New York University

Graduate School of Art and Science

Center for Data Science

Predict NYC Public School Student Population

February 15, 2017

NEW YORK UNIVERSITY

Abstract

This paper predicts student enrollment by census tract. Different from traditional projection, such as cohort analysis, we introduced two models from different perspectives. One is regression model. After proper transformation of raw data, new data matrix is built for regression analysis applying machine learning algorithms, which embeds the time series nature of the data by adding target variable in previous year as features. Another one is probabilistic graphical model which focuses on each census tract, taking advantage of time-series nature of data by incorporating transit matrix among grades. To get the initial values (kindergarten enrollments), we implement regression on accumulative sum of historical kindergarten enrollments rather than direct on raw data. After ensemble two models above, we improved the average MAE from 5.3 in our baseline model to 2.44.

Contents

1 Introduction

New York City's Public Schools are struggling with overcrowded for some schools while seats available in other schools. Therefore, NYC is looking for an improved model to better predict the future enrollments.

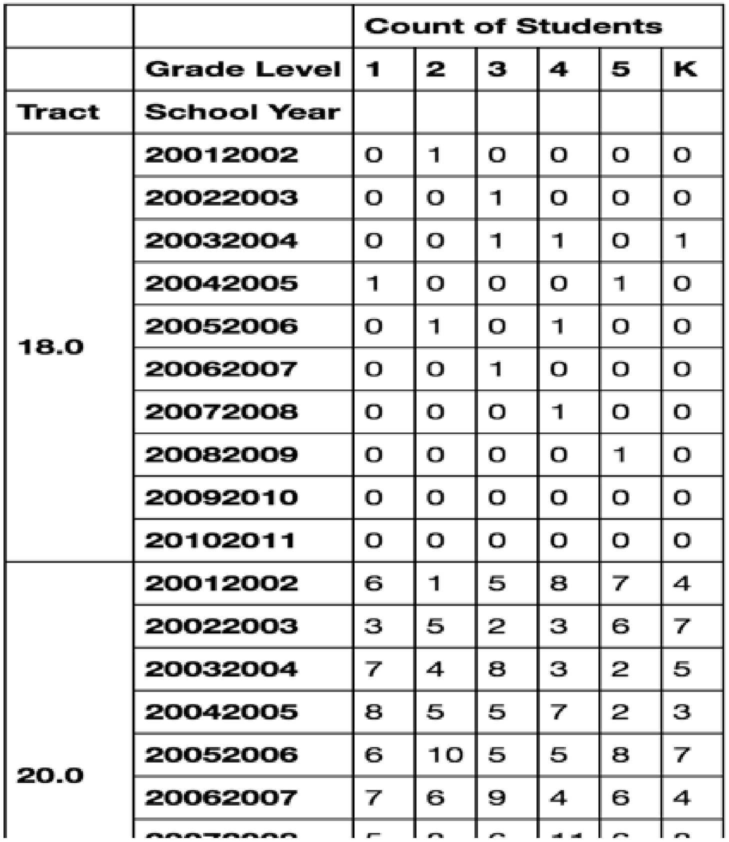
1.1 Business Understanding

More accurate prediction on students enrollment can add business value because public schools can better allocate their equipment and resources given the predicted number of enrolling students. For example, they can decide how many teachers to hire for a new school year. If they hire too many teachers, this will be a waste of money. On the contrary, if they hire insufficient teachers, the quality of education will be badly influenced. Our solution can also be deployed when the government want to get a better understanding of the local education environment and make some optimization. For example, if they want to open a new school or close an existing school, they should take a look at the trend of student enrollment in that area.

Traditional method of projection is to use birth rate to evaluate the students in kindergarten and cohort analysis to predict students in other grades. However, birth rate can not be specific to census tract level and concept may drift due to more than 5 years lagged influences. Moreover, cohort analysis performs better when data with larger magnitudes present. Therefore, a accurate and more specific model should be adopted.

1.2 Data Understanding

The underlying data science problem is predicting student enrollment by census tract, more specifically speaking, number of student enrolled in any grade between K and 5, in any school yeas between 2011-2012 and 2016-2017. And targeted schools are public schools located within Community School District 20. Since we have target variable, which is student enrollment, this is a supervised learning problem. The data instance is student enrollment by grade, school year, and census tract.



		Count of Students					
	Grade Level	1	2	3	4	5	K
Tract	School Year						
18.0	20012002	0	1	0	0	0	0
	20022003	0	0	1	0	0	0
	20032004	0	0	1	1	0	1
	20042005	1	0	0	0	1	0
	20052006	0	1	0	1	0	0
	20062007	0	0	1	0	0	0
	20072008	0	0	0	1	0	0
	20082009	0	0	0	0	1	0
	20092010	0	0	0	0	0	0
	20102011	0	0	0	0	0	0
20.0	20012002	6	1	5	8	7	4
	20022003	3	5	2	3	6	7
	20032004	7	4	8	3	2	5
	20042005	8	5	5	7	2	3
	20052006	6	10	5	5	8	7
	20062007	7	6	9	4	6	4
	[illegible]	5	[illegible]	[illegible]	11	[illegible]	[illegible]

Figure 1: Target Variable

We explore the data set includes students' number in each grade from 2001-2012 school year to 2010-2011 school year first.

To know better about the correlation between numbers of students in each grade each year, we plotted Figure 3 and Figure 4. 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 4, 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.

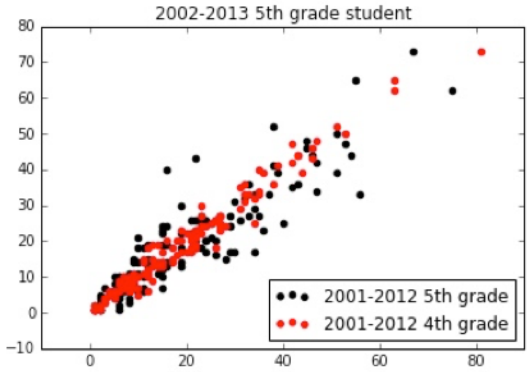


Figure 2: correlation between grade 5 and 4

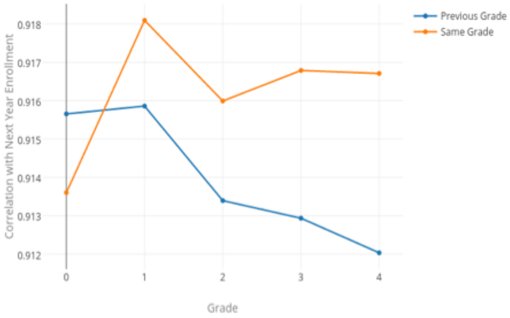


Figure 3: correlation between same grade and l
grade last year

1.3 Data Preparation

Part 1: Population distribution by age (Bayes Method)

We don't know the Age_i for year between 2001-2014, except for year 2010. The frequency of census is only once a decade. We want to use recent census data by age group and subarea for the New York City in 2010, and total population projection by age group in New York City for 2000-2014 to predict the population by age and subarea for 2000-2014.

The method we use is based on Bayes method. Denote:

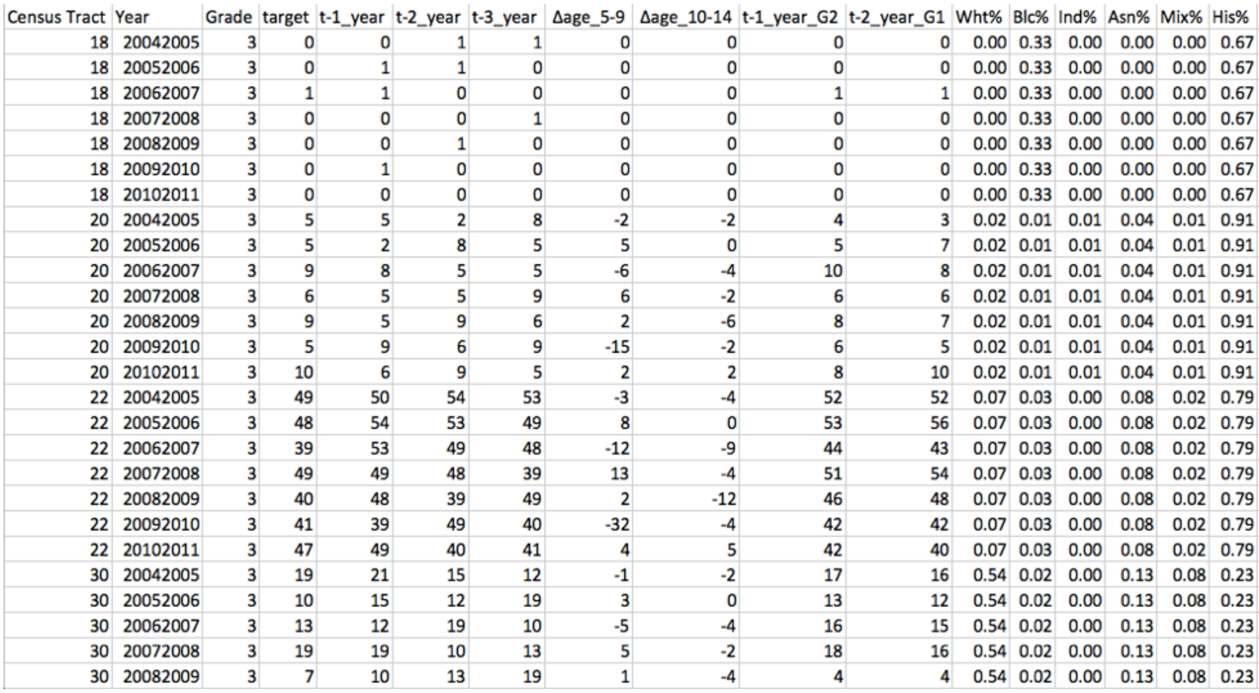
$P(x, t, s)$ = Population for subarea s, age group x, year t.

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

$$Pr(x, t, s) = Pr(t)P(s|t)P(x|s, t) \tag{1}$$

$$P(x, t, s) = P(s|t, x)P(t, x) \tag{2}$$

Since we can find estimated population for year t and age groups, which is $P(t, x)$, thus we can calculate $P(t) = \sum\limits_x P(t, x)$, we set $P(x, t, s) = P(x_0, t, s)$ as a as a initial value, where x_0 is the year of census which we have the data, then we have $P(s|t) = \frac{P(s,t)}{P(t)} = \frac{P(s,t)}{\sum\limits_s P(s,t)}$, where $P(s, t) = \sum\limits_x P(x, t, s)$. Now we update $P(x, t, s)$ through (5) by using $P(x, t)$. We already know $P(s|t, x) = \frac{P(x,t,s)}{\sum\limits_s P(x,t,s)}$, then we use equation (4) which is $P(x, t, s) = P(s, t)P(x|s, t) = P(s, t)P(x, t, s)/\sum\limits_x p(x, t, s)$ to update $P(x, t, s)$ again. We do it until the algorithm converges.



Census Tract	Year	Grade	target	t-1_year	t-2_year	t-3_year	Δage_5-9	Δage_10-14	t-1_year_G2	t-2_year_G1	Wht%	Blc%	Ind%	Asn%	Mix%	His%
18	20042005	3	0	0	1	1	0	0	0	0	0.00	0.33	0.00	0.00	0.00	0.67
18	20052006	3	0	1	1	0	0	0	0	0	0.00	0.33	0.00	0.00	0.00	0.67
18	20062007	3	1	1	0	0	0	0	1	1	0.00	0.33	0.00	0.00	0.00	0.67
18	20072008	3	0	0	0	1	0	0	0	0	0.00	0.33	0.00	0.00	0.00	0.67
18	20082009	3	0	0	1	0	0	0	0	0	0.00	0.33	0.00	0.00	0.00	0.67
18	20092010	3	0	1	0	0	0	0	0	0	0.00	0.33	0.00	0.00	0.00	0.67
18	20102011	3	0	0	0	0	0	0	0	0	0.00	0.33	0.00	0.00	0.00	0.67
20	20042005	3	5	5	2	8	-2	-2	4	3	0.02	0.01	0.01	0.04	0.01	0.91
20	20052006	3	5	2	8	5	5	0	5	7	0.02	0.01	0.01	0.04	0.01	0.91
20	20062007	3	9	8	5	5	-6	-4	10	8	0.02	0.01	0.01	0.04	0.01	0.91
20	20072008	3	6	5	5	9	6	-2	6	6	0.02	0.01	0.01	0.04	0.01	0.91
20	20082009	3	9	5	9	6	2	-6	8	7	0.02	0.01	0.01	0.04	0.01	0.91
20	20092010	3	5	9	6	9	-15	-2	6	5	0.02	0.01	0.01	0.04	0.01	0.91
20	20102011	3	10	6	9	5	2	2	8	10	0.02	0.01	0.01	0.04	0.01	0.91
22	20042005	3	49	50	54	53	-3	-4	52	52	0.07	0.03	0.00	0.08	0.02	0.79
22	20052006	3	48	54	53	49	8	0	53	56	0.07	0.03	0.00	0.08	0.02	0.79
22	20062007	3	39	53	49	48	-12	-9	44	43	0.07	0.03	0.00	0.08	0.02	0.79
22	20072008	3	49	49	48	39	13	-4	51	54	0.07	0.03	0.00	0.08	0.02	0.79
22	20082009	3	40	48	39	49	2	-12	46	48	0.07	0.03	0.00	0.08	0.02	0.79
22	20092010	3	41	39	49	40	-32	-4	42	42	0.07	0.03	0.00	0.08	0.02	0.79
22	20102011	3	47	49	40	41	4	5	42	40	0.07	0.03	0.00	0.08	0.02	0.79
30	20042005	3	19	21	15	12	-1	-2	17	16	0.54	0.02	0.00	0.13	0.08	0.23
30	20052006	3	10	15	12	19	3	0	13	12	0.54	0.02	0.00	0.13	0.08	0.23
30	20062007	3	13	12	19	10	-5	-4	16	15	0.54	0.02	0.00	0.13	0.08	0.23
30	20072008	3	19	19	10	13	5	-2	18	16	0.54	0.02	0.00	0.13	0.08	0.23
30	20082009	3	7	10	13	19	1	-4	4	4	0.54	0.02	0.00	0.13	0.08	0.23

Figure 4: Data Overview

We set the metric to be the absolute differences between independent totals by age group for each subarea and predicted population of the corresponding subarea.

$$\text{Diff} = \sum_x |FT(x,t) - \sum_s P(x,t,s)| \tag{3}$$

$FT(t,x)$ 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.

The reason for inferring population by age group is as following. Intuitively, students enrollment is closely related to population of children. For year t, if we have a predict value of $p(x,t,s)$, then we know the population of age group age_i, which is an important feature for prediction of students enrollment.

Part 2: Data matrix

Our analysis differs from grade to grade, which means we prepare different data table for each grade level. Here is an example for grade 3 as follows. Every data table we use mainly contains 5 groups.

The first group contains information for the target variable including the census tract, school year, grade level which the target belongs to and the target value.

The second group contains information about how many students of the same grade level there were in the past n_years_before years.

The third group includes the change of population in specific age group this year. To predict this, we use Bayes Method.

The forth group indicates how many students of prior grade level there were in the past years (n_grades_before).

For example, when the target variable is students number of the 3th grade in 2005-2006, and n_years_before=2, n_grades_before=2, then students number of the (3rd grade, 2003-2004), (3rd grade, 2004-2005), (2nd grade, 2004-2005), and (1st grade, 2003-2004) are treated as features. Note that for different grade level, we will set different numbers of features in the second, third and forth group.

For the fifth group, we include population ratios of different races. These ratios are calculate based on the 2000 census data because we find they do not change much over time.

The distribution of race in each census tract would be a very important and intuitive feature for the tract. There is only race information of year 2000 and 2010 available. But after comparing race distribution of 2000 and 2010, it is clear that this tract feature doesn't change much, which makes sense.

2 Baseline Model

As we can see, every instance in our data table all belongs to a specific school year and census tract. This kind of data is called panel data as they contain observations of multiple phenomena obtained over multiple time periods for the individuals. Thus, we simply consider using panel regression and include entity effect and time effect to incorporate time series nature of our data.

There are two methods to run regression on panel data and they are fixed-effect and random-effect model. The difference is whether the difference across individuals has a fixed (constant) effect or random effect. We believe that the affect generated from different census tracts and school year should be constant and adopt the fixed effect model. As for the feature we use for our baseline model, we consider the following features:

- The number of students enrolled at the same grade level.
- The change of population in age group under 5, 5 to 9 and 10 to 14.
- The population ratios of all 6 races.

We drop other features such as the number of students of prior grade in past few years because we want to make features consistent for each model. Otherwise, for example, there is no data of prior grade level available for grade K.

2.1 Panel Data Regression

Our baseline model is as follows:

$$S_{i,t} = \sum_{i=1}^{l} \alpha_i + \sum_{t=t_0}^{T} \theta_t + \beta_0 S_{i,t-1} + \sum_{i=1}^{6} \beta_i Race_i + \sum_{i=1}^{3} \gamma_i Age_i \tag{4}$$

Where S_{it} stands for the number of students enrolling in census tract i and school year t, α_i and θ_t are sets of dummy variables indicating the census tract and school year which the instance belongs to. The *Race* and *Age* variables stand for the 6 population ratios and population change of the 3 age group. Parameter $\vec{\beta}$ and $\vec{\gamma}$ are the corresponding coefficients we want to estimate.

2.2 Evaluation

A proper metric should be adopted to evaluate the performance due to the magnitudes of data in different census tracts. For example, if the true value is 0 and the predicted value is 1, relative error won't work. Moreover, if the true value is 1, the predicted value is 0, which is indeed a good prediction, but the relative error will be 100%. Therefore, we adopt Mean Absolute Error (MAE) as our metric.

We randomly split our data into two parts with a ratio 7:3. We use 70% of our data as the training set and the rest as our testing set. the MAE of all grade levels based on test data is shown in Figure 6, we also compare it with ordinary regression:

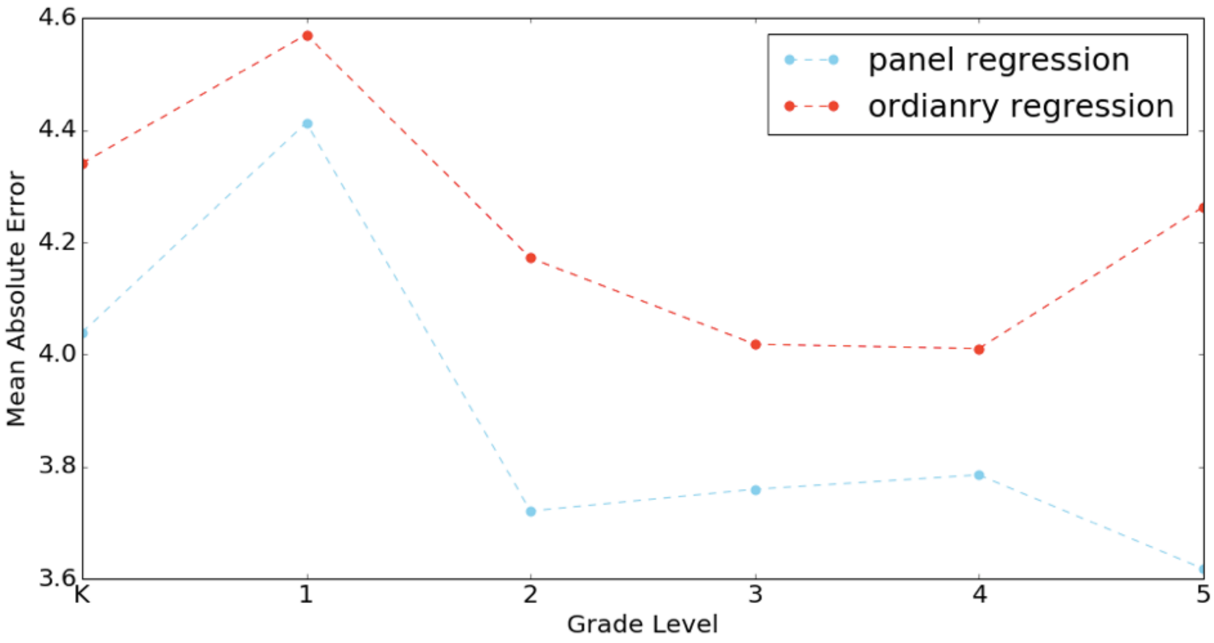


Figure 5: MAE values for all grade levels using panel regression and OLS

Meanwhile, we test the performance of our model by predicting the number of students in school year 2010-2011 for all the grades based on the training set of data in the prior ten 10 years. The average MAE between our predict data and original data is **5.305**. This means in each census tract and each grade, the average difference between predicted students number and true students number is about 5, which is a large number since in some census tract, the total number of students in a certain grade is about 7. As you can see afterwards, the baseline model will be improved when other features and models incorporated. In fact, the average MAE will be reduced to around **2.5** which is less than a half of the MAE now.

3 Improved Regression Model

To improve baseline model, we need to get a rough sense of how well a regression model can do and the relative importance of three kinds of features (students number in previous years, race distribution, and population change), a linear regression model with n_years_before=1 and n_grades_before=1 was fitted to predict students number of the 3rd grade.

There are 1413 samples when the target variable is students number of the 3rd grade.

They are split into train(70%) and test(30%) set. This result quite makes sense:

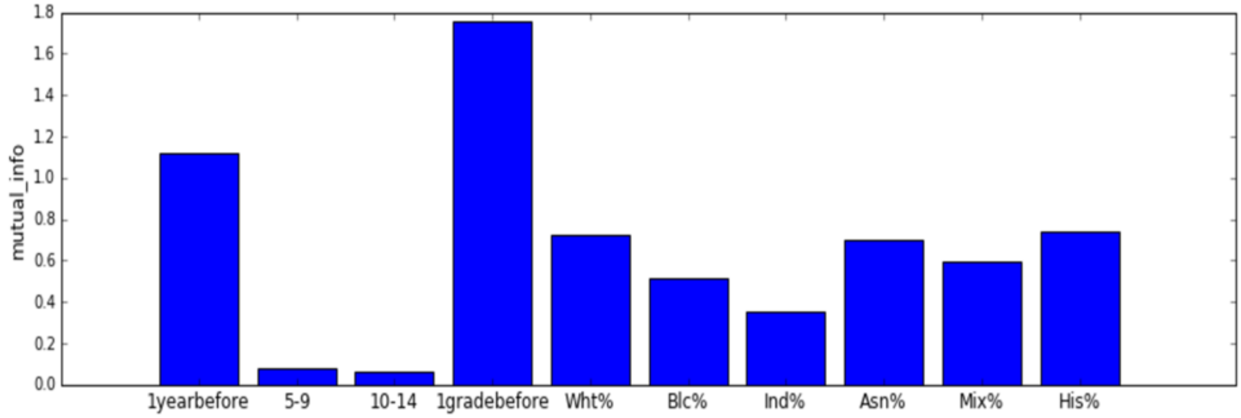


Figure 6: Mutual Information

1_year_before and 1_grade_before is the most important two, which demonstrates the time series nature of the data.

The importance of different races is at the same level. Change of population of related age group (5-9 and 10-14) is the least important.

As a result, Change of Population will be abandoned for the following analysis.

3.1 Model Selection

Before starting, 10% of instances are randomly sampled as the hold-out dataset.

step 1 Algorithm Selection

After testing several regression algorithms, Ridge Regression is found to be the optimal choice. Because its prediction accuracy is among the highest. Moreover, it's cheap to train and easy to interpret. Besides Ridge Regression, other algorithms are also considered and tested: Lasso, SVR, kNN, Back-Propagation Neural Network.

step 2 Feature Selection

After abandoning Change of Population, undecided features are n_years_before and n_grade_before. For the prediction of the 3rd grade, n_years_before is in range(5) while n_grades_before is in range(3).

step 3 Hyper-parameters Selection
The hyper-parameter for Ridge Regression would be regularization strength, alpha.

step 4 Feature Engineering
Since features of students number in previous years are more important, their root or square value would be engineered as new features. Note that the effect of feature engineering needs to be verified.

The first step is finding the optimal configuration of and feature set without feature engineering. GridSearch is applied.

Since the GridSearch technique of scikit-learn has already imbedded k-folds (k=5), so the result is trustworthy. The optimal feature set is n_years_before=4 and n_grade_before =2, which makes sense because it incorporates all the time series information. The lowest MAE is 1.854.

Repeat step1 with feature engineering. However, the result shows that feature engineering isn't improving the performance of the model.

3.2 Model Evaluation

Fit the model with the optimal configuration of feature set and alpha to 90% of not-sampled data. Then test the model on 10% sampled hold-out data. The best MAE is 1.853.Compared with the regression baseline model, MAE decreases by 7%, which is a promising result.

Then, samples in school year 2010-2011 is cut out, and six independent models are fit to the remaining samples by repeating the above steps for each grade. Those models are used for predicting students number in years 2010-2011. The MAE for school year 2010-2011 is 2.454, which is better than the baseline model

4 Model within tracts: Probabilistic Graphical Model

This model is based on data in each census tract.

Motivation : Since the data possess time-series nature, but sadly, only 10 years which are only 10 points in each census tract are given, traditional time-series models, usually deal with data in higher frequency, wouldn't be appropriate to this problem. According to the data understanding part, we see the very high correlation between number of students in higher grade and in lower grade a year before. To incorporate this information should be the most important part of the model designing. Figure 12 shows the basic idea of our probabilistic graphical model. Once the number of students in kindergarten is known, we can use the transit probability p_i (for some situations we accept $p_i > 1$) between each grades to predict number of students in other grades grade by grade. Also, we assume there is a very large pool of students outside of the system, they enroll into K-5 grades in a determined proportion.

Assumptions :

- There is no direct interaction between census tracts
- The influences of indirect factors such as birth rate, immigration, emigration, etc. on enrollment are random
- The students in grade g_i in year t only depends on students in g_{i-1} in year $1, 2, \ldots, t-1$.

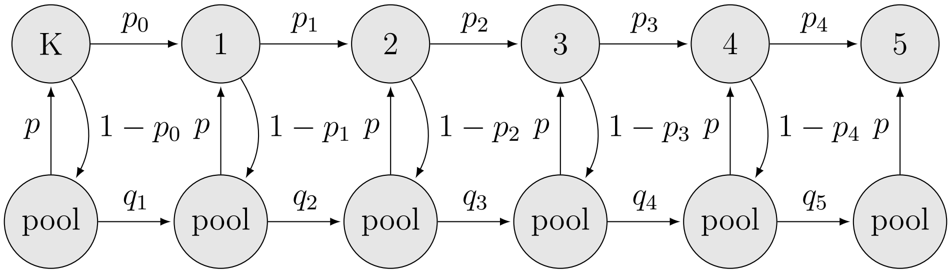


Figure 7: Probabilistic Graphical Model

4.1 Predict The Initial Value: Kindergarten Students

In this section, we will try to estimate the initial value of probabilistic graphical model Which is the population of kindergarten students. We use 2010 Census Tract 20 and 284 as our examples to illustrate the process, other tracts follow the similar process.

As we can see from Figure 8 The trends of it is quite unpredictable partly because we only have 10 points and other indirect factors (such as immigration, birth rate, etc.) which can be considered as random errors of the data. To reduce the randomness, we calculate the cumulative sum of the number of students and form them to a new time-series. Specifically, we transform a list of kindergarten students numbers in N years, say, $\{a_t^{(K)} : t = 1, 2, 3, \ldots, N\}$ to $\{S_i^{(K)} = \sum_{i=1}^{t} a_i^{(K)} : t = 1, 2, \ldots, 10\}$.

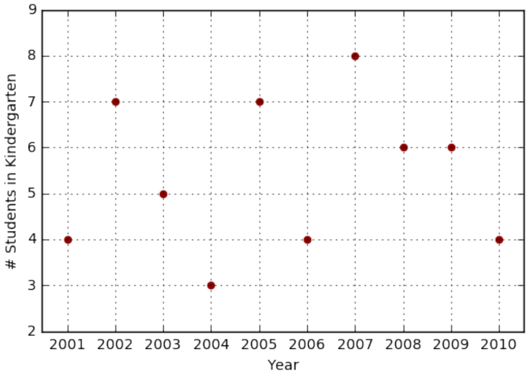


Figure 8: Number of students in tract 20

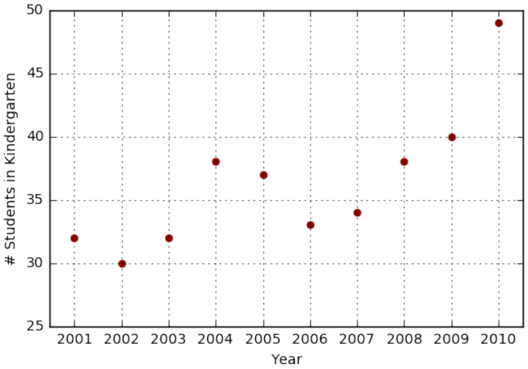


Figure 9: Number of students in tract 284

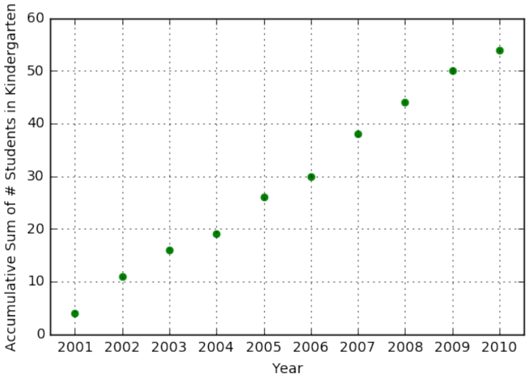


Figure 10: Accumulative sum of number of students in tract 20

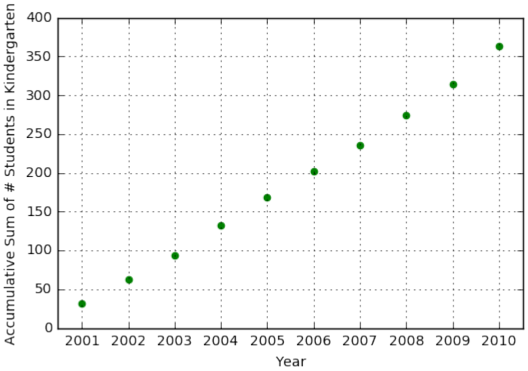


Figure 11: Accumulative sum of number of students in tract 284

After the transformation, the result is shown as Figure 10 and 11. The linear relationship is quite obvious, applying linear regression on these transformed data performs much better than directly on original data.

The regression model is as follows:

$$S_{i+1}^{(K)} = \beta A_i^{(K)} + \alpha \tag{5}$$

We also tried a second order model, which is

$$S_{i+1}^{(K)} = \beta_1 S_i^{(K)} + \beta_2 S_{i-1}^{(K)} + \alpha \tag{6}$$

The performance of (1) and (2) differs on census tracts.

4.2 State Transit

To build the model, we consider assumptions more specifically. There are three possibilities for students from grade i now will go for the next year:
(1) Grade $(i+1)$, with probability p_i, and $p_5 = 0$.
(2) Still in the grade i, or even stay down. We make assumption that the probability is 0.
(3) They will go back to the pool (either drop or move to other tracts), with probability $1-p_i$.

Similarly, as for students in grade $(i+1)$ for the next year, the status of them currently can be:
(1) In grade i this year.
(2) From other tracts.
(3) From grade $(i+1)$ or higher, with probability 0.

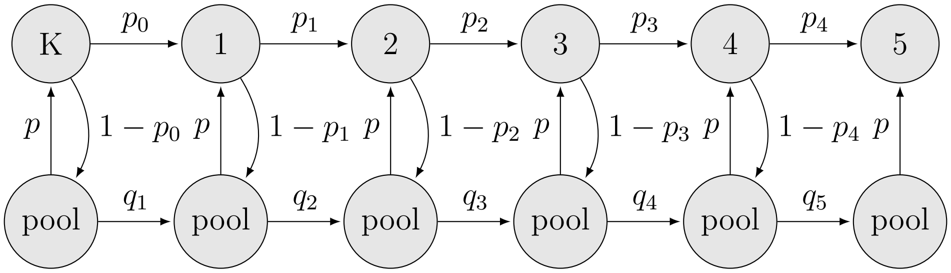


Figure 12: Probabilistic Graphical Model

Set the state vector as $(a_k, a_0, a_1, a_2, a_3, a_4, a_5, a_s)$. a_k to a_5 represents the students population. a_s represents the population of pool.

It is similar to Markov chain, since the the status of next year is related to the status of this year except for a_k, a_s. The state of kindergarten doesn't come from any state. Moreover, for state S, which represents the number of potential students from pool, is influenced by lots of factors, not only students population of this census tract. We also assume that the size of pool won't change as we can see from Figure 12.

We represent this process by matrix equations. For year i to year $(i+1)$:

$$\begin{pmatrix} p_0 & & & & & p \\ & p_1 & & & & p \\ & & p_2 & & & p \\ & & & p_3 & & p \\ & & & & p_4 & p \end{pmatrix} \begin{pmatrix} a_k \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_s \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix}$$

or

$$\begin{pmatrix} a_k & & & & & a_s \\ & a_1 & & & & a_s \\ & & a_2 & & & a_s \\ & & & a_3 & & a_s \\ & & & & a_4 & a_s \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix}$$

We don't need the exact value of these a_s. What we need is how many students will come from other census tract,which is $a_s p$. We set $a_s = 50$.

Step1 : Predict $\vec{p} = (p_0, p_1, p_2, p_3, p_4, p)^T$.

The metric we use to value the error is the l_2-norm of the difference vector $\|b_i - A_i p\|_2^2$.

Suppose we set k years data to train the model, our aim now is to $\min\limits_{p} \sum\limits_{i=1}^{k} \|b_i - A_i p\|_2^2$

$$\begin{aligned} \sum_{i=1}^{k} \|b_i - A_i p\|_2^2 &= \sum_{i=1}^{k} (b_i - A_i p)^T (b_i - A_i p) \\ &= \sum_{i=1}^{k} (p^T A_i^T A_i p - 2 b_i^T A_i p + b_i^T b_i) \\ &= p^T (\sum_{i=1}^{k} A_i^T A_i) p - 2 (\sum_{i=1} b_i^T A_i) p + \sum_{i=1}^{k} b_i^T b_i \end{aligned} \tag{7}$$

Let $A_{sum} = \sum_{i=1}^{k} A_i^T A_i$, $b_{sum} = \sum_{i=1}(b_i^T A_i)^T$
the problem becomes

$$\min_{p} \quad p^T A_{sum} p - 2b_{sum}^T p$$
$$s.t. \quad p \geq 0$$

It is a quadratic programming. We can use quadratic programming solvers to get the optimal solution.

Step2 : Predict students population of next year by using $\vec{b_i} = A_i p$

4.3 Evaluation

We have 10 data points in each grade within one census tract, which can be divided into training sets and test sets. Since we have no hyper-parameters to tune, validation sets won't be necessary, which saves costs in using data given such a small data set. The amounts of training sets are set to 1-9, 2-9, 3-9, 4-9, 5-9 thus that of test sets are the same last one years. As we can see from Figure13 and Table 1, the median MAE of all Census Tracts is quite stable, which means within each census tract, the change is small. The average MAE reduces when we have more years of data to train the model. Also, we can see the variance reduces a lot with more years data incorporated. This means with more training data, we can get a more robust model on all census tracts.

Year of Training Set	Average	Variance	Median
2005-2009	3.535	42.220	2.000
2004-2009	2.732	13.269	2.000
2003-2009	2.512	7.618	2.000
2002-2009	2.459	6.265	2.000
2001-2009	2.529	6.533	2.000

Table 1: MAE on different training sets

We also calculate the error in years started from a given year to detect the life span of the model. What we expected is the error won't increase dramatically along with the passage

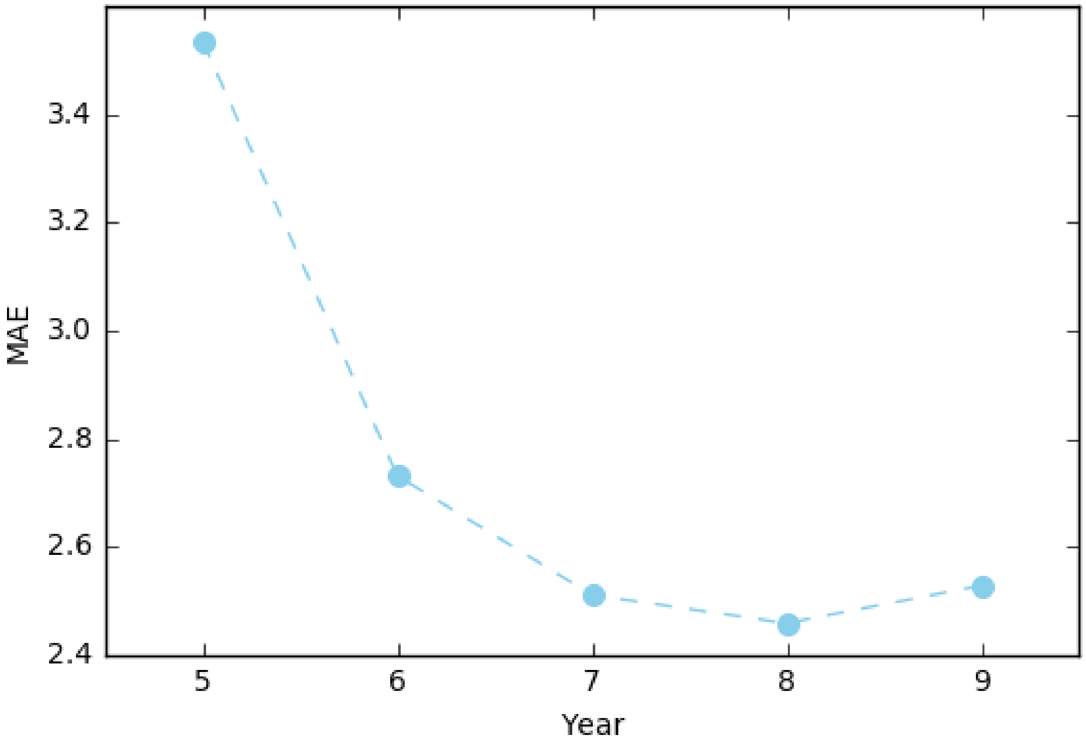


Figure 13: Average MAE for different training sets

of time. We take population data from 2001-2005 as training set, and predict the students population for the year 2006, 2007, 2008, 2009 and 2010. As we can see from Figure 14, the average MAE, as we expected, increases but not dramatically.

Moreover, we calculate the error in different region levels: from census tract to region 20 in total. What we expected is the errors could cancel each other out when we escalate the level. We use formula (9) to calculate the error due to the difference of the base rate in different levels.

$$error = \frac{mean(MAE)}{Total\ Number\ of\ Students\ in\ this\ level} \times 100\% \tag{8}$$

As we can see from Figure 15, the errors in District 20 are less than which in average error of all census tracts, this means the model performs better in a larger level.

The average MAE is about 2.5 for using 7-9 years data as training set to predict the students population next year. There are several possibilities that lead to the error:

(1) The overall population trend is hard to predict. If training data indicate the population is

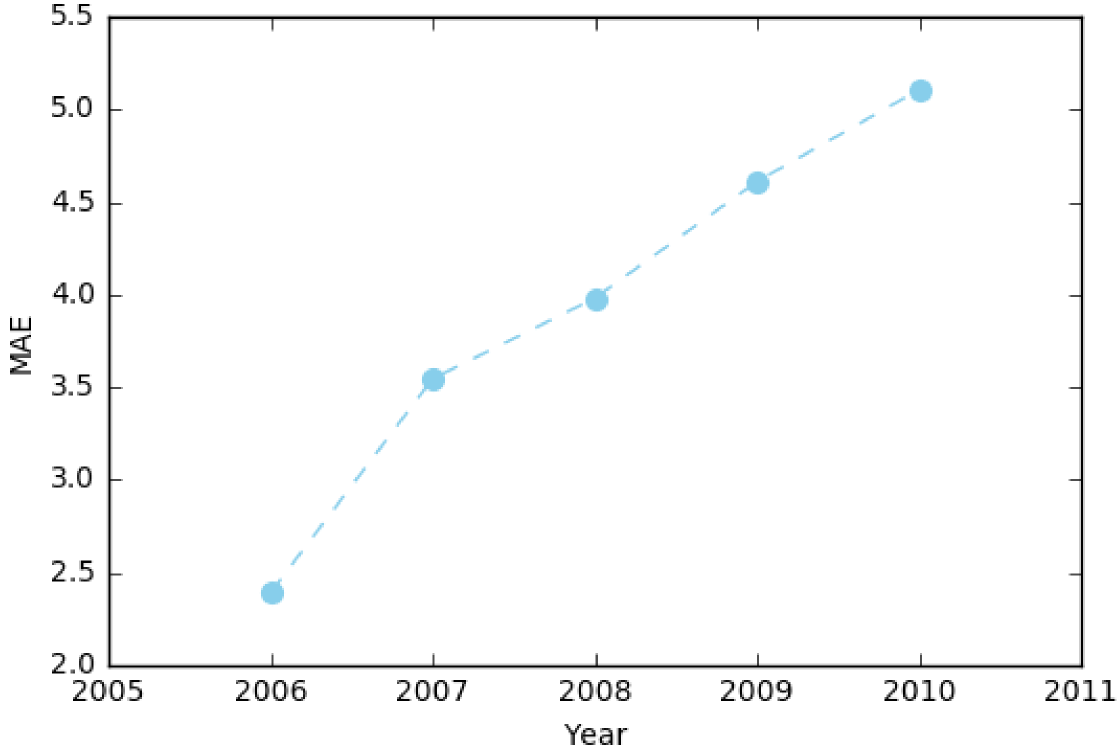


Figure 14: Average MAE for different test years

increasing, the model will predict a higher value than this year, but there is probability that the population of next year will decrease. It is a common cause of error for all the models we use.

(2) We have to predict the population of kindergarten for every census tract, but there is no such detailed and effective data like age distribution for every census tract to minimize the error.

(3) This graphical model is not the optimal choice to predict trend for a long time period.

Definitely, the model have advantages:

(1) It has high accuracy with few information required. We only need the student population to predict.

(2) The model represents the process of population change, which is clear and easy to understand.

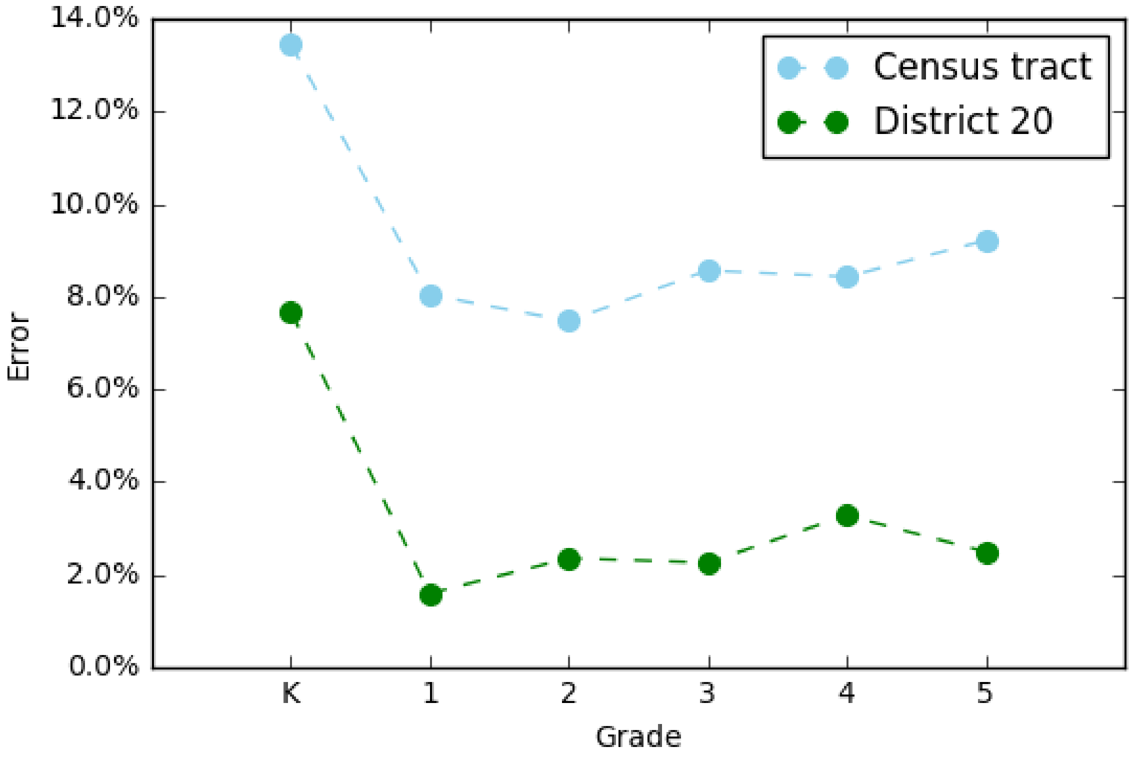


Figure 15: Average MAE for different region level

5 Model Deployment

So far, we get two improved model from different perspectives, the improved regression model is in a region level while the probabilistic graphical model is in a census tract level. Two models actually complement each other: One disadvantage of the probabilistic graphical model is that it heavily depends on the populations of kindergarten, which is quite tricky to predict when we only focus on census tract level. However, the improved regression model, by incorporating other information outside of the system such as population in age 0-10, birth rate and information inside such as population in other census tracts, may performs better. Therefore, we should always use two models for different census tracts: that is, we always select the model performs better on each tract. Through this complementary, as we can see from Figure 16 the average MAE on test data can be reduced to 2.442. Table 2 further represents the detail of the improvements in K-5 grades.

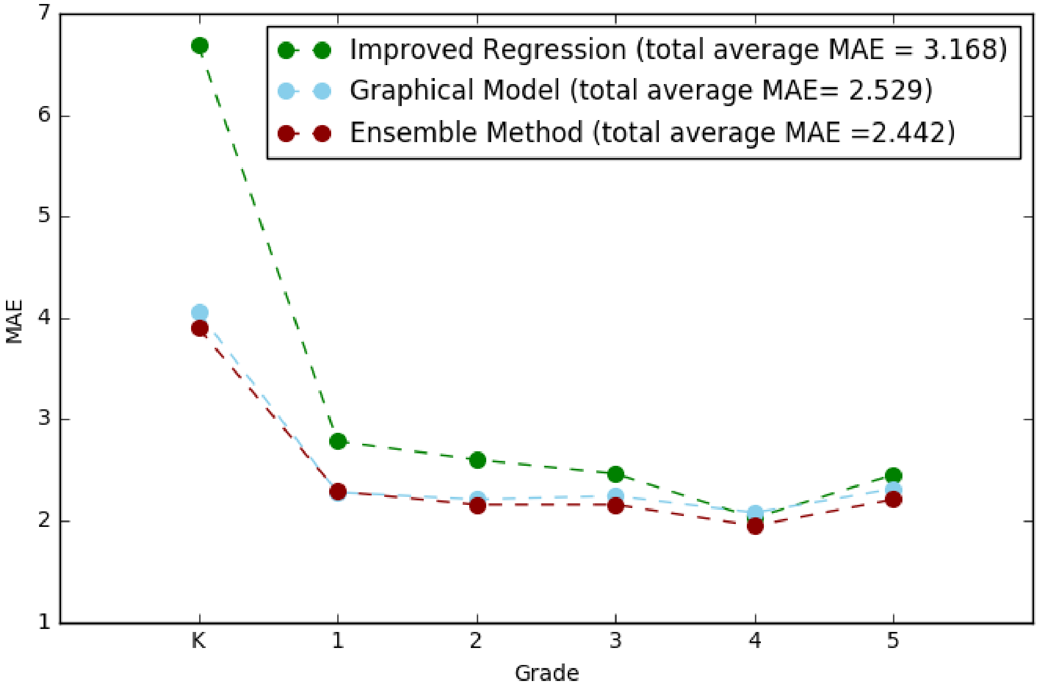


Figure 16: Average MAE for different models

Method	K	1	2	3	4	5
Improved Regression	6,687	2.783	2.600	2.461	2.026	2.452
Graphical Model	4.052	2.278	2.209	2.243	2.078	2.313
Ensemble Method	3.896	2.287	2.157	2.157	1.948	2.209

Table 2: Average MAE for different methods

The result of our model could be applied to school planning, like school budget, teaching positions, and school materials. Because the accuracy of the model will decrease when we predict it far from now ,the accuracy (MAE) of prediction need to be monitored and evaluated.

Also, the prediction of next 3 years is quite stable and trustworthy, but after 3 years prediction, the accuracy may decrease obviously. Thus the model user needs to update the model once new data come in and be extremely careful when using this model for long term planning.

The accuracy of prediction on kindergarten enrollments in probabilistic graphical model would improve when add more information such as population within age 5-8, etc. When these data become available, model should be retrained with these new features.