Sensitivity of Predictors in Educational Data: A Bayesian Network Model

Nicholas Misiunas*, Miroslava Raspopovic†, Kavitha Chandra*, Asil Oztekin○

*Francis College of Engineering
University of Massachusetts Lowell
Lowell, MA, USA

†Faculty of Information Technology
Belgrade Metropolitan University
Belgrade, Serbia

○Manning School of Business
University of Massachusetts Lowell
Lowell, MA, USA

nicholas.misiunas@student.uml.edu

Abstract

This research investigates the application of Bayesian Networks to predict causal relationships in a dataset that captures several demographic and academic features of a group of students from a four-year public university. This educational dataset is characterized by both quantitative and qualitative variables, some of which exhibit a strong pair-wise dependence. To identify this dependence, a factorial analysis of the mixed data is conducted that allows consideration of both variable types to result in a new coordinate space that captures the variance of the data with fewer dimensions. This exploratory stage enables visualization of groups of dependent variables that may be applied for predicting outcomes of interest. It also provides a validation of the results of the Bayesian network (BN) structure modeling. The BN is learnt using bootstrapped arc strength averaging to derive a graphical relationship between variables with arcs represented by a persistence parameter representative of their occurrence in the learning process. The resulting network is shown to be characterized by two major relatively independent structures; one formed by college academic performance metrics and the second by financial, housing and student demographic variables. The prediction accuracy of the BN is evaluated using evidence from pre-college and on-going college variables. The pre-college evidence is found not to be sensitive to college degree completion outcome with only 55% accuracy rate. The on-going college evidence however, improves the prediction accuracy by 75%.

Keywords: factorial analysis, mixed data, bayesian networks, education data mining

1 Introduction

According to US Department of Education, the 6-year graduation rate for the first-time enrolled full-time undergraduate students is 59%, while graduation rates are the lowest at the institutions that are the least selective on their admission criteria [1]. The degree completion rate has been one of the major factors used for analysis and research in many fields that examine the influence of educational attainment on higher earning potential and lower unemployment rates [2]. It is no surprise therefore, that statistics on graduation and retention data are often used as contributing factors for assessing the quality of academic institutions and their relative rankings. The relationship of academic and demographic factors that influence graduation outcomes has been an area of interest to educational institutions, such as universities that aim to improve their graduation rates or understand what causes their unique attrition rate.

Academic analytics (AA) is an emerging area of data mining techniques applied to monitor and anticipate students performance by combining institutional data, statistical analysis and predictive modeling that can be used to change
academic behavior [3]. Early examples of AA showed that academic success can be predicted using high school GPA and SAT scores, identifying SAT rates as being the most important feature affecting graduation rate [4]. DeAngelo et al. used high school GPA, SAT scores, on- and off-campus activities to predict graduation rates using logistics regression models [5]. The current focus in AA research is on providing timely intelligence, based on which individual academic performance can be improved. Thus, care must be taken to not only ensure that results are meaningful but also readily interpretable to stakeholders from diverse fields. AA utilizing big data collection and analysis has the potential to enhance numerous factors from the students experience, improve academic programs, and enable evidence-based decision making, by turning unstructured data into actionable information [6]. Although the focus has been mainly on using regression and structural equation models for predicting student performance, Bayesian Networks (BN) have been investigated in recent studies [7, 8]. BNs are advantageous in that their structure can be learnt from the data. This gives it an advantage over other methods, that may heavily depend on expert knowledge to develop a model, which may or may not end up being an accurate reflection of the true relationships amongst the features.

In this work, a dataset that describes selected attributes of students from five different colleges in a four-year public university is mined factorial analysis of mixed data (FAMD) and BNs. The objective is to infer the likelihood of graduation in six years conditioned on evidence provided by groups of other features such as high school performance, SAT scores, recent semester grade point average (GPA), and demographic variables. Of particular interest is to identify the sensitivity of these groups of variables, each addressing complementary attributes of the students in their ability to predict graduation status. Section 2 describes the dataset. Section 3 discusses the results of FAMD and presents the relationships between groups of variables. Section 4 presents the BNs learnt from the data and interpretation of the structure and predictions generated by the BN. Section 5 concludes the paper.

2 Data Description

The data analyzed in this paper includes \( I = 1204 \) records, each record containing thirty one features of one generation of students distributed across five colleges in a four-year state university. Each record captures: College and Major, Demographic attributes (age, gender, ethnicity, citizenship), High-School data (school attended, GPA, SAT scores), College Academic Performance variables (Full time/Part time, credits attempted, earned Fall GPA and Spring GPA), Residency Status (In-state, Out-of-State, International, Proximity program, regional exchange), Financial attributes (Need indicated, financial aid applicant, student loan, scholarship, grant recipient, parent loan), Work status, Commuting Status, and the six-year degree completion. The Major and high schools attended variables are excluded from this study since the Major carries similar information as College and the large number of names of high schools (275) attended relative to the data size does not provide enough records per school to warrant their inclusion. The features that will be considered are listed in Table 1. It is also noted that four of the categories in the Ethnicity variable that each accounted for a very small percentage of data were combined into one category, yielding five groups: White (76%), Asian (8.6%), Hispanic/Latino (6.2%), Black/African-American (3.6%), and Other (5.6%). All continuous valued variables were quantized to a finite number of levels when applied to building Bayesian networks in Section 4. SAT score values were mapped to six levels \( \{0, 800, 1000, 1200, 1400, 1600\} \) resulting in five categories, \( \{0, 1, 2, 3, 4\} \). GPAs for High school, Spring and Fall semesters and semester credits attempted and earned were also quantized into five categories. The data includes \( K_1 = 13 \) quantitative variables, and \( K_2 = 16 \) categorical variables that are composed of 48 categories, resulting in a feature set of dimension \( J = 13 + 48 = 61 \).

3 Factorial Analysis of Mixed Data

FAMD [9, 10] was devised for joint consideration of quantitative and qualitative variable types without conversion from one to the other. It operates in a manner that draws from the methodology of both principle component analysis (PCA) and multiple correspondence analysis (MCA). The data is represented in a matrix \( X \) of dimension \( I \times J \). The quantitative elements in the \( i^{th} \) row of \( X \), \( x_{ik1} \) are standardized with division by the standard deviation \( \sigma_{k1} \) for all \( k_1 \in K_1 \). Each quantitative variable \( y_{ijk2q} \) for the \( q^{th} \) category of \( k_2 \in K_2 \) is characterized by its frequency of occurrence \( p_{k2q} \) and standardized by dividing it by \( \sqrt{p_{k2q}} \). With this scaling, categories with rare occurrence are weighted higher and appear at larger distances compared to categories that have higher frequency of occurrence. The principal components are determined by singular value decomposition of the centered matrix \( \mathbf{XD}^{-1/2} - \mathbf{M} \) where \( \mathbf{D} \) is a diagonal matrix with scaling elements \( \sigma_1^2, \sigma_2^2, ... \sigma_{K_1}^2, pK_1+1, pK_1+2, ... pJ \) and each row of \( \mathbf{M} \) is the
vector of mean values of each column of \(X\). In this method, the first principle component \(z_1\) maximizes the projected variance of a linear combination of the variables by maximizing \( \sum_{k \in K_1} r^2(k, z_1) + \sum_{q \in Q} \eta^2(q, z_1) \), where \(r^2\) is the squared correlation coefficient for quantitative variables and \(\eta^2\) is the squared correlation ratio for categorical variables. The \(j^{th}\) element of the \(i^{th}\) principal component \(z_{ij}\) can decomposed as,

\[
z_{ij} = \frac{1}{\sqrt{\lambda_j}} \sum_{k \in K_2} p_{kq} \left( \frac{y_{ikq}}{p_{kq}} - 1 \right) a_{jkq} + \frac{1}{\sqrt{\lambda_j}} \sum_{q \in Q} \eta_{q} \left( \frac{y_{iq} - \eta_{q}}{\eta_{q}} \right) a_{iq}.
\]

where \(\lambda_j\) is the \(j^{th}\) eigenvalue and elements \(a_{jkq}\) are the \(k^{th}\) element of the \(j^{th}\) eigenvector. In this equation the first member is of PCA and the second is of MCA, illustrating the concepts drawn from each, in which an individual relates to those variables in which it has an above average value (and vice versa) and that an individual is at the center of the cloud of categories it possesses, respectively. FAMD results were obtained using the package FactoMineR \[11\]. The analysis redistributed the variance in the data such that when ordered, the first twenty-two PCs capture cumulatively 80% of the variance in the data. The first PC (PC1) accounts for 14.18% variance with the second (PC2) accounts for 7.77%. Fig. 1 is the relationship square for PC1 and PC2. The graph represents the squared correlation coefficient for quantitative variables and the squared correlation ratio for qualitative variables. The correlation may be found as \(S = VL^{1/2}\) and \(V\) is either the correlation coefficient in the case of a quantitative variable or the correlation ratio in case of a qualitative ratio and \(L\) is the diagonal matrix of singular values.

Both quantitative and qualitative variables appear in this figure. Qualitative variables are not decoded into their respective categories, though this can be seen in Fig. 2. PC1 is primarily correlated with college academic performance attributes [Table 1: No. 6-13, 29]. Degree completion rate (DCR) is partially correlated with PC1 and thus is partially correlated with most college academic performance attributes. An examination of the relationship of DCR to remaining PCs showed a moderate correlation with PC29 and PC31. Ethnicity and merit scholarships were found to contribute to both these PCs while the first also possesses college attributes and the second includes high school GPA. DCR is thus a variable that may be associated across multiple PCs with its correlation to college academic attributes being the most significant.

Components of PC2 that have the highest correlation are primarily SAT scores. As the PCs are orthogonal, it can be noticed that most college academic performance attributes are largely uncorrelated with SAT scores. High school GPA is uncorrelated with SAT scores and primarily uncorrelated with college academic performance but may have some relationship to DCR and the remaining variables that are also uncorrelated with both college academic performance and SAT scores.

Qualitative variables are decoded into their respective categories in Fig. 2 where the horizontal axis is \(z_1\) and the vertical axis is \(z_2\), and each category is at the center of gravity of the individuals exhibiting that category. As PC1 and PC2 are again utilized, the horizontal axis is aligned with college academic performance and the vertical with SAT scores. This is supported by the positions of DCR (DCR\_N is taking more than six years to graduate and DCR\_Y is taking six or less) along the horizontal axis. Many variables are clustered near the origin and these are thus uncorrelated with academic performance and SAT scores. These variables will appear in other PCs. DCR itself has a more complex relationship with many of the variables, illustrating that part-time students typically take longer to graduate and that the college, ethnicity, and merit variables are weakly correlated with graduating in six years. It should be noted from Fig. 2 that students who have less of financial need, are in category Ethnicity-White, Gender-Male, have high SAT scores and are intercorrelated with the Engineering and Science colleges, while students who have higher financial need, Gender-Female, are of Ethnicities Asian, Hispanic and Black/African American are intercorrelated with the Social Sciences, Fine Arts and Humanities and Business colleges. These observations could be indicators to university administrators towards the needs of particular colleges. For instance, it may be of interest to analyze the outlier in Fig. 2 which indicates that black/African Americans and permanent residents, may perform poorly on the SAT but this is not correlated with DCR. Some outliers are present, such as non-resident citizens and international students, who perform exceedingly well on the SAT and graduate quickly. The Hispanic/Latino citizen category is a potential outlier and investigation of the dataset showed it may be entered erroneously, being possessed by only a single student.

4 **Bayesian Networks**

In this section, the dependence between the quantitative and categorical variables of the dataset is further investigated using a BN. The objective is to obtain the probability of outcomes of interest conditioned on a set of observations that may be referred to as evidence. The BN represents the data as a directed acyclic graph \(G\), consisting of a set of nodes
\( V_G \), edges \( E_G \) with a joint probability density function \( p_G \). The nodes represent the \( K \) variables and the connection between two nodes is determined by data derived conditional probability tables that capture the joint probabilities of the categories of the two connected nodes [12]. Each node \( v \in V \) is characterized by the set of parent nodes \( Pa(v) \) that directly point to \( v \). Nodes that are not connected are characterized as being conditionally independent. The BN decomposes the \( K \)-dimensional joint probability distribution \( p_G \) in product form if subsets of variables are determined to be conditionally independent. In such a case, each node is completely characterized by the values assumed by its parent nodes. These local probabilities encode each node, usually in the form of conditional probability tables with values computed from the data. The joint probability density function is then \( p_G(V) = \prod_{i=1}^{K} p(v_i | Pa(v_i)) \).

The BN structure is learnt using a combination of hill-climbing, cross validation, and bootstrapping/model averaging. The R package \texttt{bnlearn} [13] is applied in this effort. Hill-climbing is a local search method that begins with an initial network structure (such as an unconnected network, randomly generated network, or a given preset) and makes changes to individual nodes. This set of local operations is performed for each node in the network and the process is repeated until a locally optimal network is found. The network score is computed from the Bayesian Information Criterion which maximizes the log likelihood function estimated from the data while discounting the size of the parameter set of the model.

Bootstrapping, which entails creation of multiple datasets known as replicates by resampling with replacement, allows exploration of the stability of the Hill-Climbing algorithm in its convergence to the optimal BN. The replicate size is set to match the size of the training dataset. A total of \( 10^4 \) replicates are generated, each of which are used to train a BN. The resulting BNs are ensembles of a random graph representation of the data, that are averaged to yield a final BN. Each arc in the averaged BN is given a weight \( w_a \) ranging from 0:1 representing its persistence in the \( 10^4 \) ensembles. The thickness of the arcs in the final graph are proportional to the weights. In the results shown only arcs with strength greater than 0.5 are retained in the averaged model.

The ensemble averaged BN obtained from bootstrapping is shown in Fig. 3. This graphical structure projects the dependence relationships between variables as learnt from the data. Degree completion rate (DCR) located at the bottom left of the graph is clearly shown to arise from the college academic performance attributes with SGPA being its parent and related hierarchically to attributes: SEAR, SATT, FEAR, FATT, and FENR along this branch of the graph with College being the parent node at the top. This structure derived independently from the BN supports one of the main results of FAMD, that showed these very attributes being correlated along PC1. A secondary relationship between DCR and HSGPA was also found in FAMD analysis that is shown in the BN, through the dependence of SGPA being the parent of both DCR and FGPA which is a parent of HSGPA. The Merit attribute which is a binary variable is also a parent of HSGPA, SAT and Financial Aid application status, a Yes/No variable type. This connection is expected as merit scholarships often are related to performance in high schools and SAT scores. The major branch on the right hand side of the graph includes mainly the various financial, housing, ethnicity and citizenship indicators and found to be independent of DCR. A parallel can be drawn from this to the result of FAMD and the categories plot. Fig. 1 also shows the close relationship between MER and FAA along PC1 which relation is captured in the BN with MER being a parent of FAA.

In order to further analyze these relationships, connections between the attributes were analyzed based on the persistence of each connection for \( w_a > 0.9 \) and those less persistent \( 0.75 \leq w_a < 0.9 \). Fig. 4 shows the four highly persistent independent BN structures that includes, connections between College and Gender, Ethnicity and Citizenship, Housing and Work, a second group that connects the financial attributes and the last two that identifies the connections in Fall and Spring college performance. The connections between financial attributes indicate that students that have financial need will most likely receive a scholarship or will take a loan. The persistent connection between Fall GPA and Spring GPA is of interest, implying that students may maintain the level of their academic performance throughout the school year.

Fig. 5 shows the graph with moderately persistent arcs, with DCR depending on spring GPA, while Fall GPA depends on the high school GPA. This indicates that even at this level of persistence, there is still a relationship between performance of students in high school and their performance at the college. In other words, in order to be able to address academic issues that can slow down the graduation rate, attention should be paid to progress indicators at the beginning of the students’ studies.

Finally, a sample of the cross-validation studies performed to predict two cases: pre-college and on-going college cases is given here. Ten-fold cross-validation is utilized, where 90% of the dataset is retained for training, whilst the remaining 10% is utilized for testing in each fold. The testing set utilizes a subset of variables to act as evidence for
which the BN predicts the probability of the outcome, using the model generated from the training set. The category with the maximum predicted probability is selected and the result compared with the outcome recorded in the data. In the pre-college case the student is assumed to be in high school and so the evidence variables are: AGE, GEN, ETH, CIT, HSGPA, SAT, FAA, NIN. In the on-going college case the given evidence is: COL, AGE, GEN, ETH, CIT, RES, HOU, SCH, MER, GRA, SLO, PLO, WOR, FGPA, SGPA. The pre-college case evidence is used to predict college academic performance and the on-going college case will predict DCR as well as HSGPA and SAT scores. Pre-college evidence results in poor prediction accuracy of college academic performance with (FGPA: 47%, SGPA: 46%, DCR: 55%). The low levels of correlation observed from FAMD results between these group of variables may support this observation. The ongoing college evidence however results in a much higher prediction accuracy of DCR (75%) and the student’s high school GPA (76%).

5 Conclusion

The results from FAMD and BN illustrate some interesting parallels between the two methods and the ways in which complex relationships amongst sets of variables may be understood. Both methods show that the college academic performance features are relatively easy to model but their relationship to the other variables is weaker and may exist through a potentially more complex relationship through intermediary features such as high school GPA. The colleges themselves are individually correlated with different demographical groups. The financial and demographic features are related to college academic performance. Efforts to predict college academic performance must therefore focus primarily on knowledge already known by the college such as semester GPA or credits taken/earned, or through high school GPA. Moreover, additional features outside the range of this dataset may provide more insight into the relationship amongst the features in the models demonstrated here.

References


Table 1: Variable Ranges and Codes

<table>
<thead>
<tr>
<th>No</th>
<th>Quantitative Variable</th>
<th>Range</th>
<th>Code</th>
<th>No</th>
<th>Quantitative Variable</th>
<th>Categories</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age [17:29]</td>
<td>AGBR</td>
<td>COL</td>
<td>14</td>
<td>College</td>
<td>5</td>
<td>COL</td>
</tr>
<tr>
<td>2</td>
<td>SAT [500:1600]</td>
<td>SAT</td>
<td>GEN</td>
<td>15</td>
<td>Gender</td>
<td>2</td>
<td>GEN</td>
</tr>
<tr>
<td>3</td>
<td>SAT Math [250:800]</td>
<td>SATM</td>
<td>RES</td>
<td>16</td>
<td>Resident</td>
<td>5</td>
<td>RES</td>
</tr>
<tr>
<td>4</td>
<td>SAT Verbal [250:800]</td>
<td>SATV</td>
<td>CIT</td>
<td>17</td>
<td>Citizen</td>
<td>5</td>
<td>CIT</td>
</tr>
<tr>
<td>5</td>
<td>High School GPA [0.5]</td>
<td>HSGPA</td>
<td>ETH</td>
<td>18</td>
<td>Ethnicity</td>
<td>5</td>
<td>ETH</td>
</tr>
<tr>
<td>6</td>
<td>Full GPA [0.4]</td>
<td>FGPA</td>
<td>HOU</td>
<td>19</td>
<td>Housing</td>
<td>2</td>
<td>HOU</td>
</tr>
<tr>
<td>7</td>
<td>Fall Attempt Credits</td>
<td>[1.20]</td>
<td>FATT</td>
<td>20</td>
<td>F.A. Applicant</td>
<td>2</td>
<td>FAA</td>
</tr>
<tr>
<td>8</td>
<td>Fall Earned Credits</td>
<td>[0.20]</td>
<td>FEAR</td>
<td>21</td>
<td>Merit</td>
<td>2</td>
<td>MER</td>
</tr>
<tr>
<td>9</td>
<td>Fall Compl. Rate</td>
<td>[0.1]</td>
<td>FCOM</td>
<td>22</td>
<td>Need Indicated</td>
<td>6</td>
<td>NIN</td>
</tr>
<tr>
<td>10</td>
<td>Spring GPA [0.4]</td>
<td>SGPA</td>
<td>AHLT Schol.</td>
<td>23</td>
<td>AHLT Schol.</td>
<td>2</td>
<td>SCH</td>
</tr>
<tr>
<td>11</td>
<td>Spring Attempt Credits</td>
<td>[1.20]</td>
<td>SATT</td>
<td>24</td>
<td>Gram</td>
<td>2</td>
<td>GRA</td>
</tr>
<tr>
<td>12</td>
<td>Spring Earned Credits</td>
<td>[0.20]</td>
<td>SEAR</td>
<td>25</td>
<td>Work</td>
<td>2</td>
<td>WOR</td>
</tr>
<tr>
<td>13</td>
<td>Spring Compl. Rate</td>
<td>[0.1]</td>
<td>SCOM</td>
<td>26</td>
<td>Student Loan</td>
<td>2</td>
<td>SLO</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27</td>
<td>Parent Loan</td>
<td>2</td>
<td>PLO</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>28</td>
<td>Full Enr. Status</td>
<td>2</td>
<td>FENR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29</td>
<td>Degree Completion 6 Yr.</td>
<td>2</td>
<td>DCR</td>
</tr>
</tbody>
</table>
Figure 1: Relationship Square for PCs 1,2

Figure 2: Rep. of Categories for PCs 1,2

Figure 3: BN without Cross Validation.

Figure 4: Threshold $> 0.9$

Figure 5: $0.75 < \text{Threshold} < 0.9$