FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice № 3

Sampling of multivariate random variables

Performed by:
Denis Zakharov,
Maxim Shitilov,
Sapelnikova Ksenia,
Vdovkina Sofia,
Academic group J4132c, J4133c

Saint-Petersburg
2022

Goals:

1. Substantiation of chosen sampling.
2. Sampling of chosen target variables using univariate parametric distributions (from practice #2) with 2 different sampling methods.
3. Estimation of relations between predictors and chosen target variables.
4. Bayesian network
5. Quality analysis.

Brief theoretical part

In this lab work we are working with a Bayesian network. This is a probabilistic graphical model that represents a set of variables and their conditional dependencies using an oriented acyclic graph. Bayesian networks are a tool for taking an event that has occurred and predicting the probability that any of several possible known causes was a contributing factor.

Structural learning is the process of using data to study the connections of a Bayesian network or a dynamic Bayesian network. There are the following structural learning algorithms: Clustering, PC, Search and Evaluation, Hierarchical, Chow-Liu, Augmented Naive Bayes Tree (TAN).

It is also possible to combine algorithms into a chain. To do this, you need to complete the work of the Structural learning wizard several times, each time importing link constraints from the previous run.

Bayesian networks are ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Efficient algorithms can perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences) are called dynamic Bayesian networks.

Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.

Results

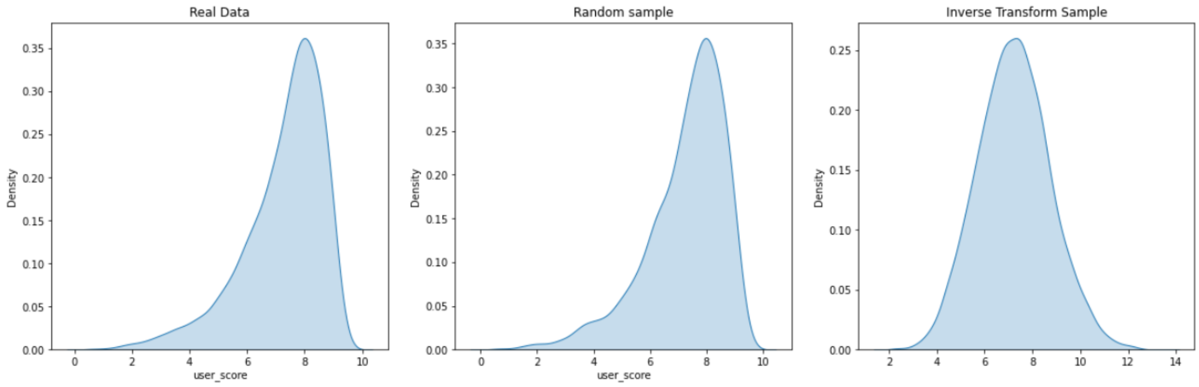
1. Choose variables for sampling from your dataset (overall – about 10 variables, 3-4 – target variables, the rest - predictors).

At this stage of laboratory work, it is necessary to select target variables for further work on laboratory work. As can be seen from the following code, the target variables are 'user_score', 'na_sales', 'other_sales'. The remaining dataset variables are predictors for the target variables.

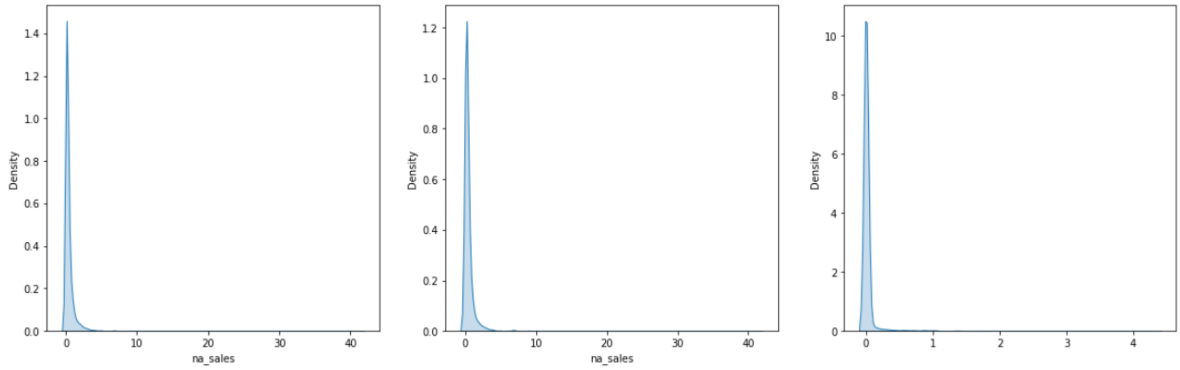
```
target = df[['user_score', 'na_sales', 'other_sales']]

predictor = df[[

    'name', 'platform', 'year_of_release', 'eu_sales',

    'jp_sales', 'critic_score', 'world_sales',

    'genre', 'rating' ]]
```

2. Using univariate parametric distributions that were fitted in Lab#2 make sampling of chosen target variables. Use these 2 different sampling methods.

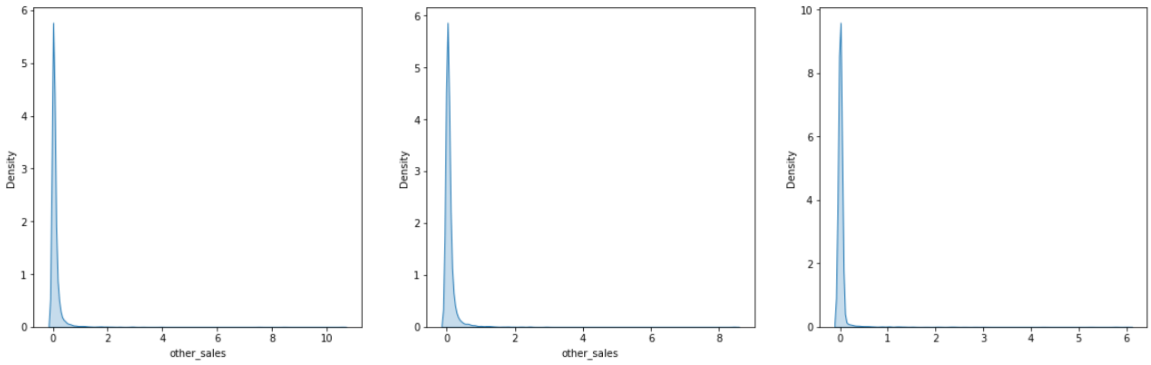
Here it was necessary to carry out a selection of target variables. For this purpose, sampling methods such as Random Sampling and Inverse Transform Sample were used. As you might guess from the name, Random Sampling returns a random selection of elements from the axis of the object. Inverse transform sampling is a basic method for pseudo-random number sampling, i.e., for generating sample numbers at random from any probability distribution given its cumulative distribution function (gamma distribution in our case).



Pic. 1 — 'user_score' variable sampling graphs for different sampling methods

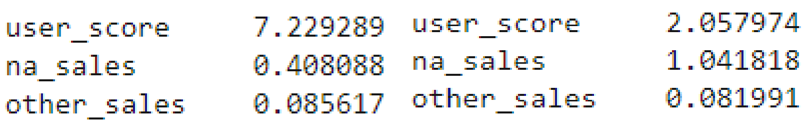


Pic. 2 — 'na_sales' variable sampling graphs for different sampling methods



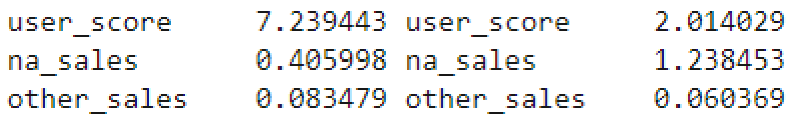
Pic. 3 — 'other_sales' variable sampling graphs for different sampling methods

Based on the graphs presented above, we can say that the Random Sampling method coped a little better, since the graphs of the results of the Inverse Transform Sample reflect the original data worse.



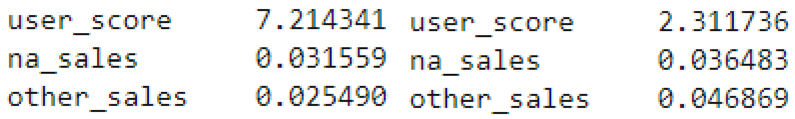
```
user_score      7.229289 user_score      2.057974
na_sales        0.408088 na_sales        1.041818
other_sales     0.085617 other_sales     0.081991
```

Pic. 4 — Mean (left) and variance (right) of the original data.



```
user_score      7.239443 user_score      2.014029
na_sales        0.405998 na_sales        1.238453
other_sales     0.083479 other_sales     0.060369
```

Pic. 5 — Mean (left) and variance (right) of the Random Sampling.



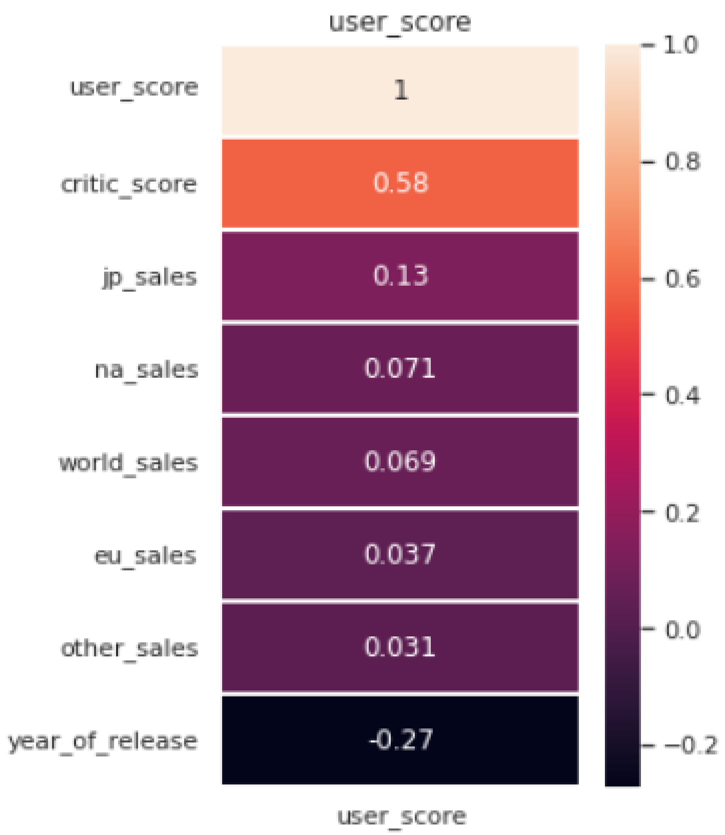
```
user_score      7.214341 user_score      2.311736
na_sales        0.031559 na_sales        0.036483
other_sales     0.025490 other_sales     0.046869
```

Pic. 6 — Mean (left) and variance (right) of the Inverse Transform Sample.

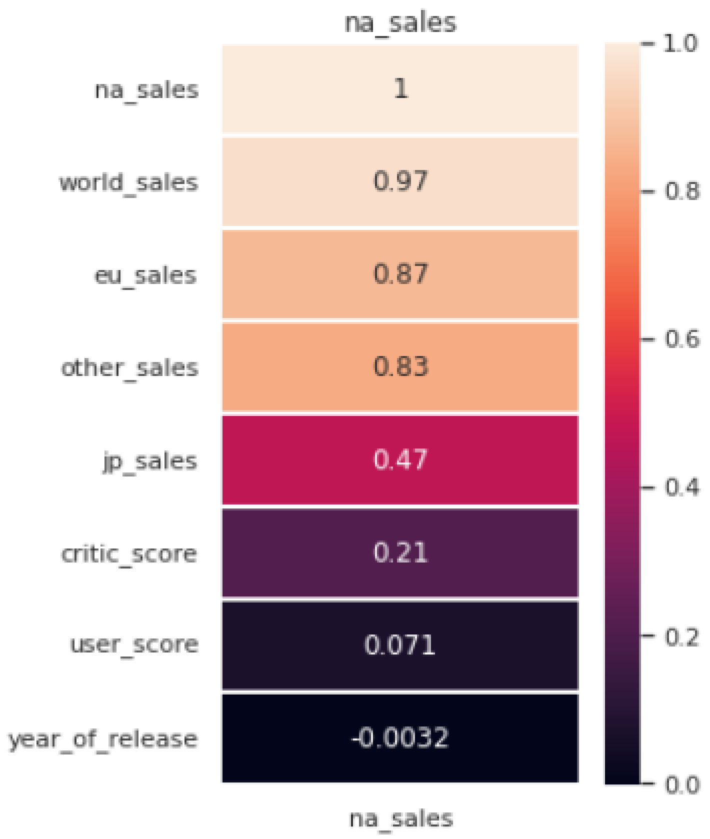
Mean and variance values show us a result similar to graphs. The results of the Random Sample are closer to the original data than the results of the Inverse Transform Sample.

3. Estimate relations between predictors and chosen target variables. At least, they should have significant correlation coefficients.

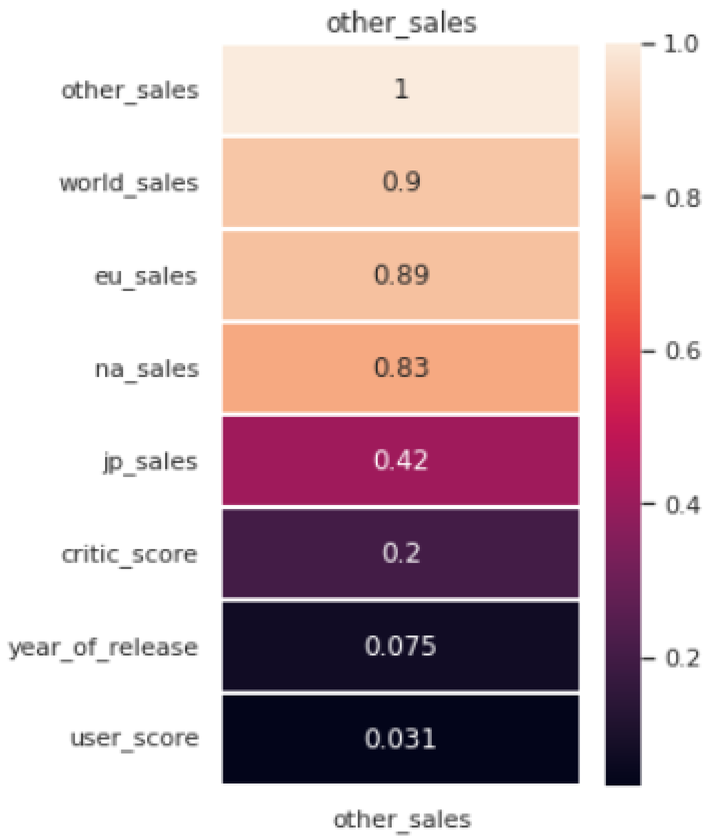
On this step of the laboratory work, it was necessary to evaluate the relationship between the predictors and the selected target variables.



Pic. 7 — Estimation of correlation coefficients for the 'user_score' variable



Pic. 8 — Estimation of correlation coefficients for the 'na_sales' variable



Pic. 9 — Estimation of correlation coefficients for the 'other_sales' variable

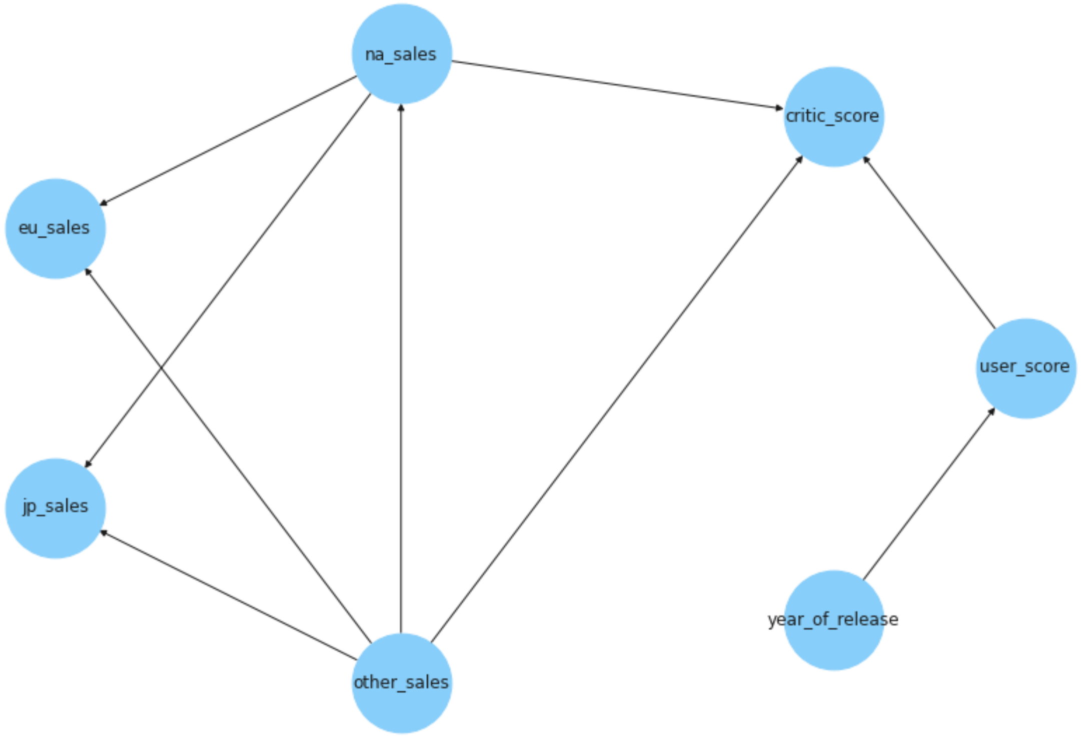
According to the resulting graphs, the most significant coefficient for the target variable 'user_score' was obtained when correlated with the predictor 'critic_score'.

The most significant coefficient for the target variable 'na_sales' was obtained when correlated with the predictor 'world_sales'.

The most significant coefficient for the target variable 'other_sales' was obtained when correlated with the predictor 'world_sales'.

4. Build a Bayesian network for a chosen set of variables. Choose its structure on the basis of multivariate analysis and train distributions in nodes using the chosen algorithm.

Knowing p-values, let us construct a Bayesian network.



Pic. 10 — Bayesian network for chosen set of variables

In the graph above, you can see the result of constructing a Bayesian network for the data set we have selected.

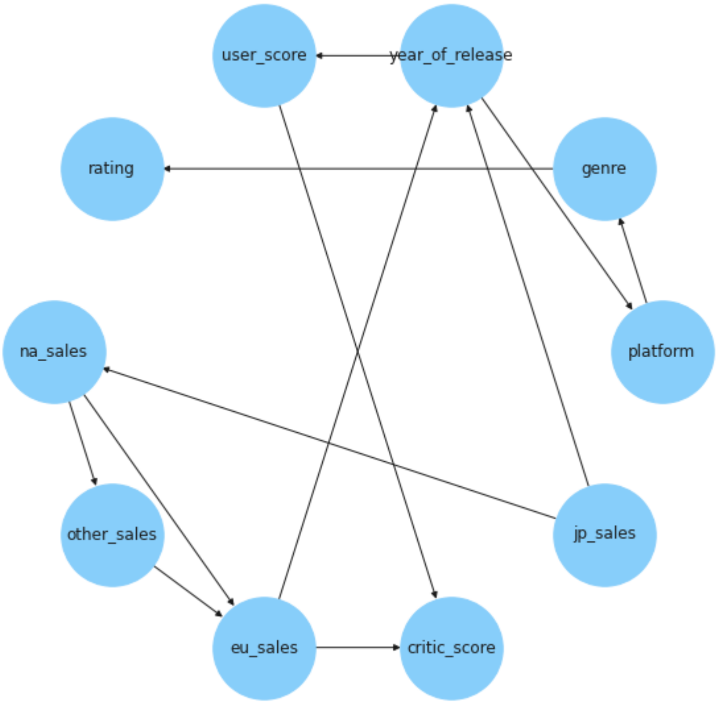
This network was created using our multivariate analysis.

5. Build a Bayesian network for the same set of variables but using 2 chosen algorithms for structural learning.

One of the algorithms that we used to build Bayesian networks is the search by climbing to the top. This is a mathematical optimization technique belonging to the family of local search algorithms. The algorithm is an iteration method that starts with an arbitrary solution to the problem, and then tries to find the best solution by step-by-step changing one of the elements of the solution.

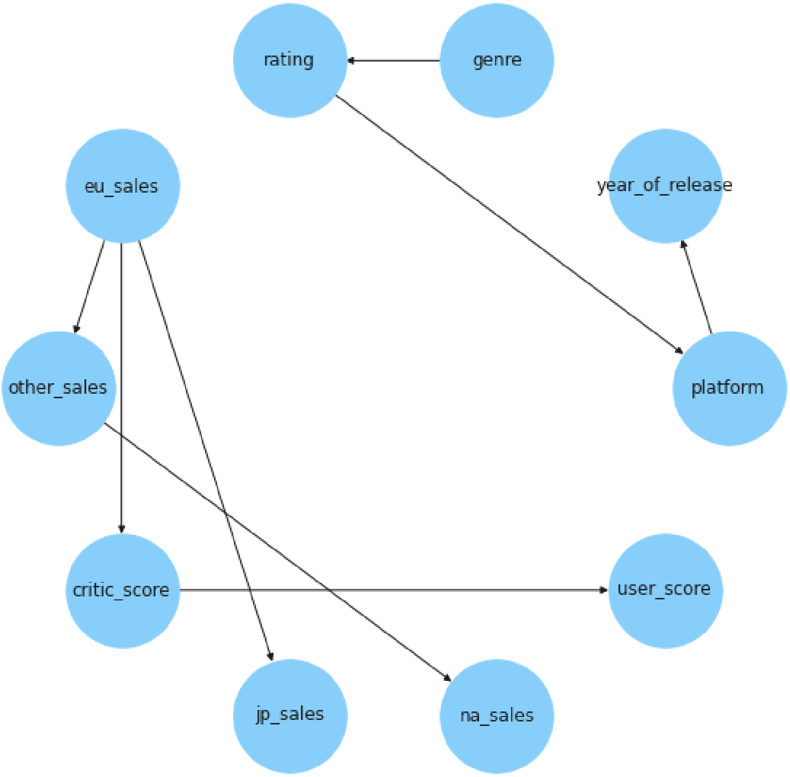
This algorithm was used to estimate three following networks.

The network on the picture 11 was obtained by using K2 structure score. This algorithm computes a score that measures how much a given variable is "influenced" by a given list of potential parents.



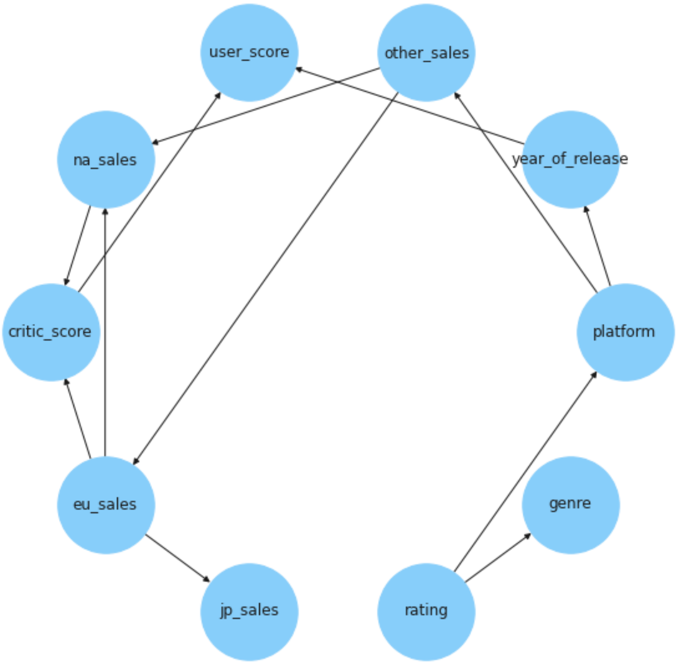
Pic. 11 — K2 structure score

The network on the picture 12 was obtained with the use of Bayesian information criterion.



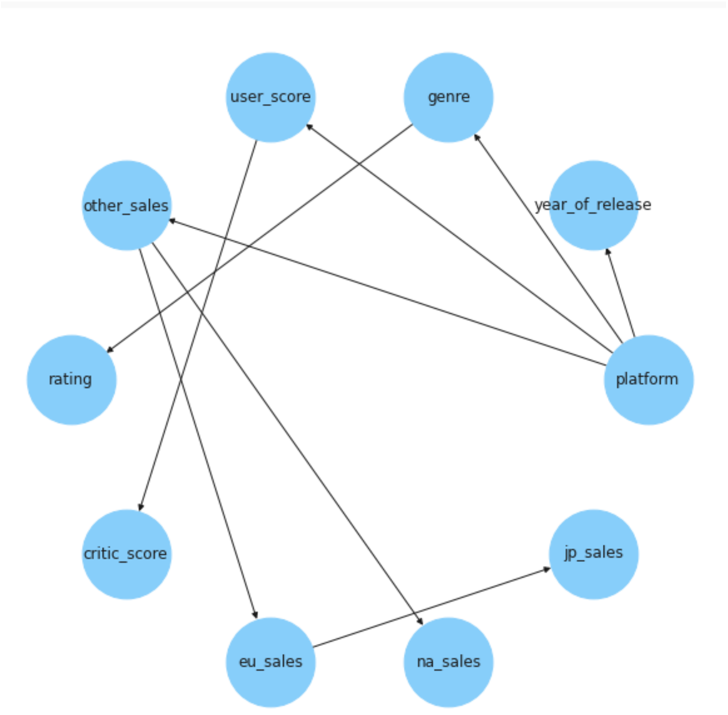
Pic. 12 — Bayesian information criterion

The network on the picture 13 was obtained with Bayesian Dirichlet equivalent uniform (BDeu) score.



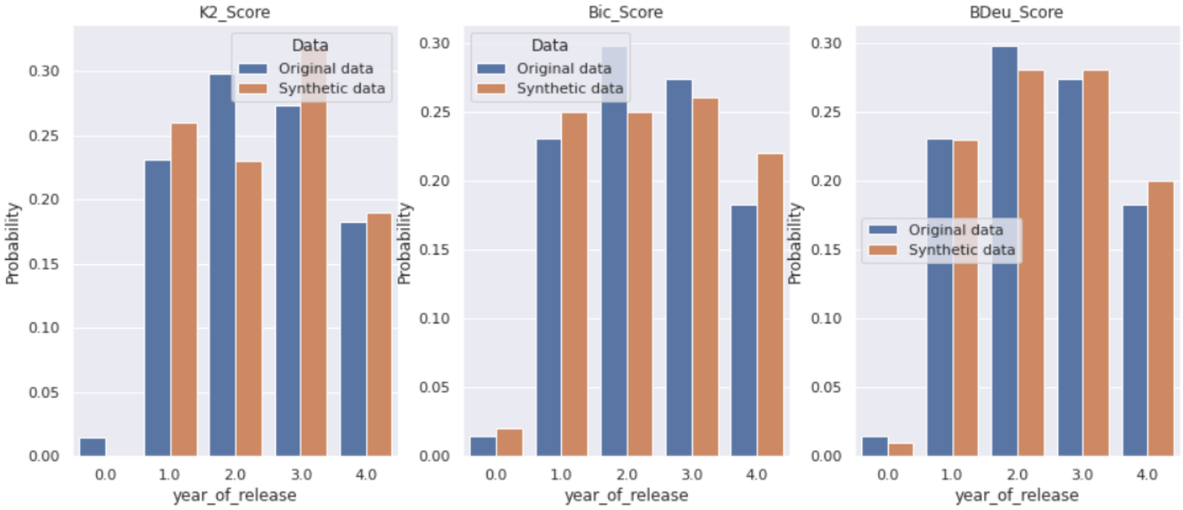
Pic. 13 — Bayesian Dirichlet equivalent uniform (BDeu) score

The following network structure on picture 14 was estimated by Tree search algorithm, that finds structure that fits best to the given data set without parameterization.

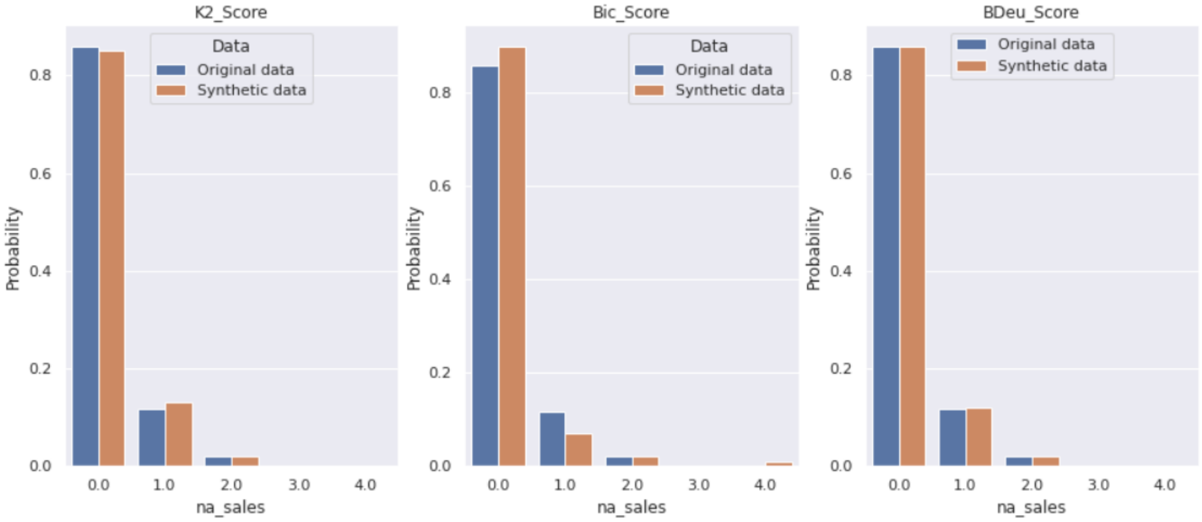


Pic. 14 — Tree search algorithm

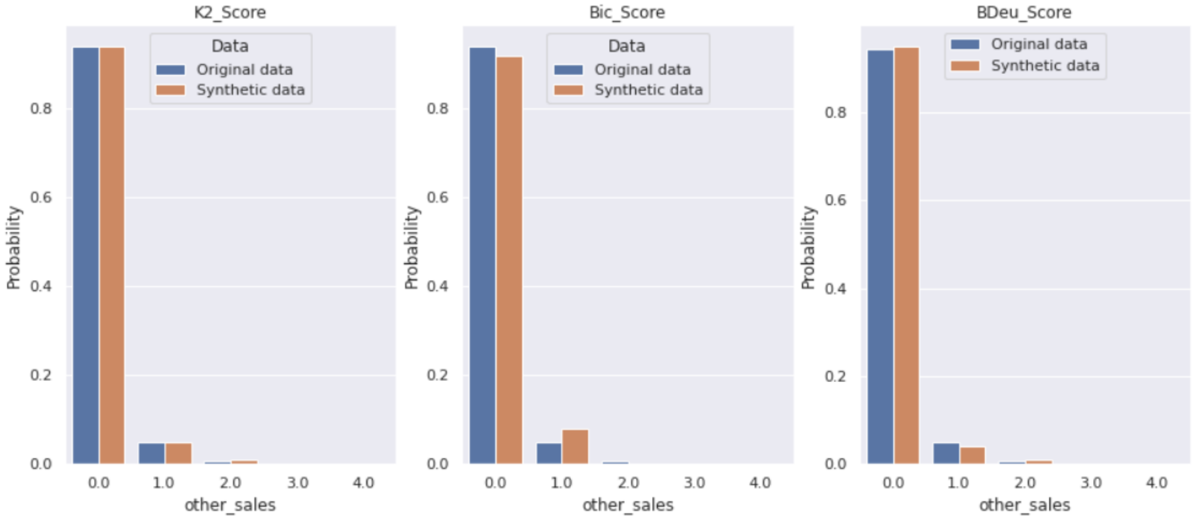
6. Analyze a quality of sampled target variables from the point of view of problem statement (e.g. prediction, gap filling, synthetic generation).



Pic. 15 — Evaluation of the target variable 'user_score'



Pic. 16 — Evaluation of the target variable 'na_sales'

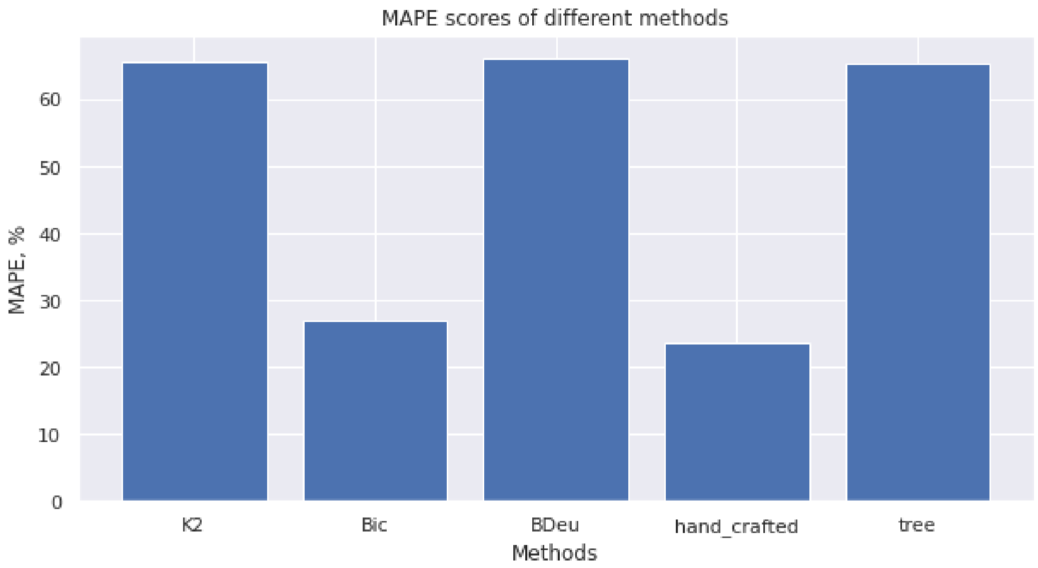


Pic. 17 — Evaluation of the target variable 'other_sales'

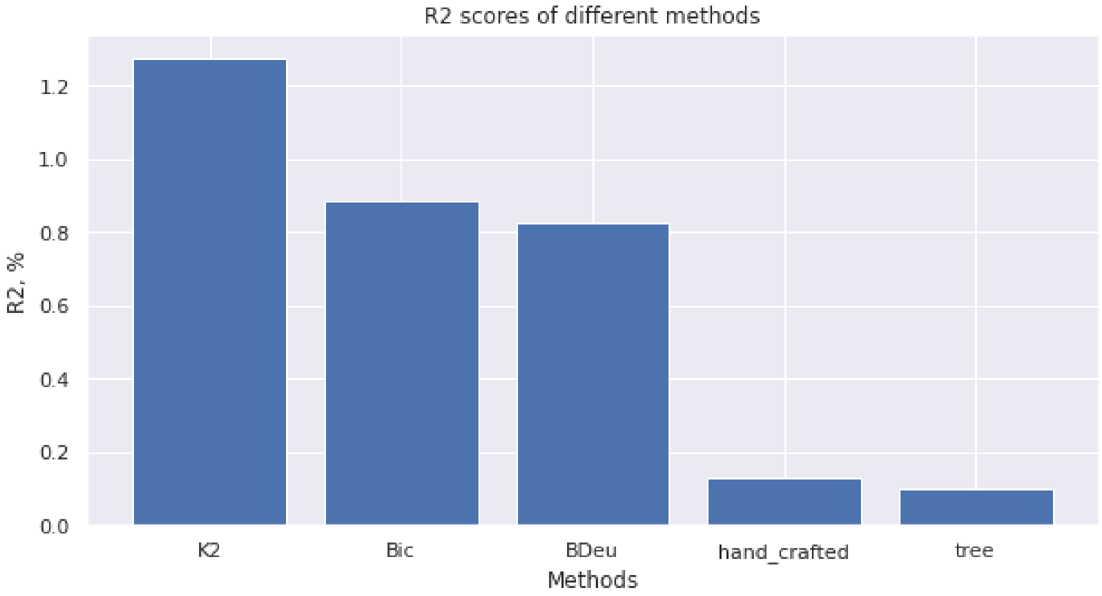
At this stage of the laboratory work, it was necessary to analyze the quality of the selected target variables from the point of view of the problem statement. To do this, several estimates were applied for each of the target variables. Exactly K2_Score, Bic_Score and BDeu_Score.

Judging by the graphs, the quality of the selected target variables satisfies all the above estimates. On average, synthetic data produces similar results to the original data. In some moments, the evaluation of the source data is higher, but there are also situations in which synthetic data copes better.

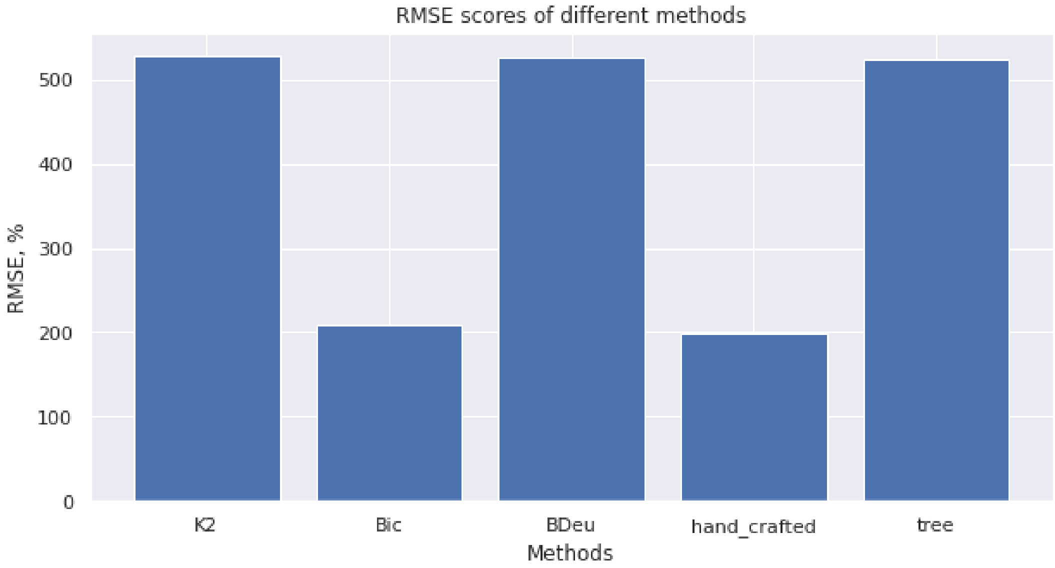
Also, let’s look at some numerical characteristics for ‘user_score’ sampling.



Pic. 18 — MAPE scores



Pic. 19 — R2 scores



Pic. 20 — RMSE scores

Conclusions

As a result of the work, we've worked with sampling of multivariate random variables. And also got our hands on building a Bayesian network for pre-chosen algorithms for tasks of structural learning.

Sourcecode

DataLore: site. – URL: https://datalore.jetbrains.com/notebook/JCNco4jwZlQCPbcDKLB3og/kRnzkLwB7jICNNOnCsqoXz (circulation date: 28.11.2022)