Using Data Mining Techniques to Predict Students at Risk of Poor Performance

Zahyah Alharbi*, James Cornford†, Liam Dolder‡ and Beatriz De La Iglesia*
*School of Computing Sciences, University of East Anglia, UK
{z.alharbi, b.iglesia} @uea.ac.uk
† Norwich Business School, University of East Anglia, UK
j.cornford@uea.ac.uk
‡ IT& Computing Service, University of East Anglia, UK
l.dolder@uea.ac.uk

***Abstract*—The achievement of good honours in Undergraduate degrees is important in the context of Higher Education (HE), both for students and for the institutions that host them. In this paper, we look at whether data mining can be used to highlight performance problems early on and propose remedial actions. Furthermore, some of the methods may also form the basis for recommender systems that may guide students towards their module choices to increase their chances of a good outcome. We use data collected through the admission process and through the students' degrees. In this paper, we predict good honours outcomes based on data at admission and on the first year module results. To validate the proposed results, we evaluate data relating to students with different characteristics from different schools. The analysis is achieved by using historical data from the Data Warehouse of a specific University. The methods used, however, are fairly general and can be used in any HE institution. Our results highlight groups of students at considerable risk of obtaining poor outcomes. For example, using admissions and first year module performance data we can isolate groups for one of the studied schools in which only 24% of students achieve good honour degrees. Over 67% of all low achievers in the school can be identified within this group.**

***Keywords*—Performance Prediction; Data Mining; Classification; Recommender System;**

I. Introduction

Nowadays, many UK universities have a specific targets for students achieving good honour degrees. Achievement in terms of good honours is often reported in league tables. For example, the Complete University guide [1] reports good honours as "the percentage of graduates achieving a first or upper second class honours degree". On the other hand, the Guardian League Tables utilises a value-added score that compares students individual degree results with their entry qualifications, to show how effective the teaching is [2]. It is also important for students to achieve a good degree as this can impact on their employment prospects [3]. It is therefore in the interest of both students and Universities to identify students at risk of not obtaining a good honours degree so that early intervention may improve their outcome.

Universities have large numbers of module choices, and it is challenging for students to familiarize themselves with all the possibilities and make appropriate choices. Appropriate choices may lead to better performance and/or higher student satisfaction. It is therefore also important to predict overall outcome and outcome in specific modules, given particular module choices. This can be the basis for a recommender system to aid students in module selection. Recommender systems are currently considered as an advisable automated solution for assisting students with their choices [4].

In this paper we attempt to use data mining techniques to predict students' outcomes based on early module performance and other student characteristics. If our methods are successful for predicting the more general problem of student good honours performance, we can then produce more granular predictions at the module level and those would form the basis for a recommender system. We hope to uncover early indicators of poor performance that may be used to target remedial action for the concerned students. We aim to investigate the available features that may be used for prediction, as well as the type of classifiers that may produce the best results.

Educational data mining is now an established field and, as such, a number of reviews have been published, e.g. [5], [6], [7]. In particular, Peña-Ayala [6] cover a number of recent work on students performance using data-mining. We review some of that work and apply best practice to our own problem.

The rest of this paper is organized as follows: Section II describes related research. Section III describes aspects that will help in improving students performance. Section IV describes the data that was used in the experiments. Section V explains briefly the sequence of experiments and the produced results. The discussion of the result is contained in section VI. Lastly, Section VII outlines the conclusion and future work.

II. Related Work

A number of studies have addressed similar issues within the educational data mining framework. For example, early warning systems were investigated by Arnold [8]. The author suggests warning weak students that they are at risk of failing to achieve satisfactory results; this is reported as a good source of motivation for them. Students are accordingly evaluated through the three commonly known traffic signal lights (green, amber and red). The signal's colour indicates the risk level for the student relative to their counterparts. The algorithm used analyses both the student's module performance and their help-seeking behaviour.

Another aspect often addressed is the prediction of student retention. Bayer et al. [9] combined social network analy-

sis with data mining techniques to identify potentially non-successful and successful students at the beginning of their studies. The aim of their study was to improve the classification accuracy of educational data when social network analysis is included.

In terms of student performance, authors have adopted different data sources. For example, Baradwaj and Pal [10] utilised a decision tree algorithm as the classification method to predict students' attainment. Their proposed model was based on students past module performances combined with their lectures and seminars' attendance, hence a measure of engagement was introduced. The study's objective was to examine students' performance and to apply early interventions for weak students and students that are most likely to drop-out.

Detecting the relationships between modules can also be helpful, particularly in terms of identifying those modules that will increase student knowledge and attainment. Bayer et al. [11] conducted another study to identify the prerequisites for difficult modules by detecting the relations between modules. Their approach focused on measuring the dependency of final grades on combinations of modules. Their method was based on the historical data on student enrolments in modules.

Vialardi et al. [12] utilised data mining techniques to predict students' future grades by using two main attributes: the difficulty of each module (taken as the average of previous students' grades), and the level of a student's knowledge before taking the module (computed from previous obtained grades in related modules).

Romero et al. [13] enriched the data available by using information from students taking particular tests. Their approach focused on applying a Class Association Rule Mining algorithm to three different matrices: a score matrix, a relationship matrix, and a knowledge matrix. These matrices were built based on the data of students' performances in their test, and on the domain knowledge provided by a pedagogue. Similarly, Wang et al. [14] have employed data mining techniques to build a prediction model for students performance. The aim of their study is to show that their prediction model is better by taking into account partial credits rather than binary credits (correct/incorrect). The partial credit is measured as the amount of assistance a student will need to solve the assigned problems. The study measured the amount of assistance through keeping track of the number of hints and the number of attempts a student requires to solve each of the assigned problems. The model was built using data retrieved from a web-based maths system for 7th to 12th grade students.

A number of models have also been used in this context. For example, Hoe et al. [15] employed a CHAID algorithm to identify the important variables that influence the performance of undergraduate students. The study examined the patterns obtained using the data of students demographics and past performances. Nguyen et al. [16] presented two classification models for predicting students' academic performances. One of the models was built using a decision tree algorithm and the other one using a Bayesian network algorithm. They used real data of undergraduate and postgraduate students at two different higher education institutions. The comparison of the two models' results shows that the decision tree classifier provided better overall accuracy. The aim of their study was to identify and assist failing students and also to determine the students that are more qualified to receive a scholarship or fellowships from their institution.

Chaturvedi and Ezeife [17] applied association rule mining to discover hidden patterns of the impact of assignments on the overall performance of students exams and total score. The objective of the study was to maintain students' motivation on successfully completing their assignments as they are considered less structured and unsupervised compared to exams and quizzes. The study was based on historical students data of University of Windsor in Canada.

Also, Siaz and Zorrilla [18] conducted a study to determine the meta-features that are more suitable for the problem of predicting student's performance. Their experiment was based on real data that related to virtual modules. The aim of the study was to build a recommender service that assists the non-expert instructor in selecting the best classifier for a given data set.

Researchers are also interested in identifying the most suitable study track for students. Al-Radaideh et al. [19] presented a classification model based on data mining techniques to assist students in taking decisions about their academic track. Their aim was to support students in making better choices about which track to enrol on after completing the foundation level. A model was developed based on the experience of students with similar academic performance and profile.

Our first aim, in conjunction with some of the studies reviewed, is to identify weak students as early as possible, i.e. those that would end up with poor outcomes. We define good performance in terms of "Good Honours" versus "Not Good Honours" outcomes because this is currently a measure generally used in the UK. The main aim is to highlight as early as possible (i.e. in year 1) groups of students that may be at risk so that targeted interventions can be proposed to improve their outcomes. Given the variety of models used in the literature with varying degrees of success and the fact that no model has emerged as the overall best, we use a number of classifiers and combine them using ensembles to establish the best possible model. Given also the literature's variation on the features to be included, we include a number of feature sets: first we attempt classification with a feature set which uses only information available at registration, then we add performance on year 1. Furthermore, we take into consideration the difficulty of each module by comparing the performance of each student with their peer group, as some studies suggested. In the future, we aim to include attributes on engagement as others studies have suggested, that are only now becoming available (e.g. engagement with library services or attendance monitoring information). We will also look at combinations of module choices in years 2 and 3 in relation to outcomes. Our approach aims to provide further evidence of best feature sets and models for classification.

III. Long term study objectives

In this section, we consider how to provide students with an appropriate intervention that may improve their overall performance. This would be the ultimate aim of this preliminary exercise.

The most significant aspect is to identify weak students that may be at risk of graduating with a lower class or abandoning their studies. Students at a high risk need particular attention and support with managing their studies if they are to graduate with higher grades. In this sense, it is important to select the attributes that closely represent the chief characteristics of the students at risk; this may include achievement in specific modules as well as personal characteristics. Some personal characteristics may suggest specific strategies. For example, if non-native or overseas students are more often associated with poor outcomes, an intervention based on additional language support may prove fruitful.

We may also identify modules that are associated with good outcomes and bad outcomes given a student profile, so that when module choices are available those modules can be suggested or discouraged respectively for students with similar characteristics and academic achievement records. The intervention in this case may be a recommender system which takes account of similar students' trajectories and achievements to recommend what may be best choices for a particular student.

We can also examine the measure of the dependencies or associations between modules. This may alert us to potential problems on related modules once a particular module is associated with a bad outcome. For example, some remedial sessions on a failed module may help students conquer related modules more successfully.

Hence, in this paper we begin our work by predicting overall good honours outcomes based on generic students' characteristics and on first year performance to inform strategies for intervention. The next step, not included in this paper, is to explore further the association between individual modules and outcomes to create a fully fledged recommender system that leads towards an improvement in good honours rates.

IV. Data Extraction and Pre-processing

The data was retrieved from a University's Data Warehouse, where information of the students and their outcomes is collected. Additionally, the data warehouse contains other important data that is required for external agencies, e.g. those collating league tables.

Initially, 19,811 records were provided, which corresponded to 984 undergraduate students that obtained their academic award throughout the years 2005 to 2013 and were registered to a specific school of study. It should be noted that we focused the initial data mining exercise on a specific school since results may only be meaningful for students undertaking the same programmes and taking similar module choices. In turn, after cleaning and filtering the data for the purpose of removing irrelevant items, the remaining data was associated with 898 students. For example, since some of the focus of this paper is to identify weak students on Year 1, we removed the data for 25 students, because their first year data was missing due to either exemptions or transferring from a different school. Additionally, for quality purpose we removed data that corresponded to 55 students, because for some reason that would require further investigation they appeared to have taken the investigated first-year modules in their second or third year of studies. We also removed 6 students on discontinued courses.

The main outcome variable was whether the students obtained Good Honours (GH) or Not Good Honours (NGH). Those in the GH class were individuals who were awarded a CLASS I*, CLASS I, or CLASS II,DIV 1 degrees. Those that achieved any other degree classifications were labelled as NGH. The grading scheme in this paper is based on British higher education system. An explanation of undergraduate grading system in the UK can be found in [20].

Next, for validation purposes we were provided with 38,608 records that corresponded to 2,214 undergraduate students that obtained their academic award throughout the years 2005 to 2013 but were registered to a different school of study associated with a different discipline. We also cleaned and filtered the data by removing data for 416 students because their first year data was missing as with the other dataset. We also removed data corresponding to 9 students either because of data linking errors or because they did not take the investigated first-year modules. The remaining data was associated with 1,789 students.

In both datasets, each student was represented by one row of data regardless of their study profile. The following information provides an overview of the data used to represent each student:

Attributes that relate to student demographics and general performance included:

- gender;
- age at entry;
- disability(Yes/No);
- level of widening participation in higher education;
- nationality;
- overall score in year 1;
- overall score in year 2;
- overall score in year 3;
- the year they obtained their academic award;
- the award class;
- fee status as in (H)ome, (O)verseas or (EU)ropean;
- and the name of the course of enrolment.

Attributes that related specifically to students' modules included:

- name of module;
- module code;
- number of students enrolled on the module;
- average mark for module computed for students registered at the same time as the current student as this was not stored in the data warehouse;
- individual mark for module;

TABLE I: Description of Performance Field

Value	Description
Fair	student mark > (module' s average mark + 5%)
Average	(student mark >= module's average mark - 5%) and (student mark <= module's average mark + 5%)
Poor	student mark < (module's average mark - 5%)

TABLE II: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the first dataset

Status	GH	NGH	Total	GH Rate
Home	433	302	735	58.9%
European	26	18	44	59.1%
Overseas	44	75	119	36.9%
Total	503	395	898	56%

- performance of student compared to his/her peers (Fair, Average, Poor) as described in Table I. The boundaries that we used in (Table I) where somehow arbitrary and we intend to experiment with different boundaries in the future.

The main objective of this study was to determine whether there were any significant patterns that could be exploited for students that completed their degree without obtaining good honours. Those could subsequently be used to suggest staged interventions for other students with similar characteristics.

V. Experiments and Results

The initial analysis of all the data for the 9 years span in both the first dataset and second dataset showed an overall Good Honours rate of 56% and 63.3% respectively. The overall GH rates for the first dataset are given in Table II and are divided by fee status. The trend of GH over the years is shown in figure 1 and is also divided by fee status into H, EU and OS students. It shows that attainment is worse for OS students with some narrowing of the gap over the years. The number of OS students has grown steadily as a proportion of the total. The number of EU students is low and hence their attainment level cannot be meaningfully assessed but is closer to that of the H students than to the OS students.

The outcome of the initial exploration for the second dataset was very similar to the outcome of the first dataset and is presented in Table III and figure 2. The attainment levels are also better in this second school for H than OS students. There is insufficient data for EU students to consider the trends in the same way. The percentage of OS students has also increased over time in this second school and the performance of both H and OS students has improved over time, although the gap remains large between both groups.

A. First Experiment: student demographics feature set

In the first phase of the experiment we used the attributes that related to student demographics and general performance, i.e. the first group of attributes, but not the attributes for specific modules. We also discounted the year averages as they will be clearly related to the outcome. Assessing each attribute independently using a Feature Selection ranking algorithm based on a Pearson chi-square test with significance level of 0.05 found that fee status, course, nationality, widening participation indicator and gender were all significant ($p<0.05$).

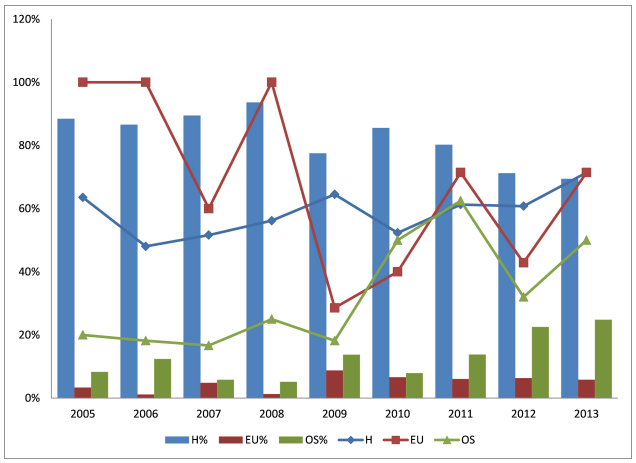


Fig. 1: GH Rate for the first dataset. The bars show the percentage of students in each fee category (Home, EU and Overseas) whereas the corresponding line charts show the percentage of GH degrees obtained for each category of students.

TABLE III: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the second dataset

Status	GH	NGH	Total	GH Rate
Home	931	370	1301	71.6%
European	39	32	71	55%
Overseas	163	254	417	39.1%
Total	1133	656	1789	63.3%

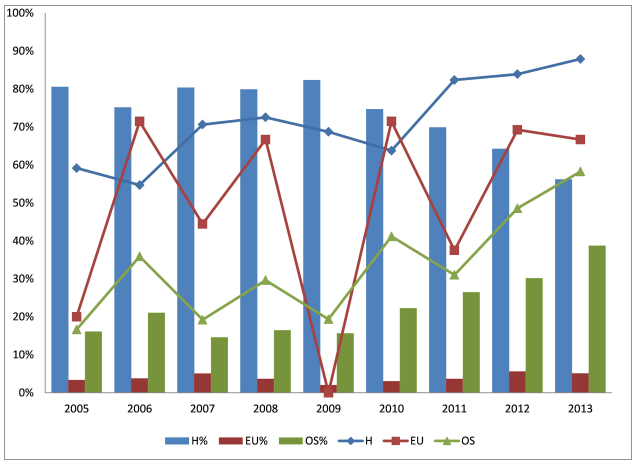


Fig. 2: GH Rate for the second dataset. The bars show the percentage of students in each fee category (Home, EU and Overseas) whereas the corresponding line charts show the percentage of GH degrees obtained for each category of students.

According to this preliminary analysis OS students have significantly worse outcomes. Some specific courses offered by the first school also had significantly worse outcomes than others. The results for nationalities which presents higher granularity than fee status cannot be taken into consideration since some countries have very low numbers of students attending which invalidate the results of the chi-square test. However, there were specific countries with substantial number of students that did have significantly lower levels of attainment and may be of concern. The widening participation attribute relates to the participation in higher education of students in a postcode relative to the HE population as a whole. The students are classed as belonging to groups 1 to 5, Non-UK or not known. A classification into the lower groups implies that the student lives in a postcode of low participation. The lowest group is associated with lower attainment. Since widening participation also reflects OS students as a rather large group (over 17% of the students belong to it) they also show lower attainment. There is little difference in the other groups in terms of attainment. In terms of gender, females have significantly higher levels of attainment than males.

Next, we used a combination of classification models to predict GH/NGH. For this we used the software IBM SPSS Modeler v 15, a well known data mining tool-kit. We used an autoclassifier which engages 9 different types of classification models and automatically selects those that perform best on the training data. The models that were tried were: logistic regression; Neural Network; Decision List; Bayesian Network; Discriminant analysis and four decision tree algorithms: C5, C&R Tree, Quest and CHAID. All algorithms used default parameters. Those selected for the first data set were a Bayesian Network, a CHAID decision tree algorithm and Logistic Regression. They were combined using an ensemble with confidence-weighted voting. Our ensemble model had an accuracy of over 65% on training data (over 58% on a test sample containing 20% of the original data). The gain chart for the ensemble model versus the selected independent models is shown in Figure 3a. Gain charts show the efficacy of the classification model calculated as the ratio between the classification results obtained with and without model[21],[22, p.212]. The accuracy of individual models was similar to the accuracy of the ensemble. It is possible using the ensemble, to chose those records which are predicted to correspond to NGH students with high prediction probability. This strategy would enable us to select the students most likely to gain NGH, so that interventions could be put in place to help them early on. Using a threshold probability of 0.5 as given by the ensemble model, we were able to select 254 students with a GH rate of 35%, considerably lower than the overall population. That group captured 165 or 41.8% of the NGH students. More precisely, 129 of the NGH group were predicted as a 2:2 class degree and the other 36 students were predicted as a lower class degree, for example 3rd or PASS class degree. Lowering the probability of the ensemble prediction to 0.3 captured a group of 294 students representing a GH rate of 38.1% still substantially lower than that of the overall population. The later threshold captured 182 or 46.1% of the NGH students (precisely, 140 of NGH group as a 2:2 class degree and 42 students as a 3rd/PASS class degree). If an intervention could change the outcome for a majority of those students from NGH to GH, it could substantially improve the overall GH rate. Note that students who obtained a 2:2 class degree, the larger group, should require less effort to help them achieve GH degree than students who obtained a lower class degree. However, in the interest of fairness the intervention should be directed to all students at risk of poor outcomes. It is plausible to think that an intervention may also be beneficial for the students that may be captured by this approach but who would have got GH degrees in the first place, i.e. the false positives (35 or 38 % of students in each scenario) as it would enable them to achieve even better outcomes. The three attributes used in all models were course, gender and fee status. Two of the models used an additional attribute: widening participation.

Then for validation purposes, we applied the above series of steps on the second dataset. We began by using only the attributes that related to student demographics. The assessment of each independent attribute using a Feature Selection ranking algorithm based on a Person chi-square test with significant level of 0.05, showed that nationality, widening participation indicator, fee status, and gender were all significant ($p<0.05$). OS students in this dataset also had significantly worse outcomes. The results for nationalities will not be taken into consideration for the same reason mentioned in the first dataset. However, the same specific countries as for the first dataset had significantly lower levels of attainment. The assessment of the widening participation attribute has shown that the lowest group of students is associated with lower attainment. The OS group within the widening participation attribute included over 26% of students, and was also associated with lower attainment. There is little difference in other groups in terms of attainment. In term of gender, females also have significantly higher level of attainment than males. However, the difference in the attribute assessment in this dataset compared to the first data set was that the course attribute was not as relevant. The reason for this is that 63% of the undergraduate students enrolled on same course, hence one of the courses that this school offers has a much higher number of students compared to the other courses.

Next, we used a combination of classification models to predict GH/NGH for the validation dataset again using the software IBM SPSS Modeler v 15. The autoclassfier selected C & R Tree, Quest and CHAID classification models for the second data set. They were combined using an ensemble with confidence-weighted voting. Our ensemble model had an accuracy of over 69% on training data (over 65% on a test sample containing 20% of the original data). The gain chart is shown in Figure 4a. It is also possible using this ensemble, to chose those records which are predicted to correspond to NGH students using high prediction probability. Using a threshold probability of 0.5 as given by the ensemble model, we were able to select 320 students with a GH rate of 35%, considerably lower than the overall population. That group captured 209 or 32% of the NGH students(159 obtained a 2:2 class degree and 50 had a 3rd/PASS class degree). Lowering the probability of the ensemble prediction to 0.3 captured a group of 443 students representing a GH rate of 39.5% still substantially lower than that of the overall population. The later threshold captured 268 or 41% of the NGH students(205 of that group had a 2:2 class degree and 63 students had a 3rd/PASS class degree). Two classifiers used the 4 available attributes: widening participation, fee status, gender and course name. One classifier, the CHAID tree, produced a very simple

model using only fee status and gender.

B. Second Experiment: adding Year 1 performance

After identifying the students that were at high risk of failing to earn a GH award class using only the first group of attributes, a second experiment considered the influence of performance on the year 1 modules on the classification. Our first dataset contained information on students enrolled on 12 different courses. Although most year 1 modules are compulsory and many of them are shared between different courses, there were 11 different modules that we needed to consider to account for all the variations. For each of those 11 modules, we considered the performance of the students with respect to their peers as defined in Table I as this could be more indicative than an absolute mark value.

Feature Selection ranking using a chi-square algorithm showed that all of the module performances were important in the classification. Furthermore, an F-test to compare the mean mark of students in the GH and NGH group for each module showed significant differences in the means with students that achieve NGH obtaining significantly lower marks on the year 1 modules. Hence poor outcomes seem to be already visible on module performance in year 1. This is an important finding since the year 1 module marks do not contribute to the overall degree classification, but are nevertheless indicative of the expected outcome.

A classification ensemble was built as in the previous experiment, but this time using the year 1 module performance attributes as well as the previous demographic attributes identified by feature selection. The autoclassifier chose a CHAID, C& R Tree and a Decision List as the classifiers and combined them to produce an accuracy over 78% on the training data (70% on the test sample). This represents a substantial improvement from the previous model. The gain chart in Figure 3b shows the evaluation of the model accuracy.

Selecting those that are predicted as NGHs with a probability greater than 0.5, as in the previous experiment, isolated a group of 355 students with a GH rate of 24.6%. There were 267 or 67% NGH students in the group. Specifically the group captures 192 who obtained a 2:2 class degree and 75 who obtained a 3rd class degree. An intervention for this group could be quite effective on the overall GH rate and quite targeted. A final assessment of those in the group showed that they had substantially lower averages for year 1, 2 ,3 and 4, as well as substantially lower averages for all year 1 modules. An F-test showed statistically significance differences ($p<0.05$) for all pairs of averages (in the selected group and all others). The mean values for years 1-4 and for all first year modules are shown for both groups in table IV.

Moreover, we applied the second phase of the experiment on the validation data set. The second dataset contained information on students enrolled on 4 different courses. Again for this school, most year 1 modules are compulsory and many of them are shared between different courses. There were 10 different modules that we needed to consider to account for all the variations. For each of those 10 modules, we also considered the performance of the students with respect to their peers.

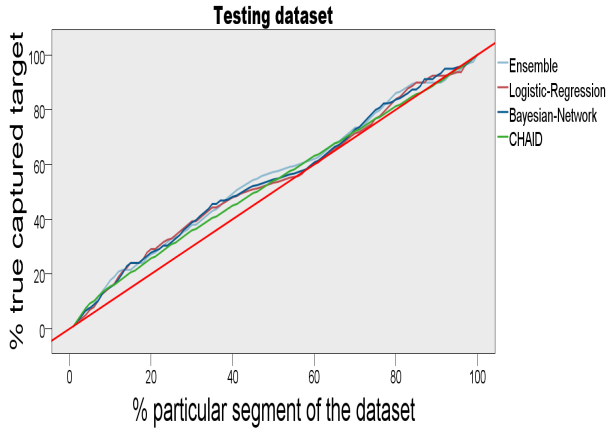
TABLE IV: Comparison of means for poor performers as selected by ensemble versus all other students in the first dataset

Attribute	Mean(Poor Performers)	Mean(others)
Module1	55.01	68.13 %
Module2	41.40	66.82%
Module3	53.98	66.22%
Module4	48.02	60.19%
Module5	65.95	74.04%
Module6	51.35	70.27%
Module7	45.49	61.02%
Module8	48.64	63.22%
Module9	42.61	56.22%
Module10	45.95	57.66%
Module11	47.24	54.23%
Year1	50.28	63.94%
Year2	51.74	61.23%
Year3	56.25	65.26%
Year4	55.31	70.57%

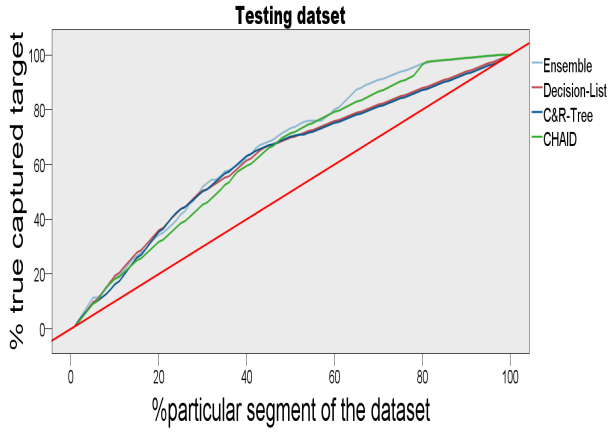
We found using the Feature Selection based on chi-square algorithm, that all of the module performances of the validated data set were important to the classification except for one module. The module that was not significant was taken by a very low number of students and was therefore discounted from the rest of the analysis. In addition, the F-test showed significant differences in the means between students that achieve GH/NGH. Those that obtained NGH had significantly lower marks on year 1 modules, even though as before, year 1 marks do not count towards degree classification.

Next, a classification ensemble was built using the year 1 module performance attributes as well as the previous demographic attributes identified by feature selection. The autoclassifier chose a Logistic Regression, a Neural Net and C& R Tree as the classifiers and combined them to produce an accuracy over 77.98% on the training data (77.96% on the test sample). This also represents a substantial improvement from the previous model for the validation data set. Figure 4b shows substantial gain for the model including year 1 performance attributes with respect to the previous model and to the baseline.

Again by selecting those that are predicted as NGHs with a probability greater than 0.5, we captured a group of 363 students with a GH rate of 19.56%. There were 292 or 45% of the NGH students. The NGH captured group included 216 students that obtained 2:2 class degrees; the remaining students in the group obtained a lower class degree. A final assessment of those in the group showed that they had substantially lower averages for year 1, 2 and 3, as well as substantially lower averages for all year 1 modules. Note that in this data set, all students completed their degree within three years, but in the first data, students may take four years to complete their degree due to year in industry variants. An F-test showed statistically significant differences ($p<0.05$) for all pairs of averages (in the selected group and all others). The mean values for years 1-3 and for all first year modules are shown for both groups in table V.

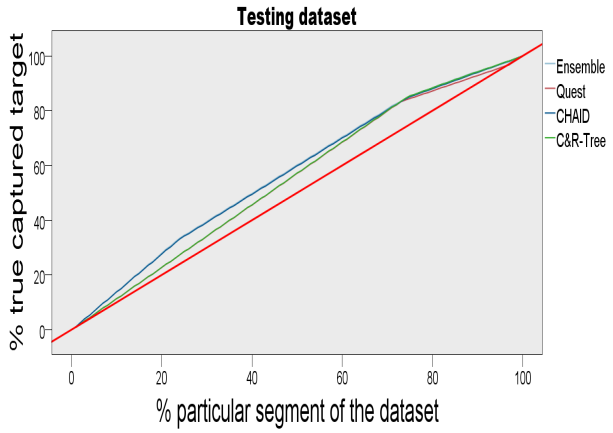


(a) First Experiment

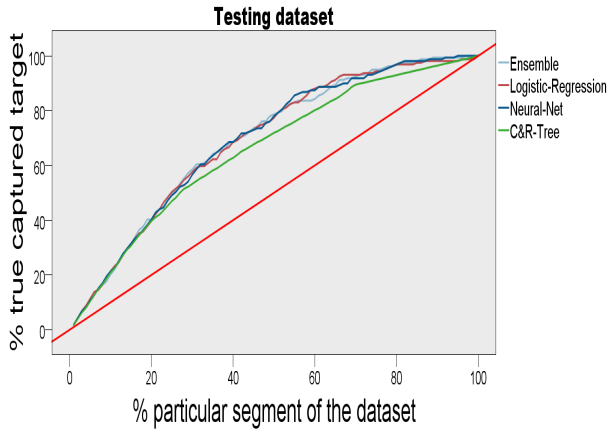


(b) Second Experiment

Fig. 3: The gain chart shows the percentage of positive predictions that the model gains for each segment of the dataset predicted. This chart is based on the testing sample from **the first dataset**. The gap between the red line (no model) and each of the remaining lines(derived models) shows the percentage of correct target selection with the derived model over a random selection of targets. Note that the dataset of x-axis is sorted by the probability of the target outcome, highest to lowest.



(a) First Experiment



(b) Second Experiment

Fig. 4: The gain chart shows the percentage of positive predictions that the model gains at each segment of the dataset. This chart is based on the testing sample from **the validation dataset**. The gap between the red line (no model) and each of the remaining lines(derived models) shows the percentage of correct target selection with the derived model over a random selection of targets. Note that the dataset of x-axis is sorted by the probability of the target outcome, highest to lowest.

VI. Discussion

We have found that the results of our preliminary analysis assessing each attribute independently using a Feature Selection ranking algorithm are in accordance with what has been found in other studies (e.g. [23],[24], [25], and [26]). Previous studies have found that Home students are associated with higher attainment than OS students ([25], [23] and [24]). In contrast, some studies [25], [27] found that there were no significant differences in the class of degree obtained by OS students compared to Home students. However, this tended to be in disciplines such as agriculture, librarianship and information science, engineering and technology, mathematical sciences or combined studies. Still, there have been significant differences in other specific disciplines [25] such as architecture, computer sciences, building and planning, social , economic and political studies, law, business and admin-

TABLE V: Comparison of means for poor performers as selected by ensemble versus all other students in the Second Dataset

Attribute	Mean(Poor Performers)	Mean(others)
Module1	43.81	58.81 %
Module2	58.00	66.65%
Module3	46.99	56.01%
Module4	44.00	54.75%
Module5	48.40	60.31%
Module6	51.05	64.29%
Module7	42.35	54.47%
Module8	54.91	58.86%
Module9	53.56	59.17%
Year1	48.60	59.09%
Year2	51.05	59.83%
Year3	55.83	64.35%

istrative studies. In the later subjects, Home students have higher levels of attainment than OS students. Additionally, our findings were consistent with other studies [23], [25] in terms of gender: female students are more likely to graduate with GH degrees than male students, although they are minorities in some disciplines such as science subjects compared to art subjects. For instance, our first dataset which relates to a science subject has 17% female students, compared to 83% male students; our validation dataset has 39% females and 61% males. Some studies [23] have found that students who come from areas with the lowest levels of participation in HE, and those who come from less affluent areas, are more likely to have lower attainment. In contrast, some other studies in [23] have found no significant difference. Those findings are in agreement with our own findings in terms of the widening of participation: Home students who come from neighbourhoods with very low participation in HE are associated with lower attainment, but there is little difference between other groups (2-low, 3-medium and 4-high). The greatest differences that other studies [23] found in terms of attainment are between students who come from different types of schools, such as comprehensive/independent schools. We did not include this attribute in our data, since we do not have this information in the University Data Warehouse. Other studies [25], [26] found that mature (21+) and/or full-time students have significantly higher levels of attainment than younger/part-time students respectively. The attribute "age at entry" was not significant in our Feature Selection assessment because 88% of students were between the ages of 17 and 21, and we excluded the attribute full/part-time because all the students in the dataset were full-time learners.

We have been able to discover groups of students that have poor performance in terms of good honours grades. Those students are identifiable with some certainty as soon as they arrive by their general characteristics, i.e. gender, course enrolled on, nationality and widening participation level. Furthermore, they are more accurately identifiable at the end of year 1 when considering their performance on different modules in that year. We expected that including attributes from module performance would improve predictive accuracy. However, we assumed that particular modules may be found to be problematic when in fact poor performance appears to affect every module.

The poor performer group show some ability to marginally improve according to their year 2 and 3 averages so targeted intervention could give them enough impulse to achieve GH degrees. If the intervention could achieve a good lift in terms of GH rates, it will also positively affect the University as it will influence league table positions.

Our discovered patterns hold for two different datasets belonging to different schools with different admission strategies and teaching different disciplines. Schools operate quite independently of one another but the same patterns have emerged from both in terms of characteristics of low attainers. We believe this gives some validation to the patterns found.

Some of the immediately obvious interventions could be targeted at the OS students who are prominent in the under achieving group (over 19% in the first dataset and over 39% in the second dataset). Providing extra English language lessons to improve their comprehension and communication skills could achieve the desired effect. Additionally, all those found to be in the selected group of predicted poor performance could be approached by their academic advisers and offered remedial sessions. Remedial sessions could run in the summer remotely to revisit areas of the course where students have done poorly. This may improve their academic knowledge and ability and prepare them to undertake the second and third years from a stronger footing. The analysis did not uncover specific problem modules as the poor performers seemed to do poorly across the board and on all modules in relation to their peers. Furthermore, our analysis could also be used to influence admission policies given the characteristics of predicted poor performers.

The next step of the analysis which is not included in this paper will be to analyse 2 and 3 year module choices and performance on those to try to discover if module choices are associated with outcomes as that could be the basis for a recommender system. Furthermore, additional attributes may also be used including measures of engagement.

In terms of classifiers, there were no overall winners as different classifiers appear to be best in different experiments but their performance was very close and any differences appeared not significant. The ensemble approach can encompass a compromise between different models. Used to target specific groups by selecting those with a high probability to belong to the target class, it represents knowledge in a usable format.

VII. Conclusion

The primary goal of this work was to predict students that are at a high risk of not achieving a good honours degree, but more importantly, to identify this as early as possible in year 1 so that interventions can be proposed. We have been able to achieve this goal with reasonable accuracy by using classification models to highlight the students that are predicted to be low achievers with high probability. Simple models built with a few attributes known at the time of registration are sufficient to identify a group containing up to 46% of the low attainers with GH rates as low as 38%. When combining this with first year performance, we were able to identify 67% of the low attainers. The group identified had a GH rate of 24.6%.

The next stage will be to recommend strategies based on this and measure performance improvements. Additionally as further research, we will evaluate performance on year 2 and 3 modules for groups of students to study the feasibility of employing recommender systems to improve GH rates and student satisfaction. It may also be possible to incorporate data on engagement (e.g. attendance monitoring, library loans) which is becoming available in the data warehouse to see its impact on prediction accuracy.

ACKNOWLEDGMENT

This work has been partially supported by the University's Business Intelligence Unit, we thank them for their assistance with data collection and processing. We also acknowledge support from grant number ES/L011859/1, from The Business and Local Government Data Research Centre, funded by the Economic and Social Research Council to provide economic, scientific and social researchers and business analysts with secure data services.

REFERENCES

[1] "Complete university league tables," February 2015. [Online]. Available: http://www.thecompleteuniversityguide.co. uk/league-tables/rankings

[2] "The guardian league table," http://www.theguardian.com/education/universityguide, March 2015.

[3] J. Vasagar, "Most graduate recruiters now looking for at least 2:1," http://www.theguardian.com/money/2012/jul/04/graduate-recruiters-look-for-21-degree, July 2012.

[4] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*. Cambridge University Press, 2010.

[5] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," in *Procedia-Social and Behavioral Sciences*. Elsevier, 2013, vol. 97, pp. 320–324.

[6] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," in *Expert systems with applications*, vol. 41(4), no. 4. Elsevier, 2014, pp. 1432–1462.

[7] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 6, pp. 601–618, 2010.

[8] K. E. Arnold, "Signals: Applying academic analytics." in *Educause Quarterly*, vol. 33(1), no. 1. ERIC, 2010, p. n1.

[9] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelinsky, "Predicting drop-out from social behaviour of students." in *International Educational Data Mining Society*. ERIC, 2012.

[10] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 2012.

[11] J. Bayer, H. Bydžovská, and J. Géryk, "Towards course prerequisites refinement," *IMEA 2012*, p. 4, 2012.

[12] C. Vialardi, J. Chue, J. P. Peche, G. Alvarado, B. Vinatea, J. Estrella, and Á. Ortigosa, "A data mining approach to guide students through the enrollment process based on academic performance," in *User modeling and user-adapted interaction*, vol. 21(1-2), no. 1-2. Springer, 2011, pp. 217–248.

[13] C. Romero, S. Ventura, E. Vasilyeva, and M. Pechenizkiy, "Class association rules mining from students' test data." in *EDM*. ERIC, 2010, pp. 317–318.

[14] Y. Wang, N. T. Heffernan, and J. E. Beck, "Representing student performance with partial credit." in *EDM*. Citeseer, 2010, pp. 335–336.

[15] A. C. K. Hoe, M. S. Ahmad, T. C. Hooi, M. Shanmugam, S. S. Gunasekaran, Z. C. Cob, and A. Ramasamy, "Analyzing students records to identify patterns of students' performance," in *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on*. IEEE, 2013, pp. 544–547.

[16] N. T. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*. IEEE, 2007, pp. T2G–7.

[17] R. Chaturvedi and C. Ezeife, "Mining the impact of course assignments on student performance." EDM, 2013, pp. 308–309.

[18] D. Garcıa-Saiz and M. Zorrilla, "Towards the development of a classification service for predicting students'performance." EDM, 2013, p. 212.

[19] Q. A. Al-Radaideh, A. Al Ananbeh, and E. M. Al-Shawakfa, "A classification model for predicting the suitable study track for school students," in *International Journal of Research and Reviews in Applied Sciences*, vol. 8(2), no. 2. Academic Research Publishing Agency, Suite 2 Islamabad Pakistan, 2011.

[20] P. Ellett, "Understanding the undergraduate grading system in the uk," http://www.hotcoursesabroad.com/study-in-the-uk/applying-to-university/understanding-undergraduate-grading-system-in-uk/, February 2013.

[21] R. K. Goran Klepac, Klepac, *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*. IGI Publishing, 2014.

[22] M. Setnes and U. Kaymak, "Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing," *Fuzzy Systems, IEEE Transactions on*, vol. 9, no. 1, pp. 153–163, 2001.

[23] T. Thiele, A. Singleton, D. Pope, and D. Stanistreet, "Predicting students' academic performance based on school and socio-demographic characteristics," *Studies in Higher Education*, no. ahead-of-print, pp. 1–23, 2015.

[24] J. T. Richardson, "The under-attainment of ethnic minority students in uk higher education: what we know and what we don't know," *Journal of Further and Higher Education*, vol. 39, no. 2, pp. 278–291, 2015.

[25] J. Morrison, B. Merrick, S. Higgs, and J. Le Métais, "Researching the performance of international students in the uk," *Studies in Higher Education*, vol. 30, no. 3, pp. 327–337, 2005.

[26] E. Smith and P. White, "What makes a successful undergraduate? the relationship between student characteristics, degree subject and academic success at university," *British Educational Research Journal*, 2014.

[27] P. Marshall and E. Chilton, "Singaporean students in british higher education: the statistics of success," *Engineering Science & Education Journal*, vol. 4, no. 4, pp. 155–160, 1995.