

**AID 311**

**Mathematics of Data Science**

Retail Transactions Dataset

**Submitted by**

Aly Mansour

320210335

CSIT – AID (2)

Section (5)

**Submitted To**

Dr. Ahmed Anter

# Abstract:

The Retail Transactions Dataset, brimming with information about purchases, customers, and products, offers a captivating glimpse into the intricate world of consumer behavior. It's like a treasure chest overflowing with gold coins, each coin representing a transaction, a customer, or a product, waiting to be unearthed and examined. By delving deeper into its contents, we can unlock a wealth of insights that can transform retail strategies and fuel business growth.

One of the most potent tools in this arsenal is market basket analysis. Imagine yourself peering over a customer's shoulder as they fill their cart. This analysis reveals the hidden patterns and secret alliances forged between products. We discover the inseparable partners like bread and milk, the complementary companions like chips and salsa, and the potential cross-sell opportunities lurking in the aisles. By understanding these partnerships, retailers can optimize product placement, suggest enticing bundles, and tailor promotions to resonate with specific customer segments.

But the journey doesn't end with products. The dataset whispers secrets about the customers themselves. Customer segmentation acts as a decoder ring, transforming seemingly random purchases into distinct personas. We can identify high-value spenders who crave luxury items, budget-conscious shoppers seeking bargains, and health-conscious consumers prioritizing organic produce. By tailoring marketing messages and product offerings to each segment, retailers can forge deeper connections, foster loyalty, and drive targeted sales.

Pricing, the lifeblood of any business, also finds its voice in this dataset. Through pricing optimization, we can witness the delicate dance between price and demand. Certain products, like designer handbags, might be relatively inelastic, meaning even a slight price increase won't deter eager buyers. Others, like everyday essentials, might be highly elastic, with even a small price hike sending customers fleeing to competitors. This knowledge empowers retailers to set optimal prices that maximize profitability while remaining competitive.

The dataset doesn't just paint a static picture of a single moment. It allows us to observe the dynamic ebb and flow of sales through retail analytics. We can track trends across products, categories, stores, and regions, identifying seasonal fluctuations, the impact of external events, and the effectiveness of marketing campaigns. Like a weather vane, this analysis helps retailers anticipate shifts in consumer preferences and adjust their sails accordingly.

But the true magic lies in the details. The dataset whispers tales of urban dwellers who prefer online shopping, suburban families who frequent supermarkets, and young professionals gravitating towards convenience stores. It reveals the influence of payment methods, the impact of promotions on specific customer segments, and the subtle variations in purchasing behavior across seasons. By meticulously analyzing these nuances, retailers can personalize their offerings, refine their marketing strategies, and create a truly customer-centric experience.

In conclusion, the Retail Transactions Dataset is more than just a collection of numbers; it's a living, breathing tapestry woven from the threads of customer choices. By delving deeper into its complexities, we can illuminate the hidden patterns, decode customer preferences, and optimize retail strategies for sustainable success. So, let the exploration begin, for within this dataset lies the potential to unlock the secrets of consumer behavior and rewrite the rules of retail engagement.

# INTRODUCTION:

The main problem this dataset seeks to address is the challenge of understanding and predicting customer behavior in a retail setting. It addresses Identifying Product Associations and Patterns, Optimizing Pricing Strategies and Analyzing Sales Trends and Performance.

Concerning the techniques used:  
We applied some feature reduction strategies like Linear Discriminate Analysis (LDA), Principle Component Analysis (PCA) and Singular Value Decomposition (SVD).  
Then we applied model implementation like:

* Naïve Bayesian
* Bayesian Belief Network
* Decision Tree
* LDA
* Neural Network
* K-NN

Then Some Model Evaluations.   
  
For the contributions made in this project, I showed some stats in the form of graphs (heat maps and scatter plots), got some accuracy scores like Naive Bayes Accuracy, Decision Tree Accuracy, LDA Accuracy and PCA Accuracy.   
  
  


# RELATED WORK:

Concerning Kaggle, there is some notebooks which barely have similar work. Some only draw graphs from the dataset and some is incomplete but there is one which has applied RANDOMFORESTCLASSIFIER model focused on the products bought and the time of the year with the highest sales record.

General Recommendation Systems and Transaction Data Mining for Recommendations

Content-Based Method:

• Recommends products based on product comparison.

• Limited by inaccessible content information.

Collaborative Filtering:

• Predicts product utility based on user ratings.

• Classified into memory-based methods (KNN) and model-based methods (Matrix factorization model).

• KNN predicts based on entire collection of previous rated products.

• Matrix factorization models map products and users into a low-dimensional latent space.

Transaction Data Mining for Recommendations:

• Rule-based models use data mining algorithms to recommend products.

• Generates spurious patterns, not relevant to recommendations.

• Rendle et al. propose a factorization model emphasizing product correlations.

• Xiang Wu et al. utilize relations of products in the same transaction for recommendations.

# METHODOLOGY:

We started by doing some data exploration and getting some basic values.  
Got some numerical values like Min, Max, Mean, Variance, etc.

We applied some feature reduction strategies like Linear Discriminate Analysis (LDA), Principle Component Analysis (PCA) and Singular Value Decomposition (SVD). We got their values in the form of matrices.

We also used Sequential model as the base for our neural network, added the input layer, hidden and output layer. Compiled it and run it. Then we used it to calculate Mean Squared error on our test.

We used Support Vector Classifier (SVC) as a classifier to get their Accuracy values and KNN classifier to print out the evaluation matrix + (Accuracy, Precision, Recall, F1 Score & ROC).

# PROPOSED MODEL:

A diagram of data processing steps

Description automatically generated

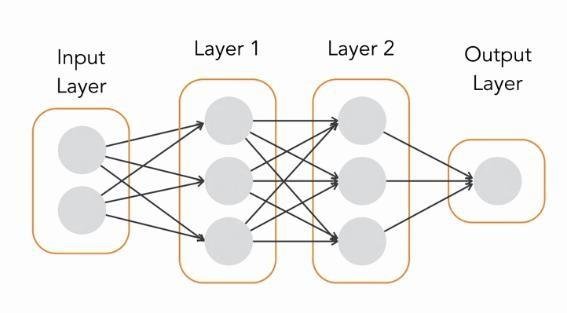
A diagram of a diagram

Description automatically generated

A diagram of a network

Description automatically generated

Sequential Model



# RESULTS:

In the data pre-processing phase, it was found that there is no null value at all in the dataset.

A screenshot of a computer screen

Description automatically generatedThese are the results of some basic computations.

We also encoded the 'Discount\_Applied' column into 1s and 0s.

A screenshot of a computer

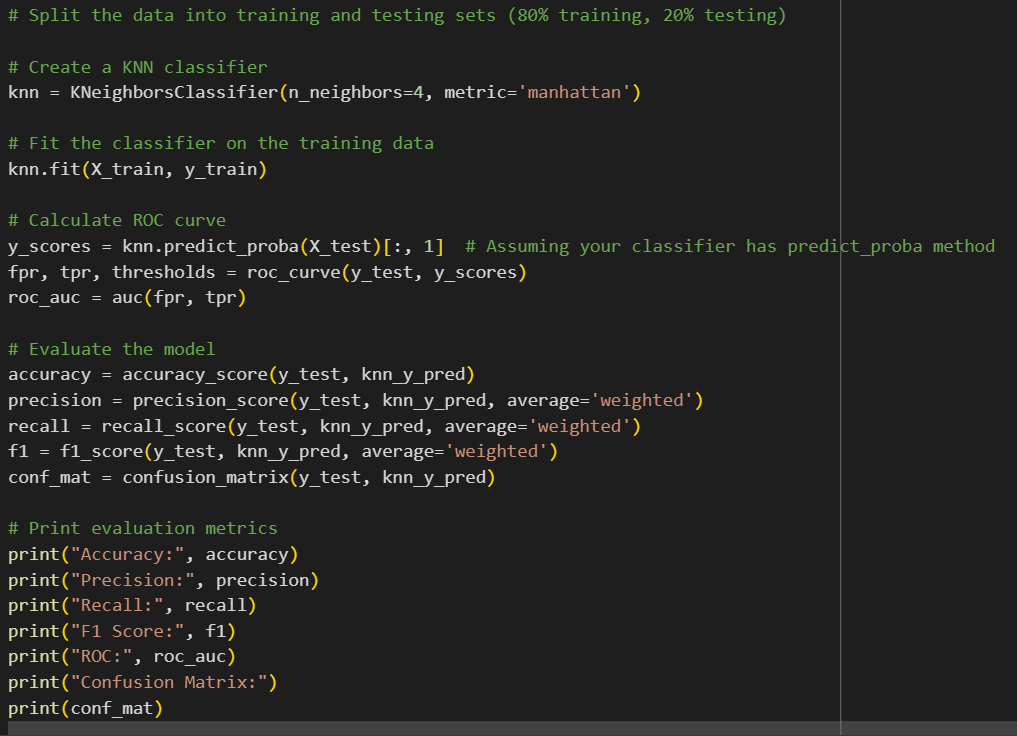
Description automatically generatedHere is the Data Analysis:

A black background with white text

Description automatically generated

A number on a black background

Description automatically generated



And for the neural network, the Sequential Model used in our work refers to a specific type of model architecture in which layers are linearly stacked on top of each other to form a neural network. This is a common and straightforward approach to designing neural networks, especially in frameworks like TensorFlow and Keras.

Concerning the conclusion, The model is underfitting bec the model has low accuracy, precision and recall. It may be underfitting, failing to capture the underlying patterns in the data.

Reference:

Yun, U. (2009). On Identifying Useful Patterns to Analyze Products in Retail Transaction Databases. *IEICE transactions on information and systems*, *92*(12), 2430-2438.

Spenrath, Y., Hassani, M., Dongen, B. V., & Tariq, H. (2020, September). Why did my consumer shop? Learning an efficient distance metric for retailer transaction data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 323-338). Cham: Springer International Publishing.

Doan, T., Veira, N., & Keng, B. (2018, November). Generating realistic sequences of customer-level transactions for retail datasets. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 820-827). IEEE.