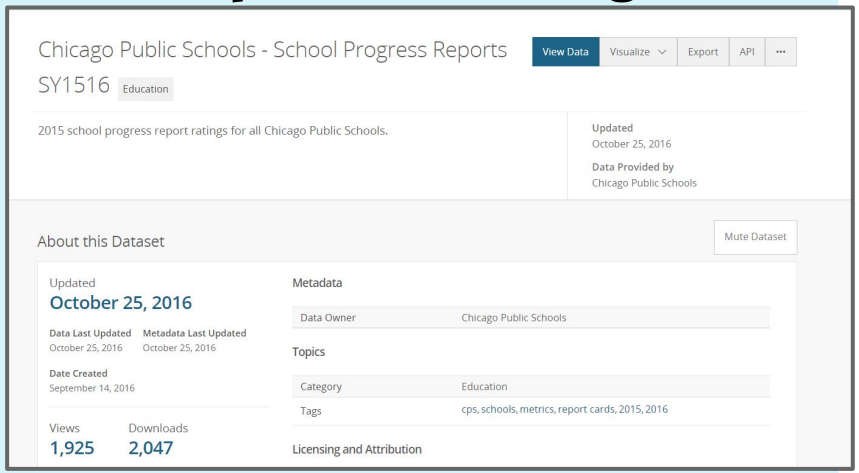
DATA602 Final Project: Piloting ML models with Chicago School Performance Data

2021 Scott Hirabayashi | Krishnamoorthy | DATA 602

Agenda

- Introduction and Project Goals
- Process Discussion and Data Issues
- Model uses and findings
- Next Steps

Data Source: Open Data Chicago



School Progress Reports 1718

Motivations

- I'm interested in school performance, particularly within High Schools.
- Schools have a limited budget, staff, and limited resources for taking action.
- A binary classification model **that is predictive of meeting or not meeting the national standards for Chicago Public Schools.**
    - ***Can we create an accurate model of school features to predict a school's graduation rates?***

Project Goals

Model:

Discover if current Educational data is sufficient for predicting whether or not a school is meeting the national standard.

Consider Features:

Look for features which may have high correlation or effect in successful models on graduation rates.

Apply to Other Data

Ultimately, I would like to use the model selection and feature selection as a guide to analyze Chicago high schools over a period of several years.

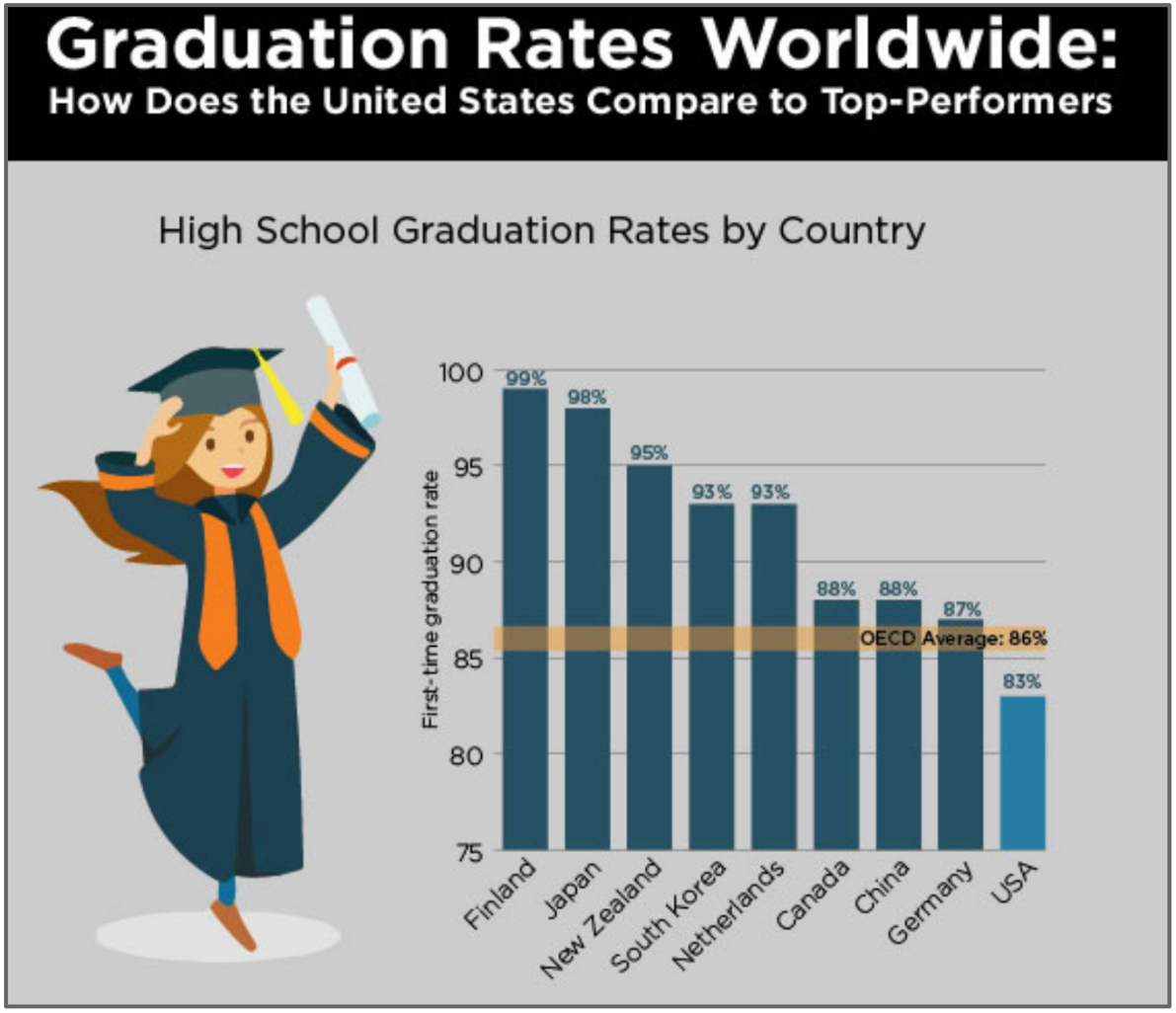
Question: What was the 4-year graduation rate for US Public High Schools in 2017?

A. 88%

B. 92%

C. 83%

D. 79%



C- 83%. It may sound high, but the US has lower graduation rates than comparable countries

https://ncee.org/quick-read/graduation-rates-worldwide/

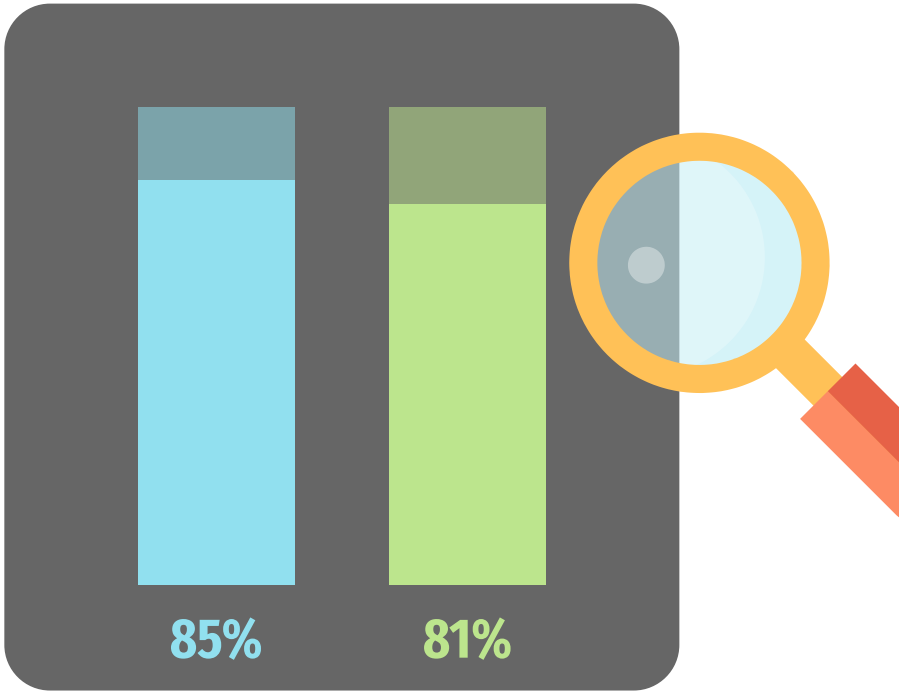
Public High School Graduation Rates (17-18)

US and Illinois

The 2017-2018 average US graduation rate was 85%.
The average Illinois 4-year graduation rate is 85% as well.

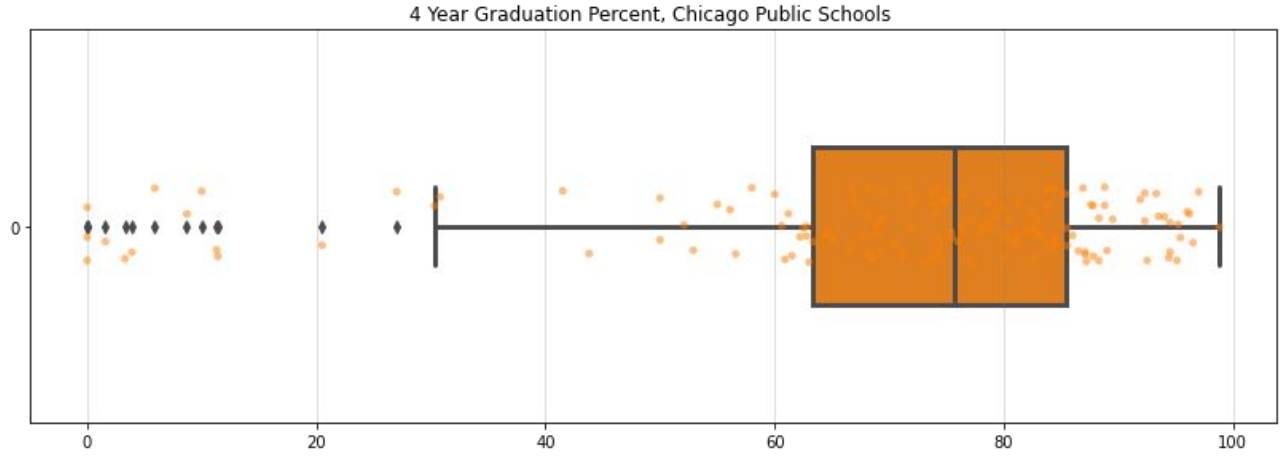
Chicago

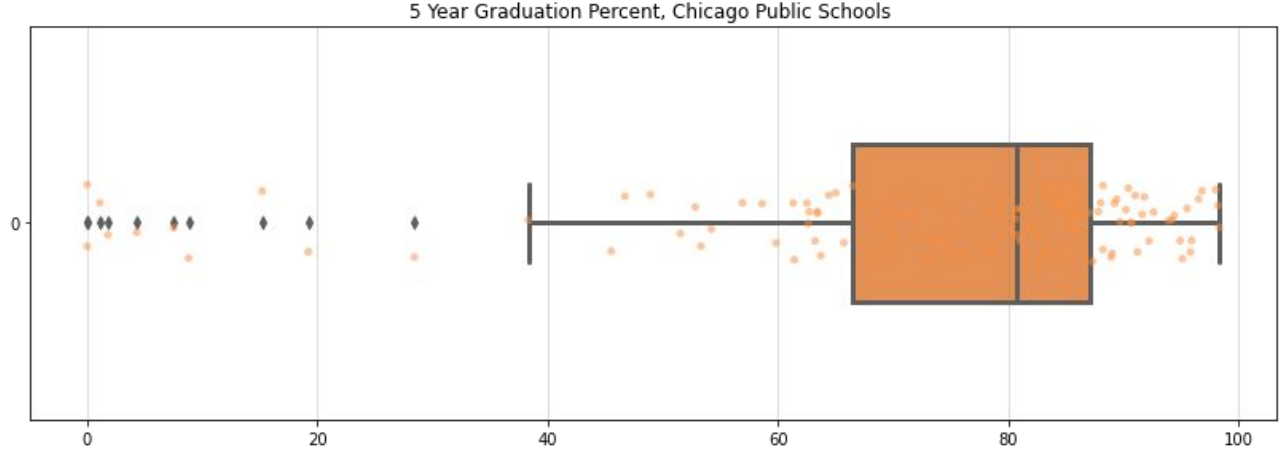
Chicago's high school graduation rate was around **81%** (80.9% with revised methodology).*



https://chicago.chalkbeat.org/2020/9/4/21422872/chicago-public-schools-reports-record-high-graduation-rate-spring-2020-amid-coronavirus-pandemic

4- and 5- year graduation rate distributions





Target Variable

The target variable is the graduation rate of a school meeting standard. If it is **85% or over, it meets standard and is encoded as 1,** otherwise it is encoded as 0.

ONLY schools that have a recorded 4 year graduation rate are considered here.

```
#Break Graduation Rates into two categories as binary classifier
categorized_graduation_rates = []

for pct in high_schools.Graduation_4_Year_School_Pct_Year_2:
    if pct >= 85:
        categorized_graduation_rates.append(1)
    else:
        categorized_graduation_rates.append(0)

categorized_graduation_rates[0:10]

```

Simple classification conversion

Procedure

Thoughts on Features

My expectations when starting this project were that there would be many predictors/features that would be obvious: high dropout rate, low teacher attendance, low student attendance/chronic truancy, low test scores, etc. would all be strong predictors of schools that were not meeting standards.

I attempted to keep as many features as possible to analyze them in future iterations.

Data Cleaning

Columns Removed

- Any columns that didn't contain at least 75% of data (>25% nulls) were dropped.
- Any columns that were descriptive or average columns were dropped.
- Any columns that were unable to be numerically converted were dropped.

Columns Cleaned

- Survey columns had values such as 'Far Above Average', 'Above Average', etc., and were encoded
- Columns were imputed with the average of the column itself.
- After all nulls were imputed, checked correlation of variables to target and kept any features greater than .1 correlation.
- Data was normalized with MinMaxScaler()..

School_ID, Short_Name, Long_Name, School_Type, Primary_Category, Address, City, State, Zip, Phone, Fax, CPS_School_Profile, Website, Progress_Report_Year, Blue_Ribbon_Award_Year, Excelerate_Award_Gold_Year, Spot_Light_Award_Year, Improvement_Award_Year, Excellence_Award_Year, Student_Growth_Rating, Student_Growth_Description, Growth_Reading_Grades_Tested_Pct_ES, Growth_Reading_Grades_Tested_Label_ES, Growth_Math_Grades_Tested_Pct_ES, Growth_Math_Grades_Tested_Label_ES, Student_Attainment_Rating, Student_Attainment_Description, Attainment_Reading_Pct_ES, Attainment_Reading_Lbl_ES, Attainment_Math_Pct_ES, Attainment_Math_Lbl_ES, Culture_Climate_Rating, Culture_Climate_Description, School_Survey_Student_Response_Rate_Pct, School_Survey_Student_Response_Rate_Avg_Pct, School_Survey_Teacher_Response_Rate_Pct, School_Survey_Teacher_Response_Rate_Avg_Pct, School_Survey_Parent_Response_Rate_Pct, School_Survey_Parent_Response_Rate_Avg_Pct, Healthy_School_Certification, Healthy_School_Certification_Description, Creative_School_Certification, Creative_School_Certification_Description, NWEA_Reading_Growth_Grade_3_Pct, NWEA_Reading_Growth_Grade_3_Lbl, NWEA_Reading_Growth_Grade_4_Pct, NWEA_Reading_Growth_Grade_4_Lbl, NWEA_Reading_Growth_Grade_5_Pct, NWEA_Reading_Growth_Grade_5_Lbl, NWEA_Reading_Growth_Grade_6_Pct, NWEA_Reading_Growth_Grade_6_Lbl, NWEA_Reading_Growth_Grade_7_Pct, NWEA_Reading_Growth_Grade_7_Lbl, NWEA_Reading_Growth_Grade_8_Pct, NWEA_Reading_Growth_Grade_8_Lbl, NWEA_Math_Growth_Grade_3_Pct, NWEA_Math_Growth_Grade_3_Lbl, NWEA_Math_Growth_Grade_4_Pct, NWEA_Math_Growth_Grade_4_Lbl, NWEA_Math_Growth_Grade_5_Pct, NWEA_Math_Growth_Grade_5_Lbl, NWEA_Math_Growth_Grade_6_Pct, NWEA_Math_Growth_Grade_6_Lbl, NWEA_Math_Growth_Grade_7_Pct, NWEA_Math_Growth_Grade_7_Lbl, NWEA_Math_Growth_Grade_8_Pct, NWEA_Math_Growth_Grade_8_Lbl, NWEA_Reading_Attainment_Grade_2_Pct, NWEA_Reading_Attainment_Grade_2_Lbl, NWEA_Reading_Attainment_Grade_3_Pct, NWEA_Reading_Attainment_Grade_3_Lbl, NWEA_Reading_Attainment_Grade_4_Pct, NWEA_Reading_Attainment_Grade_4_Lbl, NWEA_Reading_Attainment_Grade_5_Pct, NWEA_Reading_Attainment_Grade_5_Lbl, NWEA_Reading_Attainment_Grade_6_Pct, NWEA_Reading_Attainment_Grade_6_Lbl, NWEA_Reading_Attainment_Grade_7_Pct, NWEA_Reading_Attainment_Grade_7_Lbl, NWEA_Reading_Attainment_Grade_8_Pct, NWEA_Reading_Attainment_Grade_8_Lbl, NWEA_Math_Attainment_Grade_2_Pct, NWEA_Math_Attainment_Grade_2_Lbl, NWEA_Math_Attainment_Grade_3_Pct, NWEA_Math_Attainment_Grade_3_Lbl, NWEA_Math_Attainment_Grade_4_Pct, NWEA_Math_Attainment_Grade_4_Lbl, NWEA_Math_Attainment_Grade_5_Pct, NWEA_Math_Attainment_Grade_5_Lbl, NWEA_Math_Attainment_Grade_6_Pct, NWEA_Math_Attainment_Grade_6_Lbl, NWEA_Math_Attainment_Grade_7_Pct, NWEA_Math_Attainment_Grade_7_Lbl, NWEA_Math_Attainment_Grade_8_Pct, NWEA_Math_Attainment_Grade_8_Lbl, School_Survey_Involved_Families, School_Survey_Supportive_Environment, School_Survey_Ambitious_Instruction, School_Survey_Effective_Leaders, School_Survey_Collaborative_Teachers, School_Survey_Safety, Suspensions_Per_100_Students_Year_1_Pct, Suspensions_Per_100_Students_Year_2_Pct, Suspensions_Per_100_Students_Avg_Pct, Misconducts_To_Suspensions_Year_1_Pct, Misconducts_To_Suspensions_Year_2_Pct, Misconducts_To_Suspensions_Avg_Pct, Average_Length_Suspension_Year_1_Pct, Average_Length_Suspension_Year_2_Pct, Average_Length_Suspension_Avg_Pct, Behavior_Discipline_Year_1, Behavior_Discipline_Year_2, School_Survey_School_Community, School_Survey_Parent_Teacher_Partnership, School_Survey_Quality_Of_Facilities, Student_Attendance_Year_1_Pct, Student_Attendance_Year_2_Pct, Student_Attendance_Avg_Pct, Teacher_Attendance_Year_1_Pct, Teacher_Attendance_Year_2_Pct, Teacher_Attendance_Avg_Pct, One_Year_Dropout_Rate_Year_1_Pct, One_Year_Dropout_Rate_Year_2_Pct, One_Year_Dropout_Rate_Avg_Pct, Other_Metrics_Year_1, Other_Metrics_Year_2, Freshmen_On_Track_School_Pct_Year_2, Freshmen_On_Track_CPS_Pct_Year_2, Freshmen_On_Track_School_Pct_Year_1, Freshmen_On_Track_CPS_Pct_Year_1, Graduation_4_Year_School_Pct_Year_2, Graduation_4_Year_CPS_Pct_Year_2, Graduation_4_Year_School_Pct_Year_1, Graduation_4_Year_CPS_Pct_Year_1, Graduation_5_Year_School_Pct_Year_2, Graduation_5_Year_CPS_Pct_Year_2, Graduation_5_Year_School_Pct_Year_1, Graduation_5_Year_CPS_Pct_Year_1, College_Enrollment_School_Pct_Year_2, College_Enrollment_CPS_Pct_Year_2, College_Enrollment_School_Pct_Year_1, College_Enrollment_CPS_Pct_Year_1, College_Persistence_School_Pct_Year_2, College_Persistence_CPS_Pct_Year_2, College_Persistence_School_Pct_Year_1, College_Persistence_CPS_Pct_Year_1, Progress_Toward_Graduation_Year_1, Progress_Toward_Graduation_Year_2, State_School_Report_Card_URL, Mobility_Rate_Pct, Chronic_Truancy_Pct, Empty_Progress_Report_Message, School_Survey_Rating_Description, Supportive_School_Award, Supportive_School_Award_Desc, Parent_Survey_Results_Year, School_Latitude, School_Longitude, PSAT_Grade_9_Score_School_Avg, PSAT_Grade_10_Score_School_Avg, SAT_Grade_11_Score_School_Avg, SAT_Grade_11_Score_CPS_Avg, Growth_PSAT_Grade_9_School_Pct, Growth_PSAT_Grade_9_School_Lbl, Growth_PSAT_Reading_Grade_10_School_Pct, Growth_PSAT_Reading_Grade_10_School_Lbl, Growth_SAT_Grade_11_School_Pct, Growth_SAT_Grade_11_School_Lbl, Attainment_PSAT_Grade_9_School_Pct, Attainment_PSAT_Grade_9_School_Lbl, Attainment_PSAT_Grade_10_School_Pct, Attainment_PSAT_Grade_10_School_Lbl, Attainment_SAT_Grade_11_School_Pct, Attainment_SAT_Grade_11_School_Lbl, Attainment_All_Grades_School_Pct, Attainment_All_Grades_School_Lbl, Growth_PSAT_Math_Grade_10_School_Pct, Growth_PSAT_Math_Grade_10_School_Lbl, Growth_SAT_Reading_Grade_11_School_Pct, Growth_SAT_Reading_Grade_11_School_Lbl, Growth_SAT_Math_Grade_11_School_Pct, Growth_SAT_Math_Grade_11_School_Lbl,

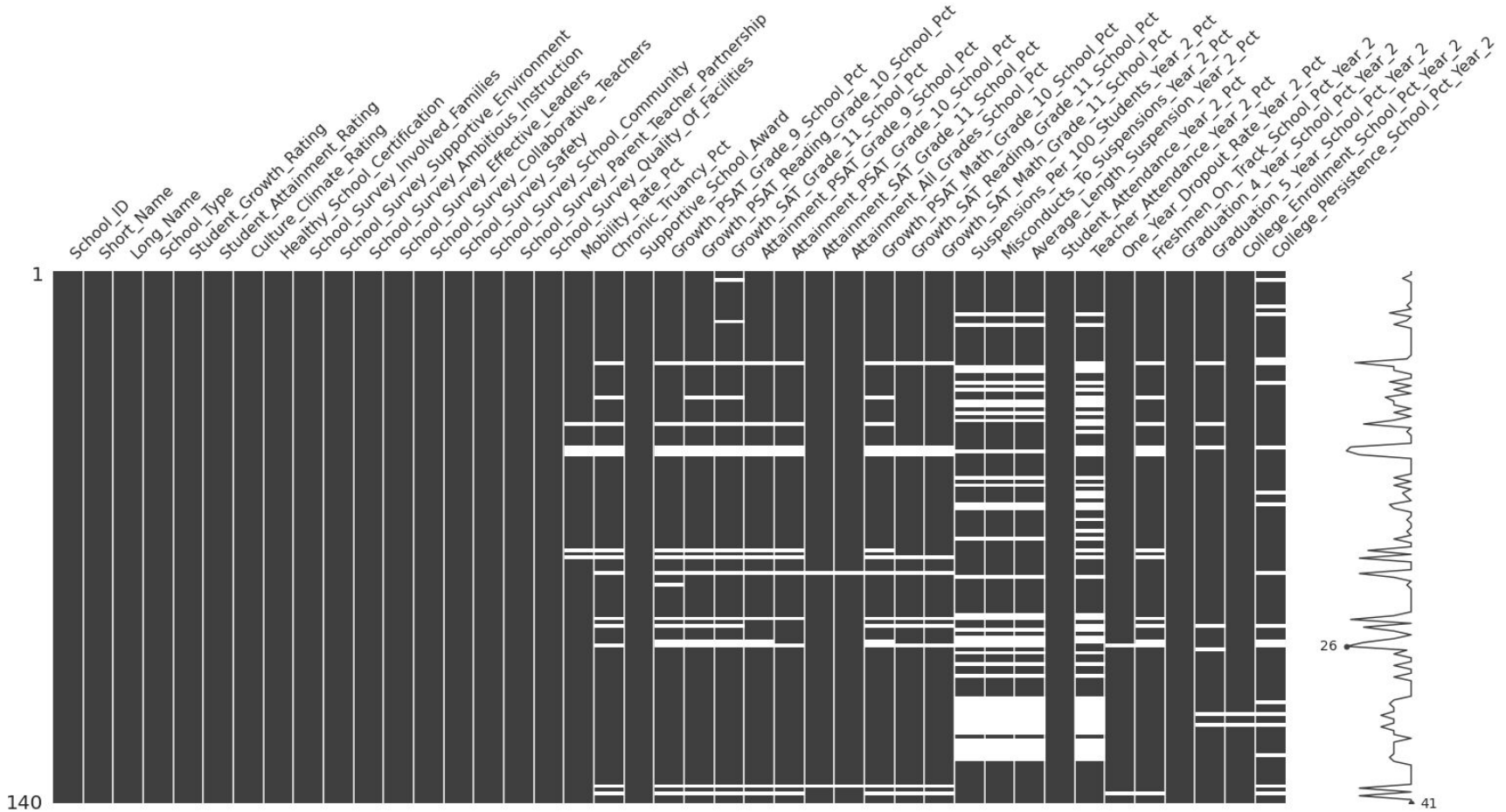
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 140 entries, 0 to 139
Data columns (total 27 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Student_Growth_Rating                  140 non-null    float64
 1   Student_Attainment_Rating              140 non-null    float64
 2   Culture_Climate_Rating                 140 non-null    float64
 3   Mobility_Rate_Pct                      140 non-null    float64
 4   Chronic_Truancy_Pct                    140 non-null    float64
 5   Growth_PSAT_Grade_9_School_Pct         140 non-null    float64
 6   Growth_SAT_Grade_11_School_Pct         140 non-null    float64
 7   Attainment_PSAT_Grade_9_School_Pct     140 non-null    float64
 8   Attainment_PSAT_Grade_10_School_Pct    140 non-null    float64
 9   Attainment_SAT_Grade_11_School_Pct     140 non-null    float64
 10  Attainment_All_Grades_School_Pct       140 non-null    float64
 11  Suspensions_Per_100_Students_Year_2_Pct 140 non-null   float64
 12  Misconducts_To_Suspensions_Year_2_Pct  140 non-null    float64
 13  Student_Attendance_Year_2_Pct          140 non-null    float64
 14  Teacher_Attendance_Year_2_Pct          140 non-null    float64
 15  One_Year_Dropout_Rate_Year_2_Pct       140 non-null    float64
 16  Freshmen_On_Track_School_Pct_Year_2    140 non-null    float64
 17  Graduation_4_Year_School_Pct_Year_2    140 non-null    float64
 18  Graduation_5_Year_School_Pct_Year_2    140 non-null    float64
 19  College_Enrollment_School_Pct_Year_2   140 non-null    float64
 20  College_Persistence_School_Pct_Year_2  140 non-null    float64
 21  survey_involved_family_num             140 non-null    float64
 22  survey_ambitious_inst_num              140 non-null    float64
 23  survey_effective_leaders_num           140 non-null    float64
 24  survey_collab_teachers_num             140 non-null    float64
 25  survey_safety_num                      140 non-null    float64
 26  target_graduation                      140 non-null    float64
dtypes: float64(27)
memory usage: 29.7 KB
```

Starting → Kept Features

Label Encoding for Survey Data

-Encodes to values,
-then treats any null values with mean.

```
def convert_ordinal_to_numerical(df_, impute_nulls = True):
    """This takes nominal categorical data and transforms it into a numerical value"""
    numerical_scores = []

    for val in list(df_):
        if val == 'FAR ABOVE AVERAGE' or val == 'FAR ABOVE EXPECTATIONS' or val == 'WELL ORGANIZED':
            numerical_scores.append(5)
        elif val == 'ABOVE AVERAGE' or val == 'ABOVE EXPECTATIONS' or val == 'ORGANIZED':
            numerical_scores.append(4)
        elif val == 'AVERAGE' or val == 'MET EXPECTATIONS' or val == 'MODERATELY ORGANIZED':
            numerical_scores.append(3)
        elif val == 'BELOW AVERAGE' or 'BELOW EXPECATATIONS' or val == 'PARTIALLY ORGANIZED':
            numerical_scores.append(2)
        elif val == 'FAR BELOW AVERAGE' or ' FAR BELOW EXPECTATIONS' or val == 'NOT YET ORGANIZED':
            numerical_scores.append(1)
        #I'll impute any missing values as average
        else: numerical_scores.append(np.nan)

    score_mean = round(np.nanmean(np.array(numerical_scores)),2)
    scores_arry = np.array(numerical_scores)

    if impute_nulls == True:
        return np.nan_to_num(scores_arry, nan = score_mean)

    else:
        return scores_arry
```


School_ID
Short_Name
Long_Name
School_Type
Student_Growth_Rating
Student_Attainment_Rating
Culture_Climate_Rating
Healthy_School_Certification
School_Survey_Involved_Families
School_Survey_Supportive_Environment
School_Survey_Ambitious_Instruction
School_Survey_Effective_Leaders
School_Survey_Collaborative_Teachers
School_Survey_Safety
School_Survey_School_Community
School_Survey_Parent_Teacher_Partnership
School_Survey_Quality_Of_Facilities
Mobility_Rate_Pct
Chronic_Truancy_Pct
Supportive_School_Award
Growth_PSAT_Grade_9_School_Pct
Growth_PSAT_Reading_Grade_10_School_Pct
Growth_SAT_Grade_11_School_Pct
Attainment_PSAT_Grade_9_School_Pct
Attainment_PSAT_Grade_10_School_Pct
Attainment_SAT_Grade_11_School_Pct
Attainment_All_Grades_School_Pct
Growth_PSAT_Math_Grade_10_School_Pct
Growth_SAT_Reading_Grade_11_School_Pct
Growth_SAT_Math_Grade_11_School_Pct
Suspensions_Per_100_Students_Year_2_Pct
Misconducts_To_Suspensions_Year_2_Pct
Average_Length_Suspension_Year_2_Pct
Student_Attendance_Year_2_Pct
Teacher_Attendance_Year_2_Pct
One_Year_Dropout_Rate_Year_2_Pct
Freshmen_On_Track_School_Pct_Year_2
Graduation_4_Year_School_Pct_Year_2
Graduation_5_Year_School_Pct_Year_2
College_Enrollment_School_Pct_Year_2
College_Persistence_School_Pct_Year_2
1
140
26
41

Train/Test Split

Train/Test Split = .7/.3

Classification Models Tested

- Logistic Regression
- Ridge Classifier
- Decision Tree
- Naive Bayes
- Random Forest
- AdaBoost
- Gradient Boost
- XG Boost

Best: Logistic and Ridge

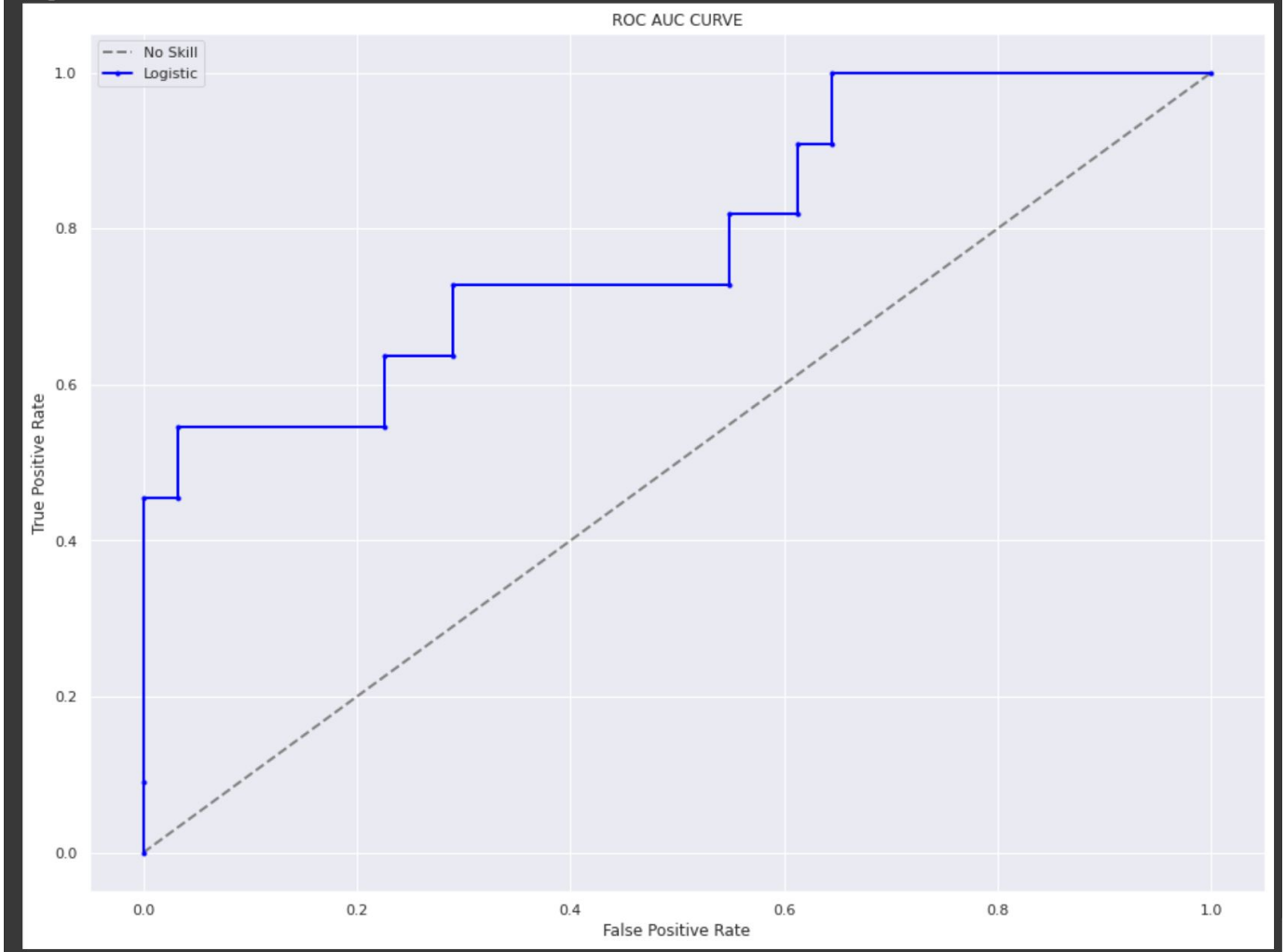
```
************************************************************
Logistic Regression
              precision    recall  f1-score   support

           0       0.83      0.97      0.90        31
           1       0.83      0.45      0.59        11

    accuracy                           0.83        42
   macro avg       0.83      0.71      0.74        42
weighted avg       0.83      0.83      0.82        42

Confusion Matrix:
[[30  1]
 [ 6  5]]
************************************************************
```

```
************************************************************
Ridge Classifier
              precision    recall  f1-score   support

           0       0.85      0.94      0.89        31
           1       0.75      0.55      0.63        11

    accuracy                           0.83        42
   macro avg       0.80      0.74      0.76        42
weighted avg       0.83      0.83      0.82        42

Confusion Matrix:
[[29  2]
 [ 5  6]]
************************************************************
```

Both had 83% accuracy on the test dataset.

ROC Curve returned a logistic score of .786, higher than a no-skill model

```
No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.786
```



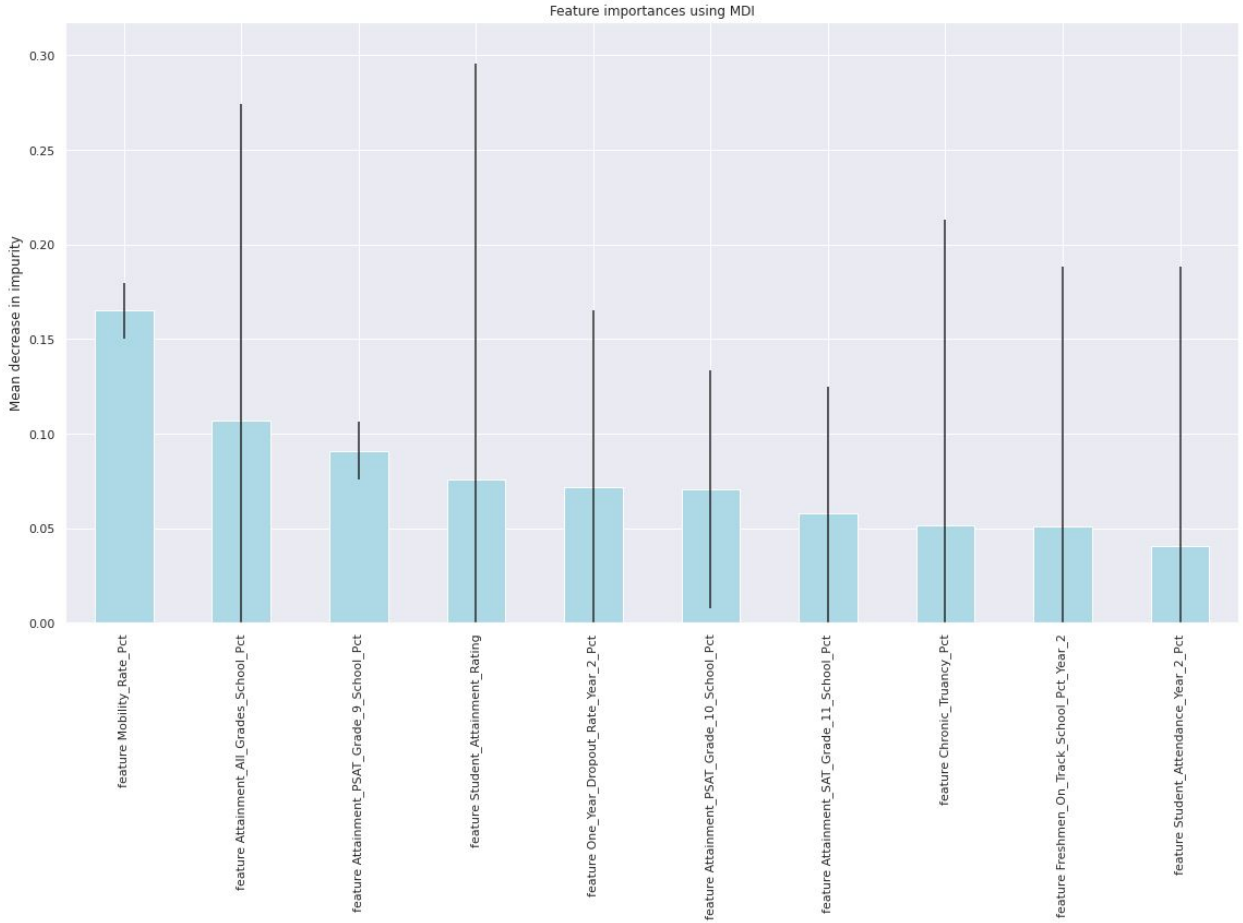
Worst: NB and Decision Tree

```
************************************************************
Naive Bayes
              precision    recall  f1-score   support

           0       0.83      0.61      0.70        31
           1       0.37      0.64      0.47        11

    accuracy                           0.62        42
   macro avg       0.60      0.62      0.59        42
weighted avg       0.71      0.62      0.64        42

Confusion Matrix:
[[19 12]
 [ 4  7]]
************************************************************
```

```
************************************************************
Decision Tree
              precision    recall  f1-score   support

           0       0.83      0.61      0.70        31
           1       0.37      0.64      0.47        11

    accuracy                           0.62        42
   macro avg       0.60      0.62      0.59        42
weighted avg       0.71      0.62      0.64        42

Confusion Matrix:
[[19 12]
 [ 4  7]]
************************************************************
```

Both had only 62% accuracy on the test dataset.

Feature Importance



Conclusions....

	Name	Predictor	Abs_predictor
0	Student_Growth_Rating	0.027877	0.027877
1	Teacher_Attendance_Year_2_Pct	0.063482	0.063482
2	survey_effective_leaders_num	-0.075418	0.075418
3	Suspensions_Per_100_Students_Year_2_Pct	-0.092485	0.092485
4	survey_collab_teachers_num	0.120262	0.120262
5	survey_safety_num	-0.217361	0.217361
6	Culture_Climate_Rating	-0.384519	0.384519
7	survey_involved_family_num	0.396385	0.396385
8	Growth_SAT_Grade_11_School_Pct	-0.398949	0.398949
9	Chronic_Truancy_Pct	-0.463141	0.463141
10	Growth_PSAT_Grade_9_School_Pct	0.469752	0.469752
11	Misconducts_To_Suspensions_Year_2_Pct	0.506284	0.506284
12	Student_Attendance_Year_2_Pct	0.590448	0.590448
13	Attainment_SAT_Grade_11_School_Pct	0.614290	0.614290
14	Attainment_PSAT_Grade_10_School_Pct	0.686115	0.686115
15	One_Year_Dropout_Rate_Year_2_Pct	-0.689593	0.689593
16	survey_ambitious_inst_num	0.866173	0.866173
17	Attainment_All_Grades_School_Pct	0.868333	0.868333
18	Mobility_Rate_Pct	-0.927564	0.927564
19	Freshmen_On_Track_School_Pct_Year_2	0.965500	0.965500
20	Attainment_PSAT_Grade_9_School_Pct	1.000486	1.000486
21	Student_Attainment_Rating	1.020248	1.020248

From the Logistic Model:

- Student_Attainment_Rating, 9th Grade PSAT Score, Freshman being on track score, and mobility rate, and attainment ratings played heavily into graduation rate.
- Interestingly (and somewhat surprisingly), student growth rating, teacher attendance, and a survey of effective leaders within schools did not.
- These findings seem somewhat antithetical to what would be expected, so more analysis is needed into each of these features.

Next Steps:

Drop more features, incorporate feedback, Summarize findings on in writing

Questions/Feedback?

Image Sources: Unsplash.com
Slides Design: Slidesgo.com