Optimizing Cross-Selling with Market Basket Analysis

**CRISP-DM**

**ML CATEGORIES**

1. **Supervised Learning: Regression & Classification; X, y**
2. **Unsupervised Learning: Clustering**
3. **Reinforcement Learning: Deep Learning**

**NON TECHNICAL PRESENTATION/ DATA REPORT (.doc/.pdf)**

**OVERVIEW & DATA UNDERSTANDING**

Background Information; What Diabetes? How is it diagnosed? Challenges involved in diabetes diagnosis? Who are the stakeholders? Proposed Solution (Analysis & Modelling)? Projected Conclusion

Project overview

In the rapidly evolving retail landscape, businesses are increasingly relying on advanced technologies to stay competitive and meet the ever-growing expectations of consumers. Retailers are now concentrating on utilizing artificial intelligence (AI) to improve all facets of their business operations, from product pricing to marketing tactics and inventory management, as technology and data-driven decision-making drive increasingly sophisticated shopping patterns. At the core of these innovations are approaches like the Market Basket Analysis (MBA), dynamic pricing optimization, and customer segmentation. AI-driven systems evaluate enormous volumes of customer data using statistical and machine learning techniques to make judgments in real time that affect retail strategy and results. The goal of this project is to develop a comprehensive solution for retail enterprises by integrating these potent strategies into an AI-Driven Smart Retail Optimization System. These approaches allow the development of more personalized and effective retail optimization systems.

Market Basket Analysis is a data mining technique that identifies associations between products bought together by customers, thus allowing retailers to understand consumer trends. Thus retailers can improve product placement, suggest related products, and raise the possibility of cross-selling. This will ultimately enhance sales.

The Dynamic Pricing Optimization provides a more intelligent method of setting prices. AI models use real-time data analysis to modify product prices in response to market trends, competition pricing, inventory levels, and demand. By promptly adapting to changes in consumer behavior and outside variables like seasonal trends or economic upheavals, businesses can maintain their competitiveness and maximize profits using this flexible pricing technique. With customer segmentation, retailers can divide their customer base into discrete groups according to a range of traits, habits, and preferences. Market segmentation allows focused marketing campaigns and tailored product recommendations to promote consumer involvement and encourage loyalty. Key techniques used in customer segmentation include clustering algorithms such as K-means clustering and hierarchical clustering.

These three elements can work together to develop an integrated, AI-powered solution that makes traditional retail a more data-driven and flexible sector. Retailers can now make better decisions, improve consumer experiences, and increase operational efficiencies in previously unachievable ways thanks to AI.

**Challenges**

Advanced AI models can be difficult to combine with retailers' old systems or incompatible databases. Smooth integration is important for real-time decision-making and implementation.

 Due to fluctuating consumer tastes, economic conditions, and rival tactics, the retail industry can be quite unstable. AI models must be able to precisely and swiftly adjust to changing circumstances.

**Stakeholders**

1. Retailers

The primary stakeholders for these systems are the retailers. They can optimize pricing strategies, customer’s engagements and product recommendations.

2.     Marketing teams

Marketing teams can utilize the insights from the designed system by targeted campaigns and promotional offers that are tailored to specific customer segments.

**Proposed solutions**

Three key elements will be integrated to create the AI-powered Smart Retail Optimization System:

1. Market Basket Analysis: Using information from previous transactions, common item sets will be found, and association rules that advocate the bundling of products will be developed. This will also inform the up-selling strategies.

2. Dynamic Pricing Optimization: Using machine learning techniques, regression models, and reinforcement learning, a pricing engine will dynamically modify product prices in response to rival pricing, customer purchase trends, and demand elasticity. The model will consider time-sensitive components like promotions, holidays, and flash sales.

3. Customer Segmentation: Customer data will be processed using clustering techniques such as K-means clustering or hierarchical clustering to create discrete groups according to preferences and purchase patterns. In addition to suggesting products for individualized suggestions, the system will advise customized marketing initiatives like loyalty programs and discounts.

**Modelling Approach**

Model Training: Train the MBA, pricing, and segmentation models using machine learning techniques such as logistic regression, decision trees, or K-means and deep learning.

Integration: Ensure that the models interact seamlessly to create a unified AI-driven AI-powered retail system.

**Projected conclusion**

The successful implementation of the AI-Driven Smart Retail Optimization System will bring a host of impactful benefits:

Maximized Revenue: By leveraging dynamic pricing and market basket analysis, retailers can optimize pricing strategies and product assortments, driving higher sales and significantly boosting profit margins.

Enhanced Customer Experience: Personalized recommendations and tailored pricing, informed by detailed customer segmentation, will create a more engaging and relevant shopping journey, fostering deeper customer loyalty.

Optimized Operational Efficiency: Streamlined inventory management and precision-targeted marketing campaigns will reduce waste, improve stock control, and ensure that resources are deployed more effectively across the business.

Sustained Competitive Edge: Using AI-powered insights, retailers can swiftly adapt to market fluctuations and emerging consumer trends, positioning themselves ahead of competitors and ensuring long-term business success.

In essence, this system will empower retailers to make data-driven decisions that improve customer satisfaction and drive sustained growth, securing a leadership position in a fiercely competitive retail landscape.

Problem Statement

Kenyan retail stores, particularly supermarkets and convenience stores, struggle with optimizing product placement and cross-selling due to a lack of data-driven insights. Key challenges include:

* **Ineffective Product Pairing** – Missed cross-selling opportunities due to limited visibility into frequently co-purchased items.
* **Suboptimal Store Layouts** – Poor product positioning reduces customer convenience and sales potential.
* **Limited Upselling Strategies** – Inability to design effective promotions, such as bundled discounts, due to insufficient purchase behavior insights.
* **Stocking Inefficiencies** – Failure to stock complementary products together leads to inconsistent availability and customer dissatisfaction.

Objectives

1. **Maximize Cross-Selling Opportunities** – Leverage Market Basket Analysis to identify frequently co-purchased products, optimize bundling, and drive incremental sales.
2. **Enhance Revenue with AI-Driven Pricing** – Implement dynamic pricing models to adjust prices based on demand, maximizing profitability and competitive advantage.
3. **Boost Customer Engagement Through Personalization** – Utilize advanced segmentation techniques to tailor marketing campaigns, improving conversion rates and customer retention.

Metrics of Success

**Accuracy Score** ≥ 85% (Correctly identified associations)

**Lift Score ≥ 1.2** – Measures the strength of association between products. A score >1.0 means buying one product increases the likelihood of purchasing another.

**Average Basket Size Growth +10%**. Tracks the increase in the number of items per transaction due to cross-selling strategies. A 10% growth means customers are buying more items in a single purchase.

**DATA UNDERSTANDING**

This dataset represents a full-year transactional analysis from **January to December 2024**, sourced directly from our client’s **internal sales system**. With **333,405 individual transactions**, the data reflects the performance of an **established, medium to large-scale retail operation** competing in a dynamic market.

To ensure confidentiality and maintain a strategic advantage, the dataset has been **carefully anonymized**, protecting:

* **Product categories and specific items**
* **Business identifiers (store chains, franchises, or brand names)**
* **Location details**
* **Sensitive customer and business information**

Only **essential transaction details**—including **date, quantity, and price**—are retained, ensuring a data-driven approach to performance analysis while safeguarding proprietary insights.

This level of anonymization underscores a commitment to **privacy, security, and competitive intelligence**, allowing for a robust evaluation of sales trends, pricing strategies, and market positioning without exposing critical business assets.

The dataset contains 333,405 rows and 7 columns, with the following structure:

1. DATE – Timestamp of transactions (needs conversion to datetime format).
2. ANONYMIZED CATEGORY – Product categories.
3. ANONYMIZED PRODUCT – Specific products within each category.
4. ANONYMIZED BUSINESS – Business names (anonymized).
5. ANONYMIZED LOCATION – Locations where transactions took place.
6. QUANTITY – Number of units sold per transaction.
7. UNIT PRICE – Price per unit (stored as string with commas, needs conversion to numerical format)

**DATA PREPARATION & ANALYSIS**

Discuss what to check on your data like missing, duplicate values, outliers, null values.   
Discuss how to deal with them? Percentage of the missing, null, duplicate values and is imputation, drop, forward/backfill the option to use

**Data Analysis**

Discussing EDA; Univariate, Bivariate & Multivariate Analysis

Have the visuals from the Technical work

**MODELING**

Mentioning the Models to use; have like 3 models (Baseline Model, 2nd Model, 3rd Model (hyperparameter tuned))

Explain (Justification) why you’re using the mentioned models.

Refers to your metric of success mentioned above

**EVALUATION**

Discussing each model and it performance

Mention the one that performed best and why

**CONCLUSIONS**

What are your findings from the analysis and modeling

**RECOMMENDATIONS**

Giving suggestions based on your findings; analysis and modeling

**NEXT STEPS**

Deployment for access to end users

Collecting more data points (a, b, c)

**TECHNICAL PRESENTATION (.ipynb)**

**OVERVIEW & DATA UNDERSTANDING**

Miniature version of what is contained in the data report

**DATA UNDERSTANDING**

Load Data, Doc String to describe the columns, .shape, .info, .describe

**DATA PREPARATION**

Technical implementation of what is captured in non-technical notebook

**Data Analysis**

Plotting the visuals; matplotlib

**MODELING**

Actual model implementation

Actually testing to see if it meets the metrics of success

**EVALUATION**

Rank the models based on their metrics of success

**CONCLUSIONS**

What are your findings from the analysis and modeling

**RECOMMENDATIONS**

Giving suggestions based on your findings; analysis and modeling; point form

**NEXT STEPS**

Deployment for access to end users

Collecting more data points (a, b, c)

**CODE QUALITY**

More of functions & OOP

Less of procedural code

**SLIDE DECK**

Team Members

Overview/Introduction

Problem Statement

Objectives

Analysis; Most important visuals

Modeling Results (Image/Diagram)

Recommendations

Conclusions

Future Work/Next Steps

**PROJECT MANAGEMENT TOOL**

Jira

Trello

Asana

DAY 1

Data Sourcing/Mining

Data Understanding  
  
DAY 2

Data Preparation   
Data Analysis

DAY 3

Modeling

Tune Modeling

DAY 4

Finalizing and Improving the Data Report

Slide Deck