CLIENT BRIEF

Predictive Analytics for Early Writing Intervention in Primary Education

Client: Data2Intel Learning Analytics
Project Duration: 6 weeks
Issue Date: September 2024

1. ABOUT DATA2INTEL

Data2Intel is an Australian learning analytics consulting service specialising in delivering data analytics solutions within the primary education sector. We partner with schools, educational institutions, and government bodies to transform educational data into actionable insights that improve student outcomes.

Our current engagement involves a consortium of forty primary schools committed to enhancing educational outcomes through data-driven interventions. This project represents a significant opportunity to demonstrate the value of predictive analytics in identifying at-risk students early and enabling timely, targeted support.

2. PROJECT CONTEXT

The Australian Education System

Australian school education operates on a K-12 system similar to many other countries:

- **Primary schools**: Years K-6 (ages 5-11)
- **Secondary schools**: Years 7-12 (ages 12-18)
- **Early years**: K-2 (Kindergarten to Year 2, ages 5-7)

Most Kindergarten students must be at least 5 years old by January of the calendar year they commence schooling.

NAPLAN and Early Years Assessment

The National Assessment Program (Literacy and Numeracy (NAPLAN)) provides annual assessments for students in Years 3, 5, 7, and 9, serving as a key benchmark for foundational literacy and numeracy skills.

Research consistently demonstrates that the early years of schooling (K-2) have a profound influence on:

- Future academic performance
- Long-term employability
- Overall wellbeing
- Career progression

The foundational skills acquired during these formative years set the stage for long-term educational and professional success. However, assessment at this age presents unique challenges. Given the young age of students (5-7 years), evaluations often include dialogue or interview-based assessments administered by trained teachers, rather than solely traditional "pen and paper" tests.

3. THE BUSINESS CHALLENGE

Our school consortium has identified a critical need: **predicting which students in the early years (Years 1-2) are at risk of underperforming in writing by Year 3**, as measured by NAPLAN results.

Early identification enables:

- Timely interventions during the crucial foundational years
- Targeted resource allocation to students who need it most
- Proactive support mechanisms rather than reactive remediation
- Better educational outcomes and long-term student success

Currently, schools lack predictive tools to identify at-risk students before Year 3 NAPLAN testing. By the time underperformance is detected through NAPLAN, valuable intervention time has been lost.

4. PROJECT OBJECTIVES

We need your expertise to develop a predictive analytics solution that will:

Primary Objectives:

1. **Build predictive models** to identify students at risk of underperforming in writing in Year 3 based on their Year 1 and Year 2 data
2. **Uncover data insights** from student performance, demographics, and background information
3. **Explore student segments** through clustering analytics to understand different learner profiles
4. **Provide actionable recommendations** to schools and educators for implementing targeted interventions

Key Analytical Questions:

We need you to address the following specific questions:

1. What are the socioeconomic status (SES) backgrounds of students in the dataset during Year 1 and Year 2? (For context: in 2018, Catholic system schools had a national average SES of 100, while independent schools averaged 102, according to the Australian Department of Education and Training)

2. What are students' reading skills (e.g., Burt Reading Scores) at the start and end of Year 1 and at the start and end of Year 2?

3. What are students' writing skills at the start of Year 1 (WritingVocab-01-SOY)? Is there a relationship between this and Year 3 writing risk?

4. Are students' literacy skills and numeracy skills related? Are there relationships between these and their Year 3 writing risk?

5. What disability conditions are present in the dataset? Are there relationships between these conditions and Year 3 writing risk?

6. What other insights might inform early interventions to improve students' writing skills?

5. AVAILABLE DATA ASSETS

We are providing access to a comprehensive dataset covering **2,000 students across over forty schools**, spanning five continuous years from **2016 to 2020**.

Dataset: LA4Schools.csv

The dataset includes:

Student Performance Metrics:

- Reading skills assessments (including Burt Reading Scores) at multiple time points in Years 1 and 2
- Numeracy skills assessments during Years 1 and 2
- Writing vocabulary assessments at the start of Year 1
- Year 3 writing risk indicator (derived from NAPLAN results)

Student Demographics & Background:

- Socioeconomic status (SES) indicators
- Family background information
- Disability conditions

Data Characteristics:

- 2,000 student records
- 5 years of longitudinal data (2016-2020)

- Multiple assessment points throughout Years 1 and 2
- Formative assessments are validated and consistent across schools
- Assessments include dialogue/interview-based evaluations appropriate for young learners (ages 5-7)
- The target variable (Year3_Writing_At_Risk) is determined by NAPLAN results

Data Quality: The dataset has been curated specifically for this project. Records with incomplete information have been removed. However, you should still conduct standard data quality checks and preprocessing as part of your methodology.

Documentation: Full data descriptions and variable definitions are provided in the accompanying **LA4Schools Data Description** document.

6. SCOPE OF WORK

6.1 Business Analysis

- Develop business and data understanding using the **Business Analysis Core Concept Model (BACCM)** framework
- Frame the business problem clearly within its educational context
- Articulate the value proposition of this predictive analytics solution

6.2 Exploratory Data Analysis (EDA)

- Prepare and explore the dataset
- Conduct data quality checks and preprocessing
- Perform statistical analysis and create visualisations
- Address the six specific analytical questions listed in Section 4
- Generate insights to inform feature selection and model development

6.3 Machine Learning Model Development

Supervised Learning (Predictive Models):

- Develop **two different predictive models** to identify students at risk of writing underperformance
- Evaluate model performance using appropriate metrics
- Compare models based on defined selection criteria
- Recommend the optimal model for deployment

Unsupervised Learning (Clustering):

- Conduct **clustering analysis** to identify student segments
- Justify the number of clusters selected
- Interpret cluster characteristics to provide actionable insights

6.4 Solution Recommendations

- Interpret and discuss model results
- Recommend a deployment strategy
- Identify opportunities for future analytics engagements
- Address the ethical and legal implications of the models
- Suggest improvements and next steps

7. DELIVERABLES

You will deliver your work to two different stakeholders within Data2Intel, each requiring a different reporting approach:

7.1 Technical Analytics Report (Part A)

Audience: Dr Alok Sinha, Director of Data and Insights (Technical stakeholder)

Format: Comprehensive technical report, maximum 2,000 words

Required Components:

- Cover page with report title, consultant details
- Table of contents
- Executive summary (max 200 words)
- **Main Report Sections:**
    1. **Introduction:** Business problem, context, and value proposition
    2. **Approach:** Overview of ML approach, problem types, prediction targets
    3. **Data Preparation & EDA:** Data sources, quality, cleansing, preprocessing, statistical analysis, visualisations, and key insights
    4. **Model Development & Evaluation:**
        - Two predictive models with performance metrics
        - Model comparison and selection criteria
        - Clustering analytics results with cluster justification
    5. **Solution Recommendation:** Result interpretation, model recommendation, future engagement opportunities
    6. **Technical Recommendations:** Development environment details, machine process diagram, data preprocessing pipeline, maintenance suggestions
- References (APA 7th edition)
- Optional appendices

Additional Technical Files:

- Python notebook with detailed comments for the deployment team
- PDF version of the Python notebook

7.2 Business Consultancy Report (Part B)

Audience: Sally Tran, Director of Education and Engagement (Non-technical stakeholder)

Format: Executive business report, maximum 1,000 words

Required Components:

- Cover page with report title, consultant details
- Table of contents
- Executive summary (max 100 words)
- **Main Report Sections:**
  1. **Introduction:** Business understanding using the BACCM framework
  2. **Insights from EDA:**
     - Answers to the six analytical questions
     - Additional insights (data quality observations, clustering insights)
  3. **Proposed ML Solution:**
     - Selected machine learning model
     - Performance interpretation
     - Pros and cons discussion
  4. **Recommendations & Conclusions:**
     - Business applications
     - Benefits to stakeholders and value proposition
     - Implications for business processes and decision-making
     - Recommendations for further improvements
     - Ethical and legal considerations
- References (APA 7th edition)
- Optional appendices

8. TECHNICAL REQUIREMENTS

Tools & Environment:

- Python for machine learning development
- Standard ML libraries (scikit-learn, pandas, numpy, matplotlib/seaborn, etc.)
- Document your development and testing environment

Code Requirements:

- Well-commented Python code suitable for handoff to the deployment team
- Clear documentation of preprocessing steps
- Reproducible analysis pipeline

Reporting Standards:

- Professional presentation
- Clear, concise language appropriate to each audience
- APA 7th edition referencing for all sources
- Proper citations for any external resources, code, or methodologies

9. SUCCESS CRITERIA

What "Good" Looks Like:

For Technical Report (Part A): ✓ Rigorous EDA with meaningful insights
✓ Well-justified model selection and comparison
✓ Clear interpretation of model performance
✓ Practical deployment recommendations
✓ Clean, well-documented code

For Business Report (Part B): ✓ Clear answers to all six analytical questions
✓ Business-friendly language (minimal technical jargon)
✓ Actionable recommendations grounded in data insights
✓ Thoughtful discussion of ethical/legal implications
✓ Clear articulation of stakeholder benefits

Overall: ✓ Demonstrates understanding of the educational context
✓ Balances technical rigor with practical applicability
✓ Provides genuine value to schools and educators
✓ Considers real-world implementation challenges

10. PROJECT TIMELINE & DELIVERABLES DUE

Project Duration: 6 weeks
Final Deliverables Due: Friday, October 4, 2024, 8:00 PM AEST / 3:30 PM IST

Submission Requirements:

- Technical Analytics Report (Part A) in PDF and Word formats
- Business Consultancy Report (Part B) in PDF and Word formats
- Python notebook (.ipynb)
- PDF version of Python notebook

11. ETHICAL & LEGAL CONSIDERATIONS

As you develop this solution, please give careful consideration to:

- **Student privacy:** Handling sensitive educational and demographic data
- **Fairness and bias:** Ensuring models don't perpetuate or amplify existing inequities
- **Transparency:** Making model decisions interpretable for educators
- **Data protection:** Compliance with Australian privacy laws and educational data regulations
- **Ethical use of predictions:** How schools will use at-risk predictions responsibly
- **Informed consent:** Data collection and usage considerations

Your Business Report (Part B) should address these implications explicitly.

12. QUESTIONS & SUPPORT

For questions about:

- **Data definitions and technical specifications:** Contact Dr Alok Sinha, Director of Data and Insights
- **Educational context and business requirements:** Contact Sally Tran, Director of Education and Engagement

We look forward to working with you on this important initiative to improve educational outcomes for primary school students across Australia.

Data2Intel Learning Analytics
Transforming educational data into student success