DEVELOPMENT OF SOFTWARE COMPONENTS FOR EXPLORATORY DATA ANALYSIS: TOWARDS A FLEXIBLE PIPELINE

Shchogolev K., Cherednichenko O.
National Technical University «Kharkiv Polytechnic Institute», Kharkiv

Exploratory Data Analysis (EDA) is a crucial step in the data science workflow, enabling users to uncover patterns, detect anomalies, and test hypotheses prior to formal modeling. Despite its importance, EDA often requires a strong programming background and familiarity with data processing libraries and tools, making it less accessible to non-technical users. The development of configurable, modular software components for EDA is therefore a highly relevant and timely challenge, especially in light of the growing democratization of data analytics.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have led to the creation of several automated and semi-automated tools aimed at simplifying data analysis. Platforms such as Google's AutoML, Microsoft Azure ML Studio, and open-source libraries like Auto-sklearn and DataPrep offer partially automated solutions for EDA, often incorporating intelligent feature selection, statistical summarization, and visualization capabilities. However, these tools typically assume a baseline level of technical proficiency, and they may lack the flexibility or transparency needed for custom exploratory workflows.

This study focuses on the design and implementation of a flexible, modular EDA pipeline that can be easily customized for standard data analysis scenarios without requiring advanced programming skills. The core idea is to develop a set of interoperable software components that users can configure through a simple interface or a configuration file. The system is designed to guide users through common EDA tasks such as data cleaning, summarization, visualization, and feature engineering, with the option to insert custom logic where needed.

The experimental part of this work involves evaluating the ease of use and adaptability of the proposed pipeline. Specifically, we assess the time and computational resources required to assemble and execute custom EDA workflows using the system. Feedback from target users - data analysts with limited programming experience - provides insight into the practical usability and learning curve associated with the tool.

Future development will focus on building a graphical user interface (GUI) that further lowers the entry barrier for non-programmers and supports interactive exploration. Additionally, deployment as a cloud-based service is planned, allowing users to run EDA tasks remotely without local setup or dependency management. This research aims to bridge the gap between powerful EDA capabilities and ease of use, promoting broader access to data-driven decision making.