Methodology for Training Data Science Skills Based on Competitions on the Kaggle Platform

Zaur Kh. Kalazhokov
Associate Professor, Department of Data Analysis and Machine Learning
Financial University under the Government of the Russian Federation
Moscow, Russia
ZKHKalazhokov@fa.ru

Yan T. Makoveichuk
Master's student in the Department of Data Analysis and Machine Learning
Financial University under the Government of the Russian Federation
Moscow, Russia
0000-0002-8919-7828

2023 Seminar on Information Computing and Processing (ICP) | 979-8-3503-6983-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICP60417.2023.10397418

***Abstract*— The article describes the experience of teaching students of Big Data analytics-related fields the basics of data science. The teaching methodology used at the Financial University under the Government of the Russian Federation is given, which consists in involvement of students of the group to competitions on Google Kaggle platform. In this methodology, the positive factors influencing the rapid acquisition of initial data science skills are: competitive effect; availability of prepared datasets and problem formulations from the real business sector in various directions; possibility of instant verification of the found solution; practical unlimited number of solutions. These factors contribute to the rapid understanding by students of the need to understand the essence of the problem for the best results, obligatory familiarization with the data, data cleaning and formatting, exploratory data analysis, forming a model with some basic level and its improvement, as well as interpretation.**

Keywords*— *machine learning, training of data science, Google Kaggle, Python, Azure Machine Learning, linear regression, logistic regression, decision trees, clustering, models.

I. Introduction

The process of training master's students to analyze data and create machine-learning models is fraught with some difficulties, primarily because not all students have the mathematical background and Python programming skills needed to get started in learning of data science.

The master's groups of "Applied Computer Science" and "Applied Mathematics and Computer Science" often enroll students who have completed a bachelor's degree in another specialty. Their choice of specialty in the master's program is caused by the desire to expand their knowledge; however, it is often difficult for them to immediately join the learning process at an equal level with more experienced classmates who have gained experience in programming during their undergraduate studies. Consequently, there is a task of training students with different skills, which requires from the teacher a non-trivial approach to teaching in such groups.

The purpose of this article is to present the methodology of teaching Data Science skills to graduate students of the Financial University under the Government of the Russian Federation.

II. Basic Part of the Study

The basis for teaching the discipline is the participation of students of the group in the competition on the Kaggle platform. The competition implies introducing an element of competition into the learning process and undoubtedly provides faster acquisition of data processing skills, building machine-learning models. Students learn to apply in practice the theoretical material, which the teacher explains in class, through participation in the competition.

For more experienced students who have programming skills, they are offered to perform tasks in the Python programming language. For students who do not have the appropriate programming skills, they are offered to perform their work in the Azure Machine Learning cloud service. This helps students quickly get up to speed on the topic of the class. This approach significantly lowers the entry threshold and gives newcomers more confidence in competing against more experienced classmates. But, of course, it is expected that at the end of the course beginners will also gain skills in analyzing data in Python. To that end, they are provided with explanations and solutions to similar problems in the Python programming language in parallel with their Azure ML assignments.

Working in Azure ML is a graphical interface right in the developer's browser. The interface is intuitive and has accessible documentation in English. To familiarize students with this tool, it is useful to perform with them a simple task to create a model for predicting the probability of survival on the Titanic liner. A variant of model building for this task is shown in Figure 1.

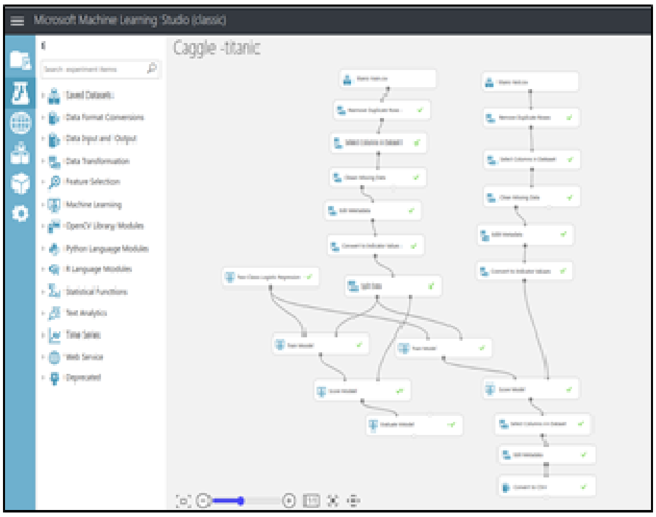


Fig. 1. An example of model building in Azure ML.

In this example, the student is introduced to the basic steps of building machine-learning models by connecting the elements of the constructor directly on his or her computer screen, without having to know programming. In the process, the student is introduced to the program interface, and practice shows that after one or two sessions, the student is already building similar models independently and improving them.

Solving such a task, students go from the very beginning of creating and testing the model to its publication on Kaggle, this removes the psychological barrier that is usually present in students before participating in such competitions.

Of course, knowledge must be provided to intuitively understand basic machine learning models such as linear regression, logistic regression, decision trees, clustering, etc., and their preferred choice for data analysis tasks.

Azure ML has a fairly wide arsenal of capabilities, allowing both simple and complex models to be built. You can see in Fig.1 that all the necessary tools are located on the left side, but you don't need all of them to get started. It is necessary to teach all the basic steps to build models.

First, you need to create and save an experiment; all created experiments remain in memory and are displayed as a list. Next, you need to be able to load your datasets, browse through them, and examine the features. Fig. 2 shows a screenshot of the trait viewing process.

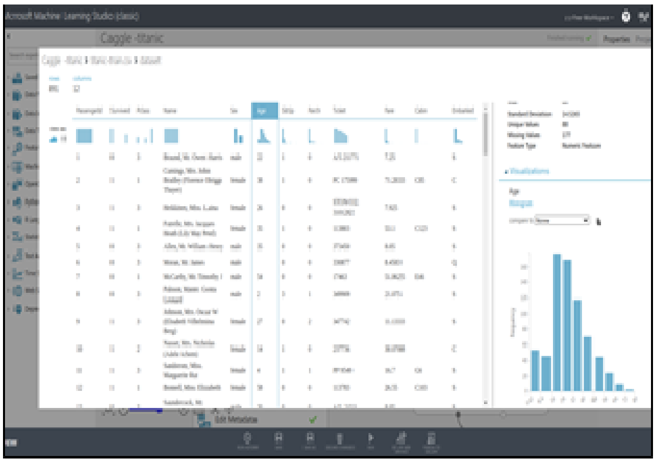
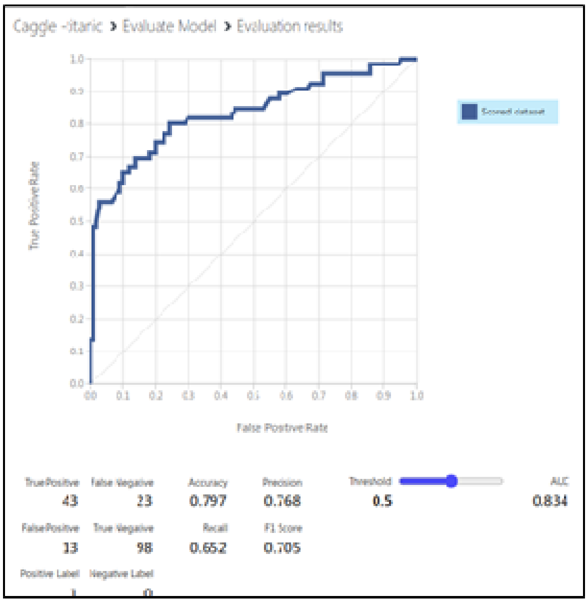


Fig. 2. Dataset Study.

Azure automatically displays all the main statistical characteristics of the trait - mean, standard deviation, number of outliers, maximum and minimum values, total number of rows, etc. A histogram is also displayed on the right-hand side to help visualize the frequency distribution. The student learns to understand the attributes presented in the dataset, to process them, removing duplicates, filling in the gaps with values appropriate in this case - in simple cases it is average, median or mode; he distinguishes different types of attributes - numerical and categorical attributes, features of work with them he draws from the teacher's explanations and from literature [1]. The search for outliers is conveniently carried out with the help of box diagrams, which can be switched to in the interface of the program (Fig. 3). It should be understood that not all outliers can be regarded as outliers.

After analyzing and processing the data, the student must build the model itself. He chooses a suitable machine learning algorithm, divides the data into training and test sets, with the help of programming language libraries performs training and evaluation of the model, the results of the model are also displayed on the screen (Fig.4).

The student is given an understanding of model quality assessment, how to navigate the quality metrics.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
43	23	0.797	0.768	0.5	0.834
False Positive	True Negative	Recall	F1 Score		
13	98	0.652	0.705		
Positive Label	Negative Label				
1	0				

Fig. 3. Box diagram.

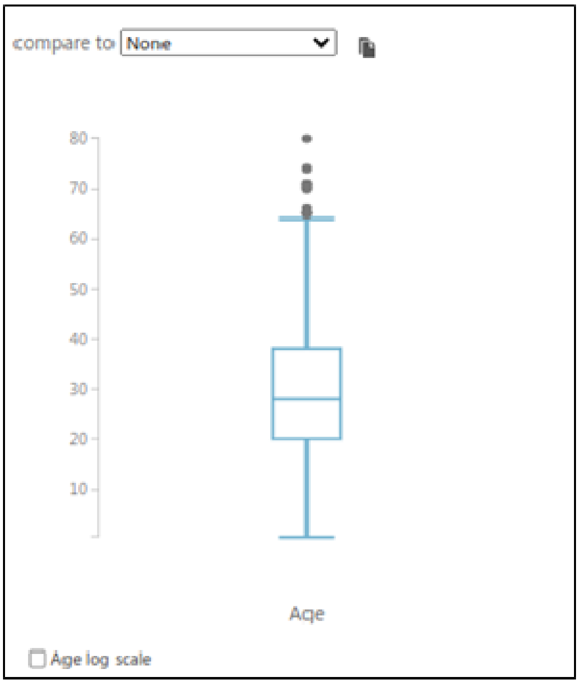


Fig. 4. Model results.

Often in their work, students confuse the proposed test set that they download from Kaggle with the test set that we get when we partition the training set. It should be understood that we get the test set when partitioning ourselves to evaluate the performance of the machine-learning model, while the test set from Kaggle is needed for its processing by the model created and predicting the result.

To avoid confusion, it is useful to process this set with Kaggle in a parallel branch and subsequently evaluate it with the created model, while preparing data in the form required by the task description, usually it is a CSV file with two columns. Azure allows you to prepare such a file also without writing any code.

Azure ML is undoubtedly a useful tool for an easy start in machine learning, but it should be understood that a real data scientist must be able to do the same operations in programming languages such as Python or R.

Practice has shown that after successfully mastering

Azure ML, many students show interest and move on to programming themselves, feeling more confident already. To do this, it is useful to review with them examples of the same problems and other interesting problems from proven books on machine learning [2].

Also, both in the development of teaching methodology, faculty members and graduate students themselves would benefit from exposure to cutting-edge research in the field [3 - 7].

While working on Azure the student gets used to the basic steps of data processing, model training, but in parallel it is useful to parse these same solvable problems with them in the Python programming language. Perhaps it is easier for beginners to work in such development environments as Jupyter Notebook or Google Colaboratory, but there are no restrictions here, someone is more comfortable to work in other environments.

III. CONCLUSIONS

Based on the above, we can draw the following conclusions.

1. Getting started in Data Science assumes a good knowledge base in programming and mathematics, but having a cloud service such as Azure ML makes the threshold of entry into the field much easier.

2. The Data Science learning methodology assumes:

- selecting an available tool (Azure ML);
- familiarity with machine learning models;
- learning the principles of working with data (cleaning, filling gaps, one-time-encoding, and other processing);
- acquiring skills in working with model libraries, training and testing them;
- practice on typical examples of dataset analysis.

3. After starting work in Azure ML, a transition to model development in advanced machine learning programming languages (Python, R) is still needed.

REFERENCES

[1] Analiz dannyh v jekonomike. Teorija verojatnostej, prikladnaja statistika, obrabotka i vizualizacija dannyh v Microsoft Excel : uchebnik / V. I. Solov'ev. - Moskva : KNORUS, 2019. (In Russian)

[2] Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition. Sebastian Raschka, Vahid Mirjalili, Packt Publishing, 2019.

[3] Mange, J. Effect of Training Data Order for Machine Learning / 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 406-407, DOI: 10.1109/CSCI49370.2019.00078.

[4] Guezzaz, A. Asimi, Y. Azrour, M. and Asimi, A. Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection / Big Data Mining and Analytics, vol. 4, no. 1, pp. 18-24, March 2021, DOI: 10.26599/BDMA.2020.9020019.

[5] Doku, R. Rawat, D. B. and Liu, C. Towards Federated Learning Approach to Determine Data Relevance in Big Data / 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), 2019, pp. 184-192, DOI: 10.1109/IRI.2019.00039.

[6] Selvi, G. et al. Automated Machine Learning Platform / 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 769-774,
DOI: 10.1109/UBMK52708.2021.9558961.

[7] Al Heeti, F. and Ilyas, M. Comparative analysis of convolutional neural network architectures for classification of plant leaf diseases / 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), 2022, pp. 1-5, DOI: 10.1109/ICMI55296.2022.9873752.