FFM-SVD: A Novel Approach to Personality-aware Recommender Systems

Student Name: Kai Widdeson
Supervisor Name: Suncica Hadzidedic
Submitted as part of the degree of MEng Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—This paper addresses and evaluates various approaches to incorporating personality data into a recommender system to improve the accuracy of recommendations. Automatic personality recognition is enabled by the LIWC dictionary and appropriate data balancing and pre-processing has been conducted. Personality-aware pre-filtering techniques are discussed and evaluated using the RPE metric developed in this project, with *non-targeted stratified personality sampling* performing best. The effectiveness of introducing personality to recommenders in these domains is evaluated against the MAP correlation score which is introduced as a predictor. This suggested that media domains are most susceptible to improvement with personality data, however, experimental results contend this. For example, the *Movies* domain was predicted to be the most influenced domain, when in actuality the *Sports* domain showed the greatest improvement. A novel personality-specific model, FFM-SVD, is proposed and shown to outperform alternative models in prediction accuracy. LightGBM, a model currently unseen in personality-aware recommenders, and a neural network recommender are also created and assessed. Real-world implementation details, ethical considerations, and future developments are also explored.

Index Terms—Machine learning, Neural nets, Performance evaluation, Personalization, Singular value decomposition

1 Introduction

This section introduces the problem domain, background material, and the motivation behind this project. Following this, a discussion of the objectives and desired contributions accompanies this project's research question.

1.1 Background and Motivation

Recommender systems (RSs) are programs which, given a set of items and users, attempt to accurately recommend items to the current user. Their aim is to maximise results satisfaction and solve the issue of information overload [5]. They employ probabilistic algorithms and techniques to infer a user's preferences and then recommend suitable items from the system's domain.

In the current digital age, there are often an almost incomprehensibly large number of options for any user. For example, the number of films and TV shows on Netflix, or the extensive selection of items available for purchase on Amazon. In many systems, user satisfaction is improved by allowing the service to recommend items to them. This allows customers to explore their chosen domains with more ease and introduce them to new items, improving the diversity of their usage. As such, an accurate recommender system can improve user satisfaction and enhance ease of browsing for its clientele.

Research focusing on improving recommender systems is therefore certainly of interest in modern times. As online domains continue to expand, and the range of items available continues to grow, recommender systems are required to evolve to maintain users satisfaction. One improvement could be the inclusion of contextual data. For example, temporal data could provide better recommendations by basing predictions on a specific time frame from the user's history. Alternatively, modal context can provide further insights to the recommender by considering a user's mood or cognitive capabilities [3]. These systems aim to incorporate additional contextual factors to provide more accurate recommendations. For example, it may be inappropriate to recommend alcohol-serving sports bars on a restaurant recommender system in a context where young children are involved. Recommenders which take advantage of contextual information are known as context-aware recommender systems (CARS). This project studies the inclusion of contextual data in recommender systems, specifically, the personality of users.

Personality information has the potential to provide numerous advantages to a recommender system. In [18], a four-category classification model is used to represent an individual's view of themselves and others in relationships. These categories were then able to predict personality traits and were also shown to strongly correlate to decision-making self-esteem and styles. This indicates a relationship between decision making and personality traits. If personality is such a crucial factor in our choices, then it follows that incorporating personality into a recommender system should improve the accuracy of the recommendations generated. In addition to this, user personality traits are overwhelmingly considered as independent variables [19]. This suggests that traits are therefore unaffected by domain, and so the resultant recommender system should, in theory, be cross-domain capable [64]. A demonstration of this is found in [63], where content across different domains was categorised and correlations between the categories and personality traits were found. If these traits are considered independent variables, then it follows that an effective personality-aware recommender system could

provide recommendations to a user in a domain in which they have no history, based on their experiences in other domains.

The benefits of a cross-domain capable system are clear when considering how it could benefit an online retailer such as Amazon. Amazon facilitates the sale of over 350 million products [17]. It would be impossible for an individual to search through any significant fraction of that collection and so a recommender system is used to suggests items which are most likely suitable. It would be of enormous benefit to a marketplace such as Amazon if they could implement an effective, accurate, and ethical personality-aware recommender system that could learn a user's preferences over all domains. This domain-invariance would allow the system to act to some extent in all scenarios, providing more value to the user as informed predictions can be made with less prior information. For example, a student who has left many reviews on music may be moving away from home and need to buy household items for which they have no purchase history. A developed personality-aware recommender system will have a personality profile for this student and will still allow for predictions to be made in this new domain, thereby allowing the online retailer to continue to provide a positive experience to the student.

Of course, there are numerous ethical implications that need to be considered when proposing incorporating data as sensitive as personality into a system. For example, are users aware of how their personalities are being collected/predicted? Can they remove usage permissions for this data if desired? Personality data is by nature very specific and personal to an individual, and so considerations associated with individuals' sensitive data need to be made such as data privacy and protection.

To develop these personality-aware systems, psychological data must first be obtained. Acquiring personality data can be conducted either explicitly, for example through questionnaires, or implicitly, perhaps by conducting sentiment analysis on past reviews, comments, or social media behaviour (see section 2). Once this data has been gathered, implementation details regarding the format of the personality data need to be chosen. Two of the most popular personality formats include the Myers-Briggs Type Indicator [53] and the Big Five Personality Domains (BFPD) [28].

Once the personality information for users has been obtained, there are numerous options with how it can be included to enhance a recommender's performance. Contextual information can be added in three paradigms: contextual pre-filtering where context is used for data selection and reduction, contextual post-filtering, where context is used to adjust traditional results, and contextual modelling, where context is incorporated directly into the recommender [4]. Personality-aware recommender systems can therefore include personality data in any of these three schemes. These approaches will be explored in the following literature (see section 2).

1.2 Research Question and Objectives

This paper's research question is: *"How can automatic personality inference be incorporated into a multi-domain recommender system to improve the recommendations generated?"*

Whilst addressing this question, potential real-world implementation details will be considered. This extends an analysis of recommendation accuracy to include discussions regarding computation time and the ease at which an approach could be combined with existing systems.

The project's objectives are as follows:

Basic

- Evaluate and select tools for personality acquisition.
- Create a basic recommender for a single domain.
- Analyse the correlations between personality traits and ratings behaviours and use these to influence the recommendations produced.

Intermediate

- Develop a text-based automatic personality recognition (APR) approach to provide personality data to the recommender system.
- Perform an offline experiment on an existing dataset to evaluate the effectiveness of incorporating personality data across multiple domains.
- Using suitable metrics, evaluate the effectiveness of incorporating personality data into recommender models and the suitability for real-world implementation.

Advanced

- Develop and compare various personality-aware pre-filtering techniques as a means of dataset reduction.
- Analyse and address ethical issues stemming from collecting and using psychological data in personality-aware recommender systems.
- Explore how neural networks and deep learning could be incorporated into personality-aware recommender systems.
- Suggest further developments to personality-aware recommender systems' architectures which could improve recommendation accuracy.

The contributions of this project include: *evaluating methods of personality acquisition* and techniques for incorporation into recommender systems, analysis of performance of the RS across *multiple domains*, research into *ethical issues and challenges* associated with using personality data, and suggesting *future developments* to improve the performance of personality-aware recommender systems.

2 Related Work

This section explores the relevant literature to this project. Similar studies, along with various approaches for personality acquisition, recommender system architecture, and evaluation approaches, are considered.

Recommender systems have an expansive selection of methods and techniques in their construction, a survey on which can be found in [9]. The literature in this field includes approaches involving probabilistic approaches, Bayesian networks, deep learning approaches, SVD models, fuzzy models, nearest neighbour algorithms, and countless others. These traditional recommenders, or 2D recommenders,

must then undergo adaptations to allow for the inclusion of contextual data and construction of a CARS.

Before the method of including personality data into a recommender is considered, the techniques for acquiring and categorising personality must be discussed. A common choice is the BFPD format of [28] and [16], alternatively known as the "Five Factor Model" (FFM), as seen in the recommenders in [54] and [7]. This approach is often favoured as the results can be quantitatively measured and represented through a simple five-by-one vector indicating a user's personality strength in each of the five following domains: *agreeableness*, *conscientiousness*, *extroversion*, *neuroticism*, and *openness to experience*.

An alternative categorisation to the FFM would be the Myers-Briggs Type Indicator (MBTI), where personalities are assigned to one of 16 distinct categories [53]. The approach in [75] involved creating an MBTI collaborative filtering (CF) recommender system with an MBTI personality-based neighbourhood approach. Experimentation revealed that the MBTI-CF can perform well on sparse data as well as offer more stable performance. Sparse data performance is critical as it is commonly seen in recommender systems where it is unlikely that the most users will have reviewed the majority of items. As MBTI places users into categories instead of estimating similarities between all users, as with the FFM approach, the issue of scalability is also improved.

It must, however, be noted that [49] claims that the MBTI scale does not account for the FFM trait of *neuroticism*, making it incomplete. This was disputed by [29] which, by building upon the work of [45], attempted to show that like the FFM, MBTI also has a circumplex structure and so is valid. The validity of MBTI is further questioned as it has been shown to be fail its test-retest reliability. Pittenger [61], noted that even over a short test-retest interval, e.g. over five weeks, as many as half of people will receive a different MBTI classification. It must also be recognised that under the MBTI categorisation, two users with relatively similar scores could be grouped separately and receive very different personality labels [61].

MBTI is from the personality-types theory, whereas FFM is from the personality-traits theory. A study in [22] compared the performance between the two, alongside other personality-traits models Eysenck and HEXACO, before proposing a hybrid personality model. The Eysenck model measures personality through Extraversion/Introversion, Neuroticism/Stability, and Psychoticism/Socialisation, whereas HEXACO extends the FFM with the additional dimension of Honesty-Humility. The incorporation of these models into a personality-aware recommender system was evaluated on the Newsfullness dataset [20]. This study found that in the cold-start phase, where little or no initial data is presented, the Eysenck and MBTI perform best. Later, once the cold-start phase had concluded, the FFM and HEXACO models were more accurate with HEXACO slightly outperforming FFM. The study proposed a hybrid personality model which demonstrated the best performance overall by combining the personality-type and personality-trait theories at different stages.

Once a personality format has been chosen, the method of personality acquisition must be considered. A social media study in [6] confirmed that the information in text, such as a user's reviews, could be used for personality feature extraction. These features were acquired through the "Extraction Methods Linguistic Inquiry and Word Count" (LIWC) text analysis technique where words were associated with psychological categories such as "social word", "positive emotions", and "negative emotions" [60]. This is a form of text-based automatic personality recognition (APR) which [21] determined to achieve acceptable accuracies that are generally higher than those provided by multi-media based or behaviour-based APRs. The sensitivity of the LIWC to identifying emotion expression has been validated in numerous studies, such as [50], and the confidence in the results produced by LIWC is high. An example of an alternative APR approach is that of an image or visual APR, as seen in [40] and [51], where profile pictures can be input into deep learning models to predict personality traits.

Instead of collecting implicit personality data through a technique such as a text-based APR, recommender systems may instead opt to use personality questionnaires to obtain data explicitly. An example of this approach can be found in the first experiment in [58], where the NEO-IPIP personality inventory is used and assessed based on the FFM. Examples of alternative personality assessment questionnaires include TIPI, IPIP, BFI, FIPI, NEO-FFI, and BFMS.

As noted by [21], whilst questionnaire-based approaches are more accurate than APR methods, APRs are much easier to conduct and can be applied to a user's existing data. This has the advantage of not needing an online experiment and allowing inference from large existing datasets. Furthermore, due to the implicit nature of APRs, they remove the threat of results being affected by numerous psychological factors associated with explicit personality acquisition such as social desirability bias (SDB). This refers to the instances where a subject may give more socially acceptable responses to questions instead of giving true answers [32] [44]. Another challenge associated with questionnaires may be the reference-group effect (RGE) [21], which considers the scenario where questionnaire responses are not based on an absolute scale but are instead derived from relative comparisons to others [65]. There are different challenges associated with implicit personality acquisition, such as validating the personality scores produced.

Dunn [23] demonstrated a different technique for implicitly gathering personality data which involved observing user's completing tasks such as playing games. This methodology would, however, still result in the drawbacks associated with online experiments as described in [21].

The study in [12] used the human personality congruency theory [31] where it is found that product-personality congruence has a positive influence on consumer preference. This theory was combined with the FFM traits and data about users' interconnections and inferred psychological traits. The result obtained provide further evidence that social media data can produce a personality prediction engine which dramatically improves user satisfaction with recommendations.

A personality acquisition methodology focusing only on the analysis of written reviews can be found in [62]. Here, a supervised model using ridge linear regression was trained on the linguistic features found in the MyPersonality [43] dataset to find FFM scores. This was then used to predict the

personality scorings for users in a different dataset. Linear regression was chosen as the personality scores are in a continuous range from one to five. This solution incorporated the personality profiles using a Kernelized Probabilistic Matrix Factorisation [77]. It revealed that the best performance could be achieved by combining the model's raw linguistic features with the calculated personality predictions.

With personality acquisition techniques considered, an exploratory analysis of contextual modelling approaches can be conducted. The personality-aware movie recommender system in [54] proposed the 50/50 recommender which combined FFM scores with a CF recommender. The CF recommender was chosen to aid the system when facing the cold-start problem. This study's evaluation revealed that users preferred the new system 3.6% more than the then state-of-the-art k-NN clustering method. They also then created the 80/20 model where the personality-calculated genre preference had a higher weighting than the rating from the standard recommender. This performed worse than the k-NN and 50/50 approaches but still provided reasonable results. This demonstrates how personality can enhance recommendations and suggests approaches for combining personality data with data from an existing recommender.

Another study using personality data to address the cold-start problem is found in [38]. In this paper, MAE was reduced and ROC sensitivity was improved when comparing three proposed CF systems, which used personality data alone or combined with ratings information, to a traditional CF recommender. This improvement was also mirrored in the study in [25], which additionally concluded that cross-domain ratings can be used in that ratings from a source domain can improve recommendations in a different target domain.

It must be noted that when testing on multiple domains, CARS can sometimes result in weaker predictive power, as illustrated by an example in [3], where it is noted that the context that a user is looking to buy a book for a child will be of no predictive value when the user seeks recommendations for work-related books. A well-developed CARS should therefore consider these different contexts. This challenge of CARS being context-specific is not present with personality-aware recommenders as the user for whom the recommendations are being generated remains the same.

Published literature reviews, such as [21] and [46], have gathered the performances of various personality-aware approaches in different domains. However, these reviews tend to simply gather domain-specific recommenders across multiple domains. Very few works in this field implement a multi-domain approach and then compares the effectiveness of the personality inclusion in each. Linear correlations between preference- and personality-based user similarities across different domains for a CF recommender are observed in [24], however, a direct comparison of the domains chosen is not presented. Personality correlations were also found between different domains and their associated genres, but again, the effectiveness of the strength of these correlations was not evaluated in an implementation [13]. These correlations could lead to a system using the CARS scheme of contextual post-filtering. The personality correlations between different domains could allow for comparisons to target users and the adjustment of results from a traditional recommender.

Another approach using FFM personality traits which has been shown to outperform a standard recommender is found in the HyPeRM study [8]. In this approach, collaborative filtering, using demographic and personality data, was combined with content-based (CB) filtering for a hybrid approach. CF is used unless the user is new in which case CB is used instead. Personality data was acquired explicitly in this work. The results of this approach show a significantly reduced standardised root mean square residual (SRMR) and root mean square error of approximation (RMSEA) indicating that the accuracy of the system has been improved by the inclusion of personality data.

An alternative approach to using personality data to improve a recommender system can be found in [72]. Different personality traits have been used to enhance the diversity of recommendations by incorporating personality as a moderative factor to adjust the diversity degree. It was found that some personality dimensions have more of an impact than others.

As it has been established that the inclusion of personality data in recommender systems has the potential to improve recommendations made, an analysis of the various recommender models must now be conducted. One collaborative model is the Funk singular value decomposition (SVD), popularised by Brandyn Webb's submission to the Netflix Prize [26]. The SVD is a matrix factorisation algorithm and Funk's variation is designed to perform well on sparse data (see section 3.5.1). SVD recommender systems have been seen to outperform numerous CF alternatives, such as in [34], where the RMSE of the SVD was compared to Slope One, KNN, Average Least Square, and Weighted Arithmetic Mean.

The SVD is a form of model-based CF where a model is trained to generate predictions, as opposed to memory-based CF where all data is used and similarity measures are recorded between items to produce recommendations. A comparative analysis between the two approaches was conducted in [2] on an e-commerce recommender, where an improved Bayesian network (model-based) and a nearest neighbour approach (memory-based) were implemented. This study is relevant to this project as the Amazon reviews dataset is used (see section 3.1), and so there should be similar trends to the e-commerce data evaluated upon. This study concluded that in the offline evaluation, model-based approaches were more accurate than memory-based, considerably faster with an average of 10x improvement, and were better at generating relevant recommendations. Where SVDs have been used in personality-aware recommender systems, they are typically a means of user pre-filtering by the use of a personality neighbourhood, as seen in [73]. They are rarely used within the actual recommendation generation with personality data. The SVD is therefore serving as a personality pre-filtering approach to be combined with a traditional 2D recommender which only considers the users and items, but not context.

An alternative set of models are tree-based approaches. Random forests (see section 3.5.2) have been consistently shown to be capable of accurately predicting missing data in sparse datasets [35]. A comparison between a random forest

approach acting on Kmeans++ clusters was conducted in [5] and reveals that adapted random forest approaches have the potential to outperform SVD alternatives as well as KNN and softmax regression. An alternative tree-based approach which does not yet seem to have been implemented with personality-aware recommender systems specifically is the LightGBM model. LightGBM has shown similar accuracies to other state-of-the-art gradient boosting approaches which are known to outperform typical tree-based models such as random forest approaches [41].

The study in [15] tested per-instance algorithm selection for recommenders using instance clustering. This involved clustering data and predicting the effectiveness of different recommenders based on cluster membership. Two of these algorithms were SVD and LightGBM, with the best performing algorithms being SVD++ and the variation LightGBM_EF, which uses additional statistical meta-features (see sections 3.5.1 and 6.2). In this study the SVD algorithm was consistently predicted to be more suitable than the LightGBM approach. This could indicate that the SVD algorithm is more suitable as a multi-domain recommender.

Due to their prevalence in recent literature, this project must also consider deep learning approaches to personality-aware recommender systems. Deep learning has been employed in personality detection, as seen with the image APR in [51], and also as a predictive model for recommendations. There are few studies relating to personality-aware deep learning recommenders, one such example can be found in [55]. In this study, a 3-layered artificial neural network (ANN) is used for friend preference classification and subsequently friend recommendation using personality differences. This ANN operates on users which have been placed into five classes of personality, and so the model is conducting multi-label classification for personality-class preference instead of forming a regression model. A survey on deep learning techniques employed in recommenders can be found in [76]. This survey includes various models such as the Multilayer Perceptron and the Restricted Boltzmann Machine, along with neural network approaches. The survey also highlights the potential for deep learning multi-domain recommenders. This is listed as a future development as deep learning's ability to learn high level abstractions makes it suitable for learning features which impact across multiple domains. The survey notes that this is a largely unexplored area.

As discussed, the existing literature in this field does not often provide a direct comparison of the impact of personality data on a recommender system when being implemented in different domains. Furthermore, whilst deep learning approaches have been employed for automatic personality recognition and multi-label classification, the literature does not address a deep-learning personality-aware regression model. As recognised by [76], further exploration into the cross-domain potential for deep learning recommenders would be of value. Despite the suitability of a random forest recommender being explored at length, no experimentation with a LightGBM personality-aware recommender has been found. It has also been seen that SVDs tend to only be implemented in personality-aware systems as a means of pre-filtering users and are not adapted for contextual modelling. This project's aims are designed to remedy these gaps in the literature. The study in [15] suggests that SVD approaches may be more suitable in multi-domain recommenders than LightGBM and this too hopes to be confirmed.

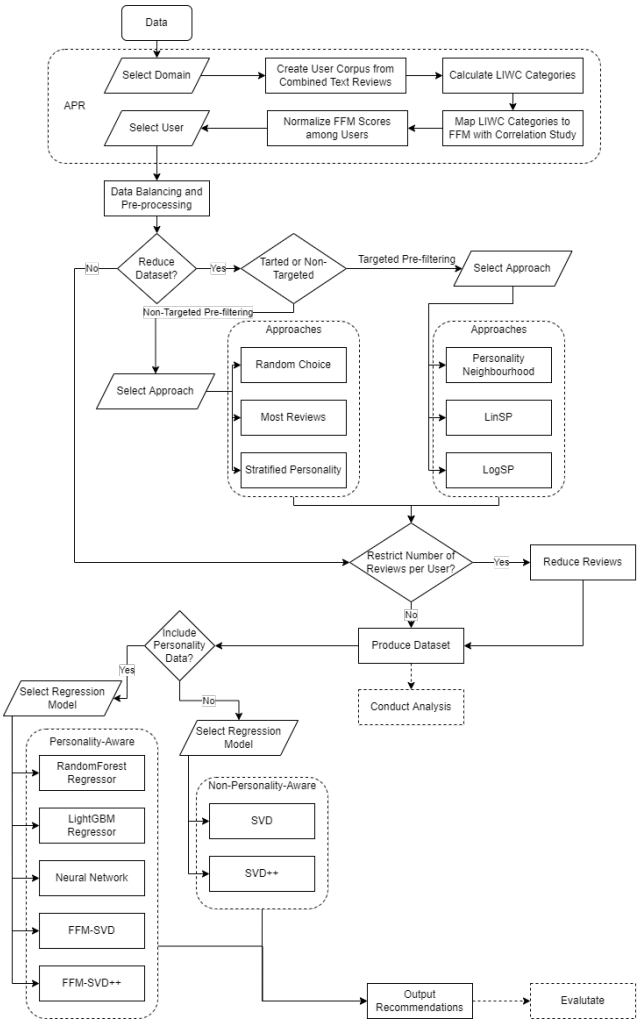

Fig. 1. Complete System Architecture

3 Solution

This section discusses the dataset chosen for this study, the analysis conducted to inform design decisions, the techniques used for automatic personality acquisition, data pre-processing and user pre-filtering approaches, regression models employed, and evaluation techniques. The full system architecture for this proposed methodology can be found in Fig. 1. User profile data is first used to automatically generate personality trait scores. The data can then be balanced and pre-processed before a pre-filtering technique is implemented to reduce the space of users on which the collaborative filtering model will act. The number of reviews can also be restricted to reduce computational cost. Following this, the resultant dataset is passed to the chosen model for the generation and outputting of recommendations

Personality data has been incorporated into both the models and the pre-filtering of this approach in an attempt to improve recommender performance. Post-filtering approaches have not been considered due to project time

TABLE 1
Chosen Domains from Dataset

Domain	Num. Reviews	Num. Users
Movies and TV	3,410,019	297,529
Music - CDs & Vinyl	1,443,755	112,395
Kindle Store	2,222,983	139,824
Video Games	497,577	55,223
Pet Supplies	2,098,325	236,987
Sports & Outdoors	2,839,940	332,447
Patio, Lawn, & Garden	798,415	103,431

limitations and implications of how different approaches could only be available to particular domains (see section 6.2).

3.1 Dataset

The dataset chosen for the implementation and evaluation of the proposed solution is the collection of over 230 million Amazon Reviews, updated in 2018, found at [56] [57]. The key information contained in this dataset includes: the reviewing user and item identifiers, the rating given on a 1-5 scale, and the review text. This therefore allows for personality acquisition through a text-based APR technique and the training of a model with the ratings given (see section 3.5). Furthermore, 5-core subsets reduce the data such that each of the users and items present have at least 5 reviews each. This dataset was chosen as it is one of the few available sources where ratings information is accompanied by review text, there is an abundance of data, and, in addition to this, the categorical segregation of the dataset enables easier testing and comparison of performance in different domains. The Amazon marketplace would also be a suitable host for a developed personality-aware recommender system (as discussed in section 1.1). From the 5-core subset, 7 different categories/domains have been chosen for analysis. These are seen in Table 1.

This selection has been chosen as it includes a variety of media domains which are expected to be impacted by personality to some degree. In addition to these, domains are included which are expected to be less affected by personality, for example, *gardening* or *Sports & Outdoors*, which include items that are designed to perform a task and therefore have less potential for personalisation.

3.2 Data Analysis

Extensive analysis has been conducted on the datasets and the APR results to best understand how to improve recommender performance. By understanding the relationships within the data, it can be determined how to include personality information to most positively impact prediction accuracy. Furthermore, an analysis of how skewed the data is will inform data pre-processing decisions to allow for increased confidence when discussing results.

The Pearson correlation is commonly used to measure the strength of the linear relationship between normally distributed variables and so will be used throughout this project, with correlations being found between all features. This measure uses information about the mean and standard

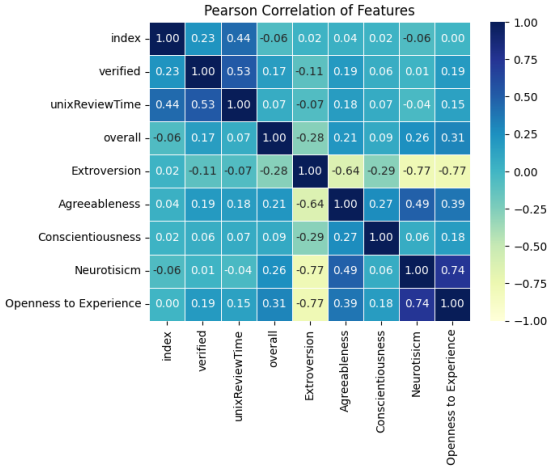

Fig. 2. Pearson Correlation Matrix of Features in *Movies* dataset (no user pre-filtering)

deviation whereas non-parametric alternative correlations use only ordinal information and scores of pairs and so use less information. As a result, these alternative measures are typically less powerful than the Pearson correlation. A commonly used alternative is the Spearman's rank correlation, however, [36] states that Spearman's rank correlation coefficient should not be interpreted as a significant measure of the strength of the associations between two variables, making it less suitable for this project's requirements. The correlation between x and y will be calculated as seen in Eq. 1, for some number of observations, n.

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad (1)$$

An example correlation matrix from the *Movies* domain, without user pre-filtering, can be found in Figure 2. The *Movies* domain has been chosen as a representative domain of the dataset as it has the most data and so correlations will be most informed (see Table 1), it has a reasonable mean corpus length, and is one of the least skewed domains (see Table 2). It was also later calculated that this domain has the highest Mean Absolute Personality (MAP) correlation, a predictor for the degree of impact that personality will have on a domain (see Section 3.6.3 and Table 6). This further indicates that this domain is a suitable representative as it should be more strongly affected by personality data.

An important insight into the APR's scores are how each personality trait differs with the rating given. This would indicate which of the traits are most significant and will best be positioned to benefit recommendation calculations. To determine this, for each category of overall score given from one to five, the personality scores of each domain for each user giving that category score were summed. This produced a *prominence* scoring for each trait for each of rating.

The results were then normalized using L2 normalization as this optimizes the mean cost. Advantages of L2 normalization include increased likelihood of unique values, reduced risk of over-fitting compared to other approaches, such as L1 which reduces the median explanation, an ability

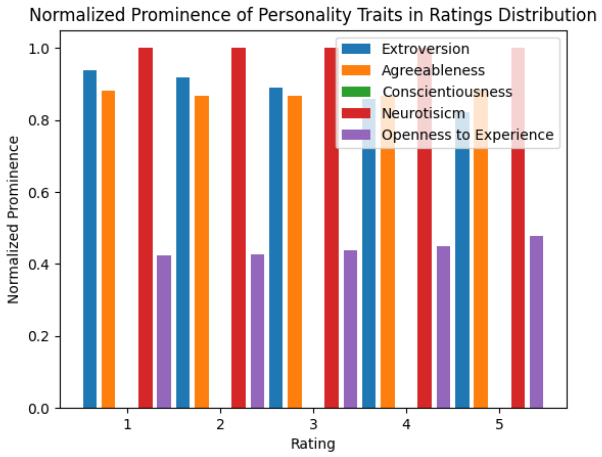

Fig. 3. Normalized Prominence of each Personality Trait within each Rating of the *Movies* domain

to learn complex data patterns, and as it is often seen to be more accurate and computationally efficient than L1 [70]. This normalization is conducted as seen in Eq. 2 and is used throughout the remainder of the project.

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{2}$$

It should be noted that in this analysis stage, ratings which were not integer values in the range of 1-5, as a result of identical review averaging (see section 3.4.1), were rounded to the nearest integer value to retain readability of the produced graphs. The results from the *Movies* Data is shown in Figure 3. Clearly, the *extroversion* trait, has a convincing negative correlation with the score given, and similarly, the *openness to experience* trait correlates positively. Other personality traits had much weaker trends and this analysis reveals that it should be expected that *conscientiousness* has a very limited impact on predictions. When this analysis of the prominence of personality traits was repeated on other domains, similar trends were shown, confirming that the conclusions made from the representative *Movies* domain should hold for all domains in the dataset.

It was expected that ratings would be more populous to either end of the ratings scale as an individual will be more likely to leave a review on an item that they have a strong sentiment towards and this has been confirmed in this analysis. The skewness of each of the domains, as well as the percentage of reviews with each of the five possible scores, can be seen in Table 2, where the skewness of the data is calculated using the percentages of values in each rating category. Typically, a skewness score less than -1 or greater than 1 indicates heavily skewed data [68]. Using this, all domains have heavily positively skewed data with a high of 1.407 for *Patio, Lawn, & Garden*. This would affect the performance of any regression models created by increasing the likelihood of models converging towards a learned, highly rated, average. To remedy this, data balancing approaches will need to be considered (see section 3.4.5).

TABLE 2
Skewness and Mean Corpus Length (MCL) across each domain.

Domain	Skewness	MCL
Movies and TV	1.279	5270.55
Music - CDs & Vinyl	1.302	9080.35
Kindle Store	1.058	8161.96
Video Games	1.308	6046.26
Pet Supplies	1.438	2246.89
Sports & Outdoors	1.352	2133.49
Patio, Lawn, & Garden	1.407	1903.94

3.3 Personality Acquisition

As seen at the start of the pipeline in Fig. 1, a text-based APR technique will be used to collect the personality data for users in this system. An APR technique has been chosen as it is the least intrusive and requires no additional work from existing users, such as personality questionnaires which would introduce ethical and validity considerations such as RGE and SDB. An alternative approach could be to find a dataset with profile images of users and attempt an image-based APR, however this would likely reduce explainability for users, introduce numerous additional ethical considerations, and produce results which have less validation from related studies.

In a real-world implementation, APR scores could be calculated, stored, and only updated for each user individually when that user's review corpus changes. This would allow for the system to retrieve personality data without computation for the majority of use-cases.

As discussed (see section 2), MBTI has numerous concerns regarding its suitability as a personality format, and other traits formats such as Eysneck and HEXACO, whilst potentially suitable, have a lack of literature which could introduce challenges such as finding LIWC correlations. Therefore the personality scale employed for this project will be the big five personality domains (BFPD) / Five Factor Model (FFM). The numerical scale nature of the FFM will allow for easier comparison between reviewers and will be better suited for traditional regression model approaches.

From each of the domains, a corpus is created for each user by combining all reviews for that user into a single piece of text. These user corpora are converted into lowercase so that they can then be operated on by the APR function. The mean corpus length (MCL) over users in each domain has been included in Table 2. It is expected that domains with a higher MCL will have more accurate APR scores, as a result of the increased amount of data available. This, however, cannot be confirmed without an online study.

The LIWC dictionary treats each word in a user's corpus as a token which may be assigned to one of the LIWC categories. The NIH study conducted in [74], found correlations between these categories and the Big Five personality traits. The NIH study used an older version of the LIWC dictionary, LIWC2001, rather than LIWC2015 which is used in this project. Unfortunately, not all of the LIWC2015 categories had an obvious mapping to LIWC2001 counterparts and so could not be included in this project as correlations were unknown. This therefore only allows a portion of a user's

corpus to be used when indicating their personality (see section 6.2).

For each user, the counts in each category are multiplied by that category's correlation scores to each of the five personality traits, as determined by [74], and then summed to produce five scores to represent the personality obtained. Scores were rounded to three decimal places and normalized which allows for all users to be comparable, regardless of the number of reviews previously left. Otherwise, a user with a larger corpus would be capable of higher personality results. Additionally, users within the same domain are likely to have similar language and so have personality trends determined by their domain, restricting the comparisons that could be made between different domains without normalization. This process also means that differences among users' values will be increased to result in more distinct variations for model fitting.

Normalization is also conducted within each personality trait. The first purpose of this is that by normalizing categories individually, a user's personality score in, for example, *extroversion*, will be relative to all other scores in the same domain. Otherwise, if neurotic tokens were more prevalent in the reviews' language, all users would score highly in *neuroticism* and would have small variations which provide little information to the model. This approach therefore ensures that there is an appropriate range of scores in each trait. The range is also reduced to be within zero and one, ensuring that different domains will produce comparable data regardless of their size.

3.4 Data Pre-Processing and Pre-Filtering

Due to the size of the datasets used, significant data reduction is required to ensure that the training of the models is computationally feasible. This gives the opportunity to further analyse the impact of personality on different pre-filtering approaches which is one of this project's *advanced* objectives.

In this subsection, data preprocessing techniques are discussed, data balancing solutions are considered, and various user pre-filtering approaches are proposed. These pre-filtering techniques are then experimentally compared and evaluated in section 4.1 and section 5.1.1 respectively. The usage of the pre-processing and pre-filtering techniques occur after the system's APR, as seen in Fig. 1.

3.4.1 Data Pre-Processing

There are numerous instances in the dataset where a user has rated an item multiple times. This will either be a result of changing opinions, multiple purchases, or slight variations in the item ordered. This can result in numerous evaluation challenges such as test data influencing the training data and multiple varying predictions being produced for the same item.

To remedy this, any user-item pairs which have more than one review are replaced with a single average score. This is then passed to the model in place of the duplicates. However, the full dataset including the duplicated reviews are still passed to the review APR. The reasoning for this is that the different reviews may contain different language which can then be used to expand the corpus for that user and provide more accurate personality scores. If not, the normalization of the results will ensure that the personality scores are unaffected.

This approach of combining reviews for identical user-item pairs could introduce some issues and room for further consideration. For example, it may be appropriate for more recent reviews to be given a greater weighting as they might be more indicative of current sentiment. This additional experimentation would be a valuable future development but is outside the scope of this project and so here the average rating is used.

3.4.2 Non-Targeted Pre-Filtering

These approaches are non-targeted as they do not take the target user's personality into consideration. In these approaches the personality scores of the current active user, for whom recommendations are being generated, do not influence which users are selected.

The first of the non-targeted approaches is where the dataset is reduced to contain only the top n users which have the most reviews, along with the chosen target user. This approach solely aims to maximise the amount of data available. Additionally, a random user selection has been implemented to observe if a random sampling approach can obtain a varied enough distribution from the dataset so as to suitably inform the models. This approach reduces selection bias but introduces some degree of sampling error [37].

Alternatively, a stratified sampling technique has been implemented. In this approach, for n desired users, each of the five categories will be required to select the $\frac{k}{2}$ users which have the highest score, as well as the $\frac{k}{2}$ users with the lowest score, where $k = \frac{n}{5}$. Users are removed from the possible list for selection after they have been selected to ensure that a single user is not added multiple times. This approach has been implemented as it allows for the greatest range over all personality traits which could potentially provide more informed data. A consideration of this approach is that, as only one trait is being focused on at a time, regardless of the scores of the other four, it is possible for some data skewing depending on how strong the inter-trait correlations are. It is hoped that with one domain being the focus for minimum and maximum values at each point, the values selected from the other four domains will be a more average distribution. This approach is a form of stratified sampling, [37], as users are being split into groups, and then an equal number of results are chosen from each group by some decision criteria. This approach reduces sampling bias [37], however, the decision criteria is domain specific and so could potentially be improved with further experimentation.

3.4.3 Personality Neighbourhood Pre-Filtering

This approach utilises the personality neighbourhood of the chosen target user. The summed absolute error between any user's five personality scores and the scores of the target user are found, and then only users whose absolute error is in the top p percent are considered. The reasoning for choosing a percentage value instead of using some threshold value, as was implemented initially, is that in the evaluation stage this will allow multiple users to be evaluated with the same number of users in their training data, improving the

validity of results. At this stage, only the total magnitude of the difference from the target user is being considered, regardless of the domains in which that difference appears.

The intent behind this approach is to reduce the dataset to those most similar to the target user. All items reviewed by this population will typically be expected to have a similar trend to the target user as they have a similar personality. Despite this, the slight discrepancies between traits still allows for a more personalised ranking of the different items. This is a form of clustered sampling, [37], which can improve efficiency over alternatives, however, it can also result in increased risk of bias as the selected neighbourhood may not be representative of the population. This could limit the amount of information obtainable by any model.

3.4.4 Linear and Logarithmic Stratified Personality Pre-Filtering

For the Linear Stratified Personality (LinSP) pre-filtering approach, all users are split into b brackets where $b = \lfloor log_2 n \rfloor$, for the n users required. Each of these brackets contain a set range of values for the summed absolute personality error between any user and the target user. Once these brackets have been created, $\lfloor \frac{n}{b} \rfloor$ users are randomly selected from each bracket, with any additional $n \ mod \ b$ being taken from the bracket most similar to the target user. This ensures an identical number of users in each pre-filtering approach, allowing for more reliable evaluations.

The rationale behind this approach is that it allows for an equal selection of users over a range of differences to the target user, allowing the model to train over a random, but guaranteed distributed, variety of users from the dataset. This approach is a form of systematic stratified sampling, [37], and aims to improve accuracy of results by reducing sampling bias whilst utilising the personality knowledge of the target user.

The Logarithmic Stratified Personality (LogSP) approach, operates identically to LinSP but where the number of users sampled from each bracket decreases as the differences between the target user and that bracket increases. Here, half of the requested number of users will be randomly selected from the lowest bracket, those users who are most similar to the target user, a quarter from the next lowest bracket, and so on. As in LinSP, any additional users are selected from the most similar bracket.

This approach attempts to keep the variety of users provided in the rationale behind LinSP, but to give priority to those users who are more similar whilst still including representatives of the most different users. This approach is a more complex form of systematic stratified sampling to LinSP, and is believe to be a novel approach created in this project.

3.4.5 Data Balancing

Following the chosen pre-filtering approach, the data needs to be balanced. As seen in the data analysis conducted before the development of this methodology, the dataset is heavily skewed towards higher ratings. Without any data balancing, initial experiments have revealed that recommender models may otherwise learn some minimal-error average to converge to. This would not reflect true learning.

Due to the small size of the minimal category, undersampling, where n items are sampled from each non-minimal category with n being the size of the minimal category, is not a viable option as the resultant data would be of an insufficient size for informed recommendations. Instead, an equal oversampling approach has been employed. More advanced techniques could improve results but this is not the focus of this project (see section 6.2). This equal oversampling approach involves duplicating items not in the maximal category until all categories are approximately the size of the maximal one. For a category n, d_n duplicates must be created and are calculated as shown in (Eq. 3), where t_n are the number of reviews with a rating n.

$$d_n = \lfloor (t_{max} - t_n)/t_n \rfloor - 1 \tag{3}$$

Initial experiments have shown this approach to stop regression models converging towards a learned average. It should be noted that ratings categories which have been created as a result of averaging identical user-item pairs, these are rating categories which may not be integers between 1 and 5, are not included in this equal oversampling as it would skew the data in favour of these averaged results. As categories are only oversampled to be *approximately* equal sizes, this should not affect the trends of the data.

3.5 Recommender Techniques

Several recommender models with and without the inclusion of personality data have been implemented for evaluation in this project. The usage of the chosen model in the recommender system follows the data pre-processing and pre-filtering and is the final stage before the outputting of recommendations and the subsequent evaluation, as seen in Fig. 1.

An example of a standard recommender technique is collaborative filtering (CF) where the recommendations generated are influenced by ratings provided by similar profiles. An alternative would be a content-based (CB) recommender, where the profile's previously reviewed items are used to attempt to infer the user's likes and create suitable recommendations based on knowledge of the items [8]. As personality data is the focus of this project, it is intuitive that users will be compared to each other when generating recommendations. As such, a CF approach has been decided upon for the methodology of this project. The models implemented are regressors and not classifiers as classification models would have no initial understanding of the relationship between similar ratings.

This section discusses the various models that have been implemented and compared for effectiveness in a personality-aware recommender system, including details about the hyperparameter tuning conducted on each. Due to time limitations of this project, only a single domain could be evaluated on in the hyperparameter tuning of the models. The *Kindle* domain was chosen as it is the least skewed domain, has one of the highest MAP correlation scores (see section 3.6.3), and also contains a suitable number of users and reviews whilst still executing in a reasonable time for the testing process. Additionally, when tuning the FFM-SVD (see section 3.5.1), only the top user (see section 3.6.4) could be evaluated against due to time restraints, as

opposed to the averaging over the top 10 users which was done with the LightGBM tuning. Both tests used *non-targeted stratified* pre-filtering. All models are tested both with and without personality data to allow for an evaluation of the effectiveness of its inclusion in each instance. For further reasoning behind the model choices made, see section 5.3. Suggestions regarding a hybrid approach with CB post-filtering are made in section 6.2.

3.5.1 SVD/SVD++ & FFM-SVD/FFM-SVD++

As discussed in section 2, the SVD model, [26] [69], has been found to be an effective form of model-based collaborative filtering with recommender systems [34]. They are reasonably efficient and have been seen to outperform alternatives. As a result of this, they have been chosen for implementation and evaluation in this project. Typical SVD models are restricted in the amount of information that they can utilise as they usually only operate on user and item identities. Therefore, this approach must be adapted if it is to incorporate personality information for contextual modelling.

First, a description of the standard SVD must be given so that later models can be understood. The SVD algorithm constructs three matrices: the *User-Item* matrix, p matrix of users to latent factors filled with random values, and the q^T matrix of latent factors by items filled with random values. Next, the *User-Item* matrix is iterated through to find true values for user-item pairs. For each true value, the dot product is found between all random values associated with the user from the p matrix and all random values associated with the item from the q^T matrix. This creates a predicted value whose formulation is shown in (Eq. 4). If the user is unknown, then the bias, b_u, and factors, p_u, are assumed to be zero, with the same applying to b_i and q_i for item i.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u \tag{4}$$

The regularised squared error between the two values, calculated in (Eq. 5) where R_{train} is the set of known user-item pairs, is found. The λ constant controls regularisation.

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2) \tag{5}$$

This error is then minimized by performing a stochastic gradient descent, seen in (Eq. 6). The randomly initialised p and q^T matrices can now be updated with improved values. This entire process is then repeated for each value in the User-Item matrix to complete one iteration of the n iterations specified [71] [42] [69]. With each iteration, the predictions are improved and the model is better fitted, producing more accurate recommendations.

$$\begin{aligned} b_u &\leftarrow b_u + \gamma(r_{ui} - \hat{r}_{ui} - \lambda b_u) \\ b_i &\leftarrow b_i + \gamma(r_{ui} - \hat{r}_{ui} - \lambda b_i) \\ p_u &\leftarrow p_u + \gamma(q_i(r_{ui} - \hat{r}_{ui}) - \lambda p_u) \\ q_i &\leftarrow q_i + \gamma(p_u(r_{ui} - \hat{r}_{ui}) - \lambda q_i) \end{aligned} \tag{6}$$

The SVD++ algorithm extends SVD by considering implicit ratings. Here, predictions are calculated as shown in (Eq. 7), where y_j terms found in I_u are the set of item factors that capture implicit ratings. These implicit ratings describe a user rating an item, regardless of the rating value given [69] [42]. This inclusion of implicit rating, where the presence of a user rating an item indicates some preference, results in stronger recommendations for items which a user has rated positively than non-rated items.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T(p_u + |I_u|^{-\frac{1}{2}} \sum_{j \in I_u} y_j) \tag{7}$$

This project now proposes the FFM-SVD and FFM-SVD++ models, as an adaptation of the SVD and SVD++ models respectively, to incorporate personality data. It should be noted that in the following description of the FFM-SVD's construction, any SVD can be replaced with an SVD++ to produce the FFM-SVD++ instead.

First, a typical SVD model is created from the 2D matrix of reviewer IDs and ASIN values, these are the item IDs. This model will be capable of predicting a target user's preferences towards items by recognising similarities between the target user's review history and their likeness to the histories of other users in its matrix.

Next, five personality SVD models are constructed. Here, a personality trait is chosen, and the model is fit on the 2D matrix of the rating given and the scoring the reviewer had in that particular trait. The five personality scores, within the range of zero and one, are rounded to n decimal places, reducing the personalities into buckets which may be then used in an SVD model. This step is crucial as initially the obtained personalities are to a high enough precision that each personality scoring resultant from the APR technique is unique, resulting in each score almost acting as an ID value and therefore producing identical results to the user-item SVD. SVD models are designed to identify the relationships between the two provided dimensions, and not within each dimension itself, making it oblivious to understanding how each personality score relates to each other. Therefore, by first pre-processing the personality scores for a domain with this bucketing, a relationship is being formed before the SVD is trained.

The test-split items already seen by the target user are then removed, for the purposes of later evaluation, and the model is fitted to the pre-processed training data. The target user's score in each domain is rounded as before, and is then passed to the model to generate predictions. The predictions produced by each of the user-item SVD, and five personality SVDs, are combined. Originally, the mean value of the SVDs was used. Improvement was then found by instead using a weighted approach. The user-item SVD, which observes relationships in common personality-oblivious CF approaches, is given the highest weighting of 1 (this is later changed). Following this, the personality SVDs are given a fractional weighting equal to their pairwise absolute Pearson correlation to the ratings.

When experimenting with the optimal size for buckets in the previous stage, rounding for the bucketing was tested between one and 7 decimal places. The approach of using weightings instead of simply finding the base mean resulted in lower optimal RMSE values and an increased standard deviation. This indicates that the correlation-weighted ap-

proach is superior as results are both more consistent and accurate, therefore providing improved recommendations. An increased standard deviation is a positive indicator that the model is not converging recommendations towards some learned average. The structure of this approach is visualised in Figure 4. Additionally, some experimentation was conducted with requiring some minimum correlation threshold for inclusion in the mean prediction, however optimal performance was found when all five personality SVDs were taken into consideration. Taking all SVDs into account will also allow for more accurate comparisons across different domains as an identical number of features will be considered. This, however, may be the cause of reduced adjusted r-squared scores for this model (see section 5.1.2).

Tuning the FFM-SVD model, due to its structure combining multiple SVDs, could not be evaluated by implementing a grid search. Instead, the parameters had to be tested individually on a single domain. This could result in optimal combinations of parameters being missed, however, the individual testing has still resulted in a tuned model which noticeably outperforms the model acting on the original parameters. All values for default parameters, as well as a further explanation of the parameters, can be found in [69]. The final experimentally chosen hyperparameters are as follows: 100 factors, 20 iterations of SGD, a 0 mean of normal distribution for factor vectors initialisation, a 0.2 standard deviation of normal distribution for factor vector initialisation, a learning rate of 0.0062, and a regularization term of 0.01.

The final two parameters to evaluate, the weighting assigned to the user-item SVD and the personality bucketing precision, are specific to the FFM-SVD approach. It has already been found to be best that each of the personality SVDs are weighted by that personality trait's correlation to the ratings. Currently, the weighting of the user-item SVD is given a weighting of 1, known hereafter as the ID importance parameter. This makes this SVD approximately as important as the other 5 SVDs combined. Various values for the ID importance were tested and experimentally evaluated. It follows that the ID importance should be resultant from the personality correlations so that it is always proportionally identically weighted across domains where the strengths of the personality correlations may differ. Due to this reasoning, the ID importance is now calculated by multiplying some γ by the sum of the correlation scores. Best results were produced with $\gamma = 2.2$, this is used for all future implementations.

Finally, the bucketing precision of the FFM-SVD needed to be determined. Testing on various numbers of decimal places for the bucketing, between 1 and 10, gave no clear linear trend. The best results were consistently within the range of 1 to 4 and so a bucketing value of 2, which gave reasonable results across all datasets, was chosen.

When attempting to tune on the FFM-SVD++, initially the tuned parameters found for the FFM-SVD were set. Unfortunately, likely as a result of the oversampling data balancing, the FFM-SVD++ computation takes far too long to be considered further in this project with an estimated fitting time of around 54 hours for the balanced *Kindle* domain. As stated, FFM-SVD++ operates identically to FFM-SVD but with the inclusion of inferred data. It is therefore

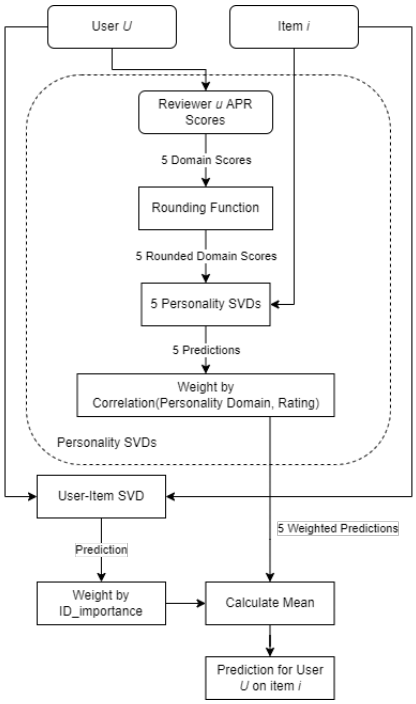

Fig. 4. FFM-SVD/FFM-SVD++ Architecture

expected that a FFM-SVD++ implementation would have better performance whilst following similar trends to the FFM-SVD approach, but it is infeasible for the remainder of this project.

The FFM-SVD proposed requires some time to fit to data, but the nature of the SVD allows predictions to be made extremely quickly once the model has been fitted. It is therefore suggested that a real-system would be able to fit an implementation of this model and store it, using it to generate predictions when needed in a timely manner and only refitting when a sufficient quantity of new data has been added to the system.

3.5.2 Tree-Based Regressors

LightGBM is a tree-based gradient boosting framework using histogram construction [47]. This model splits trees leaf-wise instead of using the level-wise splitting seen in other approaches, meaning that the leaf with the maximum delta loss is grown. The speed of the model is a result of reduced histogram building complexity by employing Gradient Based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [67].

In random forests, a random vector is sampled independently and the values retrieved are used as input parameters for an ensemble of decision trees. Averaging is then conducted over the decision trees to find best performing features, improving prediction accuracy and controlling overfitting [10]. As Random Forest often outperforms supervised learning alternatives [33], and LightGBM has been shown to consistently provide more accurate predictions than random forests (see section 2), this makes LightGBM a particularly promising choice of regression model. In addition to this, the LightGBM model gets its name from its ability to obtain predictions much quicker than alternatives, therefore

making it a suitable model for implementation in a real-world recommender system. LightGBM has been seen to perform well in recommender systems previously, forming numerous solutions in the ACM Recommender System Challenge 2019 [1], [39], and so it will be evaluated in the personality-aware recommender of this project. A random forest regression approach has been implemented, however, due to the significant computation time when operating on a reasonable amount of data, and experimental results indicating a decreased accuracy compared to LightGBM, it will not be considered further in this project's evaluation. This project's implementation does still include the random forest code for completeness.

A grid search with 3-fold cross validation was used for hyperparameter tuning. All values and explanations for default parameters can be found in [48]. The experimentally tuned hyperparameters are as follows: a learning rate of 0.3, 550 estimators, a maximum of 201 tree leaves, 100,000 subsamples for bin construction, and no minimum child weight or number of child samples.

3.5.3 Neural Network Regressor

A simple neural network regressor is implemented to demonstrate the potential for future applications of deep learning in personality-aware recommenders. This is a sequential model with three dense layers without significant hyperparameter tuning. The input layer has a dimension of size 6 to include the five personality traits and the identifier. The first two layers are given Rectified Linear Unit (ReLU) activation [11], but not the last. This is required so that the network does not behave as a linear unit. The regression is non-negative and so in theory ReLU could still be used, which converts negative values to zero, however, the linear function has been used instead as it does not limit the output. Originally, two dropout layers were implemented with probability of 0.2 in an attempt to reduce the risk of over-fitting. Following [59] which states that dropout has little to no benefit in classical regression problems of a similar nature to this project, experiments were conducted with the removal of these dropout layers. This removal, in some cases, resulted in an RMSE reduction of more than 20% and an increase in standard deviation. The dropout layers were therefore discarded.

Layer sizes were determined experimentally, and sizes of 20, 10, and an output size of 1, were chosen. The mean squared error was calculated and used as the loss for this model. An Adam optimizer has also been used as it is commonly seen in related works and has been shown to be reasonably robust to hyperparameter variations, as shown in [30]. All future results from this model have been run on 5 epochs as this has been experimentally seen to give enough time for some learning whilst also meeting time constraints of this project. Ordinarily, the loss between iterations would be tracked to create an end condition, however, the nature of the recommender system requires all users to be present in the training of the model and so this approach of only selecting a few in each iteration would introduce evaluation difficulties.

3.6 Evaluation Methods

This subsection details the evaluation methods used. This project focuses primarily on evaluating rating prediction accuracy as this is indicative of model performance within the system and allows for comparison with related and future works. By focusing on prediction accuracy, comparisons can then be made between domains and personality non-personality performances to observe the differences between the quality of predictions made. Making these comparisons allows this project to address its research question (see section 1.2). Furthermore, an analysis of prediction accuracy gives a clear insight into the feasibility of a real-world implementation.

In addition to determining prediction accuracy, with RMSE, the adjusted R-squared and standard deviation are included to observe if the model is learning effectively. The relative percentage error (RPE) for evaluating the effectiveness of results produced by various pre-filtering approaches, and the mean absolute personality (MAP) correlation for predicting the impact of personality on different domains, are both proposed for this project.

3.6.1 Model Evaluation

The root mean squared error (RMSE) is obtained for all regression models to enable a comparison and evaluation of overall performance. The RMSE is calculated as seen in (Eq. 8), where y_j is the predicted value and $\hat{y_j}$ is the true value. This is a common metric for regression models and so will also allow for comparison with future works. Furthermore, RMSE gives a higher weight to large errors than alternatives, such as the mean absolute error (MAE) which is more robust to outliers. This is useful as models which produce recommendation predictions with a large error are particularly undesirable as these items are expected to most severely impact user satisfaction.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})} \tag{8}$$

The adjusted R-squared metric (R_a^2) is also obtained for this project's evaluation. The R-squared (R2) metric is the coefficient of determination or the goodness of fit, and is the proportion of the variation in the dependent variable which is predictable in the independent variable. A value between 0 and 1 of the R2 score indicates the percentage of the variance of the data which the model can explain. This R2 score can be increased by adding independent variables to the model, but it will not decrease. As a result of this, irrelevant features have the potential to increase the R2 score despite having no benefit to the model. To remedy this, R_a^2 is calculated as shown in (Eq. 9), where n is the number of observations, k is the number of independent variables, SSr is the squared sum error of the regression line, and SSm is the squared sum error of the mean line.

$$R_a^2 = 1 - (\frac{n-1}{n-k-1} \cdot \frac{SSr}{SSm}) \tag{9}$$

The standard deviation of the predictions (Eq. 10) is used to indicate the nature of the learning that is being conducted. If an approach has a low standard deviation,

perhaps tending towards zero, it indicates that the model is learning to find some average result which reduces overall error and is the model is therefore an unsuitable predictor.

$$StdDev. = \sqrt{\frac{1}{n}\sum_{j=1}^{N}(x_j - \mu)} \quad \mu = \frac{\sum_{j=1}^{N} x_j}{N} \tag{10}$$

3.6.2 Pre-Filtering Evaluation

The RMSE scores for models acting on each domain with each pre-filtering technique are calculated. For each approach, the percentage difference between a result and the best result in a domain is found, and these values for all models are summed to produce a resultant overall score, a_r, for that pre-filtering technique. The percentage difference is used as any absolute error would give greater weighting to domains with high RMSEs.

To determine the effectiveness of the users which have been gathered by each approach, the a_r scores need to be considered with respect to the number of reviews used. The relative percentage error (RPE) score has been created for this purpose and finds the product of the normalized a_r scores and the normalized number of reviews n_r used (note that this is after the oversampling data balancing). This metric should indicate which approach provides users with the reviews that are most effective. These calculations are shown in (Eq. 11), considering model $m \in M$, domain $d \in D$, and RMSE scores s, for the score, a_r, for approach $r \in R$,.

$$RPE_r = \frac{a_r n_r}{\sum_{i \in R} a_i \cdot \sum_{i \in R} n_i}, \quad a_r = \sum_{m \in M}\sum_{d \in D} \frac{s_{m,d} - s_{best}}{\bar{s}} \tag{11}$$

3.6.3 Domain Evaluation

The correlation between personality traits and ratings are compared between different domains to indicate the degree of potential effectiveness that the inclusion of personality might provide to recommendations. This is analysed by the Mean Absolute Personality (MAP) correlation, which has been created for the purpose of this project, and is seen in (Eq. 12), where r_j is the Pearson correlation between each of the five personality traits to the ratings provided, with x as the personality variable and $\bar{x}$ being the mean, and y being the ratings variable in the sample.

$$MAP = \frac{\sum_{j=1}^{5} |r_j|}{5} \quad r_j = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{12}$$

The MAP correlation is compared to the actual percentage improvements seen in each domain to determine if it is a suitable predictor. It is also used to support a discussion on the validity of results.

3.6.4 Test User Selection

To evaluate the performance of the models in each domain, test users are found and their results averaged to reduce the effects of any outliers. All users in a study of a domain operate on the same number of pre-filtered users. However, these users will have varying numbers of reviews. As a result of this, different users may have more or less informed APR scores, potentially impacting the effectiveness of the models. It is therefore important, to conduct experiments over a range of users and combine their results, to increase confidence in the results obtained. Unfortunately, only a limited range can be used due to computational cost.

Each target user will have a different number of reviews, and so a different number of predictions. It would be unreasonable to assign the same weighting to users for which varying numbers of predictions were made. Therefore, the results from each user are weighted according to the number of predictions that the models were required to generate for them.

Initially a scoring was used such that for each user, r, the score, s_r, was given by (Eq. 13), where d was the RMSE between the user's number of reviews in each rating and a balanced distribution, and t is the total number of ratings given by a user. This approach however led to the inclusion of users with very small numbers of reviews as the ideal rating distribution was given unfair precedence for very well distributed users.

$$s_r = d * (t_{max} - t_r) \tag{13}$$

Therefore, experimental results suggest that a suitable approach is instead to select the top x% of users ranked by their total number of reviews, and then select from these the n users with the best distribution across the ratings categories. This approach ignores float values between 1 and 5 which may have been created from average identical user-item pairs, as these are rare and unlikely to affect results. After some experimentation, the top 0.1% of users ranked by their total number of reviews have been chosen from which n can be selected. Even for the *Video Games* domain with approximately 55,000 users, this top 0.1% produces 55 users to be ranked. These n chosen users therefore have a large number of reviews as well as a reasonable distribution of ratings. It is expected that n will be significantly lower than 50 due to computational limitations of testing on more users.

4 Results

This section presents the procedure, results and findings of the experimentation around the proposed methodology. Some analysis of the findings is used to inform later tests.

4.1 Comparison of Pre-Filtering Approaches

For the FFM-SVD and LightGBM models, each pre-filtering approach (see section 3.4) has been tested on all domains. These results can be found in Table 3. In brackets is the standard deviation and in bold is the best result for a domain. As the performance of the model itself is not considered at this stage, the LightGBM and FFM-SVD models do not have tuned hyperparameters. The default parameters of both the SVD and LightGBM models can be found at [69] and [48] respectively. Both models are trained on only the 5 personality traits and the item ID and results are averaged over the top 10 users (see section 3.6.4). A 70/30 train-test split acting on the target user's data has been found to be a

TABLE 3
Weighted average RMSE over Top 10 Users for a Baseline Personality LightGBM and FFM-SVD, operating on 5% of each domain's users with various pre-filtering approaches under a 70/30 train-test split, with balanced data by equal oversampling, and an FFM-SVD bucketing of 2 d.p.

Model	Approach	Movies	Music	Kindle	Video Games	Pet Supplies	Sports & Outdoors	Garden
FFM-SVD	RC	1.07682 (0.364)	0.79623 (0.254)	**0.67674 (0.217)**	1.11272 (0.337)	1.00806 (0.455)	0.80579 (0.314)	0.85139 (0.319)
	MR	**1.02658 (0.442)**	0.79607 (0.344)	0.69172 (0.283)	**1.08869 (0.425)**	**0.96184 (0.446)**	0.80100 (0.355)	**0.80316 (0.360)**
	SP	1.07790 (0.286)	0.81619 (0.221)	0.69962 (0.132)	1.12691 (0.230)	0.97676 (0.402)	0.81042 (0.246)	0.89068 (0.251)
	PN	1.03290 (0.432)	**0.78092 (0.293)**	0.69259 (0.270)	1.11071 (0.390)	1.01530 (0.471)	0.80394 (0.371)	0.83143 (0.330)
	LinSP	1.06548 (0.365)	0.81881 (0.275)	0.68620 (0.228)	1.12032 (0.315)	1.00194 (0.422)	**0.78313 (0.270)**	0.88228 (0.313)
	LogSP	1.05838 (0.356)	0.79980 (0.268)	0.69057 (0.217)	1.08936 (0.312)	1.04585 (0.385)	0.78617 (0.315)	0.84981 (0.341)
LGBM	RC	**1.22904 (0.549)**	0.98298 (0.646)	0.78860 (0.446)	1.33131 (0.805)	1.18217 (0.609)	0.98541 (0.512)	**0.87970 (0.559)**
	MR	1.25513 (0.360)	0.95447 (0.475)	0.76818 (0.282)	1.27815 (0.749)	1.19454 (0.453)	1.05970 (0.489)	0.94019 (0.568)
	SP	1.26504 (0.565)	1.09314 (0.680)	**0.73242 (0.294)**	1.35656 (0.835)	**1.04573 (0.557)**	**0.92952 (0.467)**	0.94410 (0.529)
	PN	1.24658 (0.516)	**0.94973 (0.596)**	0.79632 (0.416)	1.34940 (0.826)	1.16835 (0.514)	1.02119 (0.525)	0.88652 (0.543)
	LinSP	1.26316 (0.616)	0.99581 (0.644)	0.77372 (0.414)	**1.26235 (0.790)**	1.20233 (0.652)	1.00782 (0.571)	0.93428 (0.596)
	LogSP	1.27941 (0.572)	1.01477 (0.664)	0.78575 (0.411)	1.34781 (0.846)	1.13996 (0.619)	0.95210 (0.516)	0.88301 (0.599)

TABLE 4
RPE Scores for each Pre-Filtering Approach, calculated using FFM-SVD and LightGBM results

Approach	RPE
Random Choice (RC)	0.016564
Most Reviews (MR)	0.066095
Stratified Personality (SP)	**0.015650**
Personality Neighbourhood (PN)	0.020149
Linear Stratified Personality (LinSP)	0.020435
Logarithmic Stratified Personality (LogSP)	0.016889

reasonable separation in the literature, [27], and is used in the remainder of this project.

The RPE scores (Eq. 11) for each pre-filtering approach are presented in Table 4. The experiment in Figure 5 was also conducted, with reduced data for LightGBM to decrease computation time, to allow for an analysis of the effect of balancing on the results. A strong Pearson correlation of 0.9976 was found. This indicates that the process of data balancing, whilst capable of improving model performance (see section 3.4.5), does not affect the trends of the results. This increases confidence that the act of equal oversampling does not significantly impact the conclusions made in this project. In both sets of results the *SP* approach performed best with an RPE of 0.015650 in the full test. Therefore, all future experimentation in this project uses the *SP* approach on 5% of the users in the chosen domain.

4.2 Regression Model Performance

A comparison of the performance of the LightGBM, FFM-SVD/SVD, and Neural Network model performance on various domains, including and excluding personality data, can be found in Table 5. The personality models' performances are visualised in Fig. 6. In this analysis, all models operate on a minimal amount of data. Non-targeted models are trained only on user and item identifiers and personality-models only extend this to include the five personality traits. This ensures that any differences between model counterparts are solely a result of the inclusion of the personality data. The primary objective of this evaluation is to determine the effectiveness of the personality aspect, and not to create the most accurate recommender.

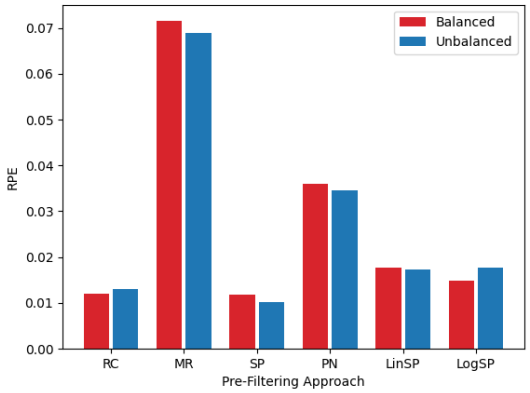

Fig. 5. RPE Comparison of Pre-filtering approaches acting on balanced and unbalanced data

As seen in Table 5, the FFM-SVD consistently outperforms the alternative approaches with even the non-personality SVD providing lower RMSEs than the personality LightGBM. The personality-aware neural network outperforms the non-personality LightGBM but the neural network approach is the least accurate of the three. The inclusion of personality appears to reduce the standard deviation in the FFM-SVD when compared to the SVD, but significantly increases it in the case of the LightGBM models.

The AR2 metric was recorded for both the personality LightGBM and FFM-SVD models in the *Kindle* domain experiment from Table 5. Scores of -0.6022 and -0.6358 were recorded for the LightGBM and FFM-SVD models respectively.

4.3 Domain Analysis

An analysis of the percentage improvement between a model's personality and non-personality counterparts, and therefore the effectiveness of the inclusions of personality in that domain, is found in Table 6. A percentage improvement measure has been used to ensure that the difference is not biased towards models which had higher initial RMSE scores and therefore greater potential for reduction.

The domain scores are then normalized and averaged to give a final domain score to compare to the MAP correlation

TABLE 5
Weighted average RMSE over Top 10 Users for each model including and excluding personality, operating on 5% of each domains users with *Stratified Non-targeted Personality pre-filtering*, SVD bucketing with 2.d.p., 70/30 train-test split, with balanced data by equal oversampling. Std. dev in brackets. Dashed line to separate results which are not considered further

Model	Movies and TV	Music	Kindle	Video Games	Pet Supplies	Sports & Outdoors	Garden
LightGBM (Pers.)	1.26504 (0.565)	1.09314 (0.680)	0.73242 (0.294)	1.35656 (0.835)	1.04573 (0.557)	0.92952 (0.467)	0.94410 (0.529)
LightGBM	1.59871 (0.265)	1.75388 (0.264)	1.38887 (0.175)	1.60864 (0.423)	1.55321 (0.228)	1.81620 (0.189)	1.78838 (0.308)
FFM-SVD	**1.07909 (0.356)**	**0.78746 (0.262)**	**0.67345 (0.174)**	1.11798 (0.303)	**0.95477 (0.420)**	**0.84098 (0.287)**	**0.93013 (0.307)**
SVD	1.11694 (0.459)	0.82569 (0.339)	0.68132 (0.220)	**1.11677 (0.403)**	0.97852 (0.519)	0.89522 (0.390)	0.94892 (0.388)
NeuralNet (Pers.)	1.567354 (0.101)	1.500775 (0.177)	1.101958 (0.127)	1.491804 (0.167)	1.647030 (0.137)	1.857303 (0.053)	1.468831 (0.053)
NeuralNet	1.608173 (0.147)	1.721087 (0.079)	1.588961 (0.026)	1.544432 (0.114)	1.589423 (0.065)	1.793896 (0.000)	1.765196 (0.112)

TABLE 6
Percentage improvement over LightGBM and FFM-SVD/SVD models for each domain, from Table 5, with associated MAP correlation score

Domain	% Improvement		Avg. of Normalized	MAP
	LightGBM	FFM-SVD		
Movies and TV	20.87%	3.39%	0.128	0.254
Music	37.67%	4.63%	0.194	0.155
Kindle	47.27%	1.16%	0.124	0.196
Video Games	15.67%	-0.11%	0.029	0.152
Pet Supplies	32.67%	2.43%	0.127	0.058
Sports & Out.	48.82%	6.06%	0.253	0.066
Garden	47.21%	1.98%	0.145	0.113

calculated for that domain. This inclusion of the MAP correlation score allows for a comparison between the predictor and the actual effectiveness.

The calculated MAP correlation scores predicted the media domains to be most affected by personality. The score of 0.254 is given to the *Movies and TV* domain, with the lowest score of 0.058 being assigned to *Pet supplies*. Contrastingly, the highest experimental result for the average of the normalized improvements was given to the non-media *Sports & Outdoors* domain and the lowest score given to the media *Video Games* domain.

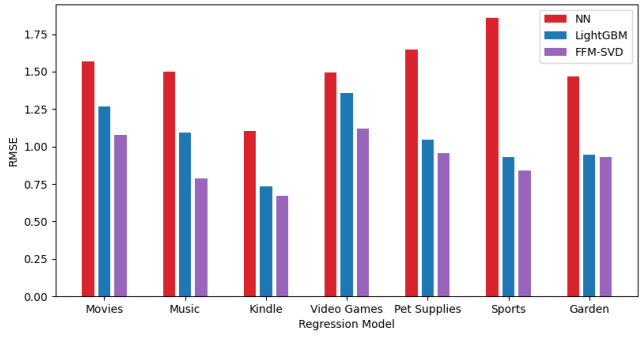

Fig. 6. RMSE comparison of personality models on each domain

5 Evaluation

This project aims to determine if and how automatic personality inference can be incorporated into a multi-domain recommender system to improve the recommendations generated. This project demonstrates that APR techniques can be combined with recommender pre-filtering approaches and regression models to positively impact recommendation accuracy. A multi-domain capable recommender system has been developed and the *SP* pre-filtering approach has been shown to use personality to outperform alternatives. Furthermore, APR scores have been used to develop the FFM-SVD which improves upon the standard SVD model. In nearly all cases, the inclusion of personality has reduced the RMSE in predictions.

This section will proceed to discuss all results found and draw conclusions. Each of the initial aims and objectives outlined (see section 1.2) will be addressed. In addition to the discussion of results, ethical implications of the methodology are considered. Finally, the validity of the methods used are discussed and are accompanied by a description of this project's evaluation challenges.

5.1 Discussion of Results

5.1.1 *Pre-filtering approaches*

As seen in Table 3, the weighted average RMSE over the top 10 users for each domain, after data-balancing, was recorded for both the FFM-SVD and LightGBM models. The RPE scores were then calculated from these results, as determined by (Eq. 11), and are seen in Table 4. The RPE results reveal the best approach to be the *non-targeted stratified personality (SP)* approach. The model obtains the most varied data from this approach as it aims to represent the entirety of the data, irrespective of the target user.

The worst performing approach, *MR*, highlights how the performance of the models is more impacted by the quality of the personality data than the quantity, a desirable outcome for this project. The *personality neighbourhood* approach was also outperformed by alternatives. This is likely a result of this approach excluding most of the range of the personality data. This will result in not gaining valuable information from contrasting users, and also making models more susceptible to smaller variations in the data which may be resultant from errors in APR predictions.

LinSP and *LogSP* being outperformed by *SP* may be due to them acting on the absolute difference to the target user, which doesn't consider the traits in which this difference appears. As a result of this, whilst a varied range of differences and similarities to the target user are obtained, there is no guarantee that among these gathered users there is a fair distribution between the personality traits. *LogSP* performed well, indicating that there is some strength to prioritising more similar users in this collaborative approach. This merits further exploration (see section 6.2).

It must also be noted that the *RC* approach performed surprisingly well. Random selection aims to obtain a randomly distributed representation of the data from which the models can fit. This hopes to eliminate bias, as hypothesised in the proposal of this pre-filtering technique.

5.1.2 Regression Model Performance

As seen in Table 5 the personality-aware LightGBM and the FFM-SVD both outperform their non-personality counterparts. Observing the percentage improvements recorded in Table 6, the LightGBM model has a mean improvement of 35.74% over the 7 domains when including personality, compared to an improvement of 2.79% between FFM-SVD and SVD. The non-personality SVD model also often has a lower RMSE than the personality-aware LightGBM model, indicating that matrix factorisation approaches may be more suitable than tree-based regressors. This conclusion of the SVD approach outperforming LightGBM concurs with the conclusions made in [15], where SVD was found to be the more suitable approach when considering per-instance recommenders acting on clustered data.

Interestingly, in the results obtained for the *Video Games* domain, the traditional SVD has a lower RMSE than the FFM-SVD. This could be a result of only tuning the parameters on a single domain as the difference is small.

It should also be noted that the personality LightGBM approach has the highest standard deviation. This suggests that it is the most likely model to be attempting to distinguish between results and not towards some learned average value. The dramatically increased standard deviation seems to be a benefit of including the personality data. However, in the hyperparameter tuning of this model, increasing the number of estimators led to decreased performance. So, whilst this increased standard deviation is indicative of attempted learning, it does not mean that further training would allow this model to produce more accurate results than the FFM-SVD. Interestingly, the FFM-SVD also has a lower standard deviation than the SVD approach. The cause of this is could be the averaging of multiple SVDs making particularly high or low predictions less probable.

Both the LightGBM and FFM-SVD approaches recorded negative AR2 scores. This indicates that features have been added which improve the model by less than expected. An interpretation of this is that the explanation towards response is negligible and there is an insignificant presence of explanatory variables. This is typically a result of too many poorly correlating features (see section 6.2).

One consideration with the FFM-SVD model is that it is inflexible to the addition of other data. For example, a best implementation would likely weight reviews by votes from other users, among other additional features. Each added feature would require an adaptation to the FFM-SVD approach. It is therefore suggested that the best implementation would perhaps be some form of hybrid scheme, joining the FFM-SVD with another model, potentially LightGBM.

The personality-aware neural network implementation is the worst performing model of those implemented, in terms of RMSE. The model outperformed its non-personality counterpart, further increasing confidence that personality-data does improve recommendation performance. It must, however, be noted that the standard deviation of the neural network approach is considerably lower than that of the other models, even having no standard deviation in the case of the non-personality-aware model in the *Sports & Outdoors* domain. This is indicative of the model converging towards some learned average. Furthermore, as stated earlier, the neural network was run on a limited number of epochs, and so clearer convergence would be expected given more training time. To stop this, a more sophisticated model and the inclusion of more of the non-personality data should be considered. This was not included in this paper so that the model would still be comparable to alternatives. Due to the low confidence that these results are representative of true learning, the neural network will not be included in future evaluations in this paper, for example, when calculating RPE scores for pre-filtering approaches.

5.1.3 Domain Analysis

Both the LightGBM and FFM-SVD models have best overall prediction accuracy on the *Kindle* domain, with the LightGBM also showing the highest percentage improvement when including personality. The overall performance is not, however, a reliable metric, due to different domains having varying amounts of data. Instead, to best analyse the effectiveness across domains, the normalized percentage improvements should be considered and compared to the MAP correlation scores.

The MAP correlation scores for each domain can be found in Table 6. The scores produced seem intuitive, with higher MAP correlations being assigned to media domains. A highest score of 0.254 was found for the *Movies and TV* domain and a lowest score of 0.058 for the *Pet Supplies* domain. This fits the project's initial reasoning that media-domains typically have more room for preference and so would be more effected by personality. Furthermore, a study in [24] on the cosine-based personality user profiles found correlations between preference- and personalty-based similarities. This demonstrated that from the domains: *movies*, *books*, and *music*, *movies* had the highest correlations, followed by *books*, and then *music*. This aligns with the MAP correlation scores presented.

Despite the intuitive MAP correlation scores, there is no convincing relationship between them and the average normalized percentage improvement. A Pearson correlation of only 0.04 between the resulting scores and the MAP correlations was found. Surprisingly, the domain most impacted by the personality data was the *Sports & Outdoors* domain. This domain was not used to tune the hyperparameters, and did not have a significantly high or low number of reviews which may have affected results. It is therefore intriguing that this domain shows the biggest improvement. This could perhaps be a result of different domains requiring different hyperparameter tuning. Alternatively, it could be due to the non-personality models being affected by using the their personality-counterparts tuned hyperparameters. More extensive testing, which is unavailable with this project's computational capabilities, is required to determine why these experimental values differ from the MAP correlations' intuitive scores.

5.2 Ethical Considerations

There are numerous ethical concerns which require consideration before the real-world implementation of personality-aware recommender systems. This subsection will aim to address some of these as an advanced objective for this project. These considerations are specifically relating to personality-aware recommender systems and not those issues seen in all recommenders such as addictiveness and the potential for an individual to harm or benefit an item by leaving fake reviews.

One system consideration is that the private nature of personality data requires it to be secured and kept in a manner in accordance with any data protection regulations, for example, GDPR legislation [52]. Additional policy approaches must be considered. This includes privacy-enhancing system architectures, such as decentralised databases, and algorithmic measures, such as encryption [52]. Furthermore, it is suggested that in a real-world implementation, the usage of personality data should be optional and withdrawable. Users must be informed of how their data is used.

Another consideration is the accuracy of the APR technique [21]. If an automatic method is being employed to approximate personality traits, which will in turn affect the users' experience, then the APR scores should be as accurate as possible. Care must be taken when assigning traits to users' personalities which may not be typically deemed as socially favourable, such as with *neuroticism*, and the visibility of such information.

If the personality-aware components of this system were to be combined with a real-world implementation, where more data about each user is present and more preferences are available, a further ethical consideration would be the possibility of user-reidentification. User re-identification is not possible from the anonymised and limited data used in this study, but should be considered if it is joined to an existing implementation. There is, however, the possibility of maliciously inferring the personality scores of users through the recommendations made [21].

All users are normalized with respect to each other and so, if the personality data is too limited, this may result in users having their personality scores incorrectly altered. This in turn could have a negative impact on individuals and so must be considered. One benefit of this normalization is that user's personalities will be distributed across all possible values. This reduces the likeliness of system bias against particular items, as all items will have the chance to correlate to some personality. If said item is not recommended to any users, this will be a result of the past experiences with the item, as desired, and should not be a symptom of system bias.

5.3 Validity and Evaluation Challenges

Where possible, 3-fold cross validation has been employed to increase confidence in the results produced and hyperparameters chosen. Unfortunately, this would have been too computationally expensive to incorporate into the FFM-SVD experimentations. To ensure reproducibility, a random seed of 42 has been set for all random components of the methodology.

When conducting the evaluation in this project, it must be considered that different domains and pre-filtering approaches will result in the selection of different users. Whilst the size of this selection can be set, the resultant number of reviews will be varying. In an attempt to maintain the validity of comparisons, any metrics used which compare across different sets of users use some form of normalization or weighting by the number of reviews present.

An additional challenge arises with the tuning of models. The personality models have been tuned and are compared to their non-personality counterparts acting on the same parameters. This could therefore provide an unfair advantage to personality-aware models. However, tuning the non-personality models by themselves would introduce numerous questions regarding the validity of the comparison, for example, with the two models training over varying numbers of iterations. Therefore, to maintain confidence in comparisons made, the personality-tuned parameters have been used in both models. It must be noted that the performance of the non-personality models could be sub-optimal. Furthermore, due to computational requirements, hyperparameter tuning could only be conducted on the *Kindle* domain, which could affect the conclusions of this study by favouring this domain in particular.

As discussed, RPE scores between unbalanced and balanced experiments have been compared and shown to produce similar results. This suggests that the equal oversampling data balancing approach and all subsequent results are valid as it has not significantly altered the trends produced in the testing of the pre-filtering approaches.

This project's system architecture is reliant on the accuracy of the chosen APR scores. The correctness of LIWC used for the APR has been discussed (see section 2) and is believed to be suitably reliable. As for the correctness specific to this project's implementations, for example, with the parsing of tokens and the normalization of categories, these scores cannot be officially validated in this paper without an online study (see section 6.2). There are, however, a couple of indicators of correctness from which the validity of this approach may be assumed. First, the data analysis conducted (see section 3.2) suggests clear trends between certain personality traits and ratings behaviours. Secondly, it has been proven that the inclusion of personality data has improved recommender performance. This improvement is in itself contributing proof that there is some validity to the scores obtained as they must contain trends of some value to which the models can fit. Furthermore, the MAP correlation scores are intuitive for each domain, further increasing confidence in the correctness of the personality scores. The choice of adopting an APR instead of using explicit acquisition has also removed certain validity considerations such as social desirability bias, [32], [44], and the reference-group effect, [21].

6 Conclusion

This section summarises the contributions of this project. Potential developments are also suggested to extend this project and field in future works.

6.1 Summary and Contributions

This project's research question regarded how personality data can be used to improve the performance of a multi-domain recommender system. A comparison of different personality formats has been conducted and the FFM model has been chosen with supported reasoning. To obtain the data for this representation, an APR approach utilising the LIWC dictionary has been implemented and proven to enhance prediction accuracy. This further validates the dictionary and the associated correlations from the study in [74]. Data analysis on the APR results was then conducted and used to inform the creation of the FFM-SVD model.

The effectiveness of incorporating personality in different domains has been evaluated and discussed, along with the potential challenges which may affect the conclusions made. The MAP correlation score was also developed and provides intuitive predictions on the usefulness of personality in different domains. These scores do not, however, align with the results and so increases the likeliness that further exploration is required to assure the validity of this study. Personality was found to be most effective in the *Sports & Outdoors* domain, followed by *Music*, and *Kindle* books respectively.

This domain analysis was conducted across several models, with the project evaluating SVD, tree-based, and deep learning approaches. The FFM-SVD has been created and shown to outperform alternatives. The optimal system found in this study involved the FFM-SVD model proposed by this paper, acting with *non-targeted stratified personality* sampling, a bucketing precision of 2 d.p., a 70-30 train test split, and with equal oversampling data balancing.

The LightGBM model has also been developed and its performance evaluated. This is believed to be the first LightGBM personality-aware recommender implementation. In keeping with the literature, [41], experimental testing during the development phase of this project suggested that the LightGBM model does produce more accurate recommendations to the commonly used random forest approach. No results were gathered for this comparison due to the random forest model being a computationally expensive approach when operating on a reasonable amount of data. This is another factor which highlights the suitability of the LightGBM model as an alternative.

An *advanced* aim of this project was to explore the usage of neural networks in recommender systems. A neural network has been constructed and the potential for deep learning personality-aware recommender systems has been discussed. The approach implemented tended to be improved by the inclusion of personality-data. However, the simple model with three dense layers resulted in convergence and so either a more sophisticated model is required or the addition of other information in the dataset is required to make the model better informed.

In addition to the models constructed, personality-aware pre-filtering techniques have been developed and the RPE metric has been created to determine the effectiveness of each approach. This RPE score was then used to confirm that the equal oversampling data balancing approach proposed in this project did not significantly alter results whilst still benefiting model performance. The *non-targeted stratified personality* approach was found to be the most effective choice of user pre-filtering. The *LogSP* approach was also proposed in this project and, whilst performing slightly worse than *SP*, still produced reasonable results.

Throughout the development of all approaches in this project, the potential for real-world implementation has informed design choices and ensured that the final solutions are capable of being combined with existing systems. In addition to this, the associated ethical considerations have been discussed at length.

This project has addressed all *basic*, *intermediate*, and *advanced* objectives proposed in section 1.2. The research question has been answered by developing an APR technique whose results inform pre-filtering approaches and multi-domain regression models.

6.2 Future Developments and Limitations

To conclude this project and complete the objectives stated in section 1.2, this section outlines possible further developments to various aspects of this project. Furthermore, comments on research which could extend the field of personality-aware recommenders are presented.

One significant development which could benefit the approaches in this project would be an updated analysis of the LIWC-trait correlations from the NIH study in [74]. These correlations involve categories from LIWC2001 which differ noticeably from those found in LIWC2015 which was used in this project. As a result of this, only a portion of the LIWC2015 categories could be used. Updating the correlations for a more recent version of the LIWC dictionary would improve the accuracy of the APR and validate the results obtained.

A machine learning approach to validate the APR scores obtained could also perhaps be introduced. For example, a comparison of the APR results to the findings in [66], where the differences between personality trait scores for different genders are explored, could be used to inform scores. Other profile factors such as age or cultural indicators could also be considered. This would introduce numerous ethical implications but is worth consideration.

Recommender performance could also be improved by observing social-ties through network analysis between users and assuming that connections may indicate similarities in personality traits. This could not be included in this project due to dataset restraints. Another potential improvement could be the consideration of temporal contextual data, for example, giving a higher weighting to more recent reviews as these may be more reflective of a user's current personality. This follows from [19] which states that a user's personality profile should not be considered as unchanging. The study in [14] notes that personality trait scores are stable over a four-year period. This could perhaps be the window of consideration with temporal context.

In this study, several pre-filtering approaches have been tested however these should be expanded upon to determine an optimal approach. For example, a variation of LogSP, V-LogSP, could be introduced where instead of importance being reduced as bracket difference increases, importance is instead first given to the brackets with lowest and highest difference, and then reduced as it works inwards towards the median bracket. An alternative approach

could be to retrieve the selections within brackets with the best distribution across the data, instead of randomly.

Equal oversampling is used to help remedy the heavily skewed dataset, however, it places considerably more importance on the reviews in minority classes. Better alternatives should be considered. These alternative data balancing approaches, if not dramatically increasing the size of the data, could also allow for the implementation and evaluation of FFM-SVD++, which is expected to improve prediction accuracy. Similarly, as discussed in [15], LightGBM_EF, could improve upon the LightGBM implementation.

The FFM-SVD is inflexible to the addition of new features without changing the models' structure. A study into potential combinations of models with the FFM-SVD to create a hybrid scheme is therefore suggested. This combination could be with the LightGBM used in this project. Alternatively, a CB recommender could be used to address performance in the cold-start scenario, as seen in [8]. Furthermore, hyperparameter tuning across all domains, instead of just the *Kindle* domain, and over a larger selection of users would increase confidence in results. This could not be conducted in this study due to computational cost.

Low AR2 scores were recorded for both the LightGBM and FFM-SVD models, suggesting that there are poorly correlating features in the personality data. The introduction of some threshold value for a personality trait to be included as a feature in the model should be considered.

Post-filtering approaches could also improve recommendations. As stated in section 2, one content-based suggestion specific to media domains is to perform natural language processing and web scraping to obtain the sub-genres associated with items and use the genre-personality correlations study in [13] to influence predictions. The personality scores of each genre could be compared to the target user's scores to estimate suitability. This approach would only be available to particular domains and so would have affected the domain analysis conducted in this project if included.

A final future development regards neural networks. First, the model outlined in this project could only be run on an extremely limited number of epochs due to the size of the data required and the time constraints of this project. A useful further study would involve more training time. Additionally, as discussed (see section 6.1), a more sophisticated approach should be considered.

REFERENCES

[1] Adamczak, J., et al. "Session-based hotel recommendations dataset: As part of the ACM recommender system challenge 2019." *ACM Transactions on Intelligent Systems and Technology (TIST)* 12.1 (2020): 1-20.

[2] Aditya, P. H., I. Budi, and Q. Munajat. "A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for E-commerce in Indonesia: A case study PT X." *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, (2016).

[3] Adomavicius, G., B. Mobasher, F. Ricci, and A. Tuzhilin. "Context-Aware Recommender Systems". *AI Magazine*, vol. 32, no. 3, Oct. (2011), pp. 67-80, doi:10.1609/aimag.v32i3.2364.

[4] Adomavicius, G., A. Tuzhilin "Context-aware recommender systems". *In Recommender Systems Handbook*; Ricci, F., Rokach, L., Shapira, B., Eds.; Springer: Boston, MA, USA, (2015); pp. 191–226.

[5] Ajesh, A., N. Jayashree, P. S. Jijin. "A random forest approach for rating-based recommender system." *2016 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE, (2016).

[6] Apriyanto, S., A. Nurhayaty. "Born In Social Media Culture: Personality Features Impact In Communication Context." *icollit* (2019): 167-175.

[7] Asabere, N. Y., A. Acakpovi, M. B. Michael. "Improving socially-aware recommendation accuracy through personality." *IEEE Transactions on Affective Computing* 9.3 (2017): 351-361.

[8] Balakrishnan, V., H. Arabi. "HyPeRM: A hybrid personality-aware recommender for movie." *Malaysian Journal of Computer Science* 31.1 (2018): 48-62.

[9] Bobadilla, J., et al. "Recommender systems survey." *Knowledge-based systems 46* (2013): 109-132.

[10] Breiman, L. "Random Forests". M*achine Learning **45**, 5–32 (2001)*. https://doi.org/10.1023/A:1010933404324

[11] Brownlee, J. "A Gentle Introduction to the Rectified Linear Unit (ReLU)" (2019) Available online at: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural\ \-networks/ (Accessed: April 2022)

[12] Buettner, R. "Predicting user behavior in electronic markets based on personality-mining in large online social networks." *Electronic Markets* 27.3 (2017): 247-265.

[13] Cantador, I., I. Fernández-Tobías, A. Bellogín. "Relating personality types with user preferences in multiple entertainment domains." *CEUR workshop proceedings*. Shlomo Berkovsky, (2013).

[14] Cobb-Clark, D. A., S. Schurer. "The stability of big-five personality traits." *Economics Letters 115.1* (2012): 11-15.

[15] Collins, A., L. Tierney, J. Beel. "Per-instance algorithm selection for recommender systems via instance clustering." arXiv preprint arXiv:2012.15151 (2020).

[16] Costa Jr, P. T., R. R. McCrae. "The Revised Neo Personality Inventory (neo-pi-r)." *Sage Publications, Inc*, (2008).

[17] Dayton, E. "Amazon Statistics You Should Know: Opportunities to Make the Most of America's Top Online Marketplace" *bigcommerce.com* Available online at: https://www.bigcommerce.com/blog/amazon-statistics/#amazon-everything-to-everybody (Accessed: April 2022)

[18] Deniz, M. "An Investigation of Decision Making Styles and the Five-Factor Personality Traits with Respect to Attachment Styles." *Educational Sciences: Theory and Practice 11.1* (2011): 105-113.

[19] Dewaele, J. "Personality: Personality traits as independent and dependent variables." *Psychology for language learning.* Palgrave Macmillan, London, (2012). 42-57.

[20] Dhelim, S., N. Aung, H. Ning. "Mining user interest based on personality-aware hybrid filtering in social networks." *Knowledge-Based Systems 206* (2020): 106227.

[21] Dhelim, S., et al. "A Survey on Personality-Aware Recommendation Systems." *arXiv preprint arXiv:2101.12153* (2021a).

[22] Dhelim, S., et al. "Big-Five, MPTI, Eysenck or HEXACO: The Ideal Personality Model for Personality-aware Recommendation Systems." *arXiv preprint arXiv:2106.03060* (2021b).

[23] Dunn, G., et al. "Evaluating interface variants on personality acquisition for recommender systems." *International Conference on User Modeling, Adaptation, and Personalization.* Springer, Berlin, Heidelberg, (2009).

[24] Fernández-Tobías, I., I. Cantador. "Personality-aware collaborative filtering: an empirical study in multiple domains with facebook data." *International conference on electronic commerce and web technologies.* Springer, Cham, (2014).

[25] Fernández-Tobías, I., I. Cantador. "On the use of cross-domain user preferences and personality traits in collaborative filtering." *International Conference on User Modeling, Adaptation, and Personalization.* Springer, Cham, (2015).

[26] Funk, S. "Netflix Update: Try This at Home" (2006) Available online at: https://sifter.org/~simon/journal/20061211.html (Accessed April 2022)

[27] Gholamy, A., V. Kreinovich, O. Kosheleva. "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation." (2018).

[28] Goldberg, L. R. "An alternative" description of personality": the big-five factor structure." *Journal of personality and social psychology* 59.6 (1990): 1216.

[29] Gonsowski, J. C. "The Myers-Briggs Type Indicator: Mapping to circumplexand Five-Factor Models, 4/27." *Strona internetowa: http://tap3x. net/EMBTI/j5gonsowski. html* (1999).
[30] Goodfellow, I., Y. Bengio, A. Courville, (2016) "Deep Learning" *MIT Press* Available online at: http://www.deeplearningbook.org
[31] Govers, P. CM, J. PL Schoormans. "Product personality and its influence on consumer preference." *Journal of Consumer Marketing* (2005).
[32] Grimm, P. "Social desirability bias." *Wiley international encyclopedia of marketing* (2010).
[33] El Guabassi, I., et al. "Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance." *International Journal of Online and Biomedical Engineering*. (2021). 17.10.3991/ijoe.v17i02.20025.
[34] Hansjons Vegeborn, V., H. Rahmani. "Comparison and Improvement Of Collaborative Filtering Algorithms." (2017).
[35] Hastie, T., R. Tibshirani, J. Friedman, "The Elements of Statistical Learning (2nd ed.)", Springer. ISBN 0-387-95284-5, (2008).
[36] Hauke, J., T. Kossowski. "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data." *Quaestiones geographicae* 30.2 (2011): 87.
[37] HealthKnowledge 1a - Epidemiology "Methods of sampling from a population" (2022) *HealthKnowledge*, Available online at: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population (Accessed: April 2022)
[38] Hu, R., P. Pu. "Enhancing collaborative filtering systems with personality information." *Proceedings of the fifth ACM conference on Recommender systems.* (2011).
[39] Jankiewicz, P, et al. "Boosting algorithms for a session-based, context-aware recommender system in an online travel domain." *Proceedings of the Workshop on ACM Recommender Systems Challenge*. (2019).
[40] Kachur, A., et al. "Assessing the Big Five personality traits using real-life static facial images." *Scientific Reports 10.1* (2020): 1-11.
[41] Ke, G., et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* **30** (2017): 3146-3154.
[42] Koren, Y. "Yahoo Research." R. Bell, C. Volinsky, AT&T Labs—Research "Matrix Factorization Techniques for Recommender Systems" *IEEE Computer Society 0018-9162/09/$26.00© 2009* IEEE (2009).
[43] Kosinski, M., D. Stillwell, T. Graepel. "Private traits and attributes are predictable from digital records of human behavior." *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, (2013).
[44] Latkin, C. A., et al. "The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland." *Addictive behaviors* 73 (2017): 133-136.
[45] Leary, T. "Interpersonal diagnosis of personality". New York: Ronald., (1957).
[46] Lex, E., et al. "Psychology-informed recommender systems". Now Publishers, (2021).
[47] LightGBM "Documentation" *LightGBM* Available online at: https://lightgbm.readthedocs.io/en/latest/ (Accessed Feb. 2022)
[48] LightGBM "lightgbm.LGBMRegressor" *LightGBM* Available online at: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html (Accessed Feb. 2022)
[49] McCrae, R. R., and P. T. Costa Jr. "Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality." *Journal of personality* 57.1 (1989): 17-40.
[50] McDonnell, M., J. E. Owen, E. OC. Bantum. "Identification of emotional expression with cancer survivors: validation of linguistic inquiry and word count." *JMIR Formative Research 4.10* (2020): e18246.
[51] Mehta, Y., et al. "Recent trends in deep learning based personality detection." *Artificial Intelligence Review 53.4* (2020): 2313-2339.
[52] Milano, S., M. Taddeo, L. Floridi. "Recommender systems and their ethical challenges." *Ai & Society 35.4* (2020): 957-967.
[53] Myers, I. B.. "The Myers-Briggs Type Indicator: Manual (1962)." (1962).
[54] Nalmpantis, O., C. Tjortjis. "The 50/50 recommender: a method incorporating personality into movie recommender systems." *International Conference on Engineering Applications of Neural Networks.* Springer, Cham, (2017).
[55] Neehal, N., M. A. Mottalib. "Prediction of preferred personality for friend recommendation in social networks using artificial neural network." *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, (2019).
[56] Ni, J. "Amazon Review Data (2018)" *UCSD*. Available online at: https://nijianmo.github.io/amazon/index.html (Accessed Nov. 2021)
[57] Ni, J., J. Li, J. McAuley "Justifying recommendations using distantly-labeled reviews and fined-grained aspects" *Empirical Methods in Natural Language Processing (EMNLP)* (2019)
[58] Nunes, M., A. S. Netto. "Recommender systems based on personality traits." *Diss. Université Montpellier II-Sciences et Techniques du Languedoc,* (2008).
[59] Özgür, A., F. Nar. "Effect of Dropout layer on Classical Regression Problems." *2020 28th Signal Processing and Communications Applications Conference (SIU)*. IEEE, (2020).
[60] Pennebaker, J. W., et al. "The development and psychometric properties of LIWC2015" (2015).
[61] Pittenger, D. J. "Measuring the MBTI... and coming up short." *Journal of Career Planning and Employment* 54.1 (1993): 48-52.
[62] Potash, P., A. Rumshisky. "Recommender System Incorporating User Personality Profile through Analysis of Written Reviews." *EMPIRE@ RecSys.* (2016).
[63] Rentfrow, P. J., L. R. Goldberg, R. Zilca. "Listening, watching, and reading: The structure and correlates of entertainment preferences." *Journal of personality* 79.2 (2011): 223-258.
[64] Ricci, F., L. Rokach, B. Shapira. "Introduction to recommender systems handbook." *Recommender systems handbook*. Springer, Boston, MA, (2011). 1-35.
[65] Schild, C., K. Ścigała, and I. Zettler. "Reference group effect." *Encyclopedia of Personality and Individual Differences* (2018): 1-3.
[66] Schmitt, D. P., et al. "Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures." *Journal of personality and social psychology 94.1* (2008): 168.
[67] Sharma, A. "What makes LightGBM lightning fast?" *Towards Data Science*, (2018) Available online at: https://towardsdatascience.com/what-makes-lightgbm-lightning-fast-a27cf0d9785e (Accessed 27/03/2022)
[68] SPC for Excel "Are the Skewness and Kurtosis Useful Statistics?" *spcforexcel.com* (2016) Available online at: https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#:~:text=If%20the%20skewness%20is%20between%20%2D1%20and%20%E2%80%93%200.5%20or%20between%20the%20data%20are%20highly%20skewed (Accessed: April 2022)
[69] Surprise Docs. "Matrix Factorization-based algorithms" (Date unknown) Available online at: https://surprise.readthedocs.io/en/stable/matrix_factorization.html#surprise.prediction_algorithms.matrix_factorization.SVD (Accessed 01/03/2022)
[70] Tyagi, N. "L2 and L1 Regularization in Machine Learning" *analyticssteps.com*, (2021) Available online at: https://www.analyticssteps.com/blogs/l2-and-l1-regularization-machine-learning (Accessed 09/04/2022)
[71] Vinceslas, M. (2019) "How Does the Funk Singular Value Decomposition Algorithm work in Recommendation Engines" *Medium, Data Driven Investor.* (2019) Available online at: https://medium.datadriveninvestor.com/how-funk-singular-value-decomposition-algorithm-work-in\\-recommendation-engines-36f2fbf62cac
[72] Wu, W., L. Chen, L. He. "Using personality to adjust diversity in recommender systems." *Proceedings of the 24th ACM conference on hypertext and social media*. (2013).
[73] Yakhchi, S., et al. "Enabling the analysis of personality aspects in recommender systems." arXiv preprint arXiv:2001.04825 (2020).
[74] Yarkoni, T.. "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers." *Journal of research in personality 44.3* (2010): 363-373.
[75] Yi, M.Y., O.J. Lee, J. J. Jung. "MBTI-Based Collaborative Recommendation System: A Case Study of Webtoon Contents." *ICCASA*. Springer, Cham, (2015).
[76] Zhang, S., et al. "Deep learning based recommender system: A survey and new perspectives." *ACM Computing Surveys (CSUR) 52.1* (2019): 1-38.
[77] Zhou, T., et al. "Kernelized probabilistic matrix factorization: Exploiting graphs and side information." *Proceedings of the 2012 SIAM international Conference on Data mining. Society for Industrial and Applied Mathematics,* (2012).