CSE 519 Project Proposal

Retail Sales Data Analysis

Objective:

To identify and suggest similar or related products that can be placed near to each other in the rack, so as to increase the sales of the products that are frequently bought together. This is achieved by building a model with the help of product embeddings, that understands the distributed representation of categorical product data and predicts similar products related to a particular product.

Background Work:

Retail data is increasing exponentially in volume, variety and value with each year. Many of the smart retailers know that these interactions hold insights for huge profit. Majority of the retail stores are already in the process of gaining insights on how to improve their sales, and provide better user experience to the customer by leveraging the power of data science techniques on their past sales data. Sales in the retail store can be increased by engaging the customers to buy multiple products. Past studies have shown that many customers tend to buy similar or related products when they are placed close to each other. So placement of products in the store plays a major factor in determining the sales. Also, placing similar or related products together will give an option for the customer to buy similar products from different brands, if stock is not available.

A local chain of retail stores is having a market basket data consisting of different products sold by each of its stores over several years. By investigating one year of the data, we can come up with directions to suggest products that can be placed near other products in order to increase sales. Being a novel objective, there is some research work done by concerned practitioners for similar problem. It has always been a challenging task to get most frequent entity which can go along with any particular set of other entities. Previous work[1] on finding similar words from a target words has used distributed word representation and found it as an efficient approach. This approach has been experimented with two algorithms namely sub sampling and negative sampling techniques with the 'Word2Vec' concepts over a large dataset without compromising training speed and performance because of higher frequencies above a certain threshold. It also showcases challenges faced in case of uncommon words and complexities in vector representations learning.

Given that our primary objective is to provide suggestions of a product based on the transaction history across the stores, it is required to explore the most related products to the products from the same invoices. A related work[2] showcases experiments on sequential transaction dataset of a Pharma retail industry by finding complementarities and similarities between products. It learns the representation of products in low dimensional space similar to the one used in Word2Vec and make suggestions by analyzing the transactions and all context products available in the same invoices. It has also worked with another approach where they calculate product to product co-occurrence score using ratio of co-occurrence probabilities of two products to figure out the similarities of complimentarity baskets and these computed embeddings i.e. Product to Product co-occurrence score are used to predict sales per year efficiently.

The objective is somewhat different than typical recommendation systems which suggest users to purchase similar items while shopping in online retail store and where they already have a significant progress. Our primary goal is to suggest similar or related items which can be placed near to other products without focusing on a single customer. However, those recommendation techniques are going to be helpful by using as a base to start from.

Dataset:

Our study is to analyse the huge volume of dataset provided by one of the famous chain of retail stores. Dataset consists of almost 17M records of sales performed by their 31 stores during 2017-2018. To begin with, we selected single store (Store #J) that have highest rating in Google. There are almost 5,80,000 sales records available for Store #J, among which 4,80,000 records involve purchase of multiple products. Store #J dataset seems to contain sufficient sales record to start the analysis. As of now, we are planning to focus on following features: item number, department code, and class code to determine unique identities along with Date, Transaction Time, Receipt Number, (Net Sale Units or Items Was Scanned), Zip Code for this study.

Exploratory Data Analysis

We tried to explore range of items and classes available in each store and tried to plot them.

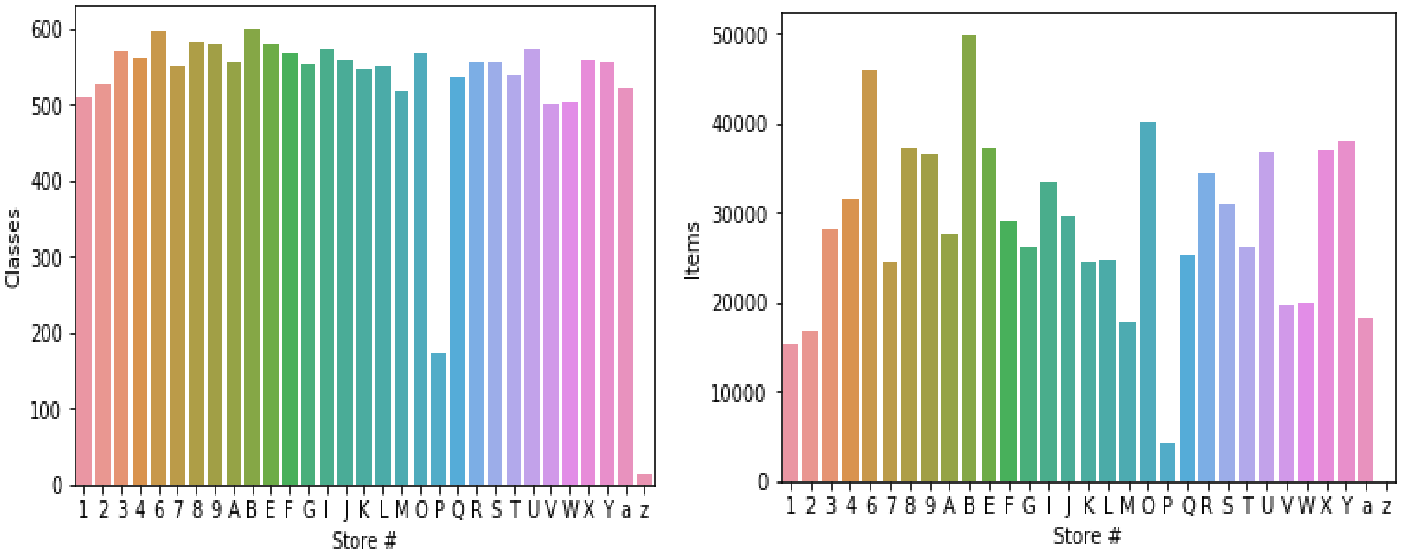
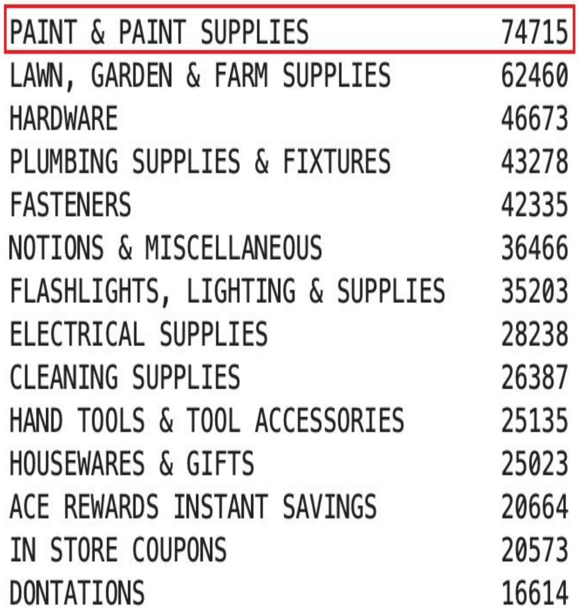


Fig 1 Plots showing range of items and classes sold by each store.

As we see, each of the stores are selling items of almost equal classes whereas types of different Items sold by each stores is not the same. So we planned to start with *Store # J* which stands in the middle of item ranges and it will be a better store to begin with compared to more or less types of items sold by other stores.

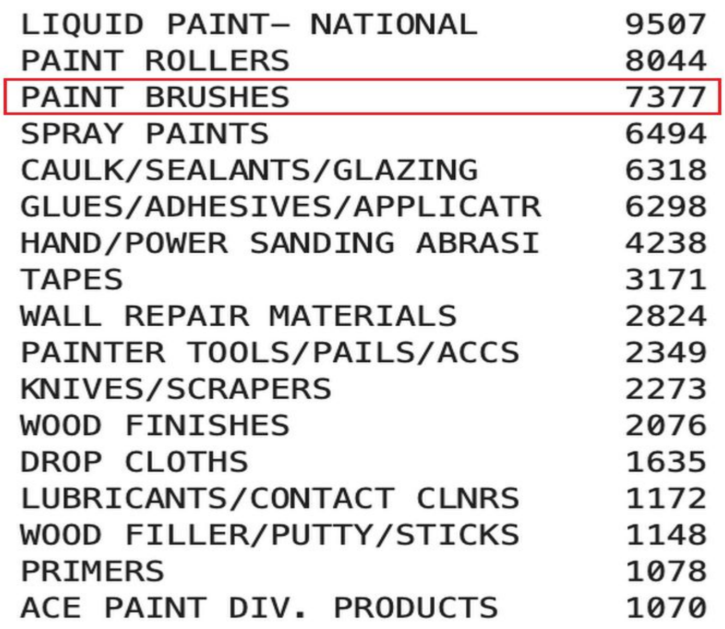
We also analysed the Department and Product Information(*in our context, we refer products to classes in the retail store dataset)* of *Store # J*.

DEPARTMENTS



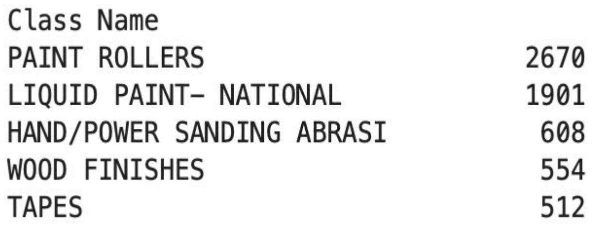
```
PAINT & PAINT SUPPLIES              74715
LAWN, GARDEN & FARM SUPPLIES        62460
HARDWARE                            46673
PLUMBING SUPPLIES & FIXTURES        43278
FASTENERS                           42335
NOTIONS & MISCELLANEOUS             36466
FLASHLIGHTS, LIGHTING & SUPPLIES    35203
ELECTRICAL SUPPLIES                 28238
CLEANING SUPPLIES                   26387
HAND TOOLS & TOOL ACCESSORIES       25135
HOUSEWARES & GIFTS                  25023
ACE REWARDS INSTANT SAVINGS         20664
IN STORE COUPONS                    20573
DONTATIONS                          16614
```

PRODUCTS OF ‘PAINT AND PAINT SUPPLIES’ DEPT



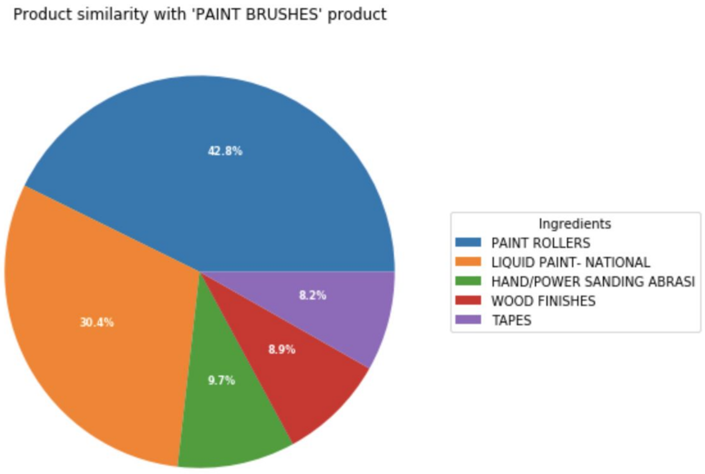
```
LIQUID PAINT- NATIONAL           9507
PAINT ROLLERS                    8044
PAINT BRUSHES                    7377
SPRAY PAINTS                     6494
CAULK/SEALANTS/GLAZING           6318
GLUES/ADHESIVES/APPLICATR        6298
HAND/POWER SANDING ABRASI        4238
TAPES                            3171
WALL REPAIR MATERIALS            2824
PAINTER TOOLS/PAILS/ACCS         2349
KNIVES/SCRAPERS                  2273
WOOD FINISHES                    2076
DROP CLOTHS                      1635
LUBRICANTS/CONTACT CLNRS         1172
WOOD FILLER/PUTTY/STICKS         1148
PRIMERS                          1078
ACE PAINT DIV. PRODUCTS          1070
```

The *‘PAINT & PAINT SUPPLIES’* department of *Store # J* is chosen for our further analysis. In the first step, our goal is to find the products similar to the product *‘PAINT BRUSHES’* of ‘Paint and Paint Supplies’’ department. We have got similar products of ‘*PAINT BRUSHES*’ by finding which products were frequently bought along with it. We have implemented this by analysing the data considering products against each transaction(*Receipt Number*) and getting the products which were billed along with ‘Paint Brushes’. The idea behind this is an assumption that the products bought together by the customers should be placed together in the rack/shelf so that their collective sale increases.



```
Class Name
PAINT ROLLERS                   2670
LIQUID PAINT- NATIONAL          1901
HAND/POWER SANDING ABRASI        608
WOOD FINISHES                    554
TAPES                            512
```

After looking at the dataset and proper querying based on the receipt/billing information, we have found that the products shown in table(*left*) are the similar products of ‘Paint Brushes’. It also shows how frequent they are bought together. The figure in turn gives the percentage of how similar these products are to the product ‘Paint Brushes’ in the*Store # J*



Challenges:

- Different stores might have different scenarios as in sells due to customers from a variety of regions.
- Clustering high dimensional categorical data is a challenge as it lacks
- Natural ordering on the individual domains.
- While analysing dataset, we have come across a lot of errors such as, in few records Item Numbers are entered incorrect to same Items, Class Codes incorrect for same Classes, etc.
- For a related product suggestion of a specific class, concerned store may not be keeping any products of that class for sale.
- We had an opportunity to visit the selected single store located nearby and realised this is not going to be direct train-predict thing but more logical decisions will have to be made in order to get suggestions to this chain of stores specifically.
- One major problem that we may hit will be inability to handle unknown products. List of products in the store may change frequently. If model encounter a new product then it may not know how to interpret or build vector for it.

Approach:

- Starting with the data cleaning, we have tried to correct Item Numbers, Class Codes as described in challenges along with fixing NaN values and datatypes. We can use Description or Names respectively from same Store to identify correct values.
- One of the stores *Store # J* is selected as base to understand the study and incorporate complexities in future.

In the initial EDA stage, we have predicted similar products by building a system based on the Costello_ace Data that finds the products frequently bought together.

In the next stage, we have extended this approach by building a *Word2Vec model*, and training its vocabulary with the product information to predict closely related or similar products to a given product.

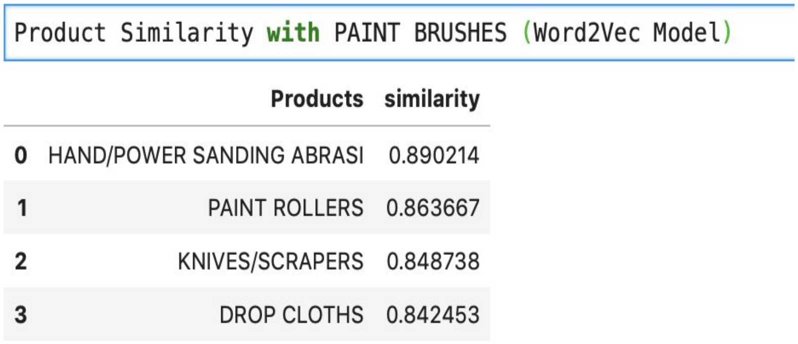
Predicting similar products using Word2Vec Model

It's not as easy to cluster categorical data compared to numerical data. If a model is trained with categorical features after one hot encoding then the model cannot understand what each category is and how each category relates to, for Example, *the word 'Albert Einstein' is equally dissimilar with the word 'Scientist' as that of the word 'Footballer'*. To overcome this issue, we need to build a model which can understand the categorical features and classifies them accordingly. This could be done by Word2Vec model, which creates learning vector representation of words called word embedding.

'You are remembered by the company that you keep' - This is the main idea behind Word2Vec model which constructs a vocabulary of all the words and put similar words together to predict the related words.

In this approach, we have built a word2vec model that accepts sentences of product information like Department Name, Cost, Class Name, Price, Location data and constructs a vocabulary of product information data, to predict similar products from the vocabulary

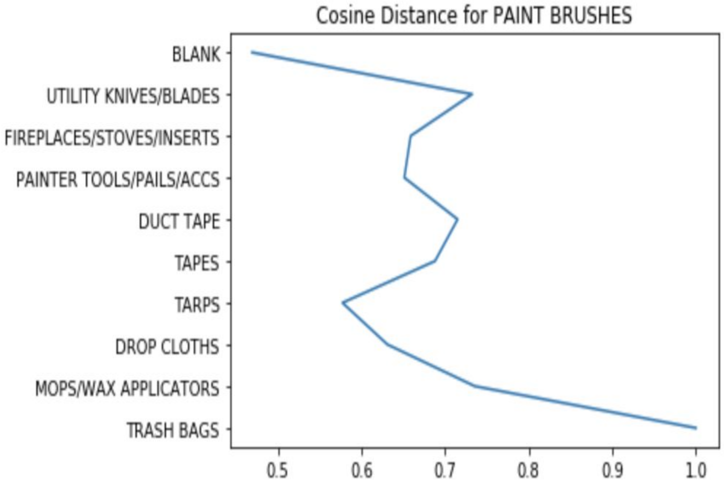
As in the previous approach , we have predicted the similar products for the product ‘Paint Brushes’



```
Product Similarity with PAINT BRUSHES (Word2Vec Model)
```

	Products	similarity
0	HAND/POWER SANDING ABRASI	0.890214
1	PAINT ROLLERS	0.863667
2	KNIVES/SCRAPERS	0.848738
3	DROP CLOTHS	0.842453

The prediction of similar products for the product ‘Paint Brushes’ using Word2Vec model is as shown in fig. By comparing it with the prediction made in the Exploratory Data Analysis section above, we can say that the model has done a reasonably good job in predicting products similar to that ‘Paint Brushes’.



The cosine distances for the product ‘Paint Brushes’ with each of the other 10 products is shown in the figure. *It can be seen that the products with greater distances, belong to a very different department* and those with the lesser distances are of the same department of that of ‘Paint Brushes’.

Based on the prediction mentioned in EDA section and Approach section, we find that some products match in both the predictions. The frequency of their togetherness in Approach 1 and cosine distance in Approach 2, confirms the similarity of these products with that product ‘Paint Brushes’.

Next milestone:

Having predicted similar products of a product within the same department, our next step is to predict similar products across different departments, which could be bought together by the customers.

Example: *The product ’Paint Brushes’ from ‘Paint and Paint Supplies’ dept could be bought together with the product ‘Sponges’ from ‘Cleaning Supplies’ dept.*

Evaluate:

We plan to evaluate the model by validating it on other same chain stores in the same region by going with clustering approach. We also plan to validate it on stores from other regions and decide how to proceed further accordingly. Results can also be compared with available machine learning systems working somewhat similar objectives.

Next Steps:

- Preprocessing and cleaning of the data - fixing anomalies
- Perform data analysis to identify relations among products and classes in a particular store.
- Analyze shifting relations between products by weathers as data move around hardware items.
- Build Machine Learning Model(s)
- Compute matrices showing product to product co-occurrence scores where higher scores mean more likely to be suggestable.

References:

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," in *proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 3111-3119*, Lake Tahoe, Nevada, 2013.

[2] L. Piciu, A. damian, N. Tapus, A. Simion and B.Dumitrescu, “Deep recommender engine based on efficient product embeddings neural pipeline”, in 2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet)