Data Science Project Protocol

Child Risk Improvement Prediction Using Machine Learning

By

Zohar Or Sharvit

Introduction

1. 360 National Program

The *360° National Program for Children and Youth at Risk* is Israel's flagship inter-ministerial initiative for identifying, supporting, and promoting the well-being of children and adolescents living in risk situations. Established in 2008 by Government Resolution No. 3070, following the *Schmid Committee Report*, the program represents a comprehensive national effort to transform how Israeli society addresses child and youth risk.

Led by the Ministry of Welfare and Social Security in collaboration with the Ministries of Education, Health, Aliyah and Integration, and National Security, and in partnership with JDC Israel and the Federation of Local Authorities, the program embodies a broad, data-driven social policy model.

Operating in more than 185 municipalities across Israel — primarily in socio-economically disadvantaged localities and neighborhoods — the program currently serves tens of thousands of children and youth each year. Its annual budget of approximately 220 million NIS is shared among the participating ministries, with about one-third dedicated to early childhood programs (*Good Start* initiative).

At its core, the 360° Program seeks not only to provide services but to strengthen community systems, promote prevention, and institutionalize cooperation between local and national government levels.

---

2. Program Goals and Value Proposition

The program's overarching goal is to **reduce the frequency, severity, and persistence of risk situations** among children and adolescents, while promoting protective environments that support their development and well-being. It seeks to achieve this by changing the societal response to children and youth at risk through shared responsibility among government ministries, municipalities, and community actors; improving the quality and accessibility of community-based services for children and their families; and building an evidence-based, continuously learning system that monitors outcomes and informs policy.

The program defines child and youth risk according to **seven life domains**, adapted from the *UN Convention on the Rights of the Child*:

1. Physical survival, health, and development
2. Family belonging and care
3. Learning and developmental skills
4. Emotional well-being and mental health
5. Social inclusion and peer integration
6. Protection from others
7. Protection from self-endangering behavior

Additional attention is given to risk-amplifying factors such as poverty, family crisis, migration, minority status, learning disabilities, and parental or child disability.

At the same time, the program emphasizes the development of personal and familial strengths as an essential part of intervention. The ***strength-based approach*** identifies internal and external resources that enable children to cope, adapt, and thrive — distinguishing between:

- **Personal strengths**: perseverance, curiosity, optimism, self-efficacy, and emotional expression.
- **Interpersonal strengths**: empathy, cooperation, friendship, and initiative.
- **Family and environmental strengths**: cohesion, community support, healthy communication, and flexible coping.

Core principles and values of the 360° Program: The 360° Program emphasize shared governance between national and local authorities, inter-ministerial collaboration, and a unified national definition of “children and youth at risk". The program empowers local authorities to design interventions tailored to their communities’ needs, while maintaining continuous monitoring through the *TAMI* (Municipal Information Infrastructure) system. It places strong emphasis on prevention, equity, and transparency in the allocation of public resources, ensuring accountability and fairness across different population groups.

Through these mechanisms, the program builds a comprehensive and accountable national infrastructure for improving the lives of children and youth at risk, serving as a model for integrated, data-driven social policy.

3. Objectives

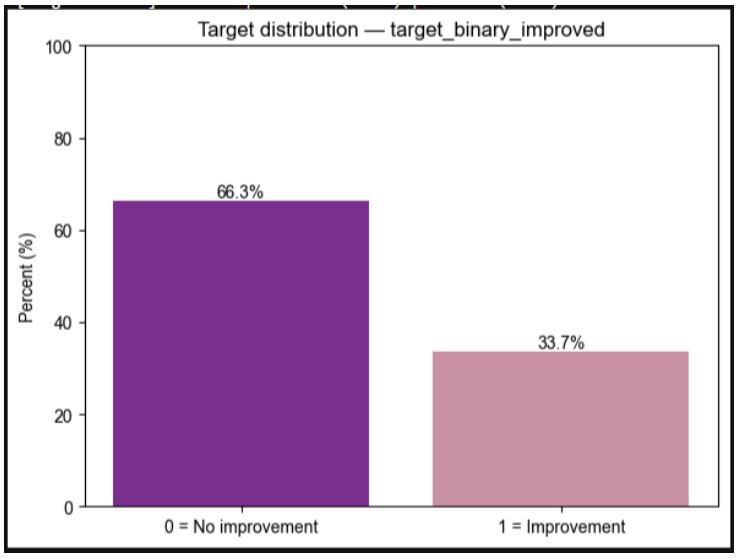
What is the probability of improvement among children and youth at risk participating in the 360° National Program?

The present project focuses on developing a predictive model that estimates the likelihood of improvement in a child's risk profile following participation in community-based interventions under the *360° National Program*.

The target variable in this study, target_binary_improved, is a binary outcome variable distinguishing between two groups of children: those whose overall level of risk improved between the first and second measurement waves (improvement=1), and those who did not improve or whose condition worsened during the intervention period (improvement=0).

Among participants in the current dataset, the majority of children (66.3%) showed no improvement, while only one-third (33.7%) demonstrated improvement in their risk status. Since the central goal of the national program is to enhance the well-being and resilience of children at risk, these findings emphasize the importance of identifying the conditions and protective factors that contribute to improvement and recovery.

Figure 1. Distribution of target variable — target_binary_improved (Percentage of children with and without improvement following participation in the program)



This project leverages machine learning techniques to analyze a comprehensive dataset integrating individual, family, program, and geographic variables. By identifying the factors most strongly associated with improvement, the project seeks to support data-driven decision-making in child welfare policy and to develop predictive models capable of flagging, at an early stage, children with a low probability of improvement, thereby enabling tailored and proactive interventions.

Project Design & Methodology

4. Data Journey

Data sources:

The analysis is based on administrative monitoring data from the *360° National Program for Children and Youth at Risk*, representing children who participated in social, educational, and community interventions between September 2023 and September 2024.

The final analytical dataset integrates information from **two structured sources**:

1. **Participant-level file:** child-level records including family background, demographic characteristics, socio-economic indicators, and detailed measures of risk and strengths.
2. **Program-level file:** registry of 88 intervention programs describing each program's organizational structure, target population, type, duration, frequency, and intended outcomes (main and secondary).

The two sources were cleaned, standardized, and merged using program identifiers (*program_code* and *program_name*) to form a unified **child-level flat file** suitable for machine learning analysis. The resulting dataset contains approximately **21,981 children** with **405 variables** capturing individual, familial, and program-level factors related to child risk, protection, and improvement.

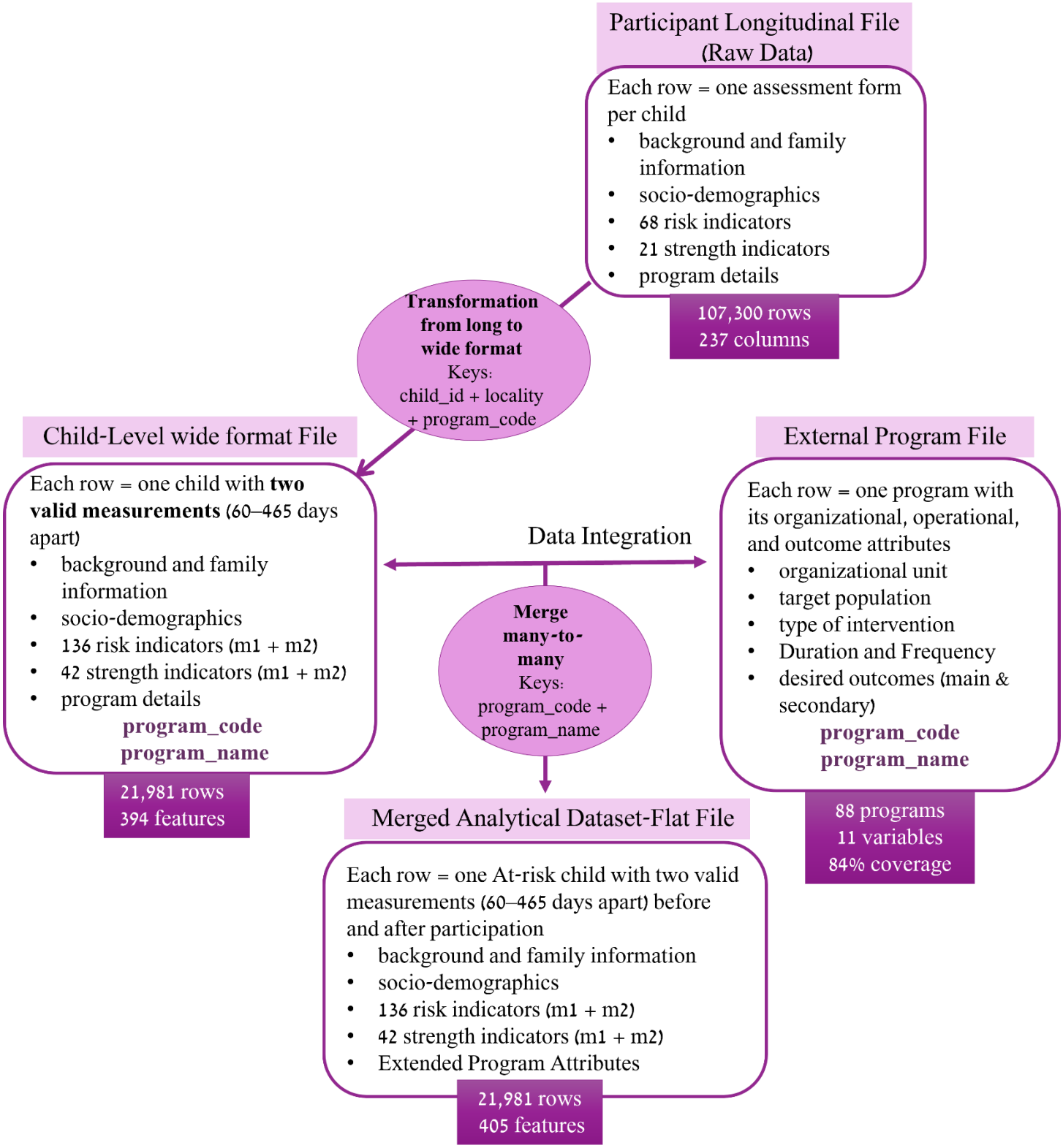
Data structure and relationships:

The following diagram illustrates the structure and integration flow of the data used in the analysis. The process combines two structured sources: the Participant Longitudinal File and the External Program File through a series of transformation and merge steps.

Records were reshaped from a long to a wide format at the child level and then merged on shared identifiers (program_code and program_name) to create a unified analytical dataset.

This final flat file enables consistent tracking of each child across two measurement waves (m1, m2) and links individual, family, and program-level information within the 360° National Program database.

Figure 2. Data Integration and Preparation Process



Relationship description of the above chart:

1. The *Participant Longitudinal File* serves as the primary child-level dataset, containing detailed information on background, socio-demographics, and standardized indicators of risk and strength for each measurement wave.
2. The *External Program File* provides complementary program-level information and was merged with the child-level data through the keys program_code and program_name.
3. Only children with two valid measurement waves (60–465 days apart) were retained, ensuring that improvement can be meaningfully assessed over time.
4. Non-informative or redundant fields (e.g., raw dates, duplicate identifiers, and system-generated variables) were excluded.
5. The resulting *merged analytical dataset* includes 21,981 children and 405 harmonized variables. The transformation from long to wide format used a composite key (child_id + locality + program_code), ensuring that each record uniquely represents one child participating in a specific program and locality across both measurement waves.

A detailed description of all variables and preprocessing steps is available in the project repository under /data/interim and /data/processed.

Time frame

The data covers two consecutive measurement waves collected between **September 2023 and September 2024** as part of the 360° National Program's municipal monitoring system.

Each child may have one, two or more valid assessment forms recorded in this timeframe. For modeling purposes, only children with **two valid measurements** separated by **60–465 days** were retained to ensure that change in risk status reflects meaningful program exposure rather than administrative timing differences.

The first measurement wave (m1) represents the child's baseline risk level at program entry, while the second (m2) represents their status following participation in community-based interventions.

All predictive features were drawn from the m1 data or static background variables to prevent temporal leakage from post-intervention information.

Outcome

The project's target variable is **target_binary_improved**, a binary indicator distinguishing between two groups:

- 1 = children whose overall level of risk improved between m1 and m2
- 0 = children who showed no improvement or deterioration

Among 21,981 children with two valid assessments, **33.7%** demonstrated improvement, while **66.3%** did not.

This distribution reveals a moderate class imbalance, which was addressed in later stages using appropriate resampling and weighting techniques.

Because the central goal of the 360° National Program is to enhance child well-being and reduce risk, this variable represents a key measure of program success and serves as the dependent variable in all subsequent modeling stages.

Data Exploratory Strategy

Initial exploration was conducted to understand data structure, quality, and relationships between variables across two measurement waves (M1, M2). Automated profiling using **ydata_profiling** revealed extensive missingness, high skewness, and dominance of low values in multiple risk indicators.

The dataset showed strong imbalance between improvement groups and varying completeness across risk domains and localities.

Preliminary statistical tests (Spearman correlations, *t*-tests, ANOVA) were used to identify redundant and influential features, highlighting the strong role of family and baseline risk indicators in predicting improvement.

Visual inspection confirmed the validity of outcome definitions and informed subsequent cleaning and feature engineering. Overall, the EDA phase established a clear understanding of the data's structure and limitations, guiding targeted handling of missing values, outliers, and imbalances in the next phase.

Making raw data into legible and essential

The dataset was standardized and refined to ensure clarity and analytical consistency. Variables were reviewed for logical coherence, naming consistency, and measurement alignment across M1 and M2.

Technical identifiers and redundant date fields were removed. Structural variables (risk, strengths, and background indicators) were unified under harmonized categories, while program codes and locality identifiers were retained to enable multi-level analyses.

Outliers

No problematic outliers were detected. All numeric features were found within plausible ranges, and no observations were excluded. Retaining full variability was important to preserve rare but valid cases (e.g., large families) that could influence child risk and improvement outcomes.

Missing Values

The dataset exhibited extensive missingness across 293 variables, including 125 with ≥80% missing and 16 entirely empty. Missingness patterns were analyzed to separate true data loss from structural skips resulting from questionnaire logic. Continuous variables were imputed using a skewness-based rule (median for |skew| > 1, mean otherwise), while categorical risk items were reclassified as 0 ("no known risk") to reflect structural absence rather than missing data. Large-scale imputation was applied to summary indicators (964,867 cells filled across M1 and M2). Two textual "strength domain" variables with >80% missing were removed, as they were not part of the data collection period, and partially missing background variables (e.g., orphan or disability status) were retained and coded as "Unknown."

The resulting dataset contained 21,981 records and 394 columns, ready for feature engineering and modeling.

Feature Engineering

New variables were engineered to capture multi-dimensional aspects of child risk, strengths, and context. Twenty-eight risk items were aggregated into seven domains, generating cross-domain indicators of improvement and intensity. Red-flag variables (severe risks) were counted to track change between M1 and M2. Geographic coordinates enabled the creation of local and district-level risk rates, as well as a composite "high-risk area" index. External program data were merged, adding organizational and intervention attributes (84% coverage). The resulting dataset expanded from 394 to 485 variables, enhancing interpretability while maintaining analytical coherence.

The spatial distribution of average child risk intensity across localities is presented in Figure 3, illustrating substantial regional variation and deviations from the national mean.

Figure 3. Average child risk intensity by locality (Measurement 2) Color gradient represents deviation from the national mean (darker = higher average risk).

Feature Selection & Encoding

To prepare the modeling matrix, categorical variables were encoded and a rigorous feature selection process was applied to prevent data leakage. Identifiers and all post-intervention (M2) or change-related variables were excluded, retaining only baseline and time-invariant predictors. Categorical fields were encoded using a hybrid strategy combining frequency encoding for services, Top-K one-hot encoding for programs and localities, and numeric aggregation for geographic risk rates. Feature selection combined univariate ANOVA F-tests with multimodel voting (Lasso, Ridge, Logistic, Gradient Boosting, Random Forest), retaining variables with ≥4 votes. The final balanced dataset included **40 predictive features** and one binary target variable (target_binary_improved), forming a clean and fully numeric matrix of 21,981 records × 41 columns, ready for model selection.

6. Models

Model Design and Evaluation

The modeling stage aimed to predict the likelihood of improvement among children participating in the 360° National Program for Children and Youth at Risk. The binary target variable (target_binary_improved) distinguishes between children who improved between the two measurement waves (1 = improvement, 0 = no improvement or deterioration).

Given the program's preventive goals, models were optimized to maximize **Recall_0**—the ability to correctly identify children unlikely to improve—thus enabling early, targeted intervention.

A series of baseline and tuned models were compared, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and CatBoost.
Baseline models established initial benchmarks: Logistic Regression achieved stable interpretability (AUC ≈ 0.77), while Decision Tree performance was slightly lower, confirming initial complexity as the dominant split variable. Ensemble methods—Random Forest and Gradient Boosting—demonstrated stronger generalization (AUC ≈ 0.78–0.79).

Subsequent hyperparameter tuning (Random, Grid, Bayesian) further enhanced performance.

- Logistic Regression (Bayesian) reached the highest Recall_0 (≈0.94) but with limited precision.
- Gradient Boosting and XGBoost (Grid-tuned) balanced precision and recall (AUC ≈ 0.79–0.80).
- CatBoost (Grid) achieved the most consistent and stable results, with Accuracy = 0.74, F1 = 0.54, ROC-AUC = 0.80, and PR-AUC = 0.67, maintaining reliable class detection across groups.

The final **CatBoost (Grid-tuned)** model was selected as the optimal solution. On the test set, it achieved ROC-AUC = 0.785, PR-AUC = 0.635, and balanced recall across classes (Recall_0 = 0.85, Recall_1 = 0.51).

At a calibrated threshold (p = 0.45), the model correctly classified 85% of non-improvers and 51% of improvers, offering both robustness and interpretability. The leaderboard (Figure 4) summarizes all tested models (Baseline, Random, Grid, Bayesian), ranked by Recall_0 → F1 → ROC-AUC.

While Logistic Regression (Bayesian) reached the highest $Recall_0$, CatBoost (Grid) demonstrated the best overall balance of accuracy, recall, and AUC, justifying its selection as the final predictive model.

Figure 4. Global Model Leaderboard – Comparison of Baseline and Tuned Classifiers

Global Leaderboard — latest per model×tuning (sorted by $Recall_0$ → F1 → ROC_AUC)

	Model_base	Tuning	Accuracy	Precision	Recall	Recall_0	F1	ROC_AUC	PR_AUC	BalancedAcc	MCC	TN_%	FP_%	FN_%	TP_%
0	Logistic Regression	Bayesian	0.692983	0.633911	0.209154	0.938665	0.314532	0.703922	0.517966	0.573910	0.222290	62.3%	4.1%	26.6%	7.0%
1	CatBoost	Grid	0.744920	0.687065	0.445045	0.897119	0.540186	0.802861	0.671121	0.671082	0.391575	59.5%	6.8%	18.7%	15.0%
2	Gradient Boosting	Grid	0.738550	0.677650	0.426126	0.897119	0.523230	0.790273	0.652779	0.661623	0.373928	59.5%	6.8%	19.3%	14.3%
3	Logistic Regression	Grid	0.731467	0.666493	0.405594	0.896941	0.504298	0.770463	0.615796	0.651268	0.354209	59.5%	6.8%	20.0%	13.7%
4	Gradient Boosting	Random	0.726418	0.654303	0.397297	0.893461	0.494395	0.766201	0.620276	0.645379	0.340714	59.3%	7.1%	20.3%	13.4%
5	Logistic Regression	Baseline	0.733769	0.664339	0.423395	0.891373	0.517181	0.772575	0.617942	0.657384	0.362331	59.1%	7.2%	19.4%	14.3%
6	CatBoost	Random	0.734304	0.658967	0.436937	0.885231	0.525460	0.779939	0.636619	0.661084	0.365615	58.7%	7.6%	19.0%	14.7%
7	Gradient Boosting	Baseline	0.732500	0.652800	0.438700	0.881600	0.524800	0.781700	0.641100	0.660200	0.361800	58.5%	7.9%	18.9%	14.8%
8	Random Forest	Baseline	0.744900	0.649300	0.527000	0.855500	0.581800	0.789800	0.657600	0.691300	0.405700	56.7%	9.6%	15.9%	17.7%
9	XGBoost	Random	0.735214	0.607434	0.603604	0.802012	0.605513	0.792325	0.653558	0.702808	0.406251	53.2%	13.1%	13.3%	20.3%
10	Decision Tree	Baseline	0.673900	0.515200	0.534200	0.744900	0.524500	0.639400	0.431700	0.639500	0.276700	49.4%	16.9%	15.7%	18.0%
11	CatBoost	Baseline	0.720700	0.568300	0.708100	0.727000	0.630600	0.800100	0.663300	0.717600	0.416700	48.2%	18.1%	9.8%	23.8%
12	XGBoost	Baseline	0.718229	0.563598	0.722523	0.716049	0.633241	0.795589	0.658789	0.719286	0.418446	47.5%	18.8%	9.3%	24.3%

Overall, the ***CatBoost model*** provides a generalizable, explainable, and policy-relevant predictive framework. This is strongly demonstrated on unseen test data, where the model **correctly identifies 85% of the children who were not expected to improve,** supporting data-driven decision-making within the 360° program.

Model Interpretation

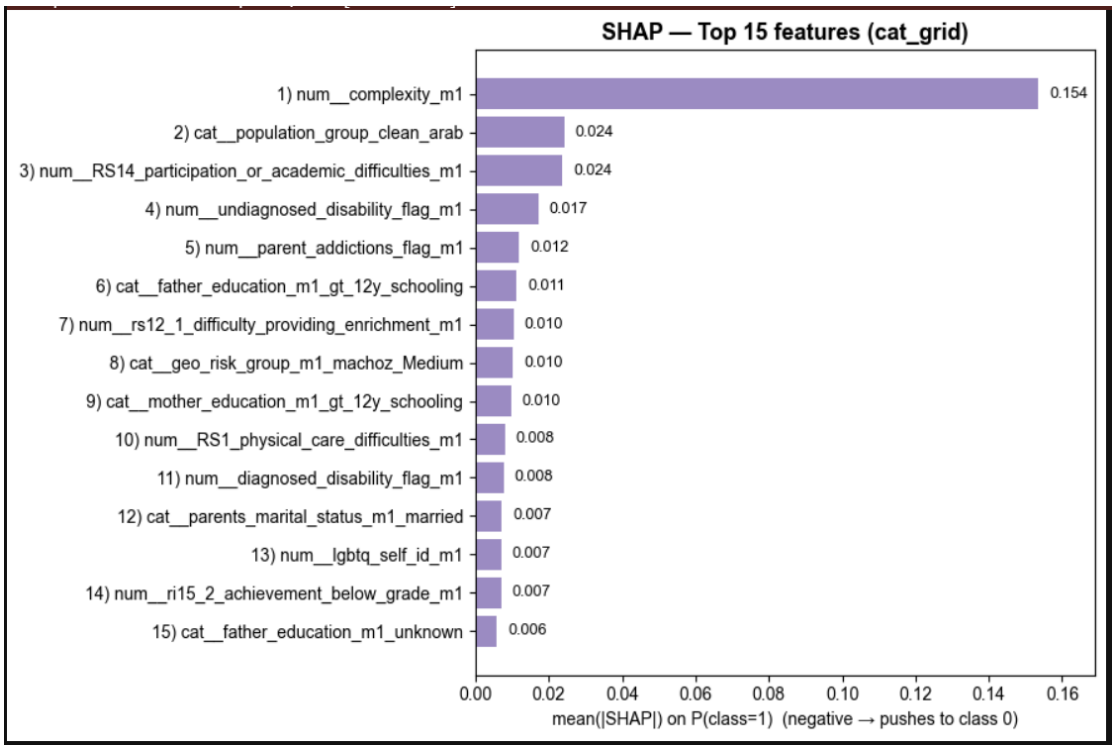
Model interpretation was conducted using **SHAP explainability analysis** to examine how each feature influenced the model's predictions.

Across all analyses, **initial complexity (complexity_m1)** emerged as the dominant predictor, indicating that children who entered the program with higher overall risk levels were more likely to show measurable improvement, suggesting that intensive interventions were particularly effective for high-complexity cases.

Additional key features included **learning difficulties** and **parental addictions** (both negatively associated with improvement), and **higher parental education** and **Arab population group** (positively associated).

Together, these variables formed a consistent and socially meaningful pattern: improvement was most common among children with complex but well-defined needs who received structured, multi-domain interventions, while stagnation or deterioration was more typical in cases of chronic family or educational difficulties.

Figure 5. SHAP Feature Importance — Mean absolute SHAP values for the top 15 predictors in the final CatBoost model. The figure presents the global ranking of the most influential features across all predictions.



Across all tuned models, Logistic Regression, Gradient Boosting, and CatBoost, the most influential predictor of improvement was consistently **initial case complexity (num_complexity_m1)**.

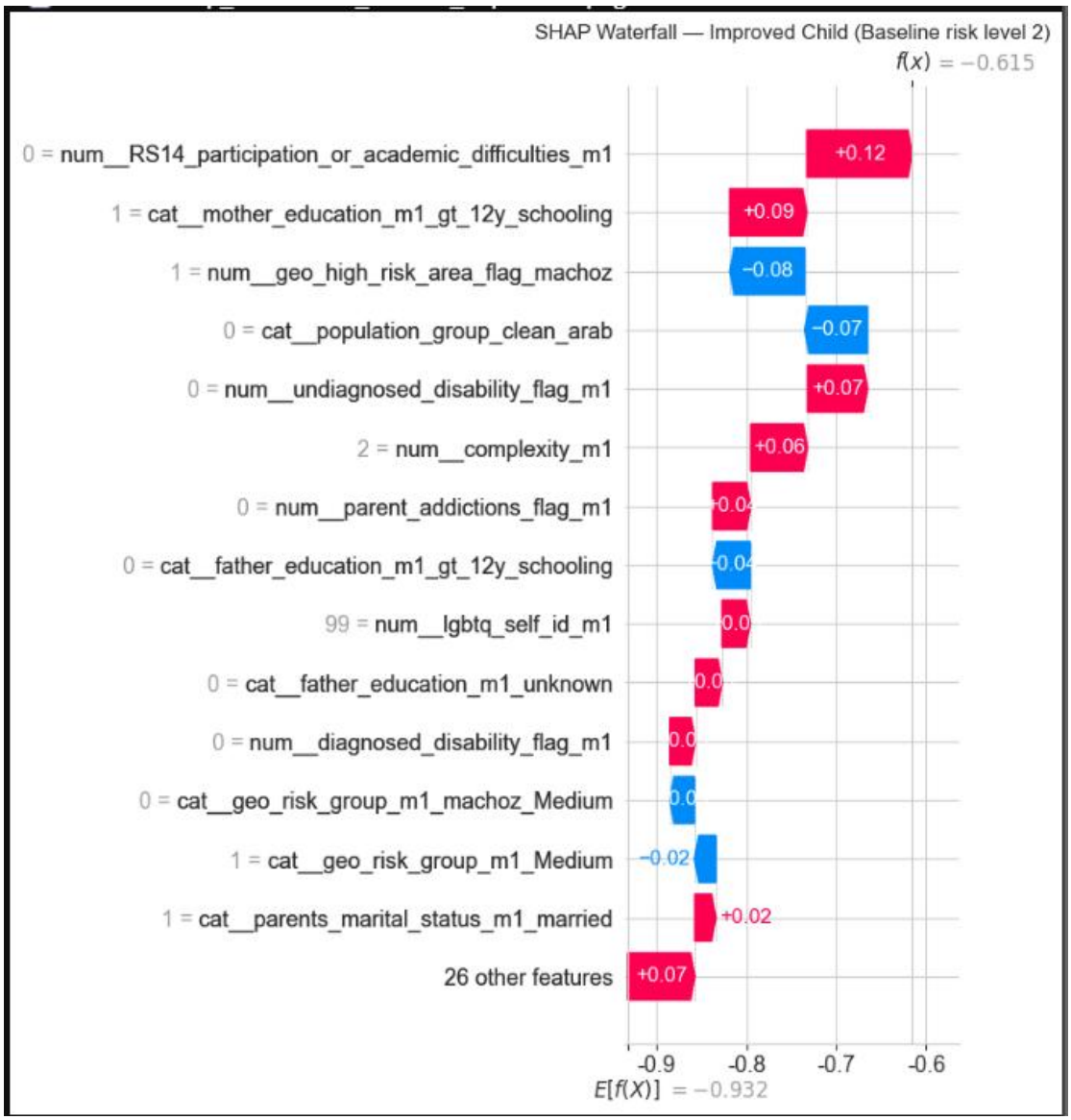
Secondary but recurring contributors included **academic or participation difficulties, population group (Arab), and parental education or disability indicators**, reflecting the combined effects of structural, socio-educational, and family factors on outcomes.

The SHAP results demonstrate strong cross-model agreement, reinforcing the robustness of these predictors in explaining improvement likelihood.

To deepen interpretation, SHAP waterfall analyses were conducted both on the **general sample** and within **specific baseline risk tiers**, illustrating how predictor effects vary by child profile.

Figure 6 presents an example of an individual prediction for a child with **low baseline complexity (risk level 2)**, showing how academic difficulties, parental education, and local area risk levels interact to determine improvement probability.

Figure 6. SHAP Waterfall Plot — Example prediction for a child with low baseline complexity (risk level 2). The plot demonstrates the transition toward interpretable subgroup analyses, revealing the differing mechanisms of change across risk levels.



Next Steps

The current modeling phase provided a solid predictive foundation. SHAP analysis, used to interpret the model's logic, confirmed that the most dominant predictive feature, by a significant margin, is the child's ***complexity level*** at program entry.

However, this analysis also revealed a "paradox": a higher baseline risk level paradoxically predicted a *greater* likelihood of improvement.

I hypothesize that the most intensive program interventions are effectively tailored to and treat these high-severity cases, driving this counter-intuitive outcome. This finding clarifies that complexity, while a highly informative predictor, is a summary indicator. This provides a strong foundation for the next stage of analysis, which is to shift **from prediction to explanation**.

The goal is to delve deeper into these distinct risk groups across the risk continuum to understand which modifiable factors, within ***risk, strengths, and program characteristics***, help or hinder their improvement. It is therefore recommended that future research focuses on identifying the "levers of improvement" for different subgroups.

Recommended analyses include:

- Developing SHAP-based explainability models **within each baseline risk tier (1–4)** to detect domain-specific contributors (testing our hypothesis).
- Applying partial dependence (PDP/ALE) and interpretable logistic models to quantify direction and effect size.
- Conducting light causal inference analyses to explore which program features and outcomes drive the strongest improvement for specific subgroups.

This transition from predictive modeling to interpretive analysis will enable actionable insights. The final objective is to provide 360° program managers with **data-driven policy recommendations** for adaptive policy design and more targeted, evidence-based interventions aimed at increasing improvement rates among children at risk.

Project Notebooks

The project is written in **Python** and includes the following Jupyter notebooks:

Stage 1 – Data Preparation

Merging longitudinal child-level files, resolving duplicates, and aligning measurement waves (M1–M2) using completeness-based logic and valid time gaps (60–465 days).

Stage 2 – EDA (Exploratory Data Analysis)

This stage was split into two parts due to its comprehensive nature:

- **Part 1 (2_EDA.ipynb):** Data profiling, descriptive statistics, and univariate distribution visualization.
- **Part 2 (2_EDA-Part2.ipynb):** Bivariate analysis and statistical relationship testing (Chi-sq, T-test, ANOVA) with the outcome variable.

Stage 3 – Data Cleansing
Treatment of missing values and structural skips, consistency checks across waves, and validation of key categorical and numeric variables.

Stage 4 – Feature Engineering
Derivation of composite indicators for risk and strength domains, geographic risk rates, and program-level attributes; creation of deltas and time-invariant aggregates.

Stage 5 – Feature Selection & Encoding
Univariate and multivariate feature ranking (L1/L2/Tree-based), leakage control, frequency and one-hot encoding, and preparation of the final modeling matrix (X = 40).

Stage 6 – Model Selection & Fine Tuning
Training of multiple classifiers (Logistic Regression, Random Forest, Gradient Boosting, XGBoost, CatBoost), cross-validation, hyperparameter tuning, and model evaluation using Recall_0, F1, and ROC/PR curves.

Thank you

Zohar Or Sharvit