From Clicks to Conscious Choices

Investigating the Effects of Carbon Footprint Data in E-Commerce Recommender Systems

FROM CLICKS TO CONSCIOUS CHOICES

INVESTIGATING THE EFFECTS OF CARBON FOOTPRINT DATA IN E-COMMERCE RECOMMENDER SYSTEMS

to obtain the degree of Master in Computer Science with Specialization in Artificial Intelligence Technology at Delft University of Technology.

by

Sneha LODHA

Born in Mumbai, India

Multimedia Computing Group, Faculty of Electrical Engineering, Mathematics and Computer Science (Faculteit Elektroniek, Wiskunde en Informatica), Delft University of Technology, Delft, The Netherlands.

Thesis committee:

Chair, Daily Supervisor	Dr. Elvin Isufi, Faculty EEMCS, TU Delft
Member,	Jie Yang, Faculty of EEMCS, TU Delft
Member,	Chirag Raman, Faculty of EEMCS, TU Delft

Keywords: Recommender Systems, CO_2 Emissions, User Study, Beyond Accuracy

ABSTRACT

One of the contributing factors to climate change is the release of gases, particularly carbon dioxide (CO2), which is amplified by the expanding E-commerce industry. E-commerce enterprises heavily depend on recommender systems as a means to incentivize consumers towards making product purchases. This master's thesis investigates the positive impacts of presenting information regarding carbon dioxide (CO_2) emissions values on user behavior and recommendation accuracy within sustainable recommender systems. Through the creation of a new dataset, CarEmissions, this study explores whether displaying emission values influences sustainable choices in recommendations. Findings demonstrate that recommendation models trained without CO_2 values consistently outperform those with CO_2 values, enhancing both accuracy and greenness. This suggests that the inclusion of CO_2 values introduces variability to user ratings, thereby influencing recommendation outcomes. Furthermore, this research examines correlations between user demographics, knowledge, and ratings, revealing insights into the lack of significant links. Additionally, it assesses the differences in recommendation quality between datasets with and without CO_2 values, highlighting the advantages of omitting CO_2 values in enhancing recommendation performance. While considering the limitations inherent to domain-specific data and convenience sampling, the thesis outlines avenues for refining data collection and exploring automated strategies for balancing recommendation accuracy and greenness. By advancing the understanding of user behavior and ethical considerations in sustainable recommender systems, this study contributes to the evolving landscape of technology-driven sustainable consumption.

Acknowledgements

This thesis was made possible through the support of many people. First and foremost, I wish to express my gratitude to my supervisor, Dr. Elvin Isufi, for his consistent guidance throughout the entire thesis process. Our meetings provided invaluable direction, and he remained consistently available for answering my questions and giving me the necessary feedback. Thank you for putting up with all the ups and downs of this journey with me. I would also like to acknowledge Jie Yang and Chirag Raman for accepting to be part of my thesis committee.

Additionally, thanks to my friends who have shared this same journey with me. Going through this process together has undoubtedly enhanced the entire experience. Finally, I'd like to extend heartfelt thanks to my family, who have been an unwavering source of support throughout my entire educational journey. Thank you for giving me every opportunity which has led me here.

Contents

1

INTRODUCTION

Awareness around climate change has been a significant concern for several decades, with the first published study dating back to 1896 by Arrhenius [1]. With increased globalization over the past years, we as consumers have seen the number of products available to buy skyrocket, but consuming items from the other side of the globe comes at a cost. Every part of the consumption process affects the CO_2 levels, from manufacturing to transportation pollution. In recent years, e-commerce has been on the rise, growing by nearly 7% since 2019 [2]. Despite the convenience and time-saving benefits it offers consumers, the rapid growth of e-commerce has significant environmental repercussions. The increased emissions and pollution resulting from the extensive operations of online retail platforms pose a serious concern. A prominent example is Amazon, the e-commerce giant, which alone accounted for a substantial carbon footprint of 71.54 million metric tons CO2 equivalent in 2021 [3].

Simultaneously, with e-commerce comes the task of recommendation. The listed items feature huge collections from myriads of producers. With this enormous amount of options, it becomes the task of the recommender system to narrow down items of interest to consumers. Recommender systems in e-commerce help turn browsers into buys, promote cross-selling of items, and gain loyalty by creating a personalized value add to the website [4]. Well-curated recommendation systems can cause up to a 35% lift in the number of items purchased by users [5]. We believe this is where we can influence consumers' decisions towards being more conscious about their purchases and, by extension, the environment.

As the awareness of climate change continues to gain prominence globally, more individuals are embracing lifestyle modifications and adopting environmentally conscious consumption patterns in their daily routines. Few studies exist in the domain of green recommender systems, and even fewer tackle the problem at the consumer level. This study investigates the potential influence of disclosing product footprint information on users' perceptions and evaluations of the products. With this data, we then train to evaluate the performance of various recommendation algorithms and use the best to nudge users to make more responsible decisions.

1

In this thesis, the main research question is: **What are the implications of showing product CO2 emission values to users for promoting sustainable choices through recommendations?**. To cover all grounds, we answer the following sub-questions:

1. *Is there a significant relationship between user demographics and knowledge, and ratings provided to items?*
2. *Does the greenness and accuracy of products recommended differ between the two datasets (CO_2 displayed vs. No CO_2) displayed?*

In the initial phase of our research, we develop a comprehensive platform for data collection and dissemination, leading to the creation of the CarEmissions dataset. This dataset holds distinct user-item interactions encompassing items with and without CO_2 information displayed, thereby resulting in three separate datasets, namely CarEmissionsCO_2, CarEmissionsNoCO_2, and CarEmissionsAll. A noteworthy attribute of the dataset is that 10% of its contents consist of identical user-item interactions, wherein one instance includes the display of CO_2 values, and the other does not. This design facilitates a within-subjects study, enabling us to gain novel insights into the impact of CO_2 information disclosure. We compare several widely used recommender systems and user study datasets to ensure the dataset's validity.

Secondly, we analyze the dataset and observe any patterns in user demographic data. We also conduct a lasso regression to determine whether there is any correlation between demographic and product attributes and the rating an item receives. A statistical analysis of the 10% overlap group is also conducted.

Thirdly, we benchmark the accuracy and greenness of the current recommender system algorithms on the new dataset. The algorithms evaluated include global average, user-based knn, item-based knn, SVD, SVD++, co-clustering, and SLIM. SVD is the overarching best-performing algorithm in terms of accuracy and greenness.

Finally, to test if the greenness of recommendations can be increased, we utilize the nudging method of re-ranking recommendations [6]. In making these recommendations as green as possible, we also focus on their effect on the recommendations' accuracy.

This thesis is organized as follows. First Chapter 2 presents the background information. Chapter 3 then investigates the current literature related to our study. In Chapter 4, we discuss the properties of the CarEmissions dataset, and Chapter 5 focuses on the creation, analysis, and correlation of the dataset. Chapter 6 evaluates the greenness of commonly used recommendation algorithms and presents the results of the nudging strategy employed. Finally Chapter 7 concludes the thesis.

2

BACKGROUND

This chapter presents the background knowledge that will be utilized in this thesis. Section 2.1 gives the various recommender algorithms and how they can be evaluated. Section 2.1.3 presents the strategy to nudge towards greener recommendations [6]. Section 2.2 shows the working of group lasso regression, and finally, Section 2.3.1 explains stratified sampling.

2.1. RECOMMENDER SYSTEMS

Giving and receiving recommendations from others is a crucial way we interact. Whether it is a restaurant to dine at, a movie to watch, or a new piece of clothing to buy, we ask the people we trust to guide us through the massive amounts of items available. Recommender systems provide an automated alternative for this [7]. These recommender systems aim to deliver the right content to the right person [8]. Similar to other machine learning tasks, recommendations are inherently predictions. Provided a set of users and a set of items, the goal of the recommender systems is to predict how relevant a specific item would be to a particular user. To train such recommender systems, known user-preference data is required. This user-preference data can fall into two categories, either implicit or explicit. Implicit data is inferred from user behavior such as clicking activity or time spent on a product page [8]. We will use explicit data in this thesis when a user rates an item, resulting in a user-item interaction.

Since no recommendation serves all customers best, several distinct recommendation techniques exist to be explored. These can be classified into two groups, namely rating-based and ranking-based recommender systems.

Rating-based. This goal of rating-based recommendation tasks is to predict the rating a user would give to a particular item [9]. These recommendation tasks learn from known rating data and aim to predict the rating for unseen user-item interactions. After the ratings have been predicted, recommendations are generated for a user by finding the items with the higher predictions.

Ranking-based. Instead of predicting numerical ratings, the goal of ranking-based systems is to predict the ranked lists that are ordered on the predicted utility for the user.

These systems are helpful when the users only consider a limited number of recommendations. Rankings can often be generated by ordering items from highest to lowest based on their predicted rating.

2.1.1. RECOMMENDER SYSTEM ALGORITHMS

Following is the explanation of commonly-used algorithms that will be explored later in the thesis.

Global Mean. The method consists of using the global mean of all ratings in the training set and the prediction of every interaction in the test set. This simplistic strategy is often used as a baseline for comparison.

User/item-knn. Nearest neighbours algorithms utilize similarities between users and items to predict ratings [8]. To predict a rating for each user-item interaction, the k-nearest neighbor algorithm for users identifies the k most similar users to the target user. It then calculates a weighted average based on these users' ratings of the item in question. Conversely, the item-knn algorithm operates similarly but instead focuses on finding k similar items. In Equation 2.1, the formula for a user-based knn prediction for rating $\hat{r}_{u,i}$ is displayed. Here u is the user, i is the item, N is the k most similar users to u, and sim() is a similarity measure. An often-used similarity measure is the cosine similarity [10].

$$\hat{r}_{ui} = \frac{\sum_{v \in N^k(u)} \text{sim}(u,v) \cdot r_{vi}}{\sum_{v \in N^k(u)} \text{sim}(u,v)} \tag{2.1}$$

Single Value Decomposition (SVD). This method assumes that we are given a matrix of ratings **R**, representing users as rows and items as matrix columns. This matrix can then be decomposed into lower dimensional matrices as shown in Equation 2.3 [11], [12]. Here U is an orthogonal left singular matrix, which represents the relationship between users and latent factors, **S** is a diagonal matrix that describes the strength of each latent factor, and **V** is a diagonal right singular matrix, which indicates the similarity between items and latent factors. The SVD decreases the dimension of **R** by extracting its latent features.

$$\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{2.2}$$

When the original matrix is reduced from rank r to rank n, where $n < r$, we can represent the resulting matrix as:

$$\mathbf{R}_n = \mathbf{U}_n\mathbf{S}_n\mathbf{V}_n^T \tag{2.3}$$

The reduced matrix $\mathbf{R}_n$ serves as the closest rank approximation to the original matrix **R** [11]. Now the unknown ratings can be predicted by filling in the missing values in the original matrix using a technique like taking the mean of known ratings. By imputing these missing values, SVD can generate predictions for unknown ratings and provide recommendations based on them. A point of note is that due to data imputation, noise is introduced into the data because these values are treated as actual ratings [13].

Additionally, SVD does not consider the structural biases in the data. An example is when the same rating value carries a different level of importance for users. This might cause some users to give structurally lower or higher ratings than others. These SVD limitations are addressed using the Improved Regularized SVD (IRSVD) in Equation 2.4.

$$r'_u i = b_u + b_i + \mathbf{p}_u^T \mathbf{q}_i \tag{2.4}$$

Here, we remove the need for imputing ratings by factorizing the rating matrix into two latent feature matrices, namely $\mathbf{p}_u^T$ and $\mathbf{q}_i$. The latent matrices of the users and items are learned by minimizing the loss of the generated predictions, and a form of gradient descent is used to find the parameters that optimize the model [14]. The bias terms b_u for user u and b_i for item i are introduced to identify the portion of ratings that individual user or item biases can explain. This thesis uses the IRSVD method for benchmarking purposes. For simplicity, this will be referred to as SVD [15], [14].

SVD++. As SVD only considers the values of known ratings to predict unknown ones, a crucial point of information can be added. A user-rated item (regardless of its value) may also provide useful information [16]. This indicates the item was important enough for the user to interact with.

$$r'_u i = \mu + b_u + b_i + \mathbf{q}_i^T \left(\mathbf{p}_u + |N(u)|^{-\frac{1}{2}} \sum_{j \in N(u)} y_j \right) \tag{2.5}$$

Here the parameters are the same as IRSVD, with new introductions such as μ, which represent the global bias, $N(u)$, which is a set of item that user u has rated, and y_j, which are the latent factors of the items in $N(u)$.

Co-clustering. This collaborative filtering technique is used to group similar users and items based on their ratings [17]. The co-clustering approach is based on the idea that users with similar preferences may also have similar item preferences and vice versa.

$$\hat{r}_{ui} = \overline{C_{ui}} + \left(\mu_u - \overline{C_u} \right) + \left(\mu_i - \overline{C_i} \right) \tag{2.6}$$

Users are assigned to cluster C_u, and items are assigned to cluster C_i and co-clusters C_{ui}. Using these, predictions can be generated using the Equation 2.6, $\overline{C_{ui}}$, $\overline{C_u}$, and $\overline{C_i}$, indicate the average ratings of those clusters.

SLIM. This technique aims to train a linear model that predicts the relevance of items for users based on their interactions and item-item similarities. The key idea is to exploit the sparsity of the user-item matrix to learn a sparse linear model that captures the underlying relationship [18].

$$\hat{r}_{ui} = \mathbf{r}_u^T \mathbf{w}_i \tag{2.7}$$

For user u and item i, $r_{u,i}$is defined in Equation 2.7. Here matrix W needs to be calculated, which is a sparse matrix of aggregation coefficients whose i-th column corresponds $\mathbf{w}_i$ and each row $\mathbf{r}_u^T$ represents the recommendation scores on all items for user u. Norm regularization introduces sparsity into the solution to calculate $\mathbf{W}$ effectively.

2.1.2. RECOMMENDER SYSTEMS EVALUATION

The next step after training recommender systems and receiving their predictions is to assess the quality of the recommendations. A commonly used goal for determining performance is accuracy. Accuracy metrics are an efficient way of computing performance and tuning a model's hyper-parameters. This thesis will use the metrics mentioned in this section to evaluate recommender systems.

RMSE. This metric operates on the rating values produced by the model to measure the accuracy [19]. Equation 2.8 $\hat{R}$ signifies all the predicted ratings. Here the error between the real and predicted value is calculated and squared for each interaction and then divided by the number of interactions. This entire term is then squared.

$$RMSE = \sqrt{\frac{\sum_{\hat{r}_{ui} \in \hat{R}} (r_{ui} - \hat{r}_{ui})^2}{|\hat{R}|}} \tag{2.8}$$

By calculating the RMSE, we obtain the root of the mean of the squared errors between the predicted and the real values. The root of the term is taken to penalize large residual terms.

NDCG. This metric measures the relevance and order of items presented in a predicted ranked list. The NDCG aims to consider both the relevance of items and their positions in the list.

$$DCG = \frac{1}{m} \sum_{u=1}^{m} \sum_{j \in I_u, v_j \leq L} \frac{2^{rel_{uj}} - 1}{\log_2(v_j + 1)} \tag{2.9}$$

Equation 2.9 presents the discounted cumulative gain (DCG) for user u, a set of items I_u, and a predicted recommendation list of length L. The DCG is calculated as the sum of the relevance scores of the items at each position, discounted logarithmically based on their position. The next step is estimating the ideal discounted cumulative gain (IDCG), representing the maximum achievable DCG value for the given list length L. This can be done by calculating the DCG of the list of recommendations ordered by ground-truth rankings. The final normalized discounted cumulative gain is obtained by dividing the DCG with the IDCG (Eq. 2.10).

$$NDCG = \frac{DCG}{IDCG} \tag{2.10}$$

GNDCG. This metric is a recent introduction in the field of green recommender systems [6]. With greenness normalized discounted cumulative gain (GNDCG), we aim to measure the greenness of the recommendation lists produced by algorithms.

$$GDCG = \frac{1}{m} \sum_{u=1}^{m} \sum_{j \in I_u, v_j \leq L} \frac{2^{g_j} - 1}{\log_2(v_j + 1)} \tag{2.11}$$

Equation 2.11 shows the modification performed on DCG. Here v_i is the rank of item i. The numerator indicates the 2^{g_i} where g_i represents the greenness of item i. The greenness of the item is taken as an exponent to place a stronger emphasis on recommending

green recipes higher in the list. Similarly to NDCG, this metric is normalized by dividing the GDCG with IGDCG, which represents the ideal greenness discounted cumulative gain. IGDCG can be found by applying DCG on the list of recommendations ordered by ground truth greenness of items.

2.1.3. NUDGING STRATEGY

This approach proposed by [6] involves a weighting model that considers predictions made for the test set. These predictions are then combined with the greenness value associated with each item during the interaction. The result is a weighted utility value, which considers both the model's predictions and the greenness of the items. The nudging strategy is defined by equation 2.12, which outlines the key components: $\mu_{u,i}$ represents the utility of the interaction, $\hat{r}_{u,i}$ corresponds to the predicted rating for the interaction (u, i), and g_i signifies the greenness value of item i. The parameter α is introduced to control the weighting between these factors. This parameter allows for a flexible balance between the two components. It is worth noting that α can be assigned a value between 0 and 1, which determines how much each factor influences the overall calculation.

$$\mu_{u,i} = \alpha \hat{r}_{u,i} + (1-\alpha) g_i \tag{2.12}$$

The nudging method utilizes predictions generated by the best-performing models. However, only the NDCG and GNDCG metrics will be employed when evaluating this method. This choice is driven by the combination of predictions, and greenness values cannot be directly compared to other approaches that rely solely on ratings.

2.2. GROUP LASSO REGRESSION

Lasso (least absolute shrinkage and selection operator) regression, formally known as L1 regularization, is a popular statistical modeling technique to estimate the relationships between variables [20]. The main idea behind the lasso regression model is to find a balance between the model's simplicity and accuracy. In the case of this thesis, lasso regression will be used for feature selection.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{2.13}$$

The lasso model starts with a standard linear regression, which assumes a linear relationship between independent variables [21]. Equation 2.13 depicts linear regression where y is the dependent variable, the β values are the parameters to be estimated, and the x values are the independent variables.

$$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{p} (|\beta_j|) \tag{2.14}$$

The goal of the lasso algorithm is to minimize the value of Equation 2.14. The first part of the equation is the residual sum of squares, and the second part introduces an additional penalty term based on the absolute values of the coefficients. Here the L1 regularization term is the sum of the absolute values multiplied by α. This α parameter

controls the amount of regularization applied. Choosing α is crucial as larger values push the coefficients to zero, whereas smaller values reduce the regularization effect, making the model closer to linear regression.

This study will use an extension of the lasso model, namely group lasso regression. Group lasso aids in feature selection by letting the researchers predefined groups of independent variables related in the regression model [22]. In traditional lasso, each model feature is treated independently without considering their relationship. Group lasso addresses this limitation by grouping related features and applying regularization at the group level [23]. It encourages sparsity within each group and across the groups, promoting the selection of entire groups of features rather than just individual coefficients.

2.3. Sampling Pipeline

2.3.1. Stratified Sampling

For this research, when collecting user-item interaction regarding cars, we are faced with the challenge that the population size of the cars dataset is too large to perform a complete analysis on. For this reason, we use stratified sampling to take a sample size of the population.

Stratified random sampling involves dividing the population into homogeneous groups based on specific characteristics referred to as strata [24]. You can choose to stratify based on multiple characteristics as long as every subject is in one stratum only. Each stratum is then sampled using proportionate sampling. This means that the sample size of each stratum is equal to the subgroup's proportion in the population as a whole. Hence subgroups that are less represented in the more significant population will be less represented in the sample.

$$\frac{Z^2(p)(1-p)}{e^2} \tag{2.15}$$

Finally, we can decide on a sample size, which must be large enough to draw statistical conclusions. Equation 2.15 is used to determine a large enough sample size, where Z is the Z-score of your desired confidence interval, p is the standard deviation, and e is the margin of error. A visual depiction of the sampling process is shown in Figure 2.1

Figure 2.1.: Visualization of stratified sampling.

2.3.2. MANN-WHITNEY U TEST

Mann-Whitney U test is a non-parametric test used to compare two population means and to test whether two sample means are equal or not [25] [26]. A non-parametric test implies that no assumptions of normality are made related to the distribution of the values. The test's null hypothesis is that there is no difference (in terms of central tendency) between the two populations. Following that, the alternative hypothesis is that there is a difference between the two populations.

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i Y_j) \tag{2.16}$$

To calculate the Mann-Whitney U test for two independent samples, the rankings of the individual values combined from both groups must be determined first. The rankings are then added up for the two groups and divided by the number of samples in that groups to determine the average rank. This is shown in Equation 2.16. The difference between the average rank of the groups shows us whether there is a possible difference between the variable. After the mean and dispersion have been estimated, the p-value for the test can be calculated using the U statistic. A p-value lower than 0.05 means the null hypothesis can be rejected.

3

Literature Review

In the previous chapter, we presented the background knowledge used in this thesis. This chapter provides an overview of the literature related to this thesis. Section 3.1 focuses on the recent trends and literature related to climate change and e-commerce activities. Section 3.2 covers popular techniques in AI used to tackle various domains of sustainability. Section 3.3 explores the different evaluation techniques, and Section 3.4 compares them. Section 3.5 delves into some novel recommender system evaluation techniques. Section 3.6 concludes the chapter.

3.1. Climate Change & E-commerce

Climate change, a critical environmental challenge, refers to long-term alterations in Earth's climate patterns, predominantly caused by human activities, such as burning fossil fuels, deforestation, and industrial processes [27]. Climate change can be measured through anthropogenic greenhouse gas increases, such as atmospheric carbon dioxide (CO_2). The global atmospheric concentration of CO_2 has increased from a pre-industrial value of 280 ppm to 421 ppm as measured in 2022 [28]. Climate change adaptation and mitigation have been a crucial research topic since the beginning of the century.

E-commerce, the electronic buying, and selling of goods and services, has witnessed exponential growth in recent years. The digital marketplace offers unparalleled convenience, enabling consumers to shop from the comfort of their homes and businesses to expand their reach [29]. The relationship between e-commerce and climate change has gained considerable attention from researchers and policymakers. As the digital retail sector grows exponentially, concerns have arisen regarding its environmental implications [30].

The environmental footprint of this industry is spread through various stages of the process. The primary contribution is transporting goods globally, which generates significant emissions due to individual decentralized shipments [31]. Packaging materials used to protect and deliver the product generate further waste and consumer energy [32]. Additionally, storing large amounts of data regarding users and products is another energy-intensive process contributing to the sector's overall footprint. Understanding the environmental implications of e-commerce in the context of climate change is crucial for developing sustainable solutions. For this reason, this thesis focuses on how the

climate impact of e-commerce can be reduced by exploring the path of sustainable recommendations to users.

3.2. SUSTAINABILITY IN ARTIFICIAL INTELLIGENCE

Throughout recent years, there have been many ways in which Artificial Intelligence is being used to combat climate change. Across many domains of the field, a nudging towards sustainability can be found.

3.2.1. COMPUTER VISION

The field of computer vision aims to enable models to interpret and understand visual information from images or videos [33]. In the sustainability domain, this technology has seen increasing demand in the AgTech industry [34]. Computer vision analyzes crop health and detects diseases, pests, and nutrient deficiencies in plants [35][36]. These capabilities allow farmers to take early targeted action instead of applying chemicals across the fields, reducing negative impacts on soil, water, and other organisms [37].

Recycling is another domain of sustainability practice in which computer vision solutions are being used. Sorting through various objects is a task in which vision-based models excel. Automated recycling systems such as TrashNet [38] can sort metal, glass, plastic, cardboard, and trash with a success rate of 86.7%

3.2.2. MACHINE LEARNING

Machine learning encompasses algorithms and statistical models that allow machines to learn from data patterns and make predictions without explicit programming [39]. Machine learning models are currently utilized to enable sustainable change in electrical systems. Electrical systems are responsible for approximately 25% of all human emissions [28]. Various prior studies facilitate the integration of low-carbon electricity sources such as solar panels and wind turbines by forecasting supply and demand [40][41][42]. Other studies are also exploring the potential of machine learning in developing new energy storage materials [43][44].

Reducing GHG emissions from various transportation systems is another application of machine learning for sustainability. Models can be used to identify, understand and forecast traffic patterns that reduce emissions produced by idle vehicles [45][46]. Other solutions for optimizing vehicle-sharing facilities such as Uber [47]. Shifting to more sustainable transportation modes has been highly researched and includes understanding user preferences and looking for ways to stimulate the use of low-emission options. To best understand user preferences, some works provide automated solutions based on user GPS and social media data [48][49]. Other works promote low-emission public transport options by forecasting traveling and arrival times of different transport models to improve their usability [50][51].

The development and upkeep of buildings are another significant focus in the discussion of sustainable emissions. Research here focuses on reducing emissions when building new structures and reducing the emissions of current facilities. Research here explored methods of forecasting the energy demand of buildings to consciously improve

energy use [52][53]. Smart home options have also been researched to predict whether rooms are occupied to use heating and lighting systems more efficiently [54].

These examples only cover a portion of the sectors where machine learning techniques are being used to encourage sustainability. [55] presents a comprehensive overview of the overlap between machine learning and climate change.

3.2.3. RECOMMENDER SYSTEMS

Merging sustainability into recommender systems is a field of research that has been starting to get a grip in recent years. As consumers become more aware of the consequences of buying through unsustainable means, a significant shift in consumer mindset has occurred. [56] proposed an energy-efficient transportation recommender system that leverages location traces to optimize taxi pickup points and parking positions. This system aims to reduce energy consumption in urban mobility. Furthermore, [57] conducted a comprehensive survey exploring the role of recommender systems in smart cities' sustainability efforts. This research showcases diverse applications, such as promoting local businesses offering healthy food options and supporting local farms. In the context of traffic management in urban cities, some systems suggest lane-switching decisions based on urban congestion while encouraging greener mobility options like cycling facilities and vehicle-sharing services.

Sustainable recommender systems have also made inroads in the travel and tourism domain. A study showcases a travel-based recommender system that tailors recommendations to tourists' diverse interests while prioritizing sustainable travel and transport methods when suggesting destinations and activities [58]. Additionally, [59] conducted a comprehensive review of recommendation systems for e-tourism, highlighting their potential to foster sustainable practices within this industry.

Sustainability considerations extend beyond travel to digital marketing, as showcased by [60] [61]. Their recommendation system for sustainable fashion products demonstrates a novel approach that prioritizes the ecological impact of products, encouraging consumers to make more conscious and environmentally friendly choices. Another noteworthy method, as presented by [62], introduces a flexible probabilistic framework to identify sustainable products and customers, enabling personalized and sustainable recommendations for future purposes. Overall, these studies collectively emphasize the multifaceted efforts made in designing recommender systems that align with sustainability goals, spanning various domains and applications.

3.3. RECOMMENDER EVALUATION

3.3.1. OFFLINE EVALUATION

Offline evaluations are attractive in the recommender systems community as they require no real user interaction; simply a dataset suffices [63]. The type of evaluation is often the first point of entry to assess the quality of recommendations produced by recommender algorithms. Recommendation tasks from various domains such as recipe recommendation [64] [65], movie recommendation [66], music recommendation [67] [68] [69]. Likely many others use offline recommendation as their primary and sole form

of assessing the validity and performance of recommendation algorithms.

Due to recommendation inherently being a prediction task, like many other machines learning evaluation strategies, offline evaluation also has a strong focus on accuracy metrics [70]. These include but are not limited to precision, recall, mean squared error (MSE), normalized discounted cumulative gain (NDCG) [71][70][63].

In recent years, the community has heavily criticized the approach of solely using accuracy metrics. [72] even goes as far as to say that focusing on accuracy alone has kept progress stagnant and hurt the development of recommender systems in some instances. As a result, many researchers have been attempting to look beyond accuracy metrics and into other forms of offline validation for recommendation algorithms. [73] work along with others propose evaluation metrics such as diversity, novelty, serendipity, and coverage to re-rank initial 'accurate' recommendations, to preserve some accuracy, but also go beyond [73], [71].

3.3.2. ONLINE EVALUATION

Online evaluation is another type of recommender validation technique that can be seen in research. Argued to be the most accurate form of testing your algorithms, online evaluation aims to test recommender systems in their deployed environment without user knowledge about the evaluation being conducted [63]. This type of evaluation study is often only in the scope of prominent corporations, with large monetary stakes as incentives. These online controlled experiments are utilized to make data-driven decisions. A few examples of well-known companies that conduct such experiments are Spotify [74], Microsoft [75], Netflix [76] etc.

Often the goal of online evaluation is to be used in the context of A/B testing, where two variants of a recommendation algorithm are being tested. A study compared the performance of different recommendation strategies on a music streaming platform using A/B testing [77]. The study found that personalized recommendations significantly improved user engagement and retention. This study describes one of the various examples.

The most popular metric in recommendation online evaluation consists of CTR (click-through rate), defined as the number of page views [63]. Other niche metrics include CPC (cost per click), namely total spend divided by the number of clicks, and RPM (revenue per thousand impressions), which is revenue divided by number of page reviews times 1000.

Claimed to be the best evaluation practice, critiques on A/B testing practices have not failed to emerge. Researchers have published several papers [78][79] highlighting the pitfalls of trusting online evaluation results. They claim that the "statistical theory of controlled experiments is well understood, ... and the difference between theory and practice is great." Minor instrumentation issues can often render metrics such as CTR brittle and unreliable. A similar study by [80] identifies common pitfalls of A/B testing in the automotive sector, which include blind adoption of good results, unclear selection of evaluation metrics and undetermined period/length of experiment.

3.3.3. USER STUDIES

Often used interchangeably with online evaluation, user studies refer to experiments designed to evaluate the performance of a recommender system from a user's perspective [63]. User studies typically contain a much smaller number of participants than online evaluation techniques. A survey by [81] reported that only approximately 25% of user studies included more than 50+ participants. Despite this, the advantage of conducting user studies lies in the rich user data the studies can collect. When conducting such a study, users know that data is being collected. Hence demographical questions can be asked with consent, along with explicit questions about user satisfaction and user perception of the system [81].

User studies can reveal fruitful findings, examples of which include: users value transparency (understanding system logic) of the recommender system, longer descriptions of products correlate more with perceived usefulness of item [82], users may prefer inclusion of some recommendations from their social network in the working of an automatic recommendation algorithm [83].

Self-Selection Bias. User studies come with their own challenges, with the main concern being the population sample. User studies often feature convenience samples from the population, which refers to samples drawn from part of the population close to hand [84]. In the study [85] where users were selected by means of sending out emails to a mailing list, convenience sampling can cause self-selection bias. Similarly, in the research by [83], participants were once again recruited through mailing lists and other shared community pages to evaluate book-recommendation algorithms. Results of user studies with samples obtained through self-selection have the risk of not being generalizable to the entire desired population.

Consumption Factor. Unlike online evaluation, in many cases, participants of user studies do not consume or experience the item being recommended. This results in the ambiguity of whether study participants would behave the same way and provide the same answers and ratings as when consuming the item. To analyze and mitigate such a factor, [86] introduced a user sincerity measure. In the study, recommended items and links to their reviews were provided, and user sincerity was calculated by measuring five implicit feedback components: visiting product reviews, time spent on reviews, printing or not the reviews, saving or not the reviews, and emailing or not the reviews. Based on the weighted factors, insincere users were removed from the study.

Along the same lines, [87] investigated the effect of actually consuming recommended songs in a user study by conducting a pre and post-consumption assessment of recommendations. Results of the survey show that user typically underestimated their liking. The study suggests that presenting adequate information about the item can help mitigate the consumption factor.

3.4. COMPARISON OF EVALUATION TECHNIQUES

Now that we have an overview and associated research of all three prominent evaluation techniques used for recommender system validation, this section will dive deep into how they compare. Studies comparing evaluation techniques are commonplace as researchers want to evaluate their methods through multiple means.

3.4.1. OFFLINE VS. ONLINE & USER STUDIES

The first point of comparison is frequently between offline evaluation techniques and online/user studies. The overarching idea behind offline evaluation is not to produce great metrics results but to have the results be translatable to a natural environment with actual users [63]. A leading researcher in this area has published many studies conducting comparative analyses of offline and online evaluations in the domain of research paper recommendation [88] [89]. The experiments evaluate a set of recommendation algorithms using offline and online evaluation to examine whether the results are correlated. Results show that the correlation between offline and online assessment is mediocre at best and concludes that offline evaluation lacks predictive power due to the ignorance of human factors.

In another study [81], a survey on 80 recommendation approaches was conducted. It was revealed that 21% of the approaches were never evaluated, and out of the estimated procedures, 70% were compared against simple and often outdated baselines instead of comparing against state-of-the-art techniques. The study's conclusions show that due to this and other shortcomings, it becomes challenging to determine which approaches are the most promising.

Coming back to the issue of the predictive power of offline evaluations, [90] claim after comparing algorithms across four dimensions (offline, online, time and non-algorithmic factors), that offline performances were not predictive of online in "the absolute and relative sense." An analogous study reporting the live evaluation of news recommendations determined that the relative performance of offline data was precisely reversed in the live system evaluation.

Contradictory to the research above, studies such as [91], [92] reports that Recommender systems with high offline scores were also the ones preferred by users and have relative predictive strength in the real world setting. Many such opposing claims can be seen in the recommender system research community. Speculation on the matter by [88] suggests that this phenomenon occurs due to sub-optimal dataset quality and differentiating domains, which recommender algorithms can be sensitive to.

3.4.2. ONLINE VS. USER STUDIES

As discussed in previous sections of this literature review, online and user studies differ in two main categories. First is user knowledge of the fact that they are part of an evaluation, and the second is the sample size of participants. Due to online experiments being expensive to conduct, academic research in recommender system evaluation often focuses substitutes this with user studies.

The study [88] concerning research paper recommendation draws to our attention that user studies strongly correlate with CTR in their research. This indicates that explicit user satisfaction is a good approximation of the acceptance rate of recommendations in an online setting. For such studies, user studies can employ convenience samples or crowd-source to obtain more users in the survey [93].

Across the literature, it has been showcased and agreed upon that the main advantage of user studies is being able to gauge user experience and provide an explanation of how and why user experience emerges from recommender systems [94][95][96]. Being

able to ask user's their explicit opinions about a plan at hand provides researchers with extra information and resources to understand their own product and algorithm better. Integrating user feedback and constructing user-centered design has been a highly crucial element in any product design [97] [98] [99]; hence a similar argument follows for recommender systems.

Having taken into account the importance of user satisfaction in recommender systems and its low correlation with accuracy metric, [100] proposes a unifying framework called ResQue, which asks users to assess the following: perceived system qualities, user beliefs as a result of these qualities, subjective attitudes, and moral intentions. The authors of the questionnaire did not stop at this but further went on to validate the use of these questions by conducting user studies with them. This tested the internal consistency and reliability of the model [100].

3.5. NOVEL EVALUATION METHODS

Through the years, novel evaluation techniques within offline and online evaluations have emerged to address various gaps in the recommender evaluation field. Advancing away from accuracy, many recent offline evaluation procedures search for new metrics to depict user behavior best. Sincerity is a metric proposed by [86], which tries to capture real intention during user studies. This metric helps to filter out users in studies that are simply not sincere in their rating, thus corrupting the collected data. Another study [101] suggested the creation of a performance metric, $\mathscr{P}$, that was developed to measure the global execution of the recommender system and the nearness to its actual goal. This is derived directly from the general objective of recommender systems and is defined as the final performance of the recommender system over the number of sessions. Here performance refers to whether the recommendation is followed or is deemed attractive to the user.

Along the same lines of validating offline recommender evaluations, [87] tested the effect of actually consuming the item being recommended rather than just looking at it on a screen, and [92] used eye-gaze tracking as an implicit indicator to validate their offline experiments. On the other hand over arching studies by [102] and [100] suggest frameworks and offer tools to redefine the evaluation research completely. A framework by [102] can conduct entire offline evaluations consisting of 13 splitting methods, 8 filtering approaches, 51 different hyperparameter optimization strategies, 50 models, and 36 metrics. The aim is to let researchers quickly test their models by configuring a simple file.

3.6. CONCLUSION

The literature review explores the relationship between climate change and e-commerce, focusing on sustainability in the context of artificial intelligence (AI) applications. This discussion highlights the growing concern regarding the environmental impact of the e-commerce industry and how AI can contribute to mitigating climate change. Integrating AI, particularly in computer vision and machine learning, offers promising solutions in various domains, such as agriculture, recycling, energy systems, and transportation,

leading to more sustainable practices.

The review also emphasizes the emerging field of sustainable recommender systems, which align with consumers' growing awareness of sustainable consumption. Sustainable recommender systems can nudge users towards making eco-friendly choices, leading to reduced environmental footprints in various sectors like transportation, tourism, and digital marketing. The discussion points out that user studies and online evaluations provide valuable insights into user experience and preference, allowing for a better understanding of recommender system performance from the user's perspective.

However, it is crucial to acknowledge the limitations and challenges of different evaluation techniques. Offline evaluations may not always reflect real-world user behavior, and online assessments can be expensive and prone to biases. User studies, while providing valuable feedback, may suffer from self-selection bias and may not capture actual consumption behavior. Researchers are continually exploring new evaluation methods, including sincerity metrics, eye-gaze tracking, and comprehensive evaluation frameworks, to address these limitations and improve the accuracy and reliability of recommender system evaluations. Overall, the literature review reveals the immense potential of AI and recommender systems in driving sustainable practices in e-commerce and various other domains.

4

CarEmissions Dataset

In the previous chapter, we discussed the literature related to the thesis. In this chapter, we discuss the CarEmissions dataset created. Section 4.1 justifies using the dataset. Section 4.2 gives an overview of the created dataset. Section 4.3 analyzes the created dataset. This includes comparison to other datasets and interaction and user properties of the dataset. Lastly, Section 4.4 concludes this chapter.

4.1. Introduction

Datasets are an essential part of developing and evaluating recommender systems. To assess whether recommendations are green and sustainable, we must look at the recommendation algorithms that generate them. The first paper in the field to enhance a dataset containing user-item preferences and the CO_2 footprints of the items was published earlier this year [6]. The CO_2 footprints of the items were added later, meaning the users providing the ratings did not know this information. In this thesis, we want to go a step further and determine how providing users with the CO_2 emissions of items affects the rating they give them and whether this can result in a greener recommender system.

We collect user preferences for a particular item category to create this dataset. The item category is cars from different brands for the following numerous reasons. Firstly, the initial cars dataset obtained contains the CO_2 emission value in terms of g/km (grams emitted per kilometer driven). This means that emission value is not gathered based on third-party information but is provided by the manufacturer. Secondly, a study [103] has shown that information such as CO_2 emissions can influence the consumer's choice of vehicle. This justifies the practical relevancy of the dataset, which is an essential factor for the RecSys research.

4.2. Dataset Overview

The final dataset consists of 136 users and 396 cars. There are a total of 9650 user-item interactions in the dataset. We collected the dataset over two months, from April 2023 to June 2023. Around 48.4% of the interactions are with cars when the CO_2 emissions are not displayed, and 51.6% of the interactions where the CO_2 emissions are shown.

The dataset also contains the age and gender information for each user. Level of education, field of work, country of origin, retention of driving license, access to car, main reasons for car use, knowledge of vehicles, and importance of emissions were kept as optional. Hence these fields exist for some users and are omitted for some. The subset of the data where CO_2 emissions are displayed will be referred to as CarEmissionsCO_2, and similarly, the other subgroup where emissions are not shown will be referred to as CarEmissionsNoCO_2. Finally, the overall dataset is CarEmissions.

4.3. Analysis

This section features an analysis of the curated dataset. This analysis consists of an initial comparison with other RecSys datasets such as Movielens100k, Movielens1M, BookCrossing, and RecipeEmissions. Furthermore, as the CarEmissions dataset features two different kinds of ratings, namely when CO_2 is or is not displayed, we also delve into how their values compare in the dataset.

4.3.1. Dataset Comparison

Table 4.1 compares the properties of commonly benchmarked recommender system datasets. We include all three versions of the CarEmissions dataset. This is because the same user-item interaction could exist in both the CO_2 and the No CO_2 dataset; hence they form two individual and distinct datasets.

The CarEmissions dataset has a smaller number of users and items on average. This, however, is fine for the dataset due to the number of interactions. Due to a high relative number of interactions, the CarEmissions dataset can achieve a significantly lower sparsity than the other datasets. The sparsity of the CO_2 and the NoCO_2 dataset are comparable to those of MovieLens100k.

The CarEmissions dataset has a mean rating lower than the other datasets, except for Book-Crossing. This is because the CarEmissions dataset stands out from the other datasets in how the data was collected. Here all users were asked to rate each car they came into contact with, whereas in the other datasets leaving ratings was optional. This causes a positivity bias in the other datasets, with users only likely to leave a rating if they enjoy the particular recipe, movie, or book. This bias is avoided in the CarEmissions dataset, thus making it more robust. This might influence the accuracy metrics such as RMSE and NDCG more negatively when compared to the other datasets. The RMSE for CarEmissions is likely to be higher because there is more spread in the training data.

Finally, looking at the median interaction per item, and user, we can see that the CarEmissions dataset is highly comparable to the MovieLens100k. They both have a more significant number of interactions per user and item, which will aid in the training phase of the recommendation algorithms. From this, we can conclude that although the CarEmissions dataset differs in some aspects, this is due to justifiable reasons which do not interfere with the validity of the dataset for the recommender system study.

Table 4.2 displayed the properties of other popular RecSys user study datasets. The respective papers did not report values missing in the table. The table shows that the number of users in CarEmissions is comparable to other user studies. The number of

	CarEmissions (All)	CarEmissions (CO_2)	CarEmissions (No CO_2)	Recipe Emissions	Movielens100k	Movielens1M	Book-Crossing
# of users	130	111	116	32090	943	6040	105283
# of items	396	396	396	5595	1682	3706	340556
# of interactions	9971	5229	4742	247219	100000	1000209	1149780
sparsity	NA.	88.11%	89.67%	99.86%	93.70%	95.53%	99.99%
minimum rating	1	1	1	0	1	1	0
maximum rating	5	5	5	5	5	5	10
mean rating	2.56	2.52	2.6	4.5	3.53	3.58	2.87
median interactions per item	25	13	12	27	27	124	1
median interactions per user	65	32	34	3	65	96	1

Table 4.1.: Properties of commonly benchmarked recommender datasets.

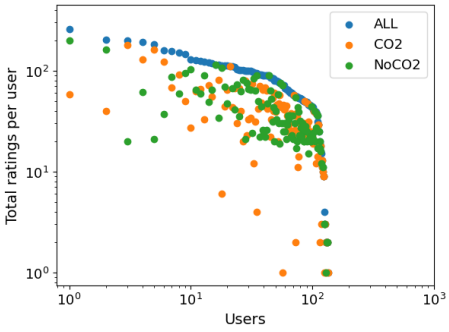
final user-item interactions and median interactions per user is much higher in CarEmissions than in its counterparts. This is because in CarEmissions users decide how many items to rate, whereas, in other datasets, the researchers determine the number of items. The mean rating given to items could not be compared, as many user study datasets conducted a ranking rather than a rating task. The ranking task refers to ranking the items provided in your preferred order.

	CarEmissions (All)	CarEmissions CO_2	CarEmissions No CO_2	Sohail et al. [86]	Ricci et al. [91]	McNee et al. [104]	Loep et al. [87]
# of users	136	122	116	10	158	138	40
# of items	396	396	396	50	-	24,000	-
# of interactions	9650	4984	4666	500	790	1380	200
Task performed	Rating	Rating	Rating	Ranking	Ranking	Ranking	Rating
Median interactions per item	24	12	12	10	-	-	-
Median interactions per user	60	30	31	50	5	10	5

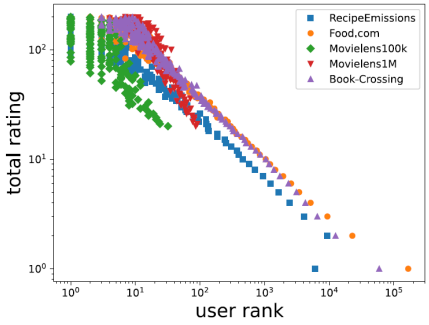
Table 4.2.: Properties comparison of most popular recommender systems user study datasets.

Figure 4.1 shows the user engagement distribution across the datasets. This figure was split into CarEmissions and other datasets for readability purposes. These figures show that all datasets, including CarEmissions, present a long-tail distribution. This implies that the majority of the users occur in a small number of interactions. CarEmissionCO_2 and CarEmission$NoCO_2$ depict a more significant spread in the distribution, indicating a more considerable variance in items rated by different users. Like Movielens, CarEmissions is also a dense dataset with more users with an average number of items rated around 30 to 40.

Figure 4.2 shows how often an item is rated. All datasets present a long-tail distribution similar to user engagement. This implies that for all recommender datasets, most items occur in few interactions except for a few. Though giving a long tail, CarEmissionsCO_2, and CarEmissions$NoCO_2$ have some notable differences from the other datasets. We observe many points densely clustered between 20 to 40 ratings per car for both datasets. This happens because the users in the study rate cars at random with a minimum of 20 cars for both CO_2 and $NoCO_2$.

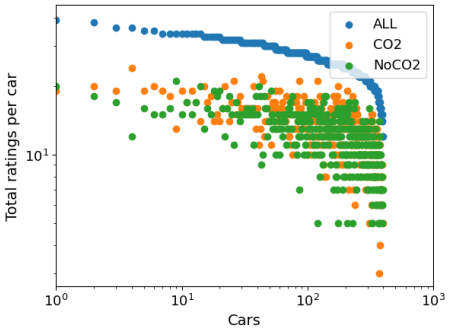


(a) CarEmissions Dataset

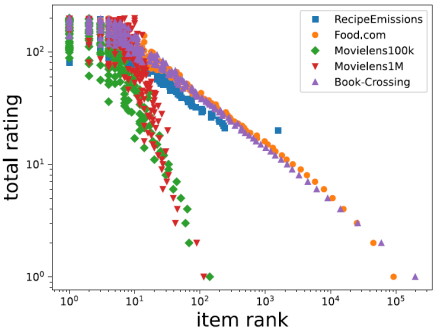


(b) Other RecSys datasets

Figure 4.1.: User engagement distribution of datasets.



(a) CarEmissions Dataset



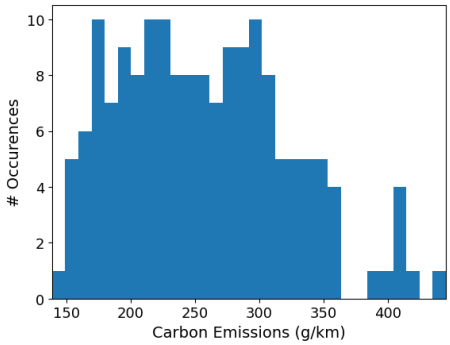
(b) Other RecSys datasets

Figure 4.2.: Item popularity distribution of datasets.

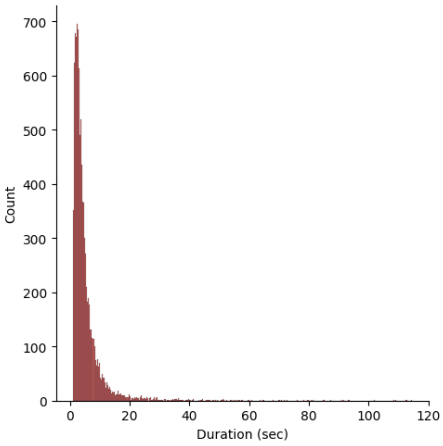
4.3.2. Interaction Properties

Now, we discuss more in-depth properties of the CarEmissions dataset to understand the data we are working with. Figure 4.3a displays the distribution of CO_2 emissions of the cars in the CarEmissions dataset. The distribution follows a bell-shaped curve, indicating a somewhat normal distribution skewed slightly to the left. This means that most cars are centered around the average to lower emissions spectrum, with a few cars having very high emissions. It is important to note that the starting value displayed for CO_2 emissions is 140. We do not normalize the emission values between the range of 1-5 to preserve the unit of measurement, which is g/km.

Figure 4.3b shows the time distribution in seconds spent per interaction. The distribution is heavily skewed on the left, indicating that most interactions take 1-10 seconds. This means that users are very quick at determining what they like about a particular car and what rating to give. This might imply that most users only look at one or two specific car characteristics to decide, as 10 seconds is insufficient to register several properties regarding an item.



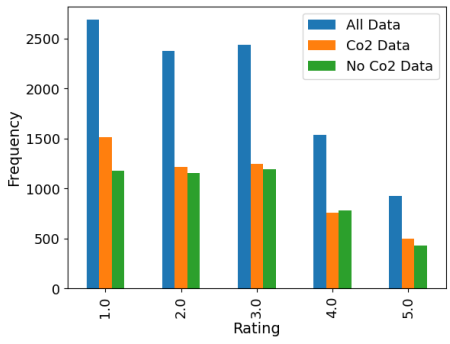
(a) Distribution of emissions



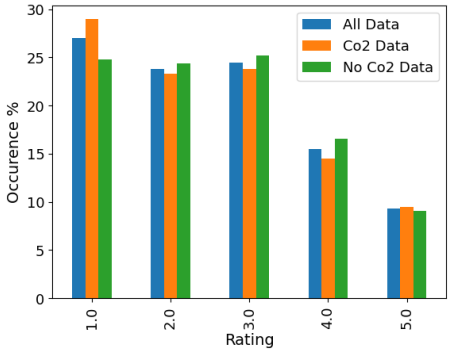
(b) Distribution duration of time (in seconds) spent per interaction.

Figure 4.3.

Figure 4.4 features the dataset's rating distribution by frequency and percentage. Figure 4.4a and 4.4b show that the most frequent rating is 1. We can also observe that when the CO_2 information is displayed, the percentage of 1 given as a rating is higher. This might indicate that upon seeing the CO_2 emissions of a car, participants are more likely to provide a lower rating. This could be attributed to the fact that even cars with lower emission values might be perceived as having high emissions, as the emissions data is not standardized within a range of 1-5. The percentage of 2 and 3 ratings are very similar, and 5 is the least occurring rating.



(a) Frequency



(b) Percentage

Figure 4.4.: CarEmissions rating distribution

Figure 4.5 depicts the emissions distribution of the cars as a function of the ratings

received. Boxplots clearly show that cars that received a rating of 1 have a higher average CO_2 emissions value. This was confirmed by a one-way ANOVA on the emission values separated by ratings. The p-value of the resulting ANOVA test was 0.003, making us reject the hypothesis that the means of the groups are the same. The cars that received the rating of 3 and 4 tend to have a lower average emissions value. This could indicate that users are inclined to give higher ratings to cars with lower emissions. It should be noted that though the rating of 3 and 4 tend to have a lower average emissions value, they include a significant amount of outliers with high CO_2 emission values.

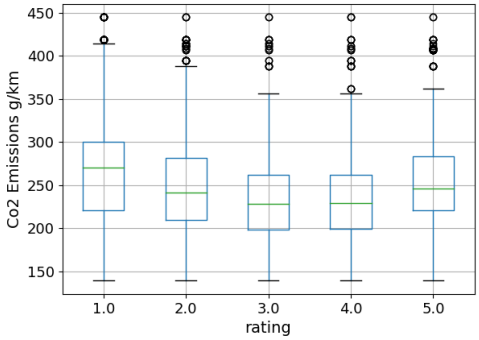


Figure 4.5.: distribution of emissions per rating

4.3.3. User Properties

In this section, we analyze the users' demographic data in CarEmissions. The dataset features 93 males, 40 females, 2 other, and 1 non-binary. Figure 4.6a shows the age distribution by different genders in the dataset. The most represented group are males between the age of 20 and 35. This occurs due to collecting a voluntary convince sample with the main population being TU Delft students. Various age ranges are also covered, with virtually no gaps between 18 to 60 years old.

Figure 4.6b displays a bar plot of the field and the highest education level attained by the participating users. The largest groups are people working in computers & technology, and engineering, with master's degrees. This, once again, is due to the voluntary sample mainly being obtained through contact with students at TU Delft. The largest groups are users with bachelor's degrees in the same fields. Apart from these, there are 1 or 2 users from various domains with varying levels of education. This ensures being able to generalize the scope of the study to a broader audience.

Figure 4.7 features various bar plots regarding user interaction with cars. This data was gathered to determine whether there is any correlation between such factors and car ratings. Most participating users own a license, and most also have car access. The largest group of users also use a car yearly. This may occur since the study was conducted in the Netherlands, where the primary mode of transportation for many is a bike. This might also indicate that users are generally more conscious of the environmental impact of using a car. Finally, most users indicate they are somewhat familiar with cars. This is

(a) Age-gender distribution of users in CarEmissions.

(b) Barplot of field and education level of users.

Figure 4.6.

desirable as it ensures that users go in with some knowledge before participating in the study.

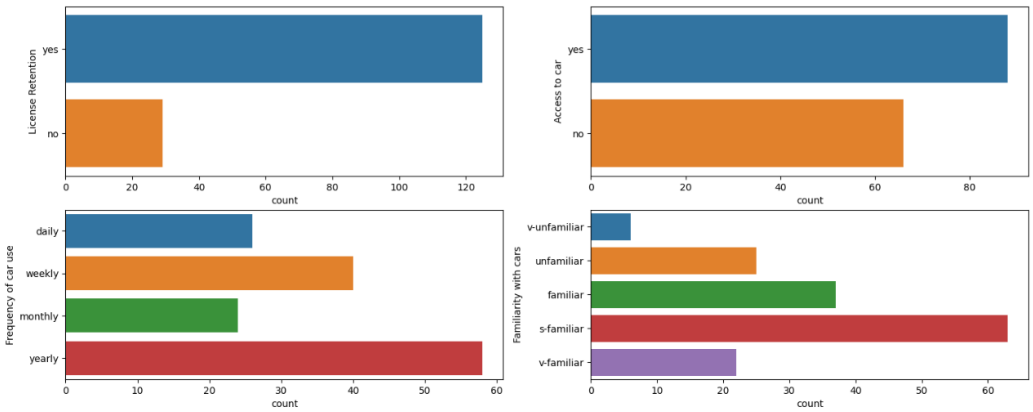


Figure 4.7.: Plots regarding general user car usage and knowledge.

4.3.4. 10% OVERLAP

CarEmissions also includes data for the same user-item interaction with and without CO_2 emissions shown to the user. This allows us to directly compare ratings a specific user gave to a particular item, with only varying the presence of the control variable.

Figure 4.8 shows the rating distribution for this 10% dataset with CO_2 displayed versus omitted. It can be observed from the plot that the mean rating of the NoCO_2 group appears to be higher than the CO_2 group. This shift in mean observed in the figure can be attributed to the CO_2 group receiving higher ratings. We conducted a Mann-Whitney test to determine whether the population differs significantly from one group to the other. The p-value of 0.1489 indicates that the population means are not significantly different.

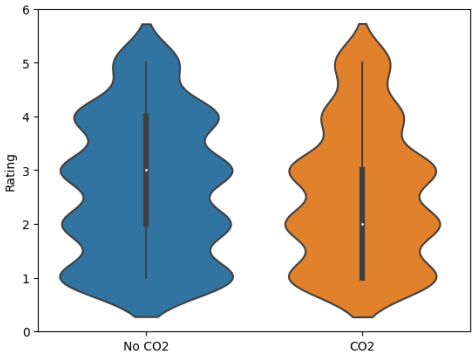


Figure 4.8.: Violin plot depicting the distribution of ratings of the 10% overlap dataset.

4.4. Discussion

We proposed a new recommender system dataset containing user-item ratings with emission values as a control variable. This dataset can be used for researching green recommender systems. It includes 136 users, 396 cars, and 9650 interactions. The dataset is intended as an innovative solution for evaluating the greenness of recommender systems by also considering user perception of emissions. In the remainder of this section, we discuss our choices and some dataset limitations.

We analyzed the CarEmissions dataset and compared it to several other commonly used datasets and other user studies from RecSys research. Comparison with other popular datasets showed similar distributions. Although car emissions show a long-tail distribution, the ratings are more spread out on a scale of 1 to 5. This is because users do not choose cars but are shown to them randomly, avoiding item popularity bias. This is likely to influence the benchmarking experiment. Since the rating values are more spread out, recommender algorithms might make it harder to accurately predict the rating. This may result in lower NDCG values when compared with other datasets.

When comparing the dataset to other user studies, we can see that CarEmissions is comparable in size and scale. The only difference is that many user studies ask their users to perform a ranking task, whereas it was a rating task for us. This is because most other user studies evaluated the output of recommendation algorithms, whereas, with CarEmissions, we aim to train green recommendation algorithms.

A limitation of CarEmissions is scalability. Representing the 'general' population is rarely possible when using volunteers for data collection. The analysis showed that most

users were males from computer and engineering backgrounds. Generalizing results obtained from such a sample user base may produce untrustworthy results. Since the sample population does not represent the real-world population, using the dataset with this knowledge is critical. This issue can be addressed in future work by conducting the study on a larger scale. An example would be posting the survey on Amazon Mechanical Turk, such that the population of users is no longer a convenience sample but instead paid participants from different backgrounds.

5

Collection & Correlation

In the previous chapter, we highlighted the properties of CarEmissions. In this section, we delve into the analysis and correlation model of the dataset. Section 5.1 highlights the collection process. Section 5.2 explains the lasso regression model created and the results. Finally, Section 5.3 concludes the chapter.

5.1. Data Collection

This section describes the creation of the dataset. The conceptual model comes first, then the experimental design, and finally the reduction of the original car dataset.

5.1.1. Original Cars Dataset

According to EU Regulation No. 2019/631, countries are required to record information for each new passenger car registered in their territory. Each year, each Member State has to submit all the information related to their new registrations. We obtained this information regarding all the new passenger cars registered in 2020. This dataset initially started out with over 7000 cars and contained the following information about each car: make, model, vehicle class, engine size, cylinders, transmission, fuel type, fuel consumption city (L/100km), fuel consumption highway (L/100km), fuel consumption combined (L/100km), and CO_2 emissions (g/km).

Upon initial inspection of the dataset, many of the cars were very similar, with slight variations between them. An example of this is the car 'ACURA ILX Compact' with one having the engine size of 2.0 and the other of 2.4. Though these differences can be considered significant, in terms of a recommender study, reducing the number of cars is an advantage as it aids with the cold start problem. We decided to reduce the number of cars through the method of stratified sampling discussed in Section 2.3.1. With this method, we can obtain a sample dataset that best represents the entire car dataset. We create distinct strata based on the make, model, and vehicle class of the cars and perform proportionate probability sampling from these. In our sample, we keep approximately 5% of our original data.

To ensure that the distributions of the dataset properties have been preserved, Figure 5.1 shows the density plots of various properties in the dataset before and after the sampling. All of the density distributions have been preserved, with no alarming changes in

the sample. After conducting the stratified sampling, the change we see in the densities is that most curves have been smoothed out. In order to go a step further in checking whether the sample is representative of the population, we conducted a Mann-Whitney test as detailed in Section 2.3.2. The results of the test did not yield any significant p-values, implying that the datasets are indeed similar. We also calculated the Wasserstein distance between the old and new distribution of variables, and the results once again indicate a negligible change between the two.

Variable	Mann-Whitney p-value	Wasserstein distance
CO_2 Emissions	0.2767	1.2873
Engine Size	0.3253	0.0875
Cylinders	0.0790	0.2064
Fuel Consumption City	0.5688	0.2337
Fuel Consumption Highway	0.3940	0.1337
Fuel Consumption Combined	0.5055	0.1806
Fuel Consumption Mpg	0.5510	0.5273

Table 5.1.: Mann-Whitney test p-values, and Wasserstein distance of variables when comparing the stratified sample to the original dataset.

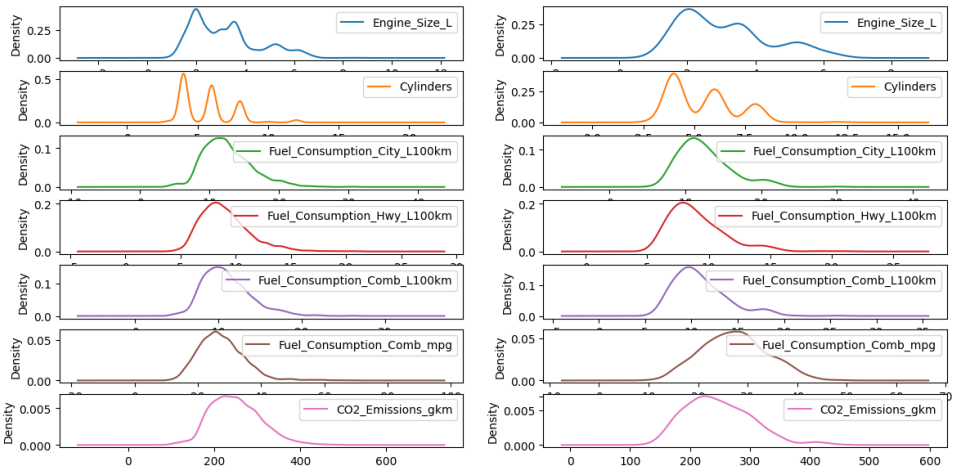


Figure 5.1.: Property density plots before and after stratified sampling.

5.1.2. RELATED CONCEPTUAL MODEL

The next goal is to gather user-item interactions data. This section introduces the conceptual model used in the study, which illustrates the information we need from users and the information that will provide us with a better understanding of their choices. A conceptual model of the variables in the collection phase can be seen in Figure 5.2. With

this model, we aim to identify the independent, mediating, moderating, and dependent variables in the study.

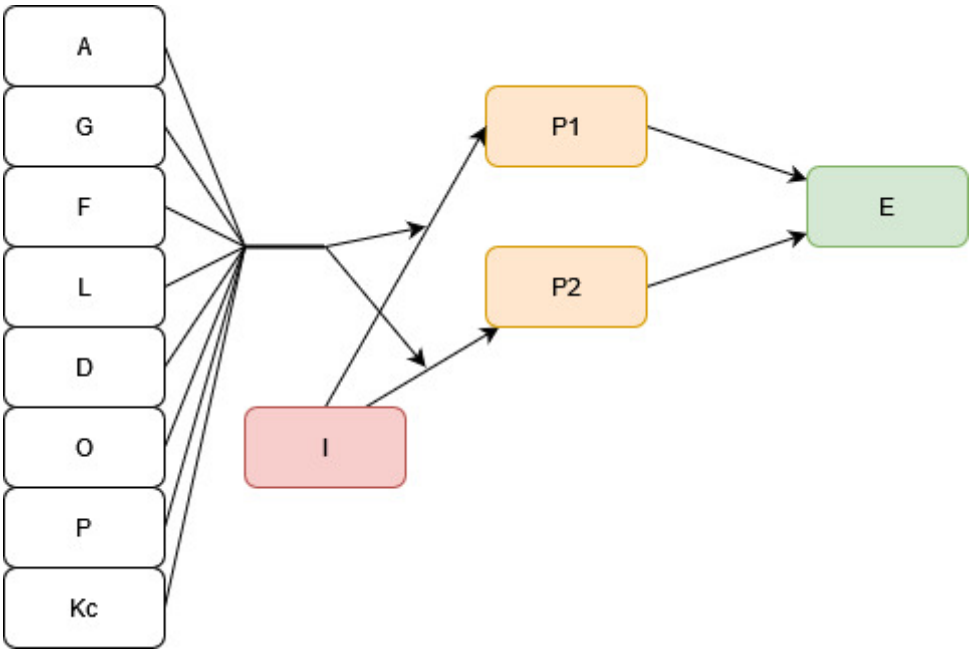


Figure 5.2.: Conceptual model for data collection protocol

Table 5.2 presents all the variables present in the model table. Here, the independent variable is the one whose effect we are trying to measure on the dependent variable. In this study, we want to determine whether there is any difference in user rating for cars given the fact that car CO_2 is displayed. The mediating variable refers to variables that depend on the independent variable and have a direct effect on the dependent variable. Finally, the moderating variable refers to variables that are likely to influence the ratings but are uncontrollable by the experimenters. This includes age, gender, level of study, knowledge of cars, and many others.

5.1.3. EXPERIMENTAL DESIGN

The experimental design of the data collection process has five distinct phases. These phases will be described in detail in this subsection.

Phase 1. Phase 1 introduces the participants to the study. On this landing page, we vaguely mention the goal of the study and what is expected from the users. Here we introduce all the car properties that the users will encounter with descriptions. Finally, at the end of phase 1, the users have to sign an informed consent form agreeing to the privacy and storage methods that this study employs.

Variable	Full Name	Type
I	CO_2 emission displayed or not	Independent
E	Assigned rating	Dependent
P1	User perception of car	Mediating
P2	User perception of brand	Mediating
A	Age	Moderating
G	Gender	Moderating
F	Field of expertise	Moderating
L	Level of education	Moderating
D	Driving license & car use	Moderating
O	Country of origin	Moderating
P	Political view category	Moderating
K_c	Prior knowledge of cars	Moderating

Table 5.2.: Overview of variables in conceptual model

Phase 2. This is the exploratory data collection phase of the study, where they have to provide some data about themselves. The information asked in this phase directly relates to the conceptual model mentioned earlier in Section 5.1.2. A snapshot of this page is presented in Figure A.1.

Phase 3 & 4. Due to the sample size being on the smaller side with 136 users, we propose a within-subjects study. This means that each participant will provide both user-car ratings with CO_2 displayed (Phase 3) and user-car ratings without CO_2 displayed (Phase 4). To eliminate first-look bias in the study, 50% of the participants will start by rating cars with CO_2 displayed, and move on to rate cars without CO_2 displayed, and the other 50% of the participants will do the same. This divides the participants into two distinct groups. Participants in group A start by rating cars with CO_2 displayed (Phase 3). In this part, they are asked to rate a minimum of 20 items before the button to move on to phase 4 is activated. There is no overlap of cars between phases 1 and 2. Participants in group B start by rating cars without CO_2 displayed (Phase 4). In this part, they are asked to rate a minimum of 20 items before the button to move on to phase 3 is activated. In phase, they rate cars with CO_2 displayed. There will be approximately be 10% overlap between cars displayed in phases 3 and 4 in this group. This is to gather any explicit change in rating of the same car when CO_2 is displayed. A snapshot of the user's front end when rating a single car is shown in figure A.2.

A 10% overlap is picked to ensure that some cars can be repeated for the same user while also ensuring this number stays low. If a user is asked to rate multiple cars that are the same, rating might start to become a frustrating task. This will compromise the quality of the dataset.

Phase 5. Finally, in the last phase of the study, once the participants have rated the number of cars they wish to purchase, some final questions regarding their view on carbon emissions are asked. This is done at the end so as not to introduce any biases during the study. Figure A.3 displays the questions asked during this phase.

5.2. MULTIVARIATE LASSO REGRESSION

We will be using group lasso regression, described in detail in Section

5.2.1. VARIABLES

Table 5.3 presents all the variables in the CarEmissions dataset regarding the rating, user, and item. Given this list of variables, we want to see whether some of them have a strong correlation with the rating a particular item receives. As already identified in Section 5.1.2, our dependent variable is rating, which means we want to see the effects of various different variables on the rating variable.

Variable	Type	Description
Review id	Integer	Unique for every user-item interaction
User id	Integer	Unique for every user
Car id	Integer	Unique for every car
Rating	Integer	Provided by user to car
CO_2 present	Boolean	Whether CO_2 shown or not for particular review
Duration	Float	Time taken for a review
Age	Integer	Age of user
Gender	Categorical	Gender of user
Highest education	Categorical	Highest education obtained by user
Field of work	Categorical	User's field of work
Location	Categorical	User's country of origin
License retention	Boolean	Whether user has license or not
Car access	Boolean	Whether use has access to car or not
Car use frequency	Categorical	How often car used by user
Car use reason	Categorical	Reason for using a car
Knowledge of cars	Categorical	Familiarity with cars
Important categories	Categorical	Aspects user deems important when rating a car
Importance emissions	Integer	Importance of CO_2 emission for user
Reason emissions	Categorical	Reason behind finding CO_2 emission important
Start stage	Boolean	User started w or w/o CO_2 emissions first
Maker model	Categorical	Car brand & model
Vehicle class	Categorical	Type of vehicle
Engine size	Float	Car engine size
Cylinders	Integer	# of cylinders in car
Transmission	Categorical	Type of gearbox
Fuel type	Categorical	Type of fuel car consumes
Fuel consumption	Float	Fuel consumed per 100km
CO_2 emissions	Integer	CO_2 emissions in g/km
Car photo	Image	Photo of car

Table 5.3.: Overview of variables in CarEmissions Dataset

5.2.2. EXPERIMENTAL RESULTS

The results of the multivariate group lasso regression show no correlation between any of the variables and rating. The strongest correlation values inspected was -0.007 between age and rating, and -0.003 between CO_2 emissions and rating. The remaining correlation values of the variables were 0. These correlation values implies that the changes in the independent variables do not lead to any predictable changes in the rating variable, at least in terms of linear relationships. To imply moderate correlation between variables, values need to be at least above 0.3.

These correlation values could have several explanations. Firstly, all users view the rating scale differently. Some user may be more optimistic with their ratings and provide higher ratings to items that they like moderately. The opposite could be the case for other users, which would rarely give a rating of 5. This difference in rating perception makes it difficult for the model to find correlations between variables.

Another reason could be the subjectivity of users when they answer any self evaluating questions. Categorical variables in the regression model such as knowledge of cars, importance emissions, important categories, and reason emissions require users to give subjective answer to their perceived self knowledge. Studies show that self-evaluation is often not a reliable measure of our actual knowledge [105]. Perceived self-knowledge could be higher or lower than reality, depending on various dynamic factors. This causes a correlation between these categorical variables, making ratings hard to predict.

5.3. DISCUSSION

In this chapter, we presented the data collection process. This process started with taking a stratified sample of the original car dataset. Then a conceptual model was created for the user study. The actual user study consisted of five distinct data collection phases. In these phases, we discussed all the information gathered from the users. We also described how the study was designed to take the control variables of CO_2 emissions into account. Then we trained a multivariate lasso regression model to determine any correlations between the independent variables and the rating variables, as presented in Table 5.3.

5.3.1. DATA COLLECTION

Collecting user data comes with challenges and unexpected problems. One such challenge was ensuring the sincierty of user ratings. Since participants in this study were not compensated in any way, the reliability of the data collected from them cannot be ensured. Ensuring that the ratings were sincere and a true reflection of their preferences is an impossible task.

Another challenge is asking users to rate items they did not actually buy. This is an ongoing problem in many recommender system studies. User ratings and buying behaviors might be very different from each other. Consider that a user likes a sports car and would give it a rating of 5, but in reality, a sports car would never be considered due to the expense. Taking such a factor into account is very difficult, as we cannot ask users to purchase items for a study. With this in mind, such user study datasets cannot be a perfect depiction of user preferences.

5.3.2. REGRESSION

The group lasso regression conducted showed poor correlation results between the independent and dependent variables. An important note for this could be that the independent variables in the data do not follow a linear relationship with the dependent variable. Multivariate group lasso a linear regression method If the relationship between variables is polynomial to some degree, our model would fail to capture this. Training other machine learning models to find correlations is outside the scope of this study but could be considered for future work.

6

Sustainable Recommender System

This chapter contains the core results of this thesis. We benchmark different commonly used recommender systems in terms of emission rates and accuracy. Section 6.1 focuses on the experimental setup of the benchmarking experiment. Section 6.2 showcases the results of the benchmarking. Section 6.3 shows the results of nudging towards greener recommendations, and finally Section 6.4 concludes the chapter.

6.1. Experimental Setup

This section describes the experimental setup for benchmarking the different recommendation algorithms. The models will all be individually trained and tuned to CarEmissions, CarEmissionsCO_2, and CarEmissionsNoCO_2.

6.1.1. Data Split

Following the common practice in machine learning algorithms, the dataset will be split into train and test sets. In order to avoid bias towards the cold-start problems, the test set only contains users and items that are present in the training set.

The test set, which is approximately 20% of the dataset, is made by sampling a random interaction from the dataset, and if the user and item from the interaction are present in the dataset, then only this interaction is moved to the test set. This sample is then removed from the original dataset, which will go on to be the train set at the end of this procedure. The final train dataset is 80% of the final dataset.

6.1.2. Hyper-parameter optimization

Hyper-parameter tuning is performed for all three CarEmissions datasets individually. This is done for all the benchmarked models, namely: ItemKnn, UserKnn, SVD, SVD++, CoClustering, and SLIM. The global average does not require any hyperparameter tuning. The grid search space for these hyperparameters is derived from another green rec-

ommender study [6]. The final values for all the hyper-parameters are mentioned in the Appendix Section B.

6.1.3. GREEN RECOMMENDER EVALUATION

After obtaining the best performing models for each partial dataset, we will utilize the nudging approach proposed by [6] detailed in 6.3. We focus on the NDCG and GNDCG values obtained from the nudging scores to effectively assess the performance and effectiveness of the method within its unique ranking context. NDCG and GNDCG will be evaluated at 10, 20, and 50 for a cohesive overview. To determine the best trade-off between prediction and greenness, we will experiment with all values of alpha, inclusive of 0 to 1, with a step size of 1.

6.2. ALGORITHM BENCHMARKING

Table 6.1 shows recommender algorithm performance in terms of NDCG at various lengths. The underlined scores in the table indicate the best-performing models for each partial dataset. SVD outperforms every other model for both CarEmissionsCO_2 and CarEmissionsNoCO_2. This is followed by SVD++. Item-knn, User-knn, and CoClustering perform the worst in these partial datasets. The likely cause of this is the higher sparsity of the dataset compared to CarEmissions. In the user-based knn algorithm, predictions for an interaction with a target item are generated using a predetermined number of similar neighbors. However, due to the sparsity, it is possible that not all neighbors of the user have rated the target item. As a result, the estimation relies on a small number of neighbors, or even a single neighbor, which reduces the accuracy of predictions. The strength of knn algorithms lies in their ability to leverage the combined data of many similar users, and the same limitation applies to item-based knn algorithms.

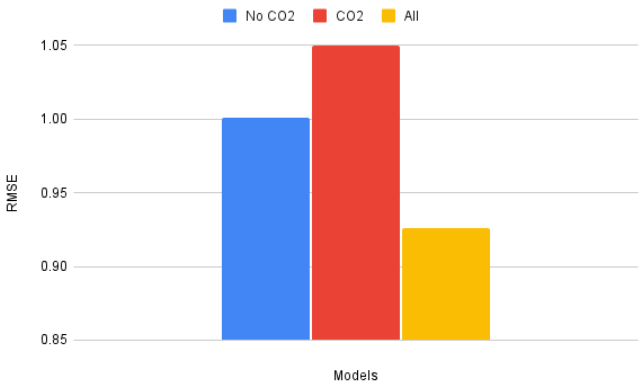
Maxtrix factorization methods such as SVD and SVD++ perform better as their factorization strategy allows them to generalize better over the entire interaction matrix. The same hold for SLIM, which is also a matrix factorization method.

It should be noted that all algorithms perform better when trained on CarEmissionsNoCO_2, than CarEmissionsCO_2. This could be because showing CO_2 emission value adds another property of the car to the mix. This might be a major influencing factor for some users but not for others, thus increasing the variance in the ratings matrix. Along with this, we can also see that the models trained on the entire CarEmissions data outperform both the CO_2 and NoCO_2 partial datasets. This is the main cause of this: double the amount of training data. By combining both the partial datasets, we increase the size of the training data. Hence, better results are an expected outcome.

SVD results. Figure 6.2 illustrates the RMSE values for two partial sets of data and the entire dataset. A lower RMSE value signifies a model with stronger predictive capabilities, as it indicates a closer match between predicted ratings and the actual ground ratings. By observing the figure, it becomes evident that the model trained on the entire dataset outperforms the model trained on the partial sets of data. One possible explanation for this superiority is the larger size of the entire dataset, which provides a greater number of training samples for the model to learn from.

	Algorithm	NDCG@10	NDCG@20	NDCG@50
CO_2 data	Global Average	0.36	0.37	0.37
	ItemKNN	0.43	0.45	0.45
	UserKNN	0.39	0.39	0.39
	SVD	**0.45**	**0.45**	**0.46**
	SVD++	0.43	0.43	0.44
	CoClustering	0.38	0.39	0.39
	SLIM	0.43	0.43	0.43
NoCO_2 data	Global Average	0.37	0.39	0.39
	ItemKNN	0.47	0.48	0.48
	UserKNN	0.46	0.47	0.47
	SVD	**0.52**	**0.52**	**0.53**
	SVD++	0.50	0.51	0.52
	CoClustering	0.46	0.47	0.47
	SLIM	0.48	0.49	0.50
All data	Global Average	0.35	0.37	0.38
	ItemKNN	0.51	0.53	0.54
	UserKNN	0.40	0.42	0.42
	SVD	**0.57**	0.56	0.58
	SVD++	0.53	0.56	0.57
	CoClustering	0.55	**0.57**	**0.59**
	SLIM	0.54	0.55	0.56

Table 6.1.: Average performance in terms of NDCG of popular recommender system algorithms trained on the CarEmissions dataset.



RMSE values for the best-performing recommender models for partial datasets

In Figure 6.1, we can observe the NDCG and GNDCG values for the trained models, providing valuable insights into their relative performance. Once again, the SVD model trained on the entire dataset emerges as the top performer, surpassing the other two

	Algorithm	GNDCG@10	GNDCG@20	GNDCG@50
CO_2 data	Global Average	0.33	0.34	0.34
	ItemKNN	0.37	0.38	0.38
	UserKNN	**0.45**	**0.45**	**0.44**
	SVD	0.42	0.41	0.41
	SVD++	0.41	0.41	0.41
	CoClustering	0.43	0.44	0.44
	SLIM	0.41	0.41	0.42
NoCO_2 data	Global Average	0.34	0.35	0.35
	ItemKNN	0.37	0.37	0.37
	UserKNN	**0.47**	0.46	0.46
	SVD	0.44	0.43	0.43
	SVD++	0.44	0.44	0.44
	CoClustering	0.46	**0.47**	**0.47**
	SLIM	0.38	0.38	0.38
All data	Global Average	0.30	0.33	0.34
	ItemKNN	0.40	0.40	0.41
	UserKNN	0.45	0.44	0.44
	SVD	**0.48**	**0.48**	**0.48**
	SVD++	0.47	0.46	0.46
	CoClustering	0.45	0.44	0.44
	SLIM	0.44	0.44	0.43

Table 6.2.: Average performance in terms of GNDCG of popular recommender system algorithms trained on the CarEmissions dataset.

models in terms of predictive accuracy. Additionally, the CO_2 model exhibits superior performance compared to the last model in the lineup.

Upon analyzing the results of the three metrics, it becomes evident that the model trained on the entire dataset showcases the best performance. This can be primarily attributed to the availability of a larger dataset, which provides the model with a more comprehensive and diverse range of training samples. With a greater volume of data to learn from, the model gains a deeper understanding of the underlying patterns and intricacies, resulting in improved predictions. Now, if we remove the model trained on the entire dataset from consideration and focus solely on the CarEmissionsCO_2 and CarEmissionNoCO_2 models, a notable trend emerges. The models trained on CarEmissionsCO_2 consistently display higher metric values across the board, suggesting their more reliable and accurate predictions in comparison to the CarEmissionsNoCO_2 models.

6.3. SUSTAINABLE NUDGING RESULTS

Sustainable nudging, as presented in Section 2.1.3, refers to nudging towards greener recommendations using existing predictions. In this section, we dive into the results of the nudging experiment.

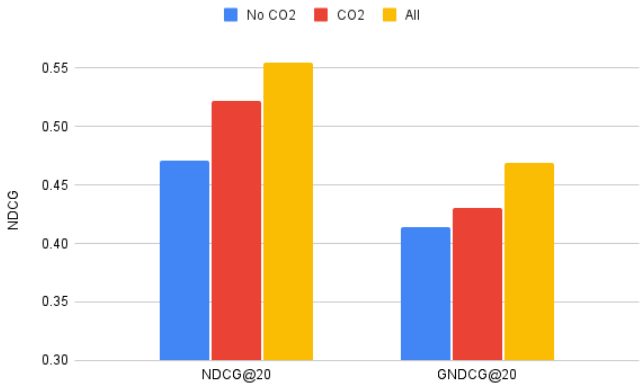


Figure 6.1.: NDCG values for best-performing recommender models for partial datasets

6.3.1. EXPERIMENTAL DESIGN

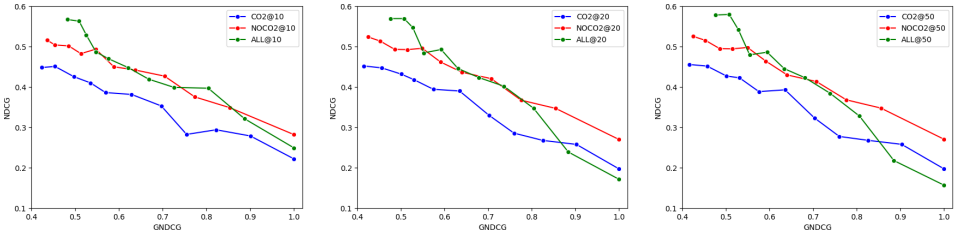
We re-use the predictions generated from the benchmarking process in the previous section. We again create rankings by ordering the interactions by their rating predictions. This means that the techniques of data processing and hyperparameter tuning remain. We measure performance using the NDCG and GNDCG. To obtain a complete overview of the influence of α on the predictions, we chose to experiment with values of alpha in the full range [0, 1], with a step size of 0.1 (i.e., [0, 0.1, 0.2 ... 0.8, 0.9, 1]).

6.3.2. EXPERIMENTAL RESULTS

In this section, we present the results of the nudging method performed. Figure 6.2a presents the nudging results of different alphas from 0 to 1 with a stepsize of 1. Here the graph depicts a trade-off between NDCG and GNDCG when varying the alpha parameter. From the results, we can see that NoCO_2 and All outperform CO_2 in terms of both NDCG and GNDCG at different alphas. Both NoCO_2 and All fluctuate with varying levels of alpha. Very similar trends can be seen in both Figures 6.2b and 6.2c.

The trade-off between NDCG and GNDCG does not follow a curve, as shown by [6]. For CarEmissions the trade-off between the two metrics is mostly linear. In Kalivert's study, it was mentioned that most items in the user-item matrix had high greenness values. With such data, it is much easier to obtain a trade-off that maintains accuracy and greenness. In the case of CarEmissions, the CO_2 values of items are normally distributed. Hence, by looking at the metric trade-off graphs, the final alpha value cannot be concluded visually.

We observe that the slope of the lines in Figure **??** is not 45 degrees, implying that in certain cases, a possible shift in alpha values may slightly decrease the NDCG while significantly increasing the GNDCG. Determining the ideal alpha value is beyond the scope of this thesis, for reasons that will be discussed in Chapter 7.



(a) SVD@10 (b) SVD@20 (c) SVD@50

Figure 6.2.: SVD performance for CarEmissions datasets.

6.4. DISCUSSION

The study in this chapter aimed to evaluate and compare various recommender system algorithms on the CarEmissions dataset. The benchmarking results revealed that SVD consistently outperformed other models, achieving the highest NDCG scores at different lengths for both CarEmissionsCO_2 and CarEmissionsNoCO_2. The lower performance of ItemKNN and UserKNN on these partial datasets can be attributed to the high sparsity of the data. The matrix factorization approaches, including SVD, SVD++, and SLIM, demonstrated superior performance due to their ability to generalize better over the entire interaction matrix.

Interestingly, the models trained on CarEmissionsNoCO_2 consistently outperformed those trained on CarEmissionsCO_2. One potential explanation for this trend is that the inclusion of CO_2 emission values in the recommendations adds an additional variable that may influence users' preferences. The variance introduced by CO_2 values might result in reduced prediction accuracy compared to the NoCO_2 dataset.

The study also introduced the concept of sustainable nudging, inspired by the work of [6]. The proposed nudging approach aimed to achieve a trade-off between prediction accuracy (NDCG) and greenness (GNDCG) by varying the alpha parameter. The results demonstrated that both CarEmissionsNoCO_2 and the combined CarEmissions dataset (all data) outperformed CarEmissionsCO_2 across various alpha values. However, unlike the results observed in the previous study, the trade-off between NDCG and GNDCG for CarEmissions followed a more linear pattern. This suggests that CarEmissions contains more diverse data with various CO_2 emission values, leading to a less predictable trade-off.

In conclusion, the benchmarking results identified SVD as the best-performing algorithm for the CarEmissions dataset. The nudging approach further highlighted the potential for improving the sustainability of recommendations by fine-tuning the trade-off between accuracy and greenness. The findings offer valuable insights into building more sustainable recommender systems that take environmental considerations into account while delivering high-quality recommendations to users. However, the study also emphasizes the importance of considering the unique characteristics of the dataset when implementing sustainable nudging, as different datasets may exhibit different trade-off patterns between NDCG and GNDCG.

7

Discussion

In this chapter, we offer a concise overview of our study and delve into both the central research query and its accompanying sub-questions. Furthermore, we address the constraints and practical factors, concluding our investigation by outlining potential avenues for future exploration.

7.1. Thesis Summary

In this thesis, we built the CarEmissions dataset and explored its impact on generating sustainable recommendations. In Chapter 1, we motivate the work done in this study and introduce the research questions. In Chapter 2, we explained the background knowledge required for this thesis. Recommendation systems, the various underlying algorithms, and evaluation metrics are presented first. Then we cover the workings of the strategy used to nudge towards sustainable recommendations. Lastly, we also explain how lasso regression and stratified sampling are performed. In Chapter 3, we reviewed existing literature that relates to our work. More precisely, we cover current sustainability trends in e-commerce, how artificial intelligence is being used to address climate change, and various different aspects of recommender systems and their evaluation.

Chapter 4 introduces the CarEmissions dataset, where we provide an overview of the dataset and compare it with other widely used datasets in the field of recommender systems. It can be inferred that the CarEmissions dataset is consistent with other datasets used for user studies in research and shares similar characteristics with popular recommender system datasets. We also create visualizations of the different properties associated with users and items found in the dataset.

Moving on to Chapter 5, we delve deeper into the process of collecting and analyzing data. We discuss the method of reducing the size of the original car dataset through stratified sampling. Additionally, we outline the phases of data collection as experienced by participants during their involvement in the study. This chapter also sheds light on the reasoning behind specific design choices made during the study. Finally, we conduct a group lasso regression to examine the potential correlation between various user demographics, car-related data, and the ratings assigned to specific cars. The results indicate that no significant correlations of this nature exist within the dataset.

In Chapter 6 we focus on using the new dataset in recommendation algorithms. We

compare the performance of global average, user-based knn, item-based knn, SVD, SVD++, co-clustering, and SLIM algorithms. The results showed that, in terms of NDCG, SVD outperforms other algorithms. In terms of the greenness of recommendations, no significant difference is seen between the algorithms. According to the nudging strategy's suggestion for enhancing the greenness of ranked lists of recommendations, we rerank the recommendations.

7.2. ANSWERS TO RESEARCH QUESTIONS

In this thesis, the main research question was: **What are the implications of showing product CO_2 emissions values to users for promoting sustainable choices through recommendations?** We also posed the following sub-questions:

1. *Is there a significant relationship between user demographics, knowledge, and ratings provided to items?*

2. *Does the greenness and accuracy of products recommended differ between the two datasets (CO_2 vs. No CO_2) displayed?*

To answer the research questions, we developed the CarEmissions dataset as described in Chapter 4. This dataset can be divided into two distinct parts, namely CarEmissionsCO_2 and CarEmissionsNoCO_2. One contains car-user interactions with CO_2 values displayed and the other without, respectively. This dataset was created as a base to determine whether showing the CO_2 values of products influences user rating behavior in any way.

This resulting dataset consisted of 136 users, 396 cars, and a total of 9650 user-item interactions. The interactions are expressed as numerical ratings between 0 and 5. We compared the dataset with several, often user recommender and user study datasets. Along with this, we also collected various user demographic data to analyze any patterns with respect to ratings given to items.

The answer to the first sub-question is analyzed in Chapter 5. Prior to collection, we expected to observe some correlation patterns between variables such as age, field of work, familiarity with cars, and the importance given to emissions with respect to ratings. Through the means of group lasso regression, we were able to investigate the correlation between user demographics and car properties on the rating given to cars. Here we conclude that there is no significant correlation between demographic variables and ratings in the CarEmissions dataset.

The second sub-question can be answered by analyzing the overarching results from Chapter 6. We benchmarked various recommender system algorithms such as global average, item-knn, user-knn, SVD, SVD++, co-clustering, and SLIM on the CarEmissions dataset. Here we tested every algorithm trained with the entire dataset (CarEmissionsAll), the part of the dataset where CO_2 emission values are displayed to the user (CarEmissionsCO_2), and the part of the dataset where CO_2 emission values are omitted (CarEmissionsNoCO_2). Models trained on CarEmissionsAll have the highest performance. Consistently, across all benchmarked algorithms, we observe that models trained with CarEmissionsNoCO_2 outperform models trained with CarEmissionsCO_2 both in terms of greenness and accuracy. The primary explanation for this is that the in-

clusion of CO_2 emission values introduces an additional variable that may influence user preferences.

To answer the overarching research question, we aim to explore the accuracy greenness trade-off and how it can be of interest to the end user receiving recommendations. Since the SVD algorithm produced the best results in terms of accuracy (NDCG) and greenness (GNDCG), we used this algorithm to dive deeper into the performance of the datasets. In Section 6.3, we employ a nudging strategy [6], where the value of the alpha parameter determines the weighting of greenness vs. accuracy of the recommendations.

We observe that SVD models trained on CarEmissionsNoCO_2 once again perform better than CarEmissionsCO_2. These results might suggest that displaying the CO_2 emission values, causes a higher variance in user-item ratings, thus making it more difficult for recommender systems to produce relevant recommendations. This does not necessarily imply that displaying emission values cannot be deemed useful. It simply implies that adding another item property, one as significant as emission values, can greatly impact user rating behavior in unexpected ways. This causes ratings to become more unpredictable, as users may interpret the same information in different ways.

To answer the main research question, in the case of CarEmissions, displaying emission values results in less relevant recommendations, both in terms of accuracy and greenness. This, however, is the first study to explore the effects of CO_2 emission values in a user-study setting. Therefore, the next section identifies some limitations, practical conditions, and future work.

7.3. LIMITATIONS & FUTURE WORK

In this section, we discuss some of the limitations and practical considerations regarding this research. We also provide directions for future work.

7.3.1. LIMITATIONS

Being one of the pioneering studies in the field of sustainable recommender systems, we were required to create our own dataset for this research. The domain of the dataset is cars, as the CO_2 emission values were accurately given by the manufacturers. The E-commerce market, however, has many more domains; therefore, a direct generalization is hard to prove.

Furthermore, the users that took part in the study were recruited through a convenience sample of connections. This convenience sample might not be representative of the general population that makes use of e-commerce recommendation engines. This also impacts the correlation models in Chapter 5 as demographic data from a convenience sample of 136 users might not contain enough variance to identify patterns.

Lastly, users were simply asked to rate items, which does not imply their intention to make future purchases. For example, a user may give a 5-star rating to a Porsche Carrera but have no intention of purchasing it in the near future. Hence, the ratings collection represents user preference or liking rather than user purchase intention.

7.3.2. FUTURE WORK

To further examine the effect of displaying emission values to nudge towards sustainable recommendations, we argue that it is important to gather more detailed data. It is important to separate the effect of providing CO_2 information from the effect of simply adding one additional piece of information. This can perhaps be done by removing a piece of information and substituting it with CO_2 information. The aim is to isolate the effect of showing emissions information from the effect of showing any additional information.

Moreover, in regards to nudging towards more sustainable recommendations, a trade-off alpha value was not clear. A method for automatically determining such a trade-off, generalizable to datasets, would be worth investigating. A possible way to achieve this could be by using evolutionary algorithm techniques to determine Pareto-optimal solutions.

Another approach to finding the right alpha trade-off would be to conduct an online evaluation of the recommendations. This would entail A/B testing recommendations with users. Each batch of users would test recommendations with a different value of alpha and indicate their satisfaction with the recommendations. Such an evaluation setup at the right scale could prove a big leap in sustainable recommendation research.

Lastly, it is important to expand data collection to a larger scale with various different categories for items. Since sustainable recommendation systems are a relatively new field, gathering more data is arguably the most important task to advance research. This would help the community of researchers develop innovative solutions without the worry of creating and benchmarking new datasets.

7.4. BROADER IMPACT

The implications of this thesis extend beyond the immediate topic of sustainable recommendation systems in the domain of car emissions and have broader significance for both the scientific community and society at large.

Scientifically, this research contributes to the emerging field of sustainable recommender systems by shedding light on the complex interplay between information provision, user behavior, and recommendation accuracy. The concept of nudging to promote sustainable choices in recommendation systems can serve as a foundation for further studies exploring behavioral economics and decision-making in the context of e-commerce and information presentation. Additionally, the dataset creation process and analysis techniques presented here can serve as a reference for researchers seeking to conduct user studies and build new datasets in the field of recommender systems.

Beyond the realm of recommender systems, the findings have implications for understanding user behavior and decision-making in the context of sustainable consumption. The research underscores the challenges of introducing new information variables to recommendation algorithms and highlights the importance of considering user response variability. This insight could inspire further investigations into the psychology of sustainable choices and the effects of information presentation on user decision-making in various domains.

From a societal perspective, the thesis raises awareness about the potential unintended consequences of displaying certain information to users. The results emphasize the

need for careful consideration when incorporating sustainability-related information into e-commerce platforms, as it can impact user preferences and influence recommendation outcomes. This has implications for businesses and policymakers aiming to promote sustainable consumption through technological interventions. Ethical concerns related to potential biases introduced by the display of CO_2 emissions values and their impact on user behavior should be taken into account when designing and deploying such recommendation systems.

In conclusion, this thesis contributes to the scientific community by advancing the understanding of sustainable recommender systems and their implications for user behavior and decision-making. The research findings can be extended beyond the specific domain of car emissions, informing future studies in diverse fields involving user preferences and recommendation algorithms. Furthermore, the societal implications underscore the importance of responsible design and implementation of technology to encourage sustainable choices while raising awareness about potential biases and ethical considerations.

A

Data Collection Platform

The user-item interactions data was collected through a website developed and hosted on a server. The website can be found at `ratingcroudsourcing.ewi.tudelft.nl`. Although services such as MTurk and others exist, creating the website allowed for a smooth participant experience, where they do not have to create any additional accounts or sign up to take participate. In addition to this, by creating the website in house we could assure that all the data was securely stored on the TU Delft servers, thus minimizing privacy risks.

The back end of the website was written in Flask which is a Python framework for web development. The front end was mainly HTML/CSS with additional JavaScript functionalities. MySQL server was used for database management and storage.

Figure A.1 shows the initial questions posed to users in order to collect some demographic data. This includes information such as age, gender, level of education, field of work/study, country of origin, retention of license, access to car, frequency of car use, reasons for car use, and familiarity with cars.

In Figure A.2 we can see a snapshot of the user front end when rating a single car. In this snapshot the CO_2 emissions score is provided, but depending on the phase this information can be omitted. The continue button is only activated once the users have rated at least 20 cars, indicating they are ready to move onto the next phase. For every single rating the time spent per user on each car is also recorded. This is to aid in further analysis of the data gather which will be discussed in the results section.

Finally, Figure A.3 shows the questions posed to users after the survey is finished. This is to gather some additional information regarding their views on sustainability. These questions are posed at the end, as to not let users be biased during the study.

A

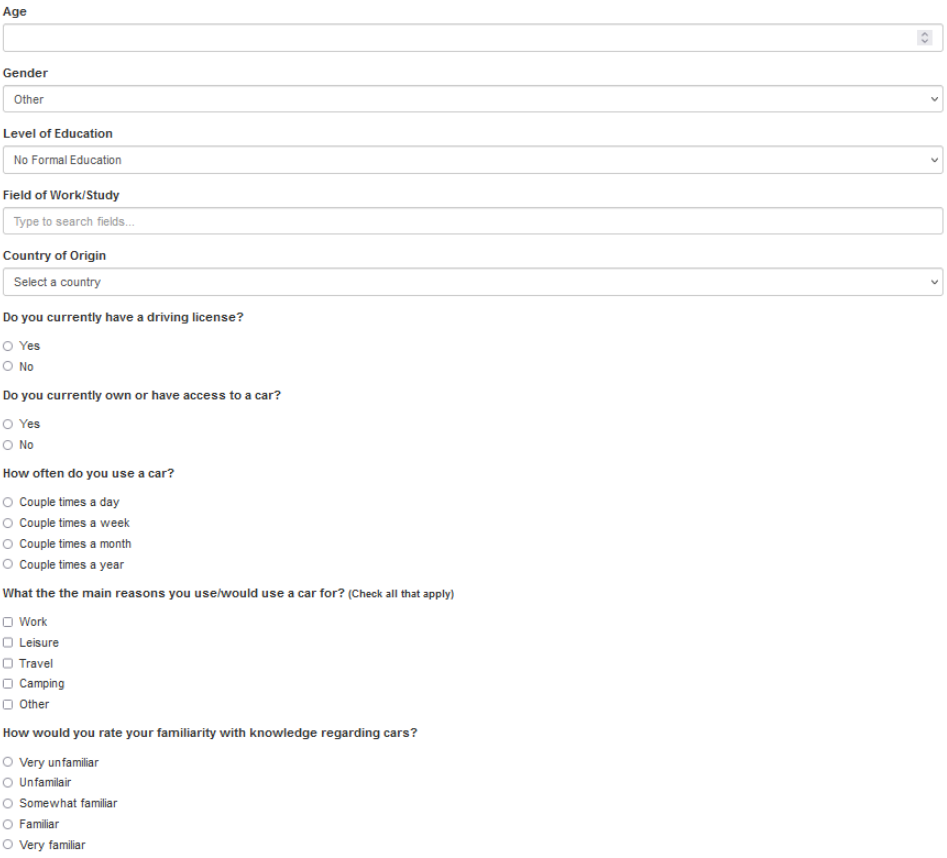


Figure A.1.: Snapshot of exploratory data collection page on the developed website

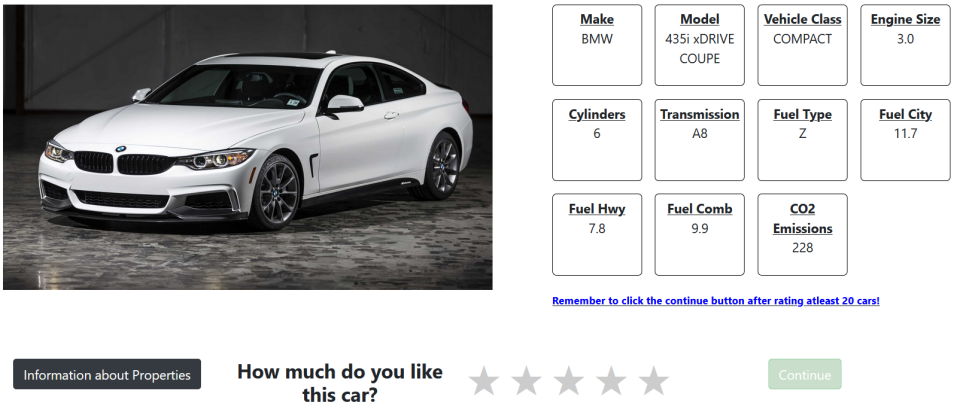


Figure A.2.: Snapshot of front end when rating a car

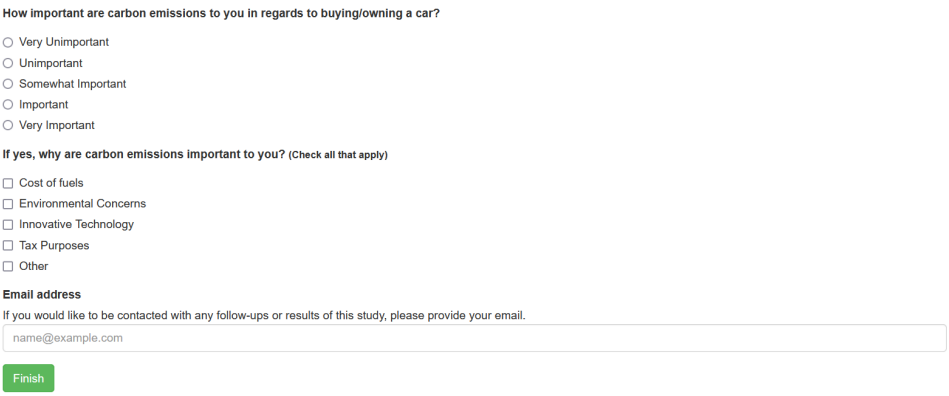


Figure A.3.: Snapshot of data collection in final phase

B

HYPER-PARAMETER SELECTION

This appendix provides the hyperparameter selection on which a grid-search is performed to optimize the algorithms' performance. Bold values are the selected, most performing, parameters.

k	20	40	60	80	100

Table B.1.: Parameter included in grid search for optimizing Item-knn and User-knn.

	CarEmissionsCO_2	**CarEmissionsNoCO_2**	**CarEmissionsAll**
k	40	60	40

Table B.2.: Selected parameter for Item-knn, for each partial dataset.

	CarEmissionsCO_2	**CarEmissionsNoCO_2**	**CarEmissionsAll**
k	60	60	40

Table B.3.: Selected parameter for User-knn, for each partial dataset.

Number of Factors	**Number of Epochs**	**Learning Rate**	**Regularization term**
20	20	0.1	0.1
50	50	0.01	0.01
100	80	0.001	0.001
150	100	0.0001	0.0001

Table B.4.: Parameters included in grid search for optimizing SVD and SVD++.

B

	CarEmissionsCO_2	CarEmissionsNoCO_2	CarEmissionsAll
Number of Factors	150	150	100
Number of Epochs	150	150	100
Learning Rate	0.01	0.01	0.01
Regularization Term	0.1	0.1	0.1

Table B.5.: Selected parameters for SVD, for each partial dataset.

	CarEmissionsCO_2	CarEmissionsNoCO_2	CarEmissionsAll
Number of Factors	150	150	100
Number of Epochs	100	150	50
Learning Rate	0.01	0.01	0.01
Regularization Term	0.1	0.1	0.1

Table B.6.: Selected parameters for SVD++, for each partial dataset.

Number of User Clusters	Number of Item Clusters	Number of Epochs
3	3	20
6	6	50
12	12	100
24	24	150

Table B.7.: Parameters included in grid search for optimizing Co-Clustering.

	CarEmissionsCO_2	CarEmissionsNoCO_2	CarEmissionsAll
Number of User Clusters	24	24	24
Number of Item Clusters	3	6	3
Number of Epochs	100	100	50

Table B.8.: Selected parameters for Co-Clustering, for each partial dataset.

L1 Regularization	L2 Regularization
0.005	0.005
0.05	0.05
0.5	0.5

Table B.9.: Parameters included in grid search for optimizing SLIM.

	CarEmissionsCO_2	CarEmissionsNoCO_2	CarEmissionsAll
L1 Regularization	0.05	0.05	0.05
L2 Regularization	0.05	0.05	0.05

Table B.10.: Selected parameters for SLIM, for each partial dataset.

BIBLIOGRAPHY

[1] S. Arrhenius. “XXXI. On the influence of carbonic acid in the air upon the temperature of the ground”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41.251 (1896), pp. 237–276.

[2] *The Surprising Case for Stronger E-commerce Growth*. en. URL: https://www.morganstanley.com/ideas/global-ecommerce-growth-forecast-2022.

[3] W. Yu, A. Hassan, and M. Adhikariparajuli. “How Did Amazon Achieve CSR and Some Sustainable Development Goals (SDGs)—Climate Change, Circular Economy, Water Resources and Employee Rights during COVID-19?” In: *Journal of Risk and Financial Management* 15.8 (2022), p. 364.

[4] J. B. Schafer, J. Konstan, and J. Riedl. “Recommender systems in e-commerce”. In: *Proceedings of the 1st ACM conference on Electronic commerce*. 1999, pp. 158–166.

[5] D. Lee and K. Hosanagar. “Impact of recommender systems on sales volume and diversity”. In: (2014).

[6] I. Kalivert. “Green Recommender Systems”. In: (2023).

[7] P. Resnick and H. R. Varian. “Recommender systems”. In: *Communications of the ACM* 40.3 (1997), pp. 56–58.

[8] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou. “Recommender systems”. In: *Physics reports* 519.1 (2012), pp. 1–49.

[9] A. M. Rashid, G. Karypis, and J. Riedl. “Influence in ratings-based recommender systems: An algorithm-independent approach”. In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM. 2005, pp. 556–560.

[10] R. H. Singh, S. Maurya, T. Tripathi, T. Narula, and G. Srivastav. “Movie recommendation system using cosine similarity and KNN”. In: *International Journal of Engineering and Advanced Technology* 9.5 (2020), pp. 556–559.

[11] B. Sarwar, G. Karypis, J. Konstan, and J. T. Riedl. “Application of dimensionality reduction in recommender system-a case study”. In: (2000).

[12] X. Zhou, J. He, G. Huang, and Y. Zhang. “SVD-based incremental approaches for recommender systems”. In: *Journal of Computer and System Sciences* 81.4 (2015), pp. 717–733.

[13] R. Mehta and K. Rana. “A review on matrix factorization techniques in recommender systems”. In: *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*. IEEE. 2017, pp. 269–274.

[14] Y. Koren, R. Bell, and C. Volinsky. “Matrix factorization techniques for recommender systems”. In: *Computer* 42.8 (2009), pp. 30–37.

[15] Y. Miao, B. Zhang, Y. Yi, and J. Lin. “Application of improved reweighted singular value decomposition for gearbox fault diagnosis based on built-in encoder information”. In: *Measurement* 168 (2021), p. 108295.

[16] W. Shi, L. Wang, and J. Qin. “User embedding for rating prediction in SVD++-based collaborative filtering”. In: *Symmetry* 12.1 (2020), p. 121.

[17] T. George and S. Merugu. “A scalable collaborative filtering framework based on co-clustering”. In: *Fifth IEEE International Conference on Data Mining (ICDM’05)*. IEEE. 2005, 4–pp.

[18] X. Ning and G. Karypis. “Slim: Sparse linear methods for top-n recommender systems”. In: *2011 IEEE 11th international conference on data mining*. IEEE. 2011, pp. 497–506.

[19] P. Cremonesi, Y. Koren, and R. Turrin. “Performance of recommender algorithms on top-n recommendation tasks”. In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 39–46.

[20] J. Ranstam and J. Cook. “LASSO regression”. In: *Journal of British Surgery* 105.10 (2018), pp. 1348–1348.

[21] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[22] L. Meier, S. Van De Geer, and P. Bühlmann. “The group lasso for logistic regression”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.1 (2008), pp. 53–71.

[23] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. “A sparse-group lasso”. In: *Journal of computational and graphical statistics* 22.2 (2013), pp. 231–245.

[24] V. L. Parsons. “Stratified sampling”. In: *Wiley StatsRef: Statistics Reference Online* (2014), pp. 1–11.

[25] P. E. McKnight and J. Najab. “Mann-Whitney U Test”. In: *The Corsini encyclopedia of psychology* (2010), pp. 1–1.

[26] T. W. MacFarland, J. M. Yates, T. W. MacFarland, and J. M. Yates. “Mann–whitney u test”. In: *Introduction to nonparametric statistics for the biological sciences using R* (2016), pp. 103–132.

[27] M. H. Dore. “Climate change and changes in global precipitation patterns: what do we know?” In: *Environment international* 31.8 (2005), pp. 1167–1181.

[28] I. P. O. C. Change. “Climate change 2007: The physical science basis”. In: *Agenda* 6.07 (2007), p. 333.

[29] A. Molla and P. S. Licker. “eCommerce adoption in developing countries: a model and instrument”. In: *Information & management* 42.6 (2005), pp. 877–899.

[30] J. J. Masele. “Towards sustainable tourism: Utilizing E-Commerce applications for minimizing impacts of climate change”. In: *Information technologies in environmental Engineering: New trends and challenges* (2011), pp. 445–459.

[31] D. Dubisz, P. Golinska-Dawson, and P. Zawodny. “Measuring CO2 Emissions in E-Commerce Deliveries: From Empirical Studies to a New Calculation Approach”. In: *Sustainability* 14.23 (2022), p. 16085.

[32] S. Escursell, P. Llorach-Massana, and M. B. Roncero. “Sustainability in e-commerce packaging: A review”. In: *Journal of cleaner production* 280 (2021), p. 124314.

[33] D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

[34] V. Kakani, V. H. Nguyen, B. P. Kumar, H. Kim, and V. R. Pasupuleti. “A critical review on computer vision and artificial intelligence in food industry”. In: *Journal of Agriculture and Food Research* 2 (2020), p. 100033.

[35] F. Raponi, R. Moscetti, D. Monarca, A. Colantoni, and R. Massantini. “Monitoring and optimization of the process of drying fruits and vegetables using computer vision: A review”. In: *Sustainability* 9.11 (2017), p. 2009.

[36] O. C. Ghergan, D. Drăghicescu, I. Iosim, and P. A. NECŞA. “THE ROLE OF COMPUTER VISION IN SUSTAINABLE AGRICULTURE.” In: *Agricultural Management/Lucrari Stiintifice Seria I, Management Agricol* 23.2 (2021).

[37] T. Esau, Q. Zaman, D. Groulx, A. Farooque, A. Schumann, and Y. Chang. “Machine vision smart sprayer for spot-application of agrochemical in wild blueberry fields”. In: *Precision agriculture* 19 (2018), pp. 770–788.

[38] D. Tomaselli. “Automated Recycling System Using Computer Vision”. In: *ECE-498: Capstone Design Project Advisor: Prof. Cotter November* 25 (2019).

[39] M. I. Jordan and T. M. Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260.

[40] R. Ahmed, V. Sreeram, Y. Mishra, and M. Arif. “A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization”. In: *Renewable and Sustainable Energy Reviews* 124 (2020), p. 109792.

[41] J. Mathe, N. Miolane, N. Sebastien, and J. Lequeux. “PVNet: A LRCN architecture for spatio-temporal photovoltaic PowerForecasting from numerical weather prediction”. In: *arXiv preprint arXiv:1902.01453* (2019).

[42] J. Maldonado-Correa, J. Solano, and M. Rojas-Moncayo. “Wind power forecasting: A systematic literature review”. In: *Wind Engineering* 45.2 (2021), pp. 413–426.

[43] Y. Liu, Q. Zhou, and G. Cui. “Machine learning boosting the development of advanced lithium batteries”. In: *Small Methods* 5.8 (2021), p. 2100442.

[44] C. P. Gomes, J. Bai, Y. Xue, J. Björck, B. Rappazzo, S. Ament, R. Bernstein, S. Kong, S. K. Suram, R. B. van Dover, *et al.* “CRYSTAL: a multi-agent AI system for automated mapping of materials’ crystal structures”. In: *MRS Communications* 9.2 (2019), pp. 600–608.

[45] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar. “Towards the deployment of machine learning solutions in network traffic classification: A systematic survey”. In: *IEEE Communications Surveys & Tutorials* 21.2 (2018), pp. 1988–2014.

B

[46] Y. Li and C. Shahabi. “A brief overview of machine learning methods for short-term traffic forecasting and future directions”. In: *Sigspatial Special* 10.1 (2018), pp. 3–9.

[47] C. Sun, N. Azari, and C. Turakhia. “Gallery: A Machine Learning Model Management System at Uber.” In: *EDBT*. Vol. 20. 2020, pp. 474–485.

[48] T. Seo, T. Kusakabe, H. Gotoh, and Y. Asakura. “Interactive online machine learning approach for activity-travel survey”. In: *Transportation Research Part B: Methodological* 123 (2019), pp. 362–373.

[49] M. Maghrebi, A. Abbasi, and S. T. Waller. “Transportation application of social media: Travel mode extraction”. In: *2016 IEEE 19th International Conference on intelligent transportation systems (ITSC)*. IEEE. 2016, pp. 1648–1653.

[50] R. Regue and W. Recker. “Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem”. In: *Transportation Research Part E: Logistics and Transportation Review* 72 (2014), pp. 192–209.

[51] M. Altinkaya and M. Zontul. “Urban bus arrival time prediction: A review of computational models”. In: *International Journal of Recent Technology and Engineering (IJRTE)* 2.4 (2013), pp. 164–169.

[52] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. Roper. “Machine learning for estimation of building energy consumption and performance: a review”. In: *Visualization in Engineering* 6 (2018), pp. 1–20.

[53] J. Kreider, D. Claridge, P. Curtiss, R. Dodier, J. Haberl, and M. Krarti. “Building energy use prediction and system identification using recurrent neural networks”. In: (1995).

[54] P. Rashidi and D. J. Cook. “Keeping the resident in the loop: Adapting the smart home to the user”. In: *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans* 39.5 (2009), pp. 949–959.

[55] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, *et al.* “Tackling climate change with machine learning”. In: *ACM Computing Surveys (CSUR)* 55.2 (2022), pp. 1–96.

[56] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani. “An energy-efficient mobile recommender system”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 899–908.

[57] L. Quijano-Sánchez, I. Cantador, M. E. Cortés-Cediel, and O. Gil. “Recommender systems for smart cities”. In: *Information systems* 92 (2020), p. 101545.

[58] K. Chaudhari and A. Thakkar. “A comprehensive survey on travel recommender systems”. In: *Archives of Computational Methods in Engineering* 27 (2020), pp. 1545–1571.

[59] H. U. Rehman Khan, C. K. Lim, M. F. Ahmed, K. L. Tan, and M. Bin Mokhtar. “Systematic review of contextual suggestion and recommendation systems for sustainable e-tourism”. In: *Sustainability* 13.15 (2021), p. 8141.

[60] H. Hwangbo and Y. Kim. “Session-based recommender system for sustainable digital marketing”. In: *Sustainability* 11.12 (2019), p. 3336.

[61] Y. Guo, C. Yin, M. Li, X. Ren, and P. Liu. “Mobile e-Commerce Recommendation System Based on Multi-Source Information Fusion for Sustainable e-Business”. en. In: *Sustainability* 10.11 (Jan. 2018), p. 147. ISSN: 2071-1050. DOI: 10.3390/su10010147.

[62] S. Tomkins, S. Isley, B. London, and L. Getoor. “Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations”. In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 214–218.

[63] A. Gunawardana, G. Shani, and S. Yogev. “Evaluating Recommender Systems”. en. In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, and B. Shapira. New York, NY: Springer US, 2022, pp. 547–601. ISBN: 978-1-07-162197-4. DOI: 10.1007/978-1-0716-2197-4_15. URL: https://doi.org/10.1007/978-1-0716-2197-4_15.

[64] C. Trattner and D. Elsweiler. “An Evaluation of Recommendation Algorithms for Online Recipe Portals”. en. In: (2019).

[65] J. Freyne and S. Berkovsky. “Intelligent food planning: personalized recipe recommendation”. In: *Proceedings of the 15th international conference on Intelligent user interfaces*. IUI ’10. New York, NY, USA: Association for Computing Machinery, Feb. 2010, pp. 321–324. ISBN: 978-1-60558-515-4. DOI: 10.1145/1719970.1720021. URL: https://doi.org/10.1145/1719970.1720021.

[66] G. Lekakos and P. Caravelas. “A hybrid approach for movie recommendation”. en. In: 36 (Jan. 2008), pp. 55–70. ISSN: 1573-7721. DOI: 10.1007/s11042-006-0082-7.

[67] A. van den Oord, S. Dieleman, and B. Schrauwen. “Deep content-based music recommendation”. In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/hash/b3ba8f1bee1238a2f37603d90b58898d-Abstract.html.

[68] M. Schedl. “Deep Learning in Music Recommendation Systems”. In: 5 (2019). ISSN: 2297-4687. URL: https://www.frontiersin.org/articles/10.3389/fams.2019.00044.

[69] X. Wang and Y. Wang. “Improving Content-based and Hybrid Music Recommendation using Deep Learning”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. MM ’14. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 627–636. ISBN: 978-1-4503-3063-3. DOI: 10.1145/2647868.2654940. URL: https://doi.org/10.1145/2647868.2654940.

[70] K. Swearingen and R. Sinha. “Interaction Design for Recommender Systems”. en. In: ().

[71] K. Krauth, S. Dean, A. Zhao, W. Guo, M. Curmei, B. Recht, and M. I. Jordan. “Do Offline Metrics Predict Online Performance in Recommender Systems?” In: arXiv:2011.07931 (Nov. 2020). arXiv:2011.07931 [cs]. DOI: 10.48550/arXiv.2011.07931. URL: http://arxiv.org/abs/2011.07931.

[72] S. M. McNee, J. Riedl, and J. A. Konstan. “Being accurate is not enough: how accuracy metrics have hurt recommender systems”. In: *CHI’06 extended abstracts on Human factors in computing systems*. 2006, pp. 1097–1101.

[73] M. Kaminskas and D. Bridge. “Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems”. In: *ACM Transactions on Interactive Intelligent Systems* 7.1 (Dec. 2016), 2:1–2:42. ISSN: 2160-6455. DOI: 10.1145/2926720.

[74] P. Chandar, B. St. Thomas, L. Maystre, V. Pappu, R. Sanchis-Ojeda, T. Wu, B. Carterette, M. Lalmas, and T. Jebara. “Using Survival Models to Estimate User Engagement in Online Experiments”. In: WWW ’22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 3186–3195. ISBN: 978-1-4503-9096-5. DOI: 10.1145/3485447.3512038. URL: https://doi.org/10.1145/3485447.3512038.

[75] S. Graham, J.-K. Min, and T. Wu. “Microsoft recommenders: tools to accelerate developing recommender systems”. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. RecSys ’19. New York, NY, USA: Association for Computing Machinery, Sept. 2019, pp. 542–543. ISBN: 978-1-4503-6243-6. DOI: 10.1145/3298689.3346967. URL: https://doi.org/10.1145/3298689.3346967.

[76] X. Amatriain and D. Agarwal. “Tutorial: Lessons Learned from Building Real-life Recommender Systems”. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys ’16. New York, NY, USA: Association for Computing Machinery, Sept. 2016, p. 433. ISBN: 978-1-4503-4035-9. DOI: 10.1145/2959100.2959194. URL: https://doi.org/10.1145/2959100.2959194.

[77] M. D. Ekstrand, J. T. Riedl, J. A. Konstan, *et al.* “Collaborative filtering recommender systems”. In: *Foundations and Trends® in Human–Computer Interaction* 4.2 (2011), pp. 81–173.

[78] R. Kohavi and R. Longbotham. “Online Controlled Experiments and A/B Testing”. en. In: ed. by C. Sammut and G. I. Webb. Boston, MA: Springer US, 2017, pp. 922–929. ISBN: 978-1-4899-7685-7. DOI: 10.1007/978-1-4899-7687-1_891. URL: http://link.springer.com/10.1007/978-1-4899-7687-1_891.

[79] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. “Trustworthy online controlled experiments: five puzzling outcomes explained”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD ’12. New York, NY, USA: Association for Computing Machinery, Aug. 2012, pp. 786–794. ISBN: 978-1-4503-1462-6. DOI: 10.1145/2339530.2339653. URL: https://doi.org/10.1145/2339530.2339653.

[80] M. Esteller-Cucala, V. Fernandez, and D. Villuendas. “Evaluating Personalization: The AB Testing Pitfalls Companies Might Not Be Aware of—A Spotlight on the Automotive Sector Websites”. In: *Frontiers in Artificial Intelligence* 3 (2020). ISSN: 2624-8212. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00020.

[81] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger. "Research paper recommender system evaluation: a quantitative literature survey". In: *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*. RepSys '13. New York, NY, USA: Association for Computing Machinery, Oct. 2013, pp. 15–22. ISBN: 978-1-4503-2465-6. DOI: 10.1145/2532508.2532512. URL: https://doi.org/10.1145/2532508.2532512.

[82] K. Swearingen and R. Sinha. "Interaction Design for Recommender Systems". en. In: ().

[83] R. Naughton and X. Lin. "Recommender Systems: Investigating the Impact of Recommendations on User Choices and Behaviors". en. In: 612 (2010).

[84] N. Mujere. *Sampling in Research*. en. Chap. 2016. DOI: 10.4018/978-1-5225-0007-0.ch006. URL: https://www.igi-global.com/chapter/sampling-in-research/www.igi-global.com/chapter/sampling-in-research/147769.

[85] S. Senecal and J. Nantel. "The influence of online product recommendations on consumers' online choices". en. In: *Journal of Retailing* 80.2 (Jan. 2004), pp. 159–169. ISSN: 0022-4359. DOI: 10.1016/j.jretai.2004.04.001.

[86] S. S. Sohail, J. Siddiqui, and R. Ali. "A comprehensive approach for the evaluation of recommender systems using implicit feedback". en. In: *International Journal of Information Technology* 11.3 (Sept. 2019), pp. 549–567. ISSN: 2511-2112. DOI: 10.1007/s41870-018-0202-4.

[87] B. Loepp, T. Donkers, T. Kleemann, and J. Ziegler. "Impact of item consumption on assessment of recommendations in user studies". en. In: *Proceedings of the 12th ACM Conference on Recommender Systems*. Vancouver British Columbia Canada: ACM, Sept. 2018, pp. 49–53. ISBN: 978-1-4503-5901-6. DOI: 10.1145/3240323.3240375. URL: https://dl.acm.org/doi/10.1145/3240323.3240375.

[88] J. Beel and S. Langer. "A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems". en. In: ed. by S. Kapidakis, C. Mazurek, and M. Werla. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 153–168. ISBN: 978-3-319-24592-8. DOI: 10.1007/978-3-319-24592-8_12.

[89] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp. "A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation". In: *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*. RepSys '13. New York, NY, USA: Association for Computing Machinery, Oct. 2013, pp. 7–14. ISBN: 978-1-4503-2465-6. DOI: 10.1145/2532508.2532511. URL: https://doi.org/10.1145/2532508.2532511.

[90] G. G. Gebremeskel and A. P. de Vries. "Recommender Systems Evaluations: Offline, Online, Time and A/A Test". en. In: ().

[91] F. Ricci, D. Massimo, and A. De Angeli. "Challenges for Recommender Systems Evaluation". In: *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*. CHItaly '21. New York, NY, USA: Association for Computing Machinery, July 2021, pp. 1–5. ISBN: 978-1-4503-8977-8. DOI: 10.1145/3464385.3464733. URL: https://doi.org/10.1145/3464385.3464733.

[92] S. Akuma, R. Iqbal, C. Jayne, and F. Doctor. "Comparative analysis of relevance feedback methods based on two user studies". en. In: *Computers in Human Behavior* 60 (July 2016), pp. 138–146. ISSN: 0747-5632. DOI: 10.1016/j.chb.2016.02.064.

[93] A. Kittur, E. H. Chi, and B. Suh. "Crowdsourcing user studies with Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. New York, NY, USA: Association for Computing Machinery, Apr. 2008, pp. 453–456. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357127. URL: https://doi.org/10.1145/1357054.1357127.

[94] Z. Munawar, N. Suryana, Z. Binti Sa'aya, and Y. Herdiana. "Framework With An Approach To The User As An Evaluation For The Recommender Systems". In: *2020 Fifth International Conference on Informatics and Computing (ICIC)*. Nov. 2020, pp. 1–5. DOI: 10.1109/ICIC50835.2020.9288565.

[95] J. A. Konstan and J. Riedl. "Recommender systems: from algorithms to user experience". en. In: *User Modeling and User-Adapted Interaction* 22.1 (Apr. 2012), pp. 101–123. ISSN: 1573-1391. DOI: 10.1007/s11257-011-9112-x.

[96] Z. D. Champiri, G. Mujtaba, S. S. Salim, and C. Yong Chong. "User Experience and Recommender Systems". In: Jan. 2019, pp. 1–5. DOI: 10.1109/ICOMET.2019.8673410.

[97] J. A. Adams. "Applying User Feedback during the Product Development Cycle". en. In: (1999).

[98] L. Serna-Mansoux, E. Chapotot, D. Millet, and S. Minel. "Study of user behaviour after eco-use feedback: the Green-Use Learning Cycle (GULC) as a new strategy for product eco-design". en. In: *International Journal on Interactive Design and Manufacturing (IJIDeM)* 8.1 (Feb. 2014), pp. 43–54. ISSN: 1955-2505. DOI: 10.1007/s12008-013-0192-1.

[99] J. Kurilova-Palisaitiene, L. Lindkvist, and E. Sundin. "Towards Facilitating Circular Product Life-Cycle Information Flow via Remanufacturing". en. In: *Procedia CIRP*. The 22nd CIRP Conference on Life Cycle Engineering 29 (Jan. 2015), pp. 780–785. ISSN: 2212-8271. DOI: 10.1016/j.procir.2015.02.162.

[100] P. Pu, L. Chen, and R. Hu. "A user-centric evaluation framework for recommender systems". In: *Proceedings of the fifth ACM conference on Recommender systems*. RecSys '11. New York, NY, USA: Association for Computing Machinery, Oct. 2011, pp. 157–164. ISBN: 978-1-4503-0683-6. DOI: 10.1145/2043932.2043962. URL: https://doi.org/10.1145/2043932.2043962.

[101] F. Hernández del Olmo and E. Gaudioso. "Evaluation of recommender systems: A new approach". en. In: *Expert Systems with Applications* 35.3 (Oct. 2008), pp. 790–804. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2007.07.047.

[102] V. W. Anelli, A. Bellogin, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, and T. Di Noia. "Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. New York, NY, USA: Association for Computing Machinery, July 2021, pp. 2405–2414. ISBN: 978-1-4503-8037-9. DOI: 10.1145/3404835.3463245. URL: https://doi.org/10.1145/3404835.3463245.

[103] R. A. Daziano, E. Waygood, Z. Patterson, and M. Braun Kohlová. "Increasing the influence of CO2 emissions information on car purchase". en. In: *Journal of Cleaner Production* 164 (Oct. 2017), pp. 861–871. ISSN: 09596526. DOI: 10.1016/j.jclepro.2017.07.001.

[104] S. M. McNee, N. Kapoor, and J. A. Konstan. "Don't look stupid: avoiding pitfalls when recommending research papers". In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 2006, pp. 171–180.

[105] D. Stein and A. M. Grant. "Disentangling the relationships among self-reflection, insight, and subjective well-being: The role of dysfunctional attitudes and core self-evaluations". In: *The Journal of psychology* 148.5 (2014), pp. 505–522.