AI Solution for Predicting Student Dropout Risk

1. Problem Definition (6 points)

Hypothetical AI Problem: Predicting Student Dropout Risk in a university.

3 Objectives:

1. Early Identification: Identify at-risk students early in the semester for timely intervention
2. Resource Allocation: Efficiently allocate academic support resources to neediest students
3. Improve Retention Rates: Increase university retention and graduation rates

2 Stakeholders:

1. Students: Primary subjects affected by interventions
2. University Administration: Decision-makers responsible for student success

1 Key Performance Indicator (KPI):

- Precision of "At-Risk" Predictions - Measures how many flagged students actually drop out, minimizing false alarms

2. Data Collection & Preprocessing (8 points)

2 Data Sources:

1. Student Information System (SIS): Demographics, GPA, test scores, course history
2. Learning Management System (LMS): Login frequency, assignment submissions, forum activity

1 Potential Bias:

- Socioeconomic Bias: Underrepresentation of low-income students who don't seek help, causing model failure for these groups

3 Preprocessing Steps:

1. Missing Data Handling: Median imputation for numerical data, new category for categorical
2. Categorical Encoding: One-Hot Encoding for text data like majors
3. Normalization: Scale numerical features to common range (0-1)

3. Model Development (8 points)

Chosen Model: Gradient Boosting Machine (XGBoost/LightGBM)

Justification: Excellent for tabular data, handles complex non-linear relationships, provides feature importance, robust to outliers

Data Splitting:

-   Training (70%): 2018-2020 data
-   Validation (15%): 2021 data (hyperparameter tuning)
-   Test (15%): 2022 data (final evaluation)

2 Hyperparameters to Tune:

1. `max_depth`: Controls tree depth to prevent over/underfitting
2. `learning_rate`: Balances model performance vs. training time

4. Evaluation & Deployment (8 points)

2 Evaluation Metrics:

1. Precision-Recall Curve (AUPRC): Ideal for imbalanced datasets like dropout prediction
2. F2-Score: Emphasizes recall to prioritize finding all at-risk students

Concept Drift & Monitoring:

-   Concept Drift: When statistical properties of target variable change over time (e.g., post-pandemic behavior changes)
-   Monitoring: Track performance metrics on recent labeled data; significant drops signal drift

1 Technical Challenge:

- Scalability & Latency: Real-time prediction demands for advisor dashboards
- Solution: Cloud auto-scaling or batch pre-computation of risk scores