CSE 519 - Project Proposal - Retail Sales Data Analysis

Introduction

The aim of every retail business is to attract new customers, retain existing customers and sell more to each customer. To ensure this, a retail business needs to offer customers the products they want at the right prices with a good consumer experience. In the ultra-competitive retail sector, traditional sources of decision-making such as executive experience and intuition are now insufficient. Leveraging data analytics to drive business and customer experience is the way forward. It helps companies to stay ahead of shopper trends by analyzing shopping behavior and act on meaningful data insights.

Problem Statement

Through this project, we are trying to provide a local chain, Costello Ace with insights that will help them with their business from the market basket data provided. The store operates 30+ stores and is spread across Long Island, Brooklyn, Staten Island, New Jersey and Maryland. The insights will help them to make customer-centric, strategic and operational decisions. More specifically we are trying to help them with the following decisions -
(1) Which products to promote?
(2) Which products to place near other products,?
(3) Which products to stock?
(4) Which to recommend to specific customers?

To gain a comprehensive view of business results and to improve operational performance across store's locations, we are also looking to analyze store level data to ensure customer experience across differential community requirements are met by identifying location specific trends and seasonal trends to maximize revenue.

Background

In this section, we are describing the significance of solving each problem and how it can help businesses and customers and then talk about background work in that space.

Product Promotion - Promotion is an essential aspect of a successful business. Its effects include brand establishment, growing within your target market segment thereby building sales and profits. Reference[1] talks about the most common strategies employed by businesses to identify products to promote. The most relevant to us are understanding the market demand and catering to it and looking at profit margins. We have added our novel approaches in detail in the Implementation section.

Product Placement Strategy - Reference **[2]** discusses about the approaches to find frequently bought items which are Frequent ItemSet mining **[4]** and Association Rule Mining **[5]**. **[6]** focusses on how Association Rule Mining is generally used in Market Basket data analysis. It also defines the set of measures to test the effectiveness of the rule we define. Association rule mining, at a basic level, involves the use of machine learning models to analyze data for patterns, or co-occurrence, in a database. It identifies frequent if-then associations, which are called association rules.

An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent.

Association rules are created by searching data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the data. Confidence indicates the number of times the if-then statements are found true. A third metric, called lift, can be used to compare confidence with expected confidence.

Association rules are calculated from itemsets, which are made up of two or more items. If rules are built from analyzing all the possible itemsets, there could be so many rules that the rules hold little meaning. With that, association rules are typically created from rules well-represented in data.

Product to Stock - From the strategies suggested in **[7]** one of the ideas we can experiment is using the time series data to analyze growth or decline of demand for products and forecast demand for the future. Other ideas in the blog discuss about relying on ground truth by testing hypothesis on a live market.

Recommendation Systems - [9] Here we provide a practical overview of recommender systems. First, three major systems are reviewed: content-based, collaborative filtering, and Nearest Neighbor. For our approach, one of the methods we are considering is the Nearest Neighbor based method as it a widely adopted method by Amazon, LinkedIn, Youtube etc. to recommend products/items. Here, we use cosine similarity for measuring the similarity between pairs of items or users. Similarity is calculated as follows -

$$similarity(a,b) = cos(a,b) = \frac{a \cdot b}{||a|| * ||b||}$$

Here we use a matrix where each element is a user vector i.e. items bought by a user then similarity between users can be calculated as cos(U1,U2).

Another approach we are considering is the Matrix factorization method, which uses Singular Value Decomposition -

Singular value decomposition (SVD) decomposes the preference matrix as

$$P_{m\times n} = U_{m\times m} \sum_{m\times n} V_{n\times n}$$

U and *V* are unitary matrices. For 4 users and 5 items, it looks like

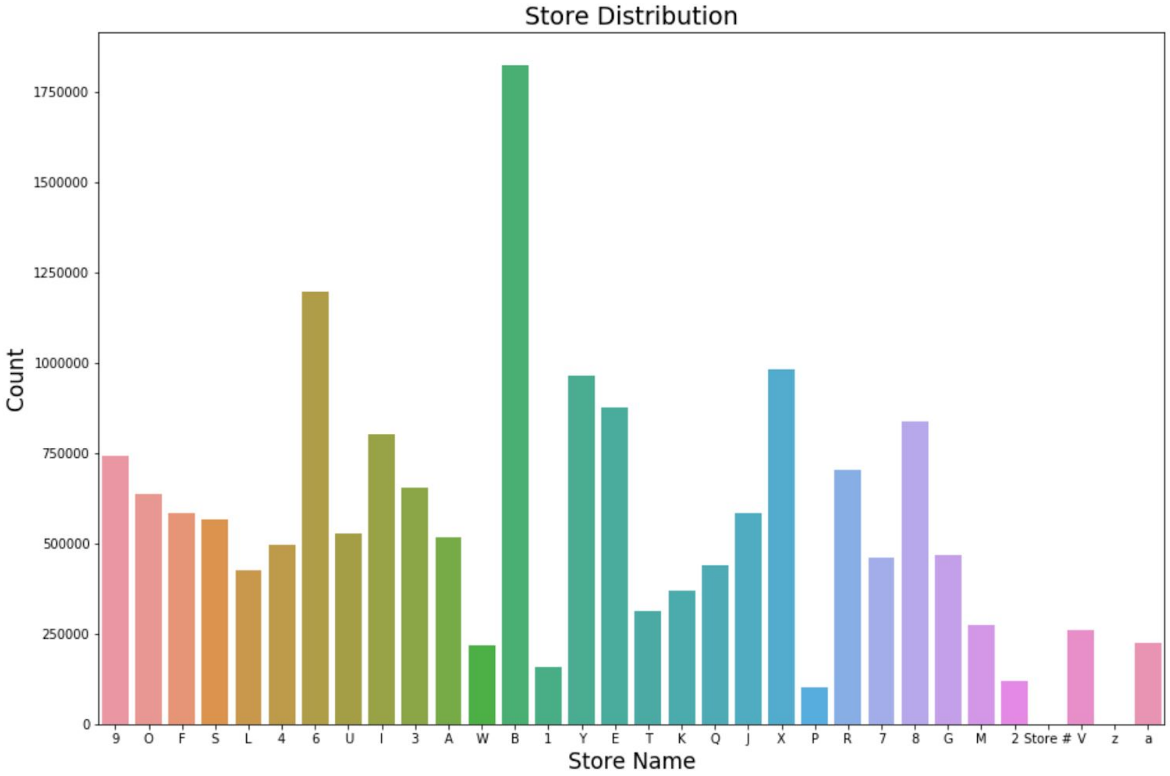
$$P_{4\times 5} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & u_{44} \end{bmatrix} \bullet \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 \end{bmatrix} \bullet \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} & v_{15} \\ v_{21} & v_{22} & v_{23} & v_{24} & v_{25} \\ v_{31} & v_{32} & v_{33} & v_{34} & v_{35} \\ v_{41} & v_{42} & v_{43} & v_{44} & v_{45} \\ v_{51} & v_{52} & v_{53} & v_{54} & v_{55} \end{bmatrix}$$

where *sigma_1 > sigma_2 > sigma_3 > sigma_4.*

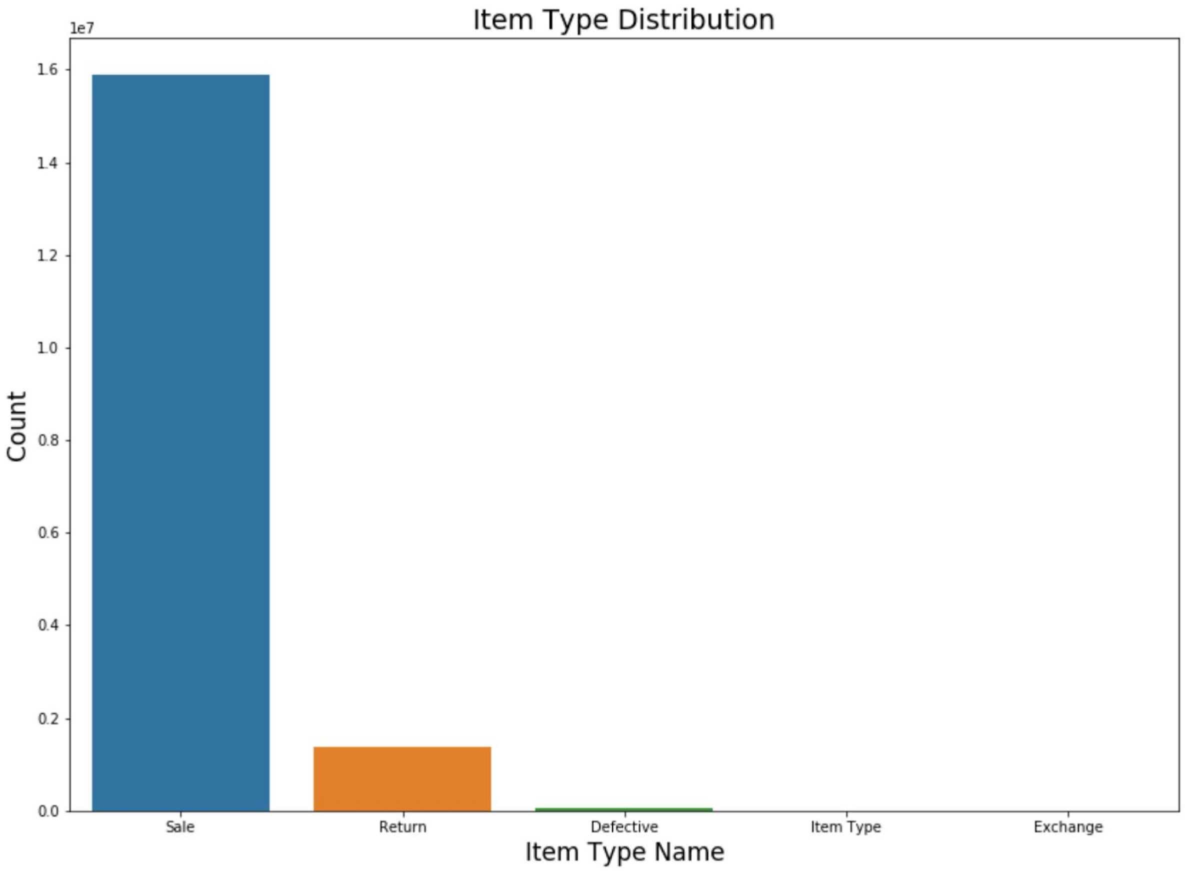
The preference of the first user for the first item can be written as

$$p_{11} = \sigma_1 u_{11} v_{11} + \sigma_2 u_{12} v_{21} + \sigma_3 u_{13} v_{31} + \sigma_4 u_{14} v_{41}$$

Preliminary Work



We plotted the number of sales happened in the various stores. The thought process behind plotting this was to find out the stores with the most or least number of sales. As we can see that stores B and 6 are having number of sales from April 2017 to Sept 2018. Also, there are stores with no sales. This is probably because there are no records for the sales in the store.



We also plotted count of sales happened for every item type. This tells us that how the distribution of sales are happened for two years. This gives the company an idea of how many items are returned or defective in the amount of items sold.

We think that no. of sales of a particular store and sales per item type will help us to analyze the store location and no of items being defective or returned in a particular store.

Dataset

- The shape of the data is (17328044,39) , i.e., 17328044 rows and 39 columns.
- **Columns** :

  'Date', 'Transaction Time', 'Customer Number', 'Receipt Number', 'Store #', 'Store Name', 'Scanned UPC', 'Item Number', 'Item Description', 'Net Sales Units', 'Net Sales', 'Cost', 'Gross Margin', 'Gross Margin %', 'Department Code', 'Department Name', 'Class Code', 'Class Name', 'Fineline Code', 'Fineline Name', 'Item was Scanned', 'MIP Promo ID', 'Promo/Discount', 'Dynamic Promo ID', 'Actual Price', 'Retail Price', 'Actual-Retail', 'Taxable', 'Tender Type', '$ Off Retail', 'Zip Code', 'Zip Plus-4', 'Loyalty ID', 'Clerk', 'Item Type', 'Line #', 'Line Item Transaction Type', 'Pricing Source', 'Return Code'

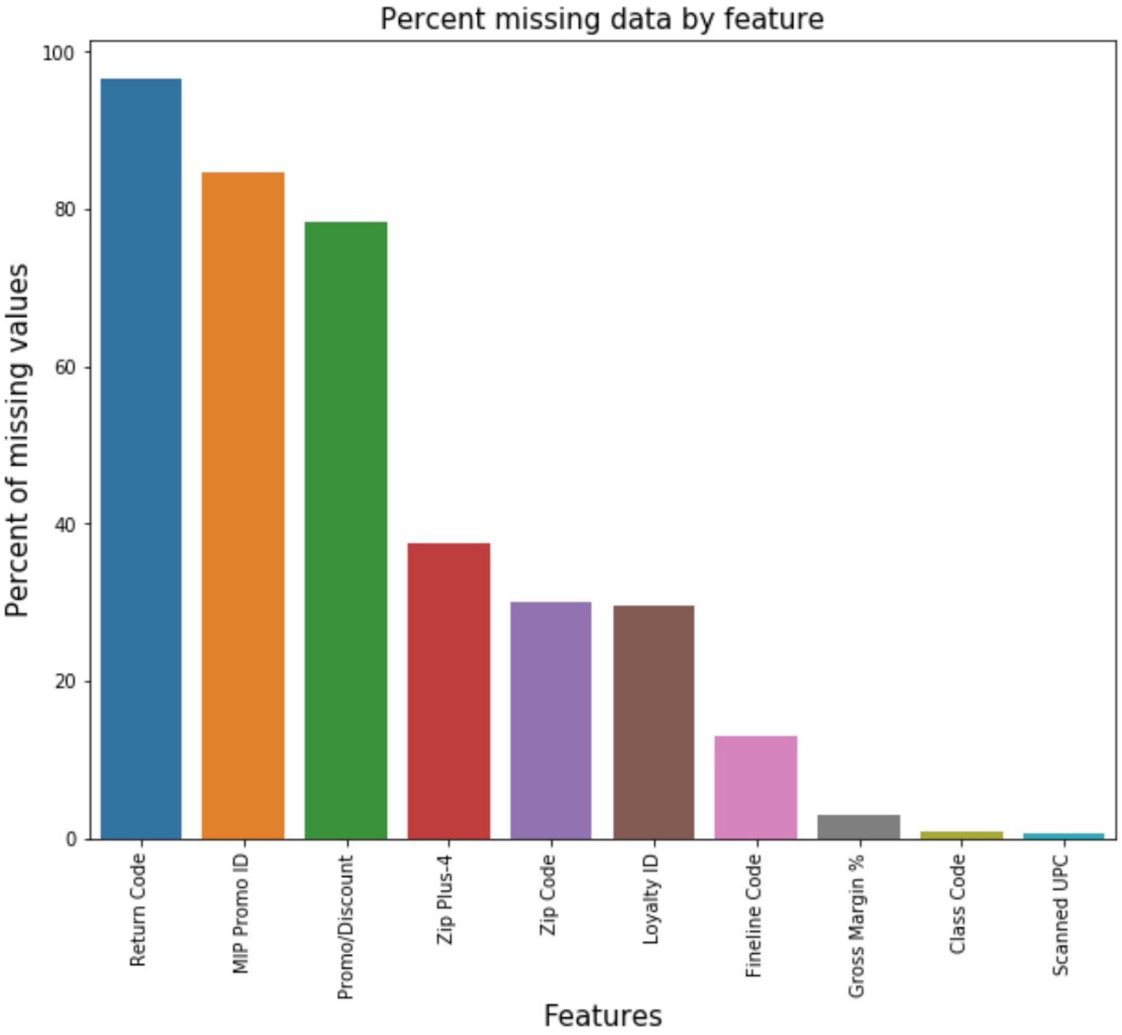
- **Column Description**

  Date - Sale date
  Transaction Time - Time of sale transaction
  Customer Number - Unique id for Customers
  Receipt Number - It is not unique throughout the dataset, so the unique id for a row is receipt number and store
  Store # - Store code
  Store Name - Name of the store
  Item Number - Unique purchased item
  Net Sales Units - No of items sold

- Null Values in Columns

Date	0
Transaction Time	1
Customer Number	63
Receipt Number	1
Store #	1
Store Name	1
Scanned UPC	121426
Item Number	1
Item Description	1
Net Sales Units	1
Net Sales	1
Cost	1
Gross Margin	1
Gross Margin %	505571
Department Code	62
Department Name	1
Class Code	144543
Class Name	4
Fineline Code	2239521
Fineline Name	1
Item was Scanned	10026
MIP Promo ID	14662753
Promo/Discount	13556315
Dynamic Promo ID	1
Actual Price	1

Retail Price	1
Actual-Retail	1
Taxable	1
Tender Type	2
$ Off Retail	1
Zip Code	5206254
Zip Plus-4	6512743
Loyalty ID	5122930
Clerk	937
Item Type	1
Line #	1



The above bar graphs show the top 10 columns having most null values. After careful consideration we are planning to drop the columns where 85% of values are null.

Implementation Details

We have segmented this section by each of the problem we are trying to solve and providing the implementation details.

1) Which products to promote?

Promotion is generally undertaken by stores to either increase sales of already known product or to create awareness about new product.

a. **Consumables** - Here we are looking at products that has been providing recurrent sales to the store. It would not matter if we continue to sell categories of product that customers are not buying. Hence it’s imperative we filter our product selection based on the pattern of products being sold to accommodate the customer’s preference.

Based on our assumptions we considered columns **StoreName** and **DepartmentCode** to know which particular category of products are highly sold at that particular location, here we consider top N products to promote, which is obtained by grouping the records by **StoreName** followed by getting count of products from each categories based on **DepartmentCode**.

b. **Profit Margin** - A product with good profit percentage but declining sales is an interesting case. It shows that a product with good margin was doing well but lost its popularity and signals a need for promotion.

   We can implement this by using columns **ItemNumber** vs **GrossMargin%**. **ItemNumber** (grouped by month) vs **Net Sales Units** units sold of that item. First plot a bar chart of **ItemNumber** and **GrossMargin%** to identify high profit products. Of the high margin products plot the time series data of units sold vs **ItemNumber** . If we identify a declining sales pattern(month over month, year over year) for any of them, we will choose it as-product to be promoted.

c. **New Product Awareness** - A new product from a category with recurrent sales is an ideal choice for promotion as people might be interested to know if the new product can provide solution better than what they already. Another way of looking at new products are products that have been promising showing growth potential.

   But, How to identify newly launched product?

Query the data frame by the year and see if a product never appeared before specific year but is appearing now in my dataframe. I.e. for X Product no entries prior to 2009 but entries after 2009, so we can infer the product was launched in 2009. The above process has to be done year wise(what range to do it for, yet to be decided.) After these products are revealed we will analyze their time series data to find the sales pattern. If we see a potential to grow i.e. an increasing sales, we will ask them to be promoted.

2) Which products to place near other products?

Cross-Merchandise - Placing goods that complement well with each other can drive impulse purchase and thus sales. For this analysis we should consider the different products a customer is buying when he comes to the store, for this analysis we consider ReceiptNumber, ItemNumber.

So from this we can know on arrival to store on a particular day what combination of products he ordered based on combination of columns **ReceiptNumber**, **ItemNumber** for a particular **StoreName** from the data set.

Here Uniqueness of the bill is by **ReceiptNumber** and **StoreName**.

3) Which products to stock?

Forecast Demand - Analyze the time series data of each product and find products with increasing sales trend either monthly/yearly. And based on percentage increase we can predict how much stock we should keep of the product for the coming month/year.
This analysis can be done on 2 levels, global level and store level.

Columns which are going to be useful here are **ItemNumber** group by **Date.**(**StoreName** is used if the analysis is specific to particular location)**.**

4) Which to recommend to specific customers?

On a higher level we can recommend a product to a customer as an answer to any one of the following questions:

- "Your selected product X goes well with product Y" (i.e. **product complementarity**)
- "Your product X was not available, so how about alternative Y?" (i.e. **product substitutability**)
- "Customers like you also tend to buy Y" (i.e. **customer similarity**)

How can Product Embedding be used to answer the above questions?
Due to the sparse nature of the product at hand i.e. a product can be coupled with all the remaining products in our selection as a recommendation for a customer, we will use product

embedding to recommend products to a customer. This approach is inspired by it’s application in NLP domain, since major data we are dealing there is sparse as well. Products are analogous to words and baskets/customers are analogous to sentences(e.g. A basket is represented with product codes).

Hence, on a higher level a product can be recommended to a customer based on the following criteria:

a. **History -** We recommend all the products that were bought by a customer in the recent pass but is not reflected in the recent sales history.

b. **Newer version of product -** We recommend newer versions of products that are already bought frequently by a customer.

c. **Best selling in that category -** Apart from single users buying history, we look at the best selling product of current category of product, bought by a customer based on aggregation of overall store basket data.

d. **Frequently bought together -** The items are typically complementary to the first item or have been purchased at the same time by a large number of previous customers.

Here to implement recommendation system we are planning to use **SVD**(Singular Value Decomposition), columns useful for this are **Customer Number, Item Number, Receipt Number, Department Code**(**Store Name** is used for analysis for a particular location).

5) Location and Seasonal trend analysis

Any predictable change or pattern in a time series that recurs or repeats over a period can be said to be seasonal. Analysis of sales pattern during the holiday season and/or other seasonal phenomenon can help predict how it affects sales. Whether sales increases or decreases during certain period can help to pre plan for manpower, sales and inventory management.

Here two external data sets are considered to analyse the problem statement. One dataset is Holidays dataset which ideally should have **‘Date’** and **‘HolidayDescription’** as columns, other dataset is for weather data with columns **‘Date’** , **‘ZipCode’**, **‘Temperature’**, **‘Weather Description’,** from weather data the datasets can be merged using column **‘ZipCode’** similarly for holidays data the datasets are merged using column **‘Date’.**

Challenges

- Cleaning without losing meaningful information, business demarcation. Data description is key to affirm our understanding of columns of data.

- After careful analysis of data, we found that all columns values are object type and most of the columns contains their names as their values.
- We need to convert object type columns to numeric columns for better analysis
- Numeric columns contain comma-separated values and contains symbols like '%' and '$'.

- The products primarily being sold are not perishables/consumables. So if a customer buys a hardware in a year, he may not buy the same hardware for years to come. Hence for the product categories in our problem definition for recommending products, we may not be able to leverage cyclical nature of consumption as is the case for FMCG products.

Validation

A naive strategy we can employ to measure our performance in answering the primary question can be:

Split the data into two sets, one from year X to Y and another from year Y to Z, where X<Y<Z. The first dataset can be used for training while the other for testing.

For product recommender systems, there are multiple other metrics which can be calculated to gauge how well we did. A summary of those metrics:

- Mean absolute error/Root mean absolute error
- Precision and recall
- Mean average precision

References

1. https://www.thebalancesmb.com/choosing-products-to-sell-2890471
2. https://datascience.stackexchange.com/questions/28611/what-is-a-product-most-frequently-bought-together-with
3. https://stats.stackexchange.com/questions/76373/how-to-figure-out-what-numbers-often-appear-together-in-a-dataset
4. http://www-ai.cs.tu-dortmund.de/LEHRE/SEMINARE/SS09/AKTARBEITENDESDM/FOLIEN/Frequent_Itemset_Mining_Methods.pdf
5. https://en.wikipedia.org/wiki/Association_rule_learning
6. https://towardsdatascience.com/association-rule-mining-be4122fc1793
7. https://hbr.org/2012/11/which-products-should-you-stock
8. http://www.cs.umd.edu/~samir/498/schafer01ecommerce.pdf
9. https://towardsdatascience.com/recommender-systems-in-practice-cef9033bb23a
10. https://towardsdatascience.com/collaborative-embeddings-for-lipstick-recommendations-98eccfa816bd