CSE 519 PROJECT PROPOSAL
Retail Sales Data Analysis

Problem Statement

Analysis and prediction of sales at Costello Ace hardware stores using the dataset from years 2017-2018.

Retail Sales Analysis is an examination of the wealth of information that the retail store proprietors have readily available. Ability to track and examine business decisions as well as customers' purchases and behaviours makes Retail Sales Analysis a powerful tool for organizations. Increasingly more Small-Medium organizations are using data-driven insights to make smarter decisions.

Background

Retail Store analysis has been a hot topic of research for the past couple of years. By virtue of such innovations and volumes of data in this domain, research has not been limited to basic analysis but has been extended to various predictive tasks as well. Some of the areas of research are:

- **Product Embeddings**

  Embeddings are dense representations of products and are traditionally used with NLPs. When learnt well, they produce similar vectors for similar items. These can be used to understand and improve product recommendations. The three most important aspects of product embeddings are:
  Complementarity: Products within the same transactions complement each other.
  Compatibility **:** Users tend to purchase products that match their preferences.
  Loyalty: A significant fraction of users repeatedly purchase the same products over time.

- **Retail Sales Prediction**

  Different times of the year might have different rates of sales, especially for some hardware products. For example, rakes might sell better during fall and shovels during winters. Identifying and predicting such patterns would be extremely valuable for the business.

- **Retail Inventory Management and Optimizations**

  Sales and predictive analysis will improve the proficiency with which the inventory is designed and managed. Optimizing inventory will maximize product availability and reduce losses to businesses by preventing product shelving.

- **Analyze and Predict Business Success Metrics**

  - ***Net Profit Margin***[1] - Net profit margin is the percentage of revenue the business makes per dollar of sales.
  - ***Store Success Rate*** - It is the Margin of Sales analysed for a store. It could also used to measure and improve employee (Sales rep/cashier) performance.

Preliminary Work

The dataset contains more than 17 million records (17,328,044). Each record lists the item sold. There are 39 parameters of sale recorded for each of the transaction.

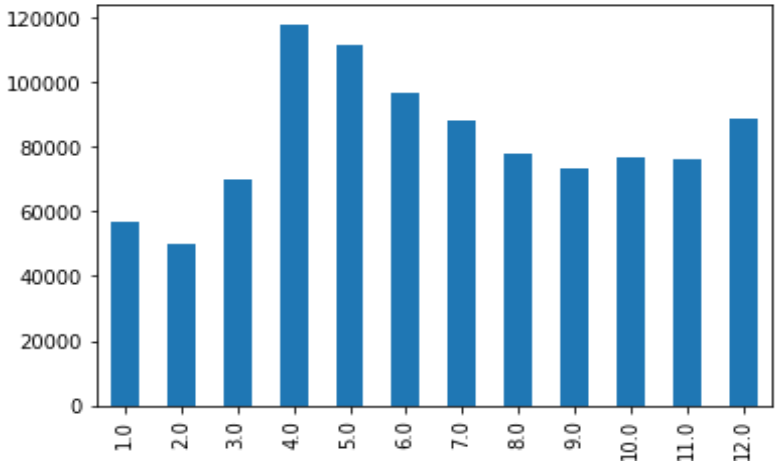
Data column characteristics

Customer Number	There are over 3.5 Lakh unique instances of customers.
Transaction Date and Time	They can be merged into a single DateTime object in Python.
Receipt Number	There are 13.2 Lakh receipt numbers. They aren't unique to a single transaction or a customer.
Store Number and Name	There are 31 Costello Ace stores along the east coast. The Store Number is mapped to the Store Name.
Item Number	The rows consist of 1.1 Lakh unique item numbers (Internal Identifier) indicating a variety of items at the hardware store. The items range from candy to barbeque grills worth $500.
Scanned UPC	UPC barcodes are standardized global identifiers, which enable products to be sold, reordered and tracked through supply chains. However, the item numbers are not mapped with Scanned UPC entries.
Class Number and Name	Classes within departments to which a particular product belong to.
Department Number and Name	Department number and name to which the product belongs to. There are 81 departments across Costello Ace stores.
Fineline Code and Fineline Name	Finelines are used to group inventory items into categories. They are updated frequently by a standard setting organization.
Promo/Discount, Loyalty ID	These are sparsely filled indicating promotional activities. Loyalty ID is assigned to the customers with repetitive purchases.
Line Item Transaction Type	Indicates the type of sale - Sale, Return, Defective and Exchange

Interesting Observations

- Checkouts are transactions consisting of items bought together by a customer. There are ~ 5.7 million unique checkouts. In a few cases, over 200 items have been purchased in a single checkout transaction indicating bulk orders.
- The transaction datetime need not be checked for timezone issues as all the store are along the same time zone.

- There are over 14 Lakhs *Return* transactions (8% of the total items) which would form a good basis for analysis on such huge margin of items returned.
- Over 4.4 Lakh records have the Cost Price lesser than Sale Price indicating that the store sold a few items at a loss.
- Analysis on monthly sales at Store X (Pasadena) and every other store shows a spike in the sales during the month of April. This could be due to the onset of summer when gardening and other repair work commences.



Monthly sales at Store X

Challenges

- The available dataset does not have a definite unique key based on which data can be cleanly processed. Interpretation of receipts is difficult because the same identifier is seen across multiple years and for multiple customers.
- Various columns have mixed data types which require careful cleaning. Furthermore, Data pre-processing is extremely limited by the hardware due to the large size of the data set.
- Sales prediction could be hampered due to the unavailability of data across multiple years.

Approach and Evaluation

As discussed in the challenges, there are repetitive receipt numbers which could be a hindrance for our analysis. To overcome this challenge, we have generated a feature by combining customer number and receipt number which we have verified to be unique.

Using the concept of Market Basket Analysis[3] based on the apriori algorithm, we try to map an item with its possible complements which can later be suggested to the customer based on the chosen item. This could be further improved by considering the department and the class of the product and matching similar ones. This would be the first step in developing a Product Recommendation Engine.

We will try to identify the retailer's keystone products, those that differentiate them in the market and could hurt the business if rendered unavailable or incompetent with the market standards. By doing this, we aim to keep the retailer informed on the keystone products and possibly come up with strategies like selling the product at a reduced profit margin while augmenting the sales volume.

As a part of predictive analysis, we will predict the business metrics for a given time period. We plan on accomplishing this task by running a regression algorithm on the same time period of data of previous years.

We can use data sold using promotional codes to increase the sales margin of certain popular items during particular seasons.

Future Work

Microsoft's *triple2vector* algorithm[2] is one best improvement that can be used to develop a robust product recommendation engine but its implementation is beyond the scope of this project because of the unavailability of libraries in python.

The results of the analysis could be better put into use if presented as an interactive application (Web based) to the Hardware Store.

The Product recommendation engine could be further applied to the e-commerce website of Costello Ace hardware stores to improve online sales.

Reference

[1] https://www.softwareadvice.com/resources/retail-data-analysis-to-boost-sales/
[2] https://www.microsoft.com/en-us/research/uploads/prod/2019/01/cikm18_mwan.pdf
[3] https://smartbridge.com/market-basket-analysis-101/