**Executive Summary**

Predicting the Enrollment of Elementary School and Secondary Schools Based on Educational Financing

General Assembly

Yu Zhang

Yuzhang1@seattleschools.org

**Introduction**

Each year the public school districts allocate billions of dollars to elementary and secondary schools based on the enrollment of the next year, by which the enrollment rate is difficult to predict. In this project, we proposed the linear regression and K nearest neighbor to help them improve the projection on the elementary and high schools. We designed the models based on the educational financing features of revenue and expenditure in the last 10 years documented data, the model show high accuracy.

**Statement of the Problems**

The school district has difficulties predicting enrollment due to the time-lapse for almost 9 months before the school starts and predicting enrollment when the last school year ends. The large discrepancies between the actual enrollment and the predicted numbers lead to the difficulties of funding management and interrupt the following up activities of instruction. Due to the budget allocation for the program and supports for elementary students, in addition to the scholarship and rewards to the high school students, accurate enrollment projections are important for the stakeholders.

The school districts are struggling to expand the enrollment, meanwhile, the trend of school reopening after the school closure means are large dispersion to project the enrollment. Armstrong et al put the enrollment project into district decision-making with the regression model for graduation and enrollment. Hussar et al use the logistic regression to analyze the expenditure to predict the cohort high school student’s data from 2019 to 2020. Developing from their method, the current study uses the features of educational financing which have a main effect on the enrollment to predict the enrollment with linear regression and K nearest neighbors. To be more specific, the research questions are:

1. What are the correlations between the features of educational financing and the elementary and secondary students' enrollments in local school districts?
2. What is the model fit for the relevant educational financing features to predict the likelihood of enrollment in the next school year?
3. What is the recommendation to the stakeholders on financing management to increase the enrollment of elementary and secondary students?

**Data Source**

The current study used the data NAEP (National Assessment of Educational Progress) data and Educational financing data from the National Finance Census (National Center for Educational Statistics) to predict the student's enrollment.

           Among the data 6 features are selected. The features are continuous variables, the missing values are dropped. The enrollment is continuous for LR and categorical for the KNN model.

**Data Analysis**

           Data pre-processing is done by selecting the most important features. Data mining techniques are applied to explore the normality of data, classification of educational financing data, and correlation of features. For KNN, a Classification of analysis is used for predictive modeling to obtain the accuracy of the model.

**Results**

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 the school’s districts. Due to the potential multicollinearity of the features, Ridge regression is used as regularization to reduce the overfitting problems of LR. The algorithm shows a slight improvement than the LR. The RMSE is 30554.39 (alpha=0.0).

KNN was used to determine the optimal number of nearest neighbors to define the class of the data. The results of cross-validation and training and testing error indicate the best K is 28-29. Therefore, the KNN model with 29 nearest neighbors performs the best results with five features. The results of the KNN model give a relatively high accuracy score is 0.886. Our result also shows the KNN classification gives higher accuracy than the LR algorithm so the educational financing management can achieve the prediction goals base on the features.

**Conclusion**

In this paper, the prediction of enrollment was solved by both the LR algorithm and the KNN model. The results indicated the spending on instruction, payments to the government, and institutional supports are more likely to increase the district enrollments. However, expanding the federal and total revenue such as allocation of financial aid and scholarship is not the best solution to maximize the enrollments for elementary and secondary students. Meanwhile, we think the KNN works well on the school district data mainly because the model has no constraint on the individual variance. From the facts that the enrollment in the consecutive 10 years data has a small variance. We believe the LR algorithm and KNN can be applied for projection enrollment for the next year.