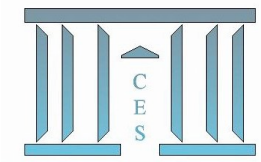
Hedonic Recommendations: An Econometric Application on Big Data

Okay Gunes

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers



Hedonic Recommendations: An Econometric Application on Big Data

Okay GUNES

2017.61

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://centredeconomiesorbonne.univ-paris1.fr/
ISSN : 1955-611X

Hedonic Recommendations: An Econometric Application on Big Data

Okay Gunes Ph.D. Economics

l'École Polytechnique, Palaiseau

Abstract

This work will demonstrate how economic theory can be applied to big data analysis. To do this, I propose two layers of machine learning that use econometric models introduced into a recommender system. The reason for doing so is to challenge traditional recommendation approaches. These approaches are inherently biased due to the fact that they ignore the final preference order for each individual and under-specify the interaction between the socio-economic characteristics of the participants and the characteristics of the commodities in question. In this respect, our hedonic recommendation approach proposes to first correct the internal preferences with respect to the tastes of each individual under the characteristics of given products. In the second layer, the relative preferences across participants are predicted by socio-economic characteristics. The robustness of the model is tested with the MovieLens (100k data consists of 943 users over 1682 movies) run by GroupLens. Our methodology shows the importance and the necessity of correcting the data set by using economic theory. This methodology can be applied for all recommender systems using ratings based on consumer decisions.

Key Words: Big Data, Python, R, Machine learning, Recommendation Engine, Econometrics
JEL: C01, C55, C80

Introduction

Conventional statistical and econometric techniques are often used to summarize, estimate, and test hypotheses in order to better analyze the nature of correlations between socioeconomic factors. Achieving better fitting results from a model determines the robustness of the answers that will be obtained from the dataset. However, supervised learning techniques such as machine learning are mostly concerned with predicting data and with the field of data mining in general. Data mining is also uses the summarization by finding interesting patterns in the data (Varian R.H, 2014). Econometricians tend to know what they are looking for before demanding the dataset. For big data analysis, the size of the data does matter, given that analysis is data-driven. Data science can essentially be considered tool-oriented research aiming to analyze large amounts of data, quickly and efficiently. It is the data scientist's decision to use convenient software and how fast to analyse the data. The main goal is to make the best possible prediction that will satisfy any need that is not necessarily based on a theory. Therefore, there is no reason to think that their aims are different, one to another. The primary purpose of economic modelling is to generate insights into complex

problems and to make estimations and predictions about rational agents' behavior (Varian, H.R., 1995). This is the task of the data scientist.

This paper will givean example of how theories regarding consumer decisions in economics can be applied to big data analysis. We chose the recommender system to demonstrate our theory, for two reasons. First we believe that recommending is purely about decision-making. Second, this study tests whether consumer behavior theories facilitate improvements to the quality of recommendations (see, Harper et al. 2005).

Recommender systems can be traced back to work in cognitive science, approximation theory, information retrieval, forecasting theories, management science and consumer choice modelling in marketing. Recommender systems emerged as an independent research area in the mid-1990s when they began to focus on recommendation problems that explicitly rely on the ratings structure. However, consumer decision theory was previously never used to improve the recommender systems. As it stands, there are three problems found within recommendation systems[1]. The first problem may be named "the measurement problem". The robustness of measuring the zero ratings of goods and services by convenient methods can be biased by the under-specification of interaction between the socio-economic characteristics of participants and the characteristics of the commodities. In fact, we cannot assume that the tastes of two individuals are the same simply because they have similar votes against other products[2]. The hypothesis is that the satisfaction gathered from the same goods and services would not be the same due to their order of preference, since similar individuals may have different tastes and socio-economic characteristics. This raises another problem later on, in that it could be argued that the observed rating values in the original data would no longer be the true. This is a kind of "rating inconsistency problem" within the data. The idea is that rating consumption of good X before consuming good Y could produce different results to consuming Y first and X subsequently. True satisfaction, as the true rating values, of an individual is supposed to be determined by the final preference order between X and Y. Given that a recommender system must consider the internal preference of each individual in order to obtain better recommendations for zero rating values[3]. Finally, recommendations can also be biased due to ignoring the recommendation produced from analysis of social interactions among other consumers.

[1] For discussions on the various limitations of the current recommendation methods see Adomavicius and Tuzhilin (2005).

[2] In fact, this problem refers to sample selection bias in econometrics, which is an issue pertaining to zero-declaration within the data set.

[3] Another criticism is that rating isn't independent from the moment of rating. For instance, the utility of consumption of a good X in the morning and in the evening would not be the same simply because the vector of the "intensity of needs" (for X with respect to the other goods) at these two moments would be different.

In this study, we propose to correct ratings with a two-step layer. First, we correct internal preferences with respect to the tastes of each individual for the characteristics of given products. These characteristics contain information about the goods or items. Bowbrick (1994) suggests that Lancaster's theory of quality in consumer demand (1966, 1971, 1975, 1979) is one of the most influential theories of consumer choice of quality in the economic literature, also very influential in marketing. To summarize, Lancaster suggests that households indeed choose the bundle of commodities that will maximize individual welfare by way of transforming these goods into characteristics. It is supposed that these characteristics became the source of utility but did not become goods, as such, that exist in the market. From a practical point of view, the Lancaster model allows us to measure how the characteristics of goods determine the value of this particular good with respect to other goods that have similar characteristics in the market.

In the second layer, we propose to make a second prediction regarding the corrected preferences from the first layer, this time using socio-economic characteristics across individuals. The hypothesis is that the recommendation occurs among individuals. The choice and preferences of individuals are also influenced by social networks, thus the ratings may influenced by other individuals' recommendations and rating values depending on criteria such as age, occupation, location etc.

The paper proceeds as follows: Section 1 briefly presents the theoretical model proposed by Lancaster. Section 2 derives the econometric specification of two-layer preference calibration. Section 3 introduces the MovieLens data set (100k) from a movie recommendation website, run by GroupLens, a research lab at the University of Minnesota. Section 4 reports the prediction results and compares them with existing methods. Section 5 concludes the paper.

1. Theory

Lancaster assumes that each consumption activity produces a fixed vector of characteristics (Z) having the n characteristics: 1,..,*n*. The individual possesses an ordinal utility function on characteristics U(*z*) and that he/she will choose a situation which maximizes U(*z*). U(*z*) is provisionally assumed to possess the ordinary convexity properties of a standard utility function. By supposing that there is a one-to-one correspondence between the vector of m goods (X): 1,…,*m* and vector of prices (P) and vector of income (K), we can write the consumer-choice program for *m*>*n* in the simpler form.

Thus, even if we have two perfectly matched individuals, the predicted values for zero rating values is biased due to a given need structure at the moment of consumption.

Max. U(*z*)
subject to $PX \leq K$
$x, z \geq 0$.

1.1 Consumption Technology

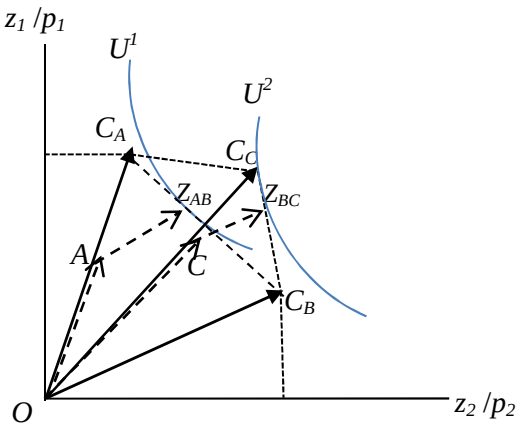
Individuals transform m good into characteristics by the vector relation **Z=BX** where B is (*m,n*) matrix. B is not a square matrix since *m>n*. B contains the *mn* coefficient a_{ij}. If a_{ij} is the quantity of characteristics *i* obtained by the consumption of one unit of good *j*, we can define C_{ij} as the amount of consumption in the form of inputs and c_{ij} is the output in the form of the amount of the characteristics *i* obtained from good *j* (see Simon, 1976).

$$c_{ij} = a_{ij} \times C_{ij}$$

The B matrix defines the consumption technology and is connected with the characteristics of the goods. The intuition is that the number of goods produces the same characteristics while one good may produce a different characteristic. B represents a set of homogeneous and linear consumption activities.

Given a price vector, this choice is a pure efficiency constraint *px=k*, and can determine a characteristics' frontier consisting of all *z* such that the value of the above program is just equal to *k*. There will be a determinate goods vector associated with each point of the characteristics' frontier. It is easy to see (see Figure 1) that the set of characteristics vectors in C-space that are preferred or indifferent to *z* transforms into a convex set in G-space if it is a convex set in C-space. Additionally, the efficiency criterion will be minimum cost.

Figure 1: Characteristics Space and Preferences



Consumption decisions depend on the characteristics of z_1 and z_2 by which it can be assumed that a feasible region of characteristics combinations is determined by the set Z and by prices p_1 and p_2 may be indicated in the characteristics space. According to the presentation of the Lancaster model given in Figure 1, the combinations that are ultimately chosen depend on the utility function and the indifference curves U_1, U_2 of the individual. C_A, C_B and C_C correspond to the maximum amount of characteristics that the consumer can buy the goods A, B and C respectively for given income budget. Let C_AC_B and $C_AC_BC_C$ are the two efficiency frontiers for given prices. The optimum is placed in Z_{AB} and the consumer buys the amount of A and B corresponding to the total characteristic vector $\overrightarrow{OA}$ and the $\overrightarrow{AZ_{AB}}$. Now suppose that the p_3, the price of C decreases. The consumer may decide to combine B and C by $\overrightarrow{0C}$ and $\overrightarrow{CZ_{BC}}$ and have a preference at ZB_C.

2. Two layer model

Let T, *y* and *x* be the target domain, outcome and feature respectively. Most of the prediction tasks are to estimate a predictor function $y:|R^n \rightarrow T$ from a real valued feature vector $x \in |R^n$. We assume there is a dataset D that consists of *n* pairs of *x* and *y* such that $D= \{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),...,(x^{(n)},y^{(n)})\}$. The common problem in recommendation is to deal with problems where almost all of the elements *y* for certain $x \epsilon D$ are zero. Cosine similarity is a another method which measures similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them (see Tan et al. 2005, Sidorov et.al. 2014 and Giller, 2012).

To show this zero-rating problem, we can use turn to survey data. MovieLens records i) which user $u \in U$ with users socio-demographic information $d \in D$ rating a movie (item) $i \in I$ with a rating $r \in \{1,2,3,4,5\}$.

- U ={Alain (A),Barbara (B) ,...}
- D={{Age of 32 (A32), Male(Am), Doctor (Ad)}, {Age of 50 (B50), Female(Bf), Artist (Ba)},...}
- I ={Toy Story (TS), Supermen (S), Star Wars (SW),...}

Let the merged data M be:

- M ={(A, A32, Am, Ad, TS, 5),(A, A32, Am, Ad, S, 2),(A, A32, Am, Ad, SW, 0); (B, B50, Bf, Ba, TS, 5),(B, B50, Bf, Ba, S, 2),(B, B50, Bf, Ba, SW, 2),…}

In this example, User A and B have similar ratings. However, A didn't watch SW and the problem remains to predict the rating of A for SW based on ratings of B.

- **First Layer**

Our interest is in knowing the preferences of A for non-zero ratings of A. Following the Lancaster model, we can argue that each item may share common characteristics and the user tries to optimize characteristics for a given budget, such as time and money.

- Z={TS={Animation, Adventure, Children's, Comedy}, S={ Action, Adventure, Fantasy}, SW={ Adventure, Fantasy, Sci-Fi },{…}}

Let the prices for the items be unitary. We also assume that the values of the items are measured by the ratings. Ratings give information about the level of satisfaction of each user after the consumption experience. We suppose that ratings are first dependent on the characteristics of the items which may be known before consumption. For a given budget of money and time, these characteristics determine the preorder preference among items. However, tags provide information about remembered utilities after consumption[4]. To show this, we have the following econometric hedonic function specification[5].

$$\ln r_{ui} = \sum_u a_{ui}.Z + \sum_{ui} a_{ui}.T + \sum_u a_{ui} D_{satisfaction} + u_{ui} \quad (1)$$

Where log *r* is the logarithm of the rating of user *u* for items *i*. Each item *i* shares common characteristics defined by vector Z with $z \in Z$. Using tags such as T is an important factor in better identifying the taste of users. The dummies from "the most important words" in tags of the items can be obtained through text mining techniques to measure common contextual words and tf-idf or Bm25 values. Furthermore, the characteristic vector can be augmented with other information sets such as the synopsis of each movie. The idea is that a drama and a comedy movie may share the same characteristic, such as the theme of love. $D_{satisfaction}$ is the relative satisfaction criterion that allows measurement of the fixed effect as the tastes of each individual. $D_{satisfaction}$={1 for each *u* and *i* having $\hat{r}_{ui} > \bar{\hat{r}}_{ui}$ and 0 for each *u* having $\hat{r}_{ui} < \bar{\hat{r}}_{ui}$ }. Where $\bar{\hat{r}}_{ui}$ is the average satisfaction of the user with respect to other users' ratings. The population regression function (1) is repeated for *n* user in

[4] The decision utility of a good or service could be different than experienced utility and remembered utility of this good or service. In fact, decision utility and experienced utility can be differentiating (Kahneman and Thaler 2006; Kahneman et al.,1997). Decision-utility (wanting or desiring) refers to a preference index describing how choices are made, as suggested by Bentham in 1789, while experienced utility refers to the measure of pleasure and pain after making a decision. In this respect, satisfaction and utility would be the same if and only if there is no sort of dissatisfaction stemming from the difference between these two utilities. Recommendation systems don't consider this specification problem in their rating analysis.

[5] The hedonic function is used by economist in order to identity the demand of good for given characteristics of goods. For the different application of hedonic function see Rosen, 1974; Bartik, 1987 ; Ohta and Griliches, 1986 ; Couton et al. 1996 and Pradier et al. 2016.

the data. The corrected rating of the users for zero-rating is finally obtained by the exponential of predicted $\hat{r}_u = \exp(\log \hat{r}_u)$.

- **Second Layer**

We can argue that social recommendation among consumers for the goods and services exists and that it depends on the socio-economic characteristics of the users, characteristics which may have in turn rearrange the preference ordering of items, hence ratings. More precisely, we suppose that the social recommendation information depends on basic socio-economic characteristics such as age, gender and occupation which determine whether or not opinions are shared through social interaction. The final true preference of users can be obtained by following the population regression function.

$$\ln \hat{r}_u = a_{u1} \ln(age) + a_{u2} \ln(age^2) + a_{u3} D_{occup} + a_{u4} D_{gender} + u_u \quad (2)$$

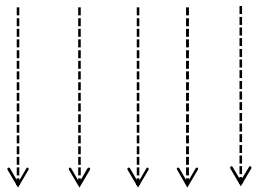
In this regression function (2), we use $ln\,\hat{r}_u$ from the first prediction of the equation (1). $\hat{r}$ is supposed to depends on age of the uses. We control the (linear or quadratic) change in tastes by age and squared age of users into the estimation function. In other words, the condition $\partial \ln \hat{r}_u / \partial \ln(age) = 0$ gives the form of change in tastes with respect to the age of users. The maximum rating for age can be computed when $f'_{age} > 0$ and $f''_{age} < 0$ (*vice versa* for the function, having a convex form: $f'_{age} > 0$ and $f''_{age} > 0$). D_{occup} represent the dummies for the occupation status and D_{gender} for gender of the users. True user ratings can therefore be computed by $\hat{\hat{r}} = \exp(\log \hat{\hat{r}})$.

Two step layers of the hedonic recommendation is summarized below

Input Layer

S = {(U, Age, Gender, Occupation, *I*, *r*)}

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{n1} & s_{nm} \end{bmatrix}$$

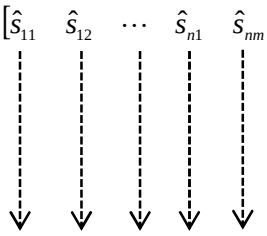


First layer (for each user)

$$\ln r_{ui} = \sum_u a_{ui}.Z + \sum_{ui} a_{ui}.T + \sum_u a_{ui} D_{satisfaction} + u_{ui}$$

$\hat{S}$ = {(U, Age, Gender, Occupation, *T*, *I*, $\hat{r}$))}

$$\begin{bmatrix} \hat{s}_{11} & \hat{s}_{12} & \cdots & \hat{s}_{n1} & \hat{s}_{nm} \end{bmatrix}$$



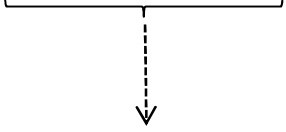
Second Layer (for all users)

$$\ln \hat{r}_u = a_{u1} \ln(age) + a_{u2} \ln(age^2) + a_{u3} D_{occup} + a_{u4} D_{gender} + u_u$$

Output

$\hat{\hat{S}}$ = {(U, Age, Gender, Occupation, T, *I*, $\hat{\hat{r}}$))}

$$\underbrace{\begin{bmatrix} \hat{\hat{r}}_{11} & \hat{\hat{r}}_{12} & \cdots & \hat{\hat{r}}_{n1} & \hat{\hat{r}}_{nm} \end{bmatrix}}$$



TEST Output

- Cosine(vector-based) similarity
- Singular Value decomposition (SVD) minimizing with Stochastic Gradient Descent
- Alternating Least Squares Method

3. Data

To test our methodology, we use the data from a movie recommendation website run by GroupLens, a research lab at the University of Minnesota. The full data set consists of 100 000 ratings by 943 users over 1682 movies. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. There is a tab-separated list consisting of user id, item id, rating, timestamp. The time stamps are unix seconds since 1/1/1970 UTC. Information about the items (movies); this is a tab separated list of movie id, movie title , release date, video release date, IMDb URL, unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western. The last 19 fields represent genres, a 1 indicates that the movie is of that genre, a 0 indicates it is not; movies can be listed under several genres at once. Demographic information about the users: this is a tab-separated list consisting of user id, age, gender, occupation and zip code.

3.1 Data Visualization

In this subsection, movie data is visualized in order to better understand the patterns of rating by age, gender occupational status of the participants.

The kernel distribution of the ratings given in Figure 2 shows that ratings 3 and 4 have the highest probabilistic distribution within the data.

Figure 2: Kernel Distribution of the Non-Zero Ratings

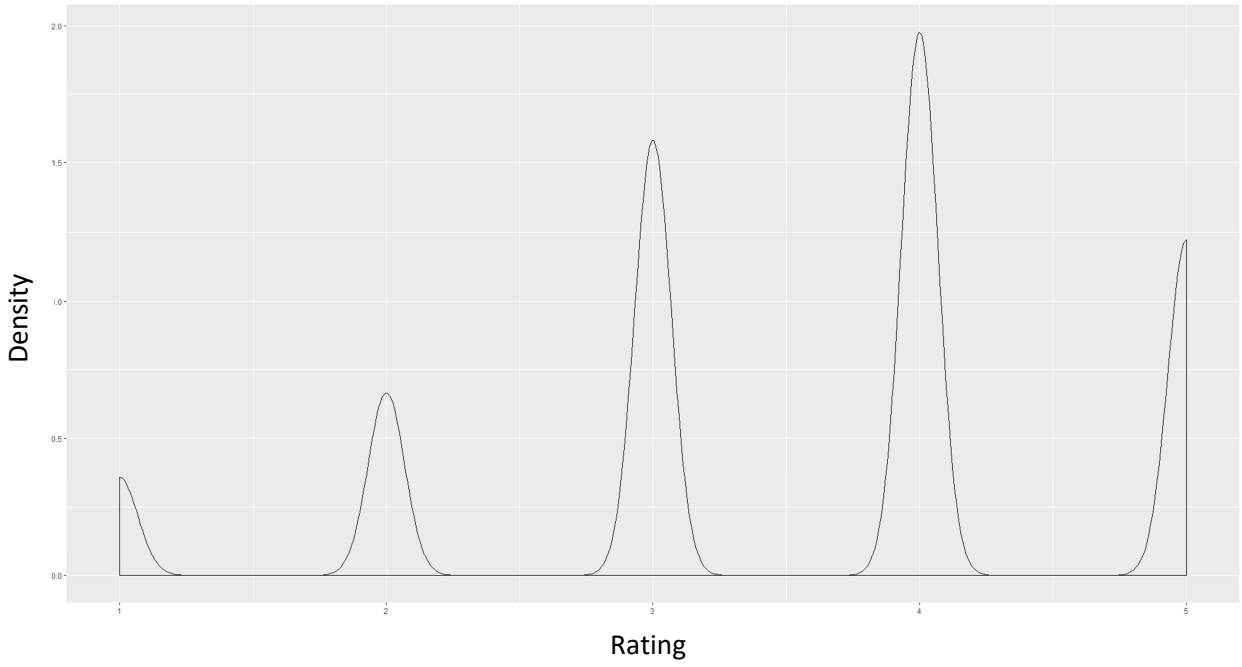
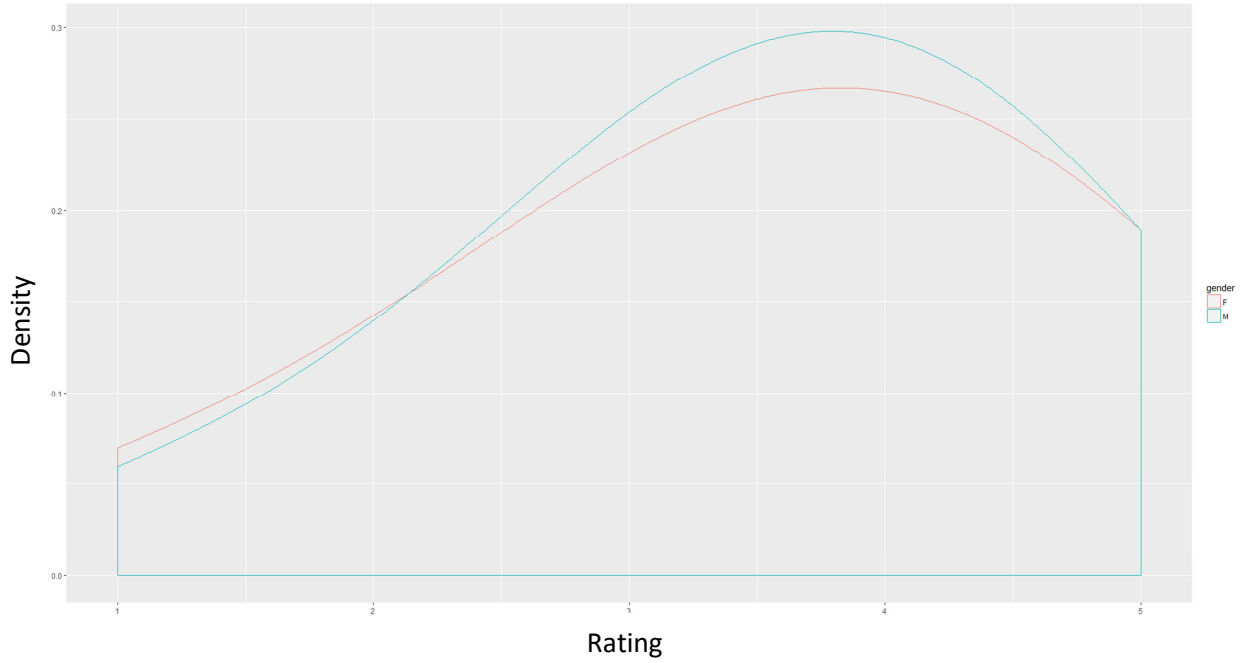


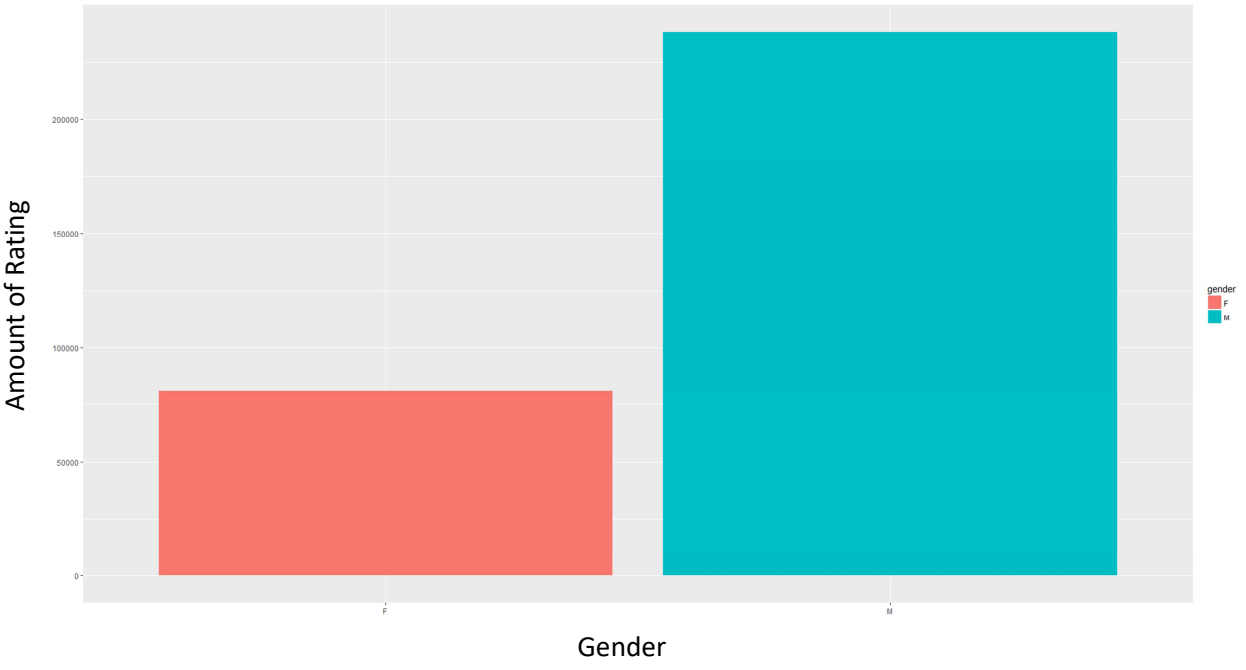
Figure 3 states that the distribution for the male population is higher after rating 2, indicating that the probability of voting for lower rates is relatively high for the female population.

Figure 3: Kernel Distribution of the Non-Zero Ratings by Gender



As it can be seen in figure 4 the tendency for participating in voting for the male population is higher than for the female population.

Figure 4: Rating by Gender



Conversely, when we look at the age distribution of votes by gender in Figure 5, it points out an intensive rating tendency between the age of 18 and 35. Figure 6 also confirms this tendency.

Figure 5: Rating by Age and Gender

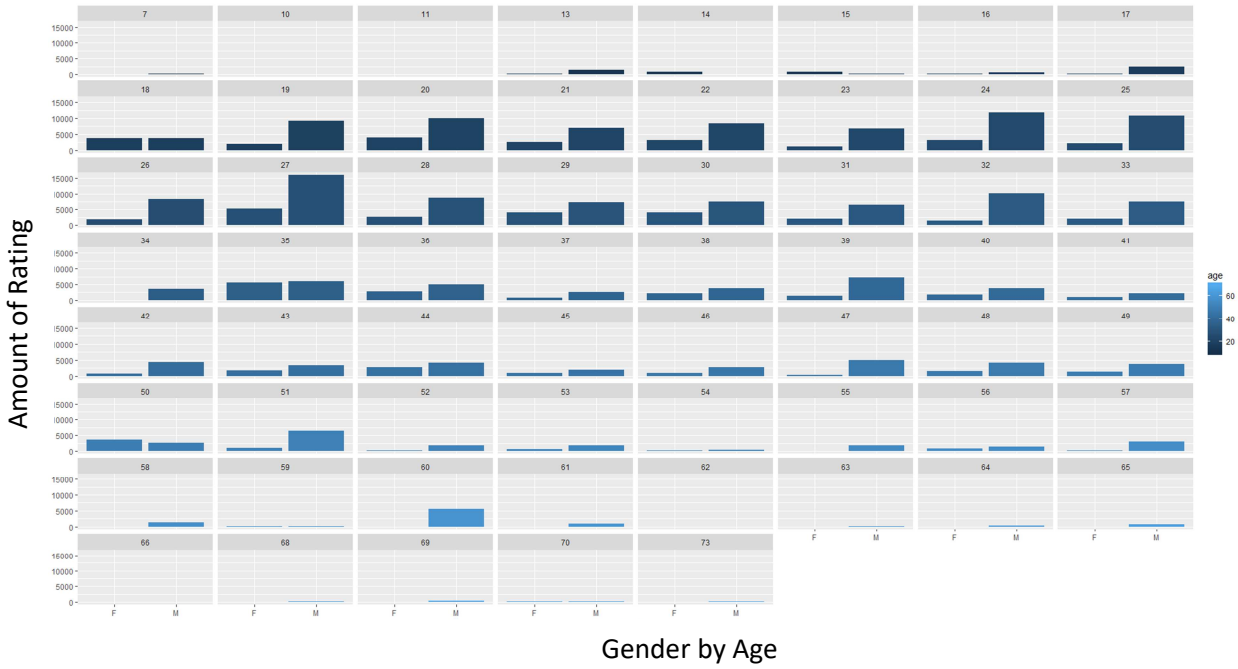
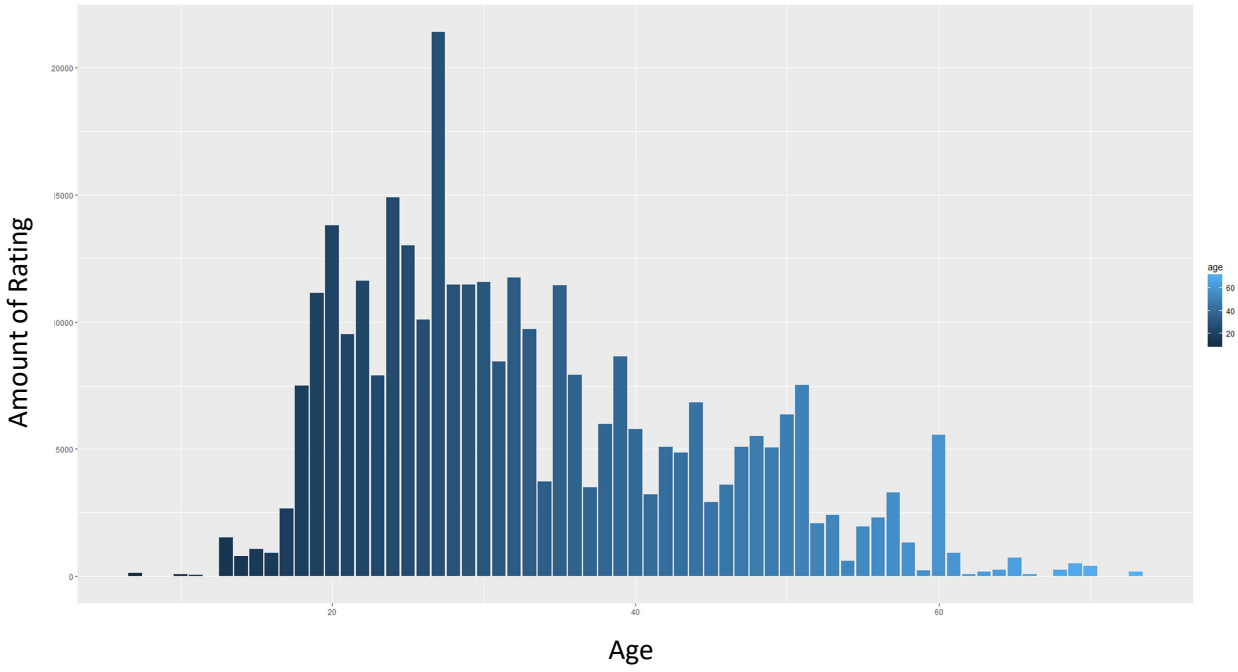


Figure 6: Rating by Age



When we look at the occupational distribution of ratings by gender from Figure 7, we can observe that the amount of ratings in the male population for the education, engineering, other,

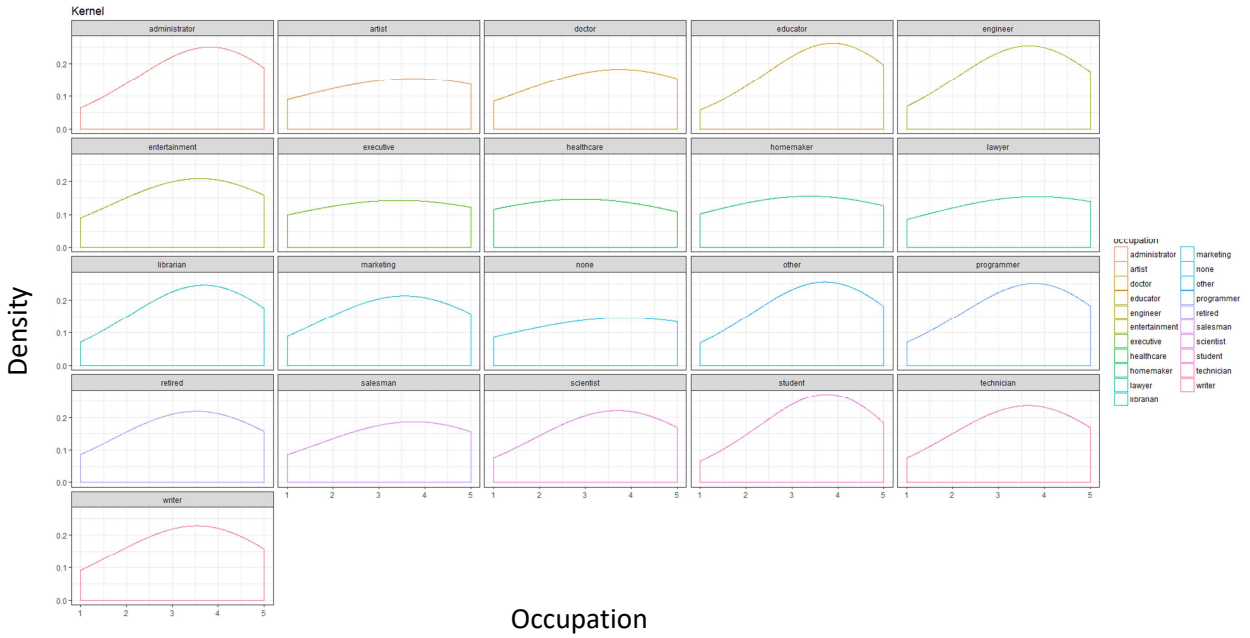
programmer, student is relatively higher than that of women. Therefore, the female population in healthcare and in library administration is more inclined to vote than the male population.

Figure 7: Rating by Age

Gender by Occupation

The kernel distribution of the ratings by occupational status is given in Figure 8. The probabilistic distribution results confirm weak distribution for artists, executives, healthcare professionals, homemakers, lawyers, and those without an occupation since the total amount of participants for these categories is low.

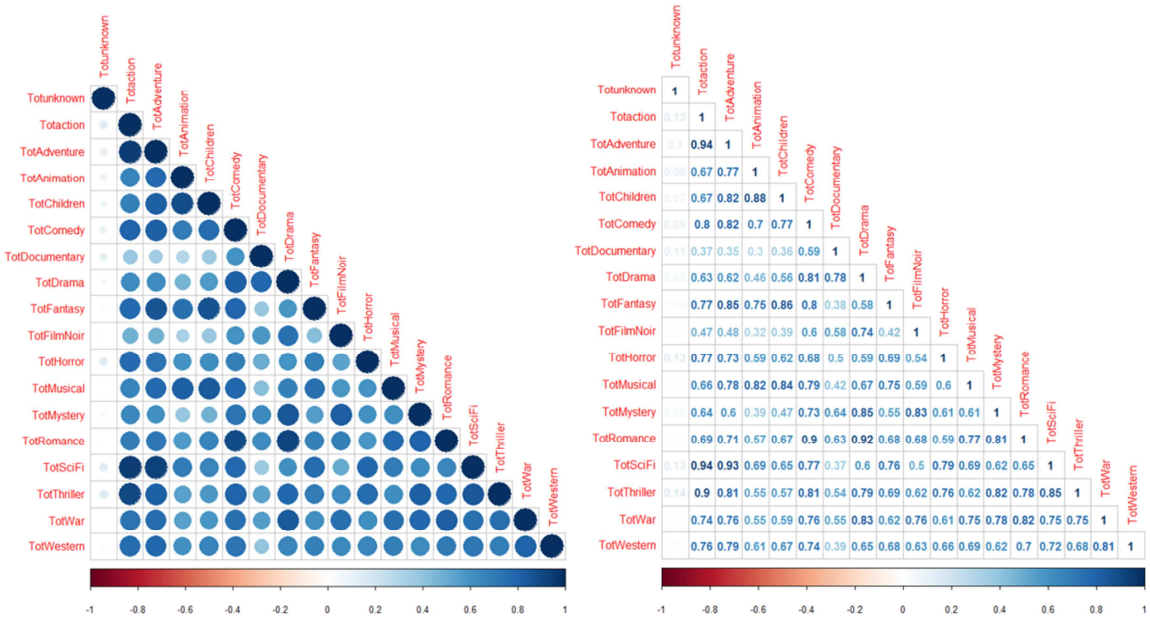
Figure 8: Kernel Distribution of the Non-Zero Ratings by Occupational Status



Occupation

Following the model description given in the theory, the correlation among the common characteristics of the items is shown in Figure 9. This descriptive analysis enables to better identify the shared characteristics of the movies in the data. For instance, the documentary is highly correlated with film noir, mystery, romance or comedy.

Figure 9: Total Characteristics Correlation



4. Results

Summary statistics of the prediction from the layers are given in the Table 1.

Table 1: Summary Statistics of the Predictions

	Log(r)	Log(r^)	Log(r^^)
count	90570	90570	90570
mean	1.190074	1.190074	1.143010
std	0.409568	0.242506	0.294705
min	0.000000	-0.553788	0.000000
25%	1.098612	1.087180	1.197483
50%	1.386294	1.222078	1.214059
75%	1.386294	1.347046	1.238191
max	1.609438	2.046746	1.378739

The summary statistics for the logarithm of the original rating prediction from the first layer ($\hat{r}$) and the re-prediction from the second layer ($\hat{\hat{r}}$) is shown in the underneath columns in Table 1. $\hat{r}$ is the prediction obtained from 943 regressions for each user for all movies. $\hat{\hat{r}}$ is the prediction using $\hat{r}$

including the zero ratings from the original rating r. We observe from Table 1 that the standard deviation is decreased by 28% between r and $\hat{r}$. Table 2 gives the estimation results for $\hat{r}$.

Table 2: Estimation Results for Second Layer ($\hat{r}$)

OLS Regression Results (Second Layer - for all users)			
Dep. Variable:	predicted	R-squared:	0.908
Model:	OLS	Adj. R-squared:	0.908
Method:	Least Squares	F-statistic:	4.053e+04
No. Observations:	90569	Prob (F-statistic)	0.00
Df Residuals:	90547	Log-Likelihood:	-37145.
Df Model:	22	AIC:	7.433e+04
Covariance Type:	nonrobust	BIC:	7.454e+04

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
occup_administrator	0.4484	0.015	29.889	0.000	0.419 0.478
occup_artist	0.4863	0.016	31.176	0.000	0.456 0.517
occup_educator	0.4445	0.015	29.163	0.000	0.415 0.474
occup_engineer	0.4554	0.015	31.375	0.000	0.427 0.484
occup_entertainment	0.4180	0.016	26.694	0.000	0.387 0.449
occup_executive	0.3171	0.016	20.342	0.000	0.287 0.348
occup_healthcare	0.1058	0.016	6.683	0.000	0.075 0.137
occup_doctor	0.5293	0.022	24.218	0.000	0.486 0.572
occup_lawyer	0.5299	0.017	30.474	0.000	0.496 0.564
occup_librarian	0.4471	0.015	29.751	0.000	0.418 0.477
occup_marketing	0.4235	0.017	25.417	0.000	0.391 0.456
occup_none	0.6057	0.018	33.895	0.000	0.571 0.641
occup_other	0.4607	0.014	32.799	0.000	0.433 0.488
occup_programmer	0.4568	0.014	31.840	0.000	0.429 0.485
occup_retired	0.3365	0.019	17.829	0.000	0.300 0.374
occup_salesman	0.4534	0.019	23.699	0.000	0.416 0.491
occup_scientist	0.5005	0.017	30.223	0.000	0.468 0.533
occup_student	0.5141	0.012	41.294	0.000	0.490 0.538
occup_technician	0.4693	0.015	31.223	0.000	0.440 0.499
occup_writer	0.3551	0.015	24.197	0.000	0.326 0.384
Dummy_Gender(M=1)	-0.0215	0.003	-7.016	0.000	-0.028 -0.015
log_age	0.0414	0.001	52.562	0.000	0.040 0.043
log_agesc	0.0827	0.002	52.562	0.000	0.080 0.086

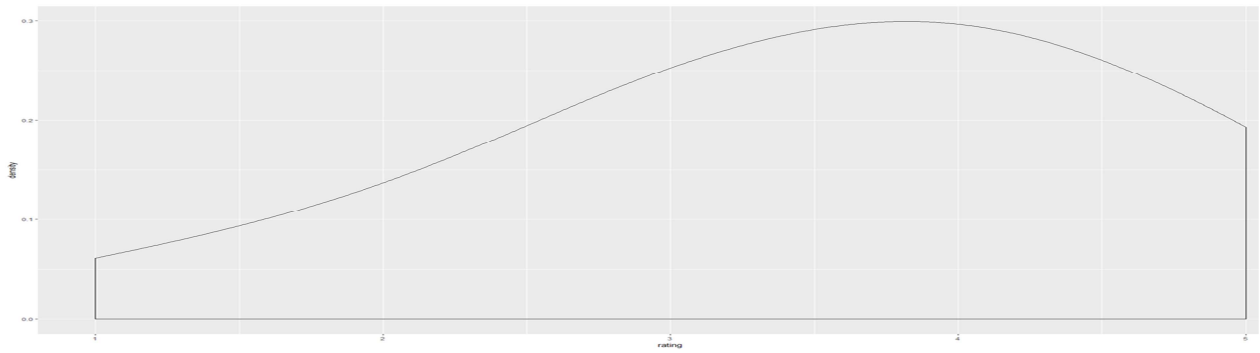
Omnibus:	26113.712	Durbin-Watson:	1.407
Prob(Omnibus):	0.000	Jarque-Bera (JB):	68555.148
Skew:	-1.570	Prob(JB):	0.00
Kurtosis:	5.881	Cond. No.	5.73e+15

The second layer is well estimated with all significant parameters in place. The residuals are normally distributed having a negative skew with a leptokurtic form. The autocorrelation in residuals is higher than 1 and lower than 3 implying that it is not a definite cause for concern. The ratings, especially for the doctors, lawyers, students, and scientists are relatively influenced by social interaction, depending on socio-economic characteristics. Therefore, the users with occupations in healthcare, as executives, as writers or who are retired have lower interactions within their social network, which in turn has a low effect on the recommendation.

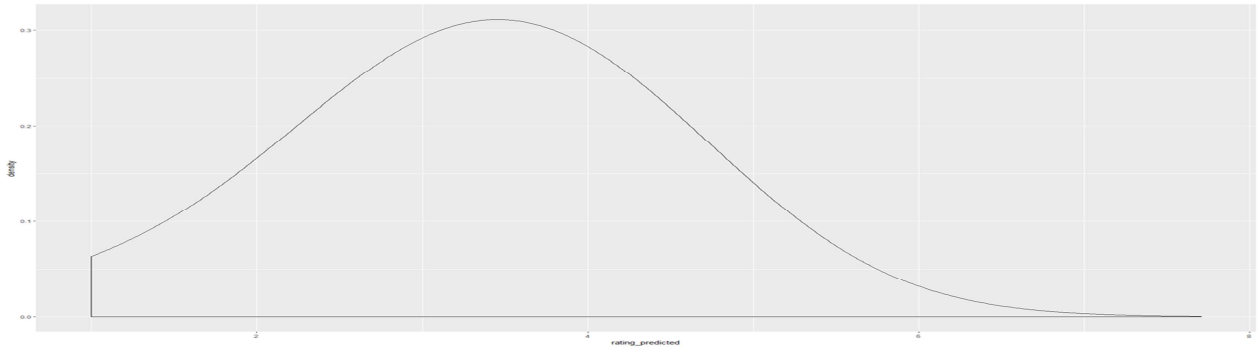
Figure 10 below shows the distribution of predictions from the layers and gives information about the changes in distribution when the inner preferences and social influences on the rating are taken into account.

Figure 10: Distribution of the Ratings.

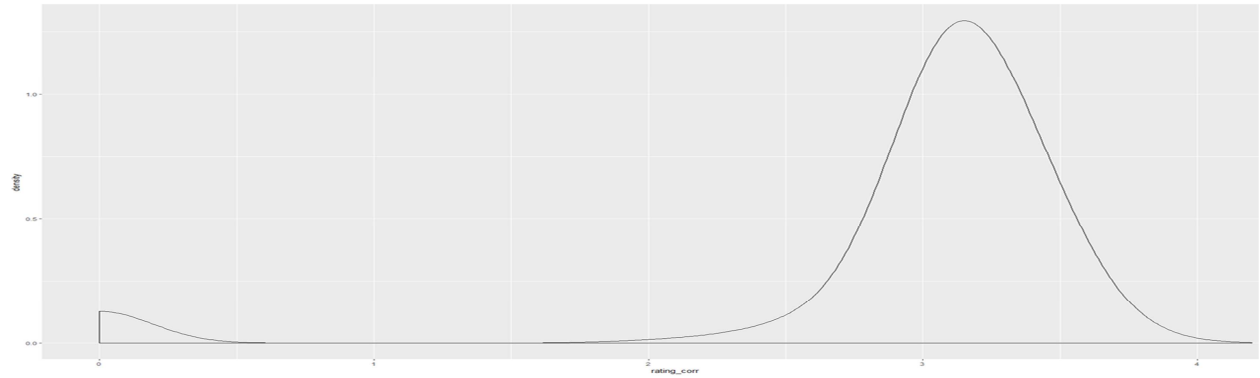
r_u



$\hat{r}_u$



$\hat{\hat{r}}_u$



The distribution for $\hat{r}_u$ has a smaller standard deviation with respect to the original non-zero ratings r_u. It had to be underlined that $\hat{r}_u$ is predicted for all films. After predicting $\hat{\hat{r}}_u$ for all movies

we replaced zero-rating movies from the original data. The true distribution $\hat{\hat{r}}_u$ has a lower mean and standard deviation than those for $\hat{r}_u$

5. Conclusion

None of the recommender system models deal with finding out the true preferences of consumers when the characteristics of goods and the consumers are being considered. In other words, it can be argued that observed ratings had to be corrected before using any recommender system. Any system must first know the internal preference structure of each consumer with respect to the common characteristics of the goods and services. Further, these internal preferences can also be determined by social interactions among consumers depending on their socio-economic characteristics. The theory is that recommendations are subject to social recommenders. In this work, we use the Lancaster theory in order to overcome the abovementioned problems. The robustness of our methodology is tested by comparing the results obtained by common recommender systems. They are listed in Table 3.

Table 3: Recommender Systems Results Comparison

Methodology	Originial Data		Corected Data	
	MAE	MSE	MAE	MSE
Cosine(vector-based) similarity	0,829	1,07	0,600	0,837
Singular Value decomposition (SVD)	2,634	8,396	1,908	4,562
Minimizing with Stochastic Gradient Descer	0,808	1,118	0,693	0,834
Alternating Least Squares Method	1,139	2,369	0,805	0,949

As it can be seen from Table 3, mean absolute error (MAE) and mean squared error (MSE) indicate that the corrected ratings give improved results for the recommendation. The main information that can be gleaned from our methodology is that the corrected references imply changes in the order of the movies already rated by consumers. This later yields better recommendations, since preferences are rearranged according to information circulated among consumers, based on their socioeconomic characteristics.

Our methodology shows the importance and the necessity of correcting the data set by using the economic theory. This methodology can be applied for all recommender systems using ratings based on consumer decisions.

6. References

Adomavicius G., and A. Tuzhilin, (2005), “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”, *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp. 734-749.

Bartik T.J., (1987), “The Estimation of Demand Parameters in Hedonic Price Models.” *Journal of Political Economy*, 95(1), pp. 81-88.

Bowbrick P., (1994), “A Refutation of the Characteristics Theory of Quality”. Available at: http://www.bowbrick.org.uk/Publications/Refuting Lancaster's Characteristics Theory.pdf (accessed: 28.04.2016).

Couton C., F. Gardes and Y. Thepaut, (1996), "Hedonic prices for environmental and safety characteristics and the Akerlof effect in the French car market," *Applied Economics Letters*, 3(7), pp. 435-440.

Graham L. G., (2012), "The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity". *Giller Investments Research Notes* (20121024/1). doi:10.2139/ssrn.2167044.

Harper F.M., X. Li , Y. Chen, J.A. Konstan, (2005), “An Economic Model of User Rating in an Online Recommender System”, *Lecture notes in computer science*, 3539-2005, pp. 307-316.

Kahneman D. and R.H. Thaler, (2006), “Anomalies: Utility Maximization and Experienced Utility”, *The Journal of Economic Perspectives*, 20(1), pp. 221-234.

Kahneman D., P.P. Wakker and S. Rakesh, (1997). “Back to Bentham? Explorations of Experienced Utility.” *Quarterly Journal of Economics*, 112(2), pp.375–405.

Lancaster, K.J., (1966), “A new approach to consumer theory”, *Journal of Political Economy*, 74, pp. 132-157.

Lancaster, K.J., (1971), *Consumer demand : a new approach*. Columbia University Press New York & London.

Lancaster, K.J., (1975), “Socially optimal product differentiation”, American Economic Review, 65(4), pp. 567-85.

Lancaster, K.J., (1979), *Variety, equity and efficiency*. Columbia studies in Economics No. 10. Columbia University Press, New York and Guildford.

Ohta, M. and Z. Griliches, (1986), "Automobile Prices and Quality: Did the Gasoline Price Increases Change Consumer Tastes in the U.S.?" *Journal of Business & Economic Statistics*, 4(2), pp. 187-198.

Tan P.-N., M. Steinbach and V. Kumar, *Introduction to Data Mining*, Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; p. 500

Pradier P., F. Gardes , X. Greffe and I. Mendoza, (2016), “Autographs and the global art market: the case of hedonic prices for French autographs (1960–2005)”, *Journal of Cultural Economics*, Springer Verlag, 2016, 40 , pp. 453-485.

Rosen, S., (1974), “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.” *Journal of Political Economy*, 82(2), pp. 34–55

Sidorov, G.; A. Gelbukh ; H. Gómez-Adorno, D. Pinto, (2014), "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model". *Computación y Sistemas*. 18 (3), pp. 491-504.

Simon, Y., (1976), «La nouvelle théorie de la demande : un panorama », *Vie et Sciences Economiques*, pp. 11-20.

Varian, H.R, (1995), *How to build an economic model in your spare time*. In Szenberg, M., ed.: Passion and Craft, How Economists Work. University of Michigan Press.

Varian, H.R, (2014), “Big Data: New Tricks for Econometrics”, *Journal of Economic Perspectives*, 28(2), pp. 3–28.