CUSTOMERS PROFILING BASED ON PSYCHOMETRIC CHARACTERISTICS

ПРОФАЙЛІНГ КОРИСТУВАЧІВ НА ОСНОВІ ПСИХОМЕТРИЧНИХ ХАРАКТЕРИСТИК

UDC 330.36.012.4

Hnot T.V.
Postgraduate Student
National University of Life
and Environmental Science of Ukraine

In today's world, it is important to know as much information about our customers as possible. Gender, age, occupation, family play an important role in this, but these characteristics don't answer deeper questions, like what customer feels when making a purchase, whether it makes sense to recommend her something, what is important for the customer: brand, quality or price. To answer these questions, we are proposing in this article to use psychometric characteristic of customers, which answer these and more other questions. Psychometric characteristics are extracted from textual information, written by the customer, and shopping patterns using state-of-the-art techniques in machine learning like XGBoost, Random Forest, LSTM models. Next, these characteristics were used to mine shopping preferences and advertisement preferences.
Key words: *machine learning, XGBoost, Random Forest, LSTM, OCEAN, profiling, psychometric/*

Сьогодні важливо знати якнайбільше інформації про наших користувачів. Стать, вік, вид діяльності, наявність сім'ї грають важливу роль, але цих характеристик недостатньо, щоб відповісти на більш глибокі питання. Наприклад, що користувач відчуває, роблячи покупки? Чи має сенс щось рекомендувати цьому користувачеві? Що важливо для користувача: бренд, якість чи ціна? Щоб відповісти на ці запитання, ми пропонуємо в цій статті використовувати психометричні характеристики користувачів, які дають відповіді на ці й на багато інших запитань. Психометричні характеристики було спрогнозовано з тексту, написаного користувачем, і патернів покупок, використовуючи найновіші техніки в машинному навчанні як XGBoost, Random Forest, LSTM моделі. Ці характеристики було використано для побудови торгівельних і рекламних уподобань.
Ключові слова: *машинне навчання, XGBoost, Random Forest, LSTM, OCEAN, профайлінг користувачів, психометрика.*

Сегодня важно знать как можно больше информации о наших пользователях. Пол, возраст, вид деятельности, наличие семьи играют важную роль, но этих характеристик недостаточно, чтобы ответить на более глубокие вопросы. Например, что пользователь чувствует, делая покупки? Имеет ли смысл что-то рекомендовать этому пользователю? Что важно для пользователя: бренд, качество или цена? Чтобы ответить на эти вопросы, мы предлагаем в этой статье использовать психометрические характеристики пользователей, которые дают ответы на эти и на многие другие вопросы. Психометрические характеристики было спрогнозировано из текста, написанного пользователем, и паттернов покупок, используя новейшие техники в машинном обучении как XGBoost, Random Forest, LSTM модели. Эти характеристики были использованы для построения торговых и рекламных предпочтений.
Ключевые слова: *машинное обучение, XGBoost, Random Forest, LSTM, OCEAN, профайлинг пользователей, психометрик.*

Introduction. OCEAN [1] stands for Big Five personalities traits or Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (Emotional range). These five traits often used to describe humans' personalities. This concept was developed in the 1980s and is widely used by psychologists to learn more about personalities and characterized them in some way.

Nowadays OCEAN concept could be used not just as a way to describe people in some psychological studies, but gives real value to the business. Knowing OCEAN of all customers, companies could personalized their approach to business, make more efficient recommendations, targeted messages and raise communication level with customers to new level.

One of the most significant examples, which demonstrates how OCEAN was used, is Trump' election campaign 2016. He hired data analytics company to support his campaign. This company, having information about millions of USA citizens, performed analytical highly targeted advertisement campaign to the most influenced electorate to confirm to vote for Trump.

So, OCEAN is a great tool we have, and in this article we will try to explain and show how we could get and use it in a retail business.

Problem statement. Having OCEAN scores of customers, we could detect extraverts or introverts within our customers and work with them as psychologist studies suggest. But how to calculate customer's OCEAN?

The most basic approach is to use surveys, which gives us an ability to calculate scores based on answers. It is pretty old approach and it works great, but the problem here is to persuade customer to fill in the form. It is not very comfortable, it takes time to answer 50–60 questions, so this approach is not the best one.

In the 2008–2009 a set of studies have appeared, which shows that OCEAN could be extracted from textual information, like tweets, Facebook posts, essays [2; 3; 4]. Having text data, written by the customer, we could analyze it and build prediction regression models to extract OCEAN scores from it. So, this approach requires having access to customers' social accounts or some textual information written by customers, like comments.

The last approach, which in details is described below, is built based on an idea to extract OCEAN from the bying patterns, like:

– whether customer prefers more expensive products or cheaper ones;

– customer always buys on weekend or weekdays;

– customer buys everything at once, or buys different categories of products at different time;

– customer likes everything she bought or dislikes everything and so on.

Research results.

Dataset for analysis. In our study we have done 3 main things:

1) Extract OCEAN scores based on comments, which customers have left on products they have bought;

2) Extract OCEAN scores based on transactional data and buying patterns;

3) Extract consumption preferences based on OCEAN.

Analysis was run on Amazon review dataset [5]. This dataset contains information about comments users have left on different products and products' metadata (like price, brand, category, etc.). For one of the customers (let's name him Edgar English), data sample is showed on Fig. 1.

In overall, this dataset contains information about almost 10,000 customers and 100,000 products. We have preprocessed it by removing some customers (outliers in a number of comments they left) and comments with less than 10 words. And also assumed that customers left comments to products they have bought (to be able to extract purchasing patterns).

OCEAN scores extracted based on textual data. As was mentioned before, OCEAN scores could be extracted using textual information, written by the customer. As we did not have exact mapping between text and five scores, we have used IBM Watson cognitive service [6] to generate labeled dataset. This service could predict personality characteristics through written text. Then LSTM [7] model was build based on this data.

Before feeding service with data, we left only last 200 words of each comment, as in a lot of cases first parts of comments were more related to products descriptions. Next, all customer's comments were split by 2400 words (large enough number to receive stable responses from service) and run through service. For customers with more than 2400 words in their comments, median values of API responses were taken to end up with final OCEAN scores.

For all customers, distributions of OCEAN scores are shown below on Fig. 2.

There are few possible reasons of so centered distributions of scores, returned by IBM API: these models were trained on skewed datasets; these models were trained on twitter text data, and here we are trying to use them for comments text data; Amazon dataset is skewed (for example, not all people like to leave comments).

To check whether service predicts not entirely random numbers, we have tested its accuracy using labeled text data of 250 users and their OCEAN, extracted based on surveys. Here we have two plots on Fig. 2: the first one shows how average correlations between real and extracted scores relate to a number of words, we send to the server; the second one shows a distribution of correlations between real and extracted scores. So, based on the first plot it is obvious, that bigger number of words yields more accurate predictions. The second plot shows that the most of observations are in the range from 0.7 to 1, that means that most of OCEAN scores we have predicted with high correlations. The average correlation between real and predicted scores for a minimum of 200 words is *0.57*.

Actually, if we compare real and extracted scores, we could observe the same patterns as with Amazon data. Predicted scores are more centered

Purchase history

Image	Title	Time	Rate	Review	Brand	Price
	Mr. Beams MB 980 Battery-Operated Indoor/Outdoor Motion-Sensing LED Ceiling Light, White	Sun Jun 15 2014	5	I use this in my bathroom above my shower. I had a leviton motion sensor installed in my bathroom, but I made the mistake of not counting for when the shower curtain was closed. So the light was always turning off on me.Rather than investing in a new system I went for this, I managed to get it on sale for $16. I have so say, this light is amazing. Not only did it save me a bunch of money on electric and a new lighting system but it actually stays on while I shower and it's bright enough! It lights up the whole shower, it's kinda like a spotlight. The light turns off after I leave the bathroom, so it works fine.It mounted perfectly to the ceiling, and seems secure.These are also great if the power goes out, one or two can light a large room. I will be investing in more for sure in the future.	Mr. Beams	23.94
	Aquis Exfoliating Back Scrubber	Sun Jun 08 2014	5	My first time with an exfoliating scrubber, and I have to say...why have I never had one before?I use it all over my body, it makes my skin so smooth and nice. This is a must have for anyone who is serious about skin care.	Aquis	10.49
	Oral-B Precision Black 7000 Rechargeable Electric Toothbrush 1 Count	Mon Jun 02 2014	1	Mine was dead on arrival. The battery compartment was broken, spent at least half an hour with a friend just to be sure I was not doing something stupid.Does not function at all.Also, the design to use the CHARGER to unscrew the battery compartment is a bad design. I will not be buying a replacement.		167.99
	Q-tips Cotton Swabs, 500 Count	Thu Apr 17 2014	5	I've been using Q-tips for years, the condensed cotton makes it much more easy to clean without coming apart - which I hate about other brands. I don't think I would use any other brand.	Q-Tips	1.35
	Maytex No More Mildew Shower Curtain Liner, Clear	Thu Apr 17 2014	5	I've had this shower curtain for a little over 5 months now, and still no mildew. However, I will be cleaning it very soon so it stays that way. If you clean the curtain at least twice a year you will find it lasts longer. If your washer has a sanitize function then you can just use that to clean it.It does as it should, keep water in the shower. No holes, no leaks, and of course... no rust.I also found that iron from the water does not stick to it like it did to the older one I had. Therefore it's not turning orange either, which I'm grateful for! Overall it's a great buy and I'm very happy with it and would buy it again, or even gift it!	MAYTEX	16.99

Previous 1 2 3 4 5 ... 11 Next

Fig. 1. Data sample for one of the customers

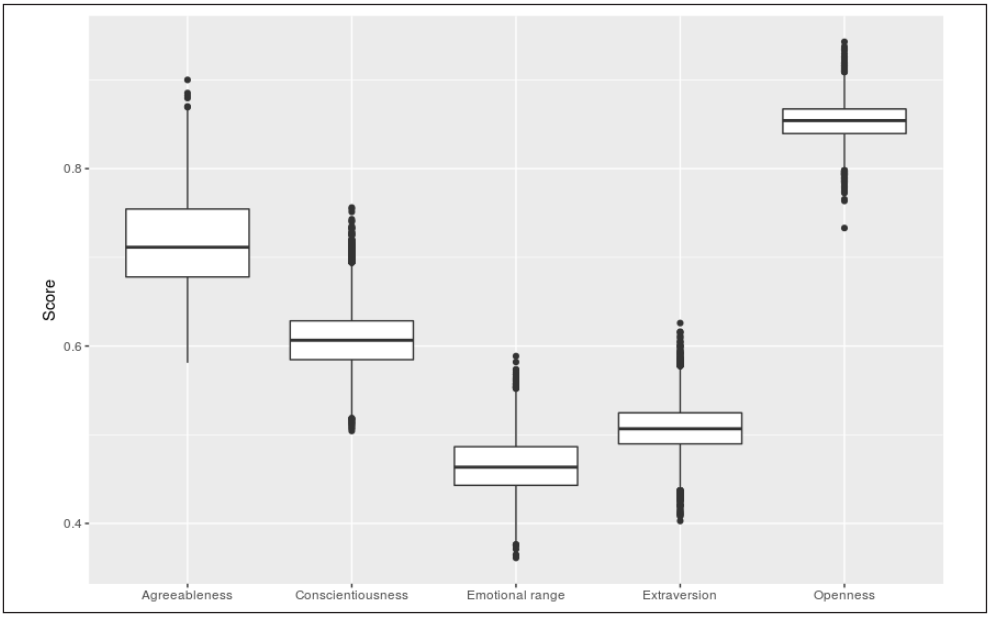

Fig. 2. Distribution of OCEAN scores

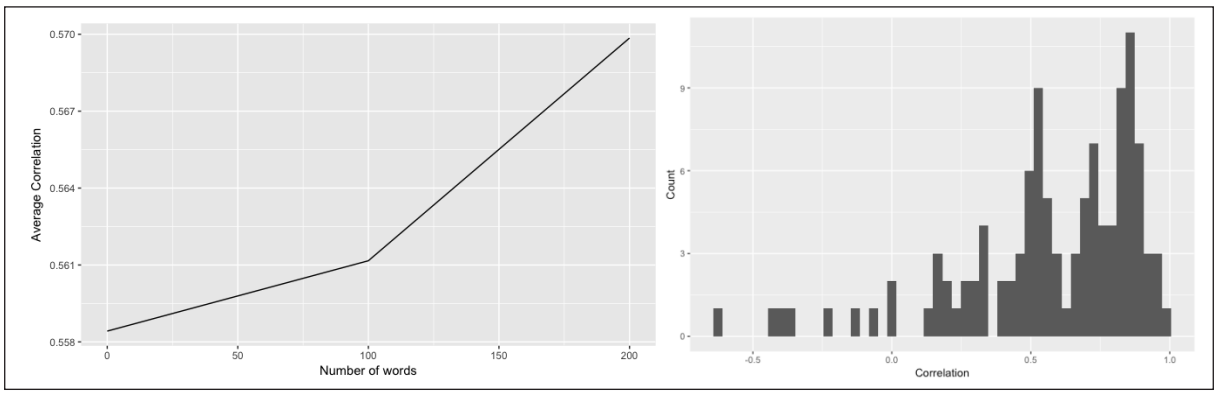

Fig. 3. Correlations between service output and real scores

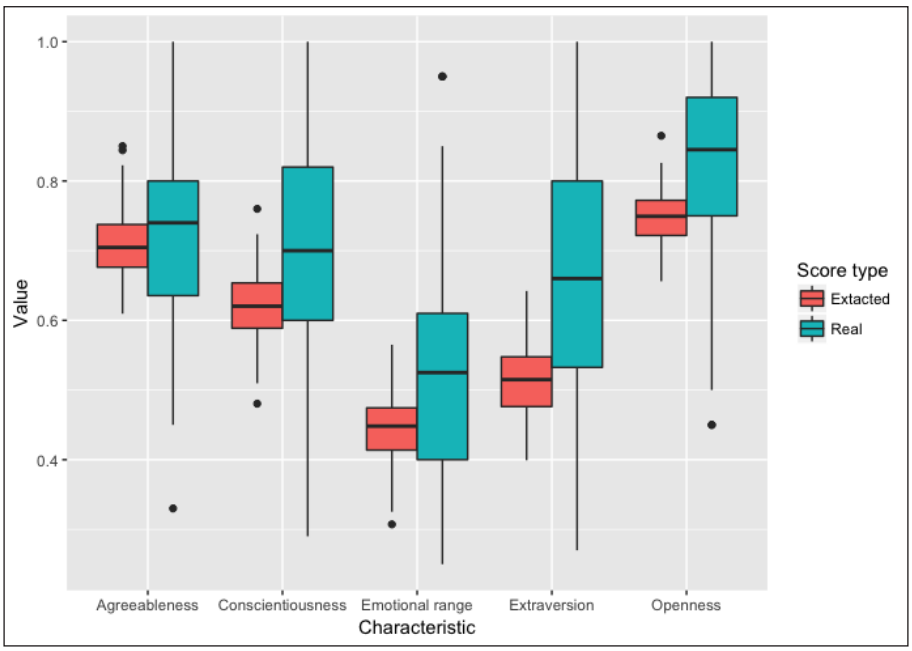

Fig. 4. Distributions of real and extracted scores

in comparison with real ones. So, probably some skewness is in service models. But, as these scores show real patterns, and we need some labeled data for transactional models, we continued analysis with them.

For Edgar English, extracted scores are shown using spider chart on Fig. 5. He has pretty high relative conscientiousness (he is efficient and organized) and low extraversion (he is reserved, reflective personality).

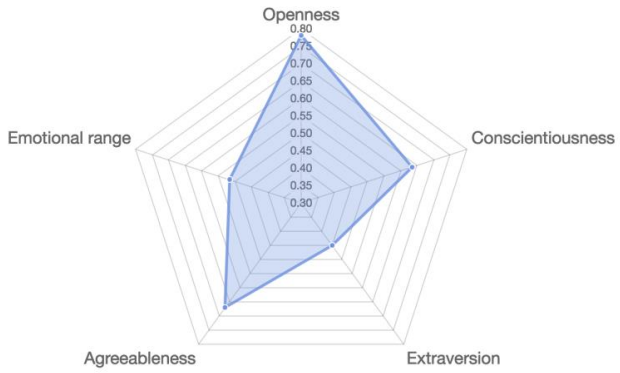

Fig. 5. OCEAN scores for one of the customer

Based on OCEAN scores we have generated set of consumption preferences, which could describe customer. There is wide range of preferences types, which could be extracted, like *purchasing portrait*:

Preference	Score
Likely to be sensitive to ownership cost when buying automobiles	High
Likely to prefer safety when buying automobiles	Low
Likely to prefer quality when buying clothes	High
Likely to prefer style when buying clothes	Low
Likely to prefer comfort when buying clothes	High
Likely to be influenced by brand name when making product purchases	Low
Likely to be influenced by product utility when making product purchases	High
Likely to be influenced by online ads when making product purchases	Low
Likely to be influenced by social media when making product purchases	Low
Likely to be influenced by family when making product purchases	Low
Likely to indulge in spur of the moment purchases	Low
Likely to prefer using credit cards for shopping	High

or *reading preferences*:

Preference	Score
Likely to read often	High
Likely to read entertainment magazines	Low
Likely to read non-fiction books	High
Likely to read financial investment books	Medium
Likely to read autobiographical books	High

OCEAN score extracted based on transactional data. In the previous part, we have shown how OCEAN scores could be extracted from textual information, written by the customer. But in a lot of cases, not all customers leave comments or there is no access to their social profiles. To deal with such situations, we have trained models to predict OCEAN based on transactions data and shopping patterns. Here we will demonstrate these models based on *Random Forest* algorithm, as it could be easily trained and tuned with sufficiently high accuracy.

The first model, *customer bought history model*, uses tf-idf representation of categories/tags of products, which customer has bought.

One product could be described by one or few tags. As in the example on Fig. 6, the toy is described by 4 tags:

Toys & Games
Tricycles
Scooters & Wagons
Ride-On Toys

Fig. 6. Tags for product

Tag “Tricycles” is in 19 products in our demonstration dataset (out of 100 000), so it is a good descriptor of a product.

Fig. 7. Products with tag “Tricycles”

On the other hand, “Toys & Games” tag is presented in 1040 products, so it is not unique and the value of feature, represented by this tag, would be lower.

In overall, Amazon review dataset has 5070 unique tags (categories) for products, so each customer could be represented by 5070 tf-idf features vector of tags of products, which he has bought.

Next, we have built prediction models to predict 5 OCEAN scores based on these features, and received accuracy, showed in Table 1.

Table 1

Accuracies for “Customer bought history” models

Characteristic	Mean RMSE	Test RMSE	Test R2
Openness	0.021	0.016	0.465
Conscientiousness	0.035	0.027	0.371
Extraversion	0.028	0.025	0.206
Agreeableness	0.050	0.027	0.712
Emotional range	0.032	0.022	0.526

“Mean RMSE” column shows average error in case prediction of mean value for each customer. It is just some base. As all models shows RMSE less

than mean, we have catched some patterns in data, related to OCEAN.

Aggreeableness and *Emotional range* could be predicted with the highest accuracy.

Next model in our ensamble – *likes model.* Having access to rates, which people have left, we have used these numbers as feature vectors. Usually, these vectors will be very sparse and huge, for example our demonstative dataset has 100 000 products, and each vector would be 100 000-value vector. As it is hard to train model with so huge vectors, we decreased dimensionality to 500 with PCA.

Test accuracies for models have increased in comparison with previus model, what tells us that customers' rates contains more information about OCEAN, than simple tags vectors.

Table 2

Accuracies for "Likes" models

Characteristic	Mean RMSE	Test RMSE	Test R^2
Openness	0.021	0.016	0.479
Conscientiousness	0.035	0.026	0.453
Extraversion	0.028	0.024	0.265
Agreeableness	0.050	0.025	0.765
Emotional range	0.032	0.020	0.621

Last model in the ensemble – buying patterns model. This model is trained based on manually generated buying patterns features. Sample list of features, which are used in our example model with Amazon comments is in Table 3.

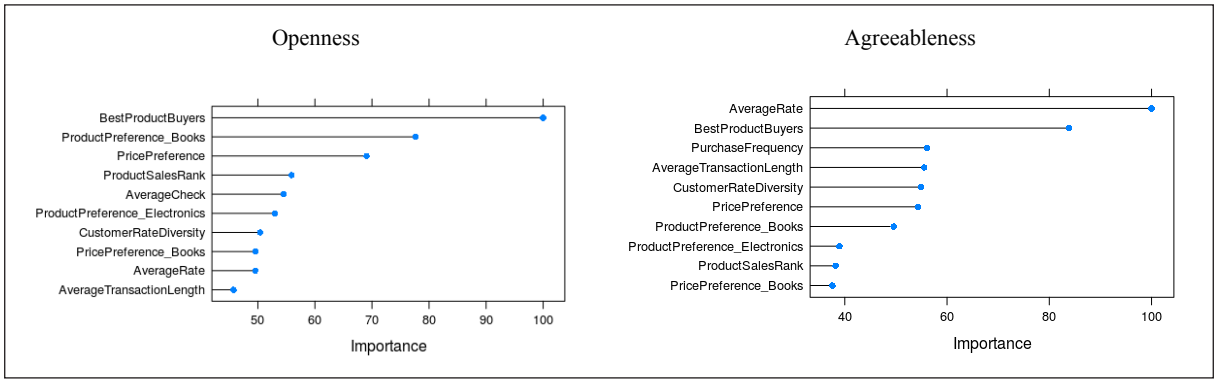

Fig. 8. Features importance for Openess and Agreeableness

Table 3

Generated list of features

Category	Feature	Description
Brand	BrandStability	Coefficient of brand change within different categories
Brand	BrandPreference	Whether customer prefers one brand or brand doesn't matter for customer
Price	PricePreference	if 1 – customer always buys the most expensive products, if 0 – the cheapest
Price	PricePreference_Category	the same as PricePreference, but detailed to each category
Products	ProductPreference_Category	Proportion of products bought from each category
Products	ProductSalesRank	Shows whether user buys product with high sales rank or not
Rate	AverageRate	Shows whether customer likes everything or dislikes everything
Rate	BestProductBuyers	Shows average rates of products customer bought (excluding his own rates)
Rate	CustomerRateDiversity	Shows whether customer opinion the same as opinions of different customers
Time	PurchaseFrequency	Average time between transactions
Time	WeekendBuyer	Proportion of weekend transactions
Time	NighBuyer	Proportions of transactions after 8PM till 6AM
Transaction	AverageTransactionLength	Shows average number of items per transaction
Transaction	AverageCheck	Average sum spent by user per transaction
Transaction	OneItemBuyer	1 if user always buys only one category per transaction, 0 if categories are different in one transaction
Products	PromotionApplied	Shows if promotions/discount codes were applied
Products	ProductSale	Customer buys products on sale

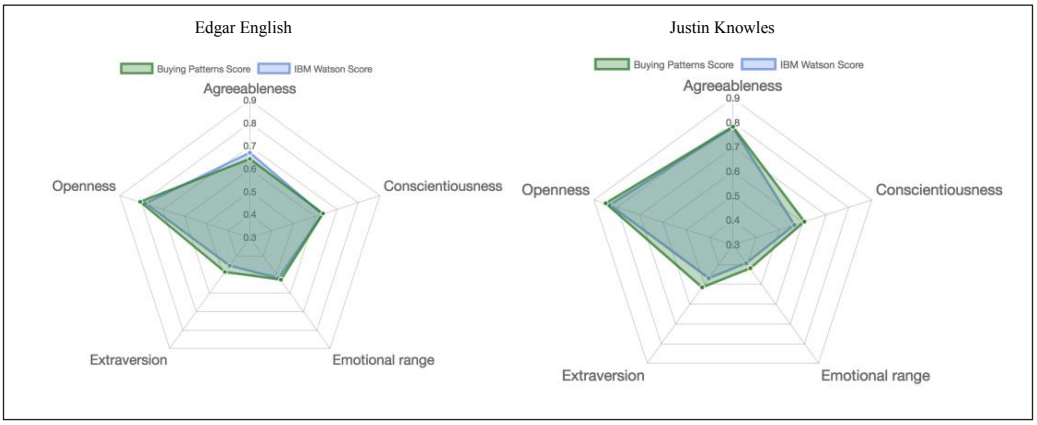

Fig. 9. Comparison of OCEAN scores based on different models

The most important feature for *Openness* prediction is BestProductBuyers, where for *Agreeableness* is AverageRate. In Amazon dataset, films and books products rates play very important roles. This is very interesting and tells us that actually what people like to read and watch tells us a lot about them.

Accuracy of last model is a little bit worse than for *likes* model.

Table 4

Accuracies for "Patterns" models

Characteristic	Mean RMSE	Test RMSE	Test R2
Openness	0.021	0.016	0.435
Conscientiousness	0.035	0.025	0.487
Extraversion	0.028	0.024	0.248
Agreeableness	0.050	0.026	0.727
Emotional range	0.032	0.022	0.525

All three models were combined in linear ensemble to incorporate all unique information from each. Final results are in Table 5.

Table 5

Accuracies for ensamle of models

Characteristic	Mean RMSE	Test RMSE	Test R2
Openness	0.021	0.015	0.50
Conscientiousness	0.035	0.024	0.53
Extraversion	0.028	0.023	0.33
Agreeableness	0.050	0.024	0.78
Emotional range	0.032	0.019	0.63

Also, we could compare extracted OCEAN based on two approaches, and results are pretty close.

Conclusion. This article shows how OCEAN scores could be extracted based on different data sources. Of course, there are a lot of ways to improve models, like using more data or more advanced machine learning. Also, for case with Amazon comments dataset, we don't have complete list of transactions, just comments and products metadata, what also decreases models' accuracies.

REFERENCES:

1. Big Five Personality traits. URL: https://en.wikipedia.org/wiki/Big_Five_personality_traits

2. Lisa A., Fast and David C. Funder. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. Journal of Personality and Social Psychology, 2008, Vol. 94, No. 2, 334–346;

3. A. J. Gill, S. Nowson, J. Oberlander. What Are They Blogging About? Personality, Topic and Motivation in Blogs. Proceedings of the Third International ICWSM Conference, 2009;

4. J. B. Hirsh, J. B. Peterson. Personality and language use in self-narratives. Journal of Research in Personality, 2009.

5. Amazon review dataset. URL: http://jmcauley.ucsd.edu/data/amazon/;

6. IBM Watson cognitive service. URL: https://www.ibm.com/watson/services/personality-insights/;

7. S. Hochreiter, J. Schmidhuber. Long short-term memory. Neural Computation 9 (8), 1735–1780, 1997.