Comparative study on interpretable E-commerce Price Prediction

Md. Fahim Shahriar
Computer Science and Engineering
Brac University
md.fahim.shahriar@g.bracu.ac.bd

Fahim Faisal Rafi
Computer Science and Engineering
Brac University
fahim.faisal.rafi@g.bracu.ac.bd

***Abstract*—In recent years, e-commerce has exploded in popularity due to the speed and convenience of exchanging goods and global services. It makes an attempt to explain the principle behind e-commerce, business models designed specifically for e-commerce, and the merits of each as well as its inherent restrictions and limits. It comes to the conclusion that engaging in e-commerce affords a number of benefits to the various stakeholders. Researchers have conducted several experiments to make E-commerce sites more feasible and accessible to the customers using machine learning and deep learning. Similarly, we have also experimented and implemented several supervised learning algorithms to make price predictions. We also have made an comparative study on machine learning and deep learning models using XAI. Among the models, Decision Tree Classifier and Random Forest Classifier both gave the best accuracy score with 95.05% and Random Forest Classifier gave the best root of mean squared error with 7.64.**

***Index Terms*—E-commerce, XAI, ANN, RNN, kNN, Naive Bayes, Decision Tree, Random Forest Classifier, SVC, Linear Regression, Ridge Regression**

INTRODUCTION

The subject of artificial intelligence (AI) and computer science known as machine learning focuses on the use of data and algorithms to simulate how people learn, hence enhancing the system's accuracy. A core part of the rapidly expanding discipline of data science is machine learning. In data mining projects, statistical approaches are used to train algorithms in order to provide classifications or predictions and to discover valuable insights. Idealistically, the actions taken as a result of these insights impact important growth metrics in applications and businesses. In addition, certain machine learning [1] algorithms have extremely specific applications; yet, the three primary approaches are still in use today. Only supervised learning techniques and classifications were used in this research. Additionally, we have decided to complete this project with an E-commerce dataset. In our research, we mainly concentrated on predicting the Selling Price of different products on an E-commerce site based on their Brand name, MRP, and Discount; by calculating the Selling Price with various algorithms and the accuracy score of those corresponding algorithms, we evaluated which classifier performs best for this type of dataset. Then we processed, cleaned up, and deleted redundant data. Once the preprocessing phase was finished, we utilized 7 different supervised machine learning classifiers and 2 deep learning models to predict and achieve the highest level of accuracy and root of mean squared error. Our goal is to compare different types of supervised learning algorithm's efficiency and how they performed on a large dataset. Scikit-Learn, also known as sklearn, which is a free machine learning package for the Python programming language that we utilized to develop the supervised machine learning classifiers in this research. We also used tensorflow to implement ANN, RNN. For data preprocessing and visualization, additional free Python tools including Pandas, NumPy, and Matplotlib were used.

LITERATURE REVIEW

kNN

KNN is a case-based learning technique that maintains all training data for classification. As a method of unmotivated learning, it is inapplicable to a variety of applications, including dynamic web mining for a large repository. Identifying select representatives to represent the complete training data set for classification is one method for enhancing its effectiveness. creating an inductive learning model from the training dataset and utilizing this model for classification (representatives). Several modern techniques, such as decision trees and neural networks, were originally designed to build such a model. Performance is one of the factors for assessing algorithms. As kNN is a simple but effective classification technique, and it is compelling as one of the most successful classification ways in our circumstance, we believe it to be one of the most effective classification methods. [2]

Naive Bayes

Naive Bayes (NB) is a well-known probabilistic technique for classifying data. It is a simple yet effective algorithm with numerous real-world applications, including product suggestions, medical diagnostics, and autonomous vehicle control. Due to the inability of real data to satisfy the assumptions of NB, modifications of NB exist to accommodate general data. The Naive Bayes approach has proven to be a practical and effective classification technique for multivariate data. However, characteristics are typically correlated, which violates the concept of conditional independence of the Naive Bayes technique and may degrade its effectiveness. In addition, datasets frequently have a high number of characteristics, which can confuse the interpretation of the results and slow down the execution of the procedure. [3]

Decision Tree

In addition to machine learning, image processing, and pattern recognition, decision trees are an effective method used in a variety of other fields. DT is a model that includes a number of fundamental tests that compare a numerical characteristic to a threshold value. The conceptual ideas underlying a neural network's inter-node connections are far easier to build than the numerical weights. DT is generally employed for grouping reasons. DT is also a popular classification technique in Data Mining. Nodes and branches make up a tree's structure. Each node represents features in a category that must be evaluated, whereas each subset specifies a possible value for each node. [4]

Random Forest Classifier

Random Forest is one of the adaptable, simple, and hyperparameter-free supervised learning methods. This categorization is excellent. In order to classify all accessible data utilizing Random Forest, a number of trees must be generated. Where the amount is greatly dependent on each data point. The quantity of breaker properties influences the minimal number of trees for each data set. Precision is significantly affected by the quantity of trees. Accuracy improves as the number of trees increases, starting with the fewest possible. There exists an optimal level of precision, after which precision is achieved even if the number of trees and precision remain unchanged. The number of breaker properties influences the accuracy of this method. The Random Forest will have low precision if the number of breaker attributes is equal to the number of accessible attributes. [5]

Support Vector Classifier

SVM, or Support Vector Machine, is one of the most widely used regression and classification techniques in Supervised Learning. It is primarily used for Classification challenges in Machine Learning. The objective of the SVM method is to build the optimal line or decision boundary that divides n-dimensional space into classes, hence enabling the classification of subsequent data points. This optimal decision limit is depicted as a hyperplane. SVM finds the extreme points/vectors that contribute to the construction of the hyperplane. These severe conditions are known as support vectors, and the corresponding technology is known as the Support Vector Machine.

Linear regression

The linear regression technique is one of the most widespread and straightforward Machine Learning algorithms. This statistical method is used to conduct predictive analysis. Spontaneous or quantitative variables, such as sales, salary, age, and product price, are predicted via linear regression.

The linear connection between a dependent variable (y) and one or more independent variables (y) is demonstrated by the linear regression method, thus the term linear regression. As linear regression displays a linear connection, it determines how the value of the dependent variable fluctuates relative to the value of the independent variable.

Ridge Regression

Ridge regression is a statistical technique for predicting the coefficients of multiple-regression models in instances where the independent variables are strongly correlated, in comparison to linear regression, which is the industry standard algorithm for regression and assumes a linear relationship between input variables and the target variable. Ridge Regression is a linear regression improvement that, during training, effectively nullifies the loss function for regularization.

ANN

The biological neural network in the human brain served as the model for the widely used machine learning technology known as artificial neural networks (ANN) [6]. For different issues, various ANN architectures will produce various solutions. Artificial neural networks are often organized in layers. Each ”node” in a layer has a ”activation function” and is formed of several linked ”nodes.” In a neural network, there are three layers: the Input layer, the Hidden Layers, and the Output layer. Some ANN variations receive inputs from neurons in the previous layer before sending the weight values of each artificial neuron as output to the following layer. This adjusts the weights across neurons using a back propagation approach to reduce the inaccuracy. This model does a decent job at picking up trends. It is quickly adaptable to new data values, although the system.

RNN

Recurrent neural networks (RNNs) are the most effective algorithm for sequential data and are the backbone of Google voice search and Apple's Siri [7]. It is one of the algorithms that helped deep learning accomplish some incredible successes over the past several years. Because it is the only algorithm in use with an internal memory, RNNs are a robust and stable type of neural network and one of the most promising ones. RNNs can be extremely accurate in predicting the future because of their internal memory, which enables them to remember key details about the input they received. They are hence the algorithm of choice for sequential data such as time series, voice, text, financial data, audio, video, weather, and many more types. When compared to other algorithms, recurrent neural networks are significantly better at understanding a sequence and its context.

XAI

The objective of an explainable AI (XAI) system is to provide reasons for its activities in order to increase human comprehension and explanatory power [8]. Designing AI systems that are more effective and understandable by humans can be done by following some common design principles. By outlining what it has done, is doing, will do, and is acting upon, the system should be able to communicate its capabilities and understandings. Every explanation is given a context based on the user's job, abilities, and expectations of the AI system. Because of this, it is impossible to describe interpretability and explainability without reference to a particular domain; instead, they must be defined in relation to that domain. The main facets of their mental process are the focus of partially interpretable models. Black box occasionally ignores ”interpretability requirements,” which are based on a certain domain. Partial explanations include saliency maps, local models that at some locations closely resemble global models, and variable importance measures.

DATASET

Dataset Description

We first generated a Pandas Dataframe after reading the dataset from the csv file. There are a total of 8 columns displayed here, and the column ”Unnamed:0” is one of them. This column is unnecessary, and we want to get rid of it during the data preprocessing stage.

Data Processing

At first we will delete the unnecessary column named “Unnamed: 0” from our dataframe.

Then we further drop all the rows that contain Nan values from our dataframe; moreover, we formatted the strings of the columns so that we can eradicate the redundant data as those data can cause various errors and problems in our research.

Data Analysis

After removing redundancy from our dataframe we plot a correlation graph in Fig. 2 and scatter matrix Fig. 1 in between a few columns.
Then we plotted a scatter plot in Fig. 3 of MRP and Selling Price based on the Category column.

We convert MRP, Discount and Selling Price as Int type data. Then, we divide the dataset, we take MRP and Discount as X and Selling Price as y. Using train_test_split from sklearn we split the dataset into two parts, train dataset and test dataset. 20% of the dataset was selected for the test dataset.

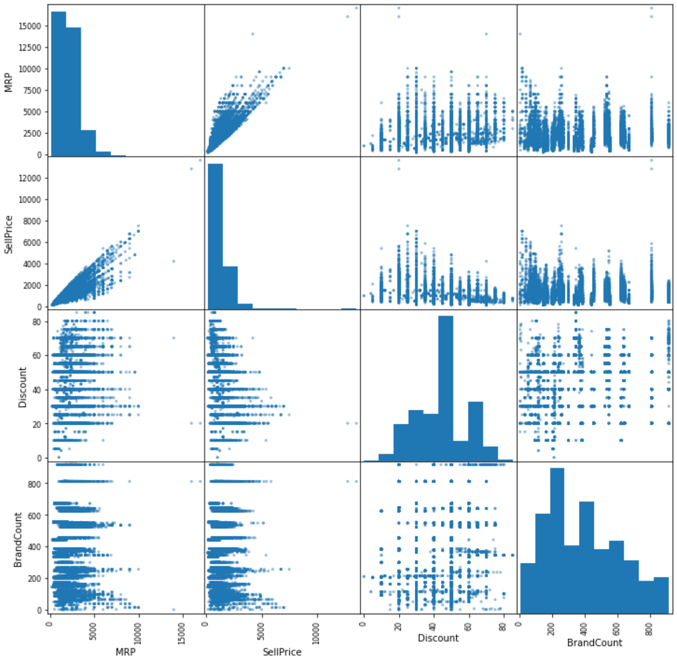

Fig. 1. Scatter Matrix

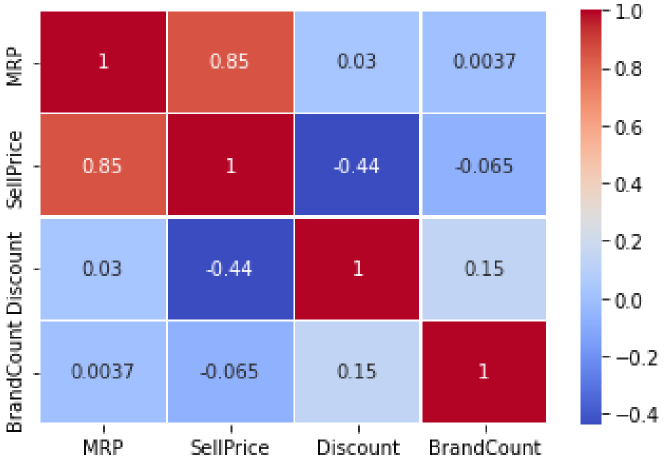

Fig. 2. Correlation Graph

METHODOLOGY

kNN

The KNN algorithm believes that similar objects are located in close proximity. KNN encapsulates the concept of similarity using arithmetic we may have learnt as children—calculating the distance between points on a graph (also called distance, proximity, or closeness). First, we populate the data. Set K to the number of selected neighbors. For each instance in the data, the distance between the query instance and the current instance is calculated. It adds the example's distance and index to the ordered collection and arranges the distances in the ordered collection of indices and distances in ascending order from smallest to largest. Selecting the initial K items

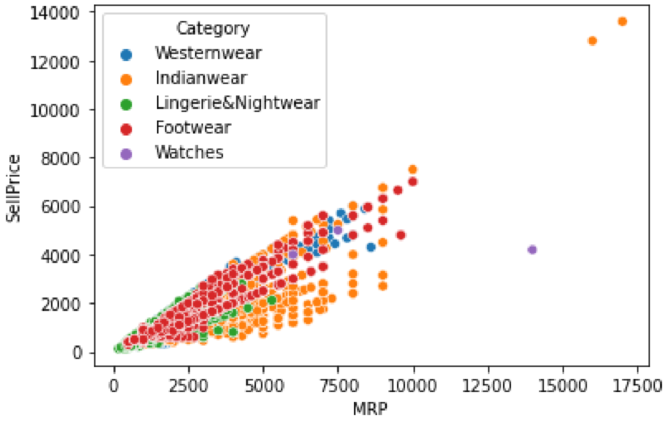

Fig. 3. Scatter Plot

from a sorted collection it obtains the labels of the K selected entries. After classification, it provides the mode of the K labels. [9]

Naive Bayes

The Naive Bayes classifier works on the principle of conditional probability, as stated by the Bayes theorem. This theorem is divided into 5 parts. First we separate the dataset by class. Then we summarize the dataset then we summarize it by classes. Afterwards, Gaussian Probability Density Function is used. Then we calculate the Class Probabilities. We can frame classification as a conditional classification problem with Bayes Theorem as follows: [10]

$P\left(yi|x1,x2,..,xn\right) \quad = \quad P\left(x1,x2,..,xn|yi\right) \quad * P\left(yi\right)/P\left(x1,x2,..,xn\right)$

Decision Tree

Decision trees use a variety of ways to determine whether or not to divide a node into two or more sub-nodes. Creation of sub-nodes increases the homogeneity of freshly produced sub-nodes. In other words, we can verify that the node's purity increases as the target value increases. The top-down greedy search method of the ID3 algorithm moves through the space of possible branches to make decision trees without backtracking. As the term suggests, a greedy algorithm always chooses the option that looks best at the time.
The root node consists of the initial set S. For each iteration, the method determines the entropy (H) and information gain (IG) of the most neglected property of the set S. The property with the lowest entropy or the greatest information gain is then chosen. The set S is then split by the chosen attribute to generate a subset of the data. As the algorithm iterates over each subset, it only evaluates previously unselected features. [11]

Random Forest Classifier

As its name implies, a random forest is comprised of multiple independent decision trees that function as an ensemble. Our model makes a prediction based on the category with the most votes. Individual trees inside the random forest emit class predictions. Ensemble approaches in this instance integrate multiple learning algorithms to obtain a greater anticipated performance than any of the individual learning algorithms could.
It chooses K data points at random from the training set and constructs the decision trees associated with the chosen data points (Subsets). Then, it selects N as the number of desired decision trees and repeats Steps 1 & 2. It determines the predictions of each decision tree for new data points and assigns the new data points to the category that received the majority of votes. [12]

Support Vector Classifier

By mapping the data to a high-dimensional subspace, SVM can categorize data points even when they are not generally linearly distinct. Once a divider between the categories has been discovered, the data are transformed so that the divider may be represented as a hyperplane. Consequently, the group to which a new record should belong may be predicted using the characteristics of the new data.
VM can get away with just the dot products between them; it actually doesn't need the actual vectors to perform its magic. As a result, we may avoid performing the time-consuming calculations for the new dimensions. We employ a mechanism known as a "Kernel Function" to do this. The Kernel's value is often set to "Linear," however it can also be different. [13]

Linear regression

Multiple Linear Regression is similar to simple linear regression but here we have more than one independent or explanatory variable. Linear Regression can be written mathematically as follows: [14]

$$Y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \beta4X4 + \beta5X5 + \beta6X6 + \epsilon$$

$$\beta0 = Y - intercept(alwaysaconstant);$$

$$\beta1, \beta2, \beta3, \beta4, \beta5, \beta6 = regression\ coefficients;$$

$$\epsilon = Errorterms(Residuals).$$

Finding the best fit line is crucial when using linear regression since it minimizes the difference between the predicted and actual values. The line with the least inaccuracy will have the best fit. To obtain the optimal values for a0 and a1 in order to locate the best-fitting line, we employ the cost function. Variable values for weights or line coefficients (a0, a1) provide a separate regression line.

Ridge Regression

An optimization procedure is used to find the model's coefficients in order to reduce the total squared error between the predictions ($\hat{y}$) and the anticipated target values (y). [15]

$$loss = sum,\ i = 0\ to\ n(yi - \hat{y}i)^2$$

A common punishment is to evaluate a model's performance according to the sum of its squared coefficient values (beta). The penalty is referred to as an L2.

$$l2\ penalty = sum,\ j = 0\ to\ p\ \beta j^2$$

Although it avoids any coefficients from being eliminated from the model by allowing their value to become zero, an L2 penalty reduces the size of all coefficients.

ANN

In the biological world, minute parts are assembled into three-dimensional brain networks [16]. These neurons appear to have almost limitless connectivity potential.

In essence, every simulated neural system has a similar topology or structure. A fraction of the neurons in such a structure interface with the current world to obtain their information sources. By creating layers of components, you can outline a framework in one of the least difficult ways possible. A functioning neural system is involved in the organization of these neurons into layers, the relationships between these layers, and the summation and exchange functions.

They can be viewed as weighted directed graphs, with neurons serving as nodes and connections between neurons serving as weighted edges. A neuron's processing portion receives a variety of messages (both from other neurons and as input signals from the external world).

At times, weighted inputs are summed at the processing element and signals are adjusted at the receiving synapses. If the threshold is crossed, it becomes an input to other neurons (or an output to the outside world), and the cycle is repeated. The weights often show how strongly the neurons are connected to one another. An output for the designed problem is obtained using the activation function, a transfer function. Suppose the desired result is either zero or one.

The weights often show how strongly the neurons are connected to one another. An output for the designed problem is obtained using the activation function, a transfer function. Consider a binary classifier where the desired output is either zero or one. As an activation function, sigmoid function might be employed.

RNN

Recurrent edges that span neighboring time steps are added to feedforward edges, making recurrent neural networks a strong superset of feedforward neural networks and giving the model a sense of time [17]. Recurrent edges, including self-connections, can cycle even when conventional edges in RNNs might not. Nodes receiving input along recurrent edges at time t get input activation from hidden nodes h(t1) in the network's prior state as well as from the current example x(t). Given the hidden state h(t) at that particular time step, the output y(t) is calculated. As a result, with the use of these recurrent connections, the input x(t-1) at time t-1 can affect the output y(t) at time t. In a straightforward recurrent neural network, we may demonstrate in two equations the computations required for computation at each time step in the forward pass:

$$h^t = \sigma(W_{hx}x + W_{hh}h^{(t-1)} + b_h)$$

$$\hat{y}^{(t)} = softmax(W_{yh}h^{(t)} + b_y)$$

XAI

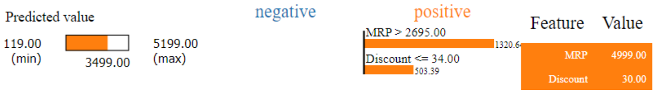

Fig. 4. Scatter Matrix

From Fig. 4 actual MRP is 4,999 and Discount is 30. now according to LIME as the price is greater than 2,695 and discount is less than 34 our model has predicted that the predicted selling price is 3,499. And we can see both MRP and Discount played a positive impact on the selling price prediction.

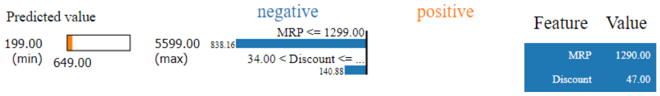

Fig. 5. Scatter Matrix

From Fig. 5 the original MRP is 1,290, while the discount is 47. According to LIME, since the price is less than 2,695 and the discount is greater than 34, our model predicts that the product will sell for 649. And we can see that both MRP and Discount had a negative effect on the predicted selling price.

Experimental Findings

The research has employed a total of nine distinct supervised learning algorithms to determine which model is most efficient, accurate, and effective when applied to a huge dataset. We have produced Accuracy Score, Recall Score, Precision Score, F1 Score, and Confusion Matrix (if required) for each classifier in order to evaluate and compare their performance on this massive dataset.

From table I we can infer that the best accuracy scores have been provided by Decision Tree classifier and Random

Models	Accuracy	Root of Mean Squared Error
KNN	90.48%	94.58
Naive Bayes	37.77%	186.97
Decision Tree	**95.05%**	21.81
Random Forest Classifier	**95.05%**	**7.64**
Support Vector Classifier	94.61%	283.20
Linear Regression	93.69%	178.40
Ridge Regression	93.69%	178.40
ANN	N/A	22.54
RNN	N/A	161.42

TABLE I
ACCURACY AND ROOT OF MEAN SQUARED ERROR OF DIFFERENT MODELS

Forest Classifier followed by Support Vector Classifier, Linear Regression, Ridge Regression and kNN Classifier. While 6 of the 7 classifier gives an accuracy score of 90% or more than 90%; but Naive Bayes Classifier fails miserably in this case and gives roughly 38% accuracy score which further proves that unlike others, Naive Bayes Classifiers is not a good option when you are working with a really big dataset.

Our main focus is on Root of Mean Squared Error. We are getting best score of 7.64 with Random Forest Classifier. Decision Tree and ANN also had a overall good outcome with 21.81 and 22.54 respectively. Support Vector Classifier had the worst outcome with a root of mean squared error of 283.20.

CONCLUSION AND FUTURE WORKS

Decision Tree classifier and Random Forest classifier both have exhibited the highest levels of accuracy with 95.05%. While, Naive Bayes Classifier fails poorly and achieves an accuracy score of 37.77%. Random Forest Classifier yields the highest possible score of 7.64. With 21.81 and 22.54 respectively, both Decision Tree and ANN achieved a favorable outcome. With a root mean square error of 283.20, Support Vector Classifier got the poorest performance.

We can come to a decision that for regression task, Random Forest Classifier gives us the best result followed by Decision Tree and ANN.

For future work, we can use more deep learning models and compare them to our used models and find the best out of them. Then, we can use the models on data from our local website and find the best selling items. This way we can understand which products to mass produce. Finally, we can create a recommendation system where we recommend similar price products to the customers.

REFERENCES

[1] IBM Cloud Education (2020) ”Machine Learning”

[2] G. Guo, H. Wang, D. Bell, Y. Bi & K. Greer ”KNN Model-Based Approach in Classification”

[3] I. Wickramasinghe & H. Kalutarage ”Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation”

[4] A. J. Myles, R. N. Feudale,Y. Liu, N. A. Woody, S. D. Brown ”An introduction to decision tree modeling”

[5] M. Huljanah, Z. Rustam, S. Utama and T. Siswantining ”Feature Selection using Random Forest Classifier for Predicting Prostate Cancer”

[6] Md. T. Sarker, S Noor, U. K. Acharjee ”Basic Application and Study of Artificial Neural Networks”

[7] N. Donges, ”A Guide to RNN: Understanding Recurrent Neural Networks and LSTM Networks” Builtin

[8] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G. Yang ”XAI—Explainable artificial intelligence”

[9] O. Harrison ”Machine Learning Basics with the K-Nearest Neighbors Algorithm”

[10] Simplilearn ”Understanding Naive Bayes Classifier”

[11] N. S. Chauhan, ”Decision Tree Algorithm, Explained”

[12] T. Yiu ”Understanding Random Forest”

[13] IBM Documentation ”How SVM Works”

[14] S. Glen. ”Linear Regression: Simple Steps, Video. Find Equation, Coefficient, Slope” From StatisticsHowTo.com

[15] J. Brownlee ”How to Develop Ridge Regression Models in Python”

[16] Analytics Vidhya ”Artificial Neural Network, Its inspiration and the Working Mechanism”

[17] C. L. Zachary ”A Critical Review of Recurrent Neural Networks for Sequence Learning”