Predicting Student Performance and Risk of Failure

Leveraging Data Analytics and Machine Learning for Educational Success

The Unleveraged Potential of Educational Data

Educational institutions possess vast amounts of student data, yet often struggle to harness its power for actionable insights. This leads to missed opportunities for student success.

Identify Struggling Students Early

Missed opportunities to intervene before academic issues escalate.

Improve Teaching Quality

Lack of data-driven feedback hinders pedagogical enhancements.

Optimize Resource Allocation

Inefficient deployment of resources due to limited foresight.

Enhance Engagement & Retention

Challenges in keeping students motivated and enrolled.

Our Project: A Data-Driven Decision Support System

This initiative focuses on building a robust system to predict student performance, empowering educators with timely, precise information.

Core Objective

To predict student performance and risk of failure by leveraging existing institutional data.

- Proactive identification of at-risk students.
- Informed decision-making for tailored interventions.
- Enhancing overall student outcomes and institutional efficiency.

Technology Stack

SQL

Data extraction and structuring

Power BI

Interactive data visualization

Python

Advanced machine learning

Phase 1: Data Preparation & Integration

The foundation of any robust analytical system begins with meticulous data collection and preparation.

1

Synthetic Dataset Creation

A comprehensive dataset was generated, simulating 10,000 students, encompassing various attributes:

- Demographics & Family Background
- Academic History (Exams, Grades)
- Attendance Records
- Teacher Performance Ratings
- Extracurricular Activities
- Disciplinary Incidents

2

Robust Data Pipeline in Snowflake

Leveraging Snowflake SQL, a streamlined data pipeline was constructed to:

- Extract raw data efficiently.
- Perform rigorous data cleaning for consistency and accuracy.
- Merge disparate data sources into a cohesive, student-level dataset.
- Ensure data integrity and readiness for analysis and modeling.

Our data pipeline ensures clean, integrated data for reliable insights.

Phase 2: In-Depth Data Analysis

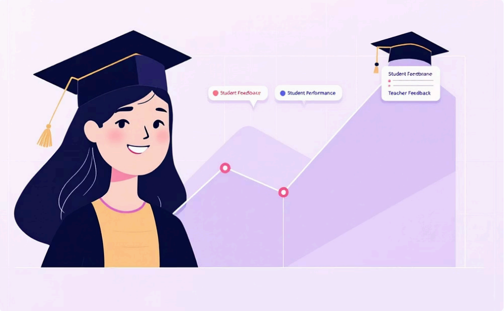
Utilizing SQL and Power BI, we explored various dimensions of student data to uncover trends and patterns.

Academic Trends

Detailed examination of performance by grade, subject, gender, family income, and parental education.

Attendance Patterns

Analysis of attendance data and its potential correlations with academic outcomes.



Teacher Effectiveness

Insights into how teacher ratings and qualifications relate to student achievement.

Activities & Discipline

Investigation into the impact of extracurricular engagement and disciplinary incidents.

This comprehensive analysis forms the bedrock for understanding the multifaceted influences on student performance.

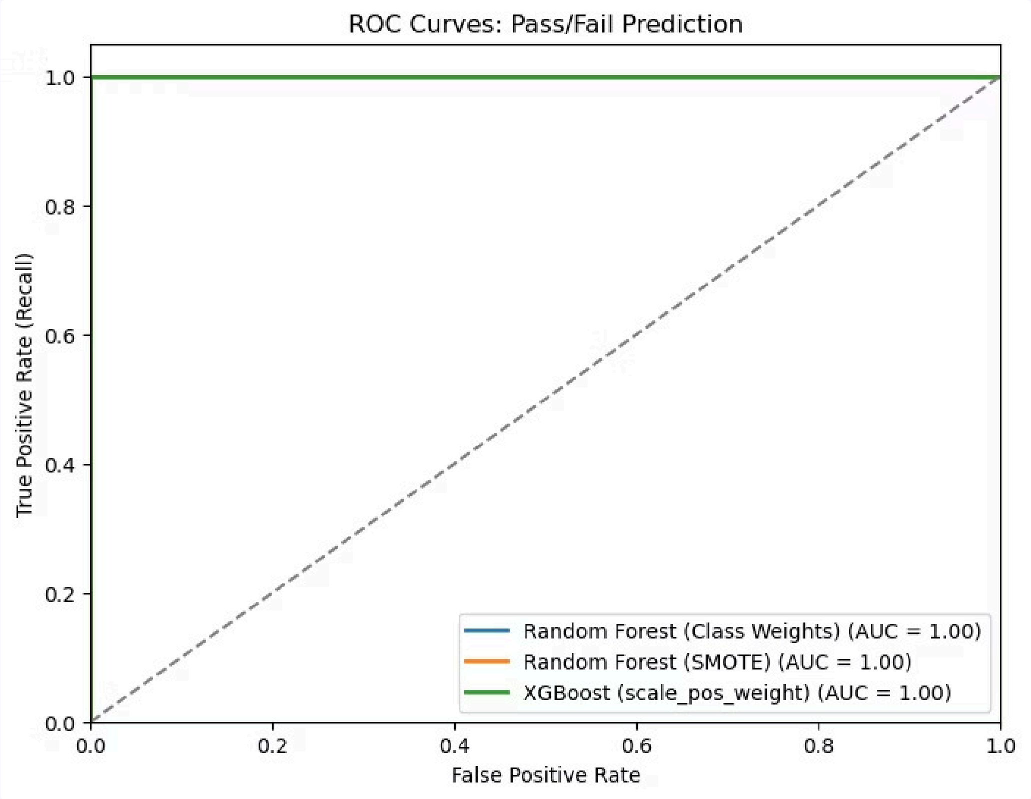
Phase 3: Predictive Modeling & Challenges

Python was employed to build predictive models, revealing critical insights and limitations.

Scenario A: With Past Scores

When past exam scores were included as features, models achieved near-perfect predictions.

- **Models:** Random Forest, XGBoost
- **Outcome:** Achieved perfect (100%) prediction accuracy.
- **Insight:** Past academic performance is an extremely strong predictor of future results.
- **Challenge:** This indicates data leakage, as "pass/fail" was defined based on these very s

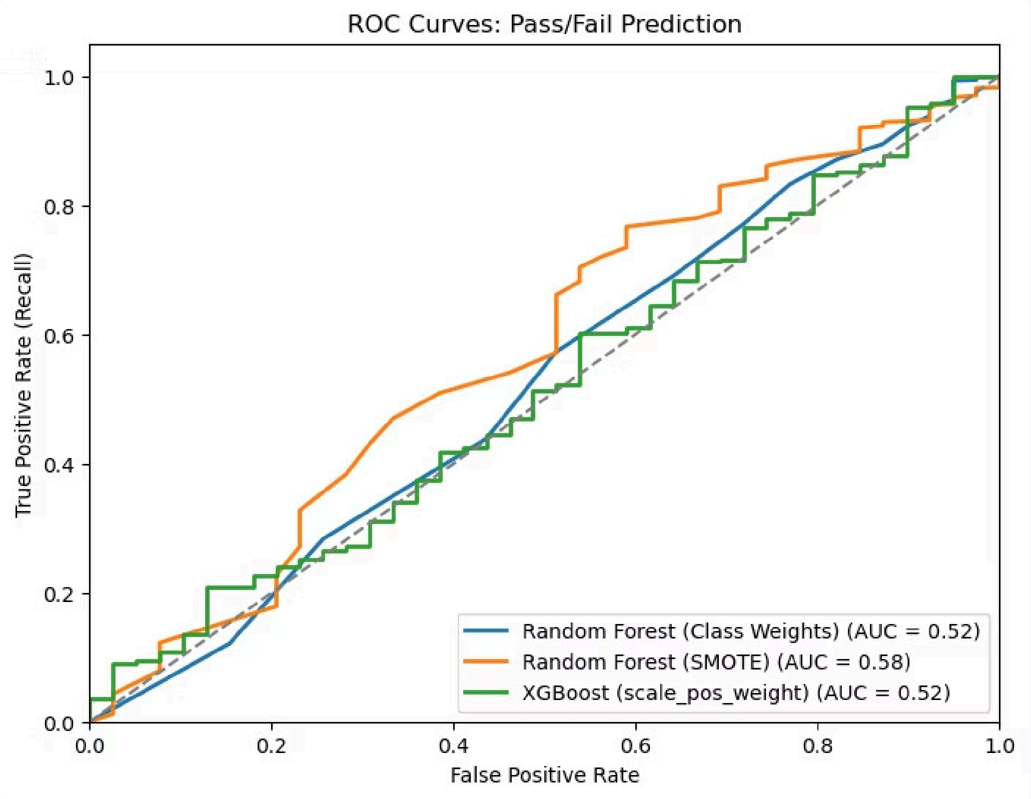


- cores.

Scenario B: Without Past Scores

Excluding past scores highlighted the true difficulty of early prediction.

- **Models:** Random Forest, XGBoost
- **Outcome:** Models struggled significantly to identify failing students.
- **Insight:** A class imbalance (98% pass rate) made identifying the 2% failure rate challenging.
- **Challenge:** Other predictors proved weak, leading to a high rate of missed true negatives.



Key Findings: Unpacking Predictive Factors

Our analysis revealed which factors truly drive student performance, and which have less direct impact.

Past Scores: The Strongest Predictor

Unsurprisingly, a student's previous academic performance remains the most reliable indicator of future success.

Minimal Direct Correlation

Attendance, teacher ratings, teacher qualifications, and participation in activities showed limited direct correlation with exam scores.

Parental Education's Slight Influence

Family income showed minimal impact on academic results, while higher parental education had a slight positive correlation.

Key Findings: Beyond Academic Metrics

We also explored the nuanced impact of extracurricular activities and disciplinary issues.

Extracurriculars & Holistic Development

While crucial for overall personal growth, social skills, and well-being, extracurricular activities did not directly correlate with improved exam scores.

Discipline: Behavior vs. Ability

Disciplinary issues were found to reflect behavioral challenges and peer influence rather than a student's inherent academic capability.

The Early Intervention Challenge

Without past exam scores, models struggled to identify failing students, highlighting the need for richer, continuous data for true early intervention.

Recommendations for Stakeholders

Translating insights into action: strategies for effective data utilization.

01

Leverage Existing Scores

Utilize past academic data for accurate predictions for students with established history.

02

Build Early Warning Systems

Integrate continuous assessment data (quizzes, homework, teacher feedback) for real-time monitoring.

03

Address Class Imbalance

Implement techniques like SMOTE or class weights in ML models to improve identification of at-risk students.

04

Focus on Student Support

Provide comprehensive resources: study skills workshops, tutoring, stress counseling, and access to learning resources.

Holistic Development & The Path Forward

A balanced approach is crucial for both academic and personal growth, guiding students to success.

Encourage Holistic Growth

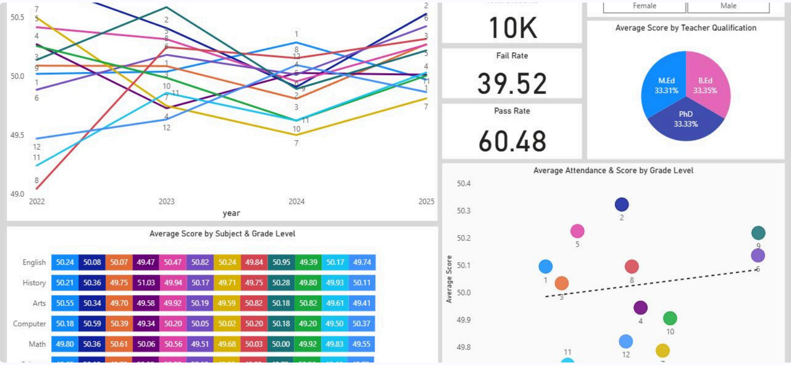
Promote sports, arts, and other activities for cognitive and personal development, recognizing their broader value.

Separate Interventions

Implement distinct strategies for academic underperformance versus behavioral issues, tailoring support effectively.

> Predicting future performance is easy with past exam data, but challenging without it. True early intervention requires schools to collect granular, continuous, and behavior-linked data beyond exams. Predictive analytics should be used as a **support tool**, not a judgment tool — guiding teachers and administrators to proactively assist students.

Dashboard



10K

Fail Rate

39.52

Pass Rate

60.48

Average Score by Subject & Grade Level

English	50.24	50.08	50.07	49.47	50.47	50.82	50.24	49.84	50.95	49.39	50.17	49.74
History	50.21	50.36	49.75	51.03	49.94	50.17	49.71	49.75	50.28	49.80	49.93	50.11
Arts	50.55	50.34	49.70	49.58	49.92	50.19	49.59	50.82	50.18	50.82	49.61	49.41
Computer	50.18	50.59	50.39	49.34	50.20	50.05	50.02	50.20	50.18	49.20	49.50	50.37
Math	49.80	50.36	50.61	50.06	50.56	49.51	49.68	50.03	50.00	49.92	49.83	49.55

app.powerbi.com

Power BI Report

Report powered by Power BI

GitHub

shijin/
StudentPerformancePre...

To analyze a student's performance based on various factors and build a predictive model to predict future performances in examinations.

1 Contributor	0 Issues	0 Stars	0 Forks

GitHub

GitHub – shijin/StudentPerformancePredictionAnalysis: To analyze a stud...

To analyze a student's performance based on various factors and build a predictive model to predict future performances in examinations. –...

Presentation by Shijin Ramesh (Data Analyst)