Personality in Computational Advertising: A Benchmark

Giorgio Roffo
University of Verona
Department of Computer Science
Giorgio.Roffo@univr.it

Alessandro Vinciarelli
University of Glasgow
School of Computing Science
Alessandro.Vinciarelli@glasgow.ac.uk

ABSTRACT

In the last decade, new ways of shopping online have increased the possibility of buying products and services more easily and faster than ever. In this new context, personality is a key determinant in the decision making of the consumer when shopping. A person's buying choices are influenced by psychological factors like impulsiveness; indeed some consumers may be more susceptible to making impulse purchases than others. Since affective metadata are more closely related to the user's experience than generic parameters, accurate predictions reveal important aspects of user's attitudes, social life, including attitude of others and social identity. This work proposes a highly innovative research that uses a personality perspective to determine the unique associations among the consumer's buying tendency and advert recommendations. In fact, the lack of a publicly available benchmark for computational advertising do not allow both the exploration of this intriguing research direction and the evaluation of recent algorithms. We present the ADS Dataset, a publicly available benchmark consisting of 300 real advertisements (i.e., Rich Media Ads, Image Ads, Text Ads) rated by 120 unacquainted individuals, enriched with Big-Five users' personality factors and 1,200 personal users' pictures.

CCS Concepts

•**Information systems** → **Computational advertising;** *Collaborative search; Test collections;*

Keywords

Recommender Systems, Computational Advertising, Ads Click Prediction, Ads Rating Prediction, Personality Traits, Data Mining

1. INTRODUCTION

Nowadays, online shopping plays an increasingly significant role in our daily lives [10]. Most consumers shop online with the majority of these shoppers preferring to shop online for reasons like saving time and avoiding crowds. Marketing campaigns can create awareness that drive consumers all the way through the process to actually making a purchase online [16]. Accordingly, a challenging problem is to provide the user with a list of recommended advertisements they might prefer, or predict how much they might prefer the content of each advert.

Past studies on recommender systems take into account information like user preferences (e.g., user's past behavior, ratings, etc.), or demographic information (e.g., gender, age, etc.), or item characteristics (e.g., price, category, etc.). For example, collaborative filtering approaches first build a model from a user's past behavior (e.g., items previously purchased and/or ratings given to those items), then use that model to predict items (or ratings for items) that the user may have an interest in by considering the opinions of other like-minded users. Other information (e.g., contexts, tags and social information) have also taken into account in the design of recommender systems [5, 18, 20].

The impact of personality factors on advertisements has been studied at the level of social sciences and microeconomics [2, 9, 35]. Recently, personality-based recommender systems are increasingly attracting the attention of researchers and industry practitioners [6, 15, 33]. Personality is the latent construct that accounts for "*individuals characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns*" [12]. Hence, personality is a critical factor which influences people's behavior and interests. Attitudes, perceptions and motivations are not directly apparent from clicks on advertisements or online purchases, but they are an important part of the success or failure of online marketing strategies. A person's buying choices are further influenced by psychological factors like impulsiveness (e.g., leads to impulse buying behaviors), openness (e.g., which reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has), neuroticism (i.e., sensitive/nervous vs. secure/confident), or extraversion (i.e., outgoing/energetic vs. solitary/reserved) which affect their motivations and attitudes [35].

To the best of our knowledge, the impact of personality factors on advertisements has been largely neglected at the level of *advert recommendation*. There is a high potential that incorporating users' characteristics into recommender systems could enhance recommendation quality and user experience. For example, given a user's preference for some items, it is possible to compute the probability that they are of the same personality type as other users, and, in turn, the probability that they will like new items [24].

Moreover, personality has shown to play an important role also in other aspects of recommender systems, such as implicit feedback, contextual information [21], affective content labeling [34]. With the development of novel techniques for the unobtrusive acquisition of personality (e.g. from social media [7, 28, 29]) this study is meant to contribute to this emerging domain proposing a new corpus which includes questionnaires of the Big-Five (BFI-

EMPIRE 2016, September 16, 2016, Boston, MA, USA.

10) personality model [25], as well as, users' liked/disliked pictures that convey much information about the users' attitudes [7].

The ADS Dataset is a highly innovative collection of 300 real advertisements rated by 120 participants and enriched with the users' five broad personality dimensions, which have been shown to capture most individual differences [4]. The user study is conducted by recruiting a set of test subjects, and asking them to perform several tasks. The subjects included in the corpus were recruited through a public platform purely dedicated to recruiting participants. The process was stopped once the first 120 individuals answered positively. The experimental protocol adopted for the data collection has been designed to capture users' preferences in a controlled usage scenario (see Section 2.1 for further details).

In this work we carry out prediction experiments performing two different tasks: *ad rating prediction* or *ad click prediction*, with the goal in mind to analyze the effect of using personality data for recommending ads. Therefore, we propose Logistic Regression (LR) [8], Support Vector Regression with radial basis function (SVR-rbf) [3], and L2-regularized L2-loss Support Vector Regression (L2-SVR) [8] as baseline systems for recommendation. We then review a large set of properties, and explain how to evaluate systems given relevant properties. We also survey a large set of evaluation metrics in the context of the property that they evaluate, and provide a library within one integrated toolbox.

Summarizing, the contribution of this work is two-fold:

Dataset: we collect and introduce a representative benchmark for computational advertising enriched with affective-like metadata such as personality factors. The benchmark allows to (i) explore the relationship between consumer characteristics, attitude toward online shopping and advert recommendation, (ii) identify the underlying dimensions of consumer shopping motivations and attitudes toward online in-store conversions, and (iii) have a reference benchmark for comparison of state-of-the-art advertisement recommender systems (ARSs). To the best of our knowledge, the ADS dataset is the first attempt at providing a set of advertisements scored by the users according to their interest into the content.

Code library: we present two broad classes of prediction accuracy measures, depending on the task the recommender system is performing: "ad rating prediction" or "ad click prediction", and provide a code library, integrating the evaluation metrics with uniform input and output formats to facilitate large scale performance evaluation. The code library and the annotated dataset are available on the *project page*[1].

The rest of the paper is organized as follows: in Section 2 we present and describe the ADS Dataset. We perform a corpus analysis investigating on the linkages between buying habits, recommendations, and personality. In Section 3, we survey a large set of evaluation metrics in the context of the property that ARSs evaluate. In Section 4 we conduct experiments for each scenario taken into account in this work, investigating on the strengths and weakness of using personality data as features for recommendation. Finally, in Section 5 conclusions are given, and future perspectives are envisaged.

2. CORPUS ANALYSIS

The corpus includes 300 advertisements voted by unacquainted individuals (120 subjects in total. Note, the data collection process is still running). Adverts equally cover three display formats: Rich Media Ads, Image Ads, Text Ads (i.e., 100 ads for each format). Participants rated (from 1-star to 5-stars) each recommended advertisement according to if they would or would not click on it (some examples are shown in the Fig.1). We labeled adverts as "clicked" (rating greater or equal to four), otherwise "not clicked" (rating less than four). The distribution of the ratings across the adverts that were scored by the users turns out to be unbalanced

[1]http://giorgioroffo.it/?ADSdataset

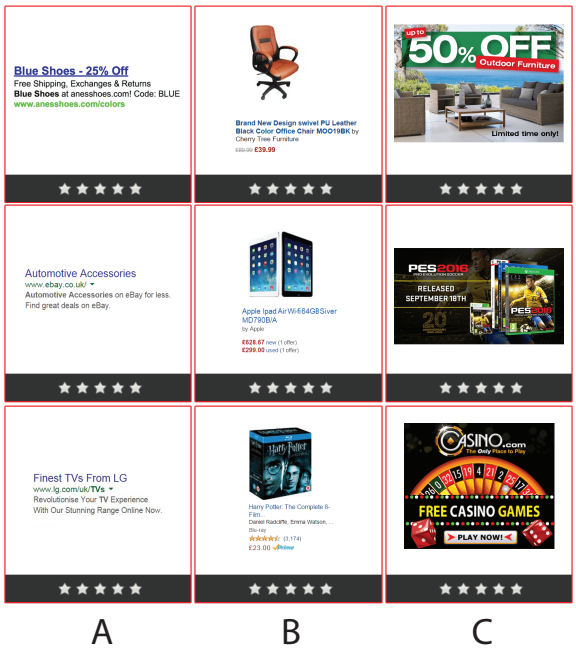

Figure 1: The figure shows three different examples for each display format. (A) Shows Text Ads that received 26.5% of the total amount of clicks. (B) Image Ads (32.7% of clicks), and (C) Rich Media Ads (40.8% of clicks).

Class Labels	Category Name	% Clicks
1	**Clothing & Shoes**	**6.2%**
2	Automotive	3.3%
3	Baby	3.3%
4	Health & Beauty	6.0%
5	Media	6.6%
6	**Consumer Electronics**	**9.2%**
7	**Console & Video Games**	**8.5%**
8	Tools & Hardware	3.0%
9	Outdoor Living	5.6%
10	**Grocery**	**7.3%**
11	Home	4.7%
12	Betting	1.6%
13	Jewelery & Watches	5.9%
14	Musical instruments	3.6%
15	Stationery & Office Supplies	5.4%
16	Pet Supplies	3.1%
17	Computer Software	5.6%
18	Sports	5.0 %
19	Toys & Games	5.1%
20	Social Dating Sites	1.0%

Table 1: ADS Dataset provides a set of 15 real adverts categorized in terms of 20 product/service categories. The most clicked categories are highlighted in green and the less clicked in red.

Group	Type	Description	References
Users' Preferences	Websites, Movies, Music, TV Programmes, Books, Hobbies	Categories of: websites users most often visit (WB), watched films (MV), listened music (MS), watched T.V. Programmes (TV), books users like to read (BK), favourite past times, kinds of sport, travel destinations.	[14, 18, 20]
Demographic	Basic information	Age, nationality, gender, home town, CAP/zip-code, type of job, weekly working hours, monetary well-being of the participant	[20]
Social Signals	Personality Traits	BFI-10: Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN)	[4, 25]
	Images/Aesthetics	Visual features from a gallery of 1.200 *positive / negative* pictures and related meta-tags	[7]
Users' Ratings	Clicks	300 ads annotated with Click / No Click by 120 subjects	[14, 23, 37]
	Feedback	From 1-star (Negative) to 5-stars (Positive) users' feedback on 300 ads	[14, 23, 37]

Table 2: The table reports the type of raw data provided by the ADS Dataset. Data of the first and last group can be considered as historical information about the users in an offline user study.

(4,841 clicked vs 31,159 unclicked).

Advert content is categorized in terms of 20 main product/service categories. For each one of the categories 15 real adverts are provided. Table 1 reports the full list of the categories used with the associated class annotations and the percentage of clicks received. At the category level, the distribution of the ratings results to be balanced (1,229 clicked vs 1,171 unclicked), where a category is considered to be clicked whenever it contains at least one clicked advert.

Inspired from recent findings which investigate the effects of personality traits on online impulse buying [2, 9, 35], and many other approaches based upon behavioral economics, lifestyle analysis, and merchandising effects [2, 19], the proposed dataset supports a trait theory approach to study the effect of personality on user's motivations and attitudes toward online in-store conversions. The trait approach was selected because it encourages the use of scientifically sound scale construction methods for developing reliable and valid measures of individual differences. As a result, the corpus includes the Big Five Inventory-10 to measure personality traits [25], the five factors have been defined as *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*, often listed under the acronyms *OCEAN*.

Recent soft-biometric approaches have shown the ability to unobtrusive acquire these traits from social media [28, 29], or infer the personality types of users from visual cues extracted from their favorite pictures [7] from a social signal processing perspective [36]. While not necessarily corresponding to the actual traits of an individual, attributed traits are still important because they are predictive of important aspects of social life, including attitude of others and social identity.

As a result, the proposed benchmark includes 1,200 spontaneously uploaded images that hold a lot of meaning for the participants and their related annotations: *positive/negative* (see Table 2 for further details). The images are personal (i.e., family, friends etc.) or just images participants really like/dislike. The motivations for labeling a picture as favorite are multiple and include social and affective aspects like, e.g., positive memories related to the content and bonds with the people that have posted the picture. Moreover, they are provided with a set of TAGS describing the content of each of them.

Finally, many other users' preference information are provided. Table 2 lists the raw data provided with the dataset, such as users' past behavior selected from a pre-defined list (e.g., watches movies, listen songs, read books, travel destinations, etc.), demographic information (like age, nationality, gender, etc.). Note, all data is anonymized (i.e., name, surname, private email, etc.), ensuring the privacy of all participants.

For further analyses related to the adverts' quality, this benchmark also provides the entire set of 300 rated advertisements (500 x 500 pixels) in PNG format.

2.1 Participant Recruitment

The subjects involved in the data collection, performed all the steps of the following protocol:

- *Step 1*: All participants have filled in a form providing, anonymously, several information about their preferences (e.g., demographic information, personal preferences).

- *Step 2*: All participants have filled the Big Five Inventory-10 to measure personality traits [25].

- *Step 3*: The participants voted each advert according with if they would or not click on the recommended ad. Ads have been displayed in the same order to all the participants.

- *Step 4*: The participants submitted some images that they like (Positives) and some others that disgust or repulse them (Negatives). Once they have uploaded their images, they also added some TAGS that describe the content of each image.

2.2 The Subjects

This corpus involves 120 English native speakers between 18 and 68. The median of the participants age is 28 (μ=31.7, σ=12.1). Most of the participants have a university education. In terms of gender, 77 are females and 43 males. The percentage distribution of household income within the sample is: 23% less or equal to 11K USD per year, 48% from 11K to 50K USD, 21% from 50K to 85K USD, and 8% more than 85K USD. The median income is between 11K and 50K USD.

In analyzing this complex data, one can observe that users' preferences are not independent of each other, they are likely to be co-expressed. Hence, it is of great significance to study groups of preferences rather than to perform a single analysis. This fact is

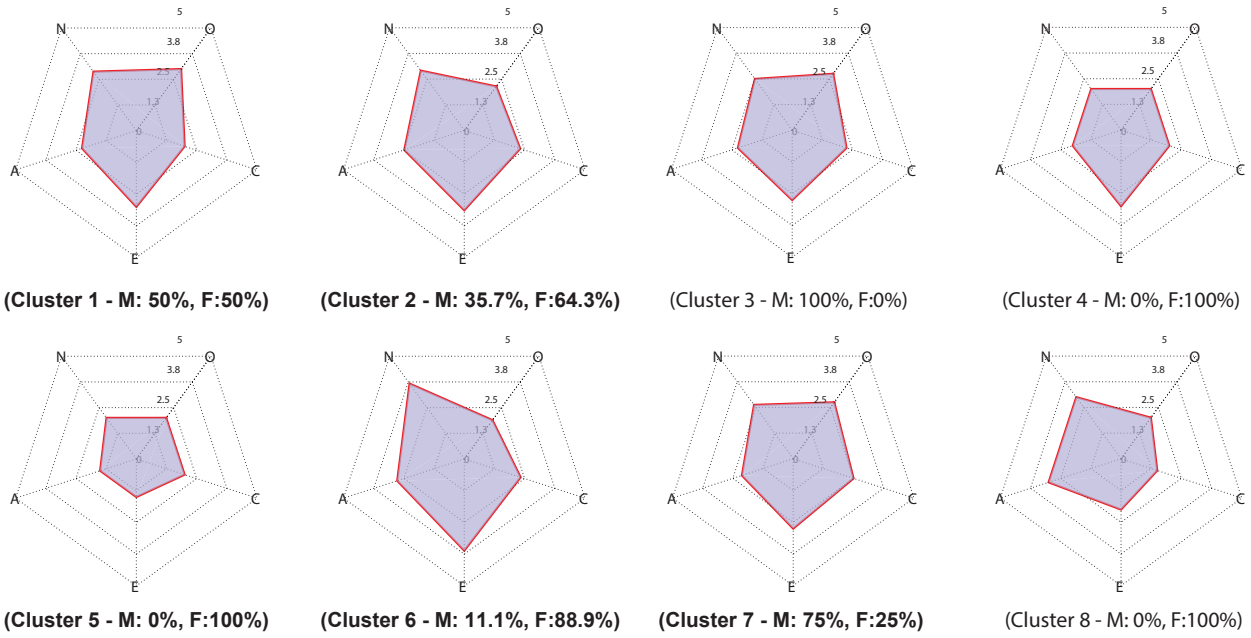

Figure 2: Spider-Diagrams for *O-C-E-A-N* Big-Five traits. The percentage of Males (M) and Female (F) belonging to each cluster is reported. We indicate in bold each instance where a statistacal significant effect (i.e., Pearson correlation at the 5% level) was found between ranks and personality factors.

also true for personality factors, analyzing subsets of data yields crucial information about patterns inside the data. Thus, clustering users' preferences can provide insights into personality of individuals which share the same preferences. We performed a statistical analysis of personality and users' preferences, linking the 5 personality factors and the most favorite users' product categories (i.e., most clicked) by means of the affinity propagation (AP) clustering algorithm [11].

AP is an algorithm that takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars. We calculated a similarity input matrix between each individual u_i considering as feature vectors v_i a binary sequence of click/no-click (i.e., v_i is 1×300). AP exchanges real-valued messages between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. Hence, the number of clusters is automatically detected, and when applied on ADS data, AP grouped the data into 8 different clusters.

Figure 2 illustrates 8 spider-diagrams, one for each cluster. Each diagram shows the average of the big-five factors regarding the subjects within the group (reported in figure as O-C-E-A-N).

Then, we ranked the most clicked categories according with samples within the group in order to compare these two variables by means of correlation obtaining interesting clues.

For instance, let us consider the cluster number 6 where 88.9% of the members are females, and 11.1% males and the average of the group members age is 28. The first 5 most clicked categories are *Baby*, followed by *Consumer Electronics*, *Stationery & Office Supplies*, *Home*, and *Jewelery & Watches*. This group is characterized by high neuroticism (see the diagram in Figure 2.(Cluster 6)), those who score high in neuroticism are often emotionally reactive and vulnerable to stress, high neuroticism causes a reactive and excitable personality, often very dynamic individuals. This group also share the highest levels of extroversion, high extraversion is often perceived as attention-seeking, and domineering.

Cluster 5 shows a subset of individuals which scores low for all the types (see the plot in Figure 2.(Cluster 5)). For instance, those with low openness seek to gain fulfillment through perseverance, and are characterized as pragmatic sometimes even perceived to be dogmatic. Some disagreement remains about how to interpret and contextualize the openness factor. The first 5 most clicked categories are *Clothing & Shoes*, *Health & Beauty*, *Jewelery & Watches*, *Outdoor Living*, and then *Consumer Electronics*. In this case the average of the group members age is 68, and the cluster contains 100% females.

Cluster Id	Avg. Age	r.1	r.2	r.3	r.4	r.5
1	**32**	**6**	**15**	**13**	**19**	**4**
2	**31**	**6**	**5**	**7**	**10**	**17**
3	22	1	7	10	4	6
4	57	6	9	10	2	1
5	**68**	**1**	**4**	**13**	**9**	**6**
6	**28**	**3**	**6**	**15**	**11**	**13**
7	**20**	**7**	**6**	**10**	**11**	**17**
8	52	3	9	19	7	4

Table 3: Top-5 ranked categories. For each cluster the table reports the average age, and the ordered list of the most clicked categories. We indicate in bold each instance where a statistical significant effect (i.e., Pearson correlation at the 5% level) was found between ranks and personality factors.

Cluster 7 is characterized by good levels of conscientiousness that is the tendency to be organized and dependable, aim for achievement, and prefer planned rather than spontaneous behavior. This cluster scores low in agreeableness, which is related to personalities often competitive or challenging people. The openness factor (>2.5) reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has. Interestingly, among the most preferred categories there are *Console & Video Games*, *Consumer Electronics*, *Grocery* and *Computer Software*.

3. EVALUATION METHODOLOGY

Research in the ARS field requires quality measures and evaluation metrics to know the quality of the techniques, methods, and algorithms for predictions and recommendations. In this section we review the process of evaluating an ARS on two main tasks: (i) measuring the accuracy of rating predictions, and (ii) measuring the accuracy of click predictions.

3.1 Scenario 1: Ad Rating Prediction

In most online advertising platforms the allocation of ads is dynamic, tailored to user interests based on their observed feedback. In this first scenario, we want to predict the feedback a user would give to an advert (e.g. 1-star through 5-stars). In such a case, we want to measure the accuracy of the system's predicted ratings. **Root Mean Squared Error (RMSE)** is perhaps the most popular metric used in evaluating the accuracy of predicted ratings. The system generates predicted ratings $\hat{r}_{u,a}$ for a test set T of user-advert pairs (u,a) for which the true ratings $r_{u,a}$ are known. The RMSE between the predicted and actual ratings is given by:

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{(u,a)\in T}(\hat{r}_{u,a} - r_{u,a})^2}. \quad (1)$$

Mean square error (MSE) is an alternative version of RMSE, the main difference between these two estimators is that RMSE penalizes more large errors, and MSE has the same units of measurement as the square of the quantity being estimated, while RMSE has the same units as the quantity being estimated. Therefore, MSE is given by

$$MSE = \frac{1}{|T|}\sum_{(u,a)\in T}(\hat{r}_{u,a} - r_{u,a})^2. \quad (2)$$

Mean Absolute Error (MAE) is a popular alternative, given by

$$MAE = \sqrt{\frac{1}{|T|}\sum_{(u,a)\in T}|\hat{r}_{u,a} - r_{u,a}|}. \quad (3)$$

As the name suggests, the MAE is an average of the absolute errors $err_{u,a} = |\hat{r}_{u,a} - r_{u,a}|$, where $\hat{r}_{u,a}$ is the prediction and $r_{u,a}$ the true value. The MAE is on same scale of data being measured.

3.2 Scenario 2: Ad Click Prediction

In many applications the recommendation system tries to recommend adverts to users in which they may be interested. For example, when items are added to the queue, Amazon suggests a set of adverts on which the user would most probably click. In this case, we are not interested in whether the system properly predicts the ratings of these adverts but rather whether the system properly predicts that the user will click on them (e.g. they perform a conversion). Therefore, we then have four possible outcomes for a recommended advertisement, as shown in Table 4.

	Recommended	Not recommended
Clicked	True-Positive (tp)	False-Negative (fn)
Not clicked	False-Positive (fp)	True-Negative (tn)

Table 4: Classification of the possible result of a recommendation of an advert to a user [22]

We can count the number of examples that fall into each cell in the table and compute the following quantities:

Precision $\frac{tp}{tp+fp}$,

Recall (True Positive Rate) $\frac{tp}{tp+fn}$.

Recall in this context is also referred to as the True Positive Rate (TPR) or *Sensitivity*, and precision is also referred to as positive predictive value (PPV).

Other related measures used include true negative rate and accuracy:

False Positive Rate (1 - Specificity) $\frac{fp}{fp+tn}$,

Accuracy $\frac{tp+tn}{tp+tn+fp+fn}$,

where true negative rate is also called *Specificity*. We can expect a trade-off between these quantities; while allowing longer recommendation lists typically improves recall, it is also likely to reduce the precision. We can compute curves comparing precision to recall, or true positive rate to false positive rate. Curves of the former type are known simply as precision-recall curves, while those of the latter type are known as a Receiver Operating Characteristic or ROC curves. A widely used measurement that summarizes the ROC curve is the **Area Under the ROC Curve (AUC)** [1] which is useful for comparing algorithms independently of application.

When evaluating precision-recall (or ROC curves) for multiple test users, a number of strategies can be employed in aggregating the results, depending on the application at hand. The usual manner in which precision-recall curves are computed in the information retrieval community [13, 27, 31, 32] is to average the resulting curves over users. Such a curve can be used to understand the trade-off between precision and recall (or false positives and false negatives) a typical user would face.

4. EXPERIMENTS AND RESULTS

In this section we show results obtained for the two types of scenarios introduced in Sec. 3. We conduct prediction experiments to explore the strengths and weakness of using personality traits as features for recommendation.

4.1 Evaluated Algorithms

Since a prediction engine lies at the basis of the most recommender systems, we selected some of the most widely used techniques for recommendations and predictions [14], such as Logistic Regression (LR) [8], Support Vector Regression with radial basis function (SVR-rbf) [3], and L2-regularized L2-loss Support Vector Regression (L2-SVR) [8]. These methods have often been based on a set of sparse binary features converted from the original categorical features via one-hot encoding [17, 26]. These engines may predict user opinions to adverts (e.g., a user's positive or negative feedback to an ad) or the probability that a user clicks or performs a conversion (e.g., an in-store purchase) when they see an ad. In Section 4, we evaluate these methods while feeding them with and without features coming from the psychometric traits.

4.2 Experimental Protocol

Let us say $X = \{\bar{x}_1, ..., \bar{x}_N\}$ is the set of observations, where the vectors $\bar{x}_i$ correspond to features coming only from the group "users' preferences" as described in Table 2 and $N = 120$ stands for the number of users involved in the experiment.

A feature is the user's selection from a pre-defined list of choices, hence, for each feature vector one element is 1 and the others are 0. Then, each column vector $\bar{x}_i$ is obtained by stacking the features on top of one another.

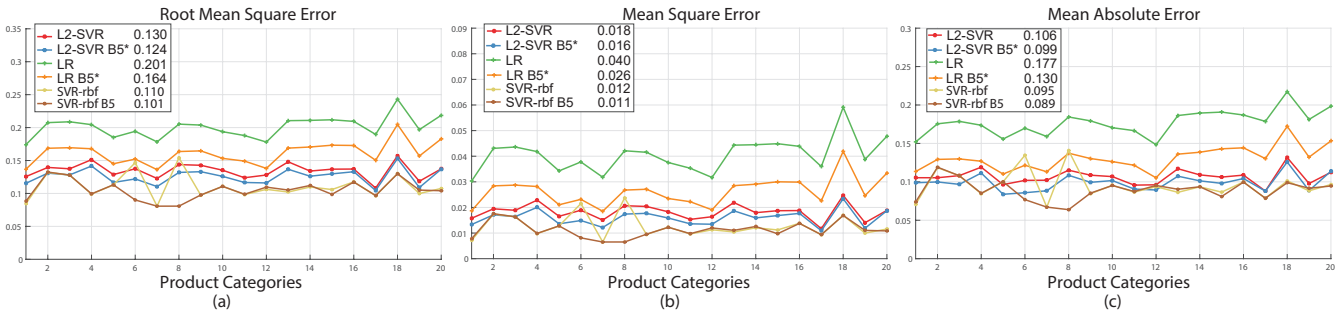

Figure 3: Measuring ratings prediction accuracy: B5 stands for Big-Five features. We indicate with an asterisk each method where B5 features, embedded into a baseline learner, shows a statistical significant effect over the baseline.

Regression is performed over the 20 product categories. The prediction problem is solved using LR, L2-SVR, and SVR-rbf, while feeding them with and without features coming from "personality traits". All experiments were performed using a k-fold approach (k = 10). In k-fold cross-validation, X is randomly partitioned into k's equal sized subsamples (the folds are the maintained the same for each algorithm in comparison). Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k - 1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used only once as the validation data. The k results from the folds can then be averaged to produce a single estimation.

Our experimental protocol includes feature selection, which represents an important pre-processing step given the sparse nature of the input data. It allows to remove many redundant features by reducing the dimensionality of the problem at hand. Hence, the representation above serves as a basis for the feature ranking and selection strategy. Ranking features allow us to detect a subset of cues which is not redundant. Accordingly, we use the training data obtained after the split as input of the infinite feature selection (Inf-FS) [30] algorithm. By construction, the Inf-FS is a graph-based method which exploits the convergence properties of the power series of matrices to evaluate the relevance of a feature with respect to all the other ones taken together. Indeed, in the Inf-FS formulation, each feature is mapped on an affinity graph, where nodes represent features, and weighted edges the relationships between them. In particular, the graph is weighted according to a function which takes into account both correlations and standard deviations between feature distributions. Each path of a certain length l over the graph is seen as a possible selection of features. Therefore, varying these paths and letting them tend to an infinite number permits the investigation of the importance of each possible subset of features.

Finally, the Inf-FS assigns a score to each feature of the initial set; where the score is related to how much the given feature is a good candidate regarding the regression task. Therefore, ranking the outcome of the Inf-FS in descendant order allows us to perform the subset feature selection throughout a *model selection stage*. In this way, we reduce the amount of features, by selecting 75% of the total. The selected features are: the number of favorite websites, T.V. programmes, sports, past times, the most watched movies and most visited websites, where we add the big-five personality traits.

4.3 Exp. 1: Ad Rating Prediction

In this section we report results for rating prediction showing that traces of user's personality can improve the prediction performance of the evaluated methods significantly. Statistical evaluation of experimental results has been considered an essential part of validation of machine learning methods. Given the user u_i, labels are assigned to each category by averaging the votes they gave to the category items such as $u_i = \{y_1, ..., y_{20}\}, y \in [1 - 5]$.

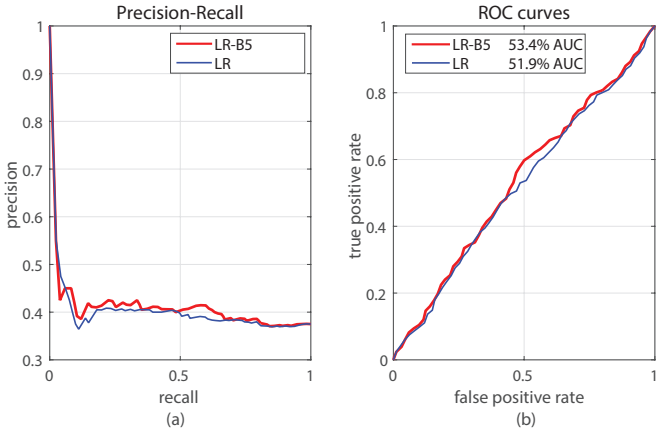

Figure 4: Comparison between LR and LR-B5: Curves show the proportion of preferred items that are actually recommended.

Figure 3 illustrates prediction results in term of RMSE, MSE and MAE plots across the categories. This first analysis shows how personality traits affect prediction performance. In order to assess if the difference in performance is statistically significant, t-tests have been used for comparing the accuracies. This statistical test is used to determine if the accuracies obtained with and without B5 are significantly different from each other (whereas both the distribution of values were normal). The test for assessing whether the data come from normal distributions with unknown, but equal, variances is the *Lilliefors* test.

Results show a statistical significant effect of personality traits while using L2-SVR (p-value < 0.05, Lilliefors Test H=0) and LR (p-value < 0.01, Lilliefors Test H=0).

As for the SVR-rbf, even if improvements in terms of prediction are not significant (B5 against no-B5), it is still interesting to notice the performance loss on categories 6 and 8, where errors go high significantly. In such a case, the B5 features do not seem to have much predictive power, however, they seem to play the role of a reliable stabilizer, but also that of an independent mediator and supporter of the regression process.

4.4 Exp. 2: Ad Click Prediction

This section shows an offline evaluation of click prediction. Along the lines of the previous experiment, a k-fold cross-validation is used. The experiment is performed at the category level, in order to

Method	ROC-AUC	Precision	Recall
L2-SVR	50.5%	39.2%	50.2%
L2-SVR B5	51.4%	39.9%	50.9%
LR	51.9%	40.3%	51.3%
LR-B5*	**53.4%**	**41.2%**	**52.1%**
SVR-rbf	48.3%	36.5%	48.8%
SVR-rb B5	50.1%	38.2%	50.2%

Table 5: Performance for ad click prediction. Big-Five features systematically contribute to the overall performance. The asterisk indicate that the method overcomes all the others.

work on a balanced distribution over the classes (1,229 clicked vs 1,171 not clicked instances), whenever a user showed their interest in a given category (i.e., the category contains at least one clicked advert) we labeled the category as "clicked" (rating greater or equal to four), otherwise "not clicked" (rating less than four). As a result, for each user we obtained a list of 20 labels representing their preference to each category. We computed precision-recall and ROC curves for each user, and then averaged the resulting curves over users. This is the usual manner in which precision-recall (or ROC) curves are computed in the information retrieval community [13, 31, 32]. Such a curve can be used to understand the trade-off between precision and recall and ROC a typical user would face.

Figure 4.(a) reports the precision-recall curves which emphasize the proportion of recommended items that are preferred and recommended. Figure 4.(b) shows the global ROC curves for LR and LR-B5, which emphasize the proportion of adverts that are not clicked but end up being recommended. The LR-B5 curve completely dominates the other curve, the decision about the superior setting for LR is easy.

The Area Under the ROC Curve is calculated as a measure of accuracy, which summarizes the precision recall of ROC curves, we report AUC, precision and recall in terms of the harmonic mean of precision and recall (F-measure) for all the methods in Table 5.

4.5 Discussions and Future Work

In this paper, we conducted a within-subject user study to investigate on the relations between users' personality related to their buying behavior and preferred item categories. A deeper analysis may involve the use of bi-clustering methods. Comparing to traditional clustering methods biclustering is not a blackbox technique. Comprehensibility is one of its main advantages, i.e. it is possible to understand why objects ended up in the same cluster.

It is worth noting that the goal of these experiments is to show how personality traits affect the prediction. In order to improve prediction accuracy, specific feature designing processes are needed so as to represent personality data and to standardize their definitions to be used as input recommender data towards to improve recommendations. In our experiments, we used a set of sparse binary features converted from the original categorical features. Moreover, many other algorithms may be used for this tasks, like the one proposed in [6, 15, 24].

For instance, the personality diagnosis [24] system is a collaborative filtering algorithm, which can be thought of as a hybrid between existing memory- and model-based algorithms. PD is fairly straightforward, maintains all data, and does not require a compilation step to incorporate new data. It is based on a simple and reasonable probabilistic model of how people rate titles.

Most of these recommender systems use to split each test user profile into sets of observed items and hidden items. The former is used as input for each recommender, the latter for performance evaluation. In our experiments, we did not use any information about the previous users' clicks, which turns out to be a more difficult task. We decided on this solution to move the focus of attention on personality data and not on other features like previous clicked ads.

5. CONCLUDING REMARKS

In this paper, we presented the ADS Dataset, a collection of 300 real advertisements rated by 120 unacquainted participants. We conducted a within-subject user study to investigate potential user issues of the personality on their buying behavior and preferred item categories.

The corpus has been collected with the main goal of studying the possible achievable benefits of employing personality traits in modern recommender systems. To obtain stronger and more relevant results for this community, appropriate and high-level features needed to be designed that carry important information for inference. In this paper, we only use raw data as sparse binary features converted from the original categorical features. We used standard techniques for recommending ads in order to show how personality traits affect the prediction, and, at the same time, set a baseline for future work.

We then reviewed a large set of properties, and explain how to evaluate systems given relevant properties. We discuss how to compare ARS based on a set of properties that are relevant for the application. Therefore, we review two main types of experiments in an offline setting, where recommendation approaches are compared with different selections of features (i.e., with and without personality traits) accordingly with our goal. We also discuss how to draw trustworthy conclusions from the conducted experiments.

Future work includes, but is not necessarily limited to, (1) feature engineering and designing for ARSs, represent personality data and standardize their definitions to be used as input recommender data towards to improve recommendations; (2) inference of personality traits and novel approaches for mapping pictures tagged as favorite into personality traits; and (3) identification of the underlying dimensions of consumer shopping motivations and personality factors.

We hope that this work motivates researchers to take into account the use of personality factors as an integral part of their future work, since there is a high potential that incorporating these kind of users' characteristics into ARS could enhance recommendation quality and user experience.

6. REFERENCES

[1] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 1975.

[2] M. Bosnjak, M. Galesic, and T. Tuten. Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach. *Journal of Business Research*, 2007.

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2011.

[4] J. V. Chen, B. chiuan Su, and A. E. Widjaja. Facebook c2c social commerce: A study of online impulse buying. *Decision Support Systems*, 2016.

[5] K. Choi, D. Yoo, G. Kim, and Y. Suh. A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. *Electronic Commerce Research and Applications*, 2012.

[6] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2003.

[7] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 213–222. ACM, 2013.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

[9] B. M. Fennis and A. T. Pruyn. You are what you wear: Brand personality influences on consumer impression formation. *Journal of Business Research*, 2007.

[10] Forrester. Online retail industry in the us will be worth $279 billion in 2015. *TechCrunch*, February 28.

[11] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[12] D. Funder. Personality. *Annual Reviews of Psychology*, 52:197–221, 2001.

[13] D. Harman. Overview of the trec 2002 novelty track. In *Text REtrieval Conference (TREC 2002)*, 2002.

[14] X. He. et al. practical lessons from predicting clicks on ads at facebook. In *Data Mining for Online Advertising*, New York, NY, USA, 2014. ACM.

[15] R. Hu and P. Pu. A Study on User Perception of Personality-Based Recommender Systems. In a. u. l. De, Bra, A. Kobsa, and D. Chin, editors, *User Modeling, Adaptation, and Personalization*, volume 6075 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010.

[16] C. Kim, K. Kwon, and W. Chang. How to measure the effectiveness of online advertising in online marketplaces. *Expert Syst. Appl.*, 2011.

[17] K.-c. Lee, B. Orten, A. Dasdan, and W. Li. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2012.

[18] S. K. Lee, Y. H. Cho, and S. H. Kim. Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. *Information Sciences*, 2010.

[19] J. C. Mowen. *The 3M Model of Motivation and Personality: Theory and Empirical Applications to Consumer Behavior*. Springer US, Boston, MA, 2000.

[20] N. nez Valdéz. et al. implicit feedback techniques on recommender systems applied to electronic books. *Comput. Hum. Behav.*, 2012.

[21] A. Odic, M. Tkalčič, A. Košir, and J. F. Tasič. A.: Relevant context in a movie recommender system: UsersâĂŹ opinion vs. statistical detection. In *In: Proc. of the 4th Workshop on Context-Aware Recommender Systems (2011*.

[22] D. L. Olson and D. Delen. *Advanced Data Mining Techniques*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[23] A. C. P. and V. M. S. Click through rate prediction for display advertisement. *International Journal of Computer Applications*, 2016.

[24] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 473–480, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[25] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 2007.

[26] M. Richardson. Predicting clicks: Estimating the click-through rate for new ads. In *International World Wide Web Conference*. ACM Press, 2007.

[27] G. Roffo, M. Cristani, L. Bazzani, H. Q. Minh, and V. Murino. Trusting skype: Learning the way people chat for fast user recognition and verification. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 748–754, Dec 2013.

[28] G. Roffo, C. Giorgetta, R. Ferrario, and M. Cristani. *Just the Way You Chat: Linking Personality, Style and Recognizability in Chats*, pages 30–41. Springer International Publishing, 2014.

[29] G. Roffo, C. Giorgetta, R. Ferrario, W. Riviera, and M. Cristani. Statistical analysis of personality and identity in chats using a keylogging platform. In *International Conference on Multimodal Interaction*. ACM, 2014.

[30] G. Roffo, S. Melzi, and M. Cristani. Infinite feature selection. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *ACM Conference on Electronic Commerce*. ACM, 2000.

[32] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

[33] M. Tkalčič, U. Burnik, and A. Košir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, 2010.

[34] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic. Affective labeling in a content-based recommender system for images. *IEEE Transactions on Multimedia*, 15(2):391–400, Feb 2013.

[35] C. A. Turkyilmaz, S. Erdem, and A. Uslu. The effects of personality traits and website quality on online impulse buying. 2015. International Conference on Strategic Innovative Marketing.

[36] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.

[37] X. Wang, W. Li, Y. Cui, R. Zhang, and J. Mao. Click-through rate estimation for rare events in online advertising. *Online Multimedia Advertising: Techniques and Technologies*, 2010.