The use of intelligent technologies of Business Intelligence Platforms and Data Science and Machine Learning Platforms for monitoring the socio-economic indicators of the administrative districts of Moscow

P A Sakhnyuk¹, T I Sakhnyuk², M V Korshikova³

¹National Research University Higher School of Economics, 20, Myasnitskaya str., Moscow, 101000
²Moscow City University, 4, 2-nd Selskohozoyastvenny street, Moscow, 129226
³Stavropol State Agrarian University, Zootekhnichesky Lane, 12, Stavropol, 355017, Russia

E-mail pav-sahnyuk@yandex.ru

Abstract. The article discusses the possibilities of studying the state of the social sphere according to the repository of the Moscow Government open data portal by administrative districts and city districts using Business Intelligence Platforms and Data Science and Machine Learning Platforms intellectual technologies. Opportunities are presented for using machine learning technologies for business analytics platforms to identify hidden patterns in order to make informed management decisions.

Keywords: data science and machine learning platforms, business intelligence platforms, machine learning, artificial intelligence, augmented analytics technology.

1. Introduction
Currently, when collecting and analyzing data obtained from databases of state statistics services, portals of open data and other sources, there is a problem of obtaining high-quality and visual information. As a result, the information available to executive authorities and individual citizens is largely expert and heuristic in nature. Therefore, there is an objective need to use intelligent business intelligence (BI) platform technologies for monitoring socio-economic indicators of administrative districts of Moscow [1, 2].

Augmented analytics technology has become the main development driver of BI platforms and platforms for data processing and machine learning. Augmented analytics technology, based on machine learning, provides insight into huge amounts of data. This technology also includes natural language processing as a way to query data and create narratives to explain drivers and graphics. Nowadays, up to half of all analytical queries in BI platforms are generated using search, natural language processing or generated automatically. Natural language processing and conversational analytics are increasing the spread of business intelligence among civilian professionals, including new classes of users, especially employees of front offices and city municipalities.
A data set of families receiving subsidies in Moscow in various sections and data on citizens’ requests to the open data portal were selected for intellectual analysis [3].

2. Visual data analysis using Business Intelligence Platforms

Figure 1 shows a collection of different visualizations created in the Power BI service.

![Dashboard number of citizens requests to the open data portal (in absolute units and per resident of administrative divisions and districts of the city)](image)

Figure 1. Dashboard number of citizens requests to the open data portal (in absolute units and per resident of administrative divisions and districts of the city)

![Results of clusterization of Moscow districts using the k-means Power BI algorithm](image)

Figure 2. Results of clusterization of Moscow districts using the k-means Power BI algorithm

Let us find “similar” objects among 146 districts of Moscow. Searching for similar objects is one of the most common tasks in data analysis. We determine the similarity based on the characteristics of the data set “the number of families receiving subsidies in Moscow in various sections”: the total number of families, low-income families, multi-child families, single-parent families, retired families, students’ families, unemployed families and the number of requests to open data portal. In Power BI clustering support provides powerful analytical capabilities and Power BI supports the k-means algorithm, the most popular clustering method. One of the most difficult tasks in clustering is to determine the number of clusters (Figure 2). To facilitate this task, Power BI provides both automatic and manual control options.
To check the results, we will perform clusterization of Moscow districts using other clusterization algorithms. As a tool for cluster analysis, we will use the Kohonen neural network, implemented in Deductor Studio analytical platform, the advantage of which in comparison with other algorithms is the ability to visually analyze multidimensional data: similar objects fall into neighboring map cells (Figure 3).

Verification of clustering results using Kohonen maps and k-means can be performed in the KNIME Analytics Platform (Figure 4), where it is easy to perform fuzzy c-means (FCM) clustering, the feature of which is that each data point is assigned to a cluster with a membership function that varies in the range from 0 to 1 [4-8]. Clusters are represented as fuzzy sets and in addition, the boundaries between clusters are also fuzzy. The degree of belonging is determined by the distance from the object to the corresponding centroids.

Figure 3. Kohonen maps for the set "Number of families receiving subsidies in Moscow in various sections" (cluster 4 shows the values for the Southern Butovo district)

Figure 4. Workflow FCM кластеризации in the KNIME Analytics Platform
The results of fuzzy clustering are shown in Figure 5. Comparing them with the results of clustering using Kohonen maps built in Deductor Studio and k-means Power BI (Figure 3, 4), we make sure that they are objective. The difference in results is observed only for the districts of Perovo, Izmailovo, Textilschiki and Pechatniki. FCM algorithm assigned them to the second cluster with a slight difference in the values of the membership function.

![Figure 5. Results of FCM clustering of Moscow districts in the KNIME Analytics Platform](image)

Figure 5. Results of FCM clustering of Moscow districts in the KNIME Analytics Platform

Corelation between the characteristics of our data set reflecting the social status of residents of city districts are investigated using Power BI dashboards. They include a correlation matrix of the studied characteristics, a dot diagram of the population of the cluster districts and the number of requests from citizens to the open data portal and a map of the city districts (Figure 6).

![Figure 6. Dashboard Power BI by Moscow districts in the context of found clusters](image)

Figure 6. Dashboard Power BI by Moscow districts in the context of found clusters

For more in-depth data research, we use the Qlik Sense platform, which combines free-form associative research, provides context-sensitive suggestions and automatically constructs...
visualizations based on artificial intelligence. The platform uses the Qlik Cognitive Engine to provide automated analysis suggestions that help you look at data in a new way by automatically generating and prioritizing analytics based on the overall data set and user search criteria. The augmented intelligence platform offers insight and automation of visual analytics by understanding questions and responding in natural language.

An analysis of the dashboard (Figure 7) shows that as the number of families of unemployed, single-parent families and families of students receiving subsidies increases, so does the number of requests from citizens to the open data portal (typical for the districts of the second cluster).

To identify key factors of influence, we use the appropriate visualizer in Power BI that will help you to understand what factors affect the metric under study (in our case, the cluster number that the Moscow region belongs to). Using machine learning in Power BI helps to explore data, allowing to perform deep analysis to automatically search for patterns, interpret results in a clear way and predict results. Using machine learning (a linear regression model), the factors of influence are ranked from the most to the least significant and a probability indicator and text description are provided to explain the influence. For example, for districts in cluster 4 (Figure 8), the most significant factor is the number of families of students receiving the grant.

For the districts of cluster 0, the most significant factor is the number of unemployed families receiving subsidies. For districts of cluster 1, the most significant factor is the number of single-parent families receiving the subsidy. For cluster 3 districts, the most significant factor is the number of families of students receiving the subsidy. For the districts of cluster 4, the most significant factor is the number of families receiving the subsidy. For the districts of cluster 5, the most significant factor is the number of unemployed families receiving subsidies.
One of the fastest ways to get an answer from data is to ask a question in natural language. This feature is provided by the Q&A function in Power BI. Q&A is an interactive option, in which often one question leads to another, because visualization opens up interesting ways to achieve the goal. In the Power BI service, the dashboard contains sheets pinned to one or more datasets, so you can ask questions about any data contained in the dataset. The Q&A tool recognizes the entered words and finds out where (in which data set) to find the answer. Q&A will also help to form a question with autofill, recalculation and other text and visual suggestions. The answer to the question is displayed as an interactive visualization and updated as the question changes.

3. Application of machine learning technologies
There is a close correlation between the number of families receiving subsidies and the number of requests from citizens to the open data portal (Figure 9). To identify hidden patterns in data and automate the main tasks of machine learning, we use the customizable artificial intelligence platform H2O Driverless AI. The platform provides automatic engineering of functions, checking and configuring models, selecting and deploying models, interpreting machine learning, creating custom scenarios in model construction, processing time series and texts, and automatically generating lines for model scoring.

Figure 8. Visualizer of searching for key power BI influence factors in the districts of cluster 4
Figure 9. Brief analytics visualizer for finding trends in a data set: scatter plot demonstrating the relationship between the number of families receiving a subsidy and the number of citizens applying to an open data portal.

H2O Driverless AI performs the function of designing input features and output variables to determine their optimal representation for building a machine learning model. Various stages of function transformations appear throughout all iterations. You can view them by hovering over iterative data points in the Variable Importance (Figure 10) section. Conversion to H2O Driverless AI is applied to columns of a dataset. Numerical, categorical, time and text converters design functions that support machine learning algorithms. Function design is a fundamental complex and expensive process in machine learning.

Figure 10. Experiment dashboard
The best model in our experiment is LightGBM, a gradient boosting framework developed by Microsoft that uses decision tree-based learning algorithms. It was specifically designed to reduce memory usage, increase learning speed and improve efficiency. Like XGBoost, it is one of the best gradient boosting implementations available [9-13]. It is also used to fit Random Forest models inside Driverless AI. The results are found in the experiment summary at the bottom right of the experiment page. Once the best predictive model is found, we can interpret it. MLI Dashboard presents various types of explanations concerning the model and its results [14-16]. All graphics on the dashboard are interactive. The number of single-parent families, families of the unemployed, families of students and large families receiving subsidies are of the greatest importance for forming a model for predicting the number of requests from citizens to the open data portal. Moreover, the contribution of these features to the model is different. Due to the high correlation with the output variable, the following indicators are: the number of families, retired families and low-income families receiving subsidies do not participate in the formation of the model. Additionally, H2O Driverless AI allows to download automatically generated documents, such as "Download experiment report" and "MLI Report", at the touch of a button. The developed model can be used to forecast the number of requests from citizens to the portal of the Moscow Government.

4. Conclusion
The results confirm the hypothesis that there is a positive correlation between the number of families receiving subsidies in Moscow in various sections and the number of requests from citizens to the open portal of the Moscow Government. This indicates that it is possible to influence the mechanism of participation of active Moscow citizens in the life of the city. The obtained objective information can be used to develop a strategy and make decisions on the development of the city.

5. Acknowledgment
The work is partially supported by the RFBR grant # 18-413-770006.

References
[1] Frolov Yu V, Sakhnyuk P A, Sakhnyuk T I 2018 Use of geanalytics tools for training bachelors and masters in the direction of "business Informatics" Bulletin Of The Russian University of peoples friendship. Series: Informatization of education 15 2 151-164 doi: 10.22363/2312-8631-2018-15-2-151-164
[2] Mamaev I I, Sakhnyuk T I, Sakhnyuk P A 2016 Analysis of the main socio-economic indicators of the districts of the Stavropol territory for 2011-2015 by means of business Analytics systems Scientific journal of Kubgau 123 09
[3] Open data portal of the Government of Moscow (data.mos.ru) https://data.mos.ru/opendata/7701236617-kolichestvo-semey-poluuchayushchih-subsidiyu-v-gorode-moske
[4] T Siddiqui, A Ausaf 2018 Data mining tools and techniques for mining software repositories: A systematic review Big Data Analytics. Springer. Singapore 717-726.
[5] Verma K, Bhardwaj S, Arya R, UL Islam M S, Bhushan M, Kumar A, Samant P. 2019 Latest Tools for Data Mining and Machine Learning International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8 9S doi: 10.35940/ijitee.I1003.0789S19
[6] R Alcalá, MJ Gacto, J Alcalá-Fdez 2018 Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017) Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 1239
[7] Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S 2016 Combining satellite imagery and machine learning to predict poverty Science 353(6301) 790–794
[8] Froelich W. 2017 Towards improving the efficiency of the fuzzy cognitive map classifier. Neurocomputing 232 83-93
[9] Mitchell R 2017 Gradient boosting, decision trees and XGBoost with
CUDA. https://devblogs.nvidia.com/parallelforall/gradient-boosting-decision-trees-xgboost-cuda/. Accessed 25 Feb 2018

[10] Mitchell R, Frank E. 2017 Accelerating the XGBoost algorithm using GPU computing PeerJ Computer Science 3:e127

[11] Shi S, Wang Q, Xu P, Chu X 2016 Benchmarking state-of-the-art deep learning software tools 2016 7th international conference on cloud computing and big data (CCBD) IEEE 99–104.

[12] Tianqi Chen and Carlos Guestrin 2016 Xgboost: A scalable tree boosting system Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794

[13] Nguyen G, Dlugolinsky S, Bobák M 2019 et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif Intell Rev 52 77–124

[14] Patrick H 2019 On the Art and Science of Machine Learning Explanations KDD ’19 XAI Workshop Proceedings

[15] Patrick H, Navdeep G, Nicholas S 2019 Proposed Guidelines for the Responsible Use of Explainable Machine Learning. arXiv:1906.03533v3 [stat.ML]

[16] Sorelle A F, Chitradeep D R, Carlos S and Dylan S 2019 Assessing the Local Interpretability of Machine Learning Models. arXiv preprint arXiv:1902.03501,