Paper ###-2020

School Violence and Safety

Thu Dinh, Alvi Mahmud, Prajakta Pai, Shelby Rangel

Oklahoma State University

# Abstract

School violence has become an emerging problem these days in the United States which not only negatively impacts children’s educational outcomes but also seriously harms their physical health and emotional well-being. School violence also results in wider social and economic costs as well as a long-term impact, since involvement in school violence at smaller ages potentially leads to future antisocial and criminal behaviors, and social and relationship difficulties. Hence, urgent action is needed to address the global problem of school violence to ensure that all children and adolescents have access to safe and non-violent learning environments. This paper aims to examine the school violence issue in the public schools of the United States, to identify the key factors leading to school violence, and to predict the level of violence at a given school. The methods used were to create new features by grouping correlated variables and adjust the bias in the rare target variable’s distribution by utilizing oversampling technique before building some classification models within SAS Enterprise Miner. The model findings indicate that by utilizing the gradient boosting model, over an 84.5% accuracy rate, 86.8% sensitivity rate, and 73.2% specificity rate at distinguishing whether the school has violent incident are achieved. The model can be further generalized to apply on higher education institutions like colleges and universities. The paper’s findings not only benefit parents in selecting a safe school for their children but also facilitate researchers, school authorities and policymakers in evaluating the effectiveness of current school policies and practices on violence issues, thereby revising strategies to prevent violence and create a safe educational environment. This is essential to achieve the goals on quality education, good health, and well-being for the young people.

# Introduction

## Background

School violence is defined as physical violence, [psychological violence](https://en.wikipedia.org/wiki/Psychological_abuse), [sexual violence](https://en.wikipedia.org/wiki/Sexual_violence), and bullying (UNESCO, 2017). Many such problems arise from antisocial behavior, family upbringing and even access to fire arms along with peer pressure. Numerous incidences of cyberbullying and phycological torture have been recorded over the years and have been observed to be increasing in number (UNESCO, 2017). Almost 9% of students have been in one or more physical fights on school property while nearly 6% had been threatened on school property between 2016 and 2017 (CDC, 2017). According to the New York times and CNN (Bosman, 2019; Wolfe and Walker, 2019), there have been 22 shootings at US schools in which someone was hurt or killed. They have occurred across the country, from Georgia to California, at elementary, middle and high schools, and on college and university campuses. All of these events could be considered school violence and might affect the students attending the school.

## Problem statement

Violence and bullying in schools violate the laws and regulations put in place by the schools to help give students both education and protection. Not only does school violence have a negative impact on the physical and mental health, and emotional well-being of those who are victimized, it also has a detrimental effect on the wider school environment (Ferrara, Franceschini, Villani, Corsello, 2019). More seriously, school violence can also lead to wider social-economic costs and long-term impact as the effects persist into adult life. In light of these potentially negative effects, the study aims at exploring the most important factors that leads to school violence and predicting the level of violence in a given school. The paper’s findings can first help parents in making better informed decisions when enrolling their kids in school by learning first hand from the schools how they prevent and control school violence. The School Survey on Crime and Safety (SSOCS) is a school-level survey where a school representative has an opportunity to answer questions about their safety procedures as well as crime incidents that happen within their school. In addition, understanding school crimes and contributing factors will also help researchers, school authorities and policymakers identify trends in crime and safety issues across time as well as any emerging problems. This, in turns, will facilitate the process of addressing school violence issues and devising strategies to prevent and tackle them so as to create a secure and effective educational environment.

# method

## data description

The School Survey on Crime and Safety (SSOCS) is the primary source of school-level data on crime and safety managed by the National Center for Education Statistics (NCES) within the Institute of Education Sciences of the U.S. Department of Education. The SSOCS is a nationally representative cross-sectional survey of about 4,800 public elementary and secondary schools which has been conducted six times, during the 1999–2000, 2003–04, 2005–06, 2007–08, 2009–10, and 2015–16 school years.

Our chosen data set for analysis only covers the latest administration of SSOCS - SSOCS 2015-2016, in which data were collected from February 22, 2016 through July 5, 2016. The data set has 2,092 rows (corresponding to 2,092 public schools that submitted completed questionnaires) with a total of 476 variables, including 230 unique variables, 50 Jackknife replicate variables which produce national estimates from the listed variables, and 196 imputation flags explaining if and how imputations were performed by the NCES. In term of variable types, there are 101 binary, 89 interval, 203 nominal, and 83 ordinal variables. The data covers such topic as school practices and programs, school security staff, school mental health services, parent/community involvement, limitations on crime prevention, disciplinary problems and actions, and other school characteristics related to school crime.

## DATA CLEANING/ validation

The Jackknife replicate variables and the imputation flags variables are irrelevant to the study’s business problem and objectives, hence they are not used in analysis. For the remaining 230 variables, there also exist some variables that are highly correlated with the target variable (VIOINC16 – Total number of school violent incidents reported) or correlated with each other, for example: INCID16 (Total number of incidents recorded), INCPOL16 (Total number of incidents reported to police), SVINC16 (Total number of serious violent incidents recorded), VIOPOL16 (Total number of violent incidents reported to police), to name a few. These variables are all dropped to avoid multicollinearity. The data set is then partitioned into 75% for training and 25% for validation. A visualization of the analysis flow implemented in SAS Enterprise Miner is provided in Appendix A.

## feature engineering

After removal of irrelevant and redundant variables, the rest input variables are grouped or combined into new features given their relationships and meanings. There are a total of 34 newly created features. Most features are a sum of two or more ordinal/binary variables, for example, two binary variables (1=Yes, 2=No) “Building access controlled locked/monitored doors” (C0112) and “Grounds access controlled locked/monitored gates” (C0114) are combined to form an ordinal variable - “Access\_Controlled” which have 3 levels (1+1=Fully controlled, 1+2=Partially controlled, 2+2=No monitoring/control). 13 of our features are created using this method.

Some of the features are created using main questions and follow-up questions. These variables/questions are mostly regarding school policies, where the main question is whether the school has a particular policy in place and the follow up question is whether the policy was used. For example, “Removal with no services available” (C0390) which is a binary variable (1=Yes, 2=No) and “Removal with no services available - action used” (C0392) which is a nominal variable (-1=Skip, 1=Yes, 2=No). These are added to form a new ordinal variable “Removal\_No\_Services” with three levels (2-1=1=No policy, 1+1=2=Policy available and used, 1+2=3=Policy available but not used). 21 of our features are created using this method. The list of all newly created features is provided in Appendix B. The detailed SAS codes and interpretations for all variables grouping are available on the project’s GitHub page (refer to Appendix E for GitHub link).

## oversampling

Since the distribution of the target variable (VIOINC16) is highly skewed with 17.35% of the schools having no violent incidents, hence we decide to apply oversampling technique so that the proportion of events and non-events gets balanced or less skewed. Accordingly, after splitting the data into a training and validation set, the minority class (target = 0) is oversampled using Proc Surveyselect with unrestricted random sampling method. We also attempt to smote sample using Proc Modeclus, however, due to a large number of categorical variables, that method is discarded. After oversampling the minority class, we append the oversampled data with the remaining training data to create a training data set with 2,416 rows and equal distribution of the target class. The detailed code is provided on the project’s GitHub page (refer to Appendix E for GitHub link). Overall, oversampling leads to better variable selection by the models and we achieve good accuracy with a simpler model. Apart from balancing the data, the weight of evidence method and variable clustering are also tried but they all lead to complex models.

# results

## exploratory analysis

We look into some demographic variables present in the dataset to see if there is visible difference in the distribution between school with no violent incidents and school with some violent incidents. Our exploratory analysis indicates that the smaller schools tend to be less violent while the bigger schools tend to be more violent. This can be intuitively understood that managing more students will be difficult. Similarly, we find school type, number of transfers in and out of the schools, and perceived levels of different disruptions having different distributions for violent and non-violent schools. A visualization for comparison of size and type between violent and non-violent schools is provided below.

|  |  |
| --- | --- |
|  |  |

Figure 1. Comparison in size and type between violent schools and non-violent schools

## Models

### Regression Models for Count Target

With a count target variable (the number of total violence incidents recorded at schools), a couple of Poisson models including Proc Genmod with negative binomial distribution and Proc HPGenselect with zero inflated negative binomial distribution are performed. The reason for choosing these distributions is due to the high dispersion in the target variable (mean=15, standard deviation=25, range = [0, 263]). The goodness of fit tests of the models (refer to Appendix C) show that the Proc HPGenselect model with stepwise selection appears to be the better model. This model has 27 effects which made meaningful interpretation cumbersome, and the Proc Genmod model has no built-in procedure for variable selection like the other model. At this stage of the research, we move to classification approach.

### Classification Models for Binary Target

We convert the target variable VIOINC16 to a binary variable where target=0 when VIOINC16=0 and target=1 when VIOINC16>0. After feature engineering and oversampling, we build Random Forest model, Logistic Regression model, and Gradient Boosting model in SAS Enterprise Miner. The purpose of choosing these models is to get a sense which variables are having the most predictability for our target and all three models we choose provide such statistics in the form of variable importance or odds ratio. The results of the models are reported below for validation data:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** |
| Logistic Regression | 78.8% | 79.7% | 74.8% |
| Random Forest | 88.3% | 95.4% | 53.5% |
| Gradient Boosting | 84.5% | 86.8% | 73.2% |

Table 1. Classification models results

Although Random Forest produces better accuracy, it has very poor specificity compared to other models, thus there is one possibility that the Random Forest algorithm maybe over fitting. Gradient Boosting has a good accuracy, sensitivity and specificity overall (refer to Appendix D for model’s classification table in the validation data), hence it is selected as the best fit model for our data set.

# generalization

The dataset contains responses from schools only, hence possible next step would be to apply the similar analysis on higher education institutions like colleges and universities. Also, the survey is not regular, as the last survey before our analyzed survey was done in 2009-10. With regular survey data available, a deeper analysis of how the violence scenario has been changing over the years could be performed to see whether certain steps introduced to prevent the violence are actually working or not.

The beneficiaries of this research can be both individual citizens and federal governing bodies. On the individual level, parents might benefit from research knowing that smaller school has less chance of violence and bullying and can identify violence prone schools based on the rate of transfer in and out of the school. Given that, parents and the community can also be more actively involved with rooting out the issues. At the state and federal level, the officials could be more lenient about school districts so that new schools which are smaller in size can be better equipped to handle the violence if it stems. Our research shows that cyberbullying is a very good indicator at predicting violence and so, schools can try and monitor the level of cyberbullying and offer students counseling to keep things from escalation. In addition, the department of education can benefit from future researches based on our study to identify whether certain policies and practices are being effective at mitigating school violence or not.

# future studies

With data available of recent past and future years, a time series analysis could be performed that will allow researchers to examine how the policies undertaken by the schools have changed and what impact it has on violence. Considering school size, demographics and incomes are largely fixed effects, this would allow future researchers to be more precise about the causality of policies and actions. In addition, future research can be done to predict transfers of students and find out the effects leading to the transfers to tackle that. Finally, with enough samples, more researches can be implemented to investigate whether such policies as providing security guards, support for mental health assessment, less restriction to school’s effort to mitigate violence, community involvement, etc. can effectively impact violence if being enacted proactively rather than reactively.

# Conclusion

The paper’s analysis and findings indicate the key factors that have a significant impact on the number of violent incidents a school endures include the number of students that attend the school, the grade level of the school, how frequently the students change classes, the amount of community/ parent contributions and their involvement with the school, the percent of students who find academic achievement important and the schools’ practices and policies on school violence. Other factors, such as the crime level in each school district and the student demographics, have minimal impact on future violent incident predictions. Given these findings, we suggest a couple of recommendations as below:

* Reduce the school size (the number of students who attend the school) and decrease the frequency of classroom changes throughout the day
* Create a protective environment by getting the community and parents more involved within the school events/ activities and providing positive role models for the students through mentoring programs
* Create awareness about cyberbullying and its effects on students within the teachers and parents and include cyberbullying as an indicator for school violence
* Implement policies and procedures that have been proven to decrease the number of violent incidents in schools that have not yet been implemented

These discoveries and recommendations could help reduce the violence within schools. Since the students of today are tomorrow’s future, creating a positive educational environment might help produce contributing members of society.

# References

Akers, R. L.; Krohn, M. D.; Lanza-Kaduce, L.; Radosevich, M. 1979. "Social learning and deviant behavior: A specific test of a general theory". *American Sociological Review*. 44 (4): 635–655.

Bandura, A. 1983. “Psychological mechanisms of aggression”. *Aggression: Theoretical and empirical reviews*. 1-40

Bosman, Julie. “Here’s a List of School Shootings in 2019”. *The New York Times*. Accessed October 21, 2019. <https://www.nytimes.com/2019/05/09/us/school-shootings-2019.html>.

CDC. 2017. “Preventing School Violence” Accessed December 4, 2019. <https://www.cdc.gov/violenceprevention/youthviolence/schoolviolence/fastfact.html>.

Ferrara, P., Franceschini, G., Villani, A., Corsello, G. 2019. “Physical, psychological and social impact of school violence on children”. *Italian Journal of Pediatrics*. 45 (76). <https://ijponline.biomedcentral.com/articles/10.1186/s13052-019-0669-z#citeas>

Jackson, M., Diliberti, M., Kemp, J., Hummel, S., Cox, C., Gbondo-Tugbawa, K., Simon, D., and Hansen, R. 2018. *2015–16 School Survey on Crime and Safety (SSOCS): Public-Use Data File User’s Manual* (NCES 2018-107). U.S. Department of Education, National Center for Education Statistics. Washington, DC. Accessed September 16, 2019. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018109>.

Pagolu, M. K., & Chakraborty, G. 2011. “Eliminating Response Style Segments in Survey Data via Double Standardization Before Clustering”. *Proceedings of SAS Global Forum*, Paper 165-2011.

Raja, V. K., & Chakraborty, G. 2016. “Improving performance of Memory Based Reasoning model using Weight of Evidence coded categorical variables”. *Proceedings of SAS Global Forum*, Paper 10961-2016.

UNESCO. 2017. *School Violence and Bullying: Global Status Report*. ISBN 978-92-3-100197-0.

Wolfe, Elizabeth**;** Walker, Christina. “In 46 weeks this year, there have been 45 school shootings”. *CNN*. Accessed October 21, 2019. <https://www.cnn.com/2019/05/08/us/school-shootings-us-2019-trnd/index.html>

# Acknowledgments

We would like to express our great appreciation to Dr. Goutam Chakraborty and Dr. Venu Gopal Lolla for their valuable guidance and excellent feedback.

# Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Thu Dinh Alvi Mahmud

Oklahoma State University Oklahoma State University

[thu.dinh@okstate.edu](mailto:thu.dinh@okstate.edu) [alvi.mahmud@okstate.edu](mailto:alvi.mahmud@okstate.edu)

Prajakta Pai Shelby Rangel

Oklahoma State University Oklahoma State University

[prajakta.pai@okstate.edu](mailto:prajakta.pai@okstate.edu) [shelby.rangel@okstate.edu](mailto:shelby.rangel@okstate.edu)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

# APPENDIX a: PROCESS FLOW

|  |
| --- |
|  |

# APPENDIX B: lIST OF new VARIABLES created from grouping

|  |  |  |  |
| --- | --- | --- | --- |
| **New variable** | **Original variables to group** | **New variable** | **Original variables to group** |
| Access\_Controlled | C0112, C0114 | Outside\_Suspend\_With\_Service | C0410, C0412 |
| Drug\_Testing | C0128, C0130 | Inside\_Suspend\_No\_Service | C0414, C0416 |
| Drilled\_Plans | C0163,C0165, C0167 | Inside\_Suspend\_With\_Service | C0418, C0420 |
| Acceptance\_Groups | C0604,C0606, C0608 | Referral\_School\_Counselor | C0422, C0424 |
| Community\_Involvement | C0204-C0218 | Inside\_Disciplinary\_Plan | C0426, C0428 |
| Law\_Enforcement\_Presence | C0610, C0648, C0612-C0618 | Outside\_Disciplinary\_Plan | C0430, C0432 |
| Law\_Enforcement\_Equipped | C0610, C0620-C0626 | BusPrivilege\_Loss\_Misbehavior | C0434, C0436 |
| Law\_Enforcement\_Participation | C0610, C0628-C0646 | Corporal\_Punishment | C0438, C0440 |
| Mental\_Health\_Assessment | C0662-C0666 | School\_Probation | C0442, C0444 |
| Mental\_Health\_Treatment | C0668-C0672 | Detention\_Saturday\_School | C0446, C0448 |
| Mental\_Health\_Efforts\_Limits | C0674-C0686 | Student\_Privileges\_Loss | C0450, C0452 |
| Teacher\_Training | C0265-C0277 | Require\_Community\_Service | C0454, C0456 |
| Removal\_No\_Service | C0390, C0392 | Disruption\_level | C0374-C0386 |
| Removal\_Tutoring | C0394, C0396 | Cyberbullying | C0389-C0393 |
| Transfer\_Specialize | C0398, C0400 | Limitation\_school\_effort | C0280-C0304 |
| Transfer\_Regular | c0402, C0404 | Parent\_participation | C0196-C0202 |
| Outside\_Suspend\_No\_Service | C0406, C0408 | Written\_Plans | C0155, C0157, C0158, C0162, C0166, C0169, C0170, C0173 |

# appendix C: Count data models results

| **Criterion** | **Genmod** | **HPGenselect** |
| --- | --- | --- |
| Log Likelihood | 64, 528.22 |  |
| Full Log Likelihood | -7,036.95 |  |
| -2 Log Likelihood |  | 14,193 |
| Pearson Chi-Square | 2,969.90 | 3,225.93 |
| AIC (smaller is better) | 14,377.90 | 14,307 |
| AICC (smaller is better) | 14,401.89 | 14,310 |
| BIC (smaller is better) | 15,236.07 | 14,629 |

# Appendix D: Gradient Boosting Model classification table

|  |
| --- |
|  |

# Appendix E: Project GitHub Link

<https://github.com/osu-msba/sss-2020-group-4/tree/master/School%20Violence%20Project>