International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 3 Issue 6‖ June 2014 ‖ PP.10-13

Predictive Models To Understand And Analyze Student Performance In Post Graduate Entrance Examinations - A Case Study.

Ananth.Y.N[1], Narahari.N.S[2]

[1]*Associate Professor & Ph.D student- Dept of Computer Science- School of Graduate Studies,JainUniversity,Bangalore-INDIA*

[2]*Professor and Head, Dept of Industrial Engineering & Management, R.V.College of Engineering Autonomous-VTU - Bangalore-INDIA*

ABSTRACT: *Post graduate admissions in India happen mostly through entrance examinations. These normally consist of multiple choice questions-While this format is suitable for many types of examinations, one has to look at the student aptitude also while devising a question paper. There has to be a right balance between the strength of the students and the difficulty level of the questions. This balance is difficult to achieve. This paper discusses some techniques to know the students strengths and to devise questions based upon them. Data mining techniques, Item analysis, Regression analysis and probabilistic techniques are used here.*

.***KEYWORDS :*** *Data mining, Item Analysis, Predictive Model, Probability ,Regression Analysis,*

1. INTRODUCTION

Post graduate admissions in India happen mainly through entrance examinations. In this competitive scenario, there are two facets- one is the examinee or the student who take the examination – he has got his own strengths and weaknesses- methods of answering –knowledge about the subject etc. The other part is the actual examination and the question paper which should cater to the multiple types of students taking the examination. While the examiner focuses on giving as tough questions as possible, the examinee tries to find as many ways and means of answering the maximum number of questions. In this regard, if there is some means by which the examiner can come to know of the strengths and weaknesses of the student community that he is addressing, then it becomes easier for him to set the question paper. This paper discusses some predictive methods to do this and try to establish a model based on which the question papers can be set. The particular case study which has been taken up is the Karnataka PGCET examination This paper includes two main parts. In the first part, the methods to know the strengths and weaknesses of the students are discussed. This includes cluster analysis, classification, Item analysis, Regression analysis – all of these applied on the marks scored by the students. The second part discusses a probabilistic model to devise multiple choice questions. The output of the first part is given as input to the second part..

II. LITERATURE SURVEY

Data mining techniques have been used in Educational Data Mining to correlate the students' data across various levels. Item Analysis and variations of Rasch models have been used in analyzing the examination scores to find out patterns of scoring.Probabilistic techniques have been used in studying the patterns of scoring in various examinations.

III. CLUSTER ANALYSIS

The basic premise of this study is that the marks that could be scored by a student in the PG- entrance examinations is directly proportional to the marks that he or she has scored in their graduate examinations. This is because the aptitude of a graduate/post graduate student would have been built over a period of several years and it is sensible to assume that the performance of the student in the graduate exams is an indicator of his future performance in the PG- entrance examinations. There are exceptions to this assumption but nevertheless it is a powerful dependency to be studied. So, in this study the marks that have been scored by the students in PG-entrance test has been considered, the correlation between the degree marks and the entrance test marks for the past three years have been studied. In the first stage, we study the dependency between the entrance test marks

and the rank scored by the students. The following snapshot shows this dependency. This snapshot has been taken with the software Rapid Miner.

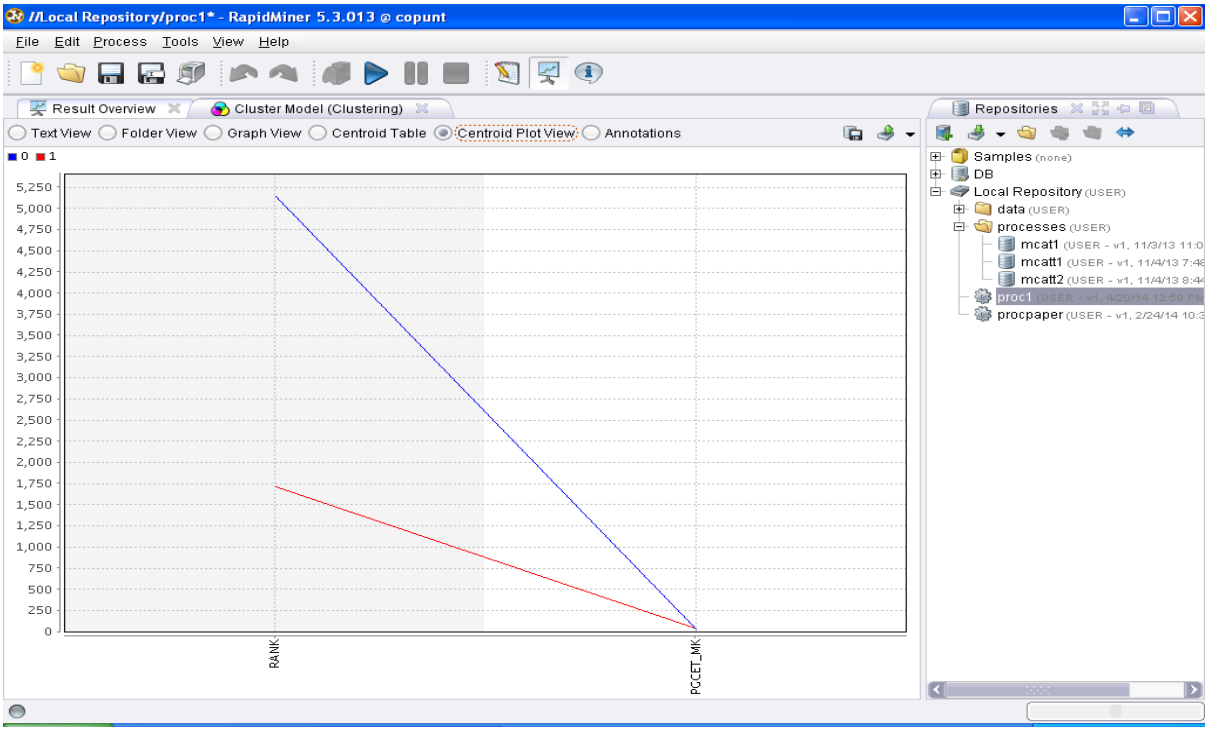

Fig.1

In the next level the dependency between the degree marks and the PG ranks have been studied.The following snapshot shows this dependency which has been taken by SPSS software.

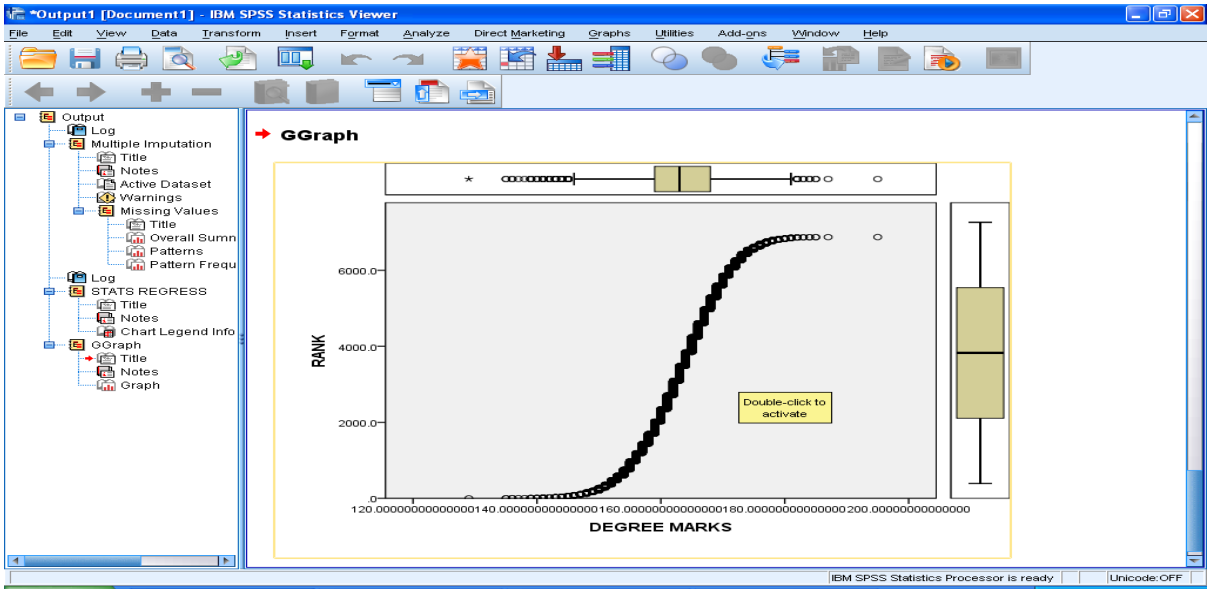

Fig.2

Observing the above we can infer that there are clusters of the students who have scored almost the same marks but have been awarded different rankings. This is of course because of the fact that the weightage for this marks is 50%-but the fact that has to be noted is that there are students with the same intelligence level who have taken up the examination –but due to some reason they have scored different marks in the PG entrance and hence the different rankings.

What we are essentially doing here is trying to analyze the clusters which have been formed based upon the distribution of marks –In the next step we can goto those individual clusters and use classification to classify those clusters into classes created for our analysis purpose. For example, we can identify clusters in which the PG entrance test marks is about 65 and the group's degree marks is around 60%.When this analysis is done for the scores over many years (in this study, it has been three years)what comes out is that students who are taking a certain marks in the degree are scoring a definite marks in the PG entrance test. This sets a kind of upper limit and a lower limit on what marks that a student can score in the PG entrance test. Although the type, number and weightage of the questions differ in these two examinations , indirectly or directly the degree marks is a strong indicator of the marks that a student can score-in other words it speaks of the aptitude and strength of the student. A predictive model can be established based upon this kind of an analysis – which will tell us given a cluster of students with a certain strength what are the strengths and weaknesses of that cluster of the students.

IV. REGRESSION ANALYSIS

The above analysis gives us the number of clusters scoring a particular range but this study can be further improved by performing a regression analysis on the scores. The independent variables here are two - namely the degree marks and the entrance test marks and the dependent variable is the ranking that has been obtained. The general model[1] that can be followed is that of multiple regression which looks like the following:

$$Y= a + b1x1 + b2x2 \quad (1)$$

Here Y= the rank scored
X1 = the degree marks
X2 = the entrance test marks
b1 and b2 are the regression coefficients-

V. ITEM ANALYSIS

The above analysis is with respect to the marks scored. Now we can refine this analysis by looking into the answers that a student gives in the actual examination. The questions that appear can be divided into many categories like knowledge based, skill based memory based and so on. In each of these categories we can further have difficulty levels assigned.[2] So by comparing the answers and trying to find their relationship with the clusters that we have found in the previous steps, we can come to a definitive picture as to what is the aptitude of a certain student. This can be studied over a set of years' data and we can use the inferences to build a model which would tell us what might be the answering level of a certain kind of clusters, before we can actually frame the question paper. In effect, this method can be used as a valid elimination technique with a certain degree of confidence.

VI. SETTING MULTIPLE CHOICE QUESTION PAPER- A DECISION PROBLEM

The problem of choosing the correct answer out of a set of answers is a "decision" problem. This problem can be described in terms of the set A = {a1,a2….an} of possible alternatives, the set B= { b1, b2….bm} of exhaustive , mutually exclusive relevant uncertain events , and the set of consequences Cij=c(ai,bj) which may possibly result. Each of these results may be quantified with a utility value uij=u(ai,bj),the possibility of uncertain events may be described with a probability distribution {p(b1),p(b2)…..p(bm)}(p(bj)>=0 ,$\sum$p(bj)= 1 over the relevant uncertain events. The optimal alternative is the one which maximizes the expected utility:

$$U\#(ai) = \sum_{j=1}^{m} u(ai,bj)*p(bj). \quad (2)$$

It is to be stressed that probability is the degree of belief.

Let {d1, d2,….dk} be the possible answers to multiple –choice questions which are all mutually exclusive and assumed to contain the true answer d*.Conventional practice says that the candidate has to mark one of them to be his choice – 1 mark is awarded for correct answer ,0 if the question is left blank and c>=0 marks to be subtracted if there is a wrong answer. From the candidate's perspective, if he has to score maximum marks, he has to maximize his answers in each of the questions. For each question the set of alternatives is {a0, a1, a2…..ak},where a0 denotes the choice of leaving the question blank, ai refers to marking the option di as the correct answer, and the score function can be written as:

$$U(a0,dj) = 0,u(aj,dj) = 1,u(aj,dl) = -c \text{ , if l is not equal to j.} \quad (3)$$

Thus if pj= Pr[d*=dj] denotes the probability which the candidate actually chooses to answer dj as the correct answer, then it follows from (1) that his expected utility for each possible alternative is given by:

$$U(a0)=0, u(aj) = (1+c)pj - c,\ j= 1,2,\ldots\ldots k \quad (4)$$

Therefore,

U(aj)>u(al) implies pj>pl

U(aj) >u(a0) implies pj>c/1+c

Thus, to maximize his score in the examination, in one particular question, the candidate has to determine his most likely answer, in other words one *mode* of his belief distribution{p1,p2…….pk} and to mark such a mode if and only if , its associated probability p*= maxjpj is larger or equal to c/1+c

It follows that the expected score in one question which acts optimally is

$$U(a)= \max(u(aj)) =\max\{(1+c)\ p^*-c,0\} \quad (5)$$

In particular, if the mark that the candidate might expect is one if he is *convinced* that the answer is correct, while the marks he might expect if he has no idea of what is the correct answer , so that his belief distribution (pj= 1/k) is uniform over all the answers is given by max{(1+c)/k-c,0} From the examiner's point of view, he has to set the paper in such a way that the penalty cost c of marking a wrong answer such that the score associated with random guessing would be zero, and this is achieved if , and only if , (1+c)/k –c=0, that is , if and only if c=1/k-1, in which case the optimal strategy for the candidate is to mark a mode if its associated probability p* is such that p*>=c/(1+c)= 1/k. Since this condition would be satisfied with any belief distribution over the possible k answers. Setting this would ensure that the candidate would always want to mark the correct answer, no matter how small the probability of marking a wrong answer. As an example, consider the PGCET exam in Karnataka where there are 4 options for a question paper and one of them has to be marked as a correct answer. Assuming that a negative mark of ½ is deducted for each wrong answer – c= ½ and 1/(1+c)=1/3,the above argument shows that the candidate's optimal strategy is to mark all answers such that p* >=1/3,that is to mark all the answers such that probability of more likely answer is at least 1/3.

Going further from this we can actually calculate a scoring rule which the examiner and the examinee can/has to adopt. The examiner 's intention is to set the questions in such a way that random guessing has to be avoided at all costs- this he can bring about by setting up the proper amount of negative marking. By combining the aspects of item analysis discussed earlier, the examiner can closely look at the options that could be set for each and every question so that the answers could be confusing/clear to the extent desired. And if we have even approximate a-priori knowledge about the student's capability to answer a certain question in a certain manner then we can combine the outputs of the regression analysis, item analysis and the above discussion into a suitable means of devising the question papers

VII. CONCLUSIONS

This paper discusses some techniques of measuring students' capability of answering in post graduate examinations when the questions are of multiple types. It also discusses techniques with which an examiner can set limits on random guessing by the students. Combining these two question papers that are appropriate can be generated.

VII. ACKNOWLEDGEMENTS

We acknowledge the support of the Vice Chancelor,Jain University and Principal - R.V.College of Engineering Autonomous-VTU – Bangalore given to write this research paper.

REFERENCES

[1] K.P.Soman,Shyam Diwakar and V.Ajay,Insight Into Data Mining –Theory and Practice(New Delhi,PHI,2012)
[2] https://www.msu.edu/dept/soweb/itanhand.html