Identifying at risk students using Biographical and Enrollment observations in the Earth Sciences

Jared Naidoo[1], Ritesh Ajoodha[1], and Ashwini Jadhav[2]

[1]School of Computer Science and Applied Mathematics.

[2]Faculty of Science.

The University of the Witwatersrand, Johannesburg.

Identifying at risk students using Biographical and Enrollment observations in the Earth Sciences

Literature Review

Overview of the chapter

This literature review explores papers, articles and journals concerned with the prediction of student performance. The review includes an introduction to the problem, problem statement and background to the problem. It also includes the factors which influence student performance, feature extraction, core principles of Bayesian Statistics and reviews of the methods used in current research. The literature review is completed with a conclusion which summarises the document.

Introduction

South Africa has doubled the number of students in higher education since 1994, and now has about 1 million students in the system. The number of students admitted to universities on a yearly basis constitutes 20% of the 18 to 23 year-old-age cohort (Jennifer M. Case & Mogashana, 2018). This fact is important because 20% of a particular age group is being fed into these institutions of higher education. Black South Africans now constitute 70% of university students (Jennifer M. Case & Mogashana, 2018).

The people of South Africa can be described as diverse. They have unique biographical, cultural, socio-economic and environmental backgrounds/differences across the population. This creates a diverse ,unique group of people across the population which includes the subset of people that attend university.

This uniqueness and diversity of students creates a problem whereby each student is subjected to their own experiences of life, their own trials and tribulations. These experiences which may be beneficial for some but tragic for others will have an effect on a students life. Students performance are influenced by their biographical associations and previous performance among other external factors (Ajoodha & Jadhav, 2019). A students home language, work-life situation, economic and environmental factors all have an impact on the students performance at university.

If we can find a way to measure a students performance based on their biographical information. We can then identify students that will experience difficulty at university and will

be at a risk of failing. We can then provide help to these students. Ultimately allowing them to pass successfully and complete their degree in the minimum time.

Research Question

Can we identify students that are at a risk of failing at university based on their biographical and enrolment information. This biographical/enrolment information includes but is not limited to aspects such as the students; gender, home-language, home-province and race.
The main ideas that I will be exploring is the relationship between a students academic performance and their individual biographical and/or enrolment information. The main question I will be asking is: Is there a specific biographical and/or enrolment observation or group of observations particular to a student that can correctly identify if the student is at risk of failing?

Influences on Student Performance

In this section I will be reviewing the primary influences of student performance. We need to understand what exactly affects the way in which a student performs at university. I will give special attention to biographical characteristics that influence student performance.
It was found that at-risk biographical profiles specifically English (29%) and South Sotho (13%) as being eminent spoken home languages (Ajoodha & Jadhav, 2019). This was found amongst the at-risk students and it presents something interesting. English making up 29%, is not surprising seeing that English is a popular language in South Africa. The interesting part is that South Sotho makes up 13% of the at risk student profiles, 13% is quite a large number for a language that is not spoken by many South Africans.
I would say that 13% indicates a possible problem that South Sotho speaking students may be experiencing. Seeing that the majority of courses are lectured in English at the University of the Witwatersrand (primary data source for this study). A lack of understanding the language would play a vital role in failure.
Another important factor to consider is race. The majority of at-risk biographical profiles have an associative Black race description (71%) (Ajoodha & Jadhav, 2019). This is an astronomical percentage of a particular race group that are at risk of failing.
Given that the study was done in South Africa using data collected from South Africans. Race would be a key characteristic to examine given the countries history. It is relevant in the study

mainly because of the restrictions placed on particular racial groups in the past. The restrictions could be affecting the descendants of these previously disadvantaged people either economically or environmentally.

This paper also notes the provinces in which these students whom are at risk are generally associated with. MP(17%), GA(14%), FS (13%) and NW (13%) as being notable provinces (Ajoodha & Jadhav, 2019). The key point or take away from this would be that there are 9 provinces in South Africa and of that 9 provinces 4 of them are making up more than 50% of students that are likely to fail in this study.

High school rank or the quality of education taught at a high school level should be looked at. The paper Campbell and P. McCabe (1984) looks into the realm of students completing a degree based on their high school marks or SAT scores.

The results indicate that students who persisted in a major in computer science, engineering, or other sciences differed from those students who left computer science for an academically dissimilar goal. These differences were related to the students' SAT math and verbal scores, their high-school rank, and their back- ground in high-school mathematics and science (Campbell & P. McCabe, 1984).

It is evident that the high school rank has played a vital role. A link can be created between the high school rank and the area of which the high school is situated. A school situated in the north of Johannesburg may perform significantly better than a school situated in a township. The location of the school could be linked to demographics such as the students high school or home address.

Of the 98 women in the sample, only 38 (39 percent) persisted in scientific and engineering majors, whereas 96 out of 158 men (61 percent) persisted (Campbell & P. McCabe, 1984). The outcome here is that generally the males were performing better than the females in areas such as science and engineering. This indicates a trend that gender stereotypes and cultural stereo types may have an impact on the students performance.

This section tried to find potential biographical information or characteristics that are key in determining a students performance. Later on we will have to do feature selection or maybe even feature scaling (removing unnecessary features). As a result proper understanding of characteristics which greatly influence the students performance is vital.

Feature Extraction

A good feature extraction out of a given data set has proven to be a key element in the success of a machine learning experiment. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information (Ramaswami & Bhaskaran, 2009). The elimination of these features that yield no value to our model can ultimately reduce computation time and improve the accuracy of our model.

The performance of prediction model highly depends on the choice of selection of most relevant variable from the list of variables used in student data set (Ramaswami & Bhaskaran, 2009).

Seeing that the performance of our model is key, we would measure this by the number of incorrect vs correctly classified students (False Positives/True Positives).

This can more generally be described as the accuracy of our experiment/model. A good feature set that is well balanced will give us better accuracy scores. Furthermore selecting the features that truly reflect the outcome will ensure that we do not feed bad or unnecessary date into our models. After all why put data into a model that the model does not need.

Feature selection is normally achieved by going through our feature set of attributes and reviewing each attribute. Various mathematical techniques can be used as well. If we are using a Decision Tree, then Information Gain can be calculated which would give us a good indication of the strength or influence that an attribute has. Some examples of techniques used for feature selection from Ramaswami and Bhaskaran (2009) are listed:

1) Correlation-based Attribute evaluation (CB),
2) Chi-Square Attribute evaluation (CH),
3) Information-Gain Attribute evaluation(IG),

My review and study of the literature found that when given a list of features, you should always analyse each feature and investigate the relevance of the feature in your model. Ask yourself the question should this feature really be in the model? Does it yield any value?

Methods used in different papers

Selected Approach

The preferred approach approach to solving this problem would be through the use of Bayesian statistics. This section explores Bayesian Statistics, Compares Decision Trees to Bayesian Networks and looks at past research concerning prediction of student performance. This is subject to change after exploring the dataset.

Bayesian Statistics

A Bayesian network can be described as a probabilistic graph model whereby the nodes represent random variables and the edges represent conditional independence assumptions. They provide a compact representation of joint probability distributions (Ajoodha, 2018). Bayesian networks are ideal as they can take an event that has occurred and calculate the probability of several known causes, we can then find the main contributing factor i.e. What caused the outcome of the situation?

When compared to other machine learning models Bayesian networks do have some advantages. They are not as data hungry as a deep learning model, they can accommodate smaller data sets and cannot be fooled by adversarial examples (Ajoodha, 2018). Another important aspect is uncertainty, Bayesian Networks are known to represent uncertainty within a problem in a much better way than say a deep learning alternative (Ajoodha, 2018). Through the use of probability theory, we can express noise and uncertainty within our problem. We can then use the inverse probability theory to make inferences on unknown quantities.

Some examples of Bayesian estimation taken from Ajoodha (2018) are that they span a range of applications including general diagnostic systems, event forecasting and machine vision.

Decision Trees vs Bayesian Networks

Decision Trees are one of the most used techniques when dealing with inductive inference. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance (Mitchell, 1997).

A decision tree is a classification method used to predict outcomes. Usually the structure of a decision tree is dependent on the problem. I.e. Decision tree structures are dependent on

various mathematical calculations. Information Gain is calculated and then a root node is assigned, followed by the remaining edges and nodes. These calculations are based on the feature set. This forms what we call the structure of the decision tree. Whereby each node represents a feature/outcome unique to that problem. Then based on a choice made at each node, we travel down the tree until we reach the end of the tree and settle on the outcome depicted by that path.

We can examine the decision tree implementation in Thai-Nghe, Janecek, and Haddawy (2007). They compared the performance of a decision tree to a bayesian network. Although the study indicates the performance of the decision tree was better than the bayesian network. What we can examine and possibly acknowledge is that they are looking at a set of features which fall out of biographical characteristics. Examples of this are features such as religion ,level of skill at English and even the rank of the institution. This set of features may not be as useful in our study. We may apply feature scaling to remove some of these unnecessary features. Their set of features contains 14 elements when compared to the set of features in Ajoodha and Jadhav (2019) there is a size difference. Ajoodha and Jadhav (2019) uses a smaller set of features. This comparison is a valid indication that we cannot assume based on Thai-Nghe et al. (2007) that a decision tree would necessarily outperform a bayesian network in all cases.

Examining the results found in various papers

In Campbell and P. McCabe (1984), they utilised the students high school marks. i.e. The students math, science and english SAT scores. They also factored in the students sex. The problem here is that this study completely excludes psychological, socio-economic and environmental factors. They gage performance using marks/scores but have not explored the possibility that these scores are affected by other external factors.

In Campbell and P. McCabe (1984) they present an important finding, Of the 98 women in the sample, only 38 (39%) persisted in scientific and engineering majors, whereas 96 out of 158 men (61%) persisted. This is key as to why biographic information must be used in studies of student performance. Reason being is that there must be some reason as to why the men were persisting with the science/engineering degrees and why these women were either dropping out or changing their area of study. What makes it more interesting is that the women completed the same subjects in high school as their male counterparts. This study takes a

statistical approach to solving the problem rather than a machine learning approach. It would be interesting to see the results produced when machine learning techniques are applied to the problem.

In Thai-Nghe et al. (2007), we note that they explore two methods of machine learning to predict student performance. The models compared were Decision Tree and Naive Bayes. In this paper the first potential problem that I noticed would be the data set. The dataset contains the classification labels of; failed, fair, good and very good. These refer to the passing/failing of a student. My main observation is that the data set is not evenly balanced, there are 9765 samples that were classified as good. 9765 out of a total of 20 492. Almost 50% of the dataset was labelled as good, leaving the remaining samples to be divided amongst the remaining three classes. This could definitely cause the models to exhibit a bias in the testing/validation phases, specifically if the testing/validation dataset contains a similar sort of bias.

In the paper Ajoodha and Jadhav (2019), when reading the graphical representations of the analysis. We can see that there is a large volume of students that are from Gauteng and at risk of failing. This raises the question; is there a problem with the education system in Gauteng? A conclusion could be that this is because the University of The Witwatersrand is based in Gauteng, hence most of their students are from Gauteng.

An important observation that I have made is that the number of black students in the experiment are larger than the other races. My concern would be that there are a a lot of samples marked as black under race vs indian, white and coloured. Could this cause some sort of imbalance within the data? Seeing that are more samples marked as black vs the other races.

The main idea behind this section was to look at the research papers objectively, to try and find any potential problems that the researchers experienced. To learn from their experiences and apply this knowledge to my own research.

Conclusion

South Africa is a unique country which boasts a diverse population. This creates a unique and diverse culture at universities across the nation. Our goal is to find out what biographical characteristics can cause these students to under-perform at university and ultimately identify students that are at a risk of failing.

We identified possible influences on student performance and then narrowed it down to biographical characteristics that can influence student performance. These characteristics include their home language, home province and race. Understanding what influences students will allow us to select accurate features when modelling our problem.

Feature extraction was then explored. We reviewed the performance of models that had feature extraction applied vs models that had no feature extraction applied. We deduced that feature extraction out of a large dataset is key to producing accurate results.

Bayesian Statistics and Decision Trees were explored. We looked at the performance of each method. Although decision trees performed better in the research we were able to justify our approach of using Bayesian Statistics to solve this problem. Mainly because we will be using a more refined dataset similar to the dataset found in Ajoodha and Jadhav (2019).

The report was concluded by examining and reviewing each method of predicting performance used in the research. Looking specifically at the data used and checking for inconsistencies such as a bias in the data or anything in the results that did not make sense. Special attention was given to the fact that we would be working with South African data, this is because South Africa experiences its own unique problems. We would need to cater for these unique problems in the forthcoming research.

The main purpose of this literature review was to look at previous research conducted around my topic and to find out what issues and difficulties other researches had experienced . I now understand all of my elected readings in great detail and feel confident enough to start writing the initial draft of my research proposal.

References

Ajoodha, R. (2018). Representation, inference and learning. *IndabaX Conference*.

Ajoodha, R., & Jadhav, A. (2019). Identifying at-risk undergraduate students using biographical and enrolment observations for mathematical science degrees at a south african university.

Campbell, P., & P. McCabe, G. (1984, 11). Predicting the success of freshmen in a computer science major. *Commun. ACM*, *27*, 1108-1113. doi: 10.1145/1968.358288

Jennifer M. Case, S. M., Delia Marshall, & Mogashana, D. (2018). *Going to university, the influence of higher education on the lives of young south africans* (Vol. 3). 4 Eccleston Place, Somerset West 7130, Cape Town, South Africa: African Minds.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Science/Engineering/Math.

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining.

Thai-Nghe, N., Janecek, P., & Haddawy, P. (2007, 11). A comparative analysis of techniques for predicting academic performance. In (p. T2G-7). doi: 10.1109/FIE.2007.4417993