Indexing the potential number of dropouts from a particular governmental school in a given year using statistical learning and logistic modelling

Vansh Gupta
Mathematics Department
The Doon School
Dehradun, India

Abstract — My research aims to perform quantitative analysis on the various factors that influence the number of dropouts from governmental schools across the country to create a 'dropout index' which predicts, based on various objective factors, how many students from a government school will drop out in that year.

Keywords — School Dropouts, Governmental Schools, Statistical Analysis, Indexing

I. INTRODUCTION

In India 40% children dropout from their school before completing their middle standard education.[1] According to UNICEF, 80 Million out of the 200 million enrolled students drop out before the completion of their elementary education in India.

My model here allows me predict the number of students who will dropout from a particular governmental school in a year.

I first identified the key region for my research. The first phase of this research involved data collection in which surveys from various schools and households where children have dropped out from school from 2010 to 2018 were taken. The second phase of this research involved raw data observation. The third phase was data processing and cleaning of the data. The fourth phase comprised a thorough data analysis and key conclusions were drawn. Based on these conclusions, in the fifth phase, an index was created. The sixth and the final phase was the implementation of this index and recording the conclusions.

I chose the state of Uttrakhand as the region of my study because there has been a consistent and alarming increase in the percentage of dropouts in the years 2014-2017. [2]

Through my model, I was able to accomplish the following:

1. Identify what factors are the most influential while calculating the number of dropouts from a particular school.
2. Use statistical learning models to train an algorithm in order to assign 'weights' to these factors.
3. Index these factors based on their 'weights'.
4. Implement this index in multiple schools across Uttrakhand to help them optimize their facilities in order to increase the student retention rate.

II. LITERATURE SURVEY

According to the research done by Ekstrom, R., Goertz, M., Pollack, J., & Rock, D. on "Who drops out of school and why?", decisions to leave school are made by individuals in response to particular circumstances, and dropout prevention programs frequently address the problems of particular individual [5]. My aim is to analyse the extent to which factors that can be controlled by the school affect the number of students dropping out from that government school. Quantitative assessment of the school's influence on the chances of a student dropping out from that school helped me create an index which predicts the number of students dropping out from that particular school.

Robert B. Pittman's study on "Social Factors, Enrollment in Vocational/Technical Courses, and High School Dropout Rates" [6] and Laurier Fortin's research titled "A multidimensional model of school dropout from an 8-year longitudinal study in a general high school population" [7] provided some interesting pointers which assisted me in creating a survey for data collection process. Since none of these studies were based in India, my research gave out some very unique insights into the dropout problem in the country. The issues of tribal influence and the prevalence of child marriage practices affecting the dropout rate in the school are some unique factors which have been accounted for in my study and are exclusive to the country.

According to Kristoff Witte & Rogge (2013)[8], the issue of student dropouts presents negative consequences at three different levels. At the private level there are costs such as higher unemployment risks (Solga, 2002) [9], lower health status (Groot & Maassen van den Brink, 2007) [10] or less educated children (Alexander et al., 2001) [11]. Second, there are higher costs to society with greater risk of criminality (Lochner & Moretti, 2004) [12], less social cohesion (Milligan et al., 2004) [13] or a lower rate of economic growth (Hanushek & Wößmann, 2007) [14]. Finally, there are fiscal consequences due to lower tax revenues, higher unemployment allowances or health costs (Psacharopoulos, 2007) [15]. My research aims to help schools reduce such 'costs' by minimizing the number of students dropping out in a given year. Also, my index could work as a metric which indicates the student retention rate of a school. Such an index can help the government identify schools which require optimization to reduce the number of dropouts in that particular school and then my index would assist the government by pin-pointing the areas where the school can improve to increase its student retention rate.

III. PHASE 1 – DATA COLLECTION

First and foremost, I created a survey which would collect all the data I would need to conduct my research. The surveys have been attached in the appendix for your reference.

I conducted two different surveys, one for the school to fill and the other for the drop-out students to fill. Please refer to AP-1.1 for the survey provided to the students and AP-1.2 for the survey provided to the schools. A few values in the data obtained for schools were obtained by the method of random selection. For example, to obtain the 'Facility Rating', 50 students at the school were *randomly* selected and asked to provide a rating between 1 and 5. The average value of these ratings were taken and rounded off to make it a *discrete* factored variable.

For the students who dropped out, I recorded their basic background information. Next, my questionnaire enquired them about the distance from their home to school and their monthly average family earnings. I then allowed a free-answer space for the students to tell me about their reasons for dropping out. This was essential for the revision of my survey which was inspired by the feedback I received to this free-response questions. First, this survey was tested on a small sample and the most common reasons of dropping out were extracted from these free-form responses and put into the second edition of the survey. I repeated this process five times to construct the most objective survey possible. After these revisions, a few prominent reasons did come up. Female students were asked if the lack of female teachers in the school could be a reason for them discontinuing their studies. Students were asked if they currently worked for money or if they passed their last exams. The satisfaction of the students with the faculty and the facilities was also quantified.

For the schools I was surveying, I designed a totally different survey sheet to get to know the school environment better. I gathered information on the number of students in the school, the number of teachers and staff in the school and calculated the student-teacher ratio. I also gathered information about the average literacy rate by calculating the mean of mother's and father's level of education. The number of washrooms (for males and females), the number of computers and labs, and the average class size was recorded and combined with the average facility ratings by the students.

IV. PHASE 2 – DATA OBSERVATIONS

After the raw data collection and processing, I shortlisted the major factors that affect the rate of dropouts in a particular academic year in a governmental school in India.

A key factor that influences the number of dropouts is the average family income of the students' families. There is a clear distinction between those who continue their education in school and those who choose to drop-out due to monetary issues. It turns out that most of the children drop out because they would rather go and support their family by working as daily laborers, surviving on odd jobs or taking over their parents' mini-businesses. Hence, the average family income (in comparison with the average education costs) in the proximity of the school proportionally affects the number of dropouts that school would have in a year.

Another factor which is very prominent when we look at the data is the proximity of the major settlements from the school. There is a clear correlation between the distance of the students' houses from their schools and their chances of dropping out from that school.

The literacy rate in the settlements around the school affects the number of dropouts to a good extent. An increased literacy rate results in less people dropping out of schools. This can be seen in the correlation plot where the columns 'Medu' and 'Fedu' are mildly red, indicating a weak negative correlation.

Tribal influence around the settlements is a crucial factor affecting the number of dropouts from a governmental school. Statistics by the Ministry of Human Resources Development [3] indicate that the dropout rate in schools among 'Scheduled Tribe Category' students is 6.86% for Uttrakhand as compared to the 2.97% dropout rate among the 'General Category' students in schools across Uttrakhand. Survey results from the school-specific survey also revealed a general trend between the tribal influence in a school and the percentage of students dropping out from the school.

Another significant trend observed in the data is that a low number of female teachers results in a high number of girls dropping out from the school. A lack of female teachers is also correlates with a lack of school facilities for women. The provision of basic amenities such as the number and the condition of toilets for women also affects the number of female teachers and the number of female students in the school. This affects the overall facility rating of a school and thus, is a strong determining factor while considering the number of dropouts in a school in a given year.

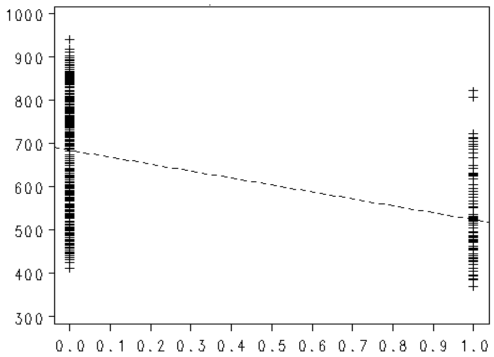
The prevalence of child-marriage practices around the region where the school is set up is also a major factor which specifically affects the number of girls dropping out from the school in a given year. This becomes very clear at a macro level when we compare the ratio of the average dropout rate of girls to boys in Rajasthan (a state where the malpractice of child-marriages is very common) to the ratio of the average dropout rate of girls to boys in the rest of the country. (1.247 in Rajasthan to 0.943 in the rest of India).[4] This pattern emerges primarily because the families of girls force them to discontinue their education so that the girls can take care of the households, while the boys are encouraged to study more to support their families and households.

V. PHASE 3 – DATA PROCESSING

Once all the data was collected, the quantifiable elements were sorted into an excel sheet and average ratings by the students were calculated and these scores were attached to the specific schools. Much of the data was converted into a either a 'binary' or a 'factored' form. For example, the data for the number of family members was converted into a binary form by labelling it as either LE3 or GT3 (less than or equal to 3 and greater than 3 respectively). The distance from home to school was converted into a factored form with 4 levels of 'traveltimes'. This data was then plotted and the correlation was calculated which considered the percentage of dropouts from the school in comparison with the severity of the factor. For example, I plotted the factor that influences the number of dropouts, with increasing severity, on the x-axis and the number of dropouts on the y-axis. Since there could be many other factors which influence the number of dropouts from the school and no single factor is solely responsible, I wanted to extract the most influential factors. Hence, for my indexation, I only considered those factors which had an absolute correlation value between 0.7 and 1.

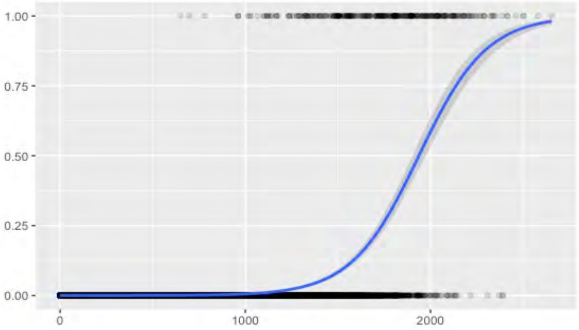
Using the primary data I gathered, I constructed a data frame and imported it into in the R Programming Language. This data was then 'cleaned' and had the unnecessary bits taken out from it. The data had a few of its constituents converted into factors with different levels, making the values *discrete* rather than *continuous.* This helps the machine learning code to understand categorical data in a much more efficient manner.

This data frame was then split into *train* and *test* data with a 70:30 ratio. The training data provides the machine with sample data to build a model on. The code was programmed to use statistical models in order to identify patterns in the data which creates an index of sorts by giving weightage to different factors which influence the chances of a student dropping out from that school. I first used a *linear regression* model to predict the chances of a student dropping out. A linear regression model constructs a linear curve which predicts the outcome based on a range of values which are not necessarily between 0 and 1, the binary results I needed for my results since the model was supposed to predict only 2 of the possible outcomes: either a student drops out or he/she does not drop out from the school.



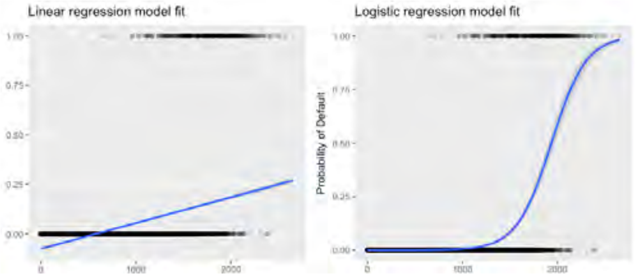
As the figure above illustrates, in binary categorical data (as in the case of a student dropping out or not), there are only 2 possibilities: 1 representing a student dropping out while 0 represents a student not dropping out. As in the case illustrated above, it would be very misleading to classify all values above 0.5 as 1 and all those below it as 0. We can see above that such a classification causes a huge margin of error. Additionally, such a regression model goes past the values of 1 and 0, giving us erroneous values which are difficult to classify.

When I came across this problem in my model, I improved upon it by implementing a logistic graph rather than a linear one.



The figure above depicts a *logistic model* graphed between the values of 0 and 1. The formula for a *logistic function*, $y = \frac{e^{f(x)}}{1+e^{f(x)}}$ where $f(x) = \hat{\beta_0} + \hat{\beta_1} X$, ensures that the value of the function stays between 0 and 1. However, this formula applies to a simple logistic model with a single variable as a predictor, thus requiring a single degree of X. The data which I wanted to process had multiple variables to be used as predictors and thus, the formula for my logistic function was much more complex with $\mathrm{p(X)} = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$ where $X = (X_1, \dots, X_p)$ are p predictors and $p(X)$ is the probability of an event happening. My algorithm constructed a similar logistic function which was then manually adjusted to fit the values between 0 and 1. My code then converted all the values below 0.5 to 0 predicting that the student will not dropout while marking all values greater-than or equal to 0.5 as 1, predicting that the student will drop out.

The comparison of the two graphs above summarises the difference between *linear* and *logistic* models.



Once the model was created, I tested it upon my test data, which was a compilation of random selection of 30% of my data set. The program was totally unfamiliar with the data and used it to predict whether the student would drop out or not. The results were then combined with the actual results and displayed on the screen.

To check the accuracy of the model, I calculated the *Mean Classification Error (MCE)* and created a *Confusion Matrix*. The accuracy, represented by $1 - MCE$, came out to be a staggering 96.85%. The *Confusion Matrix* for the test data is shown below:

	TRUE	FALSE
TRUE	108	0
FALSE	4	15

The accuracy rate according to this is comes out to be:

$$\frac{108+15}{108+15+4+0} 100\% = \frac{123}{127} \times 100\% = 96.85\%$$

Once the accuracy was tested, I drew certain conclusions based on the data.

To know the code for my statistical learning algorithm, refer to AP-1.3 in the appendix. The data for the same can be accessed through the Google Drive link in AP-1.7 in the appendix.

VI. DATA ANALYSIS AND CONCLUSIONS

Once the model was created, it was put through another program which calculated the *t-values* for each variable and the chances of a student dropping out and converted them to *p-values* to test and reject *null-hypothesis* H_0 that the variable does not affect the chances of a student dropping out and accept the *alternative-hypothesis* H_a that the variable affects the chances of the student dropping out at the 10%, 5%, 1% and 0.1% *significance levels.*

The results of the *hypotheses testing* have been put up under AP-1.4 in the appendix.

To summarise these results, the following *null-hypotheses* stand rejected at the:

10% Significance Levels –

1. Mother being a guardian
2. Age
3. Absences from school

5% Significance Levels –

1. Satisfaction with the teachers
2. Attending higher education
3. Paying for extra tuitions
4. The amount of study time
5. Father having a job in 'services'
6. Father having an 'other' job
7. Mother having a job as a 'teacher'
8. Mother having an 'other' job
9. Father's education level

1% Significance Levels –

1. Student being involved in ECAs
2. Student receiving family support

0.1% Significance Levels –

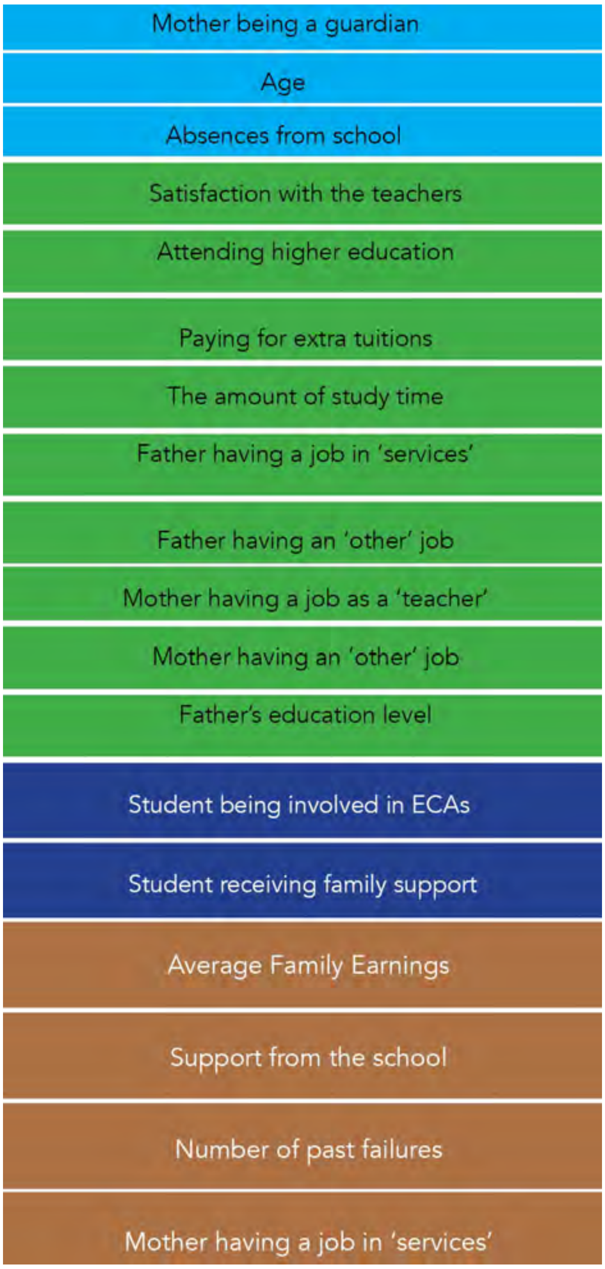
1. Average Family Earnings
2. Support from the school
3. Number of past failures
4. Mother having a job in 'services'

These *significance testing* results give us confidence intervals for our data and also create a 'hierarchy' of sorts which help us in visualizing our index. The hierarchy shows how the factors with the highest confidence interval levels (those at the 0.1% *Significance Level* have a 99.9% *Confidence Interval*, those at the 1% *Significance Level* have a 99% *Confidence Interval*, those at the 5% *Significance Level* have a 95% *Confidence Interval* and those at the 10% *Significance Level* have a 90% *Confidence Interval*).

VII. INDEXATION

While the *logistic model* accurately represents the index of whether a student would drop out of a school or not, representing it visually is a tough task. To visualise this indexing for the ease of understanding how the statistical learning model weighs the factors, I created a bar diiagram with the following weight distribution:

Factors at the 99.9% confidence interval were given a miltiplier factor of 9, those at the 99% confidence interval were given a multiplier factor of 7, those at the 95% confidence interval were given a multiplier factor of 5 and those at the 90% confidence interval were given a multiplier factor of 4.66. This can be visually represented by the bar depicted below:



In the stacked bar chart above, we can see that as the width of each plank increases, the weightage of each factor increases too and this is in accordance with the hypothesis tests we put our model through. This bar chart is helpful in visualising the weightage and summarising the conclusions of my report.

VIII. IMPLEMENTATION

To optimize the school's policies and facilities in order to reduce the number of dropouts, I developed school-specific policies by starting from the bottom of the stacked-bar chart ladder.

We couldn't do much about having children's mothers working in the services sector, however, the number of past failures was a very important factor. With the results of my research, I advised schools to increase motivation in the junior classes, revise their passing criteria and focus more on teaching younger students to reduce the average number of failures a student goes through. The schools were also advised to provide financial as well as mental support to the students, if the school was capable of doing so. This ensured that another important factor influencing the number of students dropping out from the school was taken care of. The average monthly family income is another influential factor, about which, nothing could be done on the school's side. The schools could revise their admissions policy by selecting students based on their economic background, but it would be unethical, unjust and against the objectives of a public governmental school.

The next factors that we looked upon were the ones where the hypotheses were rejected at a 1% significance level. The first one was if a student receives family support or not. To reduce the number of dropouts in this case, I suggested the schools to conduct sensitization workshops with the parents and families of the students about the importance of education and its advantages. The schools were also encouraged to expand their Extra-Curricular Activities base and some schools even decided to make it compulsory for students to take part in at least one Extra-Curricular Activity.

There are 9 levels (or factors) which influence the number of dropouts at the 5% significance level, however 5 of them are related to the students' parents' education level or job sectors, about which, not much can be done. Next up, we had the amount of study-time available to the students at home. For this, I advised the schools to hold counselling sessions with the students of the school and look at how students can find more time to study at home. The cost of extra tuitions was also seen as 'burden' on the students so the schools were advised to improve their teaching staff and hold professional training and development programs for the teachers at school since this would impact the students at multiple levels: it would improve the facility rating by the students, it would reduce the extra tuition costs that students pay, reduce the number of past failures, increase the satisfaction from teachers rating and lead to a reduction in the number of absences by a student. The students were also provided with workshops on the importance of attending higher education to motivate the students about pursuing further education. Teacher and facility satisfaction was measured at each school and the schools were advised to spend money on updating their infrastructure if they had a low facility rating. However, the schools were made aware about the fact that improving infrastructure is a secondary priority and the primary priority lies in the factors which have a lower significance level such as training teachers and sensitizing families.

The least priority was held by the age factor, the mother being a guardian and the number of absences from school. While nothing much could be done about the age and who the guardian of the student is, we realized that the number of absences from school can be reduced by implementing a 'minimum attendance policy' where the student had to attend school for a minimum number of days to be eligible to sit for the exams.

As of July 2019, this index has been able to guide schools to optimize their policies and invest their money in the most efficient manner which reduces the number of students who drop out from that particular school.

Appendix

Index:

Name

Absences

Marks

Health

① ② ③ ④ ⑤

Sex

Select ⌄

Address

○ Rural

○ Urban

Age

Family Size

Mother's Highest Education:

1. Primary
2. Secondary
3. High School
4. Undergraduate
5. Post-Graduate or Doctorate

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Father's Highest Education:

1. Primary
2. Secondary
3. High School
4. Undergraduate
5. Post-Graduate or Doctorate

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Mother's Job

Select ⌄

Father's Job

Select ⌄

Reason for choosing the school:

Select ⌄

Guardian

○ Mother

○ Father

○ Other

Does your family support your education?

○ Yes

○ No

Do you pay extra for tuitions at home?

○ Yes

○ No

Do you take part in extra-curricular activities?

○ Yes

○ No

Did you attend Nursery Education at this school?

○ Yes

○ No

How satisfied are you with the facilities and teachers at the school?

Rate: ① ② ③ ④ ⑤

What is your monthly family income (rounded to the nearest thousand)?

Did you drop out from the school?

○ Yes

○ No

Are you interested in taking up higher education?

○ Yes

○ No

Do you have access to the internet at home?

○ Yes

○ No

Rate the relations in your family

① ② ③ ④ ⑤

How many hours of free time do you get in a day?

Distance from Home to School

Study Time After School (In Hours)

Number of times you have failed a class in the past

Are you receiving financial support from the school?

○ Yes

○ No

School Address (Urban/Rural)

____

Tribal Influence (more than 40% are Scheduled Tribes) (Y/N)

____

Student to Teacher Ratio

____

Total number of students at the beginning of the year?

____

Number of students who dropped out?

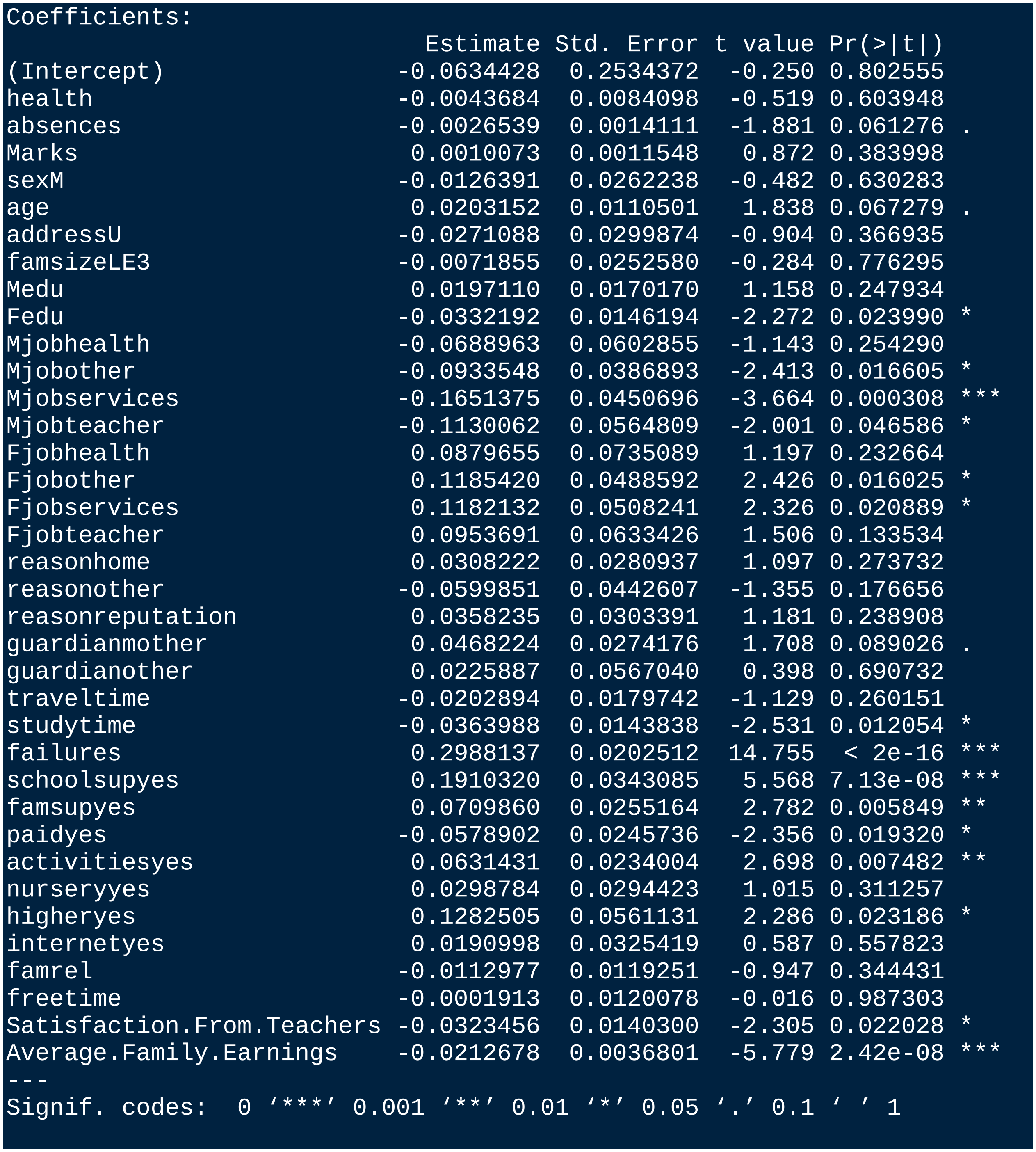
____

Code in R for the Logistical Model (AP-1.3):

```
#Calling the libraries
library(ggplot2)
library(dplyr)
library(corrgram)
library(corrplot)
library(caTools)
#Importing Data
ndf <- read.csv('dropout-data-marks.csv')
#Splitting Data into the Train and Test Sample
sample <- sample.split(ndf, 0.7)
train <- subset(ndf, sample==T)
test <- subset(ndf, sample == F)
#Creating a logistic model
log.model <- glm(Dropout ~ .,family = binomial(link='logit'),
data=train)
#Making predictions using the model on Test data
D.predictions <- predict (log.model, test, type='response')
#Adjusting the logistic curve to our data
#and converting the results to binary
results<-ifelse(D.predictions>0.5,1,0)
resultf <- cbind (results, test$Dropout)
resultf <- as.data.frame(resultf)
#Finding the Mean Squared Error
mse <- mean(sqrt((resultf$V2-resultf$D.predictions)^2))
print (mse)
#Finding the accuracy of the model
```

```
mce <- mean(results!=test$Dropout)
print (1-mce)
#Printing the predictions next to the actual test values
result
#Creating the Confusion Matrix
table(test$Dropout,results>0.5)
```

Model Summary (AP-1.4)

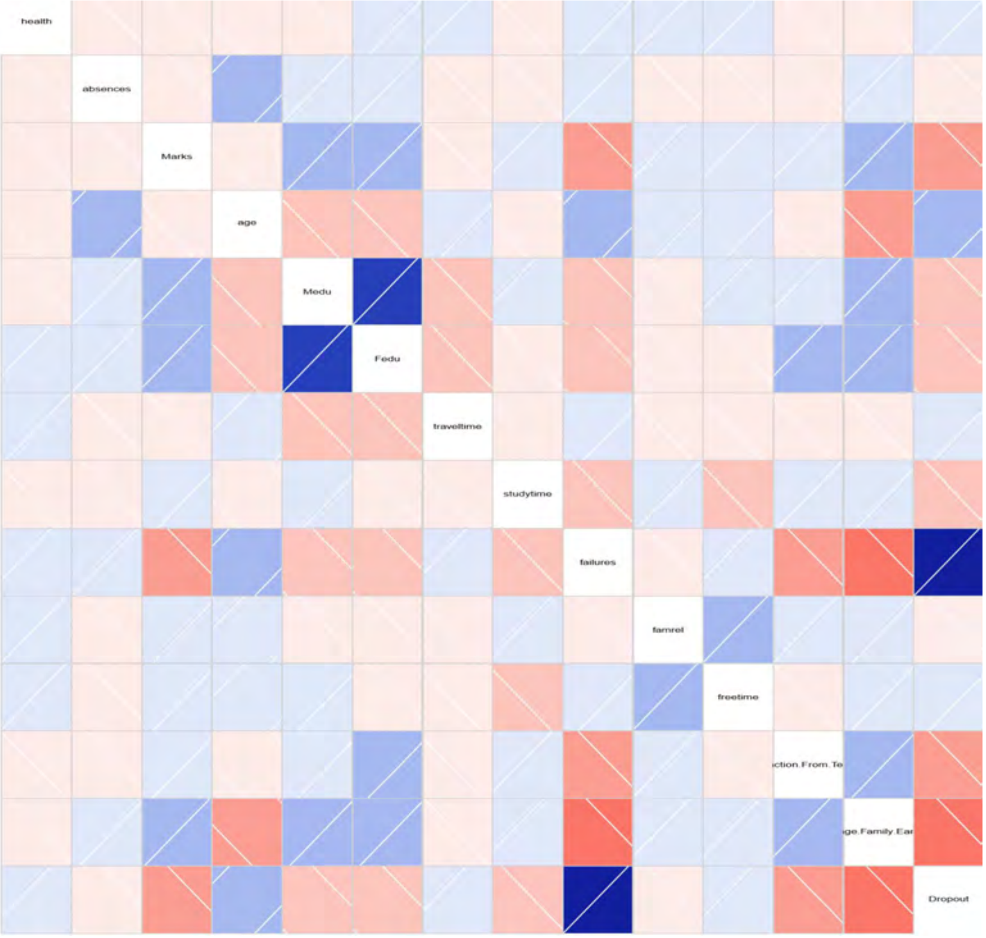


```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -0.0634428  0.2534372  -0.250 0.802555    
health                     -0.0043684  0.0084098  -0.519 0.603948    
absences                   -0.0026539  0.0014111  -1.881 0.061276 .  
Marks                       0.0010073  0.0011548   0.872 0.383998    
sexM                       -0.0126391  0.0262238  -0.482 0.630283    
age                         0.0203152  0.0110501   1.838 0.067279 .  
addressU                   -0.0271088  0.0299874  -0.904 0.366935    
famsizeLE3                 -0.0071855  0.0252580  -0.284 0.776295    
Medu                        0.0197110  0.0170170   1.158 0.247934    
Fedu                       -0.0332192  0.0146194  -2.272 0.023990 *  
Mjobhealth                 -0.0688963  0.0602855  -1.143 0.254290    
Mjobother                  -0.0933548  0.0386893  -2.413 0.016605 *  
Mjobservices               -0.1651375  0.0450696  -3.664 0.000308 ***
Mjobteacher                -0.1130062  0.0564809  -2.001 0.046586 *  
Fjobhealth                  0.0879655  0.0735089   1.197 0.232664    
Fjobother                   0.1185420  0.0488592   2.426 0.016025 *  
Fjobservices                0.1182132  0.0508241   2.326 0.020889 *  
Fjobteacher                 0.0953691  0.0633426   1.506 0.133534    
reasonhome                  0.0308222  0.0280937   1.097 0.273732    
reasonother                -0.0599851  0.0442607  -1.355 0.176656    
reasonreputation            0.0358235  0.0303391   1.181 0.238908    
guardianmother              0.0468224  0.0274176   1.708 0.089026 .  
guardianother               0.0225887  0.0567040   0.398 0.690732    
traveltime                 -0.0202894  0.0179742  -1.129 0.260151    
studytime                  -0.0363988  0.0143838  -2.531 0.012054 *  
failures                    0.2988137  0.0202512  14.755  < 2e-16 ***
schoolsupyes                0.1910320  0.0343085   5.568 7.13e-08 ***
famsupyes                   0.0709860  0.0255164   2.782 0.005849 ** 
paidyes                    -0.0578902  0.0245736  -2.356 0.019320 *  
activitiesyes               0.0631431  0.0234004   2.698 0.007482 ** 
nurseryyes                  0.0298784  0.0294423   1.015 0.311257    
higheryes                   0.1282505  0.0561131   2.286 0.023186 *  
internetyes                 0.0190998  0.0325419   0.587 0.557823    
famrel                     -0.0112977  0.0119251  -0.947 0.344431    
freetime                   -0.0001913  0.0120078  -0.016 0.987303    
Satisfaction.From.Teachers -0.0323456  0.0140300  -2.305 0.022028 *  
Average.Family.Earnings    -0.0212678  0.0036801  -5.779 2.42e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

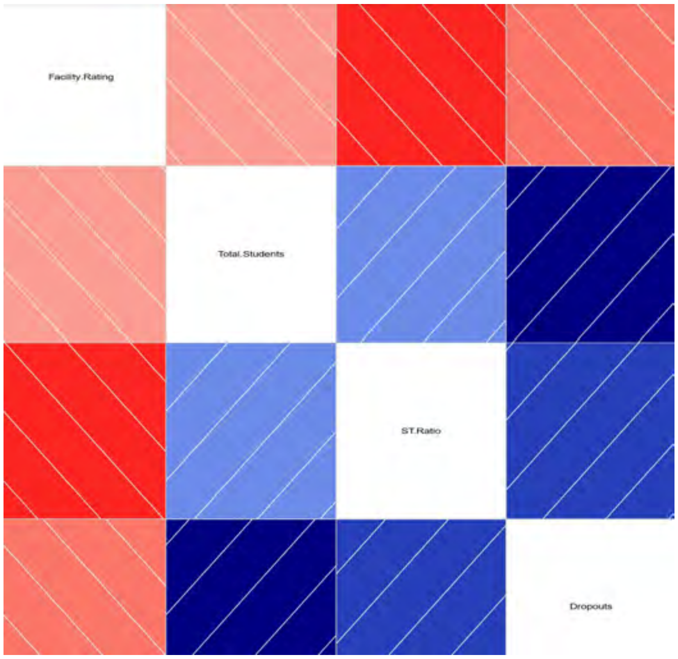
Legend for the variables (AP-1.5):

Sex - student's sex (binary: 'F' - female or 'M' - male)
Age - student's age (numeric: from 15 to 22)
Famsize - Family Size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Medu - Mother's Education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu - Father's Education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian - student's guardian (nominal: 'mother', 'father' or 'other')
traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) (converted from the distances between home and school)
studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures - number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup - extra educational support form school (binary: yes or no)
famsup - family educational support (binary: yes or no)
paid - extra paid tuition classes (binary: yes or no)
activities - extra-curricular activities (binary: yes or no)
nursery - attended nursery school (binary: yes or no)
higher - wants to take higher education (binary: yes or no)
internet - Internet access at home (binary: yes or no)
famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime - free time after school (numeric: from 1 - very low to 5 - very high)
health - current health status (numeric: from 1 - very bad to 5 - very good)
absences - number of school absences (numeric: from 0 to 93)

Data Plots and Diagrams (AP-1.6):



Correlation Plot for the different variables



Link to the data collected (AP-1.7):

https://drive.google.com/open?d=1eZMFsQvR1BsUk3taEROO0x_epMvmkUyc

Bibliography (AP-1.8):

https://www.hrw.org/news/2014/04/22/qa-talking-discrimination-and-school-dropout-rates-india

http://udise.in/Downloads/Publications/Documents/Flash_Statistics_on_School_Education-2016-17.pdf pg 307

http://udise.in/Downloads/Publications/Documents/Flash_Statistics_on_School_Education-2016-17.pdf pg 312

http://udise.in/Downloads/Publications/Documents/Flash_Statistics_on_School_Education-2016-17.pdf pg 307

5 Ekstrom, R., Goertz, M., Pollack, J., & Rock, D. (1986). Who drops out of school and why? Findings from a national study. Teachers Col- lege Record, 87(6), 356-373

6 Pittman, Robert B. "Social Factors, Enrollment in Vocational/Technical Courses, and High School Dropout Rates." *The Journal of Educational Research*, vol. 84, no. 5, 1991, pp. 288–295. *JSTOR*, www.jstor.org/stable/40539697.

7 Fortin, Laurier, et al. "A Multidimensional Model of School Dropout from an 8-Year Longitudinal Study in a General High School Population." *European Journal of Psychology of Education*, vol. 28, no. 2, 2013, pp. 563–583. *JSTOR*, www.jstor.org/stable/23421910.

8 De Witte, Kristof, and Nicky Rogge. "Dropout from Secondary Education: All's Well That Begins Well." *European Journal of Education*, vol. 48, no. 1, 2013, pp. 131–149., www.jstor.org/stable/23357050.

9 SOLGA, H. (2002) Stigmatization by negative selection. Explaining less-educated people's decreasing employment opportunities, European Sociological Review, 18, pp. 159-178.

10 Groot, W. & Maasen van den Brink, H. (2007) The health effects of education, Economics of Education Review, 26, pp. 186-200

11 Alexander, K. L., Entwisle, D. R. & Rabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school, Teachers College Record, 103, pp. 760-822

12 LOCHNER, L. & Moretti, E. (2004) The effect of education on crime: evidence from prison inmates, arrests, and self-reports, American Economic Review, 94, pp. 155-189

13 MIlLIGAN, K., MORETTI E. & Oreopoulos, P. (2004) Does education improve citizenship? Evidence from the United States and the United Kingdom, Journal of Public Economics, 88, pp 1667-1695

14 HANUSHEK, E. & Wobmann, L. (2007) The role of education quality in economic growth. The World Bank, Policy Research Working Paper 4122

15 Psacharopoulos, G. (2007) The costs of school failure — a feasibility study, European Expert Network on Economics of Education