Data Pulse

Automating Data Profiling and Machine Learning Model Selection

* Sarfaraz Ahmed Sattar Natha
Department of Software Engineering
Sir Syed University of Engineering & Technology
Karachi, Pakistan
sasattar@ssuet.edu.pk

* Abdul Moiz Chishti
Department of Software Engineering
Sir Syed University of Engineering & Technology
Karachi, Pakistan
se20f-022@ssuet.edu.pk

* Syed Abdul Aleem
Department of Software Engineering
Sir Syed University of Engineering & Technology
Karachi, Pakistan
se20f-042@ssuet.edu.pk

* Shaheer Khan Qureshi
Department of Software Engineering
Sir Syed University of Engineering & Technology
Karachi, Pakistan
se20f-003@ssuet.edu.pk

Abstract

Over the last decade, this paradigm has been on the trend for automation Tasks that are located at a higher hierarchical level in the process of machine learning Thus, the application of (ML) has received a lot of attention in recent years. due to the progress in optimization techniques that have reached highly efficient standards as far as the choice of ML models and algorithms. The fast advancement of technology, especially in artificial intelligence (AI) and other such fields where coding is dominant. Machine learning or commonly referred to as: ML, has changed the way organizations manage and analyze data. Automation information tools have now made what used to be complicated and time consuming to be easy. that become facilitating points to increase speed, precision and voluntariness. scalable data-driven decision-making. The expansion of the volume and variety of data. In this regard across industries there is a call for the sophisticated tools. for effective data analysis and the choosing of the ML model. Traditional methods require substantial experience and most of the times labor, generating barriers for organizations without specialized knowledge. To address these difficulties, Data Pulse presents an automatic system that helps to carry out a data profiling operation. and model selection, makes these improve the accessibility and efficiency of the related business processes.

1. Introduction

The introduction gives an overview of the application, highlighting its purpose of simplifying

data exploration and model building. It aims to provide users with a streamlined interface for uploading datasets, performing exploratory data analysis (EDA), and building machine learning models without extensive coding requirements.

Today, data analysis and profiling are integral to various industries, including finance, healthcare, retail, and more. Organizations leverage data to gain insights, make informed decisions, and drive innovation. However, several challenges persist:

- Data Volume
- Data Variety
- Data Quality.
- Complexity of Model Selection

2. Problem Statement

Based on the existing situation, where it is increasingly hard to work with big amounts of data, the challenge of ordering these data and being able to find something valuable in it becomes paramount. The complexity lies not not only in analysis and categorization of this data but also in deciding on which machines to utilize. mathematical learning models to get new predictions or classifications of certain values in the datasets. It has also been noted that
there is lack of efficient organisational structure. framework for the process of data profiling inautomation, the comparison of the models, and the easy availability of the developed model expertise.

retrieval impacts the application of machine learning in various fields due to the following reasons. This project thus presents the critical issues concerning data analysis and machine learning. model selection. In particular, it can help address the following questions

3. Methodology:

In alignment with all the above-highlighted challenges, we structured, designed, and developed an all-in-one solution **Data Pulse.**

The Data Pulse project addresses these challenges by automating the data profiling and machine learning model selection processes. By simplifying these tasks, Data
Pulse aims to make data-driven decision-making more accessible and efficient, allowing organizations to leverage their data without needing extensive technical expertise.

This platform performs detailed data profiling to understand the structure, quality, and key characteristics of datasets. It evaluates multiple machine learning models using various performance metrics, such as accuracy, precision, recall, and F1 score, to identify the optimal model for the given data. Users receive comprehensive reports summarizing the findings and recommendations, facilitating informed decision-making.

3.1 Purpose and Objectives

The primary objective of Data Pulse is to develop an automated platform that:

Performs Detailed Data Profiling: Analyzes datasets to understand their structure, quality, and key characteristics.

- Compares Machine Learning Models: Evaluates multiple models using various performance metrics to identify the best fit for the data.
- Generates Comprehensive Reports: Provides users with actionable insights and recommendations through detailed reports.
- Facilitates Model Deployment: Allows users to download the selected model in a deployable format, streamlining integration into existing workflows.

By achieving these objectives, Data Pulse aims to democratize access to advanced data analytics and machine learning capabilities, empowering organizations to make data-driven decisions with ease and confidence.

4. Literature Review

4.1 Literature Review

This paper specifies the literature review of Data Pulse project aiming at analyzing prior research and approaches in the areas of data profiling, EDA, and the selection of a machine learning model. and the topics related to AutoML and automated machine learning. Its goal
is to offer a detailed perspective or a base to start from in regard to a certain subject. on the status of these areas and their relevance to the creation of an end-to-end platform that helps in the generation of features and machine learning model selection process.

4.1.1 Data Profiling

Data profiling is an essential task under the data preprocessing phase and entails the analysis of data. was as follows: Organizational learning is what is hypothesized to make structure possible; the CIs generate question; the existing sources would then be used to determine the structure and content of the question as well as their relationship to the hypothesis. Techniques such as is data quality assessment, anomaly detection, and statistical analysis among others. There are different tools and frameworks that help to operate big data, including tools like Talend and tools like IBM InfoSphere, each of which is different from Apache Griffin. functionalities for data profiling. Nevertheless, these tools can be highly dependent on manual interactions and domain knowledge meaning that there is a requirement for automation.

4.1.2 Exploratory Data Analysis (EDA)

EDA includes procedures of summarizing the key features of datasets and frequently utilizes visual presentation. In this case, it seeks to identify trends, identify outliers, use / test hypotheses as well as verify assumptions. Classical Analysing the papers written by Tukey (1977), stress on the role of EDA in analysing data. Modern EDA Techniques use modern tools and libraries for visualization such as Matplotlib, Seaborn, and Plotly, that makes smart data navigation possible. Despite these advancements, integrating The biggest problem with EDA is probably the conversion of EDA into a comprehensive automated tool.

4.1.3 Machine Learning Model Selection

Appropriate selection models of machine learning is an essential component in

determining the potential of predictive analytics. It centers on comparing different algorithms on grounds process of model selection is often complex and time-consuming, necessitating automated approaches. of their assessment criteria it encompasses. such as, accuracy, precision, recall, F1-score, and ROC curves. Research by Khan, Shahidul Islam and Hoque, Abu Sayed Md Latiful. (2020) [1] the other hand supports the opinion of Amiri, Zahra and Heidari, Arash and Navimipour, Nima Jafari and Esmaeilpour, Mansour and Yazdani, Yalda (2024). [2] Comparative analyses of the various machine learning strategies are illustrated in Khan, Shahidul Islam and Hoque, Abu Sayed Md Latiful. algorithms. However, the selection of models is somewhat a laborious and lengthy procedure. necessitating automated approaches

4.1.4 Automated Machine Learning (AutoML)

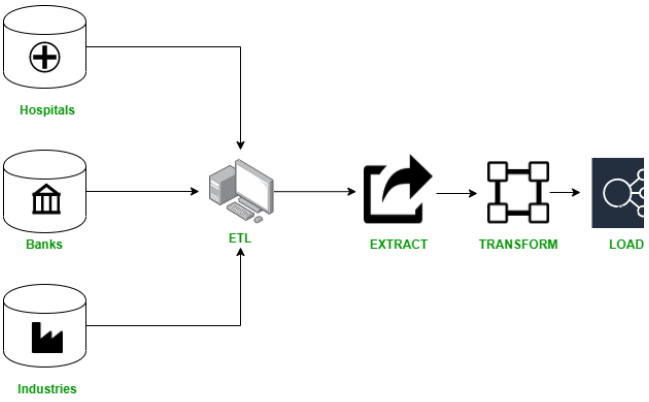
AutoML objectives to automate the give up-to-give up system of applying system learning to actual-world problems. It includes steps like facts preprocessing, function choice, version schooling, hyperparameter optimization, and model evaluation. Notable AutoML frameworks which include Google AutoML, H2O.Ai, and Auto-sklearn have made vast strides in this area. These frameworks aim to democratize system getting to know with the aid of permitting non experts to build strong fashions. However, integrating comprehensive information profiling and version selection right into a single platform remains an evolving location.

4.1.5 Integration and Deployment

The integration of educated gadget gaining knowledge of fashions into applications requires seamless deployment mechanisms. Docker and Kubernetes are generally used for containerizing and orchestrating devices studying fashions. The literature also emphasizes the importance of version interpretability and explainability, as mentioned by means of Shukla, Saurabh and Maheshwari, Arushi and Johri, Prashant. (2021) Comparative analysis of ml algorithms \& stream lit web application. 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) with the LIME framework. Ensuring that fashions aren't best performant however also interpretable is crucial for his or her adoption in sensible programs.

5. Designing and Software

Process Model



5.1 Design Methodology

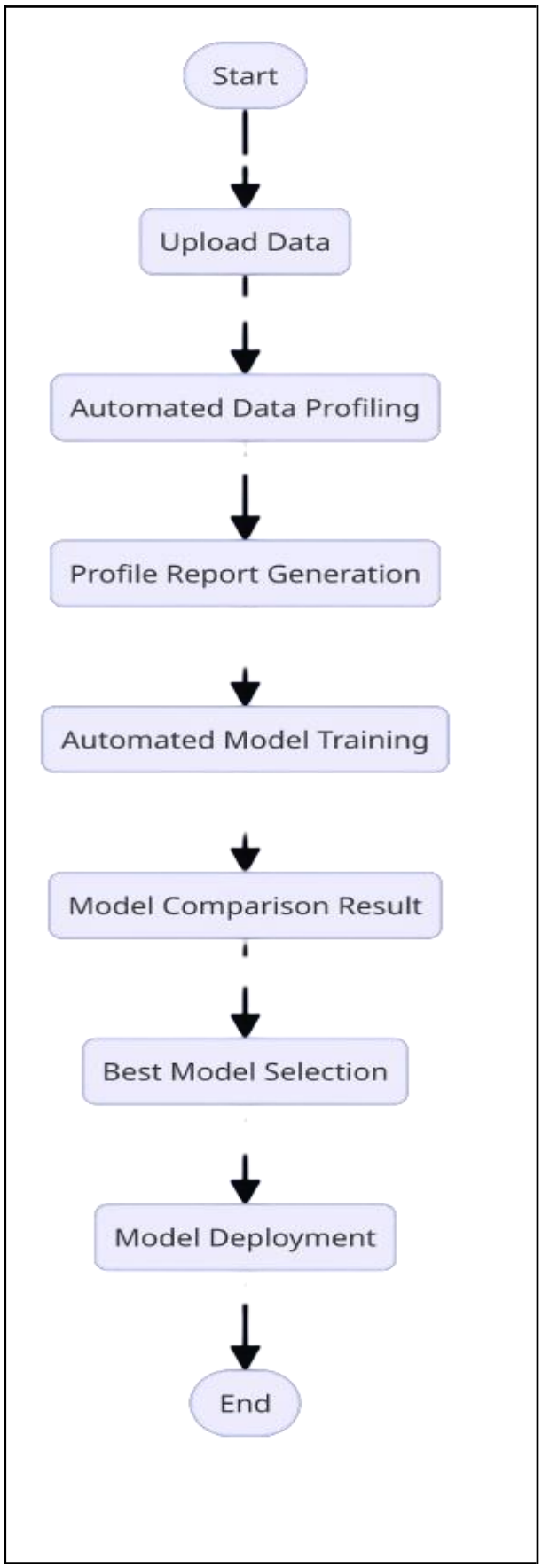
The design methodology for this undertaking follows an iterative and incremental technique, such as Agile or a hybrid Agile methodology. This approach is selected because the venture includes developing an stop-to-cease platform that calls for common feedback from stakeholders and users. Such feedback is essential to make sure that the automatic statistics profiling and system mastering model selection additives correctly meet consumer needs. An iterative approach additionally permits for flexibility in incorporating new features, refining present functionalities, and adapting to converting necessities at some stage in the development system.

5.2 Design Patterns

- **Creational Pattern:** Known also as the Factory Method, it is a concept of object-oriented programming that is used to create objects. Pattern is applied to define the process of making a category of the object. remote machine learning model objects, improving system scalability.

- **Structural Pattern:** The Facade Pattern is among the most important patterns that can be used in any technical task.
  offers a single entity for the desired service, making things easier. increase of the data profiling's complexity and model selection.

- **Behavioral Pattern:** The presented example is a specific type of Strategy Pattern.



5.4 System Flow6. System Building

The system was built using technologies such as Python and Django. Implementation challenges were addressed through solutions and workarounds, ensuring system functionality and reliability.

6.1 Data Dictionary:

The data dictionary provides a detailed description of the database schema, including:

- Tables: Names, purposes, and relationships between tables.
- Fields: Names, data types, constraints (e.g., primary keys, foreign keys), and descriptions of each field.
- Indexes: Indexes used to optimize query performance.
- Constraints: Rules that ensure data integrity, such as unique constraints, check constraints, and foreign key constraints.

By providing a clear and comprehensive data dictionary, the system ensures consistency and clarity in data management and development processes.

6.2 Testing and Evaluation

Testing methods like unit testing, integration testing, and user acceptance testing ensured system performance and reliability. The system met all performance benchmarks, excelling in specific metrics.

6.3 Results and Discussion

The Data Pulse platform successfully automated data profiling and model selection,

offering greater flexibility, scalability, and ease of use compared to existing solutions. It has significant implications for various industries by facilitating model selection and report generation.

7. References

[1] Khan, Shahidul Islam and Hoque, Abu Sayed Md Latiful. (2020). SICE: an improved missing data imputation technique. Journal of big Data, (37--100)

[2] Salekshahrezaee, Zahra and Leevy, Joffrey L and Khoshgoftaar, Taghi M. (2023). The effect of feature extraction and data sampling on credit card fraud detection. Journal of Big Data, (6--50)

[3] Amiri, Zahra and Heidari, Arash and Navimipour, Nima Jafari and Esmaeilpour, Mansour and Yazdani, Yalda (2024). Deep learning applications in IoT-based bio-and medical informatics: a systematic literature review. Neural Computing and Applications, (5757--5797)

[4] Jelena, Mitrovic and others. (2023). Comparative analysis of the traffic accidents in the territory of the city of ice for 2021 and 2022 using open data and the Streamlit application. Vojnotehni glasnik, (616--633)

[5] Shukla, Saurabh and Maheshwari, Arushi and Johri, Prashant. (2021). Comparative analysis of ml algorithms \& stream lit web application. 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), (175--180)

[6] Hong, Jiyeong and Lee, Seoro and Bae, Joo Hyun and Lee, Jimin and Park, Woon Ji and Lee, Dongjun and Kim, Jonggun and Lim, Kyoung Jae (Development and evaluation of the combined machine learning models for the prediction of dam inflow. Water, (2927)

[7] Joel, Luke Oluwaseye and Doorsamy, Wesley and Paul, Babu Sena. (2022) A review of missing data handling techniques for machine learning. International Journal of Innovative Technology and Interdisciplinary Sciences, (971--1005)