Fall 2018

IST 707 Data Analytics Project Final Report

Sandya Madhavan

**Student Performance Data Analysis**

**Introduction:**

These are the days where data mining is used across industries to analyze data, find insights and improve efficiency. This project focuses on analyzing the student performance data of two Portuguese schools. The main aim of the project is to understand the factors that contribute to student achievement and provide recommendations for improving the performance of low scoring students.

Data mining and visualization techniques were utilized to figure out insights from the dataset. After the initial exploratory analysis and significant takeaways from the dataset, predictive models were built and tested to predict the final grade of the students.

**Goals:**

The following are the major goals of the project

1. An exploratory study on the dataset.
2. Identify the attributes that affect the grades of the students.
3. Build and test predictive models to predict the final grade of students.
4. Provide Recommendations based on analysis to improve the scores of low scoring students.

**Tools:**

* R studio
* Weka

**Data Acquisition:**

The student performance dataset is obtained from UCI Machine learning Repository

Url: [https://archive.ics.uci.edu/ml/datasets/student+performance#](https://archive.ics.uci.edu/ml/datasets/student+performance)

There were two datasets, one each for Math and Portuguese subjects. Both the datasets summed up to 1044 rows and 33 attributes

Some of the attributes are listed below

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)   
2 sex - student's sex (binary: 'F' - female or 'M' - male)   
3 age - student's age (numeric: from 15 to 22)   
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)   
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

31 G1 - first period grade (numeric: from 0 to 20)   
32 G2 - second period grade (numeric: from 0 to 20)   
33 G3 - final grade (numeric: from 0 to 20, output target)

**Data Preprocessing:**

The following steps were done for data preprocessing.

1. Datasets were loaded in to R studio as CSV’s files.
2. Both the datasets were merged into a single dataframe and extra columns were added for ‘subjectname’ and ‘id’. This was done to uniquely identify an entry for a student and a subject.
3. Datatype of all fields in the dataset was reviewed and changed appropriately.
4. Dataset was clean with no N/A ‘s or blank values.
5. Fields G1, G2, G3 were converted to a scale of 0-100 from 0-20
6. ‘Category’ column was introduced for analysis purpose. Further details will be explained in section Analysis 2

**Initial Exploration:**

The loaded dataset was analyzed in R to figure out the key statistics and patterns that exists in the dataset. Some of the key findings are listed below:

1.Total no. of students are 1044.

2.56.6% of students are Female.

3.Age distribution is 15 to 22.

4.69.7% of students have Mother as Guardian.

5.59.6% percent of students have very less travel time to school.

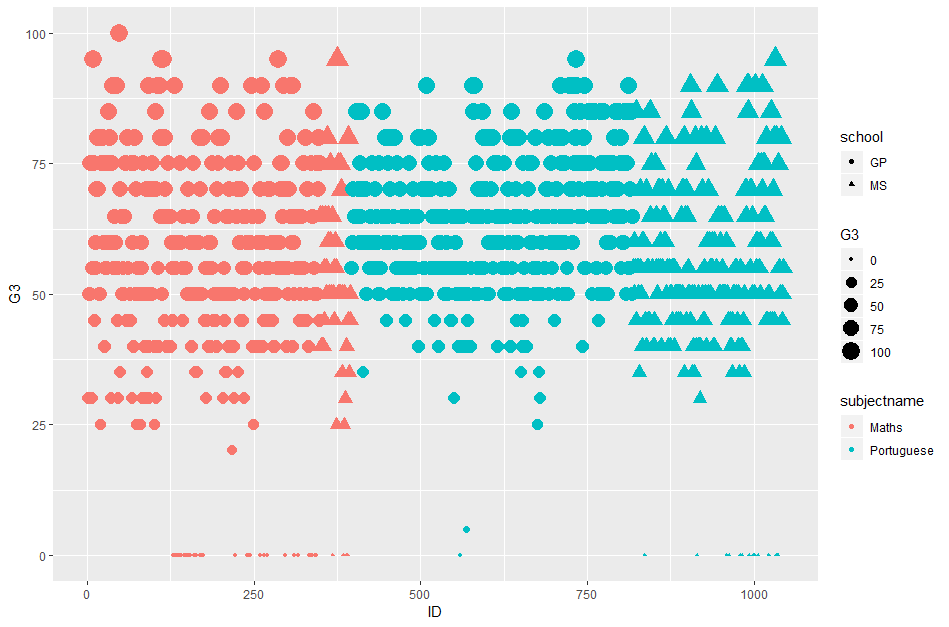
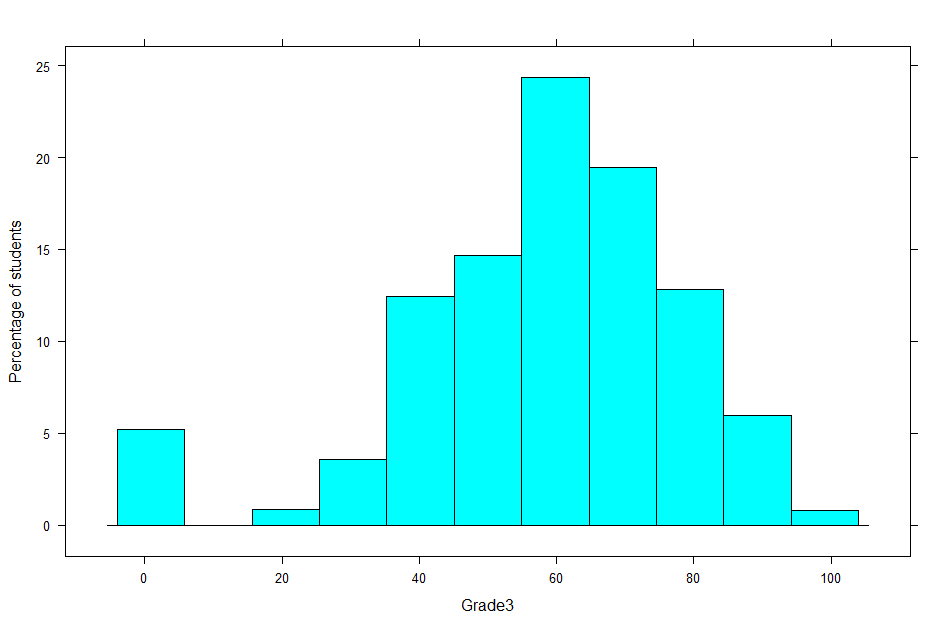
6.82.4% of the students have zero failures.

7. 88.4% student’s parents are living together.

8.64.4% of the students have no romantic interests.

9.Most students study Portuguese and in school GP.

10. Final score has a normal distribution: More than 70 percent of the students score between 45 and 75.



*Normal distribution of Grade* 3 Students distribution in school, subject and Grade 3 ranges

**Analysis 1: Significant change in score**

I was interested to know whether students had a significant increase / reduction in grades and the reasons for the same

**Assumption**: A minimum increase/decrease of 10 marks is considered a significant change.

*Question 1) Has there been a significant improvement/ reduction in Grade 3 Compared to the other tests. What is the percentage of students with significant score change?*

For this, new columns were added in the dataset to calculate the differences in grade. The following were the results after the analysis.

|  |  |  |
| --- | --- | --- |
| **Tests scores** | **Increase** | **Reduction** |
| **Grade3-Grade 1** | 22.10% | 10.20% |
| **Grade3-Grade 2** | 7% | 4.50% |
| **Grade2- Grade1** | 12.45% | 4.3% |

*Question 2) What parameters possibly affect the score change?*

A deeper analysis resulted in the following

**Score Reduction:**

* Study time is just 1- 2 hours for more than 86 percent of the students who lost marks significantly
* No School support –I.e. no extra educational support for more than 90 percent of students who lost marks significantly

**Score Improvement:**

Study time is 2-4 hours for more than 70 percent of the students who gained marks significantly

**Conclusion:**

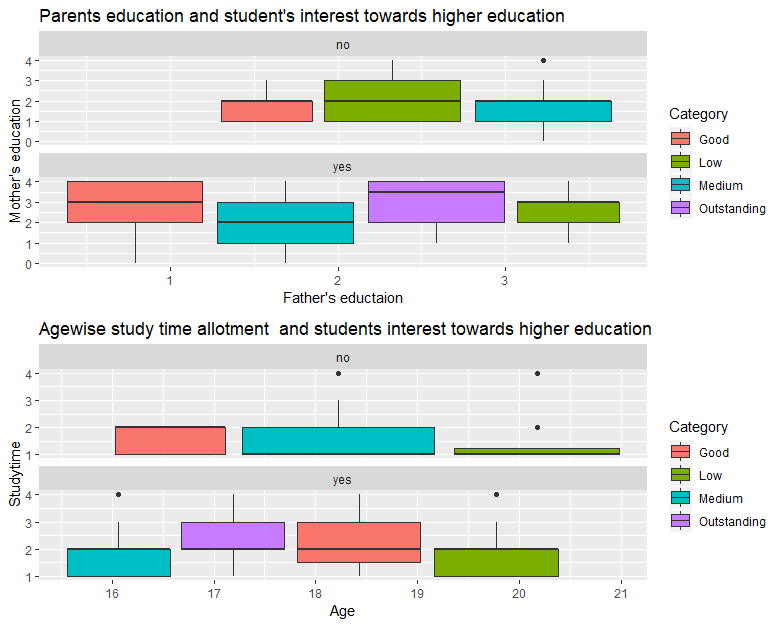
Except for the data mentioned in Increase/Reduction table, Grades are consistent in all the tests. Around 68 % of the students obtain consistent scores.

**Analysis 2: What factors influence Final score?**

In order to analyze and identify the factors that contribute to final grade, I added a new column ‘Category’ based on the following assumptions:

‘Category’: Outstanding: More than 80 marks, Good: 60-79 marks, Medium:40-59 marks, Low: less than 40 marks

Using ggplot and facetwrap, I visualized the data and found the following insights



*Graph 2: Student’s age Vs Study time -Grouped by student’s higher education interest*

*Graph 1: Father’s education Vs. Mother’s education -Grouped by student’s higher education interest*

From Graph 1:

1.All ‘Outstanding’ students are interested in higher education.

2.More than half of ‘Low’ performing students are not interested in higher education.

3.Parents education:

a)’Good’ students are not interested in higher education if their Parent’s education is levels 1-2.

b)’Good’ students are interested in higher education if at least one parent’s education is more than level 2

From Graph 2:

‘Low’ category students that are more than 19 years old and interested in Higher education don’t spend more than 2 hours of study time.

**Conclusion:**

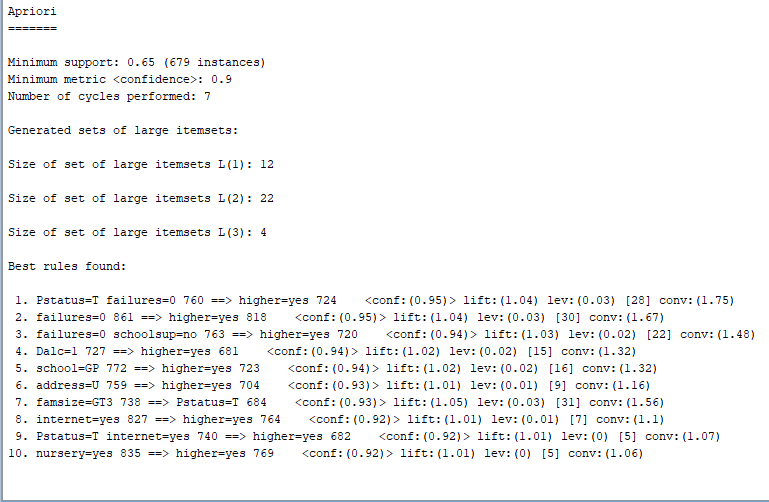
A student’s Interest in higher education, Parent’s education and student’s age affect a student’s grade significantly.

**Association rule mining:**

Weka was used to find the associations that exists in the dataset. Most of the rules had ‘higher’ as a component in it which again proved that a student’s interest in higher education plays a significant role in test grades.

Parameters: Minimum support-0.65, Confidence-0.9

*Association rules generated from Weka*



**Predictive modelling:**

The dataset was loaded into Weka to build and test regression models on the field ‘G3’ which is the final test score. ‘G3’ was predicted with three models using multiple algorithms

Model 1: G3 based on past performance i.e. G1 and G2

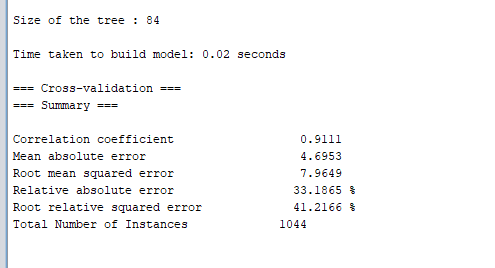
Model 2: G3 based on parameters other than G1 and G2

Model 3: G3 based on all parameters.

Algorithms chosen: Linear Regression, K-Nearest Neighbor, SVM for regression, Decision Tree.

The complete details of each model and corresponding algorithm are furnished below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Test option** | **Attribute used as class** | **RMSE** | **Parameters** | **Variables used** |
| Linear Regression | cross validation folds-10 | G3 | 7.9 | Classifiers->Functions->Linear Regression: eliminate colinear attributes-TRUE | G1, G2 |
| k-Nearest Neighbour | cross validation folds-10 | G3 | 8.25 | Classifiers->Lazy->IBK: Distance weighting-No distance weighting | G1, G2 |
| SVM for regression | cross validation folds-10 | G3 | 8.07 | Classifiers->Functions->SMO reg: filter type=normalize training data | G1, G2 |
| Decision Tree | cross validation folds-10 | G3 | 8.40 | Classifiers->Trees->REP Tree: MinNum:2 | G1, G2 |
| Decision Tree | cross validation folds-10 | G3 | 17.41 | Classifiers->Trees->REP Tree: MinNum:2 | Other demographic variables except g1 and g2 |
| Decision Tree | cross validation folds-10 | G3 | 7.96 | Classifiers->Trees->REP Tree: MinNum:2 | All variables |



*Decision Tree Model generated from Weka-Summary*

Linear regression and Decision Tree algorithms gave least RMSE’s. I prefer Decision Tree algorithm as it has considered all the variables.

**Final Conclusion and Recommendations:**

1)The following attributes affect the final grade of a student

* Study time
* Mother’s education
* Father’s education
* Interest in higher education
* Age

2)Students should be encouraged to spend more study times.

3)Students should be motivated to develop an interest in higher education for better scores.

4)70 percent of the low scoring group are in Math class. Schools may provide extra coaching assistance for Math classes.

5)At a school level, School ‘GP’ has the most ‘low’ scoring group (69%). The school may incorporate extra coaching classes.

**References:**

* <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>
* <https://datawookie.netlify.com/blog/2013/05/plotting-categorical-variables/>
* <https://stat.ethz.ch/R-manual/R-devel/library/base/html/Extract.data.frame.html>
* <https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/>

*Assumption: In this paper, the terms ‘score’ ,’marks’ and ‘grade’ are interchangeable and have the same meaning.*