# Predicting Achievement in Mathematics

Presented by: Erica, Melissa, & Jess

# Reason for Choosing Mathematics Achievement

- Two of the three members are currently classroom teachers.
- Achievement in mathematics develops critical thinking skills, is essential for many careers, promotes intellectual curiosity, enhances cognitive abilities, and can lead to innovation.
- We were interested in creating a model that could be used to predict achievement so that students could be provided early and targeted intervention.

# Description of Data Source

We chose to use a fictional dataset due to the ethical and legal concerns about using real student data. The site we used provides data solely for data science training purpose.

# Questions Guiding the Project

**Q1**: What factors best predict mathematics achievement?

**Q2**: Are there differences in math achievement by different demographic features?

# Data Exploration Phase

- Data cleaning
  - Check for missing values, outliers, & duplicates
  - Data type inspection
- Data transformation
  - Dummy coding categorical features
  - Math score conversion
    - "PASS": score ≥ 70 to 1; "FAIL": score＜70 to 0
- Data Visualizations
  - Investigate data distribution, correlations, & patterns

# Analysis Phase for Q1

- Mock Machine Learning Models
- Hyperparameter Tuning
- Optimization
- Top 3 Models

| Model | Accuracy |
|---|---|
| LogisticRegression | 0.891129 |
| GradientBoostingClassifier | 0.870968 |
| KNeighborsClassifier | 0.842742 |

# Technologies, Languages, Tools, & Algorithms

- PostgreSQL/ PGAdmin
- Jupyter Lab/ Notebook
- Python
- Google Slides
- PANDAS
- SKlearn
- Tableau
- SQLAlchemy
- Logistic Regression
- Gradient Boosting Classifer
- KNeighbors Classifier
- Google Drive

# Result of Analysis for Q1

## Optimized Logistic Regression Model: Confusion Matrix

| | Predicted Fail | Predicted Pass |
|---|---|---|
| Actual Fail | 113 (True Negative) | 8 (False Positive) |
| Actual Pass | 19 (False Negative) | 108 (True Positive) |

# Result of Analysis for Q1

## Optimized Logistic Regression Model: Classification Report

| | precision | recall | specificity | f1 | geometric mean | Index balanced accuracy (IBA) | # of samples |
|---|---|---|---|---|---|---|---|
| Fail | 0.86 | 0.93 | 0.85 | 0.89 | 0.89 | 0.80 | 121 |
| Pass | 0.93 | 0.85 | 0.93 | 0.89 | 0.89 | 0.79 | 127 |
| avg/ total | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.79 | 248 |

# Result of Analysis for Q1

Optimized Logistic Regression Model: Feature Importances

| Feature | Importance |
|---|---|
| Sex_female | 0.153848 |
| Lunch_reduced | 0.123344 |
| Test_Prep_completed | 0.106656 |
| Test_Prep_none | 0.090922 |
| Lunch_standard | 0.074234 |
| Ethnicity_group C | 0.053501 |
| Ethnicity_group A | 0.051653 |
| Ethnicity_group B | 0.051303 |
| Sex_male | 0.043731 |
| Ethnicity_group D | 0.043321 |
| PLE_bachelor degree | 0.042782 |
| PLE_some high school | 0.042444 |
| PLE_high school | 0.035400 |
| PLE_some college | 0.030720 |
| PLE_associate degree | 0.025032 |
| PLE_master degree | 0.021200 |
| Reading_Score | 0.007707 |
| Ethnicity_group E | 0.002200 |

# Analysis Phase for Q2

Tableau Interactive Dashboard

# Result of Analysis for Q2

## Overall Observations

- Math proficiency is:
  - lower for females
  - lower for students receiving free or reduced lunch
  - highest for Ethnic Group E
  - increases as Parent Level of Education increases
  - higher for students who completed a test prep course

# Recommendations for Future Analysis

- Significance testing for different demographic features
- Additional feature analysis
  - Feature binning
- Use real student data

# What We Would Do Differently

- Data Cleansing
  - Explore Manual encoding vs. dummy variable encoding to observe the impact on feature importance
- Given more time: use real student data
- Class question: any other ideas for optimization?

Thank you!
Any questions?