**Introduction**

Good morning Mike and fellow data science students, my name is Ayan Karim and today I’ll be talking about my capstone project on predicting student attendance and analyzing the factors that influence attendance at New York Public Schools.

**Model Parameters**

The way I built up models was by applying multiple baseline regression and classification models, I chose the top two models for each and then I optimized each of those models and compared them on their predictive capability and explanatory power. We’ll look at our regression model first, since our classification model builds on it.

**Interpretations of Features in Regression Models**

\_OLS Regression Model\_

The most important predictors are Economic Need Index, percent of students chronically absent, student demographic information, school quality ratings such as "Rigorous Instruction %" and finally a collection of select cities. From this information I know that demographic of the school plays a role in the attendance rate. Minority communities may have very different school participation and support systems than majority communities. It’s also reasonable to say that schools with higher need may have different attendance rates than schools with lower needs. School quality ratings indicate that support systems and educational quality of the school plays a role in student attendance. Select cities tend to have more of an impact on student attendance than others, indicating that these regions deserve another look into why they're so special. Perhaps they're underprivileged neighborhoods or extremely rich neighborhoods in which attendance tend to follow a strong trend. "Percent of Students Chronically Absent" is almost trivial in why it's important for predicting attendance, obviously schools with low attendance will have high chronical absence.

\_Gradient Boosted Regression Model\_

But the biggest take-away here is the fact that language and math proficiency is so important.

**Baseline Classifier Models**

I decided to train classification models that define and identify schools with “bad attendance.” This way, we can gain insight on what kind of indicators correlate with schools that have bad attendance, and PASSNYC can focus on developing programs at those schools.

**Important Factors related to Attendance Rate**

The most important variables consist of demographic information, multiple school performance scores, economic need index, english/ math proficiency, school income and a collection of specific cities and districts. A lot of this information is repetitive of what we've learned from our regression models. Certain demographic correlate with the economic situation, quality of education, and support systems available in certain areas. There are more variables reflecting the school's quality in this model such as "Strong Family-Community Ties Rating", however this is simply reinforcing the generic idea that the support systems and relationship with the community is very important in education and for kids going to school.

Some new information we've learned here is that school income plays an important role. This seems trivial, but it wasn't as significant in our regression models, although it was observed. Longitude and latitude are shown as important here, which reinforces the idea that certain regions may correlate with better or worse educational conditions.