

**Project Report**

on

**Advanced forecasting of demandable products prices  
using machine learning Algorithm**

**Submitted to**

**Sant Gadge Baba Amravati University**

**In partial Fulfillment of the Requirement**

**For the Degree of**

**Bachelor of Engineering in**

**Computer Science and Engineering**

**Submitted by:**

**Ms. Gayatri Zamare  
Ms. Gayatri Raghuwanshi  
Mr. Rupesh Apar  
Mr. Rupesh Dabhade  
Mr. Vaibhav Wankhade**

**Under the Guidance of**

**Prof. P. K. Bharne**



**Department of Computer Science and Engineering  
Shri Sant Gajanan Maharaj College of Engineering,  
Shegaon – 444 203 (M.S.)**

**2022-23**

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,  
SHEGAON – 444 203 (M.S.)  
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



**CERTIFICATE**

This is to certify that that Ms. Gayatri Zamare, Ms. Gayatri Raghuwanshi, Mr. Rupesh Apar, Mr. Rupesh Dabhade, and Mr. Vaibhav Wankhade, students of final year B.E. in the year 2022-23 of Computer Science and Engineering Department of this institute has completed the project work entitled “**Advanced forecasting of demandable products prices using machine learning Algorithm**” based on syllabus and has submitted a satisfactory account of his work in this report which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Computer Science and Engineering.

*P. K. Bharne*  
31/5/23.

**Dr. P. K. Bharne**  
Project Guide

*S. B. Patil*  
31/5/23  
**Dr. S. B. Patil**  
Head of Department

*S. B. Somani*  
**Dr. S. B. Somani**  
Principal

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,  
SHEGAON – 444 203 (M.S.)  
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



## CERTIFICATE

This is to certify that the project work entitled “**Advanced forecasting of demandable products prices using machine learning Algorithm**” submitted by **Ms. Gayatri Zamare, Ms. Gayatri Raghuwanshi, Mr. Rupesh Apar, Mr. Rupesh Dabhade, and Mr. Vaibhav Wankhade**, students of final year B.E. in the year 2022-23 of Computer Science and Engineering Department of this institute, is a satisfactory account of his work based on syllabus which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Computer Science and Engineering.

Internal Examiner

Date:

External Examiner

Date:

## Abstract

---

Knowing which items would be the most affordable is crucial for the organization. At this stage, categorization and prediction issues, such as price prediction, have been resolved using machine learning technology. This project seeks to produce timely and accurate price forecasts to assist the organisation in switching between neighboring markets to assist the organisation in switching between various neighbouring markets in order to sell their goods and obtain competitive rates. The data can be used by the company to make decisions regarding the timing of marketing.

The machine Learning technique allows for predicting the number of products/services to be purchased during a defined period. Demand forecasting is used in which first raw data is collected from the market, then according to the data the product prices are forecasted. This model is a catch-all phrase for the shopping process that establishes product prices in accordance with the level of supplier competition, the hour of the day, and the weather. This model will help to forecast the prices of products according to their historical data. At an organizational level, forecasts of product prices are an essential input to many decision-making activities in various functional areas such as operations, marketing, sales, production, and finance.

**Keywords:** Product prices forecasting, Machine Learning, Linear Regression, Lasso Regression, XG Boost Algorithm, Gradient Boosting Algorithm, Random Forest Regressor, Streamlit, SkLearn.

## Acknowledgement

---

*The real spirit of achieving a goal is through the way of excellence and lustrous discipline. I would have never succeeded in completing my task without the cooperation, encouragement and help provided to me by various personalities.*

*I would like to take this opportunity to express my heartfelt thanks to my guide **Prof. P. K. Bharne** , for his esteemed guidance and encouragement, especially through difficult times. His suggestions broaden my vision and guided me to succeed in this work. I am also very grateful for his guidance and comments while studying part of my seminar and learnt many things under his leadership.*

*I extend my thanks to **Dr. S.B. Patil** Head of Computer Science & Engineering Department, Shri Sant Gajanan Maharaj College of Engineering, Shegaon for their valuable support that made me consistent performer.*

*I also extend my thanks to **Dr. S. B. Somani** ,Principal Shri Sant Gajanan Maharaj College of Engineering, Shegaon for their valuable support.*

*Also I would like to thanks to all teaching and non-teaching staff of the department for their encouragement, cooperation and help. My greatest thanks are to all who wished me success especially my parents, my friends whose support and care makes me stay on earth.*

- 1) Ms. Gayatri Zamare
- 2) Ms. Gayatri Raghuwanshi
- 3) Mr. Rupesh. Apar
- 4) Mr. Rupesh Dabhade
- 5) Mr. Vaibhav Wankhade

**Final Year B. E. Sem-VIII, CSE  
Session 2022-23**

# Contents

---

<i>Abstract</i>	<i>iii</i>
<i>Acknowledgement</i>	<i>iv</i>
<i>Contents</i>	<i>v</i>
<i>List of Figures &amp; Tables</i>	<i>vi</i>
<i>Abbreviations</i>	<i>vii</i>
<b>1. Introduction</b>	<b>1</b>
1.1 Preface	2
1.2 Aim	3
1.3 Problem of statement	3
1.4 Objectives	3
<b>2. Literature survey</b>	<b>4</b>
2.1 Related Work	5
<b>3. Methodology</b>	<b>7</b>
3.1 Machine Learning	8
3.2 Types of Machine Learning Techniques	9
3.3 Machine Learning Algorithm	12
3.3.1 Random Forest Regressor	12
3.3.2 Gradient Boosting	13
3.3.3 Support Vector Machine	14
3.3.4 Decision Tree	14
3.3.5 XG Boost	15
<b>4. System Requirements</b>	<b>16</b>
4.1 Hardware Platform Used	17
4.2 Software Platform Used	17
<b>5. System Analysis</b>	<b>18</b>
5.1 Purpose	19
5.2 Project Scope	19
5.3 Organization of Project	19

<b>6. Proposed Methodology</b>	<b>21</b>
6.1 Product dataset	22
6.2 Proposed Methodology Flowchart	22
6.3 Proposed Work	23
6.3.1 Data Gathering	23
6.3.2 Data Preprocessing	24
6.3.3 Importing Libraries	30
6.3.4 Splitting Dataset	31
<b>7. Implementation</b>	<b>32</b>
7.1 Implementation For Testing and Training	33
7.2 Export and Deploy the Model	37
<b>8. Conclusion</b>	<b>41</b>
8.1 Conclusion	42
8.2 Future Work	42
<b>9. Refrences</b>	<b>44</b>

## List of Figures and Tables

---

- Figure 3.1: Machine Learning Technology
- Figure 3.2: Types of Machine Learning
- Figure 3.3: Random Forest Algorithm
- Figure 3.4: Gradient Boosting Algorithm
- Figure 3.5: Support Vector Machine Algorithm
- Figure 3.6: Decision Tree Algorithm Figure
- 3.7: XG Boost Algorithm
- Figure 6.1: Product Dataset
- Figure 6.2: Proposed Methodology Flowchart
- Figure 6.3: First Five Rows of Dataset
- Figure 6.4: Distribution of target variable
- Figure 6.5: Null values in dataset
- Figure 6.6: Company Name (Brands) with Prices
- Figure 6.7: Types of laptops with prices
- Figure 6.8: Touch-screen configuration
- Figure 6.9: Correlation price with size
- Figure 6.10: Variation in price with processor
- Figure 6.11: Variation in price with RAM
- Figure 6.12: OS Column
- Figure 6.13: Log-Normal Transform
- Figure 6.14: Split dataset
- Figure 7.1: Implement pipeline
- Figure 7.2: Random Forest Accuracy
- Figure 7.3: Gradient Boost Algorithm Accuracy
- Figure 7.4: Support Vector Machine Accuracy
- Figure 7.5: Decision Tree Algorithm Accuracy
- Figure 7.6: XG Boost Algorithm Accuracy



Figure 7.7: Front Login page code

Figure 7.8: Authenticating page

Figure 7.9: Command for URL

Figure 7.10: Front Login page

Figure 7.11: Product predictor page

Figure 7.12: Predicted price of product

## Abbreviations

---

ML - MACHINE LEARNING

SVM – SUPPORT VECTOR MACHINE

VSCoDe – VISUAL STUDIO CODE

OS – OPERATING SYSTEM

**Chapter 1**  
**INTRODUCTION**

# **INTRODUCTION**

## **1.1 PREFACE**

Effective pricing forecasting assists organizations in anticipating price increases or cuts that may impact customer demand. Previous year's data on different products are being collected and we will predict the prices of products so that we will be able to make good marketing strategies. Using machine learning the system can predict what will be the price of a particular product today or after a certain day. Due to its striking advantages over conventional methods, machine learning techniques have recently become frequently used for price prediction. ML algorithms create models using training and test data, and then use these models to make predictions [1]. A prediction algorithm will be used to predict prices. Price and arrival data information strengthens the organization's bargaining position and increases the competitiveness among dealers. The organization can switch between neighboring markets more easily when price information is provided. The information can be used by the organization to make marketing timing decisions.

The majority of machine learning (ML) algorithms that were developed within the context of data science have dominated in recent years. It has previously been used to predict time series in the financial and economic sectors. Numerous empirical studies have demonstrated that machine learning methods are more effective than time series models at forecasting various financial asset values. Among the ML techniques, Linear Regression, XG-Boost, Random Forest (RF), and Gradient Boosting Machine (GBM) etc. are widely used. All these techniques are data-driven nonparametric techniques which learn the stochastic dependency in the data.

## **1.2 AIM**

To develop Machine Learning model to forecast the prices of product using different Regression algorithm

## **1.3 STATEMENT OF PROBLEM**

If any user wants to buy any product, then our application should be compatible to provide a tentative price of that product according to the user configurations.

## **1.4 OBJECTIVE OF PROJECT**

- To convert data into an appropriate form using various pre-processing techniques for the implementation of Machine Learning algorithms.
- To find a critical feature that will influence the prices of products.
- To determine the appropriate Machine learning algorithm for sales forecasting.
- To select various performance of algorithm metrics to compare the applied Machine Learning algorithm.

**Chapter 2**  
**LITERATURE**  
**SURVEY**

## **LITERATURE SURVEY**

### **2.1 RELATED WORK**

1. Julakha Jahan Jui In et al., [1] research focuses on, two machine-learning regression-based methods for predicting flat pricing—linear regression and random forest regression—was given. Data has been scraped from a number of real estate websites using the web scrapper (Data Toolbar) software. When developing the model, seven factors that can affect flat pricing were taken into consideration. Here, the data quality has been investigated using the histogram, residual charting, and ANOVA. The linear and random forest model has been created after preprocessing the dataset. MSE, RMSE, MAE, and MRE have all been computed in order to evaluate the performance of both techniques. The measured error rate has led to the conclusion that the random forest regression model performs well.
2. Yige Wang et al., [2] research state that clearly more practical in terms of price prediction is the decision tree fitting effect using Random Forest, the order of variable importance as opposed to OLS when dealing with complicated and irregular data. Therefore, we advise choosing a random forest in these two scenarios—one in which there are many observations in the dataset and the other in which there are complex samples with noise.
3. Xinshu Li et al., [3] focus on demand for commercial housing falls into three primary categories: speculative, investment, and owner-occupied. The investment need for self-employment is to purchase and lease commercial real estate to generate rental revenue. Hypothetical demand is bought.
4. Ujjawal Sonkambale et al., [4] research machine learning techniques to predict the price of used cars based on historical data from the Kaggle and Car Dekho websites. To determine which predictions offer the best performance and accuracy, the predictions are compared and examined. Delay filters, delay lines, power amplifiers, coaxial resonators, and ceramics are index terms.
5. Subba Rao Polamuri et al., [5] this paper focuses on using ML techniques to anticipate the behaviours tracking of the stock market sensex. It compares the accuracy of various models and selects an algorithm with high accuracy. The main aim is to apply innovative work to predict the behaviour tracking of the stock market Sensex.
6. Mohamed Ali Mohamed et al., [6] This work presents a promising approach for predicting pricing for retail goods, specifically seasonal Christmas items. Machine learning-based models, such as random forest and ARIMA, are effective in predicting the prices of these items. The results demonstrate that the irregular forest model outperforms other models. The study recommends using the random forest and ARIMA models, building hybrid models, defining the problem as a time-series problem and incorporating date and time input characteristics into the suggested models.

7. Rohit Joshi et al.,[7] this research paper highlights the importance of understanding customer shopping preferences in the growing online retail industry in India. It gathered information from 124 Indian respondents spread across 18 states and constructed and verified Random Forest prediction models for a number of product categories. The results showed that the model had a high sensitivity for products like books and electronics, while having a low sensitivity for products like movies, sporting goods, and bags.
8. Ameena sherin et al.,[8] this paper states that Linear Regression, Decision Tree, and Random Forest are three unique algorithms, each of which is based on a different component of the data, that may be used to predict home values. This research study, "House Price Prediction," describes how to utilize these three algorithms to do so. The problem is that predicted prices vary depending on how accurate they are. To get around this problem, we calculate the average of the projections. As a result, it helps with mistake prevention. This average price is a fair representation of the worth of a house. Clients will be pleased with the approach since it may produce reliable results and reduce the possibility of error.
9. Liu et al.,[9] this research focus on comparing the performance of different machine learning models for prediction of housing prices. The authors compared different algorithm like Linear Regression, Decision Tree, and Random Forest Regression whereas the CNN Random Forest model gives wrong output as compared to the other models in terms of prediction accuracy, suggesting it could be a useful tool for property valuation. Future research should incorporate additional data beyond housing development to improve the model's accuracy.



**Chapter 3**  
**METHODOLOGY**

## **METHODOLOGY**

### **3.1 MACHINE LEARNING**

Machine Learning is defined as the study of computer programs that leverage algorithms and statistical models to learn through inference and patterns without being explicitly programmed. Machine learning prediction, or prediction in machine learning, refers to the output of an algorithm that has been trained on a historical dataset. The algorithm then generates probable values for unknown variables in each record of the new data. The purpose of prediction in machine learning is to project a probable data set that relates back to the original data. This helps organizations predict future customer behaviors and market changes. Essentially, prediction is used to fit a shape as close to the data as possible. With machine learning predictions, organizations use proactive decisions to avoid predicted user churn. To gain the most success with prediction in machine learning, organizations need to have the infrastructure in place to support the solutions, and high-quality data to supply the algorithm.

Prediction in machine learning allows organizations to make predictions about possible outcomes based on historical data. These assumptions allow the organization to make decisions resulting in tangible business results. Predictive analytics can be used to anticipate when users will churn or leave an organization. With this recognition, organizations have better potential to keep customers happy and satisfied.

Machine learning is closely connected to computational statistics, which focuses on generating predictions with computers; nevertheless, statistical learning is not all machine price prediction Big Mart Sales Prediction Using Machine Learning. The discipline of machine learning benefits from the study of mathematical optimization since it provides tools, theory, and application fields. Machine learning is also known as predictive analytics when it is used to solve business challenges. Machine learning

is significant because it allows businesses to see trends in customer behavior and company operating patterns while also assisting in the creation of new goods.

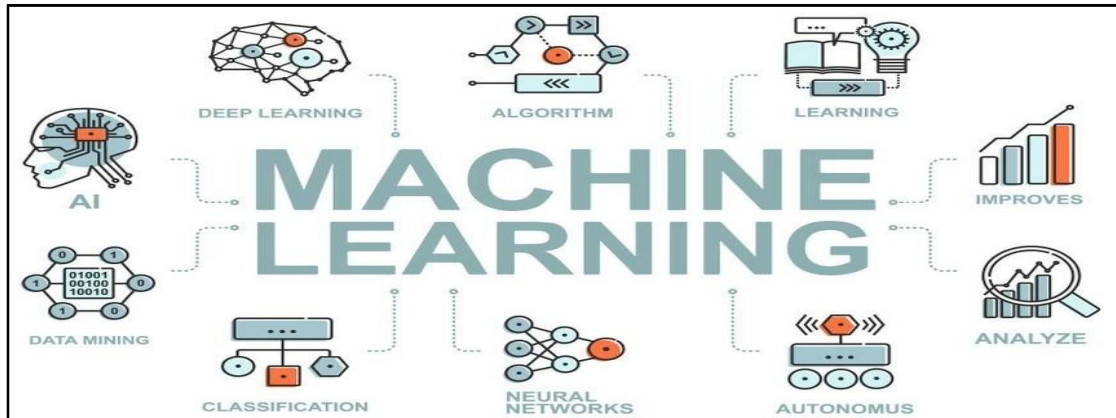


Figure 3.1: Machine Learning Technology

### 3.2 TYPES OF MACHINE LEARNING TECHNIQUES

Machine learning is broadly classified as supervised, unsupervised, semi-supervised, and reinforcement learning. A supervised learning model has two major tasks to be performed, classification and regression. Classification is about predicting a nominal class label, whereas regression is about predicting the numeric value for the class label. Mathematically, building a regression model is all about identifying the relationship between the class label and the input predictors. Predictors are also called attributes. In statistical terms, the predictors are called independent variables, while the class label is called dependent variable [2]. A regression model is a representation of this relationship between dependent and independent variables. Once this is learnt during the training phase, any new data is plugged into the relationship curve to find the prediction. This reduces the machine learning problem to solving a mathematical equation [3].

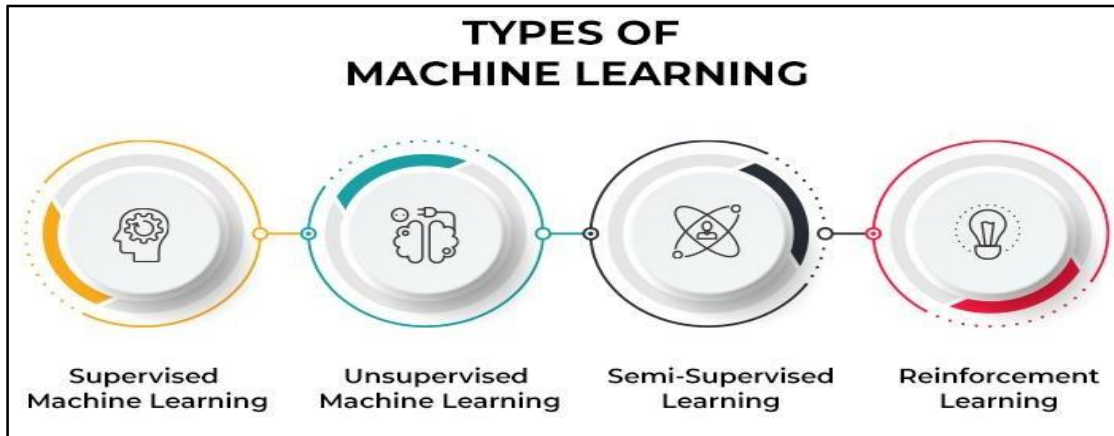


Figure 3.2: Types of Machine Learning

### A. SUPERVISED LEARNING

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

1. Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to what category new observations belong. For example, when filtering emails as 'spam' or 'not spam', the program must look at existing observational data and filter the emails accordingly.
2. Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.

3. **Forecasting:** Forecasting is the process of making predictions about the future based on the past and present data, and is commonly used to analyse trends.

#### B. SEMI-SUPERVISED LEARNING

Semi-supervised learning is similar to supervised learning, but instead uses both labeled and unlabeled data. Labeled data is essential information that has meaningful tags so that the algorithm can understand the data, whilst unlabeled data lacks that information. By using this combination, model can learn to label the unlabelled data.

#### C. UNSUPERVISED LEARNING

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analyzing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organize that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organized.

As it assesses more data, its ability to make decisions on that data gradually improves and becomes more refined.

Under the umbrella of unsupervised learning, fall:

1. **Clustering:** Clustering involves grouping sets of similar data (based on defined criteria). It's useful for segmenting data into several groups and performing analysis on each data set to find patterns.
2. **Dimension reduction:** Dimension reduction reduces the number of variables being considered to find the exact information required.

#### D. REINFORCEMENT LEARNING

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By

defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

### 3.3 MACHINE LEARNING ALGORITHMS

We have also used various algorithms in our project to train the model and they are:

1. Random Forest Regressor
2. Gradient Boosting
3. Support Vector Machine
4. Decision Tree
5. XG Boost

#### 3.3.1 RANDOM FOREST REGRESSOR

Random Forest regressor is an ensemble learning regressor model which is used for better accuracy and mostly it is used on large datasets. Because of this, we are using this regressor technique for our model as it gives more accuracy compared to other algorithms. It creates a forest to evaluate results. Random Forest builds multiple decision trees by picking the 'K' number of knowledge points from the dataset and merges them to urge a more accurate and stable prediction. The training and testing time is more for the random forest as compared to linear and lasso regression. We got the maximum accuracy for random forest, so we finalize this algorithm [4].

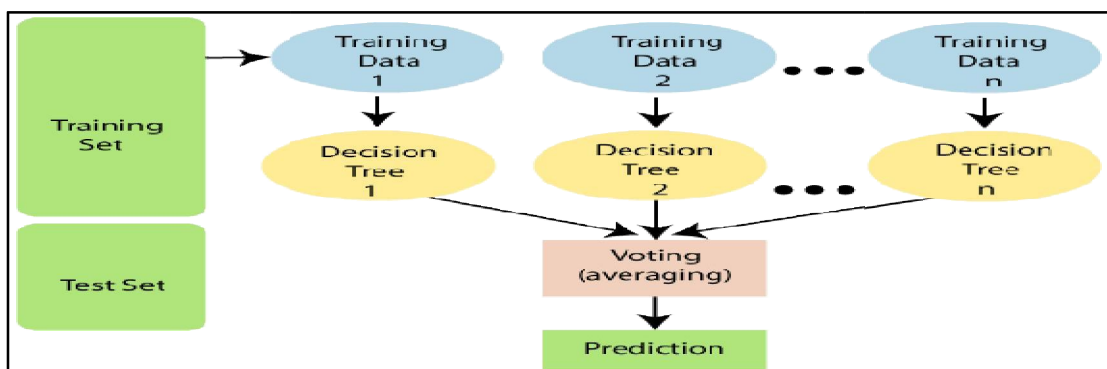


Figure 3.3: Random Forest Algorithm

### 3.3.2 GRADIENT BOOSTING

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm *helps us minimize bias error* of the model. Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

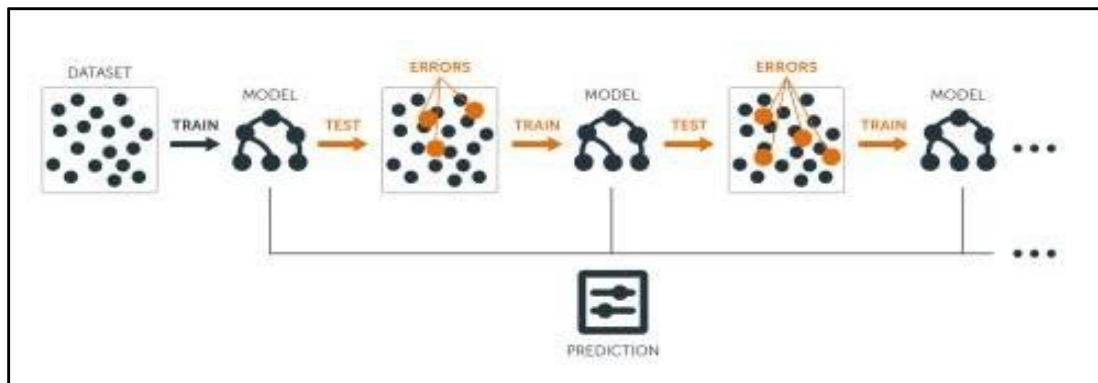


Figure 3.4: Gradient Boosting Algorithm

### 3.3.3 SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future [5]. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed a Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

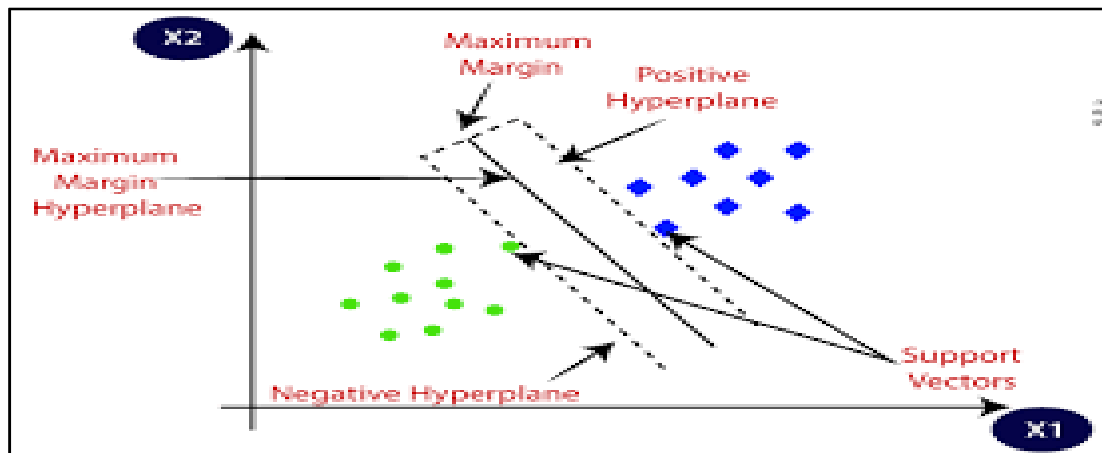


Figure 3.5: Support Vector Machine Algorithm

### 3.3.4 DECISION TREE

The decision tree regressor is a predictive machine learning model. This is the type of supervised machine learning. They make no assumptions about the errors. They are excellent at finding interactions that exist in just a part of the info . Decision tree regression observes features of an object and trains a model within the structure of a tree to predict data within the future to supply meaningful continuous output. Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.



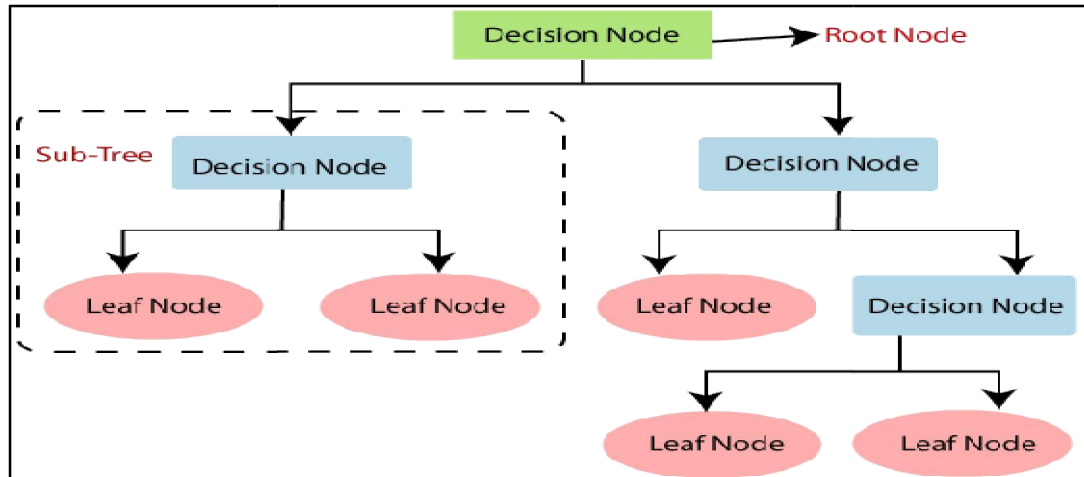


Figure 3.6: Decision Tree Algorithm

### 3.3.5 XG-BOOST ALGORITHM

XGBoost is an implementation of Gradient Boosted decision trees. XGBoost models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

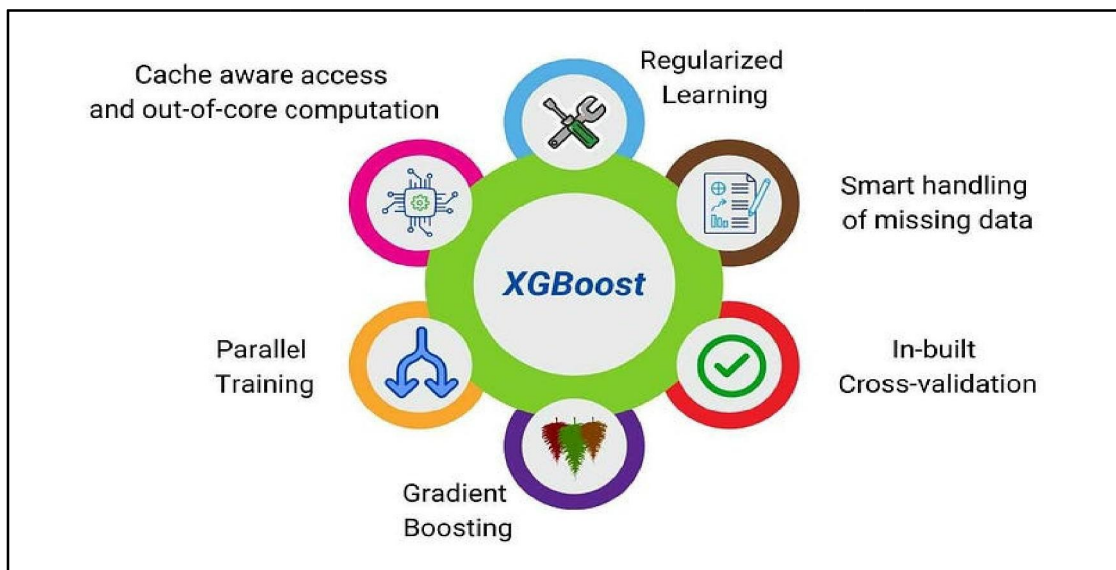


Figure 3.7: XG Boost Algorithm

**Chapter 4**  
**SYSTEM**  
**REQUIREMENTS**

## **3.2 SYSTEM REQUIREMENTS**

### **4.1 HARDWARE PLATFORM USED**

- Processor: Any Update Processor
- A Computer System

### **4.2 SOFTWARE PLATFORM USED**

- Windows 10 OS
- Jupyter Notebook
- Python3
- VScode
- Streamlit
- Python Packages
  - Numpy
  - Pandas
  - Matplotlib
  - Seaborn
  - Sklearn
  - Pickle

**Chapter 5**  
**SYSTEM**  
**ANALYSIS**

## **SYSTEM ANALYSIS**

### **5.1 PURPOSE**

The purpose of this project is to provide a tentative price forecasting of a product according to the configurations. Price prediction is done for any customer or administrator to predict the prices according to, whether is based on different configurations or different factors such as seasons, festivals, etc., The prediction model helps in decision making. The main purpose of this model is to ease the decision making for an individual.

### **5.2 PROJECT SCOPE**

1. It will Convert data into an appropriate form using various pre-processing techniques for the implementation of Machine Learning algorithms.
2. Observing and analyzing Dataset.
3. Predict and analyze prediction of different products prices.
4. This model can be implemented at various locations such as companies, stores.

### **5.3 ORGANIZATION OF PROJECT**

Chapter 1: It introduces the idea of our project along with the aim and motive of using this model. It states the statement of problem of our project and also gives the objectives of project.

Chapter 2: Literature survey provides the idea of previous work done on the idea similar to this project and also gives the basic knowledge about how the models are used in those models.

Chapter 3: After reviewing the work, the methodology which is to be used in this project is executed where it discusses the different algorithm models and their details.

Chapter 4: In this chapter, it gives the brief idea about the hardware, software and technologies which will be used in this model.

Chapter 5: System analysis refers to the purpose of the project the scope of this model and implementation scope of project.

Chapter 6: It implies to the proposed work where it refers to data gathering and performing different ML methods such as data pre-processing. Also shows how to deploy the model.

Chapter 7: In this chapter we get to know the implementation of different ML algorithms also here we measure the accuracies of those algorithms.

Chapter 8: Here we stated the conclusion derived from this project along with the future work we aim to be done.

Chapter 9: The information about the research work used in this project is given here.

**Chapter 6**  
**PROPOSED**  
**METHODOLOGY**

## PROPOSED METHODOLOGY

The methodology section reveals how our model work and how exploration is established.

### 6.1 PRODUCT DATASET:

Here is the description of dataset that has been used as input to perform regression using various algorithms. Here the dataset of laptop is used in category of electronics products for the prediction of prices which is collected form the site kaggle. The total no. of entries in dataset are more than 3k and the total no. of attributes are about 11.

The dataset taken is given by following figure:

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1303 entries, 0 to 1302
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            1303 non-null   int64
1   Company               1303 non-null   object
2   TypeName              1303 non-null   object
3   Inches                1303 non-null   float64
4   ScreenResolution      1303 non-null   object
5   Cpu                   1303 non-null   object
6   Ram                   1303 non-null   object
7   Memory                1303 non-null   object
8   Gpu                   1303 non-null   object
9   OpSys                 1303 non-null   object
10  Weight                1303 non-null   object
11  Price                 1303 non-null   float64
dtypes: float64(2), int64(1), object(9)
memory usage: 122.3+ KB
```

Figure 6.1: Product Dataset

### 6.2 PROPOSED METHODOLOGY FLOWCHART:

Here is the brief description about the flow of proposed methodology.

The methodology we employed in this proposed study is one that is frequently used in machine learning project work. Python programming language, Streamlit, Visual Studio Code, Jupyter, and Anaconda are used to implement the entire system. The following figure shows how to choose the method to utilise in the model. To forecast the price of the goods, Streamlit is being employed. With Streamlit, a free and open-source platform, users can produce beautiful machine learning and data science models and share them with others. This Python-based library was created in collaboration with ml engineers. While supervised learning develops a model that uses well-known input and output data to predict future results, unsupervised learning uncovers hidden patterns or internal structures in input data. if we need to train the data for making predictions or forecasting such as future values of a continuous variable.



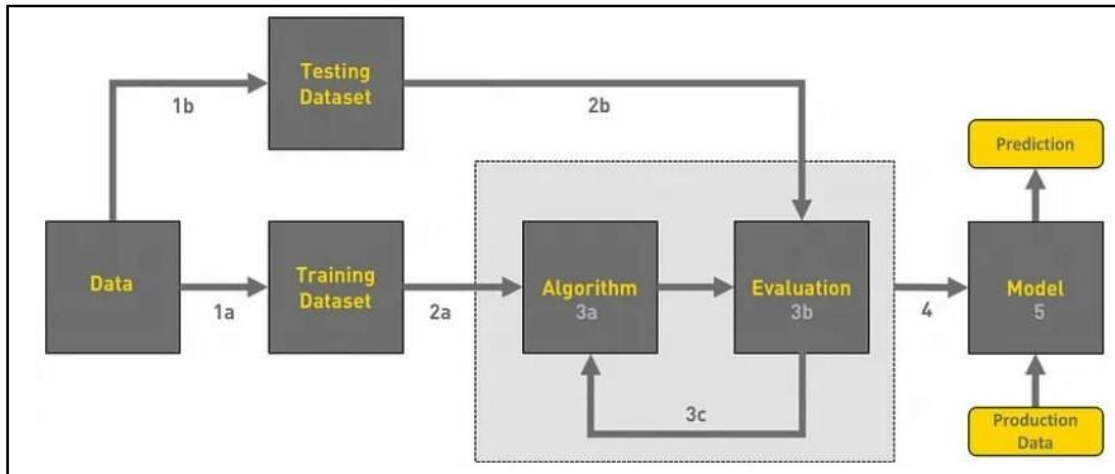


Figure 6.2: Proposed Methodology Flowchart

### 6.3 PROPOSED WORK

The model is divided into following phases i.e., Data Gathering, Data Pre-processing, Data training and testing, Data Evaluation, Exporting model. Discussed further down:

#### 6.3.1 Data Gathering

Data can be gathered from a variety of sources, including files, databases, sensors, and many other types of data sources. However, the data collected cannot be used directly for the analysis process because there may be a significant amount of missing data, extremely large values, unorganized text data, or noisy data. The process of data collection depends on the type of project we want to make. Kaggle or GitHub are some repositories that are used to collect the dataset for ML model building.

Here we have the dataset of laptop as product. Here historical data of laptop is used for the predictions. In this process we are using total 19 different company's laptops. There are total 1303 columns.

```
df.head()
```

Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price	
0	0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.6832
1	1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	47895.5232
2	2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.0000
3	3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.3360
4	4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.8080

Figure 6.3: First Five Rows of Dataset

Working on a dataset starts in our Jupyter Notebook. The first step is to import the libraries and load data. After that we took a basic understanding of data like its shape, sample, is there are any NULL values present in the dataset. Understanding the data is an important step for prediction or any machine learning project. The distribution of the target variable is skewed and it is obvious that commodities with low prices are sold and purchased more than the branded ones.

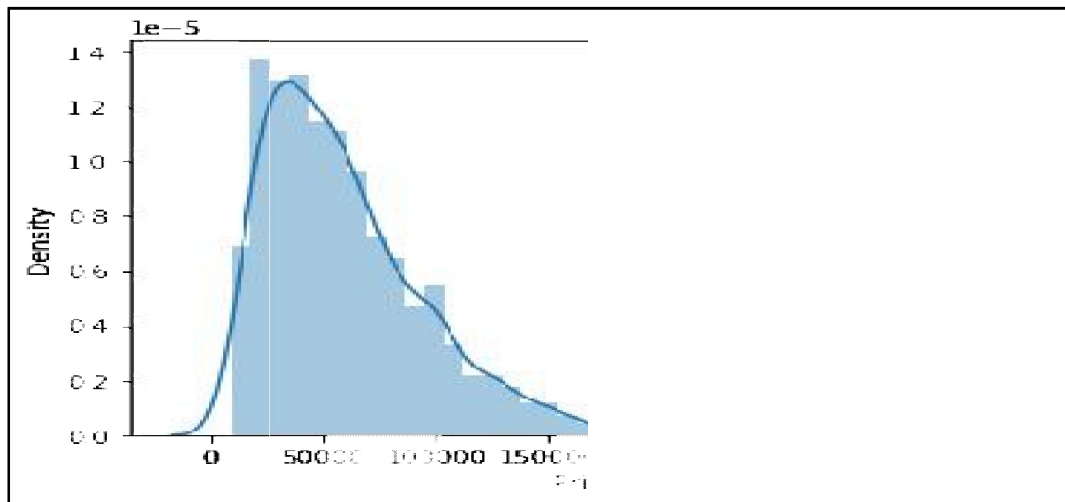


Figure 6.4: Distribution of target variable

### 6.3.2 DATA PREPROCESSING

Cleaning raw data, or transforming data that has been gathered from the real world into a clean data set, is known as data preprocessing. The goal of data pre-processing is converting raw data into clean that can be utilized to train a model. The (figure 6.3.3) shows the duplicate and null values in dataset using python functions.

```
df.duplicated().sum()
0

df.isnull().sum()
Unnamed: 0      0
Company         0
TypeName       0
Inches         0
ScreenResolution 0
Cpu            0
Ram           0
Memory        0
Gpu           0
OpSys         0
Weight        0
Price         0
dtype: int64
```

Figure 6.5: Null values in dataset

### Factors affecting prices

1) Brand Name: to understand how does brand name impacts the laptop price or what is the average price of each laptop brand? If you plot a count plot (frequency plot) of a company then the major categories present are Lenovo, Dell, HP, Asus, etc., Now, if the company relationship plot with price then we can observe that how price varies with different brands. Razer, Apple, LG, Microsoft, Google, MSI laptops are expensive, and others are in the budget range.

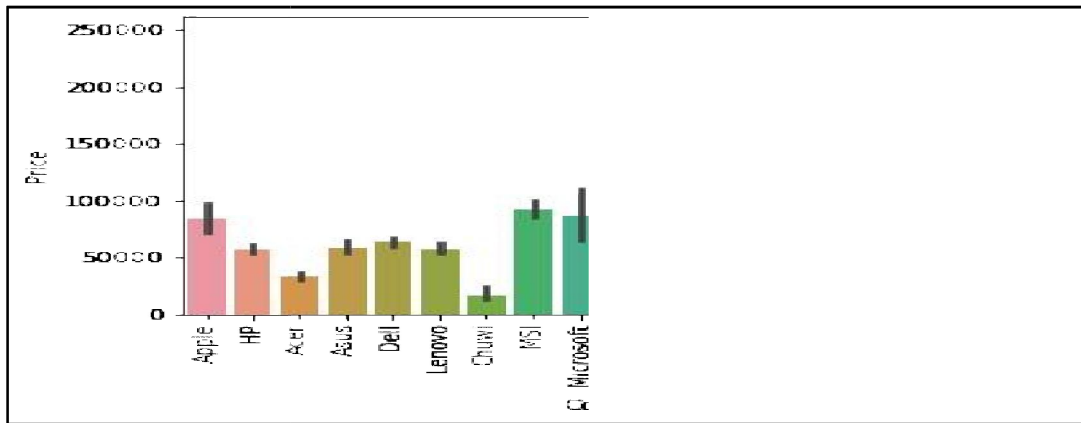


Figure 6.6: Company Name (Brands) with Prices

2) Types of laptop: Which type of laptop you are looking for like a gaming laptop, workstation, or notebook. As major people prefer notebook because it is under budget range and the same can be concluded from our data.

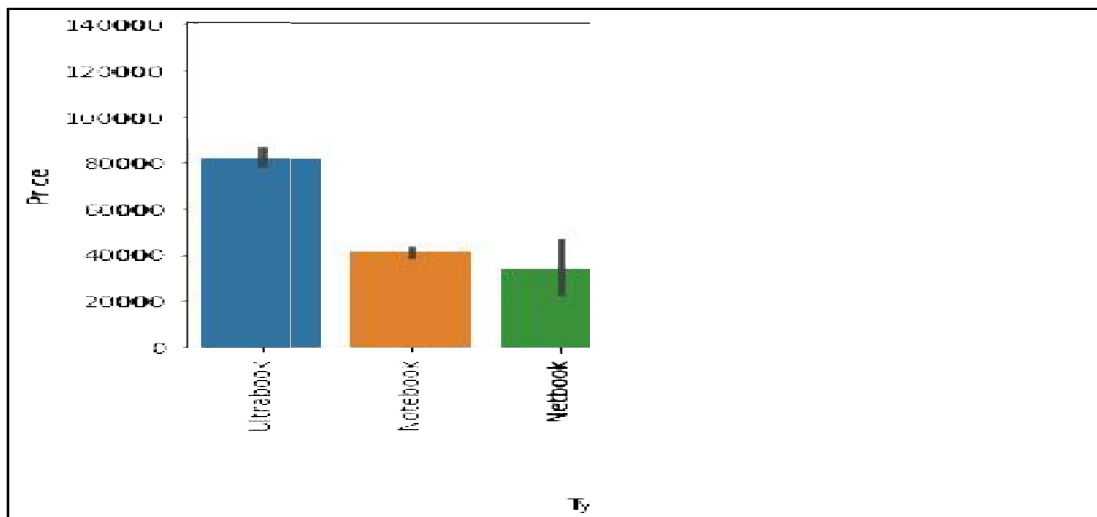


Figure 6.7: Types of laptops with prices

3) Screen Resolution: Screen resolution contains lots of information before any analysis firstly need to perform feature engineering over it (Feature engineering is a process to convert raw data to meaningful information. there are many methods that come under feature engineering like transformation, categorical encoding, etc.). If you observe unique values of the column then we can see that all value gives information related to the presence of an IPS panel, are a laptop touch screen or not, and the X-axis and Y-axis screen resolution. So extract column into 3 new columns in dataset.

- Extract Touch screen information

It is a binary variable so we can encode it as 0 and 1. one means the laptop is a touch screen and zero indicates not a touch screen.

```
data['Touchscreen'] = data['ScreenResolution'].apply(lambda x:1 if 'Touchscreen' in x else 0
```

```
sns.countplot(data['Touchscreen'])
```

```
sns.barplot(x=data['Touchscreen'],y=data['Price'])
```

If we plot the touch screen column against price then laptops with touch screens are expensive which is true in real life.

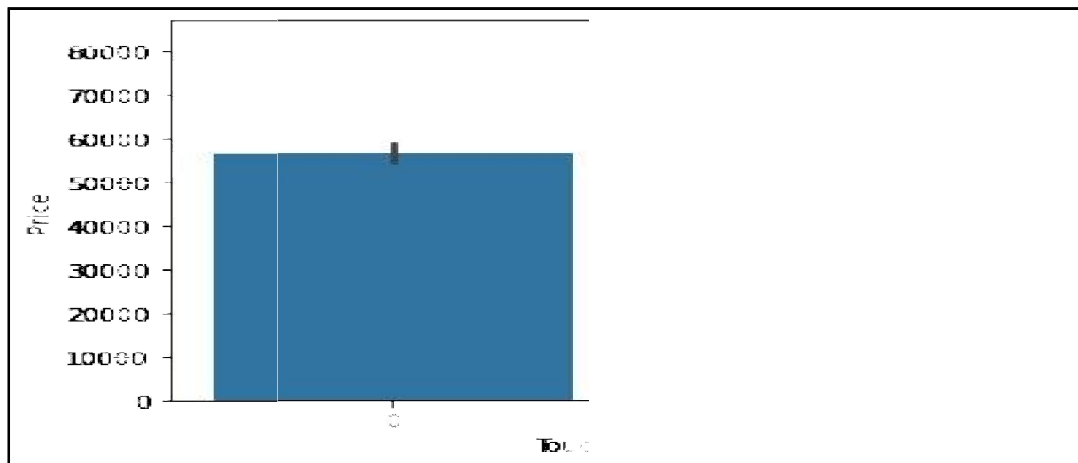


Figure 6.8: Touch-screen configuration

- Extract IPS Channel presence information

It is a binary variable and the code is the same we used above. The laptops with IPS channel are present less in our data but by observing relationship against the price of IPS channel laptops are high.

- Extract X-axis and Y-axis screen resolution dimensions

Now both the dimension are present at end of a string and separated with a cross sign. So first we will split the string with space and access the last string from the list. then split the string with a cross sign and access the zero and first index for X and Y-axis dimensions.

- Replacing inches, X and Y resolution to PPI

If you find the correlation of columns with price using the **corr** method then we can see that inches do not have a strong correlation but X and Y-axis resolution have a very strong resolution so we can take advantage of it and convert these three columns to a single column that is known as Pixel per inches(PPI). In the end, our goal is to improve the performance by having fewer features.

```
data.corr()['Price'].sort_values(ascending=False)
Price          1.000000
Ram            0.743007
X_res         0.556529
Y_res         0.552809
ppi           0.473487
Ips           0.252208
Weight        0.210370
Touchscreen   0.191226
Inches        0.068197
Name: Price, dtype: float64
```

Figure 6.9: Correlation price with size

So now we can drop the extra columns which are not of use. At this point, we have started keeping the important columns in our dataset.

4) CPU column: The CPU column then it also contains lots of information. If you again use a unique function or value counts function on the CPU column then we have 118 different categories. The information it gives is about preprocessors in laptops and speed.

To extract the preprocessor we need to extract the first three words from the string. we are having an Intel preprocessor and AMD preprocessor so we are keeping 5 categories in our dataset as i3, i5, i7, other intel processors, and AM D processors.

- How does the price vary with processors?

we can again use our bar plot property to answer this question. And as obvious the price of i7 processor is high, then of i5 processor, i3 and AMD processor lies at the almost the same range. Hence price will depend on the preprocessor.

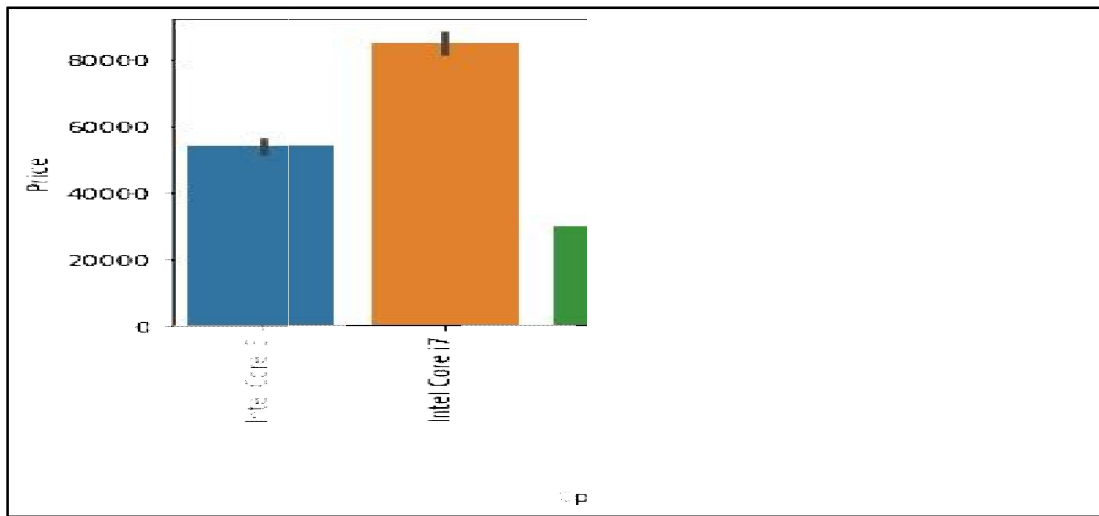


Figure 6.10: Variation in price with processor

5) Price with Ram: Again Bivariate analysis of price with Ram. If you observe the plot then Price is having a very strong positive correlation with Ram or you can say a linear relationship.

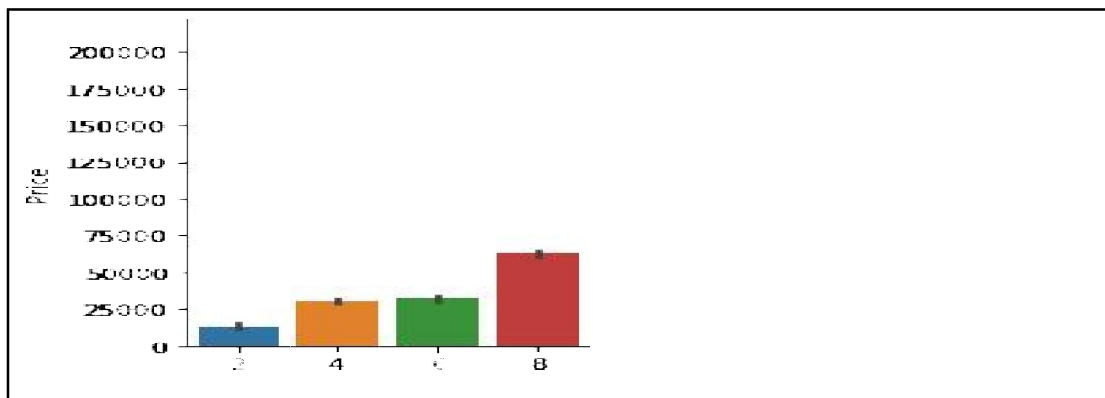


Figure 6.11: Variation in price with RAM

6) Memory Column: Memory column is again a noisy column that gives an understanding of hard drives. many laptops come with HHD and SSD both, as well in some there is an external slot present to insert after purchase. This column can disturb your analysis if not feature engineer it properly. So If the value counts use on a column then we are having 4 different categories of memory as HHD, SSD, Flash storage, and hybrid.

First, we have cleaned the memory column and then made 4 new columns which are a binary column where each column contains 1 and 0 indicate that amount four is present and which is not present. Any laptop has a single type of memory or a combination of two. so in the first column, it consists of the first memory size and if the second slot is present in the laptop then the second column contains it else we fill the null values with zero. After that in a particular column, we have multiplied the values by their binary value. It means that if in any laptop particular memory is present then it contains binary value as one and the first value will be multiplied by it, and same with the second combination. For the laptop which does have a second slot, the value will be zero multiplied by zero is zero.

7) GPU Variable: GPU(Graphical Processing Unit) has many categories in data. We are having which brand graphic card is there on a laptop. we are not having how many capacities like (6Gb, 12 Gb) graphic card is present. so we will simply extract the name of the brand.

8) OS Column: There are many categories of operating systems. we will keep all windows categories in one, Mac in one, and remaining in others. This is a simple and most used feature engineering method.

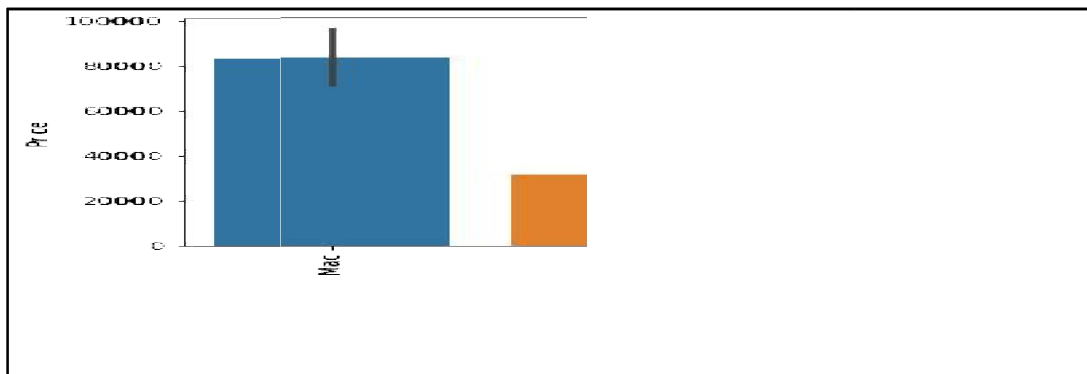


Figure 6.12: OS Column

- Log-Normal Transform : we saw the distribution of the target variable above which was right-skewed. By transforming it to normal distribution performance of the algorithm will increase. we take the log of values that transform to the normal distribution which you can observe below. So while separating dependent and independent variables we will take a log of price, and in displaying the result perform exponent of it.

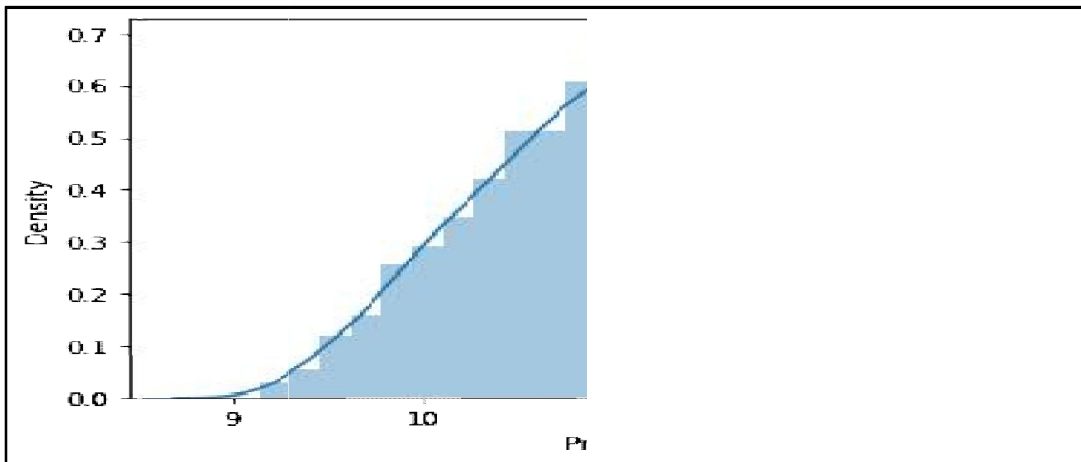


Figure 6.13: Log-Normal Transform

- **Machine Learning Modeling for Price Prediction**

Now we have prepared our data and hold a better understanding of the dataset. so let's get started with Machine learning modeling and find the best algorithm with the best hyperparameters to achieve maximum accuracy.

### 6.3.3 IMPORTING LIBRARIES

```
from sklearn.model_selection import train_test_split
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder
from sklearn.metrics import r2_score, mean_absolute_error
from sklearn.linear_model import LinearRegression,
from sklearn.tree import DecisionTreeRegressor
```



```
from sklearn.ensemble
```

```
import Random Forest Regressor, Gradient Boosting Regressor
```

```
from sklearn.svm import SVR
```

```
from xgboost import XGBRegressor
```

### 6.3.4 SPLITTING DATASET FOR TRAINING AND TESTING

Splitting the dataset is the next step in data preprocessing in machine learning. Every dataset for Machine Learning model must be split into two separate sets – training set and test set.

- **TRAINING DATA:** Training data is the data you use to train a machine learning algorithm or model to accurately predict a particular outcome, or answer, that you want your model to predict.
- **TESTING DATA:** Once your machine learning model is built (with your training data), you need unseen data to test your model. This data is called testing data, and you can use it to evaluate the performance and progress of your algorithms' training and adjust or optimize it for improved results.

```
X.head()
```

	Company	TypeName	Ram	Weight	Touchscreen	Ips	ppi	Cpu_brand	HDD	SSD	Gpu_brand	os
0	Apple	Ultrabook	8	1.37	0	1	226.983005	Intel Core i5	0	128	Intel	Mac
1	Apple	Ultrabook	8	1.34	0	0	127.677940	Intel Core i5	0	0	Intel	Mac
2	HP	Notebook	8	1.86	0	0	141.211998	Intel Core i5	0	256	Intel	Others/No OS/Linux
3	Apple	Ultrabook	16	1.83	0	1	220.534624	Intel Core i7	0	512	AMD	Mac
4	Apple	Ultrabook	8	1.37	0	1	226.983005	Intel Core i5	0	256	Intel	Mac

Figure 6.14: Split dataset

**Chapter 7**  
**IMPLEMENTATION**

## IMPLEMENTATION

### 7.1 IMPLEMENT PIPELINE FOR TESTING AND TRAINING

Let's implement a pipeline to streamline the training and testing process. First, we use a column transformer to encode categorical variables which is step one. After that, we create an object of our algorithm and pass both steps to the pipeline. using pipeline objects we predict the score on new data and display the accuracy.

**Linear Regression**

```
In [76]: step1 = ColumnTransformer(transformers=[
        ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,7,10,11])
        ], remainder='passthrough')

        step2 = LinearRegression()

        pipe = Pipeline([
            ('step1', step1),
            ('step2', step2)
        ])

        pipe.fit(X_train, y_train)

        y_pred = pipe.predict(X_test)

        print('R2 score', r2_score(y_test, y_pred))
        print('MAE', mean_absolute_error(y_test, y_pred))

R2 score 0.80732774484187
MAE 0.2101782797642877
```

Figure 7.1: Implement pipeline

In the first step for categorical encoding, we passed the index of columns to encode, and pass-through means pass the other numeric columns as it is. The best accuracy we got is with all-time catboost. But here we can use this code again by changing the algorithm and its parameters.

After this we implemented the different regression algorithms on dataset and find the accuracy with the help of statistical parameters like  $R^2$  and MSE, MAE score as follows:

- 1) Random forest algorithm
- 2) Gradient Boosting
- 3) Support Vector Machine
- 4) Decision Tree

## 5) XG Boost

## 1) RANDOM FOREST

```
In [82]: step1 = ColumnTransformer(transformers=[
        ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,7,10,11])
        ], remainder='passthrough')

        step2 = RandomForestRegressor(n_estimators=100,
                                     random_state=3,
                                     max_samples=0.5,
                                     max_features=0.75,
                                     max_depth=15)

        pipe = Pipeline([
            ('step1', step1),
            ('step2', step2)
        ])

        pipe.fit(X_train, y_train)

        y_pred = pipe.predict(X_test)

        print('R2 score', r2_score(y_test, y_pred))
        print('MAE', mean_absolute_error(y_test, y_pred))

        R2 score 0.8873402378382488
        MAE 0.15860130110457718
```

Figure 7.2: Random Forest Accuracy

## 2) GRADIENT BOOST

```
In [85]: step1 = ColumnTransformer(transformers=[
        ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,7,10,11])
        ], remainder='passthrough')

        step2 = GradientBoostingRegressor(n_estimators=500)

        pipe = Pipeline([
            ('step1', step1),
            ('step2', step2)
        ])

        pipe.fit(X_train, y_train)

        y_pred = pipe.predict(X_test)

        print('R2 score', r2_score(y_test, y_pred))
        print('MAE', mean_absolute_error(y_test, y_pred))

        R2 score 0.8825060164387656
        MAE 0.15903203631251014
```

Figure 7.3: Gradient Boost Algorithm Accuracy

### 3) SUPPORT VECTOR MACHINE

```
In [81]: step1 = ColumnTransformer(transformers=[
        ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,7,10,11])
        ], remainder='passthrough')

step2 = SVR(kernel='rbf', C=10000, epsilon=0.1)

pipe = Pipeline([
        ('step1', step1),
        ('step2', step2)
    ])

pipe.fit(X_train, y_train)

y_pred = pipe.predict(X_test)

print('R2 score', r2_score(y_test, y_pred))
print('MAE', mean_absolute_error(y_test, y_pred))

R2 score 0.8083180902289917
MAE 0.2023905942719158
```

Figure 7.4: Support Vector Machine Accuracy

### 4) DECISION TREE ALGORITHM

```
In [80]: step1 = ColumnTransformer(transformers=[
        ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,7,10,11])
        ], remainder='passthrough')

step2 = DecisionTreeRegressor(max_depth=8, min_samples_split=2,
                               min_samples_leaf=1)

pipe = Pipeline([
        ('step1', step1),
        ('step2', step2)
    ])

pipe.fit(X_train, y_train)

y_pred = pipe.predict(X_test)

print('R2 score', r2_score(y_test, y_pred))
print('MAE', mean_absolute_error(y_test, y_pred))

R2 score 0.8335652948593757
MAE 0.18564388824434946
```

Figure 7.5: Decision Tree Algorithm Accuracy

## 5) XG BOOST ALGORITHM

```
In [86]: step1 = ColumnTransformer(transformers=[
        ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,7,10,11])
        ], remainder='passthrough')

        step2 = XGBRegressor(n_estimators=45, max_depth=5, learning_rate=0.5)

        pipe = Pipeline([
            ('step1', step1),
            ('step2', step2)
        ])

        pipe.fit(X_train, y_train)

        y_pred = pipe.predict(X_test)

        print('R2 score', r2_score(y_test, y_pred))
        print('MAE', mean_absolute_error(y_test, y_pred))

        R2 score 0.8811773435850243
        MAE 0.16496203512600974
```

Figure 7.6: XG Boost Algorithm Accuracy

## 7.2 EXPORTING AND DEPLOY THE MODEL

One of the aim of our project is to create a website or application for an individual that will help the individual to select the product they want and predict the prices of that product according to their configurations. Here after calculating the accuracies of algorithm here we code for creating web application using streamlit [Streamlit is a python function that help individual with no prior knowledge of HTML, CSS, JAVASCRIPT]. Here we have done with modelling, we will save the pipeline object for the development of the project website, we will also export the data frame which will be required to create dropdowns in the website.

- STREAMLIT

Streamlit is an open-source web framework written in Python. It is the fastest way to create data apps and it is widely used by data science practitioners to deploy machine learning models. To work with this it is not important to have any knowledge of frontend languages. Streamlit contains a wide variety of functionalities, and an in-built function to meet your requirement. It provides you with a plot map, flowcharts, slider, selection box, input field, the concept of caching, etc. install streamlit using the below pip command.

pip install streamlit

create a file named app.py in the same working directory where we will write code for streamlit.

```
Users > ShreeS > Desktop > GAVATRI RAGHUVANSHI > final code > finalcode.py
from imaplib import _Authenticator
import streamlit_authenticator as stauth
import streamlit as st

# WE USED PANDA TO READ CSV FILE AND CREATE DAATAFRAME
import pandas as pd
import csv
import random

# IMPORT LINEAR REGRESSION MODEL FROM SLEARN
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor

# THIS LIBRARY IS USED TO SHOW GRAPH ON WEB BROWSER
import plotly.express as px
import yaml
from yaml.loader import SafeLoader
with open('C:\\Users\\Shrees\\Desktop\\GAVATRI RAGHUVANSHI\\Final code\\config.yaml') as file:
    config = yaml.load(file, Loader=SafeLoader)
hashed_passwords = stauth.Hasher(['abc', 'def']).generate()
# st.write(hashed_passwords)

authenticator = stauth.Authenticate(
    config['credentials'],
    config['cookie']['name'],
    config['cookie']['key'],
    config['cookie']['expiry_days'],
    config['preauthorized']
)
name, authentication_status, username = authenticator.login('Login', 'main')
```

Figure 7.7: Front Login page code

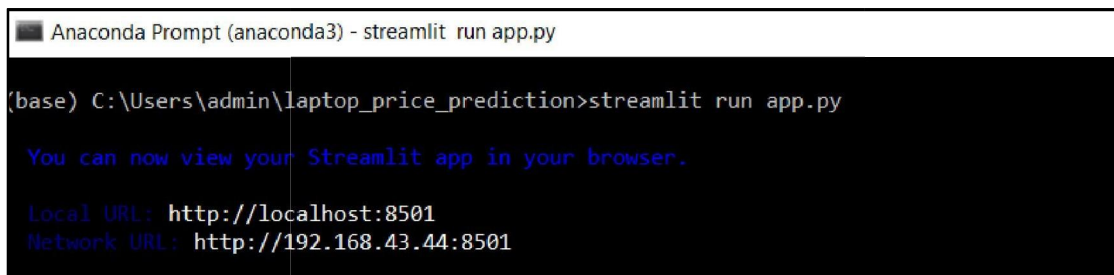
```
# CHECK AUTHNTICATION IS TRUE OR NOT

if st.session_state["authentication_status"]:
    authenticator.logout('Logout', 'main')
```

Figure 7.8: Authenticating page

First we load the data frame and model that we have saved. After that, we create an HTML form of each field based on training data columns to take input from users. In categorical columns, we provide the first parameter as input field name and second as select options which is nothing but the unique categories in the dataset. In the numerical field, we provide users with an increase or decrease in the value.

After that, we created the prediction button, and whenever it is triggered it will encode some variable and prepare a two-dimension list of inputs and pass it to the model to get the prediction that we display on the screen. Take the exponential of predicted output because we have done a log of the output variable.



```
Anaconda Prompt (anaconda3) - streamlit run app.py

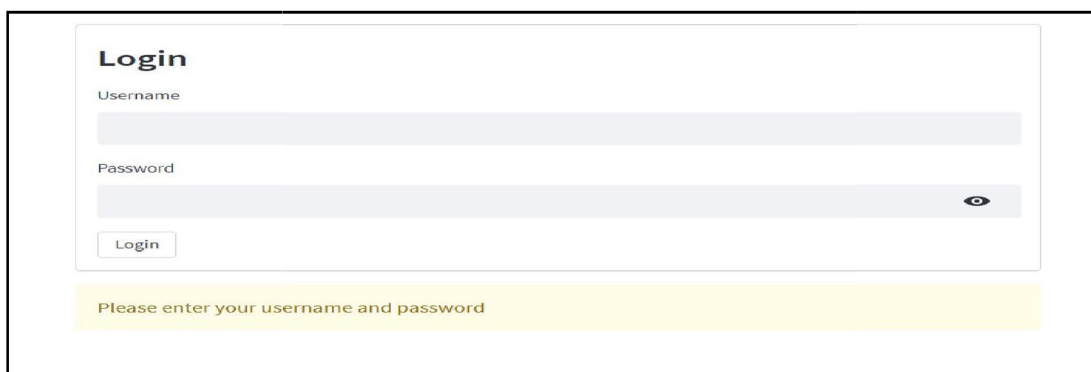
(base) C:\Users\admin\laptop_price_prediction>streamlit run app.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.43.44:8501
```

Figure 7.9: Command for URL

Now when you run the app file using the above command you will get two URL and it will automatically open the web application in your default browser or copy the URL and open it. the application will look something like the below figure.



**Login**

Username

Password

Please enter your username and password

Figure 7.10: Front Login page



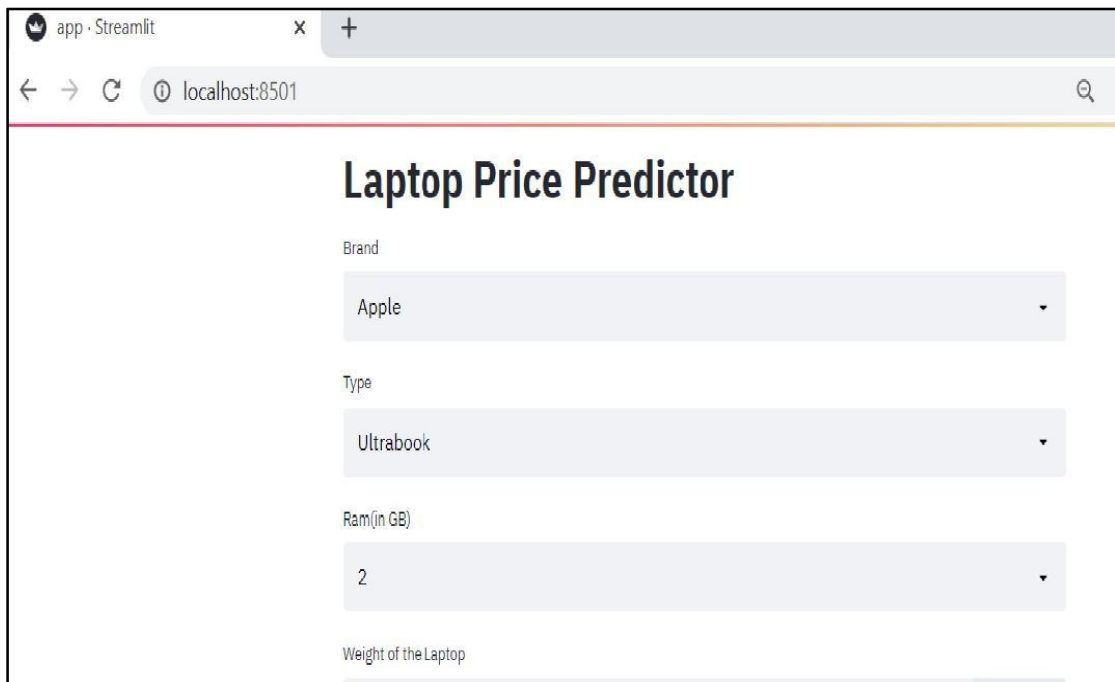


Figure 7.11: Product predictor page

Enter some data in each field and click on predict button to generate prediction. Here we get the predicted price of product according to the configuration we have enter as follow:

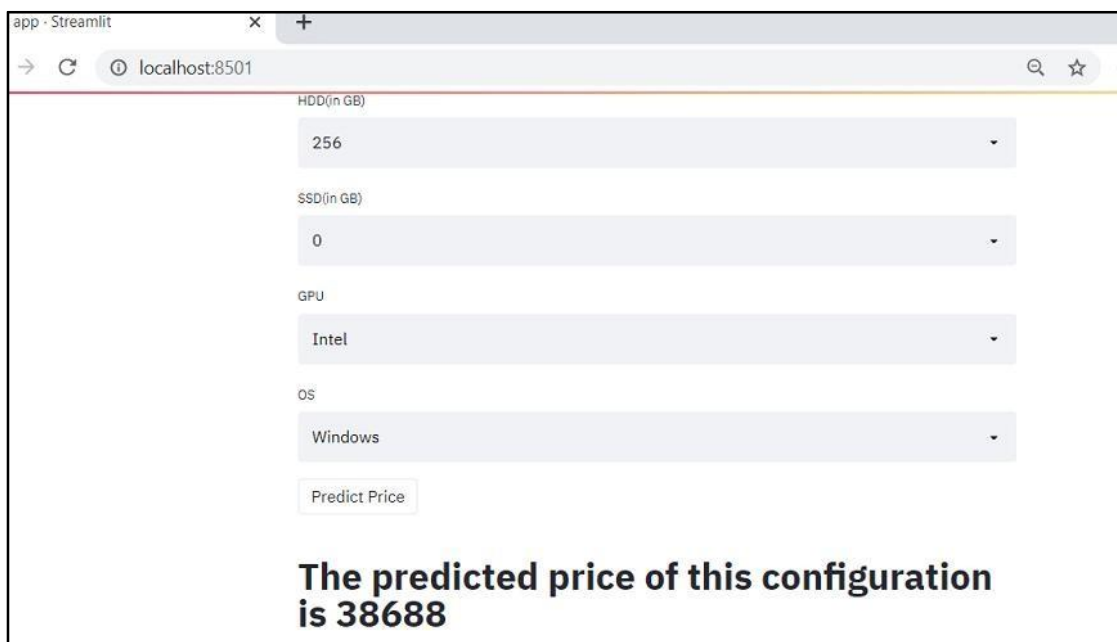


Figure 7.12: Predicted price of product

**Chapter 8**  
**CONCLUSION**

## **8.1 CONCLUSION**

- In this model, the regression models that are Linear Regression and Random Forest Regression have been used for the product price predictions.
- During the study it is observed that the linear and random forest regression show little fluctuations while changing the combinations in dataset. Here a python library streamlit is used for the frontend.
- While studying the accuracy of linear regression and random forest regression is also calculated. Hence we can conclude that the prediction algorithms i.e., Linear Regression and Random Forest Regression gives better results in the advanced forecasting of product prices.

## **8.2 FUTURE WORK**

- 1) As a part of future work, we aim to build the URL so that the user can access it automatically.
- 2) In future we aim to bind the Machine Learning model with different websites on large scale and also to upload a large dataset of different kinds of products.

**Chapter 9**  
**REFERENCES**

## REFERENCES

- [1] Julakha Jahan Jul, M. M. Imran Molla, Bifta Sama Bari, Mamunar Rashid, “Flat Price Prediction Using Linear and Random Forest Regression Based on Machine Learning Techniques”, 2020. (URL: <https://www.researchgate.net/publication/342793117>)
- [2] Yige Wang, “House-price Prediction Based on OLS linear Regression and Random Forest”.
- [3] Xinshu Li, “Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model”, 2022 Computational Intelligence and Neuroscience Volume 2022, Article ID 3590224.
- [4] UjjawalSonkambale, Suraj Kudale, Ojas Pawar, Tushar Salunke, “use car price prediction”, 2022 International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307 (www.ijsrceit.com) doi :<https://doi.org/10.32628/IJSRCSEIT>.
- [5] Subba Rao Polamuri, K.Srinivasi, A.Krishna Mohan, “Stock Market Prices Prediction Using Random Forest and Extra Tree Regression”, 2019 International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.
- [6] Mohamed Ali Mohamed, Ibrahim Mahmoud El-Henawy and Ahmad Salah, “Price Prediction of Seasonal Items Using Machine Learning and Statistical Methods”, 2021 Tech Science Press, doi:10.32604/cmc.2022.020782.
- [7] Rohit Joshi, Rohan Gupte, Palanisamy Saravanan, “A Random Forest Approach For Predicting Online Buying Behaviour Of Indian Customers”, 2018 Scientific Research Publishing, Theoretical Economics Letters, 2018, 8, 448-475(URL: <http://www.scirp.org/journal/tel>).
- [8] Ameena Sherin. V, Ms. Ankitha Philip, “House Price Prediction Using Linear Regression, Random Forest and Decision Tree Algorithms”, Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022 Vol.4, Issue.1.

[9] Mr. B. Saireddy, A. Vamshikrishna, G. Abhilash, D. Vinith Srinivas, "Car Price Prediction Using Machine Learning", 2022 International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:02/February-2022.

[10] Rola M. Elbakly, Magda M. Madbouly, and Shawkat K. Guirguis, "A Hybrid Approach for Product Price Prediction", 2022 European Journal of Engineering and Technology Research ISSN: 2736-576X DOI: <http://dx.doi.org/10.24018/ejeng.2022.7.5.2883>.

**DISSEMINATION OF  
WORK**

# Advanced Forecasting of Demandable Products Prices using Machine Learning Algorithm

Gayatri Raghuwanshi<sup>1</sup>, Gayatri Zamare<sup>2</sup>, Rupesh Apar<sup>3</sup>, Rupesh Dabhade<sup>4</sup>, Vaibhav Wankhade<sup>5</sup>  
Students, Department of Computer Science & Engineering<sup>1,2,3,4,5</sup>  
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

**Abstract:** *Knowing which items would be the most affordable is crucial for the organization. At this stage, categorization and prediction issues, such as price prediction, have been resolved using machine learning technology. This project seeks to produce timely and accurate price forecasts to assist the organisation in switching between neighboring markets to assist the organisation in switching between various neighbouring markets in order to sell their goods and obtain competitive rates. The data can be used by the company to make decisions regarding the timing of marketing. The machine Learning technique allows for predicting the number of products/services to be purchased during a defined period. Demand forecasting is used in which first raw data is collected from the market, then according to the data the product prices are forecasted. This model is a catch-all phrase for the shopping process that establishes product prices in accordance with the level of supplier competition, the hour of the day, and the weather. This model will help to forecast the prices of products according to their historical data. At an organizational level, forecasts of product prices are an essential input to many decision-making activities in various functional areas such as operations, marketing, sales, production, and finance.*

**Keywords:** Product prices forecasting, Machine Learning, Linear Regression, Lasso Regression, XG Boost Algorithm, Gradient Boosting Algorithm, Random Forest Regressor, Streamlit, SkLearn

## I. INTRODUCTION

Effective pricing forecasting assists organizations in anticipating price increases or cuts that may impact customer demand. Previous year's data on different products are being collected and we will predict the prices of products so that we will be able to make good marketing strategies. Using machine learning the system can predict what will be the price of a particular product today or after a certain day. Due to its striking advantages over conventional methods, machine learning techniques have recently become frequently used for price prediction. ML algorithms create models using training and test data, and then use these models to make predictions [1]. A prediction algorithm will be used to predict prices. Price and arrival data information strengthens the organization's bargaining position and increases the competitiveness among dealers. The organization can switch between neighboring markets more easily when price information is provided. The information can be used by the organization to make marketing timing decisions. The majority of machine learning (ML) algorithms that were developed within the context of data science have dominated in recent years. It has previously been used to predict time series in the financial and economic sectors. Numerous empirical studies have demonstrated that machine learning methods are more effective than time series models at forecasting various financial asset values.

## II. LITERATURE SURVEY

Julakha Jahan JuiIn et al., [1] research focuses on, two machine-learning regression-based methods for predicting flat pricing—linear regression and random forest regression—was given. Data has been scraped from a number of real estate websites using the web scrapper (Data Toolbar) software. When developing the model, seven factors that can affect flat pricing were taken into consideration. Here, the data quality has been investigated using the histogram, residual charting, and ANOVA. The linear and random forest model has been created after preprocessing the dataset. MSE, RMSE, MAE, and MRE have all been computed in order to evaluate the performance of both techniques. The measured error rate has led to the conclusion that the random forest regression model performs well.



Yige Wang et al.,[2] research state that clearly more practical in terms of price prediction is the decision tree fitting effect using Random Forest, the order of variable importance as opposed to OLS when dealing with complicated and irregular data. Therefore, we advise choosing a random forest in these two scenarios—one in which there are many observations in the dataset and the other in which there are complex samples with noise.

Xinshu Li et al.,[3] focus on demand for commercial housing falls into three primary categories: speculative, investment, and owner-occupied. The investment need for self-employment is to purchase and lease commercial real estate to generate rental revenue. Hypothetical demand is bought.

Ujjawal Sonkambale et al.,[4] research machine learning techniques to predict the price of used cars based on historical data from the Kaggle and Car Dekho websites. To determine which predictions offer the best performance and accuracy, the predictions are compared and examined. Delay filters, delay lines, power amplifiers, coaxial resonators, and ceramics are index terms.

Subba Rao Polamuri et al.,[5] this paper focuses on using ML techniques to anticipate the behaviours tracking of the stock market sensex. It compares the accuracy of various models and selects an algorithm with high accuracy. The main aim is to apply innovative work to predict the behaviour tracking of the stock market Sensex.

Mohamed Ali Mohamed et al.,[6] This work presents a promising approach for predicting pricing for retail goods, specifically seasonal Christmas items. Machine learning-based models, such as random forest and ARIMA, are effective in predicting the prices of these items. The results demonstrate that the irregular forest model outperforms other models. The study recommends using the random forest and ARIMA models, building hybrid models, defining the problem as a time-series problem and incorporating date and time input characteristics into the suggested models.

Rohit Joshi et al.,[7] this research paper highlights the importance of understanding customer shopping preferences in the growing online retail industry in India. It gathered information from 124 Indian respondents spread across 18 states and constructed and verified Random Forest prediction models for a number of product categories. The results showed that the model had a high sensitivity for products like books and electronics, while having a low sensitivity for products like movies, sporting goods, and bags.

Ameena sherin et al.,[8] this paper states that Linear Regression, Decision Tree, and Random Forest are three unique algorithms, each of which is based on a different component of the data, that may be used to predict home values. This research study, "House Price Prediction," describes how to utilize these three algorithms to do so. The problem is that predicted prices vary depending on how accurate they are. To get around this problem, we calculate the average of the projections. As a result, it helps with mistake prevention. This average price is a fair representation of the worth of a house. Clients will be pleased with the approach since it may produce reliable results and reduce the possibility of error.

Liu et al.,[9] this research focus on comparing the performance of different machine learning models for prediction of housing prices. The authors compared different algorithm like Linear Regression, Decision Tree, and Random Forest Regression whereas the CNN Random Forest model gives wrong output as compared to the other models in terms of prediction accuracy, suggesting it could be a useful tool for property valuation. Future research should incorporate additional data beyond housing development to improve the model's accuracy.

### III. METHODOLOGY

The methodology we employed in this proposed study is one that is frequently used in machine learning project work. Python programming language, Streamlit, Visual Studio Code, Jupyter, and Anaconda are used to implement the entire system. Figure 1 shows how to choose the method to utilise in the model. To forecast the price of the goods, Streamlit is being employed. With Streamlit, a free and open-source platform, users can produce beautiful machine learning and data science models and share them with others. This Python-based library was created in collaboration with ml engineers. While supervised learning develops a model that uses well-known input and output data to predict future results, unsupervised learning uncovers hidden patterns or internal structures in input data. if we need to train the data for making predictions or forecasting such as future values of a continuous variable i.e., temperature or prices we can use supervised learning.

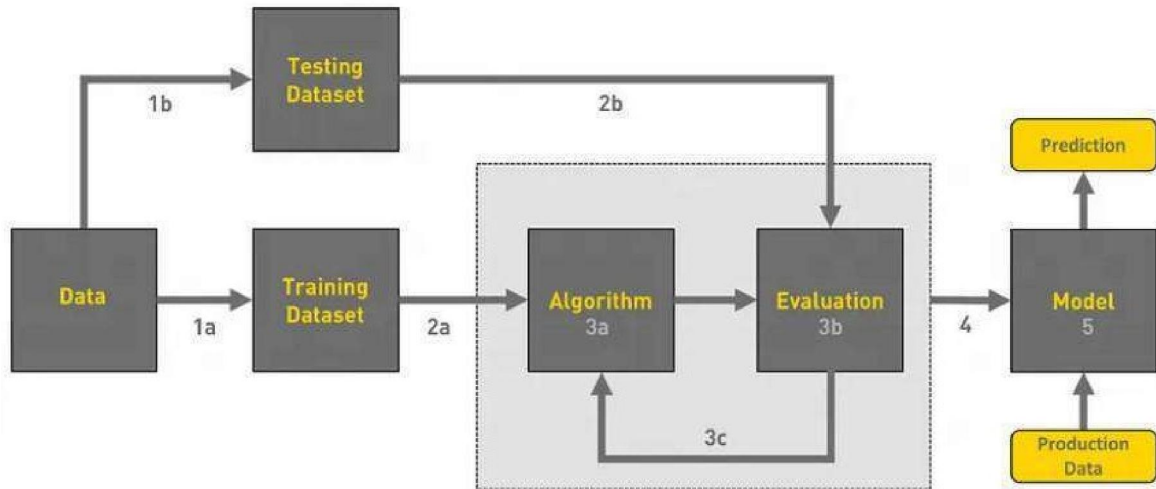


Figure 1 :Workflow of Machine Learning Model

### 3.1 Data Gathering

Data can be gathered from a variety of sources, including files, databases, sensors, and many other types of data sources. However, the data collected cannot be used directly for the analysis process because there may be a significant amount of missing data, extremely large values, unorganized text data, or noisy data. The process of data collection depends on the type of project we want to make. Kaggle or GitHub are some repositories that are used to collect the dataset for ML model building.

Here we have the dataset of laptop as product. Here historical data of laptop is used for the predictions. In this process we are using total 19 different company’s laptops. There are total 1303 columns.

```
df.head()
```

Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price	
0	0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.6832
1	1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 8000	macOS	1.34kg	47695.5232
2	2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.0000
3	3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.3360
4	4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.8080

Figure 2. Screenshot of first five rows in dataset

```
df.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1303 entries, 0 to 1302
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  ---                ---
0   Unnamed: 0            1303 non-null   int64
1   Company               1303 non-null   object
2   TypeName              1303 non-null   object
3   Inches               1303 non-null   float64
4   ScreenResolution      1303 non-null   object
5   Cpu                  1303 non-null   object
6   Ram                  1303 non-null   object
7   Memory               1303 non-null   object
8   Gpu                  1303 non-null   object
9   OpSys                1303 non-null   object
10  Weight               1303 non-null   object
11  Price                1303 non-null   float64
dtypes: float64(2), int64(1), object(9)
memory usage: 122.3+ KB
  
```

Figure 3. Screenshot of datatypes in dataset

### 3.2 Data Preprocessing

Cleaning raw data, or transforming data that has been gathered from the real world into a clean data set, is known as data preprocessing. The goal of data pre-processing is converting raw data into clean that can be utilized to train a model. The (figure 3.6) shows the duplicate and null values in dataset using python functions.

```
df.duplicated().sum()
0

df.isnull().sum()
Unnamed: 0      0
Company         0
TypeName        0
Inches          0
ScreenResolution 0
Cpu             0
Ram            0
Memory         0
Gpu            0
OpSys          0
Weight         0
Price          0
dtype: int64
```

Figure 4. Screenshot showing the duplicate/null values in dataset

#### IV. TRAINING AND TESTING DATASET

Following are the three sections that we initially divide the model into for training:-

- Training set: The training set refers to the data used to instruct the computer on handling is used to instruct the computer how to handle data. Machine learning employs algorithms to carry out the training phase. a set of data that is employed for learning, i.e., to match the classifier's parameters.
- Validation set: Cross-validation is frequently used in applied machine learning to evaluate a model's performance on untested data. The classification parameters are adjusted by synthesizing of classification are adjusted through the synthesis of unknown data from the training set.
- Test set: A collection of unobserved data that is only used to evaluate how well a fully expressed classifier performed.

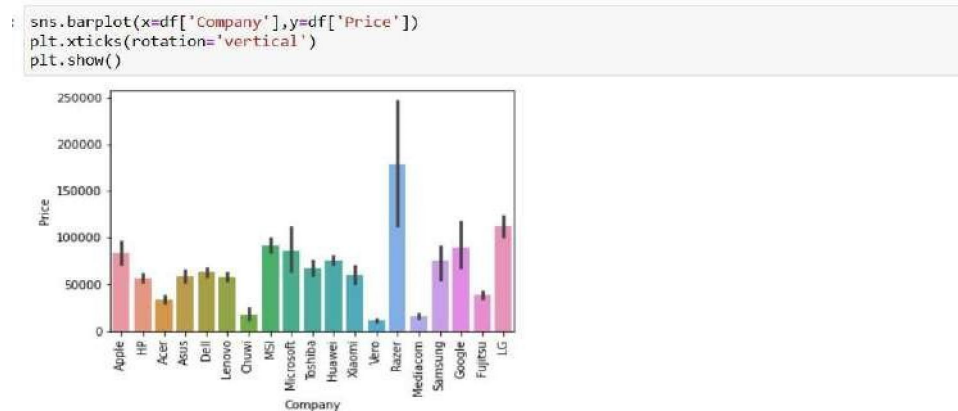


Figure 5. Barplot showing the company names along with price rate of laptop

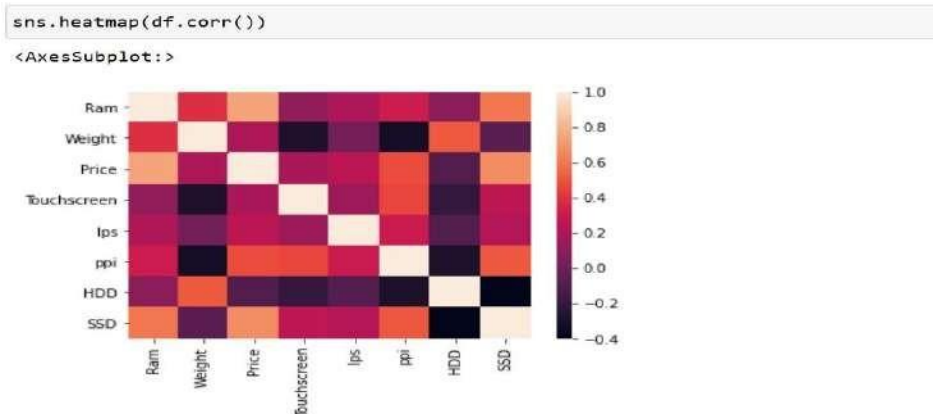


Figure 6. Heatmap for training data

## V. EVALUATION AND MODEL DEVELOPMENT

### 5.1 Choosing an Efficient Model

Here we have to choose an efficient algorithm for the model building as it is based on the amount and quality of the dataset we have used. There are mainly two algorithms that can be used for price predictions as Linear Regression and Random Forest regressor.

### 5.2 Linear Regression

It ranks among the most used machine learning regression algorithms. The output variables (future values) are predicted using a significant variable from the data set. If the labels are continuous, such as the number of planes departing from an airport each day, etc., the linear regression algorithm is utilized.

$$Y = b * x + c$$

is the formula for linear regression. 'Y' is the independent variable in the illustration above, whereas 'x' is the dependent variable. The slope of the line that gives us the output variables is referred to as "b" when you plot the linear regression, and its intercept is referred to as "c" [3]. The assumptions made by linear regression algorithms are that the relationship between the input and the output is linear. The first method was chosen because it is straightforward and takes relatively little time to train and test. Because of the feature vectors, the features were used directly without any feature mapping. For improved accuracy, we also apply regularization techniques [4].

### 5.3 Random Forest Regressor

A random forest is a meta-estimator that employs averaging to improve projected accuracy and decrease overfitting after fitting numerous classification decision trees to various dataset subsamples. The accuracy of the Random Forest Regression is high. It typically produces better results in prediction models. A data estimator for data about data is the rambling forest [7]. Numerous decision-makers were used for various subsamples of the provided data. The limit is exceeded. It increases predictability.

#### Algorithm:

Step 1: Select N chose records from the dataset.

Step 2: Create a decision tree based on N records.

Step 3a: Repeat steps 1 and 2 after selecting the number of trees from your algorithm.

Step 3b: In the case of a regression issue, each tree in the forest forecasts a value for Y (output) for a new record.

#### XG Boost Regressor

An excellent and efficient implementation of the gradient boosting approach is provided by the open-source Extreme Gradient Boosting (XGBoost) program. Effective gradient augmentation for regression-based predictive modeling is provided by the XGBoost.

#### Technology Used

- **Streamlit:** With Streamlit, a free and open-source platform, brilliant machine learning and data science web apps can be produced and distributed quickly. The use of a Python-based library is made.
- **Pandas:** Dataset manipulation is done using the Python package Pandas. It provides tools for data search, cleanup, analysis, and manipulation [8]. Pandas assist us in analysing huge data sets and coming to conclusions based on statistical concepts. Pandas can organize disorganized data sets, making them useful and readable.

## VI. CONCLUSION

In this research paper, we predict the prices of products from their historical data. In this model we used laptop as example for prediction of prices. Different regression models such as Linear Regression and Random Forest Regression Lasso Regression, etc., have been used for the product price predictions. During the study, it is observed that the accuracy is increased with decrement in MAE "Mean Absolute Error". Here the MAE count of Random forest

Regression is lower than that of other algorithms and  $R^2$  score of Random Forest regression, XGBoost, and Gradient Boost 0.8873, 8811, and 8823 resp. Hence we can conclude that the prediction algorithms i.e., Random Forest Regression, XGBoost Regressor gives better results in the advanced forecasting of product prices.

#### VII. FUTURE WORK

- As a part of future work, we aim to build the URL so that the user can access it automatically.
- In future we aim to bind the Machine Learning model with different websites on large scale and also to upload a large dataset of different kinds of products.

#### REFERENCES

- [1]. Julakha Jahan Jul, M. M. Imran Molla, Bifta Sama Bari, Mamunar Rashid, "Flat Price Prediction Using Linear and Random Forest Regression Based on Machine Learning Techniques", 2020. (URL:<https://www.researchgate.net/publication/342793117>)
- [2]. Yige Wang, "House-price Prediction Based on OLS linear Regression and Random Forest".
- [3]. Xinshu Li, "Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model", 2022 Computational Intelligence and Neuroscience Volume 2022, Article ID 3590224.
- [4]. UjjawalSonkambale, Suraj Kudale, Ojas Pawar, Tushar Salunke, "Use car price prediction", 2022 International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN: 2456-3307 ([www.ijsrcseit.com](http://www.ijsrcseit.com)) Doi: <https://doi.org/10.32628/IJSRCSEIT>.
- [5]. Subba Rao Polamuri, K. Srinivasi, A. Krishna Mohan, "Stock Market Prices Prediction Using Random Forest and Extra Tree Regression", 2019 International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.
- [6]. Mohamed Ali Mohamed, Ibrahim Mahmoud El-Henawy and Ahmad Salah, "Price Prediction of Seasonal Items Using Machine Learning and Statistical Methods", 2021 Tech Science Press, doi:10.32604/cmc.2022.020782.
- [7]. Rohit Joshi, Rohan Gupte, Palanisamy Saravanan, "A Random Forest Approach For Predicting Online Buying Behaviour of Indian Customers", 2018 Scientific Research Publishing, Theoretical Economics Letters, 2018, 8, 448-475(URL:<http://www.scirp.org/journal/tel>).
- [8]. Ameena Sherin. V, Ms Ankitha Philip, "House Price Prediction Using Linear Regression, Random Forest and Decision Tree Algorithms", Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022 Vol.4, Issue.1.
- [9]. Mr B. Saireddy, A. Vamshikrishna, G. Abhilash, D. Vinith Srinivas, "Car Price Prediction Using Machine Learning", 2022 International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:02/February-2022.
- [10]. Rola M. Elbakly, Magda M. Madbouly, and Shawkat K. Guirguis, "A Hybrid Approach for Product Price Prediction", 2022 European Journal of Engineering and Technology Research ISSN: 2736-576X DOI: <http://dx.doi.org/10.24018/ejeng.2022.7.5.2883>.

# International Journal of Advanced Research In Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal



IJARSCT

CERTIFICATE  
Of Publication

INTERNATIONAL STANDARD  
SERIAL NUMBER  
ISSN NO: 2581-9429

THIS IS TO CERTIFY THAT

Gayatri Raghuwanshi

Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

HAS PUBLISHED A REVIEW PAPER ENTITLED

Advanced Forecasting of Demandable Products Prices using Machine Learning Algorithm

In IJARSCT, Volume 3, Issue 7, April 2023

Certificate No: 042023-A2105

[www.ijarsct.co.in](http://www.ijarsct.co.in)



Crossref

DOI: 10.48175/IJARSCT-9511  
[www.doi.org](http://www.doi.org)

[www.crossref.org](http://www.crossref.org)



[www.sjifactor.com](http://www.sjifactor.com)



# International Journal of Advanced Research In Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal



IJARSCT

CERTIFICATE  
Of Publication

INTERNATIONAL STANDARD  
SERIAL NUMBER  
ISSN NO: 2581-9429

THIS IS TO CERTIFY THAT

Gayatri Zamare

Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

HAS PUBLISHED A REVIEW PAPER ENTITLED

Advanced Forecasting of Demandable Products Prices using Machine Learning Algorithm

In IJARSCT, Volume 3, Issue 7, April 2023

Certificate No: 042023-A2106

[www.ijarsct.co.in](http://www.ijarsct.co.in)



Crossref

DOI: 10.48175/IJARSCT-9511  
[www.doi.org](http://www.doi.org)

[www.crossref.org](http://www.crossref.org)



[www.sjifactor.com](http://www.sjifactor.com)



# International Journal of Advanced Research In Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal



IJARSCT

CERTIFICATE  
Of Publication

INTERNATIONAL STANDARD  
SERIAL NUMBER  
ISSN NO: 2581-9429

THIS IS TO CERTIFY THAT

Rupesh Apar

Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

HAS PUBLISHED A REVIEW PAPER ENTITLED

Advanced Forecasting of Demandable Products Prices using Machine Learning Algorithm

In IJARSCT, Volume 3, Issue 7, April 2023

Certificate No: 042023-A2107

[www.ijarsct.co.in](http://www.ijarsct.co.in)



DOI: 10.48175/IJARSCT-9511  
[www.doi.org](http://www.doi.org)

[www.crossref.org](http://www.crossref.org)



[www.sjifactor.com](http://www.sjifactor.com)





# International Journal of Advanced Research In Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal



IJARSCT

CERTIFICATE  
Of Publication

INTERNATIONAL STANDARD  
SERIAL NUMBER  
ISSN NO: 2581-9429

THIS IS TO CERTIFY THAT

Rupesh Dabhade

Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

HAS PUBLISHED A REVIEW PAPER ENTITLED

Advanced Forecasting of Demandable Products Prices using Machine Learning Algorithm

In IJARSCT, Volume 3, Issue 7, April 2023

Certificate No: 042023-A2108

[www.ijarsct.co.in](http://www.ijarsct.co.in)



DOI: 10.48175/IJARSCT-9511  
[www.doi.org](http://www.doi.org)

[www.crossref.org](http://www.crossref.org)



[www.sjifactor.com](http://www.sjifactor.com)



# International Journal of Advanced Research In Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal



IJARSCT

CERTIFICATE  
Of Publication

INTERNATIONAL STANDARD  
SERIAL NUMBER  
ISSN NO: 2581-9429

THIS IS TO CERTIFY THAT

Vaibhav Wankhade

Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

HAS PUBLISHED A REVIEW PAPER ENTITLED

Advanced Forecasting of Demandable Products Prices using Machine Learning Algorithm

In IJARSCT, Volume 3, Issue 7, April 2023

Certificate No: 042023-A2109

[www.ijarsct.co.in](http://www.ijarsct.co.in)



DOI: 10.48175/IJARSCT-9511  
[www.doi.org](http://www.doi.org)

[www.crossref.org](http://www.crossref.org)



[www.sjifactor.com](http://www.sjifactor.com)

