a. Project Review

The data mining problem is predicting student enrollment by census tract, more specifically speaking, number of student enrolled in any grade between Kth and 5th, in any school yeas between 2011-2012 and 2016-2017.

The data provided by the challenge host is number of students at grade between the Kth and the 5th, from school year 2001-2002 to 2010-2011, for each census tract (totally 157), with no extra features.

A good starting point for our project would be to predict number of students at the Kth grade, since it is the starting point for the following time series.

1	Census Tract	School Year	Grade Level	Count of Students
3351	186	20012002	K	23
3352	186	20012002	1	17
3353	186	20012002	2	18
3354	186	20012002	3	13
3355	186	20012002	4	23
3356	186	20012002	5	16
3357	186	20022003	K	26
3358	186	20022003	1	20
3359	186	20022003	2	19
3360	186	20022003	3	16
3361	186	20022003	4	12
3362	186	20022003	5	24

b. Exploratory Analysis

To know better about the correlation between numbers of students in each grade each year, we plotted the following graphs.

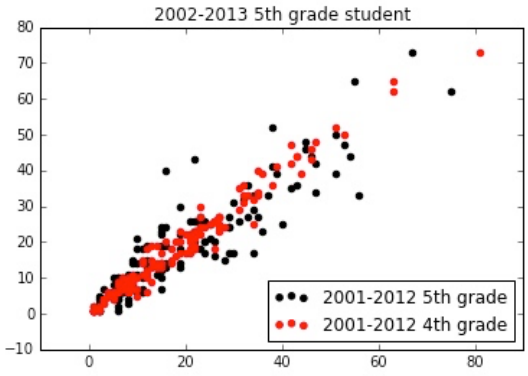


Figure 1

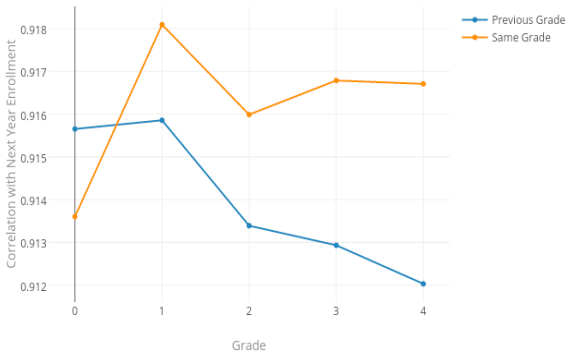


Figure 2

The black dots show the correlation between 2002 school year 5th grade student numbers and 5th grade student numbers in the previous year, while the red dots show the correlation between 2002 school year 5th grade student numbers and 4th grade in the previous year. In this case, as we can see, the 5th grade student in 2002 may more correlated to the lower grade student in the year before. This is also make a lot of sense, since usually when students finish lower grade study they will head to a higher grade study. However, this is not always the case.

As we can see from the Figure 2, if we sum up the data from all census tracts, we may carefully draw the conclusion that for most grades, same grade in this and the next year seems more correlated. This also sets up a frame of data we are going to use to predict different grade student number.

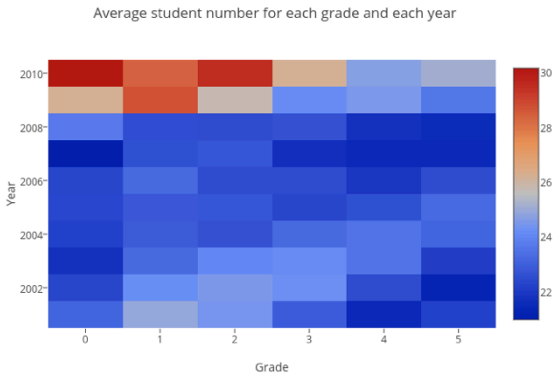


Figure 3

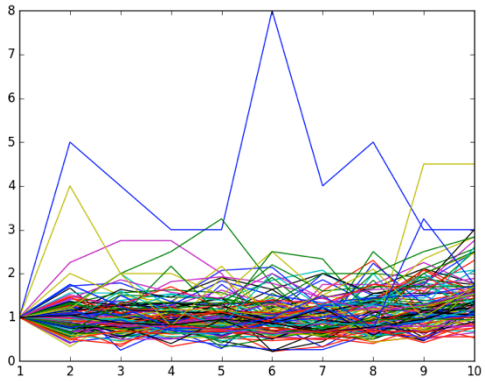


Figure 4

The heat map in Figure 3 shows the change of student numbers in a clearer way. As we can see, the average number of new students for kindergarten increases from 2001-2012 to 2010-2011. Figure 4 shows the ratio of student number in this year to the previous year in different tracts. For most census tracts, the ratios are quite stable around 1 which means the number of students didn't change to much. For those very unstable tracts, we will consider to deal with separately.

c. Prediction of population by age at each census tract

Now we have the most recent census data by age group and subarea for the New York City in 2000 and 2010, and total population projection by age group in New York City for 2000-2014, we try to predict the population by age and subarea for 2000-2014.

The method we use is based on Bayes method. According to the distribution of age and subarea in 2010, we predict the distribution of age and subarea in other years. Denote: $P(x, t, s)$ = Initial population for subarea s, age group x, year t.

To estimate $P(x, t, s)$, the population of a district within year groups x in year t, we applied some Bayesian perspectives.

$$Pr(x, t, s) = Pr(t)P(s\,|t)P(x|s, t) \quad (1)$$
$$P(x, t, s) = P(s\,|t, x)P(t, x) \quad (2)$$

Since we can find estimated population for year t and age groups, which is $P(x, t)$, thus we can calculate $P(t) = \sum_x P(t, x)$, we set $P(x, t, s) = P(x_0, t, s)$ as a initial value, where x_0 is the year of census which we have the data, then we have $P(s|t) = \frac{P(s,t)}{P(t)} = \frac{P(s,t)}{\sum_s P(s,t)}$, where $P(s, t) = \sum_x P(x, t, s)$. Now we update $P(x, t, s)$ through (2) by using $P(x, t)$ we already know and $P(s|t, x) = \frac{P(x,t,s)}{\sum_s P(x,t,s)}$, then we use equation (1) which is $P(x, t, s) = P(s, t)P(x|s, t) = P(s, t)P(x, s, t)/\sum_x P(x, t, s)$ to again update $P(x, t, s)$. We do it until the algorithm converge.

We set the matric to be the absolute differences between independent totals by age group for each subarea and predicted population of the corresponding subarea. Diff $= \sum_x |FT(x, t) - \sum_s P(x, t, s)|$. $FT(t, x)$ is the total population projection by age group in New York City for year t. We repeat the process until the metric is less than 1 or just iterates 100 times.

d. Baseline model to predict number of students at the Kth grade

1. Features and Pre-processing

We have two kinds of features:
The first one is features for census tracts, like race distribution, crime rate, high school graduation rate, etc. And we assume that they don't change with time.
The second one is features for each school year of each census tract. For now, the only available data is population.

For the baseline model, we only use race distribution and population.

For feature race, we know the proportion of several races, so we clustered it, and assigned a label to each cluster, which makes the following sense:

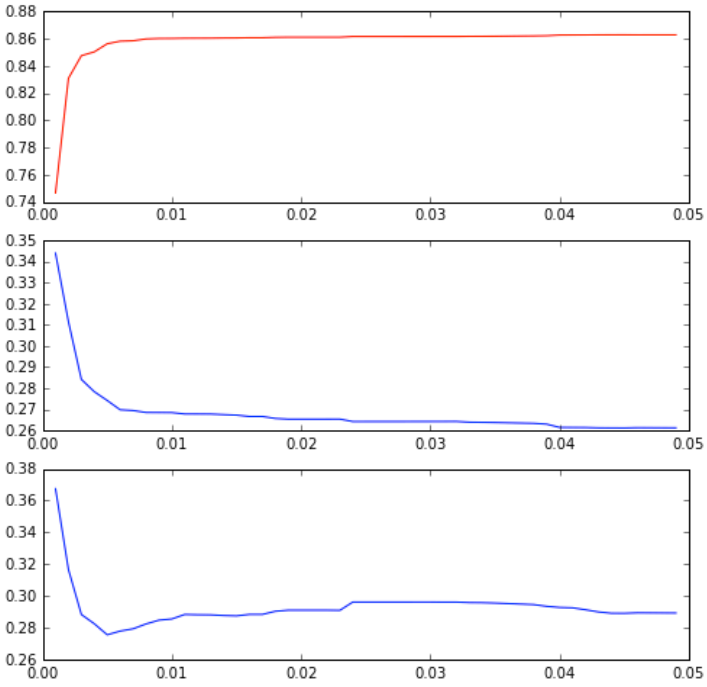
0: Mostly White
1: Half is White, others are Asian and Hispanic
2: Mostly Hispanic
3: Mostly African American
4: Mostly Asian

	tract	total	White	/ African Ame	Asian	o or More Ra	lispanic Origi	race
2	18	3	0	0.33333333	0	0	0.66666667	2
3	20	476	0.0210084	0.01260504	0.04201681	0.00630252	0.90966387	2
4	22	982	0.06924644	0.02953157	0.08248473	0.0203666	0.78716904	2
5	30	362	0.5359116	0.02486188	0.12983425	0.08287293	0.22651934	1
6	34	547	0.55941499	0	0.18647166	0.04753199	0.20658135	1
7	36	625	0.528	0.0112	0.2192	0.0416	0.1952	1
8	38	236	0.66525424	0.00847458	0.13559322	0.05508475	0.12288136	0
9	44	431	0.85382831	0.00232019	0.07424594	0.00464037	0.0649652	0

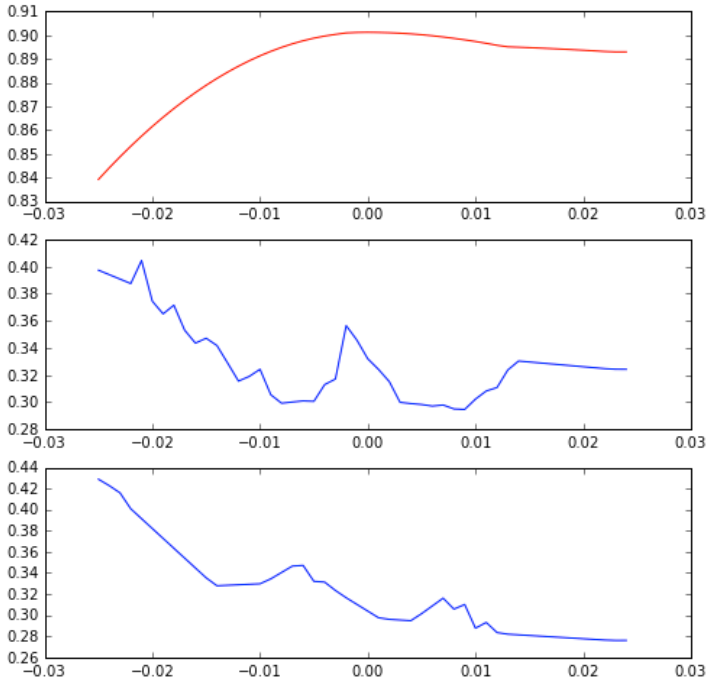
2. Modeling Process and Result

Since this is a regression problem, we construct our model in 4 different ways including lasso regression, support vector regression, neural network, and KNN. Then we compare their performance based on the score of the regression and the median relative absolute error of training dataset and testing dataset.

We found that SVM with linear kernel and Lasso Regression reached a similar and relatively better result.



SVR



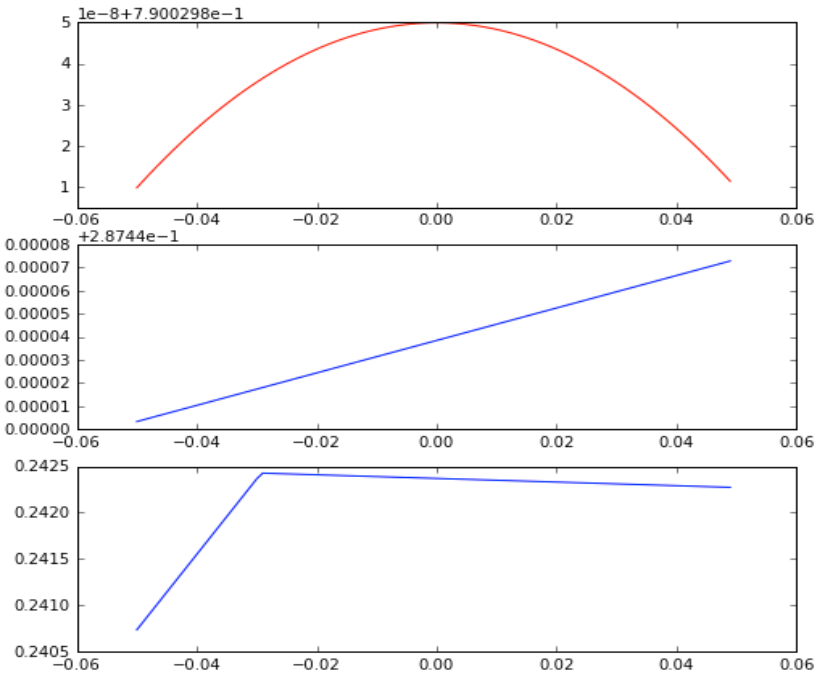
Lasso

These plots illustrate how the performance of two model changes as we tune their hyperparameter. For SVR we change the penalty parameter and for lasso regression we change the alpha value. The plot at the top shows the regression score, i.e., the R square. The plot in the middle shows the median relative absolute error of training dataset. The plot at the bottom shows the median relative absolute error of testing dataset.

Compared to the performance of other methods including KNN, neural network and SVR using other kernel types, which is unstable and shows great risk of overfitting, we can draw a conclusion here that the relationship between the target and features is likely to be linear based on the performance of the various methods we used.

While the score of regression is very high, we are also concerned whether the count of students last year will be a dominant feature that contributes to the most of the change of the target variable. Thus, we also fit a naïve model excluding any other variable except the count of students last year. In other word, we only use the information of the count of students last year to predict the count of students this year. Here are the plots of regression score, MRAE value of training and testing dataset.

From the plots above, we find that the score is greatly reduced by nearly 10%, which indicates that the other features did add some additional valuable information to our model.



Naïve Model

e. Problems and Next steps

Issue 1:
The biggest problem is the lack of detailed data for each census tract.
Issue 2:
After predicting number of students at the Kth grade, it looks like we should apply time series model to predict the following grades. But we only have 10 points (2000 to 2010) which are not enough.

Next, we will add or engineer more features for the baseline model, and apply it to predict the Kth grade for other school years.
Then, we will figure out a way to predict number of students at other grades.

Our team name is The Eastern. Members are Raochuan Fan rf1711, Haichao Wu hw1551, Sheng Liu sl5924, Weitao Lin wl1599