A Machine Learning Approach to Enrollment Prediction in Chicago Public School

YuFeng Zhuang
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District, Beijing, China
zhuangyf@bupt.edu.cn

Zuyu Gan
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District, Beijing, China
zuyu1994@163.com

Abstract— **Chicago Public School (CPS) allocates billions of dollars to hundreds of public schools including new schools in its system based on prediction of enrollment number for the next year, of which ninth grade is most difficult to be projected. In this project, we propose a method called conditional logistic regression to help them improve the enrollment projection on new high schools. We also design an ensemble model to predict ninth grade. Based on our experiments on two years, our method is better than the current method of CPS.**

Keywords-component; enrollment prediction; conditional logistic regression; ensemble model

I. Introduction

Each spring, Chicago Public School (CPS) allocates $1.8 billion to the hundreds of public schools in its system. To determine where to distribute that money, CPS must predict next year's enrollment for each school months ahead of time, then adjust budgets two to three weeks into the school year when the actual enrollment numbers are set on 20th Day. The time between making projection and obtaining actual enrollment is approximately nine months. Large discrepancies between projected enrollment and the real numbers lead to large adjustments in funding, which can disrupt teachers and students. In 2013, due to the budget allocation problem, 815 support staff, 398 tenured teachers and 510 non-tenured teachers were laid off [1]. Thus accurate enrollment projection and a frequently-updated model are important and necessary.

The problem is most severe on 9th grade, see Figure 1. The reasons are twofold. On one hand, CPS, the third largest public school system in United States, allows students to be enrolled in high schools outside their school districts, which leads to large dispersion of eighth graders when choosing their high schools and the neighborhood (or catchment) high schools struggling to attract students, so more students who live in that area go to school all over the city. On the other hand, there a lot of new high schools opening every year. In 2013, 50 schools were closed and at the same time there were 17 new high schools out of 150 in total. The trend of school opening will continue based on massive wave of public school closure, which means a certain number of schools need to be projected without any history.

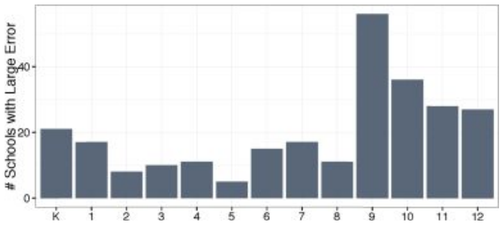


Figure 1. Number of schools with mis-projection over 30 students in 2013 for each grade. K=kindergarten, l-12=grade 1 to grade 12.

The interest about enrollment projection was mainly in the community of sociology and education. Shaw et al. [5] described several methods including cohort survival model, trend analysis (time series) etc. and reported the results of a test about their reliability. Armstrong et al. [1] put the enrollment projection problem into a decision-making framework and compared two methods — Curve Fitting (time series) and Yield from Population Components (regression). Hussar et al. [3] made projections of enrollment, graduates, teachers, and expenditures to the year 2020 using exponential smoothing and multiple linear regression for a lot of elementary and secondary schools and postsecondary degree-granting institutions. To sum up, there are mainly three approaches for school enrollment projection.

1. Cohort Survival Model: this method estimates a series of survival rates that indicate the fraction of students in one grade in a given year who "survive" to the next grade next year; if the total number of second graders in a given school is 100 and the survival rate is estimated as 0.9, then the projection made for grade 3 is 100 x 0.9 = 90.
2. Regression Model: this method employs a formula to calculate simple linear regression with several variables associated to schools such as rating, safety, attendance rate etc. and another variable representing enrollment figures; the simplest form will be linear regression.
3. Time Series Model: this method includes moving average, exponential smoothing, ARIMA model etc.; the idea is essentially fit a curve to represent the trend of time series from the history.

For above three methods, only regression model can be used for new schools; cohort survival model and time series model only use the historical enrollment and cannot be applied to predict new schools. CPS combines cohort survival model and

[1] Chicago Sum-Time News on August 21, 2013: CPS calls teacher's mom to tell him he's getting laid off

978-1-5386-0497-7/17/$31.00 ©2017 IEEE

moving average to make projection. Specifically, for the grades which are transited from previous grade (2-8, 10-12), cohort survival model is applied; while for the grades which are not naturally transited from previous grade (K, 1, 9), moving average model is used. For the new schools, CPS simply imputes with the mean enrollment of several closest school last year if they cannot obtain the information about how many students will be enrolled from the principles. Inspired by the need for better projection methods which can be applied for new high schools, we develop a machine learning approach which incorporates student and school data at individual level. Specifically, we aggregate distance from student home to school, student English level, student attendance, school race, school type etc. into a model called conditional logistic regression. To the best of our knowledge, this is the first method which predict school enrollment from individual-level data and based on our experiments, it has descent performance for new school projection comparing with several baseline methods used at present. When making projection for all high schools, we adopt an ensemble model, i.e. for new schools, conditional logistic regression is used; but for old schools, we find high correlation between last-year enrollment and this-year enrollment thus we simply make projection using last-year enrollment. The ensemble method performs much better than the current methods CPS is using as well as multiple linear regression.

For the rest of the paper, section II describes conditional logistic regression for new schools and the ensemble model for all schools. Section III provides details on the data source we use and the features we put into our model. Section IV shows the experiments and results to evaluate our models. Section V is the conclusion.

II. Model

A. Conditional Logistic Regression

Suppose there are I students and J high schools, for student i and school j , we have K interactive features $\vec{f_{ij}} = [f_{ij}^1, f_{ij}^2, \dots, f_{ij}^K,]$ such as the distance from home to school. Also for one school j, we have L school individual features $\vec{g}_j = [g_j^1, g_j^2, \dots, g_j^L,]$ such as school rating. Denote Y_{ij} which is a dichotomous random variable taking on values 0 or 1 as whether student i choosing school j. For one student, there will be only one school selected hence for one i , only one Y_{ij} in $Y_{i1}, Y_{i2}, \dots, Y_{ij}$ will be 1. The logit of student i choosing school j via conditional logistic regression model is:

$$\begin{aligned} &\mathrm{logit}(\mathrm{P}(Y_{ij} = 1 | \vec{f}_{ij}, \vec{g}_j)) \\ &= \log\left(\frac{\mathrm{P}(Y_{ij} = 1 | \vec{f}_{ij}, \vec{g}_j)}{1 - \mathrm{P}(Y_{ij} = 1 | \vec{f}_{ij}, \vec{g}_j)}\right) \qquad (1) \\ &= \sum_{k=1}^{K} \alpha_k f_{ij}^k + \sum_{k=1}^{K} \beta_l g_{ij}^l \end{aligned}$$

where $\vec{\alpha} = [\alpha_1, \dots, \alpha_K]$ is the vector of coefficients associated with interactive features and $\vec{\beta} = [\beta_1, \dots, \beta_L]$ is the vector of coefficients for school individual features.

Notice that we do not include intercept and student individual features in conditional logistic model since they will be cancelled out if we write (1) into the equation of $\mathrm{P}(Y_{ij} = 1 | \vec{f}_{ij}, \vec{g}_j)$ with simple derivation. This will be shown in Appendix.

Comparing with the form of multinomial logistic regression in (2), conditional logistic regression has much fewer parameters. If the number of features are the same i.e. $\mathrm{K} + \mathrm{L} = \mathrm{M}$, the number of parameters in conditional logistic regression is M while in multinomial logistic regression that is $\mathrm{J}(\mathrm{M} + 1)$. In our problem, J, the number of schools, is $\mathrm{O}(100)$. Hence conditional logistic regression is much cheaper for computation and will be less likely to cause the over-fitting problem. Besides, multinomial logistic regression can only incorporate student individual features but conditional logistic regression can use interactive features between students and schools as well as school features.

$$\begin{aligned} &\mathrm{logit}(\mathrm{P}(Y_{ij} = 1 | \vec{x}_i)) \\ &= \log\left(\frac{\mathrm{P}(Y_{ij} = 1 | \vec{x}_i)}{1 - \mathrm{P}(Y_{ij} = 1 | \vec{x}_i)}\right) \qquad (2) \\ &= \alpha_{0j} + \sum_{m=1}^{M} \alpha_{mj} x_i^m \end{aligned}$$

where $\vec{x_i} = [x_i^1, \dots, x_i^M]$ is the feature vector associated with student i and $\vec{\alpha_j} = [\alpha_{j0}, \dots, \alpha_{jM}]$ is the coefficient vector for school j.

Another important reason we use conditional logistic regression instead of other well-known classifiers such as KNN, SVM, random forest etc. is that all of these methods can only predict the labels existing in the training set, which means that they are only able to “discriminate” the labels based on features. But conditional logistic regression is actually “matching” features to labels. If we want to predict the enrollment number next year, classifiers like random forest cannot provide the probability of students choosing new schools.

If we denote $\vec{x_i} = [\vec{f_{ij}}, \vec{g_j}]$, the probability of student i choosing school j is:

$$\mathrm{P}(Y_{ij} = 1 | \vec{x}_i) = \frac{\exp(\vec{\theta} \cdot \vec{x}_{ij})}{\sum_{j'=1}^{J} \exp(\vec{\theta} \cdot \vec{x}_{ij'})} \qquad (3)$$

Thus the full negative log-likelihood for all students we want to minimize is:

$$\begin{aligned} \mathrm{l}(\vec{\theta}) &= -\sum_{i=1}^{I} \log \mathrm{P}(Y_{ij} = 1 | \vec{x}_i) \qquad (4) \\ &= -\sum_{i=1}^{I} \vec{\theta} \cdot \vec{x}_{ij} + \sum_{i=1}^{I} \log(\sum\nolimits_{j'=1}^{J} \exp(\vec{\theta} \cdot \vec{x}_{ij'})) \end{aligned}$$

It is natural to have automatic method for feature selection hence we add penalty term to (4). Then the optimization problem becomes:

$$\vec{\theta} = \arg\min_{\vec{\theta}} - \sum_{i=1}^{I} \vec{\theta} \cdot \vec{x}_{ij} + \sum_{i=1}^{I} \log\left(\sum\nolimits_{j'=1}^{J} \exp\left(\vec{\theta} \cdot \vec{x}_{ij'}\right)\right)$$

$$+\lambda\left(\gamma \sum_{m=1}^{M} |\theta_m| + 0.5(1-\gamma) \sum_{m=1}^{M} \theta_m^2\right) \qquad (5)$$

where γ trades off between l_1 and l_2 penalties and γ controls the extent of regularization.

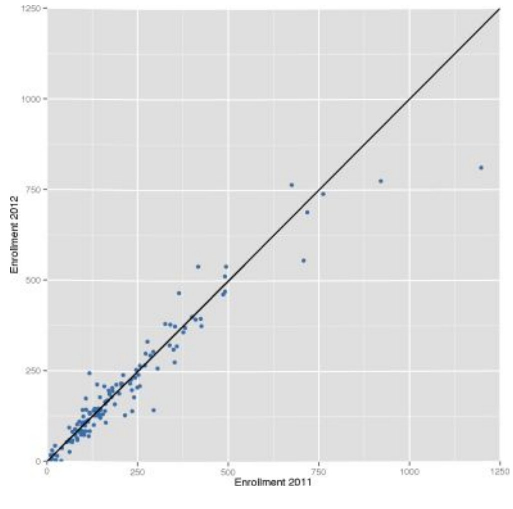
The optimization technique for above objective is cyclic coordinate descent. The convergence of this algorithm is proposed by Friedman et al. [2]. We use "clogitL1" package developed by Reid et al. [5] in R to run conditional logistic regression.

The output of the model is a matrix p_{ij} where i = 1, 2, • • • , I and j = 1, 2, • • •, J with each entry representing the probability of student i choosing school j . To obtain the predictive enrollment for each school, we sum up the probability (soft count) in each column. Specifically, for school j, the predictive enrollment is $\sum_{i=1}^{I} p_{ij}$.

B. Ensemble Model

The intuition behind conditional logistic regression is to provide a descent and interpretable way for projection of new schools. This method can also project the old schools but as we will mention in section IV, it does not perform well, which intrigues us to use ensemble models.

Figure 2 shows the correlation of enrollments in two consecutive years. You can see that the points are very close to the diagonal line except for few outliers thus the enrollments in two consecutive years should be highly correlated. Actually, the correlated coefficient is 0.959 between 2011 and 2012 and is 0.965 between 2012 and 2013.



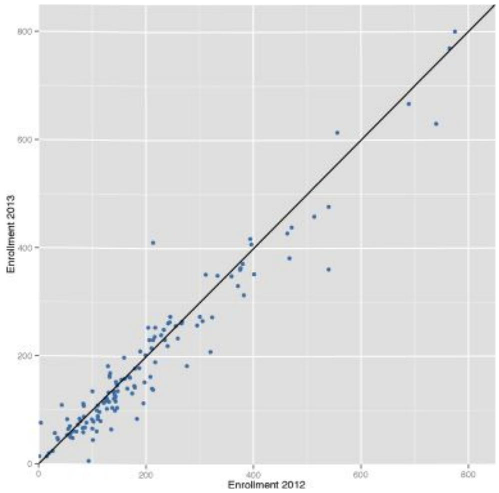


Figure 2. Correlation between enrollment last year and enrollment this year for old high schools. The x-axis is the enrollment last year; the y-axis is the enrollment this year. Top: 2012; Bottom: 2013

Therefore, our ensemble model will simply project the old schools with the enrollment number last year and use conditional logistic regression for new schools.

III. Data

Most of our data are from CPS database. CPS database is a private SQL server database which contains 400 GB data and hundreds of tables, some data are available on CPS website [2]. We also download school features such as school rating from Chicago data portal. The City of Chicago's Data Portal [3] is dedicated to promoting access to government data and encouraging the development of creative tools to engage and serve Chicago's diverse community. The site hosts over 200 datasets presented in easy- to-use formats about city departments, services, facilities and performance.

From section II, we know that conditional logistic regression needs individual school features and student/school interaction features. The followings are the list of features in our model and how they are computed:

1. Distance: radius distance from home address to high school.
2. Percentage of Same Race: percentage of the race which is the same with student race.
3. Percentage of Same ESL: if the student requires English as a Second Language (ESL), we use the percentage of students in schools who need ESL; otherwise, we use 1-percentage of students who need ESL.
4. Percentage of Same Gender: if the student is male, we use percentage of male in schools; otherwise, we use percentage of female.
5. Rating: level 1 to 3 indicates schools from good to bad.
6. Mobility: a measure of how many students are transferring in and out of a school, [add citation]
7. School Attendance Rate: average attendance rate of a school.
8. Catchment School: 1 indicates this school is a catchment school for the students; 0 indicates this school is not a

[2] http://www.cps.edu/

[3] http://www.cityofchicago.org/

catchment school for the students.

9. School Type Match: CPS schools have six types — Neighborhood, Career Academy, Charter, Contract, Magnet and Selective Enrollment; 1 indicates the high school type matches the middle school type of the student and 0 indicates the opposite.
10. Previous Enrollment: the enrollment number last year; if the school is new, then it is 0.

Among above 10 features, 5 are interactive features and 5 are school individual features. For the missing value in the features, if the feature is continuous, we impute missing data with the median; if the feature is discrete, we treat the missing data as a separate factor.

IV. Experiment

We conduct experiments on 2012 and 2013 since our features are only available on 2011, 2012 and 2013. Specifically, we train two models using the features on 2011 and 2012 to predict the enrollment on 2012 and 2013. The summary of statistics for enrollment in 2012 and 2013 is in Table I. The metric we use to evaluate our results is mean absolute error (MAE) which is the average of the error between predictive enrollment and true enrollment across all high schools.

TABLE I. Statistics for Ninth Grader Enrollment in 2012 and 2013.

Year	School	New School	Student
2012	140	8	27798
2013	150	17	28007

a. The second, third and fourth columns are referred to number of all high schools, number of new high schools, number of students

First we look at the performance on new schools and we compare conditional logistic regression with four baselines including current projection method CPS is using:

- CPS projection
- Linear regression
- Mean enrollment last year
- Mean enrollment of closest five schools

The last baseline is very similar with KNN method in machine learning with only one feature distance. The reason I choose five closest schools is that the baseline performs the best using five across the choices from one to ten. The results on 2012 and 2013 are in Figure 3. From the results, we will find that our model has only 62 MAE comparing with our best baseline CPS projection 89 in 2012. In 2013 the MAE of our model is only 49 while the best baseline linear regression is 54.

With respect to the performance on all schools, we compare four models:

- CPS projection
- Linear regression
- Conditional Logistic Regression
- Ensemble Model

The results are in Figure 4. We can see that conditional logistic regression does not have a good performance on all schools but our ensemble model has the best performance. CPS has the second best performance and our ensemble model can improve CPS projection with 3 students and 7 students on average in 2012 and 2013.

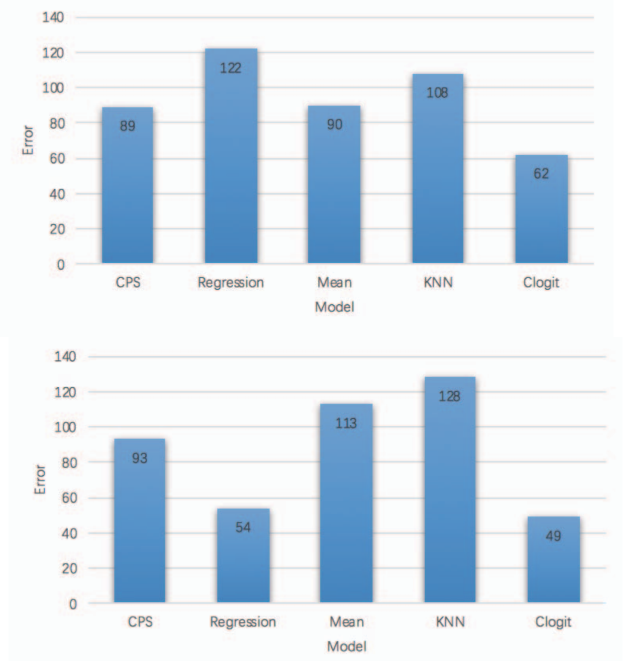


Figure 3. Performance of conditional logistic regression comparing with four baselines on new schools. CPS=CPS projection; Regression: linear regression; Mean=mean enrollment last year; KNN=mean enrollment of closest five schools; Clogit=conditional logistic regression. (Top: 2012; Bottom: 2013)

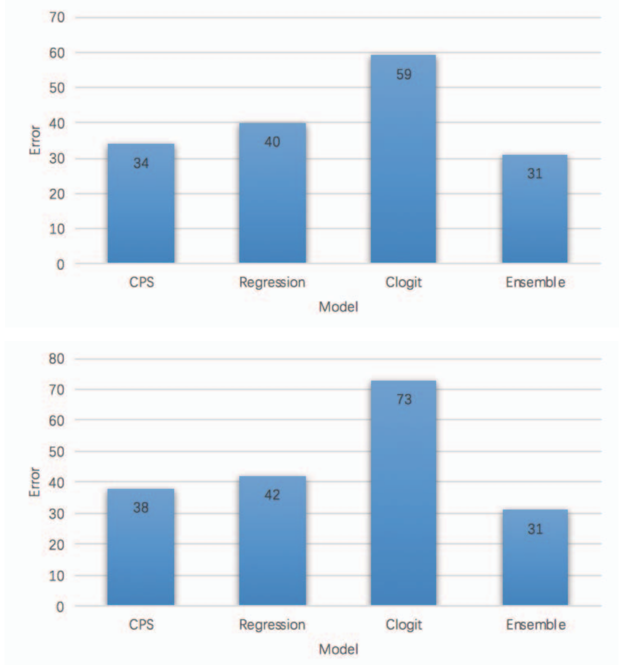


Figure 4. Performance of conditional logistic regression comparing with four baselines on all schools. CPS=CPS projection; Regression: linear regression; Clogit=conditional logistic regression; Ensemble=ensemble model. (Top: 2012; Bottom: 2013)

V. Conclusion

In this paper, we help CPS improve the enrollment projection for ninth grade, especially on new schools. We propose a novel approach "conditional logistic regression" which is the first method implemented in the field of school projection that incorporates student and school data at individual level and can be applied to new school projection. However, we also find that conditional logistic regression does not perform well on old schools. Thus we develop an ensemble method for projection of all schools. Specifically, for new schools, we use the projection from conditional logistic regression while for old schools, we simply copy the enrollment from last year based on the fact that the enrollments in two consecutive years are highly correlated. The result is better than the current projection of CPS and linear regression.

In this paper the projection on all schools is solved by the ensemble model at present but our goal is to use only one model. Conditional logistic regression should be a promising direction. We think it is not working well on old schools mainly because the model does not have the constraint on school capacity. And from the fact that the enrollments in two consecutive years are highly correlated, we believe the enrollment last year has implicitly posed the constraint of capacity on schools. Now the enrollment in the previous year is just a feature in conditional logistic regression but we can also treat it as a constraint of the model.

Acknowledgment

The authors would like to thank Yufeng Zhuang for comments on a preliminary version of this paper, and the anonymous reviewers for their helpful comments. In particular thanks to Stephen Reid and Robert Tibshirani who developed clogitl1 package[5].

References

[1] David F Armstrong and Charlene Wenckowski Nunley. Enrollment projection within a decision-making framework. The Journal of Higher Education, pp. 295-309,1981.

[2] David F Armstrong and Charlene Wenckowski Nunley. Enrollment projection within a decision-making framework. *The Journal of Higher Education*, pp. 295-309, 1981.

[3] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1): 1, 2010.

[4] William J Hussar and Tabitha M Bailey. Projections of education statistics to 2020. nces 2011-026. *National Center for Education Statistics*, 2011.

[5] Stephen Reid and Robert Tibshirani. Regularisation paths for conditional logistic regression: the clogitll package. arXiv preprint arXiv:1405.3344, 2014.