Is The Whole Greater Than The Sum Of Its Parts?

Eyal Stolov, Itay Gradenwits

March 11, 2025

Abstract

In this study, we examine whether "the whole is greater than the sum of its parts?" in the context of data integration. Specifically, we investigate whether combining multiple data sources leads to improved predictive performance and deeper insights. To streamline the data science workflow, we leverage automated machine learning (AutoML) tools for pipeline optimization and error analysis. By systematically evaluating models trained on individual and combined datasets, we assess the impact of data fusion on accuracy and interpretability. Our findings highlight the potential of integrating diverse data sources to enhance model robustness and decision-making. For the repository with the complete code and results, please click here.

1 Problem Description

In this study, we aim to improve the data integration and model evaluation stages of the data science pipeline. A key challenge in predictive modeling is determining whether combining multiple data sources leads to better outcomes. Many DS pipelines rely on isolated datasets, which may limit model accuracy and generalizability. However, integrating diverse data sources introduces complexity, such as data inconsistencies, feature redundancy, and increased computational costs.

Additionally, traditional error analysis methods are often manual and time-consuming, making it difficult to identify systematic patterns of failure in model predictions. To address these issues, we employ AutoML tools to automate pipeline optimization and streamline error analysis. By doing so, we aim to enhance the efficiency, interpretability, and overall effectiveness of the DS workflow, enabling data scientists to derive more reliable insights with reduced effort.

2 Solution Overview

2.1 Solution Proposal

What we tried, was create a fully automated machine learning (AutoML) pipeline, that can test wether adding more data was the right choice. The pipeline simplifies the process of training classification models by automatically selecting the best-performing model from multiple algorithms. The function takes a dataset, splits it into training, validation, and test sets based on specified ratios, and then runs H2OAutoML to train multiple models, optimizing for a given metric, which for us, was MSE. Additionally, we tried our best to give the user as much control over the AutoML process,including a lot of hyperparameters, like H2O cluster configuration, balance classes, training ratio, and printing final metrics. For example, an optional confusion matrix can be displayed for classification tasks, helping understand prediction distributions across classes. The function also supports plots if asked, including scoring history plots to observe training progression and feature importance plots to highlight the most influential variables in model predictions.

The function returns the best model, its performance metrics, and the AutoML leaderboard, making it a comprehensive solution for rapid model development and evaluation. This automation significantly reduces the manual effort required for hyperparameter tuning and model selection, enabling data scientists and non-experts alike to deploy optimized models with minimal intervention.

2.2 Data Used

For this study, we used **eight datasets**, grouped into four pairs, each containing common columns that allowed for integration. After preprocessing and cleaning, we connected the datasets within each pair to analyze whether data fusion leads to improved predictive performance. The four dataset pairs cover the following topics:

- **Football** – Two datasets containing statistical data on teams, players, and match outcomes.
- **Nobel Prize Winners** – Two datasets with information on Nobel laureates, including award categories and recipient demographics.
- **Video Games** – Two datasets providing sales data, ratings, and metadata related to video game releases.
- **Measures of Happiness in Countries** – Two datasets covering country-level happiness indicators, including economic and social factors.

2.3 Data Preprocessing

Data preprocessing is a crucial step in any data science pipeline, ensuring that datasets are clean, consistent, and ready for analysis. The process typically involves handling missing values, standardizing formats, and integrating multiple sources while preserving valuable information. However, automating data preprocessing is not a trivial task, as each dataset pair presents unique challenges that require manual adjustments and domain-specific decisions. In this study, we followed a structured approach to preprocess and merge the datasets. The key steps included:

1. **Removing unnecessary columns** – Eliminating irrelevant or redundant features to streamline the dataset.
2. **Matching shared column names** – Standardizing naming conventions across datasets to facilitate integration.
3. **Filtering rows** – For example, filtering only 2019 data from the economic dataset to maintain consistency.
4. **Removing unmatched rows** – Ensuring that only records present in both datasets are retained.
5. **Rounding numerical values** – Standardizing numerical features such as scores for consistency.
6. **Combining datasets** – Merging paired datasets while maintaining data integrity.
7. **Feature transformations** – Applying necessary transformations to improve model performance.

Although we aimed to follow this structured approach for all dataset pairs, **each pair required additional, dataset-specific preprocessing steps**. Differences in data structure, feature distributions, and missing values meant that a one-size-fits-all approach was not feasible. This led to the need for custom cleaning operations and manual adjustments for each pair, making full automation challenging.
To illustrate the transformation process, we include an example comparing the raw dataset headers before preprocessing with the combined dataset after integration:

Example of DataFrames:

Name	Platform	Release Date	Meta Score	User Review
Wii Sports	Wii	November 19, 2006	76	8
Mario Kart Wii	Wii	April 27, 2008	82	8
Wii Sports Resort	Wii	July 26, 2009	80	8
New Super Mario Bros.	DS	May 15, 2006	89	8

Figure 1: Video Games User Reviews Dataset

Name	Platform	Year	Genre	...	...	NA_Sales
Wii Sport	Wii	2006	Sprots	...	...	41.49
Mario Kart Wii	Wii	2008	Racing	...	...	15.85
Wii Sports Resort	Wii	2009	Sports	...	...	15.75
New Super Mario Bros.	DS	2006	Platform	...	...	11.38

Figure 2: Video Games Sales Dataset

Name	Platform	Release Date	Meta Score	User Review	...	NA_Sales
Wii Sports	Wii	November 19, 2006	76	8	...	41.49
Mario Kart Wii	Wii	April 27, 2008	82	8	...	15.85
Wii Sports Resort	Wii	July 26, 2009	80	8	...	15.75
New Super Mario Bros.	DS	May 15, 2006	89	8	...	11.38

Figure 3: Video Games Combined Dataset

2.4 Auto ML

Another question we encountered along our research, was "Are present AutoML solutions enough for making a complete ML pipeline?". We tried a few existing tools, including H2O, DataPrep and ydata_profiling. It seems AutoML solutions have made significant advancements in automating various stages of machine learning (ML) workflows in the last few years. But they are not yet sufficient to construct a fully automated end-to-end ML pipeline. Current AutoML frameworks excel in automating tasks like feature selection, model selection, and hyperparameter tuning. They streamline the process of training and evaluating models, significantly reducing the expertise required for model development. However, these solutions often fall short in handling critical pipeline components such as data preprocessing, data augmentation, and domain-specific feature engineering, which require human intuition and expertise.

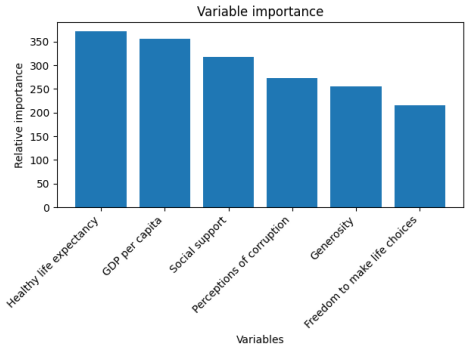
We tried making a complete pipeline that even connects the datasets automatically, but auto data preprocessing tools that are generic enough for all domains are just not developed yet.

2.5 Evaluation Metrics

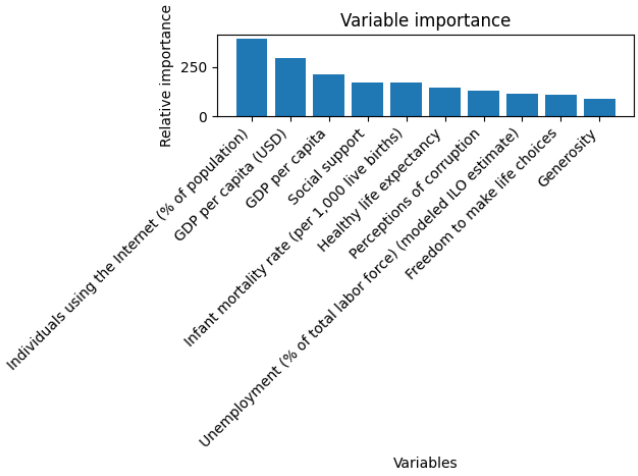
As learened in class, mean squared error(MSE) is a common metric for evaluating models, measuring the average squared difference between actual and predicted values. A lower MSE indicates a model with better predictive accuracy.
In H2O, MSE is automatically computed during model training and evaluation. When an H2O model is trained, it calculates MSE for the training, validation, and test datasets by comparing predicted values with actual outcomes. During AutoML, models are ranked based on a specified metric, and if MSE is chosen as the sorting criterion (in our case this is the chosen metric, but we added it as a user decision if the user wants to change it), H2O selects the model with the lowest MSE as the best performing model. This ensures that the chosen model minimizes prediction errors, improving its reliability in regression tasks.

3 Experimental evaluation

3.0.1 Feature Importance



(a) First image



(b) Second image

Figure 4: Two images side by side

The first plot (a) , which represents the standalone happiness model, highlights variables such as "Healthy life expectancy", "GDP per capita", and "Social support" as the most influential predictors. This indicates that well-being factors and economic prosperity directly impact happiness, as expected.
The second plot (b) , generated from the connected model that incorporates economic data, reveals a shift in variable importance. The top contributing factors now include additional economic indicators such as "Individuals using the Internet (% of population)", "Infant mortality rate", and "Unemployment rate". These new features suggest that access to technology, healthcare, and employment also play a crucial role in predicting happiness when economic data is considered. However, the impact of some previously dominant factors, such as "Healthy life expectancy" and Social support", appears to be relatively diminished.
These results imply that integrating economic data introduces new influential variables that change the model's predictive focus. While the standalone model relies on direct well-being measures, the combined dataset emphasizes broader socioeconomic factors. This highlights the importance of selecting the right feature set depending on the research objective: whether to prioritize immediate well-being determinants or a more comprehensive socioeconomic perspective on happiness.

3.0.2 Models Evaluation

Name	Standalone df MSE	Combined df MSE
Happiness	0.33	0.34
Novel Prize	0.58	0.60
Football	0.80	0.03
Video Games	0.48	0.49

Figure 5: Results

The experimental evaluation compares the Mean Squared Error (MSE) of models trained on standalone datasets versus models trained on a combined dataset within the same domain. The results indicate that for most categories, the difference in MSE between standalone and combined datasets is minimal, suggesting that merging datasets does not significantly improve predictive accuracy in these cases. For example, in the "Happiness" and "Video Games" categories, the difference between standalone and combined MSE is negligible (0.33 vs. 0.34 and 0.48 vs. 0.49, respectively), indicating that the additional data does not enhance the model's performance.

However, in the "Football" category, the MSE significantly decreases when using the combined dataset (0.80 to 0.03), showing that additional domain-related data improves predictive accuracy. This suggests that in some cases, integrating data from related sources provides more relevant information, reducing error rates and making the model more robust. Conversely, the "Novel Prize" category shows a slight increase in MSE (0.58 to 0.60), indicating that the additional data might introduce noise rather than useful patterns.

4 Related Work

5 Conclusions

Overall, the results suggest that combining datasets within the same domain can either improve, or have minimal impact on prediction accuracy depending on the context. The effectiveness of merging datasets appears to be domain-specific, requiring careful evaluation to determine whether the additional information is beneficial or redundant. The AutoML pipeline effectively identifies these differences, highlighting the need for further experimentation to optimize dataset integration strategies.

References