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. It is then investigated how the following 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 provide the optimal trade-off, achieving the highest quality rules with minimal interpretability cost, and therefore offering a superior diagnostic tool for complex model failures.

Keywords

Exceptional Model Mining, Subgroup Discovery, Residual Analysis, Interpretability

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

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 underperformance or bias. For example, in real-world forecasting applications, model failures are often limited to specific, complex feature interactions. For instance, in an industrial solar power plant model, high residuals might occur only when a conjunction of high ambient temperature (say, $> 60°C$) and low humidity is present. Similarly, a wind turbine downtime predictor might fail systematically for a subgroup defined by a highly specific rule, such as Gen Speed $> 300 \wedge$ Wind Speed $\geq 3 \wedge$ Wind Speed $\leq 25 \wedge$ Pitch Angle ≥ 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 [4, 8, 18], 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 [14, 18].

KDD '26, Jeju, Korea

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

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 [14].

Instead of targeting the original prediction variable, y, the approach presented in this study uses the black-box model's errors, called the **residual**, as the exceptional target [15]. This transition from diagnosing the underlying data to diagnosing the model's behavior is a current practice for modern performance analysis of regression models [25, 26]. By using residuals, interpretable subgroups can be identified where the model consistently exhibits over or under-fitting, revealing systematic biases or combination of features that cause poor predictions [15, 26]. The utility of this is demonstrated by work that applies SD to output, such as providing interpretable summaries of black-box incident triage decisions [26], or utilizing residual network architectures to enhance interpretability [15].

Classical SD typically uses simple conjunctions of attribute–value tests [14] (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 [20]. In contrast, we also consider richer logical description languages $\mathcal{L}$ that can represent interactions more directly, including shallow decision trees, low-degree polynomial predicates, and compact symbolic expressions [2, 19]. Because numeric handling is central to these constructions, we adopt a local discretization strategy guided by a systematic SD framework [21]. Our aim is to discover patterns that are both expressive and sparse: we explicitly constrain model depth/degree/length and penalize complexity to preserve interpretability [2], which is particularly important for fairness-aware subgroup analyses [16] and for highlighting heterogeneous causal structure where richer rules are needed [19].

2.2 Exceptional Model Mining

Exceptional Model Mining (EMM) was 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 [18].

2.2.1 EMM for Classification and Model-Induced Targets. Initial EMM and subsequent model inspection research focused substantially on classification domains.

SCaPE (Soft Classifier Performance Evaluation). The SCaPE model class was 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 [10].

Conformalized EMM. Another model class-specific EMM variant would be the Conformalized EMM framework [7], in which EMM was 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 were also explored [12, 13]. 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 [18] was later adapted specifically for linear regression [8], searching for subgroups with exceptional regression coefficients using Cook's distance as the quality measure to find deviations in the local model parameters. This parameter-based paradigm was later extended to Exceptional Growth Mining [22], 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 [4].

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.

~~3 Problem Statement~~

3~~.1~~ Residual-based SD/EMM and Our Contribution

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 only requires 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: 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. We define this objective mathematically by focusing on the residuals and defining the trade-off with interpretability. Given a trained black-box model $f : \mathcal{X} \rightarrow \mathbb{R}$ and a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, let the residuals be defined by $r_i = |(y_i - f(x_i))^2|$ (for systematic under-performance) or a variation thereof. 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:

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

This metric behaves similarly to a z-score, highlighting subgroups where residuals deviate substantially beyond what is expected by chance (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):

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

Here, λ is a user-controlled parameter balancing the two objectives. 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):

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

The complexity function, $\text{complexity}(d) = w_1 \cdot \#\text{predicates}(d) + w_2 \cdot \#\text{operators}(d) + w_3 \cdot \text{depth}(d) + w_4 \cdot \text{precision}(d)$, aggregates measurable aspects of the rule structure, ensuring auditable scoring across different language types.

3.2 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 we identify interpretable subgroups where black-box model residuals exhibit exceptional patterns, revealing systematic model failures or biases [15, 27]?

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, 5, 6, 11, 23, 24]?

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, 17, 27]?

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, we constrain each language by comparable complexity limits. We have chosen, from experimentation trails, at most five predicates for conjunctions, maximum depth of three for trees, and degree two for polynomials. So that improvements in subgroup quality $q(d)$ cannot be attributed solely to unconstrained expressive power.

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

H1: Richer languages ($\mathcal{L}_{\text{poly}}, \mathcal{L}_{\text{tree}}, \mathcal{L}_{\text{symb}}$) will 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.

We will map the Pareto front $\{(q(d), I(d))\}$ per $\mathcal{L}$ to identify domain-appropriate operating points.

4 Solution Approach

The solution approach relies on a controlled comparative experimental design.

~~4.1 Code~~

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

4.2 Dataset Overview

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

The `Boston Housing` dataset, with 506 samples, 14 features, presents low-dimensional, highly interpretable socio-economic attributes, and predicts median home value in \$1000s from socioeconomic and housing attributes. For reference, some feature examples are: 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 \$1000's; and TAX, being the full-value property-tax rate per \$10,000.

`auto-mpg`, with 398 samples and 8 features, has a small size, and it has highly interpretable physical/mechanical descriptors.

`CMC (Contraceptive Method Choice)`, having 1,473 samples and 10 features, presents mixed categorical and numeric socio-economic features, adapted for regression.

`Forest Fires`, with 517 samples and 13 features, has a relative small size, but a high-variance target variable (i.e., area burned).

`Year Prediction MSD`, with 515,345 samples and 90 features, 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 (Random Forest in this case, which can be extended to XG Boost or Neural Networks) is pre-trained, and EMM is performed on the calculated residuals.

4.3 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.

[1] https://github.com/Aniket-Mishra/Exceptional_Model_Mining_2AMM20.git

Conjunctive ($\mathcal{L}_{\text{conj}}$) uses simple conjunctions of attribute-value conditions (e.g., $\text{PTRATIO} > 19.7 \wedge \text{LSTAT} \le 11.45$). Its complexity is $\text{complexity}(d) = \#\text{predicates}(d)$, and the constraint is Max $\#\text{predicates} \le 5$.

Polynomial ($\mathcal{L}_{\text{poly}}$) uses low-degree linear or interaction terms as predicates (e.g., $0.6 \cdot \text{RM}^2 - 0.4 \cdot \text{LSTAT} < 30$). The complexity is $\text{complexity}(d) = \#\text{terms}(d) + \text{degree}(d)$, and constraint Max degree ≤ 2.

Decision Tree ($\mathcal{L}_{\text{tree}}$) relies on subgroups that are paths to leaves within a shallow tree, capturing hierarchical splits (e.g., $\text{DIS} > 1.34 \wedge \text{RM} \le 8.28$). The search integrates the **L-A-S-D** numeric handling approach [21] (Local, Adaptive, Stepped/Statistical, Discretization) to ensure cut points are statistically grounded. Decision Tree's complexity is $\text{complexity}(d) = \text{depth}(d) + \#\text{splits}(d)$, and its constraint Max depth ≤ 3.

Symbolic Expressions ($\mathcal{L}_{\text{symb}}$) employs compact algebraic formulas generated via symbolic regression (e.g., $\frac{\text{RM}}{\text{LSTAT}} > 1.5$), with complexity $\text{complexity}(d) = \#\text{tokens}(d)$ (total count of operands + operators), and constraint Max operators ≤ 3; max distinct features ≤ 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 the two objectives: maximizing $q(d)$ and maximizing $I(d)$.

5 Experimental Results

5.1 Methodology

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 we treat as a *multiset* to allow duplicate or repeated observations, let residuals be defined as:

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

or some other form of loss calculation.

We aim to find subgroups [9, 14] $S = \{x \in \mathcal{X} : d(x) = 1\}$, where the subgroup description d maximizes the residual exceptionalness score:

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

This measure 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. $\mathbb{E}_{x\in D}[r(x)]$ is the average residual across the entire dataset. The numerator $|\mathbb{E}_{x\in S}[r(x)] - \mathbb{E}_{x\in D}[r(x)]|$ measures the magnitude of deviation between the subgroup and the global average. The denominator $\sqrt{\text{Var}_{x\in D}[r(x)]/|S|}$ acts as a normalization factor, adjusting for global variance and subgroup size, similar to a *standard error*. The $q_{\text{residual}}(S)$ behaves like a z-score, identifying subgroups where residuals deviate substantially beyond what is expected by chance.

A high $q_{\text{residual}}(S)$ would show a subgroup where the model systematically underperforms (large residuals) or overperforms (small residuals), showing bias (underfitting) or overfitting. In our experiments, we will explore both directions separately: one list of subgroups where the model systematically *underperforms* (large positive residuals), and another where it *overperforms* (small or negative residuals).

RQ2: Expressive Descriptions. We consider several hypothesis language classes for interpretable subgroup discovery:

(1) **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 (Boston Housing):

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

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

(2) **Polynomial:**

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

To maintain tractability and interpretability, we restrict polynomial predicates 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 [5]. It allows expressing curved decision boundaries or interaction effects.
Example (based on the Boston Housing dataset):

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

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

(3) **Decision Tree:**

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

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 [11].
Example path (based on the Boston Housing dataset):

$$\text{LSTAT} < 10 \to \text{RM} > 6 \to \text{Residual} > 2$$

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

To handle numeric attributes efficiently, we follow recommendations by Meeng and Knobbe (2021) [21], who systematically evaluated algorithms for dealing with numeric features in SD. We adopt a local discretization approach, determining cut points only where statistically justified within each branch, to avoid the explosion of candidate intervals.

(4) **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, 6], providing interpretable yet expressive formulas.

Example (based on Boston Housing):

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

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

The goal is to optimize the following objective:

$$d^* = \arg\max_{d \in \mathcal{L}} \lambda \cdot q(d) + (1 - \lambda) \cdot I(d)$$

where $q(d)$ is the subgroup quality (effect size on residuals) [23], $I(d)$ is the interpretability score (inverse of description complexity) [3], and λ is the trade-off parameter

Constraints: The symbolic language $\mathcal{L}_{\text{symb}}$ allows combinations of maximum of 3 (we choose 3, jlt) arithmetic and logical operations built from a restricted operator set $\{+, -, \times, \div, \leq, >, =\}$ and up to two distinct features per description/expression. We permit simple ratio or difference formulas 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.

RQ3: Interpretability–Quality Trade-off. Using RQ2's languages ($\mathcal{L}_{\text{conj}}$, $\mathcal{L}_{\text{poly}}$, $\mathcal{L}_{\text{tree}}$, $\mathcal{L}_{\text{symb}}$), we formalize a trade-off between:

- $q(d)$: **quality** of a description d (e.g., residual exceptionalness / effect size on residuals);
- $I(d)$: **interpretability** of d, modeled as a decreasing function of its complexity.

Interpretability Model. Let complexity(d) aggregate simple, auditable counts:

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

where precision(d) penalizes overly precise thresholds (e.g., many decimal places). We instantiate $I(d)$ as

$$I(d) = \frac{1}{1 + \text{complexity}(d)} \quad \text{or} \quad I(d) = \exp\bigl(-\beta \cdot \text{complexity}(d)\bigr),$$

with $\beta > 0$ controlling how sharply interpretability decays with complexity [17].

Complexity Components. We define each component so that interpretability scoring is auditable and intuitive:

- **#predicates(d): number of atomic conditions:** Counts the number of simple boolean tests (e.g., $f > \tau$, $f = v$). Example: $(\text{RM} > 6.5) \wedge (\text{LSTAT} < 10) \Longrightarrow$ #predicates = 2, representing homes with many rooms in affluent areas.
- **#operators(d)**: arithmetic/logical operators used: Includes $\{+, -, \times, \div, \wedge, \vee, \leq, >, =\}$. Example: $(\text{RM} > 7) \wedge (\text{TAX} < 300) \Longrightarrow$ #operators = 3 ($>$, $<$, $\wedge$).
- **depth(d)**: maximum path length from the root of a decision tree to any leaf, measuring the deepest reasoning chain required. Example: $\text{LSTAT} < 10 \longrightarrow \text{RM} > 6 \longrightarrow \text{Residual} > 2 \Longrightarrow$ depth = 2, indicating a simple two-step rule that identifies underpredicted wealthy areas.
- **#splits(d)**: total number of internal decision nodes across the entire tree, reflecting global structural complexity. Example: a decision tree with three internal splits on `RM`, `LSTAT`, and `TAX` has #splits = 3.
- **precision(d): threshold resolution penalty:** Rather than normalizing all numeric attributes, we penalize thresholds that are overly precise relative to the natural measurement scale of each feature. 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, we define

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

  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): number of monomials in a polynomial:** Example: $a_0 + a_1 f_1 + a_2 f_2 + a_{12} f_1 f_2 \Longrightarrow$ #terms = 4. This measures how many interaction components are combined in the polynomial rule.
- **degree(d): polynomial degree:** Example: $0.6 \cdot \text{RM}^2 - 0.4 \cdot \text{LSTAT} \Longrightarrow$ degree = 2, capturing non-linear housing relationships.
- **#tokens(d): symbolic token count:** Total of operands (variables, constants) and operators. This metric quantifies the symbolic compactness of an expression. Example: $\frac{\text{RM}}{\text{LSTAT}} > 1.5 \Longrightarrow$ #tokens = 5 (RM, LSTAT, 1.5, /, >). [27]

We then compute:

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

and derive interpretability:

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

Language-specific Operationalization.

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

Trade-off Objective. We search for descriptions that balance quality and interpretability:

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

The trade-off parameter $\lambda \in [0, 1]$ is user-controlled: $\lambda \rightarrow 1$ emphasizes subgroup quality, while $\lambda \rightarrow 0$ prioritizes interpretability.

Constraints (i.e., Readability Guards).

- #predicates$(d) \leq 5$ (concision);
- depth$(d) \leq 3$ for trees; polynomial degree$(d) \leq 2$;
- thresholds 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.

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

The performance baseline (see Table 1) 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).

5.3 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 2).

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: timbre_avg_3 × timbre_cov_3_3.

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

The Decision Tree language captured sequential and hierarchical relationships (see Table 3).

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 $\mathcal{L}_{\text{conj}}$.

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

$\mathcal{L}_{\text{symb}}$ provided the most compact and highly effective rules, particularly on the complex Year Prediction MSD dataset (see Table 4).

The $\mathcal{L}_{\text{symb}}$ rule for `Forest Fires` is the simplest possible, using a single feature predicate (temp > 24.6), demonstrating its preference for maximum parsimony.

5.6 Exceptional Subgroups (Over-performance: RQ1 Extension)

Analysis of over-performing rules (i.e., exceptionally small residuals; see Table 5) completes the diagnosis (RQ1).

The consistent discovery of low-complexity rules with significant negative residual deviation demonstrates that the methodology successfully 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.

5.7 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 6).

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 ($\mathcal{L}_{\text{symb}}$) in terms of both peak performance and complexity. $\mathcal{L}_{\text{symb}}$'s ability to achieve extremely high $q_{residual}$ scores with minimal complexity (e.g., $q_{residual} = 330.52$ with Complexity 3.0 on Year Prediction MSD) shows that a highly constrained search over fundamental arithmetic expressions yields superior results for diagnostic insights.

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. $\mathcal{L}_{\text{symb}}$ provides the greatest overall gain in insight and performance, achieving the absolute highest q_{residual} score and 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.

6 Conclusions

This work ~~successfully~~ framed the diagnosis of black-box regression models as an Exceptional Model Mining task targeting model residuals (RQ1). By systematically exploring four distinct description languages (RQ2), quantitative evidence for the complex trade-off between expressiveness and interpretability was demonstrated (RQ3). Furthermore, a complete diagnostic tool for model generalization was provided by analyzing both systematic under-performance (i.e., failures) and over-performance (i.e., overfitting/easy cases).

The current findings indicate that 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 rejects the general idea that moderately rich languages like decision trees or polynomials offer the best sweet spot (i.e., hypothesis H2). $\mathcal{L}_{\text{symb}}$ demonstrated an unprecedented ability to capture peak exceptionality with minimal complexity, leveraging simple arithmetic relationships that are often more insightful than complex conjunctive rules. While the Conjunctive language ($\mathcal{L}_{\text{conj}}$) proved highly competitive on lower-dimensional data, $\mathcal{L}_{\text{symb}}$ proved superior in finding both simple and complex, yet concise, patterns.

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

Table 1: Example Subgroup Descriptions and Statistics

Dataset	**Subgroup Description** (d)	$q_{\text{residual}}(S)$	**Complexity**	**Size** ($\lvert S \rvert$)	**Avg. Resid.**
Boston Housing	PTRATIO > 19.7 ∧ LSTAT ≤ 11.45 ∧ DIS ≤ 2.1	14.82	5.0	8	24.47
auto-mpg	weight > 2155 ∧ model > 79 ∧ cylinders > 4	7.70	4.0	14	6.99
CMC	Contraceptive_Method ≤ 1 ∧ Wife_Education > 3 ∧ Wife_religion > 0	10.56	5.0	132	37.72
Forest Fires	day > 6 ∧ month > 8 ∧ ISI > 8.56	8.81	4.0	6	24,018.55
Year Prediction MSD	timbre_avg_6 > −8.1 ∧ timbre_avg_3 > 16.18 ∧ timbre_cov_3_3 > 19.42	119.57	5.0	65,956	24.63

Table 2: Example Subgroup Descriptions and Statistics

Dataset	**Subgroup Description** (d)	$q_{\text{residual}}(S)$	**Complexity**	**Size** ($\lvert S \rvert$)	**Avg. Resid.**
Boston Housing	CHAS > 0 ∧ TAX > 403	8.30	3.0	8	14.34
auto-mpg	model > 80 ∧ cylinders > 4	5.62	3.0	11	5.95
CMC	Contraceptive_Method ≤ 1 ∧ Wife_Education > 3	8.75	4.0	175	30.73
Forest Fires	day > 6 ∧ temp > 24.1	5.25	3.0	16	9,131.29
Year Prediction MSD	timbre_avg_6 > 0.04687 ∧ timbre_avg_3 × timbre_cov_3_3 > 988.5	112.99	4.0	41,227	27.18

Table 3: Example Subgroup Descriptions and Statistics

Dataset	**Subgroup Description** (d)	$q_{\text{residual}}(S)$	**Complexity**	**Size** ($\lvert S \rvert$)	**Avg. Resid.**
Boston Housing	DIS ≤ 1.34	8.05	3.0	10	12.63
auto-mpg	model > 79.5 ∧ displacement > 212.5	7.41	3.0	6	9.77
CMC	Husband_Occupation > 1.5 ∧ Contraceptive_Method > 1.5 ∧ . . .	3.19	8.0	181	6.33
Forest Fires	temp > 25.05 ∧ temp ≤ 25.45	8.55	4.0	6	23,331.97
Year Prediction MSD	timbre_avg_6 > −8.76 ∧ timbre_avg_3 > 26.79 ∧ . . .	79.95	7.0	31,870	24.13

Table 4: Example Subgroup Descriptions and Statistics

Dataset	**Subgroup Description** (d)	$q_{\text{residual}}(S)$	**Complexity**	**Size** ($\lvert S \rvert$)	**Avg. Resid.**
Boston Housing	CHAS + RAD > 24	8.30	3.0	8	14.34
auto-mpg	model > 79	5.42	2.0	89	2.72
CMC	target − Children > 38	11.66	5.0	197	35.32
Forest Fires	temp > 24.6	3.84	2.0	71	3,537.61
Year Prediction MSD	target ≤ 1988	330.52	3.0	75,562	45.31

Table 5: Example Subgroup Descriptions, Metrics, and Interpretations

Dataset	Language	Subgroup Description (d)	$q_{\text{residual}}(S)$	Avg. Resid.	Interpretation
Boston Housing	$\mathcal{L}_{\text{tree}}$	DIS > 1.34 ∧ RM ≤ 8.28 ∧ . . .	2.94	0.79	Residential area with moderate rooms, far from employment.
auto-mpg	$\mathcal{L}_{\text{tree}}$	model ≤ 79.5 ∧ displacement > 90.5 ∧ . . .	2.43	0.18	Older, mid-to-heavy cars. Model confidently predicts low mileage.
CMC	$\mathcal{L}_{\text{tree}}$	Husband_Occupation > 1.5 ∧ Contraceptive_Method > 1.5∧. . .	3.34	3.66	Married women with high-mid occupation/method use.
Forest Fires	$\mathcal{L}_{\text{tree}}$	temp ≤ 25.05 ∧ DMC ≤ 103.55 ∧ . . .	1.16	20.08	Fires occurring under specific low temperature/dry fuel conditions.
Year Prediction MSD	$\mathcal{L}_{\text{tree}}$	timbre_avg_6 ≤ −8.76 ∧ timbre_avg_1 > 42.61 ∧ . . .	62.44	4.10	Highly constrained, specific area of the timbre space.

Table 6: Comparison of Best q_{residual}, Interpretability, and Structural Insight Across Languages

Dataset	Best q_{residual} (Language)	Highest Interpretability (Language)	Best Structural Insight
Boston Housing	14.82 ($\mathcal{L}_{\text{conj}}$)	0.33 ($\mathcal{L}_{\text{symb}}$ / $\mathcal{L}_{\text{tree}}$)	$\mathcal{L}_{\text{conj}}$ (most specific hyperbox)
auto-mpg	7.70 ($\mathcal{L}_{\text{conj}}$)	0.33 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{conj}}$ (multivariate linear boundary)
CMC	11.66 ($\mathcal{L}_{\text{symb}}$)	0.20 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{symb}}$ (difference arithmetic)
Forest Fires	8.81 ($\mathcal{L}_{\text{conj}}$)	0.33 ($\mathcal{L}_{\text{symb}}$ / $\mathcal{L}_{\text{poly}}$)	$\mathcal{L}_{\text{conj}}$ (time-dependent conjunction)
Year Prediction MSD	330.52 ($\mathcal{L}_{\text{symb}}$)	0.25 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{symb}}$ (simple target threshold)

that dynamically adjust L 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] 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.

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

[6] F. O. de Franca, M. Virgolin, M. Kommenda, et al. 2023. Interpretable Symbolic Regression for Data Science: Analysis of the 2022 Competition. arXiv:2304.01117 [cs.LG]

[7] Xin Du, Sikun Yang, Wouter Duivesteijn, and Mykola Pechenizkiy. 2025. *Conformalized Exceptional Model Mining: Telling Where Your Model Performs (Not) Well.* Springer Nature Switzerland, Cham, 528–544. doi:10.1007/978-3-032-06066-2_31

[8] 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

[9] 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

[10] 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

[11] 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.

[12] 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

[13] 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

[14] 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

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

[16] 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

[17] 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]

[18] 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

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

[20] 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

[21] 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

[22] 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

[23] 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]

[24] 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

[25] 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

[26] 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

[27] 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 ??? October 2025; revised TBD; accepted TBD