Journal of Trends and Challenges in Artificial Intelligence

ASPUR PUBLISHING

Vol. 02, No. 01 (2025) 15-18, doi: 10.61552/JAI.2025.01.002 - http://jai.aspur.rs

SOLVING THE PROBLEM OF MATHEMATICAL MODELING OF THE INFLUENCE OF FACTORS ON THE CONVERSION RATE OF FUNNELS TRANSFER OF INNOVATIVE DEVELOPMENTS USING MACHINE LEARNING TOOLS

Olga Pyataeva [1]

Received 11.06.2024.
Revised 21.07.2024.
Accepted 21.08.2024.

Keywords:

Mathematical methods, forecasting, machine learning, transfer of innovative developments, conversion rate of transfer funnel.

Original research

A B S T R A C T

The article is aimed at analyzing the possibilities of solving the problem of modelling the influence of various factors on the indicator of conversion of the transfer funnel of innovative developments. In the authors' previous publications this indicator was marked as «conversion rate» (percentage of developments that have passed from the stage of «object of intellectual property, received legal protection» to the stage of «patented result of intellectual activity, put on the market»), the list of factors that influence it is determined. This article presents the result of testing the algorithm of modeling with the use of machine learning tools. The information base of the study became open information about indicators of innovation activity and patent activity in the Russian Federation. Research was asked about the possibility of using machine learning tools to construct regression problems describing the influence of factors on a feature, and to draw conclusions about the most important factors. In the course of solving the problem, the following stages of data analysis for machine learning were implemented: cleaning and formatting, preliminary analysis, selection of the most important factors, model testing on a test sample. In conclusion, it is concluded that the use of machine learning tools for this type of task provides comparable and accurate results, but uses a disproportionate amount of data as a research base, what is important when forming complex regressions and forecasts.

1. INTRODUCTION

Need to solve the problem of improving tools for forecasting the quantities of indicators, characterizing the stages of transition of innovative development from its creation to «integration» into the market (this process in previous publications was characterized using the indicator «conversion rate of funnel transfer innovation developments») required testing the possibility of using machine learning tools for the subject area under consideration (Holmström et al., 2024; Papa et al., 2022; Shanmugam et al., 2023;).

The research was based on open information about innovation activity and patent activity in the RF. The authors questioned whether machine learning tools could be used to:

- Forecasting of the conversion rate of the funnel transfer of innovative developments when moving from stage to stage (“Development” - “Processing of

[1] Corresponding author: Olga Pyataeva
Email: opyataeva@gmail.com

rights in intellectual activity" - "Processing on mother-like carrier" - "Market launch");
- Identification of the list of factors that have the greatest influence on co-efficient conversion;
- The development of a regression prediction model describing the effect of factors on the result (conversion coefficient).

2. LITERATURE REVIEW

The use of machine learning tools to solve various mathematical and economic problems is increasingly being applied in modern educational practice. This is pointed out, for example, by Jordan and Mitchell (2015), Mukhamediev et al. (2022), Nosratabadi et al. (2020), Tedre et al. (2021), Tehranian (2023), Zhong et al. (2021), and other researchers. It is emphasized that the use of machine learning methods allows to expand the range and types of data processed and to perform analysis with higher speed, minimize errors in calculations, etc.

3. METHODOLOGY

Analysis of studies in the presented area showed that problems of using machine learning tools for forecasting economic variables have not been studied enough. Only for some areas of research at present there has been the practice of using machine learning (mainly in the context of solving «point» tasks relevant to specific organizations).
However, the development of data analysis and interpretation practices and experiences with machine learning is highly relevant in view of the need to improve the quality and speed of data processing in economic forecasting (Dogan & Birant 2021; Khalil et al., 2022). The current range of machine learning methods is quite extensive (Dargan et al., 2020, Sarker, 2021).
The task of testing existing methods, forming a framework for data processing to solve research problems of forecasting and identifying the shortcomings of existing methods for their further improvement is relevant.
In the process of writing the article was used methods of analysis and synthesis, correction-regression analysis, as well as artificial intelligence methods, which imply «not a direct solution to the problem, but learning through the application of solutions to many problems».

4. PRE-ENGINEERED BUILDING (PEB)

The following phases of the course were selected to address the data analysis and processing task:
1) Data cleaning and formatting;
2) Preliminary analysis of data;
3) Selection of the most relevant factors;
4) Model validation on test sample.

The content of the work in specific stages will be presented in more detail below.

4.1 Data Cleaning and Formatting

The data presented in open sources contained «surplus» variables that had to be «cleaned up». For this purpose, the data (source table in MS Excel) was loaded into DataFrame (table).
a) The identification and cleaning of text data (format «H/D» or «Data not made available to the public») involved:
- Evaluation of their presence in the data set using the command:

```
# See the column data types and non-missing values
data.info()
```

- Replace the values of "N/D" or "Data not available in public" in the data with «not number», which allowed to change the type of these data to "float".

```
# Replace all occurrences of Not Available with numpy not a number
# Convert the data type to float
```

It was found that 4% of the rows had data in an inconsistent format.
b) The data in the non-compliant format were then filled in or excluded from the sample (which was not critical to the result, as the percentage of such data was unacceptably small).

4.2 Preliminary analysis of data

The aim of this step was to build a predictive regression model describing the influence of factors on the conversion rate of funnel transfer of innovative spin-offs. The following sub-phases were implemented in the phase:
a) Analysis of the magnitude of correlation of the conversion factor of the funnel transfer (more in detail - in the study of one of the authors of article) and various factors having a direct or indirect influence on it:
- Costs of innovation activities of organizations;
- Level of innovation activity of organizations;
- Share of organizations that have implemented technological innovations;
- Number of personnel engaged in scientific research and development;
- Average annual number of persons employed in the economy;
- GDP.

The list of factors presented is somewhat truncated. A more comprehensive list was presented in [6], but the number of factors in the course of the ongoing work has been reduced by almost half (factors with little impact on the final result have been removed). In the framework of a previously conducted study, presented factors were identified correlation coefficients that characterize the degree of interrelation between individual factors and the resulting indicator (conversion coefficient) in the model built in the MS Excel software environment.
During the test using machine learning tools, it was confirmed that the highest positive correlation with the

confirmation coefficient of the indicator «Costs of innovation activities of organizations» (+0.87), i.e., the higher the cost of innovation, the higher the conversion rate.

b) The relationship between variables «Conversion rate» and «Innovation costs of organizations» was illustrated in a graphical format (scatterplots tool) in the Python software environment.

4.3 Selection of the most relevant factors

This phase also included several sub-phases.

a) Factors that are less correlated with the result were removed. The aim of this work was to clean up the data, to create a list of the most important facts.

Some factors were found to be «superfluous»; their correlation with the environment was identified. For example, the correlation between «Annual average number of persons employed in an economy» and «Number of personnel engaged in research and development» was found (the corresponding coefficient was 0.997), although their relationship was initially obvious. This and other features that have been found to be somehow correlated are called collinar. The work methodology suggests that one of these features should be left and the other excluded, with a simplification of the model's work, which was implemented.

For the balancing of redundant data in this case the correlation coefficient was used, which allowed to remove one of the factors for the case if the coeffetive of the correlation for their ratio is above 0.8. The result was a list of 5 most important factors:

- Costs of innovation activities of organizations;
- Level of innovation activity of organizations;
- Percentage of organizations that have implemented technological innovations;
- Number of staff engaged in scientific research and development;
- GDP.

b) The model evaluation criterion was chosen

The purpose of this work was to determine whether machine learning in general is necessary. This question has proved relevant in that connection, the functionality of MS Excel also allows to use substitution data, predict using simple topics, for example, linear. The model criterion for MS Excel studies can be, for example, a conversion factor midpoint offset with an absolute mean deviation.

The sample should be classified as follows:

1. "Instructive»"

The part of data (ratio of factors and result) used for training should be selected to illustrate the connection and «train» the model.

2. Test

The test part should be defined, i.e. only factors are used and the model result must be «predicted» and then compared to the actual values.

In the data set under consideration, 60% of the records were used for training and 40% tested, which was done by the following teams in a Python programming environment:

```
# Split into 60% training and 40% testing set
X, X_test, y, y_test = train_test_split(features, targets,
test_size = 0.3,
random_state = 42)
```

The machine processing results did not show 10 points of possible conversion factor value (range from 1 to 100), i.e. an error of 10%, which confirmed that the selected model was acceptable, a The data obtained can be used to form conclusions based on them.

3. Model validation on test sample

At this stage, the model was tested on a test sample.

The necessary design of a final model, its validation (testing) and evaluation on test sample was identified. The following Python commands were used:

```
# Default model
default_model =
GradientBoostingRegressor(random_state = 42)
# Select the best model
final_model = grid_search.best_estimator_
final_model
default_pred = default_model.predict(X_test)
final_pred = final_model.predict(X_test)
```

To retrieve the analysis results:

```
print('Default model performance on the test set: MAE =
%0.4f.' % mae(y_test, default_pred))
print('Final model performance on the test set:   MAE =
%0.4f.' % mae(y_test, final_pred))
```

To form (calculate) the final figure:

Default model performance on the test set: MAE = 9.34.
Final model performance on the test set: MAE = 8.2.

The conclusion was reached that the model is superior to the basic one.

The model was illustrated by comparing the following distributions:

- the baseline values on the test sample, and
- predicted values on the test sample.

```
label = 'Values')
```

The following commands were used:

```
figsize(8, 8)
# Density plot of the final predictions and the test values
sns.kdeplot(final_pred, label = 'Predictions')
sns.kdeplot(y_test,
```

In this case, the distribution of baseline values on a test sample and predicted values on a test sample were almost identical. As a result, it became apparent that the created machine learning model is acceptable for use, its results are not valid.

5. CONCLUSIONS

Analysis and design in this study yielded the following conclusions:

The provisions considered in the article allow to draw the following conclusions.

- Machine learning methods can be used to solve economic and mathematical modelling problems

alongside traditional ones. The implementation of such methods is determined by the specific task, availability of data, ability to obtain them.

- To perform data analysis using machine learning, the project must be carried out in several stages: a) data cleaning and formatting; b) preliminary data analysis; b) selection of the most significant factors; g) testing of the model on a test sample. The list of steps can be expanded as necessary.
- The aim of the current study (modelling the influence of various factors on the conversion rate of funnel transfer of innovative developments) has been achieved. It was found that a number of previously highlighted factors, being collinear (i.e. having correlation with each other), when used together, make the model more «heavy».

The distribution of baseline values in a test sample and predicted values in a test sample were almost identical, which led to the conclusion that the created machine learning model is acceptable for use and its results are valid.

References:

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of Computational Methods in Engineering*, *27*, 1071-1092.

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060.

Holmström, J., Kostis, A., Galariotis, E., Roubaud, D., & Zopounidis, C. (2024). Stalled data flows in digital innovation networks: Underlying mechanisms and the role of related variety. *Industrial Marketing Management*, *121*, 16-26.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Khalil, M., McGough, A. S., Pourmirza, Z., Pazhoohesh, M., & Walker, S. (2022). Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption—A systematic review. *Engineering Applications of Artificial Intelligence*, *115*, 105287.

Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., ... & Yelis, M. (2022). Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. *Mathematics*, *10*(15), 2552.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., ... & Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, *8*(10), 1799.

Papa, A., Mazzucchelli, A., Ballestra, L. V., & Usai, A. (2022). The open innovation journey along heterogeneous modes of knowledge-intensive marketing collaborations: a cross-sectional study of innovative firms in Europe. *International Marketing Review*, *39*(3), 602-625.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Shanmugam, G., Rajendran, D., Thanarajan, T., Murugaraj, S. S., & Rajendran, S. (2023). Artificial Intelligence as a Catalyst in Digital Marketing: Enhancing Profitability and Market Potential. *Ingénierie des Systèmes d'Information*, *28*(6).

Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., & Pears, A. (2021). Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE access*, *9*, 110558-110572.

Tehranian, K. (2023). Can machine learning catch economic recessions using economic and market sentiments?. *arXiv preprint arXiv:2308.16200*.

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., ... & Zhang, H. (2021). Machine learning: new ideas and tools in environmental science and engineering. *Environmental science & technology*, *55*(19), 12741-12754.

Olga Pyataeva
Higher School Of Economics,
Russia.
opyataeva@gmail.com
ORCID: 0000-0001-6373-1642