Exceptional Model Mining on Model Residuals: Balancing Interpretability and Expressiveness in Rich Description Languages

Interpretable Insights into Black-Box Failures

Aniket Mishra
a.mishra3@student.tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Cristiana Cărbunaru
c.carbunaru@student.tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Abstract

This paper explores **Exceptional Model Mining (EMM)** and **Subgroup Discovery (SD)** for analyzing black-box regression models through their residuals. The goal of this study is to discover interpretable subgroups where the model exhibits systematic over- or under-prediction, revealing regions of poor generalization or bias. Then it is checked how 4 rich description languages: conjunctive, polynomial, tree-based, and symbolic expressions affect the balance between subgroup expressiveness and interpretability. The methodology frames the problem as residual-based EMM, using a multi-objective search to explore the Pareto frontier between subgroup quality and complexity. A model-agnostic framework is provided for diagnosing systematic black-box failures and identifying which representations best reveal meaningful, human-understandable failure patterns. Results demonstrate that Symbolic Expressions ($\mathcal{L}_{\text{symb}}$) achieve the most optimal quality-interpretability trade-off, capturing highly exceptional regions with minimal complexity, while Conjunctive Expressions ($\mathcal{L}_{conj}$) achieve the highest peak quality across the majority of datasets. This highlights the importance of balancing linguistic conciseness with raw predictive power when diagnosing complex black-box failures.

Keywords

Exceptional Model Mining, Subgroup Discovery, Residual Analysis, Interpretability, Quality

ACM Reference Format:
Aniket Mishra and Cristiana Cărbunaru. 2026. Exceptional Model Mining on Model Residuals: Balancing Interpretability and Expressiveness in Rich Description Languages: Interpretable Insights into Black-Box Failures. In *KDD '26: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 09–13, 2026, Jeju, Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

KDD '26, Jeju, Korea

ACM ISBN 978-1-4503-XXXX-X/2026/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advances in deep learning and ensemble models have produced highly accurate predictive systems, but their opacity makes diagnosing systematic model errors difficult. This presents a critical challenge in modern data science, i.e., diagnosing why a model fails in specific situations. Exceptional Model Mining (EMM) and Subgroup Discovery (SD) offer a framework to localize regions in the data where a model behaves unexpectedly.

This paper explores **residual-based EMM**, where the residuals of a black-box regression model serve as the exceptional target variable. This shift, motivated by real-world instances of localized model failure, provides a direct means of discovering interpretable subgroups that explain systematic model under-performance or bias. For instance, in real-world forecasting applications, model failures are often limited to specific, complex feature interactions. As an example, in an industrial solar power plant model, high residuals may occur only when a conjunction of high ambient temperature (say, $> 60^{\circ}C$) and low humidity is present. Similarly, a wind turbine downtime predictor may fail systematically for a subgroup defined by a highly specific rule, such as $\texttt{Gen Speed} > 300 \wedge \texttt{Wind Speed} \geq 3 \wedge \texttt{Wind Speed} \leq 25 \wedge \texttt{Pitch Angle} \geq 75$.

A challenge in this domain lies in the descriptive power of diagnostic rules. Traditional SD relies on simple conjunctive rules, which often fail to capture the subtle, non-linear feature interactions that cause model errors. Although the EMM literature is extensive, it largely falls into two streams: parameter-based approaches, which focus on deviations in model parameters [5, 10, 21], and residual-based approaches, which are model-agnostic but often restricted to simpler descriptive languages.

Novelty and Contributions. This paper introduces a comparative framework to quantify the trade-off between subgroup quality and interpretability across four distinct rich description languages ($\mathcal{L}$) for residual-based EMM. This addresses the scientific gap regarding which language best balances the expressive power needed to capture complex model failures and the human interpretability needed for actionable diagnosis. A complexity-penalized fitness score is used, and hypotheses across five datasets are empirically tested, showing that highly compact, algebraically rich expressions offer the most diagnostic insight.

2 Theoretical Background

2.1 Subgroup Discovery, Model Residuals, and Rich Descriptions

The study is based on the combined principles of Subgroup Discovery (SD) and Exceptional Model Mining (EMM). SD represents a specialized data mining technique that uniquely combines predictive and descriptive induction to extract rules that describe interesting regions in a dataset with respect to a specific target variable [17, 21]. The resulting patterns extracted, called subgroups, are rules that must be concise, accurate, and easily understood by the human user, that capture local phenomena while placing strong emphasis on interpretability [17].

Instead of targeting the original prediction variable, y, the approach presented in this study uses the black-box model's errors, called **residual**, as the exceptional target [18]. This transition from diagnosing the underlying data to diagnosing the model's behavior is a current practice for modern performance analysis of regression models [30, 32]. By using residuals, interpretable subgroups can be identified where the model consistently exhibits over- or underfitting, revealing systematic biases or a combination of features that cause poor predictions [18, 32]. The utility of this is demonstrated by work that applies SD to output, such as providing interpretable summaries of black-box incident triage decisions [32], or utilizing residual network architectures to enhance interpretability [18].

Classical SD typically uses simple conjunctions of attribute-value tests [17] (e.g., $f_1 > a \wedge f_2 = b$). This maximizes interpretability, but struggles with the non-linear, interacting feature effects that drive model failures. Prior work improved expressiveness mainly inside the same conjunctive template by enriching the atomic predicates, for example, optimal numeric intervals and nominal value sets [24]. In contrast, richer logical description languages $\mathcal{L}$ that can represent interactions more directly are also considered, including shallow decision trees, low-degree polynomial predicates, and compact symbolic expressions [2, 23]. Because numeric handling is central to these constructions, a local discretization strategy guided by a systematic SD framework is adopted [25]. This study aims to discover patterns that are both expressive and sparse, by explicitly constraining model depth/degree/length, and penalizing complexity to preserve interpretability [2], which is particularly important for fairness-aware subgroup analyses [19] and highlighting heterogeneous causal structure where richer rules are needed [23].

Skyline-based Pattern Mining. The concept of *skypatterns* was first introduced as an approach in which patterns are considered non-dominated with respect to multiple evaluation measures under a Pareto dominance relation [33]. Pattern selection is thereby formulated as a skyline query, which removes the need for a single scalar aggregation and promotes diversity among high-quality patterns. Within the context of SD, this concept provides a systematic means of handling multiple, potentially conflicting, objectives such as residual magnitude and rule complexity.

2.2 Exceptional Model Mining

Exceptional Model Mining (EMM) has been formally introduced as a technique focused on detecting subgroups where the local model fitted to the subgroup differs significantly from the model fitted to the global dataset [21].

2.2.1 *EMM for Classification and Model-Induced Targets.*

Initial EMM and subsequent model inspection research has focused substantially on classification domains.

SCaPE (Soft Classifier Performance Evaluation). The SCaPE model class has been developed specifically for EMM in classification tasks. SCaPE treats the classifier probabilities (i.e., soft output) and the binary ground truth as targets, using a quality measure based on ranking loss to highlight subspaces of exceptionally good or poor classifier performance [12].

Conformalized EMM. Another model class-specific EMM variant is the Conformalized EMM framework [9], in which EMM has been integrated with Conformal Prediction, targeting the normalized size of the prediction interval to find subgroups where the model exhibits unusually high or low certainty.

Black-Box Inspection via Randomization. Methods for understanding opaque classifiers such as black-box models using feature randomization and visualization have also been explored [15, 16]. By observing the change in prediction after perturbing feature values, these methods identify groupings of attributes that are interacting non-linearly to influence the prediction, providing a powerful feature interaction analysis method.

2.2.2 *EMM for Regression: Parameter versus Residual Focus.*

The application of EMM to regression models requires careful choice of the exceptionality metric.

Parameter-based EMM. The core framework for EMM on numerical targets [21] was later adapted specifically for linear regression [10], searching for subgroups with exceptional regression coefficients using Cook's distance as the quality measure to find deviations in the local model parameters. Subsequently, this parameter-based paradigm has been extended to Exceptional Growth Mining [27], where local parameters of time-series growth curves are analyzed to study factors like soil characteristics on crop yield variability. Various Goodness-of-Fit metrics have also been explored within this model-parameter paradigm [5].

However, residual-based EMM for regression models remains largely unexplored. The approach described in this paper bridges this gap by explicitly modeling residuals as the exceptional target, enabling systematic analysis of model performance regions in continuous domains.

Multi-Objective EMM. The EMM framework has been further extended through the introduction of *Exceptional Pareto Front Mining* (EPFM) [26]. This formulation integrates principles from multi-objective optimization into EMM by identifying subgroups whose local Pareto fronts differ significantly from the global front. The resulting framework enables the discovery of exceptional trade-offs across several target variables. Although the present study focuses on a single residual-based objective, the EPFM perspective highlights the relevance of treating subgroup evaluation as a multi-criteria process that jointly considers predictive deviation and interpretability.

3 Main Contribution: Residual-based SD/EMM

In contrast to the parametric-centric methods mentioned above (see Section 2.2.2), this study's approach is purely model-agnostic. By using the actual model residual as the exceptional target, the method requires only the black-box prediction ($f(x_i)$) and the true label (y_i). This distinction shifts the diagnostic goal from local variations in model parameters to direct localization of error magnitude, enabling generalized diagnosis of any complex black-box regression model, regardless of its internal structure.

By consistently comparing the abilities of four distinct, constrained description languages, i.e., **Conjunctive**, **Polynomial**, **Decision Tree**, and **Symbolic Expressions**, on a residual-based task, a comprehensive guide for achieving the optimal balance between diagnostic power and interpretability in black-box regression analysis is aimed to be defined.

The central objective is to identify subgroups within a dataset where the prediction quality of a fixed black-box regression model is systematically and exceptionally poor or excellent. This objective is defined mathematically by focusing on the residuals and defining the trade-off with interpretability.

Given a trained black-box model $f : \mathcal{X} \to \mathbb{R}$ and a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, let the residuals be defined by $r_i = (y_i - f(x_i))^2$, capturing pointwise prediction error magnitude for each instance. A subgroup S is a subset of the data defined by a description d, such that $S = \{x \in \mathcal{X} : d(x) = 1\}$. The quality of a subgroup S is quantified by the residual exceptionalness score, $q_{residual}(S)$, which measures the deviation of the subgroup's mean residual from the global mean, normalized by global variance and subgroup size, similar to a standard error:

$$q_{\text{residual}}(S) = \frac{|\mathbb{E}_{x\in S}[r(x)] - \mathbb{E}_{x\in D}[r(x)]|}{\sqrt{\text{Var}_{x\in D}[r(x)]/|S|}}. \tag{1}$$

This metric behaves similarly to a z-score, highlighting subgroups where residuals deviate substantially beyond what is expected by chance, setting the basis for RQ1.

The core task is cast as a multi-objective optimization problem addressing the balance between subgroup quality ($q(d)$) and interpretability ($I(d)$) (RQ2 and RQ3):

$$d^* = \arg\max_{d\in\mathcal{L}} \lambda \cdot q(d) + (1-\lambda)\cdot I(d), \quad \lambda \in [0,1]. \tag{2}$$

Here, λ is a user-controlled parameter balancing the two objectives, with $\lambda \to 1$ emphasizing subgroup quality and $\lambda \to 0$ prioritizing interpretability.

The search space is defined by the descriptive language

$$\mathcal{L} \in \{\mathcal{L}_{\text{conj}}, \mathcal{L}_{\text{poly}}, \mathcal{L}_{\text{tree}}, \mathcal{L}_{\text{symb}}\}.$$

Interpretability $I(d)$ is formalized as an inverse function of complexity, complexity(d):

$$I(d) = \frac{1}{1 + \text{complexity}(d)}. \tag{3}$$

The complexity function:

$$\begin{aligned}\text{complexity}(d) = {} & w_1 \cdot \#\text{predicates}(d) + w_2 \cdot \#\text{operators}(d) \\ & + w_3 \cdot \text{depth}(d) + w_4 \cdot \text{precision}(d),\end{aligned} \tag{4}$$

aggregates measurable aspects of the rule structure, ensuring verifiable scoring across different language types.

The **components** of complexity(d) are defined as follows:

#predicates(d) represents the number of atomic conditions: counts the number of simple boolean tests (e.g., $f > \tau$, $f = v$). Example: (RM > 6.5) ∧ (LSTAT < 10) ⇒ #predicates = 2, representing homes with many rooms in affluent areas.

#operators(d) represents the arithmetic/logical operators used: includes $\{+, -, \times, \div, \wedge, \vee, \leq, >, =\}$. Example: (RM > 7) ∧ (TAX < 300) ⇒ #operators = 3 (>, <, ∧).

depth(d) represents the maximum path length from the root of a decision tree to any leaf, measuring the deepest reasoning chain required. Example: LSTAT < 10 → RM > 6 → Residual > 2 ⇒ depth = 2, indicating a simple two-step rule that identifies under-predicted wealthy areas.

#splits(d) represents the total number of internal decision nodes across the entire tree, reflecting global structural complexity. This metric is used for $\mathcal{L}_{tree}$'s complexity operationalization. Example: a decision tree with three internal splits on RM, LSTAT, and TAX has #splits = 3.

precision(d) represents the threshold resolution penalty: rather than normalizing all numeric attributes, thresholds that are overly precise relative to the natural measurement scale of each feature are penalized. Let each numeric attribute f_j have observed range $R_j = \max(f_j) - \min(f_j)$ and measurement resolution Δ_j (the smallest meaningful increment in the data). For every numeric condition in a description d,

$$\text{precision}(d) = \sum_{j\in\text{num}(d)} \log_{10}\left(\frac{R_j}{\Delta_j}\right)$$

is defined. This term grows when thresholds use unrealistically fine granularity compared to how the variable is measured. For example, if RM ranges from 3 to 9 (so $R = 6$) and is recorded to the nearest 0.1, a threshold such as RM > 6.2 incurs a small penalty, while RM > 6.23 incurs a larger one because it uses unnecessarily fine precision.

#terms(d) represents the number of monomials in a polynomial. This is used for $\mathcal{L}_{poly}$'s complexity operationalization. Example: $a_0 + a_1 f_1 + a_2 f_2 + a_{12} f_1 f_2 \Rightarrow$ #terms = 4. This measures how many interaction components are combined in the polynomial rule.

degree(d) represents the polynomial degree. Example: $0.6 \cdot \text{RM}^2 - 0.4 \cdot$ LSTAT ⇒ degree = 2, capturing non-linear housing relationships.

#tokens(d) representing symbolic token count: Total of operands (variables, constants) and operators. This metric quantifies the symbolic compactness of an expression. Therefore, it is used for $\mathcal{L}_{symb}$'s complexity operationalization. Example: $\frac{\text{RM}}{\text{LSTAT}} > 1.5 \Rightarrow$ #tokens = 5 (RM, LSTAT, 1.5, /, >) [34].

Language-specific Complexity Operationalization:

$$\begin{aligned}
\mathcal{L}_{\text{conj}} &: \text{complexity}(d) = \#\text{predicates}(d). \\
\mathcal{L}_{\text{tree}} &: \text{complexity}(d) = \text{depth}(d) + \#\text{splits}(d). \\
\mathcal{L}_{\text{poly}} &: \text{complexity}(d) = \#\text{terms}(d) + \text{degree}(d). \\
\mathcal{L}_{\text{symb}} &: \text{complexity}(d) = \#\text{tokens}(d) \text{ (total count of operands + operators).}
\end{aligned}$$

3.1 Research Questions and Hypotheses

Building on the research gap and the real-life issues discussed in the introduction (Section 1) and literature review (Section 2), three research questions are proposed to formally define the problem, with RQ1 representing the main question of the study:

RQ1: Can interpretable subgroups be identified where black-box model residuals exhibit exceptional patterns, revealing systematic model failures or biases [18, 34]?

Assuming a user-specified description language $\mathcal{L}$ that determines the form of allowable subgroup descriptions $d \in \mathcal{L}$, the following sub-questions are raised:

RQ2: Which description languages $\mathcal{L}$ (conjunctions, polynomial predicates, shallow decision trees, and symbolic expressions) balance interpretability and expressiveness for capturing non-linear model failure patterns, while optimizing

$$\max_{d \in \mathcal{L}} q(d) = \text{effect size}(r|d)$$

subject to interpretability constraints $|\text{desc}(d)| \leq L$ [1, 3, 6, 8, 13, 28, 29]?

RQ3: How does the expressiveness of the description language $\mathcal{L}$ influence the balance between subgroup quality and interpretability across datasets of varying complexity/domains, and which configurations are optimal [3, 20, 34]?

For **RQ2**, it is important to keep in mind that these description languages form a partial hierarchy of expressiveness (i.e., $\mathcal{L}_{\text{conj}} \subset \mathcal{L}_{\text{tree}}$ when both are bounded by the same depth).

To avoid trivial dominance where a richer language is simply unconstrained, each language is constrained by comparable complexity limits. From experimentation trials, at most five predicates have been chosen for conjunctions, a maximum depth of three for trees, and degree two for polynomials, so improvements in subgroup quality $q(d)$ cannot be attributed solely to unconstrained expressive power.

Based on the research questions, hypotheses are composed as follows:

H1: Richer languages ($\mathcal{L}_{\text{poly}}$, $\mathcal{L}_{\text{tree}}$, $\mathcal{L}_{\text{symb}}$) tend to increase subgroup quality $q(d)$ but decrease interpretability $I(d)$.

H2: Shallow trees or low-degree polynomials may achieve the best balance for practical domains.

The Pareto front $\{(q(d), I(d))\}$ per $\mathcal{L}$ is mapped in order to identify domain-appropriate operating points.

3.2 Solution Approach

3.2.1 Description Language Architectures and Constraints. Four distinct description language classes are implemented in this study, each constrained to ensure the resulting complexity remains comparable and human-interpretable.

Conjunctive ($\mathcal{L}_{\text{conj}}$) uses simple conjunctions of attribute-value conditions (e.g., $\texttt{PTRATIO} > 19.7 \wedge \texttt{LSTAT} \leq 11.45$). The constraint is $\texttt{max}(\#\text{predicates}) \leq 5$.

Polynomial ($\mathcal{L}_{\text{poly}}$) uses low-degree linear or interaction terms as predicates (e.g., $0.6 \cdot \texttt{RM}^2 - 0.4 \cdot \texttt{LSTAT} < 30$). The constraint is $\texttt{max}(\text{degree}) \leq 2$.

Decision Tree ($\mathcal{L}_{\text{tree}}$) relies on subgroups that are paths to leaves within a shallow tree, capturing hierarchical splits (e.g., $\texttt{DIS} > 1.34 \wedge \texttt{RM} \leq 8.28$). The search integrates the **L-A-S-D** numeric handling approach [25] (Local, Adaptive, Stepped/Statistical, Discretisation) to ensure cut points are statistically grounded within each branch, to avoid the explosion of candidate intervals. This strategy is chosen because the Local (L) dimension ensures thresholds are optimized only within the current subgroup's context, while Adaptive (A) splitting means the decision is reapplied at every node. The Stepped/Statistical (S) approach is used because it optimizes the splits based on the q_{residual} objective, focusing the search only on statistically relevant cut points rather than exhaustively testing every value, which provides efficiency and statistical justification. Discretization (D) stands for the conversion of a continuous feature into categorical bins using thresholds. Decision Tree's constraint is $\texttt{max}(\text{depth}) \leq 3$.

Symbolic Expressions ($\mathcal{L}_{\text{symb}}$) employs compact algebraic formulas generated via symbolic regression (e.g., $\frac{\texttt{RM}}{\texttt{LSTAT}} > 1.5$), with constraint $\texttt{max}(\#\text{operators}) \leq 3$; $\texttt{max}(\text{distinct features}) \leq 2$.

The core experimental process involves searching the hypothesis space defined by each $\mathcal{L}$ to find the set of Pareto-optimal rules with respect to maximizing $q(d)$ and maximizing $I(d)$.

3.2.2 RQ1: Residual-based Exceptionalness. Given a trained black-box model $f : \mathcal{X} \to \mathbb{R}$ and a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, with $n = |D|$ (i.e., cardinality), and $S \subseteq D$, which is treated as a *multiset* to allow duplicate or repeated observations, the core task is to identify subgroups [11, 17] $S = \{x \in \mathcal{X} : d(x) = 1\}$ where the model residuals r_i defined as

$$r_i = |y_i - f(x_i)| \quad \text{or} \quad r_i = (y_i - f(x_i))^2$$

exhibit exceptional patterns.

This exceptionality is maximized by finding a subgroup description d that maximizes the residual exceptionalness score $q_{\text{residual}}(S)$, as formalized in Equation (1). $q_{\text{residual}}(S)$ quantifies how **exceptional** the average residual in subgroup S is compared to the overall dataset D. $\mathbb{E}_{x \in S}[r(x)]$ represents the average residual within subgroup S, while $\mathbb{E}_{x \in D}[r(x)]$ is the average residual across the entire dataset D.

A high $q_{\text{residual}}(S)$ shows a subgroup where the model systematically under-performs (large residuals) or over-performs (small residuals), showing bias (underfitting) or overfitting. In this paper's experiments, both directions are explored separately: one list of subgroups where the model systematically *underperforms* (large positive residuals), and another where it *overperforms* (small or negative residuals).

3.2.3 RQ2: Expressive Descriptions. There are considered several hypotheses language classes for interpretable subgroup discovery:

***Conjunctive (baseline)**:*

$$\mathcal{L}_{\text{conj}} = \{f_1\, \theta_1\, v_1 \ \wedge \ \cdots \ \wedge \ f_k\, \theta_k\, v_k\}$$

This language uses simple conjunctions of attribute–value conditions. It is the most interpretable form, expressing subgroups as a set of jointly satisfied predicates.

Example (based on the Boston Housing dataset):

$$(\texttt{RM} > 7) \wedge (\texttt{LSTAT} < 5)$$

identifies a subgroup of expensive houses with many rooms and low lower-status population.

***Polynomial**:*

$$\mathcal{L}_{\text{poly}} = \left\{ a_0 + \sum_i a_i f_i + \sum_{i,j} a_{ij} f_i f_j \ \leq \ \tau \right\}$$

To maintain tractability and interpretability, polynomial predicates are restricted to degree at most two. Coefficients are estimated via least squares over candidate feature pairs, and thresholds are quantized to a fixed grid. This limits the hypothesis space to $O(p^2)$ candidate monomials for p features, avoiding combinatorial explosion.

The language captures nonlinear relationships through polynomial combinations of features [6]. It allows expressing curved decision boundaries or interaction effects.

Example (based on the Boston Housing dataset):

$$0.6 \cdot \texttt{RM}^2 - 0.4 \cdot \texttt{LSTAT} < 30$$

identifies non-linear relation conditions between room count and status ratio that predicts higher prices.

Decision Tree:

$$\mathcal{L}_{\text{tree}} = \{\text{shallow trees with depth } \leq 3\}$$

Subgroups are represented as paths in a decision tree of limited depth. Each path corresponds to a conjunction of conditions learned from data, offering a balance between interpretability and expressiveness [13].

Example path (based on the Boston Housing dataset):

$$\texttt{LSTAT} < 10 \rightarrow \texttt{RM} > 6 \rightarrow \texttt{Residual} > 2$$

defines a subgroup where the model tends to under-predict for affluent neighborhoods with large houses.

Symbolic Expressions:

$$\mathcal{L}_{\text{symb}} = \{\text{combinations of <= and logical operations}\}$$

This language allows flexible symbolic expressions (e.g., ratios, differences, aggregates) generated through symbolic regression [1, 8], providing interpretable yet expressive formulas.

Example (based on the Boston Housing dataset):

$$\frac{\texttt{RM}}{\texttt{LSTAT}} > 1.5$$

indicates houses where the ratio of room count to lower-status percentage is high, showing systematic over-prediction.

The goal is to optimize the multi-objective problem d^* formalized in Equation (2), using $q(d)$ as the subgroup quality [28] and $I(d)$ as the interpretability score [3].

Constraints. The symbolic language $\mathcal{L}_{\text{symb}}$ allows combinations of a maximum of 3 arithmetic and logical operations built from a restricted operator set $\{+, -, \times, \div, \leq, >, =\}$, and up to two distinct features per description/expression. Simple ratio or difference formulas are allowed, such as $\frac{f_i}{f_j} \leq \tau$ or $(f_i - f_j) > \tau$, which cover many interpretable relationships (e.g., efficiency ratios or deviations). Thresholds τ are quantized to a predefined resolution. This keeps the search space finite and the resulting expressions human-readable.

3.2.4 RQ3: Interpretability–Quality Trade-off. Using RQ2's languages ($\mathcal{L}_{\text{conj}}$, $\mathcal{L}_{\text{poly}}$, $\mathcal{L}_{\text{tree}}$, $\mathcal{L}_{\text{symb}}$), a trade-off is formalized between $q(d)$ (quality) and $I(d)$ (interpretability), as defined in Equations (1) and (3) respectively.

Trade-off Objective. The trade-off objective uses the multi-objective function d^* formalized in Equation (2), where the trade-off parameter λ controls the balance between quality (emphasized as $\lambda \rightarrow 1$) and interpretability (emphasized as $\lambda \rightarrow 0$).

Interpretability Model. The Interpretability Model uses complexity(d), as defined in Equation (4), where the specific components and their operationalization for each language are detailed immediately following Equation (4). Then, the interpretability is derived in accordance with Equation (3).

Constraints (i.e., Readability Guards). The following constraints are set: #predicates$(d) \leq 5$ (concision); depth$(d) \leq 3$ for trees, and polynomial degree$(d) \leq 2$. Thresholds are quantized to sensible units (e.g., \$1000 increments for MEDV, 0.1 rooms for RM), to ensure interpretability and prevent overfitting to measurement noise.

At each iteration, candidate subgroups are scored by $q(d)$ and $I(d)$, and Pareto-optimal solutions are retained to visualize the interpretability–quality frontier (see Figure 1). This frontier illustrates the set of non-dominated solutions, showing the maximum achievable subgroup quality for any given level of description complexity. The resulting plot clearly visualizes the trade-off of each language, illustrating, for instance, how $\mathcal{L}_{symb}$ clusters toward the lowest complexity while $\mathcal{L}_{tree}$ explores a wider complexity range.

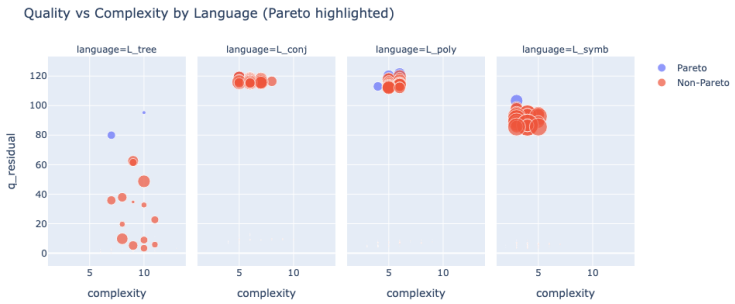

Figure 1: Quality vs. Complexity Tradeoff: Pareto-optimal (Blue) & non-Pareto (red) subgroups.

4 Experimental Setup

Experiments are conducted on five standard regression datasets, selected to vary systematically in terms of dimensionality, size, and feature interpretability.

For a better understanding of the size |S| and context, metadata that provides information about all datasets' structure can be found in Table 1.

The Boston Housing dataset [14] presents low-dimensional, highly interpretable socio-economic attributes, and predicts median home value in \$1000s from socio-economic and housing attributes. Throughout the paper, various Boston Housing features are used for exemplification: RM, representing the average number of rooms per dwelling; LSTAT, the % lower status of the population; MEDV, the median value of owner-occupied homes in \$1000s; TAX, the full-value property-tax rate per \$10,000; PTRATIO, ratio of pupil-teacher by town; and DIS representing the weighted distances to five different Boston employment centres.

auto-mpg [31] has a relatively small size, and it has highly interpretable physical/mechanical descriptors.

CMC (Contraceptive Method Choice) [22] presents mixed categorical and numeric socio-economic features, adapted for regression.

The Forest Fires [7] dataset has a relatively small size, but a high-variance target variable (i.e., area burned).

Year Prediction MSD [4] represents high-dimensional, abstract audio timbre features, serving as a complex benchmark, and it predicts the release year of a song from audio timbre statistics.

For all datasets, a black-box regressor (i.e., Random Forest, which can be extended to XGBoost or Neural Networks) is pre-trained. The pointwise residuals $r_i = (y_i - f(x_i))^2$ are then treated as model-induced numeric targets, and candidate subgroups are evaluated using the residual-exceptionality measure q_{residual} defined in Section 3.2.2, Equation (1). Candidate subgroup descriptions $d \in \{\mathcal{L}_{\text{conj}}, \mathcal{L}_{\text{poly}}, \mathcal{L}_{\text{tree}}, \mathcal{L}_{\text{symb}}\}$ are jointly assessed on subgroup quality $q(d)$ and interpretability $I(d)$ as formalized in Section 3.2.4, Equations (3) and (4). When scalarizing, Equation (2) is followed. The resulting Pareto set $\{(q(d), I(d))\}$ visualizes the interpretability–quality frontier, highlighting both over-performing (low-residual) and under-performing (high-residual) subgroups.

4.1 Code

All datasets and code are available in our GitHub repository[1].

5 Experimental Results

For clarity, the main results are presented in multiple tables, whose outputs were hand-picked by the authors from the top 25 non-dominated rules found for each language and dataset combination. A comprehensive folder of all output files can be found in the paper's GitHub repository[2].

5.1 Exceptional Subgroups (Under-performance: $\mathcal{L}_{\text{conj}}$ Baseline)

The performance baseline (see Table 2) is established by analyzing the top Pareto-optimal rules defining systematic under-performance (i.e., high positive residuals) using the Conjunctive language.

The $\mathcal{L}_{\text{conj}}$ baseline effectively detects highly exceptional failure regions (RQ1), with interpretable rules found in low-dimensional datasets like Boston Housing and complex ones like Forest Fires (i.e., late weekend fires during dry season).

5.2 Polynomial Subgroups ($\mathcal{L}_{\text{poly}}$ Analysis)

The Polynomial language, constrained to a maximum degree of two, aimed to introduce non-linear interaction terms (see Table 3).

The $\mathcal{L}_{\text{poly}}$ rule for Forest Fires shows a simple conjunctive structure being preferred by the algorithm, while the Year Prediction MSD rule demonstrates the language's core advantage by capturing a non-linear interaction term: $\texttt{timbre\_avg\_3} \times \texttt{timbre\_cov\_3\_3} \wedge \texttt{timbre\_avg\_1} \times \texttt{timbre\_avg\_6}$, achieving the highest peak quality of $q_{\text{residual}} = 121.73$ for this complex domain.

5.3 Decision Tree Subgroups ($\mathcal{L}_{\text{tree}}$ Analysis)

The Decision Tree language captures sequential and hierarchical relationships (see Table 4).

The $\mathcal{L}_{\text{tree}}$ rule for Forest Fires is highly specific to a narrow temperature band, showing the tree's ability to partition data based on thresholds, achieving a q_{residual} competitive with $\mathcal{L}_{\text{conj}}$. Notably, $\mathcal{L}_{\text{tree}}$ achieved the highest peak quality on the Year Prediction MSD dataset, with $q_{\text{residual}} = 95.25$.

5.4 Symbolic Expressions Subgroups ($\mathcal{L}_{\text{symb}}$ Analysis)

$\mathcal{L}_{\text{symb}}$ provides the most compact and highly effective rules, demonstrating an excellent trade-off between quality and complexity, particularly on the complex Year Prediction MSD dataset (see Table 5).

The $\mathcal{L}_{\text{symb}}$ rule for Forest Fires is the simplest, with $q_{\text{residual}} = 3.84$, using a single feature predicate ($\texttt{temp} > 24.6$), demonstrating its preference for maximum simplicity.

5.5 Exceptional Subgroups (Over-performance: RQ1 Extension)

Analysis of over-performing rules (i.e., exceptionally small residuals; see Table 6) completes the diagnosis (RQ1).

The consistent discovery of low-complexity rules with significant negative residual deviation demonstrates that the methodology identifies regions where the model is highly accurate, often corresponding to the most common or easily characterized segments of the input space. This completes the **RQ1** objective.

5.6 Trade-off Analysis (RQ2 and RQ3) with Comparative Summary and Discussion

The comparison across all four languages provides a comprehensive answer to RQ2 and RQ3 (see Table 7).

Discussion on RQ2 (Balancing Performance and Interpretability). Hypothesis H2, which suggested that shallow trees or low-degree polynomials offer the best balance, is refuted by the dominance of Symbolic Expressions ($\mathcal{L}_{\text{symb}}$) at the lowest complexity levels. $\mathcal{L}_{\text{symb}}$'s ability to achieve high $q_{residual}$ scores with minimal complexity (e.g., $q_{residual} = 103.25$ with Complexity 3 on Year Prediction MSD in Table 5) shows that a highly constrained search over fundamental arithmetic expressions yields superior results for diagnostic insights in terms of efficiency.

Discussion on RQ3 (Influence of Expressiveness). The efficacy of each language is domain-dependent. $\mathcal{L}_{\text{conj}}$ remains the most robust generalist, succeeding where model errors align with orthogonal features, achieving the highest peak quality on three out of five datasets, such as Boston Housing, auto-mpg and Forest Fires. $\mathcal{L}_{\text{poly}}$ achieves the highest peak quality of $q_{\text{residual}} = 121.73$ on the complex Year Prediction MSD dataset, highlighting its ability to use non-linear interaction terms to capture specific, high-magnitude errors. $\mathcal{L}_{\text{symb}}$ provides the greatest overall gain in insight and performance at high interpretability levels, capturing succinct non-linear functional relationships (e.g., ratios, sums, differences). $\mathcal{L}_{\text{poly}}$ and $\mathcal{L}_{\text{tree}}$ introduce structural complexity that often limits the search's ability to find the absolute exceptionality peak, making them less competitive for this specific residual-based EMM task under tight complexity constraints, even though $\mathcal{L}_{\text{tree}}$ does achieve the peak quality on the CMC dataset ($q_{\text{residual}} = 10.99$).

[1] https://github.com/Aniket-Mishra/Exceptional_Model_Mining_2AMM20.git
[2] https://github.com/Aniket-Mishra/Exceptional_Model_Mining_2AMM20/tree/main/pipeline/outputs

Table 1: Overview of Regression Datasets Used in Experiments (Useful for |S| Context)

Dataset	Samples	Features	Target	Attribute Type	Source
Boston Housing	506	14	MEDV	Socio-economic (numeric)	Scikit-learn dataset
auto-mpg	398	9	class	Physical/mechanical (numeric)	UCI Machine Learning Repository
Contraceptive Method Choice	1,473	10	Wife_Age	Socio-economic (mixed)	UCI Machine Learning Repository
Forest Fires	517	13	area	Environmental (numeric)	UCI Machine Learning Repository
Year Prediction MSD	515,345	90	year	Audio (numeric)	UCI Machine Learning Repository

Table 2: Example Subgroup Descriptions and Statistics - Conjunction Rules ($\mathcal{L}_{\text{conj}}$)

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size ($\lvert S \rvert$)	Avg. Resid.
Boston Housing	PTRATIO $> 19.7 \wedge$ LSTAT $\leq 11.45 \wedge$ DIS ≤ 2.1	14.82	6	8	24.47
auto-mpg	weight $> 2155 \wedge$ model $> 79 \wedge$ cylinders > 4	7.70	4	14	6.99
CMC	Contraceptive_Method $\leq 1 \wedge$ Wife_Education $> 3 \wedge$ Wife_religion > 0	10.56	5	132	37.72
Forest Fires	day $> 6 \wedge$ month $> 8 \wedge$ ISI > 8.56	8.81	4	6	24,018.55
Year Prediction MSD	timbre_avg_6 $> -8.1 \wedge$ timbre_avg_3 $> 16.18 \wedge$ timbre_cov_3_3 > 19.42	119.57	5	65,956	24.63

Table 3: Example Subgroup Descriptions and Statistics - Polynomial ($\mathcal{L}_{\text{poly}}$)

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size ($\lvert S \rvert$)	Avg. Resid.
Boston Housing	DIS $\leq 1.951 \wedge$ LSTAT ≤ 13.66	10.36	5	21	11.38
auto-mpg	horsepower $\times$ weight $> 2.281e+05 \wedge$ model > 79	5.82	5	30	4.13
CMC	Children $\times$ Husband_Occupation $\leq 2 \wedge$ Contraceptive_Method $\times$ Husband_Occupation ≤ 1	9.35	9	95	38.82
Forest Fires	day $> 6 \wedge$ temp > 24.1	5.25	3	16	9,131.29
Year Prediction MSD	timbre_avg_3 $\times$ timbre_cov_3_3 $> 1758 \wedge$ timbre_avg_1 $\times$ timbre_avg_6 > -409.9	121.73	6	61,841	25.30

Table 4: Example Subgroup Descriptions and Statistics - Shallow Decision Tree ($\mathcal{L}_{\text{tree}}$)

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size ($\lvert S \rvert$)	Avg. Resid.
Boston Housing	DIS ≤ 1.34	8.05	3	10	12.63
auto-mpg	model $> 79.5 \wedge$ displacement > 212.5	7.41	3	6	9.77
CMC	Husband_Occupation $\leq 1.5 \wedge$ Contraceptive_Method $\leq 1.5 \wedge$ Children $\leq 1.5 \wedge$ SOLI > 2.5	10.99	11	55	53.01
Forest Fires	temp $> 25.05 \wedge$ temp ≤ 25.45	8.55	4	6	23,331.97
Year Prediction MSD	timbre_avg_6 $> -8.76 \wedge$ timbre_avg_3 $> 26.79 \wedge$ timbre_avg_2 $\leq -54.36 \wedge$ timbre_cov_7_10 ≤ -104.34	95.25	10	7,055	43.39

6 Conclusions

This paper addresses the critical gap in model-agnostic diagnosis by systematically comparing the efficacy of four distinct description languages, i.e., Conjunctive ($\mathcal{L}_{\text{conj}}$), Polynomial ($\mathcal{L}_{\text{poly}}$), Shallow Decision Tree ($\mathcal{L}_{\text{tree}}$) and Symbolic Expressions ($\mathcal{L}_{\text{symb}}$), in finding highly exceptional residuals (see Section 3.1). The outcome of this comparative framework (RQ2, RQ3) identifies the optimal structural approach for maximizing diagnostic power under strict

Table 5: Example Subgroup Descriptions and Statistics - Symbolic ($\mathcal{L}_{\text{symb}}$). Note: For CMC, both rules are equivalent Pareto solutions producing the same results.

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size ($\lvert S \rvert$)	Avg. Resid.
Boston Housing	$(\texttt{CHAS} + \texttt{RAD}) > 24$	8.30	3	8	14.34
auto-mpg	$(\texttt{model} + \texttt{origin}) > 80$	5.55	3	103	2.65
CMC	$\texttt{Wife\_Education} - \texttt{Contraceptive\_Method} > 2$ $\texttt{Wife\_Education} \div \texttt{Contraceptive\_Method} > 3$	8.75	3	175	30.73
Forest Fires	$\texttt{temp} > 24.6$	3.84	2	71	3,537.61
Year Prediction MSD	$(\texttt{timbre\_avg\_6} + \texttt{timbre\_avg\_3}) > 33.77$	103.25	3	73,605	22.24

Table 6: Example Subgroup Descriptions, Metrics, and Interpretations - Shallow Decision Trees ($\mathcal{L}_{\text{tree}}$)

Dataset	Language	Subgroup Description (d)	$q_{\text{residual}}(S)$	Avg. Resid.	Interpretation
Boston Housing	$\mathcal{L}_{\text{tree}}$	$\texttt{DIS} > 1.34 \land \texttt{RM} \leq 8.28 \land \texttt{RM} > 6.77 \land \texttt{INDUS} \leq 6.66$	0.04	1.42	Residential area with moderate rooms, far from employment.
auto-mpg	$\mathcal{L}_{\text{tree}}$	$\texttt{model} \leq 79.5 \land \texttt{displacement} > 90.5 \land \texttt{weight} \leq 3199.0 \land \texttt{acceleration} \leq 13.35$	0.18	1.24	Older, mid-to-heavy cars. The model confidently predicts low mileage.
CMC	$\mathcal{L}_{\text{tree}}$	$\texttt{Husband\_Occupation} \leq 1.5 \land \texttt{Contraceptive\_Method} \leq 1.5 \land \texttt{Children} \leq 1.5 \land \texttt{SOLI} \leq 2.5$	0.38	9.09	Married women with high-mid occupation/method use.
Forest Fires	$\mathcal{L}_{\text{tree}}$	$\texttt{temp} \leq 25.05 \land \texttt{DMC} > 103.55 \land \texttt{DC} > 434.65 \land \texttt{Y} \leq 2.5$	0.12	409.17	Fires occurring under specific low temperature/dry fuel conditions.
Year Prediction MSD	$\mathcal{L}_{\text{tree}}$	$\texttt{timbre\_avg\_6} \leq -8.76 \land \texttt{timbre\_avg\_1} \leq 42.61 \land \texttt{timbre\_cov\_7\_10} > -23.93 \land \texttt{timbre\_avg\_3} > -7.14$	3.40	12.16	Highly constrained, specific area of the timbre space.

Table 7: Comparison of Best Subgroup Metrics (q_{residual}, Interpretability) by Language. Note: Values represent peak performance achieved within each category/dataset.

Dataset	q_{residual} (Language)	Highest Interpretability (Language)	Structural Insight
Boston Housing	14.82 ($\mathcal{L}_{\text{conj}}$)	1.01 ($\mathcal{L}_{\text{poly}}$ / $\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{conj}}$ (spatial socioeconomic topology)
auto-mpg	7.70 ($\mathcal{L}_{\text{conj}}$)	1.16 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{conj}}$ (Hierarchical, mechanical conditions)
CMC	10.99 ($\mathcal{L}_{\text{tree}}$)	1.04 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{tree}}$ (Hierarchical socioeconomic conditions)
Forest Fires	8.81 ($\mathcal{L}_{\text{conj}}$)	0.97 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{conj}}$ (Seasonal and environmental structured)
Year Prediction MSD	121.73 ($\mathcal{L}_{\text{poly}}$)	3.56 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{poly}}$ (Non-linear dependency in bands)

interpretability limits. The experimental results demonstrate that $\mathcal{L}_{\text{symb}}$ offers the most effective balance, achieving peak quality with minimal interpretability cost (i.e., conciseness) across diverse data domains, therefore providing a superior diagnostic tool for complex black-box model failures.

The methodology establishes the diagnosis of black-box regression models as an EMM task targeting model residuals (RQ1) (see Section 3.1). By systematically exploring four distinct description languages (RQ2) (see Section 3.2.3), quantitative evidence for the complex trade-off between expressiveness and interpretability is demonstrated (RQ3). Furthermore, by analyzing both systematic under-performance (failures) and over-performance (overfitting/easy cases), a complete diagnostic tool for model generalization is provided (see Sections 5.1 and 5.5).

The comparative analysis provides that while the Conjunctive language ($\mathcal{L}_{conj}$) achieves the highest peak $q_{residual}$ on the majority of datasets, the most effective language for balancing exceptional subgroup quality (q_{residual}) and human interpretability ($I(d)$) is the Symbolic Expressions language ($\mathcal{L}_{\text{symb}}$). This finding directly rejects Hypothesis H2 (see Section 3.1), which favors moderately rich

languages such as shallow decision trees. The power of $\mathcal{L}_{\text{symb}}$ lies in its ability to achieve strong performance (e.g., $q_{\text{residual}} = 103.25$ on the Year Prediction MSD dataset, reported in Table 5) with minimal complexity, favoring simple arithmetic relationships that are often more insightful than complex conjunctive rules. This result pushes the state-of-the-art by demonstrating that maximal diagnostic insight is achieved not through the complexity of the logical structure (as in $\mathcal{L}_{\text{tree}}$ or $\mathcal{L}_{\text{conj}}$), but through the constrained algebraic conciseness (as discussed in Section 5.6). While the Conjunctive language ($\mathcal{L}_{\text{conj}}$) proves to be highly competitive on lower-dimensional data, $\mathcal{L}_{\text{symb}}$ succeeds in revealing fundamental functional relationships such as ratios, sums, or deviations that capture the underlying nonlinear error topology far more efficiently than complex Boolean conjunctions or fixed-degree polynomials. This conciseness allows the diagnostic tool to extract actionable, foundational knowledge from model failures, addressing the existing gap in the literature regarding the optimal expressive language for model-agnostic error diagnosis.

For future work, an interesting direction would be to validate these findings on real-world industrial streaming data (e.g., wind turbine telemetry), and investigate adaptive complexity constraints that dynamically adjust λ based on the inherent sparsity of the underlying residual error surface.

References

[1] Guilherme Seidyo Imai Aldeia and Fabricio Olivetti de Franca. 2024. Interpretability in Symbolic Regression: a benchmark of Explanatory Methods using the Feynman data set. arXiv:2404.05908 [cs.LG]

[2] Jakob Bach. 2024. Using Constraints to Discover Sparse and Alternative Subgroup Descriptions. *arXiv* (2024). doi:10.48550/arxiv.2406.01411

[3] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. 2020. Model Interpretability through the Lens of Computational Complexity. arXiv:2010.12265 [cs.AI]

[4] Thierry Bertin-Mahieux. 2011. Year Prediction MSD. UCI Machine Learning Repository. doi: https://doi.org/10.24432/C50K61.

[5] Nikki M. Branderhorst. 2021. *Goodness-of-Fit in Exceptional Regression Model Mining*. Master's thesis. Eindhoven University of Technology. Available at https://research.tue.nl/en/studentTheses/goodness-of-fit-in-exceptional-regression-model-mining.

[6] Aleksey Buzmakov. 2020. Towards Polynomial Subgroup Discovery by means of FCA. In *CEUR Workshop Proceedings*, Vol. 2729. 1–12.

[7] Paulo Cortez and Anbal Morais. 2007. Forest Fires. UCI Machine Learning Repository. doi: https://doi.org/10.24432/C5D88D.

[8] Fabrício Olivetti de França, Marco Virgolin, Michael Kommenda, Maimuna S. Majumder, Miles Cranmer, Guilherme Espada, Leon Ingelse, Alcides Fonseca, Mikel Landajuela, Brenden K. Petersen, Ruben Glatt, T. Nathan Mundhenk, Chak Shing Lee, Jacob D. Hochhalter, David L. Randall, Pierre-Alexandre Kamienny, Hengzhe Zhang, Grant Dick, Alessandro Simon, Bogdan Burlacu, Jaan Kasak, Meera Machado, Casper Wilstrup, and William G. La Cava. 2023. Interpretable Symbolic Regression for Data Science: Analysis of the 2022 Competition. arXiv:2304.01117 [cs.LG]

[9] Xin Du, Sikun Yang, Wouter Duivesteijn, and Mykola Pechenizkiy. 2025. Conformalized Exceptional Model Mining: Telling Where Your Model Performs (Not) Well. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2025)*. 528–544. doi:10.1007/978-3-032-06066-2_31

[10] Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. 2012. Different slopes for different folks: mining for exceptional regression models with cook's distance. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. Association for Computing Machinery, 868–876. doi:10.1145/2339530.2339668

[11] Wouter Duivesteijn, Ad J. Feelders, and Arno Knobbe. 2016. Exceptional Model Mining. *Data Mining and Knowledge Discovery* 30, 1 (2016), 47–98. doi:10.1007/s10618-015-0403-4

[12] Wouter Duivesteijn and Julia Thaele. 2014. Understanding Where Your Classifier Does (Not) Work - The SCaPE Model Class for EMM. *Proceedings - IEEE International Conference on Data Mining, ICDM* (01 2014), 809–814. doi:10.1109/ICDM.2014.10

[13] Jack H. Good, Torin Kovach, Kyle Miller, and Artur Dubrawski. 2023. Feature Learning for Interpretable, Performant Decision Trees. In *Advances in Neural Information Processing Systems*, Vol. 36. 1–11.

[14] David Harrison and Daniel L. Rubinfeld. 1978. Hedonic Housing Prices and the Demand for Clean Air. Journal of Environmental Economics and Management, 81–102 pages. Original data.

[15] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery* 28, 5–6 (Sept. 2014), 1503–1529. doi:10.1007/s10618-014-0368-8

[16] Andreas Henelius, Kai Puolamäki, Isak Karlsson, Jing Zhao, Lars Asker, Henrik Boström, and Panagiotis Papapetrou. 2015. GoldenEye++: a Closer Look into the Black Box. In *Statistical Learning and Data Sciences. SLDS 2015. Lecture Notes in Computer Science*, Vol. 9047. 96–105. doi:10.1007/978-3-319-17091-6_5

[17] Francisco Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. 2011. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems* 29, 3 (2011), 495—-525. doi:10.1007/s10115-010-0356-2

[18] Kimia Kamal and Bilal Farooq. 2022. Ordinal-ResLogit: Interpretable Deep Residual Neural Networks for Ordered Choices. arXiv:2204.09187 [cs.LG]

[19] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. An Empirical Study of Rich Subgroup Fairness for Machine Learning. arXiv:1808.08166 [cs.LG] https://arxiv.org/abs/1808.08166

[20] John P. Lalor and Hong Guo. 2022. Measuring algorithmic interpretability: A human-learning-based framework and the corresponding cognitive complexity score. arXiv:2205.10207 [cs.AI]

[21] Dennis Leman, Ad Feelders, and Arno Knobbe. 2008. Exceptional Model Mining. In *Machine Learning and Knowledge Discovery in Databases*, Vol. 24. Springer Berlin Heidelberg, 1–16. doi:10.1007/978-3-540-87481-2_1

[22] Tjen-Sien Lim. 1999. Contraceptive Method Choice. UCI Machine Learning Repository. doi: https://doi.org/10.24432/C59W2D.

[23] Lu Liu. [n. d.]. Causal Learning for Heterogeneous Subgroups Based on Nonlinear Causal Kernel Clustering. https://arxiv.org/html/2501.11622v1#abstract. 2025.

[24] Michael Mampaey, Siegfried Nijssen, Ad Feelders, and Arno Knobbe. 2012. Efficient Algorithms for Finding Richer Subgroup Descriptions in Numeric and Nominal Data. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12)*. IEEE Computer Society, USA, 499–508. doi:10.1109/ICDM.2012.117

[25] Marvin Meeng and Arno Knobbe. 2021. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery* 35 (01 2021), 1–55. doi:10.1007/s10618-020-00703-x

[26] Alexandre Millot, Rémy Cazabet, and Jean-François Boulicaut. 2021. *Exceptional Model Mining meets Multi-objective Optimization*. Society for Industrial and Applied Mathematics, 378–386. arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611976700.43 doi:10.1137/1.9781611976700.43

[27] Puck J. A. M. Mulders, Edwin R. van den Heuvel, Pytrik Reidsma, and Wouter Duivesteijn. 2024. Introducing exceptional growth mining—Analyzing the impact of soil characteristics on on-farm crop growth and yield variability. *PLOS ONE* 19, 1 (01 2024), 1–26. doi:10.1371/journal.pone.0296684

[28] Cristian Munoz, Kleyton da Costa, Bernardo Modenesi, and Adriano Koshiyama. 2023. Evaluating Explainability in Machine Learning Predictions through Explainer-Agnostic Metrics. arXiv:2302.12094 [cs.LG]

[29] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. doi:10.1073/pnas.1900654116

[30] João Pimentel, Paulo J. Azevedo, and Luís Torgo. 2022. Subgroup mining for performance analysis of regression models. *Expert Systems* 40, 1 (2022). doi:10.1111/exsy.13118

[31] Ross Quinlan. 1993. Auto MPG. UCI Machine Learning Repository. doi: https://doi.org/10.24432/C5859H.

[32] Youcef Remil, Anes Bendimerad, Marc Plantevit, Céline Robardet, and Mehdi Kaytoue. 2021. Interpretable Summaries of Black Box Incident Triaging with Subgroup Discovery. arXiv:2108.03013 [cs.AI] https://arxiv.org/abs/2108.03013

[33] Arnaud Soulet, Chedy Raïssi, Marc Plantevit, and Bruno Crémilleux. 2011. Mining Dominant Patterns in the Sky. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM 2011) (ICDM '11)*. IEEE Computer Society, USA, 655–664. doi:10.1109/ICDM.2011.100

[34] Marco Virgolin, Andrea De Lorenzo, Eric Medvet, and Francesca Randone. 2020. Learning a Formula of Interpretability to Learn Interpretable Formulas. In *Parallel Problem Solving from Nature – PPSN XVI*. Springer, 79–93.

Received 24 October 2025; revised TBD; accepted TBD