Predicting Future School Chronic Absenteeism Rates

Capstone 3

Joseph Frasca

the problem

Absenteeism is a huge indicator of student success and also impacts school funding. If students are not in school they cannot learn or graduate. Once students and schools have high chronic absenteeism rates, it is a very difficult and costly problem to remediate.

Being able to predict a school's attendance/chronic absenteeism numbers and identify the contributing variables would allow district administrators to proactively intervene with these school's ahead of time. Saving time, money, and resources, as well as increasing student/school outcomes. We used NYC OpenData's school-level attendance data from 2013-2019. In New York State Chronic Absenteeism is defined as when a student is absent 18 or more days of the school year.

the approach

1. data wrangling

Data Definition

After loading the NYC attendance data I explored the data, each feature's unique values and checked for NaNs. There were no NaNs in the data, however later I found *'s'* objects in numeric data. I replaced these *'s'* objects with NaNs in attendance data and sent an email to NYC OpenData to find out more about what they may about. Then I changed the data types to float for any numeric columns.

I created a pandas profiling report to define/explore the data further. It seemed the variables #Total Days/# Days Absent/# Days Present/# Contributing 20+ Days/# Total Days/# Chronically Absent had large outliers, and were highly skewed. There seemed to be several independent variables that were correlated with each other and would need to be dealt with (ie. # Days Present/# Contributing 20+ Days/# Total Days). Also From the Correlation Matrix it seemed that % Chronically Absent was highly negatively correlated with % Attendance. Which makes sense - schools that have low attendance numbers in general would have high chronically absent numbers. For the target variable '% Chronically Absent' it seems most values fall in 13.5%-34.9% range

Data Cleaning

To deal with the NaNs I first explored them further. I found all 6 columns with NaNs had NaNs in the same rows.. When looking at the categorical variables, compared to the the full dataset it seemed that the data with NaNs:

- Column 'Year' - most of the NaNs are evenly spread by year, with slightly more in 2013-2014, & 2014-2016 school years

- Column 'Grade' - most of NaNs are in 'PK in K-12 Schools', middle schools, grades - 6, 9, 7, 11, 8
- 'Demographic Category' - Most of the NaNs are from Ethnicity, ELL Status (english language learners), Poverty, and SWD (students with disabilities) Status
- 'Demographic Variable' - Most of the Nans are from: Other, White, Asian, Black, ELL, Not ELL, Poverty, Not Poverty

Then I compared data with nulls vs data with no null values to try and pick up trends to learn more about the NaNs and found:

- Column 'School Name' - There are 74 schools with the same or more amount of NaNs as Non Nans
    - Highest include PS 022, Satellite Three
    - Some of the schools that I checked with the highest ratio of null to non null counts are currently closed. But not all of the schools with a lower ratio are closed
- Column 'Year' - most of the NaNs are evenly spread by year, with slightly more in 2013-2014, & 2014-2016 school years
- 'Grade' - most of NaNs are in 'PK in K-12 Schools'
- 'Demographic Category' - Most of the NaNs are from Ethnicity, ELL Status, Poverty,
- 'Demographic Variable' - There are more NaNs than non Nans for: Other, White, Asian

To gain further understanding I looked closer at school name NaN counts, and looked at the number of unique data with NaNs and compared this with the number of unique data without NaNs and found none of the unique categorical variables are missing due to NaNs. It seems that the two options for dealing with NaNs were 1) to either drop the NaNs entirely, but this would leave out a lot of data regarding Demographic Category, specifically Ethnicity, as well as certain schools. 2) Another idea would be to try and replace the NaNs based on the mean score of that variable from the certain School: certain Year: certain Demographic Category and Variable. **An email was sent to NYC OpenData on 12/14/20 to inquire about the non-numeric data (remember 's' was in each of these variables before I replaced it with NaNs). The response from NYC Open Data was such, "When data counts are low so as to potentially result in the identification of students, those numbers are redacted. the 's' indicates a low number that we have redacted." Due to this I decided to drop the NaN data from the data set.**

Feature Creation

I then created the variable 'Next Year % Chronically Absent', by adding the next year's data to the current year's. This will be our target variable in order to predict what the next year's chronically absent number may be. We will be able to do this by school, by grade, demographic category, and demographic variable. Finally I saved the wrangled data and made a few notes on things to explore further.

2. Exploratory Data Analysis

I loaded the wrangled NYC attendance data and built data profiles & tables using pandas profiling to explore the data relationships. Things I found were:

Variables with High Correlations:

- #Days Present was highly correlated with # Total Days and # Contributing 20+ Total Days
- #Chronically Absent was highly correlated with # Days Absent
- Demographic Variable was highly correlated with Demographic Category

Noticings from the Data:

- There were 4 more unique DBNs (district borough numbers) than school names.
- Some school names have more frequency than others. PS 212 and PS 253 make up 0.2% of the data, double any other school.
- There was more elementary school data than middle school data (by about 2% points)
- Ethnicity makes up the largest Demographic Category at 33.3%, double the size of the next largest
- Demographic Variable is more evenly split in each feature size than Demographic Category
- For the target variable '% Chronically Absent', most values fall in 13.5%-34.9%

Outliers

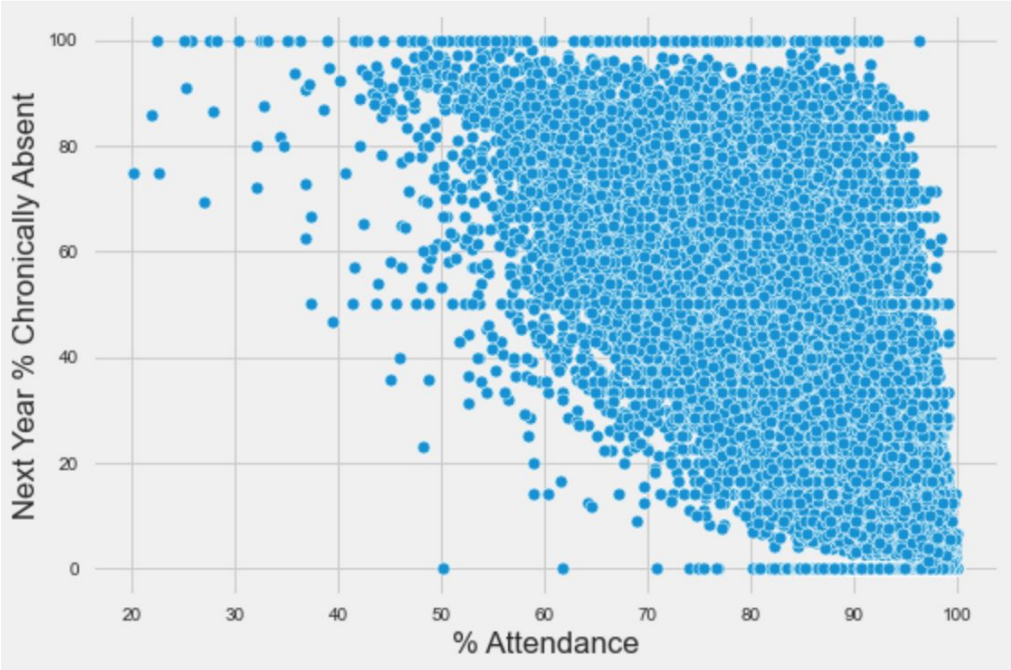
- It seems #Total Days/# Days Absent/# Days Present/# Contributing 20+ Days/# Total Days/# Chronically Absent have very large outliers, and are highly skewed.
- After deciding which of these variables we will retain (some are highly correlated). I may consider dropping outliers outside the 5th/95th percentile or 3 std above or below. I could check to see which schools they are from before dropping, if the outliers make up a large % of a few schools maybe consider inputting the 3x std./95th percentile to retain that school.

Interactions - Numeric Values

- % Chronically Absent seems to have parabolic interactions with # Total Days (peaks at 20% Chronically Absent/750k days)
- % Chronically Absent seems to have parabolic interactions with # Days Absent (peaks at 40% Chronically Absent/80k days)
- % Chronically Absent also seems to have parabolic interactions with # Days Present, # Contributing to 20+ Days, and # Chronically Absent
- #Days Absent and # Chronically Absent seem to have normal distributions in their interactions with % Chronically Absent (this seems to make sense)
- % Chronically Absent drops from 2013-14 by 2% and stays low for 2014-15 and 2015-2016 but then starts rising again until 2017-2018 and starts to level off again in 2018-2019. I wonder if this is a statistical anomaly?

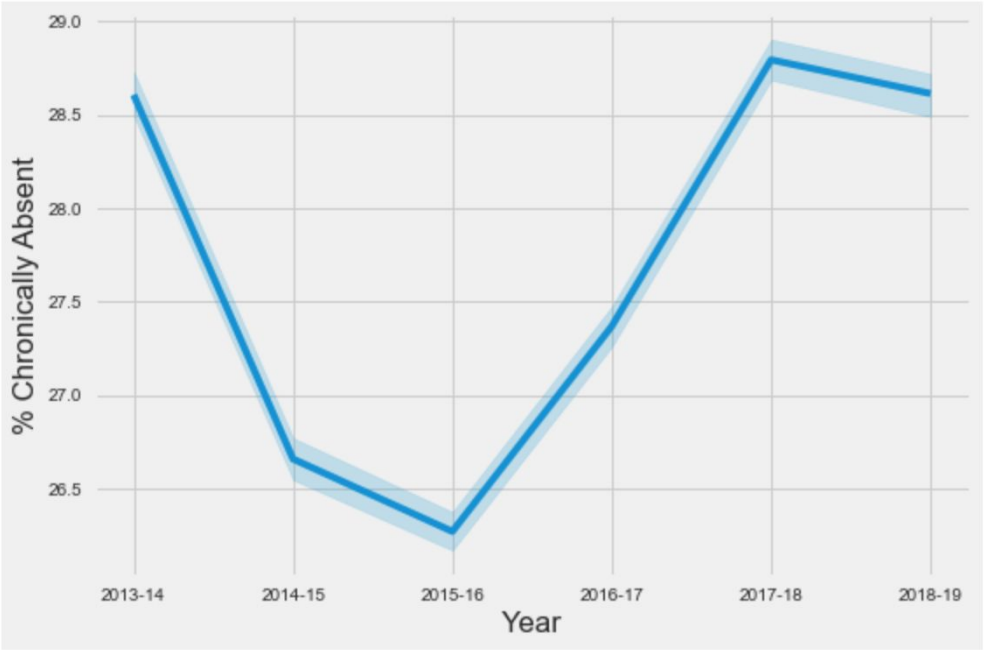
I made scatterplot for all variable interactions with target variable (‘Next Year % Chronically Absent’), below you can see ‘% Attendance’’s strong negative correlation with the target variable:

Negative Correlation between '% Attendance' & Target Variable



I then made a pairplot of each numeric variable with all other numeric variables to look at their interactions as well as a line plot of year vs. % Chronically absent.

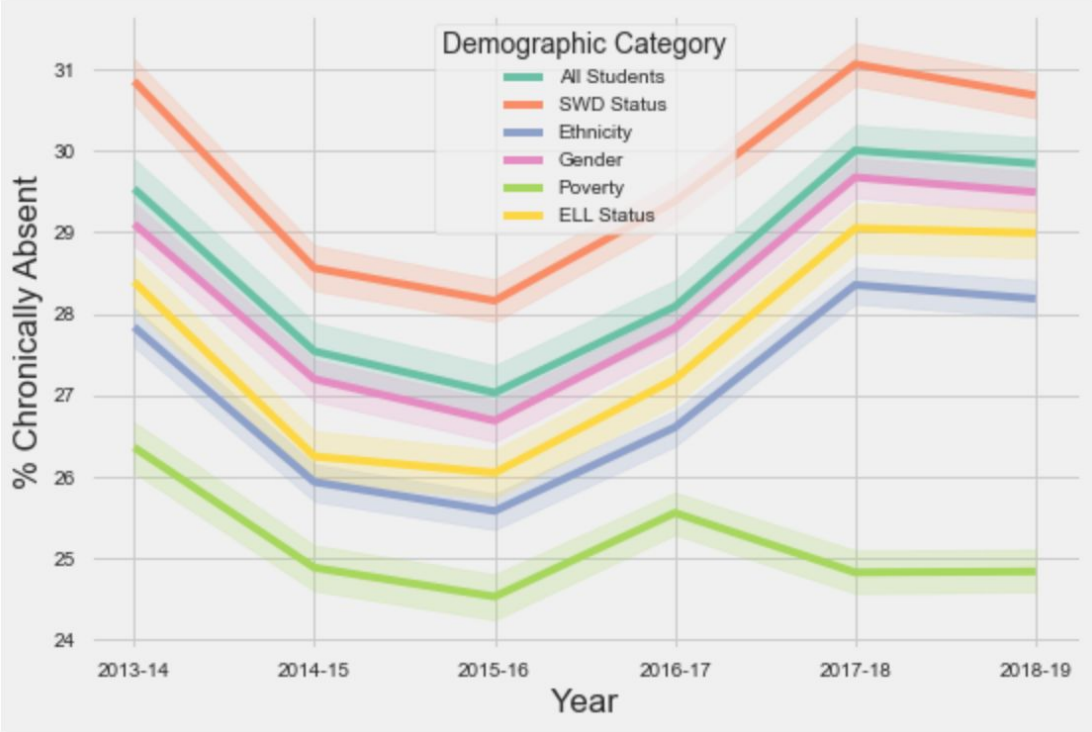
% Chronically Absent Over Time



Interactions - Categorical Variables

- The distributions of % Chronically Absent are similar across years
- There are some interesting bumps as distribution of % Chronically Absent comes to 0% across Demographic Categories. There is an increase in this area for Ethnicity and Poverty most noticeably.

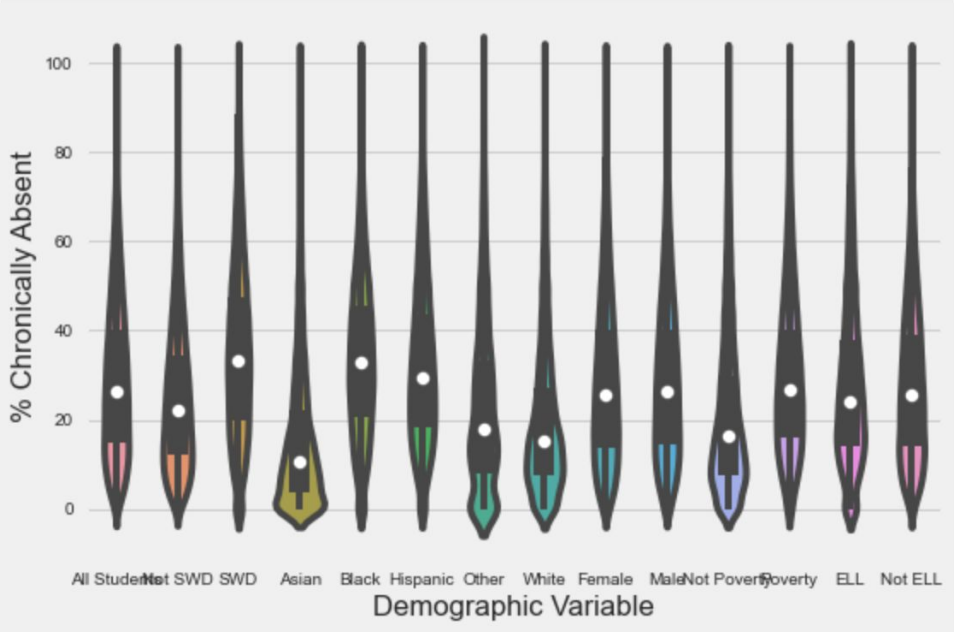
% Chronically Absent Over Time, by Demographic Category



I then made a violin plot for each of the categorical variables vs. % chronically absent. Demographic Variables and % Chronically Absent have a lot of variation.

- With Students with disabilities (SWD), Black and Hispanic students having the highest % Chronically Absent. While, Asian, and White students have the lowest % Chronically Absent.
- Not SWD has much lower % Chronically Absent than SWD, Not Poverty has much lower % Chronically Absent than Poverty.
- Male/Female have similar rates of % Chronically Absent (Females have slightly less). And ELL has very slightly lower % Chronically Absent than Not ELL
    - **These may be an interesting relationships to look further into, with the goal of closing the gaps more in the other groups**

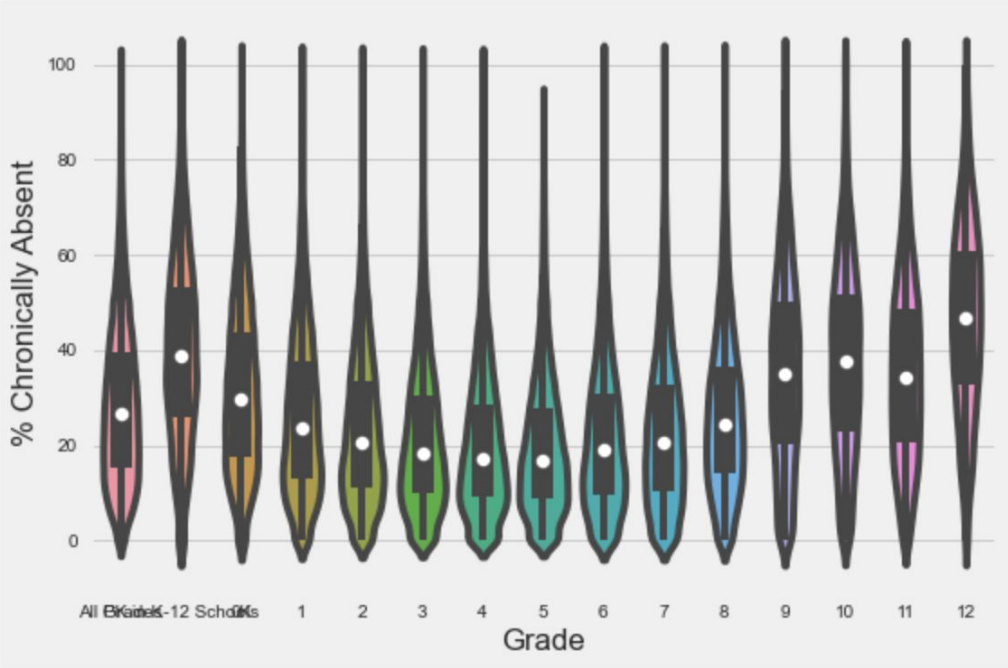
% Chronically Absent by Demographic Variable



Grade and % Chronically Absent also had a lot of variation:

- There is a parabolic relationship with the highest rates of % Chronically Absent in PreK/K and 9-12 grade (high school). 12th grade having the highest of around 45-50% chronically absent.
- The lowest % Chronically Absent can be seen from 2-7th grade (around 20% or less)

% Chronically Absent by Grade



Correlations:

- % Chronically Absent has negative correlations with # Total Days, # Days Present, # Contributing 20+ Total Days
- % Chronically Absent has positive correlations with # Days Absent, # Chronically Absent
- % Chronically Absent is strongly negatively correlated with % Attendance and strongly positively correlated with Next Year % Chronically Absent
- *There are a lot of high correlations within the current variables, it will be interesting to look at this again once it has been decided which features to drop*

From the Correlation it seems That % Chronically Absent is highly negatively correlated with % Attendance. Which makes sense. Schools that have low attendance numbers in general would have high chronically absent numbers. Using this information the following hypothesis was constructed:

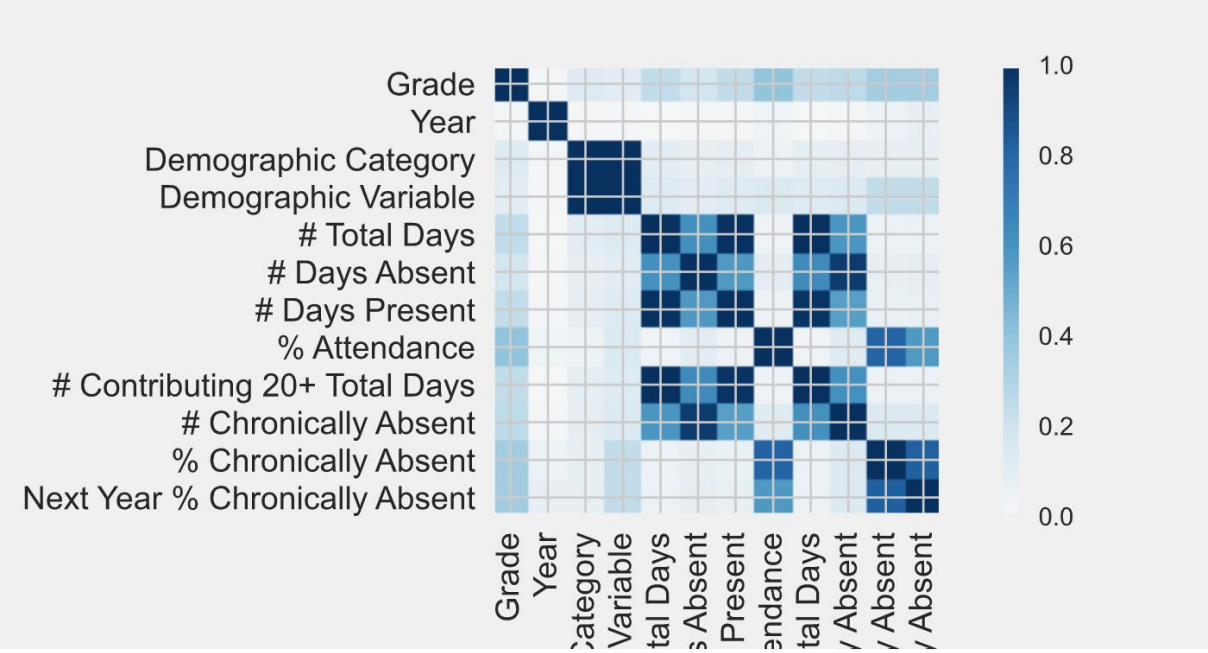
Null Hypothesis: Next Year % Chronically Absent IS NOT correlated with % Attendance and % Chronically Absent and these 2 variables can therefore not be used to model future % Chronically Absent Rates.

Alternative hypothesis: Next Year % Chronically Absent IS correlated with % Attendance and % Chronically Absent and these 2 variables can therefore be used to model future % Chronically Absent Rates.

This work can be summarized in the following Correlation Heatmap (Cramer's V (φ_c)). Where you can clearly see the high correlation between some of the categorical variables that I dealt with in preprocessing.

Cramér's V (sometimes referred to as **Cramér's phi** and denoted as φ_c)

Correlation Heatmap of Variables



Feature Selection & Engineering

In notebook 1.0 Data Wrangling the data was aggregated by year and the feature 'Next Year % Chronically Absent' was created before merging all years back together. Above see further discussion of thinking behind future feature

selection and engineering. As noted for features with high correlations, I needed to decide which features to drop.

---

3. Preprocessing & Training

After loading the wrangled data I dropped highly correlated variables that had been identified in EDA and retained the variables that made the model more generalizable (the retained features more schools would have data on):

- dropped # Total Days and #Contributing 20+ Total Days (both highly correlated with # Days Present)
- dropped #Chronically Absent (is highly correlated with # Days Absent)
- Dropped Demographic Category as Demographic Variable contains all this information and they were highly correlated

Then I defined the mean and std deviation of target variable ('Next Year % Chronically Absent') so that we could turn it into a categorical variable:

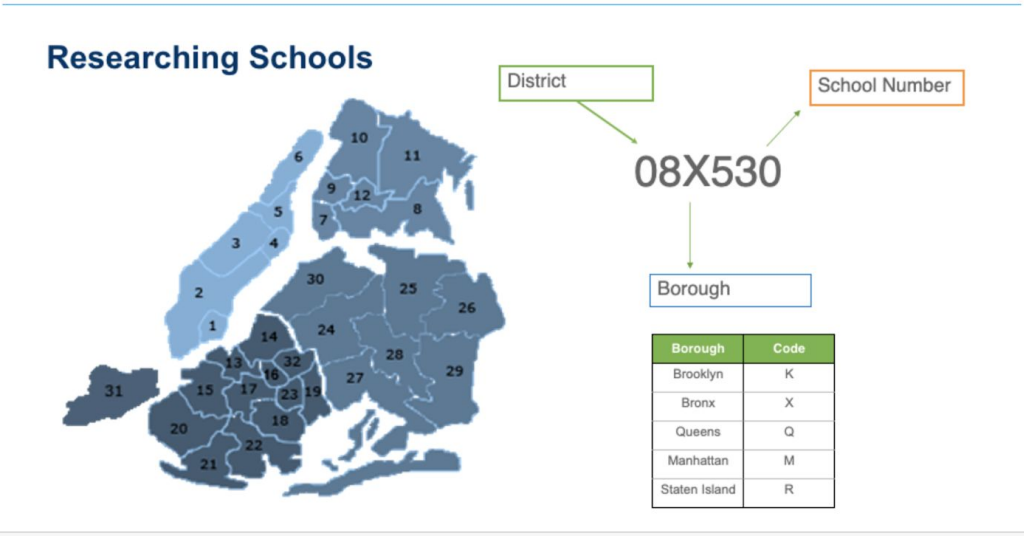
- 'High' = 1 std deviations above mean
- 'Med' = +/- 1 std from mean
- 'Low' = 1 std dev below mean

During this feature creation I noted that there were NaNs for 'Next Year % Chronically Absent' in Years pre 2018-19 (NaNs for this variable in 2018-19 were expected because we don't have data for 2019-20 yet - we are predicting it). Digging further it seems that the NaNs in 'Next Year % Chronically' Absent aren't just in 2018-19 as expected. **There are around 6% of data each year that does not have next year's data. Some of the schools I looked into were closed or had name changes. This could possibly be due to inconsistent naming of the schools year to year led to a calculation error for Next Year % Chronically Absent or these schools did not have data for that subset of students the following year. I will have to dig further into why this may be occuring in the future. This makes up around 7.8% of data.**

Feature Creation/Selection:

There were several categorical variables - over 1600 unique schools, 15 grades, 6 years and 14 demographic variables). **To reduce the number of variables in DBN (District Borough Number) and School Name, I split DBN into its 3 components, see explanation of DBN here. They are: District Number, Borough Code, and School Number. Then I retained district number and borough code while dropping school number, and borough code.** This will make our dummy categorical variables more manageable and our model more generalizable to a school's location.

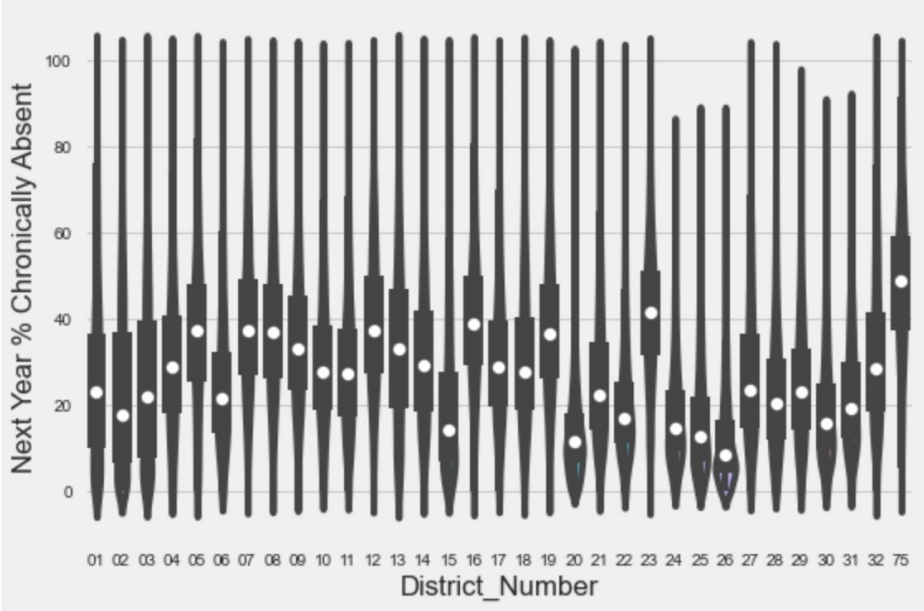
Map of NYC Districts & Breakdown of District Borough Number



Quick EDA of new variables

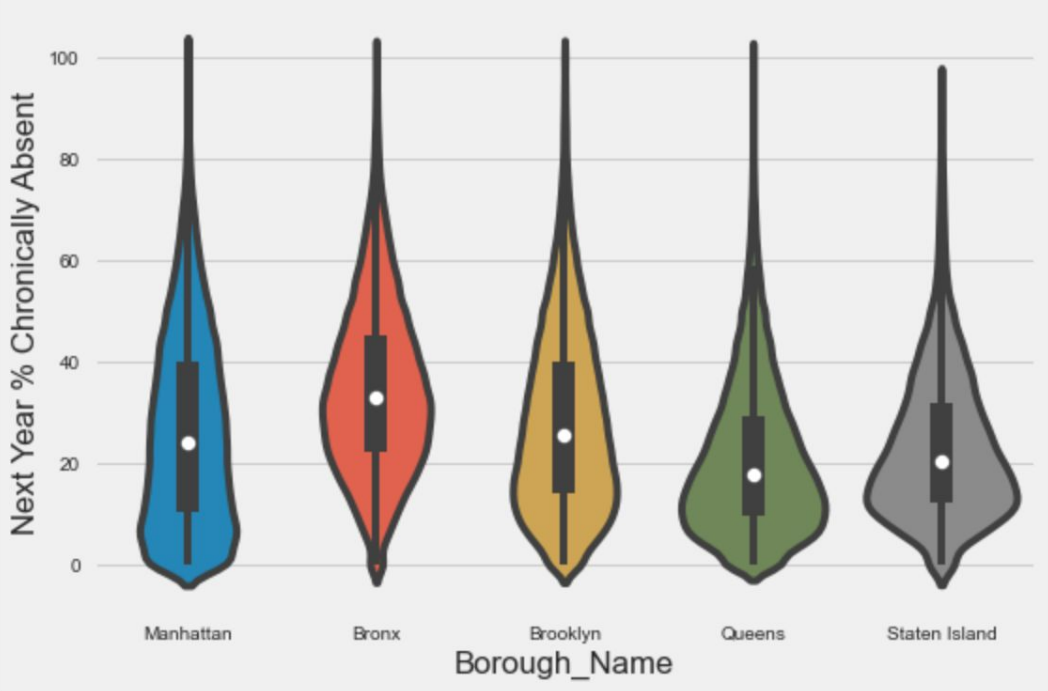
I made a scatterplot for all variable's interactions with target variables and built pandas profiling for a quick EDA. I also made violin plots of the new categorical features vs the target variable. I found there was a lot of variation by district:

Next Year % Chronically Absent by District



Zooming out you could see the variation across boroughs, Queens and Staten Island have the lowest rates of chronically absent students, followed by Manhattan and then Brooklyn (these have larger ranges than the others), while Bronx has the highest, and seems to be more normally distributed than the others:

Next Year % Chronically Absent by Borough



As noted I dropped DBN, school name, and school number in anticipation of dummy variable creation. I dropped Borough Code/Name because it is highly correlated with District Number, and District Number has more info and district Number also has more correlation with target variable. I also dropped the columns (# days absent, # days present) because I realized the "% attendance' column has this info and we don't have these variables in current daily data from NYC OpenData. I separated 2018-2019 data from 2013-18 school years for later predicting, did a final check of NaNs in the resulting dataframe (2013-2018), for now, dropped these NaNs until further research is conducted (see note of this in second paragraph of this section).

I created separate dataframes for regression (where the target variable is 0-100%) vs. classification (where the target variable is 'High', 'Medium', or 'Low') and created dummy variables for the two data frames.

I then split into testing and training datasets using Time series Train test split for the 5 school years and elected not to standardize the magnitude of numeric features using a scaler because the numeric variables only go 0-100 and the dummy variables are simply 1s and 0s.

Initial Modeling

I took an initial benchmark using a Dummy Regressor using the mean of y_train and found:

- MAEs: train = 14.425187661095018 vs. test = 14.407710596424323
- MSEs: train = 313.6617613921039 vs. test = 312.10496683625587

Then I created a simple Linear Regression Model and assessed model performance:

- R2 - train = 0.70, test = 0.67
- MAE - train = 7.03, test = 7.37
- MSE - train = 93.04 , test = 101.65

I also tried some initial classification models before saving the processed data:

Evaluation Metrics - Decision Tree Model

```
print(dtree_model.score(X_test, y_test))
print(cm)
```

```
0.7752944130903131
[[ 6442    29  6698]
 [   23  7650  4478]
 [ 2087  3686 44566]]
```

```
print(classification_report(y_test, dtree_predictions))
```

	precision	recall	f1-score	support
High	0.75	0.49	0.59	13169
Low	0.67	0.63	0.65	12151
Medium	0.80	0.89	0.84	50339
accuracy			0.78	75659
macro avg	0.74	0.67	0.69	75659
weighted avg	0.77	0.78	0.77	75659

Evaluation Metrics - K Nearest Neighbors Model

```
Accuracy: 0.7748185939544535
```

```
# Generate the confusion matrix and classification report
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[ 8049    33  5087]
 [   47  7204  4900]
 [ 3674  3296 43369]]
```

	precision	recall	f1-score	support
High	0.68	0.61	0.65	13169
Low	0.68	0.59	0.64	12151
Medium	0.81	0.86	0.84	50339
accuracy			0.77	75659
macro avg	0.73	0.69	0.71	75659
weighted avg	0.77	0.77	0.77	75659

Additional preprocessing (notebook 1.2.1)

After initial modeling and using '% Chronically Absent' as a baseline comparison for model performance (see section 4), it became clear that the model was not performing that well, so I **added 5 additional time series features**:

- **the difference between each value's '% Chronically Absent' and the 5 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)**
- **the difference between each value's '% Attendance' and the 5 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)**
- **the difference between each value's '% Chronically Absent' and the 2 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)**
- **the difference between each value's '% Attendance' and the 2 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)**
- **% Chronic Absent (which is essentially the previous year's % Chronic Absent as we are predicting the following year.**

As both '% Chronically Absent' and '% Attendance' are variables that had high feature importance (see notebook 1.3.1) and we wanted to build in additional features that may help with a model's predictive power. The addition of the variables led to around 0.6% improvement in accuracy scores across initial models.

4. Modeling

As noted above in additional preprocessing, after using **Last Year's % Chronically Absent to Predict Next Year's as a baseline comparison,** I found that this one variable led to an R2 of .64 for linear regression and an accuracy of .77 with a Decision Tree. **Meaning the previous models were not outperforming this baseline. So we tried adding 5 additional features** (see preprocessing above for more information).

- the difference between each value's '% Chronically Absent' and the 5 year averages of '% Chronically

Absent' (grouped by DBN, Grade, & Demographic Variable)
- the difference between each value's '% Attendance' and the 5 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)
- the difference between each value's '% Chronically Absent' and the 2 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)
- the difference between each value's '% Attendance' and the 2 year averages of '% Chronically Absent' (grouped by DBN, Grade, & Demographic Variable)
- % Chronically Absent (which is essentially the previous year's % Chronically Absent as we are predicting the following year.

Also the column 'Year' was dropped, since the model was not predicting future years due to our choice of 'Year' as a categorical value. Due to this, instead of a TimeSeriesSplit I used a general Train/Test split. This led to increased model performance, especially for the regression models.

After loading the processed data for regression and classification models I checked to ensure District Number was correctly an object, not integer. **This time (as noted above) I decided a more appropriate baseline comparison would be to use only Last Year's % Chronically Absent to Predict Next Year's to get a sense of how the models compared to this.** A linear regression with this supplied:

- R2 - test = 66.2
- MAE - test = 7.5
- MSE - test = 105.5

It seems that if one only used last year's % Chronically Absent you would predict Next Year % Chronically Absent correctly 66.2% of the time. So any future models would need to improve on that.

Regression Models:

<u>Linear Regression Model:</u>

- R2 - train = 0.82, test = 0.82
- MAE - train = 5.36, , test = 5.36
- MSE - train = 57.04, test = 57.00

<u>Random Forest Regression Model:</u>

- R2 - train = 0.98, test = 0.82
- MAE - train = 1.91, , test = 5.14
- MSE - train = 7.74, , test = 54.97

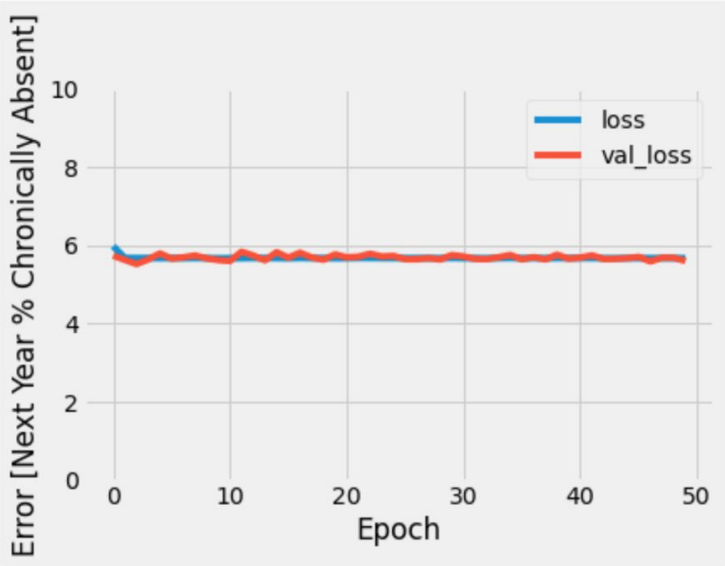
The Random Forest Regression seems to be significantly overfitting. Potentially due the amount of dummy variables?

<u>Tensorflow Deep Learning Regression Model</u>

I defined variables for a tensorflow regressor model and did a 80/20 train test split. After splitting the features from labels I

built a normalization layer and adapted it to the data. First I tried a Tensorflow Linear regression (Before building a DNN model). I built the sequential model using the Normalization layer that was adapted to the whole dataset. Then configured the training procedure and defined what would be optimized (mean_absolute_error) and how (using the optimizers.Adam). I executed the training and visualized the model's training progress - training and validation error (note this was tested up to 50 epochs and showed similar results).

Tensorflow Linear Regression - Training and Validation Error



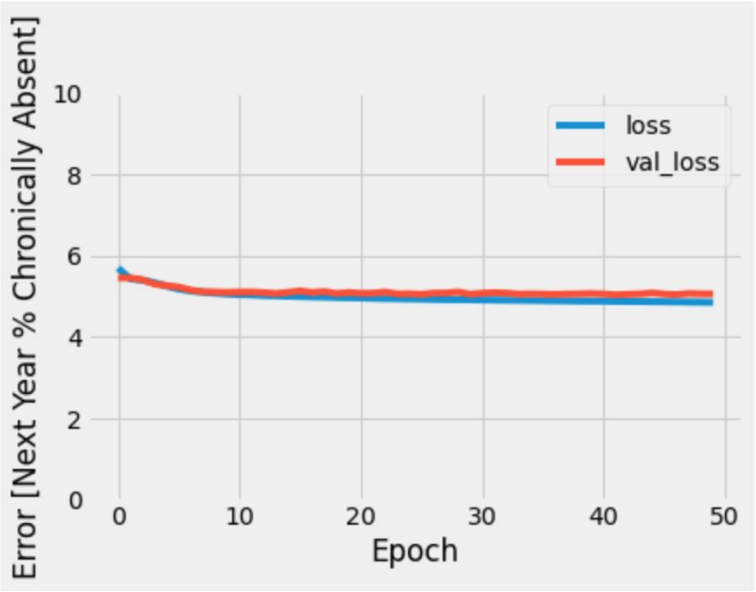
Tensorflow DNN Regression

Again I defined a function to build and compile the model, and built the sequential model with the following dense layers:

- Dense(136, activation='relu'),
- Dense(68, activation='relu')
- Dense(1)

Then checked the test-set performance (note this was tested up to 50 epoch)

Tensorflow DNN Regression - Training and Validation Error



I also looked at the errors made by the model when making predictions on the test set and also looked at the error distribution. The DNN Regression model had an R2 score of .82, MSE of 55.66, and a MAE of 5.05, making it's performance similar to the other regression models. With actually slightly better performance of MAE.

Model Performance Comparison - Regression Models

Model	R2	MSE	MAE
Baseline - Linear Regression (X = '% Chronically Absent', using last year's % Chronically Absent to predict next year's)	0.662	105.51	7.51
Linear Regression	0.818	57.00	5.36
Random Forest ***Final Model**	0.824	54.97	5.14
DNN Regression	0.822	55.66	5.05

Classification Models

Again I decided a more appropriate baseline comparison would be to use only Last Year's % Chronically Absent to Predict Next Year's to get a sense of how the models compared to this. A decision tree (max_depth =2) with this supplied an accuracy of .77. **This suggests that a classification model would have to outperform 77.4% to be considered an improvement over the baseline.** I then defined variables for classifiers and performed an 80/20 train test split then tested several models.

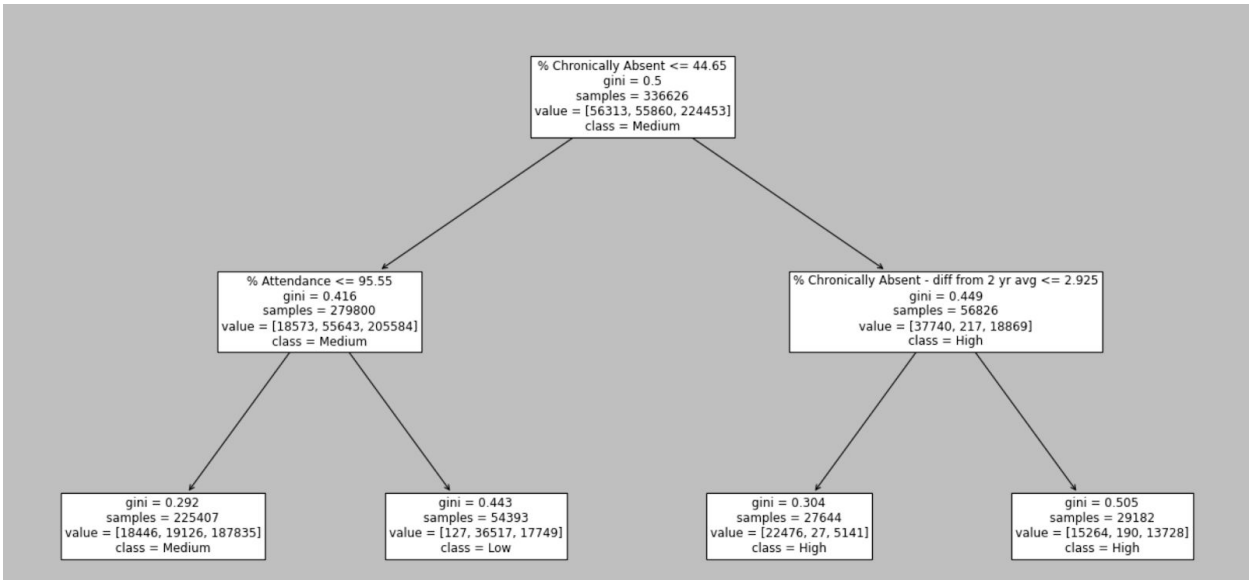
Model Performance Comparison - Classification Models

Model	Accuracy	Confusion Matrix	Classification Report
Baseline - Decision Tree (X = '% Chronically Absent', using last year's % Chronically Absent to predict next year's)	0.774	[[9326 73 9397] [33 12622 6048] [2967 6851 64892]]	precision recall f1-score support High 0.76 0.50 0.60 18796 Low 0.65 0.67 0.66 18703 Medium 0.81 0.87 0.84 74710 accuracy 0.77 112209 macro avg 0.74 0.68 0.70 112209 weighted avg 0.77 0.77 0.77 112209
Decision Tree	0.840	[[13295 8 5493] [17 12742 5944] [3470 3150 68090]]	precision recall f1-score support High 0.79 0.71 0.75 18796 Low 0.80 0.68 0.74 18703 Medium 0.86 0.91 0.88 74710 accuracy 0.84 112209 macro avg 0.82 0.77 0.79 112209 weighted avg 0.84 0.84 0.84 112209
Naive Bayes	0.516	[[15502 315 2979] [856 15871 1976] [20787 27356 26567]]	precision recall f1-score support High 0.42 0.82 0.55 18796 Low 0.36 0.85 0.51 18703 Medium 0.84 0.36 0.50 74710 accuracy 0.52 112209 macro avg 0.54 0.68 0.52 112209 weighted avg 0.69 0.52 0.51 112209
Random Forest Classifier	0.848	[[13701 7 5088] [15 13114 5574] [3366 2960 68384]]	precision recall f1-score support High 0.80 0.73 0.76 18796 Low 0.82 0.70 0.75 18703 Medium 0.87 0.92 0.89 74710 accuracy 0.85 112209 macro avg 0.83 0.78 0.80 112209 weighted avg 0.85 0.85 0.85 112209
KNN	0.838	[[13615 15 5166] [36 13279 5388] [3921 3664 67125]]	precision recall f1-score support High 0.77 0.72 0.75 18796 Low 0.78 0.71 0.74 18703 Medium 0.86 0.90 0.88 74710 accuracy 0.84 112209 macro avg 0.81 0.78 0.79 112209 weighted avg 0.84 0.84 0.84 112209

I also plotted an untuned Decision Tree (max_deph =2, accuracy: 0.778) to get a sense of how a very basic decision tree would make these choices:

Next Year's % Chronically Absent

Multi-Class Decision Tree



The plot shows that the model is simply using '%Attendance', '% Chronically Absent' & '% Chronically Absent - diff from 2 ye avg' to predict the 3 classes. Although simple it could have some informative value to schools, it simply shows that **if a school or student group within a school is Chronically Absent more than 44.65% of the time this year they will be in 'High' group of Chronically Absent next year. Also if they are at or below 44.65% Chronically Absent and above 95.55% attendance they will be in the 'Low' group next year and if their attendance is less than 95.55% then they would be in the 'Medium' group.**

So a school could use the above numbers to set goals in order to decrease Chronic Absenteeism next year, ie. all schools/student groups have % Chronically Absent above 44.65 - this would essentially move these students out of the 'High' group. Then the secondary goal would be all schools/student groups have above 95.55% attendance. This would move students from 'Medium' to 'Low' groups.

Tuning Hyperparameters

When doing hyperparameter tuning of this Decision Tree Classifier, there was not a significant gain in its performance while it's readability/explainability decreased. I tried tuning "max_depth", "max_features", "min_samples_leaf" all with randint(1, 9), and also "criterion". In the future we could try increasing the integers the model is tuned on (ie. 1-75, but we were looking for a model that was interpretable in this case). I found that the max_depth of 7 with a gini model yielded the best results.

Random Forest Classifier - It seemed that the tree depths that the Random Forest Classifier used were in the high 50s to low 60s. While trying to tune it to make it a bit more simpler (less depth), it significantly decreased performance (0.80 accuracy) in the model and significantly decreased recall (0.50% for 'High' class).

K Nearest Neighbors - the hyperparameter tuning on the KNN classifier ran for several days and did not yield any additional information. It seems KNN is slow with this many observations, since it does not generalize over data in advance, it scans historical databases each time a prediction is needed (from here).

Model Selection

I decided in the end to go with a regression model, as they showed greater improvement from the baseline comparison

compared to the classification models and regression models yielded more information, showing the actual percentage of the target variable. So I chose the **random forest regression model** as this had the best performing metrics for a regression model:

Model Metrics - `RandomForestRegressor()`

`Target Variable = 'Next Year % Chronically Absent'`

final model features	parameters	hyperparameters	performance metrics

• % Attendance
• % Chronically Absent
• '% Attendance - diff from 5 yr avg',
• '% Attendance - diff from 2 yr avg',
• '% Chronically Absent - diff from 5 yr avg',
• '% Chronically Absent - diff from 2 yr avg'
• Dummy Variables
  ○ Grade
  ○ Demographic Variable
  ○ District Number
	{'bootstrap': True,
	'ccp_alpha': 0.0,
	'criterion': 'mse',
	'max_depth': None,
	'max_features': 'auto',
	'max_leaf_nodes': None,
	'max_samples': None,
	'min_impurity_decrease': 0.0,
	'min_impurity_split': None,
	'min_samples_leaf': 1,
	'min_samples_split': 2,
	'min_weight_fraction_leaf': 0.0,
	'n_estimators': 100,
	'n_jobs': None,
	'oob_score': False,
	'random_state': None,
	'verbose': 0,
	'warm_start': False}
		{'n_estimators': 144,
		
		'min_samples_split': 2,
		
		'min_samples_leaf': 1,
		'max_features': 'sqrt',
		'max_depth': 100,
		'bootstrap': False}
			R2 (test set)
			0.824
			
			MAE (test set)
			5.14
			
			MSE (test set)
			54.97

This model had the following hyperparameters: {'n_estimators': 144, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 100, 'bootstrap': False}

This outperformed both the linear model and the tensorflow DNN Regression model. Although the DNN Regression model was similar in performance and actually slightly outperformed the random forest in it's MAE (5.05).

For **classification the best performing model in terms of overall accuracy was the Random Forest Classifier** with (84.8% Accuracy), however in many cases school districts would be most focused on the 'High' rates of Chronic Absenteeism (**Meaning higher than 1 standard deviation above the mean)**. For the 'High' class the Random forest classifier would be our first choice model as it had the highest f1-score for this class (76%) and for the reasons explained below.

However, it will depend on what a district needs to manage more, False Positives or False Negatives, in deciding which model is best in reality:

- too many False positives = may be ok for school districts with less constraints (ie. larger budgets, more time for this kind of initiative), as they may want to ensure they identified all the schools & groups that are at risk of high chronic absenteeism.
- too many False negatives = may be ok for school districts with more constraints (ie. smaller budgets, less time for this kind of initiative), as they may want to ensure a school/group they focus on will really need it.

So as noted above, the Random Forest Model would be a great choice if you want to ensure you limited False Positives (and were ok with under-identifying schools/groups that are at risk of high chronic absenteeism), because the precision is higher while the recall is lower:

'High' Next_Year_Chronic Absenteeism:

- precision: 0.80
- recall: 0.73
- f1-score: 0.76
- support: 18796

However if you wanted to ensure you limited False Negatives (and were ok with over-identifying schools/groups that are at risk of high chronic absenteeism), then the Naive Bayes Classifier (51.6% Accuracy) would be the way to go, because although the precision is low the recall is higher than other models:

'High' Next_Year_Chronic Absenteeism:

- precision: 0.42
- recall: 0.82
- f1-score: 0.55
- support: 18796

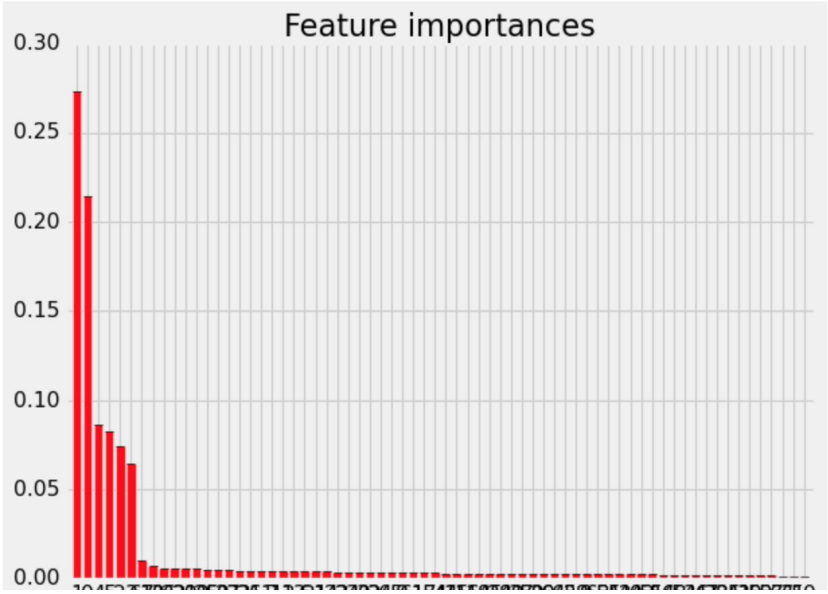
Feature Importance

As seen below, the regression model seems to be relying heavily on % Attendance and the other 4 columns with numeric values. But the classification models seem to have a slightly more even spread of feature importance across the 5 columns with numeric values.

Classification Model - Top 10 features

1. feature 1 (0.273007) - '% Chronically Absent'
2. feature 0 (0.214223) - % Attendance'
3. feature 4 (0.085994) - '% Chronically Absent - diff from 5 yr avg'
4. feature 5 (0.082235) - '% Chronically Absent - diff from 2 yr avg'
5. feature 2 (0.073907) - '% Attendance - diff from 5 yr avg'
6. feature 3 (0.064785) - '% Attendance - diff from 2 yr avg'
7. feature 67 (0.009893) - 'District_Number_75'
8. feature 10 (0.006912) - 'Grade_12'
9. feature 36 (0.005954) - 'District_Number_02'
10. feature 22 (0.005689) - 'Demographic Variable_Asian'

Random Forest Classifier - Feature Importance



Regression Model

Random Forest Regressor - Feature Importance

the findings

5. Model Predictions

After loading data, and loading both the saved regression and classification models, I refit both the Regression & Classification Models on All Available Data. Then I predicted next year's vacancy rate (2019-20) for school year 2018-19 (both with classification and regression models). I created new columns of predicted values and got labels for school name & DBN (categorical columns that were dropped prior for machine learning) back on the dataframe.

Projected 2019-20 Chronic Absenteeism Data

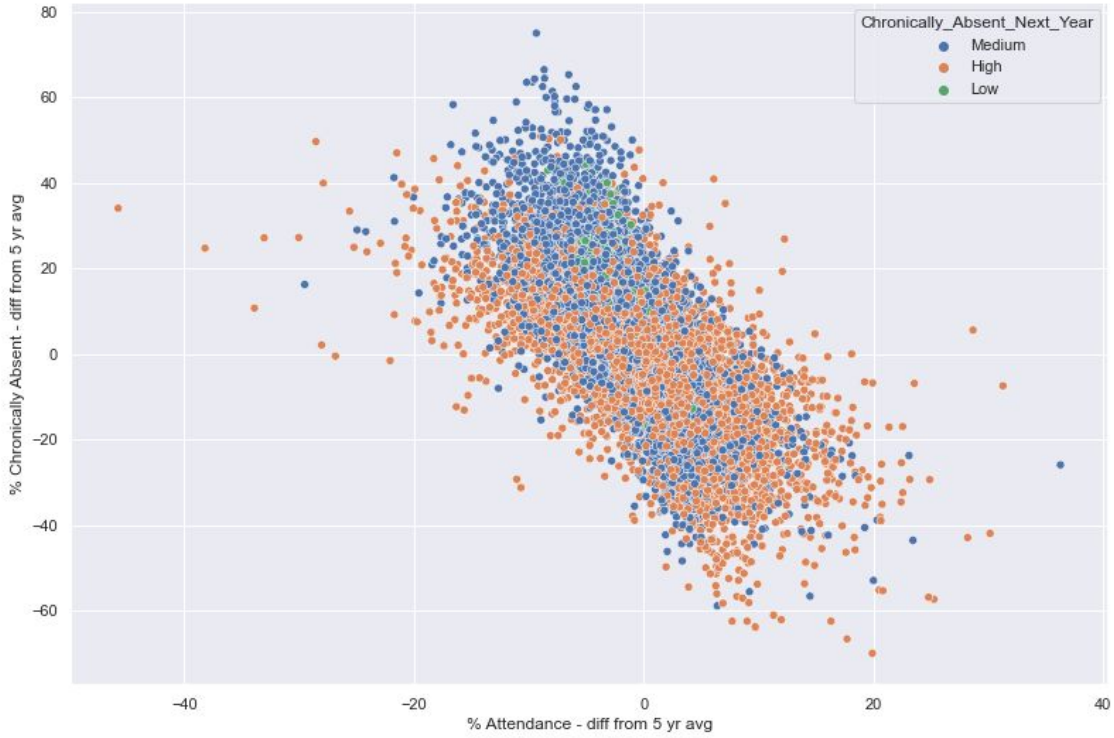
After looking at violin plots of Borough_Name vs. Next Year % Chronically Absent by class (High, Medium, Low) it seems like most of the variation is in the borough's 'Medium' class. This may be a good area to target in interventions. I did the same for Demographic Variable, District, and Grade and it again seemed like most of the variation is in the 'Medium' class.

Looking at violin plots of Demographic Variable vs. Next Year % Chronically Absent it seemed like 'Asian', 'White', 'Other'

(Race), and 'Not Poverty' students are significantly below the average. While SWD (Students with Disabilities), Black and Hispanic students seem to be significantly above the average. These findings were similar to those witnessed in earlier data in EDA, as were findings with ‘Grade’ (see visuals above in EDA section).

While looking at a scatterplot of '% Attendance - diff from 5 yr avg' vs. '% Chronically Absent - diff from 5 yr avg', colored by 'Chronically_Absent_Next_Year and scatterplot '% Attendance - diff from 2 yr avg' vs. '% Chronically Absent - diff from 2 yr avg', colored by 'Chronically_Absent_Next_Year' - these plots show there is a lot of year to year variation in the 'High' group in how they differ from their own 2 & 5 year averages. Seeming to show this group would potentially benefit from a concerted intervention.

'% Attendance - diff from 5 yr avg' vs. '% Chronically Absent - diff from 5 yr avg'



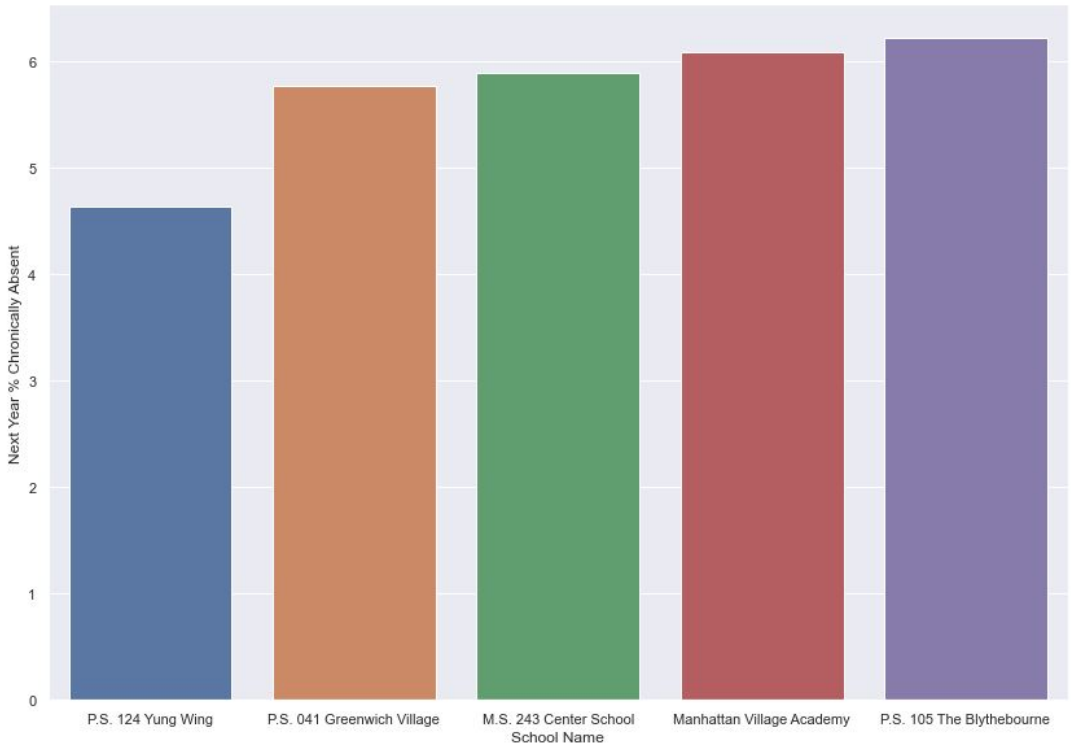
Next I look for schools/districts that have low chronic absenteeism predicted for 2019-20 school year in **variables with the highest rates of chronic absenteeism, including**:

- **grades 9-12**
- **PreK & K**
- **SWD**
- **Black, Hispanic**

These are schools that are doing well in the variables with the highest rates of chronic absenteeism. They could be studied for best practices to share with others. I made bar charts of the top performing schools, districts, and boroughs for the above populations (lowest Next Year % Chronically Absent) and then also made bar charts of the worst performing schools, districts, and boroughs for the above populations (highest Next Year % Chronically Absent).

Top Performing Schools for SWD (Students With Disabilities)

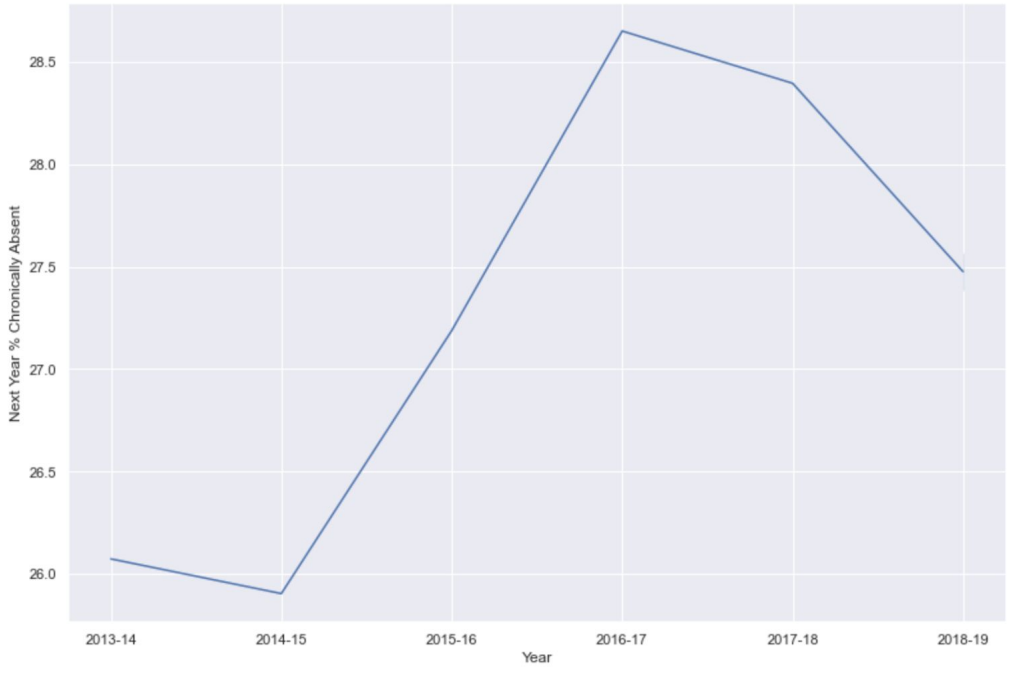
Lowest Rates of Next Year % Chronically Absent



Overall it seems like Queens/Staten Island have the lowest rates of Chronic Absenteeism for these at risk groups. Generally Manhattan is 3rd followed by Brooklyn/Bronx. Interestingly Manhattan has the highest rates of Chronic Absenteeism for black students. Possibly due to district 75? **Also the gaps between worst and top performing schools and districts are very large.**

I then looked at the variable ‘Next Year % Chronically Absent’ Across Time and made lineplots/data frames of 'Next Year % Chronically Absent' vs. Year. And also looked at this data by Grade, Demographic Variable, Borough Name, district number.

Next Year % Chronically Absent Over Time



Trends in 2019-20 seem not to be comparable with what was witnessed during covid-19 (this is understandable as it was a blackswan event). The model predicts Chronic Absenteeism to go down by almost 1%.but it most certainly went up during the pandemic. There were a few districts, demographic variables, and grades that were predicted to rise though. Also the range from top to bottom performers seems somewhat consistent over time, meaning that disparities between groups continue and may have not improved. **To test model performance in the real world it will be helpful to test 2019-20 predicting 2020-21 and see if the model data is accurate compared to the real data as it comes out.**

Further research

There are several ways to go for further research:

- I could build a model with less features (and only include the most important), as many of the features don't seem to contribute much.
- I could add more features from parent surveys and academic data, to make this data more actionable. This could also help pull apart why the exemplar schools (for high-risk categories) are doing so well.
- We could try to turn chronic absenteeism from a multi class categorical variable to a binary variable. As we are concerned mostly with whether Chronic Absenteeism is 'High' or not.
- We could test more models for regression/classification with more hyperparameter tuning.
- Also it would be good to test the model's predictions as real data comes out
- Finally it would be interesting to share the findings with NYC DOE (Department of Education)

Client recommendations

1. **Create measurable goals** and regularly & publicly report on each school. They can draw on the splits from the decision tree to determine goals (ie. all demographic groups/grades have below 44.65% chronic absenteeism, all groups have above 95.55% attendance)
2. **Look at the success schools** who have the lowest rates of chronic absenteeism for the high risk populations (groups that have the highest rates of chronic absenteeism) and **study why they are successful. Then share these best practices with other schools.**
3. Use the data to target and focus support on:
    - Target At Risk populations (Grades 9-12, PreK & K, Black and Hispanic Students, and Students with Disabilities)
    - Target Lowest Performing Districts/Boroughs/Schools and focus support on them
    - Target 'Medium' Chronic Absenteeism groups and move them to 'Low' Chronic Absenteeism. Since 'Medium' had the most variance there may be more of an opportunity to do this quickly, building momentum and experiencing success to then tackle the 'High' Chronic Absenteeism groups.