Predicting Grades with Behavioural Data

Interim Report

Ethan Kenwrick

Supervised under
Raúl Santos-Rodriguez

Department of Engineering Mathematics
University of Bristol

Introduction

Grade prediction forms the basis for the university application process in the UK. Universities themselves will advertise required results for entry to their degree programs, and students will revise and study with the aim of achieving their offers. It is the role of the teachers, however, that is mostly overlooked.

Teachers are given the near impossible task of trying to predict a students final grade; just sixteen percent of students achieved the exact grades previously predicted for them by teachers [1]. This decision, while seemingly straightforward, has a lot of implications. Not only are the academics providing supposed insight to the university recruiters, but at the same time can bring students back to reality. An aspiring Oxbridge (Oxford and Cambridge universities) medic will very quickly have to adjust their ambitions with a prediction of grade "B" in Biology. One look and university recruiters are unlikely to even give the application a second glance.

While this is the harsh reality of the application process, a familiar pattern has emerged amongst grade predictions throughout the country. For individual grade predictions, just over half of all individual grades were correctly predicted in 2009. However, of the other predictions, the overwhelming majority, 41.7%, were over predicted [2].

This does not sound like a bad thing initially. The over predictions lead to university offers. Unfortunately, if a grade is realistically out of reach for a student and they have been over predicted, it can lead to issues. Not ultimately reaching the grades could lead to rejection from the university, with them having been reassured that you should be reaching that particular level. Following on from this, a student can go through a process called clearing, getting in contact with universities to see if they can get be accepted into the university with their resulting A-level grades.

The same issues are also present on the other side of the prediction, but in a different sense. In the same year, 6.6% of grades were under predicted [2]. Following the same logic for over predicted grades, but in reverse, students are told that they will not be given predictions to get into certain universities. The arising issues in this situation is when a student goes on to overachieve. The system used for a student that overachieves is called adjustment, through which students can reach out to other universities that previously they could not apply to.

The issues with the clearing and adjustment systems are that places can disappear very quickly, leading to disappointment on both ends of the scale. Some students lose out on the place they were originally set on, and others will miss out on courses they were denied early on.

The Use of Educational Data

Data mining, otherwise known as Knowledge Discovery in Databases (KDD), is the practice of extracting useful information from huge data depositories. In recent years, this field of study has taken on a more specific approach, focusing on the area of education, culminating with the introduction of the yearly International Conference on Educational Data Mining (EDM) in 2008 [3].

This paper will focus on the use of behavioural data to create a system to aid the prediction of grades. Data provided by a school in Berkshire, containing information on students from a number of years, will be used. The data has been stored on an online system, from which it was possible to download the necessary information in the form of excel spreadsheets. The school implements a scheme where teachers score students based on their attitude and behaviour. In addition to these scores are a number of other behaviour linked pieces of information.

Current Uses of Educational Data

With EDM comes advantages over the more traditional approaches to educational research, such as laboratory experiments and design research, in particular the validity of the data. It has historically been a stumbling block for other research paradigms to balance feasibility and validity, but EDM counters this.

Using data that comes from public educational data repositories allows researchers to overcome time

consuming methods to do with subject recruitment (teachers, students etc.), study schedule and data entry [3]. In terms of validity and feasibility, the repositories aforementioned contain information and data from genuine settings, involving authentic learning exercises. This removes the need for fabricated settings and situations that would be used during predesigned laboratory sessions.

There are a wide variety of popular approaches for the use of educational data, most notably prediction and relationship mining. For prediction, the goal is relatively simple, investigating whether a particular feature of the data can be inferred though combinations of other features, for which there a three primary methods; classification, regression and density estimation.

A study was undertaken in America by the National Association for College Association Counseling, which displays a use of EDM for prediction. It involved looking to see if it was possible to use educational data to predict cumulative grades in college courses. For this particular study, solely academic achievement data was used to build the model, in contrast to the additional use of behavioural information for this investigation. Twenty six colleges agreed to partake in the study, providing high school grade point averages (HSGPA) and SAT scores [4].

The methods used for this study were focused on the area of regression analysis. The basic aim of regression is to search for and find possible relationships between variables. It includes many techniques as well as containing the ability to analysing several features, focusing on one dependent variable [5].

Through the use of these regression techniques, it was possible to find that overall, the HSGPA proved to be a slightly better predictor than the SAT scores. However, this was not the case for all analysed subgroups, perhaps highlighting the fragility and unpredictability of educational data.

The use of prediction is well documented, another example being displayed through the delivery of material through on-line resources. According to research done by Campus Computing, almost 88% of surveyed institutions reported having used a Learning Management System (LMS) for some form of course delivery[6], which have the potential to provide a wealth of educational data for EDM.

Like the previously discussed study, this particular method focused on the use of regression analysis to investigate students' learning behaviours on-line (such as their login frequency and number of questions asked), alongside their academic performance.

Interestingly, the final outcome of this particular study was not about the link between a specific behavioural trait and final grade; in fact unlike previous studies, there were no clear relationships between the grades and features used. The main finding within this particular case study revealed that combining EDM with traditional statistical analysis can be extremely powerful. It could enable a deeper understanding of a students' learning behaviours, as well as an insight into how best to adapt future delivery to maximise student potential.

Regression, while popular, is not the only method used in the area of prediction. Kotsiantis [7] investigated whether the use of different machine learning techniques could help prevent student dropout in distance learning. Alongside regression, the other five most common machine learning algorithms were used, these being Decision Trees [8], Neural Networks [9], Naive Bayes [10], Instance Based Learning [11] and Support Vector Machines [12].

The results of this study somewhat went against the trend of those previous discussed, in that it predicts those most likely to drop out with 83% accuracy. The Naive Bayes algorithm proved to be the most appropriate algorithm to use, and considering the use of basic academic performance data to do this, is worth more investigation for this study.

There has been an unwillingness to use data mining and machine learning techniques in education in recent times. Education is very much a time and place kind of system; one particular unit taught one year is not going to be the same a year on. This can greatly hinder the amount of measured learning you can do on a dataset, as many methods would typically re-

quire a standardised background [13].
Furthermore, data mining comes with drawback of privacy concerns, particularly in the area of education. Almost all educational data has some sort of link to students on the respective course, or at the respective school. Before any sort of learning or analysis can take place, anonymity becomes a priority [14]. This study looks to overcome these problematic areas. Assistance is provided by the school providing the data to help with anonymising the data, something covered in more detail in the next section.

Data Preparation

The first step in this investigation was to get hold of the data and sort it. The school providing the data has used the same system, SIMS [15], to store the data for a long time. One unfortunate issue was accessing all this stored data. Pulling out small datasets, such as grades for a single year is fine, but trying to access grades for a decade worth of students causes the program to sometimes crash.

To overcome this issue, datasets were collected into separate excel files. One for grades, one for behaviour at school, one for behavioural scores etc. This allowed the program to run smoothly, producing datasets at a reasonable rate.

Collections of the data were organised over a number of days. Initially, visits were planned to ensure the appropriate style and substance of data existed. Once these were confirmed, appointments were made for the collection. Both the running of the SIMS program and sorting of the data was done at the school, so as to cohere to the data protection scheme, explained in more detail in the next section.

The format of these datasets are easy to follow. A single student is named and the following row(s) represent the information for that same student. Due to the limited processing power of the system, a limited number of features can be extracted at any one time, so there are a large number of datasets to be brought together before any sort of analysis can be undertaken.

Anonymity

Obtaining and cleaning the data presented the challenge that comes with sensitive information. The information being accessed has to be anonymised before use. This was done under the supervision of the data protection manager of the school form which the data is from.

This was a straight forward process. The base data set used contained names of each student, both forename and surname, as well as their respective date of births. Each student was assigned a unique ID number, which was then translated across to other data sets. With this in place, information from desired datasets could be transferred to the base set by matching each students ID number.

Once the data had been collated into a single set, both forenames and surnames for every student were erased. The order of the dataset was randomised, before the ID numbers were reset to match the order of the newly shuffled data. The only copy of the original dataset (containing forenames, surnames and the original ID numbers) remained in possession of the data protection officer from the school, for the situation of more data being taken for the benefit of the project. If this situation arose, the shuffling and reindexing of the data was repeated before leaving the grounds of the school.

During the concatenation of the datasets, an occasional issue of students having the same name and date of birth arose (as the ID numbers were assigned using these personal features). In these unlikely situations, the data was manually transferred across for the appropriate student.

To reaffirm the cooperation and contentment with the school about this process, a letter has been provided by the data protection manager. This confirms they are happy that all processes are being followed and that permission was given following discussions with the schools governing board and leadership team.

Cleaning and Organising

Before transferring over the data to the base dataset, it was necessary to clean and reorganise the infor-

mation as to only take across what was needed. The first dataset to clean was the grade data.

This particular dataset contained every type of examination or test undertaken by a student during their time at the school. As the final aim is to predict the A-level grades for a student, these were the only entries required, so the rest of the information was removed for ease of use.

As was mentioned before, the data is stored online. The system is not regulated, and so entries that may mean the same thing have different names or values. Taking the subject Mathematics as an example, some students had the exam down as 'Mathematics', others as 'Maths (General)' for the same qualification. This was the case for a number of different A-levels, so care was taken to ensure no duplicates were present, and to assign all variations to a single reference name.

A similar issue that was solved by a similar issue was the presence of an exam code. Some of the older entries contained these codes, and so were removed from the dataset leaving only the examination name.

The next dataset was in a similar form to that of the grades dataset, only it listed all types of wrongdoings for each particular student. For example, missing homework, disruptive behaviour, lateness; anything deemed noteworthy by the school. For this stage of the study, the only preparation needed was to count the number of entries per student, for use for a baseline model, before concatenating it with the rest of the prepared data.

Statistical Analysis

With the data prepared and in a suitable format, some basic statistical analysis was conducted to provide an overview of the available data.

Figure 1 shows the range of exams undertaken resulting in grades. The plot clearly shows the preference toward mathematical and science based subjects, as well as Economics. However there is also a distinct lack of results in a few subjects.

Notably, there are seven subjects that have less than ten results for the dataset. The clear absence of data, in comparison to the other subjects, has the potential to cause problems within certain algorithms. For this reason, from this point onward, any particular exam with less than ten entries (arbitrarily chosen) will not be considered when algorithms are performed with the individual exam data.

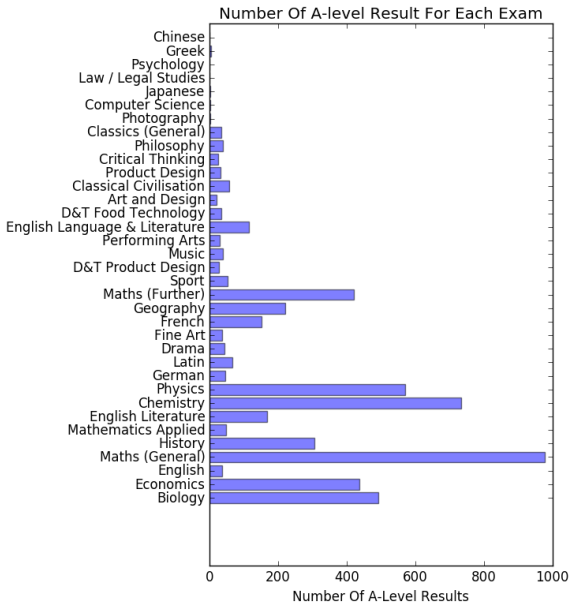

Figure 1: The number of available results for each exam taken by all students at the school

Figure 2 displays another hidden issue with the data. It is clear that there is a high bias toward the "A" grade; in fact it accounts for just under 72% of all grades recorded in the data. Running prediction machine learning algorithms on this data can prove problematic. For any result, simply predicting "A" will guarantee a 72% accuracy.

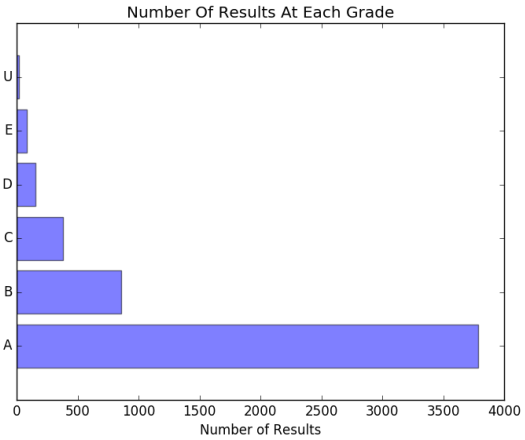

Figure 2: The number of results at each grade level for all exams taken

The problem of biased data is common, and methods exist to counter its effects. Basic ideas such as limiting the over-represented class size [16], or collecting more data are not beneficial for this dataset. With the current size of the dataset, it does not make sense to reduce it further, and access only exist to the data from the one school. Being a relatively small school (around 120s student per year) and the limit of when the data collection began, more data is unfortunately not available.

There are however, useful methods. Stratified sampling, or manually creating the training set [16], allows for an equal representation of the classes during each classifiers' training. It can also be a simple as trying many different algorithms [17]. Some may be able to pick up on slight differences that exist and work well regardless of the biased data. Many of the algorithms aforementioned in this paper will be tested, so there will be a strong variation of techniques.

A small investigation was run into the link between month of birth and academic performance. There is no surprise that there have been numerous investigations into this field, especially for younger ages [18]. Being born in August can put a student at almost a year younger than some in the same school year and have a big impact early on in ones' academic life.

As mentioned, a brief look was taken to see if these traits continued into higher education. The expectation was that, over the years, the advantage held by the older children had diminished to a near unidentifiable level.

Each students exam grades were turned into a score using the current UCAS system [19]. These scores were added and averaged per student. In addition, the students were grouped according to month of birth, and for each month the scores were also averaged. Figure 3 displays the results, with the months in order of the academic year, September through to August.

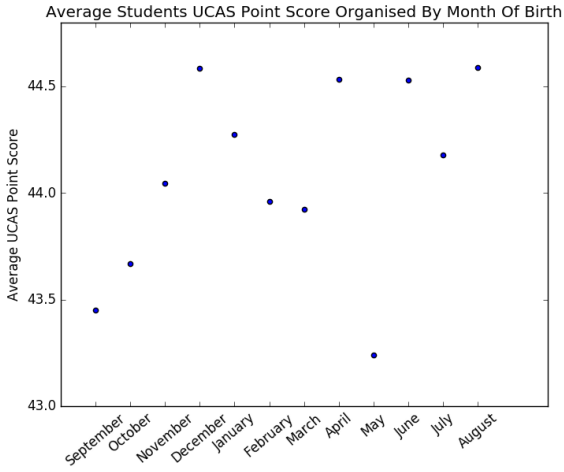

Figure 3: The average UCAS point score for student organised by their respective month of birth

An initial thought would be that in fact, the tradition is reversed, with later born students doing slightly better. However, the vertical axis, the average scores, show that there is only about a single points difference between the highest and lowest. With the UCAS point difference only 8 points per grade (i.e. "A" being worth 48 points, "B" being worth 40), the difference is negligible. This suggests that although prevalent in early school careers, for this particular dataset, the academic performance bias toward those born earlier is balanced out.

To reinforce this view, a *T-Test* was conducted. The *T-Test* is a highly used statistical analysis tool, assessing whether the means of two groups are statistically different from one another [20]. In order for the value of a *T-Test* to be significant, its corresponding *P-Value* must be less than the 0.05. In other words, you are looking for the chance that you would be wrong when making your assumption. Upon achieving this value, one could reject the null hypothesis (in this case that the means are not significantly different).

The data was split into two groups, representing the first half of the academic year and the second half, each containing the average grade score values of all students for each month. When calculated, the *T-Test* was returned at just below -0.6. However, the corresponding *P-Value* was revealed to be well above the 0.05 threshold, coming out at 0.55. This, statistically, reinforces the belief that there is no link between a students month of birth and their final A-Level grades.

Behavioural Model

As aforementioned, the key investigation in this study was the exploration of whether a link exists between a students behavioural traits and their grades. To set a baseline, a basic set up was created using the number of wrongdoings and a students average grade score, again in UCAS points.

The idea behind this baseline was to use the most basic dataset available and look for any early signs of correlation. If any relation were present, it was anticipated that with a higher grade score, the behavioural count would also increase. The results are displayed in Figure 4.

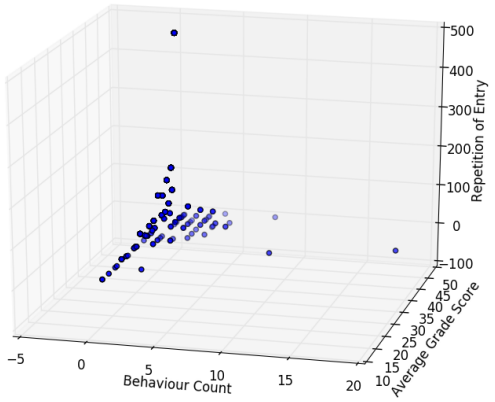

Figure 4: All students' average grade score plotted against their number of recorded wrongdoings while at school

On the graph, it is clear to see that the majority of entries have zero for their behaviour count. Just over one third of the entries are represented by the entry on the top left of the graph. This particular point already begins to hint toward the unsuitability of this particular plot.

Looking into the entries with behavioural counts, they are grouped up toward the top corner with the higher grades. This may be an unfortunate consequence of the bias in the data; as mentioned previously, around 72% of the grade entries are "A". Furthermore, for the most part the entries with any sort of non zero data for the behavioural data are singular, with no other students replicating the position.

These few points highlight clearly the inadequacies in this plot. With the heavy bias, it is difficult to create a fair view of any sort of relation that may exist. Despite this, even with the small number of values elsewhere, there is no clear evidence that the relation exists in the first place. With such basic use of the provided data, no progress is going to be made in finding the desired correlation, and more detailed and varied extraction will be required.

Work Plan

- The next stage for this project is to finish putting all the data together into a single excel

sheet. The amount of cleaning and organisation was a lot higher than was originally expected, and so I'm yet to finish putting it all together, but this should take no longer than a couple of weeks. **November 27th - December 11th**

- In addition to collating the currently held information, there is also a need to head back to the school and collect more data. Conversations are already under way with the school to arrange another trip at an appropriate time. **Meeting arranged - December 1st**
- With the data organised into a use-able format, it will be possible to start running some machine learning algorithms on it. It will be decided whether to use PCA, to reduce the number of features, or simply run the algorithms across groups of features. **December 11th - February 5th**
- Within the same time frame, there is also a need to explore the data already discussed in the report. As mentioned, the behavioural statistics revealed very little information. This particular data set contains a variety of different features that can be explored, and will be explored upon the complete concatenation of the data.
- Earlier in the report I mentioned the most popular machine learning algorithms (naive bayes, decision trees, neural networks etc.). An attempt will be made to experiment with all of these algorithms. **December 11th - February 5th**
- In order to compare the success, typically it will be easiest to use a confusion matrix, as the whole process is focussed on prediction. This will allow me to pull out measures such as precision and recall, as well as allowing the calculation of many more simply from the one table. This can open up many avenues of comparison that can and will be explored. **Will be completed upon completion of each classifier.**

References

[1] Telegraph Reporters *"Only one in six A-level students is predicted the right grades by their teachers"*,[Online] Available at: http://www.telegraph.co.uk/education/2016/12/08/one-six-a-level-students-predicted-right-grades-teachers/ [Accessed 2nd November 2017]

[2] Everett, N. & Papageorgiou, J. *"Only one in six A-level students is predicted the right grades by their teachers"*,[Online] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32412/11-1043-investigating-accuracy-predicted-a-level-grades.pdf [Accessed 2nd November 2017]

[3] Baker, R. *"Data Mining For Education"*,[Online] Available at: http://users.wpi.edu/~rsbaker/Encyclopedia%20Chapter%20Draft%20v10%20-fw.pdf [Accessed 16th October 2017]

[4] National Association for College Association Counseling *"Predicting Grades in College Courses"*,[Online] Available at: http://files.eric.ed.gov/fulltext/EJ829428.pdf [Accessed 22nd October 2017]

[5] Witten, I *"Data Mining: Practicle Machine Learning Tools and Techniques"*,[Online] Available at: https://books.google.co.uk/books?hl=en&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+algorithms+data+mining&ots=8HIMsgnCy8&sig=0IRitlvtXZ-Wx7uSkRyff1y4W6o#v=onepage&q&f=false [Accessed 9th November 2017]

[6] Abdous, M *"Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade"*,[Online] Available at: file:///Users/ethankenwrick/Downloads/Using_Data_Mining_for_Predicting_Relationships_bet.pdf [Accessed 22nd October 2017]

[7] Kotsiantis, S. *"Preventing Student Dropout in Distance Learning Using Machine Learning Techniques"*,[Online] Available at: https://s3.amazonaws.com/academia.edu.documents/31211650/KES2003.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1510260345&Signature=G%2B1UIY0MnQEZRMQAKoEUmRtzP6M%3D&response-content-disposition=inline%3B%20filename%3DPreventing_student_dropout_in_distance_l.pdf [Accessed 10th November 2017]

[8] Gupta, P. *"Decision Trees in Machine Learning"*,[Online] Available at: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed 16th November 2017]

[9] University of Wisconsin *"A Basic Introduction To Neural Networks"*,[Online] Available at: http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html [Accessed 16th November 2017]

[10] Bownlee, J. *"Naive Bayes for Machine Learning"*,[Online] Available at: https://machinelearningmastery.com/naive-bayes-for-machine-learning/ [Accessed 16th November 2017]

[11] Machine Learning for Starters *"Instance Based"*,[Online] Available at: http://trymachinelearning.com/machine-learning-algorithms/instance-based/ [Accessed 16th November 2017]

[12] Bownlee, J. *"Support Vector Machines for Machine Learning"*,[Online] Available at: https://machinelearningmastery.com/support-vector-machines-for-machine-learning/ [Accessed 16th November 2017]

[13] Romero, C *"Data Mining in E-learning"*,[Online] Available at: https://books.google.co.uk/books?hl=en&lr=&id=7WLQCwAAQBAJ&oi=fnd&pg=PA157&dq=Hamalainen,+W.,+Suhonen,+J.,+Sutinen,+E.,+%26+Toivonen,+H.+(2004).+Data+mining+in+personalizing+distance+education+courses.+In+World+conference+on+open+learning+and+distance+education,+Hong+Kong.&ots=fbIS1jDdYs&sig=B9tQtSasF00cQ_QBLB7R1_Ypq4k#v=onepage&q&f=false [Accessed 10th November 2017]

[14] Anonymous *"Advantages and Disadvantages of Data Mining"*,[Online] Available at: https://www.ukessays.com/essays/information-technology/advantages-and-disadvantages-of-data-mining-information-technology-essay.php [Accessed 16th November 2017]

[15] CAPITA *"School Information Management System"*,[Online] Available at: https://www.capita-sims.co.uk/ [Accessed 21st November 2017]

[16] Hartl, F *"Dealing With Skewed Classes"*,[Online] Available at: https://florianhartl.com/thoughts-on-machine-learning-dealing-with-skewed-classes.html [Accessed 19th November 2017]

[17] Brownlee, J *"8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset"*,[Online] Available at: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ [Accessed 19th November 2017]

[18] The Institute for Fiscal Studies *"Month of birth matters for children's well-being as well as for test scores"*,[Online] Available at: https://www.ifs.org.uk/pr/month_of_birth.pdf [Accessed 20th November 2017]

[19] UCAS *"The UCAS Tariff calculator"*,[Online] Available at: https://www.ucas.com/ucas/tariff-calculator [Accessed 20th November 2017]

[20] Trochim, W. *"The T-Test"*,[Online] Available at: https://www.socialresearchmethods.net/kb/stat_t.php [Accessed 24th November 2017]