In *CARS* We Trust: How Context-Aware Recommendations Affect Customers' Trust And Other Business Performance Measures Of Recommender Systems

Michele Gorgoglione Polytechnic of Bari, Italy m.gorgoglione@poliba.it	Umberto Panniello Polytechnic of Bari, Italy u.panniello@poliba.it	Alexander Tuzhilin Stern School, NYU atuzhili@stern.nyu.edu

ABSTRACT

Most of the work on Context-Aware Recommender Systems (CARSes) has focused on demonstrating that the contextual information leads to more *accurate* recommendations and on developing efficient recommendation algorithms utilizing this additional contextual information. Little work has been done, however, on studying how much the contextual information affects purchasing behavior and trust of customers. In this paper, we study how including context in recommendations affects customers' trust, sales and other crucial business-related performance measures. To do this, we performed a live controlled experiment with real customers of a commercial European online publisher. We delivered content-based recommendations and context-aware recommendations to two groups of customers and to a control group. We measured the recommendations' accuracy and diversification, how much customers spent purchasing products during the experiment, quantity and price of their purchases and the customers' level of trust. We aim at demonstrating that accuracy and diversification have only limited direct effect on customers' purchasing behavior, but they affect trust which drives the customer purchasing behavior. We also want to prove that CARSes can increase both recommendations' accuracy and diversification compared to other recommendation engines. This means that including contextual information in recommendations not only increases accuracy, as was demonstrated in previous studies, but it is crucial for improving trust which, in turn, can affect other business-related performance measures, such as company's sales.

1. INTRODUCTION

The use of recommender systems (RSes) in the industry has exploded over the last several years to the effect that most of the major companies either have RSes in place or have recently launched major RS initiatives.

One of the important factors affecting performance of RSes is the *contextual* information (Adomavicius and Tuzhilin, 2011). For example, when a customer is recommended a vacation package, knowledge of the traveling season constitutes an important factor. Several music streaming companies, such as LastFM and Musicovery, ask customers to specify their mood before recommending particular music by providing the users with a standard menu of well-established types of moods, such as "positive", "energetic", "calm" and "dark", and asking them to select the one(s) that suit(s) them the most at the moment. Then this mood information is used by the system to recommend only the type of music that best fits the customer's mood, mood being a contextual variable in this case. As another example, Netflix knows the location of the customers and uses the locational contextual variables, such as city and/or zip code, and time to provide context-specific recommendations of movies. Similarly, mobile recommender systems (e.g., those deployed on the Smart phones) provide more relevant recommendations to its customers when they take into account such important contextual information as the GPS-based location and time. As Reed Hastings, the CEO of Netflix, pointed out recently, Netflix can improve the performance of its RS up to 3% when taking into account such contextual information as the time of the day or location in their recommendation algorithms[1]. This observation by Hastings was echoed by industry panelists at the industrial panel held during the CARS workshop at RecSys'12 where managers from LinkedIn, Netflix, EchoNest and Telefonica reiterated the importance of contextual information and described how their recommendation engines utilize contextual information in their businesses (http://cars-workshop.org/program).

Most of the existing work on context-aware recommender systems (CARS) has focused on the accuracy metrics for measuring performance of RSes and demonstrated that knowledge of certain

[1] Watch his interview at 44:40 min at www.youtube.com/watch?v=8FJ5DBLSFe4&feature=youtu.be.

relevant types of contextual information leads to more accurate estimations of unknown ratings. Although relevant, this work constitutes only the first step towards demonstrating usefulness of contextual information because the business community is mainly interested in the economics-related performance metrics, such as changes in sale volumes, profits, and prices of purchased products, as opposed to the accurate estimations of unknown ratings.

In this paper, we focus on *business performance* metrics, such as sales volumes and money spent by customers, and examine how contextual information affects these metrics. More specifically, we examine how knowledge of contextual information leads to better business performance metrics including changes in purchasing behavior of the customers receiving recommendations. We study this problem empirically by conducting a live controlled experiment with real customers in a real industrial setting (i.e., by doing the, so called, A/B testing).

The contributions of this paper can be summarized as follows. First, we study how contextual information affects such recommendation performance measures as accuracy and diversification, and how, together, they (accuracy and diversification) affect customer trust. Second, we study the way trust affects such business performance measures, namely quantities of goods purchased and the total money spent by customers. Third, we want to demonstrate that, by transitivity, using contextual information in recommender systems can improve purchases (measured by quantity of items purchased and by the customers expenditure) because the contextual information improves accuracy and diversification, which in turn affects trust, which, finally, affects purchases. We investigate all these effects by conducting an empirical live study on the customers of a major comics book publisher in Europe. These contributions are important because *all* the prior work in the CARS field focused *exclusively* on the accuracy-based performance metrics in the empirical analysis of the offline historical data, whereas we performed “live” online experiments in this study with real customers and measured business-oriented performance metrics. Such types of experiments are done infrequently in the recommender systems community (i.e., the vast majority of empirical work still constitutes analysis of the offline historic data and the use of the accuracy-based predictive performance metrics).

It is important to point out that these findings do not necessarily prove a causal relationship between context and the business performance measures used in the research. Rather this research

demonstrates the existence of a *predictive relationship* in the sense explained by (Shmueli and Koppius, 2011). Although we cannot claim causality , we try to demonstrate the existence of correlation using extensive statistical analysis methods, which is important and novel since this is the first work establishing the relationship between context-aware recommendations and customers' purchasing behavior. In particular, we extensively tested the characteristics of our samples and data sets and applied different statistical models to our hypotheses in order to exclude the presence of biases and unobserved factors which may decrease validity of our results.

The rest of the paper is organized as follows. In Section 2, we present the prior work done on context-aware recommender systems. The methodology used during our experiment and the analysis of the results are presented in Section 3. In Section 4 we present the relation between recommendations performance and business performance. We show how context affects all the aforementioned measures in Section 5. Section 6 contains the main conclusions of this work.

2. PRIOR WORK

The effect of recommendations on the purchasing behavior of customers has been extensively studied in the case of *traditional*, non-contextual RSes. In contrast, very little research has been done on studying this effect for the *context-aware* recommender systems. In this section we first provide a review of the relevant literature in the area of non-contextual RSes. We found that research in this area has focused on examining the relationships between accuracy, diversification, trustworthiness of recommendations and levels of sales and other business related indicators. Secondly, we review similar research in the area of context-aware recommendations and show the research gap that this paper addresses.

Several scholars have studied the effect of recommendations for traditional, un-contextual RSes. Schafer et al. (2001) argued that RSes help increase sales by converting browsers into buyers, increasing cross-selling opportunities, and building customer loyalty. However, the accuracy of recommendations alone is not sufficient to explain the purchasing behavior. Trust plays a key role. Pathak et al. (2010) found that the strength of recommendations has a positive impact on sales. However, recommendations influence shoppers' decisions only when they are perceived to be objective and credible. Since retailers have full control of recommendations, it is natural for shoppers

to discount credibility of online RSes because of potential manipulation by retailers. This perception is further supported by anecdotal evidence of retailers manipulating the outcome of RSes (Flynn, 2006; Mui, 2006). Pu et al. (2011) found that, while overall satisfaction with the recommender system defined in terms of ease of use and perceived usefulness is important for usage intentions, trust in the system and choice confidence are crucial for purchasing intentions.

The effect of accuracy of recommendations on trust has also been studied. Zhang et al. (2011) showed that the customer loyalty to online stores can be increased by improving recommendations' accuracy but is not sufficient alone. Relevance, accuracy, completeness, and timeliness of recommendations have a significant effect on users' decision making and satisfaction (Bharati and Chaudhury, 2004). High accuracy of recommended items contributes to the increase of user involvement, which in turn increases user satisfaction (Hess et al., 2005). Familiar recommendations play an important role in establishing user trust in a RS (Sinha and Swearingen, 2001; Swearingen and Sinha, 2001; Swearingen and Sinha, 2002). However, Komiak and Benbasat (2006) demonstrated that the user's familiarity with the recommendations increased trust in recommender's benevolence and integrity, but not trust in its competency. Xiao and Benbasat (2007) showed that the way familiar and unfamiliar items are balanced in a recommendation list influences users' trust in perceived usefulness of, and satisfaction with RSes.

The accuracy of the predictions provided by a RS is only one of the possible variables affecting trustworthiness (Lenzini et al., 2009). Several studies have demonstrated that diversity, diversification, variety and novelty of recommendations can have an important role. Most researchers agree that consumers generally prefer more variety when given a choice (Baumol and Ide, 1956; Kahn and Lehmann, 1991). Fleder and Hosanagar (2009) demonstrated that RSes that discount item popularity in the selection of recommendable items may increase sales more than RSes that do not. Similarly, Brynjolfsson et al. (2003) showed that increased product variety made available through electronic markets can be a significantly larger source of consumer surplus gains. The concepts of novelty and diversity are often discussed as a joint function because it is important in many applications to recommend a wide range of items that customers have not seen before (Vargas and Castells, 2011; Zhang and Hurley, 2009). Following this work, in this paper we consider both

diversity and novelty as aspects of a more general concept that we call *diversification*, as explained in Section 3.3. Mcginty and Smyth (2003) found that diversity can provide significant gains if carefully tuned. Simonson (2005) showed that higher variety seeking decreases receptivity to customized offers. Ziegler et al. (2005) showed that users' overall satisfaction with recommendation lists goes beyond accuracy and involves other factors, e.g., the diversification of the result set. Similar results were found by other studies (Bollen et al., 2010; Smyth and McClave 2001; Pu and Chen 2006). Hu and Pu (2011) showed that perceived diversity significantly influences users' perceived ease of use and usefulness of the RS, positive attitudes toward the system and behavioral intentions.

Some researchers have also investigated the combined effect of accuracy and diversity. Additional recommendations of familiar products serve as a context within which unfamiliar recommendations are evaluated (Cooke et al., 2002). Liang et al. (2006) demonstrated that both the number of recommended items and the recommendation accuracy had significant effects on user satisfaction. Mcginty and Smyth (2003) highlighted the pitfalls of naively incorporating diversity-enhancing techniques into existing RSes. They pointed out that diversity should be provided adaptively rather than being enhanced in each and every recommender cycle. Knijnenburg et al. (2012) found that users may not perceive diversified recommendation sets as more diverse, but they perceive them as more accurate. Situational (context) and personal characteristics (such as trust, domain knowledge and perceived control) can mediate this perception. An important contribution was made by Adomavicius and Kwon (2012) who demonstrated the existence of a trade-off between accuracy and diversity. Ranking recommendations according to the predicted rating values provides good predictive accuracy but it tends to perform poorly with respect to recommendation diversity.

All this prior work focuses on examining relationships between accuracy, diversification, trustworthiness of recommendations and increased levels of sales and other business related indicators for the *traditional* recommender systems. Much research has been done on CARS, and Adomavicius and Tuzhilin (2011) provide a broad overview of this area. However, no research has been done on studying these effects for the *context-aware* recommender systems, especially in the context of conducting controlled live experiments in real industrial settings. Most of the research has focused on the relationship between context and accuracy. For instance, Adomavicius et al. (2005) showed that

contextual information can increase recommendation accuracy if deployed properly. Further, Panniello et al. (2009) compared several alternative context-aware methods and demonstrated that context can increase recommendation accuracy. Very little research has included other variables, such as diversity and trust, in the analysis. For instance, Panniello et al. (2012) has compared accuracy and diversity of context-aware RSes without studying the effect on customer behavior. Gorgoglione et al. (2011) studied the combined effect of accuracy and diversity on customers' trust and behavior in a live controlled experiment. They found that customers receiving context-aware recommendations spent more money compared to customers receiving different kinds of recommendations. They also compared accuracy and diversity of different recommendation engines. The authors did not propose any analytical model to explain these results and stated the need for further research.

Based on this literature review, we decided to address our research problem by analyzing the relationships among accuracy, diversification and trust of recommendations on customers' purchases and business performance for CARS. The methodology that we followed to conduct a live experiment is described in the next section.

3. METHODOLOGY

In order to study how context-aware recommendations affect customers' trust and business-related performance, we conducted a live experiment in partnership with a well-known global European publishing firm. The company's Web division mainly sells comic books and related products, such as DVDs, stickers, and T-shirts. As a part of its normal business, the company sends a weekly non-personalized newsletter to 24,364 customers. The firm agreed to send personalized recommendations of comic books via e-mail (in addition to the traditional weekly newsletter) to a sample of 360 customers as a part of our project. The 360 customers were randomly selected from the customer database. According to the privacy laws, the firm asked customers to state explicitly if they wanted to join this project to improve the customer service. The activity was presented as collaboration between the company and a university aimed at improving the customer service. The experiment participants were then randomized into three experimental treatment groups. We performed statistical tests among the treatment groups and between them and the whole population, and we found no statistically significant differences (see Section 4.2). Therefore we can conclude that the sample selection was

unbiased. Although 360 is not a large number of users, it turned out to be sufficient to demonstrate statistically significant results in our experiments, as is shown in the paper.

In the following subsections we describe the experimental design, the recommendation engines and the performance measures.

3.1. Experimental design

We followed the standard experimental design approach and split the study participants into three experimental treatment conditions. Each group received a personalized newsletter generated via a different recommendation engine: 90 users received content-based recommendations, 90 users received random recommendations and 180 users received context-aware recommendations. We decided to double the number of users included in the contextual group since a CARS needs more ratings than a traditional one to work correctly. This does not bias the results because we averaged each performance metric across customers instead of using the absolute values. Therefore, the metrics used in the study are not affected by the number of participants in each group. We measured accuracy and diversification of recommendations, the customers' trust in these recommendations and business-related metrics, namely quantity of goods purchased, average price and money spent by customers, as described further in Section 3.3 (see Table 1 for a summary of metrics). In order to make a meaningful comparison, the firm gave us access to the data pertaining to the purchasing behavior of the customers involved in the experiment in a period of twenty months before the experiment began. By comparing the data before and after the experiment, we could observe the effect of recommendations on customers' purchasing behavior and, in turn, on business performance.

We proceeded in two steps. In the first step we studied how accuracy and diversification affect customers' trust and business performance of a RS application through several Structural Equation Models (SEM) (Bollen, 1989). In the second step we studied how using contextual information can affect accuracy and diversification of recommendations by comparing different types of RSes. Finally, by combining the results of these two steps, we can logically conclude that CARS should improve both accuracy and diversification which in turn improve trust, thus, improving business performance of CARS as compared to other types of RSes. In this study, we empirically validated this conclusion and showed that this is, indeed, the case.

We did this analysis in two steps (vis-à-vis a single step) because of various biases that would adversely affect the discussion of the results. We chose a between-subjects design of the experiment because we need to measure certain aspects of the user experience (trust and purchasing behavior). In this case the experiment has to be as close to a real-world usage situation as possible and it is imperative to avoid spill-over effects and other biases (Knijnenburg, 2012). This choice entails defining different treatments, namely a group of customers receiving context-aware recommendations, while another receiving context unaware recommendations. A different choice, such as delivering context-aware and context unaware recommendations to the same group of customers, would make it possible a one-step analysis, but also increase the probability of a bias. In this case, it would be hard to separate the effects of the CARS from that of the context unaware RS because the two types of recommendation would affect each other. This is caused by "delayed response" effects (Lilien and Rangaswamy, 2003), e.g., customers react sometime after receiving the recommendation or "interference" effect (Malhotra et al., 2012), e.g., customers are influenced by both kinds of recommendations.

The detailed description of the recommender engines used in each treatment is provided in Section 3.2. The random group was used as a control group. Each subject received a weekly newsletter recommending ten comic books and was asked to provide a feedback after each newsletter. The experiment ran for nine consecutive weeks.

Before starting the experiment we asked the participants to rate a representative set of twelve comic books selected by the company. This set of comic books was representative of the whole item database, and it was the same for each user. This initial step was needed in order to build the initial user profiles and avoiding the "cold start" problem (Schein et al., 2002), given that building a pre-experiment user profile was possible only for less than 5% of users, those who had purchased more than one item per year.

After that, each subject received a personalized weekly newsletter displaying 10 recommended comic books for 9 consecutive weeks. The newsletter contained a link to a personal recommendation page displaying the ten recommended items. Five items were "recommended brand new items" selected from brand new arrivals at the firm (about 30 brand new published comic books per week),

and the remaining five were "recommended old items" selected from the arrivals in the past two months (about 250 items). As explained in Section 5.1, this does not introduce issues related to pre-imposed diversity and therefore does not bias the results. Each item was presented with the following information: title, cover image, description, a "see more details" link. The customers were invited to rate each recommended product by clicking on a (0-5) point scale. Although the users might not have read a recommended book when asked to give a rating, the information provided about the books was sufficient for the customers for making a good assessment because of the special nature of the comic books industry. Comic books usually come in series with the main characters from the series being familiar to the comic books fans. Therefore, if a customer has not read a particular recommended book, he/she can always click on the "see more details" link, read general description of the new book and then use the prior knowledge about the whole series and its main characters to form an informed opinion about how much interest he/she has in that book.

These solicited ratings were subsequently used to update the user's profile for each user (except for the random treatment, see Section 3.2). All the aforementioned settings were applied to all three treatment groups. The average response rate (i.e., users who gave feedback during the experiment) was about 65% for each treatment condition.

3.2. Types of recommender systems used in the study

During the experiment we used three different RSes: a content-based, a context-aware and a random one. We have chosen a content-based recommendation algorithm, rather than a collaborative filtering (CF) method, because it would have been difficult to generate meaningful recommendations using the CF approach since the experiment was carried out with a few participants and the user/item matrix was relatively sparse – the two conditions adversely affecting CF results.

Content-based. The content based algorithms (Pazzani and Billsus, 2007) uses characteristics of previously purchased or rated items in order to build customers' preference profiles and use these profiles to provide appropriate recommendations. For example, if user "Joe" has mainly purchased comic books containing romantic stories, the system profiles the customer as a "romantic stories" reader and provides appropriate recommendations to him. More formally, content-based systems estimate an unknown rating $u(i,s)$ of item s for user i based on the ratings $u(i,s_j)$ assigned by user i to

items $s_j \in S$ that are similar to item s (Adomavicius and Tuzhilin, 2005). In particular, let *ItemProfile(s)* for item s and *UserProfile(i)* for user i, be two vectors representing the item characteristics and the customer preference, respectively. *ItemProfile(s)* is computed by extracting a set of keywords taken from the content of s (e.g. a comic book description). *UserProfile(i)* is computed by analyzing the content of the items previously seen and rated by user i. It is defined as a vector of weights $(w_{i1}, \ldots, w_{iz})$, where each w_{ij} denotes the importance of keyword j to user i. We computed w_{ij} as an "average" of the ratings provided by user i to those items that contained the keyword $j \in Z$. In our study, we assumed that $z = 80$, thus restricting the keyword profile lengths to 80 words. Candidate items are compared with user profile and the most relevant items are recommended, where the relevance $u(i,s)$ of item s to user i is determined as the average weights of the words in common between *UserProfile(i)* and *ItemProfile(s)*. The top 10 items with the highest $u(i,s)$ are recommended to the user in the newsletter. Since we adopt a content-based engine which uses item features, we checked that each item had the same amount of information (i.e., title, sub-title and description) in order to avoid introduction of biases.

Context-Aware. The CARS developed for our experiment used the *same* content-based algorithm discussed in the previous section in order to compare the two methods on the same basis ("apples with apples"). The only difference is that we used the contextual profile *UserProfile(i,k)* of user i in context k (e.g., a gift for a parent in Fig. 1(a)) instead of the general non-contextual profile *UserProfile(i)*. We computed profile *UserProfile(i,k)* by following the pre-filtering approach (Adomavicius and Tuzhilin, 2011; Panniello et al., 2009) by analyzing the content of the items previously seen and rated by user i in context k. In particular, the contextual information k is used as a label for filtering out those items that were not rated in this context k, i.e, this method selects from the initial set of all the ratings *only* those referring to context k. As a result, *UserProfile(i,k)* contains only the data pertaining to context k. After that, the content-based algorithm is launched on *only* this selected data to produce recommendations specific to context k. Therefore, a different item can be recommended when using the contextual user profile, *UserProfile(i,k)* vs. the case when the un-contextual user profile *UserProfile(i)* is used. For example, a Spider-man comic book can be

suggested to the user “Joe” as a potential gift for a friend, while that comic book would not be recommended to the user “Joe” when he is looking for his personal collection.

In this work, we follow the representational approach to defining contextual information (Dourish, 2004). In particular, according to (Adomavicius and Tuzhilin, 2011; Panniello et al., 2009; Kwon and Kim, 2009) we define context by two *contextual attributes (variables)*: the “intent of a purchase” made by a customer and the “customer’s mood” (Figure 1). The contextual attribute “intent of purchase” distinguishes whether the user is looking for recommendations for his/her personal interest (further distinguished between recommendations for his/her collections, special issues or occasional reading) or for a gift (further distinguished between recommendations for a gift to a partner, a friend, etc.). The attribute “customer’s mood” assumes that the customer may be looking for different recommendations depending on his/her type of the mood which can be dark, energetic, positive or calm in our study. We set “intent of purchase” and “mood” as contextual variables in our study after setting up focus groups, conducting several interviews with the readers, and discussing the produced results with the company management. In particular, most of the interviewees told us that they modify their behavior depending on the intent of purchase and that their choice of reading a certain comic book is related to the emotional content and may depend on their mood. In addition, we have found similarity between these findings and several web sites settings, such as the “wish list” and the “gift options” of certain e-malls, and the “mood menu” of several music vendors and providers (e.g., see the LastFM and Musicovery examples in Section 1 of the paper). All this supports our choice of contextual variables “intent of purchase” and “mood” in the experiments and their importance in other RSes.

We also used other recommendation applications, such as music recommendations, as reference points for identifying contextual variables. When users of the contextual treatment group received the newsletter, it was requested that they specify the context in which they wanted to receive recommendations (i.e., for a personal purpose or for a gift and then for whom or what was their mood). Then recommendations were then shown to the participants only for the specified context.

Random/Control group. Unlike the content-based and context-aware approaches, the random approach does not take the user profile into consideration when recommending new products. Instead,

it randomly selects, without replacement, a set of items to recommend from the products that have not been recommended or purchased before.

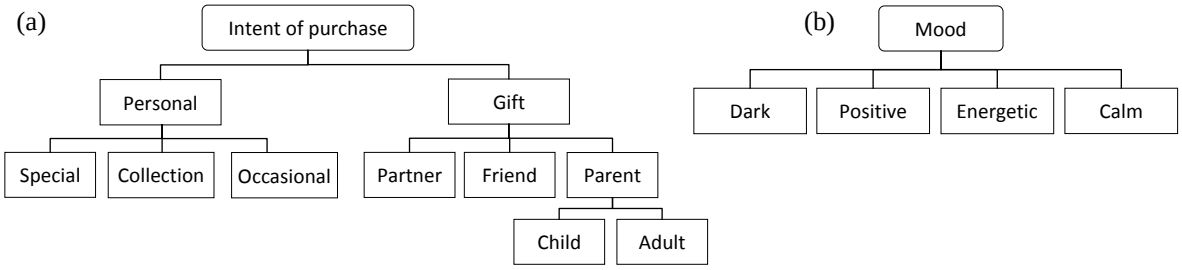

Figure 1. Hierarchical structure of context: (a) intent of purchase, (b) customer's mood.

3.3. Performance measures

We used accuracy, diversification, trust and customer purchasing activities as metrics that measure various performance aspects of our recommendation methods. We describe them in turn now.

Accuracy. Accuracy was measured by precision and by the average rating provided by customers. Among the traditional IR performance metrics, such as precision, recall and F-measure, only precision can be computed in our case, since it is not possible in a live experiment to know the ratings of the unseen items needed to compute recall and F-measure. Precision of a recommendation (Herlocker et al., 2004) for each customer in each week is measured as:

$$\text{Precision}_{\text{W}} = \frac{N_j^R}{N_j^S} \tag{1a}$$

where w is the week, N^S_j is the total number of the items recommended to the j-th customer selected (S) by the RS as items to be recommended and N^R_j is the number of items which proved to be "relevant" (R) for the j-th customer among those selected by the RS. We considered an item being "relevant" if it was rated as 3, 4 or 5 on the 0-5 scale. We decided to consider 3, 4 and 5 as "relevant" ratings instead of considering only 4 or 5 as discussed in (Herlocker et al., 2004) since our rating scale was from 0 to 5 instead of 1-5 scale as in (Herlocker et al., 2004). Precision for each customer is then computed as the average across the nine weeks of the experiment in each treatment as:

$$Pre = \frac{1}{9}\sum_{w=1}^{9} \text{Precision}_{\text{w}} \tag{1b}$$

We also measured accuracy by computing the average rating provided by each customer. We first computed the average rating provided by each user in each week as

$$\text{Average rating}_w = \frac{1}{N}\sum_{n=1}^{N} rating_{wn} \quad (2a)$$

where $rating_{wn}$ is the rating provided by the user to item n in each week w and N is the total number of items rated by the user. The overall average rating provided by each user is then

$$Rtn = \frac{1}{9}\sum_{w=1}^{9} \text{Average rating}_w \quad (2b)$$

We computed the formulas (1) and (2) using the feedback data provided by the customers after receiving each newsletter. We used *Pre* and *Rtn* variables in SEM as the two observable variables of a latent variable called “accuracy”.

An interesting alternative to the accuracy measures based on customer *ratings*, as specified in this section, would be to introduce accuracy measures based on actual *purchases*. It would be interesting to report such measurements and compare alternative accuracy metrics, those computed by considering the products rated out of what was recommended and those computed by considering the products purchased out of what was recommended. Unfortunately, the appropriate data were not available. The main reason is that only very few recommended items were actually purchased by the customers. We report the mean number of products which were both recommended and purchased by users in each group in Table 11 (Appendix 2). This is not only a limitation of the present research but is typical in most (perhaps all) of the other studies done in the recommender systems academic community. The main reason for this is a limited period of time that a researcher can use the industrial-level recommendation engines for scientific experiments. Large companies that have information on millions of customers and longitudinal data of several months or even years can successfully perform such studies. Unfortunately, they prefer not to publish their results and keep them as proprietary information. Furthermore, it would be important to conduct such study over a long time horizon in order to collect bigger datasets and computing more convincing performance metrics. However this complicates such studies even more because of the need to control for various changes in the customer population and experimental settings that occur over time. Finally, we would

like to point out that the idea of incorporating business performance oriented measures to evaluate recommender systems is related to the concept of "operational statistics" proposed by (Liyanage and Shanthikumar, 2005) in the context of certain types of inventory control problems.

Trust. At the end of the experiment, we provided the participants with a final survey in which we asked them to answer the 11 trust-related questions presented in Appendix 1. These questions asked the customers how much they agree on certain statements about the newsletter service. The purpose is to measure how much the participants trusted the received recommendations and to study whether there were differences in customers' trust across the treatments.

Trust, in general, is a multidimensional concept (McKnight et al., 2002). Many researchers have studied this multidimensional concept, and Mayer et al. (1995) demonstrated that the three most important dimensions of trust include *ability*, *benevolence*, and *integrity*. "Ability" (also referred to as "competence") represents the ability "of the trustee to do what the truster needs", "benevolence" represents the "trustee caring and motivation to act in the truster's interests", and "integrity" represents the "trustee honesty and promise keeping" (McKnight et al., 2002). All the three concepts are *interdependent*, as observed by Mayer et al. (1995). Schoorman et al. (2007) insist on the need of measuring trust by using all the three aspects because all three factors can contribute to trust. Several papers including those in the information systems and e-commerce areas (Gefen, 2002; Ganesan, 1994; Gefen and Silver 1999; Jarvenpaa et al., 1998; Gefen et al., 2003; Wang and Benbasat, 2005; Pavlou et a., 2007) have defined "trust" as a multi-dimensional variable embracing these concepts. In the recommender systems literature, Wang and Benbasat (2005) define "trust in a recommendation agent as an individual's beliefs in an agent's competence, benevolence, and integrity" and they clearly demonstrate that "all of them hold for trust in online recommendation agents". This approach based on measuring the overall level of trust that encompasses *all* the three aspects was followed by several other authors, as the review of prior literature in Schoorman et al. (2007) demonstrates. In our work, we embraced this prior research on trust and structured our questionnaire so that it is consistent with these prior concepts and definitions of trust, as shown below.

The constructs for trust were derived from prior studies, such as (Mayer et al., 1995; Beldad et al., 2010). We selected and adapted the set of questions and scales used by prior studies (Doney and

Cannon, 1997; Wang and Benbasat, 2005; Schoorman et al., 2007). Each answer was provided on the (1-5) scale:

- Four questions (from Q_2 to Q_5) are measures of "ability" and investigate the users' perceptions of whether the recommended products ("books", "recommendations" and "products") were aligned with the users' needs;
- Two questions (Q_6 and Q_7) are measures of "integrity". Integrity refers to keeping commitments and not lying, implying reliability, (McKinght et al., 2002) and, therefore, focusing on users' opinions about newsletter reliability in our context;
- Two questions (Q_8 and Q_9) are measures of "benevolence" that focus on the users' opinion about the firm's motivations.

For each question we computed the average of individual responses across the customers in each treatment.

Three additional questions were included in the survey (see Appendix 1). The first question (Q_1) was used as manipulation check. The last two questions (Q_{10} and Q_{11}) are not directly related to the measurement of trust. They were included as additional measures of purchasing behavior to be used in the case the customers involved in the experiment did not purchase anything during the 9 weeks of the experiment. We did not use Q_{10} and Q_{11} in the analyses. The fourth question in the survey (Q_4) was used as a measure of diversification as defined below. Therefore, seven measures of trust were used in total in the experiments. They were used in SEM as observable variables of a latent variable called "trust".

Diversification. We measured the recommendation diversification in our experiments by measuring both the *diversity* and the *novelty* of recommendations. As reported in Section 2, prior research considers novelty and diversity as two aspects of the more general concept of diversification (Vargas and Castells, 2011; Zhang and Hurley, 2009). We used four metrics for diversification, including three metrics for diversity and one for novelty as shown in Table 1. Each one is described below.

Diversity is defined as the extent to which the items in the recommendation list belong to different categories of items. We use "individual diversity", as opposed to "aggregated diversity" (Adomavicius and Kwon, 2012), because we want to evaluate the single individual's reaction to recommendations. Diversity is measured using the classification of diversity metrics in probability-based, logarithm-based and rank-based measures (McDonald et al., 2003). Among these metrics we selected the three most popular measures from each of the three categories, i.e., the Simpson's diversity index, the Shannon's entropy and the Tidemann & Hall's diversity index (McDonald et al., 2003) respectively. We computed the three metrics by using the data collected during all the experiments in each treatment. We used four comic book categories, according to the main classification the company uses to present its products in the web-site: 1) Marvel comics (including the well-known comic books popularized by the American publisher); 2) Manga comics (including all comic books published in Japan); 3) other comics (including all comic books popularized by either European publishers or American publishers other than "Marvel" brand); 4) bundled comics (including any kind of comic books sold in association with a DVD or other media contents). The choice of these categories was made in agreement with the company management.

The Simpson's diversity (*SD*) for each user is defined as:

$$SD = 1 - \sum_i p_i^2 \tag{3}$$

where p_i is the proportion of recommended items in the *i*-th category. The normalized Shannon's entropy (*Ent*) for each user is computed as:

$$Ent = -\sum_i p_i \log_k p_i \tag{4}$$

where p_i is the proportion of recommended items in the *i*-th category and k is the number of categories. The Tidemann & Hall's diversity index (*TH*) for each user is measured as:

$$TH = 1 - \frac{1}{(2\sum_i rp_i) - 1} \tag{5}$$

where *r* is the rank of the *i*-th category (ranked with 1 as the largest category). In order to provide each dataset with a ranking of categories, we used the number of distinct items contained in each category as defined by the relative website.

Novelty is defined as the extent to which a customer did not know the items in the recommendation list. We measured the "novelty" of recommendations using the fourth question in the final survey (Q_4), namely the extent of agreement to the statement "Personalized newsletters recommended comic books that I didn't know" (see Appendix 1).

We did not combine the four measures of diversification into one; instead, we used them in SEM as four observable variables of a latent variable called "diversification". We believe that this choice is appropriate because diversity and novelty represent different aspects of customer behavior. The two metrics are complementary rather than similar because the former captures how different items are to each other while the latter captures the customer's perception of discovering new items. Moreover, they both are required to capture "serendipity" aspects of recommendation quality.

It is clear from these definitions that the diversification metrics are also different from precision. For example, a customer can receive very precise but not diverse recommendations and vice versa. Furthermore, customer's diversification measures are not necessarily correlated with her purchase measures because (1) diversity is measured by looking at the item categories in the recommendation list, (2) novelty is measured as a customer's perception, (3) purchases are measured through the money spent, item purchased and the prices of those items. For example, a recommendation list may be very accurate but items are very similar to each other and few of them are novel. Predicting which item the customer will buy or whether her trust increase or decrease is not trivial. Our experimental settings aim at answering these kinds of questions.

Purchases. We decided to measure business performance associated with the use of a RS by measuring the purchasing behavior of customers during the experiment. In particular, we measured the following metrics: the money spent by customers, the quantity and the price of the items purchased, as we define below. These metrics were computed both during the nine weeks (2 months) of the experiment and during the 20 months before it. We selected these three metrics because (a) they are important metrics of the economic activity of a company, (b) are directly related to recommendations, and (c) we can easily measure them in our study.

We measured the purchased *quantity (Qty)* of a product per month per capita in each group during the experiment by counting the total number of products bought in each group divided by the number of months divided by the number of customers in each group:

$$Qty = \frac{\text{Number of items purchased by the users during the experiment}}{2 \times \text{number of users in the treatment}} \tag{6}$$

where 2 is the number of months (the length of our experiment).

We measured the mean monthly *expenditure* (money spent, *Mon*) per capita by the customers during the experiment in each group by using the following formula:

$$Mon = \frac{\sum_i \text{Number of items}_i \text{ bought during the experiment} \times \text{price}_i}{2 \times \text{number of users in the treatment}} \tag{7}$$

where 2 is the number of months (the length of our experiment).

We also measured the average *price (Pri)* of the products bought during the experiment by computing the average of the prices of the products bought in each period by each treatment group:

$$Pri = \frac{\text{Money spent by customers during the experiment}}{\text{Number of items bought during the experiment}} \tag{8}$$

We computed the same three measures by using prior customer data related to the 20 months before the experiment (Section 3.1). In this case the formulas (6) – (8) have to be modified by replacing the phrase “during the experiment” with “during the 20 months before the experiment” and by replacing 2 at the denominator of (6) and (7) with 20. The comparison between corresponding measures *before* and *during* the experiment is useful to observe changes in the customer purchasing behavior. We used *Qty*, *Mon* and *Pri* in SEM as three observable variables of a latent variable called “Purchases”.

Table 1 reports the measures used in the experiment. For the questions in the survey, a brief description of each question is provided in brackets.

Table 1. Summary of performance measures used in the experiment

Latent variable	Observed variable		Unit
Accuracy	*Pre*	Average precision of recommendations	%
	Rtn	Average rating provided by the user	%
Diversification	*SD*	Simpson's diversity	0-1
	Ent	Shannon's entropy	0-1
	TH	Tidemann & Hall's diversity	0-1
	Q_4	Novelty *("Recommended books that I didn't know")*	0-5
Trust	Q_2	Ability *("It is like a real expert")*	0-5
	Q_3	Ability *("Provided relevant recommendations")*	0-5
	Q_5	Ability *("I am willing to let it assist me")*	0-5
	Q_6	Integrity *("It is reliable")*	0-5
	Q_7	Integrity *("I trust it")*	0-5
	Q_8	Benevolence *("Created to help me")*	0-5
	Q_9	Benevolence *("It is a service to customers")*	0-5
Purchases	*Mon*	Money spent by customers during the experiment	€
	Qty	Quantity (number) of items purchased	#
	Pri	Average price of items purchased	€/#

4. EFFECT OF ACCURACY AND DIVERSIFICATION ON TRUST AND PURCHASES

In this section we study the effect of accuracy and diversification on customers' trust and purchases. According to the literature on RSes that is discussed below, we have developed five hypotheses about the relationships among accuracy, diversification, trust and purchases of customers receiving recommendations.

4.1. Hypotheses development

Following the literature review described in Section 2, we found that much research on RSes assumes that maximizing the recommendation accuracy leads to better economic results for the companies deploying the RS (Chen and Wu, 2007; Schafer et al., 2001; Gunawardana and Shani, 2009). Therefore, the first hypothesis can be stated as:

H1: *the recommendations' accuracy affects purchases*

In addition to accuracy, some scholars demonstrated that other performance measures are equally important. One such measure is diversification, as discussed in Section 2 and demonstrated in (Baumol and Ide, 1956; Kahn and Lehmann, 1991; Brynjolfsson et al., 2003; Fleder and Hosanagar,

2009, Adomavicius and Kwon, 2012) and by other researchers. Therefore, the second hypothesis can be stated as:

H2: *the recommendations' diversification affects purchases*

Another important factor which was shown to affect the purchasing behavior of customers is their trust in the recommender system (Pathak et al., 2010; Pu et al., 2011). Therefore, the third hypothesis can be stated as:

H3: *the customers' trust affects purchases*

Several scholars argued that customers' trust in a RS is driven by the accuracy of its recommendations (Zhang et al., 2011; Bharati and Chaudhury, 2004; Hess et al., 2005; Sinha and Swearingen, 2001; Swearingen and Sinha, 2001; Komiak and Benbasat, 2006) and by their diversification (Xiao and Benbasat, 2007; Lenzini et al., 2009; Cooke et al., 2002; Adomavicius and Kwon, 2012). Therefore, the fourth hypothesis can then be stated as:

H4: *the recommendations' accuracy and diversification affect the customers' trust*

Combining the previous hypotheses, we built the fifth hypothesis:

H5: *the recommendations' accuracy and diversification affect the customers' trust which, in turn, affects purchases*

We tested these five hypotheses by using the structural equation modeling (SEM) (Bollen, 1989). We built the SEM models for these hypotheses, as shown in Figure 2. We also tested several variants of these models, each variant using a subset of the observable variables depicted in Figure 2. We systematically varied the subset and the combinations of variables. For the sake of brevity, only the relevant results are presented in the section below.

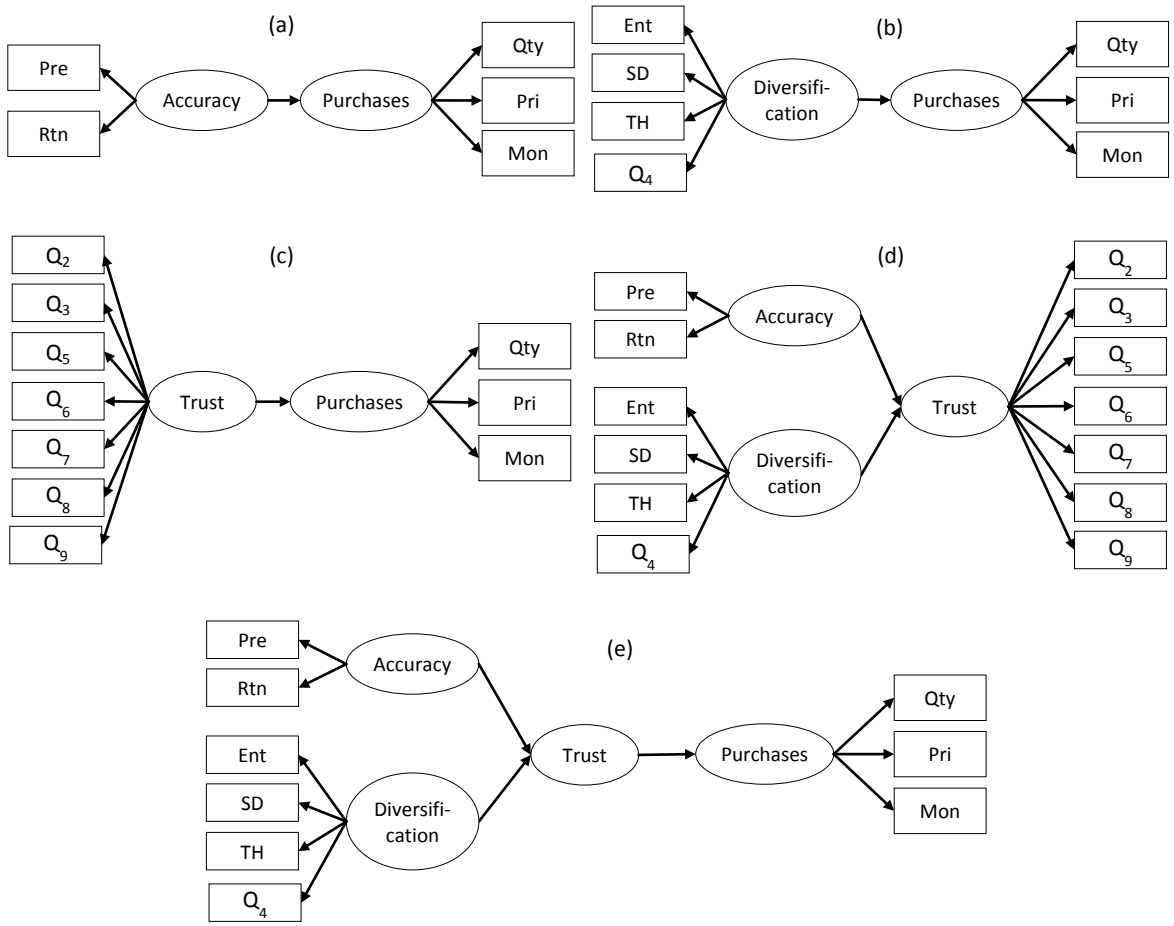

Figure 2. Structured equation models for testing hypotheses H1 (a), H2 (b), H3 (c), H4 (d), H5 (e)

4.2. Test of hypotheses

We performed several statistical tests in order to check whether there are significant differences among the three treatments and the population. In particular, we have performed t-tests and chi-square tests. We were able to perform such comparison tests with the population at large because the company that collaborated with us on this project gave us access to the demographic and the transactional data for the whole population of their customers. In particular, we selected six variables pertaining to demographic and transactional information for each treatment and the whole population: 1) age; 2) gender; 3) number of orders; 4) number of purchased items; 5) money spent on the Website; 6) year of subscription (length of the customer relationship with the online shop).

Further, we measured the distributions of each variable over the three groups that were provided content-, context-based and random recommendations respectively and over the whole population,

and performed the tests of significance described below. The tests on age, gender and subscription year (1, 2 and 6, respectively) were run on the total number of users in the three treatment groups and in the whole population. The tests on numbers of orders, number of purchased items and money spent (3, 4 and 5, respectively) were run on the subset of users who purchased at least one product before the experiment.

1) *Age*. We performed a t-test among the three treatment groups (namely content-based, context-aware and control) and we did not find any statistically significant difference (for $p>0.05$). We have also performed a t-test between the three treatment groups (namely content-based, context-aware and control) and the whole population, and we did not find any statistically significant differences either (for $p>0.05$). The results are presented in Table 10a in Appendix 2.

2) *Gender*. We performed a Chi-square test among the three treatment groups (namely, content-based, context-aware and control), and we did not find any statistically significant difference (for $p>0.05$). We have also performed a Chi-square test between the three treatment groups (namely, content-based, context-aware and control) and the whole population, and we did not find any statistically significant differences either (for $p>0.05$). The results are in Table 10a (Appendix 2).

3) *Number of orders*. We have measured the average number of orders (completed purchasing sessions) done by each user in each treatment group and by the whole population during the 20 months before the experiment (the company with which we have collaborated in this project also gave us access to this additional data). We performed t-tests in order to compare the three groups to each other and to the whole population for each variable. We did not find any statistically significant differences (for $p>0.05$). The results are presented in Table 10b (Appendix 2).

4) *Number of purchased items*. We have measured the average number of items that each user purchased (in each treatment group and in the whole population) during the 20 months before the experiment (again, we could do it because the company has also given us access to this data). We have also performed t-tests in order to compare the three groups to each other and to the whole population for each variable. We did not find statistically significant differences (for $p>0.05$). The results are presented in Table 10b (Appendix 2).

5) *Money spent on the Website*. We have measured the average number of items that each user

purchased (in each treatment group and in the whole population) during the 20 months before the experiment (again, this analysis was possible because the company has also given us access to this data). We have performed a t-test among the three groups and the whole population for each variable and we did not find statistically significant differences (for $p>0.05$). The details of this result are presented in Table 10c of Appendix 2.

6) *Subscription year*. We have performed a Chi-square test among the three treatment groups (namely content-based, context-aware and control) and we did not find any statistically significant difference (for $p>0.05$). We have performed a Chi-square test between the three treatment groups (namely content-based, context-aware and control) and the whole population and did not find any statistically significant difference (for $p>0.05$). The details of this result are presented in Table 10c of Appendix 2.

Some of the aforementioned analyses (1, 2, 6) were run on the total number of users in the groups while the other analyses (3, 4, 5) were run on the subset of users who purchased at least one product before the experiment. The reason is that the former were done to check possible differences in the *demographic* characteristics of the users, while the latter were done to check possible differences in the *purchasing behavior* of the users before the experiment. Only a subset of users in each group had purchased at least one product before the experiment and, therefore, the number of observations in Table 10a, 10b, 10c are different. These differences are, then, not caused by any phenomenon of customer attrition during the experiment. The tests ran on these tables and Figures 12 and 13 (see next paragraph and Appendix 2) show that these differences do not diminish the generalizability of the results described below.

All of the aforementioned t-tests were run as follows. We first ran the Levene's Test for Equality of Variances between groups and, based on this, we have made the corresponding assumption between equal or unequal variances between groups. Then we ran the independent sample t-test on the data. We used this procedure to make the correct assumption about variance between groups instead of assuming equal or unequal variance a priori, such as it is done by other statistical tests (e.g., Welch's t-test). Table 10a, 10b and 10c show the number of observations, the average and the standard deviation for each group and each variable. In addition, these tables show the difference

between the means and the t-value for each t-test we ran.

In conclusion, we have performed all the statistical tests described above, and none of them found any statistically significant differences between the selected groups of customers used in our study and the whole population across 6 important variables. This means that our sample is representative of the whole population of the customers of that company.

Moreover, we checked for biases in the treatment groups with respect to the propensity to trust. We found no statistically significant differences in the general trust levels of the users in the different groups, using Q_1 responses. The mean value of the answers to Q_1 in each group is shown in Table 9b in Appendix 2. Table 9b shows that the differences between groups are non-statistically significant. Thus, the users in the different treatment groups were similar in terms of propensity to trust.

We also performed the test of price distributions in order to exclude any biases which might influence customers' reactions due to significant difference in the price distributions of the products. To perform these tests, we measured (a) the distribution of the product prices in the whole catalogue, (b) the distribution of the price of recommended items. We did these tests for both the entire duration of the experiment (9 weeks) and for each week of the experiment. We then performed the analysis of the price distributions for: 1) all the products available in the catalogue in each one of the nine weeks; 2) recommended items vs. all the products available in the catalogue; 3) recommended items vs. all the products available in the catalogue among the nine weeks. These tests are described in greater detail below:

1) *Price distribution of all the products available (catalogue) in each one of the nine weeks*. We performed a t-test to compare the average price of the products in the catalogue in each couple of weeks. We did this across all the weeks. We did not find statistically significant differences ($p>0.05$). Furthermore, Figure 8 (Appendix 2) shows lines representing products price distributions for each one of the nine weeks of the experiment. As Figure 8 shows, the price distribution of the products was the same in each one of the nine weeks of the experiment (with no statistically significant differences). This allows us to conclude that *prices did not change significantly* during the nine weeks and that customers did not perceive any variation in the price of the products available which might have biased their reaction to recommendations.

2) *Price of recommended items vs. price of products available (catalogue)*. We measured the distribution of the prices of the recommended items and the distribution of the prices of the items across the whole catalogue during the entire period of the experiment. We performed a t-test to compare the two distributions and did not find statistically significant differences ($p>0.05$). Furthermore, Figure 9 (Appendix 2) shows the two distributions. As we can see in Figure 9, the two distributions are very similar (with no statistically significant differences). This allows us to conclude that there is *no bias* in the recommender systems towards recommending more of the highly priced (vs. low priced) products and that recommendations were not biased towards certain price categories.

3) *Price distribution of recommended items vs. price distribution of all the products available (catalogue) among the nine weeks*. We measured the distribution of the prices of recommended items vis-a-vis the distribution of the prices of all the items in the entire catalogue for each one of the nine weeks. We performed a t-test comparing the prices of the products in the catalogue vs. the prices of the recommended products for each week. We did not find any statistically significant differences across the two distributions ($p>0.05$). We do not present the specific data supporting this claim because of the space limitation and because this would only complement our findings reported in item (2) in the previous paragraph.

Finally, we performed statistical tests on the response rate of each treatment. These tests were useful to ensure that no bias occurred among the treatments in terms of customers' response which may influence the interpretation of the experimental results. We measured the response rate for each of the nine weeks for the three treatment groups. Figure 10 (Appendix 2) shows the corresponding distributions. We then ran the tests of statistical significance in order to check whether there are differences among the response rates of the three treatments, one per each week and we did not find any statistically significant differences among them ($p>0.05$).

After we performed all these tests described in this subsection and did not find any biases, we are ready to describe the outcomes of our tests **H1** – **H5**.

H1. We tested the hypothesis that the recommendation accuracy (measured by precision and average ratings) directly affects customers' purchases (measured by quantity and price of items purchased and by the customers expenditure). The result is that the model estimates are *not*

statistically significant. However, if the model in Figure 2(a) is modified by deleting *Pri* and *Mon* and using only the observable variable *Qty* as dependent variable, then the model is significant. Table 2 reports the results of this model where the recommendation accuracy is measured by *Pre* and *Rtn* and the purchases are measured by *Qty*.

Table 2. How recommendations' accuracy (*Pre* and *Rtn*) affects quantity of purchased items (*Qty*).

Estimate Accuracy → Quantity	4.540 (2.024)*
p-value	.846
χ^2/df	.038
RMSEA	.000
p-value close to fit	.873

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

The results demonstrate that H1 is not supported if the latent variable "purchases" is measured by all the three observable variable. H1 is supported if "purchases" is measured by "quantity" only (at 0.05 significance level). Therefore we can state that *the recommendations' accuracy directly affects the number of items customers purchase, however it does not always directly affect the money customers spend (i.e., a company's revenues)*. This result is consistent with the intuition of the scholars (Pathak et al., 2010; Pu et al., 2011) who have demonstrated that accuracy can only partially explain customers behavior because it is not the only relevant factor (as described in Section 2). This observation cannot be generalized to any application of RSes, however the result clearly shows that better accuracy does *not* necessarily result in better sales.

H2. Table 3 reports the results of the model built for testing the second hypothesis (Figure 2(b)). This model tests whether the recommendation diversification (measured by Shannon's entropy, Simpson's diversity, Tidemann & Hall's diversity and the responses to Q_4) directly affects customers' purchases (measured by *Qty*, *Pri* and *Mon*).

Table 3. How recommendations' diversification (*Ent*, *SD*, *TH*, Q_4) affects customers' purchases (*Mon, Qty, Pri*).

Estimate Diversification → Purchases	-7.431 (11.811)
p-value	.739
χ^2/df	.685
RMSEA	.000
p-value close to fit	.905

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

The results demonstrate that H2 is *not* supported. The model estimate is not significant (p>.05). We found similar results when testing several variations of the model (e.g., considering only one out of the three observable variables used to measure purchases). Therefore we can state that *the recommendations' diversification alone does not necessarily have a direct effect on customers' purchases*. The result is consistent with prior research which has proved that diversification is not sufficient to explain customer behavior and the economic performance of a company.

H3. Table 4 reports the results of the model built for testing the third hypothesis (Figure 2(c)). This model tests whether customers' trust, as measured by the answers to questions Q2 – Q9 in the survey (see Appendix 1), directly affects customers' purchases (measured by *Qty*, *Pri* and *Mon*). The results support H3. The effect of trust on purchases is significant with p<.05. Therefore, we can conclude that that *customers' trust does directly affect customers' purchases*. This result is consistent with prior literature which demonstrated that higher trust can increase sales and positively influence the shoppers' decisions.

Table 4. How trust (Q_2, Q_3, Q_5, Q_6, Q_7, Q_8, Q_9) affects customers' purchases (*Mon*, *Qty*, *Pri*).

Estimate Trust → Purchase	.374 (.188)*
p-value	.254
χ^2/df	1.152
RMSEA	.031
p-value close to fit	.753

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

Although we do not claim a causal relationship between trust and purchases, this experiment adds to prior studies the evidence that trust is directly correlated to both the quantity of items customers can buy and their price and, in turn, to the overall customers expenditure. Since trust proves to be such crucial factor to drive the business performance gained by the use of a RS, it is now important to study the drivers of trust by testing the fourth hypothesis.

H4. Table 5 reports the results of the model built for testing the fourth hypothesis (Figure 2(d)). This model tests the hypothesis that the recommendation accuracy (measured by precision and average ratings) and the recommendation diversification (measured by *Ent*, *SD*, TH and Q_4) directly affect customers' trust, as measured by the answers to the questions in the final survey (see Appendix 1).

Table 5. How recommendations' accuracy (*Pre* and *Rtn*) and diversification (*Ent, SD, TH, Q_4*) affect customers' trust (Q_2, Q_3, Q_5, Q_6, Q_7, Q_8, Q_9).

Estimate Diversification → Trust	3.146 (1.495)*
Estimate Accuracy → Trust	1.551 (.664)*
p-value	.453
χ^2/df	1.010
RMSEA	.008
p-value close to fit	.937

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

The results support H4. Therefore, we conclude that *recommendation accuracy* and *the recommendation diversification do affect customers' trust*. This result is particularly important if seen in the light of the extant literature. While prior research showed that accuracy alone cannot explain customers' behavior and diversification can be an additional important factor, our research demonstrates that it is the *combination* of higher levels of accuracy *and* diversification that can explain purchases. Therefore, the goal of a RS design should be that of maximizing *both* accuracy *and* diversification. Moreover, it is interesting to note that the effect of diversification on trust is higher than the effect of accuracy, as the estimates reported in Table 5 show.

H5. Table 6 reports the results of the model built for testing the fifth hypothesis (Figure 2(e)). This model tests whether the recommendation accuracy (measured by precision and average ratings) and the recommendation diversification (measured by *Ent*, *SD*, *TH* and Q_4) affect customers' purchases (measured by *Qty*, *Pri* and *Mon*) through trust (measured by the answers to the questions in the final survey - see Appendix 1).

Table 6. How recommendations' accuracy and diversification affect trust and in turn customers' purchases.

Estimate Accuracy → Trust	1.767 (.792)*
Estimate Diversification → Trust	3.541 (1.759)*
Estimate Trust → Purchase	.390 (.183)*
p-value	.360
χ^2/df	1.047
RMSEA	.017
p-value close to fit	.964

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

The results *support* the hypothesis. Therefore, we can conclude that the recommendations' accuracy *and* diversification affect customers' trust, which in turn affects customers' purchases. This finding does not prove that increasing both accuracy and diversification always *causes* an increase in trust which, in turn, *causes* higher level of purchases. We rather observe that these variables are correlated in terms of a *predictive relationship* (Shmueli and Koppius, 2011). Therefore, this result has immediate implications on the RS design and particularly on the issue of this research, i.e. studying how context-aware recommendations affect customer's behavior compared to other kinds of RS. In the next section these implications are discussed.

In order to make visually clear which hypotheses hold and which do not, we summarize all the results discussed above in Table 7.

Table 7. Hypotheses tested and their statistical significance

Hypothesis	**Sign. (Sign. level)**
H1 Recommendations' accuracy affects purchases	No ($p > 0.5$)
H2 Recommendations' diversification affects purchases	No ($p > 0.5$)
H3 Customers' trust affects purchases	Yes ($p < .05$)
H4 Recommendations' accuracy *and* diversification affect customers' trust	Yes ($p < .05$)
H5 Accuracy and diversification affect customers' trust which affects purchases	Yes ($p < .05$)

5. EFFECT OF CONTEXT

As mentioned in Section 3.1 we followed a two-step procedure in order to analyze how including context in recommendations affect customers' trust and purchasing behavior. The first step, analyzing the effect of accuracy and diversification on customers was discussed in Section 4. In this section, first of all, we analyze how including context in a RS affects the accuracy and diversification of recommendations and then we analyze how including context in recommendations affect customers' trust and purchasing behavior.

5.1. The effect of context on accuracy and diversification

In order to analyze how including context in a RS affects accuracy and diversification of recommendations, we compare the characteristics of the recommendations generated by the three recommendation engines with which we experimented in terms of accuracy and diversification. We

first compare accuracy and diversification separately and then we analyze and compare the combination of the two for each recommendation engine.

We first compute the average accuracy metrics across users in each treatment, i.e., content-based, context-aware and random recommendations, and compare them across the nine weeks of the experiment. Then we compute the average accuracy across the nine weeks for each single user for the three treatments and analyze their distributions.

Figure 3 reports the accuracy of recommendations for each group during the nine weeks of the experiments computed by averaging *Pre* (Figure 3a) and *Rtn* (Figure 3b) across the users in each treatment.

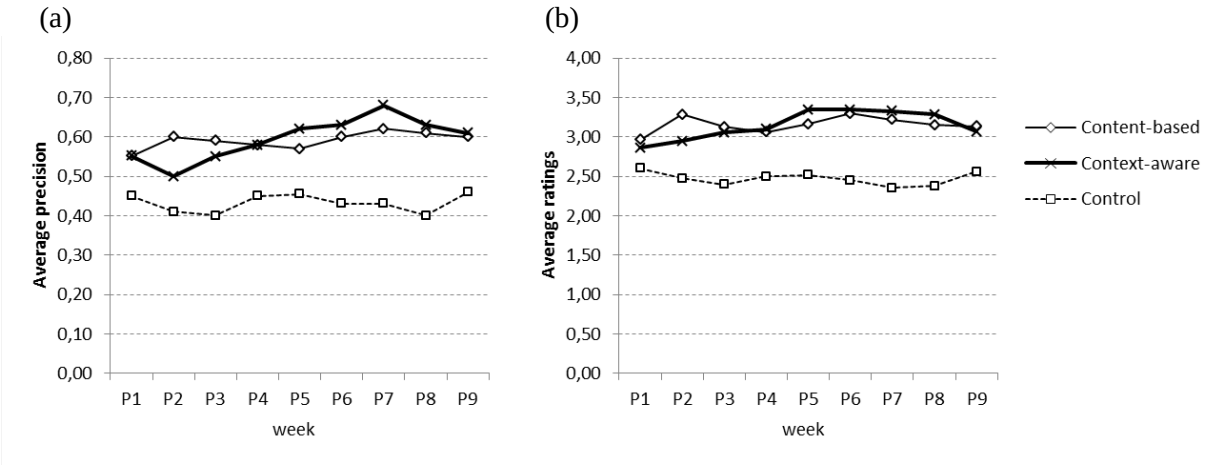

Figure 3. Precision (a) and average ratings (b) of recommendations in the nine weeks

As Figure 3(a) demonstrates, the precision of the recommendations generated by the content-based and context-aware RSes are significantly higher than that of the random recommendations, (statistically significant with $p<0.001$). To the contrary, the precision of the content-based and the context-aware RSes across the nine weeks is similar, i.e., there are no statistically significant differences found between the two groups ($p>0.5$). The precision of the CARS is slightly greater than that of the content-based RS after the fourth week. The reason is that the CARS approach takes more time to learn the preferences of customers because of the sparsity of the user-item matrix which is computed for each context with smaller amount of data. In fact, we followed the pre-filtering approach to context-aware recommendations (Adomavicius and Tuzhilin, 2011) and, therefore, split the uncontextual “User×Item” matrix into several “User×Item×Context” matrixes (one for each value

of the contextual variable). Therefore, each "User×Item×Context" matrix contains less ratings data than the uncontextual "User×Item" matrix. Therefore, there exists the following tradeoff between the uncontextual "User×Item" and the contextual "User×Item×Context" matrices. The contextual matrix has more homogeneous ratings data capturing users' context-specific experiences; but it has significantly fewer ratings. Therefore, the initial recommendations made in the first 4 weeks provide poorer recommendations vis-à-vis the uncontextual case due to the lack of the ratings data to train the model. However, the performance of the contextual matrices increases more significantly over time when more ratings become available and eventually catches up with the uncontextual case (due to the homogeneity of the context-specific ratings that better capture customer experiences than the uncontextual ratings). However, in each week after the first one, the system can make use of new users' feedbacks to update the "User×Item×Context" matrices, and the performance increases more significantly over time when more ratings become available each week. The performance eventually catches up with the uncontextual case after 4 weeks of new ratings data (due to the homogeneity of context-specific ratings that better capture customer experiences than the uncontextual ratings and due to having enough data to train the model).[2] These results are reinforced by the similar results reported in Figure 3(b) showing the average rating provided by users in each week computed by averaging *Rtn* defined in (2b) across the users in each treatment. Similarly to precision, the CARS performed slightly better than the content-based approach starting from week 4, as customers provided higher average ratings to the recommended items, while the random RS performed worse than the two personalized RSes. Only the differences between the content-based and the control groups and between the context-aware and the control groups were found statistically significant with $p<0.001$.

Figure 4 reports the distribution of recommendation accuracy across the users in each treatment. Accuracy was measured by *Pre* (Figure 4a) and *Rtn* (Figure 4b) for each customer by considering the whole set of recommendations received during the nine weeks of experiment. The graph reports the

[2] The performance decrease in the last two week is caused by the fact that the data available to the RSes to compute customers profile was reduced by the authors in order to keep the computation time low and avoid the risk of slowing down the delivery process.

percentage of customers who received a set of recommendations with a certain level of precision (a) or who provided a certain average rating from 1 to 5 (b). The differences between the three groups are statistically significant according to a t-test in all cases with $p<0.001$ except the difference between the content-based RS and the CARS when accuracy is measured by *Rtn*. These findings confirm the observation made above. Context has a significant effect on the accuracy of recommendations, given that the CARS outperforms the context unaware RS (content-based) in terms of precision.

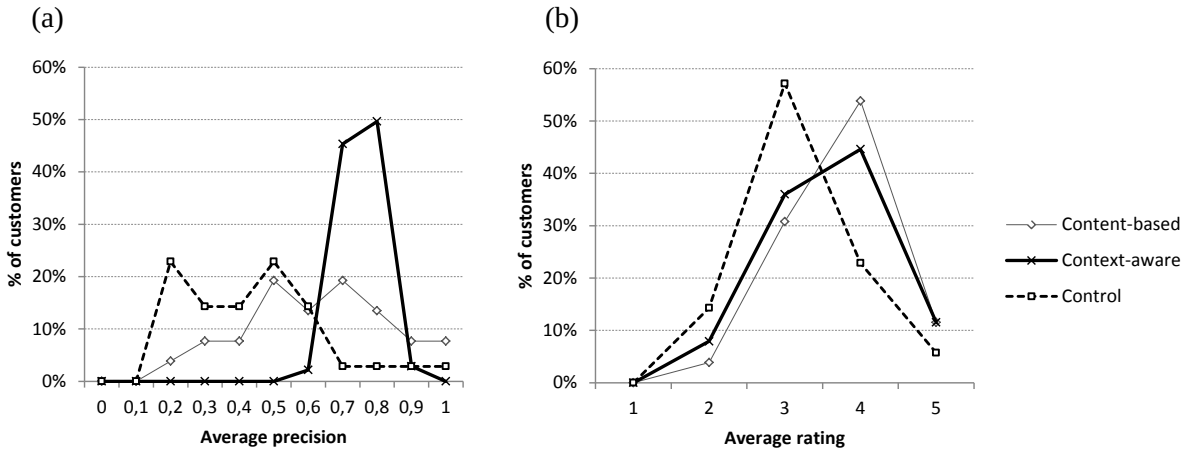

Figure 4. Distribution of (a) precision and (b) average ratings across users

Figure 5 reports three metrics of diversification for each group during the nine weeks of the experiments computed by averaging SD (Figure 5a), *Ent* (Figure 5b) and TH (Figure 5c) across the users in each treatment. The novelty metric Q_4 cannot be reported over time because it was measured only at the end of the experiment. As Figure 5 shows, the random recommender (Control) generated the most diverse recommendations, as expected, at least according to SD and *Ent* measures. *TH* present similar values, especially when random recommendations are compared to the context-aware ones, because this index is strongly influenced by the number of item categories. The recommendation diversification for the CARS is always higher than that of the content-based case. The differences are statistically significant according to a t-test in all the cases with $p<0.001$, except the difference between random and CARS when diversity is measured by TH.

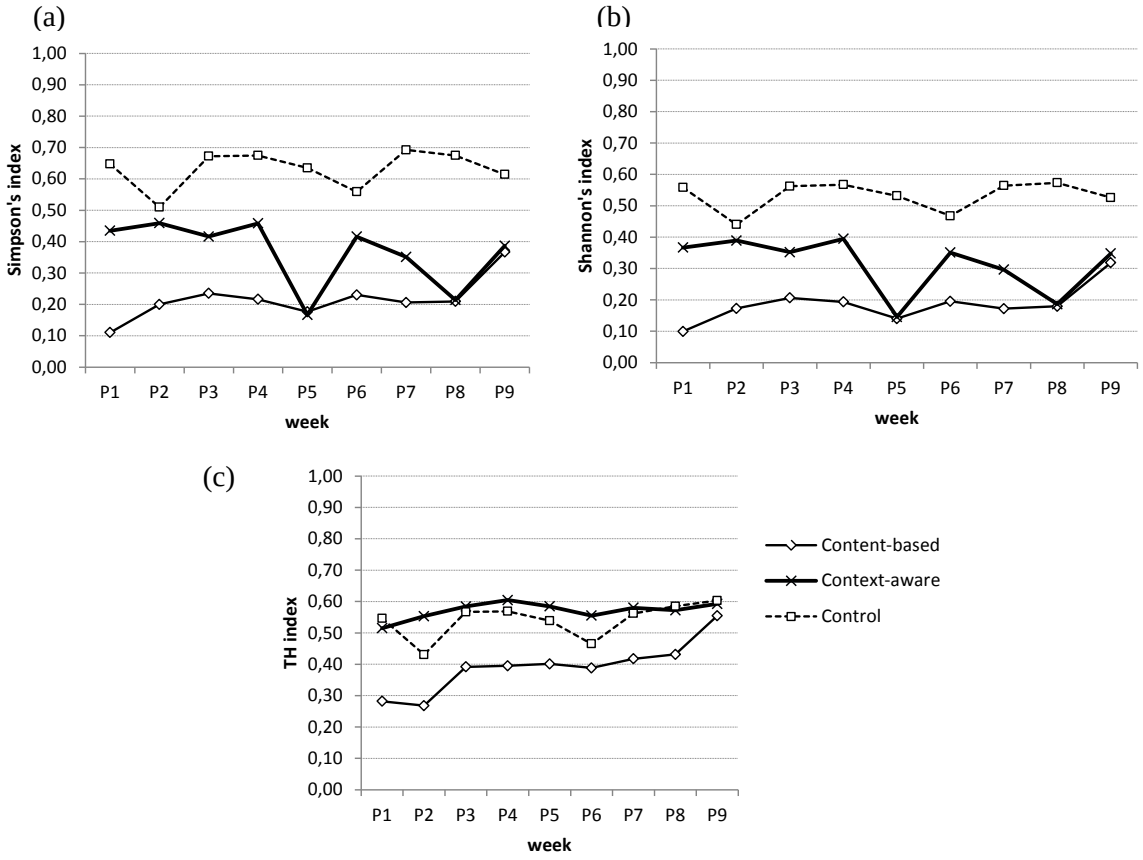

Figure 5. Diversity of recommendations measured by (a) Simpson's diversity, (b) Shannon's entropy and (c) Tidemann & Hall's diversity over the nine weeks (P1 – P9) of experiments.

Figure 6 reports the distribution of the diversification metrics for the three groups. The diversity of recommendations was measured using SD (Figure 6a), *Ent* (Figure 6b) and TH (Figure 6c). The recommendation novelty is measured by Q_4 (Figure 6d). In particular, the y-axes reports the percentage of customers who received a set of recommendations with a certain level of diversification (x-axes). All metrics are computed for each customer by considering the whole set of recommendations received during the nine weeks of experiments. It is interesting to notice that both the diversity and the novelty of the CARS approach are very similar to that provided by a random RS (Control), while the recommendations generated by the Content-based RS were much less diverse and less novel compared to the other systems. An explanation of the fact that the CARS generates almost as much diversification as a random RS is that the user can choose a different context in each week thus reaching different parts of the space of products. The CARS then generate recommendations for different context thus increasing the diversification. The differences are statistically significant

according to a t-test in all the cases with $p<0.001$. It is also interesting to notice that diversity and novelty are consistent. The recommendations generated by the random RS were perceived by customers in the Control group as the most novel, as expected, as Figure (6d) shows. The novelty of recommendations generated by a CARS (as measured by Q4) is close to that of a random RS and higher than that of a Content-based RS. The differences between the content-based and the control groups are statistically significant according to a t-test with $p<0.05$.

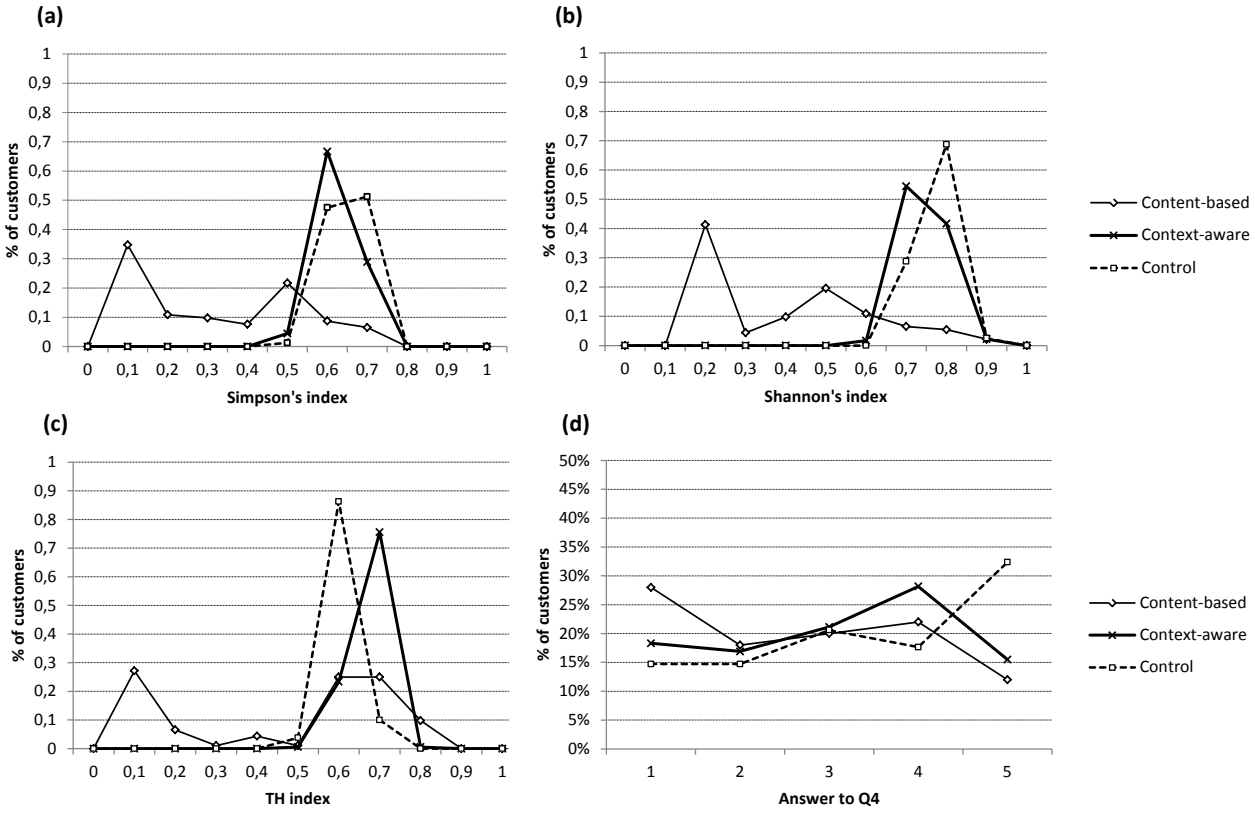

Figure 6. Distribution of (a) Simpson's diversity, (b) Shannon's entropy, (c) Tidemann & Hall's diversity and (d) novelty - Q_4 across users.

It is important to note that the experimental design, particularly the choice of including five "new" items and five "old" items in the recommendation list (see Section 3.1) did not entail a pre-imposed diversity in the recommendation list for the following reasons. First of all, the probability that a recommended item falls into a certain category (recall that we set certain product categories to measure diversity) is the same for "new" and "old" items because old items are comic books issued few months prior to our study (most comic books are published monthly as issues of a regular series). Therefore, if the proportion of new and old items in the recommendation list changes, the probability

that either the diversity or the novelty changes is very low. Secondly, we adopted the same strategy (five new items, five old items) for the three groups. Therefore, no bias would have influenced customers' behavior.

We also examined the combined effects of accuracy and diversification of the recommendations received by the users across the three treatments. For each treatment and each user, we computed the average accuracy and averaged it across the nine weeks and across users. Similarly, we computed the average diversification.

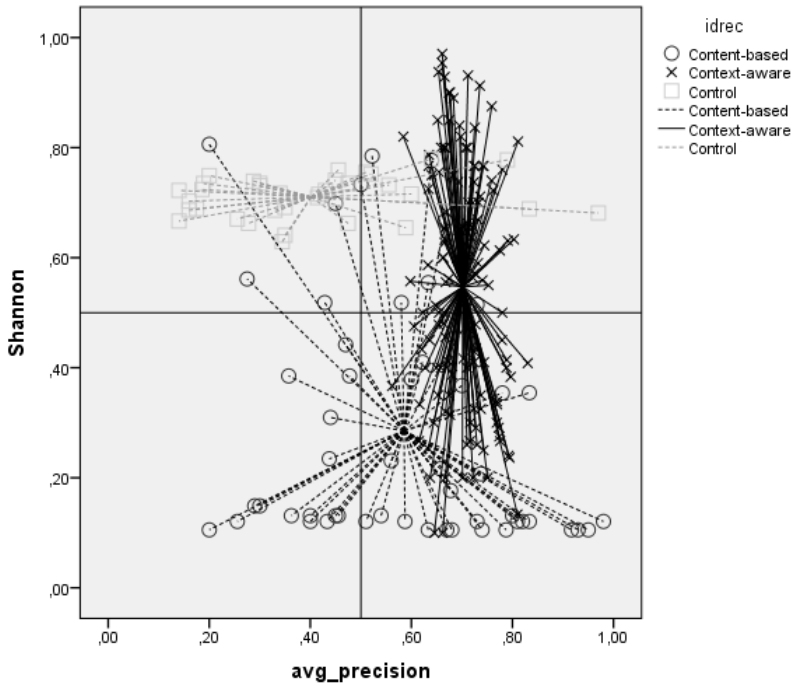

Figure 7. Accuracy (*Pre*) vs. diversity (*Ent*) of content-based (o), context-aware (×) and random (□) recommendations for each user in these three treatments.

Then we plotted accuracy (*x*-axis) against diversification (*y*-axis). We followed this procedure for each accuracy (*Pre* and *Rtn*) and diversification (SD, *Ent*, TH and Q_4) metrics combination obtaining eight plots. For the sake of conciseness, we only present one of these plots in Figure 7, where *Pre* (precision) is plotted against *Ent* (Shannon's entropy). The other graphs show similar results. In order to make the analysis consistent, we considered the subset of users for whom we have complete information, i.e., each user in this subset provided an answer to the final survey and for each of them we can compute the average precision and Shannon's entropy. Each point in Figure 7 represents a

user, and all the points representing users in the same treatment (content-based, context-aware and random recommendations) are connected to the centroid of the cluster of users in the precision–diversity space of Figure 7 belonging to the same treatment (content-based, context-aware or random clusters).

Looking at the position of the centroid of each treatment in Figure 7 and generalizing the observation to the other metrics, we can state that the content-based recommendations are characterized by high accuracy and low diversification on average, whereas the random recommendations are highly diverse but inaccurate. The context-aware recommendations are more accurate and more diversified than those generated by the content-based RS. Further, context-aware recommendations are significantly more accurate, albeit less diversified than the random ones. These differences among the centroids were significant in all the cases according to t-tests with $p<0.001$.

For the case of CARS, the range of the accuracy measure is very narrow, as shown in Figure 7, ranging from 0.55 to 0.85. The same variation range is significantly higher for both the Content-based and the Control (Random) groups. The variation span of diversity is high for both the CARS (from 0.15 to 0.95) and the content-based RS (from 0.15 to 0.80). In general, the recommendation diversification for a user in the Context-aware treatment depends on how much she is interested in exploring different contexts. The higher the interest, the more diversified recommendations are generated. The accuracy of these recommendations will be still high, independently of the user's interest in different contexts. The diversification for a user in the Content-based treatment depends on how many ratings she provides as a feedback to the RS. If the user provides few rating as feedback, the subsequent recommendations will be more diversified but less accurate. On the contrary, if the user provides many ratings as feedback, the recommendations will be more accurate but less diversified. Considering Figure 7, we can conclude that context-aware RSes provide more accurate *and* diversified recommendations than the traditional content-based RSes *and* significantly more accurate, although somewhat less diverse recommendations than the random RSes.

5.2. The effect of context on trust and purchases

As was shown in Section 4 accuracy and diversification affect trust and purchases. While improvements in accuracy and diversification *alone* do not necessarily affect customers' purchases

(hypotheses H1 and H2; see Table 7), the *combination* of higher levels of accuracy *and* diversification is effective in the sense that it affects purchases via higher levels of customer trust in recommendations (hypotheses H5). In Section 5.1 we have shown that context-aware recommendations improve both accuracy and diversification vs. the traditional content-based recommendations. Therefore, based on the combination of these results, we can hypothesize that context should affect trust and purchases. We conducted one more analysis to empirically validate this hypothesis.

We measured the actual purchasing behavior of the customers in the three treatment groups before and during the experiments. Table 8 reports the results.

Table 8. Purchasing behavior of customers in the three treatments

		Content-based	Context-aware	Control
Mon	*before*	2.03	1.95	0.91
(€)	*during*	2.38	2.50	0.94
%var		+16.9***	+28.2***	+3.6**
Qty	*before*	0.33	0.37	0.15
(#)	*during*	0.45	0.34	0.10
%var		+36.8**	-7.7***	-31.9**
Pri	*before*	6.18	5.26	6.20
(€/#)	*during*	5.29	7.31	9.43
%var		-14.5***	+38.9***	+52.1***

***Significant at p<0.01. **Significant at p<0.05.

The first two rows report the Euros spent (*Mon*) per month per customer before and during the experiment, respectively. The third row reports the percentage variation. Similar measures are reported for the number of products purchased (*Qty*) and their price (*Pri*). The two groups that received personalized recommendations increased the monthly expenditure. The increase in the context-aware group is higher than that of the content-based group, being 28.2% and 16.9% respectively. This finding shows that personalized recommendations are correlated with the increase in customer expenditures, even when a non-contextual RS, such as the one described in this paper, is employed. The monthly p.c. expenditure remained almost the same in the control group (3.6%). For the content-based group the quantity increased (36.8%) and the price of items decreased (-14.5%). On the contrary, the quantity decreased for the context-aware group by 7.7% while the average price of items increased by 38.9%. The quantity decreases by 31.9% in the control group while the few items bought have higher price (52.1%). The statistical differences were tested using the Wilcoxon test

(Barnes, 1994). It is useful to remark that the metrics presented in Table 8 are computed "per month per user" while those presented in Table 10b and Table 10c (Appendix 2), which report transactional data before the experiment, are computed "per user". The reason of the difference is explained in Appendix 2.

Because we want to demonstrate that the use of context-aware recommendations is correlated with an increase in the level of purchase, we checked that there was no significant variation in purchases before the experiment which could influence the results of our analysis. First, we measured the total sales of comic books over the period of 14 months before the start of our experiments. Figure 11 (Appendix 2) reports the results and shows that there was no general increase or decrease in the sales of comic books before the experiment. Second, we repeated this measure for the customers involved in the experiment and broke down the measurement for each group. We measured the money spent (Figure 12 in Appendix 2) and the quantity of purchased items (Figure 13 in Appendix 2). Also these figures do not show particular increasing trends in any of the three groups. Although we had the data pertaining to the 20 months before the experiment, we set 14 months as the observational period for the following reason. Fourteen months is the longest observational period for which we can track the purchases of the users involved in our experiment. If we considered the 15^{th} or the 17^{th} month before the experiment beginning, we would not find purchase data for the control group or for the other two groups, respectively. The main reason is that either users in the groups had not yet subscribed or they had just subscribed but had not purchased yet. Therefore, we set fourteen months as observational period as the best compromise between the necessity to analyze prior (pre-experimental) data and significant availability of the data on customer purchases.

Furthermore, to demonstrate that the results in Table 8 are not merely due to the (a) customer-level unobserved effects, (b) time-related unobserved effects and (3) preexisting systematic differences in trends across groups, we developed the following two econometric models and tested it on the data:

$$Purchase_{it} = \beta_i + \beta_1 \, Month + \beta_2 \, Month.CARS_i + \beta_3 \, Month.CONT_i + \varepsilon_{it} \quad (S1)$$

$$Purchase_{it} = \beta_i + \beta_1 \, Post_t + \beta_2 \, Post_t.CARS_i + \beta_3 \, Post_t.CONT_i + \varepsilon_{it} \quad (S2)$$

In particular, we used the first specification (S1) in order to study whether the changes in Table 8 can be due to customer (β_i), time (β_1) unobserved effects or to preexisting purchased trends between customers in the context-aware (β_2) and content-based (β_3) groups. We used the second specification (S2) to measure the net benefit of contextual (β_2) and content-based (β_3) recommendation after controlling for unobserved time related (β_1) and customer specific (β_i) unobserved factors. The details of the models and the results are provided in Appendix 3. The results of the analyses conducted by using S1 demonstrated there are neither customer-level unobserved effects nor time related unobserved effects that can increase the purchases across customers or preexisting declining or increasing differences in the purchase trends in the CARS and CONT groups (see Models 1, 2 and 3 in Appendix 3). The results of the analyses conducted by using S2 confirmed most of the results presented in Table 8 of the paper. In particular, as in the case of S2, this analysis also confirmed that both context-aware and content-based recommendations have a positive net effect on the total amount of money spent by users and the quantity of purchased items. The analysis did not confirm the net benefit effect on the average price of purchased items (see Models 4, 5 and 6 in Appendix 3).

It is important to remark that we measured the users' overall level of purchase rather than the purchase of "the recommended products". We could not limit the observation of purchasing behavior to the products which were both recommended and purchased because it was impossible to collect enough data (as explained in Section 3.3). For the sake of completeness, we report the mean number of products which were both recommended and purchased by users in each group in Table 11 (Appendix 2). This table reports the data about the products purchased during our study plus the next two months. We decided to extend the observation period to strengthen the validity of our analysis by providing additional supporting evidence over an extended time period. We performed a t-test for independent samples on these data and found that results are statistically significant. Further, we would like to point out that, given the extended period of time, the data presented in Table 11 is not strictly correlated to the data in Table 8 (see Appendix 2 for details). It would be interesting to demonstrate that using CARS may lead to higher purchase of "recommended products" and we should explore it further in our subsequent research.

6. CONCLUSION

There has been extensive interest developed in recommender systems (RSes) both in the industry and the academia, and many alternative approaches to providing different types of recommendations have been proposed over the last 15 years. Among them, Context-Aware Recommender Systems (CARSes) have received significant attention over the last few years (Adomavicius and Tuzhilin, 2011). Most of the work on CARSes has focused on demonstrating that the contextual information leads to more *accurate* recommendations and on developing efficient recommendation algorithms utilizing this additional contextual information. Little work has been done, however, on studying how much the contextual information affects the purchasing behavior of customers. In this paper, we went beyond accuracy and showed how context affects other crucial business-related metrics, such as expenditure of customers, quantity of purchased products, average price and levels of trust. We did it by conducting live controlled experiments with real customers in a real industrial setting of partnership with a major European publishing firm. In particular, a sample of firm's customers was split into three treatments, and each group was delivered a different kind of recommendations, namely contextual, content-based (un-contextual) and random over the period of 9 weeks. We measured accuracy and diversification of delivered recommendations, levels of trust and the business performance related to the purchasing behavior of the customers in those groups, namely quantity of products purchased, money spent and average price.

We first used these measures and the structured equation models to study how accuracy and diversification of delivered recommendations affect customer's trust and the business performance . We have shown that neither accuracy of recommendations nor diversification *alone* necessarily affect purchasing behavior of customers. The only effect we found is that accuracy affects the quantity of the items purchased. However, we also show that accuracy and diversification *together* affect customer trust which, in turn, is correlated to some of the business performance, namely quantity of products purchased and money spent. Therefore, to improve economic performance of RSes it is important to develop new methods that provide more accurate *and* diverse recommendations.

One group of such methods constitutes context-aware RSes (CARSes). This is the case because we show in the paper that CARSes provide more accurate and diverse recommendations than the

traditional content-based methods and that they also produce significantly more accurate albeit less diverse recommendations than random RSes. By combining this result with the previous one, we show that CARS dominate these two alternative recommendation methods not only in terms of the accuracy and diversification measures, but also in terms of the business-related metrics, such as trust and customer expenditure. Although we did not experiment with the algorithms based on the collaborative filtering (CF) approach, we expect that these results would be generalized to it as well. For the CF recommendations, it is well known that context increases accuracy (Adomavicius et al., 2005, Panniello et al., 2012). Regarding the hypothesis that CARS also improve diversity of recommendations for the CF systems, (Panniello et al., 2012) demonstrated that a collaborative filtering based CARS also improve diversity of recommendations, where diversity in (Panniello et al., 2012) is defined as in this paper. This implies that, even for the CF systems, CARS should improve both accuracy and diversity of recommendations. Using the argument made in this paper that accuracy and diversity affects trust, this implies that CARS should also affect customer trust for the CF-based recommender systems via increases in accuracy and diversity of recommendations arguments that should be empirically validated with real data to make them as "hard evidence"). Furthermore, once we show the increase in trust, the rest of the arguments of this paper should also follow for the CF systems.

The findings of this research have important managerial implications. As argued in the paper, it is important for recommender systems to collect and use the contextual data in most of the business applications, such as recommending movies, music, mobile and many other types of recommendations. Furthermore, we also argued that it is important to do this not only because CARS improve accuracy of recommendations, which was shown before, but also because they can contribute to the improvement of business performance measures for companies in several significant ways. Although this observation may appear obvious today, this was clearly not the case a few years ago when several leading companies, including Netflix, were slow in incorporating contextual information into their recommendation engines because it was not clear to them that such new features can improve business performance of their recommendation systems (personal communication discussions with a Netflix manager, October 23, 2009). The situation changed significantly only

recently when Netflix and others started using contextual information extensively in their recommendation engines (as Netflix's Research/Engineering Director Xavier Amatriain stated in his keynote speech at the ACM RecSys Conference in 2012)[3], and we expect that this trend will only grow in the future. In this respect, this paper provides academic evidence that contextual information contributes to the bottom line of these companies in significant ways. Therefore, we expect that this work will influence managers in the recommendation companies and will expedite the process of adopting CARS in their recommendation engines since the industry is currently thinking deeply about various ways to add contextual information to the existing recommendation engines.

One limitation of this research lies in that we have not proven causality between context and purchases but rather established a correlation between these factors. Another limitation is related to the collected data and the fact that the users rate the comic books before they read them in some cases. Although, as explained in the paper, it is a reasonable assumption for this particular application, it would be good to conduct new controlled experiments on this or other applications where ratings are elicited from the users only after they "consumed" the products.

These limitations and other considerations suggest the following possible future research directions. The first natural research direction would be an attempt to establish causality between context and purchases. Another interesting future research topic would be showing that using CARS may lead to higher purchase of "recommended products" and not only to higher overall purchase. Also, as a future work, we would like to test the results reported in this paper on other types of recommendation applications and for other types of industries. This should allow us to generalize and broaden our conclusions and perhaps identify additional factors affecting economic behavior and trust of customers besides the accuracy and diversification of recommendations studied in this paper. We would also like to conduct a bigger study involving more customers than we currently used. We would like to compare our results with different diversification techniques across a broader range of RSes than we used in this study in order to deeply test the trade-off between customer trust and recommendations diversification. Finally, an interesting future research topic would be the

[3] http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

investigation of different effects that context-aware recommendations can have on the number of items customers buy and their price.

References

Adomavicius, G. and Kwon, Y. O. 2012. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. Kn. Data Eng*. 24, 5, 896-911.

Adomavicius, G. and Tuzhilin, A. 2005. Towards the next generation of recommender systems: A survey of the state-of-the art and possible extensions. *IEEE Trans. Kn. Data Eng*. 17, 6, 734-749.

Adomavicius, G. and Tuzhilin, A. 2011. Context-Aware Recommender Systems. In: *Handbook on Recommender Systems*, Ed. Springer.

Adomavicius, G., Sankaranarayanan, R., Sen, S. and Tuzhilin, A. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Sys.* 23, 1, 103-145.

Amatriain, X. 2012. Building industrial-scale real-world recommender systems. In: Sixth ACM conference on Recommender systems, 7-8.

Amin, M. S., Yan, B., Sriram, S., Bhasin, A. and Posse. C. 2012. Social referral: leveraging network connections to deliver recommendations. In: Sixth ACM conference on Recommender systems, 273-276.

Barnes, J.W. 1994. Statistical Analysis for Engineers and Scientists. McGraw Hill, Singapore.

Baumol, W. E. and Ide, A. 1956. Variety in retailing. *Man. Sci.* 3, 1, 93-101.

Beldad, A., de Jong, M. and Steehouder, M. 2010. How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior* 26, 5, 857-869.

Bharati, P. and Chaudhury, A. 2004. An Empirical Investigation of Decision-Making Satisfaction in Web-Based Decision Support Systems. *Dec. Sup. Sys.* 37, 2, 187-197.

Bollen, D.G.F.M., Knijnenburg, B.P., Willemsen, M.C. and Graus, M.P. 2010. Understanding choice overload in recommender systems. In: The 4th ACM Conference on Recommender Systems (RecSys'10), 63–70.

Bollen, K. A. 1989. Structural equations with latent variables. Wiley Ed.

Brynjolfsson, E., Smith, M. D. and Hu, Y. 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. *Man. Sci.* 49, 11, 1580-1596.

Cooke, A. D. J., Sujan, H., Sujan, M. and Weitz, B. A. 2002. Marketing the unfamiliar: The role of context and item-specific information in electronic agent recommendations. *J. of Marketing Res.* 39, 4, 488-497.

Davis, F., Bagozzi, R. and Warshaw, P. 1989. User acceptance of computer technology: a comparison of two theoretical models. *Manag. Sci.* 35, 8, 982–1003.

Doney, P. M. and Cannon, J. P. 1997. An examination of the nature of trust in buyer–seller relationships. *J. Marketing*, 61, 2, 35-51.

Dourish, P. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing* 8, 1, 19-30.

Ekstrand, M. D., Ludwig, M., Kolb, J. and Riedl, J. 2011. LensKit: a modular recommender framework. In: Fifth ACM conference on Recommender systems,349-350.

Fishbein, M. and Ajzen, I. 1975. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley.

Fleder, D. and Hosanagar, K. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Man. Sci.* 55, 5, 697-712.

Flynn, L. J. 2006. Like This? You'll Hate That. New York Times, Jan. 23, 1.

Ganesan, S. 1994. Determinants of Long-Term Orientation in Buyer-Seller Relationship. Journal of Marketing (58), April, pp. 1-19.

Gefen, D., and Silver, M. 1999. Lessons Learned from the Successful Adoption of an ERP System, Proceedings of the 5th International Conference of the Decision Sciences Institute, Athens, Greece, pp. 1054-1057.

Gefen, D. 2002. Nurturing Clients' Trust to Encourage Engagement Success During the Customization of ERP Systems. *Omega* (30:4), pp. 287-299.

Gefen, D., Karahanna, E., Straub, D.W. 2003. Trust and TAM in Online Shopping: an Integrated Model, *MIS Quarterly* (27:1), pp. 51-90.

Gorgoglione, M., Panniello, U. and Tuzhilin, A. 2011. The effect of context-aware recommendations on customer purchasing behavior and trust. In: The fifth ACM conference on Recommender systems, 85-92.

Jarvenpaa, S. L., Knoll, K., and Leidner, D. E. 1998. Is Anybody Out There? Antecedents of Trust in Global Virtual Teams. *Journal of Management Information Systems* (14:4), pp. 29-64.

Hassenzahl, M. 2008. User experience (UX): towards an experiential perspective on product quality. In: 20th International Conference of the Association Francophone d'Interaction Homme-Machine, 11–15.

Hayes, C., Massa, P., Avesani, P. and Cunningham, P. 2002. An on-line evaluation framework for recommender systems. In: AH'2002 Workshop on Recommendation and Personalization in E-Commerce, 50–59.

Herbrich, R. 2012. Distributed, real-time bayesian learning in online services. In: Sixth ACM conference on Recommender systems, 203-204.

Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Sys.* 22, 1, 5-53.

Hess, T. J., Fuller, M. A. and Mathew, J. 2005. Involvement and Decision-Making Performance with a Decision Aid: The Influence of Social Multimedia, Gender, and Playfulness. *J. Man. Inf. Sys.* 22, 3, 15-54.

Hu, R. and Pu, P. 2011. Enhancing recommendation diversity with organization interfaces. In: the 16th International Conference on Intelligent user Interfaces, 347–350.

Kahn, B. and Lehmann D. R. 1991. Modeling choice among assortments. *J. Retailing* 67, 3, 274-299.

Knijnenburg, B. P. 2012. Conducting user experiments in recommender systems. In: The sixth ACM conference on Recommender systems, 3-4.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H. and Newell, C. 2012. Explaining the user experience of recommender systems. *User Model. User Adapt. Interact.* 22, 4-5, 441-504.

Koenigstein, N., Nice, N., Paquet, U. and Schleyen. N. 2012. The Xbox recommender system. In: Sixth ACM conference on Recommender systems, 281-284.

Komiak, S. and Benbasat, I. 2006. The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly* 30, 4, 941-960.

Kwon, O. and Kim, J. 2009. Concept lattices for visualizing and generating user profiles for context-aware service recommendations. *Exp. Sys. with Appl.* 36, 2, 1893-1902.

Lamere, P. B. 2012. I've got 10 million songs in my pocket: now what?. In: Sixth ACM conference on Recommender systems, 207-208.

Lempel, R. 2012. Recommendation challenges in web media settings. In: Sixth ACM conference on Recommender systems, 205-206.

Lenzini, G., Houten, Y.V., Huijsen, W. and Melenhorst, M. 2009. Shall I Trust a Recommendation? Towards an Evaluation of the Trustworthiness of Recommender Sites. In: ADBIS Workshop, 121-128.

Liang T. P., Lai H. J. and Ku Y. C. 2006. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings, J. *Man. Inf. Sys.* 23, 3, 45-70.

Lilien G. L. and Rangaswamy A. 2003. Marketing Engineering. Prentice Hall, Upper Saddle River, NJ, USA.

Liu, Q., Chen, T., Cai, J. and Yu, D. 2012. Enlister: baidu's recommender system for the biggest chinese Q&A website. In: Sixth ACM conference on Recommender systems, 285-288.

Malhotra N. K., Birks D. F. and Wills P. 2012. Market Research. Pearson, Harlow, UK.

Mayer, R. C., Davis, J. H. and Schoorman, F. D. 1995. An integrative model of organization trust. *Academy Man. Rev.* 20, 3, 709-734.

McDonald, D. and Dimmick, J. 2003. The conceptualization and measurement of diversity. Commun. Res. 30, 1, 60–79.

McKnight, D. H., Choudhury, V. and Kacmar, C. 2002. Developing and validating trust measures for e-commerce: an integrative typology. *Information System Research* 13, 3, 334-359.

Mcginty, L. and Smyth, B. 2003. On the role of diversity in conversational recommender systems. In: The Fifth International Conference on Case-Based Reasoning, 276-290.

McNee, S., Riedl, J. and Konstan, J. 2006. Making recommendations better: an analytic model for human-recommender interaction. In: 24th International Conference Human factors in computing systems, 1103–1108.

Mui, Y.Q. 2006. Wal-Mart Blames Web Site Incident on Employee's Error. Washington Post, Jan. 7.

Panniello, U., Tuzhilin, A., Gorgoglione, M., (2012). Comparing context-aware recommender systems in terms of accuracy and diversity. UMUAI, Special Issue on Context-aware recommender systems. 24, 1-2, 35-65.

Panniello, U., Tuzhilin, A., Gorgoglione, M., Palmisano C. and Pedone, A. 2009. Experimental comparison of pre- vs. post-filtering approaches in context-aware recommender systems. In: RecSys ’09, 265-268.

Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R. and Yin, F. 2010. Empirical Analysis of the Impact of Recommender Systems on Sales. *J. Man. Inf. Sys.* 27, 2, 159-188.

Pavlou, P., Liang, H., and Xue, Y. 2007. Understanding and mitigating uncertainty in online exchange relationships: a principal-agent perspective. *MIS Quarterly* (31:1), pp. 105-136.

Pazzani, M. J. and Billsus, D. 2007. Content-based recommendation systems. In: The adaptive web. Springer-Verlag, 325-341.

Pu, P., Chen, L. and Hu, R. 2011. A user-centric evaluation framework for recommender systems. In: The 5th ACM Conference on Recommender Systems, 23–27.

Pu, P. and Chen, L. 2006. Trust building with explanation interfaces. In: The 11th International Conference on Intelligent User Interfaces, 93–100.

Schafer, J. B., Konstan, J. A. and Riedl, J. 2001. E-commerce recommendation applications. *Data Min. Kn. Disc.* 5, 1, 115-153.

Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M. 2012. Methods and metrics for cold-start collaborative filtering. In: The 25th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 253-260.

Schoorman, F. D., Mayer, R. C. and Davis, J. H. 2007. An integrative model of organizational trust: Past, present, and future. *Academy Man. Rev.* 32, 2, 344-354.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell system Technical Journal*, 27, 379-423.

Shmueli, G. and Koppius, O. R. 2011. Predictive Analytics in Information Systems Research. *MISQ*, 35, 3, 553-572.

Simonson, I. 2005. Determinants of customers' responses to customized offers: conceptual framework and research propositions. *J. Marketing* 69, 1, 32-45.

Sinha, R. and Swearingen, K. 2001. Comparing Recommendations Made by Online Systems and Friends. In: The 2nd DELOS Workshop on Personalisation and Recommender Systems, 18-20.

Smyth, B., Coyle, M. and Briggs, P. 2012. HeyStaks: a real-world deployment of social search. In: Sixth ACM conference on Recommender systems, 289-292.

Smyth, B. and McClave, P. 2001. Similarity vs. diversity. In: The 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, 347–361.

Sundaresan, N. 2012. Recommender systems at the long tail. In: Fifth ACM conference on Recommender systems, 1-6.

Swearingen, K. and Sinha, R. 2001. Beyond Algorithms: An HCI Perspective on Recommender Systems. In: The ACM SIGIR Workshop on Recommender Systems,.

Swearingen, K. and Sinha, R. 2002. Interaction design for recommender systems. In: Designing Interactive Systems, 25–28.

Thompson, C. 2008. If You Liked This, You're Sure to Love That. The New York Times online, Nov. 21 (http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html?pagewanted=all&_r=0)

Vargas, S. and Castells, P. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In: The fifth ACM conference on Recommender systems, 109-116.

Venkatesh, V., Morris, M., Davis, G. and Davis, F. 2003. User acceptance of information technology: toward a unified view. *MIS Quarterly* 27, 3, 425–478.

Wang, W. and Benbasat, I. 2005. Trust In and Adoption of Online Recommendation Agents. *J. of the AIS*. 6, 3, 72-100.

Xiao B. and Benbasat I. 2007. E-commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly* 31, 1, 137-209.

Zhang, C., Agarwal, R. and Lucas, H. C. L. 2011. Personalized product recommendations, recommender systems, household production function, retailer learning, laboratory experiment, online product brokering. *MIS Quarterly* 35, 4, 859-881.

Zhang, M. and Hurley, N. 2009. Avoiding monotony: improving the diversity of recommendation lists. In: The 2009 ACM conference on Recommender systems, 123-130.

Ziegler, C., McNee, S.M., Konstan, J.A. and Lausen, G. 2005. Improving recommendation lists through topic diversification. In: The 14th International Conference on World Wide Web, 22–32.

Zins, A. and Bauernfeind, U. 2005. Explaining online purchase planning experiences with recommender websites. In: International Conference on Information and Communication Technologies in Tourism, 137–148.

Appendix 1. Final survey

Measure	**#**	**Question in the survey**
Check	Q_1	I usually trust people
Ability	Q_2	This personalized newsletter is like a real expert in assessing comic books
	Q_3	Personalized newsletters provided me with relevant recommendations
	Q_4	Personalized newsletters recommended comic books that I didn't know
	Q_5	I am willing to let this newsletter assist me in deciding which product to buy
Integrity	Q_6	The newsletter is reliable
	Q_7	I trust the personalized newsletter
Benevolence	Q_8	The company created the personalized newsletter to help me
	Q_9	The personalized newsletter is a service provided by the company to customers
Offline*	Q_{10}	I bought some of the recommended products offline
Price*	Q_{11}	I think the recommended products were expensive

*Not used in the analysis of results

Appendix 2. Additional experimental results

Table 9a reports the average answers to the final survey per each treatment group. Responses were given in a [1,5] scale (1 = totally disagree, 5 = totally agree). Statistically significant differences are marked with asterisks. Table 9b reports the results of analyses on Q_1 (Appendix 1). Table 10a report the results of the statistical tests run to check whether there are significant differences among the three treatments and the population in terms of demographic data (age and gender). Table 10b and Table 10c report the test results on the transactional information *before* the experiment (number of orders, purchased products, money spent and subscription year).

Table 10b and 10c show the average number of purchased items, orders and money spent *per user* before the experiment while Table 8 shows the quantity of purchased items, money spent and average price *per month per user* (both before and during the experiment). It was necessary to use the two different types of statistics (average price per month per user vs. average number of purchased items, orders and money spent per user) across these tables because these statistics were computed for different purposes. The statistics presented in Table 8 are consistent with those used in the statistical analyses described in the paper (see the Structural Equation Models in Section 4.2), while the ones reported in Table 10b and 10c were computed to check possible biases. Using the "user" as the unit of analysis in the test presented in Table 10b and Table 10c was an integral part of our analysis because the analysis aim was to check whether any bias in "customers" exist among different groups and the whole population. For this reason we computed the metrics *"per user"*. Similarly, it is important to use the "average price per month per user" statistics in Table 8 because the aim of this analysis was to show changes in customers' behavior over time among the three treatment groups. For this reason we computed the metrics *"per user per month"* in Table 8. In addition, data in Table 10b and 10c are referred to the purchasing users (i.e., those who purchased at least one product) in the three treatment groups. We used these data in order to demonstrate that there are no systematic differences in purchase behaviors of the purchasing users. We performed also the same analyses by taking all the users in the three groups (and not just the purchasing users) and we found the same results (i.e., no systematic differences in purchase behaviors of the users in the groups).

Table 9a. Results of the final survey: measures of trust (additional metrics are not displayed)

	ability				integrity		benevolence	
	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9
Content-based	3.231	3.404	2.731**	2.745	3.231	3.020**	3.500	4.118
CARS	3.169	3.458	3.056	3.141	3.417	3.222	3.694	4.028
Control	3.083	3.351	3.361**	3.200	3.611	3.472**	3.429	3.943

** Differences between values in the same column are statistically significant with $p<0.05$.

Table 9b. Results of the final survey: manipulation check and results of a t-test

	Q_1		Sign.
Content-based	2.889	Content-based vs. CARS	p=0.167233
CARS	3.125	Content-based vs. Control	p=0.494402
Control	3.054	CARS vs. Control	p=0.449062

Table 10a. Demographic data

	# of observations	avg. age (standard deviation)	gender		Difference in mean age (t-value)	Sign. gender
Content-based	90	28,48 (8.80)	10% female	Content-based vs. CARS	.63 (.483)	0.840
CARS	180	29,11 (10.76)	20% female	Content-based vs. Control	.68 (.500)	0.884
Control	90	27,80 (9.25)	18% female	CARS vs. Control	1.31 (.949)	0.947
Entire sample	24,284	27,84 (10.66)	17% female	Entire sample vs. content-based	.64 (.577)	0.876
				Entire sample vs. CARS	1.27 (1.591)	0.940
				Entire sample vs. Control	.04 (.039)	0.989

***Significant at p < .001; **significant at p < .01; *significant at p < .05

Table 10b. Transactional data (orders and items)

	# of observations	avg. orders (standard deviation)	avg. items (standard deviation)		Difference in mean Orders (t-value)	Difference in mean Items (t-value)
Content-based	26	3 (6.65)	22.76 (40.79)	Content-based vs. CARS	1 (.898)	11.41 (.863)
CARS	39	2 (1.71)	34.17 (58.51)	Content-based vs. Control	1.36 (.822)	0.94 (.873)
Control	17	1.64 (1.49)	13.82 (12.95)	CARS vs. Control	.36 (.734)	20.35 (1.412)
Entire sample	7,086	3.09 (3.22)	27.16 (41.32)	Entire sample vs. content-based	.09 (.151)	4.40 (.541)
				Entire sample vs. CARS	1.09 (2.122)	7.01 (1.055)
				Entire sample vs. Control	1.450 (1.853)	13.34 (1.330)

***Significant at p < .001; **significant at p < .01; *significant at p < .05

Table 10c. Transactional data (money spent and subscription year)

	# of observations	avg. money (standard error)	avg. year		Difference in mean Money (t-value)	Sign. year
Content-based	26	141.75 (275.76)	1.9	Content-based vs. CARS	40.76 (.494)	.990
CARS	39	182.51 (355.07)	2	Content-based vs. Control	55.75 (.765)	.996
Control	17	86.00 (145.38)	1.9	CARS vs. Control	96.51 (1.077)	.985
Entire sample	7,086	134.31 (222.87)	1.9	Entire sample vs. content-based	7.44 (.170)	.937
				Entire sample vs. CARS	48.20 (1.341)	.982
				Entire sample vs. Control	48.31 (.893)	.945

***Significant at p < .001; **significant at p < .01; *significant at p < .05

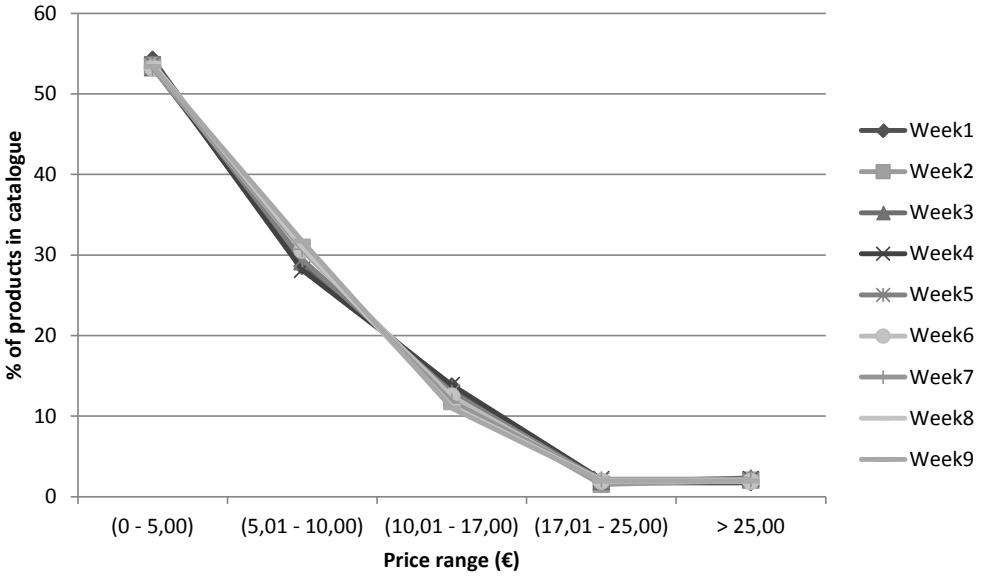

Figure 8. Price distributions of the products available at the online store in each week of the experiment

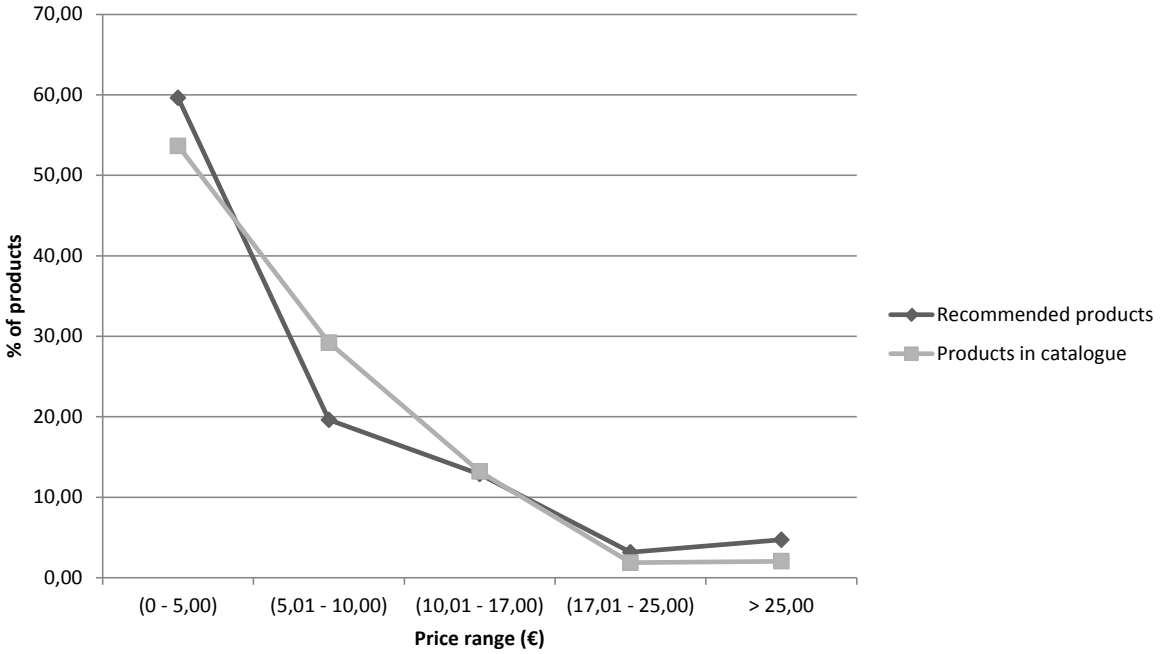

Figure 9. Distribution of price of recommended items vs. price of products available in the catalogue

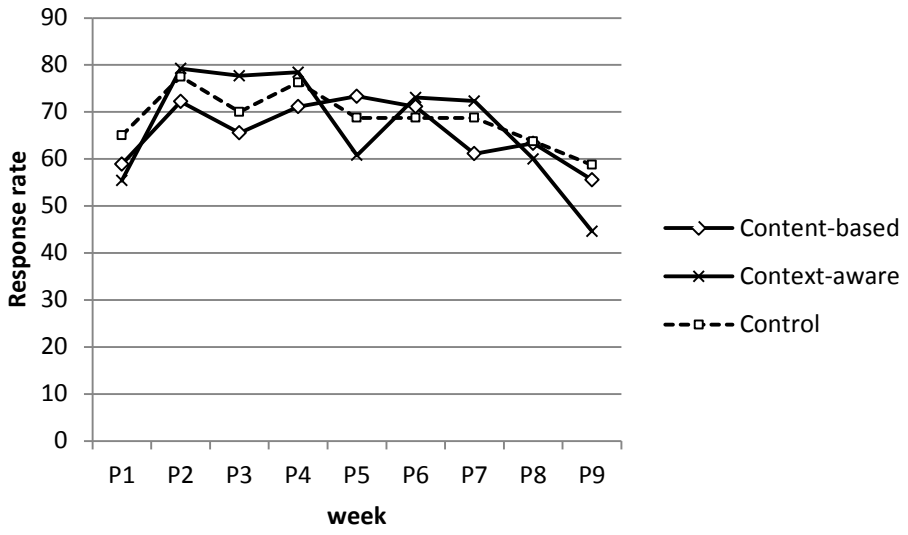

Figure 10. Distributions of response rate for the three treatment groups

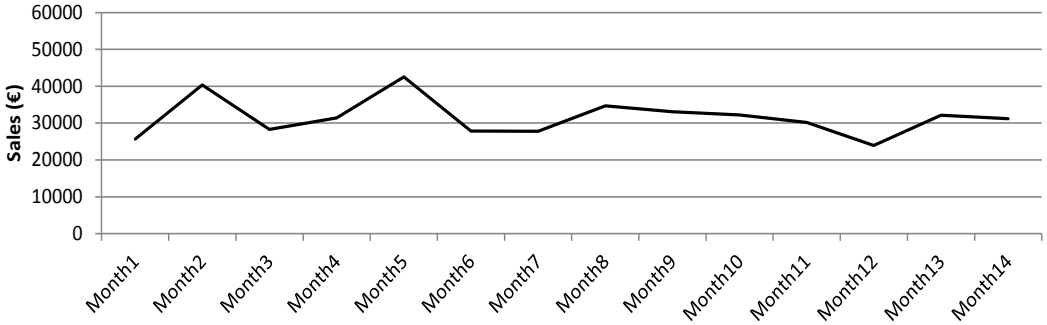

Figure 11. Distributions of sales over a period of 14 months before the experiment

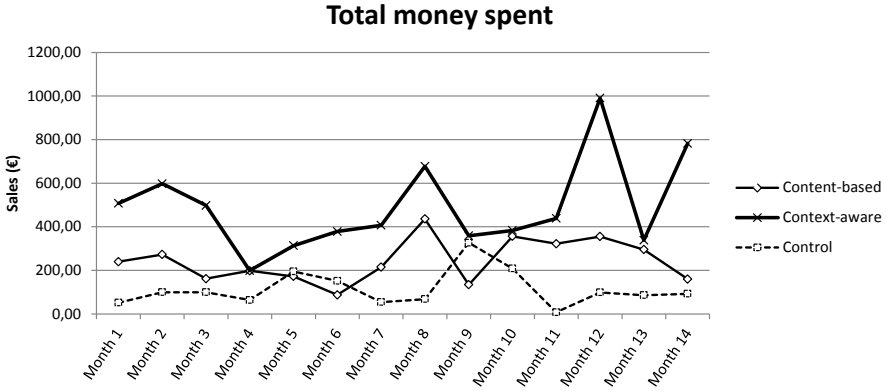

Figure 12. Distributions of sales (money spent) over a period of 14 months before the experiment for each group

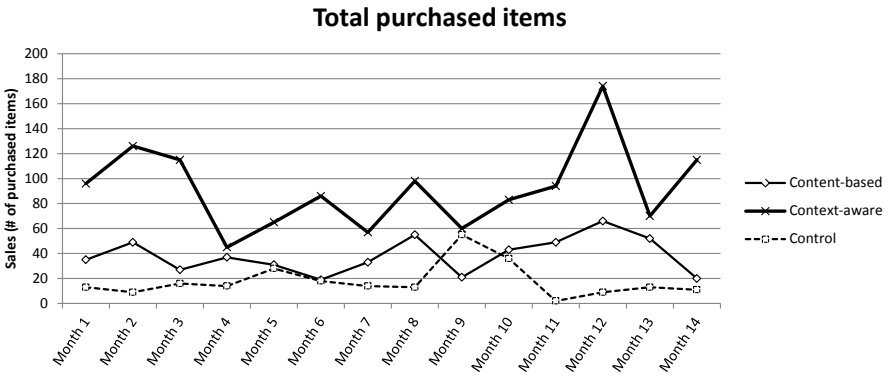

Figure 13. Distributions of sales (quantity) over a period of 14 months before the experiment for each group

Table 11. Mean number of products which were both recommended to customers and purchased by customers.

	Mean number of recommended and purchased items (standard deviation)		Difference in mean (t-value)
Content-based	3.71 (1.113)	Content-based vs. CARS	11.18 (2.835)*
CARS	14.89 (16.627)	Content-based vs. Control	2.38 (4.437)**
Control	1.33 (.577)	CARS vs. Control	13.56 (3.446)**

***Significant at p < .001; **significant at p < .01; *significant at p < .05

Table 11 reports data about the products purchased during our study (9 weeks) plus the next two months, thus having the total observational period of about 4 months. It is useful to remark that Table 8 (Section 5.2) also reports data pertaining to the users' purchasing behavior but collected over the period of 9 weeks. The numbers reported in Table 8 and Table 11 are computed in different ways because the tables were built for different purposes. The purpose of Table 11 is to strengthen the validity of our analyses by providing additional supporting evidence over an extended time period. Table 11 presents the real impact of the recommender system in terms of recommended and purchased items. Table 8 was built to show how the customer behavior changes during the experiment while our recommender system is being used. Therefore, Tables 8 and 11 are quite different and there is no strict correlation between them. Therefore, the information reported in the two tables cannot be directly compared because it refers to different observational periods.

Appendix 3. Detailed results of structural equation models and econometric models

Table 12, 13, 14, 15 and 16 report the detailed results of the SEM models used to test the five hypotheses in Section 4.2.

Table 12. Detailed SEM results for H1

Estimate Accuracy → Quantity	4.540 (2.024)*
p-value	.846
χ^2/df	.038
RMSEA	.000
p-value close to fit	.873
CFI	1
TLI	1
NFI	.999

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

Table 13. Detailed SEM results for H2

Estimate Diversification → Purchases	-7.431 (11.811)
p-value	.739
χ^2/df	.685
RMSEA	.000
p-value close to fit	.905
CFI	1
TLI	1
NFI	.996

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

Table 14. Detailed SEM results for H3

Estimate Trust → Purchase	.374 (.188)*
p-value	.254
χ^2/df	1.152
RMSEA	.031
p-value close to fit	.753
CFI	.996
TLI	.995
NFI	.972

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

Table 15. Detailed SEM results for H4

Estimate Diversification → Trust	3.146 (1.495)*
Estimate Accuracy → Trust	1.551 (.664)*
p-value	.453
χ^2/df	1.010
RMSEA	.008
p-value close to fit	.937
CFI	1
TLI	1
NFI	.971

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

Table 16. Detailed SEM results for H5

Estimate Accuracy → Trust	1.767 (.792)*
Estimate Diversification → Trust	3.541 (1.759)*
Estimate Trust → Purchase	.390 (.183)*
p-value	.360
χ^2/df	1.047
RMSEA	.017
p-value close to fit	.964
CFI	.998
TLI	.998
NFI	.964

***Significant at p < .001; **significant at p < .01; *significant at p < .05; standard errors are in parentheses.

Table 17, 18 and 19 report the detailed results of the econometric models used to confirm the validity of the results reported in Table 8 (Section 5.2).

We used the first specification (S1, see section 5.2) in order to study whether there are specific differential trends in the sales of different groups before the experiment. In particular, we built one econometric model for each variable representing purchases (money, quantity and price namely).

Model 1: $Money_{it} = \beta_i + \beta_1 Month + \beta_2 Month.CARS_i + \beta_3 Month.CONT_i + \varepsilon_{it}$

Model 2: $Quantity_{it} = \beta_i + \beta_1 Month + \beta_2 Month.CARS_i + \beta_3 Month.CONT_i + \varepsilon_{it}$

Model 3: $Price_{it} = \beta_i + \beta_1 Month + \beta_2 Month.CARS_i + \beta_3 Month.CONT_i + \varepsilon_{it}$

Table 17 reports the list and description of the variables used to build the econometric models.

Table 17. List and description of the variables used to build the econometric models.

Variable	Description
Dependent variables	
Money	Money spent by user i during month t
Quantity	Quantity of items purchased by user i during month t
Price	Average price of items purchased by user i during month t
Independent variables	
Month	Month t
$User_i$	One dummy variable for each single user
Post	Dummy variable equal to 1 if observation is during the experiment, while 0 if observation is before the experiment
CARS	Dummy variable equal to 1 if observation is pertaining an user in the context-aware treatment group, while 0 if observation is not pertaining an user in the context-aware treatment group.
CONT	Dummy variable equal to 1 if observation is pertaining an user in the content-based treatment group, while 0 if observation is not pertaining an user in the content-based treatment group.

Table 18 shows the results for the first three models.

Table 18. Results of the application of models 1, 2 and 3.

	Model 1 (money)	**Model 2 (quantity)**	**Model 3 (price)**
User	Included^{+}	Included^{+}	Included^{+}
Month	-.070 (.101)	-.012 (.017)	-.010 (.009)
CARS	.409 (1.01)	.175 (.174)	-.025 (.095)
CONT	.411 (1.11)	.121 (.192)	.107 (.105)
Constant	5.23 (1.27) ***	.836 (.218) ***	.510 (.119) ***
R-squared	0.0125	0.0130	0.0110
Prob > F	0.0005	0.0003	0.0015

***Significant at $p < .001$; **significant at $p < .01$; *significant at $p < .05$ $^{+}$ Not significant

The coefficients of the variables User and Month are not significant in each of the three models meaning that there are neither significant customer-related unobserved effects nor time-related unobserved effects that can cause the increase of purchases across customers. Finally, the coefficients of the CARS and CONT variables are not significant in all the three models meaning that there are no preexisting declining or increasing purchase trends between CARS and CONT customers.

After demonstrating that no preexisting trends or customer-, time-related bias affect results in Table 8, we used the second specification (S2, see section 5.2) in order to measure the net benefit of contextual and content-based recommendations after controlling for unobserved time related and customer specific unobserved factors. In particular, we built one econometric model for each variable representing purchases (money, quantity and price namely).

Model 4: $Money_{it} = \beta_i + \beta_1\ Post_t + \beta_2\ Post_t.CARS_i + \beta_3\ Post_t.CONT_i + \varepsilon_{it}$

Model 5: $Quantity_{it} = \beta_i + \beta_1\ Post_t + \beta_2\ Post_t.CARS_i + \beta_3\ Post_t.CONT_i + \varepsilon_{it}$

Model 6: $Price_{it} = \beta_i + \beta_1\ Post_t + \beta_2\ Post_t.CARS_i + \beta_3\ Post_t.CONT_i + \varepsilon_{it}$

In order to validate the results shown in Table 8 of the paper, we applied the aforementioned models on the data used in Table 8. The following table reports the results of these models.

Table 19. Results of the application of models 4, 5 and 6.

	Model 4 (money)	**Model 5 (quantity)**	**Model 6 (price)**
User	Included^{+}	included^{+}	included^{+}
Post	.308 (.150)	.015 (.053)	1.458 (1.227)
CARS	1.296 (.183)***	.232 (.065)***	-1.524 (1.503)
CONT	1.279 (.183)***	.266 (.065)***	-2.075 (1.503)
Constant	.772 (.150) ***	.115 (.053)*	7.081 (1.227) ***
R-squared	0.9721	0.9085	0.6336
Prob > F	0.0000	0.0002	0.3261

***Significant at p < .001; **significant at p < .01; *significant at p < .05; $^{+}$ Not significant

First, these models confirm, again, that no customer or time-related bias affects the results in Table8. Second, the models confirm most of the results in Table 8 of the paper. In particular, both CARS and CONT recommendations have a positive net effect on the total amount of money spent by users (CARS recommendations have a slightly higher effect than CONT recommendations). Moreover, both CARS and CONT recommendations have a positive net effect on the quantity of purchased items (CONT recommendations have a slightly higher effect). On the contrary, it was not possible to confirm the net benefit effect of CARS and CONT recommendations on the average price of purchased items (the model is not significant at all). Despite the latter result, we think that these models confirm our research general claim because the “money spent” by customers and the “quantity” of purchased products can represent the customers’ purchasing behavior quite well. The fact that we cannot confirm the relationship between the average price of the products purchased by customers and the kind of recommendations allows us to refine the research findings but does not imply to reject the hypotheses that a relationship exists between the purchasing behavior of customers and the kind of recommendations and that this relationship depend on trust.