Students' performance

Vinay Sanga

1 Introduction

Nowadays, Moodle is the first choice for conducting tests for the students and organising their learnings. It is essentially a learning management system that makes organising grades and course materials easier for students and the instructors alike. You might be surprised to know that some tests' grades are more important than the others and play a decisive role in calculating the overall grade of a student. What if we could judge a student's performance based on the grades obtained in some of the subjects? To do that, firstly we need to collect the data from an online learning management system and then use Machine Learning to find the best features(tests, assignments, etc.) which have a very high impact on the final grade. With this information, we can create models that can predict students' performance in a way that has the least errors or are as much close to the actual scores as possible.

2 Data Processing

The data contains details about 107 students based on the following parameters:

- Status0 - course / lectures / content related
- Status1 - assignment relate
- Status2 - grade related
- Status3 - forum related

There are 9 grades related to different quizzes, mini projects, etc. Apart from these, there are 36 logs related to the status above.

The data has no null values and there are a total of 47 features. Out of these 'Week1_Stat1' has positive correlation with every other variable.

2.1 Feature selection

There are a lot of features in this dataset, but not all of them are important for the final prediction outcome. To undertake feature selection, we used Random Forest feature importance to find the most decisive features in the dataset. The top 10 features are shown in Fig. 1.

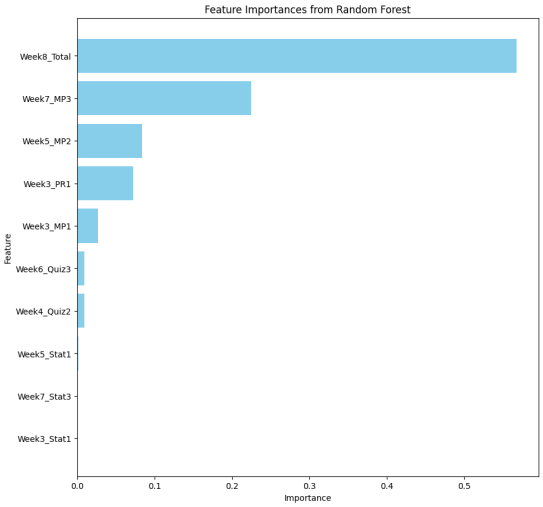

Figure 1: Top 10 features

Using this figure, we can easily see that there is a stark difference in the feature importance, with the top 5 features being prominent and others being minuscule in the importance. We can see that almost 10% of the features are actually decisive and provide significant understanding of the data.

3 Data Analysis

We used statistical measure to analyse the data as they were clear enough to draw inferences from. There were no data visualizations because the main focus of this report is on the performance evaluation of models with respect to the features.

3.1 Statistical analysis

Based on the statistical analysis of the data, we could find some interesting facts as follows:

- Week 4 Stat0 Distribution - The distribution of Week 4 Stat0 scores varies, with some students scoring close to 0 and others reaching higher values, indicating a wide range of performance in this category.

- Consistency in Week 3 Stat1 - Many students have consistently scored 0 in Week 3 Stat1, which might indicate either a lack of participation or a specific trend in this category.
- Diversity in Week 7 Stat3 - The scores show more diversity, with scores ranging from 0 to higher values, indicating that different students performed differently in this category.
- Week 2 Quiz1 Scores - Some students scored low or even 0 in Week 2 Quiz1, while others achieved higher scores, highlighting the variability in quiz performance.
- Week 5 PR2 Scores - Most students seem to have performed well in Week 5 PR2, as scores are around 5.0 for many entries.
- Variability in Week 6 Quiz3 - Week 6 Quiz3 scores show significant variation, suggesting that this assessment might be more challenging or that students' performance varied widely.
- Variability in Week 8 Scores - They vary widely among the individuals, ranging from 0.0 to approximately 99.71. This indicates that some participants performed exceptionally well, while others did not score at all or had lower scores.

3.2 Performance evaluation

We created two models using Linear Regression and Random Forest Regressor to analyse the performance. The metric used for analysis the model performance is R2 score. It can be see from figure 2 that:

- Linear Regression - The score increases till top 3 features, after which there is a sharp decline in the model performance.
- Random Forest Regressor - The score starts higher than Linear Regression but drops for top 2 features, then gradually increasing till 4 and starting to decrease after that.

This shows that Random Forest works better than Linear Regression at handling the variations in features. Also, it is overall a better model than Linear Regression for this dataset. The optimal number of features in the dataset is 3 for which both the models perform good.

4 Conclusion

There were a lot of features in the data which did not contribute to the overally efficacy of the model. It was observed that only 3 features played the determining role in predicting the final grade of the students. The scientific bottlenecks were related to the sheer number of numbers, which we overcame by using statistical analysis to perform data analysis.

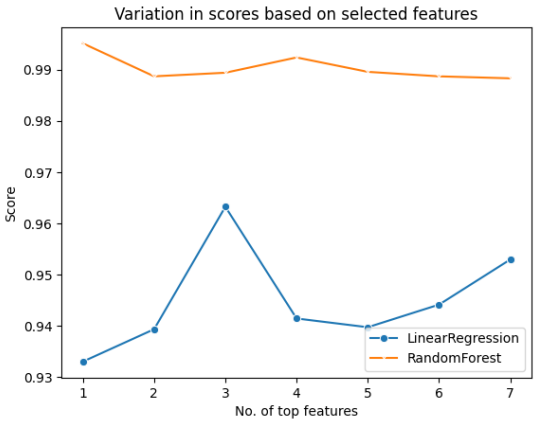

Figure 2: Scores vs top features