Predicting Suicide Ideation

Eliud Fasil, Siddharth Govindarajan, Ketan Parmar, Jim Ramsey

HMGT 6334 Project- Group 1

November 30, 2020

Table of Contents

Section 1: Introduction and Executive Summary

Teenage Suicide Ideation is a significant problem throughout the world. Public Agencies, including school systems, need better tools to help predict and identify students who are likely to be considering suicide. Reliable predictive tools will help leaders understand which students are more likely to be vulnerable to suicide ideation, permitting organizations to put in place counter measures for early intervention – well ahead of any suicide planning or action being taken.

Our healthcare data analytics project selected the topic of teenage suicide ideation in three south and southeast Asia countries – Bangladesh, Bhutan, and Myanmar. Our goal was to develop a prediction model that could identify high school students at risk of suicide ideation, and more severe suicide related acts, such as suicide planning and attempting suicide.

Section 1.1 Our Roadmap

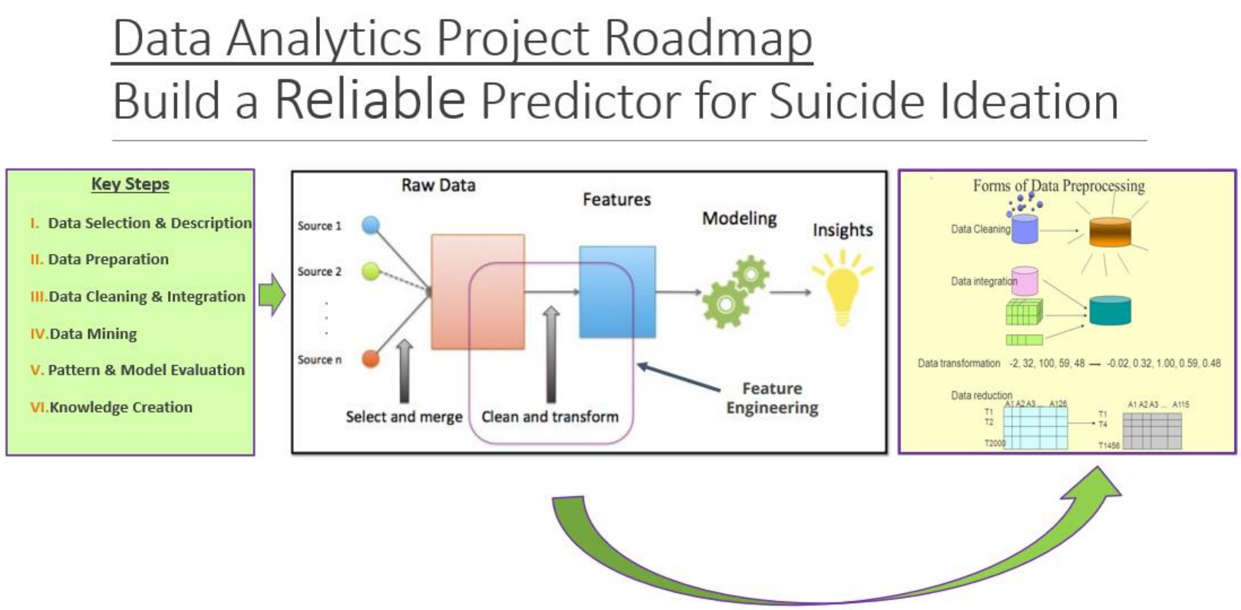


Figure A

We developed our "project roadmap" using the tools presented in our textbooks and class presentations. As noted in the Figure A above, the following were the key steps we followed for data analysis:

- Data Selection & Description
- Data Preparation
- Data Cleaning & Integration
- Data Mining

- Pattern & Model Evaluation
- Knowledge Creation

The part of the project that required the most work was that involved with the various steps of data pre-processing.

Having worked through the predictive models for the individual three countries, we found that our models individually were not as strong as we would have liked. Based on the direct feedback from Professor Ayvaci, we completed another iteration and weaved together data from all three countries to develop a common predictive tool – which yielded a more reliable outcome.

Section 2: Project Background and Motivation

In selecting a project for our Healthcare Data Analytics course, our team was drawn to pursuing one that had social impact and real-world implications. We came across a dataset on the World Health Organization's (WHO) website dealing with the topic of suicide ideation in high school students. The WHO views suicide rate reduction as one of their sustainable development goals.

A new topic for our team, we learned that suicide "ideation" is the act of "thinking" about suicide, which can then lead to more severe outcomes, such as planning for suicide, attempting suicide – and death.

We did some preliminary research on this topic from published sources (see bibliography) and decided to examine a part of the world that does not get as much attention as the United States and Western Europe. We selected WHO studies from three countries in South and Southeast Asia that while relatively close in geography, have distinct cultures and histories:

- Bangladesh
- Bhutan
- Myanmar (formerly known as Burma)

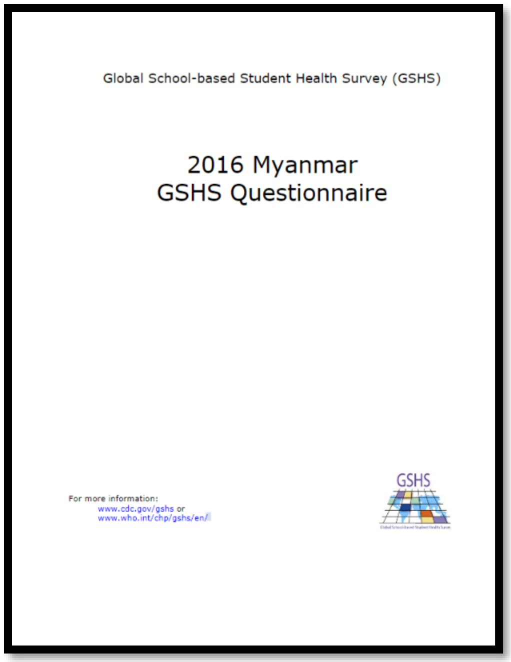
We also found in our research the causals of suicide ideation have been the topic of many studies, and that such causals are global in nature. However, what remains lacking are the necessary predictive tools that can help identify individuals from a population that may be contemplating suicide ideation and related actions. Further, existing predictive tools were found to be lacking, and what appears to be more effective is using data that is extracted from student medical records and self-completed questionnaires.

While we did not have access to specific student medical records, we were encouraged to use the WHO surveys from just a few years ago (2016) to be the source of our analytics project. Our challenge then for our project was to build a reliable predictive model using just the WHO data files that could be used to identify students who might be considering suicide. This information can equip both school administrators and public health officials with the means to develop preventive and early intervention countermeasures to reduce the occurrences of suicide ideation and related actions – and save lives.

Section 3: Data Description

Our objective was to create a prediction model that identifies students that are likely to have suicide ideation. We had the following number of observations and variables for our three countries:

Country	**Observations**	**Variables**	**Number of Considered suicide**	**Number of made plans for suicide**	**Number of Attempted suicides**
Bangladesh	2,989	135	127	171	152
Myanmar	2,838	141	245	183	240
Bhutan	7,576	141	862	1043	828



Global School-based Student Health Survey (GSHS)

2016 Myanmar
GSHS Questionnaire

For more information:
www.cdc.gov/gshs or
www.who.int/chp/gshs/en/

GSHS

Sample Report

Our data originates from the WHO (World Health Organization) website specifically from the identified countries : [Data collections - WHO]

Section 3.1: Risk Factors

Our target variables were questions related to the risk factors[2,3] of suicidal ideation.

Risk factors are characteristics of a person or his or her environment that increase the likelihood that he or she will die by suicide (i.e., suicide risk). Major risk factors for suicide can include:

- Prior suicide attempt(s)
- Misuse and abuse of alcohol or other drugs
- Mental disorders, particularly depression and other mood disorders
- Access to lethal means
- Knowing someone who died by suicide, particularly a family member
- Social isolation
- Chronic disease and disability
- Lack of access to behavioral health care

For our report, we used questions pertaining to the following topics:

Major risk factors for suicide https://www.sprc.org/about-suicide/risk-protective-factors	**Columns in our datasets that pertain to the factors mentioned in the previous column**
Misuse and abuse of alcohol or other drugs	Drug usage (4 questions)
Mental disorders, particularly depression and other mood disorders / Social isolation	Feeling and Friendship (6 Questions)
Prior suicide attempt(s)	Suicide plan and attempt (3 Questions)
Lack of access to behavioral health care	Food habits, Brushing, Washing hands, Physical activity (20 Questions)

Section 3.2: Addressing "Missingness"

There are some blank rows in a column which signifies that there was no response for that question from that student. We transformed these categorical blank values into mode of the column and numerical blank values with the median of the columns and then when we ran our classification problem.

Section 4: Models before presentation feedback

To build our models, we went through a series of steps.

- We imported the data in the R studio and performed the data cleaning process with imputation, outlier detection, and related data cleaning steps.
- We then performed exploratory data analysis (EDA) and implemented the classification algorithms on the dependent variables.

We ran our prediction models with the following regressions:

- Logistic regression
- XG Boosting
- Decision tree (was not successful in classifying the data) – Accuracy was remaining the same for all the tree depths

For each of our three countries, we selected the following questions as our <u>dependent variables</u>:

- During the past 12 months, did you ever seriously <u>consider</u> attempting suicide?
- During the past 12 months, did you <u>plan</u> about how you would attempt suicide?
- During the past 12 months, how many times did you <u>attempt</u> suicide?

Section 4.1: Data Cleaning Process

The following are steps taken during our data cleaning process:

- Converting the necessary numerical columns to factors
- Imputing the numeric NAs with Median
- Imputing the categorical NAs with Mode
- Detecting handling outliers for numerical variables
- Checking correlation for numerical variables

Column Name	Column Description
Q1	Custom Age
Q2	Sex
Q3	In what grade are you
Q4	How tall are you
Q5	How much do you weigh
Q6	How often went hungry
Q7	No fruit eating
Q8	No vegetable eating
Q9	Soft drinks
Q10	Fast food eating
Q11	Tooth brushing
Q12	Hand washing before eating
Q13	Hand washing after toilet
Q14	Hand washing with soap
Q15	Physically attacked
Q16	Physical fighting
Q17	Seriously injured
Q18	Serious injury type
Q19	Serious injury cause
Q20	Bullied
Q21	Bullied how
Q22	Felt lonely
Q23	Could not sleep
Q24	Considered suicide
Q25	Made a suicide plan
Q26	Attempted suicide
Q27	Close friends
Q28	Initiation of cigarette use
Q29	Current cigarette use
Q30	Other tobacco use
Q31	Smoking cessation
Q32	Smoking in their presence

Q33	Parental tobacco use
Q34	Initiation of alcohol use
Q35	Current alcohol use
Q36	Drank 2+ drinks
Q37	Source of alcohol
Q38	Really drunk
Q39	Trouble from drinking
Q40	Initiation of drug use
Q41	Ever marijuana use
Q42	Current marijuana use
Q43	Amphetamine or methamphetamine use
Q44	Ever sexual intercourse
Q45	Age first had sex
Q46	Number of sex partners
Q47	Condom use
Q48	Birth control used
Q49	Physical activity past 7 days
Q50	Walk or bike to school
Q51	PE attendance
Q52	Sitting activities
Q53	Miss school no permission
Q54	Other students' kind and helpful
Q55	Parents check homework
Q56	Parents understand problems
Q57	Parents know about free time
Q58	Parents go through their things

Section 4.2 : Models before presentation feedback

Section 4.2.1: Considered suicide

For the 3 countries , we ran 3 different models having considered suicide as the dependent variable.

The Confusion matrix, AUC, Accuracy, Sensitivity, Specificity and Precision are tabulated below. For Logistic regression we used, GLM package in R and for XG Boost we used Rattle.

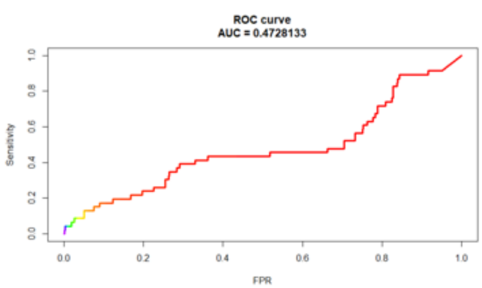
Bhutan	Logistic	XG Boost
Accuracy	88.34	89
Sensitivity	11.24	90.19
Specificity	98.21	36
Precision	44.62	98.43

Bangladesh	Logistic	XG Boost
Accuracy	93.97	94.65
Sensitivity	4.35	94.86
Specificity	98.82	0
Precision	16.67	99.76

Myanmar	Logistic	XG Boost
Accuracy	89.89	91.67
Sensitivity	16.18	92.49
Specificity	96.3	38.46
Precision	27.5	98.98

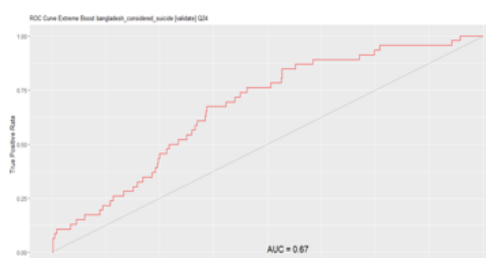
Based on the above metrics, XG Boost was better for Bhutan and Myanmar but in case of Bangladesh since the specificity is 0 for XG Boost, logistic is better in that case.

ROC for Bangladesh



AUC : 0.47

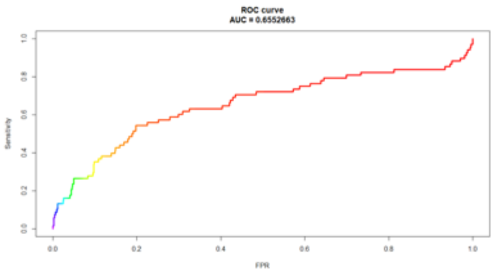
Logistic Regression



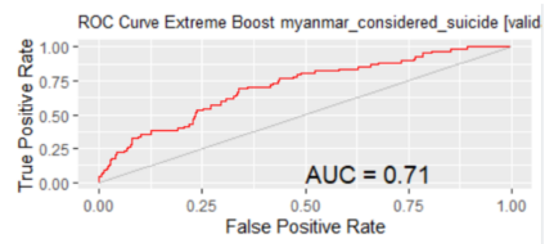
AUC : 0.67

XG Boost

ROC for Myanmar

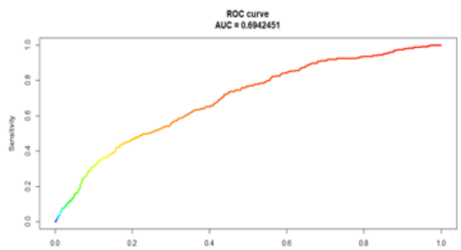


AUC: 0.66
Logistic Regression

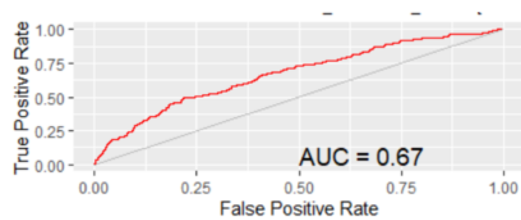


AUC : 0.71
XG Boost

ROC for Bhutan



AUC: 0.69
Logistic Regression



AUC : 0.67
XG Boost

While comparing the ROC values, for Bangladesh – Logistic it was 0.47 and not so great. But for Bhutan and Myanmar – XG Boost gave us an acceptable ROC value of 0.67 and 0.71, respectively.

Section 4.2.2: Made plans for suicide

For the 3 countries , we ran 3 different models having made plans for suicide as the dependent variable.

The Confusion matrix, AUC, Accuracy, Sensitivity, Specificity and Precision are tabulated below. For Logistic regression we used, GLM package in R and for XG Boost we used Rattle.

Bhutan	Logistic	XG Boost
Accuracy	85.65	86.23
Sensitivity	10.58	87.5
Specificity	97.6	36.84
Precision	41.25	98.16

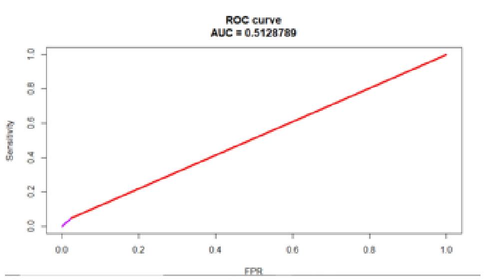
Bangladesh	Logistic	XG Boost
Accuracy	91.41	94.09

Sensitivity	5.08	94.08
Specificity	97.49	100
Precision	12.5	100

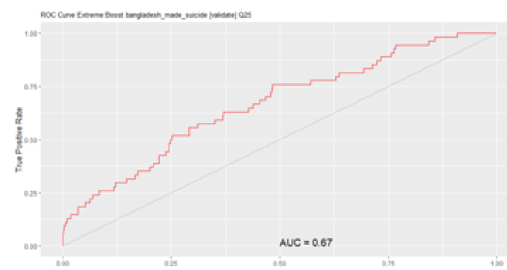
Myanmar	**Logistic**	**XG Boost**
Accuracy	88.72	93.54
Sensitivity	8.82	93.54
Specificity	95.66	50
Precision	15	99.87

Based on the above metrics, XG Boost was better for Bhutan and Myanmar but in case of Bangladesh since the precision and specificity is 100 for XG Boost, logistic is better in that case.

ROC for Bangladesh

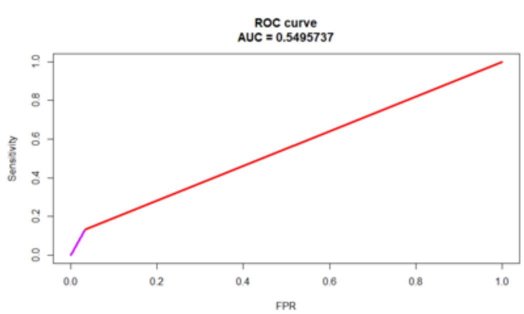


AUC: 0.51
Logistic Regression

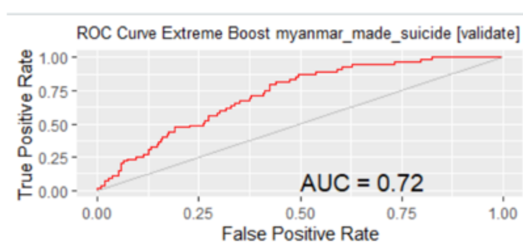


AUC : 0.67
XG Boost

ROC for Myanmar

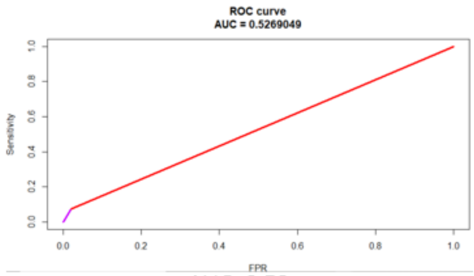


AUC: 0.55
Logistic Regression



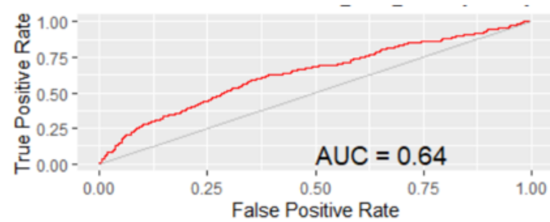
AUC : 0.72
XG Boost

ROC for Bhutan



AUC: 0.53

Logistic Regression



AUC : 0.64

XG Boost

While comparing the ROC values, for Bangladesh – Logistic it was 0.51 and not so great. But for Bhutan and Myanmar – XG Boost gave us an acceptable ROC value of 0.64 and 0.72, respectively.

Section 4.2.3: Attempted suicide

For the 3 countries , we ran 3 different models having attempted suicide as the dependent variable.

The Confusion matrix, AUC, Accuracy, Sensitivity, Specificity and Precision are tabulated below. For Logistic regression we used, GLM package in R and for XG Boost we used Rattle.

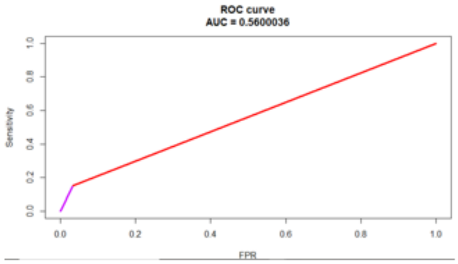
Bhutan	Logistic	XG Boost
Accuracy	88.12	89.4
Sensitivity	13.89	90.12
Specificity	97.38	45.95
Precision	39.77	99.02

Bangladesh	Logistic	XG Boost
Accuracy	93.08	93.98
Sensitivity	15.38	94.38
Specificity	96.62	42.86
Precision	17.14	99.53

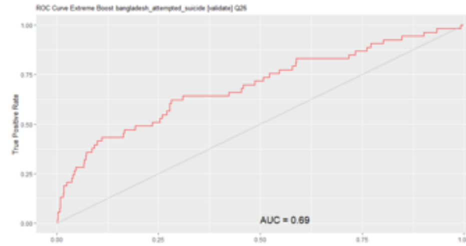
Myanmar	Logistic	XG Boost
Accuracy	88.64	90.85
Sensitivity	13.92	91.25
Specificity	96.5	33.33
Precision	28.95	99.48

Based on the above metrics, XG Boost was better for all 3 countries.

ROC for Bangladesh

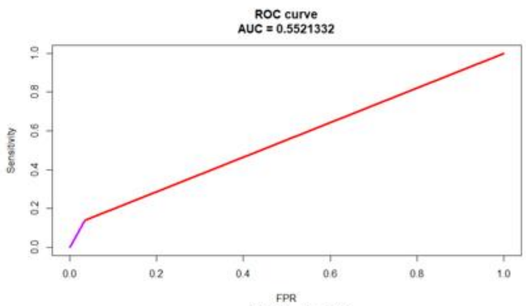


AUC: 0.56
Logistic Regression

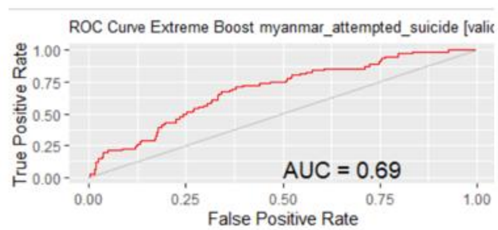


AUC : 0.69
XG Boost

ROC for Myanmar

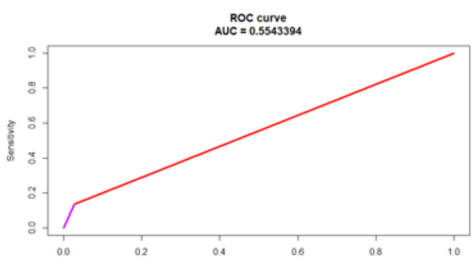


AUC: 0.55
Logistic Regression

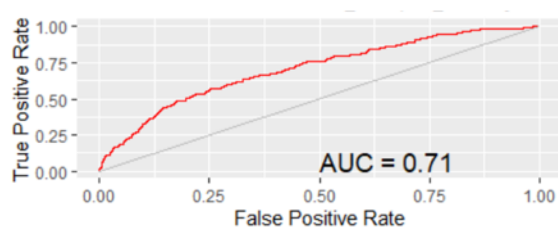


AUC : 0.69
XG Boost

ROC for Bhutan



AUC: 0.55
Logistic Regression



AUC : 0.71
XG Boost

While comparing the ROC values, for all 3 countries we have an acceptable ROC values – 0.69 for Bangladesh, 0.69 for Myanmar and 0.71 for Myanmar.

Section 4.2.4 : Seriousness of suicide

The seriousness of suicide is built below based on the conditions :-

Condition value	Condition
0	If a person has never attempted suicide nor made plans nor considered about it
3	Attempted suicide
2	Made suicide plans
1	Considered suicide plans

Down sampling was required since the output of dependent variables were not constant across. Multinomial logit was run, and the result is shown below.

Country	Distribution of data	Confusion Matrix
Bhutan	0 1 2 3 217 148 264 828	prediction_model2 0 1 2 3 0 7 13 11 6 1 11 9 13 16 2 11 10 7 9 3 15 12 13 13
Myanmar	0 1 2 3 14 18 2566 240	prediction_model2 0 1 2 3 0 2 1 2 2 1 0 1 0 2 2 0 1 2 0 3 2 1 0 0
Bangladesh	0 1 2 3 32 52 65 152	prediction_model2 0 1 2 3 0 0 2 2 3 1 2 4 0 4 2 4 3 5 0 3 3 0 2 2

There is a huge amount of misclassification rate.

Section 5: Data cleaning after presentation feedback

Section 5.1: Data Cleaning Process

The following are steps taken during our data cleaning process:

- Converting the necessary numerical columns to factors
- Imputing the numeric NAs with Median
- Imputing the categorical NAs with Mode
- Detecting handling outliers for numerical variables
- Checking correlation for numerical variables

Based on the parameters – considered suicide, made plans for suicide and attempted suicide we created a new column which is set as 1 if at least one of the above 3 parameters were 1 else 0.

After making the above change we had,

- 2109 students in at least one of the 3 cases.
- 11294 students not in any of the 3 cases.

We ran the logistic regression and XG boost on this new modified dataset.

Section 5.2: Important variables in the dataset

The questionnaire covered a wide range of questions from the lifestyle of the student, their home life, and their personal experiences. We filtered through the topics and selected the questions that most directly relates to our dependent variables and removed any columns that were completely missing.

Below is the list of questions and what each is asking:

- Q15 Physically attacked
- Q16 Physical fighting
- Q17 Seriously injured
- Q20 Bullied
- Q22 Felt lonely
- Q23 Could not sleep
- Q27 Close friends
- Q32 Smoking in their presence
- Q33 Parental tobacco use
- Q34 Initiation of alcohol use
- Q35 Current alcohol use
- Q36 Drank 2+ drinks
- Q37 Source of alcohol
- Q38 Really drunk
- Q39 Trouble from drinking
- Q40 Initiation of drug use
- Q49 Physical activity past 7 days
- Q54 Other students are kind and helpful
- Q56 Parents understand problems
- Q58 Parents go through their things

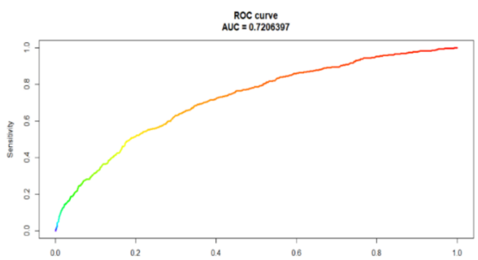
Section 6: Models after presentation feedback

Section 6.1 : Running the model with all variables - no down sampling

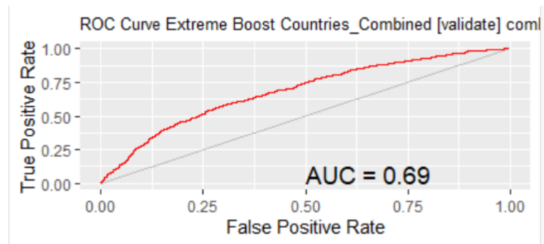
All variables were used for data classification, no down sampling with Logistic regression and XG Boost.

	Logistic	XG Boost
Accuracy	0.8440298507	0.8296443671
Sensitivity	0.1080246914	0.07716049383
Specificity	0.9854685647	0.9742069374
Precision	0.5882352941	0.3649635036

ROC Curve



Logistic : 0.72



XG Boost : 0.69

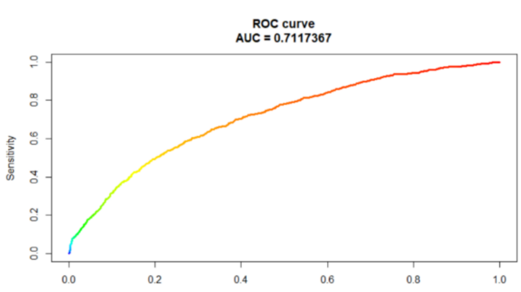
For the parameters – Accuracy, Sensitivity, Specificity, Precision and ROC curve we have logistic better than XG Boost. But the sensitivity value is very low.

Section 6.2: Running the logistic model with important variables - no down sampling

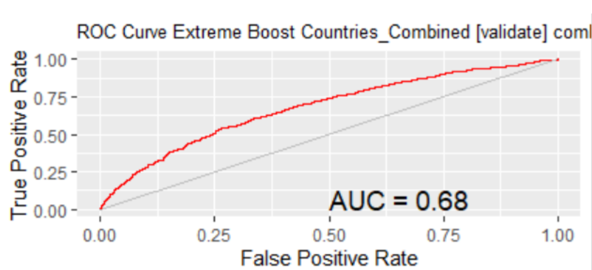
Important variables mentioned in Section 5.1.2 were used for data classification, no down sampling with Logistic regression and XG Boost.

	Logistic	XG Boost
Accuracy	0.8410447761	0.8348669485
Sensitivity	0.08796296296	0.0987654321
Specificity	0.9857651246	0.9762822413
Precision	0.5428571429	0.4444444444

ROC Curve



Logistic : 0.71



XG Boost : 0.68

For the parameters – Accuracy, Sensitivity, Specificity, Precision and ROC curve – except sensitivity we have logistic better than XG Boost.

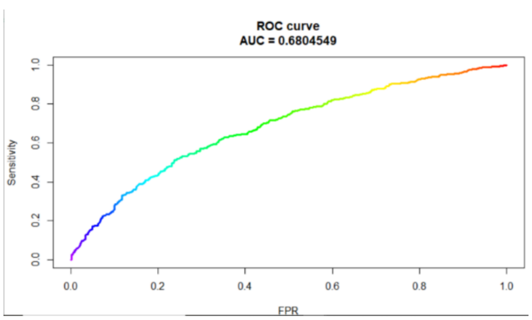
But the sensitivity value is very low.

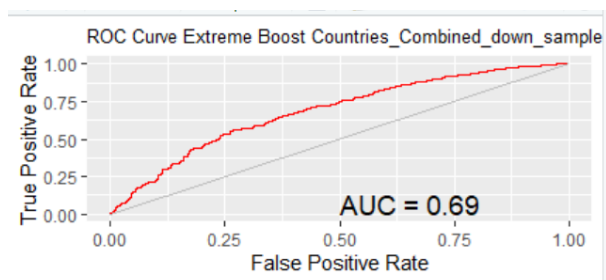
Section 6.3: Running the logistic model with all variables - down sampling

All variables were used for data classification, no down sampling with Logistic regression and XG Boost.

	Logistic	XG Boost
Accuracy	0.6376582278	0.6366508689
Sensitivity	0.6265822785	0.6603475513
Specificity	0.6487341772	0.6603475513
Precision	0.640776699	0.6434494196

ROC Curve





Logistic : 0.68 XG Boost : 0.69

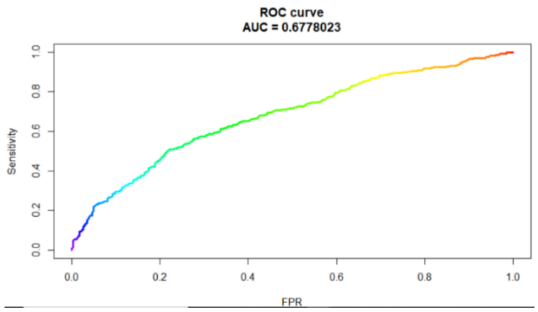
For the parameters – Accuracy, Sensitivity, Specificity, Precision and ROC curve we have XG Boost better than logistic and sensitivity is not as bad as the above case without down sampling.

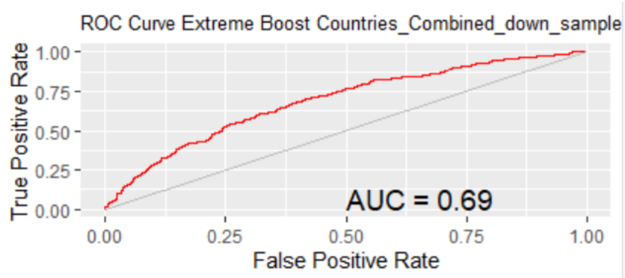
Section 6.4: Running the logistic model with important variables - down sampling

Important variables mentioned in Section 5.1.2 were used for data classification, down sampling with Logistic regression and XG Boost.

	Logistic	XG Boost
Accuracy	0.6352848101	0.640600316
Sensitivity	0.6075949367	0.5939968404
Specificity	0.6629746835	0.6872037915
Precision	0.6432160804	0.6550522648

ROC Curve





Logistic : 0.68

XG Boost : 0.69

For the parameters – Accuracy, Sensitivity, Specificity, Precision and ROC curve we have XG Boost better than logistic and sensitivity is not as bad as the above case without down sampling.

Section 7: Our Findings

Initially we tried considered suicide, made plans for suicide and attempted suicide separately for 3 countries and our results were not so great due to the dependent variable being quite imbalanced.

Based on the feedback given by the professor, we tried combining the 3 countries and 3 of the above conditions together and we had good results when we down sampled the results.

We now have an analytics tools that can predict the students who are more likely to consider suicide, make plans for suicide or attempt suicide. With the appropriate amount of data preparation, this analytic tool can be the starting point for other similar WHO studies in different countries.

There was a lot of work to do to get the data ready. There is quite a variation in in the questionnaires in terms of questions not answered.

Logistic and XG Boost was better in their respective cases and for down sampling the results with combined dataset, we had XG boost doing a better classification work.

A note - if the dataset is not having equal number of 0's (Never tried any of the 3 cases) and 1's (Tried at least one of the 3 cases), an appropriate data down sampling would be required.

The most challenging part was cleaning the dataset as we had more than 50 columns for each of the country. Getting the data right was the most important job of this project.

The most interesting part was how each model's parameters were performing notably different than the other model. This was quite the learning experience to learn how data "behaves" as we prepare and analyze it.

The biggest value added to us was what we could learn about data analytics from each other working as a team and expanding our understanding of how to apply data mining tackling a real-world problem.

Section 8: Managerial Implications & Conclusions

Despite the significant efforts required to manage the questionnaire data, this project did yield a working data analytics model that can effectively help to identify students who are thinking about suicide, planning suicide and may even be attempting suicide.

This has implications (and benefits) for public health officials and other government leaders, including those in education. Through reliably predicting students "at risk," countermeasures can be put in place for early intervention to reduce such occurrences.

One recent study[1] found that prediction models incorporating both health record data and responses to self-report questionnaires substantially outperform existing suicide prediction tools.

Using this data analytics project, as a foundation, refinements to predicting teenage suicide can be made by:

- Weaving in student health record data (if available) with similar self-reported questionnaire to strengthen the predictive model
- Taking the analysis further, identify the factors that most influence suicide ideation for a given population

References

[1]Gregory E. Simon et al (October 2018). "Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records," the American Journal of Psychiatry, 175.10 appi.ajp.2018.17101167 (psychiatryonline.org)

[2]"Risk and Protective Factors in Racial/Ethnic Populations in the U.S.," Suicide Prevention Resource Center, October 2020 (https://www.sprc.org/about-suicide/risk-protective-factors)

[3]Lavoie, Amy (February 7, 2008). "Suicide Risk Factors Consistent Globally," The Harvard Gazette Suicide risk factors consistent globally – Harvard Gazette

Burke, J. (2013). Health Analytics: Gaining the Insights to Transform Health Care (Vol. 71). John Wiley & Sons.

Viswanathan, V. (2014). Data Analytics with R: A Hands-on Approach. InfoVista Incorporated.