Kartik Mathur

Mkartik477@gmail.com

Prepared for Elaine Race

Director of Education and Engagement  
Data2Intel

ANALYSIS OF NUMERACY RISK IN PRIMARY SCHOOLS

Business Report

Contents

[Executive Summary 1](#_Toc200110100)

[Business Understanding 2](#_Toc200110101)

[EDA Key Insights 2](#_Toc200110102)

[Student Demographics Analysis 3](#_Toc200110103)

[Numeracy Skills by Risk 3](#_Toc200110104)

[Literacy Skills by Risk 4](#_Toc200110105)

[Relation between Literacy and Numeracy 4](#_Toc200110106)

[Disability Conditions by Risk 5](#_Toc200110107)

[Additional Insights for Early Intervention 6](#_Toc200110108)

[Clustering Analysis 9](#_Toc200110109)

[Proposed Machine Learning Solution 11](#_Toc200110110)

[Recommendations and conclusions 13](#_Toc200110111)

# Executive Summary

This report presents exploratory data analysis (EDA) and predictive modelling to identify primary school students at risk of underperforming in numeracy. We uncover patterns affecting academic outcomes using demographic, literacy, and numeracy data and recommend targeted interventions.

The selected machine learning model offers predictive capabilities to support early intervention strategies for educators. Key ethical considerations and future improvements are also discussed.

# Business Understanding

The following is our approach to the problem using the BACCM framework.

* Change: Apply concrete, data driven methods to identify risk factors, enabling targeted interventions that enhance student success rates.
* Need: Develop early identification techniques to proactively support students through personalised interventions.
* Solution: Predictive models to accurately profile students and assess the significance of potential risk factors.
* Stakeholders: At-Risk students, primary schools, teachers, policymakers.
* Value: Achieve optimal improvement in educational outcomes with minimum intervention.

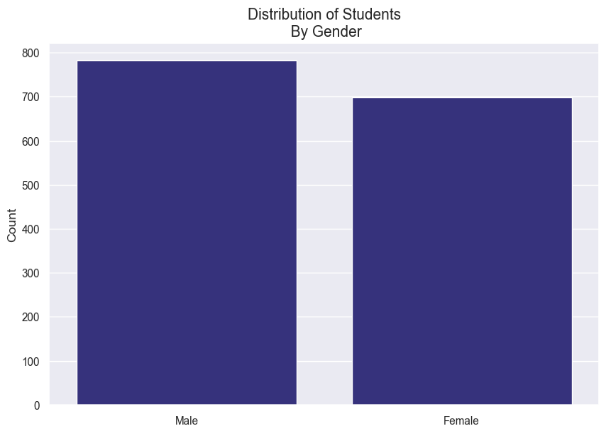
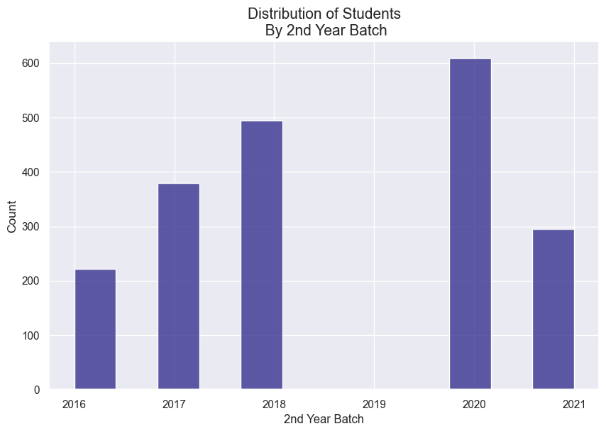
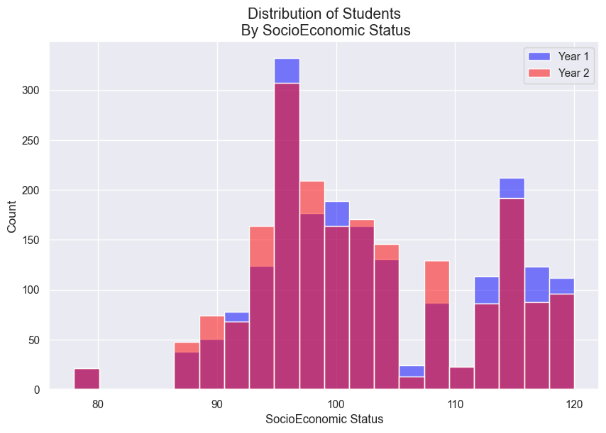
# EDA Key Insights

This section presents responses to the six questions mentioned in the brief, analysing student demographics, numeracy and literacy skills, disabilities, and their relationships with Year 3 numeracy risk, offering insights for early intervention

All analysis was done after removing abnormal entries during processing.

### Student Demographics Analysis

A graph of a distribution of students

AI-generated content may be incorrect.

Based on the Age, Gender, Batch, and Socio-Economic status of the students, the data seems to be diverse, with the notable absence of any students from the 2019 batch.

### Numeracy Skills by Risk

**A graph of numbers and numbers

AI-generated content may be incorrect.**

As expected, students at risk generally scored lower during their first year. Despite this, they perform similarly to the others in their second year.

### Literacy Skills by Risk

*A graph of different colored squares

AI-generated content may be incorrect.*

By the end of their second year, students at risk achieve similar scores to those who are not. They have shown greater improvement in their Literacy Skills over the course of their first two years in comparison.

### Relation between Literacy and Numeracy

A screenshot of a graph

AI-generated content may be incorrect. --------

The students’ literacy skills, and numeracy skills for a given year are highly correlated with each other. Literacy skill and numeracy skill are however, not strongly correlated.

A table of numbers with text

AI-generated content may be incorrect.

The table above shows the correlation between each academic column and the risk of underperforming in numeracy in year 3. Each column has a weak, inverse correlation with the target variable.

### Disability Conditions by Risk

*A graph of disability and risk

AI-generated content may be incorrect.*

The body of students at risk have a much lower amount of those suffering from a physical or social-emotional disability. However, the proportion of those suffering from a sensory disability is much higher.

*A graph showing different colored squares

AI-generated content may be incorrect.*

The above graph shows that students suffering from a sensory disability also have a much lower probability of receiving NCCD funding compared to other disabilities.

### Additional Insights for Early Intervention

Comparing the academic performance of students according to whether they are at risk reveals a clear disparity.

**A graph of numbers and numbers

AI-generated content may be incorrect.**

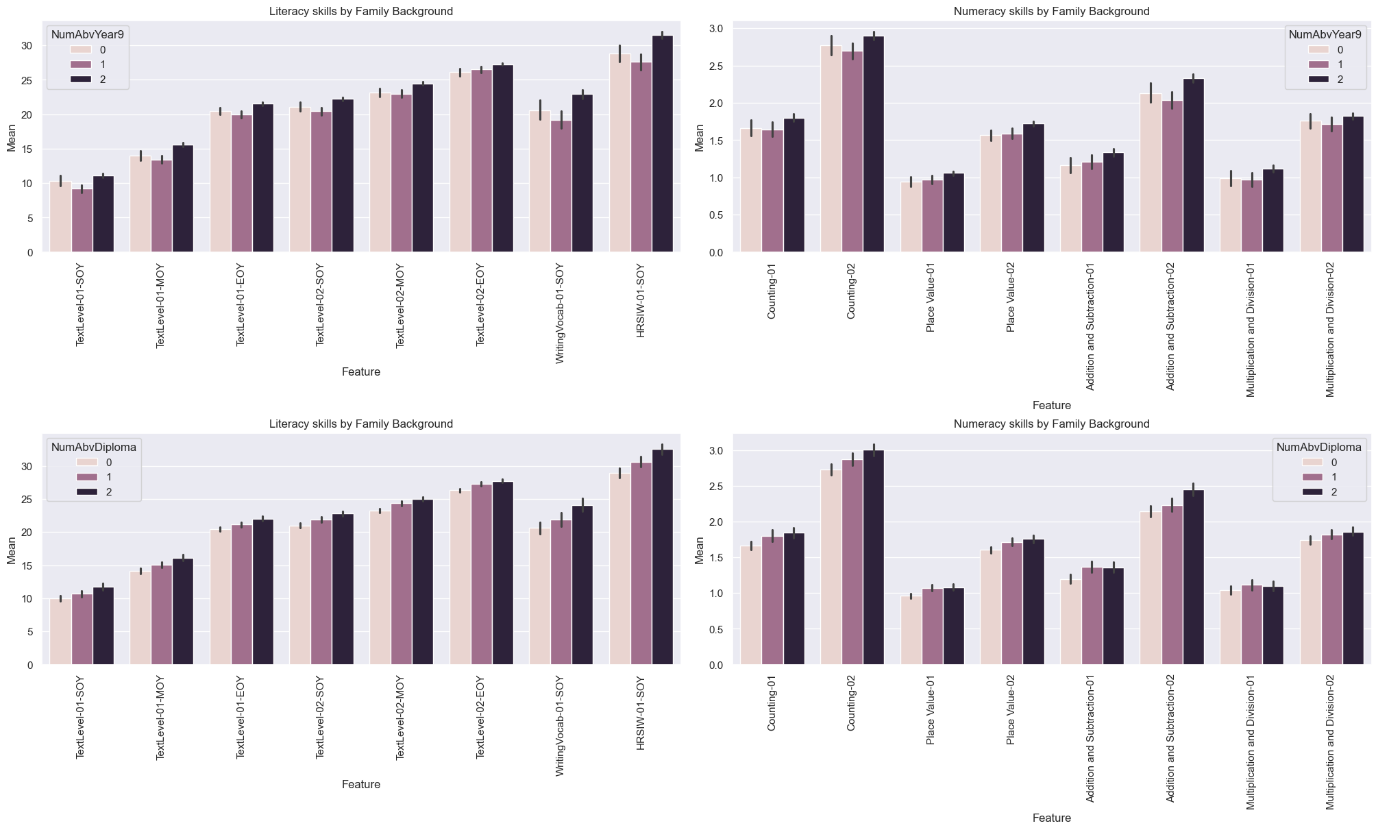
Additionally, when conducting machine learning analysis, we find the importance of each column according to the model.

The table below shows the top 10 most important columns according to our machine learning model.

A table with text and numbers

AI-generated content may be incorrect.

7 of the 10 most important columns relate to academic performance. Of the remaining, 2 columns relate to the academic background of the student’s parents. It is clear from this data that academic performance is the greatest indicator of whether the student is at risk of underperforming in numeracy in year 3.



The above graph shows the academic performance of students according to their family background, specifically the education of their parents.

We observe that students who have more educated parents achieve higher test scores, likely due to the high quality of personalized education that their parents can provide.

A graph of a number of parents above year

AI-generated content may be incorrect.

From the above graph, showing the probability of a student being at risk according to the number of parents of a certain educational level, having less educated parents can significantly increase risk.

### Clustering Analysis

A diagram of a cluster of dots

AI-generated content may be incorrect.

The plot above shows the results of the clustering analysis. Each data point has been separated into distinct, non-overlapping clusters.

A graph of a bar chart

AI-generated content may be incorrect.

Cluster 0 contains most of the students at risk of underperforming at numeracy in year 3. Clusters 1 and 2 seem quite like each other in this respect. Investigating other columns will reveal the distinctions.

To further understand each cluster, we will look at the means of each column according to cluster.

A graph of multiple colored rectangular objects

AI-generated content may be incorrect.

The graph above shows the students’ literacy skills by cluster. As expected, students belonging to cluster 0 tend to have lower academic performance. A distinction between clusters 1 and 2 also arises, with cluster 2 having higher academic performance on average.

A graph of numbers and numbers

AI-generated content may be incorrect.

The numeracy skills of each cluster are consistent with their literacy skills.

# Proposed Machine Learning Solution

The most effective machine learning model in this case is Logistic Regression. Specifically, the model pipeline outlined below:

logreg\_pipe\_1 = Pipeline(

        [

            ("resample", RandomUnderSampler(random\_state=42)),

            ("scaling", StandardScaler()),

            ("pca", PCA(n\_components=8, random\_state=42)),

            ("classifier", LogisticRegression(

C=0.1,

solver='newton-cg',

random\_state=42

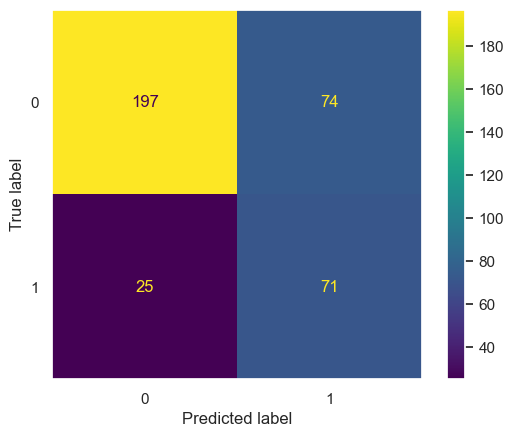
)

)

        ]

    )

The model was trained on a sample of 80% of the data provided and tested on the remaining 20%. The results of that test are as follows.



Above is a confusion matrix which compares the model’s predictions with reality. In this case, the label “1” represents students who are at risk and the label “0” represents those who are not.

As outlined in the technical report, this model correctly identified 73.9% of the at-risk students marking a performance of 208% that of an unoptimized model.

Given this, the model has the following Pros and Cons

**Pros:**

* Effective at identifying students who are at risk.
* Highly scalable performance.
* Computationally inexpensive due to its simplicity
* Highly interpretable, providing data on the most important columns.

**Cons**

* Assumes linear relationships, potentially missing out on small nuances.
* Highly sensitive to outliers in the training data, requiring thorough preprocessing
* Prioritizes identifying students at risk, potentially leading to many false positives.

### Suggested Improvements

The model would be greatly improved with the following changes

* Correct data regarding the real numeracy skills of the students in year 3 for more accurate training data
* More samples to capitalize on the model’s scalability and refine predictions.

# Recommendations and conclusions

The model places the most importance in the following features:

1. **HRSIW Assessment Results**

A student’s performance in the HRSIW assessment is indicative of their listening comprehension skills. Since the most common method of instruction is oral, students weak in listening comprehension are unlikely to reap the expected benefits of daily instruction.

Such students can be accommodated by introducing additional methods of instruction that do not rely solely on phonemic awareness.

1. **Family Educational Background**

Parents often provide educational support to their children. The quality of this support is informed by the parents’ own skills in teaching and learning. Students who do not have highly educated people in their family miss out on this additional support and personal attention.

This can simply be rectified through intentional efforts by teachers to identify these students and prioritize their attention on them.

1. **Basic Numeracy Skills**

According to the model, deficits in basic skills like place value, and addition and subtraction, are more indicative of risk.

This means that prioritizing foundational concepts over more complex ideas is likely to be much more effective than solely focusing on the year 3 material.