Analyzing Factors of NUST MS Admission Policy/Process: A Step towards Uniform National Admission Policy for the Universities of Pakistan.

By

Dure Adan Ammara

00000317630

Supervisor

Dr. Zamir Hussain

Department of Computational Science and Engineering

Research Centre for Modelling and Simulation (RCMS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

August 2021

Analyzing Factors of NUST MS Admission Policy/Process: A Step towards Uniform National Admission Policy for the Universities of Pakistan.

By

Dure Adan Ammara

00000317630

Supervisor

Dr. Zamir Hussain

A thesis submitted in conformity with the requirements for

the degree of *Master of Science* in

Computational Science and Engineering

Department of Computational Science and Engineering

Research Center for Modelling and Simulation (RCMS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

August 2021

Declaration

I, *Dure Adan Ammara* declare that this thesis titled "Analyzing Factors of NUST MS Admission Policy/Process: A Step towards Uniform National Admission Policy for the Universities of Pakistan." and the work presented in it are my own and has been generated by me as a result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a Master of Science degree at NUST
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at NUST or any other institution, this has been clearly stated
3. Where I have consulted the published work of others, this is always clearly attributed
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
5. I have acknowledged all main sources of help
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Dure Adan Ammara,

00000317630

Copyright Notice

This thesis is dedicated to my beloved parents.

Acknowledgments

Firstly, I would thank Allah the Almighty, the Most Gracious, and the Most Merciful for all of His blessings.

I would like to express my deepest appreciation to my supervisor, Dr.Zamir Hussain, for his unparallel support throughout each phase of the procedure along with inspiring my interest in the domain of Data Analysis & Statistics.

This study would not have been completed without the kind assistance of the ICT directorate of NUST who helped us to access the archival admission data.

I would also like to extend my gratitude to my co-supervisor, Dr.Adnan Maqsood, for being so cooperative all the time.

Many thanks to my GEC members, Dr.Rizwan Riaz, and Dr.Ammar Mushtaq, for their acute feedback that pushed my work to more refinement.

In addition, I am extremely grateful to my family for always being there for me and believing in my abilities. Finally, special thanks to Zunera, for simulating discussions on my work as well as a delightful intrusion.

Table of Contents

1 Introduction 1

1.1 ACAD: 2

1.2 GAT: 2

1.2.1 GAT-GENERAL 3

1.2.2 GRE 4

1.3 INT: 5

1.4 Personal Motivation 5

1.5 Research Question 6

1.6 Objectives: 6

1.7 Scope of the study: 7

1.8 Structure of thesis 7

2 Literature Review 9

2.1 Rational of Admission Process 9

2.2 Integrant of a selection process 10

2.3 Assembly of admission processes in our territory 13

2.4 NUST PG Admission Criteria 15

3 Methodologies 17

3.1 Data Collection 17

3.2 Analysis 18

3.2.1 Data Pre-processing 18

3.2.2 Descriptive Analysis 18

3.2.3 Comparative Analysis 19

3.2.4 Model Development 19

3.2.5 Model Evaluation 27

4 Results and Discussion 30

4.1 Data Collection 30

4.2 Data Preprocessing 31

4.3 Descriptive Analysis 32

4.4 Comparative Analysis 40

4.4.1 Statistical testing 43

4.5 Model Development 46

4.5.1 Multiple Linear Regression (MLR) 46

4.5.2 Binary Logistic Regression (BLR) 72

5 Summary, Conclusions & Recommendations 85

5.1 Summary 85

5.1 Conclusions and Findings 86

5.2 Future recommendations 87

References 89

Appendices 98

List of Abbreviations

ACAD	Previous Academics
AD	Anderson-Darling test
ANOVA	Analysis of variance
BLR	Binary logistic regression
CI	Confidence Interval
COV	Coefficient of variation
df	Degree of freedom
GEC	Guidance and Examination Committee
GAT	Graduate Assessment Test
GRE	Graduate Record Examination
GMAT	Graduate Management Admission test
HEC	Higher Education Commission
HL	Hosmer-Lemeshow test
INT	Interview
ICT	Information and communication technology
LR	Likelihood ratio
LL	Log likelihood
MAE	Mean absolute error
MLR	Multiple linear regression
MAPE	Mean absolute percentage error

MMI	Multi mini-interviews
MLE	Maximum likelihood estimation
NTS	National Testing service
ETS	Educational testing service
SE	Standard error
SS	Sum of squares
SD	Standard deviation
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
OR	Odds ratio

List of Tables

Table 2.1 Summary of Admission Variables 12

Table 3.1 Available Data 18

Table 3.2 Error Measures 27

Table 3.3 Confusion Matrix 28

Table 3.4 Performance Evaluation Matrix 29

Table 4.1 Combined Data Overview 31

Table 4.2 Descriptive Statistics of Combined Data 33

Table 4.3 ACAD yearly descriptive statistics 37

Table 4.4 GAT yearly descriptive statistics 38

Table 4.5 INT yearly descriptive statistics 39

Table 4.6 Descriptive Statistics of Admitted students 42

Table 4.7 Descriptive Statistics of Not-admitted students 43

Table 4.8 Statistics of F-test 44

Table 4.9 Statistics of the two-sample t-test 45

Table 4.10 Statistics of independent sample t-test 46

Table 4.11 Correlation Matrix 47

Table 4.12 Combined results for MLR Model 48

Table 4.13 ACAD_dataset Correlation Matrix 51

Table 4.14 ACAD_dataset combined results for MLR analysis 53

Table 4.15 GAT_dataset Correlation Matrix 54

Table 4.16 GAT_dataset combined results for MLR analysis 56

Table 4.17 INT_dataset Correlation Matrix 57

Table 4.18 INT_dataset combined results for MLR analysis 59

Table 4.19 Assessment measures for validation 60

Table 4.20 ACAD(1)_dataset Correlation Matrix 61

Table 4.21 ACAD(1)_dataset combined results for MLR analysis 62

Table 4.22 GAT(1)_dataset Correlation Matrix 63

Table 4.23 GAT(1)_dataset combined results for MLR analysis ... 64

Table 4.24 INT(1)_dataset Correlation Matrix ... 65

Table 4.25 INT(1)_dataset combined results for MLR analysis ... 66

Table 4.26 Assessment measures for validation ... 67

Table 4.27 Normality Assessment of Error term ... 68

Table 4.28 ln_INT_dataset Correlation Matrix ... 69

Table 4.29 ln_INT_dataset combined results for MLR analysis ... 70

Table 4.30 Case Processing Summary ... 72

Table 4.31 Classification Table [a,b] ... 73

Table 4.32 Variables in the Equation ... 73

Table 4.33 Omnibus Tests of Model Coefficients ... 73

Table 4.34 Model Summary ... 74

Table 4.35 Classification Table [a] ... 74

Table 4.36 Variables in the Equation ... 75

Table 4.37 Confusion Matrix ... 75

Table 4.38 Omnibus Tests of Forward LR Model Coefficients ... 76

Table 4.39 Forward LR Model Summary 77

Table 4.40 Forward LR Classification Table [a] 77

Table 4.41 Forward LR Variables in the Equation 78

Table 4.42 Backward LR Variables in the Equation 79

Table 4.43 Models Evaluation (a) 83

Table 4.44 Models Evaluation (b) 83

Table 4.45 Models Evaluation (c) 84

List of Figures

Figure 2.1 Heterogeneous Variables around the globe. 13

Figure 3.1 Least square--Fitted Line.. 21

Figure 3.2 Sigmoid Function 23

Figure 4.1 Year-wise applicants count.. 34

Figure 4.2 Histogram of ACAD by Years 35

Figure 4.3 Histogram of GAT by Years 36

Figure 4.4 Histogram of INT by Years.. 36

Figure 4.5 Histogram of ACAD by Status.. 41

Figure 4.6 Histogram of GAT by Status... 41

Figure 4.7 Histogram of INT by Status 41

Figure 4.8 Scatter plot of Merit VS ACAD 49

Figure 4.9 Scatter plot of Merit VS GAT 49

Figure 4.10 Scatter plot of Merit VS INT... 50

Figure 4.11 Scatter plots of ACAD_dataset 52

Figure 4.12 Scatter plots of GAT_dataset.. 55

Figure 4.13 Scatter plots of INT_dataset... 58

Figure 4.14 Normality plot ln_INT_dataset.. 71

Figure 4.15 Predicted & Observed Values for the initial model (a)... 81

Figure 4.16 Predicted & Observed Values for the initial model (b)... 82

Abstract

Universities introduce the process of selection or screening for the applicants with the objective to select the “BEST” among available. Therefore, this process must be transparent, proficient, balanced, and complete. Studies highlighted a heterogeneous set of variables that individually or collectively can be observed as a process; though, it is almost impossible to define a universal criterion. This study investigated the effectiveness, balance, and completeness of different variables as an admission process, followed by a leading national university (National University of Sciences and Technology (NUST)) at the postgraduate level. The process consists of three variables: i) previous academic record of an applicant (ACAD) ii) marks obtained in graduate record examinations (general) or graduate assessment test (general) (GAT) (a test conducted by the national testing service of Pakistan for Higher Education Commission of Pakistan) and iii) interview (INT) conducted by the concerned school/institution/center of NUST. Moreover, the current weightages of ACAD, GAT, and INT in the merit calculation are 25%, 50%, and 25%, respectively. Since this is an empirical analysis, therefore, an archival student’s admission data, spanning over seven years has been used in this study. The information concerning these mentioned variables of 13094 applicants has been provided by the ICT directorate of NUST. Based on the literature review, the span and size of the sample used for analysis are sufficiently large to derive significant conclusions regarding the process. Descriptive and inferential analysis has been used to observe general trends of variables and for comparison of the performance of admitted against not admitted students. Moreover, multiple linear regression (MLR) & binary logistic regression (BLR) models have been used to develop predictive models for merit (being continuous variable) and status, i.e. admitted vs not admitted (being categorical variable). The results showed that admitted students significantly differ in performance relative to not admitted

students, primarily influenced by the INT scores with a marginal difference between GAT and ACAD scores. The results of the process of development of predictive models showed that the linear method is not suitable for this purpose due to the lack of a linear relationship between dependent and independent variables. Therefore, binary logistic regression considering the status of an applicant is a suitable alternative. Results showed that the three variables (ACAD, GAT, and INT) are not balanced (as compared to subjective weights assigned to them) and complete (lack in predictive ability). Therefore, there is a need for revision of weightages and inclusion of other relevant factors like popularity of a program, financial status of the applicant, place of residence, hostel facility, etc., These results provide useful insight for the choice of variables to be observed as a process of admission for postgraduate students, not only for NUST but for other national and international universities. Further research considering other factors and case studies for different universities can pave the way towards the uniform national admission process especially at the postgraduate level.

Chapter One

1 Introduction

"Effective admission procedures are a critical component of an institution's ability to fulfill its mission and goals, and on a greater scale, of the capacity of tertiary education to contribute to a nation's economic and social goals." (University Admission Worldwide, Helms 2008) [1]

Academia is the central pillar for a nation's substantial development. It is expected to produce skilled graduates for the relevant industry and provide efficient solutions to the everyday problems of society through innovative research. With research and innovation being the leading focus, postgraduate students are the fuel of any research lead university; therefore, the procedure of selection for the "best" intake among available requires dynamic and evident features. The strategy of various universities includes ways and means for the induction of best research students and subsequently to enhance their academic and professional strengths as they contribute majorly towards uplifting the ranking and reputation of a university [2]–[5].

A common problem in the developing countries is that they usually adopt systems or policies of the developed countries without correlating them with local constraints; hence are rendered useless with achieving no desired results. Therefore, there is a strong need to create an evidence-based, practical, and indigenous process encompassing a thorough approach: available information and analysis of the given sources and limitations. Keeping the goal above in mind, this study analyzes the admission process for the postgraduate students practiced by a leading national university of Pakistan (National University of Sciences & Technology (NUST)). The aim is to scrutinize this policy and provide NUST with a strategic proposition regarding its enrollment policy. The analysis focuses on checking the adequacy of different characteristics/factors/variables of the policy, developing a predictive model to evaluate the completeness of factors in the process, and establishing a comparative evaluation of the variables that influence the admission policy.

The admission policy of NUST for postgraduate students depends on three variables: i) previous academic record of a student (ACAD), ii) marks obtained in graduate record examinations (general) or graduate assessment test (general) (GAT) (a test conducted by national testing service of Pakistan for HEC) and iii) interview (INT) conducted by NUST. A brief detail of each variable is as follows:

1.1 ACAD:

The score of the variable ACAD, out of 25, has been allotted, for each applicant, based on the details provided in Appendix I. Through Appendix I, it has been observed that the minimum ACAD marks are 9 out of 25. Cumulative grade point average or percentage obtained in the terminal degree/transcript of the applicant has been used to obtain the score of ACAD using the last column (Marks allotted) illustrated in Appendix I. The weightage of this variable in the existing admission policy is 25 %.

1.2 GAT:

This variable represents test scores, i.e., Graduate Record Examinations (GRE) and Graduate Assessment Test general (GAT-GENERAL). Maximum marks for this variable are 100, and the minimum expected characters are 50 in the current followed admission policy for the postgraduate students. The weightage of this variable in the existing admission policy is 50%.

1.2.1 GAT-GENERAL

GAT-GENERAL is a test conducted by the National Testing Service (NTS) of Pakistan, which is mandatory for admissions in MS / MPhil and Ph.D. programs in almost all universities of Pakistan. This test consists of 100 multiple-choice questions. There are usually three sections of the paper namely:

I. **Verbal Reasoning**: This section consists of questions to analyze English comprehension and understanding of the language. The questions usually cover sentence completion, analogy, and critical reading skills.

II. **Quantitative Reasoning**: This section aims to assess the mathematical thinking process, the ability to put logic and reasoning together and solve problems in a quantitative setting. There is a balance of questions requiring basic knowledge of arithmetic, algebra, geometry, and statistics. All of these areas are usually covering the basic and essential high school level content.

III. **Analytical Reasoning**: This section is designed to test analytical skills. Mostly, this part has logical puzzle questions with given limitations. These questions have only one correct answer, with weights varying for various disciplines.

This test's total marks are 100, and a candidate with at least 50 marks is considered qualified. For further details, see http://www.nts.org.pk/Products/GATGEN/gat-g-introduction.php

1.2.2 GRE

GRE is an international test conducted by Educational Testing Service (ETS) by the USA, mandatory for admissions in masters or doctoral degree programs almost globally. There are broadly three sections of the paper namely:

I. **Verbal Reasoning**: This section measures the ability to draw conclusions and analyze important points by understanding the meanings of words. The questions usually cover text completion, reading comprehension and, sentence equivalence. [6]

II. **Quantitative Reasoning**: This section aims to assess the mathematical thinking process, the ability to put logic and reasoning together and solve problems in a quantitative setting. There is a balance of questions requiring basic knowledge of arithmetic, algebra, geometry, and statistics. Questions are usually of the comparison type, MCQs, and numeric entry. [6]

III. **Analytical Writing**: This section is designed to analyze creative writing skills and the ability of focused and effective writing supported by relevant examples. There are two tasks in the categories of “Analyze an Issue” and “Analyze an Argument”. [6]

The total marks of this test are 340. For further details, visit https://www.ets.org/gre/revised_general/about

1.3 INT:

To evaluate an applicant’s suitability, an interview is conducted in the concerned institution/school/center of NUST. The total marks of INT are 25. The procedure used to obtain the marks of INT for each candidate is based on a Performa including various attributes assessed by the interview committee (usually consisting of three to five members of the department/center/school). The detailed Performa is available in Appendix II. The weightage of this variable in the existing admission policy is 25 %.

Based on the details above, the merit of an applicant is thus calculated as:

$$\boldsymbol{Merit\ =\ ACAD\ score + (0.5) * GAT\ score + INT\ score} \quad \textbf{1-1}$$

These all are the variables used by NUST for calculating the merit of an applicant. The study will use the data to identify any existing patterns and clear any ambiguity regarding discrimination of admitted and not admitted students with the effectiveness of the stated variables. The results of the study would provide valuable guidelines to national/international universities seeking to revamp their admission procedure to attract the best lot of students at the postgraduate level.

1.4 Personal Motivation

We all have been gone through some sort of screening process. There are plenty of examples, e.g., School/College/University admission process or Job screening processes, and many more. However, do we ever wonder that how this process is examining us? Is it objective or just a subjective one? This is the real motivation behind this work to create awareness among all of us. The initial step here is to analyze the admission process at NUST. We are evaluating patterns, if any, of how students are being selected at the MS level and the difference between average performances of applicants admitted vs not admitted.

1.5 Research Question

The admission process for PG students at NUST aims to select ‘THE BEST’ among available. So far, there is no such scientific and empirical finding available to prove its adequacy. Thus, there is a need to answer a genuine question that: Is the admission process for PG students at NUST adequate in terms of its variables being assessed? Adequacy will be checked by answering this sub-question:

1. Are the three variables effective (i.e. fairly discriminating the admitted and not admitted students), balanced (i.e., compatible with their subjective weightages), and complete (in terms of their predictive ability of assigned class as admitted against not admitted student)?

1.6 Objectives:

Based on the stated questions above, the main objectives of the study are:

i. To analyze the general trends and tendencies of three stated variables, i.e., ACAD, GAT, and INT, used to calculate an applicant’s merit or to decide status.

ii. To investigate whether there exists a significant difference in the average performance of applicants concerning the stated variables for the calculation of merit or observing the status of applicants as admitted or not admitted. This will help to understand the effectiveness of these variables in the process.

iii. To evaluate the predictive ability and completeness of the variables being followed by developing predictive models using linear and binary logistic regression.

1.7 Scope of the study:

The proposed study will proceed with the following limitations:

i. The study results would be limited to the stated variables only (extracted through provided data), as there may exist different factors in the admission process of other universities, for instance, research proposal, allotted number of seats, the popularity of any specific program, etc.

ii. The findings of the analysis would be based on the secondary data already collected by the principal investigator of the study. The addition of further data and experiences of the concerned officials of leading public/private universities/institutes of the country and/or Higher Education Commission of Pakistan (as being a regulatory body for procedures/systems related to higher education) will depend upon the availability and appropriateness of the data/information and financial resources, etc.

iii. This study will only focus on the investigation of the statistical/empirical significance of the process.

1.8 Structure of thesis

The rest of this thesis is assembled as follows. In chapter 2 there is a literature review concerning the related studies worldwide and a link of empirical verification of this work. Chapter 3 discusses the applied methods, both conceptual and theoretical, used to collect, analyze, and modeling purposes. Chapter 4 provides results and their discussions to infer what is at the base of the process under focus and thus it was concluded in the 5th chapter.

Chapter Two

2 Literature Review

The major focus of this study is to empirically analyze different characteristics of the admission process for postgraduate students at NUST. With respect to global enrollment processes, we have a variety of heterogeneous variables to evaluate any applicant. There are studies around the globe illustrating the significance or insignificance of different variables considered as screening of the students. Summaries of few relevant international studies are provided in the later sections. However, concerning Pakistan, to the best of our knowledge, there is no publicly available scientific and empirical study analyzing the admission process of any university or degree awarding institute.

Before starting, we should know the formal definition of an admission process. It is defined as the screening process or student selection process used to filter desired applications from a pool of applicants when demand exceeds supply [7].

2.1 Rational of Admission Process

While talking about any admission process, we have the image of some sort of filtering process in our minds to get into the desired institution or organization. The rationale behind this selective admission is that there are more applicants than the available slots. This limited capacity is due to sparse resources and attracting the most prepared lot of applicants [8]. Moreover, the primary purpose of this process is to predict academic achievement to facilitate the applicant fully [1]. Thus, to handle the vast number of sundry applicants there is a need for an efficient admission process [9].

2.2 Integrant of a selection process

To comprehend the admission process used in higher education, there is a need to zoom into this process. The admission process comprises those factors that are essential to rank and select the desired lot of applicants [7]. Several qualities may come to mind while thinking of the chosen research student, e.g., a balanced combination of cognitive (GPA, aptitude test scores, etc.) and non-cognitive skills (Interviews, psychometric questionnaires, situational judgment tests, etc.) [10], [11]. However, the question mark is still active on the question of how to measure those skills effectively.

From the very beginning, researchers have indulged in selecting relevant factors that should be in the screening process to get " the best " lot of applicants. Universities and research institutes have the main aim to enroll those applicants who can perform well. The previous academic record is the first and easiest way to do that [12]. Initially, the studies of Ingram, Goldberg & Alliger, researchers of the 19^{th} and 20^{th} centuries, states that GRE/GMAT along with the cumulative grade point average (CGPA) are the most common tools to screen the highest-scoring individuals [13], [14]. Researches then headed towards finding the relation between these two factors. Robertson and Nielson found out that there is a significantly ($p\text{-value} < 0.05$) weak correlation between GRE and CGPA with a correlation coefficient = $r = 2.9$ [15]. On the contrary, the findings of Harvancik & Gordon states that the relation between these two factors is significantly moderate ($r = 0.48$, $p\text{-value} < 0.05$) [16]. The trend of heterogeneity persists with the other two findings, with one, by Milner, King, and McNeil, is showing a statistically weak correlation ($r = 0.238$) [17] and the other one by Thornel and McCoy's study, which suggests the significantly moderate/high correlation between these two variables with a coefficient of correlation = r = 0.43 [18].

In the 21st century, this research took a pace towards the possibility of significant supplementary factors as there are other considerable variables instead of just these two (GRE/GMAT and CGPA) in the domain of the admission process. When we search for " the best " candidate, we usually look for the qualities that must be in that applicant to justify the position fully. In terms of a postgraduate student, these qualities would be intelligence, critical thinking, self-motivation, interpersonal skills, problem solving and exploration skills, personal integrity, etc. [10]. The main question that arose and is still open is how to adequately measure all of these qualities in a candidate [10]. Thus, universities settle on merit-based policies to adopt a "fair" selection procedure. However, the definition of fairness is still vague [19].

Now, nearly all of the ongoing admission policies are working on the ground rule of merit. To check the validity of such admission measures in terms of admitted vs non-admitted applicants, Moruzi and Norman analyzed factors stated in Table 2.1. They concluded that admitted students have a significant difference in scores as compared to non-admitted students. However, they added that the mandatory test scores, Licencing Examination, were comparable [10]. After that, a study was specified on the admitted students only, based on the performance of factors in Table 2.1. Students were enrolled through different techniques, and then it turns out the students with high GPA outperformed all other students [12]. Here, again there is inconsistency in terms of the significance of the admission variables. Some studies focus on the importance of GPA, while others are discouraging mandatory tests based on low measuring capability. Moreover, the other two pieces of research about the predictive capability of the factors follow the same pattern as above. Thomas, Barbara, and Razack, and many other researchers have analyzed and concluded that Multi Mini Interviews (MMI) is the most significant factor among all others [20]–[22]. Whereas the finding of Abdulmohsin & Abdulaziz is in favor of the General Aptitude Test (GAT) in terms of its high predictive ability among previous grades and Scholastic Achievement Admission Test (SAAT) [23].

Table 2.1 Summary of Admission Variables

Author's (Year of publication)	Variables
Moruzi and Norman (2002)	i. GPA ii. Autobiographical submission iii. Simulated Tutorial iv. Personal Interview v. Licensing Examination
Nienke R. S, Anke M. T, J.C. Borleff and , Janke C.S (2014)	i. GPA ii. Course Credit (First Year) iii. Test Scores
Thomas, Barbara, and Razack (2017)	i. GPA ii. Personal Statements iii. MMI iv. Reference Letters
Abdulmohsin and Abdulaziz (2020)	i. SAAT ii. High school grade iii. GAT

To broadly categorize, some of the findings favor combining cognitive criteria and non-cognitive criteria to access the applicant's potential [11]. At the same time, some have their focal point on either of the requirements, as mentioned before. Some of the extreme findings are also at hand, by Nitza .D and Dan .S, that there is no actual structural relation between these admission policies with the student's potential to succeed [24]. The details of the aforementioned studies provide a variety of heterogeneous variables as shown in Figure 2.1. Therefore, the convergence to a universal process is not easy.

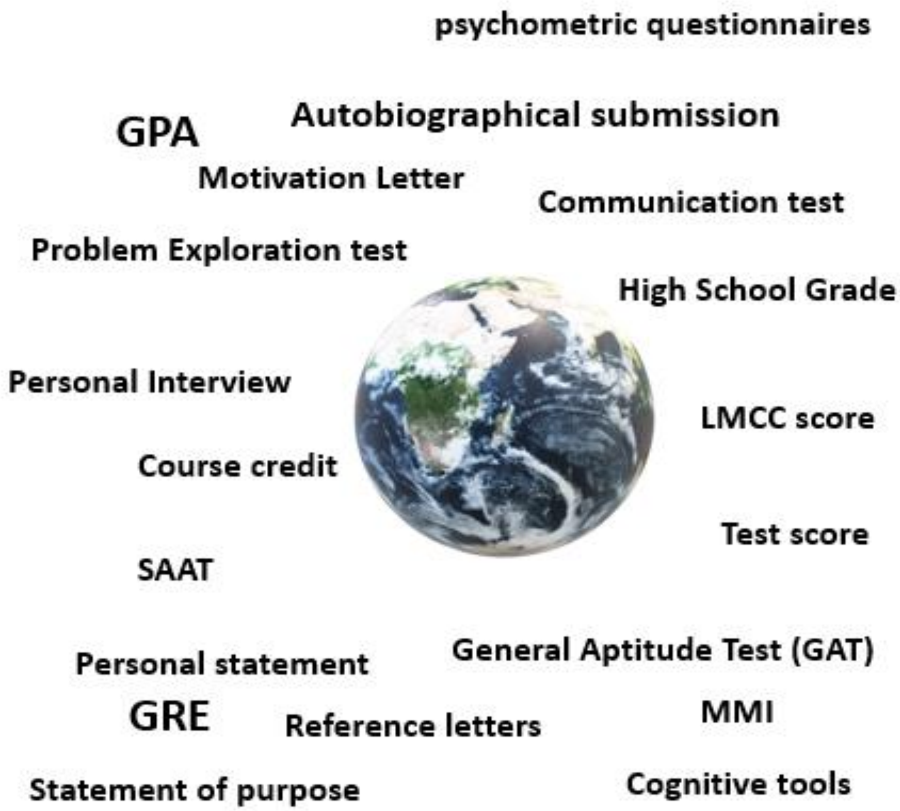


Figure 2.1 Heterogeneous Variables around the globe.

2.3 Assembly of admission processes in our territory

The above discourse about the selection of the most optimized blend of variables was on a global level. If we narrow it down to our national group, we are suffering from a shortage of research on this genre because here, the trends are of the same version despite the local constraints. Moreover, nearly all of those policies and decisions were data-driven because we all are surrounded by computational development in today's world. Following the best practices of the developed countries is a common problem in developing countries. Adopting best practices of the renowned institutes of the developed countries without correlating with the indigenous environment may not produce desirable results for developing countries, especially Pakistan. Therefore, there is a strong need to establish evidence-based practical indigenous solutions based on appropriate analyses using good information, keeping in view the local limitations.

Our primary concern here is the selection criteria used by universities and research institutes to enroll postgraduate students. To satisfy society's challenges, industry, country, etc., universities are heavily dependent on their output of research and innovation [25]–[27]. After the reforms in the higher education sector by the Government of Pakistan, Higher Education Commission (HEC) - established in 2002 - comes with the mandate to promote higher education, research, and development in the country to enhance the quality of teaching and research. The mission of HEC is to Facilitate Institutions of Higher Learning to serve as an engine of socio-economic development in Pakistan. In the current policy of HEC for ranking of various universities and higher education institutes of Pakistan, the research component is given 41% weightage [28]. This shows that research is a key component in the rankings of universities, and for this reason, the postgraduate students are of paramount significance. Therefore, the process of selection for "the best" intake among available requires dynamic and evident features. The strategy of various universities includes ways and means for the induction of adequate research students and subsequently to enhance the academic and professional strengths of these students as they contribute majorly towards uplifting the ranking and reputation of a university [2]–[5].

Just like all other countries, Pakistan is also using the screening process for enrolling the desired candidates. Many renowned institutions have their admission criteria; however, the source of their objectivity is not yet known because here, in Pakistan, there is not a single publicly available empirical study on this topic. Universities here are following the already established policies for enrolling students; however, the basis for their approach is not yet known. In a nutshell, there is no such connectivity-based and data-driven research available at the national level yet.

2.4 NUST PG Admission Criteria

NUST is one of the top universities and is a research lead institute in Pakistan. NUST is following the same pattern, aforementioned in the above section, of research in picking the admission variables for the postgraduate applicants. Three main variables have been measured in the current admission policy to filter out " the best " candidates. The three variables are ACAD. GAT, INT (details of all of these variables are available in the 1st chapter of Introduction). This study will be focusing on the NUST MS admission process to find out the objectivity of this criteria. The university somehow quantifies these variables to determine the merit of the applicants. A fundamental question concerning the procedures in practice is: Are we accurately and impartially discriminating between the admitted and not admitted students? The analysis of the study will be based on secondary data of the stated factors of more than ten thousand applicants intended for admission in various disciplines programs over more than five years. The study will focus on the general trends and tendencies of the stated factors, comparative analysis of the performance of admitted students against not admitted students, and provide helpful guidelines based on predictive models of these factors for the calculation of merit of an applicant. These results would be a step towards uniform admission policies for postgraduate students in the universities of Pakistan.

The span and size of the sample used for analysis are sufficiently large to derive conclusive shreds of evidence regarding the stated factors. The results will provide valuable guidelines to national universities seeking to adopt their academic regulations to attract the best students at the postgraduate level.

Chapter Three

3 Methodologies

This section deals with the step-by-step approach of processes and details of all the methods used for the analysis of this study. This work is indulged with the data of PG students of NUST to check the objectivity of the current admission policy. The aim is to gain an understanding of the trends and tendencies of the variables being used. The philosophy of this project is empirical and experimental based on the objective to check the behavior of attributes along with the development of the predictive model. Moreover, the approach is quantitative.

3.1 Data Collection

This study is based on the archival method, which involves describing data that existed before the survey. Data of all the applicants of various schools/centers/departments of NUST located at the H-12 sector of Islamabad (the capital city of Pakistan) for the period 2008 to 2014 has been used for the current study. Before 2008, the complete record of all the applicants was not available in electronic form. An applicant was defined as a person who had applied in any postgraduate program (MS, MPhil) at any department of NUST regardless of its acceptance at NUST. The complete information regarding the variables of interest of the applicants is available with the information and communication technologies (ICT) directorate of the NUST (Table 3.1). The department of ICT, NUST has merged this data by collecting it from all the departments and schools of NUST for universities' records. Data consists of 13094 values with 5458 admitted students and 7636 not-admitted students. The span and size of the sample used for analysis are sufficiently large to derive conclusive pieces of evidence regarding the stated factors. The results will provide valuable guidelines to national universities seeking to regulate their academic processes to attract the best students at the postgraduate level.

Table 3.1 Available Data

<table>
<tr><th>Rows</th><th colspan="6">Columns</th></tr>
<tr><td rowspan="3">13094</td><th>YEAR</th><th>ACAD</th><th>GAT</th><th>INT</th><th>MERIT</th><th>STATUS</th></tr>
<tr><td rowspan="2">Categorical Variable</td><td colspan="4">Continuous Variable</td><td>Binary Categorical Variable</td></tr>
<tr><td colspan="3">Dependent Variable</td><td colspan="2">Independent Variable</td></tr>
</table>

3.2 Analysis

These are the steps that have been followed throughout the analysis:

3.2.1 Data Pre-processing

It is beneficial to preprocess the data for getting excellent and authentic results before digging into the analysis [29]. Data were analyzed for any blank values and outliers with Minitab® -19.1 and Excel. Before analysis, the upper and lower bounds of all of the three variables were checked to make sure that all data values are in the desired range.

3.2.2 Descriptive Analysis

After performing the pre-requisite of data filtering, descriptive analysis of the sorted data was initiated in the Minitab® -19.1 to discover patterns in the data in terms of the variables. ACAD, GAT & INT were then analyzed to get their trends and tendencies. For in-depth analysis and comparison, the complete data set of the applicants has been divided into two groups (admitted students and not admitted students). Descriptive statistics of each of the three were analyzed, and the histograms measured the central tendency, dispersion, and shape. In addition to examining trends of combined data, data sets of admitted and not admitted were also analyzed separately to explore possible significant differences of each character with their descriptive measures.

3.2.3 Comparative Analysis

The whole of the complete dataset was then divided to investigating the discrimination power of the variables concerning the status of applicants. It is essential to test the significance of their differences (if any) concerning their variables i.e., ACAD (admitted) VS ACAD (not admitted) and vice versa. Two types of tests have been used to check the differences of admitted VS not-admitted student data. Hypothesis testing concerning the difference of variances with F-test and hypothesis testing concerning the difference of means with independent sample t-test and two-sample t-test has been used. Year-wise graphical representation of admitted VS not-admitted students was also used to understand how the rate has changed each year.

3.2.4 Model Development

Predictive analytics is a valuable tool for using archival data for developing and evaluating the model. In light of the third objective of this study, there is a need to probe the predictive power besides the completeness of the three variables.

If we look at the available variables, both independent and dependent, then there are two main combinations (Table 3.1). First, the combination of all the continuous independent variables with the continuous dependent variable (Merit). The second one is the combination of all the continuous independent variables with the binary categorical variable (Status). In each of the cases, the most prominent model is Multiple Linear Regression (MLR) [30] and Binary Logistic Regression (BLR) respectively [31].

Let's look at the model development phase for each of these models:

3.2.4.1 Merit score prediction of an applicant:

Regression analysis is a widely used statistical technique for predictive purposes and for measuring the relationship between variables. It tells the information regarding one variable while keeping all other variables fixed [32]. According to the merit formula (1-1). Merit is a continuous numeric variable. Therefore, with the three independent variables (ACAD, GAT, INT) that are also continuous, we are directed towards the MLR. MLR is a statistical technique used to model the relationship between the response variable (Y) and two or more explanatory variables (x_i) [32]. The general regression equation (3-1) is:

$$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon} \quad \mathbf{3\text{-}1}$$

Here, Y is the dependent variable with n independent variables: $x_1 + x_2 + \cdots + x_n$. Along with independent variables, Y depends on the parameters (constants and coefficients): $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ and error term ε [33].

Before applying MLR, there is always a need to check for its pre-requisites [34]. As the name implies, there should be a linear relationship between independent and dependent variables. Thus, on the combined dataset, Pearson correlation was used to test linearity between variables and checked the multicollinearity in Minitab® -19.1. After this test, MLR was applied on the same data and without constant terms to check the significance of constants and coefficients in the regression equation.

Method of least squares was used to get the best-fitted line in MLR [35]. This method is by default set in Minitab® while doing MLR analysis. The least-square method works on the principle to minimize the sum of squared residuals (error = difference between observed and fitted values). This straight line is the visual form of the least square method that passes through the given data points [36].

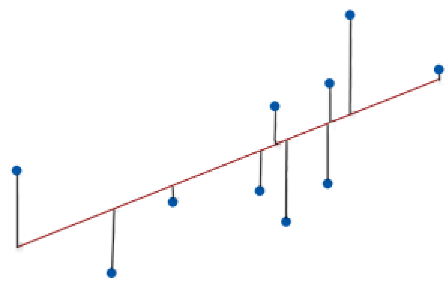


Figure 3.1 Least square--Fitted Line

3.2.4.2 Status (admitted / not-admitted) prediction of an applicant:

Among many machine learning techniques/algorithms, supervised learning is a tenet of the predictive model for this study. Predictive modeling, nowadays, is a versatile analytical technique. This work aims to look for patterns and test the significance of the variables by developing a model based on BLR, a well-known statistical technique, which falls under the supervised learning algorithm [37]. There is one categorical dependent variable (y = status of applicant = admitted or not admitted) in this study with three continuous numerical independent variables (x_1 = ACAD, x_2 = GAT, x_3 = INT). Here, the basic justification of using Binary Logistic Regression instead of Linear Regression is the dichotomous nature of our response variable and its ability to describe the relationship between predictors and binary response variables [38]. BLR, in the end, tells us about the probability of an applicant being admitted [37]. Mathematical representation of that probability model, Sigmoid function / logistic function, is:

$$\boldsymbol{E(y) = \frac{1}{1 + e^{[-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)]}}} \qquad \textbf{3-2}$$

Here E(y) represents the probability of being admitted, with parameters x_1 , x_2 , x_3 on which y is depending, and intercept b_0 & regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$. [37]. Generally, the logistic regression equation can also be represented as:

$$\boldsymbol{Ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} + \in \qquad \textbf{3-3}$$

Here, p is representing the probability of being admitted or not admitted [39].

The sigmoid function returns values between or equal to 0 and 1, just as the definition of probability. When the input variables are moving towards infinity, the output of this function tends towards/equal to 1. When the input variables move towards - infinity, this function’s work tends towards/equal to 0 [40].

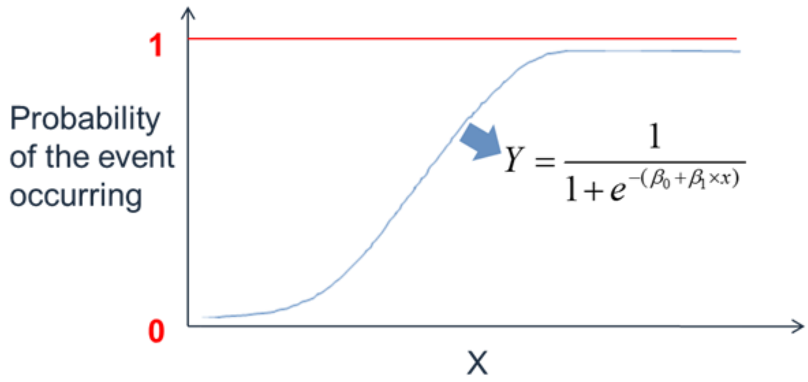


Figure 3.2 Sigmoid Function

For estimating the values of intercepts & coefficients, the maximum likelihood approach (MLE) is used because of the assumptions of BLR. Moreover, our sample size is sufficiently large, so MLE is the precise estimator in this case. The process of calculation is genuinely based on maximizing the likelihood function for the paraments. The likelihood function is useful here as it points out the chance or probability of the sample to be a function of possible values of parameters. Thus, the importance of parameters is determined in this way by maximizing the likelihood function.

Odds express the likelihood of the occurrence of an event relative to the likelihood of the non-occurrence of that event [41]. The classic definition of odd of an event is:

$$\boldsymbol{odd = \frac{p}{1-p}} \quad \textbf{3-4}$$

With p = probability of occurrence of an event and $1 - p =$ chance of non-occurrence of that event. Odds ratio (OR) is about comparing the odds of two occasions. Its interpretation depends on the baseline category. The category that is coded 0 naturally be treated as the baseline category. Its values are ranging from 0 to infinity. OR from 0 to below one are referring that the event is “less likely” to happen in contrast with the baseline group, OR equals to 1 shows that there is no difference and odds are the same for both of the groups; however, the OR value greater than one is showing that the event is “more likely” to happen in contrast with the baseline group [41].

Confidence intervals (CI), as per the definition, are the interval estimators of the actual value of the odds ratio. As our sample size is large enough, comparing it with sample sizes of related studies, it is concluded that the information we got from CI is accurate because CI uses the normal distribution of the data. In this study, we’ll be using the 95% CI, which means that there is 95% confidence about the occurrence of the odd ratio within that interval.

The Wald test has been used for testing the significance of three defining variables (ACAD, GAT, INT) in the model. This technique is suitable here due to the binary nature of the logistic model. The Wald test is behaving as the t-test just because the dataset is large enough [42]. This means that it deals with a null hypothesis that there is no difference between the expected and observed values. Thus, the higher Wald’s statistic value the more it is appropriate [43]. Every Wald statistic has a p-value associated with it. Less p-value (below the significance level) is the indication of significance [44]. The Wald test is functioning based on the chi-squared test, also known as the Wald Chi-Squared test [45].

Chi-square (χ^2) is a test that is used to check the significance of two categorical variables. The p-value is the main lead to point out the importance between them. Chi-square statistic is significant if the associated p-value is under the significance level (either 5% or 1%). Thus telling us about the significant association between those two variables [46]. Wald test works as an approximation of likelihood ratio (LR) test whereas the Wald test is more versatile among both of them [45]. LR test is used to pick the most suitable model for "good fit" among the two nested models. LR test uses likelihood function and log-likelihood function for this task however, log-likelihood function is the most common statistic used in popular statistical software for its interpretation. The null hypothesis of this test is about the smaller model (model with fewer explanatory variables), i.e., " smaller model is the best-fitted model " [47]. In SPSS this log-likelihood is used after multiplying it with -2 and thus -2 log-likelihood (-2LL) interpreted as deviance. -2LL is commonly used to see how well the BLR model fits with the data [48]. In other words, deviance tells us about the amount of unexplained variation in our binary logistic model. This value needs to be low for a model to predict the binary outcome [48].

Another better test for the goodness of fit, was found in the output of logistic regression in SPSS, named the Hosmer-Lemeshow (HL) test. This test explains how well the data fit the model by considering the match between observed and predicted values. HL test is also using deviance (-2LL) [49]. Its output table consists of a chi-squared value with the p-value. The null hypothesis of this test is that there is no difference between observed and expected values everywhere. Thus, small p-values are the sign of a " poor fit " model. However, a large p-value is not the assurance of the " good-fit " model. Even after all of the provided information, the HL test is not a strong candidate to opt for while checking the goodness of fit of a binary model because it lacks the power to address the overfitting issues and arbitrary choices of bins [50].

While performing BLR in SPSS, there are different types of methods in the stepwise regression approach, i.e., forward likelihood ratio, backward likelihood ratio, forward Wald, backward Wald, etc. These hierarchal methods are working on a specified principle. The forward likelihood ratio method and forward Wald method are working on the rule that initially they picked the most significant explanatory variable based on specific criteria and then added it to the null model, a model having only the constant value. It keeps on adding those explanatory variables till the maximum significance is achieved. On the contrary, all of the backward methods eliminate those explanatory variables from the full model (model having all of the variables) that are not contributing to the model's significance or have less significance [51].

Normalization in statistics is quite essential when dealing with a large diversity of variable scales. Many normalization techniques are operational. The simplest of all MIN-MAX scaler techniques. The formula for this normalization technique is:

$$\boldsymbol{X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}} \quad \textbf{3-5}$$

Here, this formula is scaled the data between 0 and 1, sometimes between -1 to 1. The maximum value is mounted at one and the minimum value at 0. However, this method is sensitive to outliers and is not able to handle any of them. To avoid this case, the frequently used standardization technique came to play. Z-Score is an efficient normalization method.

$$\boldsymbol{Z = \frac{X - \mu}{\sigma}} \qquad \textbf{3-6}$$

Here, μ and σ are the mean value and standard deviation of the variable respectively. This method standardizes the values around 0. Thus, the values below the mean = 0 will be projected towards the negative side, and the values more remarkable than the mean will then be normalized to the positive side [52].

3.2.5 Model Evaluation

After making models, intending to check the significance of the variables being used to enroll applicants in a postgraduate program at NUST, it is essential to evaluate the performance of the developed model. For the merit model's evaluation, root mean square error (RMSE), mean absolute percentage error (MAPE) and, mean absolute error (MAE) will be used [53], [54],[55].

Table 3.2 Error Measures

S. No.	**Measures**	**Mathematical Representations**
1	**RMSE**	$\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{n}}$
2	**MAPE**	$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert(Y_i - \widehat{Y_i})\rvert}{\lvert Y_i\rvert}\right) * 100$
3	**MAE**	$\frac{\sum_{i=1}^{n}\lvert(Y_i - \widehat{Y_i})\rvert}{n}$

The general way for the BLR model's evaluation, known as classification table or confusion matrix used for the assessment of any predictive model extracted from [37], [39]. It has four outcomes as the performance measures: True Positive (TP), True Negative (TN), False Positive (FP) and, False Negative (FN).The general way for the BLR model's evaluation, known as classification table or confusion matrix used for the assessment of any predictive model extracted from [37], [39]. It has four outcomes as the performance measures: True Positive (TP), True Negative (TN), False Positive (FP) and, False Negative (FN).

Table 3.3 Confusion Matrix

Test Outcome		**Predicted**	
		Positive (1)	**Negative (0)**
Observed	**Positive (1)**	TP	FP
	Negative (0)	FN	TN

There are three classic performance measures for any predictive analytical model: Accuracy, Specificity & Sensitivity. From these measures, the performance of the models is then being compared and evaluated. Definitions with mathematical representations of these three performance measures are described below:

Table 3.4 Performance Evaluation Matrix

S. No.	Metric	Description	Equation
1	**Accuracy**	It is used to predict the classification correctly.	$\frac{TP + TN}{TP + TN + FN + FP}$
2	**Sensitivity**	It is used to predict the proportion of YES's that are correctly identified.	$\frac{TP}{TP + FN}$
3	**Specificity**	It is used to predict the proportion of NO's that are correctly identified.	$\frac{TN}{TN + FP}$

Chapter Four

4 Results and Discussion

The primary aim of this research is to do an empirical evaluation of the employed policy. This work will then provide valuable guidelines for the universities of Pakistan to adopt a uniform admission policy for postgraduate students. This chapter deals with the results and their discussions of all the analyses done to check the objectivity of the current followed postgraduate NUST admission policy. This analysis had provided an in-depth understanding of ACAD, GAT, and INT through the applied descriptive and inferential methodologies. Secondly, it also handles the interpretations related to predictive models using linear and logistic regression to check the statistical significance and insignificance of the stated variables.

4.1 Data Collection

The study was carried out at the H-12 campus of the National University of Sciences and Technology (NUST), Islamabad, Pakistan. To observe the empirical nature of the currently followed admission policy at the postgraduate level, data for the MS applicants at all schools and departments were collected from the ICT directorate of NUST. The ICT department has collected all this data from individual schools and departments for the university's record. Data was of seven years from 2008 to 2014. Before 2008 the data of all the applicants were not available in the electronic format. Dataset consists of 13094 values with six primary columns, i.e., YEAR, ACAD, INT, GAT, MERIT, STATUS (Table 3.1).

4.2 Data Preprocessing

Before diving into the formal descriptive and comparative analysis, it is always productive to look at the general nature of the data. Data were preprocessed to look for any missing and out-of-bound values by using Minitab® -19.1 and Excel. Each column of 13094 values was examined. There were three out of the accepted range of ACAD values. The minimum possible marks for ACAD are 9 and thus, those three typographical errored rows were eliminated from the data. Moreover, there were two such errors in the INT variable too. A person was marked as admitted with a zero mark in INT. Therefore, all of these five errored rows were eliminated, and then we left out with a total of 13089 complete observations(n) to move to the next step. Now we have a sorted data (Table 4.1).

Table 4.1 Combined Data Overview

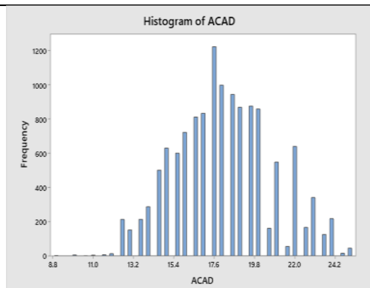
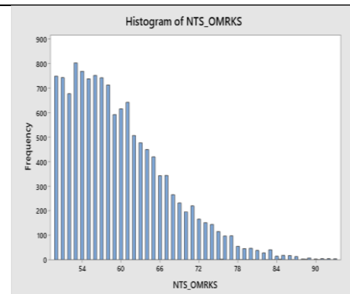
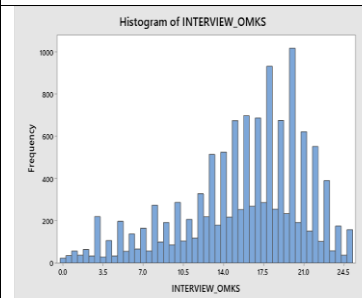
S. No.	Independent Variables	n	Minimum	Maximum
1	**INT**	13089	0	25
2	**GAT**	13089	50	93
3	**ACAD**	13089	9	25

The minimum score for GAT is 50 as mentioned in the first chapter. However, the maximum score for GAT is 93 instead of 100. After this step, there is a need to dig into the marrows of data to understand the trend.

4.3 Descriptive Analysis

One of this research objectives is to look for trends and tendencies of the variables being used in the PG admission process at NUST. Here, Table 4.2 depicts the summary of descriptive statistics of all three elements. This shows that ACAD mean is 18.04 with a standard deviation (SD) of 2.65, and the coefficient of variation (COV) is 14.69. Mean and median are nearly equal showing that the ACAD data is following a relatively normal distribution. GAT mean is 59.86 with an SD of 7.60, and COV is 12.70. The mean is not coinciding with the median here. Moreover, the value of skewness = 0.98, and the histogram shows that GAT tends towards the minimum value=50. INT mean is 15.65 with an SD of 5.41 and COV=34.57. The skewness value = -0.73, and the histogram shows that the INT score tends its maximum value=25. After comparing the COV of ACAD, GAT & INT it can be concluded that INT is the most inconsistent variable. Since higher the COV less consistent is the performance of the applicant. In other words, the performance of the applicants in INT is more deviated around its mean score. On the other hand, the GAT variable is the most consistent one with the least COV which implies that the performance of the applicants is more consistently around the mean score i.e., almost 60.

Table 4.2 Descriptive Statistics of Combined Data

S. No.	Measures	ACAD	GAT	INT
1	**Mean**	18.04	59.86	15.65
2	**SD**	2.65	7.60	5.41
3	**COV**	14.69	12.70	34.57
4	**Skewness**	0.19	0.98	-0.73
5	**Kurtosis**	-0.30	0.79	-0.01
6	**Q1**	16	54	12.89
7	**Median**	18	58	16.69
8	**Mode**	17.5	53	20
9	**Q3**	19.5	64	20.00
10	**Histogram**			

After the general overall view, the dataset was analyzed year-wise to understand the flow of applicants and the rate of change within each characteristic. From the year-wise analysis, it has been observed that the number of applicants(n) has been increased continuously (Figure 4.1). The reason may be that NUST was shifted to its main campus in 2008 and now acquiring land of about 707 acres with various program offerings. Furthermore, the number of admitted students has decreased during the last few years. It shows that with an expansion in offerings of different degree programs in various disciplines, the popularity and demand, so as the applicants, have been increased. Therefore, as a need of time, the admission procedure needs assessment analysis to address the shortcomings of the adopted approach, if any.

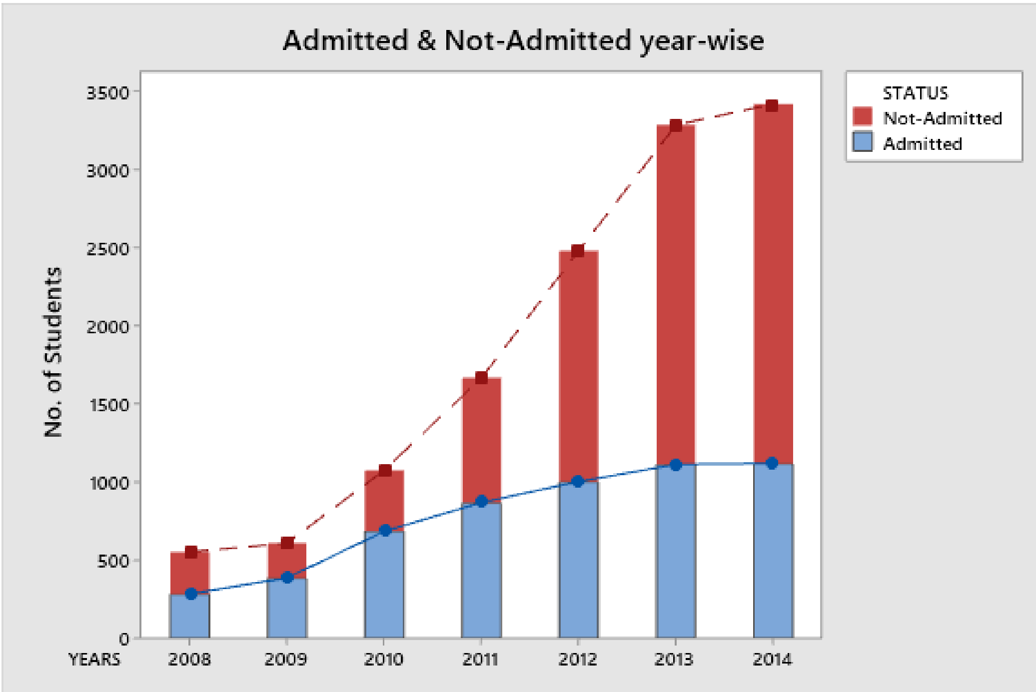


Figure 4.1 Year-wise applicants count

Through Table 4.3, It has been observed that the average ACAD score is roughly decreasing while the dispersion tends to decrease with the increase in the number of applicants each year. It may be because of an increase in the range(Max-Min) of ACAD data. Mostly ACAD variable tends to its mean value, just as the trend in overall data. Table 4.4, shows that the GAT scores each year are positively skewed because it tends towards its lower limit i.e., 50 scores. Table 4.5 shows that the average score of INT tends to decrease with the increase in applicants number each year. The score of ACAD in 2009 is most inconsistent whereas the score of ACAD in 2012 is the consistent one based on COV. More or less, GAT has the same level of consistency in each year; however, 2012 is considered the most inconsistent year, supporting the highest SD of 8.49, whereas 2009 is conceded as a consistent year because of a lower level of SD in it. Direct relation has been observed between the range of INT and its degree of consistency/inconsistency, i.e., if the range is high, so does the inconsistency in 2014 INT values. Except for 2008, the tendency of the INT variable is mainly towards the upper bound of it. It has been observed from Table 4.4 that 75% of applicants have less than 70 GAT scores in each year, and only a few students had scored between 70 to 95, and no one has a total score on the test.

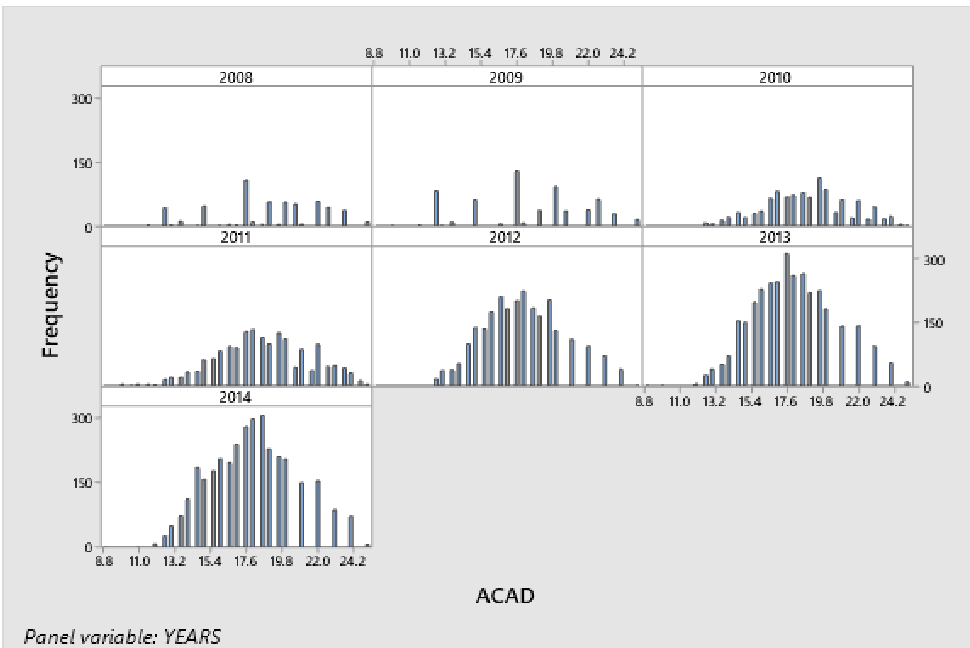


Figure 4.2 Histogram of ACAD by Years

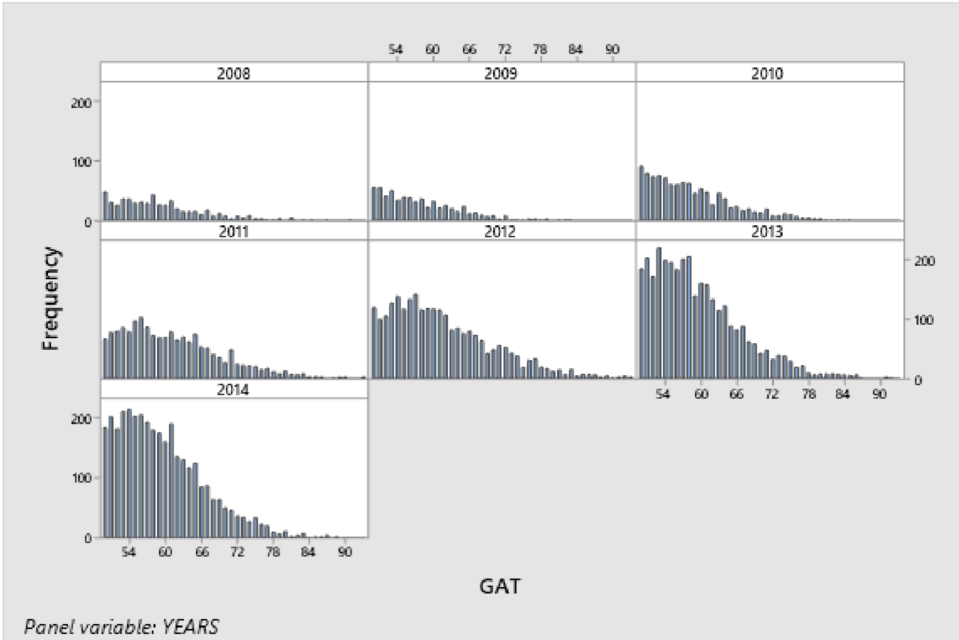


Figure 4.3 Histogram of GAT by Years

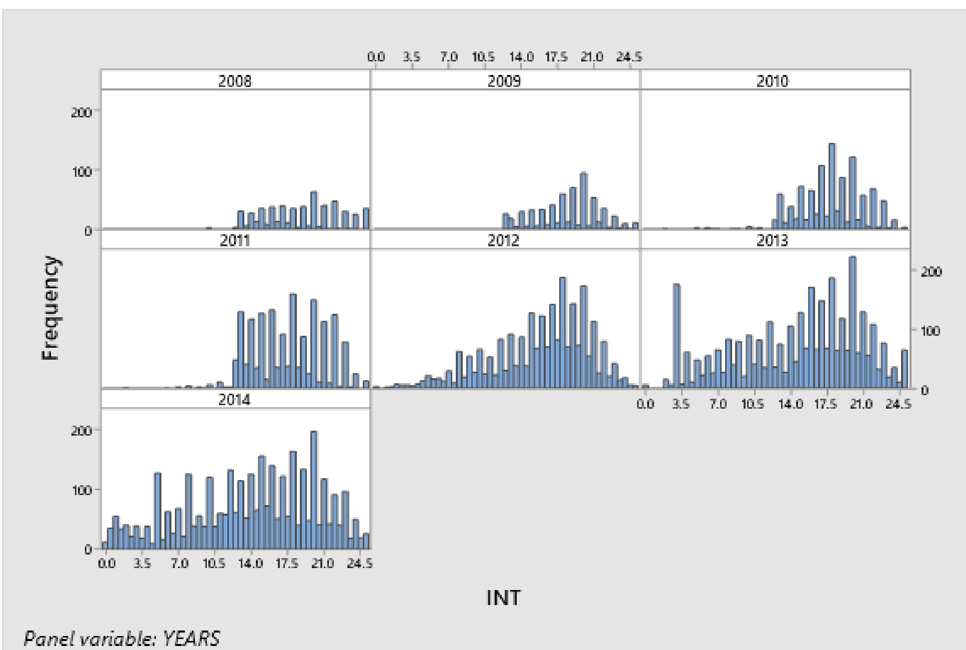


Figure 4.4 Histogram of INT by Years

Table 4.3 ACAD yearly descriptive statistics

S. No.	ACAD	2008	2009	2010	2011	2012	2013	2014
1	n	552	609	1076	1668	2481	3284	3419
2	Max.	25	25	25	25	25	25	25
3	Min.	11.5	10	12.5	10	12.5	9	9
4	Mean	18.92	18.39	18.79	18.59	17.77	17.79	17.77
5	SD	3.22	3.51	2.56	2.78	2.41	2.46	2.55
6	COV	17.00	19.07	13.64	14.96	13.53	13.84	14.37
7	Skewness	-0.41	-0.25	-0.01	-0.04	0.35	0.35	0.28
8	Kurtosis	-0.54	-0.85	-0.43	-0.39	-0.11	-0.02	-0.19
9	Q1	17.50	15.00	17.00	16.50	16.00	16.00	16.00
10	Median	19	19	19	18.50	17.50	17.50	18.00
11	Q3	22	20.5	20.5	20.5	19.5	19.5	19.5

Table 4.4 GAT yearly descriptive statistics

S. No.	GAT	2008	2009	2010	2011	2012	2013	2014
1	n	552	609	1076	1668	2481	3284	3419
2	Max.	91	83	85	93	93	92	90.5
3	Min.	50	50	50	50	50	50	50
4	Mean	59.03	57.69	58.38	61.17	61.44	59.46	59.44
5	SD	7.25	6.45	6.96	8.07	8.49	7.32	7.08
6	COV	12.28	11.18	11.92	13.19	13.82	12.31	11.91
7	Skewness	1.07	1.05	1	0.81	0.89	1.02	0.92
8	Kurtosis	1.22	1.04	0.59	0.31	0.44	0.94	0.68
9	Q1	53	53	53	55	55	54	54
10	Median	58	56	57	60	60	58	58
11	Q3	63	62	62.75	66	66	64	64

Table 4.5 INT yearly descriptive statistics

S. No.	INT	2008	2009	2010	2011	2012	2013	2014
1	n	552	609	1076	1668	2481	3284	3419
2	Max.	25	25	25	25	25	25	25
3	Min.	9.92	12.5	2	2	0.13	0	0
4	Mean	18.75	18.45	17.89	17.50	15.76	14.69	13.89
5	SD	3.52	3	3.09	3.39	4.72	6.01	6.23
6	COV	18.77	16.26	17.26	19.39	29.96	40.90	44.87
7	Skewness	0.04	-0.22	-0.48	-0.04	-0.68	-0.44	-0.36
8	Kurtosis	-0.95	-0.55	1.01	-0.52	-0.04	-0.78	-0.76
9	Q1	16	16.15	16	15	13	10	9.33
10	Median	19	19	18	17.67	16.75	16	14.93
11	Q3	21.92	20.29	20	20	19.25	19.66	19

4.4 Comparative Analysis

For in-depth analysis and comparison of discrimination ability of variables, the complete data set of the applicants has been divided into two groups concerning their status (admitted and not admitted). The admitted dataset has n=5455 values whereas the not admitted data set has n=7634 values. Descriptive measures of each of the three were analyzed along with the histograms to measure the central tendency, dispersion, and shape of each of the data sets. Table 4.6 shows the descriptive analysis of admitted students, whereas Table 4.7 shows the descriptive statistics of not admitted applicants for three (ACAD, GAT, INT) variables. Here, in Table 4.6 & Table 4.7, the behavior of variables is mostly the same as the combined statistics. The ACAD variable in admitted students and not admitted students follows the normal distribution because the mean is coinciding with the median having small skewness of 0.1 and 0.23, respectively. The GAT variable in both data sets shows its tendency towards the minimum value = 50 with a skewness of 0.85 and 1.07, respectively. The corresponding summary statistics of ACAD and GAT for admitted and not admitted students are quite close. In Table 4.6, INT variable having mean =17.70 & SD=3.93 is showing tendency towards its maximum value = 25 with skewness = -0.74 & kurtosis=0.88. However, in Table 4.7, INT has a high SD of 5.83 and is widely dispersed. The mean and median of INT for admitted students is higher than the mean and median of not admitted students, while the SD is low along with the COV. It seems that INT is creating a gap between scores of merits for admitted and not admitted students. In both these tables, the performance of applicants in INT shows the most inconsistent behavior and the performance of applicants in GAT is consistent. Moreover, the level of inconsistency is high in INT. The level of consistency is high in the GAT of admitted student’s data set compared to the not admitted student’s data set. From this analysis, we can conclude that, among the three variables, INT is the most powerful variable in terms of discriminating between the applicants based on their status. Below histograms are the graphical representation of the above description.

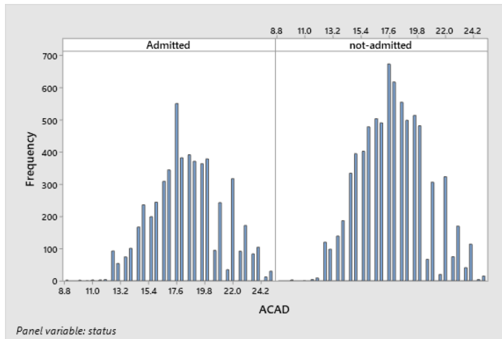


Figure 4.5 Histogram of ACAD by Status

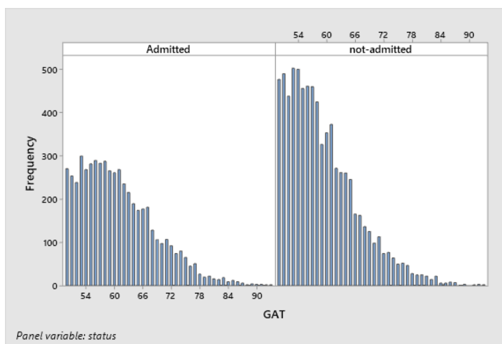


Figure 4.6 Histogram of GAT by Status

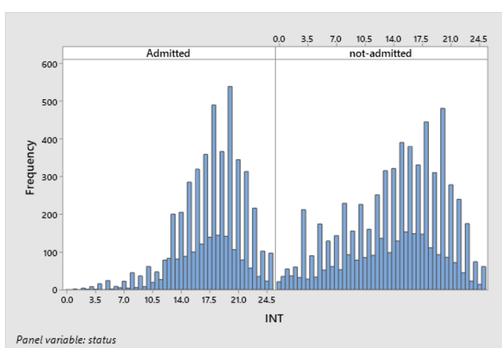


Figure 4.7 Histogram of INT by Status

Table 4.6 Descriptive Statistics of Admitted students

S. No.	Measures	ACAD	GAT	INT
1	**n**	5455	5455	5455
2	**Max.**	25	93	25
3	**Min.**	9	50	1
4	**Mean**	18.35	60.80	17.70
5	**SD**	2.72	7.87	3.93
6	**COV**	14.85	12.95	22.18
7	**Skewness**	0.10	0.85	-0.74
8	**Kurtosis**	-0.34	0.5	0.88
9	**Q1**	16.5	55	15.20
10	**Median**	18	59	18
11	**Q3**	20	66	20.25

Table 4.7 Descriptive Statistics of Not-admitted students

S. No.	Measures	ACAD	GAT	INT
1	**n**	7634	7634	7634
2	**Max.**	25	93	25
3	**Min.**	10	50	0
4	**Mean**	17.82	59.19	14.19
5	**SD**	2.57	7.33	5.83
6	**COV**	14.44	12.39	41.12
7	**Skewness**	0.23	1.07	-0.44
8	**Kurtosis**	-0.25	1.06	-0.62
9	**Q1**	16	54	10
10	**Median**	17.5	58	15
11	**Q3**	19.5	63	19

4.4.1 Statistical testing

We have analyzed the data under the lens of descriptive measures. Now, it is also essential to formally access the statistical differences of these variables within each group Two types of statistical tests have been used to check the differences of admitted VS not-admitted student data. Hypothesis testing regarding the difference of variances with F-test and hypothesis testing concerning the difference of means with two-sample t-test used. Considering the central limit theorem, the F-test has been used at a 5% level of significance to look at the equality of variance between admitted and not admitted data sets regarding each characteristic. Measures in Table 4.8 illustrating that for each of the variables: ACAD, GAT, and INT, the variance of applicants admitted and not admitted are not equal, at a 5 % level of significance.

Table 4.8 Statistics of F-test

σ_1 : standard deviation of Admitted data σ_2 : standard deviation of Not-Admitted data Ratio : σ_1 / σ_2		Null hypothesis H_o : $\sigma_1^2 / \sigma_2^2 = 1$ Alternative hypothesis H_1 : $\sigma_1^2 / \sigma_2^2 \neq 1$		
Variables	**F-value**	**df1**	**df2**	**p-value**
ACAD	1.12	5454	7633	0.000
GAT	1.15	5454	7633	0.000
INT	0.45	5454	7633	0.000

To formally test the hypothesis that the mean score of applicants admitted by the NUST is equal to the mean score of applicants not admitted for each variable, an independent sample t-test [56] has been used. Table 4.9 provides the results of the independent sample t-test for testing of equality of two population means. The results illustrate that for each of the variables: ACAD, GAT, and INT, the means of applicants admitted and not admitted are not equal, at a 5 % level of significance. INT is showing the highest difference among all variables being observed.

Table 4.9 Statistics of the two-sample t-test

μ_1: mean of Admitted data μ_2: mean of Not admitted data Difference : $\mu_1 - \mu_2$				Null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ Alternative hypothesis $H_1 : \mu_1 - \mu_2 \neq 0$	
Variables	**t-value**	**df**	**difference**	**95% of CI for difference**	**p-value**
ACAD	11.07	11336	0.5226	(0.4300,0.6151)	0.000
GAT	11.87	11226	1.610	(1.344.1.876)	0.000
INT	41.19	13042	3.5157	(3.3484,3.6829)	0.000

To check the closeness of the 95% confidence interval (CI) of each characteristic, in each data set of admitted & not admitted, a one-sample t-test is used. From Table 4.10 it can be observed that findings of this test support the conclusions of the two-sample t-test that the difference between 95% CI of INT is highest (upper bound difference = 3.542, lower bound difference = 3.489 as compared to the 95% CI of GAT and ACAD. ACAD with the lowest upper and lower bound difference.

Here, all of the three variables are statistically and significantly differentiating the performance of applicants based on their status. However, the results from the descriptive figures are not in line with this test. Thus, from this part of the analysis, we can sum up that ACAD and GAT are not playing a significant role in discriminating applicants in two subgroups (Admitted and Not admitted) even though they have a statistically significant difference. Moreover, INT is the most significant factor in this domain.

Table 4.10 Statistics of independent sample t-test

	Admitted			Not Admitted		
Variables	**t-value**	**p-value**	**95% CI**	**t-value**	**p-value**	**95% CI**
INT	0.04	0.967	(17.598,17.806)	-0.05	0.959	(14.056,14.317)
GAT	-0.01	0.992	(60.590,61.008)	-0.02	0.988	(59.024,59.353)
ACAD	-0.11	0.912	(18.274,18.418)	0.11	0.909	(17.766,17.881)

4.5 Model Development

One of the objectives of this research is to access the predictive ability and completeness of the three variables. Merit score and status of any applicant are the two main things to consider while talking about the admitted and not admitted applicants. Therefore, two different statistical techniques were used to address both of the aspects individually.

4.5.1 Multiple Linear Regression (MLR)

The definition of merit, mentioned in the previous chapter, states that merit is a continuous dependent variable so does the three independent variables (Table 3.1). Therefore, MLR is the first pole star that came to sight in this scenario. Before the development of the model, it is a prerequisite to check the degree of correlation between the variables. The correlation matrix is illustrated in Table 4.11.

Table 4.11 Correlation Matrix

	INT	GAT	ACAD	Merit
INT	1	0.189**	0.202**	0.815**
GAT	---	1	0.124**	0.630**
ACAD	---	---	1	0.515**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

This correlation matrix is providing two main particulars:

1. All of the independent variables have a low correlation with each other. Thus, there is no outrageous multicollinearity issue.

2. There is a strong positive and significant linear relationship between the dependent and independent variables. Which is suggesting that all of the independent variables are significant in terms of merit prediction. Moreover, INT-MERIT has the highest correlation coefficient value. Thus, merit is highly dependent on INT for prediction.

After this affirmation, applying MLR is the most suitable option here. The developed model is provided in the following equation.

$$\boldsymbol{Merit = 0.000000 + INT + 0.5 * GAT + ACAD} \qquad \textbf{4-1}$$

Table 4.12 Combined results for MLR Model

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
Constant		0	0	*	*
INT		1	0	*	*
GAT		0.5	0	*	*
ACAD		1	0	*	*
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	291422	874265	*	*
INT	1	357162	357162	*	*
GAT	1	180926	180926	*	*
ACAD	1	87452	87452	*	*

Table 4.12 is showing uncertain results of MLR analysis. The standard error is 0 and there is not a single value for the F-test and t-test. These all are pointing towards the homogeneity and repetition of observations within data. For further investigation, scatter plots were generated. As if we look at the figures below, there are many distinct merit values on a single ACAD value (Figure 4.8), and the same goes on for INT (Figure 4.9) & GAT (Figure 4.10). After these dubious values, we decided to give the average values a try.

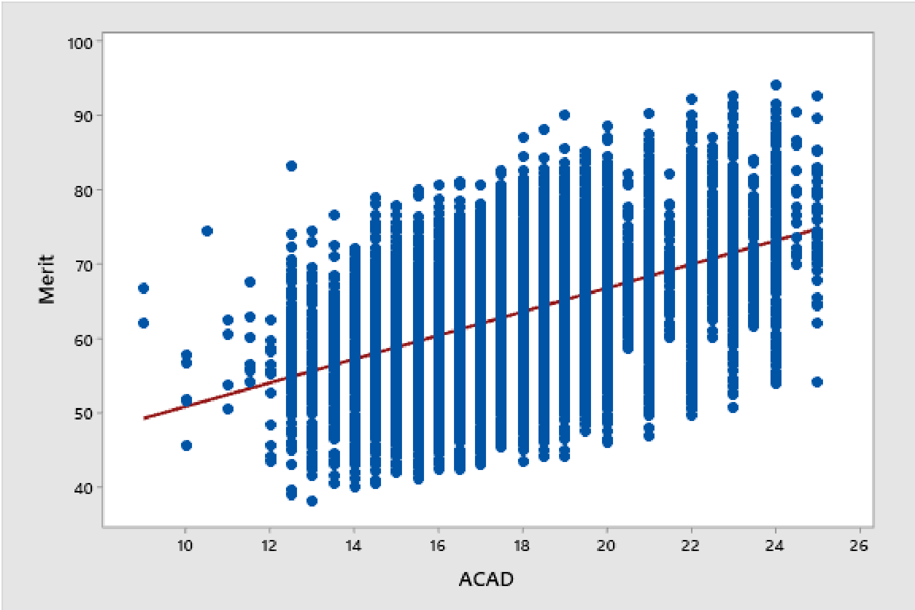


Figure 4.8 Scatter plot of Merit VS ACAD

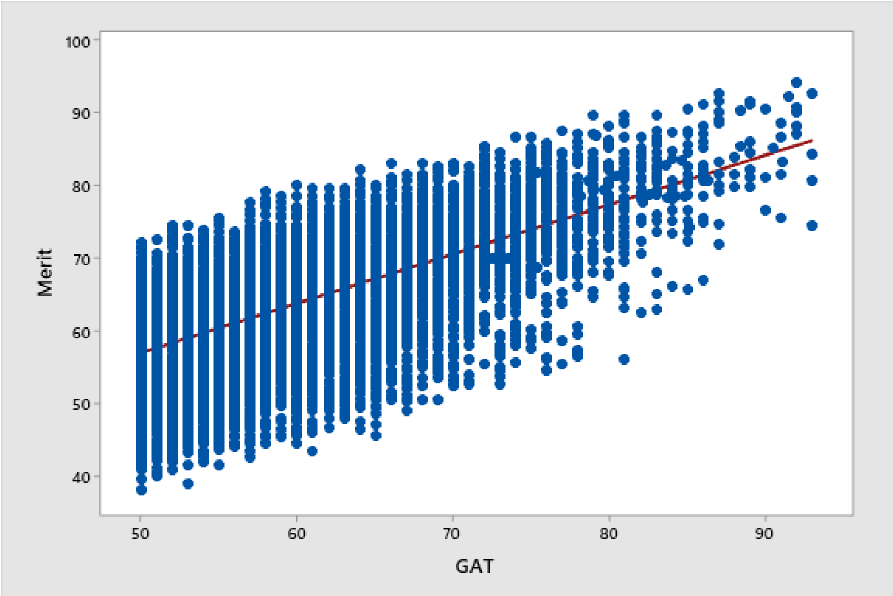


Figure 4.9 Scatter plot of Merit VS GAT

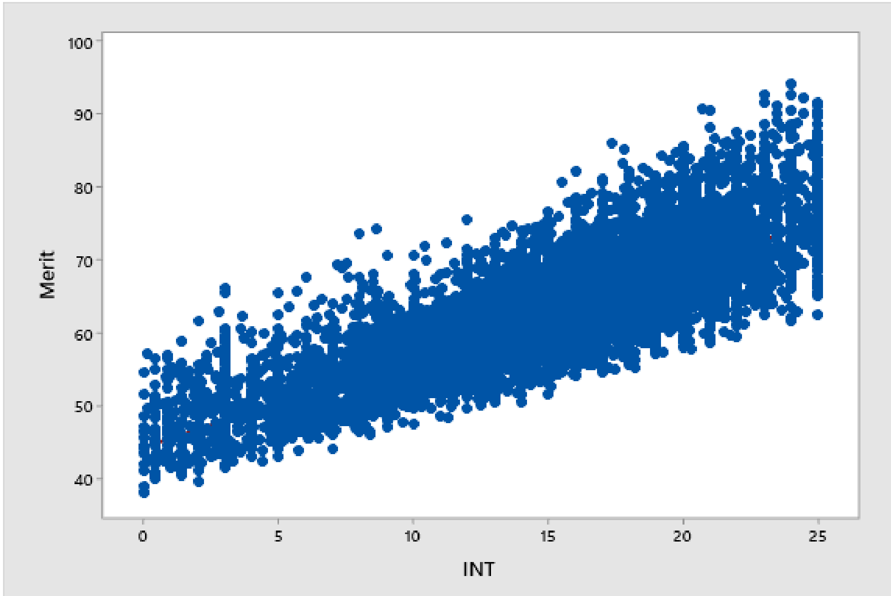


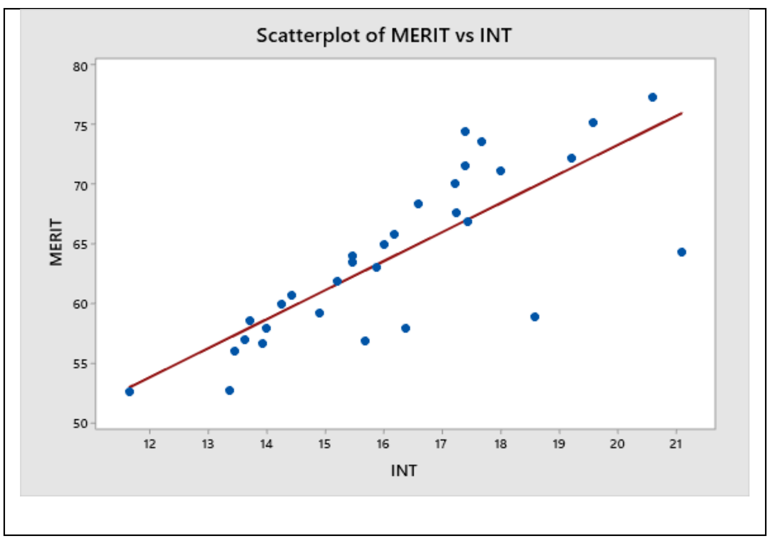
Figure 4.10 Scatter plot of Merit VS INT

After visualizing this scenario of dependent variable vs all independent variables, we took the averages of all of those merits, INT, GAT values with the same ACAD values using MS Excel and Minitab. After doing this, we left out 32 values in ACAD_dataset. We repeat the same procedure on GAT, and we left out 67 values in GAT_dataset. INT variable had so much variation in it thus we decided to have a threshold of 0-9 values because of the bulk of applicants' scores. We repeated the same average procedure on all values except those that fall under this range, i.e., we took the average of all merit, GAT, ACAD that have the same INT values within the scope of 10-25. Therefore, we left out a total of 2275 values in INT_dataset. Now, we have three more datasets. We decided to run MLR on each of these datasets to find out the most optimal model. Firstly, we check the correlations of all variables in the ACAD_dataset.

Table 4.13 ACAD_dataset Correlation Matrix

	INT	GAT	ACAD	Merit
INT	1	0.335**	0.450**	0.788**
GAT	---	1	-0.146	0.463**
ACAD	---	---	1	0.771**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

Table 4.13 is showing the magnitude of a linear relationship between dependent and independent variables in ACAD_dataset. It can be seen that there is a high positive and significant relation between merit and all of the independent variables except GAT. There is a significantly moderate/ low correlation between GAT & INT and ACAD & INT. However, there is an insignificantly low linear relation between GAT & ACAD. We also checked the scatter plot of these variables (Figure 4.11).



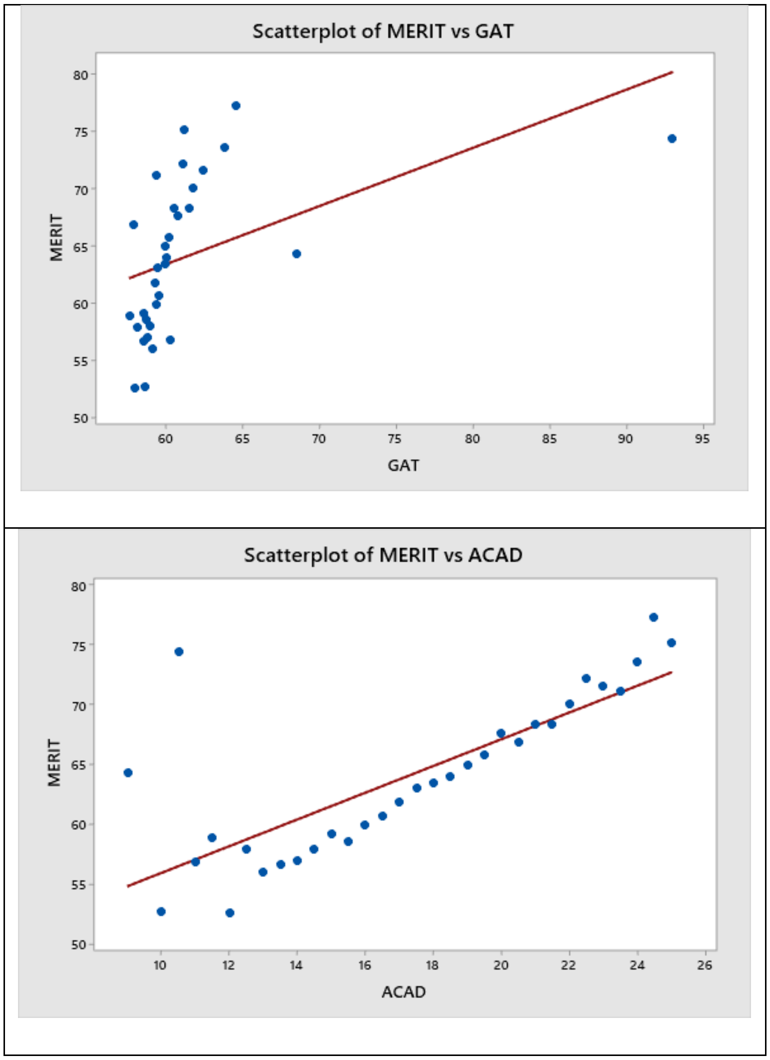


Figure 4.11 Scatter plots of ACAD_dataset

After this, we ran the MLR on this data. The constant term (0.00113) on that MLR model was insignificant (t-value = 0.33, p-value = 0.746) at 5% significance level. Then we re-ran it without considering the constant term. Thus, we got this regression equation:

$$\boldsymbol{Merit = (1.00012 * INT) + (0.499973 * GAT) + (0.99999 * ACAD)} \qquad \textbf{4-2}$$

Table 4.14 ACAD_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
INT		1.00012	0.00017	5955.6	0.000
GAT		0.499973	0.000036	13971.2	0.000
ACAD		0.99999	0.00007	13777.7	0.000
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0.0016714	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	44195.6	132587	1.58206E+10	0.000
INT	1	99.1	99	3.54691E+07	0.000
GAT	1	545.3	545	1.95193E+08	0.000
ACAD	1	530.3	530	1.89825E+08	0.000

From Table 4.14, it can be observed that there is consistency in this model. Initially, all of the coefficients are significant so does the overall adequacy of the model. We repeat the same procedure on GAT_dataset and INT_dataset.

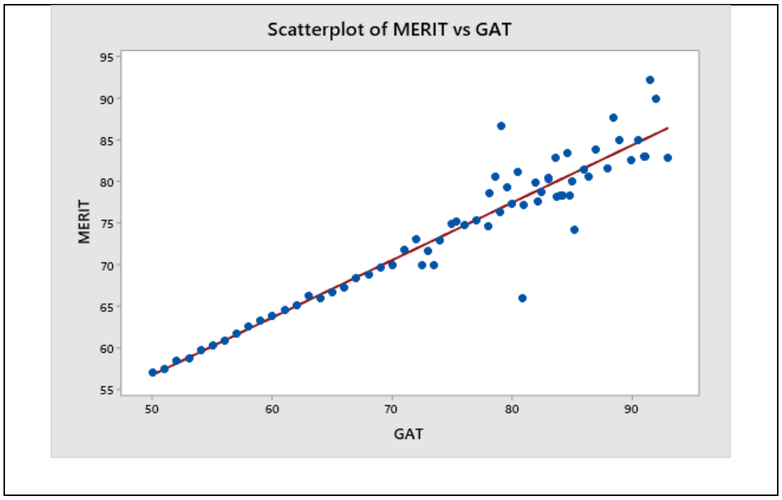
SPSS correlation matrix (Table 4.15) for GAT_dataset shows a high positive and significant linear relation between Merit & INT, Merit & GAT, but there is a substantial relation between

Merit & ACAD. Also, there is not any strong linear relationship between independent variables except GAT & INT. However, the correlation coefficient is significant between them.

Table 4.15 GAT_dataset Correlation Matrix

	INT	**GAT**	**ACAD**	**Merit**
INT	1	0.610**	0.131	0.784**
GAT	---	1	0.286*	0.949**
ACAD	---	---	1	0.431**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) * Correlation is significant at 0.05 level (2-tailed) --- Repetitive correlation coefficient values				

We confirmed this finding with the help of a scatter plot between all these variables in this dataset (Figure 4.12).



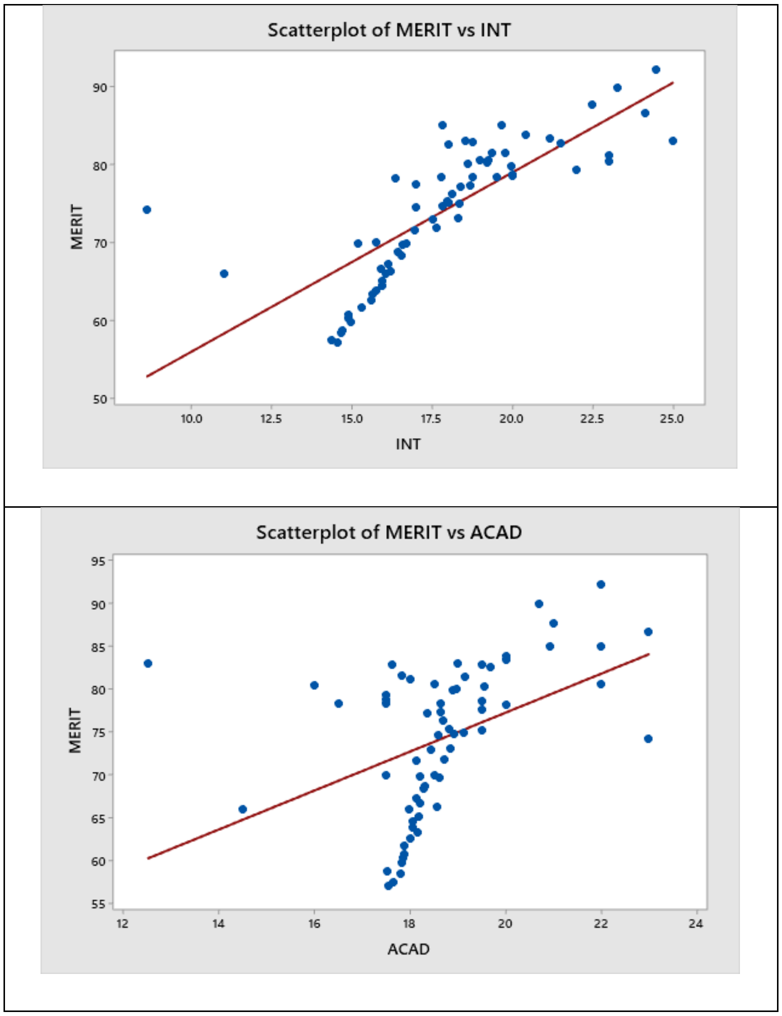


Figure 4.12 Scatter plots of GAT_dataset

After this, we ran the MLR on this data. The constant term (0.0046) on that MLR model was insignificant (t-value = 0.08, p-value = 0.933) at 5% significance level. Then we re-ran it without considering the constant term. Thus, we got this regression equation:

$$\boldsymbol{Merit} = (\mathbf{0.999983} * \boldsymbol{INT}) + (\mathbf{0.499995} * \boldsymbol{GAT}) + (\mathbf{1.00005} * \boldsymbol{ACAD}) \qquad \textbf{4-3}$$

Table 4.16 GAT_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
INT		0.999983	0.000180	5562.87	0.000
GAT		0.499995	0.000048	10433.05	0.000
ACAD		1.00005	0.00016	6342.40	0.000
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0.0035849	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	124784	374353	9.70986E+09	0.000
INT	1	398	398	30945505.88	0.000
GAT	1	1399	1399	1.08849E+08	0.000
ACAD	1	517	517	40226023.92	0.000

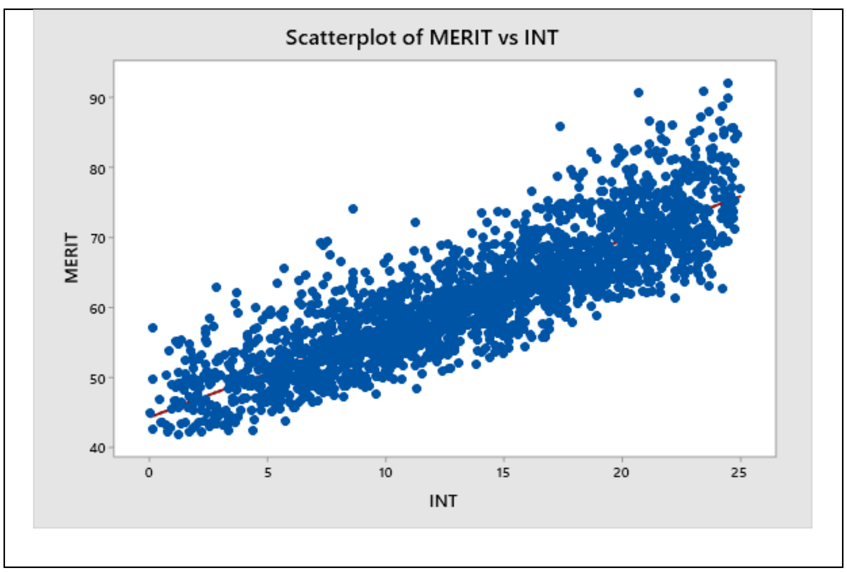
From Table 4.16, it can be observed that there is consistency in this model. Initially, all of the coefficients are significant so does the overall adequacy of the model.

SPSS correlation matrix (Table 4.17) for INT_dataset shows a high positive and significant linear relation between merit and all independent variables. Also, there is no outrageous multicollinearity among the independent variables.

Table 4.17 INT_dataset Correlation Matrix

	INT	GAT	ACAD	Merit
INT	1	0.249**	0.259**	0.861**
GAT	---	1	0.133**	0.630**
ACAD	---	---	1	0.517**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

We confirmed this finding with the help of a scatter plot between all these variables in this dataset (Figure 4.13).



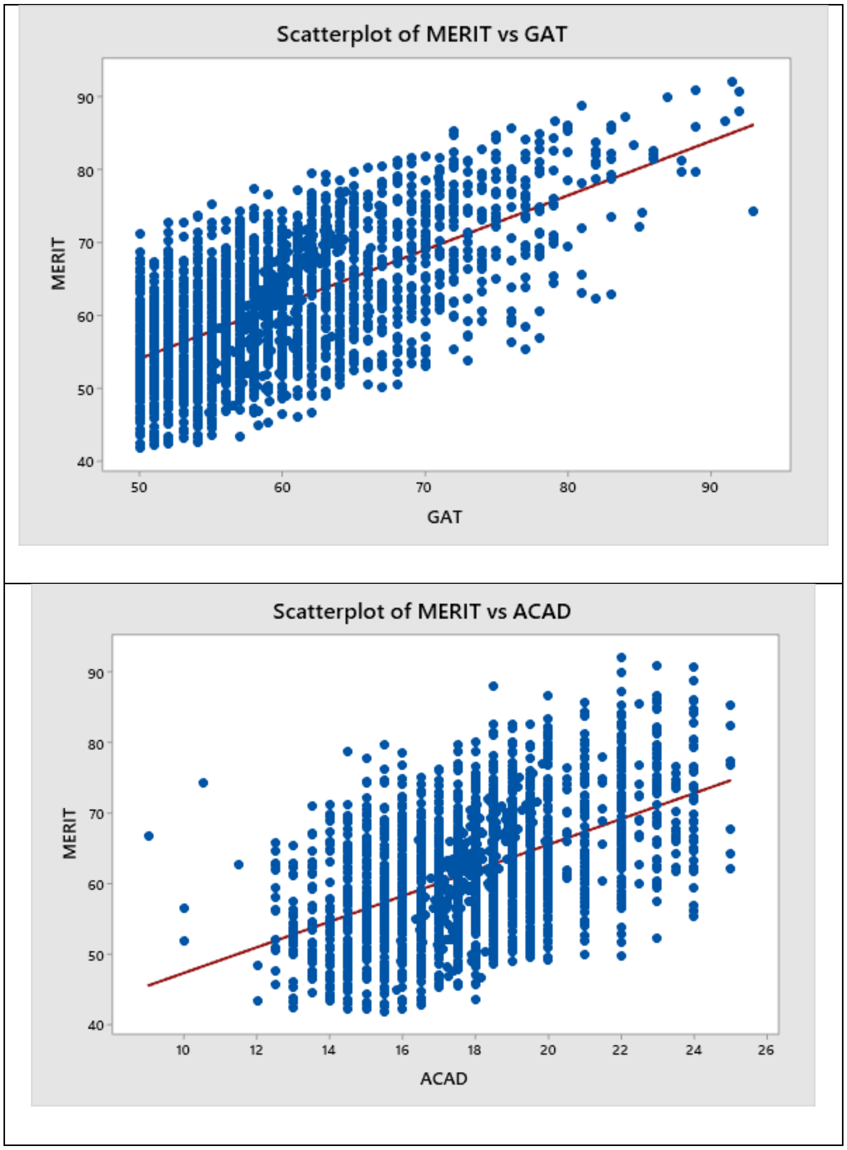


Figure 4.13 Scatter plots of INT_dataset

After this, we ran the MLR on this data. The constant term (0.0001790) on that MLR model was insignificant (t-value = 0.78, p-value = 0.433) at 5% significance level. Then we re-ran it without considering the constant term. Thus, we got this regression equation:

$$Merit = (\mathbf{1} * INT) + (\mathbf{0.499998} * GAT) + (\mathbf{1} * ACAD) \qquad \textbf{4-4}$$

Table 4.18 INT_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
INT		1	0	255995	0.000
GAT		0.499998	0.000002	229551	0.000
ACAD		1	0.000010	136470	0.000
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0.0010795	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	2945348	8836044	2.52760E+12	0.000
INT	1	76365	76365	6.55337E+10	0.000
GAT	1	61403	61403	5.26938E+10	0.000
ACAD	1	21702	21702	1.86241E+10	0.000

From Table 4.18, it can be observed that there is consistency in this model. Initially, all of the coefficients are significant so does the overall adequacy of the model. However, all of the MLR results for equations (4-1,4-2,4-3) were superficial in terms of model fitting i.e., R-sq = 100%. Therefore, to see how robust the model is, we took three forecasting measures (RMSE, MAPE, MAE). We found the magnitudes of these measures, of ACAD_dataset MLR model (4-2), GAT_dataset MLR model (4-3) and, INT_dataset MLR model (4-4), according to the formulas that have been mentioned in the previous chapter. From Table 4.19, we can conclude that the

INT_dataset MLR model has the lowest RMSE, MAPE and, MAE values thus showing the most optimized model among the other two.

Table 4.19 Assessment measures for validation

S. No.	Datasets	RMSE	MAPE	MAE
1	**ACAD_dataset**	0.0015915	0.0013596%	0.0008820
2	**GAT_dataset**	0.0035045	0.0015351%	0.0011596
3	**INT_dataset**	0.0010828	0.0005118%	0.0003174

As a possible solution for these superficial results, all of the influential observations were eliminated. At each MLR model of three datasets, influential values are picked up by the software. Those influential values were based on high standardized residual (R) and Hi-Leverage points (X) [57]. We eliminated all of those outstanding values for each of the datasets to see if the model improves. For ACAD_dataset there were 6 out of 32 values were detected. After eliminating those 6 (18.75%) values from this data, we repeated the same procedure as before.

Table 4.20 ACAD(1)_dataset Correlation Matrix

	INT	GAT	ACAD	Merit
INT	1	0.621**	0.872**	0.926**
GAT	---	1	0.765**	0.792**
ACAD	---	---	1	0.988**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

By comparing with the initial ACAD_dataset, it has been observed that the correlation coefficient between merit and all of the independent variables is high after eliminating those outliers or high leverage values (Table 4.20). The extracted MLR equation with coefficient (0.00223) was insignificant with (t-value = 0.37, p-value = 0.716). Thus, we ran the MLR without considering the constant term.

$$\boldsymbol{Merit} = (\mathbf{1.00005} * \boldsymbol{INT}) + (\mathbf{0.499994} * \boldsymbol{GAT}) + (\mathbf{0.999983} * \boldsymbol{ACAD}) \qquad \textbf{4-5}$$

Table 4.21 ACAD(1)_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
INT		1.00005	0.000100	9554.7	0.000
GAT		0.499994	0.000018	27750.8	0.000
ACAD		0.999983	0.000046	21766.4	0.000
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0.0004767	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	34863.7	104591	1.53448E+11	0.000
INT	1	20.7	21	9.12915E+07	0.000
GAT	1	175.0	175	7.70109E+08	0.000
ACAD	1	107.6	108	4.73778E+08	0.000

Table 4.21 and Table 4.14, show that there is no difference in terms of the individual and overall significance of both models. However, the value for S, Adj MS, and Adj SS were reduced after eliminating values.

In the GAT_dataset, there were five outliers or high leverage values detected out of 67 values. We eliminated those 5 (7.46%) values and repeated the same procedure. From Table 4.15 and 4.22, It has been observed that the correlation coefficient increased in its values between merit and independent variables after the elimination of those values.

Table 4.22 GAT(1)_dataset Correlation Matrix

	INT	**GAT**	**ACAD**	**Merit**
INT	1	0.777**	0.476**	0.882**
GAT	---	1	0.497**	0.971**
ACAD	---	---	1	0.620**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

We ran the MLR on this data and gain the constant term (0.00120) was insignificant (t-value = 0.49, p-value = 0.629). The detail of the MLR model(4-6), without the constant term, on the GAT(1)_dataset is:

$$\boldsymbol{Merit = (1.00004 * INT) + (0.500004 * GAT) + (0.999948 * ACAD)} \qquad \textbf{4-6}$$

Table 4.23 GAT(1)_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
INT		1.00004	0.00011	9413.04	0.000
GAT		0.500004	0.000023	21844.15	0.000
ACAD		0.999948	0.000075	13249.42	0.000
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0.0013261	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	115054	345161	6.54285E+10	0.000
INT	1	156	156	88605342.80	0.000
GAT	1	839	839	4.77167E+08	0.000
ACAD	1	309	309	1.75547E+08	0.000

Table 4.23 and Table 4.16, show that there is no difference in terms of the individual and overall significance of both models. However, the value for S, Adj MS, and Adj SS were reduced after eliminating values.

In the INT_dataset, 117 outliers or high leverage values were detected out of 2275 values. We eliminated those 117 (5.14%) values and repeated the same procedure. From Table 4.17 and 4.24, It has been observed that the correlation coefficient increased in its values between merit and independent variables after the elimination of those values.

Table 4.24 INT(1)_dataset Correlation Matrix

	INT	**GAT**	**ACAD**	**Merit**
INT	1	0.254**	0.279**	0.862**
GAT	---	1	0.172**	0.635**
ACAD	---	---	1	0.543**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

We ran the MLR on this data and found the results as the initial model. The standard error appeared to be zero thus, there are not any t-values, F-value, and p-values. The details of this MLR model(4-7), on the INT(1)_dataset is:

$$\boldsymbol{Merit} = (\mathbf{1} * \boldsymbol{INT}) + (\mathbf{0.5} * \boldsymbol{GAT}) + (\mathbf{1} * \boldsymbol{ACAD}) \qquad \textbf{4-7}$$

Table 4.25 INT(1)_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
INT		1	0	*	*
GAT		0.5	0	*	*
ACAD		1	0	*	*
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0	100%	100%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	2784352	8353057	*	*
INT	1	71208	71208	*	*
GAT	1	54664	54664	*	*
ACAD	1	19238	19238	*	*

Table 4.25 and Table 4.18, show that there is a difference in the MLR results. After eliminating those unusual values, the model is showing zero standard error.

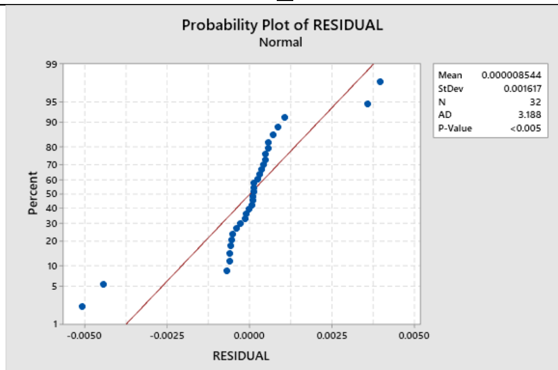
Again, all of these results, from refined models (4-5,4-6,4-7), are superficial. As mentioned above the assessment of these models is very essential thus, we took the same three measures and found the error magnitude of the latest three models (4-5,4-6,4-7). From Table 4.26, we can conclude that apart of INT(1)_dataset, ACAD(1)_dataset has the lowest RMSE, MAPE and, MAE values thus showing the most optimized model among the others. After comparing it with Table 4.19, It is evident that the error terms have fewer values after the elimination of outliers or high leverage values from all of the datasets.

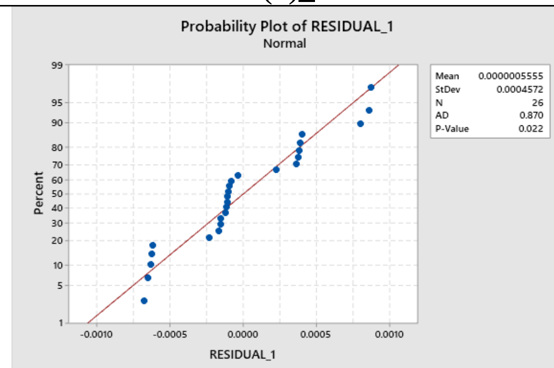
Table 4.26 Assessment measures for validation

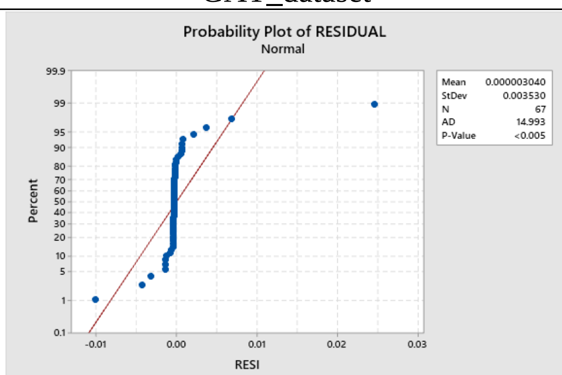
S. No.	Datasets	RMSE	MAPE	MAE
1	**ACAD(1)_dataset**	0.0004486	0.0005567%	0.0003569
2	**GAT(1)_dataset**	0.0012951	0.0007454%	0.0005645
3	**INT(1)_dataset**	0	0%	0

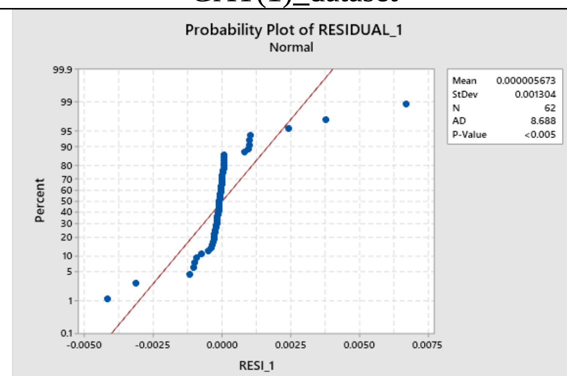
After the formulation of all of these models under MLR analysis, some assumptions need to be considered. Suppose, if we talk about the normality of the residual (error) term. Then, from Table 2.27, It can be seen that only the ACAD(1)_dataset has the normal error term among all other models.

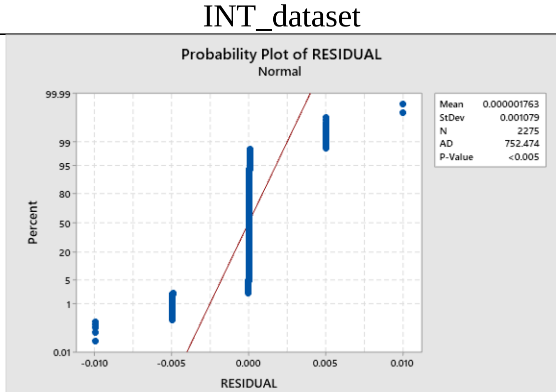
Table 4.27 Normality Assessment of Error term

<table>
<tr><td>ACAD_dataset</td><td>ACAD(1)_dataset</td></tr>
<tr><td>

p-value < 0.005</td><td>

p-value = 0.022</td></tr>
<tr><td>GAT_dataset</td><td>GAT(1)_dataset</td></tr>
<tr><td>

p-value < 0.005</td><td>

p-value < 0.005</td></tr>
<tr><td>INT_dataset</td><td>INT(1)_dataset</td></tr>
<tr><td>

p-value < 0.005</td><td>S.E = 0</td></tr>
</table>

The normality of the data is of the utmost importance in any analysis. Thus, we decided to apply the natural log transformation on the INT_dataset, which showed the most optimized results in the initial run. The goal for applying this transformation is only to check whether the data turned towards normality or not. Therefore, we used the Minitab for this task and applied it to all of the four columns (INT, GAT, ACAD, MERIT). SPSS correlation Table 4.28 is showing that all variables are behaving efficiently.

Table 4.28 ln_INT_dataset Correlation Matrix

	INT	**GAT**	**ACAD**	**Merit**
INT	1	0.191**	0.198**	0.787**
GAT	---	1	0.125**	0.611**
ACAD	---	---	1	0.507**
Merit	---	---	---	1
** Correlation is significant at 0.01 level (2-tailed) --- Repetitive correlation coefficient values				

We ran the MLR model on this data and at this time the constant term is significant. The developed MLR model is:

$$\boldsymbol{ln}(\boldsymbol{Merit}) = \mathbf{0.5882} + (\mathbf{0.14545} * \mathbf{ln}(\boldsymbol{INT})) + (\mathbf{0.54249} * \mathbf{ln}(\boldsymbol{GAT})) + (\mathbf{0.32904} * \boldsymbol{ln}(\boldsymbol{ACAD})) \quad \textbf{4-8}$$

Table 4.29 ln_INT_dataset combined results for MLR analysis

Coefficients:					
Terms		**Coef**	**SE Coef**	**t-value**	**p-value**
Constant		0.5882	0.0281	20.90	0.000
ln(INT)		0.14545	0.00121	120.62	0.000
ln(GAT)		0.54249	0.00631	86.03	0.000
ln(ACAD)		0.32904	0.00529	62.20	0.000
Model Summary		**S**	**R-sq**	**R-sq(adj)**	
		0.0353288	94.09%	94.08%	
ANOVA					
	df	**Adj MS**	**Adj SS**	**F-value**	**p-value**
Regression	3	15.0391	45.1172	12049.38	0.000
ln(INT)	1	18.1587	18.1587	14548.85	0.000
ln(GAT)	1	9.2371	9.2371	7400.77	0.000
ln(ACAD)	1	4.8281	4.8281	3868.28	0.000

From Table 4.29, It can be seen that there is no difference in terms of the individual and overall significance of this model. However, the residual term is not normal here, with a p-value of less than 5% (Figure 4.14).

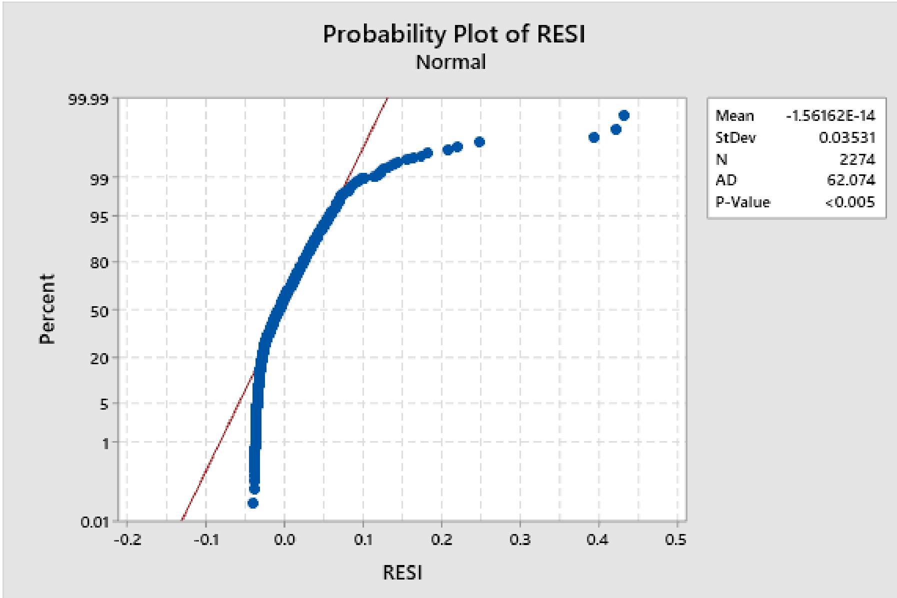


Figure 4.14 Normality plot ln_INT_dataset

The above-mentioned results concerning the development of the MLR model considering merit as dependent variables reveal the following facts:

i. There does not exist a strong linear relationship between dependent and independent variables. Most of the models are showing superficial results.

ii. Assessment analysis of the developed models is showing statistically insignificant results besides data trimming and transformation of variables. Hence, there is a question mark on the statistical as well as the practical use of these models.

Therefore, we can conclude that the given variables are not suitable for the development of predictive models in a linear framework. Hence, an alternate approach would be the use of logistic regression considering the status of the applicant as dependent variables with the given set of three independent variables.

4.5.2 Binary Logistic Regression (BLR)

The definition of status, mentioned in the previous chapter, states that status is a categorical (dichotomous) dependent variable and there are three continuous independent variables (ACAD, INT, GAT). Therefore, BLR is an apt technique. Before actually diving into BLR there is a need to check the general assumptions. All of those pre-requisites by [58] were reviewed, and then we moved forward with SPSS for the model development. The predictive model is developed to predict the status based on the aforementioned input variables.

At first, all of those independent variables were added along with a dependent variable. From Table 4.30, It can be seen that the total number of complete observations(n) used in this model is 13089.

Table 4.30 Case Processing Summary

		n	Percent
Selected Cases	Included in Analysis	13089	100.0
Unselected cases		0	.0
	Total	13089	100.0

The automatic status encoding by the SPSS is 0 for Admitted and 1 for Not-admitted. Thereby, *p* is representing the probability of being not admitted in the BLR model. The output consists of a null model (Block 0: having no explanatory variable) and then the model with all explanatory variables (Block 1). Table 4.31 tells us about the model prediction solely without including variables and the overall percentage is 58.3, “ a little bit better than tossing a coin” [43]. However, Table 4.32 is showing that this null model, having a constant (B) term only, is significant with a p-value = 0.000 < 0.01. This statistical significance is only since the data size is large enough, although it is predicting only 58% accurately. It has been observed that our large sample size is triggering the high levels of statistical significance for relatively minor effects in several cases [43].

Table 4.31 Classification Table [a,b]

<table>
<tr><th colspan="3" rowspan="3">Observed</th><th colspan="3">Predicted</th></tr>
<tr><th colspan="2">STATUS</th><th rowspan="2">Percentage Correct</th></tr>
<tr><th>Admitted</th><th>Not-Admitted</th></tr>
<tr><td rowspan="3">Step 0</td><td rowspan="2">STATUS</td><td>Admitted</td><td>0</td><td>5455</td><td>.0</td></tr>
<tr><td>Not-Admitted</td><td>0</td><td>7634</td><td>100.0</td></tr>
<tr><td colspan="2">Overall Percentage</td><td></td><td></td><td>58.3</td></tr>
<tr><td colspan="6">a. Constant is included in the model.
b. The cut value is 0.500</td></tr>
</table>

Table 4.32 Variables in the Equation

		B	**S.E**	**Wald**	**df**	**Sig**	**Exp (B)**
Step 0	**Constant**	.336	.018	358.894	1	.000	1.399

After analyzing the null model, we moved towards the main model with all of the explanatory variables. Here, the "Model" row of Table 4.33. shows the comparison between the null model and this model based on significance, the improvement over the baseline model. Thus, this model is significantly better because the chi-square value is highly significant (chi-square = 1496.869, p-value < 0.000). As we added all the explanatory variables at once instead of any stepwise approach, that is why the same values of "Step" and "Block". This model is explaining approximately 14.5% of the variance in the outcome (Table 4.34).

Table 4.33 Omnibus Tests of Model Coefficients

		Chi-square	**df**	**Sig.**
Step 1	**Step**	1496.869	3	.000
	Block	1496.869	3	.000
	Model	1496.869	3	.000

Table 4.34 Model Summary

Step	-2 Log likelihood	Nagelkerke R square
1	16283.893[a]	0.145
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.		

This model is behaving more accurately than the baseline model. Thus, this statement is supported by the correct overall percentage of this model that is 62.3%, and there is an increase of almost 6.86% from the null model's correct percentage (Table 4.35).

Table 4.35 Classification Table [a]

Observed			Predicted		
			STATUS		Percentage Correct
			Admitted	Not-Admitted	
Step 1	STATUS	Admitted	2510	2945	46.0
		Not-Admitted	1997	5637	73.9
	Overall Percentage				62.3
a. The cut value is 0.500					

After looking at the overall model significance it is also essential to look at each explanatory variable's significance. Table 4.36 is showing INT as a highly significant variable (Wald = 1053.527, p-value < 0.000) and ACAD is the least significant one (Wald = 10.207, p-value < 0.02). Coefficients (B) for all the explanatory variables are significant and negative, indicating that one unit change in these variables will decrease the odds of falling into the Not-Admitted group. The Exp(B) column is showing the odds ratio of the corresponding variables. There is approximately a 13% less chance to be Not-Admitted with the increase of one mark in INT. GAT and ACAD have their odds ratio very close to 1, which shows roughly no or negligible association between independent and dependent variables [59]. If there is an increase in the mark of GAT and ACAD, then the chance of being Not-Admitted decreased by approximately 1% and 2%, respectively.

Table 4.36 Variables in the Equation

		B	S.E	Wald	df	Sig.	Exp(B)	Confidence Interval	
								Lower	Upper
Step 1[a]	INT	-0.135	0.004	1053.527	1	.000	0.873	0.866	0.881
	GAT	-0.011	0.003	20.923	1	.000	0.989	0.984	0.993
	ACAD	-0.023	0.007	10.207	1	.002	0.977	0.963	0.991
	Constant	3.616	0.189	367.196	1	.000	37.188		
a. Variable(s) entered on step 1: INT, GAT, ACAD.									

After running the BLR model it is necessary to evaluate the model to see the error trend and magnitude. We used the confusion matrix for the evaluation of basis error measures. We used SPSS and Excel to obtain the details of the confusion matrix (Table 4.37). Then we calculated the Sensitivity, Specificity, and Accuracy of the model with all of the variables. It turns out the model is more accurately predicting applicants that are Not-Admitted but less accurately predicting the Admitted ones (Table 4.37).

Table 4.37 Confusion Matrix

		Status	
Predicted Group		Positive=**1**=NOT-admitted	Negative=**0**=admitted
	Positive=**1**=NOT-admitted	**TP** = 5637	**FP** = 2945
	Negative=**0**=admitted	**FN** = 1997	**TN** = 2510
Performance measures for evaluation			
	Sensitivity	**Specificity**	**Accuracy**
	73.84%	46.01%	62.24%
TP = No. of correct classifications predicted as NOT-Admitted (positive).			
TN = No. of correct classifications predicted as Admitted (negative).			
FP = No. of applicants that are incorrectly predicted as NOT-Admitted when it is Admitted.			
FN = No. of applicants that are incorrectly predicted as Admitted when it is NOT-Admitted.			

Initially, we considered all of the explanatory variables and ran the model. We headed towards the stepwise approach to see the possible refinements in it. In SPSS, some hierarchical approaches, i.e., Forward (Backward) Likelihood Ratio, Forward (Backward) Wald, etc. We initiated with a Forward stepwise Likelihood Ratio (LR). It worked on adding the most significant variable initially and then kept on adding the following most crucial variable and so on till the full significance of the model was achieved.

Table 4.38 Omnibus Tests of Forward LR Model Coefficients

		Chi-square	df	Sig.
Step 1	**Step**	1462.985	1	.000
	Block	1462.985	1	.000
	Model	1462.985	1	.000
Step 2	**Step**	23.674	1	.000
	Block	1486.659	2	.000
	Model	1486.659	2	.000
Step 3	**Step**	10.210	1	.001
	Block	1496.869	3	.000
	Model	1496.869	3	.000

The null model of forward LR is the same as the null model initially without the stepwise approach model. In Table 4.38, there are three steps because this model is achieving its maximum significance in three steps with three variables. We only have three explanatory variables, which means that our model is most significant after considering these independent variables. This model initially had low chi-square value (chi-square = 1462.985, df =1, p-value < .000) and then increased in the chi-square value (chi-square = 1496.869, df =3, p-value < .000). Thus, indicating that the model is most significant if it treats all of the explanatory variables. Table 4.39 is showing consistent results. In each step, the - 2LL value is increasing as well as the R-square value. By taking only the most significant variable, the model is explaining approximately 14% of the outcome. In contrast, after considering all of the three significant variables, the model explains approximately 14.5% of the outcome.

Table 4.39 Forward LR Model Summary

Step	-2 Log likelihood	Nagelkerke R square
1	16317.777[a]	0.142
2	16294.103 [a]	0.145
3	16283.893[a]	0.145
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.		

Table 4.40 Forward LR Classification Table [a]

Observed			Predicted		
			STATUS		Percentage Correct
			Admitted	Not-Admitted	
Step 1	STATUS	Admitted	2518	2937	46.2
		Not-Admitted	2014	5620	73.6
	Overall Percentage				62.2
Step 2	STATUS	Admitted	2495	2960	45.7
		Not-Admitted	1999	5635	73.8
	Overall Percentage				62.1
Step 3	STATUS	Admitted	2510	2945	46.0
		Not-Admitted	1997	5637	73.8
	Overall Percentage				62.2
a. The cut value is 0.500					

The correct overall percentage in step 1 (Table 4.40) was 62.2 concerning the most significant variable. In contrast, after adding the second most crucial variable, the correct percentage reduces by the difference of 0.1, i.e., 62.1%. Finally, this percentage increases by the same difference to 62.2% after combing all of the three independent variables. This value is the same as the initial model without any stepwise approach (Table 4.35).

Table 4.41 Forward LR Variables in the Equation

		B	S.E	Wald	df	Sig.	Exp(B)	Confidence Interval	
								Lower	Upper
Step 1[a]	**INT**	-0.141	0.004	1201.783	1	.000	0.868	0.862	0.875
	Constant	2.607	0.070	1386.373	1	.000	13.554		
Step 2[b]	**INT**	-0.138	0.004	1116.046	1	.000	0.872	0.865	0.879
	GAT	-0.012	0.002	23.628	1	.000	0.988	0.983	0.993
	Constant	3.275	0.155	446.417	1	.000	26.444		
Step 3[c]	**INT**	-0.135	0.004	1053.527	1	.000	0.873	0.866	0.881
	GAT	-0.011	0.003	20.923	1	.000	0.989	0.984	0.993
	ACAD	-0.023	0.007	10.207	1	.001	0.977	0.963	0.991
	Constant	3.616	0.189	367.196	1	.000	37.188		

a. Variable(s) entered on step 1: INT.
b. Variable(s) entered on step 2: GAT.
c. Variable(s) entered on step 3: ACAD.

From Table 4.41, it can be seen that initially INT (Wald = 1201.783, p-value < 0.000) was added as the most significant variable among all other variables. At the second step GAT (Wald = 23.628, p-value < 0.000) was added as the second most significant variable with INT (Wald =1116.046, p-value < 0.000). The third step was initiated because there was some room for achieving the full significance. Thus, it added all of the three explanatory variables in the final model INT (Wald = 1053.527, p-value < 0.000), GAT (Wald = 20.923, p-value < 0.000), and, ACAD (Wald = 10.207, p-value < 0.002). The odds ratio in the Exp(B) column is telling the same story as Table 4.36.

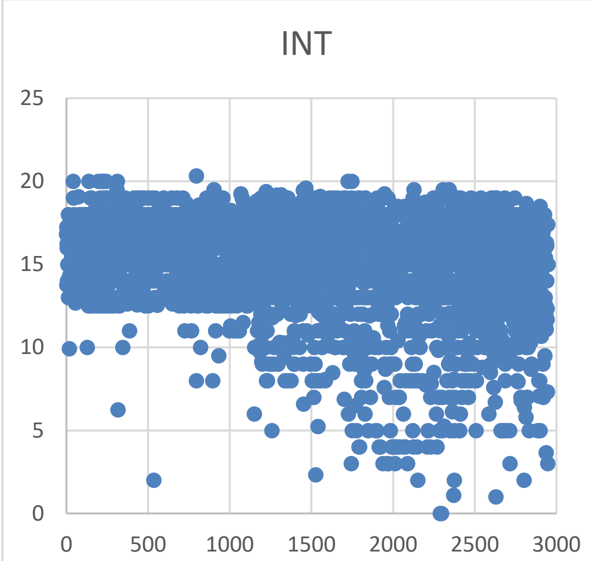
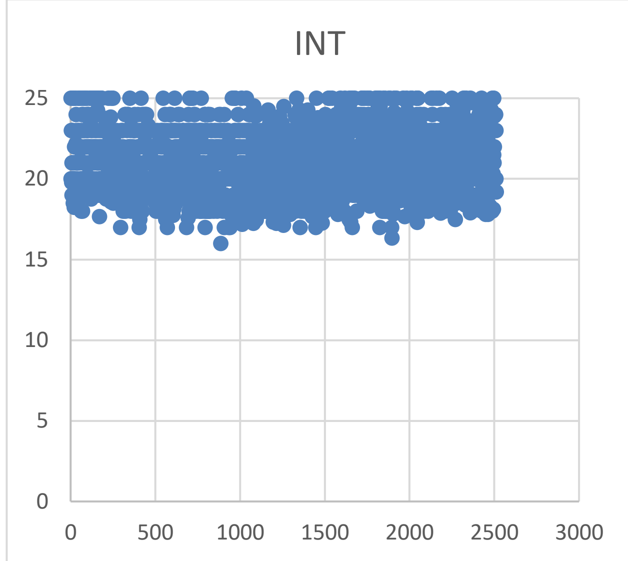
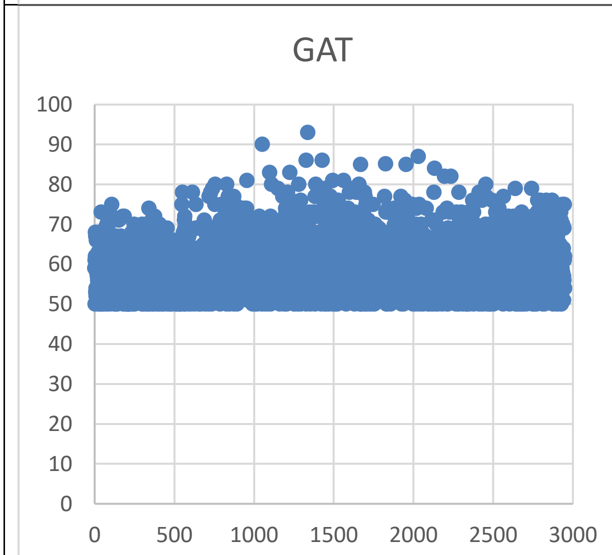
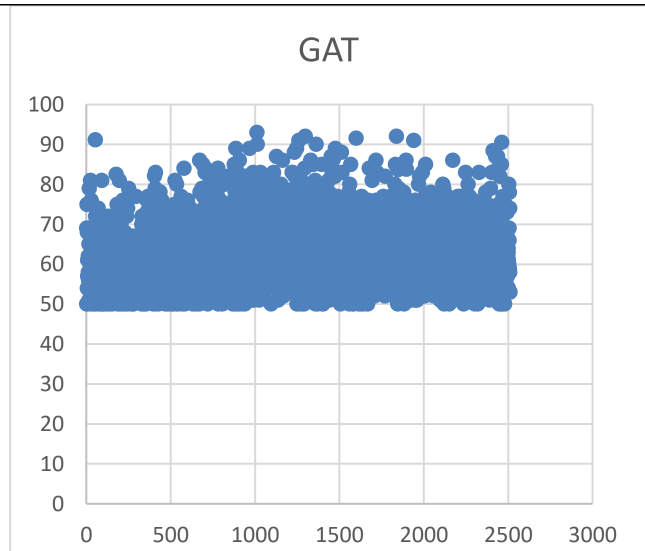
After running the model by forward LR, we ran the backward LR method to see the behavior of the variables in it. There was no difference in any value except the principle on which the backward LR is working, initially picking all of the explanatory variables and then eliminating those that are not contributing towards utmost significance. Thus, it terminated on the first step with all of the three variables (Table 4.42). We re-ran the model with the forward and backward

Wald method to look for possible refinement in our model. Thus, we got the same results as of forward LR and backward LR models.

Table 4.42 Backward LR Variables in the Equation

		B	S.E	Wald	df	Sig.	Exp(B)	Confidence Interval	
								Lower	Upper
Step 1[a]	**INT**	-0.135	0.004	1053.527	1	.000	0.873	0.866	0.881
	GAT	-0.011	0.003	20.923	1	.000	0.989	0.984	0.993
	ACAD	-0.023	0.007	10.207	1	.001	0.977	0.963	0.991
	Constant	3.616	0.189	367.196	1	.000	37.188		
a. Variable(s) entered on step 1: INT, GAT, ACAD.									

After applying different stepwise methods, we dug into the evaluation Table 4.37 for understanding the behavior of False Positive and False Negative values. We used Excel to plot those observations. It can be seen from Figure 4.15 that INT is showing a different trend in FP & TN. Applicants are correctly predicted when roughly they are above the threshold of 15 scores and wrongly predicted when the scores are approximately below 20. From Figure 4.16, it can be seen that INT is behaving differently here too. Applicants predict correctly when the score is roughly below 20 and predicted otherwise when the score is approximately above 15. Whereas, almost the trend for GAT & ACAD is the same in both the categories.

FP	TN
INT 25 20 15 10 5 0 0 500 1000 1500 2000 2500 3000	INT 25 20 15 10 5 0 0 500 1000 1500 2000 2500 3000
GAT 100 90 80 70 60 50 40 30 20 10 0 0 500 1000 1500 2000 2500 3000	GAT 100 90 80 70 60 50 40 30 20 10 0 0 500 1000 1500 2000 2500 3000

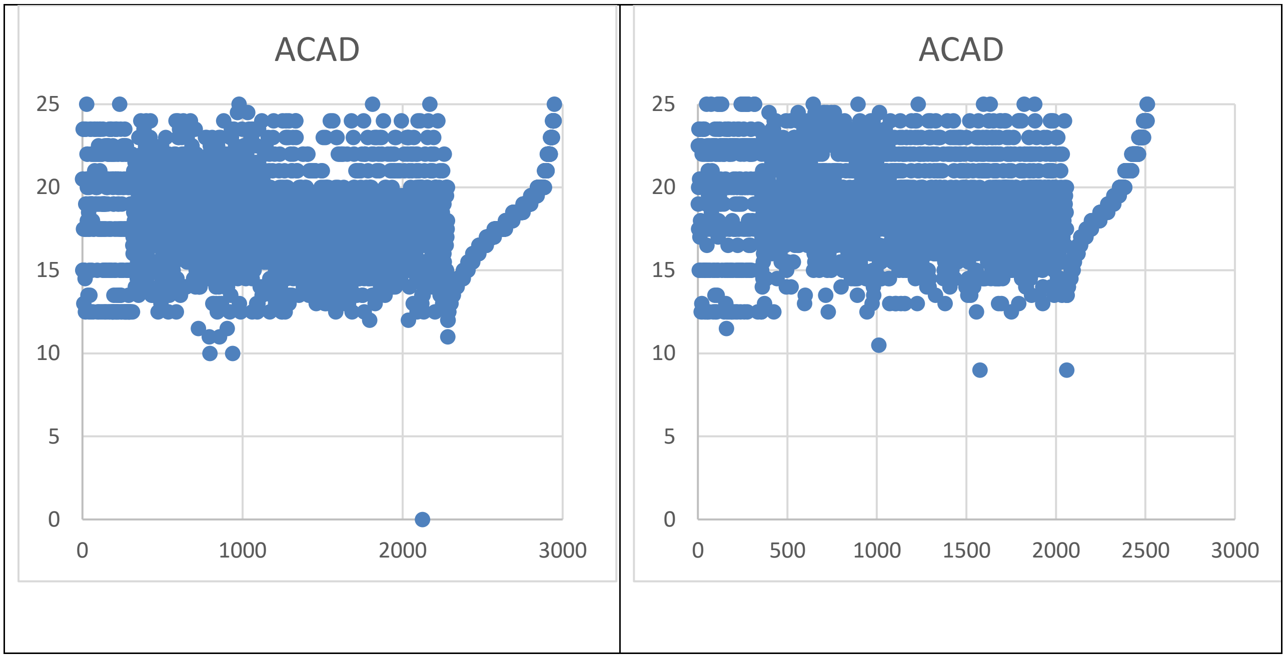
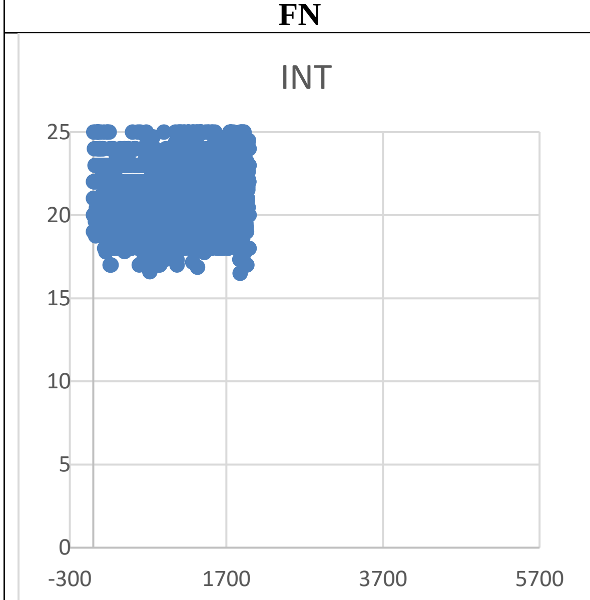
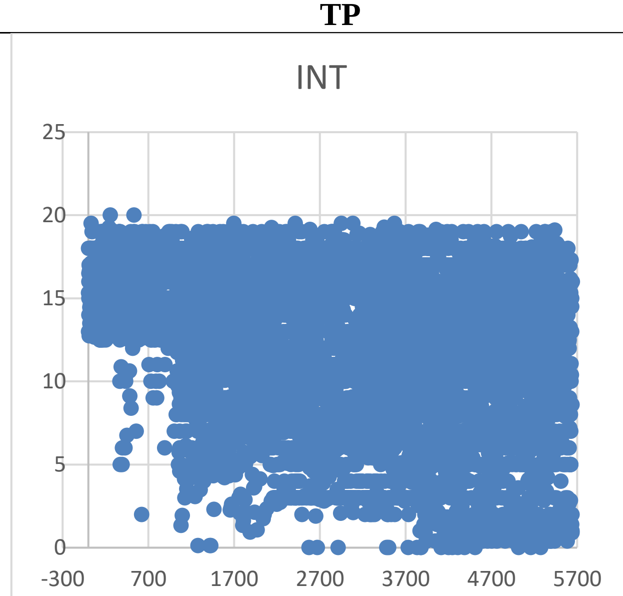


Figure 4.15 Predicted & Observed Values for the initial model (a).

FN	TP
	

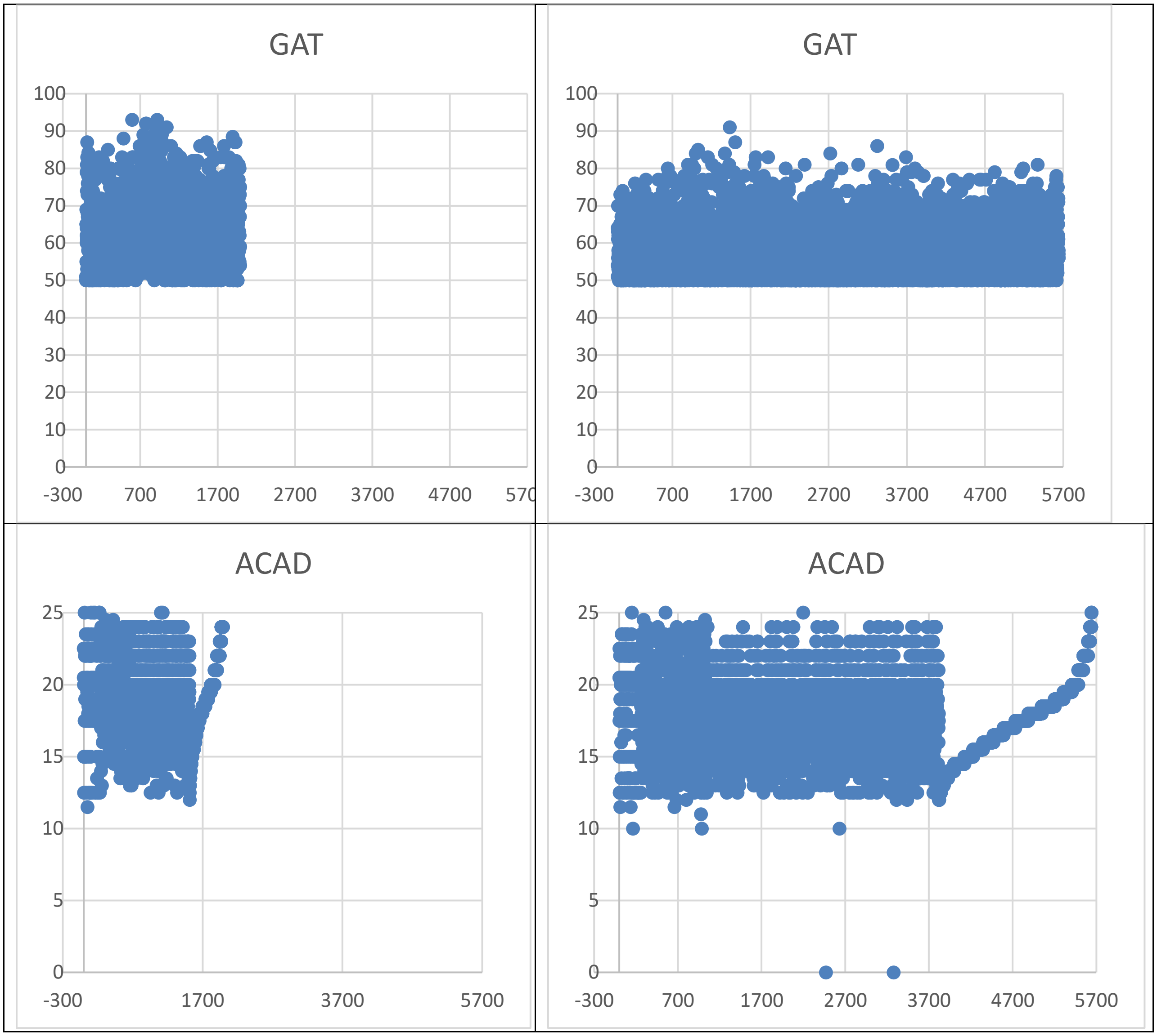


Figure 4.16 Predicted & Observed Values for the initial model (b).

After observing the scatter plots, we decided to run the model with a single explanatory variable i.e., INT with Status (INT-Model), GAT with Status (GAT-Model) and, ACAD with Status (ACAD-Model), to see if we can add some strength to the overall accuracy. we ran the model and calculated the accuracy for each of them. The initial model has the maximum overall accuracy, followed by the INT model accuracy (Table 4.43). This concludes that the initial model is so far the most accurate one.

Table 4.43 Models Evaluation (a)

S. No.	Measures	Initial Model	INT-Model	GAT-Model	ACAD-Model
1	**Sensitivity**	73.84%	73.62%	93.82%	95.49%
2	**Specificity**	46.01%	46.16%	9.04%	7.37%
3	**Accuracy**	62.24%	62.17%	58.48%	58.77%

After running individual models, we realized the need to consider the pairwise model approach, i.e., INT_GAT model, INT_ACAD model and, ACAD_GAT model. we ran the models to check whether these combinations can increase the overall accuracy. After comparing these values, it can be concluded that by considering all of these three variables we can get the most from our model (Table 4.44).

Table 4.44 Models Evaluation (b)

S. No.	Measures	Initial Model	INT_GAT Model	INT_ACAD Model	GAT_ACAD Model
1	**Sensitivity**	73.84%	73.81%	73.76%	90.03%
2	**Specificity**	46.01%	45.74%	45.88%	14.70%
3	**Accuracy**	62.24%	62.11%	62.14%	58.64%

The results mentioned above are running the BLR model with the natural values of the explanatory values. To observe any change in the model accuracy by removing the scale effect, we decided to run normalized/standardized techniques on the variables under observations. Using SPSS, two main techniques(MIN-MAX scaler, Z-scores) were used, as mentioned in the previous chapter. After normalizing all of these three explanatory variables, we ran the BLR model, checked the accuracy of both models, and compared it with the original non-normalized values. It was observed that the model accuracy did not increase by normalizing and standardizing the values of independent variables. Thus, it can be concluded that our model's overall performance is not associated with any scaling effect (Table 4.45). Thus, the most optimized BLR model we have with the most accuracy is:

$$\boldsymbol{log\left(\frac{p}{1-p}\right) = 3.62 - (0.14 * INT) - (0.01 * GAT) - (0.02 * ACAD)} \qquad \textbf{4-9}$$

In a nutshell, this BLR model is showing far more stable results than that of MLR. The main findings from this modeling part are somehow in line with the analysis that we have done before. In other words, we can say that here the appropriateness of BLR hovers over the MLR. In the assessment phase of the final model, the accuracy of the model was acceptable. However, the numbers of sensitivity and specificity were raising the question of completeness of the variables.

Table 4.45 Models Evaluation (c)

S. No.	**Measures**	**Initial Model**	**MIN-MAX Model**	**Z-Score Model**
1	**Sensitivity**	73.84%	73.84%	73.84%
2	**Specificity**	46.01%	46.01%	46.01%
3	**Accuracy**	62.24%	62.24%	62.24%

Chapter Five

5 Summary, Conclusions & Recommendations

A literature review of the past studies in the second chapter showed that there exist a variety of heterogeneous variables for an efficient enrollment process. In the 3rd Chapter, a stepwise approach has been used to analyze NUST postgraduate admission process. Forth chapter describes the trends and tendencies, discrimination power, predictive ability, and completeness of the three variables considered for the calculation of merit or decision of status of an applicant. This chapter presents a brief summary of the work done, major conclusions, recommendations for future research, and implications.

5.1 Summary

Universities are practicing certain criteria for screening students at different levels. A variety of literature is available on the enrollment processes, screening characteristics, and completeness with respect to the predictive ability of the model articulated from the process. Most of them provided comparisons and combinations of variables which provide useful guidelines for the development of screening processes for the universities. The literature reviewed as a base to this study either had its focus on the relationship between different variables used in the admission process or predictive power or completeness of the process as a whole. A major missing link is a debate or support of subjective process using empirical analysis. One of the novelties of this research is that it evaluates a subjectively employed policy at the postgraduate level through empirical analysis. This debate is an initial step towards the introduction of data-driven policies or processes for the screening of students, especially at the postgraduate level. Future research will open avenues of a data-reflective and uniform admission policy in the universities, especially for Pakistan.

5.1 Conclusions and Findings

Based on the empirical analysis of evaluating a subjectively employed policy by a leading national university of Pakistan consisting of ACAD, GAT, and INT scores of the applicants, major findings are provided below:

i. Descriptive analysis of the variables showed positive skewness in GAT scores with most consistency around average, symmetric behavior of ACAD scores, and negative skewness in INT scores with least consistency around average. Hence, students are performing better in INT scores but with the highest variation. Interestingly, with respect to GAT scores, the majority of the students (75% of the students have GAT scores from 50 to 64) performing within a smaller range towards the lower limit of the variable. These findings contradict the assigned subjective weightage of GAT scores in the admission process.

ii. Comparison of descriptive statistics of three variables with respect to the status of applicants showed comparable performances in ACAD scores as well as GAT scores. Note that the averages of ACAD and GAT scores are quite closer to each other (for admitted and not admitted students) but the difference between averages of the two categories for each variable is statistically significant. Therefore we can conclude that ACAD and GAT scores are not drawing a clear line between admitted and not admitted applicants. However, INT scores are playing a prominent role in the distinction of applicants in this domain, with consistent results in both descriptive and inferential analysis.

iii. Correlation analysis breaks various myths of a strong linear relationship between ACAD, GAT, and INT scores as in fact the correlation is weak though significant. For instance, a myth that applicants with high ACAD scores usually have high GAT scores and perform excellently in INT is not supported by the analysis.

iv. For the development of the predictive models and to evaluate the completeness of independent variables, linear methods with the least-squares estimation method are not suitable due to the functional relationship between the three independent variables (ACAD, GAT, and INT scores) and merit being the continuous dependent variable. Therefore, BLR models, as an alternative in non-linear framework, have been used. The results of BLR models showed that INT scores is the most significant variable. Moreover, the change of one score of INT changes the log odds to 13 % with respect to the defined category. For the current information, the maximum possible accuracy achieved is almost 63% for the prediction of the status. Therefore another main conclusion is that these three variables, ACAD, GAT and INT scores are incomplete in terms of predicting the status of an applicant. Hence there is a need to consider other important characteristics of the applicants.

5.2 Future recommendations

Some recommendations for future work include:

1. Dimension reduction methods like principal component analysis can be used to calculate new weightages of the existing variables.

2. Predictive power and completeness of the model can be improved by incorporating various other characteristics of the applicant like gender, age, availability of hostel facility, financial status, the popularity of a program, etc.
3. Use of other machine learning models like Artificial Neural Network, Support Vector Machine, Random Forest, Decision Trees, etc., can also be employed to check the adequacy and completeness of the existing and newly incorporated variables.

References

[1] A. Chadi and M. de Pinto, "Selecting successful students? Undergraduate grades as an admission criterion," *Appl. Econ.*, vol. 50, no. 28, pp. 3089–3105, 2018, doi: 10.1080/00036846.2017.1418072.

[2] J. Kim, "The Functions and Dysfunctions of College Rankings: An Analysis of Institutional Expenditure," *Res. High. Educ.*, vol. 59, no. 1, pp. 54–87, 2018, doi: 10.1007/s11162-017-9455-1.

[3] C. C. Morphew and C. Swanson, "University Rankings," *Univ. Rank.*, pp. 185–199, 2011, doi: 10.1007/978-94-007-1116-7.

[4] G. Buela-Casal, O. Gutiérrez-Martínez, M. P. Bermúdez-Sánchez, and O. Vadillo-Muñoz, "Comparative study of international academic rankings of universities," *Scientometrics*, vol. 71, no. 3, pp. 349–365, 2007, doi: 10.1007/s11192-007-1653-8.

[5] D. A. Gioia and K. G. Corley, "The rankings game: Managing business school reputation," *Corp. Reput. Rev.*, vol. 3, no. 4, pp. 319–333, 2000.

[6] "GRE General Test Structure (For Test Takers)." https://www.ets.org/gre/revised_general/about/structure/ (accessed Feb. 05, 2021).

[7] G. Harman, "Student selection and admission to higher education:Policies and practices in the Asian region," *High. Educ.*, vol. 27, no. 3, pp. 313–339, 1994, doi: 10.1007/BF01432073.

[8] J. Vulperhorst, C. Lutz, R. de Kleijn, and J. van Tartwijk, “Disentangling the predictive validity of high school grades for academic success in university,” *Assess. Eval. High. Educ.*, vol. 43, no. 3, pp. 399–414, 2018, doi: 10.1080/02602938.2017.1353586.

[9] J. Zimmermann, A. von Davier, and H. R. Heinimann, “Adaptive admissions process for effective and fair graduate admission,” *Int. J. Educ. Manag.*, vol. 31, no. 4, pp. 540–558, 2017, doi: 10.1108/IJEM-06-2015-0080.

[10] C. Kulatunga-Moruzi and G. R. Norman, “Validity of Admissions Measures in Predicting Performance Outcomes: A Comparison of Those Who Were and Were not Accepted at McMaster,” *Teach. Learn. Med.*, vol. 14, no. 1, pp. 43–48, 2002, doi: 10.1207/S15328015TLM1401_10.

[11] S. M. Lucieer, K. M. Stegers-Jager, R. M. J. P. Rikers, and A. P. N. Themmen, “Non-cognitive selected students do not outperform lottery-admitted students in the pre-clinical stage of medical school,” *Adv. Heal. Sci. Educ.*, vol. 21, no. 1, pp. 51–61, 2016, doi: 10.1007/s10459-015-9610-4.

[12] N. R. Schripsema, A. M. van Trigt, J. C. C. Borleffs, and J. Cohen-Schotanus, “Selection and study performance: Comparing three admission processes within one medical school,” *Med. Educ.*, vol. 48, no. 12, pp. 1201–1210, 2014, doi: 10.1111/medu.12537.

[13] 1999 Terranova, Pablo Lombeida, “No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title,” *ALI ASADZADEH (2017)the role Tour. Environ. its Gov. law.Electic J. Biol. 13.*, no. 3, pp. 1–8, 1999, doi:

10.5860/choice.41-2927.14.

[14] E. L. Goldberg and G. M. Alliger, “Assessing the Validity of the Gre for Students in Psychology: A Validity Generalization Approach,” *Educ. Psychol. Meas.*, vol. 52, no. 4, pp. 1019–1027, Dec. 1992, doi: 10.1177/0013164492052004026.

[15] M. Robertson, W. N.-A. Psychologist, and undefined 1961, “The graduate record examination and selection of graduate students.,” *psycnet.apa.org*, Accessed: May 01, 2021. [Online]. Available: https://psycnet.apa.org/record/1962-07107-001.

[16] M. Harvancik and G. Golsan, “Graduate Record Examination Scores and Grade Point Averages: Is There a Relationship?.,” 1986, Accessed: May 01, 2021. [Online]. Available: https://eric.ed.gov/?id=ED270682.

[17] M. Milner, J. S. Mcneil, and S. W. King, “The GRE: A question of validity in predicting performance in professional schools of social work,” *Educ. Psychol. Meas.*, vol. 44, no. 4, pp. 945–950, 1984, doi: 10.1177/0013164484444015.

[18] J. G. Thornell and A. McCoy, “The Predictive Validity of the Graduate Record Examinations for Subgroups of Students in Different Academic Disciplines,” *Educ. Psychol. Meas.*, vol. 45, no. 2, pp. 415–419, 1985, doi: 10.1177/001316448504500229.

[19] T. Pitman, “Understanding ‘fairness’ in student selection: are there differences and does it make a difference anyway?,” *Stud. High. Educ.*, vol. 41, no. 7, pp. 1203–1216, 2016, doi: 10.1080/03075079.2014.968545.

[20] A. Thomas, M. E. Young, B. L. Mazer, S. E. Lubarsky, and S. I. Razack, "Reliability and validity of the multiple mini interview (MMI) for admissions to an occupational therapy professional program," *Br. J. Occup. Ther.*, vol. 80, no. 9, pp. 558–567, 2017, doi: 10.1177/0308022617713980.

[21] S. I. Haider, M. F. Bari, and S. Ijaz, "Using multiple mini-interviews for students' admissions in pakistan: A pilot study," *Adv. Med. Educ. Pract.*, vol. 11, pp. 179–185, 2020, doi: 10.2147/AMEP.S246285.

[22] G. Hoad-Reddick and T. V. Macfarlane, "An analysis of an admissions system: Can performance in the first year of the dental course be predicted?," *Br. Dent. J.*, vol. 186, no. 3, pp. 138–142, 1999, doi: 10.1038/sj.bdj.4800043a.

[23] A. Alkushi and A. Althewini, "The Predictive Validity of Admission Criteria for College Assignment in Saudi Universities: King Saud bin Abdulaziz University for Health Sciences Experience," *Int. Educ. Stud.*, vol. 13, no. 4, p. 141, 2020, doi: 10.5539/ies.v13n4p141.

[24] N. Davidovitch and D. Soen, "Predicting Academic Success Using Admission Profiles," *J. Int. Educ. Res.*, vol. 11, no. 3, pp. 125–142, 2015, doi: 10.19030/jier.v11i3.9361.

[25] R. Arambewela and J. Hall, "An empirical model of international student satisfaction," *Asia Pacific J. Mark. Logist.*, vol. 21, no. 4, pp. 555–569, 2009, doi: 10.1108/13555850910997599.

[26] B. L. Delgado-Márquez, M. Á. Escudero-Torres, and N. E. Hurtado-Torres, "Being highly

internationalised strengthens your reputation: An empirical investigation of top higher education institutions," *High. Educ.*, vol. 66, no. 5, pp. 619–633, 2013, doi: 10.1007/s10734-013-9626-8.

[27] E. Hazelkorn, *Rankings and the reshaping of higher education: The battle for world-class excellence*. Palgrave Macmillan, 2011.

[28] H. I. G. Her and E. At, "Ranking 2014 of Pakistani Higher Education Institutions (HEIs)," 2014.

[29] L. R. William Eberle, Douglas Talbert, Erik Simpson, "3_Using Machine learning and predictive models to assess admission policies and standards.pdf." 2013.

[30] "Data considerations for regression - Minitab Express." https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/multiple-regression/before-you-start/data-considerations/ (accessed Aug. 25, 2021).

[31] "Data considerations for Binary Logistic Regression - Minitab Express." https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/before-you-start/data-considerations/ (accessed Aug. 25, 2021).

[32] P. Aleksandar, P. Silvana, and Z. Valentina, "Multiple Linear regression model for predicting bidding price," no. October, 2015, [Online]. Available: https://www.researchgate.net/publication/282646773_Multiple_Linear_regression_model_for_predicting_bidding_price.

[33] G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," *Procedia - Soc. Behav. Sci.*, vol. 106, pp. 234–240, 2013, doi: 10.1016/j.sbspro.2013.12.027.

[34] J. W. Osborne and E. Waters, "Four assumptions of multiple regression that researchers should always test," *Pract. Assessment, Res. Eval.*, vol. 8, no. 2, pp. 2002–2003, 2003.

[35] "Least Squares Method Definition." https://www.investopedia.com/terms/l/least-squares-method.asp (accessed Mar. 13, 2021).

[36] "Ordinary least squares - Statistics By Jim." https://statisticsbyjim.com/glossary/ordinary-least-squares/ (accessed Mar. 13, 2021).

[37] L. Hatzilygeroudis, A. Karatrantou, and C. Pierrakeas, "PASS: An expert system with certainty factors for predicting student success," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3213, no. September, pp. 292–298, 2004, doi: 10.1007/978-3-540-30132-5_43.

[38] "Overview for Binary Logistic Regression - Minitab Express." https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/before-you-start/overview/ (accessed Jan. 13, 2021).

[39] J. Cirelli, A. M. Konkol, F. Aqlan, and J. C. Nwokeji, "Predictive analytics models for student admission and enrollment," *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, vol. 2018, no. SEP, pp. 1395–1403, 2018.

[40] Caner Erden, “Logistic Regression using Minitab | by Caner Erden | Towards Data Science,” 2020. https://towardsdatascience.com/logistic-regression-using-minitab-d58a80ec548a (accessed Jan. 13, 2021).

[41] P. D. F. Prof. Steve Strand, Dr. Stuart Cadwallader, “4.2 An Introduction to Odds, Odds Ratios and Exponents,” *Econ. Soc. Res. Counc.*, 2011.

[42] “Wald Test: Definition, Examples, Running the Test - Statistics How To,” *Agresti A. (1990) Categorical Data Analysis. John Wiley and Sons, New York.*, 1990. https://www.statisticshowto.com/wald-test/ (accessed Apr. 02, 2021).

[43] “4.12 The SPSS Logistic Regression Output.” https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod4/12/index.html (accessed Apr. 02, 2021).

[44] “Walds Test- National Center for Research Methods,” 2011.

[45] A. Agresti, “3Rd-Ed-Alan_Agresti_Categorical_Data_Analysis.Pdf,” *International encyclopedia of statistical science*, vol. 47, no. 4. pp. 755–758, 2013.

[46] P. D. F. Prof. Steve Strand, Dr. Stuart Cadwallader, “Chi-Squared (Economic & Social Research Council).”

[47] S. Glen, “Likelihood-Ratio Tests (Probability and Mathematical Statistics) -.” https://www.statisticshowto.com/likelihood-ratio-tests/ (accessed Apr. 15, 2021).

[48] P. D. F. Prof. Steve Strand, Dr. Stuart Cadwallader, “-2LL (Economic & Social Research

Council, ESRC),” p. 2011, 2011.

[49] P. D. F. Prof. Steve Strand, Dr. Stuart Cadwallader, “Hosmer Lemeshow Test (Goodness of fit)- E & S Research Council.,” *Edeb. A J. Middle East. Lit.*, no. 131, pp. 5–16, 2006.

[50] S. L. David W.Hosmer, *Applied Logistic Regression*, vol. 3rd editio. .

[51] P. D. F. Prof. Steve Strand, Dr. Stuart Cadwallader, “4.8 Methods of Logistic Regression,” 2011.

[52] “Normalization | Codecademy.” https://www.codecademy.com/articles/normalization (accessed Apr. 18, 2021).

[53] M. S. R. Khan, Z. Hussain, and I. Ahmad, “A comparison of quadratic regression and artificial neural networks for the estimation of quantiles at ungauged sites in regional frequency analysis,” *Appl. Ecol. Environ. Res.*, vol. 17, no. 3, pp. 6937–6959, 2019, doi: 10.15666/aeer/1703_69376959.

[54] A. Alghamdi, A. Barsheed, H. Almshjary, and H. Alghamdi, “A Machine Learning Approach for Graduate Admission Prediction,” *ACM Int. Conf. Proceeding Ser.*, pp. 155–158, 2020, doi: 10.1145/3388818.3393716.

[55] “RMSE: Root Mean Square Error - Statistics How To.” https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/ (accessed Feb. 16, 2021).

[56] H. E. Daniels, “Exact saddlepoint approximations,” *Biometrika*, vol. 67, no. 1, pp. 59–63,

1980, doi: 10.1093/biomet/67.1.59.

[57] "Fits and diagnostics table for Fit Regression Model - Minitab." https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-regression-model/interpret-the-results/all-statistics-and-graphs/fits-and-diagnostics-table/ (accessed Mar. 17, 2021).

[58] H. Midi, S. K. Sarkar, and S. Rana, "Collinearity diagnostics of binary logistic regression model," *J. Interdiscip. Math.*, vol. 13, no. 3, pp. 253–267, 2010, doi: 10.1080/09720502.2010.10700699.

[59] "Interpreting Odds Ratio - Senguptas Research Academy." https://senguptasresearchacademy.com/odds-ratio/ (accessed Aug. 26, 2021).

Appendices

Appendix I

Scheme of allocation of Marks for the variable ACAD (out of 25)

CGPA **(in terminal degree/transcript)**	**Or**	**Percentage** **(in terminal degree/transcript)**	**Marks Allotted**
4.00		98.00 - 100.00	25.00
3.90 - 3.99		95.00 - 97.99	24.00
3.80 - 3.89		92.00 - 94.99	23.00
3.70 - 3.79		89.00 - 91.99	22.00
3.60 - 3.69		86.00 - 88.99	21.00
3.50 - 3.59		83.00 - 85.99	20.00
3.40 - 3.49		80.00 - 82.99	19.50
3.30 - 3.39		79.00 - 79.99	19.00
3.20 - 3.29		77.00 - 78.99	18.50
3.10 - 3.19		75.00 - 76.99	18.00
3.00 - 3.09		73.00 - 74.99	17.50
2.90 - 2.99		71.00 - 72.99	17.00
2.80 - 2.89		69.00 - 70.99	16.50
2.70 - 2.79		67.00 - 68.99	16.00
2.60 - 2.69		65.00 - 66.99	15.50
2.50 - 2.59		63.00 - 64.99	15.00
2.40 - 2.49		60.00 - 62.99	14.50
2.30 - 2.39		58.00 - 59.99	14.00
2.20 - 2.29		57.00 - 57.99	13.50
2.10 - 2.19		56.00 - 56.99	13.00
2.00 - 2.09		55.00 - 55.99	12.50
- - -		54.00 - 54.99	12.00
- - -		53.00 - 53.99	11.50
- - -		52.00 - 52.99	11.00
- - -		51.00 - 51.99	10.50
- - -		50.00 - 50.99	10.00
- - -		Less than 50	9.00

Note: Percentage will only be valid if cumulative grade point average (CGPA) is not mentioned in terminal degree/transcript.

Appendix II

MS Individual Interview Proforma

Roll No.____________________Name:_________________________________Degree _______________

No.	Parameter	Total Marks Allocated	Total Marks Earned
1	**Personality**		
	a. Appearance	1	
	b. Mannerism	2	
	c. Emotional stability / Maturity	2	
	Sub-Total	**5**	
2	**Communication Skills**		
	a. Writing Skill (must ask candidate to write statement of purpose)	3	
	b. Fluency in expression	2	
	Sub-Total	**5**	
3	**Motivation / Zeal / Commitment**		
	a. Commitment to complete MS	2	
	b. Employbility after graduation or already employed	1	
	c. Potential for success	1	
	d. Leadership Qualities/ Teamworker	1	
	Sub-Total	**5**	
4	**Knowledge of Applied Discipline**		
	a. Degree of knowledge & expression of interest in the applied program	1	
	b. No of core / elective courses taken in UG, relevant to applied program.	1	
	c. Situational Awareness/ General knowledge	1	
	d. Ability to apply existing knowledge	1	
	e. Ability to apply new concepts	1	
	Sub-Total	**5**	
5	**Research Aptitude**		
	a. Research experience / publication	2	
	b. Rapidity in thinking & reasoning	1	
	c. Commitment to intense learning	1	
	d. UG project/ MS thesis	1	
	Sub-Total	**5**	
TOTAL SCORE		**25**	

Remarks (if any):

Signature:_________________ Name:________________ Date:_________________