery
Upto ₹19,900 Off on Exchange
No Cost EMI from ₹11,786/month
Add to Compare
APPLE 2023 MacBook Pro M2 Max - (32 GB/1 TB SSD/macOS Ventura) MNWA3HN/A
Apple M2 Max Processor
32 GB Unified Memory RAM
Mac OS Operating System
1 TB SSD
40.64 cm (16 Inch) Display
1 Year Limited Warranty
₹3,49,900 Assured
Free delivery
Upto ₹17,900 Off on Exchange
Bank Offer
Add to Compare
27°C Haze
Search
ENG US
01:38 11-05-2023



	Product Name	Processor	Ram	OS	Storage	Display	Warranty	Price	Rating
0	HP Victus Ryzen 7 Octa Core 5800H - (8 GB/512 ...	AMD Ryzen 7 Octa Core Processor	8 GB DDR4 RAM	64 bit Windows 11 Operating System	512 GB SSD	40.89 cm (16.1 inch) Display	1 Year Onsite Warranty	₹62,990	3.9
1	Lenovo IdeaPad Slim 3 Intel Core i3 11th Gen -...	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	Windows 11 Operating System	256 GB SSD	39.62 cm (15.6 inch) Display	2 Year warranty	₹33,990	4.2
2	ASUS Vivobook 15 Core i3 11th Gen - (8 GB/512 ...	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	64 bit Windows 11 Operating System	512 GB SSD	39.62 cm (15.6 inch) Display	1 Year Onsite Warranty	₹38,990	4.2
3	ASUS Vivobook 15 Core i5 11th Gen - (8 GB/512 ...	Intel Core i5 Processor (11th Gen)	8 GB DDR4 RAM	64 bit Windows 11 Operating System	512 GB SSD	39.62 cm (15.6 inch) Display	1 Year Onsite Warranty	₹47,990	4.3
4	Lenovo IdeaPad Gaming 3 Ryzen 5 Hexa Core 5600...	AMD Ryzen 5 Hexa Core Processor	8 GB DDR4 RAM	64 bit Windows 11 Operating System	1 TB HDD|256 GB SSD	39.62 cm (15.6 inch) Display	1 Year Onsite Warranty + 1 Year Premium Care +...	₹56,990	4.1
5	realme Book (Slim) Core i3 11th Gen - (8 GB/25...	Stylish & Portable Thin and Light Laptop	14 inch 2K QHD, IPS LCD Display (400nits peak...	Finger Print Sensor for Faster System Access	Light Laptop without Optical Disk Drive	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	₹36,999	4.4
6	HP 14s Intel Core i3 11th Gen - (8 GB/256 GB S...	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	64 bit Windows 11 Operating System	256 GB SSD	35.56 cm (14 inch) Display	1 Year Onsite Warranty	₹38,490	4.2
7	Lenovo IdeaPad 3 Intel Core i5 12th Gen - (16 ...	Intel Core i5 Processor (12th Gen)	16 GB DDR4 RAM	64 bit Windows 11 Operating System	512 GB SSD	39.62 cm (15.6 inch) Display	2 Year warranty	₹60,990	4.1
8	HP 14s Intel Core i3 11th Gen - (8 GB/512 GB S...	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	64 bit Windows 11 Operating System	512 GB SSD	35.56 cm (14 inch) Display	Microsoft Office Home & Student 2021	₹40,490	4.3
9	HP Ryzen 5 Hexa Core 5500U - (8 GB/512 GB SSD/...	AMD Ryzen 5 Hexa Core Processor	8 GB DDR4 RAM	64 bit Windows 11 Operating System	512 GB SSD	39.62 cm (15.6 inch) Display	Microsoft Office Home 2019 & Office 365, HP Do...	₹44,999	4.3

CHAPTER-6

Conclusion

CONCLUSION

"If programming is magic, then web scraping is wizardry," said Ryan Mitchell. The presence of the Internet led to increasing source of information that can be accessed so that information seeking activities become the most common activities performed and became one of the activities that took quite a bit. The Internet will be remembered as the first place where we can collect huge amounts of data without spending a lot of energy or money. Whether in e-commerce or e-marketing, the use of the technique of web scraping will be the key to success as it will provide insight into the targeting market and help decision makers.

ACKNOWLEDGMENT

On this assignment, we gave it our best shot. This would not have been possible without the assistance and goodwill of a large number of people and organisations. We want to show our thanks to everyone. We are grateful to **Mr Rajeev Kumar.** for her direction and regular monitoring, as well as for supplying us with the essential project information and her help in seeing the project through to completion. We'd like to thank a member of the Galgotias University community for his exceptional cooperation and support in helping us complete this project. We would want to express our gratitude to the people in the business for their time and attention in this way. Our gratitude goes out to all of our project partners as well as those who volunteered their time and expertise to assist us.

REFERENCES

[1] Chaulagain, R.S., Pandey, S., Basnet, S.R., & Shakya, S. (2017, November). Cloud based web scraping for big data applications. In 2017 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 138-143). IEEE.

[2] Hajba, G.L. (2018). Website Scraping with Python: Using BeautifulSoup and Scrapy. Apress, Berkeley, California.

[3] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi and Tao Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", IEEE Trans. On Knowledge and Data Engineering, 2015

[4] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: an experiment on online product reviews," IEEE Intelligent Systems, vol. 25, no. 4, pp. 46–53, 2010 [5] Renita Crystal Pereira and T Vanitha, "Web Scraping of Social Networks", International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, pp. 237-239, Oct. 2018

[6] Glez-Peña, D.; Lourenço, A.; López-Fernández, H.; Reboiro-Jato, M.; Fdez-Riverola, F. Web scaping technologies in an API world. Brienfings Bioinform. 2014, 15, 788–794. [CrossRef] [PubMed]

[7] Saurkar, A.V.; Pathare, K.G.; Gode, S.A. An overview on web scraping techniques and tools. Int. J. Future Revolut. Comput. Sci.Commun. Eng. 2018, 4, 363–367.

[8] Gojare, S., Joshi, R., & Gaigaware, D. (2015). Analysis and design of selenium webdriver automation testing.

[9] Hillen, J. Web scraping for food price research. Br. Food J. 2019, 121, 3350–3361.

[10] A. Osmadi, E. M. Kamal, H. Hassan, and H. A. Fattah, "Exploring the elements of housing price in Malaysia," Asian Soc. Sci., vol. 11, no. 24, pp. 26–38, 2015, doi: 10.5539/ass.v11n24p26.

[11] Socher, R. (2018). Ai's next great challenge: Understanding the nuances of language. Harvard Business Review

[12] Kukreja, H., N, B., S, S. C., & S, K. (2016). AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK. Journal of Electrical & Electronics Engineering, School of Engineering & Technology, Jain University, 27-29.

[13] Vicario- Becerra, R., Alaminos, D., Aranda, E., & Fernández-Gámez, M. A. (2020). Deep recurrent convolutional neural network for bankruptcy prediction: A case of the restaurant industry. Sustainability, 12(12), 5180

[14] Cao, Q., and M. J. Schniederjans. 2004. "Empirical study of the relationship between operations strategy and information systems strategic orientation in an e-commerce environment." Review of. International Journal of Production Research 42 (15):2915-39. Chakravarty, Anindita, Y

[15] Chen, Mao, and Jaswinder Pal Singh. 2001. Computing and using reputations for internet ratings. Paper presented at the Proceedings of the 3rd ACM conference on Electronic Commerce.

Publication

Gmail Tanishq Pundir <tanishqpundir2@gmail.com>

Acceptance Notification 5th IEEE ICAC3N-23 & Registration: Paper ID 488
1 message

Microsoft CMT <email@msr-cmt.org> Tue, May 16, 2023 at 12:11 AM
Reply-To: Vishnu Sharma <vishnu.sharma@galgotiacollege.edu>
To: Tanishq Pundir <tanishqpundir2@gmail.com>

Dear Tanishq Pundir,
galgotias university

Greetings from ICAC3N-23 ...!!!

Congratulations....!!!!!

On behalf of the 5th ICAC3N-23 Program Committee, we are delighted to inform you that the submission of "Paper ID- 488 " titled " Product Price Prediction by Machine Learning Method through Web Scrapping " has been accepted for presentation and further publication with IEEE at the ICAC3N- 23 subject to incorporate the reviewers and editors comments in your final paper. All accepted papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

For early registration benefit please complete your registration by clicking on the following Link: https://forms.gle/8e6RzNbho7CphnYN7 on or before 20 May 2023.

Registration fee details are available @ https://icac3n.in/register.

You must incorporate following comments in your final paper submitted at the time of registration for consideration of publication with IEEE:

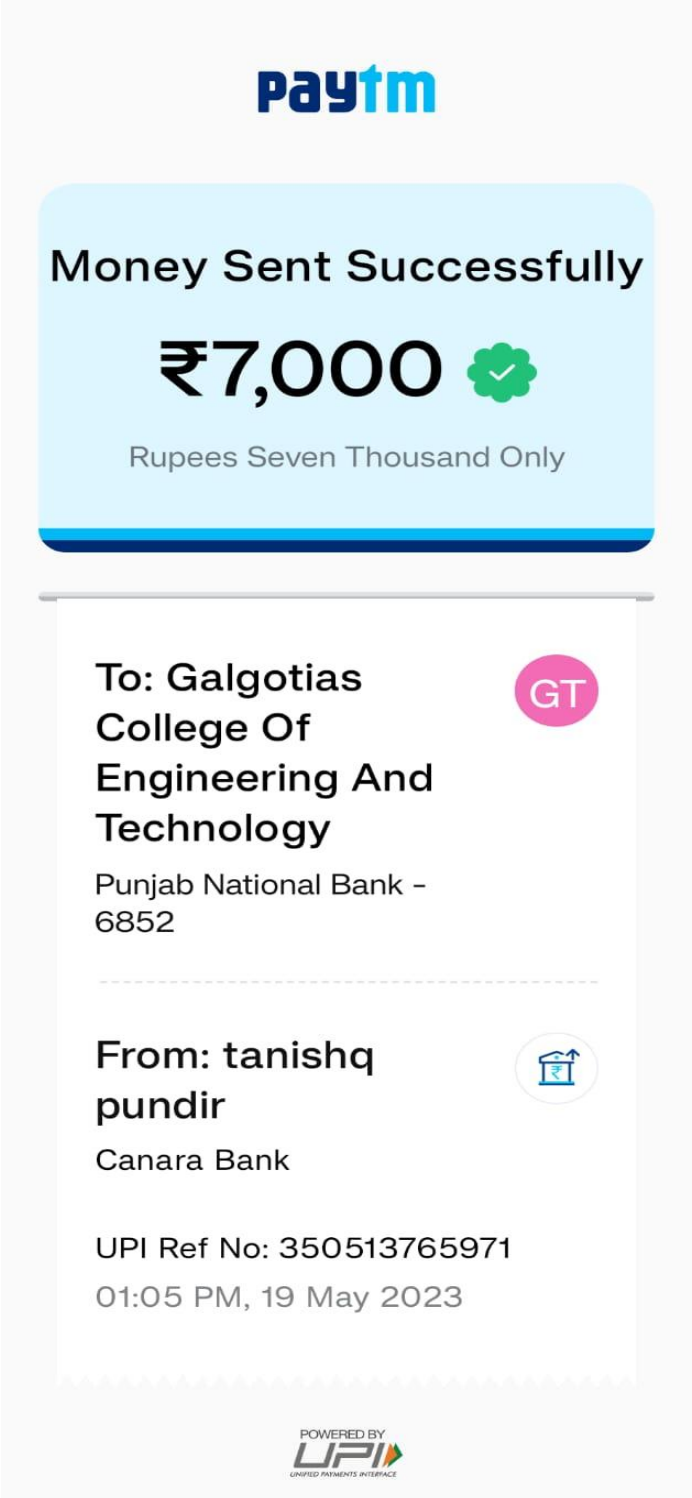
The title chosen "Product Price Prediction by Machine Learning Method through Web Scrapping " is relevant.
The formatting of paper is not proper. Formatting must be strictly as per template. Otherwise it will not be published.
Author list formatting is not proper. All authors information must be complete and should be in proper format and as per the sequence desired.
Citation of references within the content is not proper. Make only relevant citation. All references must be cited in content properly.
References are not in proper format. Format and assign number to the references properly.
An overview of paper is desired to eradicate typo and grammatical error.
Formatting and Quality of figures must be good.
All figure and tables must be properly captioned and numbered as per IEEE conference template.
Add a comparison table in literature review section with the work already done in this filed.
Conclusion and result section needs to be improved and require better explanation.

Editor Comments/Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Transfer of E-copyright to IEEE and Presenting paper in conference is compulsory for publication of paper in IEEE.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences. Plagiarism must be less then 20% (checked through Turnitin).
5. Change in paper title, name of authors or affiliation of authors will not be allowed after registration of papers.
6. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
7. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp) or write us at icac3n23@gmail.com.

Regards:
Organizing committee
ICAC3N – 2023

Acceptance


Paytm
Money Sent Successfully
₹7,000
Rupees Seven Thousand Only
To: Galgotias College Of Engineering And Technology
GT
Punjab National Bank - 6852
From: tanishq pundir
Canara Bank
UPI Ref No: 350513765971
01:05 PM, 19 May 2023
POWERED BY
UPI
UNIFIED PAYMENTS INTERFACE

Payment