BUDT 737: Enterprise Cloud Computing and Big Data

Project Title: Retail Insights - Graph Analytics for Enhanced Customer Experiences

Team 14

Members: Aniket Chafekar
Shiv Mohan Lanka
Yash Shetty

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	Signature
1)	Aniket Chafekar	Aniket Chafekar
2)	Shiv Mohan Lanka	L Shiv Mohan
3)	Yash Shetty	Yash Shetty

II. Executive Summary

In this retail analytics project, our aim was to uncover valuable insights for a retail store by understanding the dynamics of transactions between workers and customers. To achieve this, we utilized advanced data analysis techniques, starting with the implementation of PageRank mechanisms through SQL graph frames in PySpark. This approach allowed us to identify the most influential transactions and workers within the store. By employing unsupervised machine learning, specifically K-means clustering, we segmented the data into five distinct clusters based on prediction values, shedding light on patterns and relationships among different customers and workers.

One of the key highlights of our study involved the implementation of collaborative filtering, a powerful recommendation system. This allowed us to predict the likelihood of a customer purchasing a particular item, even if they had not made such a purchase before. The integration of frequent pattern mining further enriched our analysis by unveiling associations between items frequently bought together, providing valuable insights into product bundling opportunities.

As a result, our research not only provides a comprehensive understanding of customer-worker transactions but also equips the retail store with actionable intelligence to enhance customer experience, optimize worker interactions, and drive strategic decision-making. The combination of graph analytics, machine learning, and pattern mining offers a holistic perspective, making this study a valuable asset for any retail establishment seeking data-driven excellence.

III. Data Description

The datasets used in this study were sourced from stack overflow, providing detailed records of transactions between workers and customers. The data were meticulously collected and anonymized to ensure privacy and comply with ethical standards.

What the Data Are:

Edges Dataset:

Variables: src (source worker/customer), dst (destination worker/customer), relation (type of relationship(buys/manages)), item (purchased item), rating (customer rating).

Units: Categorical (worker or customer for src and dst, buys or manages for relation, item names, and ratings).

Nodes Dataset:

Variables: id (worker/customer ID), name (worker or customer).

Units: Categorical (worker or customer for name).

Items Dataset (for Frequent Pattern Mining):

Variables: item (combination of items bought together).

Units: Categorical (item names separated by semicolons).

Test Dataset (for Collaborative Filtering):

Variables: src (worker/customer ID), item (item to predict purchase probability).

Units: Categorical (worker or customer for src, item names).

Sample Size (n) and Number of Variables (k):

Edges Dataset: n = 1500, k = 5

Nodes Dataset: n = 500, k = 2

Items Dataset: n = 1750, k = 1

Test Dataset: n = 17, k = 2

This data is super interesting because it helps us understand how customers and workers interact in the store. By figuring out what customers like and how workers influence purchases, we can make shopping experiences better. We use this info to predict what customers might want, making it easier to stock products and plan promotions.

III. Research Questions

- Identify the key transactions that significantly impact customer satisfaction and purchasing behavior. Which transactions contribute the most to positive customer experiences? How can we optimize these interactions to enhance overall satisfaction?
- Who are the most influential actors in driving engagement and sales? How can we leverage their strengths to enhance overall worker performance?
  Predict and cater to customer preferences for personalized shopping experiences.
- What products do customers prefer, and how can we strategically position them in the store?Optimize inventory management based on predicted customer preferences.
- How can we efficiently stock products aligned with customer demand? Are there specific items that should be prioritized or bundled together for increased sales?
- What insights can be gained to tailor marketing efforts to customer preferences? How can we use the data to create promotions that resonate with our target audience?

IV. Methodology

In our retail analytics project, we employed a multi-faceted methodology combining graph analytics, machine learning, and other advanced data mining techniques to extract meaningful insights. Here's a concise summary of the methodologies used and the rationale behind them:

- **Graph Analytics (PageRank Mechanism):**
  Utilized SQL graph frames in PySpark to implement the PageRank mechanism.
  This technique allowed us to identify influential transactions and workers within the retail store. PageRank is well-suited for capturing the significance of relationships and transactions, providing a foundation for understanding key interactions driving customer satisfaction and worker impact.
- **Applied K-means** clustering to segment data into distinct clusters based on prediction values.K-means clustering enabled us to categorize customers and workers into groups, offering insights into patterns and relationships. This technique is effective for discovering inherent structures in the data, guiding decisions related to worker optimization and enhancing customer experiences.
- **Implemented collaborative filtering** to predict customer preferences for items not yet purchased. This allowed us to predict the likelihood of a customer buying a particular item, even if they had not made such a purchase before. This personalized recommendation system is crucial for tailoring the shopping experience, driving sales, and improving overall customer satisfaction.
- **Frequent Pattern Mining:** Employed frequent pattern mining to identify associations between items frequently bought together. This technique provided insights into item relationships, supporting strategic decisions related to product bundling and placement. By understanding which items are commonly purchased together, the retail store can optimize inventory and enhance cross-selling opportunities.

V. Results and Finding

RQ1:

Relation: buys

116, worker, 14.036168656614949, 116, 51, buys, pencils, 4,

116, worker, 14.036168656614949, 116, 145, buys, laptop, 3,

Relation : manages

116, worker, 14.036168656614949, 1, 168, manages, online, 30,

140, worker, 13.697565136254159, 14, 107, manages, offline, 40,

RQ2:

From the above transactions, leveraging the worker id 116's influence in the customer and worker sections, promoting new items like new models in laptops, should be done through him. We can enhance the overall satisfaction by leveraging the clustering plots constructed between rating and pagerank. Analyzing such a graph gives us an opportunity on where/whom to focus our marketing and advertising on, so that the overall performance (rating is improved).

RQ3:

Transaction: 410, paper weights, 6.0, 4.439337, 267, pens, 8.0, 3.6358492.

Having a high prediction means that the customer is more likely to rate it highest. So, using this output customer/worker in the above mentioned transaction for advertising purposes helps in a good review of the products.

RQ4:

Bundling paper clips with stickers (where confidence=0.33816425120772947,lift=1.1883281919950333, support=0.12) in the store attracts and reminds more customers in purchase of the items. Similarly, grouping stickers paper clips (where confidence=0.42168674698795183, Lift=1.1883281919950333, support=0.12) and the combining mouse and mouse pad (where confidence=0.3882113821138211, Lift=1.4454679121259297, Support=0.10914285714285714) also might increase chances of customers purchasing the consequent products.

RQ5:

With the use of above information, new products can also be advertised together by setting up a threshold lift/confidence. For example if confidence threshold is 0.2 and lift threshold is 1.1, grouping headphones, mouse, mouse pad together in the store and in the marketing campaigns, helps increase sales.

VII. Conclusion

This study leveraged advanced data analysis techniques, including graph analytics, machine learning, and frequent pattern mining, to extract valuable insights from transaction data between workers and customers in a retail store. The implementation of the PageRank mechanism through SQL graph frames in PySpark identified influential transactions and workers, while K-means clustering segmented the data into distinct clusters, revealing patterns and relationships among customers and workers.

The collaborative filtering approach enabled the prediction of customer preferences for items they had not previously purchased, facilitating personalized recommendations and enhancing the shopping experience. Furthermore, frequent pattern mining unveiled associations between items frequently bought together, supporting strategic decisions related to product bundling and placement.

The findings from this project provide the retail store with actionable intelligence to optimize worker interactions, enhance customer satisfaction, and drive strategic decision-making. By leveraging the insights gained from graph analytics, machine learning, and pattern mining, the store can tailor marketing efforts, improve inventory management, and create a more engaging and personalized shopping experience for customers.

In summary, this study offers a comprehensive understanding of customer-worker transactions and equips the retail store with data-driven strategies to achieve operational excellence, increase sales, and foster long-term customer loyalty.