Personalized Recommendations of Upcoming Sport Events

Sebastian Cornelius Bergh

Supervisor: Assoc. Prof. Dr. Mehdi Elahi
Co-supervisors: Dr. Lars Skjærven and Astrid Tessem

Master's Thesis
Department of Information Science and Media Studies
University of Bergen

May 31, 2023

Scientific environment

This study takes place within the Department of Information Science and Media Studies at the University of Bergen. It is a part of Work Package 2 of the MediaFutures center, which concentrates on user modeling, personalization, and engagement. The research is conducted in collaboration with the media platform TV 2.

Acknowledgements

I would like to express my sincere appreciation to my supervisors, Dr. Mehdi Elahi, Dr. Lars Skjaerven, and Astrid Tessem, for their invaluable guidance, motivation, and support throughout the entire thesis. I am really grateful to have had the opportunity to work under their supervision, benefiting from their knowledge and mentorship. I am particularly thankful for the exceptional level of support, assistance, and active engagement they have provided during every phase of this research. Their expertise and dedication have significantly contributed to the success of this thesis. I would also like to thank MediaFutures and TV 2 for granting me the unique opportunity to work on this thesis. The chance to work with real data from TV 2 and deploy my project on their website has been invaluable in gaining practical insights and enhancing the relevance of my research.

This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through The Centers for Research-based Innovation scheme, project number 309339.

Sebastian Cornelius Bergh
Bergen, May 2023

Abstract

Recommender systems have emerged as essential tools for enhancing user engagement and content discovery in various domains, including the sports industry. In the context of sports viewing, personalized recommendations have become increasingly significant, enabling users to easily connect with their favorite sports teams, explore new content, and broaden their viewing preferences. Collaborative filtering (CF) stands out as a popular recommendation algorithm that analyzes the similarities and patterns in user-item interactions. By examining the behavior and preferences of a group of users, CF identifies similar users and recommends items that have been positively received by those with similar tastes. Applying CF to sports recommendations presents an opportunity to introduce users to new sports events enjoyed by their peers. However, recommending upcoming live sports events introduces unique challenges, such as limited availability and the need to strike a balance between catering to users' favorite sports and introducing them to new content.

This master thesis aims to address these challenges through the development of a personalized recommendation system for upcoming sports events using CF. The system will analyze user viewing history to provide tailored recommendations that facilitate content discovery and enable users to easily locate their preferred sports events. The research objectives include identifying the most suitable collaborative filtering model for sports content recommendation, investigating the factors that influence sports fans' preferences for specific types of live sports events, and evaluating the effectiveness of personalized recommendations compared to non-personalized approaches. The proposed system is implemented and A/B tested on TV 2 Play, one of Norway's largest digital streaming platforms, with the ultimate goal of enhancing user experience and engagement by delivering personalized and relevant recommendations for sports content. This research contributes to the field by proposing a novel collaborative filtering recommender for sports based on user viewing sessions, exploring effective strategies for recommending upcoming live sports events, and assessing the system's performance in terms of accuracy and user satisfaction.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

According to a report by Grand View Research, the global sports analytics market size was valued at USD 1.9 billion in 2020 and is expected to grow at a compound annual growth rate (CAGR) of 21.0% from 2021 to 2028. This growth is being driven by the increasing demand for data-driven decision-making in the sports industry (*Research*, 2021). However, while data analytics has made significant strides in various aspects of sports, there is still a crucial area that requires attention and improvement - personalized recommendations for sports viewers.

Personalized recommendations have become increasingly important for sports viewers seeking seamless engagement with their favorite sports teams and athletes, as well as the ability to discover new and exciting content. Recommendation systems that leverage user data and preferences play a pivotal role in providing tailored content and recommendations, enhancing the user experience, and increasing engagement and retention on streaming platforms (*Aggarwal*, 2016). Without a reliable recommender system, users may not be exposed to different sports they may enjoy, limiting their viewing options. However, a well-designed recommender system can introduce them to sports that align with their preferences, as well as further diversify their viewing experience (*Elahi et al.*, 2021).

Collaborative filtering (CF) stands out as one of the most popular recommendation algorithms. By analyzing the preferences of users who share similar tastes, CF predicts a user's preference for a particular item (*Jannach et al.*, 2010). One specific CF technique, utilized by the Alternating Least Squares (ALS) algorithm, involves leveraging matrix factorization to reveal latent factors within user-item interactions and capture the underlying data structure. By decomposing the user-item matrix into user and item embeddings, ALS effectively models user preferences and captures key characteristics of the items (*Kuroda et al.*, 2020). Leveraging collaborative filtering techniques for sport recommendation such as ALS offers an exciting opportunity for users to be introduced to sports events that similar peers have previously enjoyed, increasing the likelihood of their enjoyment as well (*Aggarwal*, 2016).

However, there are several challenges and considerations when it comes to recommending upcoming live sports events. The live aspect of the recommendation, where only a limited number of items are available at any given time, adds complexity to the recommendation process (*Turrin et al.*, 2014). Additionally, there is a fundamen-

tal question of how to strike the right balance between recommending users' favorite sports, which we know they already enjoy, and introducing them to new sports to expand their preferences. Sports fans also tend to have diverse interests and might be interested in watching events from a variety of different sports and leagues. Without a recommendation system to guide them, it might be difficult for them to discover new content and engage with the full range of sports events available (*Petander*, 2019).

Hence, the primary objective of this thesis is to address these challenges and provide users with a more engaging and rewarding viewing experience through the development of a personalized recommendation system using collaborative filtering for upcoming sports events. By analyzing user viewing history and preferences, the recommendation system will offer personalized recommendations that help users discover new content and easily find their favorites. Throughout this thesis, we will explore the methodology, evaluate the effectiveness of the system, and assess its impact on user engagement and satisfaction.

1.2 Problem statement

Although there has been extensive research in the field of recommender systems, including separate research on both the recommendations of linear shows and sports content, there is still a significant gap in the literature when it comes to recommending upcoming live sports events. This gap highlights a novel and challenging problem that needs to be addressed. One of the unique challenges of recommending live sports events is the temporal and dynamic nature of the data, which requires real-time and personalized recommendations to users (*Turrin et al.*, 2014). Additionally, sports present further complexities, such as users' strong biases towards certain teams or sports, the need for fresh and relevant content, and the limited lifecycle of specific sports (*Petander*, 2019).

To address this issue, this thesis aims to develop a sports recommender system for TV 2 Play[1], one of the largest media platforms in Norway. The system leverages collaborative filtering techniques to recommend upcoming live sports events to users based on their previous viewing history. The end goal is to improve the user experience on the TV 2 Play platform by providing users with personalized and relevant sports content.

1.3 Objectives / Research questions

In order to explore and tackle the specific challenges presented by the problem statement, the thesis aims to answer the following research questions:

RQ1: Which collaborative filtering model is best suited for recommendation of sports content?

RQ2: Which factors influence sports fans' preferences for specific types of live sports events?

[1] https://play.TV2.no/

1.4 Contribution

The main contributions of the thesis are the following:

- Proposing a novel sports-based collaborative recommendation technique based on user viewing sessions of sports content. The implementation of the proposed approach can be found in the MediaFutures Github repository [2].
- A comprehensive offline evaluation of the proposed recommendation approach, including comparisons with different baselines on accuracy and beyond accuracy metrics.
- Developing and deploying a sports-based collaborative recommender on one of Norway's largest digital streaming platforms (TV 2 Play) for A/B testing.

1.5 Thesis outline

- **Chapter 2: Background** Provides an overview of the literature and core concepts relevant to this thesis. Section 2.1 delves into the background knowledge of recommender systems. Section 2.2 discusses the use of these systems in the media industry. Section 2.3 explores the use of recommender systems in generating sports recommendations, while section 2.4 centers on creating linear TV recommendations. Section 2.5 gives an overview of relevant evaluation and online testing of recommender systems. Finally, section 2.6 summarizes the previous related work and highlights key differences from the work in this thesis.
- **Chapter 3: Methods** Outlines the specific techniques and procedures employed to address the research questions. Section 3.1 presents the dataset provided by TV 2. Section 3.2 provides an overview of the recommendation algorithms used for the offline evaluation. Section 3.3 describes the recommender algorithm used for the online experiments, while Section 3.4 explains the technical details for the offline evaluation. Section 3.5 describes the experiment design, including both the offline evaluation and the online experiments. The offline evaluation is detailed in subsection 3.5.1, while the online experiments are described in subsection 3.5.2.
- **Chapter 4: Results and Discussion** Details the analysis conducted and the results from both the offline evaluation and the online experiments. Section 4.1, Experiment A: Exploratory Analysis, provides an overview of the exploratory analysis with a specific focus on time, sport, and tournaments. Section 4.2, Experiment B: Offline Evaluation, outlines the process of selecting hyperparameters and presents the results of the offline evaluation. Section 4.3, Experiment C: Online evaluation, presents the results from both experiments employed for A/B testing on TV 2 Play.
- **Chapter 5 Conclusion and Future Work** Provides an overview of the thesis by discussing its main contributions, results, limitations, and future work. The chapter is divided into four sections. Section 5.1 summarizes the research carried out

[2] https://github.com/sfimediafutures/MA_Sebastian-Cornelius-Bergh

in the thesis. Section 5.2 discusses the key contributions of the thesis. Section 5.3 presents the results obtained based on the research questions set out. Finally, Section 5.4 highlights the limitations of the thesis and discusses potential directions for future research.

Chapter 2

Background

This section provides an overview of the literature and core concepts relevant to this thesis. Section 2.1 delves into the background knowledge of recommender systems. Section 2.2 discusses the use of these systems in the media industry. Section 2.3 explores the use of recommender systems in generating sports recommendations, while section 2.4 centers on creating linear TV recommendations. Section 2.5 gives an overview of relevant evaluation and online testing of recommender systems. Finally, section 2.6 summarizes the previous related work and highlights key differences from the work in this thesis.

2.1 Recommender systems

The ever-increasing volume and diversity of data available online, including videos, articles, and images, have created a significant challenge for users in discovering relevant content. To address this issue, recommender systems have proven to be invaluable in assisting users to discover content that aligns with their interests and preferences (*Bobadilla et al.*, 2013; *Elahi et al.*, 2018; *Jannach et al.*, 2010).

The idea of recommender systems occurred in the early 1990s with the objective of assisting online users in discovering more relevant and engaging content (*Jannach et al.*, 2010; *Schafer et al.*, 1999). Recommender systems have demonstrated their effectiveness in diverse forms and domains. These systems utilize various data sources to infer customer interests, including both implicit and explicit feedback. Explicit feedback is more direct such as likes, dislikes, and ratings. Whereas implicit feedback is even easier to collect and comes from user actions such as viewing a particular video, or purchasing a specific product on a website (*Aggarwal*, 2016).

Recommender systems have become ubiquitous on the web, appearing in various forms, such as video-sharing platforms like YouTube, where they suggest related videos, to news sites that recommend other relevant articles. While their primary objective is to boost revenue for the merchant, the ways in which they achieve this objective are diverse. For instance, for YouTube, the click-through rate (CTR) is crucial as it generates more ad revenue. For e-commerce sites like Amazon, recommending products that customers are likely to purchase is key. However, the importance of providing relevant recommendations goes beyond mere revenue generation. Recommending the right item to the customer can also improve user satisfaction, leading to better cus-

tomer retention. In fact, personalized recommendations have been shown to increase the CTR on one of the most popular news sites Forbes by approximately 37 percent (*Kirshenbaum et al.*, 2012).

To achieve this broader business-centric goal of maximizing revenue, recommender systems typically have four goals, as described by *Aggarwal* (2016): *Relevance, Novelty, Serendipity,* and *diversity*. These goals go beyond merely increasing short-term revenue and aim to improve the overall user experience and engagement. By providing relevant, novel, and diverse recommendations, users are more likely to stay on a platform or visit it more frequently, ultimately contributing to increased revenue in the long term (*Aggarwal*, 2016).

Relevancy is perhaps the most apparent goal of recommender systems and involves their ability to suggest items that are relevant for the user (*Aggarwal*, 2016). As the user is more likely to engage with items that are interesting to them, recommending relevant items is essential.

The goal of *Novelty* in a recommendation system refers to its ability to suggest items that the user has not engaged with in the past (*Aggarwal*, 2016). In addition to helping users discover new things, this can also lead to an increase in sales diversity.

If a user consistently consumes the same items and the recommendations they receive reinforce this behavior, there may be undesired effects such as filter bubbles. To prevent this, the importance of *serendipity* in a recommendation system cannot be overstated. Serendipity refers to the system's ability to recommend something unexpected for the user (*Aggarwal*, 2016). Unlike novelty, serendipity involves recommending something truly surprising, not just something the user has not previously engaged with. This adds an element of randomness for the user to discover new things.

The final goal mentioned is to increase recommendation *diversity*. This can be achieved by introducing some variety in a typical suggested top-k items list. Such lists can often contain very similar items, so by adding some diversity, there is a greater chance that the user may like at least one of the recommendations (*Aggarwal*, 2016).

2.1.1 Types of recommender systems

Recommender systems are employed across many domains, resulting in variations in the data used to generate predictions. This has led to the development of several distinct types of recommender systems, primarily differentiated by the basis for their recommendations. The three primary approaches are *content-based*, *collaborative*, and *hybrid*.

Content-based (CB) recommender systems utilize a user's past preferences to recommend items that share similar content to what the user has liked before (*Aggarwal*, 2016; *Deldjoo et al.*, 2015; *Pazzani and Billsus*, 2007). They analyze item attributes and suggest similar items based on user interests. These systems rely on item data and user profiles created from implicit or explicit feedback. Additional data sources like social media profiles or purchase histories can also be incorporated to enhance personalized recommendations. Modern CB systems can also exploit audio-visual features automatically extracted from the images or videos (of items) and incorporate them into the recommendation process *Rimaz et al.* (2019, 2021). This integration of audio-visual elements brings several advantages, including addressing the challenge of cold start

problems(*Cheng et al.*, 2019).

Collaborative Filtering (CF) is a technique used to filter the most promising items out of a large set, where users implicitly collaborate by providing feedback. Collaborative recommendation approaches utilize the past preferences and interests of users to recommend items that people with similar tastes have enjoyed in the past (*Aggarwal*, 2016; *Elahi et al.*, 2018; *Koren et al.*, 2021). The main idea behind this approach is that if two users share similar interests and have liked the same items in the past, it is likely that they will have similar interests in the future (*Jannach et al.*, 2010). For example, if user A and user B have similar taste in movies and watch a lot of the same ones, but A has recently watched a movie that B has not, it would make sense to recommend the movie A has watched to B (*Jannach et al.*, 2010). It is important to note that while CF does not require any knowledge of the item itself, it does depend on having a large amount of data on user preferences and item ratings. Additionally, CF can suffer from the "cold start problem," where new users or items do not have enough ratings to generate meaningful recommendations. Therefore, it is essential to balance the strengths and limitations of CF with other approaches such as content-based filtering to provide more accurate and diverse recommendations (*Cantador et al.*, 2010; *Hazrati and Elahi*, 2021; *Jannach et al.*, 2010).

The two recommendation approaches mentioned above have their respective advantages and limitations due to their use of different sources of data. However, using these approaches in isolation may lead to situations where the recommendation process is limited. As a solution, a *hybrid* approach has been developed to combine the best of both worlds. Hybrid recommender systems are designed to exploit more knowledge available in different data sources, reducing the limitations of isolated systems (*Aggarwal*, 2016; *Burke*, 2002; *Elahi et al.*, 2023; *Kvifte et al.*, 2022). According to *Aggarwal* (2016), there are three primary ways to create a hybrid recommendation system. The first approach is *ensemble design*, which combines the results from different algorithms into a single, more robust output. For instance, a content-based and a collaborative recommender could be combined to produce a single rating output. The second approach is *monolithic design*, where the recommendation algorithm uses different data types, such as item attributes and user feedback, to make recommendations. This approach combines the strengths of both content-based and collaborative filtering methods to produce a single output. In this approach, the recommendation algorithm may use item attributes to find items that are similar in content to the user's past preferences, and then use collaborative filtering to recommend items that other users with similar preferences have liked. The third approach is *mixed systems*, which allows the user to compare and choose between the different recommendations provided by the different algorithms. Mixed systems can be designed to present recommendations from different sources in various ways, such as ranking items based on a combination of recommendations or displaying items recommended by each algorithm in separate sections. It is important to note that hybrid recommender systems need to be carefully designed to avoid introducing bias and ensure an effective combination of the different approaches.

2.1.2 Challenges and ethical considerations

Despite the benefits of recommender systems outlined thus far, it is crucial to acknowledge the significant challenges they present, including algorithmic biases, lack of trustworthiness, and ethical considerations (*Elahi et al.*, 2022; *Wang et al.*, 2022). These challenges have a profound impact on shaping user preferences and influencing choices (*Klimashevskaia et al.*, 2022). The impact is far-reaching beyond the domains of online streaming and e-commerce, and they may be implemented in contexts that involve moral considerations such as healthcare, insurance, and the labor market. Failing to address these ethical issues in the design, deployment, and use of recommender systems may lead to opportunity costs, public distrust, and even backlash against their use in general (*Milano et al.*, 2020).

User *privacy* is a primary ethical challenge in recommender systems, with personal data collection and usage posing risks of breaches and rights violations (*Friedman et al.*, 2015; *Milano et al.*, 2020). Collaborative filtering techniques can raise systemic privacy concerns by constructing accurate user profiles even with limited data (*Milano et al.*, 2020). Addressing these privacy concerns is crucial in the design, deployment, and use of recommender systems. Another important consideration is *fairness*. Unfairness can be defined as the presence of bias, prejudice, or favoritism towards certain individuals or groups based on their inherent or acquired characteristics (*Burke*, 2017; *Elahi et al.*, 2021; *Ge et al.*, 2021). The lack of fairness can profoundly affect individuals and society, highlighting the need to address it in the design and development of recommender systems, despite its subjective and challenging nature to define. As recommender systems continue to play an increasingly important role in people's lives, the need for *transparency* in these systems has become more apparent. Transparency entails making recommendations understandable and explainable to users. Lack of transparency can lead to mistrust and dissatisfaction with the system. In the context of recommender systems, transparency can be achieved by providing explanations about how the system works and how recommendations are generated. This can be particularly important for ensuring fairness in the system. Research on transparency in recommender systems has shown that users are more likely to trust and use a system that provides transparent explanations (*Balog et al.*, 2019; *Elahi et al.*, 2021). Additionally, *manipulation* is a significant ethical concern in recommender systems, especially on social media and news platforms, where filter bubbles and biases can be reinforced (*Milano et al.*, 2020). Certain user groups can manipulate the system by generating positive feedback and driving up the system's rate of recommendations for specific items. This can be particularly problematic in the context of news recommendation systems and social media platforms, as demonstrated by the Cambridge Analytica scandal in 2018 and external interference in US political elections in recent years (*Milano et al.*, 2020). To address this, various approaches have been proposed to promote *diversity* in recommendations, but striking a balance between relevance and diversity remains a challenge (*Milano et al.*, 2020).

2.2 Media Recommendation

Content recommendation has become increasingly popular for online consumers on modern media sites, providing users with suggested videos, articles, and personalized experiences (*Albanese et al.*, 2013; *Elahi et al.*, 2021; *Yu et al.*, 2006). These systems not only enhance user satisfaction but also offer significant business value for providers. For instance, *Gomez-Uribe and Hunt* (2016) notes that Netflix, one of the most popular media services, considers its recommender system a core component of its business. According to the paper, personalization and recommendations save Netflix more than 1 billion dollars annually as of 2015. Moreover, personalized recommendations help distribute viewing across a wider range of videos, including those in the long tail, thus benefiting both users and content providers.

These recommender systems are increasingly automated, often determined by AI algorithms with the goal of helping consumers discover relevant content more easily. However, the highlighting or filtering of information that comes with such systems can lead to undesired consequences, such as filter bubbles and the spread of misinformation (*Fernández et al.*, 2021). The lack of editorial control may unintentionally amplify false or misleading information. Other ethical considerations discussed in section 2.1.2 are also particularly pertinent to media content recommendation. Privacy, for instance, is a crucial issue, given that these systems gather vast amounts of user data to facilitate personalized recommendations. Transparency is also important, ensuring users have insight into how recommendations are generated and addressing biases or inappropriate content (*Elahi et al.*, 2021).

2.3 Sport recommendation

The task of enabling users to discover and engage with relevant sports content from a vast catalog is a critical challenge for sports media distributors (*Petander*, 2019). Such discovery may involve finding live games of the user's favorite team or discovering new sports to follow. To address this challenge, a well-designed sports recommender system can offer personalized and relevant recommendations to the user, resulting in increased fan engagement, better content discovery, and higher revenue from targeted advertising.

However, live sports present unique challenges for recommender systems. Fans typically prefer fresh content, such as last night's game or even better, a game currently being broadcasted. A sports recommender system must be able to provide real-time recommendations based on live events, such as unexpected upsets or changes in the schedule. Additionally, the lifecycle of sports, including seasons, leagues, and tournaments, significantly influences a sports fan's interests. For example, a user who supports a team that is knocked out of a tournament may lose interest in following that competition (*Petander*, 2019).

2.4 Linear TV recommendation

Linear TV refers to programs that air at a scheduled time on a specific channel. As a result of this specific format, there can occur challenges when trying to provide recommendations. Unlike standard video-on-demand (VoD) recommendations, recommendations for linear TV need to take into consideration the fact that programs are scheduled at specific times (*Kim et al.*, 2018). The catalog of items is very dynamic, with different items being available at different times. In addition, the user's consumption patterns are strongly affected by both time context and channel preferences (*Turrin et al.*, 2014).

Turrin et al. (2014) describes several areas where recommending TV programs for linear TV can be more challenging than conventional VoD recommender systems. Some of the most important are *Dynamic catalog of items*, *Time-constrained catalog of items*, and that *a user cannot watch different TV channels simultaneously*.

Dynamic catalog of items: In services like VoD, the available content is updated very rarely, maybe a few movies are uploaded every day or week. For linear TV, however, each program is scheduled at a specific time, making it available only for a specific time interval. This makes the catalog of items constantly change and at the same time leads to constant new item problems as many upcoming TV programs never have been watched in the past (*Turrin et al.*, 2014).

Time-constrained catalog of items: Not only does the catalog constantly update, but the time in which they are available is also limited. In standard VoD users have the option to select and view content whenever they want. For the recommendation of linear TV programs, however, the recommendation must take into account that it should only consider programs that are transmitted within a certain time period after or during the moment of recommendation.

A user cannot watch different TV channels simultaneously: For linear TV, different potentially attractive programs for the user may run at the same time on different channels, forcing the user to make a choice. In standard recommender systems, this is not a problem as items always are available and users can consume several items at the same time. When analyzing viewing habits on linear TV, a recommender system must therefore consider that some programs may not have been watched because it was scheduled at the same time as a more attractive program, and not solely because the user did not find it interesting (*Turrin et al.*, 2014).

2.5 Offline and online evaluation of recommender systems

Performance evaluation is a crucial part of recommender systems and determining what constitutes a good recommender system is a key problem in this evaluation process. Evaluating the performance of recommendation algorithms serves as the basis for algorithm selection and it is therefore essential to evaluate these algorithms on various datasets to obtain the optimal parameters before deploying to the online system (*Cheng and Liu*, 2017). *Offline evaluation* and *online experiments* are two key methods used in the evaluation of recommender systems.

Offline evaluation involves collecting datasets of user behaviors in advance, such as choices or ratings on items, to simulate interactions between users and recommender

systems. These datasets can be randomly sampled from real user behavior logs or obtained from certain time stamps of the log. It is important that these collected datasets closely resemble the true user interactions. The basic method of offline evaluation draws inspiration from machine learning and typically involves dividing the dataset into training and test sets and constructing recommendation models based on the training data, then evaluating their performance on the test data (*Cañamares et al.*, 2020). The advantages of offline evaluation include its low cost and quick evaluation of different recommendation algorithms, as it doesn't need interaction from real users. However, it is limited in evaluating factors like serendipity or novelty and is mainly focused on prediction accuracy or Top-N precision of recommendations (*Cheng and Liu*, 2017). Overall, the main goal of offline evaluation is to compare the performance of the recommendation algorithms with the use of some metrics and filter out unsuitable algorithms, and be left with some candidate algorithms. Then the more costly online experiment can be carried out for further evaluation and optimization.

Online experiments involve large-scale testing on already deployed recommender systems. This method evaluates or compares different recommender systems based on real tasks performed by real users and therefore provides the most realistic testing results among the evaluation methods (*Cheng and Liu*, 2017). Some of the advantages of online experiments are that they allow for the entire performance of the recommender system to be evaluated, such as long-term business profit and user retention, rather than solely relying on single metrics. Therefore, Online experiments can be used to understand the impact of evaluation metrics on the overall performance of the system. In many cases, designers of the system wish to influence user behavior through recommender systems, and online experiments enable the evaluation of the systems' influence on user behavior. Factors like users' intentions, familiarity with items, trust in the system, and UI design play a role in the actual effect of recommender systems (*Cheng and Liu*, 2017). When conducting online evaluations, problems that should be taken into consideration include random sampling of users to ensure a fair comparison, consistency in influencing factors when focusing on specific metrics and avoiding situations where recommender systems recommend too many unrelated items, leading to reduced user trust (*Gunawardana et al.*, 2022).

2.6 Related research and key differences

The problem of providing recommendations in a linear setting has shown to be more challenging in several areas including the time-constrained nature of the items in the catalog and the dynamic updates to the catalog based on the TV schedule (*Turrin et al.*, 2014). In their work, *Turrin et al.* presents a time-based recommender that takes into account these challenges. Previous studies such as *Kim et al.* (2018) present a time-aware recommender but with a focus on standalone TVs. *Sanchez et al.* (2012b) describes a recommender system for sports videos with a focus on audiovisual consumption. The article also addresses some important considerations when recommending sports content, but does not focus on any live content. The article *Sanchez et al.* (2012a), explores recommender systems specifically designed for sports videos in the context of large-scale events like the Olympic Games. Although their focus is primarily on video recommendations during large-scale events, they have valuable insight into

utilizing implicit user information to provide recommendations. *Ding et al.* (2017) propose a recommender system that combines multiple linear regression and collaborative filtering techniques to recommend football videos based on user preferences. While their emphasis is on football videos, their use of collaborative filtering, along with the utilization of implicit user behavior data, aligns with some of our research objectives.

While several works have focused on both recommendations for live content and recommendations for sports content, there still remains a significant gap in the literature related to the recommendation of live upcoming sports events. This gap is noteworthy considering the popularity and frequent viewership of sports events on linear TV. While previous research has primarily focused on recommendations for live programs and on-demand sports content, there has been limited exploration of how recommendation systems can deliver timely and relevant suggestions for upcoming sports events. Furthermore, it is also worth noting that existing research that has focused on recommendations for live programs or on-demand sports content has often required explicit user feedback.

To fill the gap in the literature, this thesis focuses on both the recommendation of upcoming sports events and the challenges of providing recommendations in a linear TV setting. This study will conduct a large-scale test on one of Norway's biggest streaming platforms, using a novel approach that does not require explicit user feedback.

Chapter 3

Methodology

The Methodology chapter outlines the specific techniques and procedures employed to address the research questions. Section 3.1 presents the dataset provided by TV 2. Section 3.2 provides an overview of the recommendation algorithms used for the offline evaluation. Section 3.3 describes the recommender algorithm used for the online experiments, while Section 3.4 explains the technical details for the offline evaluation. Section 3.5 describes the experiment design, including both the offline evaluation and the online experiments. The offline evaluation is detailed in subsection 3.5.1, while the online experiments are described in subsection 3.5.2.

3.1 Dataset

The study utilized various datasets provided by TV 2. The datasets were divided into four types based on the type of information they contained. These included data about viewing sessions, metadata pertaining to sports events, participant-specific information, and general information about a given sporting event. The data itself comprises information such as the start and end times of viewing sessions, the type of event, the sport, participants, and the tournament or cup the event was a part of. It was collected over approximately three and a half months, from the end of June 2022 to the middle of October 2022. This subsection provides a description of how these datasets were manipulated and combined based on different usage areas in both the exploratory analysis and model-building phases. Due to the sensitive nature of the dataset, only an approximation of the number of users and interactions is included.

To provide a better understanding of the initial datasets used in this study, Table 3.1 displays their sizes in terms of observations and features.

Table 3.1: Size of datasets

Dataset	Observations	Features
Viewing sessions	> 1 000 0000	9
Metadata	6 838	7
Participants	1 768	8
Events	4 200	10

For the exploratory analysis, the datasets were manipulated in a range of different

ways. Two distinct datasets were created for the purpose of generating *K-means* clusters. One dataset was designed to cluster users based on their sports viewing patterns, while the other aimed to cluster users based on their tournament viewing patterns. Both datasets were aggregated by user id, with the watch time of each sport or tournament serving as the features. Table 3.2 provides an example sample of the dataset used for clustering users based on sports viewing patterns. The dataset used for clustering users based on tournament viewing patterns followed a similar construction approach but with tournaments serving as features instead of sports. The rest of the datasets generated for the explorative analysis are detailed in Section 4.1

Table 3.2: Sample of the aggregated dataset on user id and sport. Showing the duration each user has watched each sport in seconds

User Id	Basketball watched	Cheerleading watched	E-sport watched	...	Fotball watched
1	13,000 sec	0 sec	5,000 sec	...	0 sec
2	0 sec	10,000 sec	0 sec	...	21,000 sec
3	0 sec	0 sec	5,000 sec	...	0 sec
4	0 sec	6,000 sec	7,000 sec	...	0 sec
5	0 sec	3,000 sec	0 sec	...	8,000 sec
6	0 sec	0 sec	0 sec	...	16,000 sec
7	15,000 sec	0 sec	0 sec	...	0 sec
8	0 sec	10,000 sec	6,000 sec	...	0 sec
9	0 sec	0 sec	0 sec	...	4,000 sec

For the model building, two separate datasets were created - one for sports and one for tournaments. The sports dataset was aggregated based on user id and sport, with the total watch time for each user on each sport serving as a feature. Similarly, the tournament dataset was aggregated on user id and tournament, showing the total watch time for each user on each tournament. A description of the aggregated sports and tournament datasets can be found in Table 3.3.

Table 3.3: Dataset description

Dataset	Users	Items	Interactions
Sport	> 100 000	20	> 400 000
Tournament	> 100 000	195	> 600 000

To illustrate the dataset structure, an example sample of the sports dataset is shown in Table 3.4, which is aggregated based on user id and sport. This dataset was used to train and evaluate the models for sports. Likewise, the corresponding dataset for tournaments was used to train and evaluate the models for tournaments.

Before being used in the training of the models, the watch time in seconds was normalized using a min-max scaler. This normalization technique transformed the data to a standardized range, ensuring that the watch time values were adjusted proportionally and constrained within a specific range, allowing for fair comparison and accurate model training (*Patro and Sahu*, 2015).

Table 3.4: Sample of the aggregated dataset on user id and sport. Showing the duration each user has watched each sport in seconds

User Id	Sport	Watch time in seconds
4	Fotball	17,000 sec
4	Sjakk	5,000 sec
4	Sykkel	9,000 sec
11	Sykkel	6,000 sec
11	Basketball	10,000 sec
25	Fotball	9,000 sec
25	Sjakk	7,000 sec
25	Sykkel	11,000 sec
25	Basketball	2,000 sec
25	Ishockhey	4,000 sec
25	Poker	20,000 sec
25	MyGame	3,000 sec

3.2 Recommendation algorithms

Table 3.5 presents the primary recommender algorithm and baselines used in this study. These models, excluding the Random and Most Popular approaches, were obtained from the Python library Implicit[1]. The primary recommender model employed in the experiment was a modified version of the Alternating Least Squares (ALS) algorithm, which is a pure collaborative filtering approach specifically optimized for implicit datasets. This algorithm was chosen as the primary model based on recommendations from the industry partner and its suitability for large-scale applications. The ALS algorithm is based on matrix factorization, a technique that represents user-item interactions in a low-dimensional space (*Hu et al.*, 2008). This algorithm assumes that both users and items have certain latent factors that influence their behavior and attempts to estimate these factors using observed user-item interactions.

Table 3.5: Recommendation algorithms used

Recommender algorithm	Type
Alternating Least Squares (ALS)	Pure CF
Bayesian Personal Ranking (BPR)	Pure CF
Logistic Matrix Factorization (LMF)	Pure CF
Most Popular	Non-personalized
Random	Non-personalized

The ALS algorithm is an iterative approach that alternates between fixing one set of factors (either user or item) and solving for the other set. This process is repeated until the algorithm converges to a solution. The algorithm selects the best set of factors by solving a least squares problem for each iteration, hence the name "Alternating Least Squares." One of the strengths of the ALS algorithm is its ability to handle sparse data, which is common in recommendation systems where users only interact with a

[1] `Implicit`: Fast Python Collaborative Filtering for Implicit Datasets. Available at: https://github.com/benfred/implicit

small subset of available items. This makes it suitable for datasets with a large number of users and items, which may not have complete information about all user-item interactions (*Hu et al.*, 2008).

In the offline evaluation, four baseline algorithms, namely Bayesian Personal Ranking (BPR), Logistic Matrix Factorization (LMF), Most Popular, and Random, will be compared to the ALS algorithm. The Most Popular algorithm is a non-personalized model that recommends the top N most popular items to all users, while the Random is also non-personalized and recommends N random items from the catalog to each user.

Similar to the ALS algorithm, the BPR algorithm is also a pure CF algorithm optimized for implicit datasets and utilizing matrix factorization. However, unlike ALS, BPR uses an optimization criterion, BPR-Opt, and a specific algorithm, LearnBPR, for optimization (*Rendle et al.*, 2009). A key distinguishing feature between these two algorithms is that BPR predicts user preferences for all possible item pairs, instead of predicting exact ratings for individual items. This unique approach allows BPR to be optimized for ranking tasks, enabling it to effectively perform in scenarios where the primary goal is to produce item rankings. (*Rendle et al.*, 2009)

LMF, like ALS and BPR, also utilizes matrix factorization to estimate latent factors that influence user-item interaction. However, it differs in its use of logistic regression to model the probability of a user interacting with an item. This allows the model to handle missing data more effectively and potentially leads to improved accuracy in certain scenarios (*Johnson*, 2014).

In contrast to ALS, BPR, and LMF, the popularity and random models are simple non-personalized models included purely as simpler baselines. While all three CF algorithms use matrix factorization, they differ in their optimization criteria, approaches, and specific techniques, which can impact their performance in different recommendation scenarios.

3.3 Recommendation for online evaluation

The ALS model, which was compared to the baseline models, was the model used for generating recommendations in the online evaluation. In order to enhance its performance, a grid search was conducted to identify the optimal combination of hyperparameters, including *factors*, *iterations*, and *regularization*. Hyperparameters are values that cannot be learned directly from the data and must be set before training the model. However, they can have a significant impact on the model's performance (*Belete and D H*, 2021).

The *factors* hyperparameter represents the number of latent factors used to represent each user and item in the recommendation model (*Bennett and Lanning*, 2007). These factors capture the underlying features or characteristics of users and items that influence their preferences. A higher number of factors can improve the accuracy of the model, but it can also increase computational complexity and training time, so the optimal number of factors depends on the complexity of the dataset (*Bennett and Lanning*, 2007).

The *iterations* hyperparameter is the number of times the ALS algorithm iterates over the training data to optimize the model. Each iteration updates the latent factors for all users and items, based on the observed ratings and the predicted ratings from the

previous iteration (*Koren*, 2008). Increasing the number of iterations can improve the accuracy of the model, but also increase the computational time.

The *regularization* hyperparameter is used to control the amount of regularization applied to the model during training. Regularization is a technique used to prevent overfitting, which is when the model fits too closely to the training data and fails to generalize well to new data (*Bennett and Lanning*, 2007). A higher regularization parameter will increase the amount of penalty applied to the model for larger weights, thus reducing overfitting. However, setting the regularization parameter too high can also result in underfitting, where the model is too simple to capture the underlying patterns in the data (*Bennett and Lanning*, 2007).

3.4 Technical details

For the offline evaluation, the experiments were run in Python 3.9.13. The hardware used was an Intel(R) Core(TM) i5-1035G1 CPU and 8GB RAM.

3.5 Experiment design

This section describes the evaluation approach used to assess the performance of the recommendation algorithms applied in this study. The evaluation consists of two main parts: offline evaluation and online experiments.

The offline evaluation involves conducting an exploratory analysis of the dataset and applying evaluation metrics to measure the performance of the recommendation algorithms. The exploratory analysis aims to identify any patterns, trends, biases, or limitations that may impact the results. The evaluation metrics used will provide a quantitative measure of the performance of the different recommendation algorithms and baselines.

The online experiments involve deploying the selected recommendation approaches on the TV 2 Play website for A/B testing to collect implicit user feedback on their performance. This data will be used to evaluate the effectiveness of the system in a real-world setting. The online experiments will be evaluated through measures such as click-through rate, views, and clicks, to measure the success of the approaches compared to TV 2 Play's current solutions.

3.5.1 Offline evaluation

This section is divided into 2 parts: Experiment A: Exploratory analysis and Experiment B: Quality of recommendations. Apart from the exploratory analysis in experiment A, the two datasets used were the aggregated datasets mentioned in section 3.1, namely the sports dataset aggregated on user id and sport, as well as the tournament dataset aggregated on user id and tournament. The sports dataset contains over 100 000 users, 20 unique sports items, and over 400 000 interactions. The tournament dataset contains over 100 000 users, 195 unique tournament items, and over 600 000 interactions.

Due to uncertainties regarding the inclusion of teams to influence recommendations in the online experiments, it was decided not to include teams in the exploratory analysis and the quality of recommendations sections. However, as the decision changed and teams were eventually included, the approach taken to integrate teams into the online experiments will be elaborated in the online evaluation part.

Experiment A: Exploratory analysis

This part of the offline evaluation delves into an extensive exploratory analysis of the dataset employed in this study, with the aim of gaining a thorough understanding of its characteristics. This analysis covers several aspects of the dataset, including *time*, *items*, and *users*.

Time

The presence of timestamps in viewing sessions allows for the analysis of the duration that users spend on various items, as well as the identification of patterns in their viewing behavior. A heatmap was generated using seaborn's[2] heatmap() method to display the correlation between the day of the week and user activity. This color-coded representation of similarity values between pairs of items visually exposed the similarity between items and helped detect any patterns or trends.

Items

A combination of visualization techniques was used to measure item similarity and gain a deeper understanding of the items in the dataset. The relationships between items were visualized using a dendrogram imported from the Python library Scipy[3]. The dendrogram represents items as leaves of a tree-like structure, with branches indicating the distance between items or clusters. The height of the branches corresponds to the level of similarity or distance between the items, with lower branches indicating greater similarity.

To investigate the similarity between the items in more detail, a heatmap from Seaborn was employed. The heatmap provided a color-coded representation of the similarity values between pairs of items. Additionally, pyplot's histogram from Matplotlib[4] was used to give an overview of the distribution of similarity between the items. This histogram helped give an understanding of the overall distribution of item similarities and identify any outliers or unexpected patterns.

Users

K-means clustering was applied to group users, with the aim of identifying patterns and similarities among the users. The clustering process involved dividing users into a specified number of clusters based on the similarity of their viewing history. To visualize

[2] `Seaborn`: Seaborn is a Python data visualization library based on matplotlib. Available at: https://seaborn.pydata.org/

[3] `Scipy`: SciPy is a free and open-source Python library used for scientific computing and technical computing. Available at: https://scipy.org/

[4] `Matplotlib`: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Available at: https://matplotlib.org/

the clusters, *principal component analysis* (PCA) was used to reduce the dimensionality of the item vectors and enable plot visualization of the clusters for better analysis and interpretation. PCA is a mathematical algorithm that serves the purpose of reducing the dimensionality of the data while preserving most of the variations in the dataset (*Ringnér*, 2008). This reduction is achieved by identifying directions, called principal components, along which the variation in the data is maximal. By using only a few principal components, each sample can be represented by a relatively small number of variables, which facilitates visualization and helps to identify similarities and differences between samples (*Ringnér*, 2008). In this case, the reduced-dimensional representation allows for easier visualization of the data, making it possible to identify clusters and patterns in the data. To visualize the PCA implementation and the clusters, sci-kit learn was used.

Another aspect of the exploratory analysis involved using the majority vote technique to identify the most and second-most preferred items for each user based on their viewing history. This was done for both the tournament and sports datasets and was used in addition to PCA and k-means clustering to give the groups clearer and deeper meaning. These preferences also provided insight into the popularity of the different items.

Experiment B: Quality of recommendations

To assess the accuracy of the recommendations, a set of standard accuracy metrics were employed. These included precision@K, recall@K, average precision@K, NDCG@K, Hit-rate@K, Reciprocal rank@K, ROC-AUC, and PR-AUC. In addition to these accuracy metrics, catalog coverage was also used as a beyond-accuracy metric to evaluate the coverage of the recommended items across the entire catalog.

Accuracy metrics

The metrics precision at top K recommendations *(Precision@K)* is frequently used to evaluate the effectiveness of a system in accurately identifying relevant items. To calculate $P@K$, the system identifies the top K recommended items for each user, while also ensuring that the items have corresponding ratings within the test set T (*Schedl et al.*, 2018). For each user u $P_u@K$ is computed as:

$$P_u@K = \frac{|L_u \cap \hat{L}_u|}{|\hat{L}_u|} \tag{3.1}$$

Where L_u is the set of relevant items for user u in the test set T and $\hat{L}_u$ denotes the recommended set containing the K items in T with the highest predicted ratings for the user u. The overall $P@K$ is then computed by averaging $P_u@K$ values for all users in the test set.

Recall at top K recommendations *(Recall@K)* looks at what proportion of the test items would have been retrieved with the top K recommended list (*Schedl et al.*, 2018). For each user u $R_u@K$ is defined as:

$$R_u@K = \frac{|L_u \cap \hat{L}_u|}{|L_u|} \tag{3.2}$$

Where L_u represents the set of items in the test set T that are relevant to the user u, while $\hat{L}_u$ refers to the recommended set comprising the K items in T with the highest predicted ratings for user u. The overall $R@K$ is calculated by averaging $R_u@K$ values for all user in the test

Mean average precision at top K recommendations *(MAP@K)* is a rank-based metric that computes the overall precision of the system at different lengths of recommendation lists. MAP is computed as the arithmetic mean of the average precision over the entire set of users in the test set (*Schedl et al.*, 2018). Average precision for the top K recommendations *(AP@K)* is defined as:

$$AP@K = \frac{1}{N}\sum_{i=1}^{K} P@i \times rel(i) \tag{3.3}$$

where *rel(i)* is an indicator signaling if the i^th recommended item is relevant, i.e., *rel(i)*=1, or not, i.e., *rel(i)*=0; N is the total number of relevant items.

'Hit Rate' at K *(Hit@K)* is a simple yes or no metric that looks at whether any of top K recommended items were in the test set for a given user u (*Wang et al.*, 2015). Hit@K is defined as:

$$Hit@K = \max_{i=1..K} \begin{cases} 1, & r_i \in T \\ 0, & otherwise \end{cases} \tag{3.4}$$

The average of this metric across users is typically called 'Hit Rate'.

Reciprocal rank at K *(RR@K)* only looks at the rank of the first recommended item that is in the test set, and outputs its inverse (*Chapelle et al.*, 2009). RR@K is defined as:

$$RR@K = \max_{i=1..K} \frac{1}{i} \;\; s.t. \;\; r_i \in T \tag{3.5}$$

The average of this metric across users is typically called "Mean Reciprocal Rank".

Normalized discounted cumulative gain at K *(NDCG@K)* is a metric used to evaluate the quality of ranking in recommender systems. The ranking of recommendations for given user u is based on the predicted rating values, which are sorted in descending order (*Schedl et al.*, 2018). $DCG_u@K$ is defined as follows:

$$DCG_u@K = \sum_{i=1}^{K} \frac{r_{u,i}}{log_2(i+1)} \tag{3.6}$$

Where $r_{u,i}$ is the true ratings found in test set T for the item ranked at position i for user u, and K being the length of the recommendation list. Since the rating distribution depends on the users' behavior, the *DCG* values for different users are not directly comparable. Therefore, the cumulative gain for each user should be normalized. This is done by computing the ideal *DCG* for user u, denoted as $IDCG_u$, which is the DCG_u value for the best possible ranking, obtained by ordering the items by true ratings in descending order. Normalized discounted cumulative gain at K for user u is then calculated as:

$$NDCG_u@K = \frac{DCG_u@K}{IDCG_u@K} \tag{3.7}$$

The overall normalized discounted cumulative gain at *K (NDCG@K)* is then computed by averaging $NDCG_u@K$ over the entire set of users.

ROC-AUC stands for area under the receiver-operating characteristic curve. While the metrics above only looked at the top K recommended items, this metric looks at the full ranking of items instead, and produces a standardized number between zero and one in which 0.5 denotes random predictions (*He and Ma*, 2013). To calculate the ROC-AUC, true positive rate (TPR), aka. recall and false positive rate (FPR) is needed. TPR can be defined as:

$$TPR = \frac{TP}{TP+FN} \tag{3.8}$$

FPR can be defined as:

$$FPR = \frac{FP}{FP+TN} \tag{3.9}$$

The area under the curve can be calculated using the trapezoidal rule which is defined as:

$$\int_a^b f(x)dx \tag{3.10}$$

PR-AUC stands for area under the precision-recall curve. While *ROC-AUC* provides an overview of the overall ranking, the focus is often only on how effectively it retrieves test items within top ranks. In this regard, the area under the precision-recall curve can offer a more informative assessment, although it should be noted that this metric lacks standardization and its minimum value does not reach zero (*He and Ma*, 2013). To calculate the PR-AUC, precision is needed in addition to recall (3.9) and the

trapezoidal rule (3.10). Precision can be defined as:

$$Precision = \frac{TP}{TP+FP} \tag{3.11}$$

Beyond accuracy metric

The diversity of the recommender systems was evaluated using catalog coverage, a beyond-accuracy metric. This metric measures the proportion of unique items in the catalog that are recommended to users, providing insights into the diversity of the recommendations (*Ge et al.*, 2010). Higher catalog coverage indicates a wider variety of items being recommended, which is crucial for exposing users to a diverse range of products or content. Encouraging exploration of a wider range of items can lead to increased user satisfaction, underscoring the importance of evaluating and optimizing recommendation algorithms based not only on accuracy metrics but also on beyond-accuracy metrics like catalog coverage (*Ge et al.*, 2010). The catalog coverage score was calculated for all users with a recommendation of 10 items, and it was used to measure the proportion of unique items that were recommended. The catalog coverage is calculated as:

$$CatalogCoverage = \frac{|U_{j=1...N} I_L^j|}{|I|} \tag{3.12}$$

Where I_L^j is denoted as the set of all items contained in the list L returned by the j^{th} recommendations returned to users. N is the total number of recommendations observed during the measurement time, and I is the set of all available items, i.e., the catalog

3.5.2 Online experiments

Two separate A/B tests were conducted on TV 2 Play's website: Online experiment 1: Recommendation of sports and Online experiment 2: Recommendation of upcoming sports events the next seven days.

Online experiment 1: Recommendation sports

Implementation

In the first online experiment, we implemented the collaborative filtering ALS model to personalize the "Idretter" feed on TV 2 Play's "Sport" page. This feed displays a variety of sports to users (shown in Figure 3.1). Clicking on a specific sport redirects users to a dedicated page that showcases upcoming broadcasts for that sport.

TV 2's current method follows a non-personalized approach, prioritizing the order of items in the list solely based on time. This means that the sports scheduled to be

broadcasted next appear first on the list. This approach is effective, especially for popular sports aired during prime time. While the collaborative filtering approach may not necessarily outperform this baseline in terms of numerical metrics or click-through rates, it has the potential to suggest a more diverse range of items. Consequently, it can introduce viewers to new sports and expand their interests.

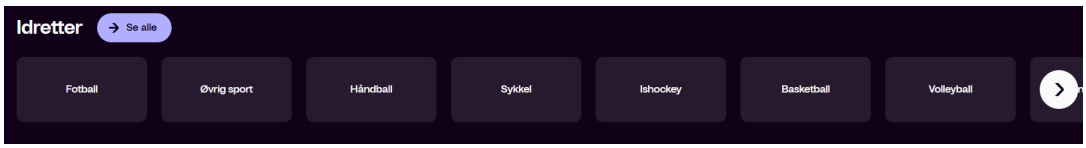

Figure 3.1: Screenshot from TV 2 Play showing the kind of list a user would be presented with. The title translates to "Sports"

Experiment design

An A/B methodology was employed to conduct an online experiment on TV 2 Play between March 30th and April 12th. The experiment aimed to compare user interactions between TV 2's existing time-based approach and our collaborative filtering approach. The recommendation list was available under the "Sport" category on TV 2 Play's online platform. The experiment was presented to 50 percent of users, while the remaining 50 percent received TV 2's time-based approach. User interactions were measured using views, clicks, and click-through-rate (CTR).

Online experiment 2: Recommendation of upcoming sports events the next seven days

Implementation:

In the second experiment, the collaborative filtering ALS model was applied to the "Sport for deg de neste 7 dagene" feed on TV 2 Play's "Sport" page. This feed displays personalized recommendations for upcoming sports events that will be broadcasted within the next seven days (as depicted in Figure 3.2).

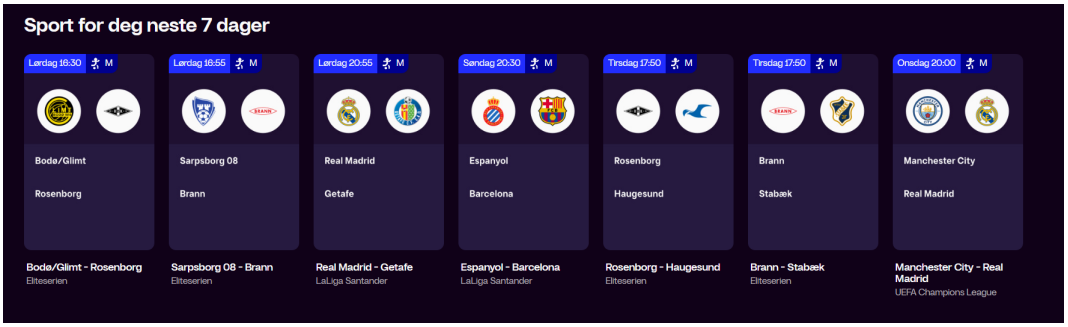

Figure 3.2: Screenshot from TV 2 Play showing the kind of list a user would be presented with. The title translates to "Sport for you the next 7 days"

Currently, this feed is personalized for each user based on their implicit favorites, as well as their explicit favorites. The explicit favorites are determined by a feature TV 2 Plays offers, which lets users follow their favorite team. These teams are then deemed

as explicit favorites by the recommender system. However, these favorites are not widely used. The implicit favorites are determined in a similar fashion to the majority vote, where the teams, tournaments, and sports are deemed as implicit favorites if the watch time for the user is over a certain threshold.

Unlike online experiment 1, the approach for experiment 2 does not propose a pure CF approach but rather incorporates the ALS model to filter out items. Simplified, the steps of this approach can be described as the following for each user (see also pseudocode in Algorithm 1):

- The program iterates through each sports events item that is set to be scheduled during the next seven days.
- If the item contains one of the user's explicit favorite teams, the item will automatically be featured in the feed.
- For each item the system calculates a total score for the item using the ALS model and if the total score is over a certain *threshold*, the item is included in the recommended list for that user. The total score is made up of three different scorings from the ALS model, with scores based on sports, tournaments, and teams. This score is calculated as:
  - The ALS model trained separately on sports, tournaments, and teams is utilized to provide a list of all sports, tournaments, and teams, and a corresponding score.
  - For the items that have assigned tournaments and teams, the scores of these two are used to calculate the total score. These are sports like ”fotball" and ”håndball" where users tend to often follow certain tournaments or clubs, making these the most important factors when creating the score.
  - Some items for example ”sykkel" and ”sjakk" do not have any assigned tournaments or teams. For these items, users tend to follow the sport itself more than individual tournaments and teams. Therefore, only the ALS-score for sport is considered to create the score for these items.
  - When calculating the score for items that only have sports, the user's ALS-score for that sport is added to the score.
  - When calculating the score for items that have tournaments and teams, the score from the teams is added to the score, in addition to the score from the tournament multiplied by 0,2. The score of the teams is weighted more as it's deemed more crucial for the user's preferences.
- The threshold which determines whether an item should be included in the final list is calculated as:
  - The min score is calculated based on the drop-off within the recommendations for each user. In other words where there is a significant drop in score from one item to the next. Specifically, this drop-off is defined when the score of the next item is less than 0.1 times the previous item.
  - This calculation was made on the basis of iterating through all the users in the test set and calculating where the drop-off occurs.

Algorithm 1 Process of recommending sports events based on ALS scoring

```
1: sport_items = Sports items the next 7 days
2: returned_sport_events = empty list of items to be presented in the feed
3: ALS_sport = ALS-score of sports
4: ALS_tournament = ALS-score of tournaments
5: ALS_team = ALS-score of teams
6:
7: for item in sport_items do
8:     score = 0
9:     if item == explicit_favorite then
10:        returned_sport_events.append(item)
11:    end if
12:    if tournament and team NOT IN item then
13:        ALS_score_sport = ALS_sport[item]
14:        score += ALS_score_sport
15:        if score > min_score_sport then
16:            returned_sport_events.append(item)
17:        end if
18:    else
19:        if tournament IN item then
20:            ALS_score_tournament = ALS_tournament[item]
21:            score += ALS_score_tournament *0.2
22:        end if
23:        if Team IN item then
24:            ALS_score_team = ALS_team[item]
25:            score += ALS_score_team
26:        end if
27:        if score > min_score_tournament_team then
28:            returned_sport_events.append(item)
29:        end if
30:    end if
31: end for
32: return returned_sport_events
```

Experiment design

An A/B methodology was employed to conduct an online experiment on TV 2 Play between May 12th and May 28th. The experiment aimed to compare user interactions between TV 2's existing implicit favorite approach and our collaborative filtering approach. The recommendation list was available under the "Sport" category on TV 2 Play's online platform. The experiment was presented to 50 percent of users, while the remaining 50 percent received TV 2's implicit favorite approach. User interactions were measured using views, clicks, and click-through-rate (CTR).

Chapter 4

Results and Discussion

This chapter details the analysis conducted and the results from both the offline evaluation and the online experiments. Section 4.1, Experiment A: Exploratory Analysis, provides an overview of the exploratory analysis with a specific focus on time, sport, and tournaments. Section 4.2, Experiment B: Offline Evaluation, outlines the process of selecting hyperparameters and presents the results of the offline evaluation. Section 4.3, Experiment C: Online evaluation, presents the results from both experiments employed for A/B testing on TV 2 Play.

4.1 Experiment A: Exploratory analysis

Experiment A involves an exploratory data analysis that aims to gain a deeper understanding of the datasets used in this study. The analysis is divided into three parts: Time, Sports, and Tournaments. Each part focuses on a different aspect of the data and utilizes different datasets that were tailored to the specific area of analysis.

In the upcoming subsections, a detailed overview of each part of Experiment A will be provided. This overview will include information about the datasets used, the visualization techniques employed, and the key findings.

4.1.1 Time

The dataset used to create the heatmap in Figure 4.1 was aggregated on users, with the amount of time they watched content on different days of the week serving as features. This heatmap shows the correlation between the different days that users watch items.

The days with the highest correlation are Tuesday and Wednesday, as well as Saturday and Sunday. As we can see from the barplot in Figure 4.2, these are also the most watched days. The reason for this is primarily due to the fact that major football matches are often scheduled on these days, and as we will demonstrate later, a significant number of users watch a substantial amount of football.

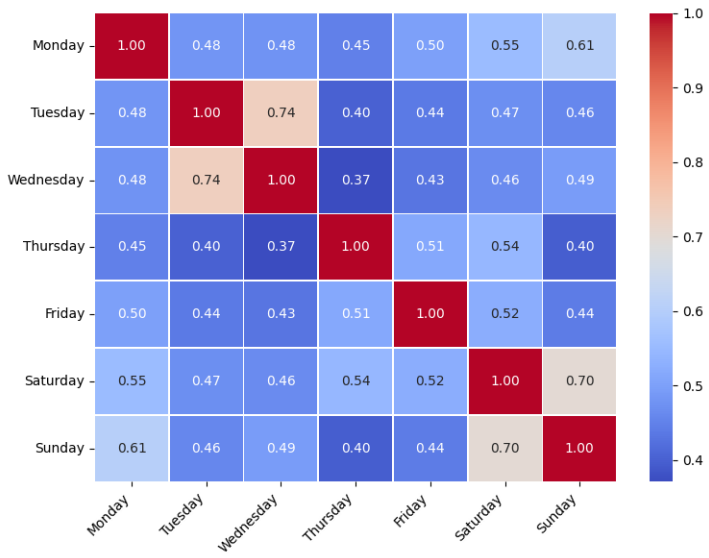

Figure 4.1: Heatmap of correlation on days of the week

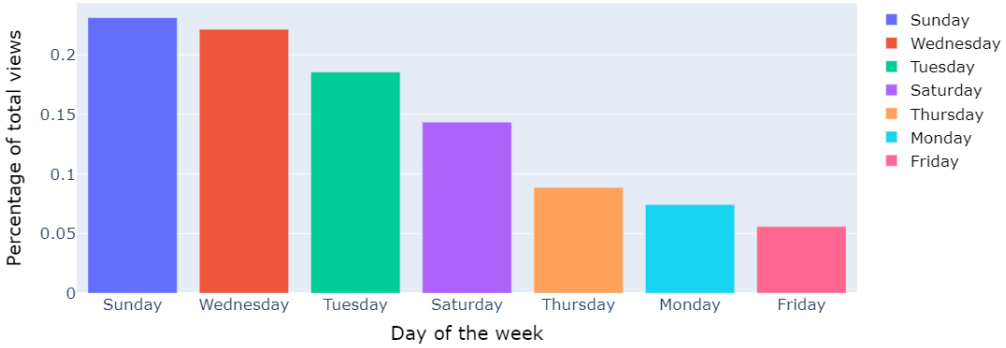

Figure 4.2: Barplot of most watched days of the week

4.1.2 Sport

Item similarity

Item similarity was calculated using the trained ALS model, identifying the 20 most similar items for each item in the dataset. This enabled the capturing of the item-item similarity of every sports item in the dataset. The resulting histogram of item similarity, depicted in Figure 4.3, reveals a highly right-skewed distribution. This suggests that most sports items in the dataset are not very similar to one another, indicating significant variation in their characteristics.

To gain further insights, the item similarity was visualized using a heatmap (Figure 4.4) and a dendrogram with three hierarchical clusters (Figure 4.5), based on a pairwise distance matrix. The heatmap and dendrogram showcased specific pairs of sports items that exhibited higher similarity scores. Notably, ”E-sport" demonstrated a high similarity with ”Programmer”, ”Svømming" showed similarity with ”Tennis”, and ”Cheerleading" exhibited similarity with ”Friidrett”.

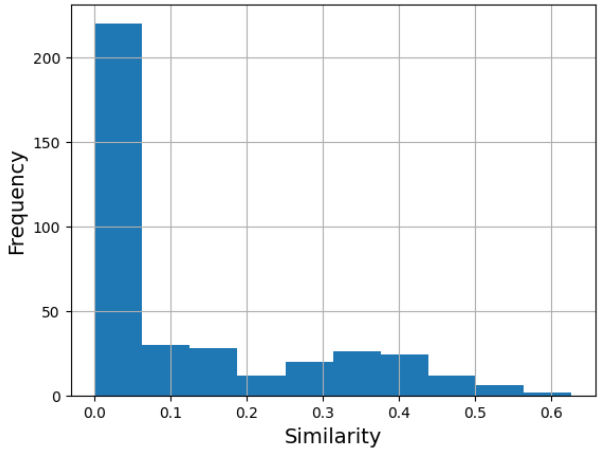

Figure 4.3: Histogram of similarity on sport

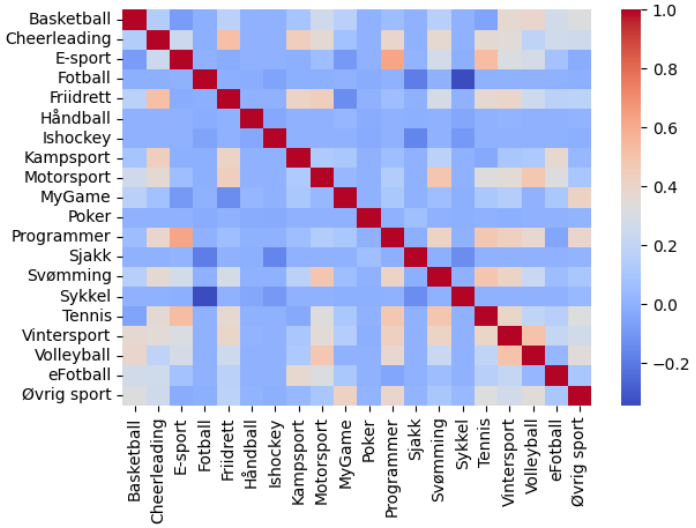

Figure 4.4: Heatmap of similarity on sport

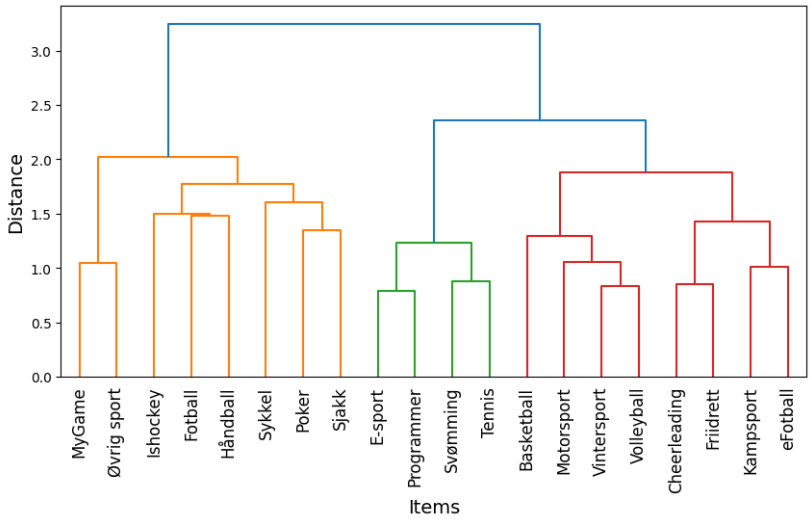

Figure 4.5: Dendogram of similarity on sport

Majority vote

Since the datasets did not include predefined explicit favorites for users, the majority vote approach was employed to determine implicit favorites and second favorites for each user. This involved calculating the watch time in seconds for each sport and identifying the sport with the highest watch time as the first favorite, and the sport with the second-highest watch time as the second favorite. The distribution of users with different sports as their first favorite is illustrated in Figure 4.6, while Figure 4.7 showcases the distribution of second favorites.

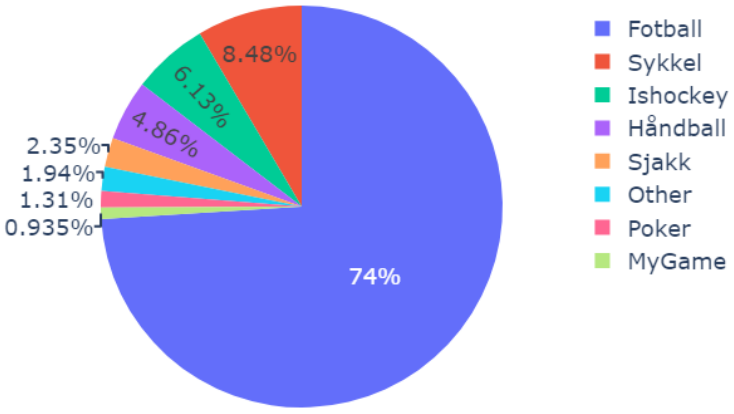

Figure 4.6: Pie chart of favorite sport

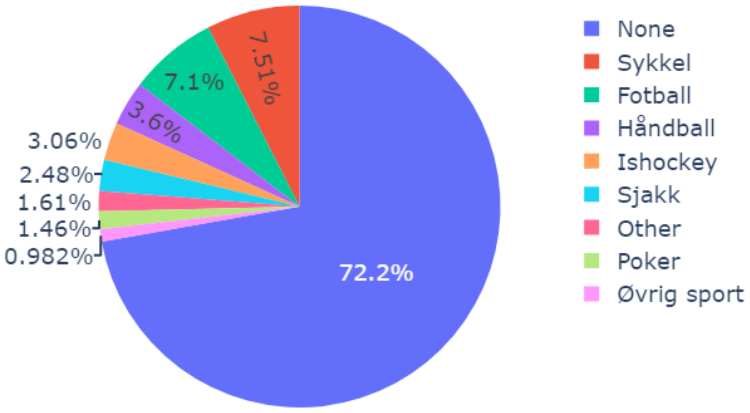

Figure 4.7: Pie chart of second favorite sport

The pie chart presented in Figure 4.6 reveals that football holds the top spot as the

most preferred sport, with over 73% of users having it as their favorite. Following ”football", the preferences decline for sports like ”sykkkel", ”ishockey", and ”håndball". The remaining sports exhibit considerably lower levels of preference among users. Conversely, the pie chart representing second favorite sports in Figure 4.7 demonstrates that over 72% of users have chosen "None" as their second favorite. This indicates that these users have primarily watched only one sport, which is their implicit favorite. This observation suggests that a significant portion of users either exclusively focus on a single sport or have limited engagement and explore fewer sports.

Clusters

To cluster the users, K-means were performed using Scikit-Learn library. The algorithm was used to group similar users together into 8 distinct clusters. The number of clusters was determined using the Elbow curve (see figure 4.8), which suggested that 8 was an optimal number. The data contained 20 dimensions, which made it difficult to visualize. Therefore, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data. The PCA produced two principal components that were used to create a two-dimensional scatter plot, which was used to visualize the clustering results as shown in figure 4.9. There is a lot of overlap between the likes of cluster 2 and cluster 6 which is expected as the similarity between those users is high.

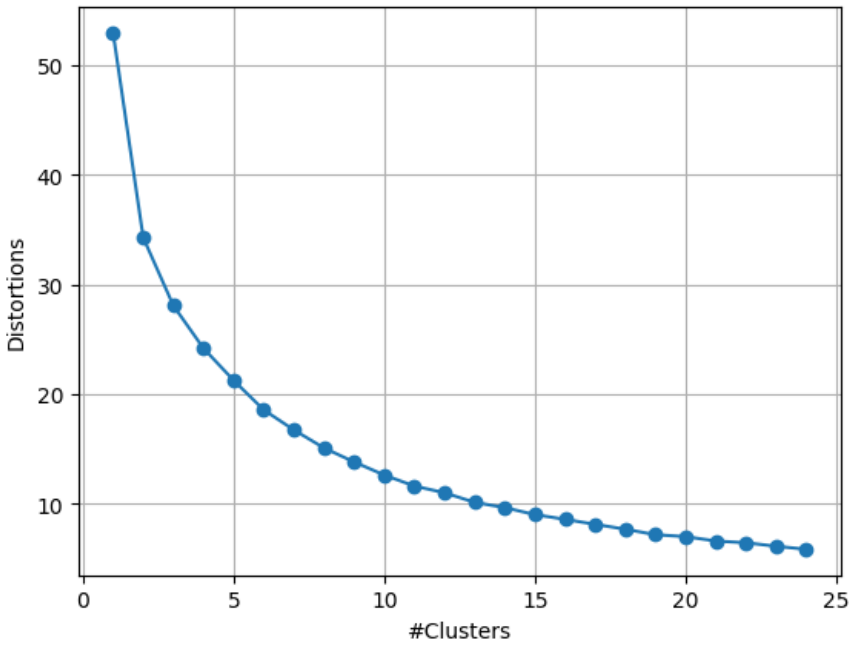

Figure 4.8: Elbow curve for distortion between clusters

To explore the K-Means clusters, each row in Table 4.1 provides a breakdown of each of the 8 clusters. Additionally, the table presents the favorite and second favorite sports obtained for the users through the majority vote approach, along with the average viewing time. The favorite sports indicate the clusters' preferred sports, and the average viewing time provides insights into the users' activity levels, which impact the assigned cluster. The amount of users in each cluster is represented as a percentage of the whole dataset.

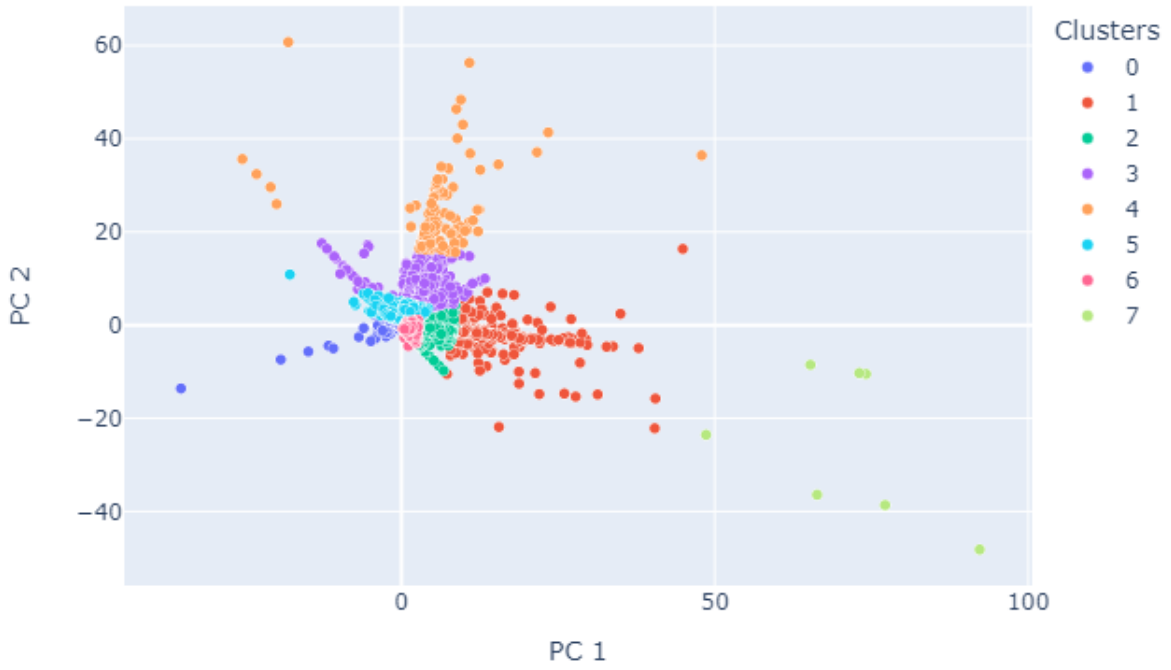

Figure 4.9: KMeans clustering users with 2 principle components

Table 4.1: Overview of clusters with additional information

Cluster	Amount of users in % of total users	Favourites (% of users in cluster)	Second Favourites (% of users in cluster)	Average viewing time on favourites
Cluster0	69.1%	Football (67.9%) Sykkel (10.8%)	None (78.6%) Football (7.6%)	1.53 h 0.49 h
Cluster1	2.4%	Ishockey (97.7%) Fotball (0.6%)	None (42.1%) Fotball (34.5%)	15.1 h 1.68 h
Cluster2	1.5%	Fotball (99.9%) Håndball (0.1%)	None (38.4%) Sykkel (24.7%)	48.39 h 0.51 h
Cluster3	0.8%	Sykkel (99.4%) Sjakk (0.2%)	Fotball (44.0%) None (34.0%)	38.29 h 1.73 h
Cluster4	0.3%	Sjakk (98.5%) Sykkel (0.6%)	Fotball (41.9%) None (26.5%)	52.25 h 3.39 h
Cluster5	19.4%	Fotball (97.2%) Sykkel (1.2%)	None (65.1%) Sykkel (12.5%)	9.58 h 0.45 h
Cluster6	6.5%	Fotball (99.6%) Sjakk (0.1%)	None (52.1%) Sykkel (18.5%)	23.05 h 0.35 h
Cluster7	0.1%	Fotball (99.5%) Sykkel (0.5%)	None (32.0%) Sykkel (29.1%)	118.2 h 4.92 h

As indicated in the majority vote, ”fotball" is the most watched item by a distance. This is also reflected in the table, where it appears as the favorite in most of the clusters. However, further analysis of the K-Means clusters reveals more insight. For instance, while both cluster 2 and cluster 5 are dominated by ”fotball", cluster 2 has more active users who spend more time watching sports than cluster 5. Other clusters like cluster 1 are dominated by "ishockey”, while cluster 3 is dominated by ”sykkel".

The clusters help identify the different types of users existing in the datasets. Some clusters, like cluster 7 have a small number of users who are highly interested in specific sports, representing really passionate users. In contrast, other clusters, like cluster 0 consist of a larger group of users with more diverse viewing habits, indicating a broader range of interests beyond specific sports.

4.1.3 Tournament

Item similarity

The calculation of item similarity for tournament items was performed similarly to the item similarity of sports items. However, compared to the sports dataset, the tournament dataset has many more features, with 195 unique tournaments. This means that the visualization will be presented slightly differently. A histogram of the item similarity is shown in Figure 4.10. The distribution is heavily right-skewed, meaning that a majority of the items are not very similar. However, some items reach close to 1 in similarity. Compared to sports items, there are many more similar items in the tournament dataset. This is also shown in the heatmap in Figure 4.11 and as a dendrogram with three hierarchical clusters based on pairwise distance matrix in Figure 4.12.

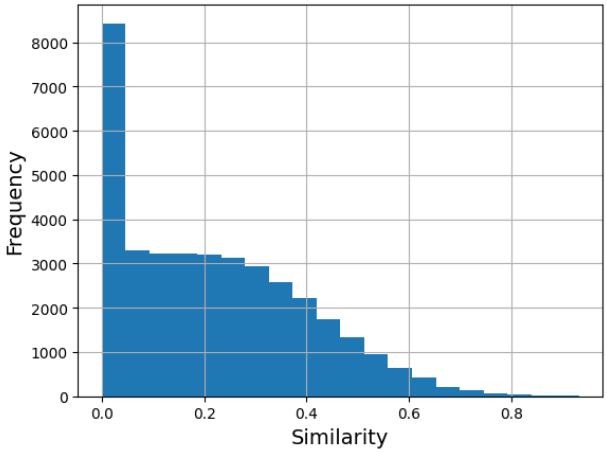

Figure 4.10: Histogram of similarity on tournaments

The dataset includes several tournaments that exhibit high similarity. For instance, the heatmap in Figure 4.11 highlights the pair of items with the highest similarity score of 0.949, namely “Regionscupen G15” and “TrønderEnergi-serien Gutter 16”. These are two age-specific handball tournaments, which explains their similarity.

The dendrogram also shows that many tournaments have short pairwise distances. This makes sense given that many of the tournaments are age-specific tournaments within the same sport (as shown in Figure 4.11), and other tournaments where the same teams participate in different tournaments.

A combination of these figures exhibits a similarity pattern where most items are not very similar but are more spread than the similarity between sports items. The most similar items also show very similar tendencies.

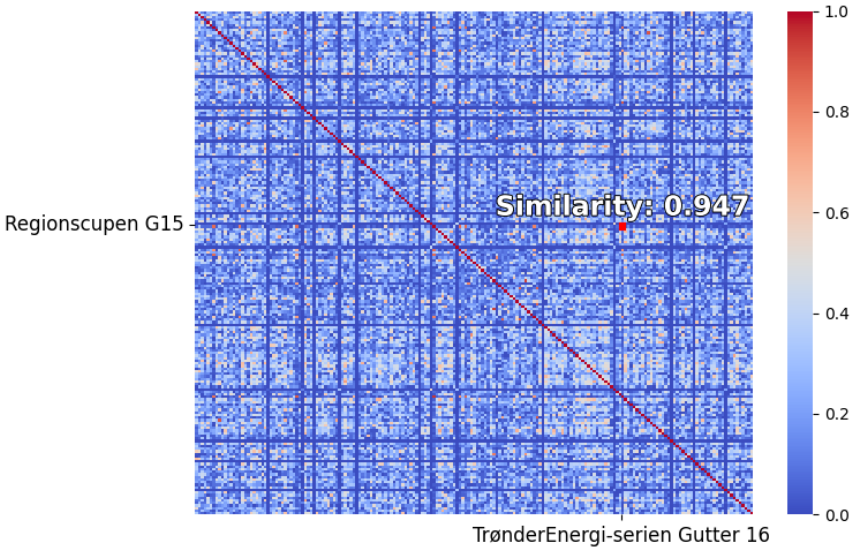

Figure 4.11: Heatmap of similarity on tournaments

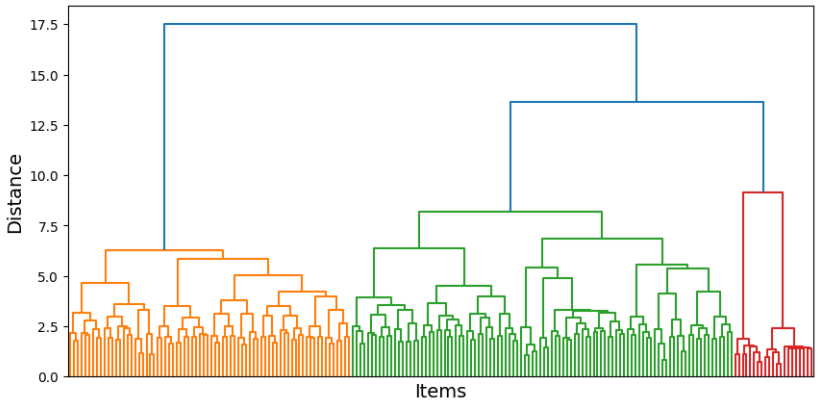

Figure 4.12: Dendogram of similarity on tournaments

Majority vote

A majority vote approach was also utilized for tournaments to define implicit favorites and second favorites for each user. Figure 4.13 displays the distribution of the number of users that have different tournaments as their first favorites, while Figure 4.14 shows the distribution of second favorites. The two favorites were determined by the amount of watch time of each tournament by users.

The favorites were as expected mostly football tournaments, with ”UEFA Champions league" dominating at 35.5%. ”ishockey" and ”håndball" were the other sports with a tournament making an appearance in the 8 most popular.

In terms of second favorite tournaments, the majority of users had only watched one tournament, resulting in the highest percentage being attributed to ”None" at 43.8%. The other most popular tournaments were largely consistent with the favorites. Compared to the majority vote for sports, the percentage of users having ”None" as their second favorite is considerably lower, indicating a tendency among users to concentrate on a single sport. This suggests that a large part of users tend to focus on a single sport and watch multiple tournaments within that sport. There is still a relatively high

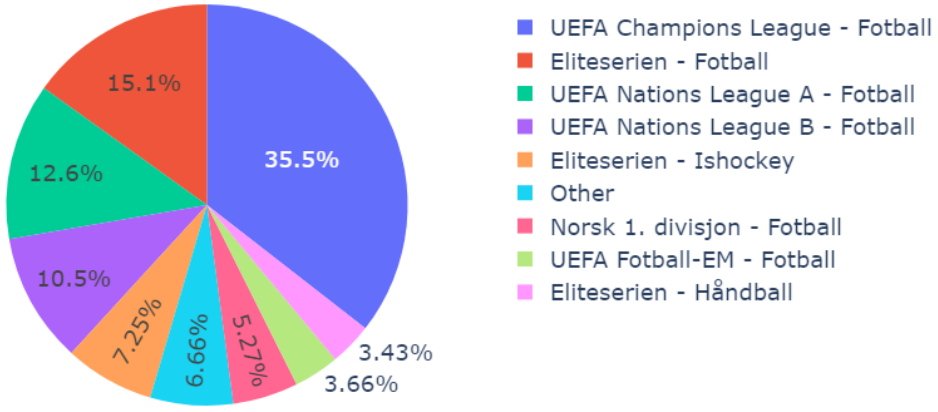

Figure 4.13: Pie chart of favorite tournaments with corresponding sport

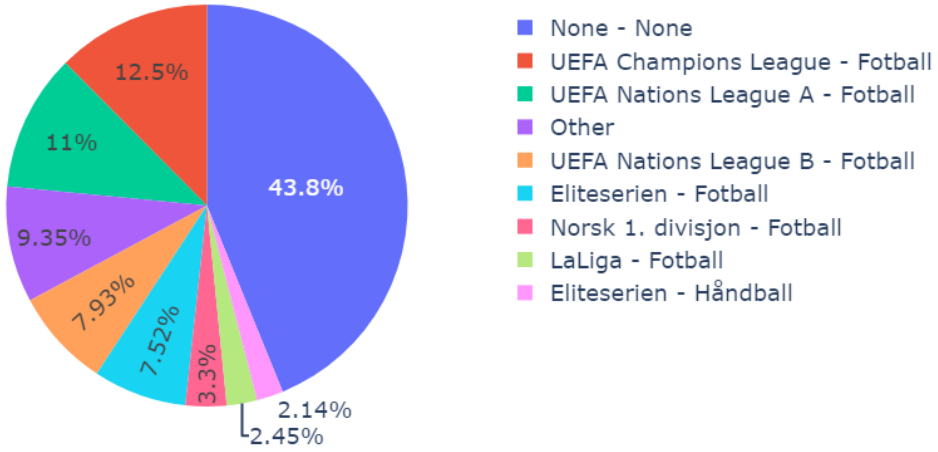

Figure 4.14: Pie chart of second favorite tournaments with corresponding sport

percentage of users in the dataset who have only watched one tournament. This further emphasizes the presence of a significant proportion of less active users in the dataset.

Clusters

To cluster the users on tournaments, K-means were performed using the Scikit-Learn library. The algorithm was used to group similar users together into 9 distinct clusters. The number of clusters was determined using the Elbow curve (see figure 4.15), which suggested that 9 was an optimal number. The data contained 195 dimensions, which made it difficult to visualize the clusters. Therefore, Principal Component Analysis

(PCA) was applied to reduce the dimensionality of the data. The PCA produced two principal components that were used to create a two-dimensional scatter plot, which was used to visualize the clustering results as shown in figure 4.16.

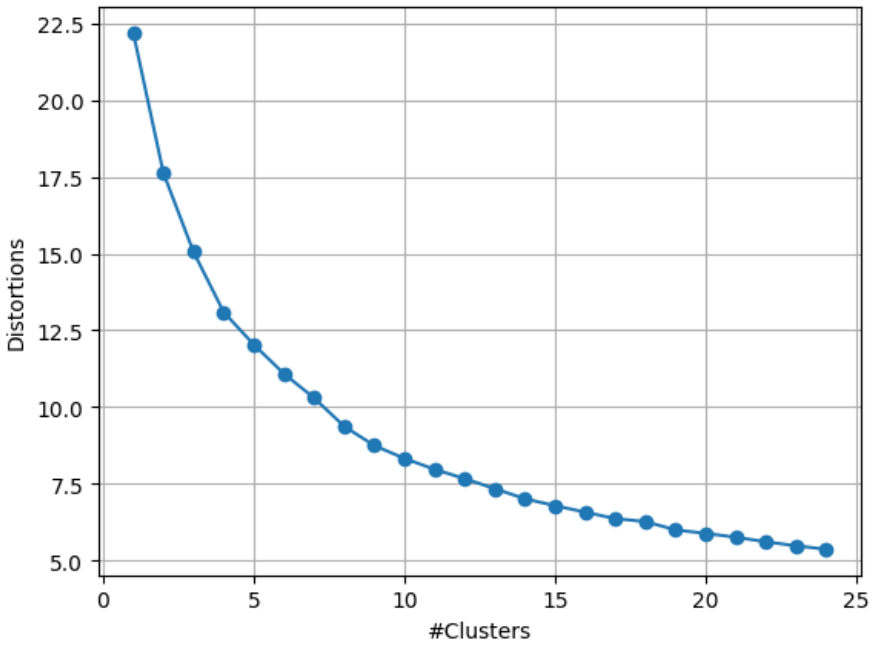

Figure 4.15: Elbow curve for distortion between clusters

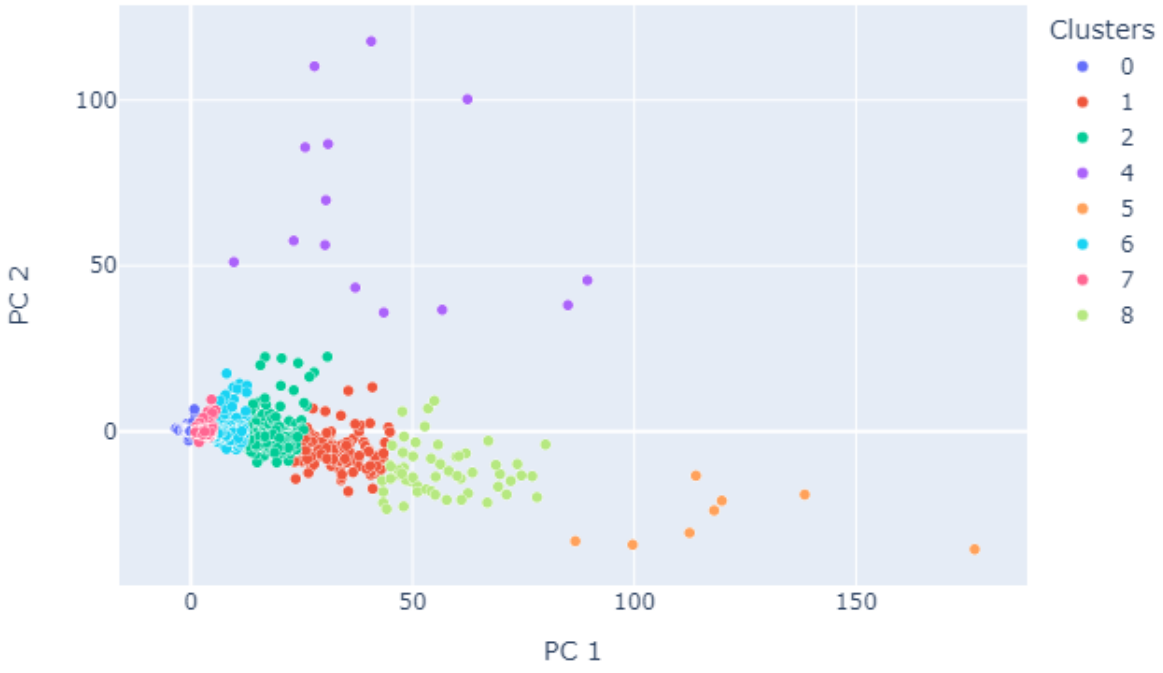

Figure 4.16: KMeans clustering users with 2 principle components

Table 4.2: Overview of clusters with additional information

Cluster	Amount of users in % of total users	Favourites (% of users in cluster)	Second Favourites (% of users in cluster)	Average viewing time on favourites
Cluster0	14.38%	UEFA Champions League - Fotball (92.75%) Eliteserien - Fotball (2.02%)	UEFA Nations League A - Fotball (23.72%) None (22.43%)	5.96h 0.68h
Cluster1	70.36%	UEFA Champions League - Fotball (28.21%) UEFA Nations League B - Fotball (14.51%)	None (54.82%) UEFA Champions League - Fotball (9.75%)	0.58h 0.34h
Cluster2	4.72%	Eliteserien - Fotball (97.31%) UEFA Champions League - Fotball (1.66%)	UEFA Champions League - Fotball (46.12%) UEFA Nations League A - Fotball (11.87%)	11.15h 2.64h
Cluster3	2.89%	Eliteserien - Ishockey (99.07%) 1. divisjon, menn - Ishockey (0.23%)	None (41.16%) UEFA Champions League - Fotball (9.02%)	14.12h 0.27h
Cluster4	0.001%	Eliteserien - Fotball (100%)	UEFA Champions League - Fotball (66.67%) Norsk 1. divisjon - Fotball (33.33%)	524.93h
Cluster5	0.95%	Eliteserien - Fotball (96.61%) Norsk 1. divisjon - Fotball (1.60%)	UEFA Champions League - Fotball (46.48%) Norsk 1. divisjon - Fotball (23.74%)	30.51h 5.22h
Cluster6	2.75%	UEFA Nations League A - Fotball (88.76%) UEFA Champions League - Fotball (4.67%)	UEFA UEFA Champions League - Fotball (33.98%) UEFA Nations League B - Fotball (24.33%)	9.20h 2.83h
Cluster7	1.69%	Norsk 1. divisjon - Fotball (96.70%) Eliteserien - Fotball (1.47%)	UEFA Champions League - Fotball (37.90%) Eliteserien - Fotball (27.14%)	13.63h 2.53h
Cluster4	2.23%	UEFA Champions League - Fotball (92.79%) UEFA Nations League A - Fotball (3.71%)	UEFA Nations League A - Fotball (43.17%) Eliteserien - Fotball (22.53%)	16.23h 5.05h

To explore the K-Means clusters, each row in Table 4.2 provides a breakdown of each of the 9 clusters. Additionally, the table presents the favorite and second favorite obtained for the users through the majority vote approach, along with the average viewing time. The favorite tournament indicates the clusters' preferred tournaments, and the average viewing time provides insights into the users' activity levels, which impact the assigned cluster. Cluster 1 is the largest cluster with 70.36 percent of the users, and represents less active users. The average viewing time of their favorites as well as the second favorite of the users being "None", indicated that this is a cluster made mostly of not-very-active users. Cluster 2 and Cluster 5 consist of viewers that mostly watch "Eliteserien", but with different degrees of activity in terms of viewing time. Cluster 3 consists mostly of active "ishockey" viewers, while cluster 4 consists of few but very active "Eliteserien" watchers.

4.2 Experiment B: Offline evaluation

Experiment B is divided into two sections, 4.2.1 Sport and 4.2.2 Tournament. Each of these sections will cover the grid search, evaluation metrics, and beyond evaluation metrics utilized when conducting the offline evaluation of the algorithms.

4.2.1 Sport

Grid search

To determine the best parameters for the ALS model in context of sports recommendations, a grid search was conducted. The grid search explored various values for the factors, iterations, and regularization parameters. Specifically, the factors parameter was varied with values of 2, 4, 6, 8, 12, 16, 29, 24, and 32, the iterations parameter was tested with values of 10, 20, 30, 40, and 50, and the regularization parameter was experimented with values of 0.1, 0.01, and 0.001.

The outcomes of the grid search are represented in Figure 4.17, which depicts a heatmap showcasing the Precision@5 accuracy measures for different combinations of hyperparameters. The y-axis represents the number of iterations, while the x-axis

displays the factors and regularization values. Analysis of the heatmap reveals that a regularization value of 0.1 consistently outperformed the alternatives of 0.01 and 0.001 across most scenarios. Notably, the combination of 0.1 regularization, 30 iterations, and 32 factors yielded the highest precision among all the tested hyperparameter combinations for the ALS model.

However, considering the dataset's limited size of only 20 sports items, employing 32 factors would be impractical as it could lead to adverse effects such as overfitting. In agreement with the industry partner, the number of factors was therefore adjusted to 4 while maintaining 32 iterations and a regularization value of 0.1, resulting in a more suitable configuration for the model.

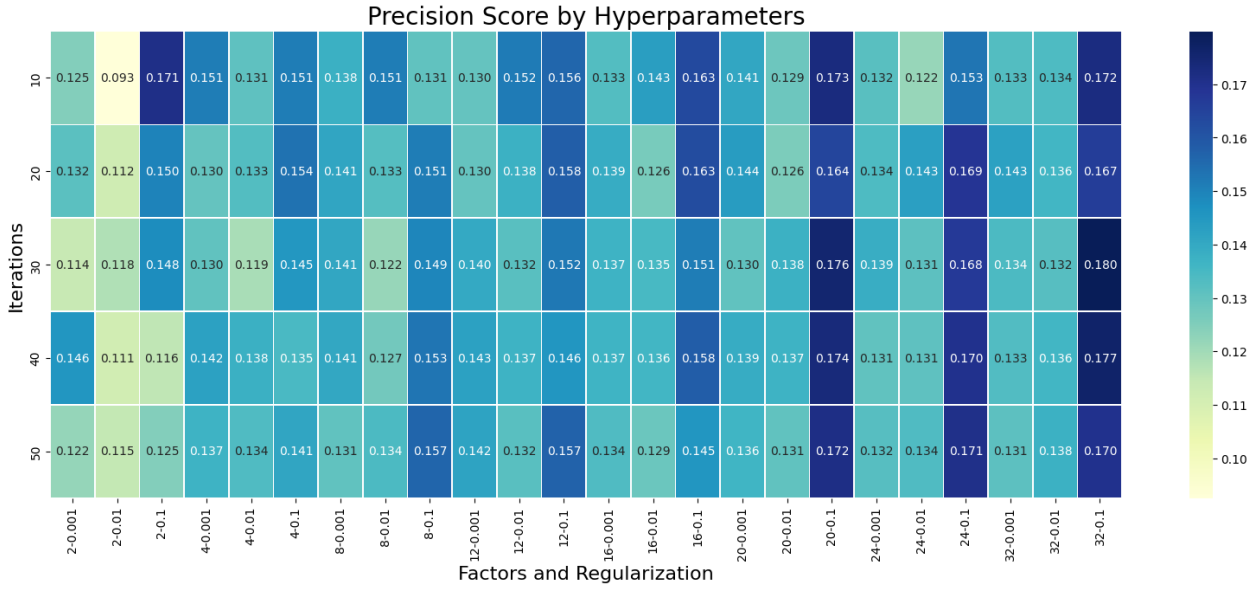

Figure 4.17: Heatmap of result from grid search on sports

Evaluation metrics

The quality of recommendations for sports items was measured using several accuracy metrics. Shown In table 4.4 are the accuracy results for the main ALS algorithm compared to four baseline models.

Table 4.3: Evaluation metrics sports

Evaluation Metrics	Random	Popularity	ALS	BPR	LMF
P@10	0.054	**0.102**	0.091	0.061	0.102
R@10	0.532	**0.996**	0.892	0.600	0.996
AP@10	0.198	**0.732**	0.617	0.370	0.461
NDCG@10	0.275	**0.797**	0.684	0.423	0.590
Hit@10	0.538	**0.997**	0.894	0.607	0.996
RR@10	0.201	**0.735**	0.620	0.374	0.463
ROC_AUC	0.501	**0.938**	0.836	0.565	0.880
PR_AUC	0.230	**0.733**	0.623	0.397	0.461

As shown in table 4.3, the random model, as expected, has the lowest performance across all the metrics. However, the popularity model surprisingly outperforms the CF models (ALS, LMF, BPR) in every metric. While it is expected that the popularity model, which recommends popular items to all users without considering their individual preferences, would score high in metrics like P@10 and Hit@10, it is unexpected that it would excel in all metrics. CF models, leveraging user behavior patterns and item features, are typically expected to perform better in metrics such as R@10, NDCG@10, or ROC_AUC by providing users with more relevant and tailored recommendations (*Kluver et al.*, 2018). However, as shown in *Ji et al.* (2020), recommendations based on popularity have been proven effective for users with limited interactions with a system. This finding is supported by the exploratory analysis, which revealed that many users in the dataset had a relatively low number of interactions.

When it comes to the P@10 of the CF models, ALS had a score of 0.091 and LMF had a score of 0.102 outperforming BPR which had a score of 0.061. Much the same goes for R@10, where ALS and LMF achieve high scores with 0.892 and 0.996, respectively, while BPR lags with only 0.600.

Contrary to expectations, BPR also achieves a lower NDCG@10 score of 0.423 compared to ALS (0.684) and LMF (0.590). This is unexpected considering that BPR is a ranking-oriented model that should theoretically outperform ALS and LMF in ranking-related metrics.

Given that the catalog consists of only 20 sports items, it is expected that most models would achieve a high score in terms of catalog coverage. As expected, the ALS, LMF, and Random algorithms achieved a perfect score of 1, indicating that they recommended all items in the catalog at some point. The BPR fell slightly behind with a score of 0.95, meaning that it failed to recommend one of the items from the catalog. Most popular only got a score of 0,5 which is expected as it only recommends the 10 most popular items from the catalog every time.

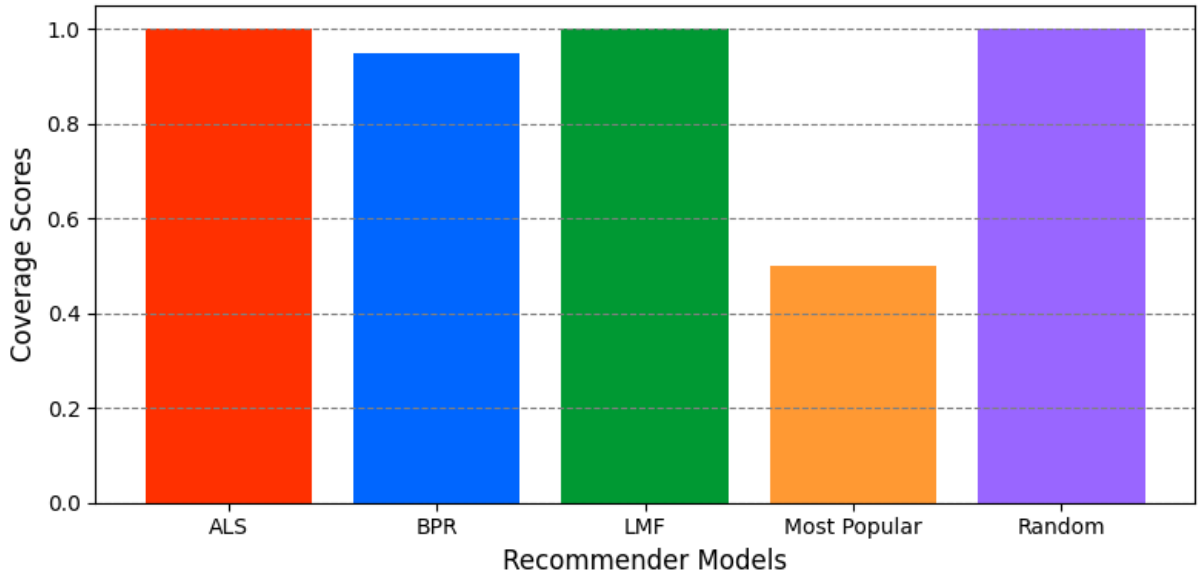

Figure 4.18: Catalog coverage of sports

Overall, the results indicated that ALS and LMF are the two best-performing models for personalized recommendations. Although LMF outperforms ALS in terms of

P@10, R@10, Hit@10, and ROC_AUC by a small margin, ALS performed better in the remaining metrics, including catalog coverage. Popularity, on the other hand, performs well across the board but lacks personalization in its recommendations.

The unexpected dominance of the popularity model across all accuracy metrics does however raise questions about the dataset or evaluation methodology employed. It suggests a potential bias in the dataset towards popular items, which aligns with the findings from the exploratory analysis of sports in section 4.1.2, revealing certain sports to be overwhelmingly dominant.

4.2.2 Tournament

Grid search

To determine the optimal parameters for the ALS model in the context of tournament recommendations, a grid search was performed. The grid search involved exploring various values for the factors, iterations, and regularization parameters. Specifically, the factors parameter was tested with values of 2, 4, 6, 8, 12, 16, 29, 24, and 32, the iterations parameter was examined with values of 10, 20, 30, 40, and 50, and the regularization parameter was experimented with values of 0.1, 0.01, and 0.001.

The results of the grid search are depicted in Figure 4.19, which presents a heatmap illustrating the Precision@5 accuracy measures for different combinations of hyperparameters. The y-axis represents the number of iterations, while the x-axis displays the factors and regularization values. Analysis of the heatmap indicates that a regularization value of 0.1 consistently outperformed the alternatives of 0.01 and 0.001 across most scenarios. Notably, among all the tested hyperparameter combinations for the ALS model, the combination of 0.1 regularization, 50 iterations, and 2 factors yielded the highest precision.

Considering that the tournaments dataset consists of 195 unique items, which is significantly larger than the sports dataset, it was determined that utilizing only 2 factors may not be sufficient. Therefore, in collaboration with the industry partner, it was agreed to adjust the number of factors to 8 while keeping the iterations at 50 and the regularization value at 0.1.

Evaluation metrics

In this experiment, the quality of tournament item recommendations was evaluated using extensive accuracy metrics. Table 4.4 presents the accuracy results for the main ALS algorithm compared to four baseline models.

As shown in the evaluation metrics for tournaments (table 4.4), there are a lot of similarities to the evaluation metrics of sports presented in table 4.3. As expected, also in this table the random model performs the poorest across all metrics. The popularity model, also here achieves the highest score in every metric. This recurring dominance of the popularity model raises further questions about the datasets or evaluation methodology employed.

When considering P@10, Popularity, ALS, and LMF demonstrate relatively similar performance, with Popularity achieving the highest score of 0.110, followed closely by LMF at 0.109, and ALS at 0.104. BPR lags significantly behind with a score of

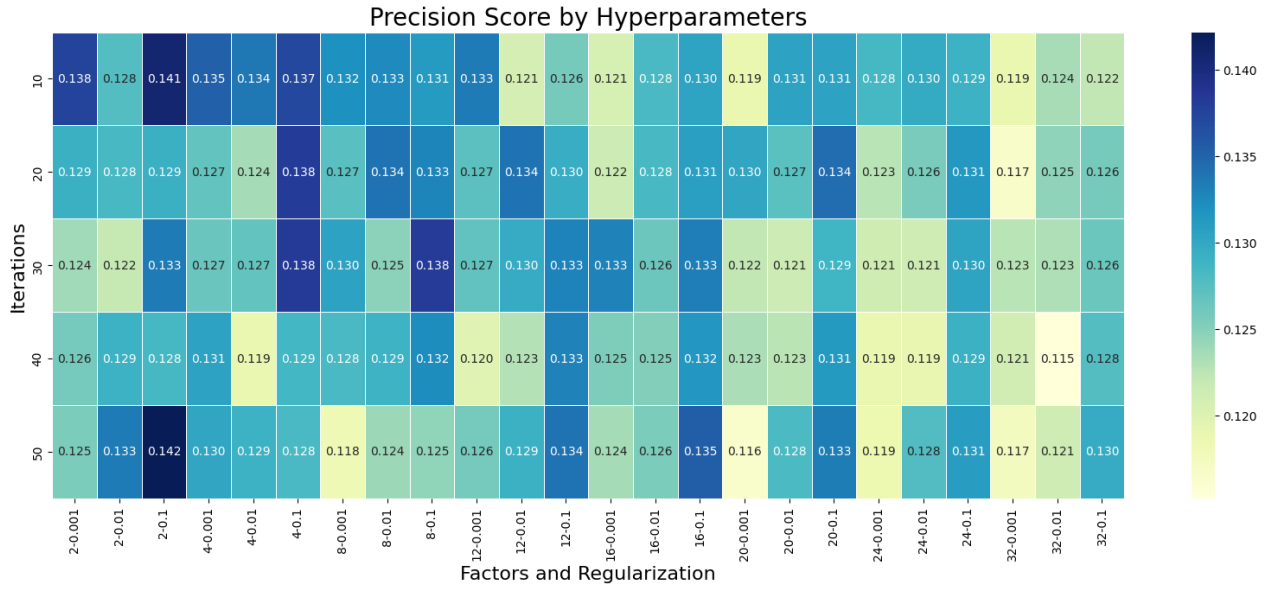

Figure 4.19: Heatmap of result from grid search on tournaments

Table 4.4: Evaluation metrics tournaments

Evaluation Metrics	Random	Popularity	ALS	BPR	LMF
P@10	0.003	**0.110**	0.104	0.047	0.109
R@10	0.030	**0.909**	0.862	0.403	0.904
AP@10	0.005	**0.576**	0.521	0.305	0.360
NDCG@10	0.011	**0.664**	0.611	0.341	0.492
Hit@10	0.037	**0.931**	0.887	0.450	0.929
RR@10	0.006	**0.607**	0.550	0.345	0.377
ROC_AUC	0.501	**0.981**	0.936	0.608	0.969
PR_AUC	0.022	**0.583**	0.529	0.313	0.367

0.047, while the random model performs the worst with a score of 0.003. A similar pattern emerges with R@10, where Popularity, LMF, and ALS exhibit comparable scores. Popularity leads with a score of 0.909, closely followed by LMF at 0.904 and ALS at 0.862. BPR trails significantly behind at 0.403, while the random model has the lowest score of 0.030.

Also here, BPR scores lower on the ranking-based metric NDCG@10, than the other CF models, with a score of 0.341. ALS and Popularity achieve the highest scores at 0.611 and 0.664, respectively, while the random model has the lowest score at 0.011.

Regarding the metrics AP@10, RR@10, and PR_AUC, Popularity achieves the highest score, closely followed by ALS. BPR and LMF exhibit somewhat lower scores than the other two, with LMF slightly outperforming BPR.

Given the larger catalog of tournament items, consisting of 195 compared to the sports catalog with 20 items, it is anticipated that the models will exhibit lower overall catalog coverage scores on tournaments. As illustrated in Figure 4.20, the random algorithm achieves the highest score of 0.78. This is in line with expectations since the random algorithm suggests items randomly, eventually covering a significant portion of the catalog. The BPR model achieves a score of 0.69, while the ALS and LMS models

perform similarly, scoring 0.36 and 0.33, respectively. As anticipated, the most popular algorithm has the lowest score of 0.05, as it only recommends the top 10 popular items in each instance.

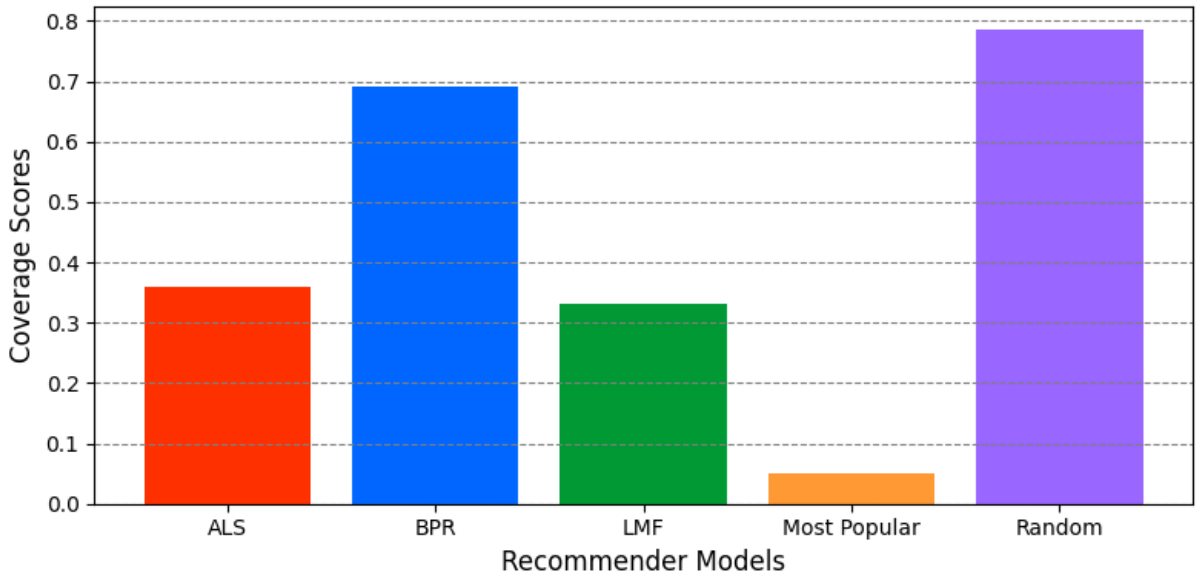

Figure 4.20: Catalog coverage of tournaments

Overall, the results of the evaluation metrics for tournaments align with the tendencies observed in the sports evaluation metrics, deeming ALS and LMF as the most suitable models for the recommendation of tournaments. However, in terms of where LMF outperforms ALS, the margin is less significant in the evaluation metrics for tournaments compared to the evaluation metrics of sports. ALS still comfortably outperforms LMF in metrics such as AP@10, NDCG@10, RR@10, and PR_AUC, while also achieving a slightly better score in catalog coverage. Popularity consistently performs well across all metrics, but it does not provide personalized recommendations, as evident in the catalog coverage metric.

4.3 Experiment C: Online experiments

This section is divided into two parts. The first part focuses on Experiment 1, which involves personalization of sports in the "Sport" feed on TV 2 Play. The second part covers Experiment 2, which revolves around the recommendation of upcoming sports events in the "Sport de neste 7 dagene" feed on TV 2 Play.

Experiment 1: Online experiment 1: Recommendation of sport

To evaluate the quality of experiment 1, an A/B testing was conducted over a 14-day period, from March 30th to April 12th. The model utilized the user's viewing history to generate personalized recommendations through CF, resulting in an ordered list presented to the user (an example of which can be seen in Figure 4.21). The results of the experiment are displayed in Figure 4.22, comparing the baseline approach to the CF model in terms of views, clicks, and click-through-rate

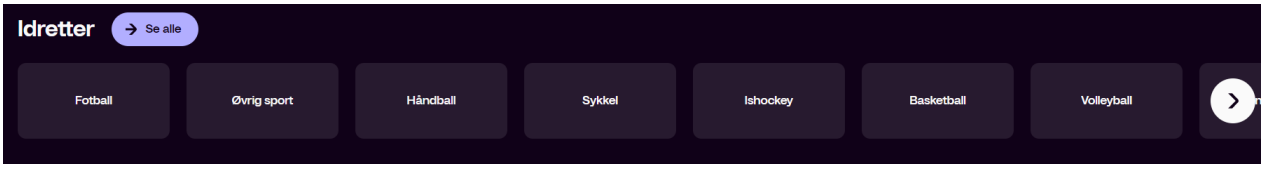

Figure 4.21: Screenshot from TV 2 Play showing the kind of sports feed a user would be presented with in online experiment 1

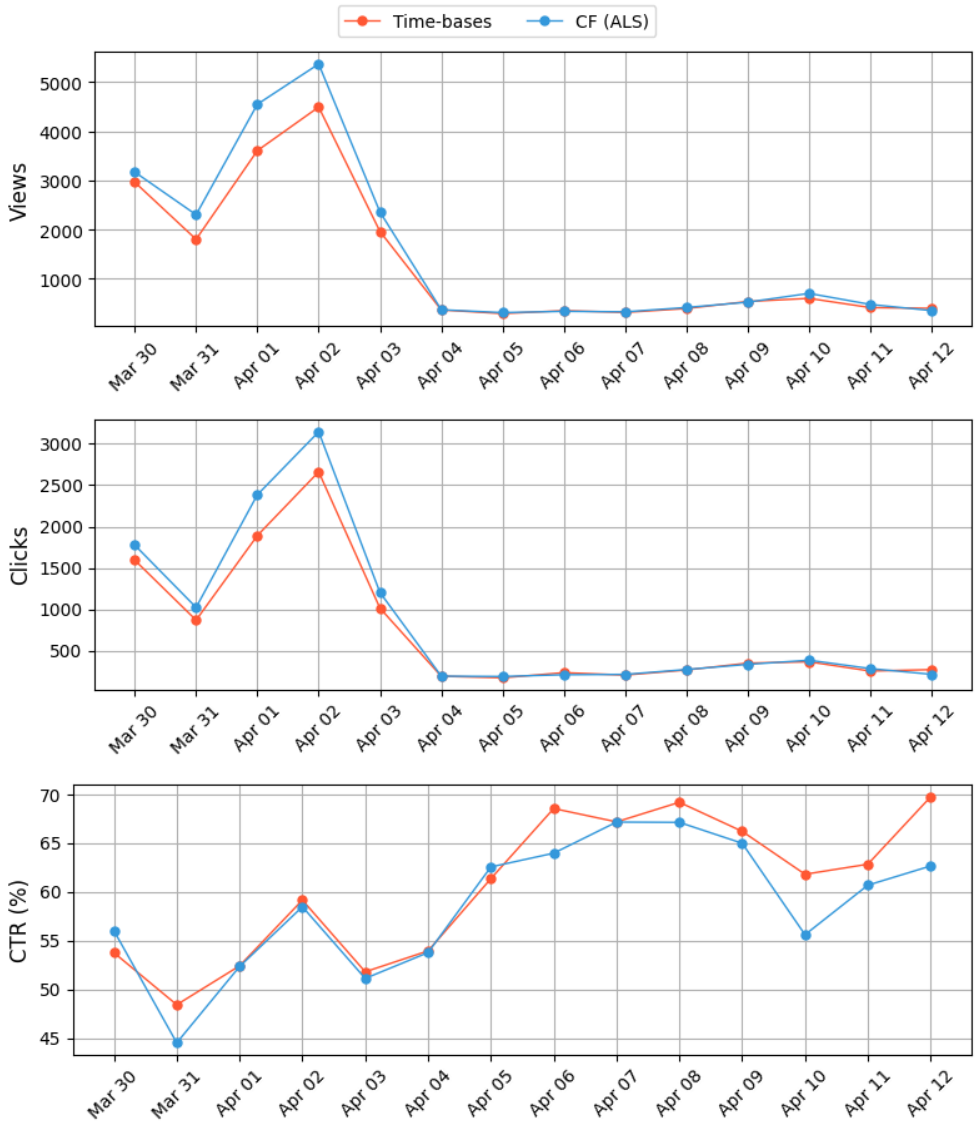

Figure 4.22: Plot from the 14-day online experiment

Table 4.5: Statistics from online experiment 1

Metric	CF (ALS)	Time-based
Views	219,394	199,551
Clicks	79,996	78,259
CTR	37,48%	39,22%

In terms of performance, the strong baseline slightly outperforms the CF approach.

As shown in Figure 4.22, the difference between the two approaches in terms of CTR is minimal on most days. The day with the largest discrepancy is April 10th, which coincides with the start of the Norwegian football season. This could have influenced the results, as many users may have visited the website to find the football games, thus increasing the click-through rate on the baseline approach which would present the football item first as it is next up in the schedule.

Although the CF approach underperforms slightly in comparison to the baseline in terms of click-through rate, it is worth noting that there may be additional benefits that are not immediately apparent from this limited analysis. Specifically, the model has the potential to offer a more diverse range of recommendations to users, thereby potentially introducing them to previously unknown sports. This represents a potentially valuable contribution to user engagement and satisfaction, but a more comprehensive analysis of these potential benefits is precluded by the limitations of the available data.

Furthermore, it is worth noting that the click-through rates of the CF approach and the baseline are quite close, with the CF achieving almost 37.5% and the baseline a little over 39%. While this suggests that the baseline approach is currently a viable option, the potential benefits of a more diverse list of recommendations cannot be discounted. Further exploration of this approach could give valuable insights into user preferences and behaviors, ultimately leading to a more personalized and satisfying user experience.

Online experiment 2: Recommendation of upcoming sports events the next seven days

To assess the quality of online experiment 2, an A/B testing was conducted on TV 2 Play over a span of 16 days, from May 12th to May 28th, 2023. As explained in section 3.5.2, the personalized recommendations for this experiment were generated using the ALS model. The ALS model scored all upcoming sports events for the following seven days and filter out the ones under a certain threshold, considering the sports, tournaments, and teams of each item.

A notable distinction between the personalized recommendations from the ALS approach and the implicit favorite baseline, is the number of items presented to the user. The ALS approach tends to provide a larger selection of items, often including those also recommended by the implicit favorite approach. Figure 4.23 illustrates an example of ALS recommendations, while Figure 4.24 shows an example of recommendations from the implicit favorites approach. The results of the experiment are depicted in Figure 4.25, comparing the baseline approach to the collaborative filtering model in terms of views, clicks, and click-through rate.

Regarding performance, the views displayed in Figure 4.25 and the total views presented in Table 4.6 illustrated that the ALS approach increases the visibility of the feed to a larger user base, resulting in a higher view count. However, it performs less favorably in terms of CTR, achieving 14.88% compared to the baseline with 19.12%. Figure 4.25 reveals that May 13th and May 16th were the days with the most active users in terms of views and clicks. These days coincided with matches involving most of the Norwegian football teams in the top tier, as well as some big Spanish football teams. Therefore it is unsurprising that the implicit favorite approach performed better these days in terms of CTR, as it naturally recommended users' implicit favorites, and football enthusiasts tend to have preferred teams or matches that they are inclined

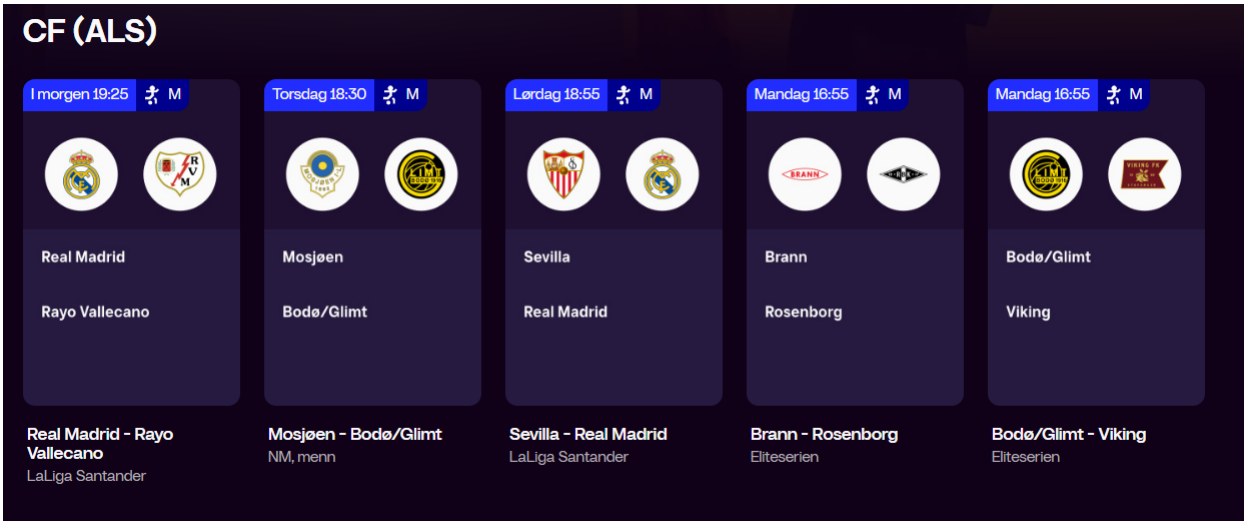

Figure 4.23: Screenshot from TV 2 Play showing the kind of feed a user would be presented with if they were in the B group of online experiment 2

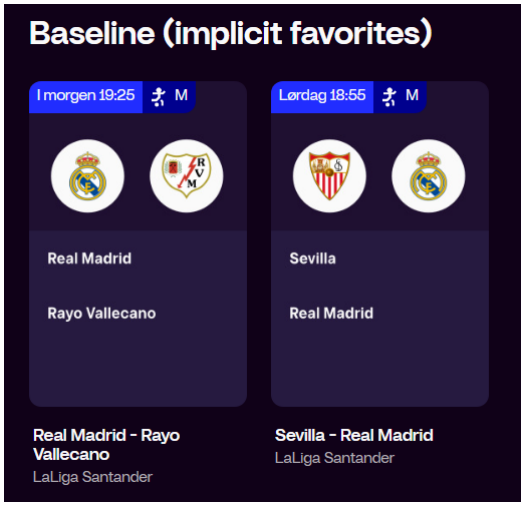

Figure 4.24: Screenshot from TV 2 Play showing the kind of feed a user would be presented with if they were in the A group of online experiment 2

to watch.

Table 4.6: Statistics from online experiment 2

Metric	CF (ALS)	Implicit favorites
Views	44,767	35,281
Clicks	6,661	6,745
CTR	14,88%	19,12%

Although the ALS approach underperforms compared to the baseline in terms of CTR, it did increase the visibility of the feed in terms of views with a large margin, generating almost 10.000 more views. Similar to experiment 1, it's also important here to recognize that there may be additional benefits not immediately apparent from this limited analysis. Specifically, the ALS model has the potential to offer users a more diverse range of recommendations, potentially introducing them to new and previously

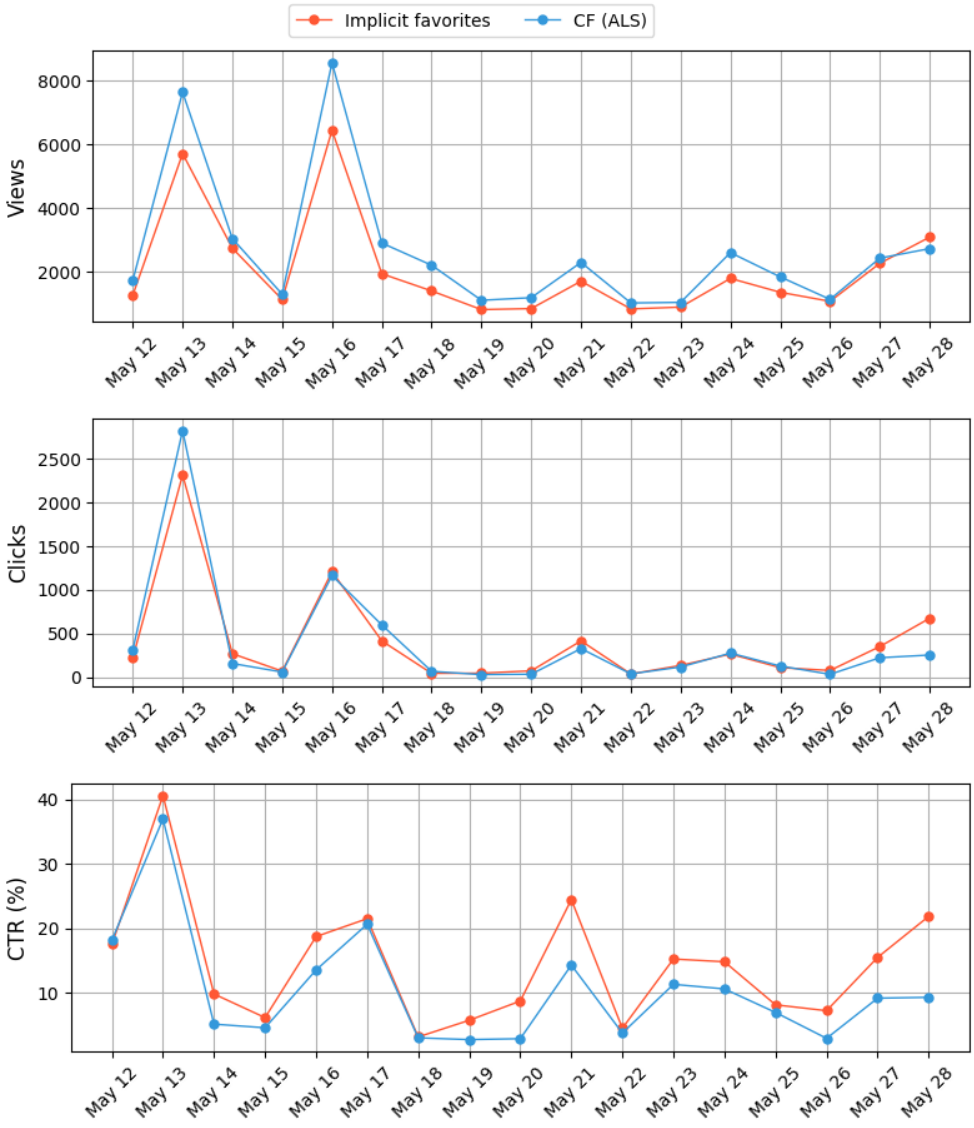

Figure 4.25: Plot from the 16-day online experiment

unknown sports, tournaments, or clubs. However, also in this case, a more comprehensive analysis of these potential benefits is hindered by the limitations inherent in the available data.

Chapter 5

Conclusion and Future Work

This chapter provides an overview of the thesis by discussing its main contributions, results, limitations, and future work. The chapter is divided into four sections. Section 5.1 summarizes the research carried out in the thesis. Section 5.2 discusses the key contributions of the thesis. Section 5.3 presents the results obtained based on the research questions set out. Finally, Section 5.4 highlights the limitations of the thesis and discusses potential directions for future research.

5.1 Summary

In this thesis, a collaborative filtering technique has been employed to generate recommendations for users both on upcoming live sports events, as well as recommendations of pure sports. The research follows a structured approach, encompassing key steps such as literature review, data analysis, model development, offline evaluation, and A/B testing.

The literature review in Chapter 2 establishes the research context and presents the state-of-the-art in the field. It provides an overview of existing methodologies and challenges, laying the foundation for the research.

An extensive analysis of the data sets provided by TV 2 was conducted, covering important aspects of the data. This exploratory analysis uncovered valuable insights into the underlying characteristics and patterns within the data, serving as a basis for informed model development.

The offline evaluation of the developed model is presented in Chapter 4, comparing its performance against four baseline models. Various evaluation metrics are used to assess the effectiveness of the models, enabling meaningful comparisons.

Two A/B tests were conducted on the streaming platform TV 2 Play, giving real-world data on the performance of the approaches. Chapter 3 provides more detailed information on how the different approaches were constructed, and what factors were considered in their creation. The thesis concludes with an evaluation of the approaches' online performance, which is elaborated in Chapter 4, incorporating the insights and outcomes derived from the A/B tests.

5.2 Main contributions

This thesis advances the field of live sport recommender systems in the following ways:

- *Proposing a novel sports-based collaborative recommendation technique based on user viewing sessions of sports content:* Throughout Chapter 3, different approaches and considerations when recommending sports content were addressed. Two resulting approaches were created utilizing the ALS technique. The first approach uses the ALS to personalize and sort the sports feed on TV 2 Play. The second approach, not only considers sports but also incorporates the ALS to consider tournaments and teams. In this approach, the ALS is used more as a filtering tool, which filters out items if their score from the model is under a certain threshold.

- *A comprehensive offline evaluation of the proposed recommendation approach, including comparisons with different baselines on accuracy and beyond accuracy metrics:* Chapter 3 outlines the evaluation metrics employed to measure the performance of the ALS model compared to a set of baseline models. Chapter 4 presents a thorough evaluation of recommendation quality, incorporating both offline evaluation and online experiments along with exploratory analysis. In evaluating recommendation quality, accuracy and beyond accuracy metrics were utilized to compare the ALS model to the baseline models. The baseline models consist of a mixture of other collaborative models such as LMF and BPR, as well as a most popular model and a random model. The exploratory analysis delved deeper into the characteristics of both users and items. The items were examined for similarity using histograms, dendrograms, and heatmaps. The popularity distribution of items was further investigated using a majority vote to identify favorite and second favorite items. The users were explored using K-means clustering, with the elbow method used to determine the optimal number of clusters. Additionally, PCA was employed to visualize user similarity.

- *Developing and deploying a sports-based collaborative recommender on one of Norway's largest digital streaming platforms (TV 2 Play) for A/B testing:* Chapter 3 proposes the implementation and experiment design of the two different online experiments. Their performance is further evaluated in Chapter 4. The first experiment was deployed on TV 2 Play and A/B was tested between March 30th and April 12th, while the other was deployed between May 12th and May 28th. In the first experiment, 50 percent of users in the A/B testing were presented with items sorted based on the presented approach, while the other group was presented with TV 2's purely time-based list of items. In the second experiment, 50 percent of the users were presented with TV 2's implicit favorites approach, while the other 50 percent were presented with the CF approach, considering sports, tournaments and teams. In both experiments, the quality of the systems has been evaluated with real-world data with an A/B test on TV 2 Play, where they are measured in terms of views, clicks and CTR.

5.3 Conclusion

In conclusion, this thesis has addressed the challenge of personalizing sports recommendations in real time, taking into account the live nature of sports events. Extensive exploratory analysis was conducted on a dataset provided by TV 2, one of Norway's largest media companies, to propose a collaborative filtering recommender system. The system was evaluated both offline and online through an A/B test.

The results of the offline experiments in Section 4.2.1 showed that the ALS and LMF models outperformed the other baselines when trained on sports data. The popularity model scored the highest in terms of accuracy, but when taking into account the beyond accuracy model, the catalog coverage, it always recommended the 10 most popular items, failing to enable users to discover new content. The LMF was outperformed by the ALS in terms of AP@10, TAP@10, NDCG@10, PR@10, ROC_AUC, and PR_AUC, while both models covered all items of the small catalog when looking at the catalog coverage. Even though the performance of both models was close, the ALS was deemed the most suitable model compared to the baselines.

Regarding the models trained on tournament data as shown in Section 4.2.2, a similar conclusion can be drawn. The ALS and LMF were also the best-performing in terms of accuracy, apart from the popularity model. The P@10 and R@10 on ALS and LMF had less difference, while the ALS still outperformed the LMF in the same metrics as when the models were trained on the sports data. The ALS also slightly outperformed the LMF in terms of catalog coverage, deeming it the most suitable for the models trained on tournament data as well. As a conclusion of RQ 1, the offline evaluation demonstrated that, in this setting, the ALS is the most suitable collaborative filtering method compared to the other baselines.

Regarding RQ 2, the factors that influence the user's preference for specific types of live sports events were proposed in Chapter 3.4, as well as detailing how these were incorporated into online experiment number 2 in Section 3.4.2. The results presented in Section 4.3, show that the experiment underperforms compared to the implicit favorites baseline in terms of CTR. It did, however, have a positive effect on increasing the visibility of the feed, reaching a larger user base. This outcome highlights the potential of the approach despite its underperformance. With the limited available analysis data, it would be premature to completely dismiss this approach. Fine-tuning the weighting of tournaments and teams, or adjusting the threshold configuration, could also potentially enhance the model's performance.

The results of the online experiment 1, looking purely at personalization and recommendation of sports, are detailed in the first part of Section 4.3. The results show that the performance is very similar in terms of CTR, but more data would be needed to see if the diversity is better.

5.4 Limitations and future work

The present study has certain limitations and provides opportunities for future research. For instance, a larger dataset collected over a longer period of time could offer deeper insights into user behavior and patterns, ultimately improving the accuracy of the models trained on this data.

Furthermore, while the first and second online experiments provided valuable insights into user views, clicks, and click-through rates, there are additional factors that could be explored to gain a more comprehensive understanding of user engagement. For instance, conducting further research to analyze diversity and other relevant factors may uncover new insights and broaden the scope of the analysis. By considering these additional dimensions, a deeper and more comprehensive understanding of user behavior can be achieved, opening up insights that have the potential to improve the systems.

The inclusion of teams in the online experiment was a point of uncertainty and was not thoroughly explored in the exploratory analysis. Further research could emphasize more focus on this, and potentially discover new valuable insight on this aspect that further could improve the way teams are incorporated, enhancing its impact on the approach.

The online experiments, especially experiment number 2 has a lot of room for further exloration, by experimenting with different weights considering sports, tournament, or experimenting with different approaches to calculating the drop-off, this approach could have the potential to perform much better.

In Chapter 3, the grid search conducted to optimize the hyperparameters of our recommender system provided valuable insights before deploying the system for A/B testing. However, it is important to acknowledge that additional experimentation with the hyperparameters could have provided further benefits. These types of experiments are inherently time-consuming and computationally expensive. In future research, it could be worthwhile to conduct a more comprehensive grid search, exploring a wider range of hyperparameters and potentially considering alternative parameters.

Bibliography

Aggarwal, C. C. (2016), *Recommender Systems - The Textbook*, 1-498 pp., Springer. 1.1, 2.1, 2.1.1

Albanese, M., A. d'Acierno, V. Moscato, F. Persia, and A. Picariello (2013), A multimedia recommender system, *ACM Transactions on Internet Technology (TOIT)*, *13*(1), 1–32. 2.2

Balog, K., F. Radlinski, and S. Arakelyan (2019), Transparent, scrutable and explainable user models for personalized recommendation, in *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 265–274. 2.1.2

Belete, D., and M. D H (2021), Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results, *International Journal of Computers and Applications*, *44*, 1–12, doi:10.1080/1206212X.2021.1974663. 3.3

Bennett, J., and S. Lanning (2007), Matrix factorization techniques for recommender systems, *Netflix Prize Documentation*. 3.3

Bobadilla, J., F. Ortega, A. Hernando, and A. Gutiérrez (2013), Recommender systems survey, *Knowledge-based systems*, *46*, 109–132. 2.1

Burke, R. (2002), Hybrid recommender systems: Survey and experiments, *User modeling and user-adapted interaction*, *12*, 331–370. 2.1.1

Burke, R. (2017), Multisided fairness for recommendation, *arXiv preprint arXiv:1707.00093*. 2.1.2

Cantador, I., A. Bellogín, and D. Vallet (2010), Content-based recommendation in social tagging systems, in *Proceedings of the fourth ACM conference on Recommender systems*, pp. 237–240. 2.1.1

Cañamares, R., P. Castells, and A. Moffat (2020), Offline evaluation options for recommender systems, *Information Retrieval Journal*, *23*, 387–410. 2.5

Chapelle, O., D. Metlzer, Y. Zhang, and P. Grinspan (2009), Expected reciprocal rank for graded relevance, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, p. 621–630, Association for Computing Machinery, New York, NY, USA, doi:10.1145/1645953.1646033. 3.5.1

Cheng, M., and P. Liu (2017), *Privacy Aspects of Recommender Systems*, pp. 1246–1256, Totem Publisher. 2.5

Cheng, Z., X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli (2019), Mmalfm: Explainable recommendation by leveraging reviews and images, *ACM Trans. Inf. Syst.*, *37*(2), doi:10.1145/3291060. 2.1.1

Deldjoo, Y., M. Elahi, M. Quadrana, and P. Cremonesi (2015), Toward building a content-based video recommendation system based on low-level features, in *E-Commerce and Web Technologies: 16th International Conference on Electronic Commerce and Web Technologies, EC-Web 2015, Valencia, Spain, September 2015, Revised Selected Papers 16*, pp. 45–56, Springer. 2.1.1

Ding, J., Y. Wang, Q. Wang, and Y. Cao (2017), Football video recommendation system with automatic rating based on user behavior, in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5, doi:10.1109/CISP-BMEI.2017.8301933. 2.6

Elahi, M., M. Braunhofer, T. Gurbanov, and F. Ricci (2018), User preference elicitation, rating sparsity and cold start. 2.1, 2.1.1

Elahi, M., D. Jannach, L. Skjærven, E. Knudsen, H. Sjøvaag, K. Tolonen, Ø. Holmstad, I. Pipkin, E. Throndsen, A. Stenbom, et al. (2021), Towards responsible media recommendation, *AI and Ethics*, pp. 1–12. 1.1, 2.1.2, 2.2

Elahi, M., D. Jannach, L. Skjærven, E. Knudsen, H. Sjøvaag, K. Tolonen, Ø. Holmstad, I. Pipkin, E. Throndsen, A. Stenbom, et al. (2022), Towards responsible media recommendation, *AI and Ethics*, pp. 1–12. 2.1.2

Elahi, M., D. K. Kholgh, M. S. Kiarostami, M. Oussalah, and S. Saghari (2023), Hybrid recommendation by incorporating the sentiment of product reviews, *Information Sciences*. 2.1.1

Fernández, M., A. Bellogín, and I. Cantador (2021), Analysing the effect of recommendation algorithms on the amplification of misinformation, *arXiv preprint arXiv:2103.14748*. 2.2

Friedman, A., B. P. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky (2015), *Privacy Aspects of Recommender Systems*, pp. 649–688, Springer US, Boston, MA, doi:10.1007/978-1-4899-7637-6_19. 2.1.2

Ge, M., C. Delgado-Battenfeld, and D. Jannach (2010), Beyond accuracy: Evaluating recommender systems by coverage and serendipity, RecSys '10, p. 257–260, Association for Computing Machinery, New York, NY, USA, doi:10.1145/1864708.1864761. 3.5.1

Ge, Y., S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, et al. (2021), Towards long-term fairness in recommendation, in *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 445–453. 2.1.2

Gomez-Uribe, C. A., and N. Hunt (2016), The netflix recommender system: Algorithms, business value, and innovation, *6*(4), doi:10.1145/2843948. 2.2

Gunawardana, A., G. Shani, and S. Yogev (2022), *Evaluating Recommender Systems*, pp. 547–601, Springer US, New York, NY, doi:10.1007/978-1-0716-2197-4_15. 2.5

Hazrati, N., and M. Elahi (2021), Addressing the new item problem in video recommender systems by incorporation of visual features with restricted boltzmann machines, *Expert Systems*, *38*(3), e12,645. 2.1.1

He, H., and Y. Ma (2013), Imbalanced learning: Foundations, algorithms, and applications, *Imbalanced Learning: Foundations, Algorithms, and Applications*, doi: 10.1002/9781118646106. 3.5.1, 3.5.1

Hu, Y., Y. Koren, and C. Volinsky (2008), Collaborative filtering for implicit feedback datasets, in *2008 Eighth IEEE International Conference on Data Mining*, IEEE. 3.2, 3.2

Jannach, D., M. Zanker, A. Felfernig, and G. Friedrich (2010), *Recommender Systems: An Introduction*, Cambridge University Press, doi:10.1017/CBO9780511763113. 1.1, 2.1, 2.1.1

Ji, Y., A. Sun, J. Zhang, and C. Li (2020), A re-visit of the popularity baseline in recommender systems, in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 1749–1752, Association for Computing Machinery, New York, NY, USA, doi: 10.1145/3397271.3401233. 4.2.1

Johnson, J. (2014), Logistic matrix factorization for implicit feedback data, in *NIPS Workshop on Distributed and Large Scale Machine Learning*. 3.2

Kim, N.-r., S. Oh, and J.-H. Lee (2018), A television recommender system learning a user's time-aware watching patterns using quadratic programming. 2.4, 2.6

Kirshenbaum, E., G. Forman, and M. Dugan (2012), A live comparison of methods for personalized article recommendation at forbes.com, in *Machine Learning and Knowledge Discovery in Databases*, edited by P. A. Flach, T. De Bie, and N. Cristianini, pp. 51–66, Springer Berlin Heidelberg, Berlin, Heidelberg. 2.1

Klimashevskaia, A., M. Elahi, D. Jannach, C. Trattner, and L. Skjærven (2022), Mitigating popularity bias in recommendation: Potential and limits of calibration approaches, in *Advances in Bias and Fairness in Information Retrieval: Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers*, pp. 82–90, Springer. 2.1.2

Kluver, D., M. D. Ekstrand, and J. A. Konstan (2018), *Rating-Based Collaborative Filtering: Algorithms and Evaluation*, pp. 344–390, Springer International Publishing, Cham, doi:10.1007/978-3-319-90092-6_10. 4.2.1

Koren, Y. (2008), Factorization meets the neighborhood: A multifaceted collaborative filtering model, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, p. 426–434, Association for Computing Machinery, New York, NY, USA, doi:10.1145/1401890.1401944. 3.3

Koren, Y., S. Rendle, and R. Bell (2021), Advances in collaborative filtering, *Recommender systems handbook*, pp. 91–142. 2.1.1

Kuroda, M., Y. Mori, and M. Iizuka (2020), Initial value selection for the alternating least squares algorithm, in *Advanced Studies in Classification and Data Science*, edited by T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, and M. Vichi, pp. 227–239, Springer Singapore, Singapore. 1.1

Kvifte, T., M. Elahi, and C. Trattner (2022), Hybrid recommendation of movies based on deep content features, in *Service-Oriented Computing–ICSOC 2021 Workshops: AIOps, STRAPS, AI-PA and Satellite Events, Dubai, United Arab Emirates, November 22–25, 2021, Proceedings*, pp. 32–45, Springer. 2.1.1

Milano, S., M. Taddeo, and L. Floridi (2020), Recommender systems and their ethical challenges, *AI & Soc*, *35*, 957–967, doi:10.1007/s00146-020-00950-y. 2.1.2

Patro, S. G. K., and K. K. Sahu (2015), Normalization: A preprocessing stage, *CoRR*, *abs/1503.06462*. 3.1

Pazzani, M. J., and D. Billsus (2007), Content-based recommendation systems, *The adaptive web: methods and strategies of web personalization*, pp. 325–341. 2.1.1

Petander, H. (2019), Personalised recommendations within sport. 1.1, 1.2, 2.3

Rendle, S., C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme (2009), Bpr: Bayesian personalized ranking from implicit feedback, in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press. 3.2

Research, G. V. (2021), Sports technology market size, share & trends analysis report by technology (device, smart stadium, esports), by application (soccer, baseball, basketball, ice hockey), by region, and segment forecasts, 2021 - 2028, *Grand View Research*. 1.1

Rimaz, M. H., M. Elahi, F. Bakhshandegan Moghadam, C. Trattner, R. Hosseini, and M. Tkalčič (2019), Exploring the power of visual features for the recommendation of movies, in *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 303–308. 2.1.1

Rimaz, M. H., R. Hosseini, M. Elahi, and F. B. Moghaddam (2021), Audiolens: audio-aware video recommendation for mitigating new item problem, in *Service-Oriented Computing–ICSOC 2020 Workshops: AIOps, CFTIC, STRAPS, AI-PA, AI-IOTS, and Satellite Events, Dubai, United Arab Emirates, December 14–17, 2020, Proceedings*, pp. 365–378, Springer. 2.1.1

Ringnér, M. (2008), What is principal component analysis?, *Nature biotechnology*, *26*, 303–304. 3.5.1

Sanchez, F., M. Alduan, F. Alvarez, J. M. Menendez, and O. Baez (2012a), Recommender system for sport videos based on user audiovisual consumption, *IEEE Transactions on Multimedia*, *14*(6), 1546–1557, doi:10.1109/TMM.2012.2217121. 2.6

Sanchez, J., A. Moreno, A. Vellido, and S. Grau (2012b), Multimodal recommendation of sports video content, *IEEE Transactions on Multimedia*. 2.6

Schafer, J. B., J. Konstan, and J. Riedl (1999), Recommender systems in e-commerce, in *Proceedings of the 1st ACM conference on Electronic commerce*, pp. 158–166. 2.1

Schedl, M., H. Zamani, C.-W. Chen, M. Elahi, E. Gómez, B. Hidasi, D. Jannach, N. Lathia, C. Musto, T.-N. Nguyen, et al. (2018), Current challenges and visions in music recommender systems research, *International Journal of Multimedia Information Retrieval*, *7*(2), 95–116. 3.5.1, 3.5.1, 3.5.1, 3.5.1

Turrin, R., A. Condorelli, P. Cremonesi, and R. Pagano (2014), Time-based tv programs prediction. 1.1, 1.2, 2.4, 2.6

Wang, S., X. Zhang, Y. Wang, H. Liu, and F. Ricci (2022), Trustworthy recommender systems, *arXiv preprint arXiv:2208.06265*. 2.1.2

Wang, X., Y. Guo, and C. Xu (2015), Recommendation algorithms for optimizing hit rate, user satisfaction and website revenue, in *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, p. 1820–1826, AAAI Press. 3.5.1

Yu, Z., X. Zhou, D. Zhang, C.-Y. Chin, X. Wang, and J. Men (2006), Supporting context-aware media recommendations for smart phones, *IEEE Pervasive Computing*, *5*(3), 68–75. 2.2