REPURCHASING INFORMATION: AN IN-DEPTH STUDY ARTIFICIAL INTELLIGENCE PROJECT

Nick Wils
Master of Computer Science: Data Science and Artificial Intelligence
University of Antwerp
student number: 20162945
Email: nick.wils@student.uantwerpen.be

Abstract—**This paper explores the enhancement of a H&M recommender system for a kaggle competition by focusing on repurchase patterns across three sprints. Sprint 1 reveals a substantial 16.4% of transactions involve repurchases, identifying predictive opportunities such as sizing-related patterns. Sprint 2 establishes a personalized baseline, combining repurchase data and popularity metrics, achieving a personal record Kaggle score of 0.02149. Sprint 3 explores candidate generation methods based on repurchase data, average spending, and a combination of methods. Despite challenges, the study underscores the significance of repurchase information in this competition, while acknowledging the need for further exploration and refinement, particularly in candidate generation.**

1. Introduction

In the pursuit of refining the H&M recommender system for a kaggle competition, this research mostly focused on repurchase patterns. The exploration traversed six pivotal research questions, each examined and analyzed through a series of three sprints.

This paper continues as follows: Section 2 navigates through the background and related work. Section 3 explains the overarching methodology, providing insight into the structured framework employed throughout the project. The heart of the research lies in Sections 4 up to 6, where three chronological sprints unfold, each dedicated to solving a primary and secondary research question. Each research question will have a methodology, result and discussion. Section 7, the conclusion consolidates the findings, shedding light on the successes, challenges, and potential avenues for future exploration.

2. Overall Methodology

The iterative process employed in each sprint followed a structured methodology, but could differ based on the needs of the research question. The approach can be summarized into the following steps:

Step 1: Analysis — Analyzing the training data and probing the test set. Including, if needed, a limited literary study.

Step 2: Find a Starting point — Begin by identifying a base for the implementation, often leveraging a portion of the Radek notebook or adopting the most straightforward and simple implementation available.

Step 3: Incremental Improvements — Iteratively implement enhancements to the baseline. Utilize relevant metrics such as the Kaggle score, MAP@12, precision@12, and recall@12 to guide the process. This step involves experimenting with different strategies and algorithms.

Step 4: Combine Implementations — When appropriate, combine different implementations to create a comprehensive solution.

Step 5: Reflect and Evaluate — Regularly reflect on the progress made, evaluating whether the implemented strategies address the research questions effectively. This step involves a critical examination of what worked well and what did not, leading to insights that guide future iterations.

Step 6: Formulate New Research Questions — Identify the most interesting and promising areas for further exploration based on the insights gained. Formulate new research questions that extend the scope of the investigation, addressing aspects that were not covered in the initial questions.

2.1. Utilization of Kaggle

The development environment chosen for this research was Kaggle due to its advantageous features. Kaggle provides a platform with ample RAM resources, built-in version control, and the convenience of storing intermediate results. The ability to easily access and call upon stored data without local downloads simplifies the experimentation process, contributing to a more efficient and effective research workflow.

3. Background and Related Work

The foundation of this research lies in the utilization of Radeks notebook, providing a starting point for the exploration and enhancement of store recommender systems. [1] Radeks notebook includes a Warmup Notebook which in a modified version played a crucial role in data preprocessing. [2]

4. Sprint 1, Data Analysis

In the first sprint two research questions were explored. The first idea was investigating whether people repurchased articles and if this information would help prediction making. Secondly the effect of seasonality on the buying trends of customers regarding fabric and color was investigated.

4.1. RQ1: To what extent can leveraging item repurchase data, item similarity, and customer ownership information enhance baseline predictions of customer purchases?

The underlying objective of this research question is to find out whether a customer's previous purchase of an item correlates with the likelihood of repurchasing the same item or a closely related one.

4.1.1. Methodology. To address this question, An analysis of repurchase patterns was conducted, employing metrics such as repurchase rate, average time between purchases of the same item, item-specific repurchase rate, and customer-specific repurchase rate. Additionally questions such as: "Who makes repurchases? What is repurchased?" are answered. To evaluate the impact of repurchasing information on predictions, the Radek notebook was utilized, generating three prediction versions: a baseline incorporating the full predictions, a modified version with keeping only predictions for items the customer already owns purchased in the last 10 weeks (other predictions were changed to 00000), and a final version with all predictions, excluding those purchased in the last 10 weeks (which were set to 00000). As an extension, exploration to customers purchasing items similar to those they already owned was performed.

4.1.2. Results and discussion repurchases. Analyzing repurchase information revealed some interesting insights. Here are the key findings:

- Average repurchase rate: 16.4%
- Average time between purchases: 5.37 days
- Instances of bulk purchases: 570 in 1 day
- Numerous instances of 2-3 purchases in 1 day, likely indicating different sizes
- Number of unique articles bought: 104,547
- Number of unique articles bought by the same customer on different days: 63,079
- Percentage of repurchasing articles on different days among all bought articles: 60.336%
- Number of unique customers buying: 1,362,281
- Number of unique customers buying the same article on different days: 377,478
- Percentage of repurchasing customers on different days among all buying customers: 27.709%

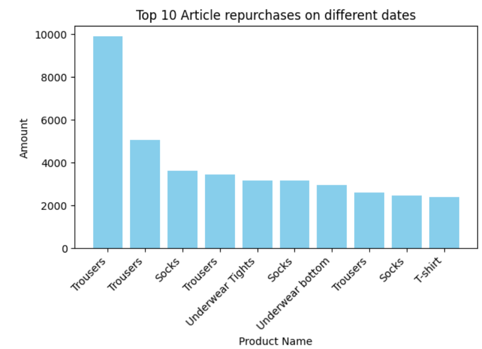


Figure 1. Types of top 10 most repurchased articles.

These findings underscore the significance of repurchase patterns, with a substantial 16.4% of transactions being repurchases. In these repurchases 3 patterns were uncovered:

- Bulk purchases: Typically associated with events or projects.
- A few instances of the same item bought in a single day: Possibly indicating purchases in multiple sizes, with the unused sizes being returned.
- Items repurchased with some days in between: Possibly due to sizing issues where the customer sent back the original item and purchased a different size.

The last category holds particular interest in this use case, as it suggests a predictive opportunity: customers are likely to repurchase an item that they initially bought toward the end of the training data. This insight could be valuable for refining future predictions and recommendations.

Notably, there emerged a significant amount of bulk purchases that requires careful consideration in the recommendation system, as they may skew statistics related to popular items.

In the Figure 1 the categories of the 10 most repurchased articles are shown. This reveals that mostly trousers, leggings and socks are repurchased. Looking beyond the top 10 you see the most of the other categories.

To assess the significance of repurchase information in improving predictions, I utilized, as described in the methodology, the Radek notebook to make predictions. The goal is to evaluate the impact of repurchase patterns on the accuracy of the model's predictions. Three distinct versions were generated from the same predictions for submission to Kaggle:

- Baseline (Base Radek): This version incorporated the full set of predictions without any modifications.
  Kaggle Score: 0.02005
- Repurchase Information Only (Only owned articles): Predictions for items that customers did not purchase in the last 10 weeks were set to 00000.
  Kaggle Score: 0.01564
- Excluding Recent Repurchases (Only new articles): Predictions for items purchased by customers in the last 10 weeks were set to 00000.
  Kaggle Score: 0.00434

The comparison of these versions revealed compelling insights into the impact of repurchase data on the model's performance. Notably, approximately 78% of the baseline score was attributed to predictions involving repurchases. This version only used 8% of the prediction space.

4.1.3. Results and discussion item similarity. Moving beyond repurchase patterns, the exploration extended to item similarity as a potential factor influencing customer preferences. Initially, attempts were made to implement different algorithms such as cosine similarity. However, limitations in available system resources, particularly RAM, needed to be addressed.

Two options were considered to address these limitations: pruning the article dataset to focus on frequently purchased items with selected columns or exploring item similarity within different articles of the same product type. The latter option, focusing on similar products within the same type, was chosen due to its relevance to the repurchasing context—customers buying the same item in different colors.

To gauge item similarity, all combinations of customer_id and product_code were analyzed. The following key statistics were derived:

- Total unique combinations: 24,414,690
- Combinations with at least 2 different article_ids: 4,946,329 (20.259%)
- Combinations with at least 2 different article_ids on at least 2 different dates: 1,708,722 (6.999%)

These findings indicate that among customers who bought at least one article within a product code, approximately 20% of the time, they also purchased another article of the same product code. Notably, in around 7% of customer-product code combinations, customers bought different articles of the same product type on different days. This presents a valuable opportunity to leverage owned items in making predictions, offering a potential avenue for refining the recommendation system.

4.2. RQ2: What is the impact of seasonality on the colors and materials people buy, and can this information be leveraged to enhance prediction scores?

The idea here is that people might have different clothing preferences regarding fabrics and colors depending on the weather and season. Given that the test data week (23/09 - 29/09) falls within autumn, the focus lies on identifying potential shifts in buying behavior during this seasonal transition.

4.2.1. Methodology and results. To validate this seasonality hypothesis, an initial investigation involved consulting fashion blogs [3] and academic studies [4]. These sources affirmed that indeed, people tend to gravitate towards different colors and fabrics depending on the season.

Analyzing the training data, fabric information was extracted from the product descriptions. While the most popular fabrics and colors remained consistent, a notable increase was observed in woolen, dark red, and gray items during autumn.This finding is in line with the paper and fashion advice.

The next step involved incorporating this seasonal information into the predictive model. Initial attempts included creating features directly from fabric types, but these did not yield significant improvements. Subsequently, a binary feature denoting "autumn themed color or fabric" (1 if true, 0 otherwise) was introduced. This feature was then used to boost the initial sample weights, leading to varying degrees of success. The most optimal results were achieved with a threefold boost of the sample weights, resulting in a Kaggle score improvement from 0.02005 to 0.02009.

4.2.2. Discussion. While the exploration into seasonal variations in buying preferences provides some insights, it introduces a challenge concerning timeliness. Seasonal changes are unlikely to occur abruptly within a single week, potentially limiting the immediate relevance of this information in the context of the one-week test data. It is likely that the model already incorporates a degree of seasonality by focusing on recent weeks during training. It may be more suitable to explore the impact of seasonality on a more diverse time scale, incorporating a broader range of temporal data during model training. But this falls outside the scope of the initial research question.

5. Sprint 2, Improved Baseline

Building upon the insights gained from the initial research questions, the decision was made to advance with the exploration of repurchase information as it exhibited promising potential, particularly following its success in the Radek notebook. While the application of seasonality-based color and fabrics might have potential, the decision to prioritize repurchase information in this sprint was motivated by its demonstrated effectiveness.

In this sprint, the focus was placed on applying the knowledge gained from the previous analysis. For this second sprint the choice was made to exploit repurchase data through a personalized baseline.

5.1. RQ1: Is a combination of recently purchased items and the most popular items a more effective baseline for predicting future purchases? What is the optimal time frame for recently purchased items to be useful in predictions?

This research question encompasses two crucial components: popularity and owned items. The strategy was to enhance popularity initially and subsequently determine the optimal combination of owned items and popularity.

5.1.1. Methodology. The foundation for both for popularity was established with the Radek bestsellers. Improvements began by redefining popularity, acknowledging the impact of bulk purchases. The definition of popularity changed from most items bought to items bought by the most unique customers—an approach more aligned with this use case. Further enhancements to popularity involved exploring different timeframes and categorizing customers into groups, aided by pruning the testweek.

The personalized baseline was then constructed by combining the last 12 purchases of each customer with the 12 most popular items, eliminating duplicates, and selecting the first 12 items as predictions. This approach prioritizes repurchase predictions while supplementing them with popular items.

To determine the optimal combination, experimentation was conducted with different timeframes for the repurchase component (ranging from 1 to 10 weeks). Delving deeper into repurchase analysis, consideration was given to utilizing owned items only if they belonged to a category that is typically repurchased. Finally an attempt was made to further improve the baseline with time decaying popularity.

In evaluating the effectiveness of the personalized baseline, two key metrics were utilized: MAP@12 implemented in the Radek notebook and the Kaggle score. Notably, due to computational constraints associated with calculating the repurchase list for each customer, the focus primarily remained on optimizing the Kaggle score. The computational demands, particularly in excluding the last week for test purposes, influenced the metric selection.

5.1.2. Results. The Radek implementation of popularity yielded a Kaggle score of 0.00703, whereas my implementation demonstrated improvement with a score of 0.00738. To determine the optimal timeframe for incorporating repurchase information, various timeframes were tested. Results indicated a downward trend in the Kaggle score from 0.00738 using the last week to 0.00564 using the last month and further declining to only 0.00315 using the last year. This trend is shown in Figure 2.

To enhance the baseline, customers were grouped based on age buckets, the following buckets worked best: [age<25] [25<=age<50] [50<=age] with a score of 0.00747.

Further experimentation involved exploring various customer groupings including, excluding people with only 1 purchase from popularity. The most significant enhancement was achieved by splitting the customers up into buying solely women's clothing / men's clothing / baby clothes and people buying from multiple categories. I achieved these buckets by extracting words like man, woman, baby in the section column of the articles dataset. This resulted in a score of 0.00739.

Analyzing repurchase information exclusively, predictions were made using solely owned items within different

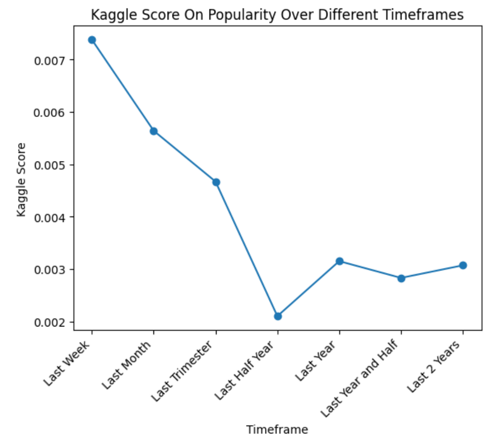


Figure 2. Kaggle score of popularity on different timeframes.

timeframes. The Kaggle scores for these experiments can been seen in Figure 3a.

The critical information here is the slope that is observed around 4 weeks, indicating that repurchase information's significance diminishes beyond this timeframe.

Subsequently, my implementation of popularity without groups was added to repurchase information in different timeframes as shown in Figure 3b.

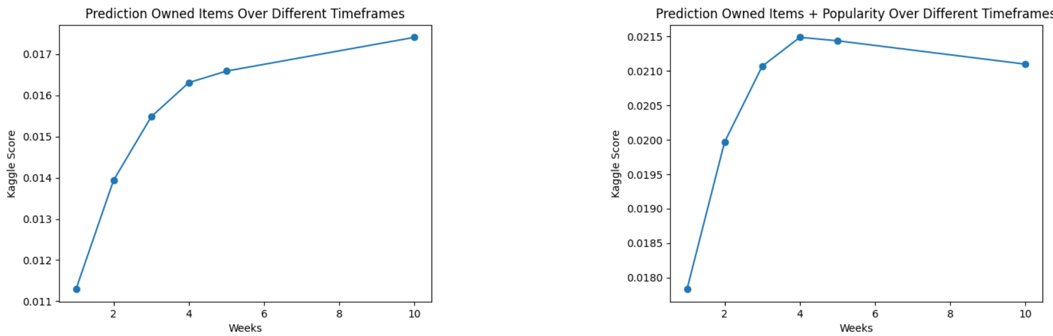
These results indicated that, up to 4 weeks, repurchase information remained more influential than popularity in improving predictions. After this the added owned items from the 5th week and beyond work worse than popularity.

The next experiment was utilizing the grouped popularity instead of general popularity. This resulted however in a slightly lower score when combined with owned articles. 0.02111 was the best grouped popularity attempt,which was achieved with the age buckets. This result was surprising as grouped popularity worked better in a vacuum.

Further optimization was explored by filtering owned times, only including those that belonged to categories that are typically repurchased. The top 20 article categories were extracted and owned items that did not belong to one of those categories were not included in the prediction. However, this approach did not yield improvement, achieving a score of 0.02108 with the best settings.

In an effort to refine the baseline, time-decaying popularity was considered. This involved bucketing customers based on their weekly purchase patterns. While adapting a Kaggle notebook that implemented this approach and already had a score of 0.0216, adding my popularity and repurchase information improved the score to 0.02153. [5]

5.1.3. Discussion. This research question explored the synergy between repurchase information and popularity, revealing that repurchase information holds substantial predictive power, especially within the initial 4 weeks. A personal record of 0.02149 on Kaggle was achieved utilizing solely my code. While the different grouped iteration did not have



(a) Only owned articles

(b) Owned articles and popularity

Figure 3. Kaggle score of owned article with and without popularity on different timeframes.

added value to the final baseline, they might still be useful in modeling or candidate generation.

5.2. RQ2: Does augmenting customer profiles with items that have the same product_id, lead to better predictions? When filling the prediction space, should it prioritize similar items or popular items for improved accuracy?

The aim is to improve the baseline by utilizing the analysis of customers buying different items with the same product code

5.2.1. Methodology and results. The approach involved adding different articles with the same product code to the list of owned articles in predictions, with subsequent filling based on popularity. Initially, attempts were made to exclusively use similar items, resulting in a best score of 0.00319. Subsequent efforts to enhance this by incorporating repurchased items and popularity were unsuccessful.

5.2.2. Discussion. The specific implementation for improving predictions using item similarity did not yield the desired outcomes. Further research and experimentation may be required to identify a refined approach that balances the influence of item similarity with other predictive factors.

6. Sprint 3: Candidate Generation

Building upon the success of the previous sprint, the focus shifted to exploring the integration of repurchase information in a different phase of the prediction-making process: candidate generation. Two research questions were formulated to guide this exploration and since these questions are very related, they will be discussed together.

6.1. Research Questions

To what extent does generating candidate recommendations based on repurchase data and popularity contribute to improved prediction accuracy?

What specific measures of popularity, such as overall purchase frequency, or bought by most users, proves to be most effective in enhancing the accuracy of candidate recommendations?

6.1.1. Methodology. These research questions used Radeks candidate generation and model as a base. Utilizing the same model model, different iterations of candidate generation were implemented. Evaluation metrics such as Radek's MAP@12, my own precision@12 and recall@12, and the Kaggle score were used to compare and assess the effectiveness of the various candidate generation methods.

The first step involved implementing candidates based on whether they were previously purchased. Subsequently, the average amount customers spent was added as a factor in the candidate generation process. The final iteration merged Radek's candidate with my own.

Furthermore, my own implementation of popularity was introduced and utilized both in candidate generation and as a feature in the model. This was then combined with the different candidate generation methods developed during the sprint.

6.1.2. Results. The implementation of candidate generation proved to be a challenging endeavor. While the application of repurchase information in candidate generation did not yield significant improvements, Table 1 illustrates the best results achieved for each type of candidate generation method. For all candidate generation methods, except those using Radek's candidate generation, my own implementation of popularity was employed. Radek's bestsellers were used in those exceptions as they demonstrated superior performance.

6.1.3. Discussion. Despite my efforts, the exploration into incorporating repurchase information into candidate generation did not result in a notable improvement. The success observed in the baseline, coupled with the significant number of repurchases identified in the analysis, suggests the potential for improvement in this area. While a breakthrough was not achieved by the deadline, the results highlight the

TABLE 1. CANDIDATE GENERATION RESULTS

	MAP	Precision	Recall	Kaggle
repurchases	0.013108	0.003898	0.020388	0.01164
repurchases + avg. spending	0.000134	0.000095	0.000496	0.01161
Radek + repurchases + avg. spending	0.006866	0.005364	0.021700	0.02043
Radek	0.016931	0.008483	0.038930	0.02046

complexity of creating an effective and enhanced candidate generation method. Moving forward, the pursuit of a more dynamic and flexible implementation may prove to be more fruitful. Further research into refining the approach to candidate generation, especially one that accommodates the dynamic nature of repurchase patterns, remains an avenue worth exploring.

7. Conclusion

The journey through the Kaggle competition is an exploration into the intricacies of customer behavior. Beginning with an analysis of repurchase patterns, the initial sprint focused on uncovering insights into customer purchasing habits. The findings highlighted the significance of repurchase information, with a notable 16.4% of transactions being repurchases. The identification of patterns such as bulk purchases and repurchase patterns presented opportunities for refining future predictions.

A second analysis delved into the impact of seasonality on purchasing preferences, specifically in relation to colors and fabrics. Insights from fashion blogs and studies were corroborated with an analysis of training data, revealing a notable surge in autumn-specific items. While seasonality proved relevant, the question of timeliness emerged, suggesting that its influence may not be instantaneous and may require consideration in the context of more diverse temporal data during model training.

The subsequent sprints delved into leveraging repurchase information in different aspects of the predictive modeling process. The personalized baseline, introduced in Sprint 2, demonstrated the power of combining repurchase data and popularity metrics. The success of different implementations of popularity in different settings showcases that the best scoring implementation might be different depending on the combinations.

The exploration extended to candidate generation in Sprint 3, where the challenge of incorporating repurchase information persisted. Despite rigorous efforts, a breakthrough in developing an improved candidate generation method remained elusive.

The research questions posed throughout the project led to valuable insights and considerations. While the personalized baseline revealed the importance of repurchase information within a limited timeframe, the attempt to augment customer profiles with items of the same product_id proved challenging.

In conclusion, the project provided valuable insights into customer purchasing behavior, the influence of seasonality, and the challenges of incorporating repurchase information into different stages of the recommender system. While certain goals were achieved, such as the successful implementation of a personalized baseline, the project also revealed areas that require further exploration and refinement such as candidate generation.

References

[1] M. Gorelli, “Radek’s lgbmranker starter-pack,” *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/code/marcogorelli/radek-s-lgbmranker-starter-pack

[2] ——, “Radek’s lgbmranker starter pack warmup,” *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/code/marcogorelli/radek-s-lgbmranker-starter-pack-warmup

[3] the concept wardrobe, “True autumn: A comprehensive guide,” *the concept wardrobe*, 2019. [Online]. Available: https://theconceptwardrobe.com/colour-analysis-comprehensive-guides/true-autumn-a-comprehensive-guide

[4] K. B. Schloss, R. Nelson, L. Parker, I. A. Heck, and S. E. Palmer, “Seasonal variations in color preference,” *Wiley Online Library*, vol. 41, no. 6, pp. 1589–1612, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/cogs.12429

[5] M. Bhattacharyya, “Time decaying popularity benchmark [0.0216],” *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/code/mayukh18/time-decaying-popularity-benchmark-0-0216/