ALY 6040  
Module 4: Students Scores Data

Professor: Justin Grosz

Chongfan Bao CPS in Informatics

Northeastern University

06/18/2021

**Abstract**

For this project, I will explore the student scores dataset in order to learn what the reason can lead to the failure of student on final grade. At the beginning, I will implement exploration data analysis to get general understanding to the data. Based on the understanding, I will try to build logistic regression model and support vector machine model and calculate their accuracy to judge if the model is better enough to predict the students’ scores.

*Keywords:* Logistic regression, Support vector machines (SVM), Exploration data analysis(EDA)

**Introduction**

Children's grades are always the most important thing for parents as well as the school board. Parents want their children to get as higher as possible score so that they go to a famous college and receive a decent job. The school board would like to increase the school fame by high acceptance rate and good education. Therefore, it is essential to supervise the scores of students to help students pass their exam and get a high score. In this dataset, there are many students failed to pass their final exams, so the school board want to know what kind of factors affect the students’ scores. Then, based on these reasons, they find solution to the best of their ability. I will implement exploration data analysis firstly to research potential problems or relations. I am also going to build logistic regression model and support vector machine model to predict the student if they can pass the final exam or not.

**Data overview**

According to figure 1, I observe that there are 395 observations and 31 variables among 13 numeric variables and 18-factor variables. Based on my experiences, I will select part of the variables to conduct EDA to explore the relationship between final grades with other variables. In the beginning, I want to know how many male and female students of different ages in the dataset. From figure 2, I notice that 16 years old students are the most. Only male students are in the school when they are over 20 years old. After this, I decided to look for a relationship with the final grade. In figure 3 shows the relation between the study time and final grade. I can find that no matter how many houses the student study a week. The students who are failed in the final exam are more than people who pass the exam. This is an interesting thing. Normally, if the student works hard, they are easier to pass the exam. So, we can discuss further in future work. From image 4, I also notice that students who failed in the final exam are the most when no one absents the course. If there are at least 24 students absent from the class in the school, the fail rate is 100 percent. Then, I discuss the situation under extra education support and family education support. Based on figure 5 and 6, similarly, even if the parents help their children study, they still fail in the final exam. Finally, I analyze the relationship between final grade with students’ desire for education and health status. I observer from figure 7 and 8 that when the students don’t want to accept higher education, they must fail the final exam. Most students are the healthiest failed the final exam. All in all, most students in the school failed to pass the class no matter how to work hard they are. Hence, I think maybe the final exam or the course is too difficult or the teacher is too strict for scores.

# Data preprocessing and modeling

Given exploration data analysis, I have a general understanding of this data. It is times to implement data cleaning and build model for further research. Initially, I inspect the missing values. Luckily, I find there is not any missing value. Them, I look if there is duplicate data. It doesn’t have any duplicate values either. In the end, I try to check the outliers. From image 10, I observe that the “abcenses”, “failures”, “studytime”, and “famrel” variables have outliers. Particularly, the “ absences” variable have the most outliers. Nevertheless, I find that there is no need to clean up these so-called outliers. According to the data explanation file, I find all data are located the specified range. So, I don’t need to do any data cleaning.

I decide to plot a correlation matrix to look which variable has a great impact on “final grade”. However, the data include many factor variables so that I have to convert the factor to numeric. I leverage the one-hot encode method to transfer all factor into binary variables and keep same format for numeric variables. Another good reason is that the logistic regression model is better for binary variable. From image 11, I notice that the correlation of variable is not so obvious. Only the “workday alcohol consumption (Walc)” and “weekend alcohol consumption (Dalc)”, “ mother's education (Medu)” and “father's education (Fedu)” exit little strong correlation. That we discussed variables with “final grade” before is also not so clear.

Given this, I prefer to implement feature engineering to find the rank of influence of variable to the final grade. According to the K- nearest neighbors' algorithm, I start to build our model and estimate each feature influence on the final grade. Then, I will build the logistic regression to predict the outcomes. Firstly, I change “Fail” values of the “final grade” variable to 0, “Yes” to “1”. Then, I again convert the “0” and “1” as factors. This is logistic regression requires binary variable. I fit a logistic regression model and obtain a vector of the predicted probability of each observation based on the model. Next, I create vector of class predictions based on probability values and assign to class 0 if the probability is <= 0.5, and assign to class 1 if the probability is >0 .5. This is in order to get the accuracy rate and error rate of the model so that I can judge if the model is better enough to predict the student if they can pass the final exam or not.

Based on feature selection outcomes, I am ready to select variables that have a great impact on final grade to build support vector machine model. At the beginning, I separate the data into training data and test data. The former is for training model, the latter is for testing the accuracy of the model. I will use the svm() function to fit the model with specified values of the parameters. For example, here I fit an SVM with a radial kernel, using cost=1 and gamma=1. Next, I leverage the tune() function performs 10-fold cross-validation to select from a range of specified parameter values. Finally, I will draw predictions on the test set using the best model obtained from the training set and access the accuracy.

**Data Analysis**

In the figure 12, I notice that the “Medu” has the greatest impact on “final grade” followed by “failures” and “Walc”. I also find that “famsup” and “ paid” seldom affect the “ final grade”, which means even if the parents help their children and extra paid classes within the course subject, most of them still cannot get a good scores.

In the figure 13, I conduct logistic regression on all variables and final grade. I can know that when the probability is <= .5, there are 316 values; if the probability is > .5, there are 79 values. Finally, I find our model has 80.5%1 accuracy rate and 19.49% error rate, which means it could be a good model to predict the students’ final grades.

In figure 14, when I use the training data set to train the SVM model, I know that there are 197 support vectors and 2 classes. After conducting 10-fold cross validation, I find that when cost parameter is equal to 0.01, the model has best performance. Its error value is 0.22. So, we choose the best model to test, and find that the support vectors become 101 and 71.72% accuracy rate and 28.28% error rate.

**Conclusion**

In this module, I find that most student could fail to pass the test even if they spend at least 10 hours on study every week or they participate in extra education support or get help from their parents. In summary, the number of students failed the final grade is bigger than the pass in Math and Portugal courses. This is a strange and interesting discovery. Due to limited information, I cannot find the precise reasons. Furthermore,the school board want to increase the pass rate that they can recruit students whose parents have high-level education and less drink alcohol. Or, they should strictly keep the absences rate at low level.

**References**

Classifying data using Support Vector Machines(SVMs) in R. Retrieved from https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/

James, L., (2018). Support Vector Machines in R. Retrieved from https://www.datacamp.com/community/tutorials/support-vector-machines-r

**Appendix A**

**Figure1:** *Data summary*

Text

Description automatically generated

**Figure2:***The odor variable distribution under the circumstance of the class variable*

*Chart, histogram

Description automatically generated*

**Figure3:** *Relationship between study time and final grade*

*Chart, bar chart

Description automatically generated*

**Figure4:** *Relationship between absences and final grade*

*Chart, histogram

Description automatically generated*

**Figure5:** *Relationship between school support and final grade*

*Chart, bar chart, box and whisker chart

Description automatically generated*

**Figure6:** *Relationship between family support and final grade*

*Chart, bar chart

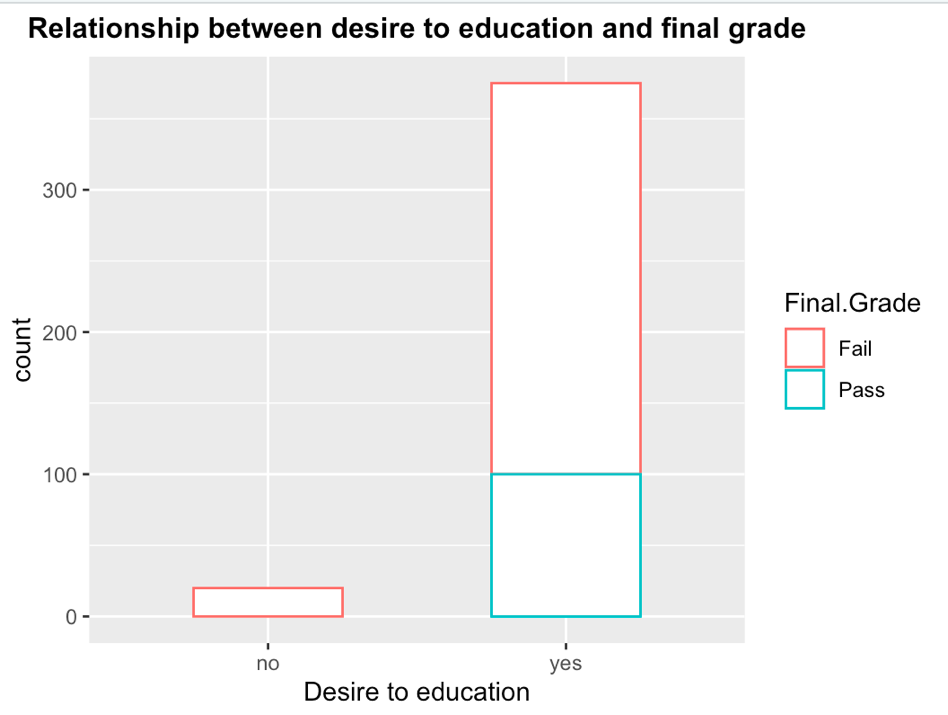
Description automatically generated*

**Figure7:** *Relationship between desire to education and final grade*

*Chart, bar chart

Description automatically generated*

**Figure8:** *Relationship between health and final grade*

**

**Figure9:** *Missing value and duplicate check*

**Table

Description automatically generated with medium confidence**

**Figure10:** *Outliers inspection*

*Graphical user interface

Description automatically generated*

**Figure11:** *Identify the toxicity of mushrooms by mushroom characteristics*

*Chart, scatter chart

Description automatically generated*

**Figure12:** *Feature rank*

*Chart

Description automatically generated*

**Figure13:** *Logistic regression*

*Table

Description automatically generated*

*Graphical user interface, text

Description automatically generated with medium confidence*

Appendix B

R code

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated with medium confidenceText

Description automatically generatedGraphical user interface, text, application

Description automatically generated