Quantifying Variation in American School Safety with Explainable Machine Learning:

An Application of Machine Learning Feature Importances for the Social Sciences

Kagen Yick Liang LIM

Academic Year 2020–2021

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts in Quantitative Methods in the Social Sciences,

Columbia University in the City of New York

Abstract

Social scientists, methodologists and policy analysts often work with datasets that contain numerous potential explanatory variables, which contain a mix of data types (e.g., survey items measured on different scales). This presents unique challenges for considering these variables within the same model. An approach utilizing machine learning feature importances, a key metric of explainable machine learning, is illustrated through an investigation of variation in American school safety along markers of student disadvantage (e.g., students with low grades, or students who do not regard academic achievement as important). This study is based on the 2017–2018 School Survey of Crime and Safety (SSOCS) conducted by the United States Department of Education, which captures perceptions and responses from school leaders, on the safety conditions of their respective schools. After an illustrative example based on a conventional linear regression approach, random forest regressors are fit and tuned on a theoretically-relevant subset of features in this dataset. SHapley Additive exPlanation (SHAP) values were then calculated from the tuned random forest regressor model, as feature importance metrics for each feature and observation in the data. Results indicate that schools with a higher percentage of disadvantaged students and/or schools that implemented random sweeps were, on average, less safe. On the other hand, schools that implemented parental involvement strategies were, on average, more safe. The value of this explainable machine learning approach for the school safety literature, policy analysis and wider social and behavioural research is also discussed, along with limitations and future directions on how SHAP values could be employed for social science research.

Keywords: Explainable Machine Learning, Linear Regression, Random Forests, SHapley Additive exPlanations, American School Safety

Acknowledgements

I would like to thank Dr. Michael D. Parrott, who was my supervisor for this M.A. thesis. His patient and dedicated guidance was essential in the development of the core arguments and approaches taken in this thesis. I would also like to thank Dr. Elena B. Krumova for her guidance through the development of this manuscript. I deeply appreciate the help of Sabrina Y., Garda R., Zhi Kang C., and Calvin H. for reviewing earlier versions of this manuscript.

The work performed in this thesis would not have been possible without the many talented individuals and teams who have worked on developing essential open-source software in the R (R Core Team, 2017) and Python (van Rossum & Drake, 2009) languages. I accessed R through the RStudio IDE (RStudio Team, 2020). In the R language, I used the tidyverse package extensively for reading the data in and data cleaning (Wickham et al., 2019). I used the psych (Revelle, 2021) package for generating summary statistics. I used the lmtest package for regression diagnostics (Hothorn et al., 2020). Regressions and likelihood ratio tests were conducted with base R functions. All summary statistics and regression tables were generated with the stargazer package (Hlavac, 2018). I accessed Python through the Google Colab IDE. In the Python language, I used the pandas (McKinney, 2010; The pandas Development Team, 2020) and NumPy (Harris, Millman, & van der Walt, 2020) APIs extensively for data cleaning and statistical computing. Pandas was also used to generate a scatter matrix plot. I used the scikit-learn API extensively for machine learning (Pedregosa et al., 2011; Buitink et al., 2013). I used the ELI5 API to calculate permutation feature importances (Korobov & Lopuhin, 2021). The SHAP API was central to both the conceptualization and the implementation of the analyses in this thesis (Lundberg & Lee, 2017). The matplotlib API was a dependency in the SHAP API to generate the main figures of this thesis (Hunter, 2007).

Table of Contents

Introduction

In a classic paper published in *Statistical Science*, Breiman (2001) defined two cultures of statistical modelling — data modelling, which presumes that some given model (e.g., linear or logistic regression) can approximate the relationship between given inputs and outputs, and algorithmic modelling, which makes little attempt to account for the relationship between inputs and outputs but focuses on the extent that algorithmic models, given some inputs, can correctly predict some given outputs. Breiman was prescient in his delineation of algorithmic modelling as a distinct field of research and practice. Algorithmic modelling was less recognized during the point of Breiman's writing (i.e., at the turn of the century), but has since grown astronomically and is better known today as machine learning and artificial intelligence. However, Breiman's depiction of two distinct cultures may no longer be the most accurate picture. Governments and consumers are increasingly aware of the dangers of the unquestioned applications of machine learning algorithms, particularly when these models are seen as 'black boxes' (Adadi & Berrada, 2018). This is coupled with the widespread availability of personal data which can be easily scraped, collected and used in machine learning models for a wide variety of use cases, such as targeted marketing or recommender systems (e.g., Netflix). Hence, there are increasing demands on researchers, governments and large technology corporations to develop tools for explainable artificial intelligence (also known as XAI), that give insight into these machine learning models. Some of these developments include seminal work that has been used to great effect in biomedical applications, like the application of SHapley Additive exPlanations (SHAP) values to predict hypoxaemia during surgery (Lundberg et al., 2018). These bring machine learning and artificial intelligence models closer to the utility of interpretable data models like linear regressions that researchers have turned to for insights on explaining relationships between variables.

This convergence of the two modelling cultures is coupled with unique opportunities for researchers in the social sciences to work with data at an unprecedented scale (see Salganik, 2017). Massive open datasets with a large number of variables relevant to social research, like the American Census or the General Social Survey by the National Opinion Research Centre at the University of Chicago, are some examples of data with both a substantial number of dimensions and a large sample size. Such datasets are ripe for the application of machine learning models, which are able to learn from high-dimensional datasets with a large number of features.

In this thesis, I apply explainable machine learning to social scientific survey data. I begin with an illustrative example on a conventional social scientific modelling approach, using linear regression modelling. I evaluate the value of this classic approach. Consequently, I focus on the use of the random forest regressor, a non-linear, ensemble tree machine learning model which requires minimal preprocessing or feature engineering, and how it can be used for explaining relationships between independent variables and a given dependent variable. Insights are derived from an explainable machine learning approach, which concurrently considers a large number of mixed-type variables, partially overcomes multicollinearity concerns, and provides insights for quantitative inference through the use of SHAP values (Lundberg & Lee, 2017). Comparisons between results obtained from the linear regression and explainable machine learning models are also made, to compare insights derived from the novel explainable machine learning approach with the conventional social scientific linear regression approach, with a focus on interpreting convergences and divergences between the approaches in the results, and discussing the value of the novel explainable machine learning approach.

The analyses are illustrated through a topical case study on American School Safety. In recent years, high-profile school shootings in the United States (U.S.) have brought the issue of school safety to the forefront of international and national attention. Mass shootings, as devastating as they are, compose merely one kind of school safety threat; there are many potential scenarios that might result in severe, long-term or even fatal consequences on school-going children and teenagers across the U.S. In a more sinister vein, the prevalence of such incidents might even be unequally distributed across the school-going population. One postulation, in line with the differential opportunities theory of crime and deviance, might be that students of lower social class experience school environments that are less safe, given the existence of delinquent subcultures in their schools (Cloward & Ohlin, 1960). The following research question is hence of value, to test the validity of the differential opportunities theory in accounting for socio-economic variation, within the domain of school safety: how might markers of student disadvantage, approximating differences in social class, explain school safety within American schools, specifically safety concerns with regard to violent incidents? The dataset utilized is the most recent edition (i.e., 2017–2018) of the School Survey of Crime and Safety (SSOCS), a cross-sectional study of American K-12 public schools by the U.S. Department of Education; the samples within these surveys are selected to be representative of the population of K-12 public schools across the U.S. (U.S. Department of Education, n.d.).

This thesis is hence a contribution to extant literature in two ways. First, as a topical contribution toward school safety, by explaining the uneven distribution of school safety threats across social class, therefore informing education and security policymaking. This adds functional value and precision to the prevention of school safety incidents. Second, as a methodological contribution of applying explainable machine learning techniques to the social sciences. This joins a growing body of work among social scientists who are cognizant of this trend, and are beginning to deploy machine learning models as alternatives to general linear models or in conjunction with these more conventional linear models, even for problems where explanation and interpretability are valued (see Athey & Imbens, 2019 and Molina & Garip, 2019 for reviews).

Literature Review for Case Study: American School Safety

This review lays the foundation for a literature-informed selection of explanatory or input variables relevant to school safety, by bringing together some segments of the extant literature: (a) school safety, within the larger disciplinary context of crime, deviance and social class and (b) school safety as studied and implemented by researchers and policymakers and (c) quantitative modelling that has previously utilized past or most recent data from the SSOCS.

School Safety: Crime and Deviance Perspectives. This section will focus on the theoretical conceptions underpinning school safety. There has been some empirical work done on the relationship between socioeconomic class and crime, but the direction of this relationship is uncertain. Tittle, Villemez and Smith (1978) reviewed official statistics published by local governments and data that captured respondents' perceptions of crime, and claimed that there was no class variation in criminality. A later response paper (Braithwaite, 1981), then pointed out that the earlier Tittle et al. (1978) study had left out two major categories of data in their analyzes – victimization survey data (i.e., surveys which capture crimes that might not have been reported to law enforcement authorities, but nevertheless were crimes committed against individuals) and direct observational studies (i.e., fieldwork that involves the direct observation of phenomena). With this additional evidence considered, Braithwaite (1981) concludes that there is a link between social class and criminality in the literature. His conclusion is supported by Thornberry and Farnworth (1982), which also establishes that *adults* of lower social class seem to commit more crimes than those of higher social class. This study, however, fails to establish the same inverse pattern between social class and *juvenile* delinquency/criminality. Thus, even though there is growing evidence that social class is indeed an important factor to consider in *adult* criminality, there is no evidence yet for *juvenile* criminality. This is a crucial gap to consider, especially for a study of school safety, where teenage offenders and delinquency are of primary concern.

School Safety: Policy Research and Implementation. Beyond theoretical conceptions of juvenile behaviour, this section will focus on scholarly work that studied how school policies have shaped safety outcomes. One framework can be found in Robinson, Leeb, Merrick and Forbes (2016), which suggests two different groups of constructs within school contexts – *relationships* and *environments* – that are essential to safe schools. This segmentation will serve to structure this review.

Indeed, the social networks and relationships that students have matter, with regard to school safety outcomes. Teachers and school leaders are certainly important parts of this network. Lenzi et al. (2017) have also shown that a given school's sense of community and teacher support are constructs that are inversely related to students' perceptions of danger or risk at school; elsewhere, Astor, Benbenishty and Estrada (2009) establish the leadership of the school's principal as a core organizational element in forming safe schools. Students' friendships are also important considerations. In line with the aforementioned differential opportunities theory, students who have more friends that possess delinquent tendencies are more likely to have a higher self-perceived risk of criminal offending themselves (Yuan & An, 2017). In a study conducted on third to fifth graders, researchers also found that if one's peers at school are victimized by safety incidents at schools, one is more likely to have mental health repercussions, manifested as internalizing symptoms; perceived school safety moderates the relationship between peer victimization and internalizing symptoms (Fite et al., 2019). Beyond the physical boundaries of the school, relationships with parental stakeholders has also been shown to be essential. Lesneskie and Brock (2017) found that schools with more parental involvement tend to, on average, experience lowered levels of violence.

One's physical environment is also essential, perhaps especially in the context of school safety, since safety features can involve modifications to one's built environment. Importantly, the presence of many security or safety features (e.g., presence of dedicated school resource officers, metal detectors) may not be associated with school safety. In fact, schools that are unsafe are often characterized by a large number of visible security features (see Reingle Gonalez, Jetelina & Jennings, 2016 for a review). This might be a case of bi-directional causality, given that unsafe schools might be more likely to spend more on security. Nevertheless, there remain marked inequalities between students who feel safe at school and those who do not. Perumean-Chaney and Sutton (2013) also found that the presence of visible safety measures at school are associated with lowered perceptions of safety, but importantly also found that students who attended schools with possibly less resources or status (e.g., larger class sizes and schools with more reports of behavioural problems) felt less safe than students who attend schools with markers of higher status.

School Safety: Quantitative Modelling of Past SSOCS Data. This section focuses on reviewing specific quantitative models that have employed previous editions of SSOCS Data. This section will, hence, inform the quantitative models in this study. There are a number of studies that have utilized previous editions of the SSOCS. The SSOCS has been conducted during the 1999–2000, 2003–2004, 2005–2006, 2007–2008, 2009–2010, 2015–2016 school years. Over the years, the multiple editions of the SSOCS have considered and quantitatively modelled similar constructs as reviewed above, namely the *relationships* and *environments* and their effects on school safety. This is not meant to be an exhaustive review, but in order to represent the progression of research on the SSOCS, a chronological approach is taken in this section to sequentially represent how each year of the SSOCS data has been used in cross-sectional study designs.

Nickerson and Martens (2008) and Chen (2008) used the 1999–2000 SSOCS data with a cross-sectional study design. Through their exploratory factor analyses, Nickerson and Martens (2008) found four different factors, constituted by groups of school safety mechanisms and policies that were implemented by schools, which they termed as four different school approaches. These were named and interpreted from the items that primarily loaded onto each factor. These were the (a) security/enforcement approach (i.e., items that had this approach as their primary factor loading include using security cameras, detention, and transferring students out of their present school) the (b) crisis plans approach (i.e., items that had this approach as their primary factor loading include having a written plan for shootings, and having a written plan for school riots) , the (c) educational/therapeutic approach (i.e., items that had this approach as their primary factor loading include formal violence prevention training, training staff and teachers in crime prevention) and (d) the control approach (i.e., items that had this approach as their primary factor loading include implementing a strict dress code for students, performing metal detector checks). Following which, Nickerson and Martens (2008) ran two separate hierarchical linear regressions on two dependent variables: school crime and school disruption. For both regressions, they first entered seven demographic variables (i.e., student enrollment, percentage of students with low socio-economic status (SES), percentage of students with special educational needs, neighbourhood crime statistics, percentage of racial minority students, students' grade level and the urbanicity of the school) in the first step and then before inputting three of the four school approaches they uncovered in the second step of the regressions; only (a) educational/therapeutic, (c) security/enforcement and (d) control approaches were considered, as the (b) crisis plans approach was understood as a reactive mechanism to incidents, rather than something that could actively act to prevent them. They found that security enforcement approaches were statistically significant in explaining the variance in school crime or disruption, along with the enrollment, schools' grade level and percentage of students with low education. Notably, the variable capturing the proportion of students of low SES was not statistically significant in these regressions. Chen (2008) was similarly interested in one of Nickerson and Martens' (2008) dependent variables, on school crime, and finding explanatory variables that might account for this outcome variable. Chen (2008) was specifically interested in defining school contexts through a structural equation modelling approach. He fit a structural equation model to the data, with school crime as the dependent variable. He included the following variables: school enrollment, urbanicity, community crime where students live, student socio-economic status, student misbehaviour (i.e., the frequency of bullying and classroom disorder), school safety measures, and serious penalties for students. Chen (2008) generally found that the effects of school location (i.e., urbanicity) and characteristics (i.e., student socio-economic class) were mediated by what he defined as ‘school climate’ factors, which were proxies for school environments, like student misbehaviour and school size. Controlling for school characteristics, like Nickerson and Martens (2008), Chen (2008) found that a harsher approach toward student behaviour was positively and significantly linked to heightened school crime.

Han (2010) utilized the 2003–2004 SSOCS data with a cross-sectional study design. Han (2010) was specifically interested in urban schools’ uniform policies (i.e., the main independent variable) on students’ disruptive or problematic behaviours (i.e., the dependent variable), and further subsetted the data to focus on schools with more than 50 percent minority students. The independent variables Han (2010) considered from the SSOCS, besides whether schools implemented mandatory uniform policies, were factors like the presence of student crime prevention programs, teacher training programs, parental involvement programs, the involvement of the community, along with demographic variables, the percentage of students who were below the 15th percentile of standardized tests scores, and a measure of school size. Han (2010) ran four separate linear regressions, for elementary schools, middle schools and high schools and all schools considered together in a single regression, and found that implementing a uniform policy has a statistically significant, positive relationship with problem behaviours, net of all other variables, for high schools but for elementary schools, middle schools and for all schools considered in the same regression, a mandatory uniform policy had a statistically significant, negative relationship with student problem behaviors. Notably, the variable for the percentage of underperforming students in a given school was statistically significant and positive for the regression subsetted to elementary schools only, and for the regression with all schools considered together, but not for the regressions that considered either middle or high schools in isolation; but one must again note that Han (2010) subsetted the data to consider only urban schools with more than 50 percent minority students, and if the entire SSOCS sample — representative of the national population — was considered, we might expect these findings might be different.

Han and Akiba (2011) utilized the 2005–2006 SSOCS data with a cross-sectional study design, to study socio-economic variation in severe disciplinary action. The dataset was subsetted to consider only middle and high schools, with the reasoning that elementary schools were less susceptible to disciplinary problems. Using multiple logistic regression analyses with expulsion, transfer to specialized schools and suspensions as separate regressions, and all three kinds of severe punishments considered together in a fourth regression, they found that, net of control variables student problem behaviour, school level and school size, schools that served more disadvantaged students were, on average, more likely to deliver more severe punishments. In their operationalization of 'disadvantaged students', Han and Akiba (2011) considered the percentage of (a) students with special educational needs (b) students who were academically underperforming (i.e., below the 15th percentile on standardized tests), (c) African American students, (d) Hispanic students and (e) students in poverty.

Cuellar (2018), Gastic and Johnson (2015) and Lesneskie and Block (2017) utilized the 2007–2008 SSOCS data with a cross-sectional study design. Cuellar (2018) focused solely on the high schools within the sample, and in a similar vein to the approach taken by Nickerson and Martens (2008), Cuellar (2018) conducted factor analysis to uncover latent constructs for how schools have attempted to stem school safety incidents. Cueller (2018) uncovered five main latent constructs of interest, which were interpreted as different approaches that schools have taken — (a) Physical (e.g., metal detector checks, having locked gates), (b) Interactionist (e.g., individual mentoring for students, promoting sense of community), (c) Legal (e.g., having a sworn law enforcement officer, drug testing for students), (d) Parental Involvement (e.g., having parents participate in open houses, or having parents participate in parent teacher conferences), and (e) Community Involvement (e.g., having social service agencies involved in the school, having civic organizations involved in the school). Cuellar (2018) also considered demographic variables like school size, the percentage of minority students, urbanicity and the crime rate in the school's neighbourhood. Some of his key findings, using negative binomial regressions for differents kinds of incidents, include how parental involvement was negatively associated with incidences of violence, physical attacks, and possession of firearms, net of other variables; community involvement was negatively associated with the number of firearm or explosives incidences, net of all other variables, but interactionist strategies were found to be positively associated with violent incidents and firearm incidents, net of all other variables. Gastic and Johnson (2015) focused specifically on the use of metal detectors in schools. Unlike most of the studies in this review, the number of violent incidents was used as an explanatory variable here, with the binary variable of whether a school required daily metal detectors checks of students as the main dependent variable. The use of this binary dependent variable necessitated the use of a logistic regression model, which Gastic and Johnson (2015) applied. The urbanicity/location of the school, grade level of school, school size and the proportion of the students who are students of colour, were entered as control variables. Gastic and Johnson (2015) found that schools that were high-violence and with high minority populations, were on average, and net of all other control variables, more likely to apply daily metal detector checks as a security measure. This suggests disproportionality in the distribution of daily metal detector checks, according to socio-economic divides. Lesneskie and Block (2017), as reviewed briefly in the section above too, utilized the data to investigate the specific contributions of parental and community involvement to the stemming of school violence. They utilized variables which captured school security (e.g., use of metal detectors, use of security cameras), school climate (e.g., teacher support, sense of community at the school), parental involvement and community involvement, as well as control variables for racial composition, school size, crime around the school's area and urbanicity of the school. Net of all the variables considered, negative binomial regression analyses revealed schools with higher levels of community and parental involvement indeed experienced less violence on average.

Literature-Informed Modelling: A Synthesis. Based on this review, a number of salient variables for consideration in a quantitative model for school safety have been uncovered. All prior quantitative models reviewed have considered some measure of whether students were disadvantaged (e.g., low-SES, academic achievement, or minority students), which happens to be the primary interest of this present study. In terms of additional explanatory or control variables, school size and the use of security features or enforcement were present in all these models too. A number of additional explanatory variables: neighborhood crime, grade level, location (i.e., urbanicity of the school), educational approaches taken by the school, parental involvement and community involvement, were also considered by a number of the reviewed studies.

The Present Study

The present study investigates variation in American school safety, along various measures of student disadvantage (e.g., students with lower academic achievement) primarily through an explainable machine learning approach, but an illustrative example of a regression modelling approach is also explored, in order to derive some results based on conventional social scientific modelling. However, while beta coefficients of explanatory variables in a regression model, with their associated variance and significance levels would be of great interest in a conventional quantitative social science study that solely uses regression based models, in the present study a feature importance measure, SHAP, is of primary interest. This metric is particularly valuable because it enables an extremely localized and granular estimation of how each observation across the dataset, across each feature, contributes to the value of the model's prediction (i.e., a greater or small number of school safety incidents). The use of SHAP values also means that conventional ways that scientific hypotheses are phrased, with a frequentist approach to hypothesis testing as the philosophical basis, cannot be used. At present, methods to determine statistical significance values for SHAP have not been developed, although some promising future directions to do so will be discussed in the final part of this thesis. In one conference paper that has also applied SHAP values to social science or policy research, the researchers used SHAP values to study changes in caste differences for the labour force participation (i.e., a binary variable for yes or no) and the type of work (i.e., a binary variable for blue-collar or white-collar) performed by women in India. They relied on a non-directional conditional hypothesis on the differing importance of caste in predicting labour force participation, over respondents of different ages (Joshi & Joshi, 2019). I take reference from this approach, but additionally form directional expectations for relationships between explanatory variables and school safety.

Expectations. I have three expectations. One deals with student disadvantage and socio-economic variation stemming from differential opportunities theory, while the other two stem from salient environmental and relational features that were recognised as important in the extant literature.

Expectation 1. I expect that schools with more disadvantaged students, on average, experience lowered school safety. Expectation 1 would be falsified if there is either no relationship or a positive relationship between markers of student disadvantage and school safety.

Expectation 2. I expect that the presence of high-security environmental features (e.g., presence of metal detectors), on average, would be associated with lowered school safety. Expectation 2 would be falsified if there is no such relationship, or if there is a positive relationship between the high-security environmental features and school safety.

Expectation 3. I expect that the presence of parental involvement, on average, would be linked to increased school safety. Expectation 3 would be falsified if there was no such relationship, or if there is a negative relationship between parental involvement and school safety.

Survey Data Description. The dataset used is the 2017–2018 School Survey on Crime and Safety (SSOCS) (Padgett et al. 2020), a cross-sectional study of Grade K-12 public schools, across the U.S.. This is the public-access dataset; some variables perceived as sensitive were omitted from this version of the data, such as the percentage of minority students or direct measures of socio-economic class in a given school, which would have otherwise been of interest for this present study. The survey is administered by the United States Department of Education. The SSOCS was sent out to 4803 schools (i.e., 1170 primary, 1704 middle, 1748 high and 181 combined schools), based on a stratified random sample of all public schools, which constitute the sampling frame[1]. 66 schools out of this initial sample, however, were ineligible

[1] School level, locale and enrollment size were used to stratify the entire sampling frame (n=84418). The percentage of white students, and the region of the school were taken into account as well, to order each sub-sample within each stratum, before random samples were taken from each strata. (Padgett et al., 2020, p. F-2).

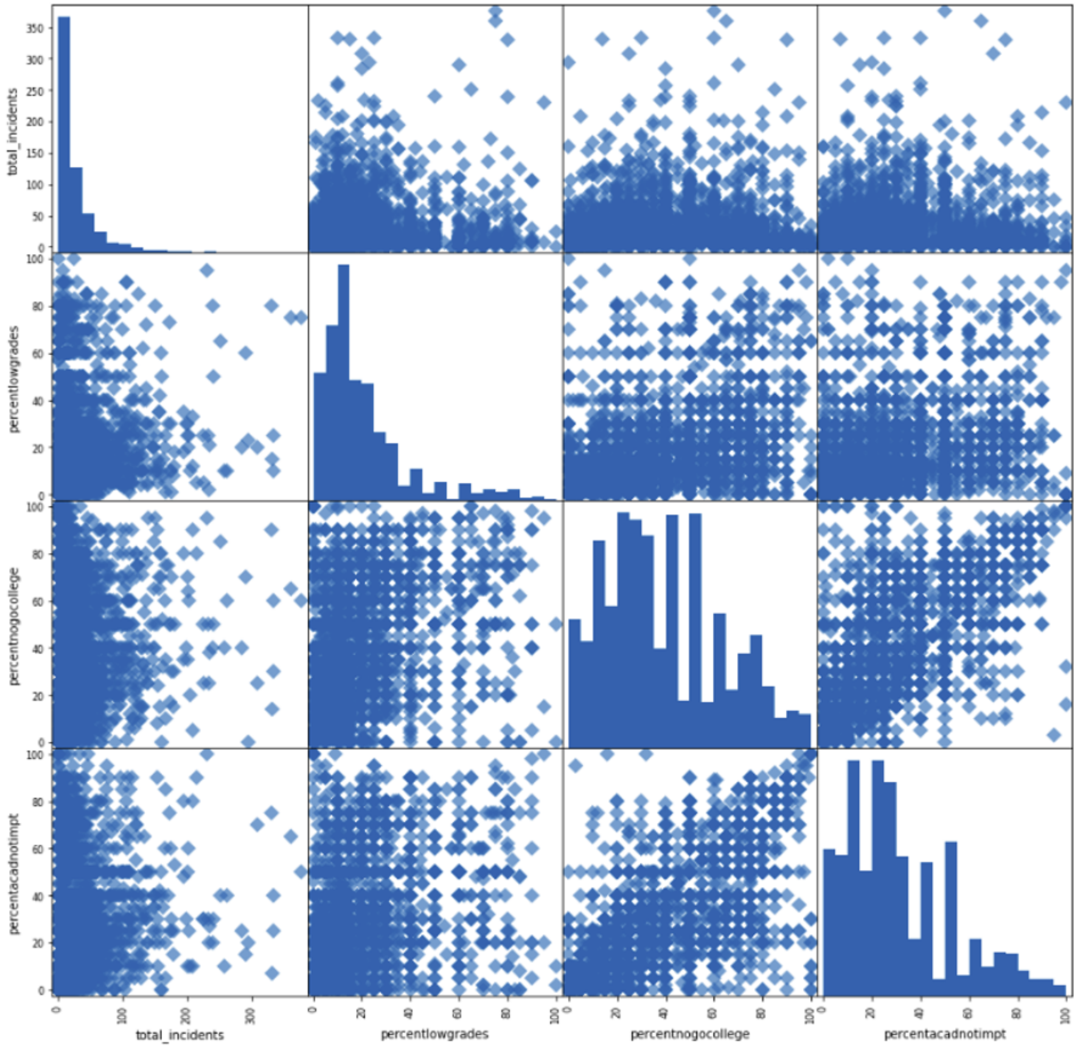
(i.e., 22 primary, 26 middle, 11 high and 7 combined schools)[2]. The response rate was 58.3%, and accordingly 2762 schools were represented (i.e., 671 primary, 975 middle, 997 high and 119 combined schools) in the final sample. Most of the time, the respondents to this survey were school principals (*n* = 2229), but vice-principals or disciplinarians (*n* = 339), school-level staff (*n* = 118), security staff (*n* = 16), and district superintendents or district staff (*n* = 12) also provided information for some schools.

This dataset contains variables which capture school characteristics for the 2017–2018 school year. Three key items are of interest, as proxies for the socio-economic conditions within the school: the percentage of students who a) are below the 15th percentile on standardized tests (*percentlowgrades*), b) are unlikely to proceed to college (*percentnogocollege*), and c) consider academic achievement to be unimportant (*percentacadnotimpt*). These are all ratio variables (i.e., participants filled in a percentage that ranged from 0% to 100%). There are a number of variables which might approximate school safety (e.g., number of violent incidents, number of disciplinary incidents), at various levels of granularity. The most broad and encompassing one is a composite variable, capturing the total number of incidents recorded during the year (*total_incident*). This is also a ratio variable. This composite variables capture range of incidents including rape, sexual assault, robbery, physical attack/fight, theft, possession of a firearm/explosive device, possession of a knife or sharp object, distribution, possession, or use of illegal drugs (Padgett et al., 2020, p. 41). A scatter matrix of the three socio-economic variables and one of the candidate target variables, *total_incident*, is presented as Figure 1.

[2] These schools were no longer in existence, at the point that researchers were reaching out to them. They might have closed down, or had a change in status (e.g., school merger, privatization). (Padgett et al., 2020, p. 27)

Figure 1

Scatter Matrix of Socio-Economic Variables and total_incident



Notably, there appears to be a positive correlation between *percentnogocollege* and *percentacadnotimpt*, suggesting that these two variables could share some overlapping variance in the data. Importantly, both *total_incidents* and *percentlowgrades* are right-skewed, which accounts for the bulk of the data at smaller *total_incidents* values, for the bivariate relationships between *total_incidents* and each of the three socio-economic variables. This could mean that there is no clear linear relationship between these variables, and non-linear ways to conceptualize this relationship are needed.

There are some additional variables of interest within the dataset. Notably, in line with the literature review above, this dataset contains features regarding students' relationships and environments. The dataset includes 21 security measures that schools have implemented, as part of their efforts to enhance school safety.[3] These are categorical variables; respondents answer with 'yes' or 'no', with regard to whether these measures have physically been implemented to shape students' environments. Critically, the implementation of some security measures (e.g., presence of metal detectors) could also vary along socio-economic divides, informing us about socio-economic variation. Additionally, there are 12 additional variables on parental and community involvement in school, which provide some indication of whether parental or community input was sought within that specific school. 10 of these are categorical variables; respondents answer with 'yes' or 'no', with regard to whether parental or community engagement was executed using these specific means.[4] 2 of these are ordinal variables, in which

[3] The variables are: a) require visitors to sign or check in and wear badges (*visitor_badge*), b) control access to school buildings during school hours (e.g., locked or monitored doors, loading docks) (*control_building*) (c) control access to school grounds during school hours (e.g., locked or monitored gates) (*control_ground*), d) require metal detector checks on students every day (*require_check*), e) perform one or more random metal detector checks on students (*random_check*), f) equip classrooms with locks so that doors can be locked from the inside (*lock_inside)*, g) close the campus for most or all students during lunch (*close_lunch*), h) perform one or more random sweeps (e.g., locker checks, dog sniffs) for contraband (*random_sweep*), i) require drug testing for students participating in athletics or other extracurricular activities (*drug_testing*), j) require students to wear uniforms (*wear_uniform*), k) enforce a strict dress code (*dress_code*), l) provide school lockers to students (*school_lockers*), m) require clear book bags or ban book bags on school grounds (*book_bags*), n) have panic button(s) or silent alarm(s) that directly connect to law enforcement in the event of an incident (*panic_button*), o) provide an electronic notification system that automatically notified parents in case of a school-wide emergency (*emergency_notif*), p) provide a structured anonymous threat reporting system (*anonymous_report*), q) require students to wear badges or picture IDs (*student_id*), r) require faculty and staff to wear badges or picture IDs (*staff_id*), s) use one or more security cameras to monitor the school (*security_camera*), t) provide two-way radios to any staff (*radio*), u) prohibit non-academic use of cell phones or smartphones during school hours (*prohibit_phone*).

[4] The variables are: a) whether there is a formal process to seek parental input on school safety and student disciplinary matters (*disciplinary_process*), b) whether there is parental training or assistance with regard to students' problem behaviours (*parent_training_assist*), and whether c) parent groups (*parenthelp_drug*), d) social service agencies (*socialshelp_drug*), e) juvenile justice agencies (*juvhelp_drug*), f) law enforcement agencies (*lawhelp_drug*), g) mental health agencies (*mhhelp_drug*), h) civic organizations (*civichelp_drug*), i) private

the respondent provides some estimate (i.e., *0-25%*, *26-50%*, *61-75%*, *76-100%* or *does not offer*) of the proportion of students whose parents were engaged in either an open house event (*open_house*) or in a routinely scheduled parent-teacher conference (*parent_teacher_conf*).

This dataset also contains features regarding educational initiatives or approaches taken by schools, in relation to steps taken to ensure that students receive programs that are intended to alleviate violent incidents. This consists of 8 categorical variables; respondents answer with ‘yes’ or ‘no’, with regard to whether these approaches was executed using these specific programs or initiatives[5].

There are further demographic control variables in the dataset, which are of interest. There are variables on the crime level in the area where the students live (*crimes_students_residence*) and the crime level in the area in which the school is located (*crimes_school_area*). These are ordinal variables that share three levels: *high*, *medium*, and *low* crime. *crimes_students_residence* has an additional option for *mixed* levels of crime. These variables could also explain school safety, since crime from these areas could also relate to school safety within schools. Additionally, there is a variable for the geographic location of the school. This variable had four levels: *city, suburb*, *town* and *rural*; it was recoded into a binary indicator variable *urban*, such that rural schools would be accorded a value of 0, whereas if the

corporations (*privatehelp_drug*) or j) religious organizations (*relighelp_drug*) were involved in schools’ efforts to promote safe, disciplined and drug-free schools.

[5] The variables are: a) whether schools implemented some form of violence or harm prevention curriculum or training for students (e.g., conflict resolution) (*prevention_curriculum*), b) whether schools implemented some form of socio-emotional learning for students (e.g., social skills or anger management) (*sel*), c) whether schools used variants of behavioral or behavior modification intervention for students (e.g., the use of positive reinforcements) (*behav_mod*), d) individual mentoring or coaching of students by adults (*mentoring*), e) whether peer students were involved in peer mediation (*peer_mediation*), f) whether student peers were involved in student courts, to address student conduct issues or minor offences (*student_court*), g) whether student peers were involved in restorative circles (e.g., ‘peace circles’) (*restorative_circles*) or h) whether there were programs that were intended to socially integrate fellow students, and provide a sense of community (*promote_community*).

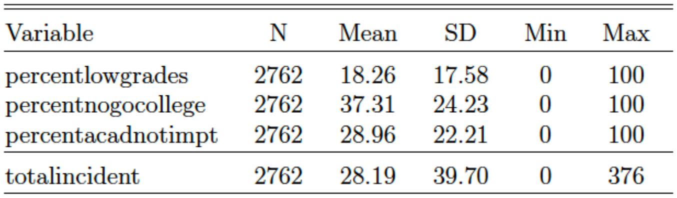
school had an was located in a city, suburb or town, it would be accorded a value of 1, to indicate that the school was located in an urban area.

Two more variables, *school_size* and *school_type,* were also included. These were variables that accounted for the size of student enrollment within a given school, and the grade level for a given school. *school_size* had four levels (i.e., *<300*, *300-499*, *500-999*, *1000+*), while *school_type* had four levels (i.e., *Primary, Middle, High, Combined*).

In total, there are 49 independent variables of interest for this study. With regard to *school_type*, there are 671 elementary schools, 1094 middle or combined schools and 997 high schools that constitute the sample of 2762 schools. With regard to *school_size*, in this dataset of 2762 schools, there are 286 schools with less than 300 students enrolled, 605 schools with 300-499 students enrolled, 1042 schools with 500-999 students enrolled, and 829 schools with over 1000 students enrolled. Descriptive statistics of the other 47 relevant independent variables, and the dependent variable, *total_incident*, are presented in Tables 1 to 6.

Table 1

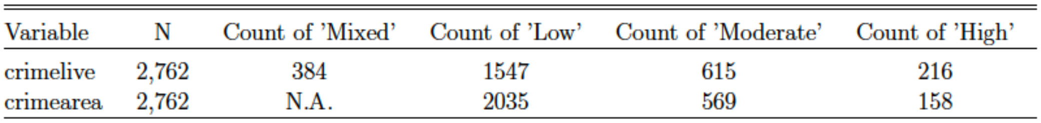
Descriptive Statistics of Socio-Economic and School Safety Variables



Variable	N	Mean	SD	Min	Max
percentlowgrades	2762	18.26	17.58	0	100
percentnogocollege	2762	37.31	24.23	0	100
percentacadnotimpt	2762	28.96	22.21	0	100
totalincident	2762	28.19	39.70	0	376

Table 2

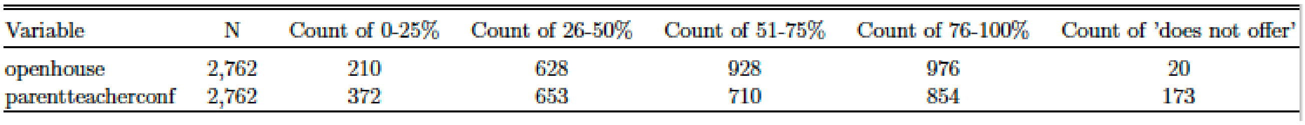
Descriptive Statistics of Neighborhood Crime Control Variables



Variable	N	Count of ’Mixed’	Count of ’Low’	Count of ’Moderate’	Count of ’High’
crimelive	2,762	384	1547	615	216
crimearea	2,762	N.A.	2035	569	158

Table 3

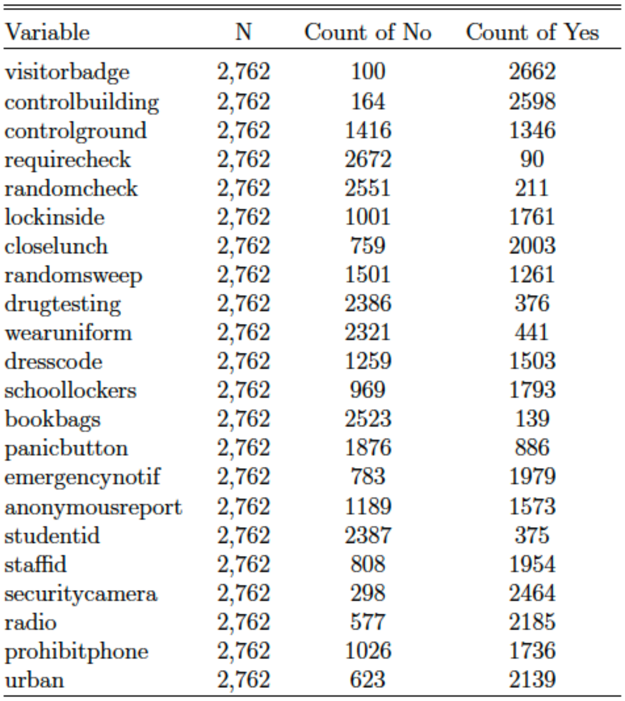
Descriptive Statistics of Parental and Community Involvement Ordinal Variables



Variable	N	Count of 0-25%	Count of 26-50%	Count of 51-75%	Count of 76-100%	Count of 'does not offer'
openhouse	2,762	210	628	928	976	20
parentteacherconf	2,762	372	653	710	854	173

Table 4

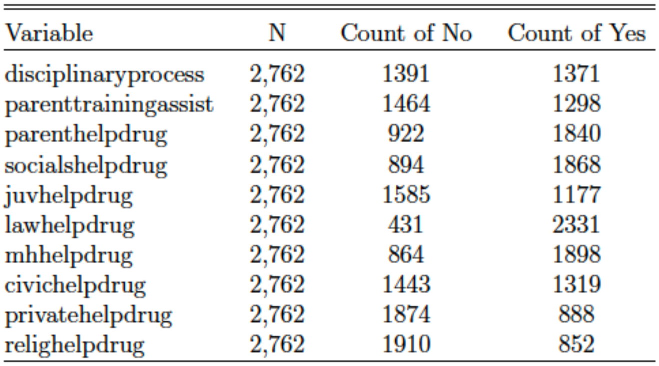
Descriptive Statistics of Security/Environmental Variables and Urbanicity Control Variable



Variable	N	Count of No	Count of Yes
visitorbadge	2,762	100	2662
controlbuilding	2,762	164	2598
controlground	2,762	1416	1346
requirecheck	2,762	2672	90
randomcheck	2,762	2551	211
lockinside	2,762	1001	1761
closelunch	2,762	759	2003
randomsweep	2,762	1501	1261
drugtesting	2,762	2386	376
wearuniform	2,762	2321	441
dresscode	2,762	1259	1503
schoollockers	2,762	969	1793
bookbags	2,762	2523	139
panicbutton	2,762	1876	886
emergencynotif	2,762	783	1979
anonymousreport	2,762	1189	1573
studentid	2,762	2387	375
staffid	2,762	808	1954
securitycamera	2,762	298	2464
radio	2,762	577	2185
prohibitphone	2,762	1026	1736
urban	2,762	623	2139

Table 5

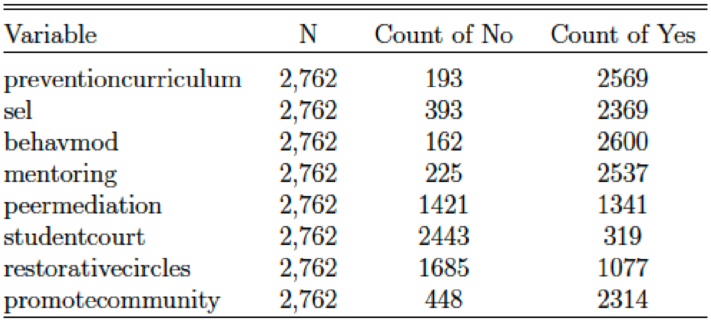
Descriptive Statistics of Parental and Community Involvement Binary Variables



Variable	N	Count of No	Count of Yes
disciplinaryprocess	2,762	1391	1371
parenttrainingassist	2,762	1464	1298
parenthelpdrug	2,762	922	1840
socialshelpdrug	2,762	894	1868
juvhelpdrug	2,762	1585	1177
lawhelpdrug	2,762	431	2331
mhhelpdrug	2,762	864	1898
civichelpdrug	2,762	1443	1319
privatehelpdrug	2,762	1874	888
relighelpdrug	2,762	1910	852

Table 6

Descriptive Statistics of Educational Approaches Variables



Variable	N	Count of No	Count of Yes
preventioncurriculum	2,762	193	2569
sel	2,762	393	2369
behavmod	2,762	162	2600
mentoring	2,762	225	2537
peermediation	2,762	1421	1341
studentcourt	2,762	2443	319
restorativecircles	2,762	1685	1077
promotecommunity	2,762	448	2314

Illustrative Example of Conventional Social Scientific Linear Regression Approach

In this section, I flesh out an illustrative example of a typical social scientific modelling approach, utilizing linear regression based modelling, to derive some findings based on the conventional, well-developed approaches that have been employed by social scientists. This is meant to both provide some inferential indications of the multivariate linear relationships between some key independent variables with the outcome variable, which will also feature in the explainable machine learning model, as well as illustrate the value of a complementary explainable machine learning approach.

In line with a theory-based approach to modelling, linear regressions are sequentially built to test for exclusion restrictions, in order to determine whether groups of added variables have no effect on the dependent variable (Wooldridge, 2013, p.143). Specifically, this is achieved by partial effects *F*-test(s) that compare a restricted model, with a subset of explanatory variables from the unrestricted model, against an unrestricted model. This procedure tests for the null hypothesis that the variables that differ between the unrestricted and restricted models explain no additional variance in the dependent variable, and hence do not improve the overall model fit when included in the model.

The literature review has surfaced a few groups of explanatory variables for consideration, in modelling school safety: variables relating to student disadvantage, parental involvement and community involvement variables, environment variables that measure the security features or enforcement present in the school and educational approaches taken by the school. A small subset of these features are considered in this illustrative example. The demographic control variables for school type, school size, neighbourhood crime, grade level and urbanicity were also involved.

Methods

Feature Selection and Engineering. Specified variables were recoded, in order to aid in their interpretation. Initially, *crimes_students_residence* had four levels, *high*, *medium*, *low* and *mixed* levels of crime. *Mixed* was recoded into the same category as *medium* levels of crime, for ease of interpretation. *School_type* initially had four levels, *elementary*, *middle*, *combined*, and *high* schools, with *combined* schools being schools that could either be elementary-middle combined schools or middle-high combined schools; they were recoded such that *combined* schools were in the same category as *middle* schools.

All other variables involved, across all variable types were recoded; binary categorical variables were recoded so that the presence of a given variable was denoted by a higher value (*1 = presence*, *0 = absence*), while ordinal and continuous/ratio variables were recoded to ensure that higher levels of the variable were measures for higher values of the construct.

Given the visible high degree of collinearity between *percentnogocollege* and *percentacadnotimpt* in Figure 1, a formal Pearson's correlation coefficient test was conducted. Indeed, there was a statistically significant and strong linear relationship between these ratio variables ($r = 0.69$, $p<.001$); the statistical significance of this value is sufficient evidence at the 1% significance level to reject the null hypothesis that they are not correlated at all. The inclusion of both variables is likely to cause or worsen multicollinearity problems for the model. Hence, between these two variables, only *percentacadnotimpt* was considered going forward, since *percentnogocollege* and *percentacadnotimpt* not only measure similar constructs (i.e., someone who does not consider academic results to be important would logically also be less likely to pursue higher education). There was no similar problem between *percentlowgrades* and *percentacadnotimpt* ($r = 0.31$, $p<.001$), or between *percentlowgrades* and *percentnogocollege* (*r*

= 0.36, *p*<.001); these correlations were statistically significant, which means that there is sufficient evidence at the 1% significance level to reject the null hypothesis that they are not correlated at all. However, the correlation coefficients between *percentlowgrades* and *percentacadnotimpt* and between *percentlowgrades* and *percentnogocollege* were substantially lower than the correlation between *percentnogocollege* and *percentacadnotimpt*. This indicates that *percentlowgrades* likely captures a slightly different facet of student disadvantage.

Finally, given the univariate distribution visualizations in Figure 1 of the ratio variables of interest, which now constitute *percentlowgrades*, *percentacadnotimpt* and the dependent variable, *total_incidents*, it is noticeable that all these variables are visibly skewed, and do not approximate a normal distribution. At the same time, for each of these variables, the minimum value of these variables is zero (see Table 1), which represent schools that are extremely high achieving for *percentlowgrades*, *percentacadnotimpt*, or are extremely safe for *total_incidents*, which make a plain logarithmic transformation of these variables impossible as *log*(0) is undefined. Hence, a *log*(1+*x*) transformation was applied to these variables, where *x* represents the initial value of the variable before the transformation; such a transformation ensures that all transformed values are either zero (since *log*(1) =0) or a value greater than zero, and to ensure that the univariate distribution of the variables more closely approximates a normal distribution.

Base Model. The main variables of interest in this study, markers of student disadvantage, were considered along with the *school_size* control variable, which was considered by a number of previous studies. The base model is represented by this equation:

$$y_{totalincident} = \beta 1 X_{\text{percentlowgrades}} + \beta 2 X_{\text{percentacadnotimpt}} + \beta 3 X_{\text{school\_size}} + \alpha 1$$

Adding Control Variables. During the second iteration, key control variables were added to the base model, based on the literature review, including variables which approximate the crime levels around students' residences (*crimes_students_residence*) and around the school (*crimes_school_area*). The *school_type* variable was also added; as this was a categorical variable with three levels, the level of the variable with the lowest value, *elementary*, was assigned as the reference category, with dummy variables for *high* schools and *middle_combined* schools included in the model. The model is represented by this equation:

$$y_{totalincident} = \beta 1 X_{\text{percentlowgrades}} + \beta 2 X_{\text{percentacadnotimpt}} + \beta 3 X_{\text{school\_size}} + \beta 4 X_{\text{crimes\_schools\_area}} + \beta 5 X_{\text{crimes\_students\_residence}} + \beta 6 Z 1 X_{\text{school\_type=High\_School}} + \beta 7 Z 2 X_{\text{school\_type=Middle\_Combined\_School}} + \alpha 2$$

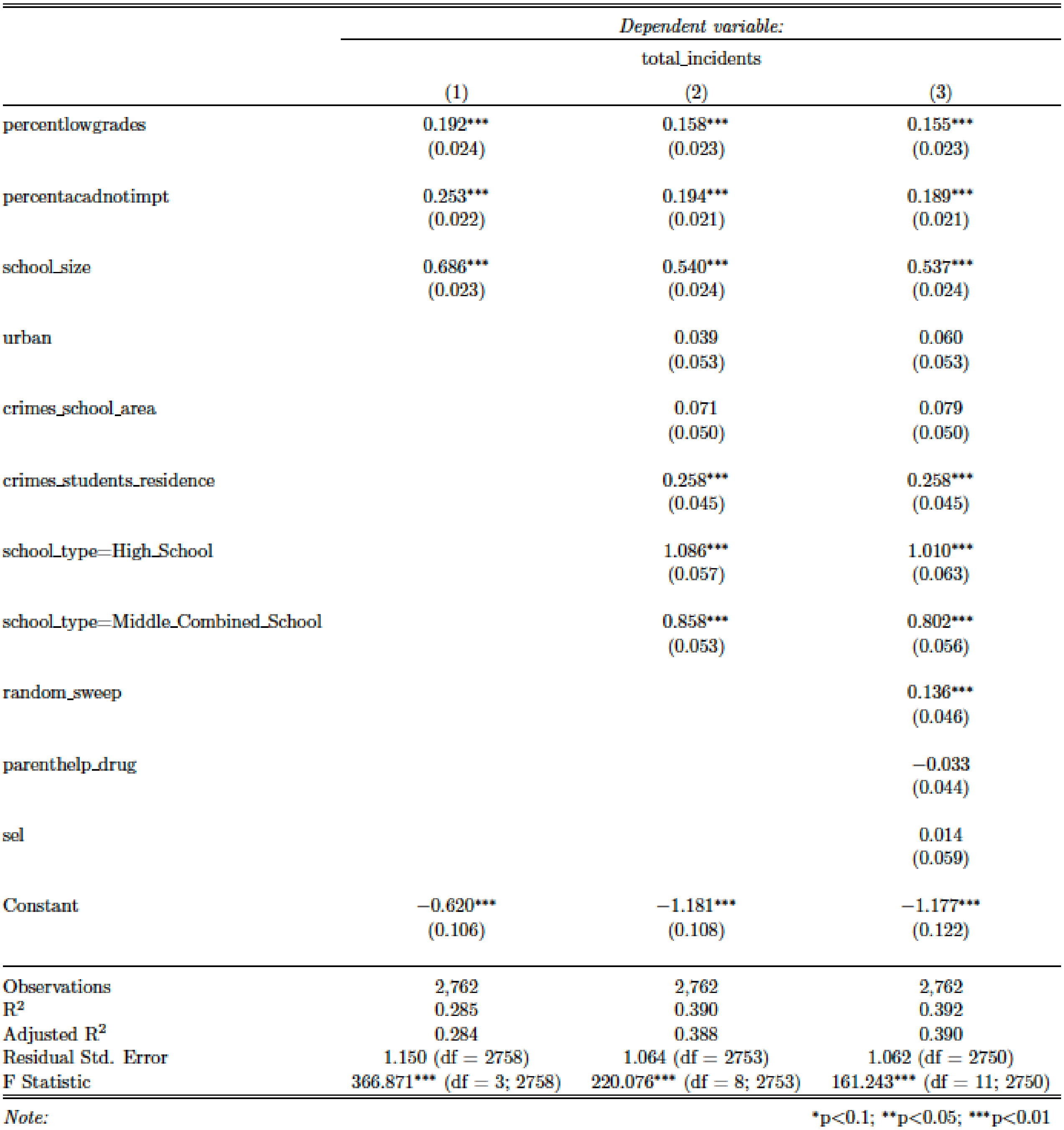
Adding Additional Explanatory Variables. At the final iteration of this illustrative example, additional explanatory variables were added, based on the literature review. *random_sweep* is one of the 21 categorical variables which indicate the high-security features put in place by the school. *parenthelp_drug* is one of the five variables which approximate the level of parental involvement in a school. *sel* is one of the eight variables that measure the educational approaches and climate within a school, capturing the presence or absence of socio-emotional learning. All of these variables are binary categorical variables. The model with control variables as well as explanatory variables considered is represented by this equation:

$$y_{totalincident} = \beta 1 X_{\text{percentlowgrades}} + \beta 2 X_{\text{percentacadnotimpt}} + \beta 3 X_{\text{school\_size}} + \beta 4 X_{\text{crimes\_schools\_area}} + \beta 5 X_{\text{crimes\_students\_residence}} + \beta 6 Z 1 X_{\text{school\_type=High\_School}} + \beta 7 Z 2 X_{\text{school\_type=Middle\_Combined\_School}} + \beta 8 Z 3 X_{\text{random\_sweep}} + \beta 8 Z 3 X_{\text{parenthelp\_drug}} \beta 8 Z 3 X_{\text{sel}} + \alpha 3$$

Results and Discussion

Table 7

Results of Illustrative Linear Regression Models



	Dependent variable:		
	total_incidents		
	(1)	(2)	(3)
percentlowgrades	0.192***	0.158***	0.155***
	(0.024)	(0.023)	(0.023)
percentacadnotimpt	0.253***	0.194***	0.189***
	(0.022)	(0.021)	(0.021)
school_size	0.686***	0.540***	0.537***
	(0.023)	(0.024)	(0.024)
urban		0.039	0.060
		(0.053)	(0.053)
crimes_school_area		0.071	0.079
		(0.050)	(0.050)
crimes_students_residence		0.258***	0.258***
		(0.045)	(0.045)
school_type=High_School		1.086***	1.010***
		(0.057)	(0.063)
school_type=Middle_Combined_School		0.858***	0.802***
		(0.053)	(0.056)
random_sweep			0.136***
			(0.046)
parenthelp_drug			−0.033
			(0.044)
sel			0.014
			(0.059)
Constant	−0.620***	−1.181***	−1.177***
	(0.106)	(0.108)	(0.122)
Observations	2,762	2,762	2,762
R^2	0.285	0.390	0.392
Adjusted R^2	0.284	0.388	0.390
Residual Std. Error	1.150 (df = 2758)	1.064 (df = 2753)	1.062 (df = 2750)
F Statistic	366.871*** (df = 3; 2758)	220.076*** (df = 8; 2753)	161.243*** (df = 11; 2750)
Note:			*p<0.1; **p<0.05; ***p<0.01

Partial *F*-Tests. Importantly, it should be noted that Model (1) is nested within Model (2), which in turn is nested in Model (3). This makes it possible to test for exclusion restrictions, between the three models. A partial *F*-test was hence conducted between the three models, which revealed that the inclusion of the control variables in Model (2) significantly increased model fit from Model (1) (*p*<.001) at the 1% significance level, and that the inclusion of some additional, theoretically-relevant explanatory variables in Model (3) significantly increased model fit from Model (2) (*p*=0.03) at the 5% significance level. There is support for the inclusion of the groups of variables at each iteration.

Interpreting Results. Based on this skeletal model, with just 11 variables, some key main effects have emerged, as shown in Table 7. Notably, the coefficients of both *percentlowgrades* and *percentacadimpt* were statistically significant at the 1% significance level, which means it is very unlikely that these variables have no relationship with the outcome variable of interest. The coefficients of *percentlowgrades* and *percentacadimpt* are 0.155 and 0.189 in Model (3); as these variables and the dependent variable are *log*(1+*x*) transformed variables, they are interpreted as such: for schools that are 1% higher on *percentlowgrades* or *percentacadimpt*, we can expect a 0.155% higher value of *total_incidents* or 0.189% higher value of *total_incidents* respectively, on average and net of all other variables in Model (3). This is broadly suggestive evidence in support of the expectations in this study, in that schools with a higher proportion of these disadvantaged students would, on average, experience more incidents and thus lower levels of school safety. The coefficient values of these variables, in Model (3), were noticeably lower than either Model (2) or Model (1), which indicates that some of the variables added in Models (2) and (3) might mediate between *percentlowgrades* and *total_incidents* or between *percentacadimpt* and *total_incidents.* Notably, two control variables, *school_size* and *crimes_students_residence*, are ordinal variables with coefficients of 0.537 and 0.258 in Model (3), that are both statistically significant at the 1% significance level; their coefficients in Model (2) were similar, and were also statistically significant at the 1% significance level. This means that for each one-unit increase in a level of either *school_size* (e.g., from *<300* students to *300-499* students) or *crimes_students_residence* (e.g., from *low_crime* to *medium_crime*), there is on average, and net of all other variables, a 53.7% and 25.8% increase in *total_incident* respectively. Two control variables, *urban* and *crimes_school_area*, do not have statistically significant coefficients; neither are two explanatory variables, *parenthelp_drug* and *sel*. Another control variable, *school_type*, was dummy coded into two variables, one for *school_type=High_School*, one for *school_type=Middle_Combined _School* which have coefficients of 1.010, and 0.802 respectively. These are interpreted with respect to the reference category, *school_type=Elementary_School*, which is not included within the regression to prevent perfect collinearity between the three indicator variables; on average and net of all other variables, high schools have 101% more *total_incident* compared to elementary schools, while middle and combined schools have 80.2% more *total_incident* compared to elementary schools. Finally, the coefficient of *random_sweep* is 0.136, and is statistically significant at the 1% significance level; as this is a binary categorical variable for the presence or absence of random sweeps in schools, this coefficient means that schools which implement random sweeps, on average and net of all other variables, have 13.6% more *total_incident* than schools that do not.

Model Diagnostics. A Breusch-Pagan test was conducted on Model (3), revealing that there was heteroscedasticity in the model (*p*<.001).

Evaluation of this Approach. The illustrative example has shown the typical social scientific modelling workflow, which makes some points of value apparent. This approach remains a choice approach for social scientists because of its interpretability, and how beta coefficients can be used to garner insights; in this base model, there is early evidence that there is variation in school safety due to variation in levels of student disadvantage and variation according to whether schools implement the high-security environmental feature, *random_sweep*.

However, there are also additional problems which could be addressed with alternative approaches. First, there is heteroscedasticity in the linear regression model. This presents a problem for the interpretation of the standard errors in the model, and the subsequent interpretation of significance values, because this indicates that the error variances are not normally distributed. While this part of the problem could be addressed through the application of robust standard errors, it still does not address a deeper underlying issue, that a linear model may indeed not be a good fit for this specific phenomena at all. Linking this back to Breiman's (2001) framing of linear regressions as data models that capture 'actual relationships' between independent and dependent variables, there is actually no basis to claim that ordinary least squares regression lines best approximate multivariate relationships. These relationships may instead be non-linear in nature, which may not be as accurately captured by Pearson's correlation coefficients or linear regression beta coefficients. This was an insight gleaned from the scatter matrix, Figure 1, one that has persisted in spite of the fact that logarithmic transformations were applied to the skewed dependent variable *total_incidents* and the two variables *percentacadnotimpt* and *percentlowgrades*. Second, and crucially, this problem is likely to be worsened as additional variables are added to the regression model. Typically, social scientists might think about dimensionality reduction or factor analytic approaches at this point, in order to reduce the number of variables that need to be considered in the models. In the context of a large dataset with a mix of variable types, as the present SSOCS data is, typical dimensionality reduction approaches like principal component analyses (i.e., which can only be applied to continuous data) cannot be used to overcome the problems of multicollinearity. Neither can factor analytic approaches, as the application of these approaches assume that all variables are continuous in nature.

Explainable Machine Learning Approach

The gaps raised above motivate the primary approach used in this section, utilizing explainable machine learning tools for social scientific insights. Specifically, the random forest regressor, chosen for its non-linearity and effectiveness, was used in conjunction with SHAP values to gain insights on how given features relate to school safety. A recursive feature elimination approach is used to partially address collinearity concerns between variables.

Methods

Feature Preprocessing and Engineering. The preprocessing steps employed in the illustrative example are also implemented in this segment. The only exception was the *school_type* variable, which was dummy coded in the illustrative example, but for the purposes of feeding it into the random forest regressor as a single variable, was label encoded instead (*1 = elementary, 2 = middle_combined, 3 = high*) into three levels of *school_type.*

In this segment, 48 independent variables were of interest[6], along with one dependent variable. *total_incidents*, as a proxy for school safety, remains the dependent variable of interest (i.e., y). All other explanatory features/variables of interest were incorporated into a data frame (i.e., X). The data frame was then subject to a train test split into *X_train, X_test, y_train, y_test* sets, with the test set consisting of 20% of the dataset and the train set consisting of 80% of the dataset. A random state of 42 was used for the replicability of this analysis.

Base Model. A base model was established, in order to set a baseline comparison against the models formulated during the model experimentation process. An ensemble tree model, namely an untuned bagged trees regressor native to scikit-learn version 0.22.2, was used (Pedregosa et al., 2011). This base model had 100 *n_estimators* (i.e., constituent decision trees

[6] For the same reason as in the illustrative example above (i.e., a strong and significant linear relationship between *percentnogocollege* and *percentacadnotimpt*), *percentnogocollege* was also not considered in this segment.

within the bagged trees regressor), with the *min_samples_leaf* (i.e., proportion of the training data required at each leaf node of a decision tree) and *min_samples_split* (i.e., proportion of the training data required to split a node of a decision tree) parameters set to their default values of 1 and 2 respectively. A random state of 42 was also set for the replicability of this analysis.

Hyperparameter Tuning. This process was meant to optimize and tune the base model, based on the training data, *X_train* and *y_train*. Two cross validation (CV) tools were used, RandomizedSearchCV and GridSearchCV, from the *model_selection* module of scikit-learn (Pedregosa et al., 2011).

First, the RandomizedSearchCV function was used to define and narrow the search space, with cross-validation conducted on models with randomly selected sets of parameter values. The following candidate parameters were used for the RandomizedSearchCV: for *n_estimators*, 10 values were entered for consideration, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000; for *max_depth* (i.e., the maximum number of levels in the decision tree) 11 values were entered for consideration, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110; for *min_samples_split*, three values were entered for consideration, 2, 5, 10; for *min_samples_leaf*, three values were entered for consideration, 1, 2 or 4; for *bootstrap* (i.e., whether the selected samples for training each tree are bootstrapped), the boolean values of *True* or *False* were entered for consideration. It is important to note that these values are not independent of one another. For instance, *max_depth* is related to *min_samples_split*, since the lower the threshold for resulting a split in a decision tree, the deeper each decision tree can go, resulting in a random forest model with more parameters overall. This parameter grid was fit to a random forest regressor (i.e., *max_features* as the square root of the total number of features) with a random state of 42 for replicability; this was done with 5-fold cross validation, over 100 iterations. This means that 100 random combinations of features from this given parameter grid were tested. This process produced a model with 800 *n_estimators*, 2 for *min_samples_split*, 2 for *min_samples_leaf*, max_depth of 30 and *bootstrap* as *False* as the best performing set of model parameters.

With these values to inform and narrow the potential search space, the GridSearchCV function was used to perform grid search cross validation on the values generated by the RandomizedSearchCV process, as an additional step to ensure that a well validated model was used. GridSearchCV differs from RandomizedSearchCV because all possible permutations of parameters, given the parameter grid introduced, are assessed and subject to cross-validation. The following parameters were used for the RandomizedSearchCV: for *n_estimators*, 4 values were entered for consideration, 700, 800, 900 or 1000; for *max_depth* (i.e., the maximum number of levels in the decision tree) 3 values were entered for consideration, 30, 40, 50; for *min_samples_split*, three values were entered for consideration, 2, 3,4; for *min_samples_leaf*, three values were entered for consideration, 2, 3, 4. The exact same best parameters were found from this GridSearchCV process, namely 800 *n_estimators*, 2 for *min_samples_split*, 2 for *min_samples_leaf*, max_depth of 30 and *bootstrap* as *False.*

Model Evaluation. This tuned random forest regressor was then evaluated on held-out test data, *X_test* and *y_test*. This procedure evaluates the performance of the tuned random forest regressor model on unseen test data that was not present during the training of the model. This provides an indication of how well this model can be generalized to new data. This tuned model had a root mean squared error (RMSE) value of 1.042 when evaluated on the held-out test set. This was compared against the base model, which had a RMSE value of 1.055 when evaluated on held-out test data. This validated that the tuned random forest regressor did indeed improve model predictions, making it our best model from this model experimentation process.

Recursive Feature Elimination with Permutation Feature Importance. A key problem with regard to calculation of feature importance metrics is the problem of correlated features (Toloşi and Lengauer, 2011). Correlated clusters of variables, with shared variance, would lower the estimated importance of individual variables that constitute such clusters, no matter what feature importance metrics are used. It is hence important to take this into account, and address this problem. A metric of feature importance, permutation feature importance, was used to remove variables which contribute least of the variance to the prediction. This is a global measure of how important a given feature is, by quantifying the decrease in a model's cross-validated performance when this feature is randomly permuted. At the same time, there remain concerns of collinearity between groups of variables through the use of permutation feature importance, and it is likely that in a single stage recursive feature elimination process, these groups of related variables might be removed at the same time, since it follows that correlated features would share similar variance with the dependent variable. Hence, recursive feature elimination was conducted with multiple stages on the optimized random forest regressor trained on *X_train and y_train*, with only one variable with the lowest permutation importance removed at each stage, to ensure that variables which might be related to the removed variable were not also removed at the same time. This was conducted sequentially until half the original variables remained, namely: *'percentlowgrades', 'percentacadnotimpt, 'school_size', 'urban', 'crimes_students_residence', 'crimes_school_area', 'school_type', 'control_ground', 'require_check', 'random_sweep', 'wear_uniform', 'dress_code', 'anonymous_report', 'student_id', 'prohibit_phone', 'open_house', 'parent_teacher_conf', 'disciplinary_process', 'juvhelp_drug', 'lawhelp_drug', 'mhhelp_drug', 'privatehelp_drug', 'relighelp_drug', 'peer_mediation'*. Given this set of 24 features, held-out evaluation of the held-out test data, *X_test and y_test* was once again conducted in order to check whether features that were critically associated with *total_incident* were removed. The RMSE of this restricted model when evaluated on the test set, with 24 instead of 48 features, was 1.044, a marginal increase in RMSE when compared to the RMSE of the unrestricted model with 48 features, which was 1.042.

Final Model. Given that the previously tuned model was already the best cross-validated model on the *X_train* and *y_train* data from the RandomizedSearchCV and GridSearchCV procedures, and performed better than a baseline model when evaluated on the held-out *X_test and y_test* data, these indicate that the tuned hyperparameters for the random forest regressor were suitable for the entire dataset. I took reference from Joshi and Joshi (2019), a paper that also used SHAP for social science, which collapsed across train and test sets to calculate SHAP values on the entire dataset. In the same vein, I removed the train test split segmentation in the data for the present study. Then, the random forest regressor, with the best hyperparameters obtained through cross-validation, was refitted on the entire dataset with 2762 observations, with 24 variables that were obtained through the recursive feature elimination procedure.

Calculation of SHAP Values. SHAP values were calculated from the final model, utilizing the Kernel SHAP approximation method of the SHAP package in Python (Lundberg & Lee, 2017). SHAP values are based on a game-theoretic method of calculating feature importances, with each constituent feature in a model treated as a 'player'. In a non-linear model like the random forest regressor, SHAP values are calculated based on the average marginal contribution of each constituent feature to output predictions, based on all possible feature orderings (Lundberg & Lee, 2017, p.5). These values represent the unique contribution of how each original value of the independent variable relates to the outcome variable, vis a vis a base value prediction, which refers to the average prediction of the random forest regressor given the inputs and output predictions. The base value prediction has a SHAP value of 0. Each variable value which contributes to a higher prediction of the dependent variable, *total_incident*, is represented with a positive SHAP value, with higher SHAP values referring to higher numbers of *total_incident*, and hence lower school safety. Conversely, each variable value which contributes to a lower prediction of the dependent variable, *total_incident*, is represented with a negative SHAP value, with higher SHAP values referring to lower numbers of *total_incident*, and hence higher school safety. In other words, each of the 2762 observations have 24 SHAP values, each representing one independent variable in the model; 66288 SHAP values were calculated for all observations, across all features. The base value prediction of the random forest regressor was a value of 2.65 for *total_incidents*[7], which refers to a value of 13.2 actual school safety incidents[8]. Each of the 66288 SHAP values can be directly added or subtracted from the base value of 2.65, to determine how each observation's variable values contribute to output values of *total_incident*, providing an explanation for how each output value of *total_incident* in the dataset was derived.

Results and Discussion

The results for the SHAP output will be discussed under two headers: *magnitude* (i.e., which independent variables are most important in their association with the dependent variable), and *direction* (i.e., how do different values of the independent variable relate to the dependent variable). These mirror some statistical properties that make linear regression beta coefficients highly desirable and interpretable.

[7] *total_incident* was subject to a log(1+*x*) transformation, and should not be directly interpreted as 2.65 school safety incidents.

[8] Given log(1+*x*) = 2.65, where *x* is the original value of the number of school safety incidents, *x* = *exp*(2.65) - 1 = 13.2.

Magnitude. As established in the section above, variables' SHAP values can be positive or negative. To understand which variables have the *greatest* impact on *total_incident*, regardless of polarity, the absolute value of all 66288 SHAP were generated, and averaged. A subset of 10 variables with the highest mean absolute SHAP values are presented in Figure 2.

As validation that these variables were indeed the variables which contributed the most of the variance to the output variable, the random forest regressor model (i.e., with the best parameters generated through the hyperparameter tuning process above) was fit only to a subset of the data with these 10 variables. The RMSE of this model with 10 variables, when validated on held-out test data, was 1.076, which was slightly higher than the model trained on 24 variables which had a RMSE of 1.044.

The descriptive statistics for the variables with the top 10 highest absolute SHAP values for the random forest regressor trained on 24 variables obtained through recursive feature elimination, are presented in Table 8. Figure 2 further visualizes the top 10 variables with the highest mean absolute SHAP values, for the random forest regressor trained on 24 variables obtained through recursive feature elimination.

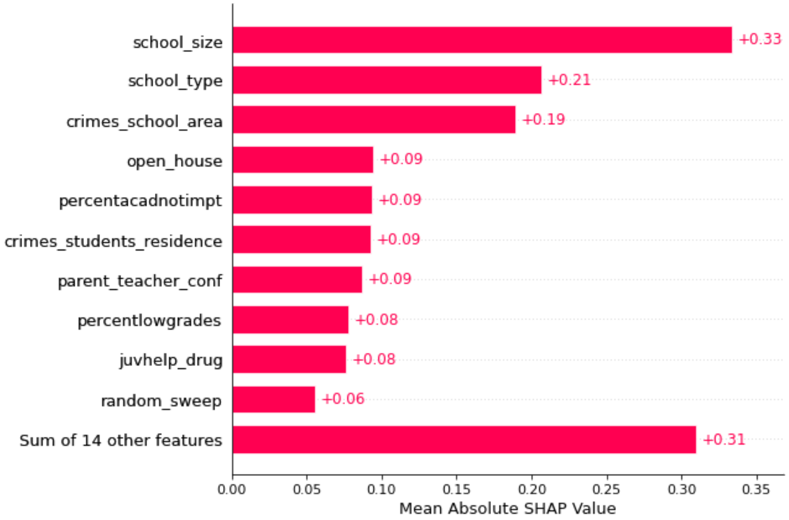
Table 8

Descriptive Statistics of SHAP Values for Top 10 Variables with Highest Mean Absolute SHAP Values

Variable	N	Mean	SD	Min	Median	Max	Mean Absolute
schoolsize	2,762	-0.0335	0.389	-0.947	-0.122	0.755	0.334
schooltype	2,762	-0.0473	0.278	-0.895	0.0802	0.256	0.206
crimesschoolarea	2,762	0.0127	0.236	-0.653	0.0185	0.904	0.189
openhouse	2,762	0.00517	0.113	-0.360	0.0286	0.309	0.0945
percentacadnotimpt	2,762	0.0152	0.117	-0.634	0.0464	0.360	0.0933
crimesstudentsresidence	2,762	0.0119	0.107	-0.201	-0.0366	0.408	0.0928
parentteacherconf	2,762	0.00420	0.102	-0.265	0.00910	0.327	0.0868
percentlowgrades	2,762	0.00154	0.106	-0.512	0.00422	0.345	0.0776
juvhelpdrug	2,762	-0.0147	0.0848	-0.237	-0.0276	0.210	0.0759
randomsweep	2,762	-0.00801	0.0659	–0.203	-0.00392	0.193	0.0554

Figure 2

Top 10 Variables with Highest Mean Absolute SHAP Values



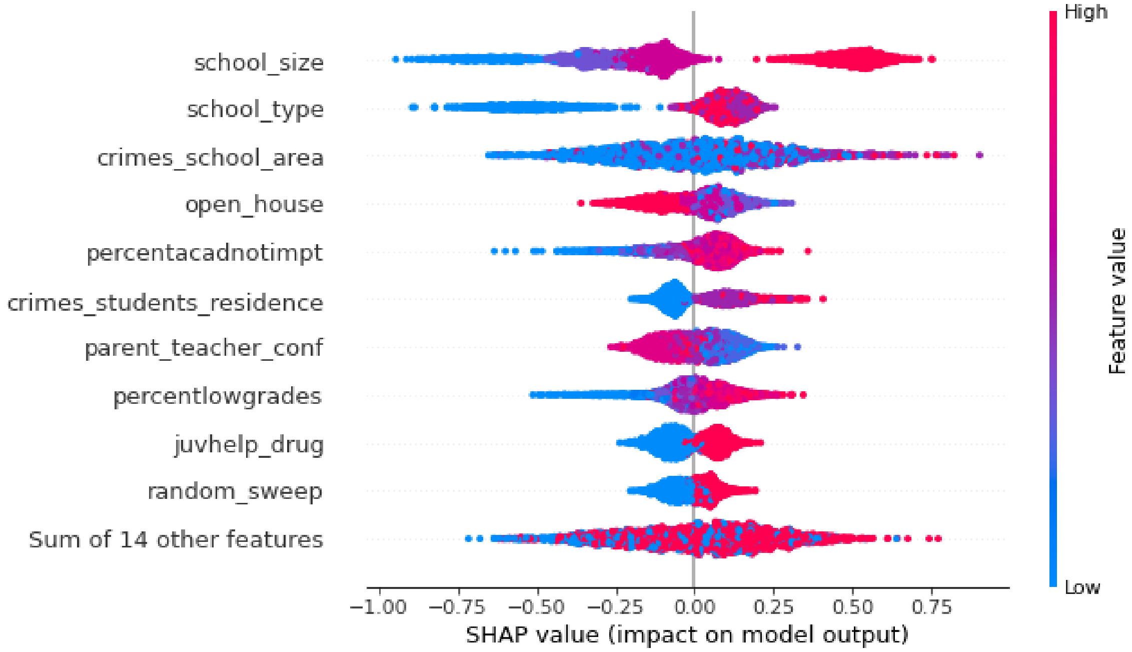
From Figure 2 and Table 8, it is evident that a number of control variables explain a large amount of the variance in the data, *school_size*, *school_type*, *crimes_school_area* and *crimes_students_residence*, which have mean absolute SHAP values of around 0.334, 0.206, 0.189 and 0.0928 respectively. These control variables aside, parental involvement variables *open_house* and *parent_teacher_conf*, along with markers of student disadvantage *percentacadnotimpt* and *percentlowgrades*, have mean absolute SHAP values of around 0.0945, 0.0868, 0.0933 and 0.0776 respectively. One binary variable for community involvement, *juvhelp_drug*, and one binary variable for security measures, *random_sweep*, form the last of these top 10 variables, with mean absolute SHAP values of 0.0759 and 0.0554 respectively. The interpretation of these quantities is key. One can understand the mean absolute SHAP value for each variable as the average net change that the variable contributes to the output value of *total_incident* for this random forest regressor. The mean absolute SHAP value of *percentacadnotimpt is* around 0.0933. This means that, on average, *percentacadnotimpt* changes the value of *total_incident* by 0.0933 (i.e., changes the number of school safety incidents by around 0.0978 incidents[9]), net of all other 23 variables considered in the random forest regressor. All other values can be interpreted in a similar way. While this might be useful as a first metric to filter out some variables with considerable influence on the output value, it is not an ideal metric because it does not provide further information about how the independent variables relate to *total_incident*. This metric alone cannot be used as evidence for the directional expectations that have been generated for this thesis, based on the literature.

[9] Given that log(1+x) = 0.0933, x = exp(0.0933) - 1 = 0.0978.

Direction. An alternative visualization which more comprehensively displays the full information provided by the calculation of SHAP values is displayed as Figure 3. These are the exact same ordering of 10 variables with the highest mean absolute SHAP values from the random forest regressor trained on 24 variables, as presented in Figure 2, but there are key differences between Figures 2 and 3. In Figure 3, the *actual* SHAP values are presented on the *x*-axis, and each observation is presented for each variable, colored by its *original* value (i.e., each variable is represented by 2762 observations). In other words, this provides a visual that shows how higher or lower *original* values of each independent variable is associated with variation in the values of the dependent variable, *total_incident*, for this random forest regressor. Based on Figure 3, one can see that with the exception of *crimes_school_area* and the last row, which is an amalgamation of 14 variables that have lower mean absolute SHAP values than the top 10, all other variables' SHAP values seem to vary systematically according to different levels of their original values. This facilitates the interpretation of these variables.

Figure 3

Beeswarm Plot of Top 10 Variables with Highest Mean Absolute SHAP Values



For *school_size*, *percentacadnotimpt*, *crimes_students_residence*, *percentlowgrades*, *juvhelp_drug* and *random_sweep*, it appears that higher values of the original variable values are associated with higher SHAP values, which relate to higher quantities of *total_incident*, while lower values of the original variable values are associated with lower SHAP values, which relate to lower quantities of *total_incident*; for *juvhelp_drug* and *random_sweep*, given that these are binary categorical variables, there are only two levels, the higher level being the presence of such a variable, and the lower level being the absence of such a variable. This gives insights into how each of these variables relate to school safety. Schools that have more students (i.e., higher values of *school_size*) and/or have more students who experience forms of disadvantage (i.e., higher values of *percentacadnotimpt, percentlowgrades*) and/or have more students who live in less safe neighborhoods (i.e., higher values of *crimes_students_residence*) and/or have juvenile agencies involved in drug prevention efforts (i.e., presence of *crimes_students_residence* and/or have random sweeps on students' belongings in their schools (i.e., presence of *random_sweep*), are on average, and net of all other variables considered in the regressor, less safe than schools that have lower values/levels of these variables.

For *open_house* and *parent_teacher_conf*, higher values of the original variable values are associated with lower SHAP values, which relate to lower quantities of *total_incident*, while lower values of the original variable values are associated with higher SHAP values, which relate to higher quantities of *total_incident*. In other words, schools that do implement more frequent open houses (i.e, higher values of *open_house*) or have more frequent parent teacher conferences (i.e., higher values of *parent_teacher_conf*) are, on average and net of all other variables in the regressor, more safe than schools that have lower values/levels of these variables.

For *school_type*, a different pattern from the ones described above is observed. It appears that for the second highest level of *school_type* (i.e., middle or combined schools), there are on average, and net of all other variables in the model, more school safety incidents than the highest level of *school_type* (i.e., high school). Both high schools and middle or combined schools have more school safety incidents than the lowest level of *school_type* (i.e., elementary school). *crimes_schools_area* does seem to have the least cleanly delineated pattern, among the other 9 variables described here, but from Figure 3 it still appears that schools with more crime around their areas (i.e., higher values of *crimes_schools_area*) are associated with positive SHAP values more so than negative SHAP values, and are hence less safe.

Based on these findings, it is possible to derive insights relevant to the three expectations formed for this thesis. The interpretation of the SHAP values of *percentlowgrades* and *percentacadnotimpt* is in support of Expectation 1, on the association between higher student disadvantage and lowered school safety. The interpretation of the SHAP value of *random_sweep* is in support of Expectation 2, on the association between the presence of a high-security environmental feature and lowered school safety. The interpretation of the SHAP value of *open_house* and *parent_teacher_conf* is in support of Expectation 3, on the association between higher parental involvement and higher school safety. These were obtained, in joint consideration with the key control variables that were considered in the extant literature, and with additional explanatory variables considered concerning educational approaches for students and community involvement.

General Discussion

In this thesis, an explainable machine learning approach was used to consider 48 theoretically relevant variables from the 2017 SSOCS. The results of this approach lend support to three expectations formed based on extant literature, that student disadvantage, parental involvement and high-security environmental features are key variables in association with school safety. Importantly, this approach established, with the use of SHAP values for local interpretability and an optimized non-linear random forest regressor on the dataset, that higher levels of student disadvantage and a high-security environmental feature, *random_sweep*, was linked to lower school safety, while higher parental involvement was linked to higher school safety. Magnitude was determined on a cut-off for the top 10 variables with the highest absolute mean SHAP values, and direction was determined based on the association between original variable values and the variance in SHAP values along the variance in the original variable values. These results are discussed in comparison to the results derived from the illustrative example of linear regression analyses, as a contribution to the school safety literature, and then as a contribution to social scientific methodology.

Comparison of Linear Regression and Explainable Machine Learning Results

As the proposed explainable machine learning approach is a novel approach to social scientific research, due caution is required in the interpretation of results derived from this approach. Hence, I compare between the findings of the novel explainable machine learning approach, employing SHAP values, with the illustrative linear regression example. The linear regression results reflect the insights that would be derived from a conventional social science approach; a comparison against these results provide a sense of the reliability of these findings These results are discussed in Figure 4, which summarizes the convergences and divergences in the results between the explainable machine learning and linear regression approaches taken in this thesis.

Figure 4

Comparison of Linear Regression and Explainable Machine Learning Findings

	Linear Regression Approach (Beta Coefficient and Significance Levels of Variables)	**Explainable Machine Learning Approach (Top 10 Variables with Highest SHAP Values)**	**Comparison of Results**
Student Disadvantage	Coefficients of *percentlowgrades* and *percentacadnotimpt* are statistically significant and positive (i.e., higher values of these variables are linked to more school safety incidents)	*percentlowgrades* and *percentacadnotimpt* are among the variables with the top 10 highest mean absolute SHAP values, and larger values of these variables are linked to higher SHAP values (i.e., more school safety incidents)	Convergence on direction and importance
Security/ Environmental Variables	Coefficient of *random_sweep* is statistically significant and positive (i.e., higher values of *random_sweep* are linked to more school safety incidents)	*random_sweep* is among the variables with the top 10 highest mean absolute SHAP values, and larger values of *random_sweep* are linked with higher SHAP values (i.e., more school safety incidents)	Convergence on direction and importance
Parental and Community Involvement	Coefficient of *parenthelp_drug* is not statistically significant, but it is positive (i.e., higher values of *parenthelp_drug* are linked to less school safety incidents)	*parenthelp_drug* was not among the variables with the top 10 highest mean absolute SHAP values, but *open_house* and *parent_teacher_conf* were, and larger values of these other variables are linked with smaller SHAP values (i.e., less school safety incidents)	Divergence in importance but convergence in direction for parental and community involvement variables

Educational Approaches	Coefficient of *sel* is not statistically significant, but it is positive (i.e., higher values of *sel* are linked to less school safety incidents)	*sel* was not among the variables with the top 10 highest mean absolute SHAP values, and neither were any other variables relating to Educational/Intervention Approaches	Convergence on importance
Demographic Controls	Coefficients of *school_size*, *school_type*, *crimes_students_residence* and *school_type* are significant and positive (i.e., higher values of these variables are linked to less school safety incidents); coefficient of *crimes_school_area* is not significant but is positive (i.e., higher values of *crime_school_area* are linked to less school safety incidents); coefficient of *urban* is not significant	*school_size*, *school_type*, *crimes_students_residence, school_type* and *crimes_school_area* are among the variables with the top 10 highest mean absolute SHAP values, and larger values of these variables are linked with higher SHAP values (i.e., more school safety incidents). *urban* was not among the variables with the top 10 absolute SHAP values	Divergence in importance for *crimes_school_area;* convergence in importance and direction other variables

From a summarized comparison of these results in Figure 4, it can be observed that there are convergences between the explainable machine learning and linear regression approaches. It appears that there are convergences between the findings on student disadvantage variables and security/environmental variables, and variables on education/intervention approaches. Specifically, both linear regression and explainable machine learning approaches have uncovered that the presence of high-security features, like *random_sweep*, are linked to higher levels of school safety incidents. Both approaches have also shown that higher levels of student disadvantage are also linked to higher levels of school safety incidents. Furthermore, both approaches have suggested that relative to other variables, and net with other theoretically-relevant variables included in the models, educational approaches (e.g., *sel*, a variable which captures whether a given school implements socio-emotional learning) seem to be less important in accounting for the variance in the number of school safety incidents. These lend support to the interpretation of these variables, according to the metrics generated through the explainable machine learning approach. These are also broadly in use of SHAP values employed for application on this set of school safety data.

However, there are also divergences when the two approaches are compared. In the linear regression, the coefficient of *parenthelp_drug*, a variable which captures whether parents were involved in drug prevention efforts for students, was not statistically significant, but was positive. This interpretation that higher levels of a parental and community involvement would lead to lower school safety incidents, is similar to the interpretation of results obtained from the explainable machine learning approach, where two variables (i.e., *parent_teacher_conf* and *open_house*) with high absolute SHAP values would also be interpreted in a similar manner (i.e., higher values of these variables are associated with less school safety incidents). With regard to key demographic variables that were surfaced within extant literature, the coefficient of *crimes_school_area*, a control variable which captures the level of crime around each school, was not statistically significant in the linear regression model; however, it did surface as one of the variables with the highest absolute SHAP values through the machine learning approach. In both approaches taken, however, higher levels of *crimes_school_area* remain associated with higher quantities of school safety incidents. Further, there was a broad convergence in the direction and importance of the remnant demographic variables included. For instance the coefficient of *urban*, a variable that captures whether a school is located in an urban area, was not statistically significant in the linear regression model, and was also not among the top 10 variables with the highest absolute SHAP values.

Two points of comparison can be made about the utility of these approaches. First, the issue of variable selection for the illustrative linear regression model, is one possible explanation for why parental and community involvement was not found to significantly explain the variance in *total_incident*, unlike the importance of parental and community involvement variables found through the explainable machine learning approach. Empirically, for any given number of independent variables which could represent some latent construct (e.g., ‘parental and community involvement’), selecting the most representative variables of that latent construct is often a challenge for linear regression modelling. While one could also overcome this issue with recursive or forward selection techniques for linear regression approaches, the explainable machine learning approach attempted in this thesis is one way a non-linear regressor could be used to meaningfully select the variables which account most of the variance in the outcome variable of interest. Second, it must be established that there are open questions to the comparability of these two approaches, since the metrics in question are completely different. In the case of linear regression, statistical significance tests on the beta coefficients were conducted. On the other hand, for explainable machine learning, a rough criterion for the top 10 variables with the highest absolute SHAP values was chosen. By comparison, the linear regression coefficients provide far more precision in terms of understanding the relationship between given independent variables and the outcome variable. Both absolute SHAP values and statistical significance tests on regression coefficients, however, do provide an indication of how much variance each independent variable contributes to the dependent variable. In light of this, the fact that there are substantial convergences across the main variable groups concerned in this thesis, across both approaches, is beginning evidence that the explainable machine learning approach has been useful in capturing relationships between a given outcome variable and a number of theoretically-relevant and statistically-important independent variables, in this context of examining the correlates of school safety incidents.

Student Disadvantage and School Safety Policy

These findings are a contribution toward the conceptual understanding of school safety by policy analysts and social scientists in this area. The finding that schools with higher markers of student disadvantage are on average, less safe, lends credence to the differential opportunities theory. Specifically, the markers of student disadvantage used in this study pertain to students' academic performance, and their attitudes towards academic achievement. Moreover, this study found that, given associations between parental involvement and school safety, such policies should be given due consideration by schools. The findings of this study should motivate a further investigation of high-security environmental features, like *random_sweep*, which are associated with schools that are less safe. Finally, even in terms of descriptive demographics of schools, there is evidence of inequality in school safety, given that larger schools, schools with more students who live in unsafe neighborhoods, schools in unsafe neighborhoods, as well as middle and combined schools experience lower levels of school safety.

This finding joins existing quantitative models that have found effects of student disadvantage on school safety (e.g., Han, 2010; Nickerson & Martens, 2008). Unlike both studies, however, this study additionally considered 46 other control or explanatory variables within the same machine learning model, to account for the effects of student disadvantage, net of the many potential causes of school safety. Unlike Han (2010), the present study's model was not restricted to a subset of the population (i.e., schools with more than 50 percent minority students), but was run on the full 2017-18 SSOCS dataset that was representative of the demographics within the United States. Unlike Nickerson & Martens (2008), which used a factor analytic approach to make sense of a range of school safety characteristics and features, the present study focused on the variables with the highest absolute SHAP values, therefore variables that contributed the highest variance in their association with the dependent variable, as the most crucial variables in relation to the output values of the dependent variable. This does have shortcomings, as the following section will describe.

Future Directions and Limitations. First, it is important to establish that these findings are associational, and not causal in nature. This does not have anything to do with the metrics. Instead, this limitation is due to the context in which the data was collected, since the SSOCS was collected as an observational survey. It is, however, likely the case that future studies on school safety will continue to be observational in nature, unless opportunities arise to conduct natural experiments or develop quasi-experimental designs on panel data; as an example, if a school experiences a change in demographics over time (e.g., due to a school merger between a school with a vastly different school in terms of the percentage of disadvantaged students), it might be possible to employ a difference-in-difference design or regression discontinuity design to quantify the average treatment effect of a change in the percentage of disadvantaged schools. The value of explainable machine learning to analyze such sequential or panel data would be a direction for further exploration. Secondly, the approach taken in this study uses one or two variables to make claims about latent constructs like parental involvement or student disadvantage; by comparison, the factor analytic approach taken by Nickerson and Martens (2008) to develop 5 separate factors for school safety approaches (i.e., each factor consisting of the contributions of multiple school safety variables) seems more rigorous. Nevertheless, the aforementioned point of the difficulty of performing factor analysis on mixed-type data, as is the case for our dataset, still holds. Further, there is also value in using the variables that we have eventually settled on, since these were the variables that contributed the greatest amount of variance to *total_incident*. Additional points of value of the current thesis' approach will be further expounded on in the subsection on the value of recursive feature elimination below.

Explainable Machine Learning for the Social Sciences

The value of the explainable machine learning approach taken in this thesis, has potential applications to social science or policy topics beyond school safety, since it can be readily used for any given problem with a large number of independent variables, in relation to a dependent variable. This methodological property is of great potential value for policy analysts and social scientists, who often need to analyze stochastic human or behavioral data. Such data tends to have multiple concurrent associations or causes, which then requires more complex statistical modelling that can consider a host of variables in unison. At the same time, there remain multiple open questions about the present work, which future research should address.

Why use the Random Forest Regressor, not Other Models? A random forest regressor was chosen for use in this dataset, because of its modest size in terms of observations (i.e., 2762), and the random forest regressor model's ability to perform well with relatively little tuning (see Hastie, Tibshirani & Friedman, 2009 for a review). However, given that both permutation feature importance and SHAP fall into a class of model agnostic measures, technically any machine learning model can be used in place of the random forest regressor (see Molnar, 2021 for a review). Other estimators, such as regularized linear estimators like Lasso or Ridge Regressions, could also be used in the model experimentation process. Notably, however, models like the Lasso and Ridge regression need to be applied with caution since standardization of variables, especially in use cases with mixed-type variables, would be an issue. The comparative flexibility of the random forest regressor, and the minimal need for preprocessing remains a good reason to utilize this model in future use cases, where mixed-type variables are involved.

In particular, if the sample size of a given dataset is far larger, it may be worthwhile to conduct model experimentation with neural networks, to optimize the performance of the model that will be analyzed with SHAP. This is of high importance, since SHAP essentially is an explication of what a given machine learning model has learnt from the data. The better a machine learning model's fit on the data (e.g., the lowest possible RMSE given one's model experimentation), the better the insights we can derive from applying explainable machine learning for social scientific interpretations.

Why use Recursive Feature Elimination, not Hierarchical Clustering? In this study, recursive feature elimination with permutation feature importance was chosen to minimize the shared variance between potentially related variables. Another approach, as recommended by the documentation of the scikit-learn developers,[10] is the use of hierarchical clustering, an unsupervised machine learning method, to cluster groups of related variables together, before calculating permutation importance on one given variable per cluster (Buitinck et al., 2013). This is indeed a more principled way to avoid multicollinearity, but again, hierarchical clustering needs to be done on a dataset with continuous variables; given mixed-type datasets like the SSOCS, it would not be possible to do this. Additionally, it is also noteworthy that even when hierarchical clustering is applied, one variable would still be selected per cluster to represent that given cluster to avoid the same problems of multicollinearity when calculating permutation importance. While features that are removed through recursive feature elimination with permutation importance were obtained through a very different way, the goals are in fact highly similar to the approach recommended by the scikit-learn developers, to reduce the effect of correlated variables.

Are SHAP Values Useful? The utility of the SHAP values, in providing both magnitude and direction of each effect, has been explained above, in the interpretation of results in the 'Explainable Machine Learning Approach' segment of this thesis. However, one must note that the present analysis was based on the assumption that only the variables with the highest absolute SHAP values were most worthy of consideration. This meant that there were 14 other variables, with lower absolute SHAP values, that were not directly considered but certainly also factored into the overall calculation of the SHAP values for each variable. In principle, there is no reason why 10 variables was the most suitable cutoff for consideration in SHAP. This was an arbitrary number to consider the variables that contributed the most variance to the output variable's values. However, as noted above, the RMSE of the model with 10 variables, when

[10] This specific page within the documentation of the scikit-learn API is of interest: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html.

validated on held-out test data, was 1.076. This was only slightly higher than the model trained on 24 variables which had a RMSE of 1.044. This does show that the top 10 variables with the highest absolute SHAP values do account for a large amount of the variance in the dependent variable.

Nevertheless, further research is required to more fully utilize all of the SHAP values that are provided by all variables, with some ideas on how to do so presented in the future directions segment; but for the purposes of this thesis, the descriptive statistics for all variables, including the 14 other variables that were not examined closely in this thesis are attached as Appendix B, as resources for further exploration; the entire dataset of SHAP values is also available, with reference to Appendix A.

Future Directions and Limitations. Given the new nature of this methodology, there are abundant opportunities to advance the use of explainable machine learning for the social sciences, using SHAP, beyond the approaches that were attempted within the main body of this thesis.

The first possibility is the development of an approach that would bring the application of SHAP values closer to conventional hypothesis testing, under a frequentist framework that is more familiar to the wider quantitative social sciences. Notably, the fact that one SHAP value is generated per observation, per feature, means that the entire dataset is represented as SHAP values, with reference to a baseline; each SHAP value, again, represents the value that each individual observation contributes to the dependent variable, for a given model. This raises the possibility of using classic non-parametric sampling techniques, like bootstrapping, to construct 95% confidence intervals for specific sub-samples of interest (e.g., on SHAP values for schools without random sweeps, and SHAP values for schools with random sweeps). This can be followed with conventional statistical significance tests, through a comparison of means.

The second important possibility is the application of SHAP values to uncover feature interactions. Interactions are difficult to uncover in the social sciences; in the absence of extant literature on specific conditional relationships, there is no principled way to include interaction effects for the social sciences. A data mining approach using SHAP values, however, would make it possible to discover theoretically useful interaction effects. The TreeExplainer method for tree-based models like gradient boosted trees and the random forest model has been proposed by Lundberg et al. (2020). This approach of calculating SHAP interactions values is intended to tease apart main effects of variables and potential non-linear interactions between these variables. This is an additional extension of the current work in the thesis, that would make it even more useful for social scientists, policy analysts and methodologists.

Conclusion

Through the use of SHAP values applied on a tuned random forest regressor, this thesis has provided some directions on the utility of explainable machine learning on the social sciences, and contributed to the theoretical understanding of the relational, environmental, and demographic factors that underlie school safety. This approach, along with promising future extensions, also provides researchers and policy analysts with a way to consider numerous theoretically-relevant variables within the same model. It provides quantitative metrics for the order of independent variables that are most important in explaining the variance in the dependent variable, along with the direction of the non-linear relationship between these independent variables and the dependent variable. The use of explainable machine learning in this thesis is an instance of how the latest and best in data science has potential value for analyzing school safety incidents. This approach can be generalized and applied for research on topics in the wider social sciences, as one additional technique that has value for modelling multivariate associations.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Astor, R. A., Benbenishty, R., & Estrada, J. N. (2009). School violence and theoretically atypical schools: The principal's centrality in orchestrating safe schools. *American Educational Research Journal, 46*(2), 423–461. https://doi.org/10.3102/0002831208329598

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*(1), 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Barnes, T. N., Leite, W., & Smith, S. W. (2017). A Quasi-Experimental Analysis of Schoolwide Violence Prevention Programs. *Journal of School Violence*, *16*(1), 49–67. https://doi.org/10.1080/15388220.2015.1112806

Braithwaite, John. 1981. The Myth of Social Class and Criminality Reconsidered. *American Sociological Review*, *46*(1), 36. https://doi.org/10.2307/2095025.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv, *1309.0238*. https://arxiv.org/abs/1309.0238

Chen, G. (2008). Communities, Students, Schools, and School Crime: A Confirmatory Study of Crime. *Urban Education*, *43*(3), 301–318.

Cloward, R. A., & Ohlin, L. E. (1960). *Delinquency and Opportunity: A theory of delinquent gangs.* Free Press.

Cuellar, M. J. (2018). School Safety Strategies and Their Effects on the Occurrence of School-Based Violence in U.S. High Schools: An Exploratory Study. *Journal of School Violence*, *17*(1), 28–45. https://doi.org/10.1080/15388220.2016.1193742

Fite, P. J., Poquiz, J., Díaz, K. I., Williford, A. & Tampke, E. C. (2019). Links Between Peer Victimization, Perceived School Safety, and Internalizing Symptoms in Middle Childhood. *School Psychology Review, 48*(4), 309–19. https://doi.org/10.17105/SPR-2018-0092.V48-4

Gastic, B., & Johnson, D. (2015). Disproportionality in Daily Metal Detector Student Searches in U.S. Public Schools. *Journal of School Violence*, *14*(3), 299–315. https://doi.org/10.1080/15388220.2014.924074

Han, S. (2003). A Mandatory Uniform Policy in Urban Schools : Findings from the School Survey on Crime and Safety : 2003-04 University of Missouri Student Problem Behaviors : Fre- quency and Types Problem Behavior and Student Characteristics : Minority Status and Achieve. *International Journal of Education Policy and Leadership*, *5*(8), 1–13.

Han, S., & Akiba, M. (2011). School Safety , Severe Disciplinary Actions , and School Characteristics : A Secondary Analysis of the School Survey on Crime and Safety. *Journal of School Leadership*, *21*, 262–292. https://doi.org/10.1177/105268461102100206

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D.,Wieser, E; Taylor, J; Berg, S; Smith, NJ; Kern, R; Picus, M; Hoyer, S; van Kerkwijk, MH; Brett, M; Haldane, A; del Río, JF; Wiebe, M; Peterson, P; Gérard-Marchant, P; Sheppard, K; Reddy, T; Weckesser, W; Abbasi, H; Gohlke, C; Oliphant, TE. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York: Springer.

Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, C., Mitchell, D. (2020). lmtest: Testing Linear Regression Models. R package version 0.9-38. https://CRAN.R-project.org/package=lmtest

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Joshi, K., & Joshi, C. K. (2019). Working women and caste in India: A study of social disadvantage using feature attribution. arXiv, 1905.03092. https://ideas.repec.org/p/arx/papers/1905.03092.html

Korobov M. and Lopuhin K. (2019). ELI5. Python package version 0.11.0. https://pypi.org/project/eli5/

Lenzi, M., Jill S., Furlong, M. J., Mayworm, A., Hunnicutt, K. & Vieno, A. (2017). School Sense of Community, Teacher Support, and Students' School Safety Perceptions. *American Journal of Community Psychology, 60*(3–4), 527–37. https://doi.org/10.1002/ajcp.12174

Lesneskie. E. & Block, S. (2017). School Violence: The Role of Parental and Community Involvement, *Journal of School Violence*, *16*(4), 426-444. https://doi.org/10.1080/15388220.2016.1168744

Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765–4774.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, *45*(1), 27–45. https://doi.org/10.1146/annurev-soc-073117-041106

Molnar, C. (2021). Interpretable machine learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Nickerson, A. B., & Martens, M. P. (2008). School violence: Associations with control, security/enforcement, educational/therapeutic approaches, and demographic factors. *School Psychology Review*, *37*(2), 228–243. https://doi.org/10.1080/02796015.2008.12087897

Padgett, Z., Jackson, M., Correa, S., Kemp, J., Gilary, A., Meier, A., Gbondo-Tugbawa, K., & McClure, T. (2020). *School Survey on Crime and Safety: 2017–18 Public-Use Data File User's Manual* (NCES 2020- 054). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Perumean-Chaney, S. E., & Sutton., L. M. (2013). Students and Perceived School Safety: The Impact of School Security Measures. *American Journal of Criminal Justice, 38*(4), 570–88. https://doi.org/10.1007/s12103-012-9182-2.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Reingle Gonzalez, J. M., Jetelina, K. K. & Jennings, W. G. (2016). Structural School Safety Measures, SROs, and School-Related Delinquent Behavior and Perceptions of Safety: A State-of-the-Art Review. *Policing*, *39*(3), 438–54. https://doi.org/10.1108/PIJPSM-05-2016-0065.

Revelle, W. (2019). psych: Procedures for Personality and Psychological Research. R package version 1.9.12. http://personality-project.org/r, https://personality-project.org/r/psych-manual.pdf

Robinson, L. R., Leeb, R. T., Merrick, M. T., and Forbes, L. W. (2016). Conceptualizing and Measuring Safe, Stable, Nurturing Relationships and Environments in Educational Settings. *Journal of Child and Family Studies*, *25*(5), 1488–1504. https://doi.org/10.1007/s10826-015-0332-2.

RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/.

Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press. Open review edition.

Tanner-Smith, E. E., Fisher, B. W., Addington, L. A., & Gardella, J. H. (2018). Adding Security, but Subtracting Safety? Exploring Schools' use of Multiple Visible Security Measures. *American Journal of Criminal Justice*, *43*(1), 102–119. https://doi.org/10.1007/s12103-017-9409-3

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300

The pandas Development Team. (2020). Pandas. Python package version 1.2.4. https://doi.org/10.5281/zenodo.3509134

Thornberry, T. P., and Margaret F. (1982). Social Correlates of Criminal Involvement: Further Evidence on the Relationship Between Social Status and Criminal Behavior. *American Sociological Review*, *47*(4), 505–18.

Tittle, Charles R., Wayne J. Villemez, and Douglas A. Smith. (1978). The Myth of Social Class and Criminality : An Empirical Assessment of the Empirical Evidence. *American Sociological Review*, *43*(5), 643–56. https://doi.org/https://doi.org/10.2307/2094541.

U. S. Department of Education (n.d.). School Survey on Crime and Safety (SSOCS) — Overview. Retrieved February 03, 2021, from https://nces.ed.gov/surveys/ssocs

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*, 1686. https://doi.org/10.21105/joss.01686

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach*. Mason, OH: South-Western Cengage Learning.

Yoo, W., Mayberry, R., Bae, S., Singh, K., Peter He, Q., & Lillard, J. W., Jr (2014). A Study of Effects of MultiCollinearity in the Multivariable Analysis. *International Journal of Applied Science and Technology*, *4*(5), 9–19.

Yuan, Y., and An., W. (2017). Context, Network, and Adolescent Perceived Risk. *Social Science Research* 62, 378–93. https://doi.org/10.1016/j.ssresearch.2016.08.018.

Appendix

Appendix A — Data and Code Availability Statement

Statistical computing was done in R for data preprocessing, feature engineering and for the illustrative example of social scientific linear regression modelling. The code for the machine learning models is in Python. All of my code is currently hosted on a GitHub repository: https://github.com/kagenlim/MA-Thesis. If I might be able to help in any way, I would be most happy to do whatever I can; please reach out to me at kagen.lim@columbia.edu and I will be happy to have a conversation about this piece of work.

Appendix B — Full Set of SHAP Values

These are the descriptive statistics of the SHAP values for the 24 variables that were isolated by the recursive feature elimination procedure. They have been sorted in descending order, according to their mean absolute SHAP values.

	Count	**Mean**	**SD**	**Min**	**25th_ Percentile**	**Median**	**75th_ Percentile**	**Max**	**Mean_ Absolute**
school_size	2762	-0.033	0.389	-0.947	-0.295	-0.122	0.415	0.755	0.33401
school_type	2762	-0.047	0.278	-0.895	-0.023	0.08	0.131	0.256	0.20647
crimes_school_area	2762	0.013	0.236	-0.653	-0.158	0.019	0.17	0.904	0.18946
open_house	2762	0.005	0.113	-0.36	-0.082	0.029	0.089	0.309	0.09453
percentacadnotimpt	2762	0.015	0.117	-0.634	-0.035	0.046	0.095	0.36	0.09333
crimes_students_residence	2762	0.012	0.107	-0.201	-0.074	-0.037	0.102	0.408	0.09276
parent_teacher_conf	2762	0.004	0.102	-0.265	-0.078	0.009	0.086	0.327	0.08677
percentlowgrades	2762	0.002	0.106	-0.512	-0.049	0.004	0.07	0.345	0.07756
juvhelp_drug	2762	-0.015	0.085	-0.237	-0.082	-0.028	0.063	0.21	0.07595
random_sweep	2762	-0.008	0.066	-0.203	-0.059	-0.004	0.044	0.193	0.05541
mhhelp_drug	2762	0.008	0.052	-0.174	-0.026	0.027	0.042	0.116	0.04473
anonymous_report	2762	-0.002	0.054	-0.237	-0.041	0.007	0.04	0.151	0.04421

urban	2762	0.006	0.043	-0.239	0.003	0.02	0.031	0.096	0.03441
prohibit_phone	2762	-0.003	0.038	-0.12	-0.031	-0.01	0.023	0.115	0.03166
control_ground	2762	-0.003	0.032	-0.14	-0.021	0.002	0.019	0.113	0.02515
dress_code	2762	0	0.032	-0.136	-0.018	0.001	0.02	0.144	0.02451
relighelp_drug	2762	0.004	0.032	-0.074	-0.015	-0.005	0.014	0.202	0.0223
student_id	2762	0.003	0.034	-0.11	-0.011	-0.006	0	0.277	0.01721
disciplinary_process	2762	0.001	0.02	-0.086	-0.011	0.002	0.014	0.069	0.01596
peer_mediation	2762	0.002	0.019	-0.071	-0.009	0.002	0.013	0.083	0.0143
privatehelp_drug	2762	0	0.019	-0.066	-0.01	-0.001	0.008	0.152	0.01344
lawhelp_drug	2762	-0.003	0.025	-0.191	-0.001	0.003	0.006	0.095	0.0112
wear_uniform	2762	0	0.018	-0.143	-0.003	0.002	0.005	0.164	0.00919
require_check	2762	-0.001	0.008	-0.146	-0.001	0	0	0.046	0.00165