Cairo University
Faculty of Computers & Artificial Intelligence
Operations Research & Decision Support Department

Availa

The Graduation Project Submitted to
The Faculty of Computers and Artificial Intelligence,
Cairo University In Partial Fulfillment of the Requirements for
the bachelor's degree

In

Operations Research and Decision Support

Presented By

Reham Rashad El-Bayoumi (20200190)
Mariam Tarek Mohammed (20200524)
Mariam Ahmed Gharib (20200517)
Dina Alaa Helmy Abd El-Gawad (20201231)

Under Supervision of:

Prof. Dr. Ihab El-Khodary

CAIRO UNIVERSITY
July 2024

Availa

Reham Rashad El-Bayoumi (20200190)
Mariam Tarek Mohammed (20200524)
Mariam Ahmed Gharib (20200517)
Dina Alaa Helmy Abd El-Gawad (20201231)

Supervised By:

Prof. Dr. Ihab El-Khodary

CAIRO UNIVERSITY
July 2024

ABSTRACT

The disorganized display of goods in e-commerce poses significant challenges, leading to customer frustration and reduced sales for business owners. The E-commerce Optimization Tool addresses these issues by leveraging machine learning and data analytics to enhance product placement and dynamic pricing. The tool's primary objectives are to increase consumer satisfaction and maximize profitability by balancing customer preferences and operational efficiencies.

Market Basket Analysis (MBA) and Dynamic Pricing Optimization are key methodologies employed. MBA uses the Apriori algorithm to uncover purchasing patterns, enabling businesses to optimize product placement.

Dynamic Pricing adjusts prices based on market demand and customer behavior, utilizing machine learning models such as Linear Regression, Bayesian Ridge, Random Forest, and Decision Tree.

The integration of MBA and Dynamic Pricing generates association rules that inform pricing strategies, offering personalized product suggestions and discounts. This approach improves the shopping experience, increases sales, and enhances revenue. By providing dynamic, data-driven product placement and pricing, the E-commerce Optimization Tool significantly boosts both customer satisfaction and business profitability.

DECLARATION

We hereby declare that our dissertation is entirely our work and genuine / original. We understand that in case of discovery of any PLAGIARISM at any stage, our group will be assigned an F (FAIL) grade, and it may result in withdrawal of our bachelor's degree.

Group members:

Name Signature

Mariam Tarek Mohamed Amin ______________________

Mariam Ahmed Gharib Mohamed ______________________

Reham Rashad Ahmed Mohamed ______________________

Dina Alaa Helmy Abd El-Gawad ______________________

Contents

List of Figures

1. Chapter1: Introduction

1.1. Introduction

The world is growing day by day, with people and their desires to own and buy. On the other hand, there are people whose main job is to sell. As a result, people became from some point of view "who sells and who buys". A person buys to achieve his desire, and a person sells to achieve his gain... and this cycle continues indefinitely. Hence the problem of the availability of a fair purchase with a satisfied buyer impression. The solution to this problem is for the purchasing markets to display the products in the best way that would satisfy the buyer, add discounts from time to time according to certain factors, and the price must also be compatible with specific factors such as the demand for this product and the quantity required by this buyer, and according to its basic cost, including the seller’s lowest profit.

However, there has become a major problem between the unjustified greed of the seller, neglecting to satisfy the buyer, and the buyer’s never-ending desires, while facing the inflation of purchasing power due to the nature of life. In the face of this problem, the efforts of scientists and statisticians have united, trying to obtain algorithms, techniques, and ideal solutions to the problem of the complexity of the purchasing process.

The availability of a fair purchase with a satisfied buyer impression has been and continues to be a major problem, as it is a complex and important area of research. That is why we found many fields that are constantly interested in this problem because there is a solution to it, for example, fields such as procurement analytics, operations research, supply chain management, and behavioral economics

1.2. Problem Domain

Due to the change in customer behavior and increased Internet penetration, the e-commerce sector has witnessed rapid growth in the number of consumers and purchases. Businesses are always looking for ways to improve their operations and consumer happiness in this cutthroat market.

This chart shows global retail e-commerce sales from 2014 to 2026, with growth expected to continue in 2026.

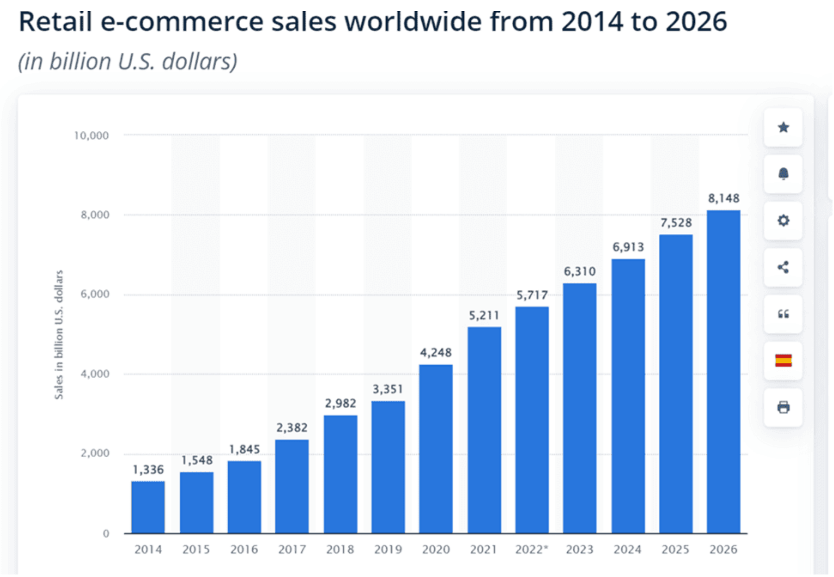


Figure 1(Global Retail E-commerce Sales from 2014 to 2026 (in Billion U.S. Dollars))

Therefore, market basket analysis and dynamic pricing are important methods that have emerged; Both can be used to maximize revenue, improve sales, and enhance customer satisfaction.

1.3. Problem Statement

The current practice of displaying goods in a random and disorganized manner poses significant challenges for both customers and business owners. Customers struggle to locate products efficiently, leading to frustration and a suboptimal shopping experience. This difficulty in product discovery not only diminishes customer satisfaction but also adversely affects demand, ultimately resulting in reduced sales for the business owner.

To address these issues, it is essential to implement a strategic approach to product placement and organization. By optimizing the display of goods, we aim to enhance the ease of product discovery for customers, thereby increasing their satisfaction and encouraging repeat business. Furthermore, an organized and logical product display can significantly contribute to maximizing the business owner's profit by boosting sales and improving overall operational efficiency.

1.4. Proposed System:

1.4.1. Aims and Objectives:

The primary objective of the E-commerce Optimization Tool is to transform the market by leveraging advanced technologies such as machine learning and data analytics to address key challenges faced by businesses, particularly those related to product organization and customer satisfaction. The tool aims to enhance consumer satisfaction by providing customized solutions in essential areas such as optimized product placement and dynamic pricing.

Achieving this goal necessitates balancing multiple considerations, such as customer preferences and operational efficiencies, to establish a system that maximizes profitability, reduces costs, and fulfills all necessary requirements and needs.

To achieve this aim, the objectives of the E-commerce Optimization Tool include:

Developing Market Basket Analysis and Consumer Behavior Modeling:

- Gain deep insights into customer preferences and purchasing habits.
- Recommend products that may purchase together to ensure customers can easily locate desired items, enhancing their shopping experience.

Implementing a Dynamic Pricing Optimization System:

- Utilize predictive analytics and optimization models to determine the most effective pricing strategies.

The overarching goals of the E-commerce Optimization Tool are to improve the shopping experience for consumers and maximize operational efficiency. To accomplish these aims and objectives, the tool employs a variety of techniques and algorithms, such as predictive analytics, machine learning models, and optimization strategies, to identify optimal solutions that address all necessary constraints and objectives.

1.4.2. Proposed System Features:

- Analyze products and find relationships between them using Apriori algorithm in the Market Basket Analysis technique.
- Use these relations to make suggestions to customers about products they usually buy, enhancing customer satisfaction through personalized recommendations.
- Utilize these insights in dynamic pricing to predict the prices of products that are commonly sold together.
- Implement dynamic pricing to create offers and discounts for products that are usually sold together. When a customer buys one product, they offer the remaining related products at a reduced price, increasing overall profit for the store owner.
- Provide businesses with deep insights into customer preferences and purchasing habits through advanced market basket analysis and consumer behavior modeling.
- Enable businesses to tailor their product offerings and marketing strategies to increase customer loyalty by offering a recommendation engine based on past purchasing behavior and current trends.
- Consequently, improve the warehouse system indirectly by ensuring that inventory is quickly rotated, reducing the time products remain in storage and lowering costs for the owner.
- Consequently, implementing these features will accelerate product turnover, reducing inventory costs and increasing overall efficiency.

1.5. Development Methodology:

The e-commerce optimization tool helps sellers improve the buying and selling process for consumers by creating better dynamic pricing for products and improving the appearance of products on the platform.

Data Collection and Data Preprocessing:

Data Collection: This data was found on Kaggle, which is data for a Turkish e-commerce website, that includes many products in different categories.

Data Cleaning: It encodes the categorical data, changes the data format so that there are uniform standards and removes the non-numerical parts in numerical attributes. Furthermore, it drops the unimportant qualities so that it becomes easy to work with the model.

Feature Engineering: Create a new feature called 'demand' to be able to have better understanding into customer behavior and efficient pricing.

Data Interpretation:

Exploratory Data Analysis (EDA): Descriptive analysis to understand the quality of information, capturing any of the outliers and evaluating skewness of the figures. Descriptive statistics of the variables of the categorical and numerical variables were computed.

Correlation Matrix: This would be developed as a way to find the relationship among various attributes that would identify critical factors that drive price and customer behavior.

Association Rule Mining:

Association Rules: Association rules were generated prior to running the Apriori algorithm for the identification of relationships between the items to a considerable extent that are bought together. The rules provided the first insight into the purchase pattern by the customer.

Given the data, run the Apriori algorithm to mine market basket information at a much finer level, after which it would be easy to notice what things are typically purchased together. Such an implementation would hence lead to deeper insight into customer purchasing behavior and what they like.

Implement Dynamic Pricing Model

Machine Learning Models: Of the few tested and being tested, some include Decision Tree, Bayesian Ridge, Linear Regression, Random Forest, XGB and CatBoost.

Model Training: All these have been cleaned and preprocessed and are finally ready to be dumped into the models that will maximize revenue and satisfaction using optimal pricing strategies.m

Selecting Models and Evaluation Metrics:

Performance Evaluation Metrics: The models were evaluated on R-squared, mean squared error, sqrt mean squared error, and mean absolute error.

Model Selection: Cross-evaluate the models; based on this, select the best model for deployment.

1.6. Resource Requirement:

The resource requirements for Availa can vary depending on the size of the market customers and sellers, and the number of unique products categories. Some of the key resources required for Availa are:

Hardware: our tool will require hardware resources such as servers, storage devices, and networking equipment to host and run the software.

Software: For the system to support the Availa tool, software resources including operating systems, databases, and web servers are needed.

Development Team: The development team will require personnel resources such as project managers, software developers, data analysts, ML engineers and quality assurance engineers to design, develop, and test the Availa tool.

Data: The system will require data resources such as transaction dates, data for each product in each transaction, product data, customers' data, and sellers' data.

Training: The tool will require training resources to ensure that each user can deal with any software our tool was added to, and each seller and customer will be satisfied with the pricing strategies

Support: The tool will require ongoing support resources such as technical support staff to address any issues or concerns that arise.

Budget: The system will require financial resources to cover the cost of hardware, software, personnel, data, training, and support.

Overall, the resource requirements for the Availa tool can be significant, but the benefits of such a tool can outweigh the costs. By efficiently helping customers to purchase what they need and achieve fair pricing, that will solve the availability of a fair purchase with a satisfied buyer impression problem.

1.7. Report Layout:

1. **Introduction:** This section provides background information about the E-commerce Optimization Tool, including the purpose, objectives, and scope of the project.

2. **Data Collection and Cleaning:** This section describes the data sources used, the methods for collecting data, and the processes for cleaning and preprocessing the data to ensure its quality and usability.
3. **Data Exploration:** This section presents the initial exploration of the data, including key metrics, patterns, and insights discovered through exploratory data analysis (EDA), supported by visualizations.
4. **Model Development:** This section details the selection, training, and validation of machine learning and optimization models, as well as their implementation to achieve the project’s objectives in areas like pricing optimization.
5. **Conclusion:** This section summarizes the key findings and outcomes of the project, discusses its contributions to the e-commerce market, and suggests areas for future work and potential improvements.
6. **References:** This section lists all the sources cited in the report, as well as any additional resources or tools used, following a consistent citation style.

2. Chapter2: Background/Existing Work

2.1. Introduction

The concept of fair purchase and buyer satisfaction is integral to developing effective pricing strategies within the e-commerce market. Fair purchase refers to the perception that the price paid by the consumer is just and reasonable, while buyer satisfaction encompasses the overall contentment of the customer with their purchasing experience. Achieving both fair purchase and buyer satisfaction is crucial for maintaining a healthy business environment, as it fosters trust, encourages repeat business, and enhances brand loyalty.

In the competitive landscape of e-commerce, businesses must navigate the delicate balance between setting prices that maximize profitability and ensuring customers feel they are receiving fair value for their money. The challenge lies in implementing pricing strategies that are both fair and satisfying to buyers, which in turn drives sustainable growth and long-term success.

This project aims to find the best balance between these two essential aspects. By leveraging advanced data analytics and machine learning, the E-commerce Optimization Tool seeks to develop dynamic pricing models that prioritize consumer satisfaction. This balance is key to creating a thriving e-commerce ecosystem where businesses can prosper while maintaining strong, positive relationships with their customers.

2.2. Overview of project

This project aims to revolutionize the e-commerce market by developing an advanced E-commerce Optimization Tool. The core focus of the project is to utilize market basket analysis and dynamic pricing strategies to address the key challenges faced by businesses in this industry. Market basket analysis will provide deep insights into customer purchasing patterns, helping businesses understand which products are often bought together and how to effectively bundle or promote these items. This analysis will enable businesses to tailor their offerings more precisely to meet customer preferences.

Dynamic pricing strategies, on the other hand, will allow businesses to adjust their prices in real-time based on several factors such as market demand, competition, and inventory levels. By implementing these strategies, businesses can optimize their pricing to maximize revenue while keeping fairness and customer satisfaction.

The expected outcomes of the project include improved inventory management, increased profitability, enhanced customer satisfaction, and a stronger competitive position in the market. The project looks to create a more efficient and dynamic e-commerce environment, where businesses can thrive, and consumers enjoy a better shopping experience.

2.3. Limitations of project

Much like any other technology, the E-commerce Optimization Tool has multiple advantages and certain constraints, too. Knowing the project limits helps me understand best how to improve and set proper expectations. In this section, data availability, computational complexity, and possible risks with dynamic pricing are described as the most important limitations. These are rather vital moving forward, which will indicate how workable and effective the tool may be.

1. **Data Availability and Quality:** For the E-commerce Optimization Tool, the concept of good data is fundamental. What lies behind this report is a dataset consisting of

millions of records with 27 attributes—a feature that is hard to come by. Getting such a vast and detailed dataset into place requires much work. So much work was needed, ensuring the data quality is maintained at very high levels in all aspects—complete and correct data. This was much needed to ensure our price and product recommendations were reliable and effective for improving our e-commerce strategy.

2. **Computational Complexity:** We encountered during the project computational complexity. The market-basket analysis is done using the Apriori algorithm. At the same time, the dynamic pricing is carried out by machine learning models that include Linear Regression, Bayesian Ridge, Random Forest, Decision Tree, XGB Model, CatBoost models. It is a complex and time-consuming job to process the data and train the models because of its voluminous nature. Hence, it is essential that the system adequately copes with this complexity to maintain performance.

3. **Dynamic Pricing Risks:** Dynamic pricing is full of associated risks that we need to handle with care. Fast price changes, according to market and customer behavior, impact customer perception of us if not readily transparent. Otherwise, he/she will show indecisiveness or dissatisfaction. It may further provoke competitors' reaction, which may result in price wars, hurt profitability, and stability. It is necessary to balance the risks against the rewards of dynamic pricing within the project.

4. **Technical Limitations:** We are pulled down by some of the technical limitations, Accuracy of predictive models and algorithms that get applied, for example, in dynamic pricing or market basket analysis, might differ. Although enhanced techniques improve prediction, they also are not faultless. Sometimes they also make errors. We need to overcome these limitations if our tools are to work correctly and effectively.

2.3.1. Innovations in project

Integration of Market Basket Analysis and Dynamic Pricing

To enhance our E-commerce Optimization Tool, we have implemented an innovative method that combines Market Basket Analysis and Dynamic Pricing for better results. This new combination uses the advantages of both approaches to offer customized and convincing pricing strategies that boost customer interactions and boost revenue.

Market Basket Analysis is a method of data mining that recognizes items commonly bought in conjunction by consumers. By utilizing the Apriori algorithm, we uncover association rules that expose these combinations of products. These observations help us grasp buying trends and customer likes.

Dynamic Pricing includes changing the prices of goods instantly according to factors like market demand, customer actions. Using machine learning techniques such as

Linear Regression, Bayesian Ridge, Random Forest, and Decision Tree models, we can predict ideal prices that increase both customer contentment and profits.

The way our innovative integration operates is as follows:

- Generation of Association Rules:
  The transaction data is analyzed with the Apriori algorithm to generate rules that show which products are commonly purchased together.

- Dynamic Pricing Application

  When a consumer is ready to buy a single item from a specific group of products, the pricing system adjusts accordingly.
  The system suggests related products to the customer at a reduced price. Our predictive pricing models are utilized to calculate these discounts in a dynamic manner, making sure they are both appealing and financially beneficial.

- Enhancement of the overall customer experience.

  This method offers clients personalized suggestions and appealing deals on items they are likely to desire.
  By providing discounts on interconnected items, we improve the chances of more sales, elevating the overall shopping experience and increasing revenue.

2.3.2. Design of project

The problem that our project would address is to ensure fair purchases that ensure a satisfied buyer impression. It would be performed by the amalgamation of market basket analysis with dynamic pricing strategies. Data will be collected, and analyzed, and specific algorithms implemented to predict and then adjust the price dynamically considering customer behavior and market trends.

Technical Approach:

Data Collection: We will use real data from e-commerce websites selling everything from A to Z, including good and useful futures which help us. I will mention the nature and form of data in the next point.

Data Preprocessing: Clean and preprocess the dataset to handle missing values, outliers, and inconsistencies:

- Normalization and scaling of data for uniformity in analysis.
- And that is done in the next points:
- Handle missing values
- Handle outliers
- Encodes the categorical data
- Modify the data format
- Dropping of non-numerical parts in numerical attributes
- Remove unimportant qualities
- Create new important features

Market Basket Analysis: Apply Apriori algorithm to come up with frequent itemset and association rules. In this way, it would be possible to understand the relationship between various products and the trend of customers purchasing habits. Use results of market basket analysis in identifying products that are often purchased together and recommend complementary products to customers.

Dynamic Pricing Strategies: Dynamic pricing is based on demand, customer behavior, and market conditions to set the price of a product or service. Predicting the optimal pricing points for maximum revenue with the least decline in customer satisfaction can be done with machine learning algorithms—for example, regression analysis. Elasticity of demand would refine the prices in real-time with any alteration in customer demand and competitor pricing.

Data will be used; we need data will help in four points:

- Know the frequent items set and how the customer deal with it, in which case.
- Know how the product be priced according to the frequency.
- Guide to the data to know the demand on this item according to time.
- Take info about the pricing before and after considering the factors like the demand and the amount to achieve later the fair market.

2.4. Existing Work:

1. **Amazon:** In 2000, Amazon may have violated the Robinson-Patman Act by charging different customers for the same goods at different times, thus engaging in pricing discrimination. Amazon did not stop the practice, even though the company apologized publicly and refunded nearly 7,000 customers after this episode drew criticism.

   Certain in-demand commodities were claimed to have experienced a fourfold increase in price during the COVID-19 pandemic, drawing unfavorable attention. Amazon blamed a few vendors for driving up costs for necessities like masks and hand sanitizers, but the company denied any manipulation of this kind. Nevertheless, the cost of critical products sold by Amazon' had also increased significantly. According to Amazon, this was caused by a software error.

2. **Uber:** Uber riders saw their fares increase eight times over after the 2013 storm that hit New York. Public personalities reacted negatively to this episode, with Salman Rushdie publicly denouncing this action, among others. Following this event, the business began capping the maximum amount that surge pricing might reach in an emergency in 2015. In certain areas, drivers have been known to wait to take rides until surge pricing drives rates up to a level that meets their needs.

3. **Wendy's:** Wendy's declared in 2024 that it will explore dynamic pricing in select American locations in 2025. Plans to alter menu boards were disclosed to stakeholders along with this pricing method Due to this action, the corporation faced a great deal of criticism online. Wendy's responded by saying that the planned implementation was restricted to price reductions during times of low traffic.

4. **Netflix:** Market Basket Analysis helps Netflix understand the content viewed by subscribers and provide better recommendations. The dynamic-pricing strategy adopted by Netflix involves market-oriented location pricing. That would increase gradually but surely by 10-12% with tiered pricing raises. You can see reasonable price adjustments going forward, making a balance between subscriber growth and profitability.

5. **Instacart:** Instacart, the same-day grocery delivery and pickup service, uses the market basket analysis to fathom customer shopping behavior in forming sets of items commonly purchased together to suggest reorders based on purchase history. Dynamic pricing involves surge pricing, real-time adjustment of price, which maximizes revenue in this manner through tie-up with Oversight for AI-powered price experiments. This strategy keeps convenience, pricing, and CX absolutely in balance to optimize the grocery shopping experience.

3. Chapter3: Methodology

3.1. Market Basket Analysis

3.1.1. Introduction

We produce enormous amounts of data in our daily lives. Although this data is constantly expanding, not all the information is meaningful. Utilizing the stored data to extract relevant information is essential. Data mining, also known as Knowledge Discovery in Data (KDD), is the process of obtaining valuable information from a big dataset. Data selection, preprocessing, transformation, mining, and interpretation are all part of this process.

Another name for market basket analysis is association rule mining. Large volumes of data are kept in databases during this process from a variety of industries, including marketing, banking, and medicine. For example, if a customer purchases bread from a grocery store, there is a chance that they will also purchase milk or jam. This will reveal the customer's purchase behavior.

Customers can easily purchase things using this method, and the shop can gain valuable insight into customer purchasing patterns and use this knowledge to make decisions based on the most popular items.

3.1.2. Market Basket Analysis Methodology

The Market Basket Analysis (MBA) technique has become a powerful instrument that provides a methodical, data-driven approach to identifying patterns in consumer purchase behavior. MBAs are now useful in a variety of industries, having originated in the retail industry and given practical insights on product relationships, cross-selling opportunities, and consumer trends. The understanding that businesses functioning in the competitive environment of today cannot depend just on intuition and traditional market research techniques motivates this endeavor. Strategic information obtained through a thorough Market Basket Analysis can guide decision-making, enabling businesses to develop tailored product offerings, and launch focused marketing efforts.

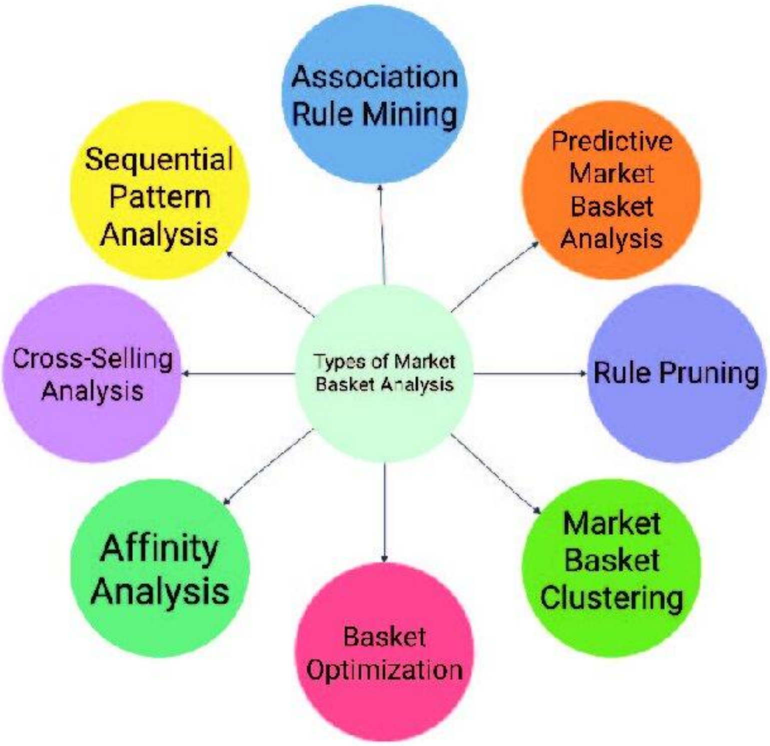


Figure 2(Different types of Market Basket Analysis utilized in Business)

Market Basket Analysis (MBA) is a data analytics process that finds trends in the things that customers typically buy together. There are several MBA options, and each offers unique perspectives to understand consumer behavior and guide business strategy. Primary forms consist of:

1. **Association Rule Mining:** Finding interesting associations between dataset elements is the core of the cardinal MBA technique
2. **Sequential Pattern Analysis:** This method looks at purchase orders over an extended period and goes beyond simple association rules
3. **Cross-Selling Analysis:** This variation makes the most of cross-selling strategies by emphasizing tandem item purchases.
4. **Affinity Analysis:** This technique, which is similar to association rule mining, stresses the strength of the links between the parts.
5. **Basket Optimization:** The goal is to maximize product placement in retail locations or online by utilizing identified item linkages
6. **Market Basket Clustering:** This technique finds segments with shared tastes by clustering customers according to their past purchases.
7. **Rule pruning**: is the process of fine-tuning association rules to identify the most pertinent and helpful information
8. **Predictive MBA:** Using past transaction data and pattern recognition to predict future consumer behavior.

To maintain a competitive edge, companies must perform a Comprehensive Market Basket Analysis (MBA) to interpret nuanced details, overlooked patterns, and intricate relationships that influence consumer decisions. Understanding the complex relationships between the objects in baskets is not only an academic endeavor but also a crucial strategic need for businesses hoping to achieve long-term success. The driving force is the belief that companies can gain a competitive edge by using cutting-edge data analytics and complex algorithms to solve the riddles of customer behavior. The research findings could help businesses better focus their marketing campaigns, streamline inventory control, and customize goods and services to the unique needs and preferences of their target group.

Businesses of all sizes can benefit from this by understanding the intricate processes influencing consumer decisions through an extensive examination of linked purchase data.

acquire useful market intelligence to fight tactically in a volatile environment. This study is more than just a theoretical undertaking; it aims to give company leaders useful insights and the tools they need to quickly adjust to changing consumer trends.

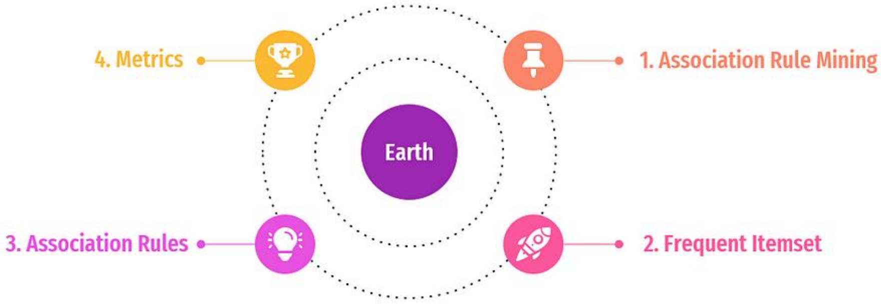


Figure 3(How Market Basket Analysis Work)

3.1.3. Data Mining

Thus far, data mining techniques have proven effective in a variety of industries. For example, in the field of marketing, health insurance, fraud detection, etc.

Furthermore, data mining techniques can be used to stop fraud and abuse in other fields. In the healthcare industry, doctors identify acceptable methods and best practices, pharmaceutical companies may improve their consumer service, and patients receive safer and more reasonably priced medical care Data mining has been incredibly useful recently in the marketing industry, particularly for consumer segmentation and basket analysis.

Customer segmentation is breaking down the entire client pool into smaller consumer segments, each of which has the same customers. This segmentation method is helpful for locating and classifying clients according to their characteristics.

3.1.4. Association Rules Mining

Another name for association rule mining, or association analyses, which are frequently used in market basket analyses, is affiliation analysis, or association rule extraction. Currently, AR is the most convenient way for analyzing market basket datasets with many sales transactions. A recorded transaction ledger contains a list of items for every transaction. In general, a transaction is an arrangement, a contract, or a tiny amount of a sale. In the marketing industry, a typical purchase is a group of products acquired from a sales store. Typically, a certain transaction's whole set of details is recorded into the database for storage. containing the costs of the items, their quantities, some user data... etc.

The association between things in a single supermarket or E-Commerce is known as the Association Rule. Let us talk about a few concepts before discussing the support and confidence values. Let I be a set of n attributes called items, I = {i1, i2, . . ., in}. Let D = {t1, t2, . . ., tm} which is a set of m transactions (the database). Every transaction in D has a distinct transaction ID and includes a portion of the items in I. A connotation of the type A → B, where A, B ⊆ I and A ∩ B = Ø, is defined as a rule. In terms of supermarket data, the rule A → B states that purchasing A (which might be an individual item or a series of goods) implies purchasing B (which can also be an individual item or a set of items). Another way to interpret this rule would be to say that a customer is more likely to purchase item B in the future if they purchase item A. Furthermore, rules are frequently limited to the one thing in the consequent (rule's right-hand side).

The rule of association the learning the result matrix is usually large and dispersed (this varies depending on the AR technique used), but this necessitated the use of specific evaluation metrics. Two indices—the strength of the rule and its presence or nature—generally characterize the existence, intent, and potency of an association rule. Support and confidence are the association rule's most well-known quality metrics. These two metrics are typically used to extract association rules.

3.1.4.1. Support:

The percentage of transactions in the dataset that make up the item or itemset helps the item or itemset. Certain rules might not be very useful in certain situations. Because it would not be profitable to market things that are "seldom bought together," a rule with little support may also have limited commercial value. displays the support value for the ratio of revenues between sales slips in database D that contain A and B.

$$\text{Support} = \frac{number\ of\ transactions\ with\ items(s)}{number\ of\ transactions}$$

3.1.4.2. Confidence:

The rule's strength is indicated by its confidence value. It is defined as the events that show the right-hand transaction is included in the left-hand transaction. An interaction statute's presumption of causation is not always true. Rather, it displays a distinct co-occurrence between the objects that came before and after, together with the algorithm for determining the confidence value.

$$\text{Confidence} = \frac{Support(X\&Y)}{Support(X)}$$

3.1.4.3. Lift:

Lift is a statistic that evaluates the degree of correlation between two item sets (A and B) while accounting for the likelihood that B may occur randomly or as expected given A. It is employed to ascertain whether there is a more substantial correlation between A and B than would be predicted by chance.

When A is present, there is a positive impact on the probability that B will be purchased (lift larger than 1), indicating a stronger correlation between A and B than would be predicted by chance.

If the lift is 1, then there is no additional relationship between A and B than what would be predicted by chance.

A lower correlation between A and B is shown by a lift value of less than 1, which implies that the existence of A has a detrimental impact on the probability of B being purchased.

$$\text{Lift} = \frac{Support(X\&Y)}{Support(X)*Support(Y)}$$

3.1.4.4. Leverage:

Leverage quantifies the discrepancy between the frequency at which two things (X and Y) occur together in transactions as seen and the frequency that would be predicted if X and Y were unrelated to one another. To put it another way, it measures the degree to which the combined incidence of X and Y differs from what would be predicted by chance.

- When X and Y occur together more frequently than one would anticipate by chance, a positive leverage value suggests a favorable link.
- When the leverage value is zero, it indicates that there is no correlation and that the co-occurrence of X and Y is as expected under independence.
- A negative leverage value denotes a negative association or avoidance since it means that X and Y together occur less frequently than would be predicted.

$$Leverage\ (X \rightarrow Y) = Support(X\&Y) - Support(X) * Support(Y)$$

3.1.4.5. Conviction:

The ratio of the observed frequency of X occurring without Y to the expected frequency of X occurring without Y is known as the conviction. It basically measures the degree to which the existence of X requires the absence of Y. An uncommon occurrence of A without B is shown by X high conviction value, which also shows how strongly dependent the rule X => Y is.

1. Since X's presence has no effect on Y's presence or absence, a conviction value near 1 indicates that there is no relationship between X and Y.
2. significant dependence between the presence of X and the presence of Y is implied by a high belief value (> 1), which suggests a significant link.

3. A low conviction value (<1) denotes a weaker relationship, indicating that the presence of X is not highly dependent on the presence of Y.

$$Convictions\ (X \rightarrow Y) = \frac{Support(X)Support(\bar{Y})}{Support(X\&\bar{Y})}$$

3.1.4.6. Zhang’s Metric:

Zhang's metric is a metric that evaluates the degree of correlation—whether positive or negative—between two items by accounting for both of their co-occurrence and non-occurrence. It is especially helpful when we want to know how item A's presence or absence influences the possibility that item B will be present in a transaction.

In practical terms, Zhang's measure can be used to evaluate the co-occurrence and non-co-occurrence of two items or itemset in transaction data to determine their relationship. Negative values imply a negative association or avoidance, whilst positive values show a good association. A value near zero signifies the absence of any noteworthy correlation.

To determine Zhang's measure, divide the numerator by the denominator. This guarantees that the value of the metric lies between -1, which denotes a strong negative correlation, and 1, which indicates a strong positive association, with 0 denoting no association.

Remember that Zhang's metric can be a useful tool in situations where traditional measures like confidence and lift may not fully represent the depth of item relationships. It provides a nuanced way to quantify association and dissociation between items.

$$Zhang(A \rightarrow B) = \frac{Confidence(A \rightarrow B) - \ Confidence(\bar{A} \rightarrow B)}{\mathrm{Max}[Confidence(A \rightarrow B), Confidence(\bar{A} \rightarrow B)]}$$

$$\mathrm{Confidence}\ = \frac{Support(X\&Y)}{Support(X)}$$

3.1.5. Apriori Algorithm

As opposed to the "soul mate" algorithm, the Apriori Algorithm [Agrawal and Srikan 1994, Agrawal et al. 1996] defines the market basket by considering every transaction in the database. Using association rules, the market basket can be represented as a left and a right side, Left→Right. For example, the rule {B, C} ⇒{A} should be understood as follows given an item set {A, B, C}: if a client bought {B, C}, he would presumably buy {A}. When the following rule was discovered—"on Thursdays, grocery store customers often purchase diapers and beer together"—it gained popularity. This method was first applied in pattern recognition [Berry and Linoff 1997].

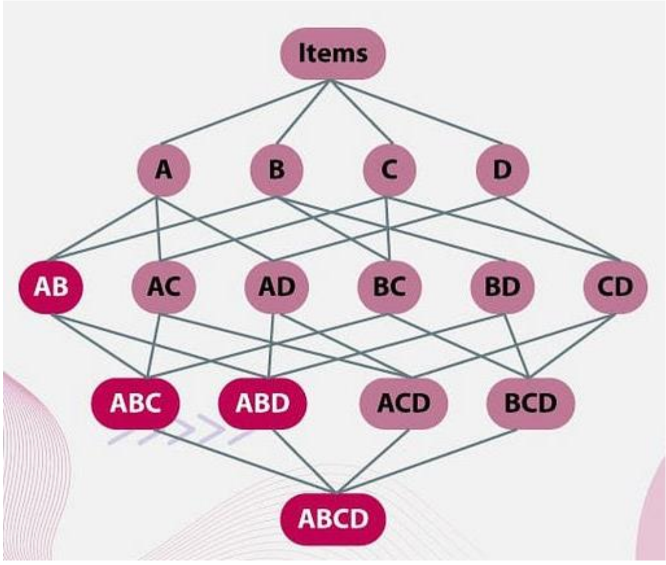


Figure 4(Lattice Structure of Itemset Combinations in Market Basket Analysis)

The support measure and the confidence measure are two metrics that can be used to assess the association rules. Assume that {A, B} is an itemset and that the association rule is A⇒B. The probability P(A∩B) or the relative frequency is the same as the support measure. The conditional probability of B given A, P(B|A), which is equal to P(A∩B)/P(A), provides the confidence measure. According to the rule "diapers=>beer," which has a 50% support measure and an 85% confidence measure, 50% of transactions involve the purchase of both diapers and beer, and 85% of those who make diaper purchases also make beer purchases, or P(beer | diapers) =85%.

Commercial products like SAS Institute's Enterprise Miner [SAS 2000] included the Apriori algorithm. This method takes a table containing buy transactions as input. Every transaction has the following information: transaction, item; it also includes the things that were purchased. The minimal support (min sup) and maximum number of acquired items (max_k) of a certain basket are the definitions of input parameters. This algorithm was built using the SAS Institute's Enterprise Miner package, with default values of max_k=4 and min sup=5%. Because it will eliminate the pointless search branches, the minimum support value (min sup) is crucial to the algorithm.
A parameter called minsup is employed to regulate the exponential algorithm's combinatorial development. Because Apriori only employs one value for the minsup, certain basket sizes might be tiny, and others might be excessive.

The Apriori algorithm generates sets of market baskets in its first step. The set of often purchased items that total k items is designated as Ik. First, the algorithm generates I1 by filtering the items whose frequency is higher than the minsup. The Ik+1 candidates, such as Ik⊆Ik+1, are generated for every Ik in the subsequent phases. The method eliminates the baskets that are less than the minsup for each Ik+1 candidate. When the cycle reaches Imax_k, it is over.

The Apriori algorithm creates sets of market baskets and Left-to-Right association rules in the second step. The confidence and support measures are computed for every rule.

The Apriori algorithm produces outputs that are simple to interpret and can reveal a variety of unusual trends. However, it could be challenging to comprehend the results due to the large volume of association rules. The algorithm's computational times when searching for huge item sets are a second issue because of its exponential complexity.

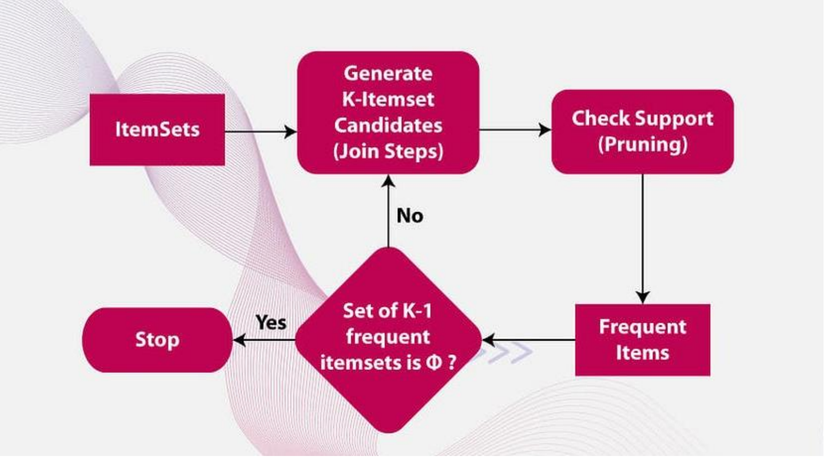

ItemSets
Generate K-Itemset Candidates (Join Steps)
Check Support (Pruning)
No
Frequent Items
Set of K-1 frequent itemsets is Φ ?
Yes
Stop

Figure 5(Apriori algorithm steps in Data Mining)

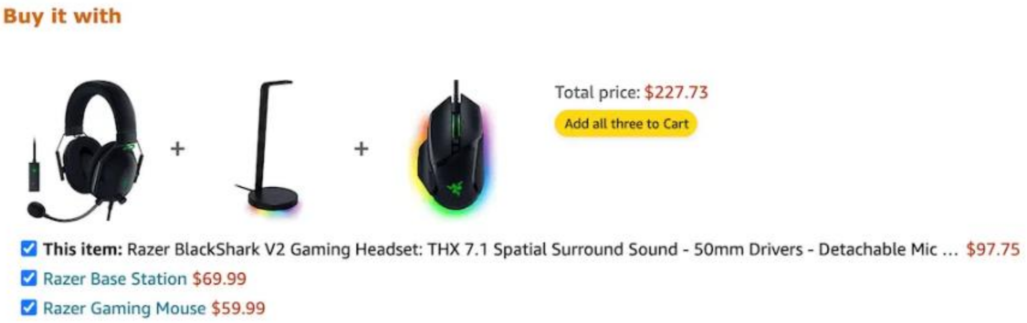
3.1.6. Cross-Selling

Cross-selling is the practice of selling more goods or services to current clients or potential clients. It can strengthen your company interaction with your customers and try to offer a more thorough solution to their concerns.

The basic concept behind cross-selling is that a consumer is only presented with an additional product or service if they have already bought the base product but have not completed the buy transaction. Therefore, as complementary goods are only supplied together with the purchase of the core product, they are not risky to offer. Thus, to optimize profit, financial institutions are prepared to employ this strategy (Baumgarten, Widz, Białokozowicz & Dietl, 2006). Cross-selling makes it possible to boost a customer's revenue and deepen their relationship. Cross-selling is the term for the strategy used to maximize earnings from a single customer.

Using this strategy also lessens the requirement for and associated expenses with client acquisition. Because cross-selling can lead to a good customer relationship and lower the chance of the client leaving in favor of a competitor, financial institutions are more ready to use it (Szczepaniec, 2003). Cross-selling serves the following aims (Paradecki, 2008, p. 343): boosting turnover while fixed costs remain constant; boosting customer loyalty; decreasing client fluctuation; fostering a relationship with the client; making the best use of distribution channels; and lowering risk and uncertainty.

For example, online shops such as Amazon often display related products on the product page you are now viewing. When you are checking out, they might also offer you complementary products. As an illustration, consider this:



Buy it with

Total price: $227.73

☑ **This item:** Razer BlackShark V2 Gaming Headset: THX 7.1 Spatial Surround Sound - 50mm Drivers - Detachable Mic ... $97.75

☑ Razer Base Station $69.99

☑ Razer Gaming Mouse $59.99

Figure 6(Amazon Cross-selling Example)

We utilized Market Basket Analysis techniques to derive key patterns in customer purchase behavior, including

Association Rule Mining: We derived associations based on support and confidence measures using the Apriori algorithm.

Cross-Selling Analysis: We derived frequently co-occurring pairs of items with association rules and confidence measures to enable cross-selling.

Affinity Analysis: In this, we leveraged the calculations and used the Zhang metric for measuring the strength of linkages between items. This marked strong associations of products.

Rule Pruning: Filtering rules by consequent support, leverage, and Zhang's metric to highlight the most significant and actionable insights.

Apriori Algorithm: the algorithm applied as it efficiently manages enormous databases like our millions of records and 27 attributes. Its role was to identify frequent itemsets and generate valuable association rules along with calculating support and confidence metrics for developing effective cross-selling strategies.

3.2. Dynamic Pricing

Dynamic pricing has become an effective strategy for pricing decision optimization across a range of businesses, especially e-commerce. Dynamic pricing algorithms, in contrast to conventional static pricing methods, allow firms to instantly modify their rates in response to many factors such market conditions, competitor pricing, customer demand, and individual client preferences. With this dynamic strategy, businesses may set prices that correspond with customers' willingness to pay and adjust to changes in market dynamics, allowing them to maximize value.

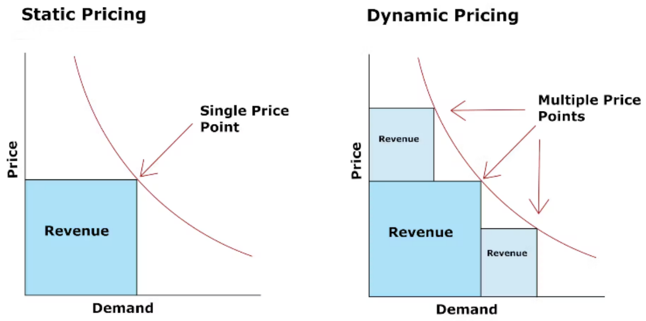


Figure 7(Dynamic vs Static pricing revenue difference)

The prevalence of dynamic pricing is attributed to advancements in technology, particularly the increased availability of customer demand data, sophisticated decision support tools, and the influence of emerging technologies. The increase in demand data availability in the retail sector (Khan, 2023, 2022, 2021, 2021) has given firms the opportunity to more effectively analyze consumer behaviors and make more informed price decisions. Similarly, dynamic pricing tactics have affected lower call rates, increased competition, and better network infrastructure in mobile communication.

The implementation of direct-to-customer business models, improved inventory decisions, and improved production process coordination have all been examples of how dynamic pricing has affected the automotive sector. Dynamic pricing has been successful because of the strengthening of network connections made possible by the widespread use of the internet. This has helped both buyers and sellers by lowering menu prices and offering an integrated customer information database.

When implementing dynamic pricing, it is important to consider factors including customers' willingness to pay different rates, the availability of a segmented market, fair play regulations, and making sure that revenue costs surpass the costs of segmentation and enforcement. This tactic works especially well in businesses with high fixed costs and low variable costs. Implementing short-term cycles like temporary and permanent markdowns, re-pricing depending on rivals' prices, and modifying prices during high and low demand scenarios are some examples of how to use dynamic pricing strategies. Dynamic pricing provides flexibility to a wide range of industries, including E-Commerce, retail, airlines, hotels, electric utilities, mobile communication systems, automobiles, athletic events, car rentals, and insurance.

3.2.1. Types of Dynamic Pricing:

1. **Segmented pricing:** This kind of approach is predicated on offering various prices to various client segments. The categories are split according to factors like geography and demography. Most people believe that segmented pricing is unjust, and as a result, there are several lawsuits and disgruntled customers. As example:
   - Students can often get a cheaper plane ticket because they are less likely to be earning money at the time of purchase.
   - A product may cost more in wealthy regions where consumers may afford to purchase it.
   - If you purchase a product online rather than in-store, it may be less expensive.

- Because consumers who utilize coupons are usually price sensitive, a product with a coupon may be less expensive.

2. **Time-based Pricing:** When pricing for goods and services fluctuate during the day, this approach is employed in such industries. For instance, certain businesses charge more for same-day delivery. Taxi transportation costs are very expensive during the night. For example:
   - Holiday travel is more common; thus, airfares and train tickets are more expensive during this time of year.
   - When schools and universities reopen in September, stationery prices may increase.
   - Older products are typically priced cheaply by retailers and e-commerce businesses to entice people to purchase them before they expire or go out of style.

3. **Peak Pricing:** Market data is needed for peak pricing. It is comparable to time-based data in that companies set prices based on periods of high demand; however, they also use competition data, including availability or inventory. As Example:
   - A company may raise prices if it knows that other companies are running low on inventory.
   - Electricity providers may raise their rates for increasing power consumption during heatwaves.
   - When there are fewer drivers present, Uber often raises the price of rides.

4. **Penetration Pricing:** A company using this pricing strategy is trying to break into a new market. To draw clients, the corporation mostly uses this method by offering lower prices than its rivals. For example, Swiggy gives greater savings when the company expands into a new city. This draws clients to place orders with Swiggy.

5. **Competitive pricing:** The process of pricing business goods and services in line with market and rival prices is known as competitive pricing. According to a 2019 survey, 70% of participants said that the primary reason they chose to shop with a specific online retailer was their reasonable cost. Companies can determine prices:
   - over the pricing of their rivals to suggest that their good or service is superior to theirs.
   - lower than rivals' prices in the expectation that more buyers will buy their goods after comparing costs across several suppliers.
   - at a rival's price to prevent a loss. Companies that set their prices in line with the market typically use a variety of marketing strategies to set their goods apart from those of their rivals.

6. **Bulk pricing:** Offering a cheaper price or a discount to clients who want to purchase a product in large quantities is known as bulk pricing. For example:
   - Retailers can entice customers to purchase multiple items of the same product by offering a buy-2-get-1 bundle.

- Travel agents provide cheaper travel packages that include hotel, car rental, and flights than if you were to purchase each item separately. This approach promotes purchasing a package rather than a single service.

For the data containing the customer orders that would work as an indicator of the demand for a product and their selling prices, we could apply time-based pricing effectively. We could also implement bulk pricing strategies based on the results from the market basket analysis and cross-selling to provide incentives, such as discounts and promotions that might entice customers to buy more products from a package.

3.2.2. Machine Learning:

Machine learning (ML) is a branch of artificial intelligence (AI) that gives computers the capacity to autonomously learn from data and historical events, finding patterns to help them make predictions with little help from humans.

Without explicit programming, computers can function independently due to machine learning techniques. Applications for machine learning are fed new data and can learn, grow, evolve, and adapt on their own.

Large data sets can yield useful information through machine learning, which uses algorithms to find patterns and learn through iterations. ML algorithms do not rely on any fixed equation that may be used as a model; instead, they use computation techniques to learn directly from data.

Throughout the "learning" processes, the performance of machine learning algorithms improves adaptively as more samples become available. Deep learning, for instance, is a branch of machine learning that trains machines to imitate natural human behaviors like learning from examples. Compared to traditional ML algorithms, it offers more efficient parameters.

Machine learning is not a new idea; it dates to the Enigma Machine's employment during World War II. However, it is only recently that complicated mathematical computations can now be applied automatically to an increasing amount and variety of data.

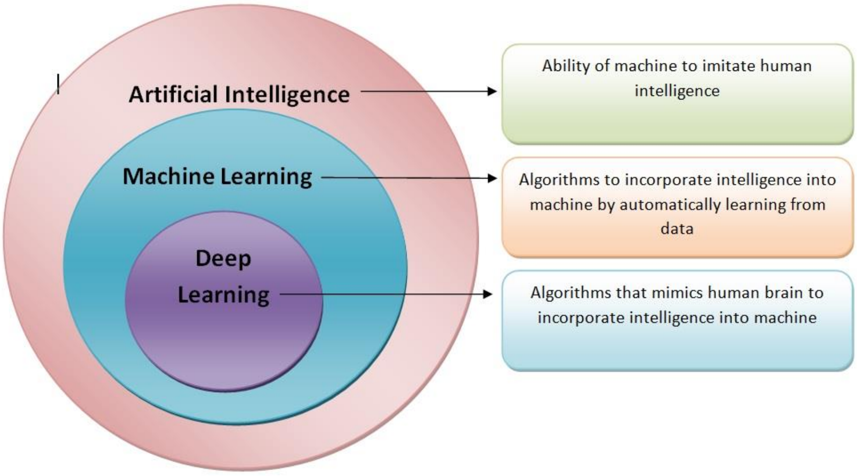

Artificial Intelligence
Machine Learning
Deep Learning
Ability of machine to imitate human intelligence
Algorithms to incorporate intelligence into machine by automatically learning from data
Algorithms that mimics human brain to incorporate intelligence into machine

Figure 8(Artificial Intelligence, Machine learning and Deep Learning)

3.2.2.1. Machine learning life cycle

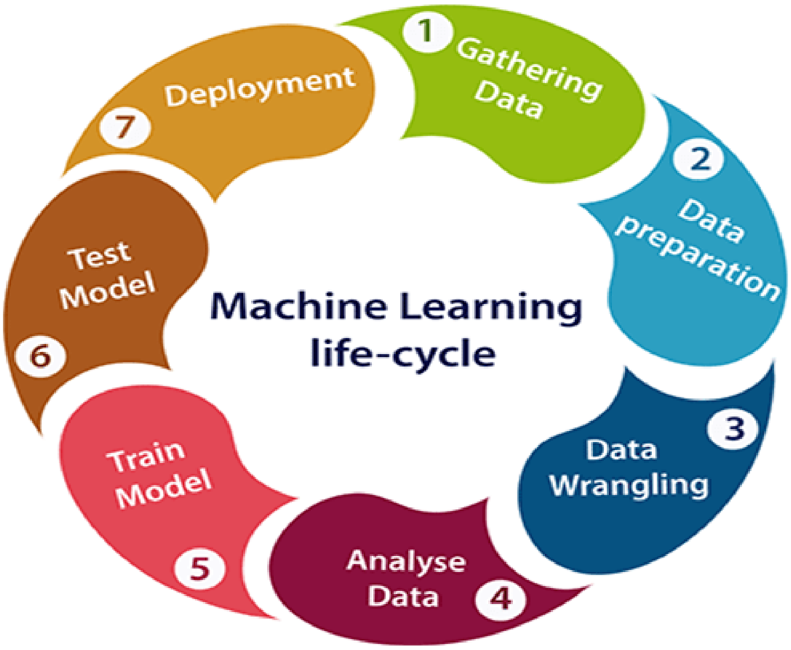


Figure 9(Machine learning life cycle)

Computer systems may now automatically learn without explicit programming due to machine learning. However, how is a machine learning system implemented? Thus, the machine learning life cycle can be used to explain it. A machine learning project can be built efficiently by following a cycle known as the machine learning life cycle. The life cycle's primary goal is to resolve the issue or undertaking.

The seven main steps in the machine learning life cycle are listed below:

- Gathering Data

- Data preparation
- Data Wrangling
- Analyze Data
- Train the model
- Test the model
- Deployment

1. **Gathering Data:**

Data gathering is the initial and crucial step in the machine learning life cycle. It includes:

- Finding the right places to gather data, such as files, databases, the internet, or mobile devices, is known as "identifying data sources."
- Data Collection: Compiling the information from these sources.
- Integrating Data: Creating a single, integrated dataset by combining data from multiple sources.

For machine learning predictions to be precise and effective, both the quantity and quality of the data acquired are critical.

2. **Data preparation:**

We must prepare the data for additional actions after we have collected it. The process of organizing our data and getting it ready for use in machine learning training is known as data preparation.

In this stage, we first aggregate all the data, and then we randomly sort the data. Two procedures can be further separated out of this step:

- Data investigation: It is employed to comprehend the type of facts that we must deal with. We must comprehend the properties, structure, and caliber of data. An efficient result is the result of a deeper comprehension of the facts. This contains outliers, general trends, and correlations.
- Pre-processing of data: Preprocessing the data in preparation for analysis is the next stage.

3. **Data Wrangling:**

Is the process of preparing raw data for analysis by cleaning and formatting it into a format that may be used. This action consists of:

- Cleaning data involves identifying and eliminating issues including noise, duplicate data, incorrect data, and missing values in order to address quality issues.
- Choosing the appropriate variables for analysis is known as variable selection.
- Data transformation is the process of putting data into the right format.

Efficient data wrangling is essential because it guarantees the accuracy and consistency of analytic results.

4. **Analyze Data:**

The data has now been cleaned and prepped and is ready for analysis. This action includes:

- choice of techniques for analysis
- building models
- Study the outcome.

This step's objective is to create a machine learning model that will study the data using a variety of analytical methods and evaluate the results. First, the type of problem must be determined. Next, machine learning techniques such as classification, regression, cluster analysis, association, etc., must be chosen. Finally, the model must be built using prepared data and evaluated.

5. **Train the model:**
   The model must now be trained to increase its performance and provide a better solution to the problem.
   We train the model with a variety of machine learning algorithms using datasets. A model must be trained for it to understand the different features, rules, and patterns.

6. **Test the model:**
   We test our machine learning model once it has been trained on a particular dataset. Using a test dataset, we assess the accuracy of our model in this stage. By testing the model, we can determine its accuracy % in relation to the project's or problem's requirements.

7. **Deployment:**
   Deployment, the final stage of the machine learning life cycle, involves implementing the model in an actual system.
   We implement the above-prepared model in the actual system if it is delivering the required correct results at a reasonable pace. However, we will first determine whether the project is utilizing the available data to improve performance before releasing it. The deployment step is comparable to creating a project's final report.

3.2.2.2. Supervised Machine Learning:

In supervised learning, machines are trained with properly "labelled" training data and then predict the output based on that data. Given that the data is marked, some input data has already been assigned the appropriate output.
The supervisor that trains the computers to accurately print the output in supervised learning is the training data that is given to them. It uses the same idea that a student would learn under a teacher's guidance.
Giving the machine learning model accurate input and output data is known as supervised learning. Finding a mapping function to connect the input variable (x) and the output variable (y) is the goal of a supervised learning algorithm.

How Supervised Learning Works?

Supervised learning trains models using labeled data, where each data point has a category or value assigned (the label).
The model learns the relationship between the input data and the labels. After training, the model can then predict the labels for new, unseen data.
Think of it like training a friend to identify shapes. You show them examples of squares, triangles, etc., labeled accordingly. Once trained, they can identify new shapes based on what they learned.

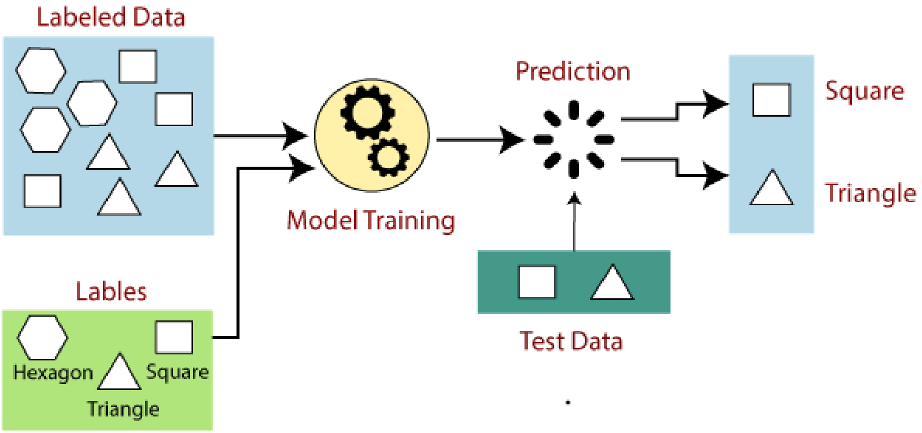


Figure 10(Supervised Machine Learning flow)

Supervised learning classified into two categories of algorithms.

1. Classification: When the output variable is a category, such "Red," "Blue," "Disease," or "No disease," then there is a classification difficulty.
2. Regression: Regression problems arise when an output variable, such "Weight" or "Dollars," has a real value.

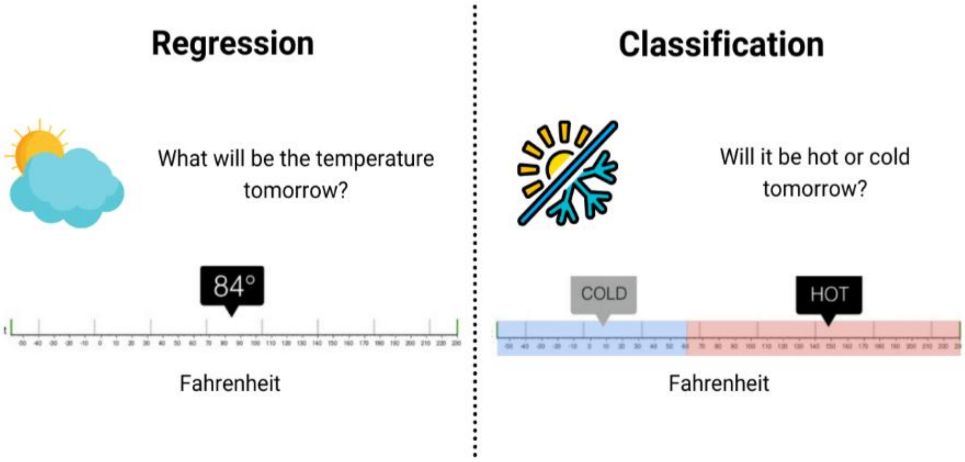

Regression
What will be the temperature tomorrow?
84°
Fahrenheit
Classification
Will it be hot or cold tomorrow?
COLD
HOT
Fahrenheit

Figure 11(Regression vs classification)

3.2.2.3. Unsupervised Machine learning:

Unsupervised learning, as the name implies, is a machine learning technique in which training datasets are not used to supervise models. Rather, the models themselves extract the insights and latent patterns from the provided data. It is comparable to the process of learning that occurs in the human brain when something is learned. Unsupervised learning is a subset of machine learning in which models are trained on unlabeled datasets and then permitted to operate unsupervised on the data.

Unlike supervised learning, where we know the input data but no matching output data, unsupervised learning cannot be immediately applied to a regression or classification task.

Unsupervised learning aims to extract the dataset's underlying structure, classify the data based on similarities, and encode the dataset in a compressed manner, Here is an Example:

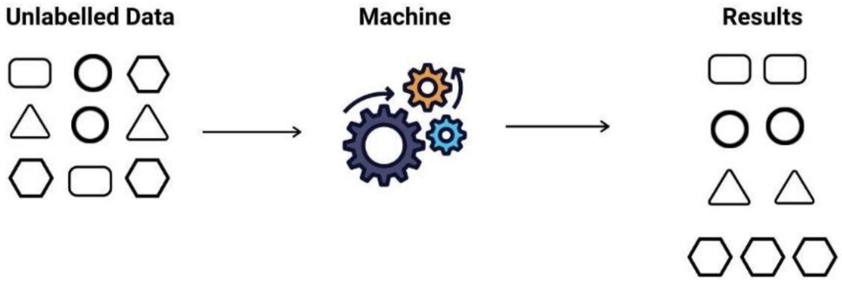


Figure 12(Unsupervised Machine learning flow)

The unsupervised learning algorithm can be further categorized into two types of problems:

1. Clustering: The process of clustering things into groups so that those with the greatest similarities stay in that group and have little to no similarities with other objects in the group is known as clustering. By identifying patterns among the data objects, cluster analysis groups them into groups based on if such patterns exist.
2. Association: An unsupervised learning technique called an association rule is used to determine the connections between variables in a big database. It finds the collection of elements that exist in the dataset. An association rule improves the efficacy of marketing strategy. For example, consumers who purchase X (bread, for example) also frequently buy Y (butter/jam). Market Basket Analysis is an example of an association rule in action.

3.2.2.4. Ensemble learning:

A general meta-approach to machine learning called ensemble learning combines the predictions of several models to improve predictive performance.

While the number of ensembles you can create for your predictive modeling challenge seems limitless, the discipline of ensemble learning is dominated by three techniques. To the extent that each is an area of study that has given rise to numerous more specialized techniques, rather than algorithms in and of themselves.

The three primary categories of group learning techniques are boosting, stacking, and bagging.

- In bagging, many decision trees are fitted to various samples of the same dataset, and the results are averaged.
- In stacking, a variety of model types are fitted to the same set of data, and a different model is used to determine the optimal way to combine the predictions.
- Boosting is the process of gradually adding ensemble members to improve the predictions from earlier models and produce a weighted average of the outputs.

3.2.3. Machine Learning Models:

3.2.3.1. Decision Tree Model:

Decision tree is a non-parametric supervised learning technique for classification and regression tasks. A hierarchical tree structure with a root, branches, internal nodes, and leaf nodes represents it. This model describes a decision with all possible consequences in a conditional control statement; consequently, it is easy to understand and interpret. The decision trees are very flexible, and their multiple applications could be shown through the many feature-based splits from the root node to its leaf nodes.

Decision Tree Terminologies:

- Root Node: The initial node that consists of the dataset, from which other nodes stem based on the characteristics or conditions.
- Decision Nodes: They are the intermediate nodes that occur because of a split in the root node, and they represent other decisions or conditions.
- Terminal Nodes (Leaf Nodes): These are nodes without further splitting and represent the final classification or outcome.
- Sub-Tree: A part of the decision tree where only a few of the decisions or outcomes are shown.
- Pruning: Removal of selected nodes to avoid overfitting and to simplify the model.
- Branch / Sub-Tree: A path of specific decisions and their outcomes within the tree. Parent
- Node and Child Nodes: a parent node is split into sub-nodes (or child nodes), where the parent can be a decision or a condition, while child nodes are further decisions or outcomes.

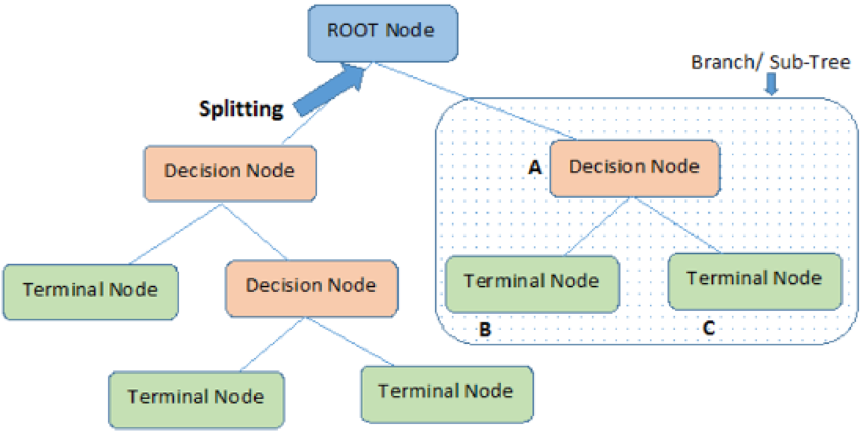


Figure 13(Decision Tree algorithm)

3.2.3.2. Linear Regression:

One of the simplest and most widely used machine learning methods is linear regression. It's a statistical technique for forecasting analysis. For continuous/real or numerical variables like sales, salary, age, product price, etc., linear regression generates predictions.

The term "linear regression" refers to a procedure that displays a linear relationship between one or more independent (y) variables and a dependent (y) variable. Given that it displays a linear connection, linear regression determines how the value of the dependent variable varies in response to the value of the independent variable.

The link between the variables is represented by a skewed straight line according to the linear regression model. Look at the picture below:

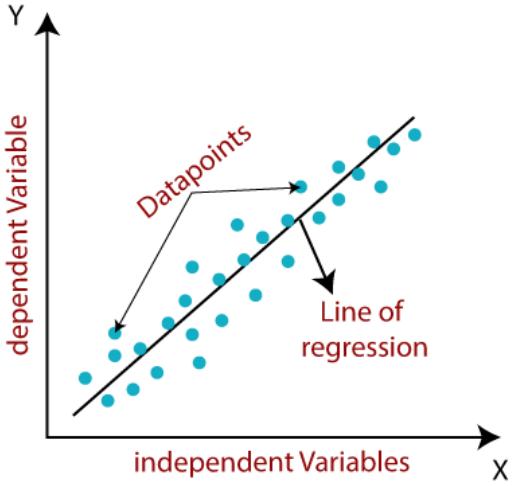


Figure 14(Linear Regression Model)

3.2.3.3. Bayesian Ridge Model:

The Bayesian Ridge model is a linear regression technique that incorporates ridge regression in the context of Bayesian inference. It further regularizes to guard against overfitting; it treats model parameters as random variables governed by probabilistic distributions to allow quantification of the uncertainty in predictions. Bayesian ridge optimization is then a procedure that integrates prior knowledge with updated information and, hence, results in a very robust and flexible regression model, which inherently deals with overfitting and offers a probabilistic interpretation of results.

3.2.3.4. Random Forest Model:

By using both feature randomness and bagging to produce an uncorrelated forest of decision trees, the random forest algorithm is an extension of the bagging technique. Feature randomness, sometimes referred to as "the random subspace method" or "feature bagging" (link is external to IBM.com), produces a random selection of features that guarantee low correlation between decision trees. There is a significant distinction between random forests and decision trees. Random forests merely choose a portion of those features, whereas decision trees consider all potential feature splits.

How it works?

The three primary hyperparameters of random forest algorithms must be set prior to training. These consist of the size of the nodes, the count of trees, and the quantity of characteristics sampled. Regression and classification issues can then be resolved using the random forest classifier.
Each decision tree in the ensemble of decision trees used in the random forest technique is made up of a bootstrap sample, which is a sample of data taken from a training set with replacement. One-third of the training sample is designated as test data; this is referred to as the out-of-bag (oob) sample

Feature bagging is then used to introduce a further randomization, increasing dataset variety and decreasing decision tree correlation. The prediction's determination will change depending on the kind of problem. The individual decision trees in a regression job will be averaged, and in a classification work, the predicted class will be determined by a majority vote, or the most common categorical variable. Lastly, that prediction is confirmed using cross-validation using the oob sample.

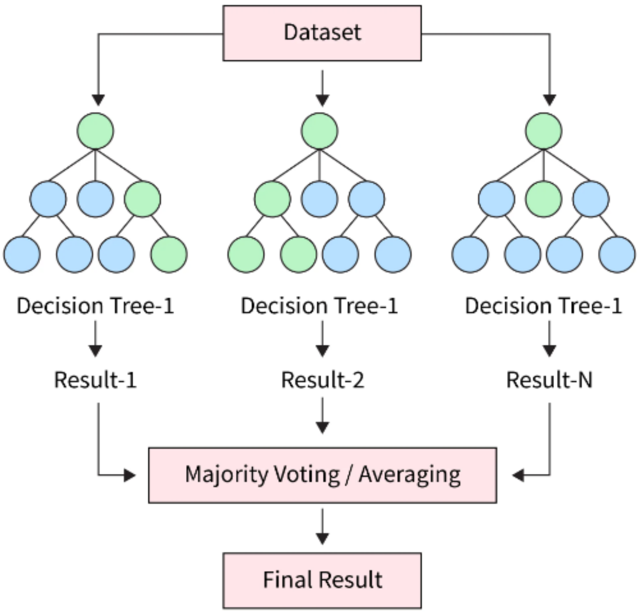


Figure 15(Random Forest Model)

3.2.3.5. XGB Model:

Extreme Gradient Boosting, or XGBoost, is an ensemble learning machine learning technique. For supervised learning tasks like regression and classification, it's in. By iteratively merging the predictions of several different models, frequently decision trees, XGBoost creates a predictive model.

For the algorithm to function, weak learners are gradually added to the ensemble, with each new learner concentrating on fixing the mistakes caused by the previous ones. Throughout training, it minimizes a predetermined loss function by using an optimization approach called gradient descent.

The XGBoost Algorithm's key characteristics are its capacity to manage complex relationships in data, regularization strategies to avoid overfitting, and the integration of parallel processing for effective computation.

3.2.3.6. CatBoost Model:

CatBoost is a supervised machine learning method that is used by the Train Using AutoML tool and uses decision trees for classification and regression. CatBoost uses gradient boosting (the Boost) and operates with categorical data (the Cat), as its name implies. Several decision trees are built iteratively during the gradient boosting procedure. Better outcomes are produced by each succeeding tree, which enhances the output of the preceding tree. For a quicker implementation, CatBoost enhances the original gradient boost technique.

A disadvantage of other decision tree-based techniques is that, usually, they require pre-processing of the data to convert categorical string variables to numerical values, one-hot encodings, and other formats. This is not the case with CatBoost. Without requiring any preprocessing, this technique can immediately ingest a mix of categorical and non-categorical explanatory variables. It performs preprocessing as an algorithmic step. CatBoost encodes categorical characteristics using an approach known as ordered encoding. In order to determine a value to replace the categorical feature, ordered encoding takes into account the target statistics from every row before a data point.

CatBoost's usage of symmetric trees is another unique characteristic. This indicates that all the decision nodes utilize a similar split condition at every depth level.

CatBoost has the potential to be faster than other techniques like XGBoost. It keeps several of the previous algorithms' features, including regularization, cross-validation, and support for missing values. Both tiny and huge amounts of data work well with this approach.

Based on the nature of the data and the nature of the work, the CatBoost Model was chosen, which in turn is. It efficiently prevents overfitting through built-in regularization techniques and offers competitive performance with minimum tuning of parameters. CatBoost is scalable, supports feature importance analysis, and integrates well with Python and other data science tools, making it a powerful choice for dynamic pricing and various machine learning applications.

4. Chapter4: Implementation

4.1. Data Overview

The dataset under consideration, sourced from Kaggle, pertains to an e-commerce platform based in Turkey. This dataset is a comprehensive compilation that captures the diversity and breadth of the products offered on the site. It includes a multitude of products spanning various categories, providing a rich and detailed view of the e-commerce landscape. Each entry in the dataset corresponds to a specific product and encompasses a wide array of attributes and features that describe the product in detail.

Our dataset is quite extensive, comprising a total of 1,000,000 entries. This large volume of data ensures a robust and representative sample of the e-commerce site's inventory. Each entry is detailed across 27 columns, with each column representing a different aspect or characteristic of the product. The dataset's comprehensive nature makes it an invaluable resource for a variety of analytical purposes, including market analysis, trend identification, and sales predicting.

Specifically, the dataset includes information on 27,000 Turkish supermarket items available across 81 stores, with each city in Turkey having a store. It also contains data on 20,000 real Turkish customer names and addresses, along with a count of 369.4 different brands.

The breadth and depth of this dataset provide a fertile ground for data exploration and insight generation, making it a powerful tool for anyone looking to understand the dynamics of the e-commerce sector in Turkey. Whether you are a data scientist, a business analyst, or a market researcher, this dataset offers a wealth of information that can be leveraged to gain a deeper understanding of product performance, customer preferences, and market trends.

- **Description of the data:**

ID	A unique identifier for each record in the dataset. It is commonly used as a primary key.
ORDERID	Identifier for a specific order. It associates order details with a particular transaction.
ORDERDETAILID	Unique identifier for each order detail. It may be used to link specific details (items, quantities, prices) to an order.
DATE	The date when the order was made, or the transaction occurred.
USERID	Identifier for the user who placed the order. It may be a foreign key linking to a user table.
USERNAME	The username of the user who placed the order.
NAMESURNAME	The full name of the user who placed the order.
STATUS_	The status of the order, indicating whether it is processed, shipped, delivered, etc.
ITEMID	Identifier for a specific item in the order.

ITEMCODE	A code associated with a specific item, often used for inventory, or tracking purposes.
ITEMNAME	The name or description of the item.
AMOUNT	The quantity or number of items ordered.
UNITPRICE	The price of a single unit of the item.
PRICE	The price of the item multiplied by the quantity (AMOUNT). It represents the subtotal for that item.
TOTALPRICE	The total price of the order, including all items and any additional costs.
CATEGORY1, CATEGORY2, CATEGORY3, CATEGORY4	Different levels or types of categories to classify the item. It is a way of organizing and grouping related items.
BRAND	The brand of the item.
USERGENDER	The gender of the user who placed the order.
USERBIRTHDATE	The birthdate of the user.
REGION, CITY, TOWN, DISTRICT	Geographic information indicating where the user is located.
ADDRESSTEXT	The detailed text describing the user's address.

4.2. Data Preprocessing

Before starting the implementation, it was crucial to ensure that our data was ready for use through various data analysis processes. This preparatory phase was essential for understanding the data and the relationships within it, setting the stage for meaningful analysis and accurate insights. The data preparation process involved several key steps: collecting, cleaning, exploring, and modeling the data.

First, we began with data collection, gathering all relevant data from various sources. This initial step ensured that we had a comprehensive dataset to work with. However, raw data often contains numerous issues that can compromise analysis, necessitating thorough cleaning.

Data cleaning was the next critical step. This process involved identifying and rectifying several common problems. Duplicate data was removed to prevent redundant information from skewing results. Outdated data, which could lead to inaccuracies, was also eliminated. We addressed incomplete data by either filling in missing values using appropriate techniques or discarding records that lacked critical information. Incorrect data, which could distort analysis, was corrected based on reliable sources. Inconsistent data formats were standardized to ensure uniformity across the dataset.

Following cleaning, we moved on to data exploration. This stage was pivotal in gaining a deep understanding of the dataset. We performed exploratory data analysis (EDA) to uncover patterns, trends, and relationships within the data. Visualization techniques and summary statistics were employed to provide a clear picture of the data's characteristics and distribution.

Finally, with a clean and well-understood dataset, we proceeded to the modeling phase. This involved selecting appropriate models and algorithms to analyze the data and derive actionable insights. The thorough preparation ensured that our models were built on solid, reliable data, enhancing their accuracy and robustness.

4.2.1. Data cleaning

In the process of preparing our dataset for analysis, we undertook several essential data cleaning tasks to ensure the accuracy and reliability of our data.

First, we addressed the issue of **missing values**, which can compromise the integrity of the analysis. We applied a combination of techniques, such as imputing missing values with the mean or median of the respective columns, and in cases where data was critically lacking, we opted to remove those records entirely.

Next, we **handled duplicate** data by identifying and removing redundant entries. This step ensured that each data point was unique and no repeated information would skew our results.

To further streamline our dataset, we carefully examined each column and dropped those that were deemed irrelevant or **useless** for our analysis. By reducing clutter, we focused on the features that truly mattered, making our dataset more manageable and relevant.

Additionally, we paid close attention to **outliers**, recognizing that extreme values could distort our findings. We employed strategies to identify these outliers and decided on appropriate actions, such as removing them or transforming the data to mitigate their impact.

4.2.2. Data Transformation

- **Corrected Calculation in TOTALPRICE Column:** One of our primary tasks during data preparation was to meticulously review and correct the calculations in the TOTALPRICE column. This involved ensuring that all values accurately reflected the intended financial calculations, addressing any discrepancies or errors that could potentially affect the integrity of our financial analysis. By systematically verifying and rectifying these calculations, we aimed to enhance the reliability and precision of our dataset, laying a solid foundation for accurate financial insights and decision-making.

- **Transformed USERGENDER to Binary Numerical Format:** Another critical aspect of our data preparation involved transforming the USERGENDER column from categorical to numerical format using binary encoding. This transformation was essential for facilitating more straightforward statistical analysis and machine learning processes. By assigning binary values (0 and 1) to different gender categories, we standardized the representation of gender data across our dataset. This standardization not only streamlined computational processes but also enabled more efficient model training and analysis, ensuring that gender could be effectively integrated into predictive models and statistical analyses without ambiguity or inconsistency.

4.3. Data Exploration

In the Data Exploration phase of our analysis, we engaged in a thorough examination of our dataset to uncover meaningful insights and patterns. Our approach began with calculating descriptive statistics such as mean, median, and standard deviation to understand the central tendencies and variability of our variables. We then employed various visualization techniques, including histograms, box plots, and scatter plots, to visualize the distributions and relationships within the data. These visualizations provided a clear overview of our data's characteristics and highlighted potential outliers or anomalies. Additionally, we explored correlations between different variables using correlation matrices and scatter plots, identifying significant relationships and potential dependencies. This phase allowed us to identify data trends, understand feature distributions, and gain a deeper understanding of our dataset's structure and nuances. The insights gained from this exploration phase informed subsequent steps in our analysis, guiding the selection of appropriate models and strategies to derive actionable insights and make informed decisions.

4.4. Data Visualization:

Data visualization plays a crucial role in our analysis by transforming raw data into visually accessible insights that are intuitive to interpret and analyze. By employing various graphical representations such as histograms, box plots, scatter plots, and heatmaps, we effectively communicate complex datasets in a way that is both meaningful and actionable. These visualizations not only enhance our understanding of data distributions, trends, and relationships but also highlight outliers or patterns that may not be immediately apparent from numerical summaries alone. By presenting data visually, we can easily compare values across different variables or categories, enabling us to identify correlations, anomalies, and trends more effectively. This visual approach not only simplifies complex data but also supports more informed decision-making processes by providing stakeholders with clear, concise representations of key insights and findings. Ultimately, data visualization serves as a powerful tool in our analytical toolkit, empowering us to derive deeper insights and make strategic decisions based on a comprehensive understanding of our dataset.

4.4.1. Pie Chart for Total Sales by Category:

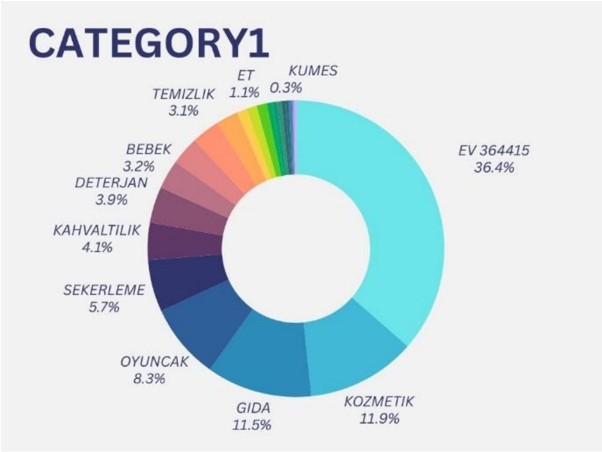


Figure 16(Products Category Distribution)

To understand revenue generation across different product categories (CATEGORY1), we use a Pie Chart for Total Sales by Category. This chart visually represents the distribution of sales among categories, highlighting which categories generate the most revenue. It helps businesses allocate resources effectively and focus marketing efforts where they are most likely to succeed. By identifying top-performing categories, businesses can align their strategies with market demand and pinpoint areas for growth. This visual tool is essential for tracking sales performance and guiding strategic business decisions.

4.4.2. Scatter plot for Price vs. Unit Price

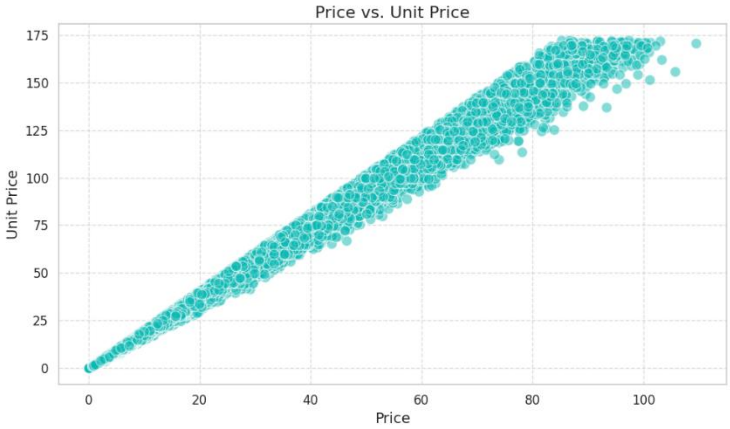


Figure 17(Price vs. Unit Price)

We created an enhanced scatter plot to visualize the relationship between Price and Unit Price in our dataset. This plot utilizes dots to represent individual data points, where each dot's position on the graph corresponds to its respective Price and Unit Price values. This visualization technique helps us identify any patterns or trends between Price and Unit Price, supporting deeper insights into pricing strategies or cost structures within our dataset. By examining this plot, we can better understand how changes in Price relate to changes in Unit Price, aiding in strategic decision-making and potential optimizations in pricing strategies.

4.4.3. Histogram of the distribution of Unit Price within the dataset.

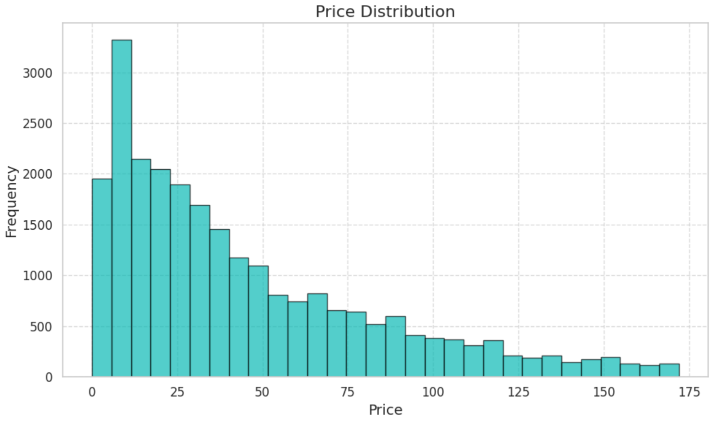


Figure 18(Price Distribution)

We generated an enhanced histogram to visualize the distribution of Unit Price within our dataset. This visualization technique allows us to quickly grasp the distribution pattern of Unit Price within our dataset, highlighting common price ranges and potential outliers. By examining this histogram, we gain insights into the variability and concentration of Unit Price values, which can inform pricing strategies or product segmentation decisions. This graphical representation is instrumental in understanding the distributional characteristics of Unit Price and supports data-driven decision-making processes.

4.4.4. **Line Chart for Sales Over Time**

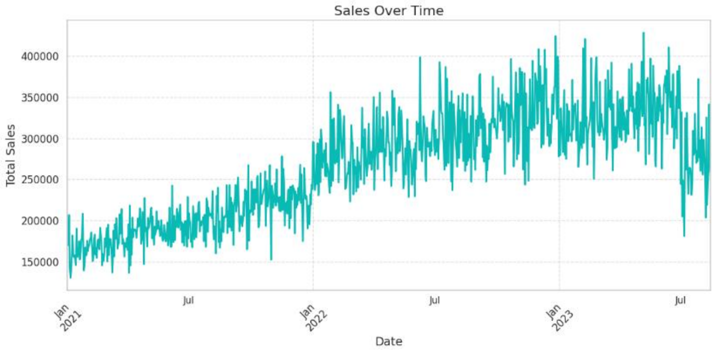


Figure 19(Sales Over Time)

The **Line Chart for Sales Over Time** offers a clear visualization of how sales figures fluctuate over a specified period. By displaying trends and patterns in sales data, it helps businesses identify seasonal variations, peak periods, and potential downturns. This chart is crucial for making informed decisions on marketing strategies, and sales predicting, ultimately aiding in optimizing overall business performance.

4.4.5. Stacked Bar Chart for Category Sales by Region

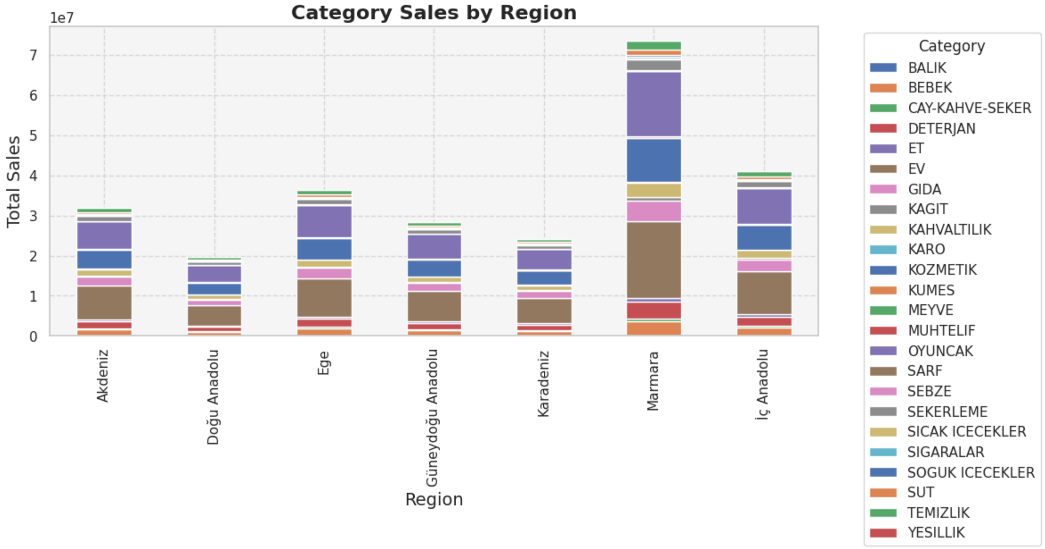


Figure 20(Category Sales by Region)

The **Stacked Bar Chart for Category Sales by Region** visually represents the total sales of different product categories across various regions. This chart helps businesses understand regional preferences and the distribution of category sales. By comparing sales performance across regions, companies can tailor their marketing and inventory strategies to meet local demands more effectively. It also aids in identifying strong and weak markets, facilitating targeted growth efforts and resource allocation.

4.4.6. Scatter Plot for Price vs. Amount

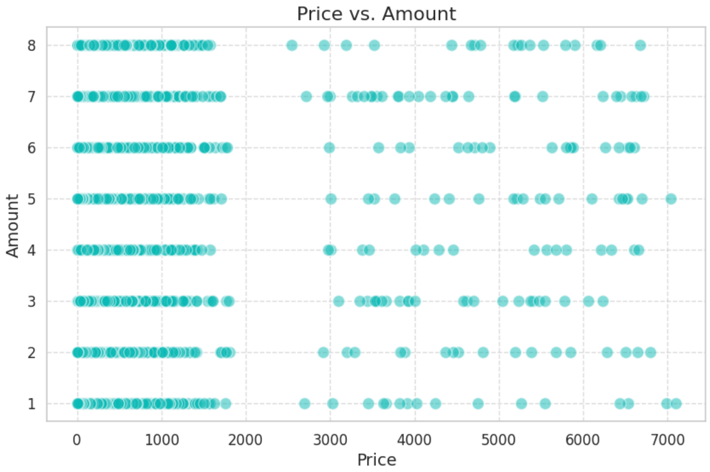


Figure 21(Price VS Amount Scatter plot)

The **Scatter Plot for Price vs. Amount** provides a visual representation of the relationship between product price and quantity sold. By plotting individual data points, it helps identify patterns, trends, and potential outliers in sales behavior. This chart is essential for understanding how price impacts demand, enabling businesses to optimize pricing strategies. Additionally, it can highlight any anomalies that may require further investigation, ensuring a comprehensive understanding of sales dynamics.

4.4.7. Correlation Heatmap Between Numerical Columns.

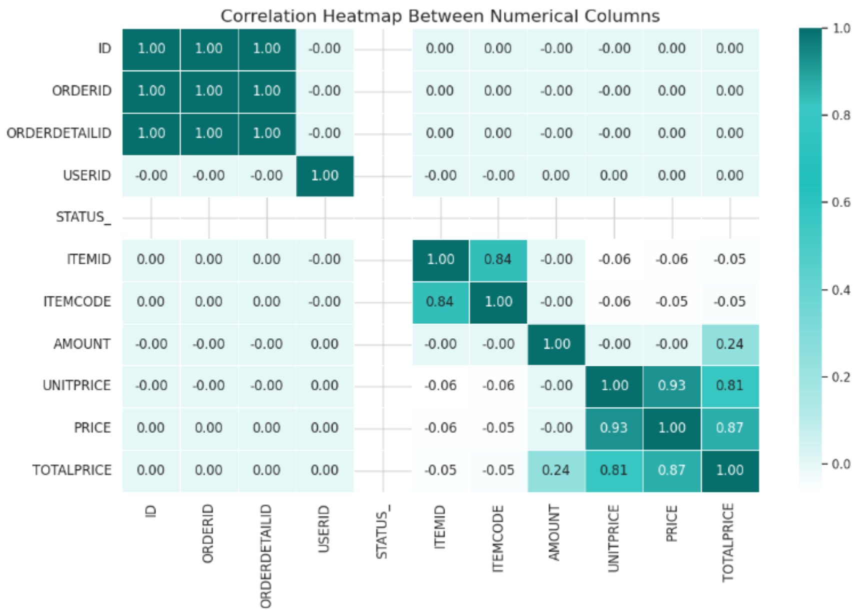


Figure 22(Heatmap Between Numerical Columns)

This correlation heatmap helps us understand the relationships between different numerical variables in our dataset. For instance, if PRICE and TOTALPRICE show a high correlation, it indicates that as the price of items increases, the total price does too. This insight can guide our data analysis and modeling efforts, ensuring we consider the most influential factors in our business decisions.

4.4.8. Line plot For Monthly Sales Trend

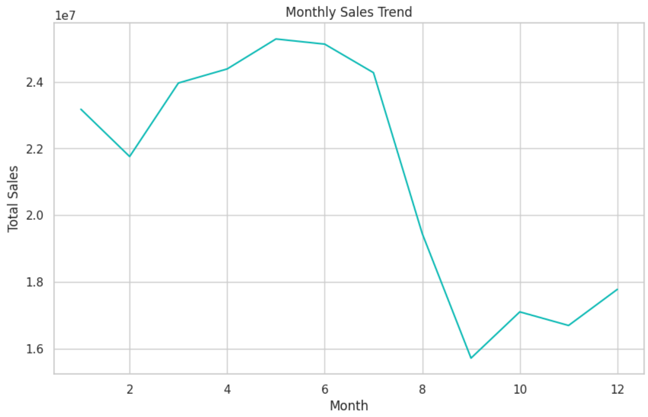


Figure 23(Monthly Sales Trend)

The "Monthly Sales Trend" graph illustrates the variation in total sales across different months. We can pinpoint peak sales periods and months where sales may decline, enabling more strategic planning. Understanding these trends helps optimize marketing efforts, predict future sales, and improve overall business performance.

4.4.9. Area Chart for Cumulative Sales Over Time

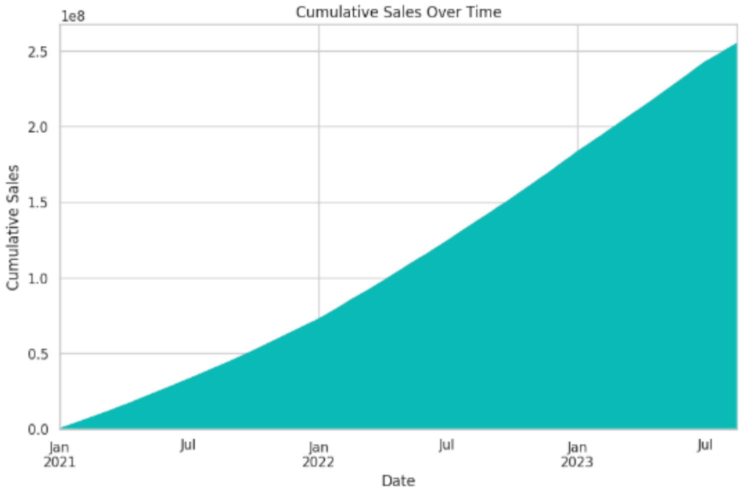


Figure 24(Cumulative Sales Over Time)

We utilized an area chart to visualize the cumulative sales progression over time based on our dataset. The chart illustrates the accumulation of total sales (TOTALPRICE) across various dates, providing a clear representation of sales trends and growth patterns. The cumulative sales are computed by aggregating the total sales values over each date and then cumulatively summing them up. This visual representation is instrumental in identifying sales trends, seasonal fluctuations, and overall growth trajectories, thereby supporting strategic decision-making and performance analysis based on historical sales data.

4.5. Data Manipulation

After cleaning and exploring, it is time to manipulate the data to obtain the most value out of it. This is done by combining the data and including all the different sources to narrow the range of private data. We all have the main data inputs which are product categories. We will summarize the entire data based on the weekly or monthly data and calculate the average sales profit for each category in each week or each month. Where the difference between each period is almost constant, replacing it with fixed numbers starting from zero and calculating the correlation between all the inputs to take it as a basis in the model stage. The data is divided for implementation by taking part of it for training and testing to reach a more accurate prediction by calculating the error.

4.6. Market Basket Analysis Implementation

4.6.1. Introduction

In this chapter, we will perform market basket analysis to uncover patterns and associations within a dataset of transactions. Market basket analysis helps identify which items are frequently purchased together, providing valuable insights for businesses to optimize product placement, marketing strategies, and inventory management. The steps involved in this analysis include detecting the least frequent items, recovering transaction item sets, encoding these transactions, calculating support metrics, analyzing the distribution of item counts per transaction, adding and merging categories to explore combined effects, and finally, computing confidence metrics to understand the strength of associations between items. Each step will be explained in detail along with the corresponding results to illustrate the process and findings.

4.6.2. Applied steps of Market Basket Analysis (MBA)

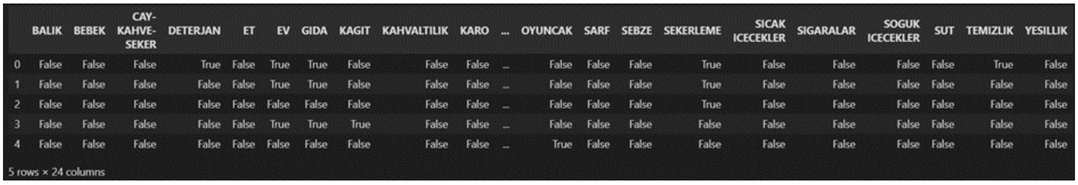
- **Detecting the Least Frequent Items:**
  This step counts how often each "itemid" appears in the dataset and displays the ten least frequent items. The value_counts() method sorts the counts in descending order, and tail (10) retrieves the bottom ten items. The result is a series that shows item IDs with their respective counts, ranging from 5 to 10.

  Each line shows an itemid and the number of times it appears in the Data Frame column. For example:
  - **itemid** 8762 appears 10 times.
  - **itemid** 2449 appears 10 times.
  - **itemid** 1354 appears 5 times, and so on.

- **Recovering Transaction Itemsets:**
  Transactions are grouped by "orderid" to collect unique items in category1 for each order. This creates a list of items purchased in each transaction.
  The output shows the first 5 transactions (orders) and the unique categories of items within each order, which can be useful for market basket analysis, association rule mining, or understanding customer purchasing patterns.
  Each line corresponds to a unique orderid and lists the unique categories found in that order. For example:

- **Order 1** contains items from the categories: EV, TEMIZLIK, SEKERLEME, DETERJAN, GIDA.
- **Order 2** contains items from the categories: EV, SEKERLEME, KOZMETIK, GIDA.
- **Order 3** contains items only from the category: SEKERLEME.
- **Order 4** contains items from the categories: EV, KAGIT, GIDA.
- **Order 5** contains items only from the category: OYUNCAK.

- **Encoding Transactions:**
  An encoder converts the transactions into a one-hot encoded format, where each category becomes a column. Rows represent transactions, and cells indicate the presence (True) or absence (False) of a category in the transaction.
  A value of 'True' indicates that the category is present in the transaction, while 'False' indicates its absence.
  For example, the first row indicates that the first transaction includes items from the categories DETERJAN, EV, GIDA, and SEKERLEME.



	BALIK	BEBEK	CAY-KAHVE-SEKER	DETERJAN	ET	EV	GIDA	KAGIT	KAHVALTILIK	KARO	...	OYUNCAK	SARF	SEBZE	SEKERLEME	SICAK ICECEKLER	SIGARALAR	SOGUK ICECEKLER	SUT	TEMIZLIK	YESILLIK
0	False	False	False	True	False	True	True	False	False	False	...	False	False	False	True	False	False	False	False	True	False
1	False	False	False	False	False	True	True	False	False	False	...	False	False	False	True	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	False	False	False	False	False	False
3	False	False	False	False	False	True	True	True	False	False	...	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	True	False	False	False	False	False	False	False	False	False

5 rows × 24 columns

Figure 25(First Transaction)

  This format is useful for machine learning algorithms that require numerical input, and it is commonly used in market basket analysis to find associations between items.

- **Calculating Support Metrics:**
  The mean of each column (category) is calculated to determine the support metric, which is the proportion of transactions that include the category. Each value in the output tells us how prevalent a category is across all transactions. Higher values indicate more commonly purchased categories, while lower values indicate rarer categories. For example, the category EV is the most common, appearing in 81.01% of transactions, followed by GIDA at 42.22% and KOZMETIK at 41.08%, whereas categories like KARO and MUHTELIF are among the least common, each appearing in less than 0.4% of transactions.

- **Distribution of Item Counts per Transaction:**
  This step counts the number of items in each transaction by summing the Boolean values across columns for each row. The distribution shows most transactions contain between 1 and 5 categories, with the highest number (43,705) having exactly 3 categories, while very few transactions contain 7 or more categories.

- **Adding and Merging Categories:**
  A new column Ev-GIDA is created first as the intersection (&) of EV and GIDA, then as the union (|). The support values for these new columns are calculated. The

intersection shows that 35.01% of transactions contain both EV and GIDA, while the union shows that Ev-GIDA appears in 88.22% of transactions.

- **Computing Confidence Metrics:**
  This step computes confidence metrics for association. Confidence indicates the likelihood of GIDA given EV and vice versa. The confidence of EV -> GIDA is about 43.22%, meaning that 43.22% of transactions containing EV also contain GIDA. Conversely, the confidence of GIDA -> EV is about 82.93%, meaning that 82.93% of transactions containing GIDA also contain EV.

4.6.3. Application of the Apriori Algorithm

We will now expand on this basic analysis. In market basket analysis, the Apriori algorithm is a widely used technique for mining frequently occurring itemsets and producing association rules. We can further filter and select the most important itemsets and rules by establishing minimum support and confidence levels. This allows us to gain deeper insights into client purchasing behavior and facilitates more efficient decision-making.

- **Applying the Apriori Algorithm with a Minimum Support Threshold of 0.01:**
  In this step, we apply the Apriori algorithm to the one-hot encoded transaction data with a minimum support threshold of 0.01. This threshold means that only itemsets appearing in at least 1% of the transactions will be considered frequent. The Apriori algorithm identifies these frequent itemsets and their corresponding support values. We then print the first 100 frequent itemsets along with their support values. For instance, the first few rows show itemsets with support values such as 0.116707 and 0.051146, corresponding to individual items or combinations of items that meet the minimum support requirement. This initial application of the Apriori algorithm helps us understand which items are commonly purchased together at a higher threshold.

- **Applying the Apriori Algorithm with a Minimum Support Threshold of 0.001:**
  Next, we lower the minimum support threshold to 0.001 to identify more frequent itemsets, including those that appear in at least 0.1% of transactions. Additionally, we use the `use_colnames=True` parameter to ensure that the itemsets are labeled with their category names instead of numerical indices. This more granular analysis reveals a broader range of frequent itemsets. For example, itemsets such as `(BALIK)` with a support of 0.005410 and `(EV, GIDA)` with a support of 0.350101 are identified. This lower threshold allows us to capture less frequent, yet potentially significant, itemsets that could provide deeper insights into purchasing patterns.

- **Applying the Apriori Algorithm with a Two-Item Limit:**
  Finally, we apply the Apriori algorithm with the same minimum support threshold of 0.001 but impose a limit of two items per itemset using the `max_len=2` parameter. This constraint focuses the analysis on pairs of items, making it easier to interpret the associations between items. By printing the frequent itemsets, we observe pairs such as `(EV, GIDA)` with a support of 0.350101 and `(EV, KOZMETIK)` with a support of 0.340887. This targeted approach helps in identifying and

understanding the relationships between pairs of items, which can be particularly useful for cross-promotional strategies and optimizing product placements.

4.6.4. Association Rules Calculations

By following the past steps, we can calculate and refine association rules that reveal significant and meaningful relationships between items in the dataset

- **Calculating Association Rules with Support Threshold of 0.0001:**
  In this step, we calculate the association rules from the frequent itemsets with a support threshold of 0.0001. This means that we consider itemsets appearing in at least 0.01% of the transactions. The association rules are generated using the 'support' metric, and the first 20 rules are displayed. The results include columns such as antecedents, consequents, support, confidence, lift, leverage, conviction, and Zhang's metric. And here is the key Metrics:
  - **Antecedents & Consequents**: The itemsets on the left-hand side and right-hand side of the rule.
  - **Antecedent Support**: The proportion of transactions that contain the antecedent.
  - **Consequent Support**: The proportion of transactions that contain the consequent.
  - **Support**: The proportion of transactions that contain both the antecedent and consequent.
  - **Confidence**: The probability of the consequent given the antecedent (support of both / support of antecedent).
  - **Lift**: The ratio of the observed support to that expected if the antecedent and consequent were independent (confidence / consequent support).
  - **Leverage**: The difference between the observed support and the expected support if the antecedent and consequent were independent.
  - **Conviction**: A measure of the **rule's implication, considering the directionality (1**- consequent support) / (1 - confidence).
  - **Zhang's Metric**: A measure of the strength of the association.

  For instance, the rule (BALIK) -> (EV) has a support of 0.004510 and a confidence of 0.833647, indicating that 83.36% of transactions containing BALIK also contain EV. This step helps identify relationships between items based on their co-occurrence in the dataset.

- **Calculating Association Rules with Confidence Threshold of 0.01:**
  Next, we calculate the association rules using a confidence threshold of 0.01. Confidence measures the probability that the consequent is present in transactions where the antecedent is present. The rules with confidence values greater than 0.01 are displayed. For example, the rule (BALIK) -> (EV) has a confidence of 0.833647, meaning that in 83.36% of the cases where BALIK is purchased, EV is also purchased. The association rules reveal that items such as BALIK frequently co-occur with EV, GIDA, and KOZMETIK, while BEBEK shows strong associations with CAY-KAHVE-SEKER, DETERJAN, and EV, indicating significant purchasing patterns and positive associations among these categories.

This step helps focus on the strength of the association between items rather than just their frequency.

- **Filtering Rules by Consequent Support Above 0.095**
  We further refine the association rules by selecting only those with a consequent support greater than 0.095. Consequent support indicates how frequently the consequent appears in the dataset. By filtering for higher consequent support, we ensure that the rules include consequents that are more significant in the context of the entire dataset. For instance, items like BALIK, BEBEK, and SOGUK ICECEKLER have notable associations with commonly purchased categories such as EV, GIDA, KOZMETIK, and SUT. These patterns can be useful for understanding customer behavior and optimizing product placements and marketing strategies. This filtering step narrows down the rules to those with more prominent consequences.

- **Filtering Rules by Leverage Higher Than 0.0:**
  Leverage measures the difference between the observed frequency of the antecedent and consequent appearing together, and the frequency expected if they were independent. We select rules with leverage greater than 0.0, indicating a positive association between antecedents and consequents. For example, the rule (BALIK) -> (EV) has a leverage of 0.000127, signifying that the items appear together more frequently than expected by chance. This step highlights associations with a meaningful degree of co-occurrence.

- **Filtering Rules by Zhang's Metric Greater Than 0:**
  Finally, we filter the rules using Zhang's metric, which evaluates the significance of an association. We select rules with a Zhang's metric greater than 0, ensuring that only statistically significant associations are considered. The final set of rules represents item relationships that are not only frequent and confident but also statistically relevant. For instance, rules with higher Zhang's metric values indicate stronger and more meaningful associations

4.7. Dynamic Pricing Implementation

4.7.1. ML Models

As we mentioned in the previous chapter DP is a (), so we started by searching for a suitable model to predict the new price of a product based on some factors determined by our data, and we find 6 models may lead to good and related results and here is the next 6 ML models we choose to test it and check if it may lead to good result:

1. **Decision tree model** is a machine learning method that predicts continuous numerical values by building a model that resembles a tree. Decision tree regression is concerned with estimating **numerical outcomes**, as opposed to classification problems where the output is categorical.

2. **The linear regression model** is a statistically supervised learning method that uses one or more independent features to establish a linear connection to predict the quantitative variable.
   It assists in determining: → Whether an independent variable predicts the dependent variable well; → Which independent variable significantly influences the dependent variable's prediction.

3. **Bayesian linear regression model** is a model that creates linear regression by utilizing probability distributions as opposed to point estimates. It is believed that the response, y, will come from a probability distribution rather than being assessed as a single value.

4. **Random forest model** is a popular machine learning algorithm that belongs to the supervised learning subfield. It can be applied to ML issues involving both classification and regression. Its foundation is the idea of ensemble learning, which is the process of merging several classifiers to solve a challenging issue and enhance the model's functionality.

5. **XGBoost model** is a machine learning algorithm in the class of ensemble learning and, more specifically, the gradient boosting framework that further uses decision trees as base learners and incorporates additional regularization techniques to improve model generalization. XGBoost is widely known for computational efficiency. Its efficient processing provides insightful feature importance analysis and seamless handling of missing values. That makes it an algorithm of choice for various tasks ranging from regression, classification, and ranking to other various data science challenges.

6. **CatBoost model** is a supervised machine learning method utilized by the Train Using AutoML tool. It leverages decision trees for classification and regression. The name CatBoost comes from two major features: it works with categorical data (Cat) and uses gradient boosting (Boost). Gradient boosting iteratively builds many decision trees. Every tree builds up from the result of the previous one, so their relative improvements are better. CatBoost improves the primary gradient boost method for faster implementation

4.7.2. Data Splitting (Train/ Test)

After we choose the best models that may deal with our dataset and match our goal lets prepare the data for models training and testing
To train and test the model we need to split the data in two ways

1. Features and Target
   We have to choose the main features to use as X and target as Y, in our project we take columns (**“demand”**, **“Amount”**, and **“unit price”**) as features for model training so the model will use the demand in product and the amount that this transaction has from this product, and last is unit price, because the indication of the

new price will depends on the actual price of the product and the demand for the product on time and amount, and then we look to our target, generally we use the model to predict the price so we will use it also as the model target form column (**“Price”**)

2. Training and Testing
   We need to split the data to 2 main parts; first part is data for training the model and second part is data for testing the model after training
   There are many ways to split data like 50:50, 60:40, 70:30, 80:20, and 90:10 train-test ratios for the splits. For us, we choose the 70:30 strategy in data splitting 70% of the data for training and 30% of data for testing the model, and after that our data is finally ready for training and testing chosen ML models and let's go on train and test the data.

4.7.3. Training Models

For models training we must go into 2 main steps in each one training:

1. Initialize the model
2. Train the chosen model.

4.7.4. Testing Models

For testing the models, we will use 4 terms commonly used in regression analysis and machine learning to evaluate the performance of predictive models.
Here are:

- **Mean Absolute Error (MAE)** represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.
- **Mean Squared Error (MSE)** represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.
- **Root Mean Squared Error (RMSE)** is the square root of Mean Squared error. It measures the standard deviation of residuals.
- **R-squared** represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

The model works better if it has a lower value for MAE, MSE, and RMSE, which implies a higher accuracy of the regression model. However, a higher value of R square is considered desirable.

4.7.5. Comparing between models

After testing each model separately, we will start by comparing between models to see which one has the minimum values for MAE, MSE, and RMSE. And having a maximum value for R-squared means the model is the most related one, and this model must be chosen.

4.7.6. Applying the right model

Applying the chosen model and using it for your work, starting with calculating the score for the model and checking if there is any way to enhance this score. We then apply it to price prediction in our next steps, and we end up here with a dynamic pricing model.

5. Chapter 5: Results

5.1. Integration between MBA and Dynamic Pricing

Relate between the two main topics (MBA, and DP) in our project will help us to achieve our goal, in enhancing sales levels with increasing customer satisfaction, here is to ways we employed to serve our goal

5.1.1. Identifying Complementary Products

MBA Insight: MBA will be able to pick up pairs or sets of products that go together in sales.
Dynamic Pricing Application: This information is useful to retailers in developing bundled pricing strategies.
Example: If bread and butter are closely purchased together, dynamic pricing can be exercised along with a discount when both items are bought, hence luring higher sales volumes.

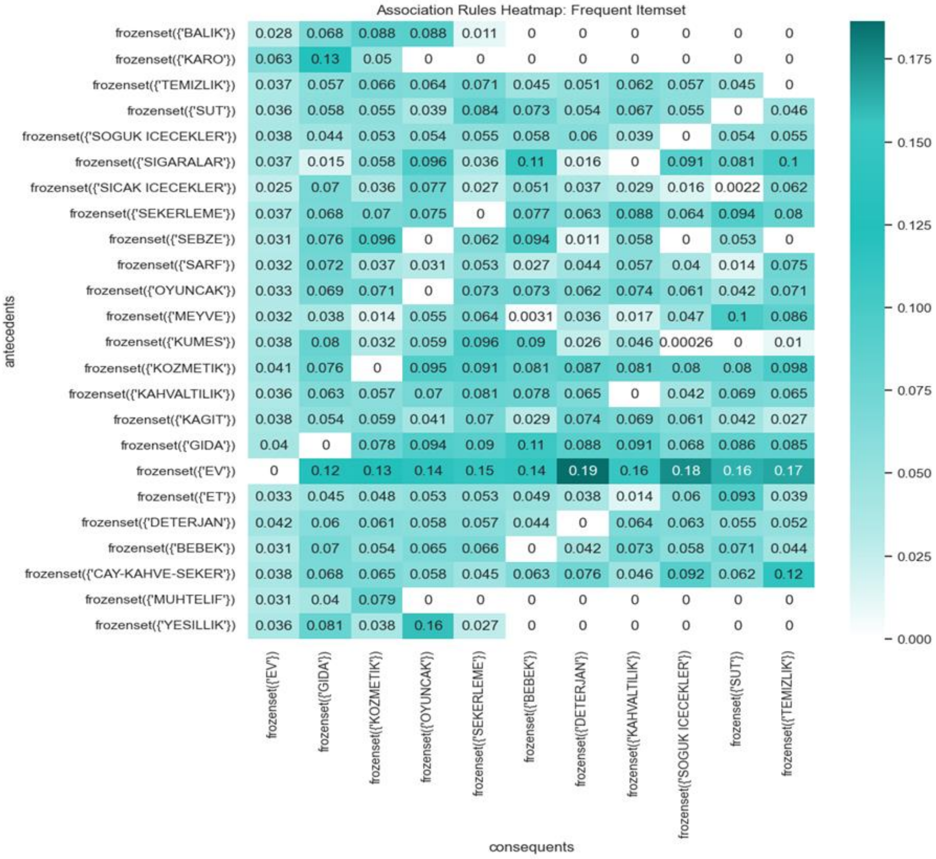
5.1.2. Optimizing Cross-Selling Opportunities

MBA Insight: MBA provides insight into which pair of products are usually bought together.
Dynamic Pricing Application: While the customer is making the purchase, the system can dynamically alter prices or may even indicate discounts to the relevant ones identified through MBA, so that the opportunity for cross-selling is enhanced more.
Example: If bread and butter are usually sold together, but this customer only buys one of them, dynamic pricing can be practiced in conjunction with recommending the other to him with a discount in the price of the second if he will still go forward and takes the first, thus attracting higher sales volumes.

5.2. Market Basket Analysis Result

5.2.1. Visualization Patterns

5.2.1.1. Heatmap for association rules: frequent itemset



	frozenset({'EV'})	frozenset({'GIDA'})	frozenset({'KOZMETIK'})	frozenset({'OYUNCAK'})	frozenset({'SEKERLEME'})	frozenset({'BEBEK'})	frozenset({'DETERJAN'})	frozenset({'KAHVALTILIK'})	frozenset({'SOGUK ICECEKLER'})	frozenset({'SUT'})	frozenset({'TEMIZLIK'})
frozenset({'BALIK'})	0.028	0.068	0.088	0.088	0.011	0	0	0	0	0	0
frozenset({'KARO'})	0.063	0.13	0.05	0	0	0	0	0	0	0	0
frozenset({'TEMIZLIK'})	0.037	0.057	0.066	0.064	0.071	0.045	0.051	0.062	0.057	0.045	0
frozenset({'SUT'})	0.036	0.058	0.055	0.039	0.084	0.073	0.054	0.067	0.055	0	0.046
frozenset({'SOGUK ICECEKLER'})	0.038	0.044	0.053	0.054	0.055	0.058	0.06	0.039	0	0.054	0.055
frozenset({'SIGARALAR'})	0.037	0.015	0.058	0.096	0.036	0.11	0.016	0	0.091	0.081	0.1
frozenset({'SICAK ICECEKLER'})	0.025	0.07	0.036	0.077	0.027	0.051	0.037	0.029	0.016	0.0022	0.062
frozenset({'SEKERLEME'})	0.037	0.068	0.07	0.075	0	0.077	0.063	0.088	0.064	0.094	0.08
frozenset({'SEBZE'})	0.031	0.076	0.096	0	0.062	0.094	0.011	0.058	0	0.053	0
frozenset({'SARF'})	0.032	0.072	0.037	0.031	0.053	0.027	0.044	0.057	0.04	0.014	0.075
frozenset({'OYUNCAK'})	0.033	0.069	0.071	0	0.073	0.073	0.062	0.074	0.061	0.042	0.071
frozenset({'MEYVE'})	0.032	0.038	0.014	0.055	0.064	0.0031	0.036	0.017	0.047	0.1	0.086
frozenset({'KUMES'})	0.038	0.08	0.032	0.059	0.096	0.09	0.026	0.046	0.00026	0	0.01
frozenset({'KOZMETIK'})	0.041	0.076	0	0.095	0.091	0.081	0.087	0.081	0.08	0.08	0.098
frozenset({'KAHVALTILIK'})	0.036	0.063	0.057	0.07	0.081	0.078	0.065	0	0.042	0.069	0.065
frozenset({'KAGIT'})	0.038	0.054	0.059	0.041	0.07	0.029	0.074	0.069	0.061	0.042	0.027
frozenset({'GIDA'})	0.04	0	0.078	0.094	0.09	0.11	0.088	0.091	0.068	0.086	0.085
frozenset({'EV'})	0	0.12	0.13	0.14	0.15	0.14	0.19	0.16	0.18	0.16	0.17
frozenset({'ET'})	0.033	0.045	0.048	0.053	0.053	0.049	0.038	0.014	0.06	0.093	0.039
frozenset({'DETERJAN'})	0.042	0.06	0.061	0.058	0.057	0.044	0	0.064	0.063	0.055	0.052
frozenset({'BEBEK'})	0.031	0.07	0.054	0.065	0.066	0	0.042	0.073	0.058	0.071	0.044
frozenset({'CAY-KAHVE-SEKER'})	0.038	0.068	0.065	0.058	0.045	0.063	0.076	0.046	0.092	0.062	0.12
frozenset({'MUHTELIF'})	0.031	0.04	0.079	0	0	0	0	0	0	0	0
frozenset({'YESILLIK'})	0.036	0.081	0.038	0.16	0.027	0	0	0	0	0	0

Figure 26(Associations Rules: Frequent Itemset)

Declaration of the Heatmap

Rows (Antecedents):

- Each row represents a different antecedent, which is the item or set of items on the left-hand side of the association rule. For example, (BALIK), (EV), (KAHVALTILIK), etc.

Columns (Consequents):

- Each column represents a different consequent, which is the item or set of items on the right-hand side of the association rule. For example, (EV), (KOZMETIK), (OYUNCAK), etc.

Cell Values (zhangs_metric):

- The value of each cell represents the zhangs_metric of the association rule, which is the proportion of transactions in the dataset that contain both the antecedent and the consequent. Higher values indicate stronger associations.

Color Intensity:

- The intensity of the color in each cell represents the magnitude of the support value. Darker colors indicate higher support values, meaning stronger associations. The color gradient on the right-hand side serves as a legend indicating the support value range.

Interpreting the Heatmap

High zhangs_metric Associations:

- **Frozen items like 'BALIK' (fish) and 'KARPUZ' (watermelon)**: These pairs have relatively high association values, suggesting that customers who buy 'BALIK' are also likely to buy 'KARPUZ' together.
- **'EV' (home) and several other categories**: The category 'EV' has high association values with 'KOZMETIK' (cosmetics), 'KAGIT' (paper), 'GIDA' (food), 'DETERJAN' (detergent), 'MUTFAK' (kitchen), and 'YESILLIK' (greens).

Medium zhangs_metric Associations:

- There are several moderate association strengths between various frozen items and other categories. For example, 'SEBZE' (vegetables) has moderate associations with 'KAGIT', 'SUT' (milk), and 'KOZMETIK'.
- 'MEYVE' (fruits) and 'EV' also show moderate associations with several other categories, indicating that customers often buy these items together.

Low zhangs_metric Associations:

- Lighter cells represent weaker associations. For example, the cell for (KAHVALTILIK) as an antecedent and (YESILIK) as a consequent has a lower zhangs_metric value.

Zero zhangs_metric:

- Cells with a value of 0 are either white or light, indicating no transactions containing the antecedent and the consequent.

Noticeable Patterns

Strong Self-Associations:

- Items like `(EV)` show strong self-associations, as indicated by the darkest cells along the diagonal, signifying high zhangs_metric for transactions containing the same item multiple times or the item being very frequent in the dataset.

Frequent Antecedents:

- Antecedents like `(EV)` and `(GIDA)` have several darker cells in their rows, indicating they frequently appear in association rules as antecedents.

Frequent Consequents:

- Consequents such as `(EV)` and `(KOZMETIK)` have darker cells in their columns, indicating they frequently appear in association rules as consequents.

5.2.1.2. Scatter plot for leverage vs. confidence

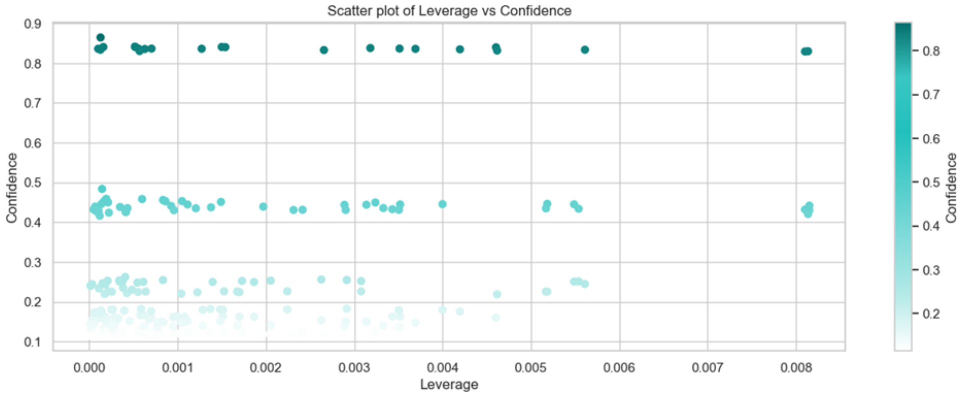


Figure 27(Leverage vs. Confidence)

<u>Declaration of the Scatter plot</u>

X-axis (Leverage):

- Leverage measures the difference in the probability of the antecedent and consequent appearing together in the dataset compared to if they were statistically independent.
- Higher leverage values indicate a stronger association beyond what would be expected by chance.

Y-axis (Confidence):

- Confidence measures how often the consequent is found in transactions that contain the antecedent.
- Higher confidence values indicate that the consequent is frequently observed when the antecedent is present.

Color Intensity:

- The color intensity of each point represents the confidence value, with darker colors indicating higher confidence. The color gradient on the right-hand side serves as a legend indicating the confidence value range.

<u>Interpreting the Scatter Plot:</u>

High Confidence, Low Leverage

- Points at the top of the plot but close to the y-axis (low leverage) indicate high confidence but low additional information beyond what would be expected by chance. This means that while the rule is reliable (high confidence), it may not be particularly surprising or informative (low leverage).

High Confidence, Moderate to High Leverage

- Points towards the top right of the plot (both high confidence and higher leverage) indicate strong, reliable, and informative rules. These rules are both highly reliable and significantly better than chance.

Low to Moderate Confidence

- Points lower on the y-axis indicate rules with lower confidence, suggesting that the consequent is not as frequently observed when the antecedent is present.

Sparse High Leverage

- There are fewer points with high leverage values, indicating that strong and informative associations are less common.

This scatter plot is useful for identifying and evaluating the strength and reliability of association rules in the dataset. High confidence rules (darker points) towards the top right of the plot are particularly strong and informative, while clusters of points at various confidence levels and low leverage indicate common but less surprising associations. The distribution of points provides insight into the overall pattern and strength of associations in the dataset.

5.2.1.3. Pareto chart for support of association rules

With access to the latest steps in applying the market basket analysis

- We sort the association rules based on their support values in descending order.
- Then calculating the cumulative support as a percentage. It divides each support value by the sum of all support values to get a relative frequency, then multiplies by 100 to convert it to a percentage. The ‘np.cumsum’ function computes the cumulative sum of these percentages.

So, the resulting graph has two key components:

- **Bar Plot (Left Y-Axis):**
  - Each vertical bar represents the support of an association rule.
  - The bars are sorted in descending order of support.
  - The y-axis on the left shows the support values.

- **Cumulative Distribution Function (Right Y-Axis):**
  - The dashed line represents the cumulative support percentage.
  - The y-axis on the right shows the cumulative support percentage.
  - The cumulative support line starts at 0% and increases as more rules are added, eventually reaching 100%.

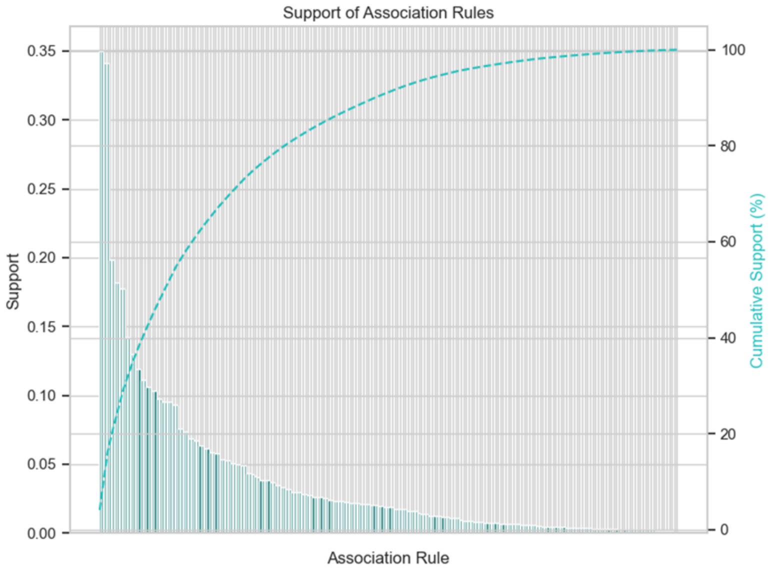


Figure 28(Support of Association Rules)

Insights

- The graph shows that a few association rules have high support, as indicated by the taller bars on the left.
- As more rules are considered, the cumulative support quickly rises, indicating that the top rules account for a significant portion of the total support.
- The cumulative support curve starts to plateau towards the end, suggesting that additional rules contribute less to the overall support.

5.2.1.4. Scatter plot for confidence vs. association rules

Graph Interpretation

The scatter plot represents the relationship between the support and confidence of association rules:

- **Support (x-axis):**

  This measures how frequently the itemset in the rule appear in the dataset. Higher support indicates that the itemset are more common.

- **Confidence (y-axis):**

  This measures the reliability of the inference made by the rule. Higher confidence indicates a stronger association between the itemset.

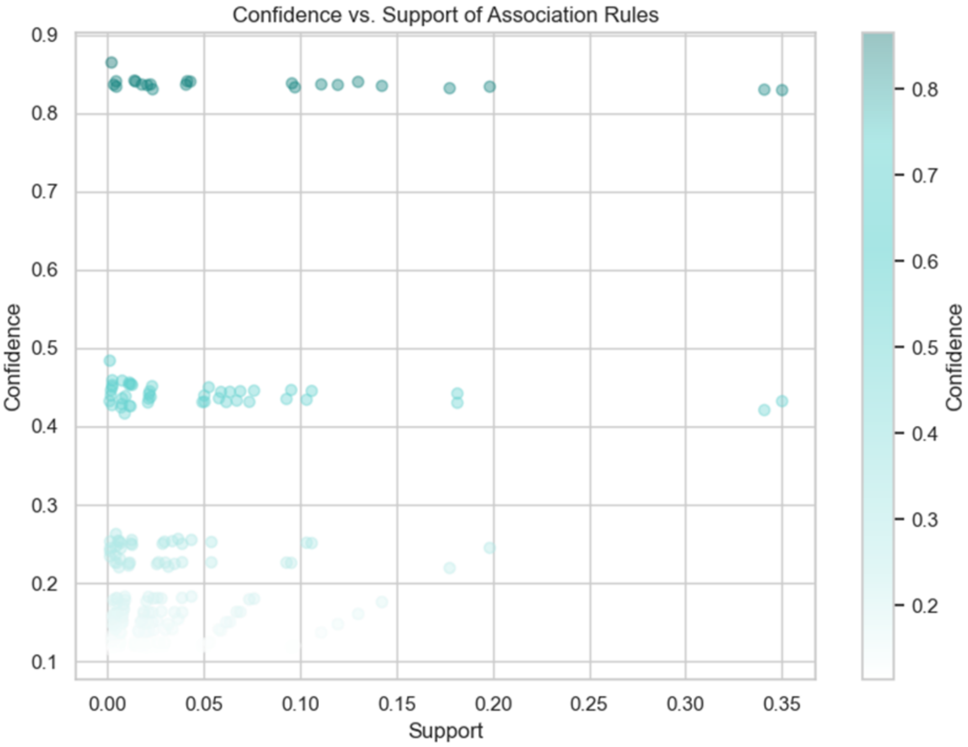


Figure 29(Confidence vs. Association Rules)

Insights

Clusters of Points:

- There are clusters of points at different levels of confidence, suggesting that some rules consistently have higher or lower confidence values.
- Most of the rules have low support values, as indicated by the concentration of points towards the left side of the plot.

Color Coding:

- The color of the points, which corresponds to the confidence value, varies across the plot. The color bar on the right helps to interpret these values.
- Darker colors (higher confidence) are scattered throughout the plot, indicating that confidence does not directly correlate with support.

Additional Observations

- Some rules with high support have varying confidence levels, showing that frequent itemsets do not always imply strong associations.
- Conversely, some rules with low support can have high confidence, indicating strong associations that occur less frequently.
- This visualization helps in understanding how support and confidence of association rules are distributed and their relationship, aiding in the selection of meaningful rules for further analysis.

5.2.1.5. Network diagram for cross-selling recommendations

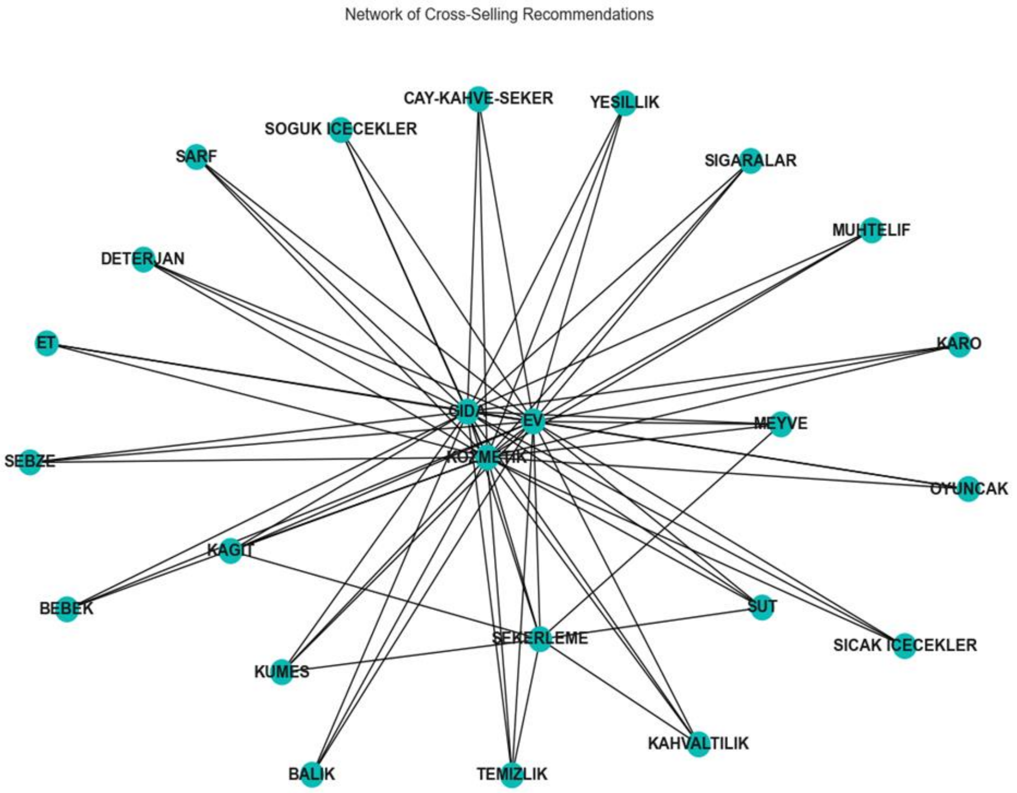


Figure 30(Cross-Selling Recommendations)

- **Filtering a set of association rules:** to identify potential cross-selling opportunities, focusing on rules where both the antecedents and consequents consist of single items. It then sorts these rules by confidence and support and selects the top 75 rules for visualization.
- The create_network_diagram function is defined to create a network diagram of these top cross-selling recommendations using NetworkX. The graph is undirected, where each node represents a product category, and edges between nodes indicate a recommended cross-sell relationship based on the association rules.
- Nodes are added to the graph with descriptive labels, and edges are created between antecedents and their consequents.
- For instance, if a customer buys one product, the connected product is recommended as a potential additional purchase. The network is designed to show which products are frequently bought together, helping businesses identify effective cross-selling strategies.
- Nodes are colored uniformly to represent the cross-selling rule type, and the layout ensures that the nodes are spread out for better visibility of the connections. The graph provides an intuitive way to see the relationships between different product categories and identify key items that are central to multiple cross-selling opportunities.

- Here is an example to the results of the network diagram representing cross-selling recommendations:
  Customers who bought 'KARO' (Tile) also bought 'EV'(Home).
  Customers who bought 'DETERJAN'(DETERGENT) also bought 'EV'(Home)..
  Customers who bought 'TEMIZLIK' (CLEANING)also bought 'EV'.
  Customers who bought 'SEKERLEME' (candy)also bought 'GIDA'(food).
  Customers who bought 'SICAK ICECEKLER'(hot drinks) also bought 'GIDA'.

5.2.2. Cross-Selling Insights from Market Basket Analysis

5.2.2.1. Cross-Selling Recommendations Printout

- Filtering the top 75 cross-selling recommendations and printing them in a readable format showcasing items that can be cross-sold effectively.
- This snippet iterates through the ‘top_cross_selling’ Data Frame, extracts the antecedent and consequent items for each rule, and prints a recommendation statement. The result is a list of statements such as "Customers who bought 'KARO' also bought 'EV'," indicates that customers who purchased 'KARO' tend to also buy 'EV'. This pattern is repeated for each product pair listed, providing insights into frequent purchase combinations.

5.2.2.2. Product Recommendation Functions for Cross-Selling Analysis

- **Check Product Function**
  - This function filters the Data Frame ‘df’ for rows where category1 matches the provided item and returns the unique name of the product. It's useful for validating and retrieving product names based on category.
- **Recommend Product Function**
  - The ‘recommend_product’ function suggests products to cross-sell based on association rules.
  - This function takes a product category and recommends up to five related products using sorted association rules. It checks for antecedents matching the input product and retrieves corresponding consequents. If multiple rules suggest the same consequent product, duplicates are removed using a dictionary conversion. Finally, it prints the top recommended products related to the input item. It then validates and prints the product names.
    - **For example**, calling ‘recommend product("YESILLIK")’ results in suggesting products that are frequently purchased alongside 'YESILLIK' which are:
      - {OYUNCAK, GIDA, KOZMETIK, EV, SEKERLEME}
      - This means customers who buy 'YESILLIK' often also purchase these other products, based on the association rules derived from transaction data.
    - **Another example,** calling ‘recommend product("EKERLEME")’ results in suggesting products that are frequently purchased alongside ' EKERLEME' which are:
      - {SUT, KAHVALTILIK, TEMIZLIK, BEBEK, OYUNCAK, KOZMETIK, GIDA, SOGUK ICECEKLER, DETERJAN, EV}.

- This means customers who buy ' EKERLEME ' often also purchase these other products, based on the association rules derived from transaction data.

5.2.3. Mean Price Calculation

- Computes the average price of items in a specific category, useful for pricing strategies, identifying pricing trends and understanding product value.
- Creating a temporary Data Frame containing only the ‘category1’ and ‘price’ columns, filters it for the "EV" category, and calculates the mean price. The prices for products in this category are displayed, and the mean price is stored in variable x can help in understanding customer spending behavior and preferences within the 'EV' category, aiding in personalized marketing and promotions.
- And such insights are integral to the dynamic pricing algorithm, allowing for more effective and responsive pricing strategies.

5.3. Dynamic pricing results

As we mentioned in the previous chapter, we go on in the next steps for get ready with dynamic pricing model:

- Choosing the models that may be useful
- Split the data
- Train the models
- Test the models
- Compare between models
- Choose the best model using models’ comparison criteria (MSE, MAE, RMSE, and R-square)
- Apply the chosen model
  Let’s go step by step checking the result of each point

5.3.1. Choosing the models that may be useful

After searching for models deeply and checking the pros and cons of each one we find 6 models (mentioned in the previous chapter) that may be useful and help us with the data nature, and our target with our features. Here are the 6 chosen models with her powerful point we choose it for:

Linear Regression

It is good to establish a baseline model. It is helpful for the relationship between unit price, amount, demand, and new price to be approximately linear, due to which it provides sensible predictions.

Bayesian Linear Regression

Gives point estimates and confidence intervals for predictions; for the task of price optimization, this is critically important. That is, one needs to understand the uncertainty around price predictions to make proper pricing decisions and manage risk.

Decision tree

Useful to capture complex relationships and interactions between unit price, amount, and demand.

Random forests
It provides a robust and powerful model for predicting new prices through the capture of complex relationships and interactions.

XGBBoost
It is a very powerful application for attaining the highest accuracy when making price predictions, especially when the data set is huge and relatively complex.

Catboost
It has in-built mechanisms to avoid overfitting, basically ordered boosting, gradient-based feature selection, or better generalization of details in not-shown data, which are very important in price prediction.

After choosing the model, let’s go on the data preparation for training and testing.

5.3.2. Split the data

We work on splitting the data for features and target our target Is predicting the price, according to our features → unit price (original price for item), demand, and the amount that taken in same transaction from same item.

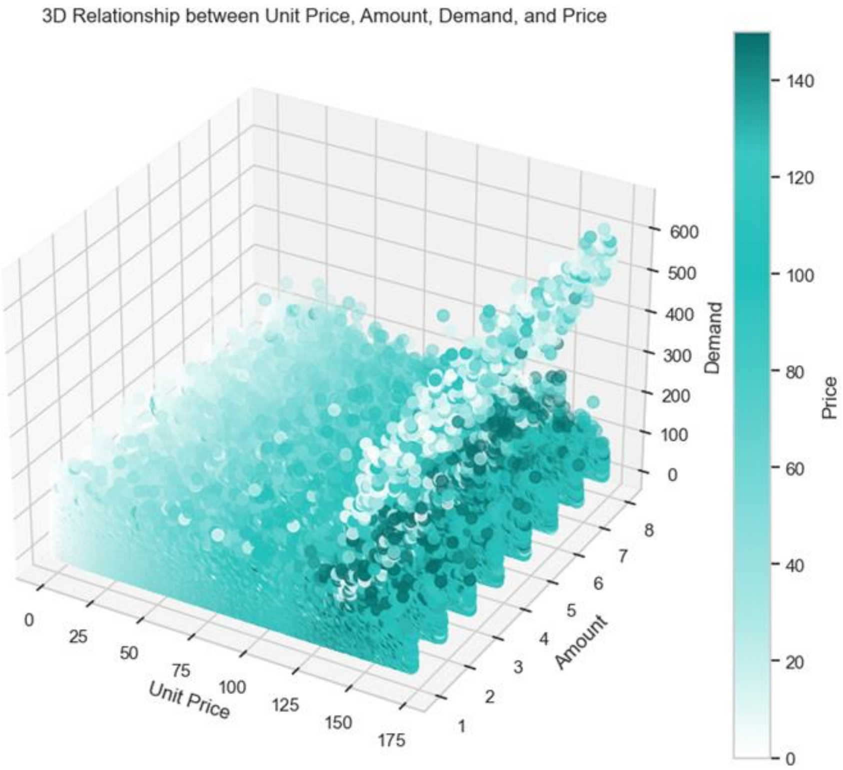


Figure 31(Relationship between UnitPrice, Amount, Demand and Price)

Then we split the data for training and testing, choosing strategy 70/30 where 70 percent of data for training and 30 percent for testing.

5.3.3. Results of Training Models

Each model was successfully trained, and this is its result

5.3.4. Results of Testing Models

As mentioned, we use MSE, RMSE, MAE, and R-square to test each model, and here are the results for each model:

1. **Decision tree**
   MSE is = 221.10913810991096
   RMSE is = 14.869739006112749
   MAE is = 9.443972887746218
   R-square is = 0.7979741392220726

2. **Linger regression**
   MSE is = 218.3422998006615
   RMSE is = 14.776410247440394
   MAE is = 9.269733688602258
   R-square is = 0.8005021798803542

3. **Bayesian linear regression model**
   MSE is = 218.34229892754442
   RMSE is = 14.776410217896105
   MAE is = 9.269741908522063
   R-square is = 0.8005021806781152

4. **Random forest model**
   MSE is = 218.73857728191004
   RMSE is = 14.78981329435602
   MAE is = 9.392603415584544
   R-square is = 0.8001401039393031

5. **XGBoost model**
   MSE is = 205.26683830093552
   RMSE is = 14.327136430596852
   MAE is = 8.803151502445408
   R-square is = 0.812449136876938

6. **CatBoost model**
   MSE is = 205.1027475948813
   RMSE is = 14.321408715447
   MAE is = 8.798734691965825
   R-square is = 0.8125990653982991

After training and testing, we must use only one model, so we must compare between models and choose the best one and apply it.

5.3.5. Comparing between models

The result that coming from training compounded in the next figure

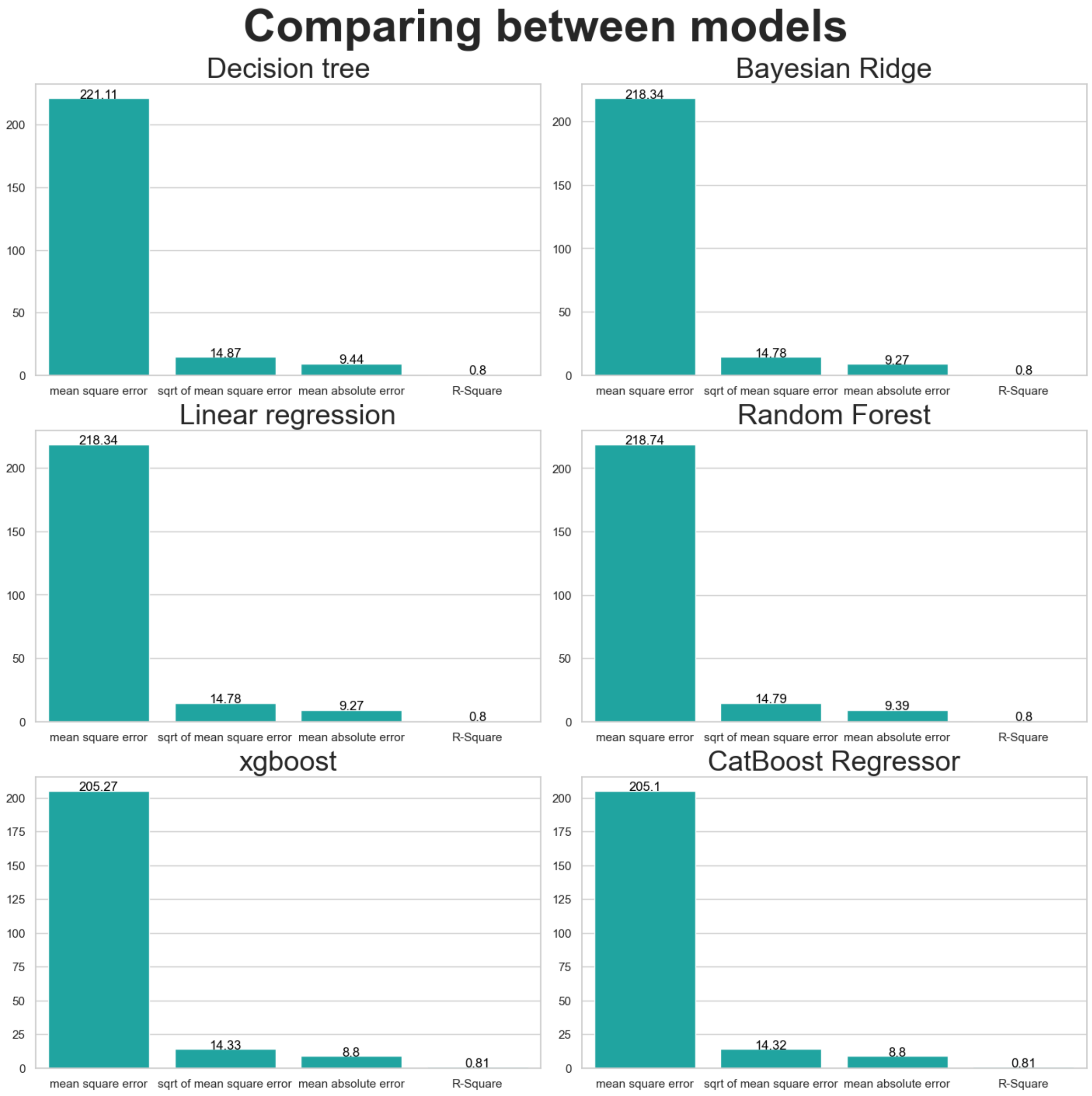


Figure 32(Comparing between models)

For us we chose CatBoost Regression model because it gives best results for MSE, MAE, RMSE, and R-square.

MSE is = 205.1027475948813
RMSE is = 14.321408715447
MAE is = 8.798734691965825
R-square is = 0.8125990653982991
Score percentage = 81.25990653982991

5.3.6. Applying the right model

We can call this step the result step because it is the result of all the previous steps and helps us reach our goal and start dynamically pricing products successfully. Deeply, we start to calculate the chosen model accuracy and it's **81.2599** and we compare the price before and after using the model to check if the model works well in pricing and the next figure shows the difference between before and after the model in pricing for product with **Product ID = 13492** (Actual vs Predicted)

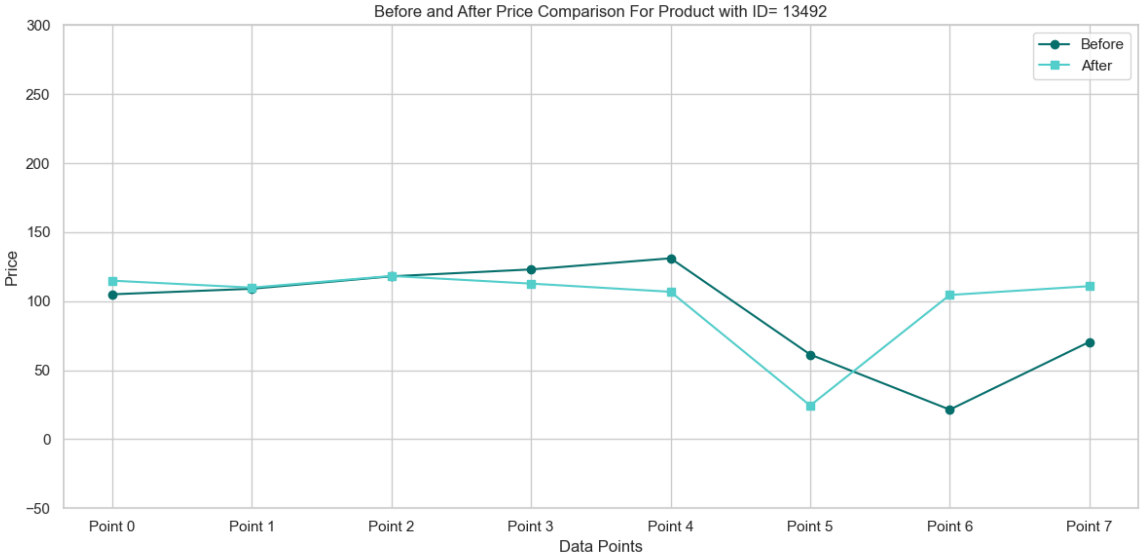


Figure 33(Before and After Price Comparison for Product with ID= 13492)

Then finally the model works well and can achieve our objectives.

5.4. Integration between MBA and Dynamic Pricing Results

In this step we go in some steps as next

- Display the result of MBA to Know the items that may be sold together.
- Check if the item in the next 2 cases:
  - If usual 2 items sold together, we recommend the 2 items in one bundle for all customers
  - If usual 2 items sold together and the customer take one of them, we recommend the second one for him with percentage discount

Example using product ID = 7438 from category EV and product ID = 18171 from category KARO, the original prices for both are 32.9 and 50.0, if the first product will be sold 5 items in 2021-04-19 the price of the product will be 20.5 and in same day the second one price will be 43.5, we have two ways to recommend products in that case

- Buy both items in bundle with price 20.5+43.5 = 46 and the original price was 32.9+50=83, so he can buy it in bundle with discount 23%
- Recommend the second one to him with discount 13% and price 43.5

6. Chapter 6: Conclusion and Future Work

6.1. Conclusion

In this project, we developed **Availa**, a comprehensive E-commerce Optimization Tool designed to address key challenges in the e-commerce sector. The primary objectives were to enhance dynamic pricing and conduct detailed market basket analysis. Through the application of advanced technologies such as machine learning and predictive analytics, AVAILA aims to transform the e-commerce landscape by improving both seller profitability and customer satisfaction.

Key accomplishments of this project include:

Dynamic Pricing Implementation: We developed a robust dynamic pricing system that adjusts prices in real-time based on market demand, competition, and consumer behavior. This system ensures competitive pricing strategies that maximize revenue while maintaining customer satisfaction.

Market Basket Analysis: By analyzing customer purchasing patterns, we provided valuable insights into product associations and customer preferences. This analysis helps businesses optimize product placement and improve overall customer experience.

Regarding the models, the CatBoost model emerged as the best performer with an R^2 score of 0.812599. Its ability to handle categorical features efficiently and its robustness against overfitting make it the ideal choice for AVAILA's dynamic pricing and market basket analysis modules. The close performance of the XGBoost model highlights the strength of gradient boosting techniques, while the Bayesian Ridge and Linear Regression models provide simpler, yet effective, alternatives. The Random Forest and Decision Tree models, although useful, were less accurate in this context.

So, we can say that The E-commerce Optimization Tool bridges the gap between seller profitability and customer satisfaction by leveraging advanced technologies and data analytics. Through dynamic pricing, our tool enables businesses to adjust prices in real-time based on market demand, competition, and consumer behavior, ensuring competitive pricing strategies that maximize revenue while maintaining customer satisfaction. Market basket analysis further empowers businesses with data-driven insights into customer purchasing patterns and product associations. This analysis allows for the optimization of product placement, and enhanced cross-selling opportunities. By providing a comprehensive understanding of customer behavior and market dynamics, our tool equips businesses with the knowledge and strategies needed to thrive in the competitive e-commerce landscape. The combination of these powerful features ensures that businesses can not only increase their profitability but also deliver a superior shopping experience that fosters customer loyalty and long-term success.

6.2. Future Work

While AVAILA has achieved significant milestones, there are several areas for future work to enhance its capabilities and address emerging challenges in the e-commerce industry. The following are key areas for future research and development:

- **Integration of More Advanced Machine Learning Models**: Future work can explore the integration of more sophisticated machine learning models to further improve the accuracy and efficiency of dynamic pricing, and market basket analysis.
- **Expansion to Other Markets**: Future work can focus on adapting and expanding the tool to other regional and global markets.
- **Enhanced User Interface and Experience**: Improving the user interface and experience of AVAILA will make it more accessible and user-friendly for businesses of all sizes. Future work can focus on developing intuitive dashboards and visualizations to present insights and recommendations effectively.
- **Sustainability Considerations**: As sustainability becomes increasingly important, future work can explore ways to incorporate sustainability metrics into AVAILA's optimization algorithms. This addition will help businesses make environmentally conscious decisions while maintaining profitability.
- **Advanced Fraud Detection and Prevention**: Enhancing the vendor credibility verification system with advanced fraud detection and prevention techniques will further ensure the safety and reliability of the marketplace.
- **Collaboration with Industry Stakeholders**: Collaborating with industry stakeholders, including e-commerce platforms, businesses, and regulatory bodies, can provide valuable insights and ensure that AVAILA remains relevant and effective in addressing real-world challenges.

7. REFRENCES

- https://www.kaggle.com/datasets/omercolakoglu/1m-rows-turkish-market-sales-dataset/data
- Hermina CI, Aishwaryalakshmi B, Gopalakrishnan B (2022) Market basket analysis for a supermarket. ResearchGate. Retrieved from https://www.researchgate.net/publication/365489098_MARKET_BASKET_ANALYSIS_FOR_A_SUPERMARKET
- Anthonio Joshua (2023) Customer Spending Pattern Analysis with Market Basket Analysis to Produce Product Strategy and Price Bundling at Kopi Soe Mekarwangi Bandung. Research of Social Science Economics and Management 3(1): 36-48. Retrieved from https://www.researchgate.net/publication/373544512_Customer_Spending_Pattern_Analysis_with_Market_Basket_Analysis_to_Produce_Product_Strategy_and_Price_Bundling_at_Kopi_Soe_Mekarwangi_Bandung
- Patil B, Khot L, Desai S (2022) A Study on Market Basket Analysis Using Apriori Algorithm. Gogte Institute of Technology. Retrieved from https://www.researchgate.net/publication/361372486_A_STUDY_ON_MARKET_BASKET_ANALYSIS_USING_APRIORI_ALGORITHM
- Singha K, Parthanadee P, Kessuvan A, Buddhakulsomsiri J (2024) Market Basket Analysis of a Health Food Store in Thailand: A Case Study. International Journal of Knowledge and Systems Science 15(1): 1-14. IGI Global. Retrieved from https://www.researchgate.net/publication/375566996_Market_Basket_Analysis_of_a_Health_Food_Store_in_Thailand_A_Case_Study
- Dio R, Dermawan AA, Putera DA (2023) Application of Market Basket Analysis on Beauty Clinic to Increasing Customer's Buying Decision. SinkrOn 8(3): 1348-1356. Retrieved from https://www.researchgate.net/publication/372037805_Application_of_Market_Basket_Analysis_on_Beauty_Clinic_to_Increasing_Customer's_Buying_Decision
- Dogan O (2023) Market Basket Analysis with Statistically Improved Association Rules Considering Product Details. University of Padua. Retrieved from https://www.researchgate.net/publication/369034639_Market_Basket_Analysis_with_Statistically_Improved_Association_Rules_Considering_Product_Details
- Mohd Rosli NAH, Teo NHI (2022) Market Basket Analysis using Apriori Algorithm: Grocery Items Recommendation. Advanced International Journal of Business, Entrepreneurship and SMEs 4(14): 01-09. Retrieved from https://www.researchgate.net/publication/367200162_MARKET_BASKET_ANALYSIS_USING_APRIORI_ALGORITHM_GROCERY_ITEMS_RECOMMENDATION
- Priyanyo AH, Arifa A (2022) Implementation of Market Basket Analysis with Apriori Algorithm in Minimarket. Jurnal Teknik Informatika (Jutif) 3(5): 1423-

1429. Retrieved from https://www.researchgate.net/publication/367610066_IMPLEMENTATION_OF_MARKET_BASKET_ANALYSIS_WITH_APRIORI_ALGORITHM_IN_MINIMARKET

- Patil B, Khot L, Desai S (2022) A Study on Market Basket Analysis Using Apriori Algorithm. Gogte Institute of Technology. Retrieved from https://www.researchgate.net/publication/361372486_A_STUDY_ON_MARKET_BASKET_ANALYSIS_USING_APRIORI_ALGORITHM
- Omol E, Onyango D, Mburu L, Abuonji P (2024) Apriori Algorithm and Market Basket Analysis to Uncover Consumer Buying Patterns: Case of a Kenyan Supermarket. Buana Information Technology and Computer Sciences (BIT and CS) 5(2): 51-63. Retrieved from https://www.researchgate.net/publication/381302082_Apriori_Algorithm_and_Market_Basket_Analysis_to_Uncover_Consumer_Buying_Patterns_Case_of_a_Kenyan_Supermarket
- Arora Y, Bhateja N, Goswami V, Kukreja R, Rajput A (2022) Market Basket Analysis using Apriori Algorithm. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) 10(3): 62-66. Retrieved from https://www.researchgate.net/publication/361817199_Market_Basket_Analysis_using_Apriori_Algorithm
- Hwang SB, Kim S (2006) Dynamic Pricing Algorithm for E-Commerce. In: Sobh T, Elleithy K (eds) Advances in Systems, Computing Sciences and Software Engineering. Springer, Dordrecht, pp 149–155. Retrieved from https://link.springer.com/chapter/10.1007/1-4020-5263-4_24
- Bertsimas D, Perakis G (2006) Dynamic Pricing: A Learning Approach. In: Lawphongpanich S, Hearn DW, Smith MJ (eds) Mathematical and Computational Models for Congestion Charging. Applied Optimization, vol 101. Springer, Boston, MA, pp 45–79. Retrieved from https://link.springer.com/chapter/10.1007/0-387-29645-X_3
- https://www.geeksforgeeks.org/apriori-algorithm/
- https://www.aporia.com/learn/machine-learning-for-business/dynamic-pricing-models-types-algorithms-and-best-practices/
- https://www.pricefx.com/learning-center/dynamic-pricing-strategy-types-the-top-10-with-examples/
- https://app.datacamp.com/learn/courses/market-basket-analysis-in-python
- https://research.aimultiple.com/dynamic-pricing/#different-modules-of-dynamic-pricing