CSE 519 Progress Report - Retail Sales Analysis

1) Objective

The project aims to analyse the sales data of a chain of retail stores to create meaningful insight into buying patterns and provide recommendations to customers.

2) Background

Online stores usually profit from dynamic recommendation systems to encourage customers to buy more. Brick and mortar stores can try to mirror this by placing frequently bought together products closer in the store. The store also needs to manage its inventory smartly, making changes according to the weather or time of the year.

The combined aim of our tasks is to predict the behaviour of the customer and predicting what they might want to buy next. To understand this, we tried to look at some existing research on buying trends of customers. The store needs to consider these properties to maximise profits which can be achieved by offering subscription models or providing email recommendations to customers.

3) Dataset

The dataset initially provided had only 2 years of data, 2017-18. Another additional data set for the years 2015-16 was later provided. Most of our progress report uses only the 2017-18 data unless explicitly mentioned otherwise.

The dataset had a lot of scope for pre processing, there were some erroneous values as well. Eg. after every month's data the header row is repeated. So there are 23 header rows in the data which need to be removed. Customer Number for 56 entries is missing but it has the same Loyalty ID and thus it is safe to assume this to be the same customer across all rows. We have imputed the customer number as 404 for these entries.

Other null value handling included either dropping rows when in extremely small amount and with no meaning. We have dropped 50 rows with missing Department Code and 4 rows with missing Class Name. There exist certain null values in class and fineline codes which are left null but their corresponding name columns mention them as BLANK and thus we have considered them as BLANK. Similarly customer zip codes had values like 'NY', '-' and so on, such records were limited in relation to the overall dataset and hence were dropped for the purpose of location-based analysis.

Dataset details for 2017-18

There were a little above 17 million rows in the data. For the embedding task we have dropped rows which do not have a unique Customer Number, hence we have considered rows only with non null Loyalty ID. (This comes out to be a little over 12 million rows)

Embedding dataset

We had to further put some restrictions when training the item embedding. The restrictions are of the kind: a customer is considered only if he has shopped a certain amount of times. This will be further explained in the approach section.

4) Proposed work

In the project proposal we proposed to pursue 2 categories of tasks namely -

A) Recommendation System
B) Time Series Analysis

Overview of work done so far:

A) Recommendation System
   a) Built Matrix Factorization models for Customer and Item embeddings
   b) Customer Embedding used to find similar customers from their Loyalty ID.
   c) Similar customers can be plotted on a map of New York.
   d) Find similar/frequently-bought-together items for any Item numbers.

B) Time Series Analysis
   a) Inventory management: Predict demand for future year, looking at the available data of an item.
   b) Selling prediction: Predict gross margins for future year, looking at the available data of an item.

C) Location Analysis
   a) Customers plot on New York map
   b) Stores plot on New York map with profitability color coding
   c) Customer distribution and store plot on map
   d) Similar customers plot on map, found using Customer Embeddings

5) Approach

A) Recommendation Systems

To create a similar Product recommendation system, we first need to generate Product embeddings. For creating the customer-product interaction matrix, we first filtered out on Products and Customers occurrences. We removed the Return transactions otherwise it would lead to double counting of certain products. Further we selected only those Products and Customers which occurred at least 120 and 60 times respectively. This gives us total 7443 unique items and 8392 users.

We then created the interaction matrix *S*, where an element *S[i][j]* = number of times customer *'i'* purchased product *'j'*. We did a train:test split for this interaction matrix in 80:20 ratio.
Mathematically, we have:

1. a set C of 'customers';
2. a set P of 'products';
3. an interaction matrix S of size |*C*| x |*P*|.

We have to find two matrices of *K* 'latent features', *W (|C|* x *K matrix)* and *I (|P|* x *K matrix)* such that their dot-product approximates *S*:

$$S \approx C \times I^T = \widehat{S}$$

$$\widehat{s}_{ab} = i_b^T \cdot c_a$$

Where i_b and c_a are the vectors representing 'a' item embeddings and 'b' customer embeddings in K dimension. We experimented with two loss functions, Bayesian Personalized Ranking(BPR), Weighted Approximate-Rank Pairwise(WARP), in which WARP clearly outperformed BPR. We applied Adagrad to optimize the loss function, since it handles sparse data well.

The product embeddings we get are then utilized to query similar products, by using the cosine similarity measure. The customer embeddings we get here is utilized in the Location Analysis (Part-C).

B) Time Series Analysis

The aim of this analysis is to forecast data depending on current trends. We currently have 2 years of data available which in itself is insufficient to get a precise prediction of the trend or seasonality but we have tried our best to forecast the demand and profits.

There are several methods used to achieve this such as ARIMA which stands for Autoregressive Integrated Moving Average. We have used a method that is mentioned in the paper Forecasting at scale by Taylor SJ, Letham B. 2017. They have proposed a modular regression model with interpretable parameters that can be intuitively adjusted by analysts with domain knowledge about the time series.

The reason for using this method is that the data can be easily verified by people at Costello's Ace by looking at it as they have several people with domain knowledge.

C) Location Analysis

We are primarily using geographical plots to help visualize our results from exploratory analysis and the customer similarity obtained from customer embeddings. In addition, the plots can also provide insights into geographical customer distributions to make decisions like new store openings or expansions or location-wise promotions, etc.

6) Results

A) Recommendation Systems

We demonstrate the effectiveness of our item embeddings by showing how they can be used to search for 'similar' items. We have retrieved top five items by calculating cosine similarity between the items. In Figure 8 we query six items amongst the most popular items, for example if we query "Powerade Fruit Punch" item, we get "Red Bull Energy Drink" , "Vanilla Coke" , "Ice Tea" as the most similar items. We have identified that in this dataset cosine scores above 0.95 produce more meaningful results. These results depicts that the model has been successful in learning the underlying buying patterns of the customers.

Query: POWERADE FRUIT PUNCH

	Recommended Products	Score
1	RED BULL ENRGY DRNK8.4OZ	0.963856
2	VANILLA COKE 20OZ	0.961257
3	GOLD PEAK SWEET ICE TEA	0.95884
4	FUZE ICE TEA	0.951142
5	RED BULL ENRGY DRNK 12OZ	0.950947

Query: CHERRY COKE 20OZ

	Recommended Products	Score
1	SPRITE 20OZ	0.971938
2	CANDY SNICKERS 3.29 OZ	0.965966
3	FUZE ICE TEA	0.955276
4	MINUTE MAID LEMONADE	0.952316
5	SEAGRAMS GINGER ALE	0.949839

Query: ACE TOP SOIL 40#

	Recommended Products	Score
1	TOP SOIL 40#	0.96367
2	COSTELLOS HIGH TRAFFIC MIX 3LB	0.952285
3	EZ SEED SUN&SHADE 10#	0.937675
4	LED ACE A19 60W EQ DAY	0.93296
5	MULCH LI COLOR ENH BLK MULCH	0.92528

Query: PAINTBRS CHIP 1.5WT BRSL

	Recommended Products	Score
1	PAINTBRSH CHIP2"WHT BRSL	0.958074
2	PRO GROUT SAW	0.952567
3	MSKG TAPE SHARP 1.41x60	0.937321
4	PAINTBRSH CHIP 3"WT BRSL	0.919744
5	4" FABRIC ROLLER W/ HANDLE	0.916718

Query: MOP FLOOR HARDWOOD BONA

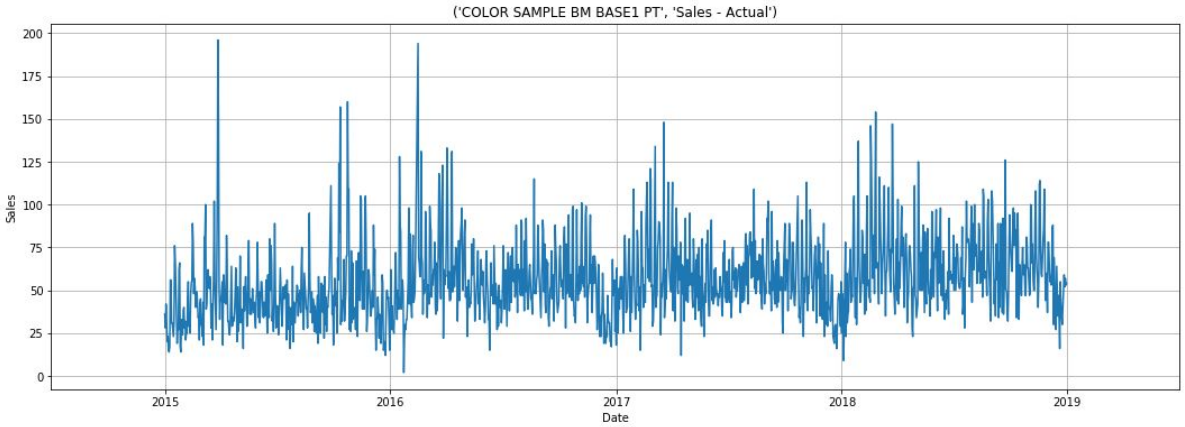
	Recommended Products	Score
1	TILEX MOLD/MILDEW 32OZ	0.974818
2	PLEDGE MLT-SURF AERO	0.960355
3	BUCKET BISQUE 15QT	0.959346
4	SOFTSCRUB LMN CLNSR 26OZ	0.946663
5	O-CEL-O HANDY 4 PACK	0.939224

Query: KEY KWIKSET KW1-ACE250PK

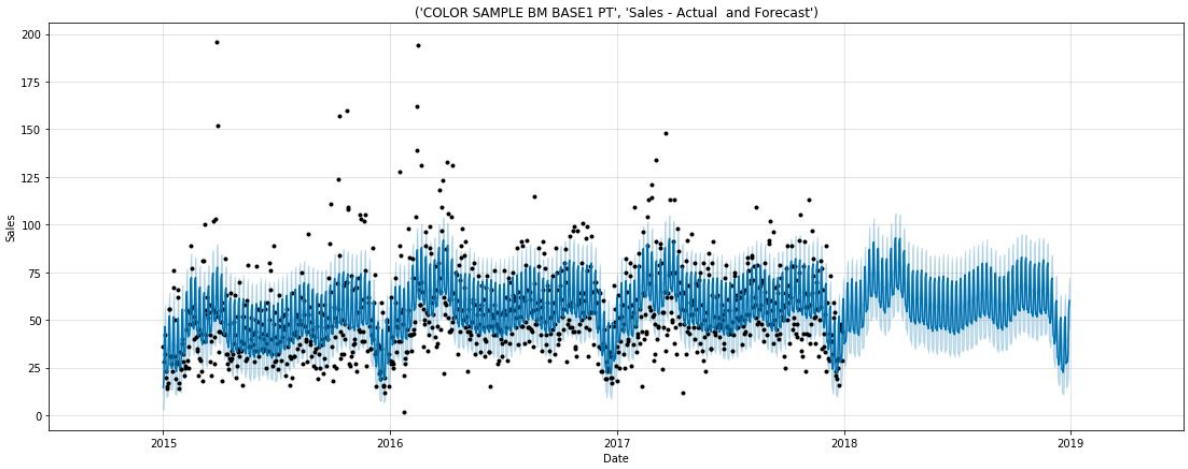
	Recommended Products	Score
1	KEY SCHLAGE SC1-ACE250PK	0.963829
2	KEY SCHLAGE SC1250PK	0.911381
3	MASKING TAPE1.88X60YD GP	0.904061
4	ECON HACKSAW 10" ACE	0.893545
5	GLOVES LATEX LG 2PAIR	0.879998

B) Time Series Analysis

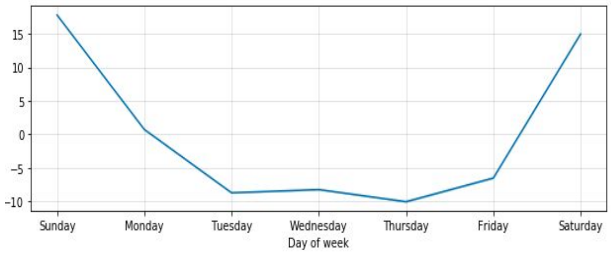
We have data of the years 2015-2018 available with us. Below is a plot for the total sales of the item: COLOR SAMPLE BM BASE1 PT over the 4 year period.

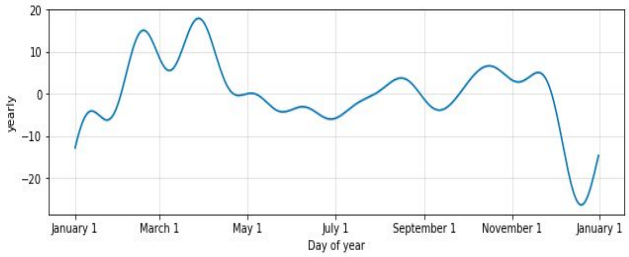


Now we have used the data of the years 2015-2017 and predicted for the sales for the year 2018.
As we can see the forecast lines up well with the actual sales seen in above figure.

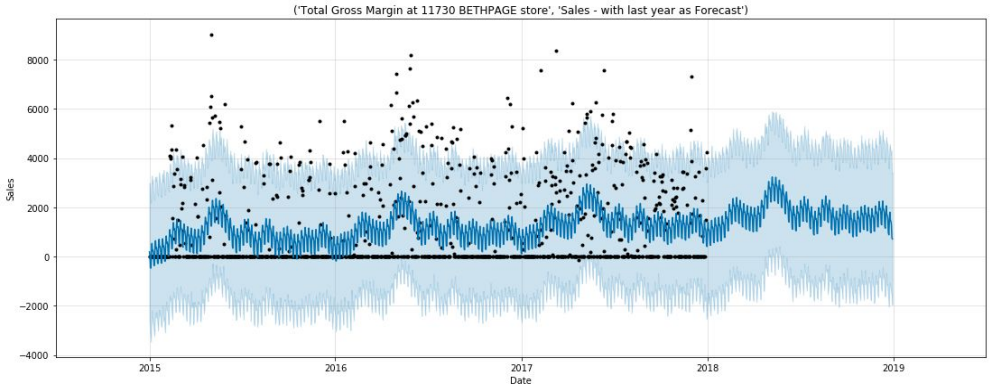


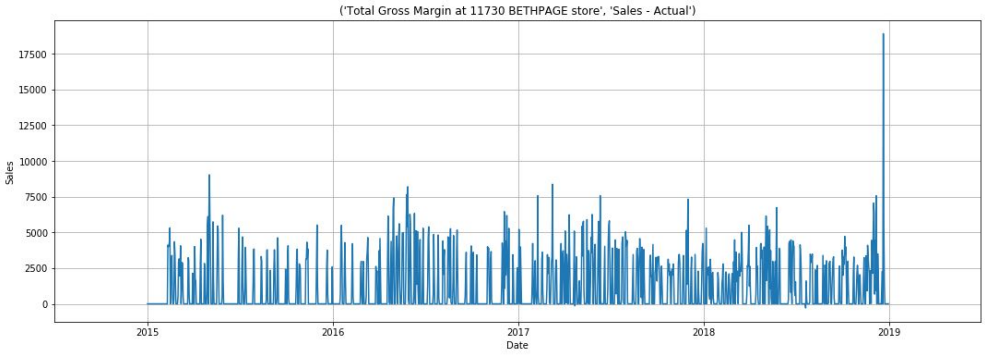
There are certain trends that can be observed as well. Below we can see that the trends for total time period, weekly sales and yearly sales. The y-axis represents the trend and not the actual values of amount of sales. We are able to retrieve the forecasted sales for any, day, month or entire year. The sales data analysed in all these graphs is the total sales of the item across all stores. We can filter out data for each store and give specific predictions for specific stores at a point in time.





We can also forecast the gross margins of a store over a time period. This can be mainly used to see if a store is underperforming than the expected margins, this should not be treated as a prediction of the gross margin.



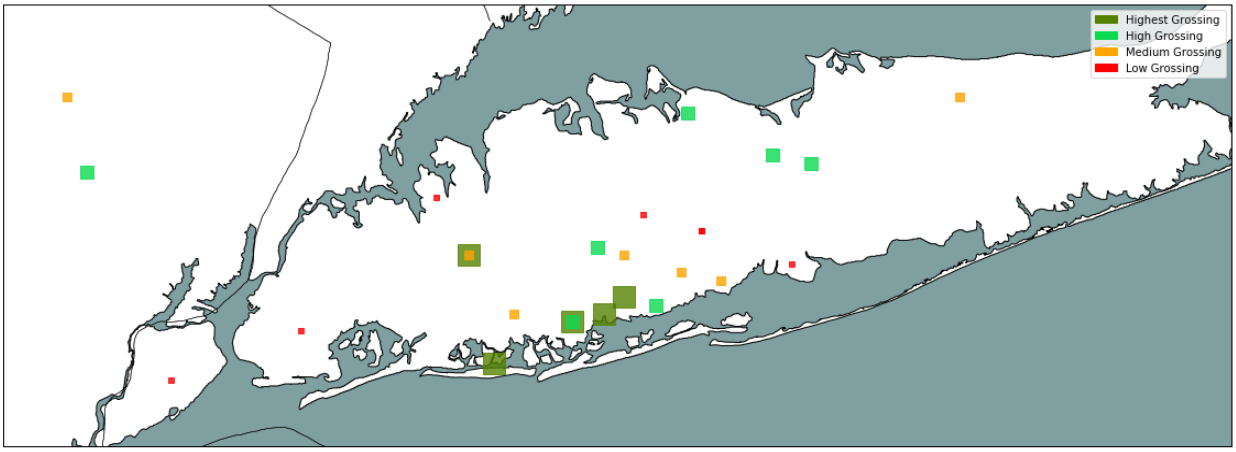


C) Location Analysis

I. Customer base and Store Distribution geographically

The graph below shows the density distribution of Costello's Ace customers in the New York region. The size of the bubbles represent the density of customers at each zip code and the minimum number of customers per zip code to be displayed on the map is chosen to be 50.

The graph below plots the stores based on their gross margins earned. We can see a close correlation with the distribution of customers shown above. The color and size of store markers help identify the store classification by gross margin. In addition, some zip codes have more than one stores hence the overlapping store locations can be seen.



II. Customer base distribution for stores differing in gross margin

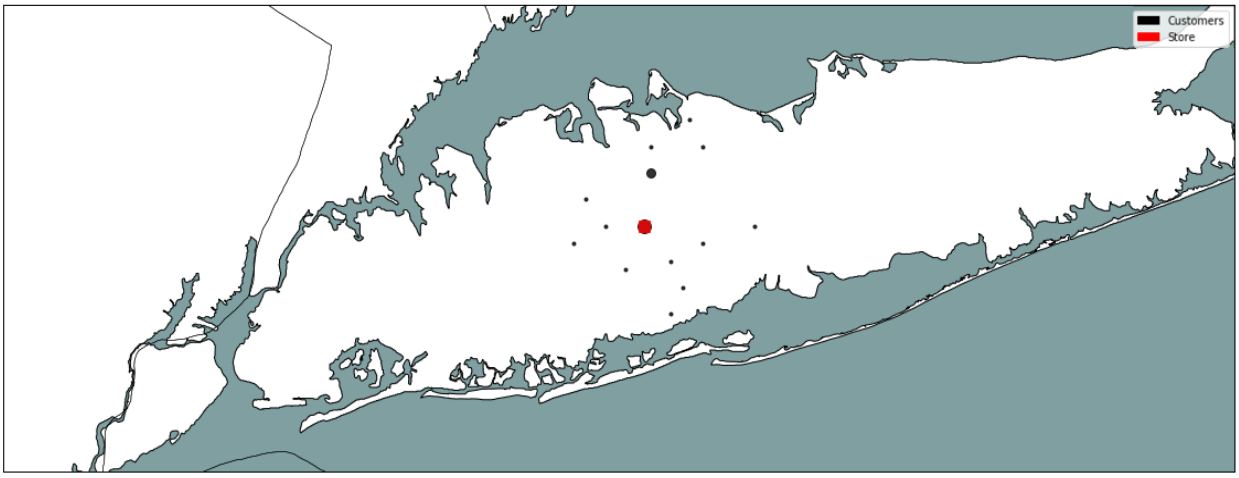
Following graphs show the distribution of customers for a representative store from each category by gross margin earned.
Inferences

1. We can see that a majority of customers reside in the immediate neighborhood;
2. For the higher grossing stores, a significant fraction travels from further distance whereas for low grossing stores, the customer base is much closely clustered;
3. These inferences can be used to decide the advertising radius based on location;

ISLAND PARK (Highest grossing store example)

NESCONSET (High grossing store example)

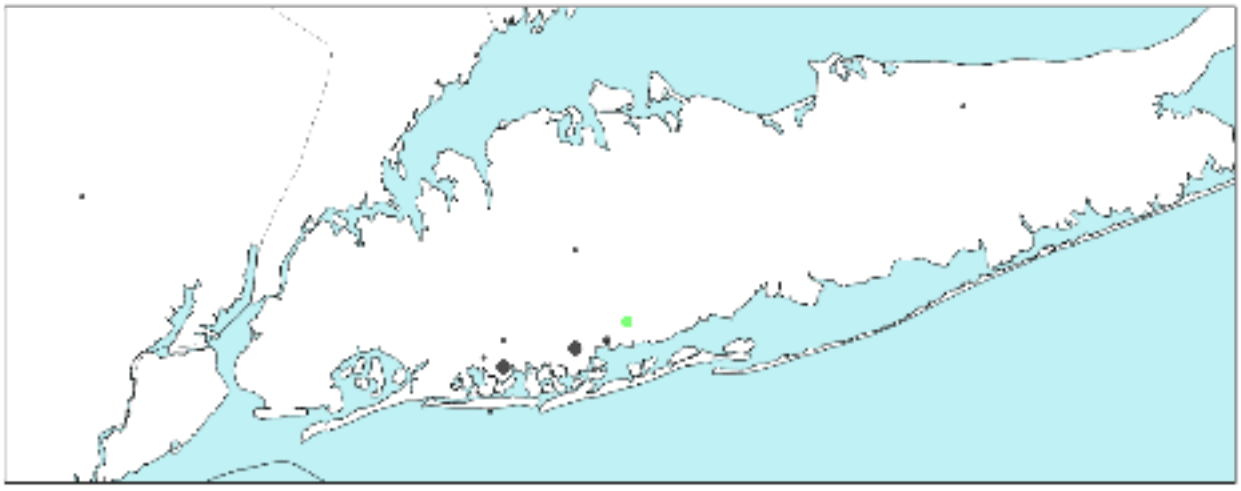
LINCOLN PARK (Medium grossing store example)

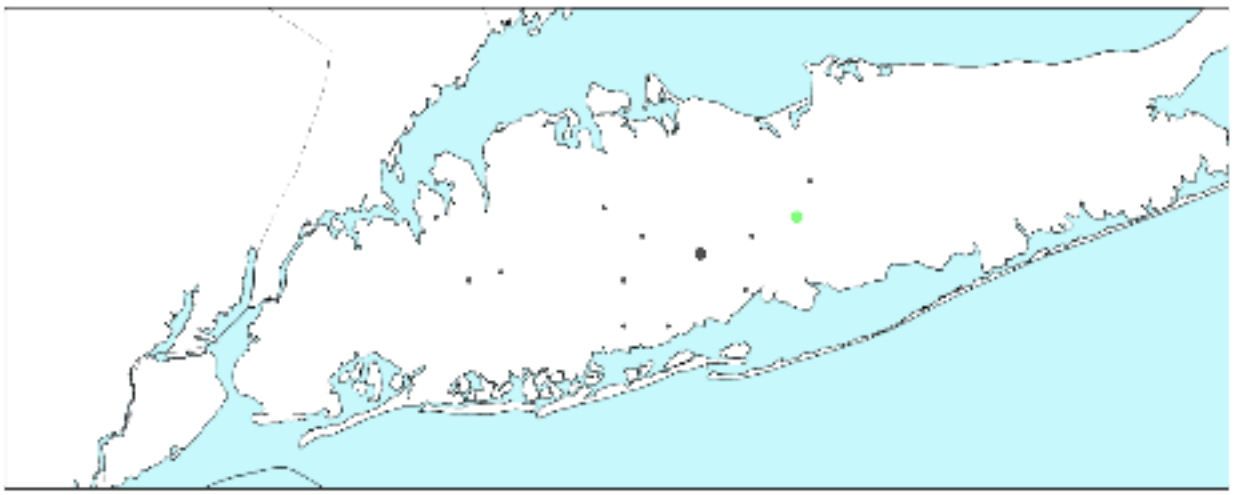


MELVILLE (Low grossing store example)

III. Customer Similarity based on Customer Embeddings

The below 2 plots are a demonstration of our results from customer embeddings. Each plot shows a chosen customer (green circle) and the geographical distribution of customers similar to it (black circles, varying in size proportional to density)





7) Future Work

A) Recommendation System

a) Current model is a collaborative filtering based one, we will incorporate content-based model, item features like 'Cost', 'Department Came', 'Class Name', and Fineline Name' to create a hybrid recommendation model;
b) Current model only captures similar items, in the next iteration we will be creating model for deriving complementary items;
c) Further we will try to apply Tri-gram based Triple2vec algorithm for further generating customer and item embeddings and compare the results;

B) Time Series Analysis

a) The current model used is a basic linear predictor that predicts the trend in sales. It does not take into factor any other features that are available;
b) We can use additional features such as weather data, promotions and train more complex models (LSTMs and neural nets);
c) However, it is unclear if this will provide significant gains over the baseline which is doing a good job at predicting the seasonal trends of products.

C) Location Analysis

a) We will explore ways to publish interactive plots with filtering to provide run-time insightful visualizations on the data
b) Recommend locations for new stores or expansion based on geographical customer distribution

8) References

- Jose, Cijo, and François Fleuret. "Scalable metric learning via weighted approximate rank component analysis." *European conference on computer vision*. Springer, Cham, 2016.
- Rendle, Steffen, et al. "BPR: Bayesian personalized ranking from implicit feedback." *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009.
- Fionda, Valeria, and Giuseppe Pirró. "Triple2Vec: Learning Triple Embeddings from Knowledge Graphs." *arXiv preprint arXiv:1905.11691* (2019).
- Taylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2