Vending Machine Sales Analysis Using Machine Learning

Submitted by

Ishtiana Haque

201-35-2994

Department of Software Engineering

Daffodil International University

Supervised by

Ms. Nusrat Jahan

Assistant Professor

Department of Software Engineering

Daffodil International University

This Thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering

APPROVAL

APPROVAL

This thesis titled on **"Vending Machine Sales Analysis Using Machine Learning"**, submitted by **Ishtiana Haque (ID: 201-35-2994)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

9/7/24

-- **Chairman**

Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

9/7/24

-- **Internal Examiner 1**

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **Internal Examiner 2**

Md. Rajib Mia
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **External Examiner**

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering
Jagannath University

DECLARATION

Declaration

I announce that I am rendering this study document under Nusrat Jahan, Assistant Professor, Department of Software Engineering, Daffodil International University. I therefore, state that this work or any portion of it was proposed here therefore for Bachelor's degree or any graduation.

Supervised By

15.07.24

....

Nusrat Jahan
Assistant Professor
Department of Software Engineering
Daffodil International University

Submitted By

Ishtiana Haque

....

Ishtiana Haque
ID: 201-35-2994
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, I sincerely appreciate and am grateful to Almighty God for His divine grace, which made it possible for me to successfully complete the final year thesis.

I express my gratitude and our deep obligation to Ms. Nusrat Jahan, Assistant Professor, at Daffodil International University in Dhaka's Department of Software Engineering. To complete this assignment, our supervisor must have deep knowledge of and a strong interest in the topic of "Machine Learning." The completion of this assignment has been made possible by his unending patience, academic guidance, ongoing encouragement, constant and energetic supervision, constructive criticism, insightful counsel, reviewing numerous subpar versions, and fixing them at every stage.

I want to sincerely thank Dr. Imran Mahmud, the faculty head for software engineering, for his helpful assistance in getting our project completed.

Moreover, I would also like to thank all the other teachers as well as the staff from Daffodil International University's department of software engineering who have been with us every step along the way- your support has meant everything!

In closing though I must say that no one has done more than my parents; without them this never would've happened!

ABSTRACT

Modern vending machines offer a wider range of products and services, including fresh food, electronics, beauty products, and even rental services like bicycles or umbrellas. they provide convenient access to goods and services in locations where traditional stores might not be feasible or available. I got curious about this and decided to dive into vending machine data. I wanted to see them from both the customer's view and the perspective of the people who manage them, using data to guide my investigation. I gather data from vending machines across different locations. The locations include a library, a mall, an office location, and manufacturing locations. I have used Regression (XGBRegressor) analysis to predict sales volume based on factors such as product price, location demographics, and time of day. I also used k-Nearest Neighbors (kNN) for analyzing mapped quantity sold in vending machine data involves predicting the quantity of products sold based on the characteristics of neighboring data points. Additionally, I also used Linear Regression which is employed to identify linear relationships between sales and various influencing factors, while a Decision Tree provided a visual representation of decisions, and their possible outcomes based on different variables affecting sales. Assess metrics such as accuracy, precision, recall, and R-squared value to measure the model's effectiveness in predicting vending machine behavior. According to the conclusions of this research, Vending machine sales thrive on technology, diverse offerings, and easy payments. Meeting consumer preferences and managing inventory efficiently is key to success.

Table Of Contents

APPROVAL i

DECLARATION ii

ACKNOWLEDGEMENT iii

ABSTRACT iv

CHAPTER 1 1

1.INTRODUCTION 1

1.1 BACKGROUND 1

1.2 MOTIVATION OF THE RESEARCH 2

1.4 RESEARCH QUESTIONS 4

1.5 RESEARCH OBJECTIVE 4

1.6 RESEARCH SCOPE 4

CHAPTER 2 5

2. LITERATURE REVIEW 5

2.1 INTRODUCTION 5

2.2 PREVIOUS LITERATURE 5

CHAPTER 3 8

3. RESEARCH METHODOLOGY 8

3.1 RESEARCH METHODOLOGY 8

3.2 DATA COLLECTION 8

3.3 DATA PREPROCESSING 9

3.4 XGBRegressor 10

3.5 Linear Regression 12

3.6 k-nearest neighbor (kNN) 13

3.6 Decision Tree 14

3.9 Evaluation Methods 14

3.9.1 Accuracy: 14

3.9.2 Precision: 14

3.9.3 Recall (Sensitivity): 15

3.9.4 F1 Score: 15

3.9.5 Confusion Matrix: 15

3.9.6 Mean Squared Error 16

3.9.7 R^2 value 16
CHAPTER 4 17
4. RESULTS AND DISCUSSION 17
4.1 INTRODUCTION 17
4.2 RESULT DISCUSSION 17
4.2.1 XGBRegressor 18
4.2.2 Linear Regression 19
4.2.3 k-nearest neighbor (kNN) 20
4.2.4 Decision Tree 21
4.2.5 Heatmap 22
CHAPTER 5 23
5. CONCLUSION AND LIMITATIONS 23
5.1 CONCLUSION 23
5.2 LIMITATIONS 23
CHAPTER 6 24
REFERENCES 24

List Of Figure

Fig. 3.1: Figure of methodology -- 10

Figure3.2: A general architecture of XGB 11

Figure 3.3 General architecture of Linea Regression 12

Figure 3.4:.k-nearest neighbor (kNN) 13

Figure 3.5: Confusion Matrix 15

Figure 4.1:Output of XGBRegressor 18

Figure 4.2 Output Of Linear Regression 19

Figure 4.3: Result Of Knn 20

Figure 4.4 Output of Decision Tree 21

Figure 4.5: Heatmap 22

CHAPTER 1

1.INTRODUCTION

1.1 BACKGROUND

Vending machines are everywhere these days! It's kind of amazing how they sell all sorts of things — tasty snacks to neat gadgets. They offer more friendly service than regular in many places. Plus, owners can save money on setup & labor costs. These machines provide what customers want, whether it's food, drinks, CDs, newspapers, cleaning supplies, fun toys, or even postage stamps (Feng-Cheng Lina, 2011) [1]. So, what exactly is a vending machine? Well, it's a simple machine that gives out food, drinks, & other goods when folks pay with cash or with credit card. Nowadays, you can find them in all sorts of public spots like shopping malls, airports & offices. These self-service devices make it easy for companies to sell stuff and for customers to grab what they want (Mnyakin, 2020) [2]. Vending is really all about convenience! Just drop some coins in and push a button. But underneath this ease lies a complicated mix of how people behave as consumers & the economics of the market. The machines automatically give out whatever products people choose. Although collecting money and keeping the machines stocked is done by people, the sales happen all on their own (Hidetaka Sakai*, 1999) [3].

For this Thesis, we're looking at real sales data from vending machines in Central New Jersey between January and August 2022. This dataset shows sales for four different types of snacks: food items, carbonated drinks, non-carbonated drinks & water. You'll see familiar products like Sun Chips, Coca-Cola, Bai Lemonade or Poland Springs Water making appearances in these transactions. It's like they each play their part in the busy world of vending machine sales. We placed the vending machines in four different locations: a mall, a library, a corporate office & a manufacturing plant (Vending Machine Sales, 2022) [4]. Every spot attracts its own unique customers! For example—people at the library tend to go for healthier snacks because that place is focused on wellness.

With this sales data from various locations handy, we're excited to find out general patterns about sales and build machine learning models that help us make better decisions about what affects sales of vending machines. Today's tech can store huge amounts of data! But figuring out how to process that information meaningfully is just as important as keeping it safe (Natalya V. Razmochaeva1, 2019) [5]. We have cool machine-learning methods to tackle these challenges. Our research investigates how city living (Hidetaka Sakai*, 1999) [3], product placement, and marketing strategies affect how vending machines operate here in Central New Jersey.

1.2 MOTIVATION OF THE RESEARCH

In today's world, convenience means a lot. Vending machines are a big part of our daily lives. They help us grab a quick meal or drink when we're busy & on the move. You can find them everywhere – from bustling shopping malls to quiet office buildings in places like Katowice. These machines serve different people in various locations.

As vending machines continue to pop up worldwide, understanding what makes them sell well becomes more important. More and more technologies are being used in these machines to offer personal & customized services. Recently, many new technologies have come along that give vending machines cool new features, according to Elliot (2007). Now, you can find options like cashless payments, machine monitoring from afar, security measures, and even planning menus based on data (Feng-Cheng Lina, 2011) [1].

This study aims to uncover what drives vending machine sales in Central New Jersey. We will look at data from places like libraries, malls, offices, and even factories (Vending Machine Sales, 2022) [4]. The goal here is simple: gather important insights by analyzing trends and how users behave at these spots. This could help improve vending operation efficiency, increase sales, & make customers happier in the vending machine industry today and in the future.

Looking at real data from vending machines in Central New Jersey can show us several things: what products people like best, how a machine's location affects its sales, & what factors influence product choices. When vending machines first start operating, there isn't any sales data available yet because they are new to the location. We don't know what customers will choose initially.

So, when setting up these machines, it's super helpful to pick products that meet as many customers' needs as possible (Polina S. Deryabina, 2021) [6]. As time passes, these vending machines can learn about customer habits & preferences. This learning helps fine-tune the experience for everyone so that each person can get what they want and where they want it!

1.3 PROBLEM STATEMENT

In conventional retail stores, clients can see, touch, and feel the items they recently purchased, and with the assistance of staff, they can select from a wide variety of items. A distributing machine acknowledges cash or other insertable shapes of installment in trade for an assortment of commodities from its supporters. Since distributing machines are considered mechanical gear, most items and innovative advancements in this range have been generally needs-aligned and centered on the concept of installment handling (Bajaj, 2020) [7].

Indeed, even though distributing machines give fast exchanges all over and anytime at lower operational costs, not at all like distributing machines, shopper benefit faces the issue of client dissatisfaction against moment exchanges. A few calculations are utilized by numerous analysts, to be specific, SVM calculation, KNN calculation, SPSS, etc. Beyond any doubt, you see vending machines all over, but there's a part underneath the hood that we might take for allowed. Data preprocessing is a critical component of data analysis. Data Pre-Processing is what determines the applicability of some machine learning methods. Distributing machine administrators persistently fight with challenges to adjust item assortments, menu estimating methodologies, and machine situations to extend deal execution and offer to buyer trading occasions. In any case, the proprietor cannot track the distributing machine's deals employing a telemetry framework. To confirm buys and renew the refreshments, the proprietor ought to buy and visit the distributing machine area. The proprietor must bring each refreshment that's accessible at the location at the same time so that it can be refilled. The fetching of labor and transportation is raised by this issue (Bajaj, 2020) [7]. In this ponder, we collected information from different places around Central Unused Shirt, such as libraries, shopping centers, workplaces, and a fabricating location, to discover designs for distributing machine deals (Distributing Machine Deals, 2022).

We are inquisitive as to what drives individuals to select what they get from the distributing machine and the effect the course of action and arrangement of vending machines have on these choices. We need to form vending machines way better for clients and way better for sellers. Furthermore, we need to assist distributing machine administrators and partners to more ideally stock their distributing machines, set costs that make sense for client procurement, and choose areas for maximum sales. Administrators may discover it supportive to know a distributing machine's deal volume when deciding the best estimating strategy for their products. On occasion, on the off chance that a machine reliably offers a part of a specific item, the administrator may well be able to raise the cost to some degree and still have solid deals. (Mnyakin, 2020) [2]

1.4 RESEARCH QUESTIONS

Q1: How our proposed machine learning algorithms effectively utilized to optimize sales performance and revenue generation in smart vending machines?

Q2: Which machine learning algorithms have demonstrated the most effectiveness in smart vending machines?

1.5 RESEARCH OBJECTIVE

The main goal of this research is to understand the factors that impact vending machine sales and to provide actionable insights for vending machine operators. Specifically, the objectives are as follows:

- Optimize sales performance and revenue generation in smart vending machines by accurately predicting.
- Identify key factors influencing sales performance and revenue generation in vending machines.
- Evaluate the effectiveness of machine learning-driven strategies in increasing sales volume and profitability.

1.6 RESEARCH SCOPE

The research's main scope is as follows:

- Increased sales volume.
- Reduced operational costs.
- Ability to adapt quickly to changing market conditions, ensuring competitiveness.
- Identification of factors that can enhance vending machine sales.

CHAPTER 2

2. LITERATURE REVIEW

2.1 INTRODUCTION

A literature review is an investigation that looks at what is known about a subject from the existing body of knowledge. These types of reviews involve looking at previous studies, research papers, conference proceedings, books and articles related to this topic to get a complete understanding of where we are situated within it. By doing so, researchers can also bring together collective findings and identify trends, gaps or points that need further investigation – ultimately enabling them synthesize findings more coherently. This critical examination identifies limitations inherent in any study area but should be followed with proposing strategies for overcoming such weaknesses during subsequent research projects designed within those parameters thereby helping improve relevance as well as quality output anticipated from such undertakings while contributing new insights and knowledge transfer that is expected to benefit the concerned community.

The implications of vending machines are plentiful, consisting of sales dynamics, patron conduct, and tech advancements. Traditional It becomes difficult for forecast systems to deal with big data, and it will help in improving the accuracy of sales forecasting. These issues could be overcome. [2] .It is applied with the help of various data mining techniques. The incorporation of machine mastering (ML) and statistics analytics to the operation of vending machines has enabled improvements and improvements to person revel in. This evaluation of the literature specially investigates an array of research which has contextualized the aforementioned areas.

2.2 PREVIOUS LITERATURE

Vending machines are automated machines that offer products like snacks, beverages, and lottery tickets. They are valuable since they save time and physical labor. Relatively few studies of this kind have been made on vending machine sales behavior. Findings from previous research in vending machine sales provide substantial information about aspects of consumer behavior, location analysis, technological innovation, and operational efficiency. To be competent enough and to generate more, on the revenue front, business organizations are always looking for an improved model or technique of data mining and maintenance of critical data (Cheriyan, 2018) [8]. We also want to see if any new technologies for vending machines have increased levels in research and are affecting vending sales and user experience, such as cashless vending machine systems, vending machine inventory management software, and other vending technology.

(Natalya V. Razmochaeva1, 2019) [5]researched the automation of retail sales management processes. The regression methods they used were linear regression, ridge regression, and isotonic regression. They acquired data on product sales from product sales databases for evaluation. Sales data by period/service frequency. In addition, they catch ideas on possible approaches to reduce the dimensionality of sales product data and make use of machine learning to solve data processing issues.

(Hidetaka Sakai*, 1999) [3], Minoru Higashihara, Masashi Yasuda, and Masato Oosumi crafted energy-saving canned beverage vending machines using fuzzy logic. They applied fuzzy logic and multiple regressive models to increase forecasting accuracy.

(Mnyakin, 2020) [2] explores factors of vending business sales volume, trying to optimize it. They test their hypotheses by applying a multivariate regression model. Estimates effects on vending machine sales of sales literature. The researchers surveyed 251 vending machine operators (Mnyakin, 2020) [2]. Sales volume influence factors were evaluated by utilizing regression analysis. Vending machine sales and customer base increase through effective marketing strategies.

(Meagan L Pharis, 2016) [9] worked on healthy vending initiatives that resulted in healthier items at new vending locations and healthy item sales increasing without decreasing total sales. This included measuring sales volume, revenue, and the percentage of healthier products being purchased. Pricing strategies and promotions were used to drive sales in their healthier snack and beverage categories. Sales volume and revenue information from contracted vendors are collected by them.

(Feng-Cheng Lina, 2011) [1] proposes a framework for a localized product recommendation system for automatic vending machine applications. The aim is to provide recommendations of suitable localized products to customers in different locations. They proposed a hybrid technique by combining a meta-heuristic approach, clustering technique, classification, and statistical method. In this approach, an intelligent system was implemented that:

Analyze product attributes and decide localized products based on transaction data. To demonstrate the feasibility and effectiveness of the proposed approach, we applied the system to several automatic vending machines owned by an information service company in Taiwan.

(Cheriyan, 2018) [8]In this paper, they analyzed the concept of sales data and sales forecasts. The In the later part of the research work, different techniques and measures for sales predictions have been explained. The performance evaluation based on which the best predictive model is suggested for the sales trend forecast. The results are summarized in terms of the reliability and accuracy of efficiency:

Techniques taken for Prediction and Forecasting. The studies found that the Gradient Boost Algorithm is the best-fit model, which shows maximum accuracy in forecasting and future sales prediction.

(Park, 2015) [10] worked on the operation of smart vending machine systems under conditions of product substitution in cases of stock-outs faced by customers. They propose a simple and intuitive procedure for estimating substitution probabilities, formulate a comprehensive optimization mathematical model of the operation problem of smart vending machine systems, and finally design a two-phase heuristic approach based on this mathematical model. The computational

Experiments demonstrate a considerable economic gain of the smart vending machine system over the traditional one.

CHAPTER 3

3. RESEARCH METHODOLOGY

3.1 RESEARCH METHODOLOGY

We have applied KNN, Regression for our work on a dataset collected from Kaggle dataset.

3.2 DATA COLLECTION

Data collection the dataset for this study was collected in Central New Jersey from vending machines over 8 months periods spanning from January 2022 to August 2022. [3] The vending machines were used in 4 distinct places: a mall, a library, a corporate office, and a manufacturing plant. This wide range of kids allowed us to get a more well-rounded view of consumer habits within different environments.

This consists of heaps of houses. Status denotes whether the processing of machine data was successful. A Device ID is a digital fingerprint that is related to the asset of each vending machine. The vending machine location, along with the friendly name for each machine, is included. Specific products vended by the dataset categorized as Carbonated, Food, Non-carbonated, or Water. Transaction IDs are unique IDs, and TransDate is the date and time when the transaction is done. It also has the type of transaction, whether cash or credit. The dataset also contains information like how each product was vended (i.e., "by coil number"), how much the product cost, and how many were vended in the same transaction.

The other included are the Mapped Fields: Mapped Coil #, Mapped Price, and Mapped Qty Sold, which are associated with what the actual data sold computations. LineTotal - Pivot this to get the total sale amount per transaction, and then get the sum of all transactions to show the total was recorded on the credit card TransTotal. For example, if a user buys a drink for $3 and the twist for $1.50, the TransTotal should be $4.50. Processed Date, this information is related to the date the transaction was going through SeedLive, a static interface for electronically aggregating all transactions.

This included tracking every single transaction at the vending machines, recording each product sold in detail (product identifier/category), the sell price, and the transaction details. We use this extensive data source to perform a very deep analysis of vending machine sales as it relates to a variety of factors such as location, product type, and transaction method as they influence sales performance. Through this data, we hope to reveal actionable insights and unlock best practices for vending machine operations and product owners to enhance sales and their overall vending experience with vending consumers.

3.3 DATA PREPROCESSING

Data was collected from various vending machines located in Central New Jersey, covering the period from January 2022 to August 2022 [3]. The dataset includes detailed transaction records such as product types, prices, locations, and specifics about each transaction. Data preprocessing is an important step before data analysis. For some machine learning methods, only after preprocessing can they be implemented [1]. Now, after data preprocessing, to have a clear understanding It was an exploratory analysis in response to the nature of our data. Conducted [2]. This data required thorough preprocessing to ensure its quality and enhance the accuracy of our analysis. The dataset comprises about 9617 records, each containing information on transactions made at the vending machines. Each record includes annotation files indicating the exact details and labels of the transactions. These annotations are stored in XML files, and the data is organized into 17 features. Representing different product categories. This structured and annotated data facilitated the development of accurate analytical models. The dataset was split into training and validation sets to train and validate the models. 80% of the data was used for training and 20% for validation. The training feature matrix has 7693 rows and 17 columns, indicating there are 7693 data points in the training set. The testing feature matrix has 1924 rows and 17 columns, indicating there are 1924 data points in the testing set.

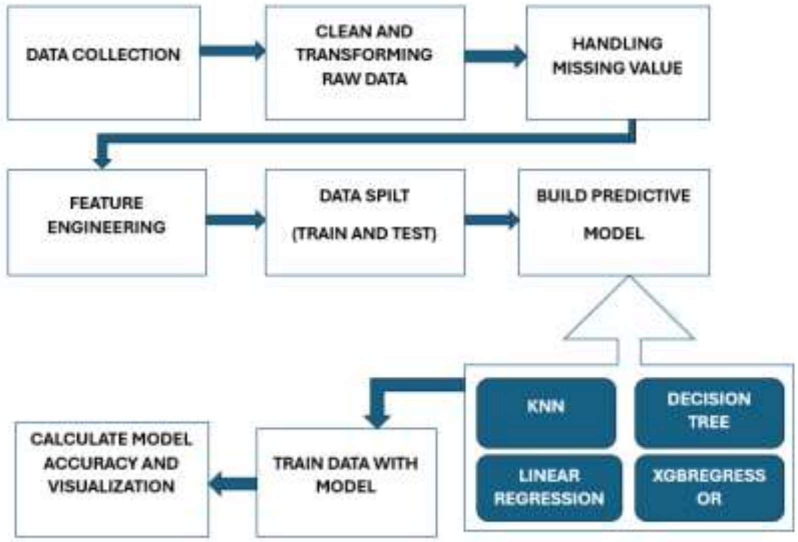

Fig. 3.1: Figure of methodology

The diagram (see Fig. 3.1) illustrates a workflow how sales analysis using a vending machine with machine learning would work: beginning with data collection, then cleaning and transforming raw data, and handling missing values to assure data quality. After feature engineering, the dataset is divided for training and test sets towards the creation of predictive models like KNN, Decision Trees, Linear Regression, or XGBoost. Train the model and finally evaluate its performance in terms of accuracy, with visualization showing sales patterns and trends for vending machines.

3.4 XGBRegressor

XGBRegressor plays a pivotal role, which will be an important constituent in our whole arsenal for predictive modeling with respect to the vending machine sales analysis thesis. XGBRegressor had superior results when the task was to forecast the sales trends for vending machines, one of the most robust regression algorithms in the XGBoost library. The $R2$ value is a direct measure of the model's performance in explaining variance in sales data, and it is a very important metric.

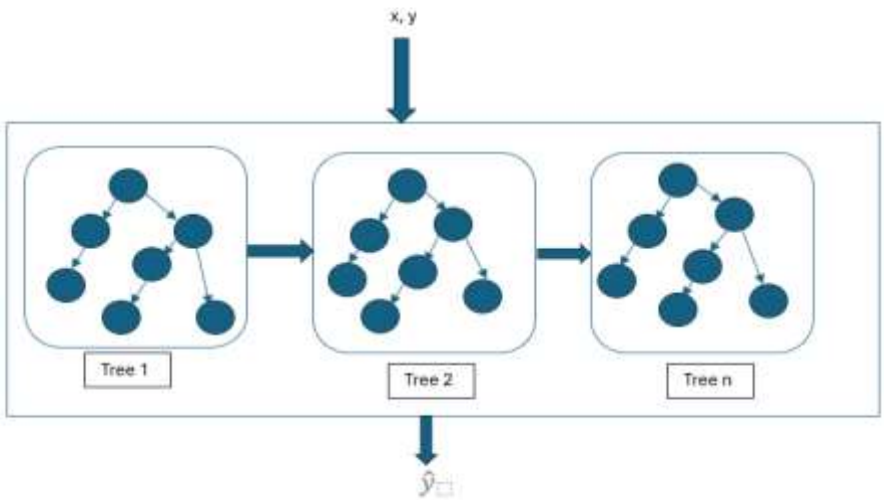

Figure3.2: A general architecture of XGB

This figure shows the prediction process of an ensemble model of many decision trees, like XGBoost. The input data includes features x and target values y, and then the data is fed through a series of decision trees: Tree 1, Tree 2..., Tree n. Each tree makes a prediction output, f1, f2..., fn. These tree outputs are added to compute the final prediction outcome. The combined prediction y^ is calculated by taking the sum of all tree outputs, formally written as y^=∑k=1nfk(x), where y^ is the sum of the prediction of the ensemble of trees. Then we can check the model's goodness of fit using the R-squared value. R2: also known as the coefficient of determination, R2 measures how much of the dependent variable variation can be accounted for based on the independent variables. It is an indicator of how well the model's predictions match that of the actual data, where a value closer to 1 reflects a better fit. The performance metric R2 value is calculated from the final prediction and the target values and is a key performance metric to judge how well the predictions are being made by an ensemble model.

3.5 Linear Regression

The relationship between the input (x) and the output (y) is captured in a linear equation in the simplest case of a single input feature and is given by y=$\beta 0$+$\beta 1$x, where y is the predicted output, $\beta 0$ is the intercept, and $\beta 1$ is the coefficient that represents the slope of the line. The focus is for the Continuous response where the variable gives continuous, Categorical predictors [2] .This equation can be generalized to multiple input features as y=$\beta 0$+$\beta 1x1$+$\beta 2x2$+…+βnxn, where each βi is the weight or coefficient for the corresponding feature xi.

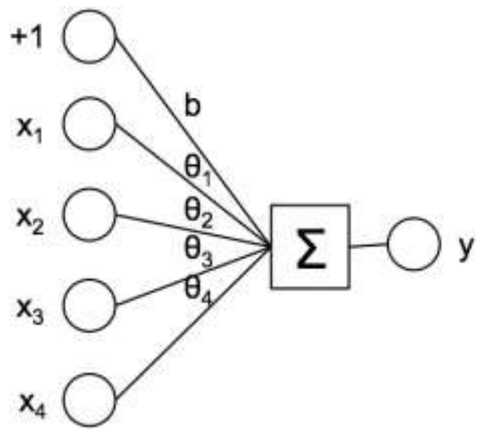

Figure 3.3 General architecture of Linea Regression

A type of regression is when a single output is predicted with a single input. It does so by setting the coefficients such that the mean square error is minimized, thereby providing a best-fitting line that can correctly predict the target variable and the independent variables.

$$\text{MSE} = \frac{1}{n}\sum_{i=I}^{n}(y_i - \hat{y}_i)^2$$

Training a linear regression model means calculating the values of the coefficients β that, in turn, minimize the cost function. MSE (Mean Squared Error) - It is the average of the squares of the errors between the predicted and actual values; we used it for this study.

3.6 k-nearest neighbor (kNN)

K-Nearest Neighbors operates by holding on to all available data and making a prediction based on the k number of similar data points in the feature space. To achieve this, KNN compares patterns and trends in new sales data with data that was previously recorded. The same goes for each new data point, and the algorithm will calculate the distance to all points in the dataset find the k most similar points, and then impute the values of the dependent variable from these closest points. One of the reasons we employed KNN to make our vending machine sales predictions is its strength in the ability to generalize predictions without making assumptions about the data distribution of the underlying data.

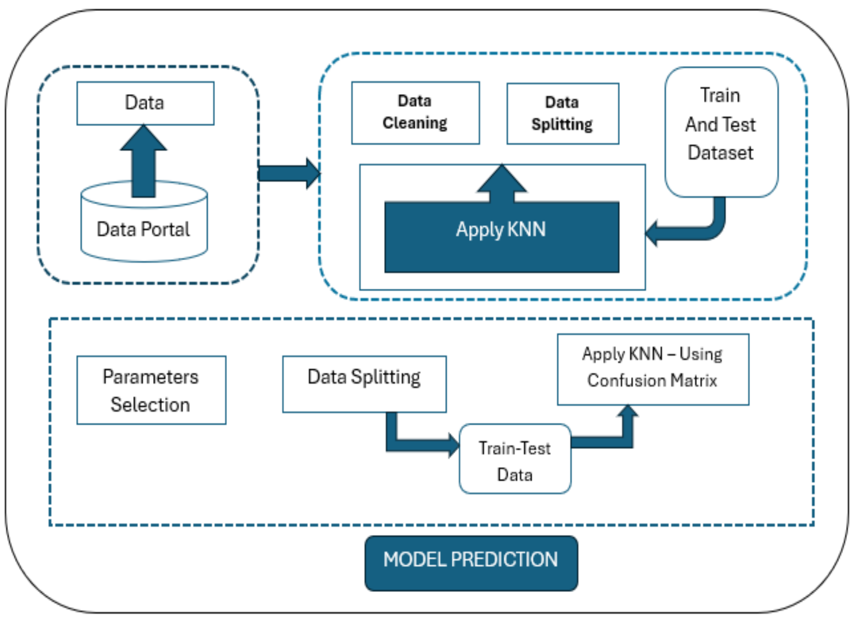

Figure 3.4: k-nearest neighbor (kNN)

The diagram shows (Figure 3.4) the machine learning process and can be applied to the K-Nearest Neighbors (KNN) algorithm. Data is first collected by a data portal; then, it goes through a data-cleaning process to make the data accurate. Then, that data is split into training and testing datasets. Clean data undergo KNN, and finally, confusion matrices are used for the model evaluation. It selects parameters and again splits data into training and testing data. Finally, the process ends by making a final prediction.

3.6 Decision Tree

It learns many different trees, successively splitting the data based on the most informative features. The decision tree is referred to as a recursive classifier or partition of the instant space. It happens to be a strong form of multiple variable analyses and happens to be a robust data mining tool. It has applications found across various domains. This approach represents factors involved in achieving a predetermined goal and the corresponding factors to achieve the goal and the ways and means of implementation. For each node of the tree, the algorithm chooses the feature that creates the purest children nodes, meaning the algorithm will try to reduce impurity or otherwise increase the amount of information it gains from the data. Basically, a decision tree is a structure laid in architecture, having a root node to the farthest leaf, representing the whole dataset. From there, the algorithm selects the feature and threshold that best breaks the data into two subsets. Now the decision tree will process the features like Product detail, number of products, total transaction, and other points available with the "RQty" feature. The decision points in the tree will now be conditioned by the amount sold (e.g., "If RQty <= 1 then go left, else go right"). The visualization of the decision tree shows in what ways the quantity sold affects sales performance and if there is a non-linear relationship between quantity and sales.

3.9 Evaluation Methods

3.9.1 Accuracy:

Calculate the number of correctly classified instances out of all instances. Works well for balanced datasets but may be biased otherwise!

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})}$$

3.9.2 Precision:

It measures the proportion of true positive predictions out of all positive predictions. This considers the correctness of the positive predictions alone and helps in scenarios where the cost of false positives is extreme.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

3.9.3 Recall (Sensitivity):

It measures the proportion of the number of true positive predictions against that of all actual positives. It focuses on capturing all the positive instances and is useful when the cost associated with a false negative is high.

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

3.9.4 F1 Score:

It gives a balance between recall and precision, and it provides a balance between the two, especially when one is working with classes that are unevenly distributed.

$$\text{F1 Score} = \frac{2\times\text{Precision}+\text{Recall}}{\text{Precision}\times\text{Recall}}$$

3.9.5 Confusion Matrix:

It is just a confusion matrix recording the number of true positives, true negatives, false positives, and false negatives of the predictions given by a classification model. In this way, it presents us with a notion of how our model behaves within every class represented.

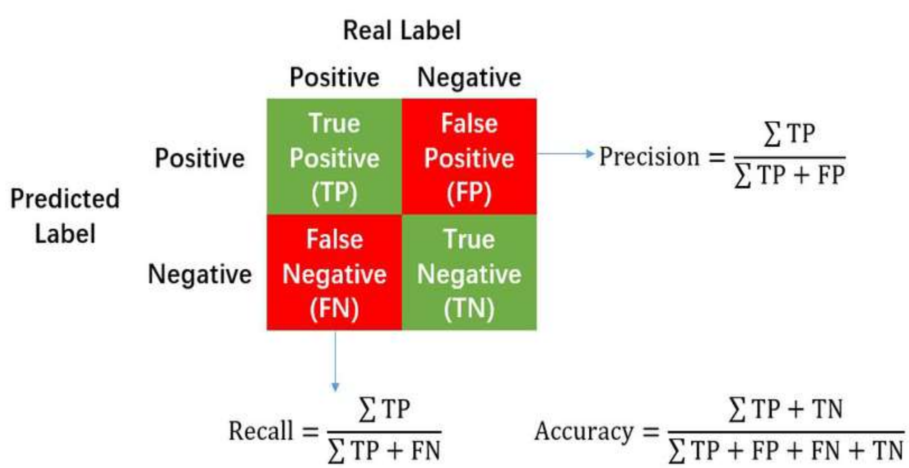

Figure 3.5: Confusion Matrix

3.9.6 Mean Squared Error

Now you have something to compare your models with a baseline metric, the mean squared error. The simple linear regression model simply squares the difference between the real values and the estimated ones and then averages them.

3.9.7 R^2 value

R2 is referred to as the statistical measure, more commonly called the Coefficient of Determination. Typically, it turns out to be the first number mentioned when results are communicated after regression analysis has been carried out. It ranges from 0 to 1; where 0 would mean that the model does not explain any variances that are there in the target variable and 1 indicates that the model explains all the variances perfectly.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

A model with an R-squared of 0.8787 would then suggest that about 87.87% of the variance in the dependent variable is predictable from the independent variables involved in the model, as represented in XGBRegressor.

CHAPTER 4

4. RESULTS AND DISCUSSION

4.1 INTRODUCTION

In the following section, based on data collection and preprocessing previously discussed, we discuss the results of using various models for Vending machine sales analysis over detection techniques. Based on the sales data, models were selected for detecting patterns, and anomalies and predicting potential insights which are buried deep in the sales data.

4.2 RESULT DISCUSSION

We fitted our dataset using KNN, Linear Regression, XGBRegressor, and Decision Tree models, and then evaluated each model by different parameters and showed it graphically. These estimations help users to see how good each model is in predicting sales in vending machines. By carefully evaluating the models using statistical parameters such as R Squared value and Mean Squared Error, and visualizing the weights and biases of measurement, we indeed got a clear view of the performance of each model. "Artificial Intelligence/Machine Learning (ML) is used to optimize the performance standard based on sample data or past Experience" according to Ethem Alpaydin.

4.2.1 XGBRegressor

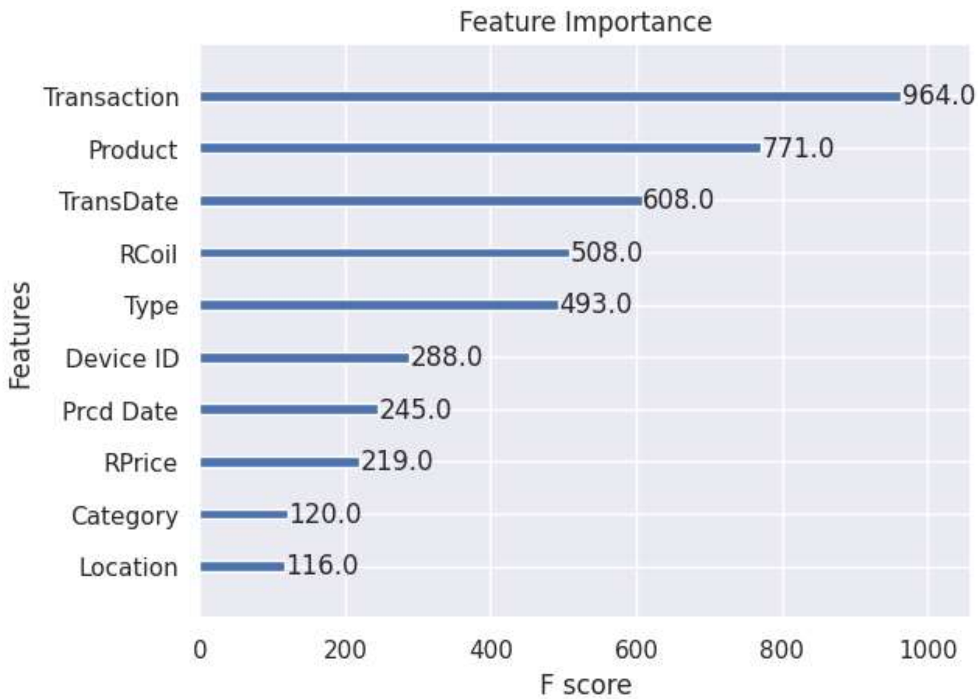

Figure 4.1: Output of XGBRegressor

The bar chart "Feature Importance" describes the importance of various features in a model. Of these, "Transaction" is the most influential feature with an F score of 964.0, whereas "Location" is the least, with an F score of 116.0. Other important ones are "Product" at 771.0 and "TransDate" at 608.0. The chart ranks these features through their impact on the model in a very organized way. The more important the feature, the longer its corresponding bar.

4.2.2 Linear Regression

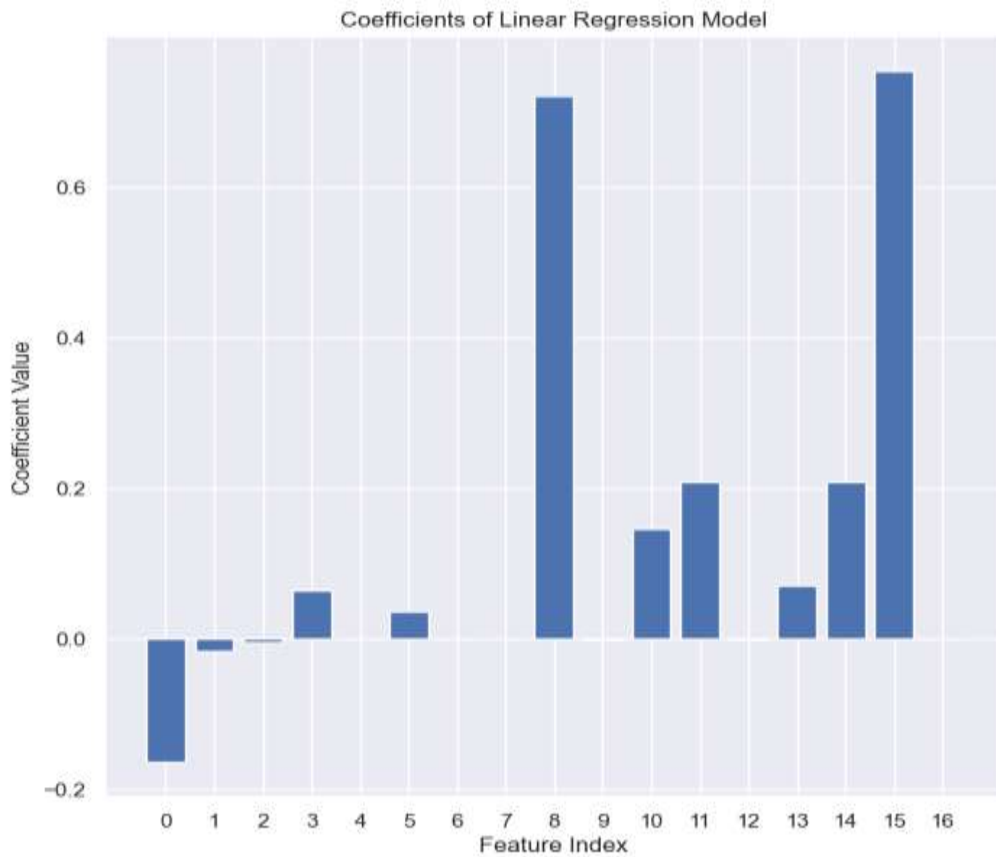

Figure 4.2 Output of Linear Regression

The Bar chart "Coefficients of Linear Regression Model" expresses the coefficients of different features in a linear regression model. On the x-axis, there is the index of a feature, and on the y-axis—the corresponding coefficient values. The plot shows that there exist large positive coefficients for some features, for example, indexed 8 and 16, while the coefficient of the feature at index 0 was a very small negative. Level height shows the importance of each feature. The higher the bars, the more positive/negative each contribution towards the model's predictions.

4.2.3 k-nearest neighbor (kNN)

```
accuracy_score on train dataset :  0.9872232952013074
Precision: 0.9746098345881277
Recall: 0.9872232952013074
F1 Score: 0.9808760162399354
```

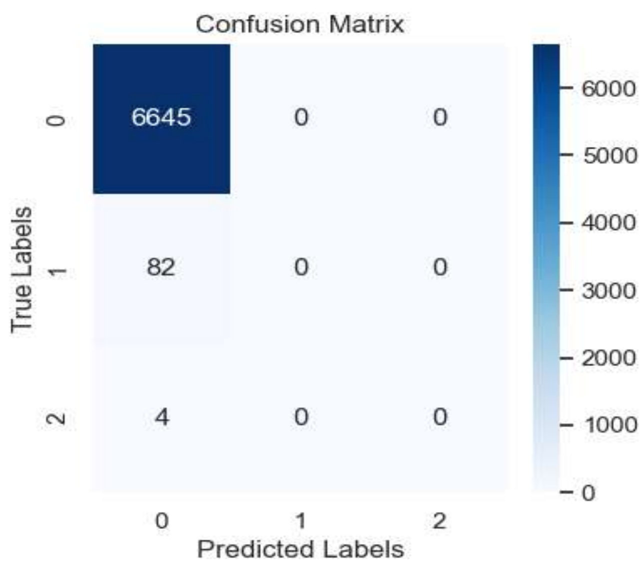

Figure 4.3: Result Of Knn

As shown in the confusion matrix in the first image, the model has predicted all the samples as class '0' (6645 correctly) but failed to predict any samples as class '1' or '2'. The second image gives the following performance metrics of the model on the training dataset. Accuracy: 0.9872, Prec.: 0.9746, Recall: 0.9872, F1 Score: 0.9809. It means that basically the performance of the model is good, but from the confusion matrix, there is high bias towards class '0', missing the prediction of classes '1' and '2'.

4.2.4 Decision Tree

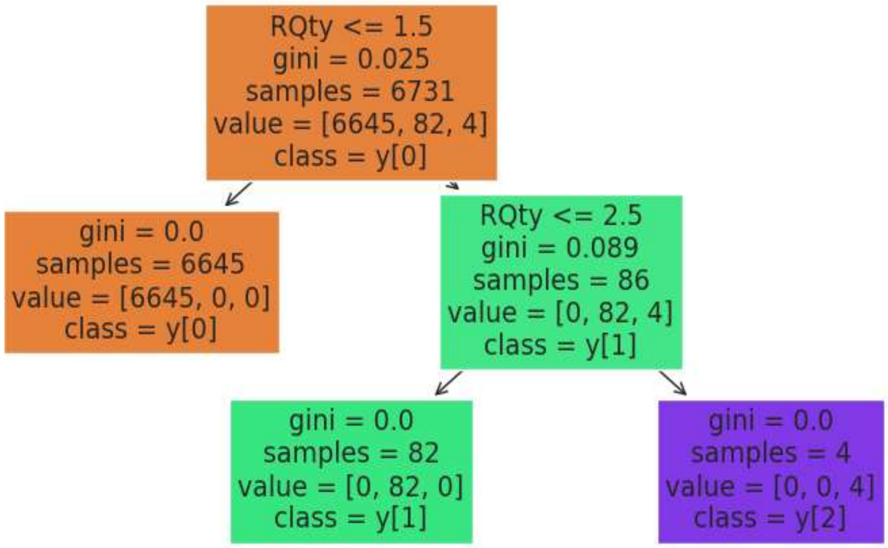

Figure 4.4 Output of Decision Tree

The following shows a classification decision tree: The root node splits using 'RQty <= 1.5' with a 0.025 Gini impurity and 6731 samples. This grows three-terminal nodes. The leftmost terminal node borrows 6645 samples classified as class '0' with a Gini impurity of 0.0. The right branch is further divided at 'RQty <= 2.5', growing two-terminal nodes. One node provides a classification of 82 samples as class '1' and the other as class '2' with 4, both with a Gini impurity of 0.0.

4.2.5 Heatmap

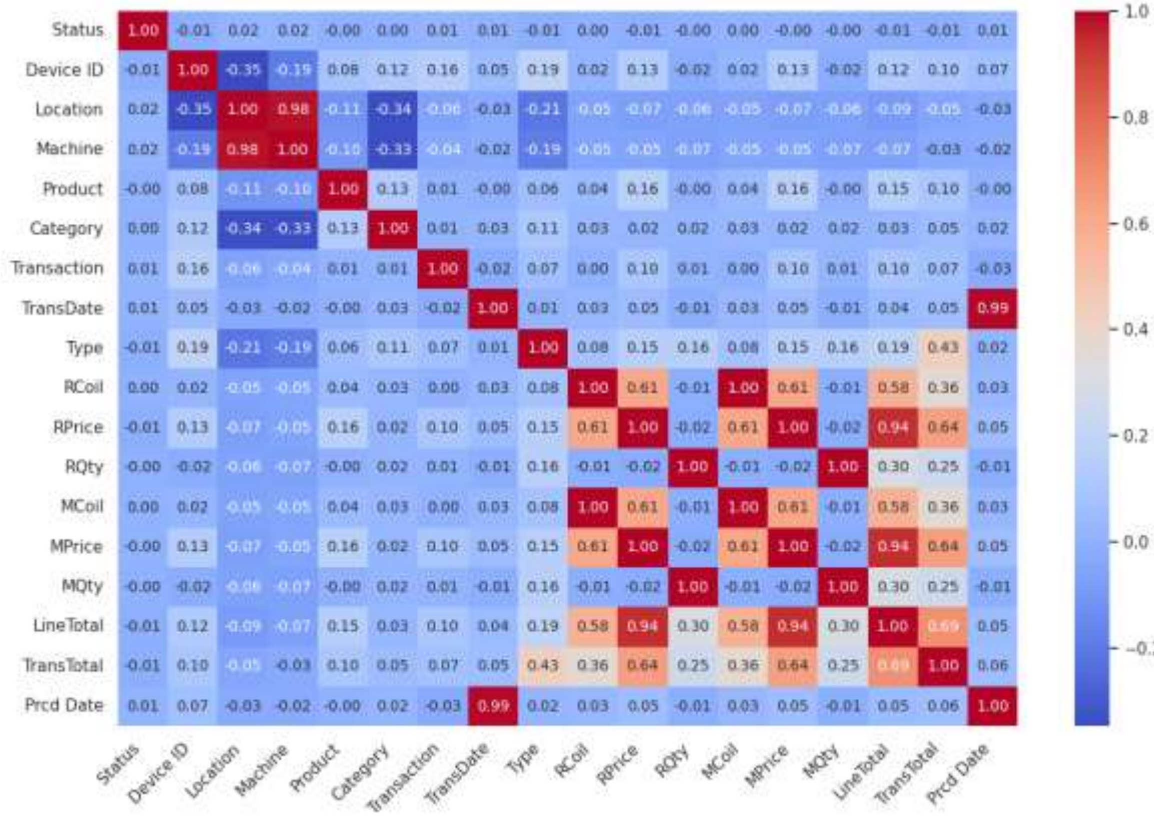

Figure 4.5: Heatmap

The heatmap shows the correlation between different variables. This correlation color intensity ranges from weak/ negative correlations, shown in dark colors, to strong positive correlations in light colors. Some important observations to note are the strong positive correlations of variables like RPrice and LineTotal, MPrice, and LineTotal, which show that these prices greatly influence total line and transaction values. Other correlations, such as RQty with Transaction, are much weaker, reflected by the darker colors.

CHAPTER 5

5. CONCLUSION AND LIMITATIONS

5.1 CONCLUSION

Vending machine sales data analysis using various machine learning models has returned useful information in terms of sales predictions and bringing efficiency into the operation of vending machines. Of all the models estimated in this study, XGBRegressor revealed the best predictive power, evidenced by its high R-squared value and the comparatively very low Mean Squared Error. Therefore, XGBRegressor would be a good candidate for modeling vending machine sales based on the features we have. The vending machine data itself enabled us to focus on both the customer side (the user experience) and the operational side, managing the actual vending machines. Places that I looked to collect data included libraries, malls, offices, manufacturing sites, and so on to understand what really influenced sales in volume and product quantity. I have used a few of these metrics to evaluate the effectiveness of each model: accuracy, precision, recall, and R-squared value. These performance metrics served as a foundation for making judgments about how well models predict vending machine behavior. Decision Trees offer insights into sales patterns and consumer preferences for the vending machine operator to arrive at an informed decision in deciding on the management of inventories and products.

5.2 LIMITATIONS

This sample dataset by this study may not be indicative of all possible scenarios since it was drawn by design from a relatively small number of vending machine locations. Note further that the models as developed currently are only on these features and most likely do not capture all the variables that could be associated with vending machines' overall sales and therefore do not reside in the files. We implemented models of various complexities: Linear Regression, Decision Tree, XGBRegressor, KNN, etc. Less complex models like Linear Regression or Decision Trees could not realize complex relationships between the data. Further, I'll try to collect less biased and bigger datasets, use more features that may influence sales, and validate the models in real-world experiments.

CHAPTER 6

REFERENCES

[1] D. M. K. Natalya V. Razmochaeva1, "Data Presentation and Application of Machine Methods for Automating Retail Sales," pp. 1444-1448, 2019.

[2] S. Cheriyan, "Intelligent sales prediction using machine learning techniques," *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE,* pp. 53-38, 2018.

[3] "Vending Machine Sales," Central New Jersey, 2022.

[4] M. Mnyakin, "An Empirical Investigation of the Determinants of Vending Machine Sales Volume," *Journal of Artificial Intelligence and Machine Learning in Management,* vol. 2, 2020.

[5] H. N. M. H. M. Y. O. Hidetaka Sakai*, " Development of a fuzzy sales forecasting system for vending machines," *Computers & Industrial Engineering,* pp. 428-449, 1999.

[6] H.-W. Y. C.-H. H. T.-C. W. Feng-Cheng Lina, " Recommendation system for localized products in vending machines," *Expert Systems with Applications,* 2011.

[7] V. M. I. V. M. ,. V. S. B. Polina S. Deryabina, "Challenges and Perspectives of Recommender Systems in Vending Machines," pp. 284-289, 2021.

[8] P. C. Mary V. Brown, P. Matthew Flint and B. James Fuqua, "The Effects of a Nutrition Education Intervention on Vending Machine Sales on a University Campus," *Journal of American College Health,* pp. 512-516, 2014.

[9] Y. T. K. 1. Yu Cao1, "A Real-Time Application for the Analysis of Multi-Purpose Vending Machines with Machine Learning," pp. 1-17, 2023.

[10] L. C. W. M. 3. Meagan L Pharis, "Sales of healthy snacks and beverages following the implementation of healthy vending standards in Cityof Philadelphia vending machines," *Public Health Nutrition,* pp. 1-17, 2016.

[11] T. S. Nilani Ratnasria, *Vending Machine Technologies: A Review Article,* pp. 160-166, 2022.

[12] P. W. F. a. J. R. F. Norm Borin, "A Model for Determining Retail Product Category Assortment and Shelf Space Allocation," pp. 1-26.

[13] Y.-B. P. a. H.-S. Park, "Smart vending machine systems: operation and performance," *European J. Industrial Engineering,* pp. 409-431, 2015.

[14] M. M. K. B. a. M. R.-Š. Bohanec, "Explaining machine learning models in sales predictions," *Expert system with application,* 2017.

[15] P. R. R. S. S. V. S. &. S. N. Bajaj, "Sales prediction using machine learning algorithms.," *International Research Journal of Engineering and Technology (IRJET),* pp. 3619-3625., 2020.

[16] Z. (. M. Huo, "Sales prediction based on machine learning," *2nd International Conference on E-Commerce and Internet Technology (ECIT) ,* pp. 410-415, 2021.

[17] M. K. Y. S. P. U. S. S. V. &. B. S. Gurnani, "Forecasting of sales by using fusion of machine learning techniques," *International Conference on Data Management, Analytics and Innovation (ICDMAI),* pp. 93-101, 2017.

[18] Z. Shilong, "Machine learning model for sales forecasting by using XGBoost.," *IEEE international conference on consumer electronics and computer engineering (ICCECE) ,* pp. 480-483, 2021.

[19] C. S. M. P. P. &. G. S. Wu, "Comparison of different machine learning algorithms for multiple regression on black friday sales data.," *IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) ,* pp. 16-20, 2018.

[20] R. Odegua, "Applied machine learning for supermarket sales prediction," *project: predictive machine learning in industry.,* 2020.

[21] M. A. R. S. S. &. Z. M. Soltaninejad, "Using Machine Learning Techniques to Forecast Mehram Company's Sales," *A Case Study. Journal of Business and Management Studies,,* pp. 42-53, 2024.

[22] S. K. C. S. &. J. T. Sharma, "Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach.," *Information Systems and e-Business Management, ,* pp. 261-284., 2019.

[23] S. G. G. T. &. U. S. Kohli, " Sales prediction using linear and KNN regression.," *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI ,* pp. 321-329, 2019.

[24] G. &. N. N. Behera, "A comparative study of big mart sales prediction.," *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, ,* pp. 421-432, 2019.

[25] R. Odegua, " Applied machine learning for supermarket sales prediction.," *project: predictive machine learning in industry..*

[26] F. A. M. A. A. A. A. A. F. S. M. &. G. E. A. A. Rao, "BMSP-ML: big mart sales prediction using different machine learning techniques.," *AES International Journal of Artificial Intelligence,,* pp. 12(2), 874., 2023.

