Sergio Figueroa

Professor Batarseh

CIS 3252

Fall 2019

Student Performance Analysis

Introduction

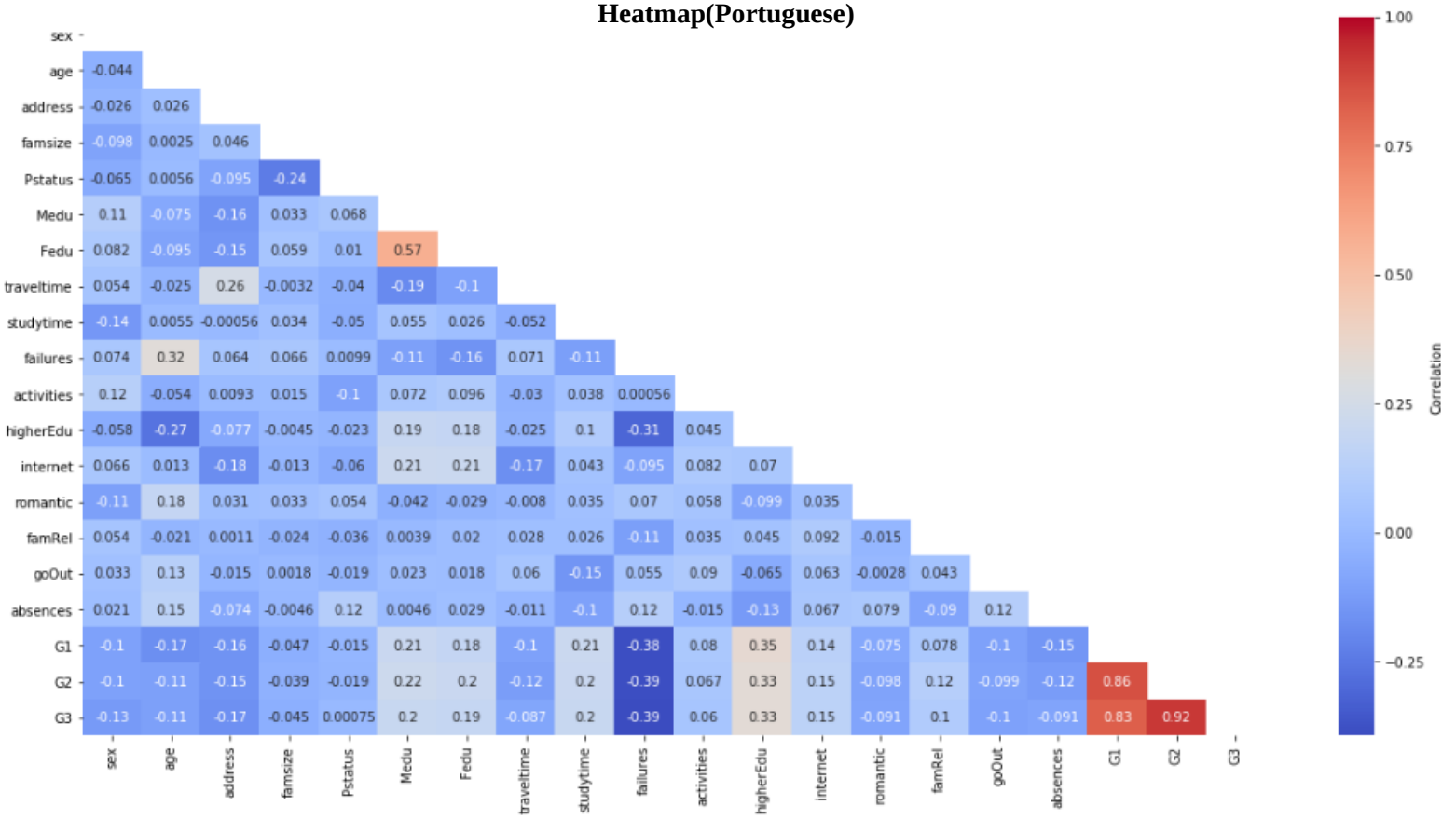
Education plays a critical role in changing someone's life; it is the foundation on which we build everything else. The knowledge and insight that a proper education provides is often the key to success later in life. Unfortunately, the education system is exceptionally complex which makes it difficult to remedy a faulty one. There are countless factors in play that can hinder a child's education, from the degree of coursework in a school to their relationship at home. A common approach to improving a school that is not up to par is to fund money into better text material and faculty – an easy fix, but not an efficient one. The cost of new teacher salaries and textbooks is much greater than an afterschool tutoring program, and it may not even alleviate the problem. With the rise of interest in BI tools and Data Mining, we can gather data around the world to extract valuable information from, such as trends and patterns, and use it to gain insights into better strategic decision making.

As a student that has experienced difficulties in my education, I understand the impact that demographic and social factors have on a child's schooling. I grew up in poor conditions with very few resources and relied on my education to bring me success in life. Any opportunity to better myself that my public school offered, I took it – GATE and C.H.A.M.P. were just a few of them. It is important to understand the severity that a situation has on a child's upbringing and work to improve it. School systems can learn more efficient methods of child development through an analysis of these factors and how they impact a student's performance.

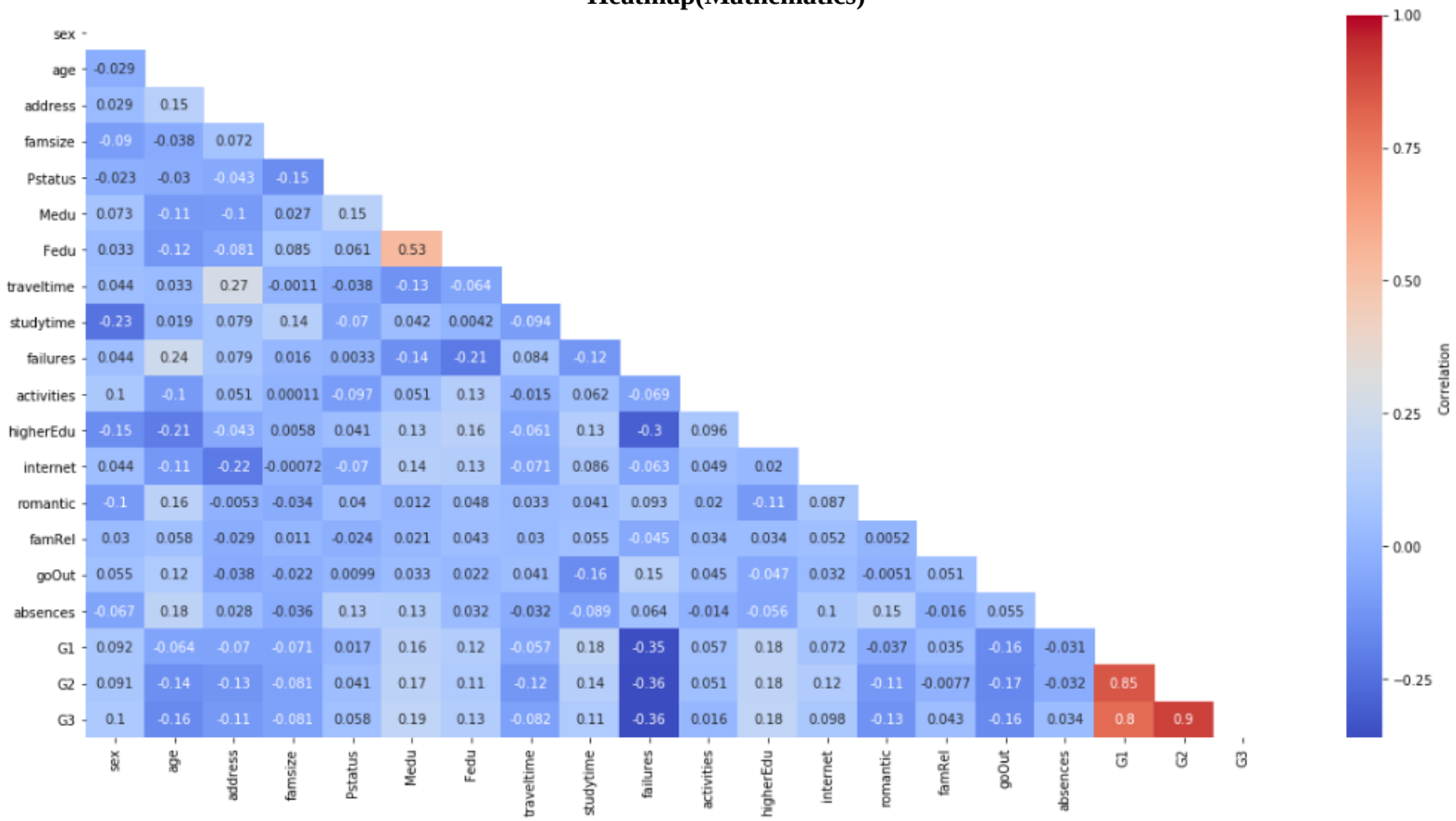
In this paper, I will use a dataset collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal by the University of Minho[1]. This data was built from two sources: school reports and questionnaires. The surveys were designed with closed questions based on several demographic (parent's education/marriage status), social/emotional (relationship status), and school related variables (failures, absences) that may affect a student's performance (test scores). During the cleaning process, some features were left out – this was mainly for clarification and ease-of-use. The 5-level classification (1 – very bad to 5 – very good) was replaced with a binary classification (0 – 1 to 3, 1 – 4 to 5) and several attributes were dropped due to poor implementation (parent's job, student's guardian, reason of school, extra paid class). This way, a multivariable regression model can be fit to predict a student's test scores and we can evaluate the impact some attributes have on each other and in the target variable. The final version of the dataset includes 20 attributes for both sets, 395 records for Mathematics scores and 649 for Portuguese language scores. These two subjects are core subjects in most public education systems and each student is evaluated in three periods (G3 is the final grade). The goal of this study is to answer questions such as: Who is likely to have better attendance? How likely can our model predict scores? Which factors affect student achievement the most? Analyzing this data using BI techniques will help schools make better decisions in order to improve student education.

Data Analysis

Heatmap(Portuguese)



Heatmap(Mathematics)



Portuguese

```
Summary Statistics for Studytime:
        Avg: 0.2033898305084746
        Stdev: 0.40252007074704804
        Min: 0.0
        Max: 1.0
```

Mathematics

```
Summary Statistics for Studytime:
        Avg: 0.23291139240506328
        Stdev: 0.42268626153803235
        Min: 0.0
        Max: 1.0
```

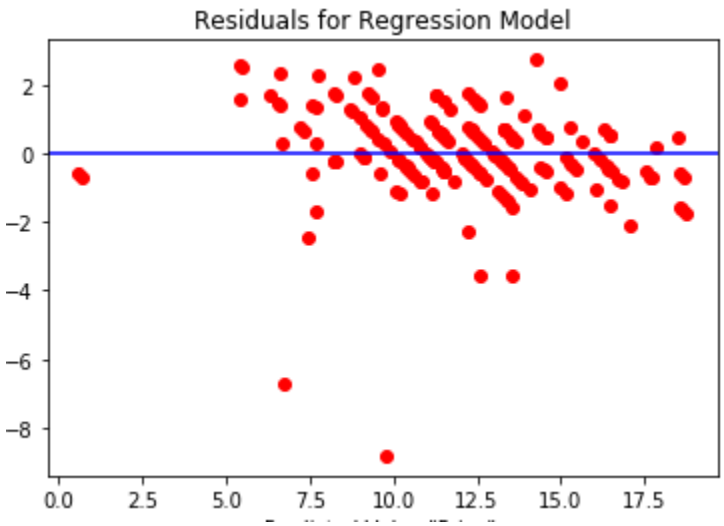
With so many factors in play, it is difficult to read into which ones affect the student's performance more than others. Generally, we will find that the students with more failures in previous classes will have lower test scores – this is also true for students with no interest in higher education and lower amounts of study time. However, in both data sets for mathematics and Portuguese there was an alarming correlation between urban/rural addresses, relationship status, time spent going out, absences, and performances on scores. The correlation between previous scores (G1, G2) is exceptionally high, but it is more important to predict G3 from demographic/social/school factors; previous scores will always be the best way of predicting final scores.

Methods

The methods used are multivariable regression after preprocessing. Much of the source data was in five-level classification, which can only be done with high performance tools. Since I am limited to Python, I cleaned up the data into binary classification in order to implement a regression model. This is best suited for this data because it is easy model. The main component of my analysis is the heatmap – this way we can clearly see which attributes are highly correlated between each other.

Results

Portuguese



```
Test Score: 0.8248849252037962
```

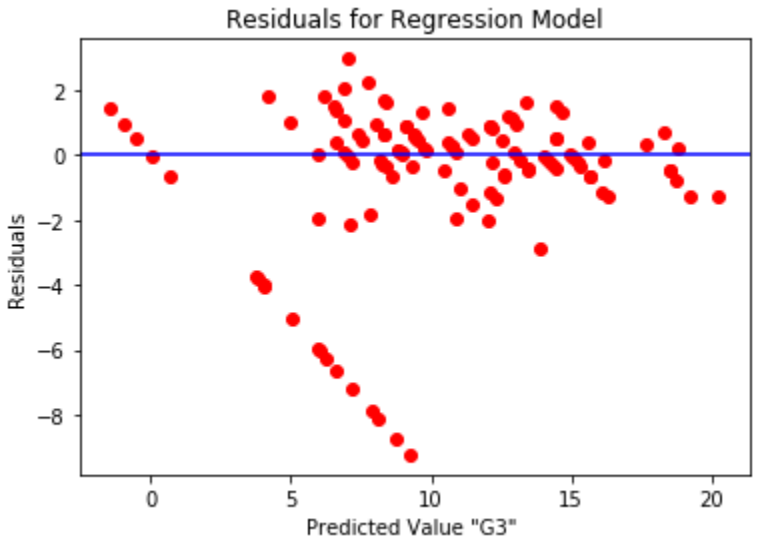
```
from sklearn import metrics
```

```
#Two very important metrics for multi-variable regression
print('MSE :', metrics.mean_squared_error(yTest,price_y_pred))

print('RMSE :', np.sqrt(metrics.mean_squared_error(yTest,price_y_pred)))
```

```
MSE : 1.7194941792796157
RMSE : 1.3112948483387006
```

Mathematics




```
Test Score: 0.7919400289474993
```

```
print('MSE :', metrics.mean_squared_error(yTest,price_y_pred))

print('RMSE :', np.sqrt(metrics.mean_squared_error(yTest,price_y_pred)))
```

```
MSE : 5.810810416734684
RMSE : 2.410562261534575
```

My analysis shows that the regression model used in this analysis is a good fit in predicting a student's G3 score. With an accuracy of 80% for mathematics and Portuguese, along with very low MSE and RMSE values, it is safe to say that a student's performance can be predicted by the included attributes.

Discussion

I believe the following results were found due to the high impact that demographic/social/school factors have on a student's performance. In my analysis, I found a connection between certain demographic and social factors – students who lived in rural areas had a much higher travel distance, absences, failures, and different social statuses. The ones who lived so far away ended up spending much more time out with friends, which may have been why the average study time was so low. Not many students were interested in a higher education, mostly due to their parent's level of education, and unfortunately, we cannot change this. What we can change, is the recognition of a college degree. The importance of a higher education should be instilled in these students.

Conclusion

Who is likely to have better attendance? How likely can our model predict scores? Which factors affect student achievement the most? These were the questions I examined within my paper. The impact of these factors stands out in the performance of these children, and it is important for schools to evaluate this information to make better decisions. The lack of study time was quite alarming to me, and I would advise a better tool for allowing these students more time devoted to their performance.