Presented by Don Kipchumba

Predicting At-Risk Students Using Machine Learning

Trends and challenges

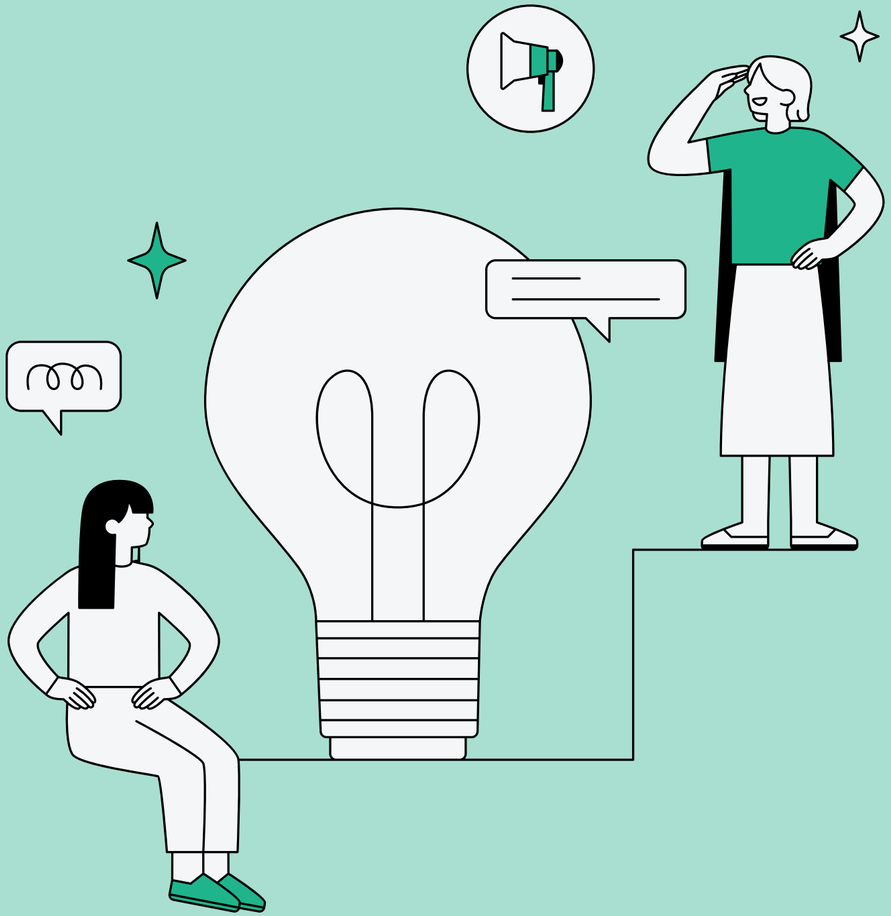
BUSINESS PROBLEM

Why This Project Matters

- Schools struggle to identify struggling students early
- Late intervention leads to poor academic outcomes
- Resources for academic support are limited

Business Goal

- Predict whether a student will pass or fail
- Enable early, data-driven intervention



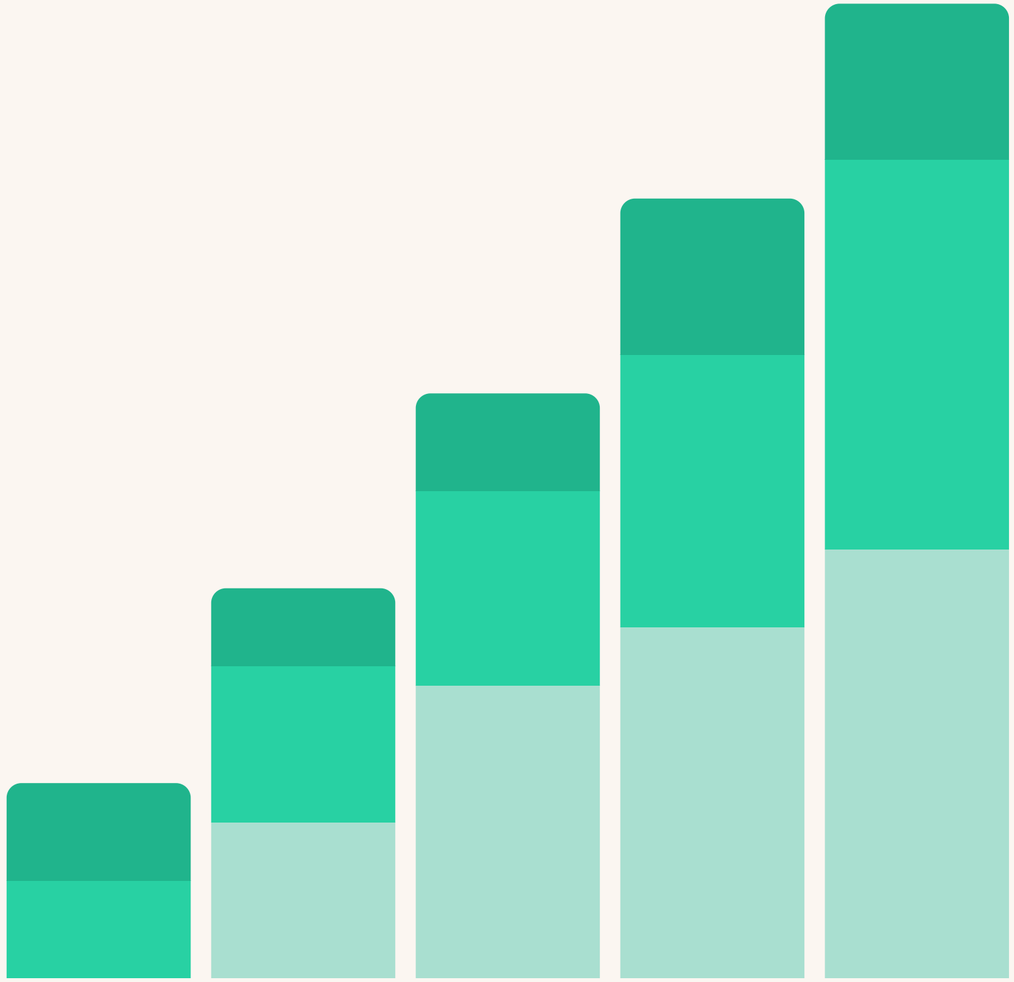
DATA OVERVIEW

Dataset Summary

- 1,000,000 student records
- Simulated academic performance dataset

Key Features

- Weekly self-study hours
- Attendance percentage
- Class participation
- Total score
- Final grade (target variable)



DATA PREPARATION

Data Preparation Steps

- Verified no missing values
- Removed unnecessary identifiers
- Created binary target:
  - Pass = 1
  - Fail = 0
- Split data into:
  - Training
  - Validation
  - Test sets

Key Observations

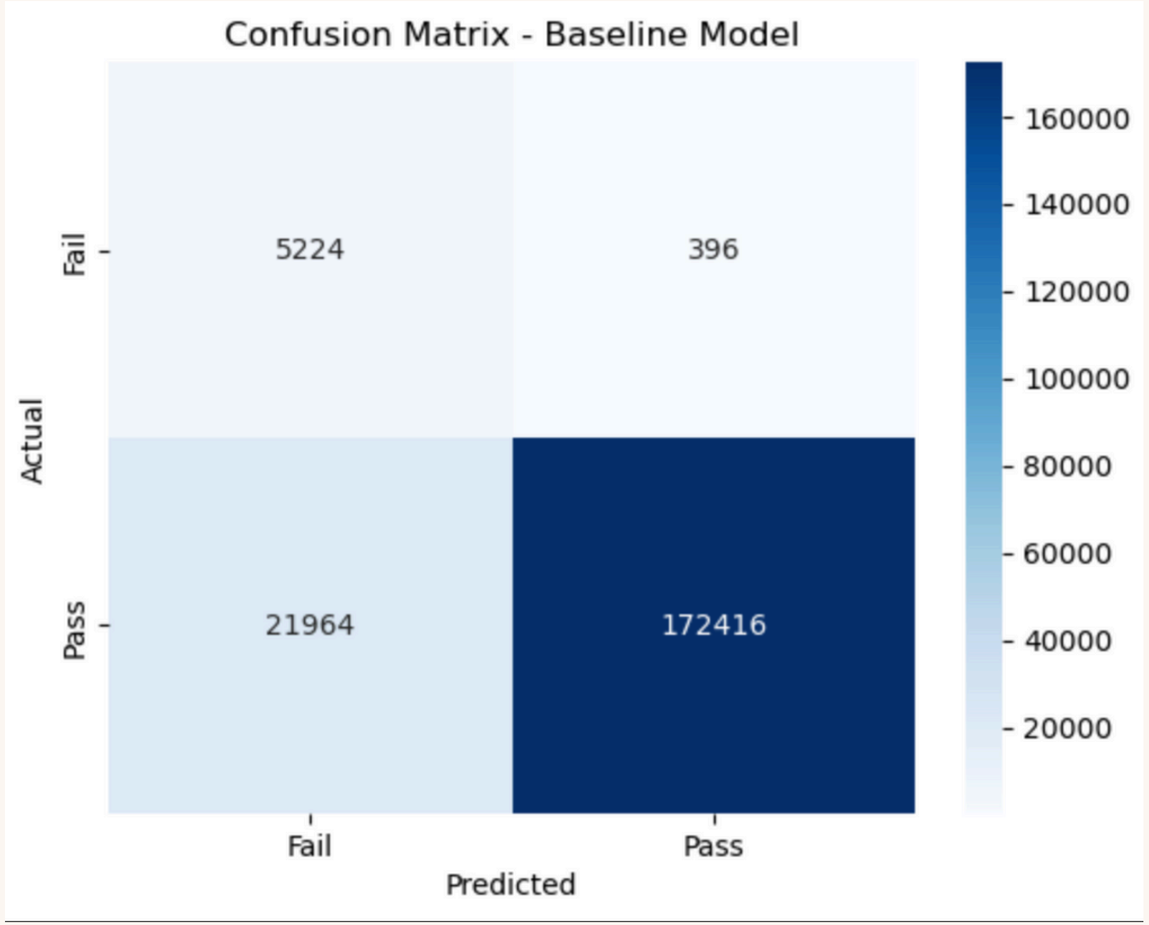
- 97% of students passed
- Dataset is highly imbalanced
- Strong correlation between:
- Study hours and performance
- Attendance and success
- Failing students are rare but critical to detect

BASELINE MODEL

Baseline Approach
- Logistic Regression
- Simple, interpretable benchmark

Performance
- Accuracy: 89%
- Recall (Fail): 93%
- Recall (Pass): 89%

Limitation
- Poor precision for failing students
- High false-positive rate



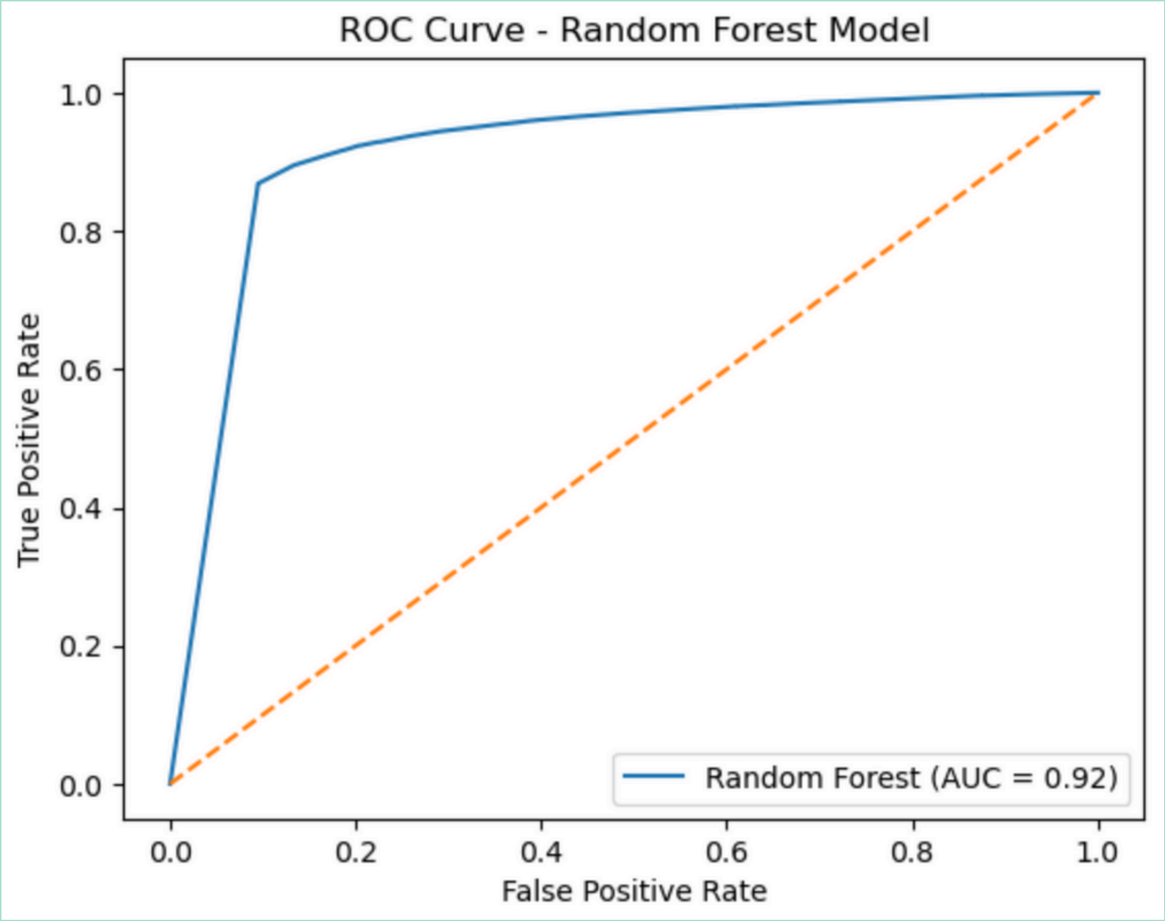
FINAL MODEL (RANDOM FOREST)

Why Random Forest

- Captures non-linear relationships
- Handles feature interactions well
- Better suited for complex patterns

Selected as Final Model

- Based on validation performance
- Improved generalization



MODEL EVALUATION

Final Model Performance

- Accuracy: 97%
- Recall (Pass): 99%
- Recall (Fail): 27%
- ROC-AUC: ~0.88

Why These Metrics

- Recall prioritized to reduce missed students
- ROC-AUC evaluates overall classification quality

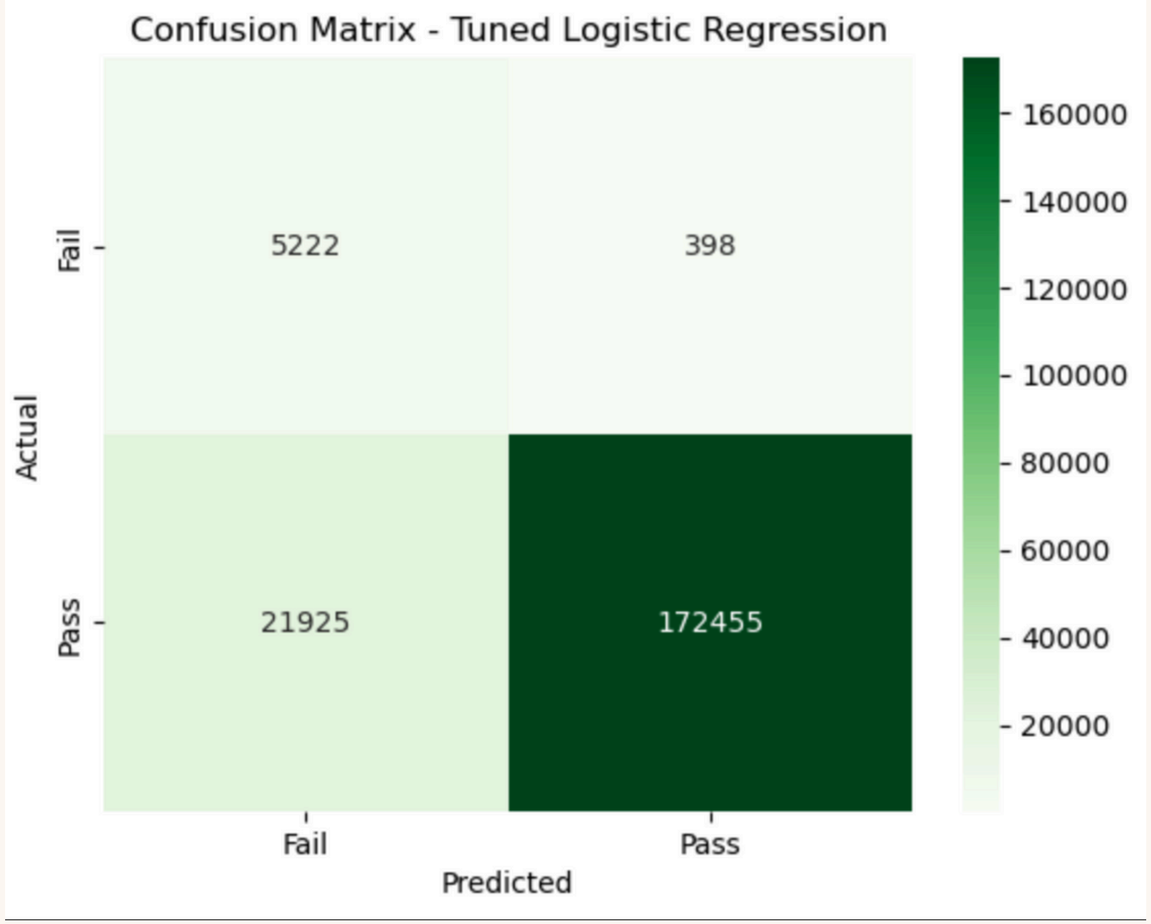
CONFUSION MATRIX

What the Model Does Well

- Excellent at identifying students who will pass
- Strong overall predictive power

Current Limitation

- Struggles to detect failing students
- Class imbalance affects sensitivity



LIMITATIONS & FUTURE WORK

Current Limitations

- Highly imbalanced dataset
- Synthetic data

Next steps

- Add real academic data
- Include socio-economic features

CONCLUSION

- Data can predict student performance effectively
- Early intervention is achievable
- Model supports better academic decision-making

Presented by Don Kipchumba

Thank you very much!