Paper 2026-2014

SAS® Data Mining for Predictor Identification: Developing Strategies for High School Dropout Prevention

Wendy B.Dickinson, Ph.D., Director of Art Education
Ringling College of Art + Design
2700 North Tamiami Trail
Sarasota, Florida 34234
Morgan C. Wang, Professor and Director of Data Mining
TC II 203, Department of Statistics
University of Central Florida
Orlando, Florida 32816-2370

Executive Summary:

The high school dropout problem has been called a national crisis. Nearly one-third of all high school students leave the public school system before graduating (Swanson, 2004), and the problem is particularly severe among minority students (Greene & Winters, 2005; U.S. Department of Education, 2006). Educators, researchers, and policymakers continue to work to identify effective dropout prevention strategies. One effective approach is to identify high-risk students at early stage then provide corresponding interventions to keep them in school. One of the strength of Educational Data Mining is to reveal hidden patterns and predict future performance by analyzing accessible student data. These predictive algorithms generated by predictive modeling then can be constructed as an early warning system. You might see various early warning systems adopted by schools, districts, and states. However, because individual schools and districts have various combinations of races, genders, social-economic statues, it is impossible to use a set of standardized predictors and obtain satisfactory predictive results. In addition, analyzing limited number of variables and limited historical data cannot generate accurate models. Not mention the predictive model might not consider interactions among predictors. The strength of data mining is the capability of analyzing a large amount of data and variables. Multiple analytic strategies (including model comparisons) can be applied to maximum the model performance. In the section of future goals, we proposed a debuted data mining framework which will construct an early warning and trend analysis systems with components of data warehousing, data mining, and reporting systems at the levels of individual students, individual schools, individual Counties and the whole state.

The purpose of this report is to summarize analysis results of a pilot study, which analyzed high school student data in Dade and Hillsborough Counties from years 2008 to 2010. After data cleaning, the 2008 dataset contains 88 schools in Dade and 72 schools in Hillsborough (160 schools in total). The 2009 dataset contains 68 schools in Dade and 37 schools in Hillsborough (105 schools in total). The 2010 dataset contains 74 schools in Dade and 34 schools in Hillsborough (108 schools in total). Variables in FCAT reading levels, genders, races, and free lunch ratios were adopted to construct a model for dropout rate prediction.

In order to obtain better analysis outcomes, model performance were compared with the following combinations: (a) use 2008 data for training and 2010 data for validation; (b) use 2009 data for training and 2010 data for validation; and (c) use 2008 data for training and 2009 data for validation.

Below summarizes major findings:

1. The model used 2009 data for training and 2010 for validation has the best performance. The model can explain about 80.31% of the dropout rate variance. The results show two main effect variables—Percentages of Students in Levels 1 and 2 Reading—have negative and positive correlations with the dropout rate respectively. Other variable interactions between races, reading levels, and percentage of free or reduced lunch also have positive or negative correlations with the dropout rate.
2. Both 2009-training-2010-validation and 2008-training-2009-validation perform better than 2008-training-2010-validation. The results indicate models using the previous year for training and the current year for validation have better performance.

Project goals and future goals:

Project goals: The main goal of this pilot study is to identify important predictors which can explain or predict variances of dropout rates (i.e. Why individual schools have higher or lower dropout rates) in Florida's high schools.

Future goals: Eventually, we would like to develop an early warning system which can identify at-risk students at the early stage in order to provide corresponding interventions. The early warning system will contain the following components:

- The whole system will be a distributed data-mining framework which contains country-level data warehousing server and state-level data warehousing server
- County-level data-warehousing server synchronizes Students Information Databases from all high schools in the County every week and stores data in data warehousing (Grid) format. The data warehousing server has two major components: data mining and reporting components.
    - Data Mining Component: Each of schools in the Country will have customized algorithms which are generated by analyzing students' historical data. Then these algorithms will be in charge of minoring and tracking individual students in order to identify at-risk students at early stage.
    - Reporting Component: The major strength of the Data Grid format is the powerful reporting and data visualization capabilities. The reporting component can provide real-time, snap-shot or trend reports at the student-level, school-level, and Country-level based on user's needs.
- State-level data-warehousing server only synchronizes important variables with all County-level data-warehousing servers. Similar to the County-level server, the State-level server also contains the data mining and the reporting components.
    - Instead of identifying at-risk students, the data mining component here is to identify at-risk school which might need additional supports. Customized algorithms are also generated by analyzing historical data of individual high schools.
    - Reporting Component: The reporting component in the State-Level data warehousing server aims to provide real-time, snap-shot or trend reports at the school-level, country-level, and state-Level based on user's needs.

Data set overview

The main purpose of this pilot study is to identify important predictors which can explain or predict variances of dropout rates (ie. Why individual schools have higher or lower dropout rates) in Florida's high schools. Therefore, we collected school data in Dade and Hillsborough Counties to conduct a pilot study. The data source contains the following data files:

- dade_county_school_level_data_2008_2010_processed.xls: The dataset contains reading test levels of high school students in Dade County from years 2008 to 2010. The reading test levels were grouped by multiple categories such as school, race, gender, and free or reduce lunch, etc.
- Hillsborough_FCAT_REading_school_level_data_2008_2010.xls: The dataset contains reading test levels of high school students in Hillsborough County from years 2008 to 2010. The reading test levels are grouped by multiple categories such as school, race, gender, and free or reduce lunch, etc.
- 2008_2009_dropout rates by race.xls: The data contains dropout rates of Florida's high schools in 2008. Dropout rate were grouped county, school, and race.
- dropout2009.xls: The data contains dropout rates of Florida's high schools in 2009. Dropout rate were grouped county, school, and race.
- 2010_drbyracebyschl.xls: The data contains dropout rates of Florida's high schools in 2010. Dropout rate were grouped county, school, and race.

Date Cleaning and Processing

The purpose of this study is to identify key predictors of dropout rate from collected datasets. First, we have dropout rates data from 2008 to 2010, and data of reading test levels in Dade and Hillsborough. Therefore, the strategy is to analyze data in Dade and Hillsborough as a pilot study to demonstrate relationships between the predictors and the student dropout rates. The other analysis strategy is to compare model performances by using the following combinations: (a) use 2008 data for training and 2010 data for validation; (b) use 2009 data for training and 2010 data for validation; and (c) use 2008 data for training and 2009 data for validation.

Derived variables were generated for model training and validation in order to obtain better analysis results. Table 1 only lists variables collected or generated for the analysis. The overall dropout rate was adopted as the dependent variable. County and School IDs were excluded from the analysis. The rest of variables are independent variables.

Table 1 List of Variables

Variable Name	Description	Role
County_ID	County ID	ID
School_ID	School ID	ID
PCT_LV1	Percentage of students in level 1 reading	Independent
PCT_LV2	Percentage of students in level 2 reading	Independent
PCT_LV3	Percentage of students in level 3 reading	Independent
PCT_LV4	Percentage of students in level 4 reading	Independent
PCT_LV5	Percentage of students in level 5 reading	Independent
PCT_White	Percentage of White students	Independent
PCT_Black	Percentage of Black students	Independent
PCT_Hispanic	Percentage of Hispanic Students	Independent
PCT_Other	Percentage of students other than White, Black, and Hispanic	Independent
PCT_Female	Percentage of female students	Independent
PCT_FreeLunch	Percentage of students with free or reduced lunch	Independent
Total_Dropout	Overall dropout rate	Dependent

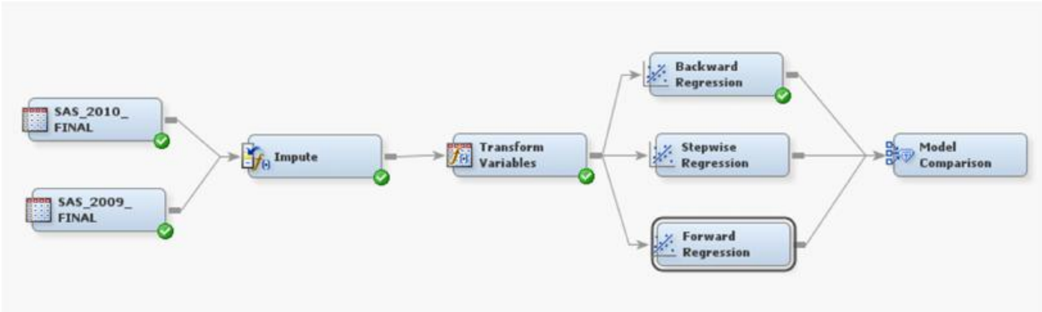
The Dade's FCAT reading dataset contains 380, 416, and 434 schools in years of 2008, 2009, and 2010 respectively. The Hillsborough's FCAT reading dataset contains 245, 255, and 261 schools in years of 2008, 2009, and 2010 respectively. However, the raw dropout rate datasets only contains 131 (2008), 114 (2009) and 118 schools (2010) in Dade Country, and 99 (2008), 71 (2009) and 67 (2010) in Hillsborough Country. In addition, observations without FCAT reading levels and/or overall dropout rate were removed from the analysis. After data cleaning, the 2008 dataset contains 88 schools in Dade and 72 schools in Hillsborough (160 schools in total). The 2009 dataset contains 68 schools in Dade and 37 schools in Hillsborough (105 schools in total). The 2010 dataset contains 74 schools in Dade and 34 schools in Hillsborough (108 schools in total).

Predictive Modeling

SAS Enterprise Miner 7.1 and SAS 9.3 are the major analytic tools for this study. Figure 1 shows an example of the analytic flows for dropout rate prediction. The model use 2009 data (SAS_2009_Final) for model training and 2010 data (SAS_2010_Final) for model validation. The rest of analytic flows (2008 training & 2009 validation; 2008 training & 2010 validation) are similar to the analytic flow in

Figure 1.

Figure 1 An Example of Analytic Flow



Missing value imputation

The dependent variable, Total_Dropout, contains two observations with missing values in the 2009 dataset. These observations with 0 dropout rates were stored in blanks in the raw datasets. Therefore, these missing values were imputed with 0 (see Figure 2).

Figure 2 Results of missing value imputation

Imputation Summary

Variable Name	Number of Missing for TRAIN	Imputed Variable	Impute Value	Role
TotalDropout	2	IMP_TotalDropout	0	TARGET

Variable Transformation

In order to obtain better analysis outcomes, all interval variables were transformed with the max normalization method in the SAS Enterprise Miner. Figure 3 shows formulas for variable transformations in the 2009 dataset. Variables in 2008 and 2010 datasets were also transformed with the max normalization method based on distributions of individual variables.

Figure 3 Results of variable transformation

Computed Transformations
(maximum 500 observations printed)

Input Name	Role	Input Level	Name	Level	Formula
PCT_Black	INPUT	INTERVAL	PWR_PCT_Black	INTERVAL	(max(PCT_Black-0, 0.0)/0.953002611)**0.25
PCT_FreeLunch	INPUT	INTERVAL	EXP_PCT_FreeLunch	INTERVAL	exp(max(PCT_FreeLunch-0.0526315789, 0.0)/0.8807017544)
PCT_Hispanic	INPUT	INTERVAL	SQRT_PCT_Hispanic	INTERVAL	sqrt(max(PCT_Hispanic-0.0287206266, 0.0)/0.9712793734)
PCT_LV1	INPUT	INTERVAL	SQRT_PCT_LV1	INTERVAL	sqrt(max(PCT_LV1-0, 0.0)/0.7962962963)
PCT_LV2	INPUT	INTERVAL	SQR_PCT_LV2	INTERVAL	(max(PCT_LV2-0.0731707317, 0.0)/0.3813747228)**2
PCT_LV4	INPUT	INTERVAL	SQRT_PCT_LV4	INTERVAL	sqrt(max(PCT_LV4-0, 0.0)/0.4)
PCT_LV5	INPUT	INTERVAL	SQRT_PCT_LV5	INTERVAL	sqrt(max(PCT_LV5-0, 0.0)/0.2764227642)
PCT_Other	INPUT	INTERVAL	SQRT_PCT_Other	INTERVAL	sqrt(max(PCT_Other-0, 0.0)/0.1945080092)
PCT_White	INPUT	INTERVAL	PWR_PCT_White	INTERVAL	(max(PCT_White-0, 0.0)/0.7894736842)**0.25

Predictive Modeling

Regression was adopted as the major algorithm for predictive modeling. Three variable selection methods, backward, stepwise, and forward were used for model computation. The validation error was used as the model selection criterion. The model comparison node is applied to select the best model based on the smallest validation errors. Factor interactions and polynomial terms were also enabled to improve the model performance.

Results and Findings

2009 training & 2010 validation

Results of model comparison show that Stepwise Regression is the best model based on validation errors. It can explain about 80.31% of the dropout rate variance (see Figure 4).

Figure 4 Results of Stepwise Regression using 2009 Data for Training and 2010 Data for Validation

Model Fit Statistics

R-Square	0.8031	Adj R-Sq	0.7774
AIC	233.4789	BIC	217.2675
SBC	267.9804	C(p)	235.6443

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.7020	3.5862	-2.98	0.0036
SQRT_PCT_LV1	1	83.4780	8.9909	9.28	<.0001
SQR_PCT_LV2	1	-34.8663	11.9046	-2.93	0.0043
EXP_PCT_FreeLunch*PWR_PCT_White	1	11.2008	2.6442	4.24	<.0001
EXP_PCT_FreeLunch*SQRT_PCT_LV1	1	-26.6470	3.3656	-7.92	<.0001
EXP_PCT_FreeLunch*SQRT_PCT_LV5	1	-2.5842	1.4401	-1.79	0.0760
EXP_PCT_FreeLunch*SQR_PCT_LV2	1	15.2802	4.3302	3.53	0.0007
PCT_LV3*PCT_LV3	1	26.0191	19.0599	1.37	0.1755
PCT_LV3*SQRT_PCT_LV1	1	-99.4225	9.9417	-10.00	<.0001
PCT_LV3*SQR_PCT_LV2	1	82.7556	17.9432	4.61	<.0001
PWR_PCT_White*SQRT_PCT_LV1	1	-42.0813	7.4787	-5.63	<.0001
SQRT_PCT_Hispanic*SQRT_PCT_LV1	1	30.9673	3.4467	8.98	<.0001
SQRT_PCT_Hispanic*SQR_PCT_LV2	1	-26.7241	5.1789	-5.16	<.0001

The results show two main effect variables—Percentages of Students in Levels 1 and 2 Reading—have negative and positive correlations with the dropout rate respectively. Other variable interactions between races, reading levels, and percentage of free or reduced lunch also have positive or negative correlations with the dropout rate.

2008 training & 2010 validation

Results of model comparison show that Stepwise Regression is the best model based on validation errors. It can explain about 69.66 % of the dropout rate variance (see Figure 5).

Figure 5 Results of Stepwise Regression using 2008 Data for Training and 2010 Data for Dalidation

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	75.814820	12.635803	58.15	<.0001
Error	152	33.028237	0.217291		
Corrected Total	158	108.843056			

Model Fit Statistics

R-Square	0.6966	Adj R-Sq	0.6846
AIC	-235.8751	BIC	-244.6383
SBC	-214.3927	C(p)	383.2391

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.3127	0.1134	20.39	<.0001
PWR_PCT_LV5	1	0.5173	0.1875	2.76	0.0065
EXP_PCT_FreeLunch*PCT_LV3	1	-2.3621	0.2207	-10.70	<.0001
EXP_PCT_FreeLunch*PCT_LV4	1	-2.2608	0.2412	-9.37	<.0001
LOG_PCT_Other*SQRT_PCT_Hispanic	1	3.2892	0.8697	3.78	0.0002
LOG_PCT_Other*SQRT_PCT_White	1	-7.3714	2.1069	-3.50	0.0006
SQRT_PCT_Black*SQRT_PCT_White	1	-1.1098	0.4216	-2.63	0.0093

The results show one main effect variable—Percentages of Students in Levels 5. Other variable interactions between races, reading levels, and percentage of free or reduced lunch also have correlations with the dropout rate.

2008 training & 2009 validation

Results of model comparison show that Stepwise Regression is the best model based on validation errors. It can explain about 77.43 % of the dropout rate variance (see Figure 6).

Figure 6 Results of Stepwise Regression using 2008 Data for Training and 2010 Data for Dalidation

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	84.281233	6.483172	38.27	<.0001
Error	145	24.561824	0.169392		
Corrected Total	158	108.843056			

Model Fit Statistics

R-Square	0.7743	Adj R-Sq	0.7541
AIC	-268.9660	BIC	-284.3415
SBC	-226.0014	C(p)	261.8310

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.0938	0.1247	16.79	<.0001
PWR_PCT_LV5	1	2.5590	0.4947	5.17	<.0001
EXP_PCT_FreeLunch*PCT_Female	1	-1.2281	0.2673	-4.59	<.0001
EXP_PCT_FreeLunch*PCT_LV3	1	-1.8570	0.2283	-8.13	<.0001
EXP_PCT_FreeLunch*PCT_LV4	1	-1.0702	0.4223	-2.53	0.0123
EXP_PCT_FreeLunch*SQRT_PCT_Hispanic	1	0.1591	0.0882	1.80	0.0734
LOG_PCT_Other*SQRT_PCT_Hispanic	1	1.7597	0.8991	1.96	0.0522
LOG_PCT_Other*SQRT_PCT_White	1	-5.8881	2.3439	-2.51	0.0131
PCT_Female*PCT_Female	1	4.4818	0.7773	5.77	<.0001
PCT_Female*PCT_LV4	1	-6.7172	1.5140	-4.44	<.0001
PCT_Female*PWR_PCT_LV5	1	-3.1541	0.8221	-3.84	0.0002
PCT_LV4*SQRT_PCT_White	1	3.9568	1.1698	3.38	0.0009
PWR_PCT_LV5*SQRT_PCT_White	1	-1.9408	0.5003	-3.88	0.0002
SQRT_PCT_Black*SQRT_PCT_White	1	-0.7419	0.3840	-1.93	0.0553

The results show one main effect variable—Percentages of Students in Levels 5. Other variable interactions between races, reading levels, and percentage of free or reduced lunch also have correlations with the dropout rate.

Overall, models using the previous year for training (such as 2008 training & 2009 validation, and 2009 training & 2010 validation) have better performance than the model with 2008 training & 2009 validation.