Recreating ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning

Aishwariya Ranjan
MS CS
Arizona State University
Tempe, AZ, USA
aranja11@asu.edu

Anvita Lingampalli
MS CS
Arizona State University
Tempe, AZ, USA
alingam1@asu.edu

Andrea Esgar
MS CS
Arizona State University
Tempe, AZ, USA
aesgar@asu.edu

Vihang Pancholi
MS CS
Arizona State University
Tempe, AZ, USA
vpancho1@asu.edu

Aditi Ganapathi
MS CS
Arizona State University
Tempe, AZ, USA
aganap12@asu.edu

ABSTRACT

Although automated machine learning (AutoML) has shown great promise in streamlining the model generation process, current AutoML solutions lack the convenience of use and user control due to the hours of tedious work required to tune hyperparameters and hyperpartitions. In order to overcome these constraints, the initial ATMSeer system was created with a multi-granularity visualization interface that enables users to track the AutoML process in real time, examine the models that are being searched, and adjust the search space.

In this work, we offer our replication of the ATMSeer dashboard, with the goal of introducing the same user control and transparency concepts into AutoML workflows. Important visualization elements seen in our prototype implementation include thorough model inspection panels, hyperparameter tradeoff scatter plots, and algorithm distribution histograms. These interactive visual aids aim to provide users with a thorough grasp of the AutoML search procedure, enabling them to make more educated choices when finalizing the model choice.

We introduced a correlation heatmap as an extension to the existing dashboard to allow users to identify highly correlated features and drop them from the dataset. This would ease the load on the model and save time in data engineering as well.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CSE 578: Data Visualization, Spring 2024, Tempe, Arizona, USA

KEYWORDS

Automated Machine Learning, ATM Hyperparameter Tuning, Dashboard, Data Visualization

ACM Reference format:

Aishwariya Ranjan, Anvita Lingampalli, Andrea Esgar, Vihang Pancholi, and Aditi Ganapathi. 2024. Recreating ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning. In the Data Visualization Final Project (CSE578-Spring 24). ASU, Tempe, AZ, USA, 4 pages.

1 Introduction

Automated Machine Learning (AutoML) has emerged as a valuable tool to simplify model generation in machine learning. Traditionally, determining the optimal models and hyperparameters for a given task has been laborious and time-consuming. AutoML was developed to help with this challenge by automating the tuning of the hyperparameters and selecting the most suitable model for a given project. While AutoML has a lot of potential, the current solutions fall short of providing users with a high level of convenience and control. Even after using AutoML, users often struggle to identify the best model for their specific needs, which can lead to suboptimal results.

To address these challenges, the ATMSeer system was created. This system provides a visualization interface to help users better understand the AutoML process in real time. It allows users to inspect the models, see what hyperparameters are performing the best, and fine-tune the search space to their preferences. The purpose of ATMSeer was to enhance transparency and user control within AutoML workflows.

This study aims to replicate and extend the functionality of the ATMSeer dashboard. Our objective is to create the same level of transparency and user control as created by the ATMSeer system. Some interactive visualization elements that were created to help with this are model inspection panels, hyperparameter tradeoff scatterplots, and algorithm distribution histograms. Our extension plan for this dashboard is to include a correlation heatmap. The heatmap allows users to identify and eliminate features in the datasets to alleviate the computational burden on the model and give a better understanding of the dataset features. We aim to enhance the usability and effectiveness of AutoML workflows, allowing a more efficient and informed model-generation process.

1.1 Paper Summary

The paper chosen to implement, titled "ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning,"[1] introduces ATMSeer, an interactive visualization tool created to increase the efficiency of AutoML processes. Addressing the challenges associated with navigating the model search space in AutoML, the paper acknowledges users' tendencies to distrust automatic results and overspend search budgets.

ATMSeer helps to mitigate these challenges by empowering users to interactively refine the AutoML search space and analyze results using a multi-granularity visualization approach. The creation of the dashboard was informed by interviews with machine learning experts and validated through two case studies, expert interviews, and a user study involving 13 end users.

The tool enables users to monitor the AutoML process, analyze searched models, and refine the search space in real time. It identifies three decisions in the workflow: updating the search space, modifying the computational budgets, and deciding on the model to choose. Evaluations, including the user study, confirm the usability and effectiveness of ATMSeer with interactive visualization elements and real time monitoring capabilities enhancing the informed generation of models. ATMSeer helps to make AutoML workflows more efficient, enabling users with advanced proficiency in machine learning to make superior model selections.

2 Visualization Design

The three primary parts of the ATMSeer interface are the AutoML Profiler, the Overview Panel, and the Control Panel. The Overview Panel offers high-level statistics about the current search process, while the Control Panel lets users manage the AutoML workflow. The AutoML Profiler, the system's central component, has several interactive visualizations that let users examine the distribution of searched algorithms, weigh the performance trade-offs of various hyperparameter setups, and examine the specifics of each model.

The ATMSeer system attempts to address the issues of trust and control that frequently impede the adoption of AutoML in real-world applications by including these visualization techniques. We will look more closely at the ATMSeer visualization design and its guiding principles in the parts that follow by basing our analyses on Munzner's Nested Model.

2.1 Domain Abstraction

ATMSeer is intended for machine learning experts who have had difficulties in the past when developing machine learning models due to laborious and prone to error manual searches. The platform is designed with specific user groups in mind: data scientists who work on model selection and hyperparameter tuning; decision makers who concentrate on strategy development and resource allocation; and machine learning researchers who explore model efficiency and AutoML correctness.

2.2 Data Abstraction

In general, each model in the AutoML process can be treated as a multivariate data point with the following abstraction, as shown in Table I.

Table I: Each feature of the AutoML process visualized on the dashboard

Feature	Type
Algorithm	Categorical {Multi-layer Perceptron, k-Nearest Neighbor, Gaussian Regression, Stochastic Gradient Descent}
Hyperpartition	A set of categorical variables
Hyperparameter	A set of quantitative variables
Performance	Quantitative

More details on the selected datasets and the data abstraction associated with them will be provided in the upcoming Datasets section.

2.3 Task Abstraction

The dashboard created from this project shows high-level insights into the AutoML process through the general summary and top models, which include the best performance score, total number of models, coverage metrics, and performance distribution. Some of its specific tasks have been described below.

2.3.1 Focus Mode. Shows matching algorithms and hyperpartitions to help users focus on the best models.

2.3.2 Decision Support. Helps users decide which models to choose, how much money to spend, and whether to keep going with the AutoML process.

2.3.3 Granular Analysis. Provides three granularity levels to assess AutoML performance in-depth: algorithm, hyperpartition, and hyperparameter.

2.3.4 Overview Mode. Lists out the top-k best performant models for a quick overview when an in-depth analysis is not required for the user.

2.4 Idiom Abstraction

The ATMSeer dashboard utilizes several types of interactive visualizations to provide users with transparency and control over the AutoML process. Table II gives an overview of the idioms used and their purpose.

Table II: Section-wise detail for the dashboard

Dashboard Section		Idiom	Description
a. Control Panel		• Dropdowns • Upload button • Run/resume button	Allows users to upload a new dataset or select an existing dataset and create or resume an AutoML process.
b. Overview Panel	**b1.**	• Text summary • Histogram	High-level statistics and performance of all tried models are summarized using metrics.
	b2.	Text list	List of top k models for users to compare and choose.
c. AutoML Profiler	**c1.**	Histogram	Algorithm-wise view
	c2.	Progress bars	Hyperpartition-wise view
	c3.	• Scatterplots • Area plots	Hyperparameter-wise view

The Algorithm Distribution Histogram presents the distribution of different algorithm classes, such as multilayer perceptron, decision trees, and kNN, along with their respective hyperparameter values. Hyperparameter Tradeoff Scatter Plots allow users to investigate performance tradeoffs between different hyperparameter configurations in the searched models. The Model Details Panel offers comprehensive information about individual models, including the algorithm, hyperparameter settings, and performance metrics. These multi-granularity visualizations provide a thorough understanding of the AutoML search process, offering high-level overviews as well as detailed insights at the model level. ATMSeer aims to increase the transparency of the AutoML system and enable users to make informed decisions about refining the search space.

2.5 Extension Plan

The selected extension plan for the existing dashboard was a correlation heatmap for the selected dataset. In data analysis and machine learning, a correlation heatmap between dataset characteristics can be a useful tool for feature reduction and selection. By displaying the direction and degree of correlations between various attributes, this visualization helps researchers find pairs of strongly linked variables. Because strongly linked features may convey comparable information, their presence may imply redundancy within the collection.

Reducing the feature space will simplify the model and lessen the chance of overfitting by choosing one feature from each pair of highly linked characteristics. By concentrating on the most informative features and lowering processing demands, this method can enhance model performance and efficiency.

3 Datasets

The expected type of dataset is a comma-separated value (.csv) file with quantitative features and one categorical feature defining the class of the data point. Currently, the dashboard consists of the following two pre-loaded datasets.

1. The Dry Bean Classification dataset
2. Pollution dataset

These datasets have been discussed in detail in the upcoming subsections.

3.1 Dry Bean Classification Dataset

This is a **static tabular dataset** sourced from Kaggle[2]. Using a high-resolution camera, pictures of 13,611 grains of seven distinct registered dried beans were captured. The grains yielded a total of 16 features: 12 dimensions and 4 shape types. The features are used to determine the target variable, class, which contains the 7 types of beans used in the dataset.

The data abstraction for 5 of the Dry Bean Dataset's 17 features has been performed in Table III shown below.

Table III: Selected features and their type from the dry bean dataset

Feature	Variable Type
Area	Quantitative
Perimeter	Quantitative
Eccentricity	Quantitative
Roundness	Quantitative
Class	Categorical {Seker, Barbunya, Bombay, Cali, Dermosan, Horoz, and Sira}

3.2 Pollution Dataset

This is a **static tabular dataset** used in the original ATMSeer paper, sourced from OpenML[3]. The pollution dataset was created to see the relation between the instabilities of regression estimates with air pollution and mortality. The dataset has a total of 16 quantitative features, including the average annual precipitation, average household size, percentage of people employed in white collar occupations, percentage of families with income less than $3000, and the total age-adjusted mortality rate per 100,00 which is the target variable.

The data abstraction for 5 of the Pollution Dataset's 16 features, for example, is explained in Table IV below.

Table IV: Selected features and their type from the pollution dataset

Feature	Variable Type
Precipitation	Quantitative
Education	Quantitative
Density	Quantitative
Humidity	Quantitative
Mortality	Quantitative

4 Results

The replication and extension of the ATMSeer dashboard has returned promising results that enhances user control and transparency within AutoML workflows. Through the incorporation of interactive visualization elements and real time monitoring capabilities, users can make more informed decisions and have better model performance. With our extension plan of the correlation heatmap, users can determine the important and irrelevant features in a dataset allowing them to better refine the search space.

It's noted that conventional AutoML systems accommodate around 8 to 15 algorithms. These systems typically explore a range of 100 to 400 models during the process. ATMSeer stands out by using a categorical color scheme that effectively represents various machine learning algorithms, ensuring clarity for users. ATMSeer's visualization capabilities highlight how it works well for different real world machine learning tasks. Ultimately, compared to baseline AutoML approaches, this dashboard has a notable improvement in understanding the process, as well as computational efficiency and user satisfaction.

4.1 Case Study: Dry Bean Classification

In the first case study, we focused on a classification task using the Dry Bean dataset. The objective was to use our replicated dashboard to explore various model options, hyperparameter configurations, and feature selections to optimize the classification accuracy of different types of beans.

The correlation heatmap extension improved feature selection by identifying highly correlated features and redundant ones to know what features to keep and remove. This helps to improve model efficiency and reduce overfitting. Through interactive visualizations like hyperparameter tradeoff scatterplots and algorithm distribution histograms, we were able to compare the performance of different machine learning algorithms with multiple hyperparameter configurations.

By applying these usages of the dashboard, we observed that the best classifier was k-Nearest Neighbor with $k = 11$, leaf size = 17, and $p = 2$ when tested against the F-Score metric. We were able to look at multiple different algorithms with many hyperparameters, and easily see which models would fit best to our dataset task.

5 Conclusion

We presented our replication of the ATMSeer dashboard in this study, a visualization-driven solution intended to tackle the shortcomings of user control and search space in AutoML workflows. Our prototype implementation incorporates a suite of interactive visualizations that let users examine the searched models, monitor the AutoML process, and narrow down the search space—all while drawing inspiration from the original ATMSeer research.

Users are given a thorough knowledge of the intricate AutoML search process by the main visualization elements, which include the detailed model inspection panels, hyperparameter tradeoff scatter plots, and algorithm distribution histograms. Our reconstructed ATMSeer dashboard seeks to solve the shortcomings of current AutoML tools by giving users the ability to search the search space, compare model performance, and guide the AutoML system toward desired conclusions.

By expanding on the fundamental research and design tenets of the initial ATMSeer system, we want to support the continuous endeavors to render AutoML more reliable and accessible to a diverse group of practitioners and subject matter experts.

5 Future Scope

Many of the essential visualization and interaction capabilities of the redesigned ATMSeer dashboard are included in the current prototype, but there are a few crucial areas that still offer room for growth and development. The inability to stream and monitor in real time is a significant drawback. To fully achieve the transparency and control objectives of the ATMSeer architecture, it would be imperative to integrate real-time data updates and seamless model search monitoring. Furthermore, the lack of an automated pipeline for prepping datasets in the current implementation has restricted the system's capacity to handle a variety of user datasets. Creating a solid preprocessing and data ingestion module would improve the ATMSeer dashboard's usefulness and generalizability significantly.

ACKNOWLEDGMENTS

The authors would like to sincerely thank Professor Chris Bryan for his excellent advice and knowledge during the development of the ATMSeer dashboard. We also sincerely appreciate Shubham Chawla, the teaching assistant, for his unceasing efforts in offering timely and constructive feedback that has been crucial in improving the technical parts of our work. The authors also thank the course graders for their invaluable contributions, whose in-depth assessments have strengthened the general quality of the project and helped us discover areas for development.

REFERENCES

[1] Wang, Q., Ming, Y., Jin, Z., Shen, Q., Liu, D., Smith, M. J., ... & Qu, H. 2019. ATMSeer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), May 2019, pp. 1-12.

[2] Koklu, M. and Ozkan, I. A. 2020. Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. *Computers and Electronics in Agriculture* 174 (2020), 105507. DOI: https://doi.org/10.1016/j.compag.2020.105507.

[3] McDonald, G.C. and Schwing, R.C. 1973. Instabilities of Regression Estimates Relating Air Pollution to Mortality. *Technometrics* 15, 3 (1973), 463–481.