Product Recommendation Package Based on Customer Transaction History and Product Description

Application of basket analysis and Natural Language Process (NLP) techniques to identify sales growth opportunities with real sales data provided by Triad Component Group (TCG)

Team 29

Phan, Richie C rphan6@gatech.edu
Kim, Joonsoo jkim3497@gatech.edu
Morrill, Chris cmorrill6@gatech.edu
Azzu, Shawn sazzu3@gatech.edu
Hutchins, Rebecca J rhutchins8@gatech.edu

Abstract

This research project utilized actual data furnished by a current employer of one of our team members, namely Triad Component Group (TCG). TCG is a small business based in San Diego, California, and has been around for over 25 years. The company's focus is on manufacturing and distributing electrical components with its three sub companies: OptiFuse, Switch Components, and Vortex. TCG specializes in the sale of fuses, switches, circuit breakers, and their respective accessories. Customers range from Tesla and GM to local and small "mom and pop" locations. These customers rely on TCGs products as a key part of their supply chain to continue producing. In the past year OptiFuse and Switch Components have celebrated the launch of their new websites.

With the launch of the new websites, TCG is interested in improving the customer experience through updates to the product recommendation and search systems. It is believed that improved customer experience will translate into increased sales revenue and profits for the companies.

To accomplish these business goals, several models were built to improve the product recommendation and search features deployed on the TCG websites and other customer communication channels. The models fall into two broad categories: basket analysis and natural language processing approaches. The completion of this project could see direct implementation into the TCG websites, and its utilization can help drive sales for the companies.

Introduction

At present, the TCG sites have an item recommendation system which is based solely on individual expert understanding and intuition. This means that the recommendation system is currently a manual process by which an individual familiar with the company products and the industry manually selects which products should be shown as recommendations. The company would like to implement a more robust, data driven, and empirical approach to item recommendations to display for customers visiting their website. This sort of recommendation system can be a win-win for both TCG and its customers as it can both increase sales and make it easier for customers to find things they need and make purchases from an individual supplier. Our group plans to utilize sales data from all three sub-companies as well as item-specific details to identify an algorithmic approach to item recommendation. An improved item recommendation system can benefit TCG by optimizing cross-selling opportunities and improving the customer website interaction experience.

To improve the customer experience and drive an increase in sales, this analysis seeks to answer the following questions:

1) Utilizing sales and item data, is there an effective way to recommend items to customers based on the current items they are viewing or past purchase history with the goal of driving sales?
2) Can a Natural Language Processing (NLP) model(s) be developed, built using details and descriptions of orders and respective items to generate product recommendations, that has potential to improve sales and product search on Triad Component Group's website?

Data Sources

We obtained permission from TCG to use the data for research purposes. We also assured that any Personally Identifiable Information (PII) was appropriately eliminated and encoded. The customer identity was encoded with a unique identifier called Customer Number to protect the privacy of the customer while still allowing analysis of the data.

Figure 1 shows the Entity Relationship Diagram (ERD), a visual representation of the system scope and the inter-relationships amongst entities. For ease of understanding, the ERD does now show all attributes of each table

but instead prioritizes the attributes important to understanding relationships within the SQL Database. The green tables are the raw data. The blue tables are culminated, also available on SQL (created by Axiom), but a conjunction of multiple tables.

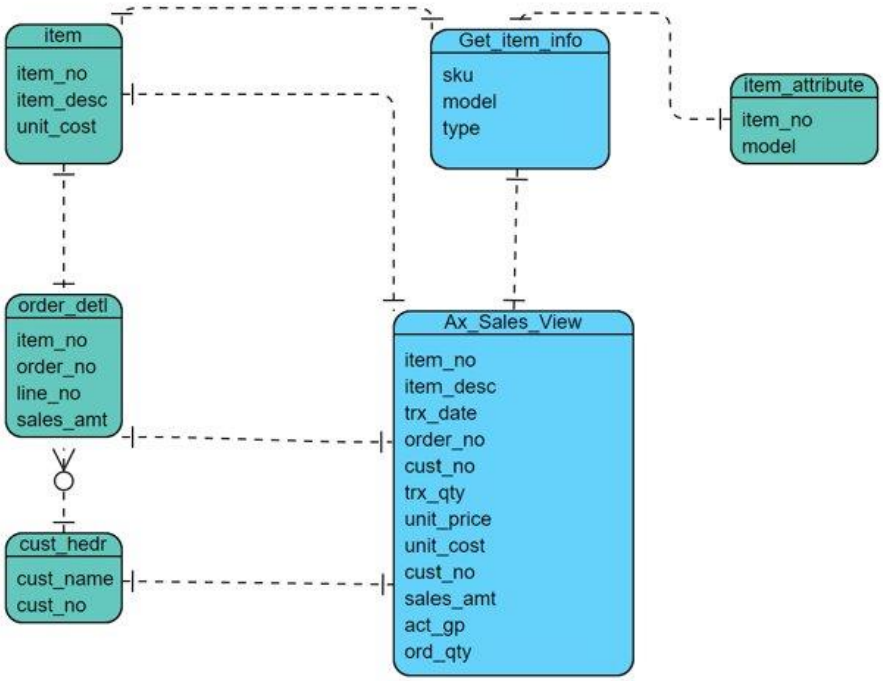


Figure 1. TCG Database Entity Relationship Diagram

The first data set (Ax_Sales_View) is a table which contains Sales Information for the company's entire usage of its CRM (customer relations manager) system EPDS. As of 6 Mar 2023 there are 211 columns and 231,644 observations ranging in time from 2005 to present, composed of 76,470 unique order records from 1,361 customers. Accompanying this data is Get_item_info, which contains make and model specific data that comes directly from the company's two websites, Switch Components and OptiFuse.

Exploratory Data Analysis

The two main tables utilized were the Get_item_info and Ax_Sales_View tables. The Get_item_info table contains item specific information that makes each product identifiable and helps to categorize each product. The main categories for Triad's Optifuse division include Fuses, Fuse Accessories, and Circuit Breakers. Within each category, attribute information, such as, Type, Automotive (Y/N), Housing/Style, Marine (Y/N), Reset Type, Equipment Type, and more is stored; this information is used mostly for filtering processes on the website. However, we can utilize this information to find/identify links in our products and apply this information to our models. The Get_item_info contains the following amount of information for each main category:

- Total Fuses: 3,287; Number of Distinct Fuse Models: 150
- Total Circuit Breakers: 627; Number of Distinct Circuit Breaker Models: 317
- Total Fuse Accessories: 428; Number of Distinct FA Models: 177

The Ax_Sales_View table contains sale specific information for orders for the company that are made and stored directly to our CRM (Customer Relationship Manager) system EPDS, as opposed to being made on E-commerce

platforms. Ax_Sales_View has the following characteristics:
Number of Observations: 231,644

- Number of Attributes/Features: 211
- Number of Unique Customers: 1,360
- Number of Unique Items Sold: 6,351
- Number of Orders: 76,470
- Average line numbers per order: 2.65
- Initial Date: 2005-06-22
- Final Date (up to): 2023-03-03

For this data set Customer and Salesperson information was either removed or anonymized to maintain the integrity of "confidential" or "private" information. Some key and interesting information can be gleaned from this data set and visualized:

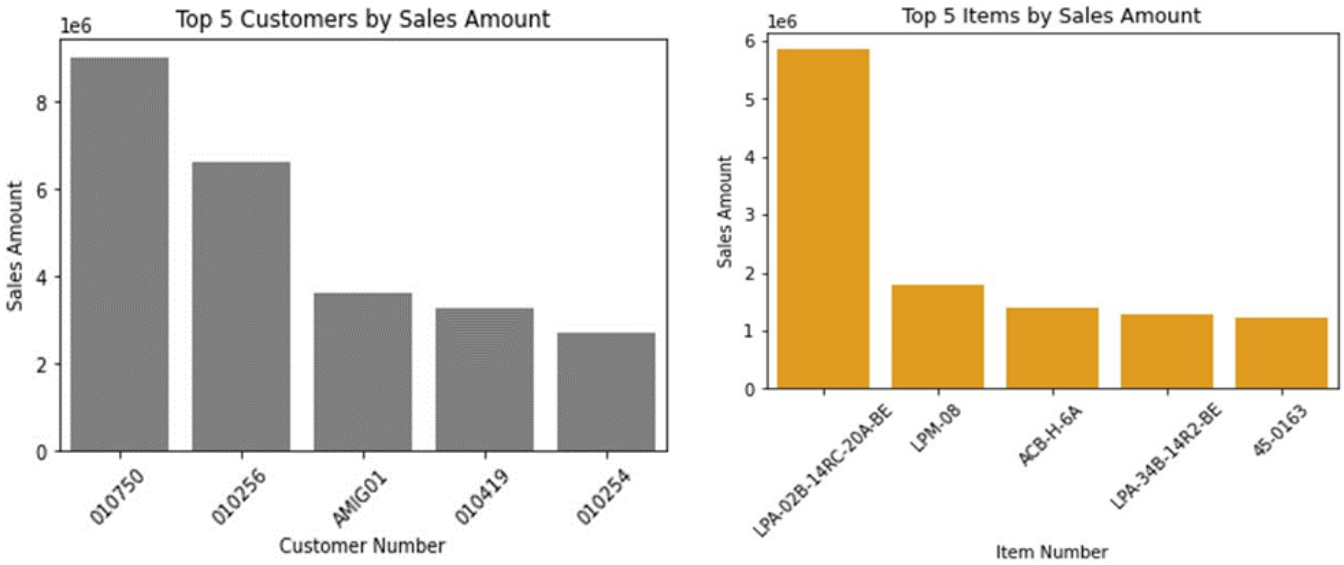


Figure 2. TCA: Top 5 Customers (Numbers) by Sales Amount

This figure showcases the top 5 customers by sales amount, here it can be identified that the customer 010750 is one of the top customers and is followed by 010256, there is then quite a gap from the next 3. Furthermore, the top items sold overall can be examined. This figure shows the top 5 items sold by sales amount; here the item number LPA-02B-14RC-20A-BE is the most sold by far and it is followed by LMP-08 which is closely followed by ACB-H-6A. It is important to understand the top customers and top items sold as they can have a large influence on what we see in the model outputs. Visualizing them and interoperating these results can help in understanding customer behavior, identifying popular items, as well as cross and up selling opportunities. This also brings to light the ideas and concepts of weighing profitability of an item and how influential the sale of an item may be - especially in relation to its supplementary and complementary items.

Proposed Methodology Overview

We utilized batch/basket analysis, K-means clustering and natural language processing (NLP) techniques to build models that can inform product recommendation decisions on the TCG websites and in other customer communication channels such as marketing emails. Note that our team's approach was to integrate three analytical models aligned with business objectives of maximizing cross-sell opportunities and product searching

capabilities. Also having a pair of analytics processes with different methods in parallel not only diversifies analytics approaches but also creates fallback mechanism.

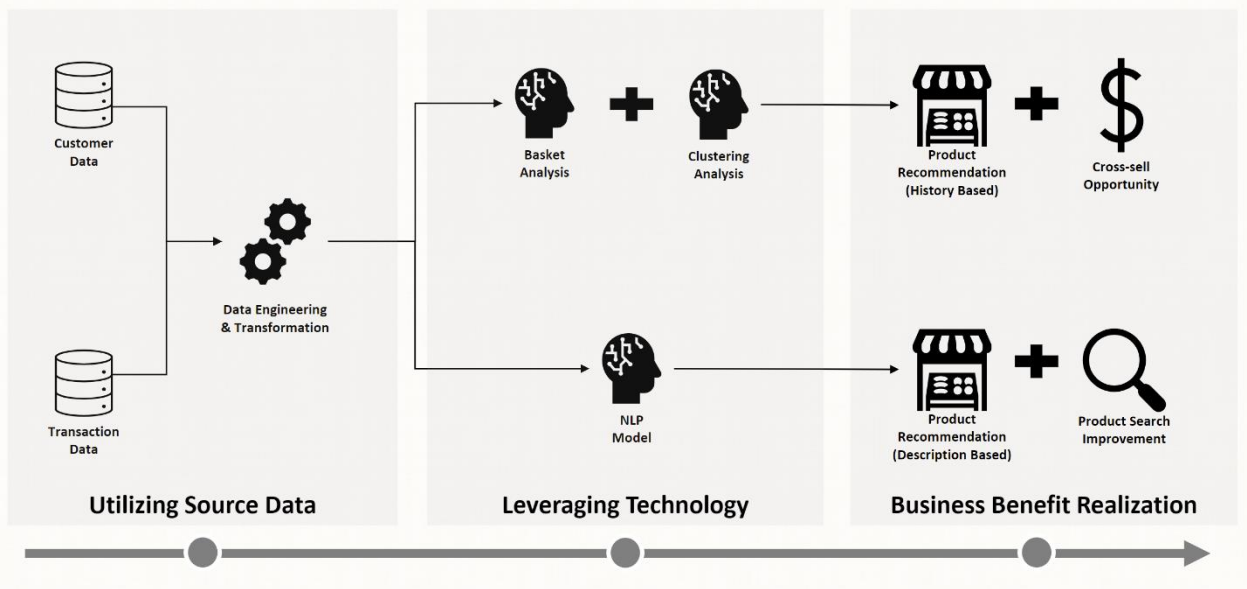


Figure 3. TCA marketing analytics tool workstream summary

Proposed Methodology I – Basket Analysis + Clustering

Basket analysis is used to identify items that are frequently purchased together. Although some product combinations may be obvious (e.g., peanut butter and jelly), there is the potential for unexpected product combinations that are less obvious. Basket analysis can be used to confirm expected product correlations and discover hidden correlations by applying a set of statistical rules. The most common statistical rule for basket analysis is association rule mining which is a process to identify relationships and interdependencies between items purchased by customers.

In the case of TCG, there are at least two levels of association rule mining that could occur. One level is at the purchase order level, where correlations can be discovered between items purchased at the same time (in the same order, but different line number). The second is at the level of the customer, where we can look at items purchased over the lifetime of a customer. This will provide additional information to inform a recommendation system for TCG. Enhanced analysis could include adding weights to the items with larger profit margins to further refine the recommendation system for TCG with a goal of increasing both revenues and profits.

Similar to other statistical techniques, basket analysis alone does not assert causality – but it may reveal useful interactions and tendencies that can be exploited.

This analysis is performed by first creating a sparse, binary matrix of all items purchased for each order. To build association rules, we consider three primary metrics:

1) Support – does a pair of products appear together enough to merit consideration? This is the percentage of all orders where both Product A and Product B appear.
2) Confidence – codependence or conditional probability; out of the orders in which Product A appears, how often do we see Product B?
3) Lift – strength of association compared to random chance given each product's support. Values of lift > 1 indicates higher probability of purchasing items together.

The result of this analysis is a list of associated sets – for each identified pair of products, we can report the 'directional' metrics; when Product A is purchased, what is the confidence that Product B will also be purchased? We can then combine those with the lift of that pair to understand the overall strength of that particular association.

Although we were able to create association rules for the main products, the basket analysis has a limitation in that it doesn't propose a way to categorize similar product pairs based on their metrics of association or potential profitability. The model output itself cannot recommend any actions that could generate profits for the company, like the combination of peanut butter and jelly. To address this issue, we integrated a K-means clustering model that uses the output metrics of the association model (e.g., support and lift) and additional product characteristics, such as the product category. Finally, we mapped the individual segments identified by the clustering model with the profitability data of individual products, resulting in a finalized model that quantifies the fiscal benefits of basket analysis for the company.

Proposed Methodology II – NLP Content-Based Recommender

Natural language processing relies on rule-based modeling of the human language, statistics, and machine learning to enable a computer to understand text like the way a human would understand the nuances of language. NLP can be useful for a company like TCG to improve the returned search results when a potential customer searches for products through a query on the website and to improve the recommendations provided in marketing emails by considering the customer purchase history.

The employed NLP methods are all Content-Based Recommender systems using unsupervised learning. We used content-based filtering to recommend items to people based on the attributes of the items and customers using the packages TFIDF, CounterVectorizer, Spacy, and KNN. Each package uses slightly different methods for the language processing but follows a general model to draw relationships between two documents. The NLP process used for the recommender involves extracting numerical features from the two text documents and then using a similarity function to rate the similarity between the two documents. Then the top N comparisons are taken based on these scores to create the output of the recommenders. The Search recommender compares a "search" document and the item description dataset(df_item) while the Customer History recommender compares a customer dataset (df_cust) with text data on all items purchased by a customer and compares that to the item dataset. There are four recommenders built for both the search recommender and customer history recommender (total of 8) as shown in the diagram below.

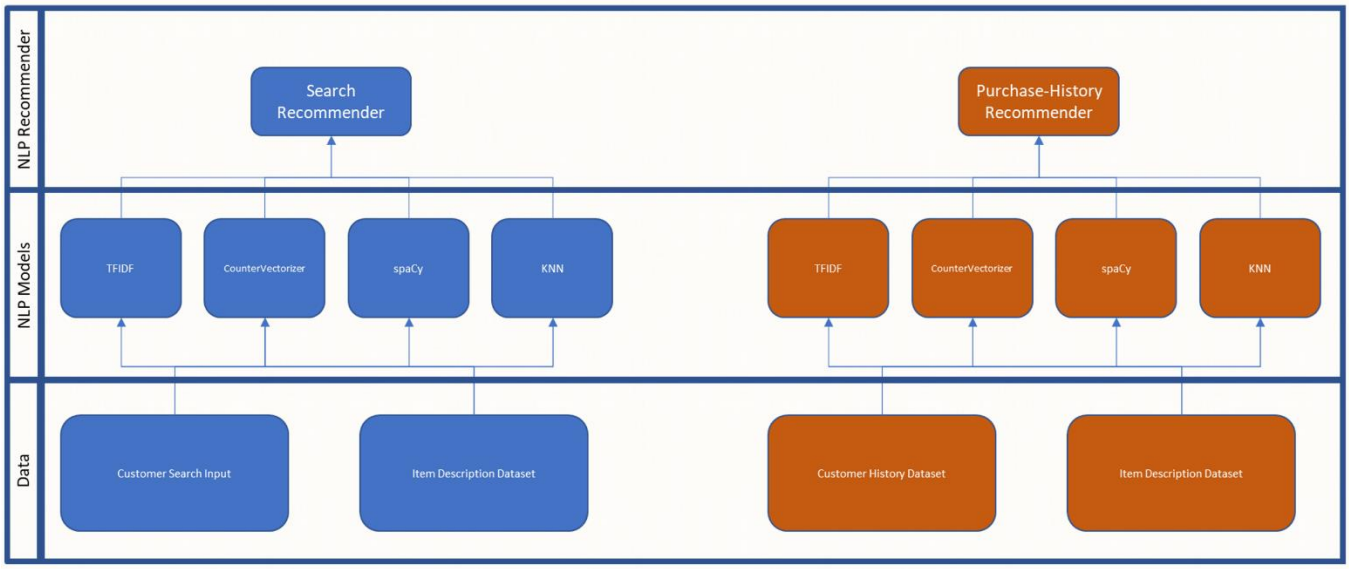


Figure 4. NLP Content-Based Recommender structure overview diagram

To create the two datasets used in the NLP model, the item dataset and customer dataset, specific text features were taken to form a single Text column and ID Column. The item dataset was constructed by combining the Getitem_info dataset and the Switch Components Desc dataset. Contents of text columns such as description, Category, Sub-category, Fuse type, Accessory type, Keywords, and Fuse Housing/Style were concatenated to form one column called text and another column that consisted of PN (Part Number) to use as a key. For the customer history recommender, we made a dataset for customers called df_cust. This was made from the ASV 2 less dataset and was grouped by customer ID. Subsequently, a combination of all item descriptions and product category created a combined text column.

TFIDF Method

TFIDF stands for Term Frequency Inverse Document Frequency, which is the method used to vectorize the two text documents before running them through similarity function. Term frequency is the number of times a word appears in a document. It is divided by the total number of words in that document. This will give the document its own term frequency. Below is the equation.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad idf(w) = log(\frac{N}{df_t})$$

Figure 5. TF function formula (left) and IDF function formula (right)

Inverse Document Frequency is the log number of documents divided by the number of documents that contains the word (w). It determines the weight of rare words across all documents. The equation is represented in Figure 5. The TF is multiplied by the IDF to create vectors for the documents. These two vectors are then put into a cosine similarity function to create a score out of 1, with a score closer to 1 representing high similarity. Cosine similarity is a common NLP metric used to show the similarity between documents. It measures the cosine angle between two vectors in a multi-dimensional space.

CounterVectorizer Method

CounterVectorizer uses a Bag-of-words approach to vectorize two documents. A list is created from the two documents that turn each word into an element connected to a count of occurrences. This is used to extract features and create a vector out of the documents. Cosine similarity is then used on the vectors to create a similarity function, where the score is out of 1, with more similarity closer to 1.

SpaCy Method

SpaCy has its own unique way of document vectorization. First through tokenization, the package breaks down the input text to individual words or tokens. Part-of-speech (POS) tagging SpaCy assigns each token a part-of-speech tag (e.g. noun, verb, adjective) based on its context within the sentence. Dependency parsing determines the syntactic relationships between the tokens, creating a dependency tree that represents the grammatical structure of the sentence. Named entity recognition (NER) identifies and labels named entities in the text, such as people, organizations, and locations. Lemmatization, the process of grouping together different inflected forms of the same word, reduces each token to its base form (e.g. "am", "are", "is" all become "be"). Stop words removal removes frequently used words that do not carry much meaning (e.g. "the", "a", "an"). Then finally it will Vectorize the text content of the documents. SpaCy's similarity function is constructed by sentence embedding (averaging of word embedding), then using cosine similarity to derive a score out of 1, like the previous methods.

KNN Method

This method does not have its own vectorization method. The vectorization method was taken from TFIDF and KNN clustering was used to get the 5 nearest neighbors. The distance is calculated and indexed from the query taken as input to find the 5 nearest neighbors. Scores closer to zero means higher similarity in this case.

Incorporating Profitability

The company noted that they wanted to incorporate profitability into our NLP model. It was considered to add weights to give better scores to more profitable items. The issue with just adding weights to all the scores was that it risked recommending items the customer was not interested in simply because the item had a larger profit margin. This would frustrate customers in the search process and would likely lead to lower sales as customers left the website due to the difficulty in finding what they want. We decided to focus more on returning relevant results and inferred that if we give the people what they want, profitability will follow and have an indirect correlation. To add some form of profitability incorporation, we took the top 5 results, and reordered the results at the point of showing them to the customer. Specifically, the NLP method would provide the top 5 results based on the analysis described above. Before showing those results to the customer, the results are reordered based on the profit margin.

Analysis and Results I – Basket Analysis + Clustering

Identifying association rule sets turned out to be more difficult than the classic peanut butter and jelly example. Our initial attempt with a minimum support of .01 (meaning a pair had to show up on at least 1% of all orders) returned only two rules. Many of the products are clearly associated, but due to the sheer volume of products (and the fact that the average order contained only 2.65 unique products) we needed to relax this constraint. In addition, to make the output of our analysis as applicable as possible, we first narrowed down our input and considered only orders from 2019 or later, as the company indicated this is the period most relevant to current sales.

After performing this analysis at the order-level, it was determined that a number of potentially important associations might be left out. This is due to the fact that many orders contained a single product; if the same customer places an order for Product A today and Product B tomorrow, they are recorded as separate orders and will not show up in our basket results.

So, we 'flattened' the data and performed the same analysis but at the customer level instead. In this way we are more likely to capture important information about the customers we serve. This can help us understand patterns and trends across time periods or industries and even plan our supply/inventory.

This client-level analysis returned 23,224 product pairs that exhibited lift > 1 across all client purchase histories. This provides great potential to not only fine-tune a recommendation system tailored to specific business needs, but also to analyze and understand profitability patterns across product categories.

First, by mapping product data – average cost, sale price, order quantity, etc. - to our association rules, we can easily customize and prioritize recommendations according to business objectives. For instance, when a customer places an individual item into their cart, recommended pairings might be sorted by:

- Confidence – highest 'directional' probability
- Profitability – average margin (dollars or percentage)
- Combination – products meeting minimum confidence threshold, sorted by profitability
- Value – combine price, profit margin and order quantity into single 'value' metric
- Need – weight products by category, region, inventory levels, holding cost

This is just a sampling of the opportunities available; domain knowledge can be used to create any custom scoring metric that can be tested and tuned over time.

These recommendations are not simply limited to individual products; the same process can be applied to any customer in the system. The resulting recommendations indicate which products a customer may be interested in based on their purchase history and might prove useful in marketing emails to encourage follow-up purchases.

Another interesting discovery that emerged from this analysis is the ability to perform a 'reverse' search. That is, for any given product, we can use association rules to identify customers which 'should' purchase that item. An example of this approach began with a particularly profitable item - LPM-08 - which exhibits an average profit of approximately $1.50 and an average order quantity of over 1000. By searching for clients which had purchased its top antecedent, Customer 010069 was identified for purchasing only the first half of a popular pairing. Upon further investigation, we learned that this customer purchased all four items which recommend LPM-08 with over 80% confidence and yet purchased neither it nor any substitutes! This presents a clear call for action: is this customer purchasing that item elsewhere, and if so, what can we do to earn that business? Or, if they do not actually need that item, what can we learn about the way our products are being used? This same process can be applied to every product in our inventory with cases meeting specified criteria being flagged for review by sales and/or product teams.

We also utilized a K-means clustering algorithm that incorporated output metrics obtained from our association model, including support, confidence, lift, leverage, and conviction, in addition to the categorical variable of product type. The clustering model generated 3 optimal clusters for 23,224 product pairs based on mathematical distances. The 'lift' metric from the association model exhibited the most significant signal among all input variables for clustering. Moreover, we merged the output of the K-means clustering model with the average gross profit (GP) margins between the antecedent and consequent product pairs. This enabled us to

identify that product pairs belonging to cluster 1 displayed the highest per-unit profit overall, which indicated a clear opportunity for the TCG to pursue cross-selling initiatives.

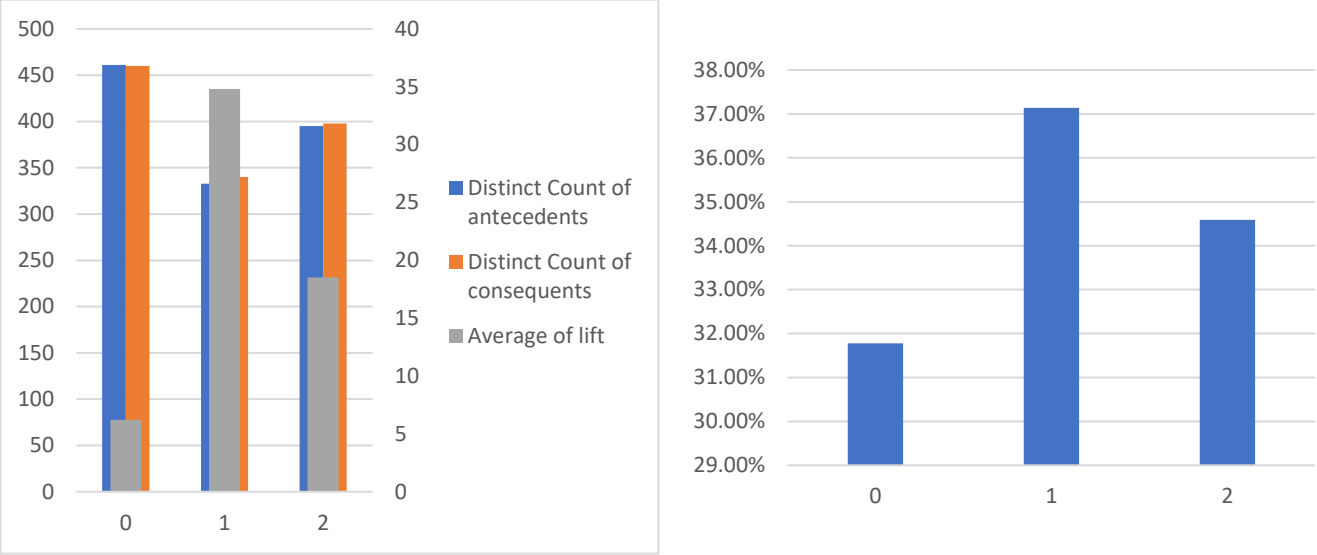


Figure 6. Distribution of antecedent-consequent by cluster (left) and average Gross Profit (GP) margin by products in each cluster (right)

Analysis and Results II – NLP Content-Based Recommender

The Search Recommender was tested by inputting the word "screw" into the model. TFIDF, CounterVectorizer, and KNN gave equivalent results, while spacy gave a different result. The three models return two types of screws sold in different quantities, while SpaCy produced items related to screws. The same occurred when inputting words like "blade" or "switch".

TFIDF

	CustomerID	PN	Description	score	GP%
0	010001	LCB-01B-12B	blade style circuit breakers line circuit brea...	0.682838	0.764927
1	010001	LCB-01B-14B	blade style circuit breakers line circuit brea...	0.682838	0.764923
2	010001	LCB-01B-16B	blade style circuit breakers line circuit brea...	0.682838	0.764242
3	010001	PC-1A-DC-1-RL	screw push button spst red lead electromechani...	0.295497	0.749836
4	010001	PC-1B-DC-1-RL	screw push button spst red lead electromechani...	0.295497	0.749836

CounterVectorizer

	CustomerID	PN	Description	score	GP%
0	010001	LCB-01B-12B	blade style circuit breakers line circuit brea...	0.586987	0.764927
1	010001	LCB-01B-14B	blade style circuit breakers line circuit brea...	0.586987	0.764923
2	010001	LCB-01B-16B	blade style circuit breakers line circuit brea...	0.586987	0.764242
3	010001	PC-1A-DC-1-RL	screw push button spst red lead electromechani...	0.345403	0.749836
4	010001	PC-1B-DC-1-RL	screw push button spst red lead electromechani...	0.345403	0.749836

SpaCy

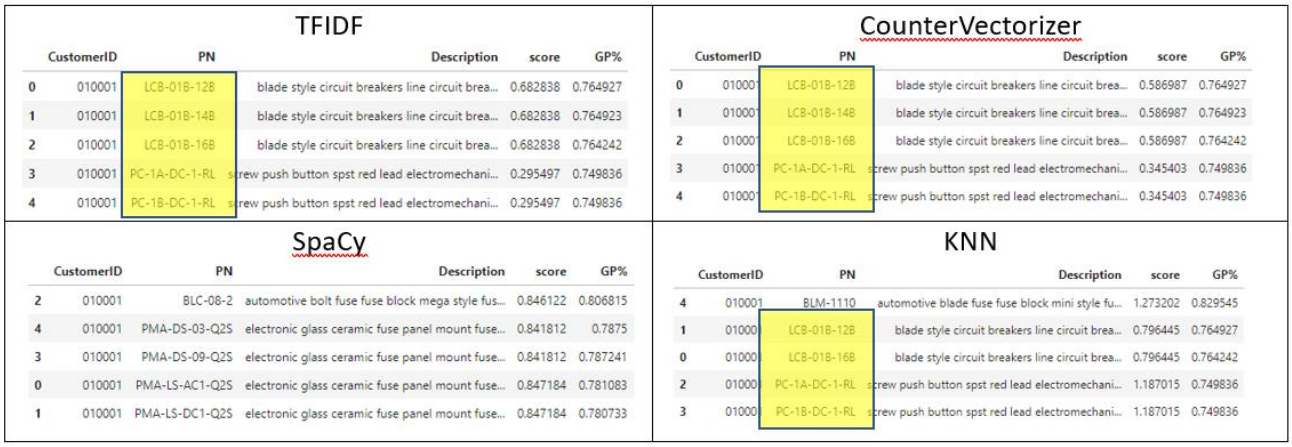
	CustomerID	PN	Description	score	GP%
2	010001	BLC-08-2	automotive bolt fuse fuse block mega style fus...	0.846122	0.806815
4	010001	PMA-DS-03-Q2S	electronic glass ceramic fuse panel mount fuse...	0.841812	0.7875
3	010001	PMA-DS-09-Q2S	electronic glass ceramic fuse panel mount fuse...	0.841812	0.787241
0	010001	PMA-LS-AC1-Q2S	electronic glass ceramic fuse panel mount fuse...	0.847184	0.781083
1	010001	PMA-LS-DC1-Q2S	electronic glass ceramic fuse panel mount fuse...	0.847184	0.780733

KNN

	CustomerID	PN	Description	score	GP%
4	010001	BLM-1110	automotive blade fuse fuse block mini style fu...	1.273202	0.829545
1	010001	LCB-01B-12B	blade style circuit breakers line circuit brea...	0.796445	0.764927
0	010001	LCB-01B-16B	blade style circuit breakers line circuit brea...	0.796445	0.764242
2	010001	PC-1A-DC-1-RL	screw push button spst red lead electromechani...	1.187015	0.749836
3	010001	PC-1B-DC-1-RL	screw push button spst red lead electromechani...	1.187015	0.749836

Figure 7. Search Recommender outputs for "Screw", yellow box indicates similarities between recommendations

For the customer history results, we had a similar occurrence when looking for recommendations for customer 010001. TFIDF, CounterVectorizer, and KNN produced slightly more distinct results than the Search Recommender but were similar. SpaCy was quite different again in this recommender this case as well.



TFIDF

	CustomerID	PN	Description	score	GP%
0	010001	LCB-01B-12B	blade style circuit breakers line circuit brea...	0.682838	0.764927
1	010001	LCB-01B-14B	blade style circuit breakers line circuit brea...	0.682838	0.764923
2	010001	LCB-01B-16B	blade style circuit breakers line circuit brea...	0.682838	0.764242
3	010001	PC-1A-DC-1-RL	screw push button spst red lead electromechani...	0.295497	0.749836
4	010001	PC-1B-DC-1-RL	screw push button spst red lead electromechani...	0.295497	0.749836

CounterVectorizer

	CustomerID	PN	Description	score	GP%
0	010001	LCB-01B-12B	blade style circuit breakers line circuit brea...	0.586987	0.764927
1	010001	LCB-01B-14B	blade style circuit breakers line circuit brea...	0.586987	0.764923
2	010001	LCB-01B-16B	blade style circuit breakers line circuit brea...	0.586987	0.764242
3	010001	PC-1A-DC-1-RL	screw push button spst red lead electromechani...	0.345403	0.749836
4	010001	PC-1B-DC-1-RL	screw push button spst red lead electromechani...	0.345403	0.749836

SpaCy

	CustomerID	PN	Description	score	GP%
2	010001	BLC-08-2	automotive bolt fuse fuse block mega style fus...	0.846122	0.806815
4	010001	PMA-DS-03-Q2S	electronic glass ceramic fuse panel mount fuse...	0.841812	0.7875
3	010001	PMA-DS-09-Q2S	electronic glass ceramic fuse panel mount fuse...	0.841812	0.787241
0	010001	PMA-LS-AC1-Q2S	electronic glass ceramic fuse panel mount fuse...	0.847184	0.781083
1	010001	PMA-LS-DC1-Q2S	electronic glass ceramic fuse panel mount fuse...	0.847184	0.780733

KNN

	CustomerID	PN	Description	score	GP%
4	010001	BLM-1110	automotive blade fuse fuse block mini style fu...	1.273202	0.829545
1	010001	LCB-01B-12B	blade style circuit breakers line circuit brea...	0.796445	0.764927
0	010001	LCB-01B-16B	blade style circuit breakers line circuit brea...	0.796445	0.764242
2	010001	PC-1A-DC-1-RL	screw push button spst red lead electromechani...	1.187015	0.749836
3	010001	PC-1B-DC-1-RL	screw push button spst red lead electromechani...	1.187015	0.749836

Figure 8. Customer History Recommender of customer 010001, yellow box indicates similar products between recommendations

Analysis and Results III – Basket Analysis Package + NLP Integration

Basket analysis was compared with NLP Customer History recommender to compare the top 5 results. Initially customer 010001 was chosen for this comparison. The Basket Analysis 'Top 5'approach returns zero recommended items for this customer. There are two potential causes: customer 010001 has not purchased an item exhibiting lift > 1 with any other items or has already purchased all recommended products. This is an important reminder that although a useful tool, a recommendation system based purely on purchase history has its limitations. Loosening tuning/cutoff parameters may return recommendations for this customer but at the risk of reduced overall confidence in the model.

This is why we propose implementing multiple models for a comprehensive analytics-based solution which captures the strengths and potential synergy between the two methods. The NLP models will return items based on any search term or purchase history, which may capture patterns missed by basket analysis. Basket analysis may provide better ability to understand overall customer tendencies/trends and perform directional analysis. Each provides the opportunity to match potential recommendations with profitability or other desired data. NLP recommendations could also be weighted using association rules and clustering output. Recommendations from each method can be tested and compared using A/B testing on websites and/or marketing emails. There are multiple ways to combine the two methods to enhance our solutions.

Conclusions

We propose several applications using the results of this basket analysis:

- Gain a better understanding of our clients' patterns and needs
    - Analyze pairings across categories to understand which types of pairings provide the most value or contain the greatest potential to increase profits.
    - Identify loss-leaders (individual or category level) and products which may exhibit potential to become valuable loss-leaders.
- Improve recommendations:
    - Cross-selling: when Product A is added to a client's cart, recommend Product B based on strength of association rules – similar to 'customers also purchased' recommendations on many websites. As discussed previously, these can be weighted by profitability or other metrics.

    - Recommend products based on client purchase history – incentive to return for complementary products or consider company as alternative to current supplier.
- Increase sales through targeted intervention:
    - Identify customers that are only purchasing half of a popular pairing. Are they sourcing the other half elsewhere? If so, how can we earn that business?
    - Push desired products through discounts of high-confidence antecedents.
- Boost efficiency of marketing costs:
    - Tailor campaigns to each customer and spend less money advertising products that customers are not likely to purchase.
- Supply chain planning:
    - Predictive analysis – if we determine products are complementary and a recently onboarded client purchases a large quantity of Product A, will we need more Product B in 3 weeks?
    - If Product D is overstocked, identify and offer targeted discounts on Products A/B/C which exhibit high directional lift.

Scientific Research Question 1

Utilizing sales and item data, is there an effective way to recommend items to customers based on the current items they are viewing or past purchase history in a way to help drive sales?
Using the past 5 years of transaction history, we can return the top recommended product pairings for any individual item or customer. These recommendations can be tailored to address specific business outcomes. Clustering similar items also allows us to better understand profitability trends and patterns across product categories.

Scientific Research Question 2

Can we develop (a) Natural Language Processing (NLP) model(s) to improve sales and product search on Triad Component Group's website by utilizing details and descriptions of orders and respective items?
Using the text from product descriptions coupled with either a search query term or customer purchase history as input, we can return the top recommended product pairings. The product recommendation order can be tailored to promote products with higher profit margins.

TCG Management Feedback

"…First of all, I think that the team did a great job, and I would like to thank you for your work. I believe that the conclusions you have drawn from the two analyses will be very useful for our company.
I found the basket analysis specially interesting. It provides great information to better understand our customers' purchasing patterns and we will be using its findings for different business purposes. The most immediate will be to update our website Product Recommendation Engine.
We have two product recommendation blocks, on a product detail page:

- *You can also use: Customers are shown substitute products to keep customers who have already landed on the product page browsing even after they realize that the item is not available/suitable for them*
- *Use it with: Customers are reminded of accessories and complementary items to increase the average order value (cross-selling)*

So far, these recommendations have been made based on our conceptual knowledge of the product and its applications. Thanks to this analysis we will be able to improve these recommendations based on how our customers are actually purchasing the products.
To do this, we will have to improve/complete the analysis identifying if the associated products are substitute or complementary products…" - ***Oihane (Marketing Manager, Triad Components Group)***

Lessons Learned

Aside from the various techniques and their application (which I enjoyed learning and using). A lesson I learned from this course was how to effectively display the approach, data, methods, and findings of data analysis tools and techniques in application. I feel that the structure given and the one I utilized for my assignments have helped me to develop these skills to present my work in a concise yet effective manner. A suggestion I would have for the TAs/Course would be to have a lecture (or set of lectures) that dive into the structure of the reports and provide explanations and real-world use cases in perhaps different professional industries. One of the great aspects of this program is the diversity of professionals and fields; it would be great for those of us beginning our careers and even those having developed theirs's to see the way (and impact) that reports like this are applied and used! – ***Shawn***

I have also appreciated the emphasis this course placed on understanding and explaining the use of each model. Of course, there may be times where we simply seek the most accurate prediction possible, but in almost every application we will be better served by approaching with an understanding of the strengths, weaknesses, assumptions and limitations of the methods we choose to use. Being able to grasp and convey a simple example of what is actually being performed will undoubtedly aid us as we propose projects, select models and explain results to others. – ***Chris***

The aim of this paper is to explore techniques that can be applied to real-world problems by incorporating various modeling techniques. This objective presented a significant challenge from a project management perspective, but we were able to successfully overcome obstacles by prioritizing tasks based on project requirements and maintaining open communication channels among team members. Leveraging the diverse backgrounds and expertise of our team members from various industries proved to be a crucial factor in the development of a robust final product. I recommend that instructors and TAs encourage students to collaborate with individuals from different professional backgrounds to add practical benefits to the students. - ***Joonsoo***

*This course has helped me reinforce a process for approaching an analytics problem and how to properly convey my findings in a report format. For me, it is easy to learn how to solve a set problem, but more difficult to express and draw conclusions from the answers found. Having to explain my approach and reasoning to solving the problems presented gives me a deeper understanding of the models and what the solutions mean. Also doing our project would a real-world issue from a company has set the framework for how I can tackle analytical problems at my workplace. Understanding the business requirements and fitting your approach to the customer/business will not always fit the textbook method. In the end there is a balance between textbook methods and the requirements of the business/customer. -****Richie***

It was quite interesting to work on a real world problem that could provide value to a company. One lesson learned was the importance of trying different methods that achieve a similar goal. For example, in the case of the NLP methods, three provided similar results, but the fourth model provided divergent results. To further understand this phenomenon it would be useful to perform A/B testing to understand how implementation of the models actually impacts sales. - ***Rebecca***

References

Raschka, Sebastian. *Mlxtend Package Documentation*, https://rasbt.github.io/mlxtend/.

Bogart, Ben. "The 'Frequently Bought Together' Recommendation System." *Medium*, Towards Data Science, 5 Oct. 2021, https://towardsdatascience.com/the-frequently-bought-together-recommendation-system-b4ed076b24e5.

Olivares, Armand. "Building NLP Content-Based Recommender Systems." *Medium*, Medium, 7 July 2019, https://medium.com/@armandj.olivares/building-nlp-content-based-recommender-systems-b104a709c042.

"Spacy 101: Everything You Need to Know · Spacy Usage Documentation." *SpaCy 101: Everything You Need to Know*, https://spacy.io/usage/spacy-101.

M, Darla. "How TF-IDF Works." *Medium*, Towards Data Science, 17 Feb. 2021, https://towardsdatascience.com/how-tf-idf-works-3dbf35e568f0.

Zhou, Victor. "A Simple Explanation of the Bag-of-Words Model." *Medium*, Towards Data Science, 11 Dec. 2019, https://towardsdatascience.com/a-simple-explanation-of-the-bag-of-words-model-b88fc4f4971.