2017

An Exploration Study of using the Universities Performance and Enrolments Features for Predicting the International Quality

Aeshah Althagafi
Technological University Dublin

An Exploration Study of using the Universities Performance and Enrolments Features for Predicting the International Quality

Aeshah Althagafi

D14124341

A dissertation submitted in partial fulfilment of the requirements of Dublin Institution of Technology for the degree of

M.Sc. in Computing (Data Analytics)

2017

Declaration

I certify that this dissertation which I submit hereby for an examination for the award of MSc in Computing (Data Analytics) is entirely my own work and has not been copied from someone's work without acknowledging or citing properly.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institution of Technology and has not been submitted in whole or part for an award in any other Institution or University.

The work presented in this dissertation conforms to the principles and requirements of the Institution's guidelines for ethics in research.

Signed: **_____Aeshah_____**

Date: **03 January 2017**

ABSTRACT

Quality ranking systems are crucial in the assessment of the academic performance of an institution because these assessment systems give details about how different learning institutions deliver their services. Education quality is also of paramount importance to the students because it is through quality education that these students develop skills that are needed in the job market. Besides, education enhances a student's academic and reasoning capacities.

When universities are subjected to ranking systems, they are likely to improve their quality to be ranked high in the system. When the university administrators are exposed to ranking, competition gears up. Through competition, the quality of education also improves and through that the general education system improves.

In addition, with rapid technological progress, increased human mobility and economic growth, the concept of quality assessment at the national level has shifted to an international level and now the evaluation of higher education quality is being conducted on the basis of international standards and comparisons. In the present context, a global ranking of a university has a significant influence on attracting research funding and academic talent. Universities are expected to collaborate and compete on an international level, and it is no longer enough to achieve excellence within any national group. It is therefore, not surprising that there is a rising tendency among universities to become centres of "World class excellence".

The findings of this study indicated that teaching, citations, income, number of students are key predictors for predicting the international outlook of universities. Also, it showed that geography is a significant contributor that recognized when it was added to the models for assessing the quality of the worldwide universities.

Keywords: *education quality, performance indicators, regression, tuning SVM, international outlook.*

ACKNOWLEDGEMENT

I would like to thank all of the lecturers that have contributed to my learning experience during the MSc in Computing. In particular, ***Luca Longo*** for answering all my questions regard this project and unlimited help that he provided to me.

Also, I would like to thank my supervisor ***Svetlana Hensman*** who helped me during this semester and supporting me to achieve my dissertation.

.Many thanks for all the support that I have received from my family, in particular, my parents for their unfaltering support and encouragement. Special thank for my husband ***Majed*** who helped me for completing my degree, his support has been invaluable.

Contents

Table of Tables

Table of Figures

Table of Equations

ABBREVIATIONS

THER	Time Higher Education Ranking
CWUR	Centre World Universities Ranking
QS	Quacquarelli Symonds
THE -QS	Times Higher Education-Quacquarelli Symonds
CRISP-DM	Cross-Industry Process for Data Mining
ML	Machine Learning
MLR	Multiple Linear Regression Regression
SVM	Support Vector Machine
SVR	Support Vector Regression
RMSE	Root Mean Square Error
R^2	R square
MAE	Mean Absolute Error
CV	Cross Validation
LOOCV	Leave One Out Cross Validation
SPSS	Statistical Package for the Social Sciences
VIF	Variance Inflation Factor
HEECAT	Higher Education Evaluation and Accreditation Council of Taiwan

CHAPTER 1 - INTRODUCTION

Overview of Project Area

Quality ranking systems are crucial in the assessment of the academic performance of an institution because these assessment systems give details about how different learning institutions deliver their services (Dill & Soo, 2005). Education quality is also of paramount importance to the students because it is through quality education that these students develop skills that are needed in the job market. Besides, education enhances a student's academic and reasoning capacities. The parameters used in ranking the universities are therefore vital as they form the basis of a metric that can be used to compare students from various universities (Zineldin, Akdag &Vasicheva, 2011).

When universities are subjected to ranking systems, they are likely to improve their quality to be ranked high in the system. When the university administrators are exposed to ranking, competition gears up. Through competition, the quality of education also improves and through that the general education system improves (Dill, 2006).

In addition, with rapid technological progress, increased human mobility and economic growth, the concept of quality assessment at the national level has shifted to an international level and now the evaluation of higher education quality is being conducted on the basis of international standards and comparisons (Rust & Kim, 2014). In the present context, a global ranking of a university has a significant influence on attracting research funding and academic talent. Universities are expected to collaborate and compete on an international level, and it is no longer enough to achieve excellence within any national group. It is, therefore, not surprising that there is a rising tendency among universities to become centres of "World class excellence" (Hazelkorn, 2006).

Machine learning algorithms constitute an important area of research and are widely used in the financial sector (Zhu et al., 2016), medicine (Shipp et al., 2002), information technology (Sebastian, 2002), etc. However, this area is still unused in the evaluation of the international quality of the higher education. The application of machine learning techniques to this field would aid in avoiding various biases that can be identified in the present methods used for ranking universities. A literature review highlighting these biases has been presented in Chapter 2. In this work, two machine

learning algorithms, namely multiple linear regression and Support Vector Machine will be deployed to assess the international quality of the universities.

1.1 Background

Given the market-oriented, global education scene decisions pertaining to universities such as a choice by a student or funding from government agencies are often determined by the relative merit of the university as compared to its international counterparts. The current systems for ranking universities consider multiple factors to assess and then rate or rank the quality of education in universities. The parameters usually used include the following: the quality of teaching, scholarly publication by the faculty and the students, citations, income and number of students enrolled. Students are relying more on a university ranking to make their decision on where to study to achieve their educational goals. A ranking system allows researchers and policy makers identify high ranked institutions that are likely to be more productive and hence producing better graduates, teaching, researchers and contribute more to the society as a whole.

At present, most of the university ranking studies are carried out by media based entities such as the Times Higher Education Supplement (THES), and World University Ranking and the biases conducted by the authors can greatly influence the final ranking (Buela et al., 2007). There is a scope for developing a scientific and unbiased method for ranking institutions of higher learning. This research attempts to fill this gap in the literature review by using Machine learning algorithms to predict the international quality. It involves conducting experiments and analysing the correlation between the international quality of the universities and these two groups of features. The first group, namely institutional features, contain characteristics related to universities such as teaching, research, geography, the level of English, etc. This research also examines the correlation between the international quality and the second group of features that are related to the student enrolment such as a number of students, staff to student ratio, etc. This research project aims to build different regression and support vector machine models, and then compare the accuracy of predicting of the international quality of universities using both groups of predictors.

1.2 Research Problem

The importance of education to any society cannot be underestimated. Over the years, there has been a great increase in the number of universities worldwide which caused learning institutions to become more competitive and more eager to enrol students. The advent of a global international society has further increased the need for a scientific tool for ranking universities. While some attempts have been made to objectively rank institutions of higher learning such as Shanghai Jiao Tong Ranking (Liu & Cheng, 2005), there is still a pressing need to develop scientific and unbiased approaches for the same.

This research aims to investigate the predictive power of two important indicators, namely, geography and level of English spoken with the two groups of features that mentioned above (features related to the institution and human involved in the learning process). It also examines their impact on the dependent variable "international quality of university". This attempt is made to answer the research question: What are the factors that affect the international quality of the universities?

1.3 Research Hypotheses and Objectives

1.3.1 Hypotheses

The aim of this study is to identify the factors that influence the international quality of universities. Thus, to identify these factors, the following hypotheses are developed and tested in this dissertation to reach the most significant set of factors.

H1: International quality of the universities is affected by the teaching score of the universities

H2: International quality of the universities is affected by the research score of the universities

H3: International quality of the universities is affected by the university income.

H4: International quality of the universities is affected by the citation score of the universities.

H5: International quality of the universities is affected by the number of students enrolled in the universities.

H6: International quality of the universities is affected by the number of the international students enrolled in the universities.

H7: International quality of the universities is affected by the ratio of the female students enrolled in the universities.

H8: International quality of the universities is affected by the ratio of staff to students enrolled in the universities.

H9: International quality of the universities is affected by university location.

H10: International quality of the universities is affected by the level of English which is spoken or used in the learning process.

H11: The accuracy of the multiple linear regression model increases when selecting the significant predictors from the two groups of variables that are related to the institution performance and human element compared with the accuracy of Multiple Linear regression using one group of predictors only.

1.3.2 Objectives

The objectives of this research are summarized in the following points:

- To perform a thorough review of all the available methodologies for the assessment of the universities quality at international level.
- To select and add suitable features to be used for the assessment.
- To analyse the relationships between different features.
- To select the suitable ML algorithms and compare them using relevant evaluation metrics. In this work, two popular evaluation metrics are used, namely Root Mean Squared Error (RMSE), and R-squared (R2).

1.4 Research Methodology

This dissertation uses available data to analyse the importance of various factors in university ranking. No first hand data was collected. Instead, the data available from reputable sources regarding university characteristics is utilized. Therefore, the research methodology can be characterised as secondary data analysis. The parameters used to study university ranking can be numerically quantified, and these quantities are then used to assess university ranking. This structured and data driven approach makes the

research quantitative rather than qualitative. Unlike exploratory research, the secondary data was collected and analysed to derive usable statistical relationships. As the objective of this study is to identify specific indicators associated with university ranking, a down top inductive approach of reasoning is followed rather than a deductive one.

Hypotheses are postulated and tested using available data. The results are based on hypothesis testing and experiments which use data related to different indicators for the investigation of universities assessment.

To sum up, this research is a **secondary** and **quantitative** research, an **empirical** investigation that uses the **inductive reasoning** approach for understanding and selecting the appropriate features and uses statistical models for analysing the **available data**.

1.5 Gaps and motivations

It is observed that there are no published studies that investigate the relationships between different indicators which have been considered in the assessment statistically. In the literature review, various biases in the approaches to the evaluation of the international quality were identified. For example, higher weights are assigned to specific indicators while other significant indicators are often ignored without clearly stating the motivation behind the choice of those weights.

It is necessary to carry out research that provides a methodology which can analyse the existing issues and provide an analysis driven methodology of using the indicators (teaching, research, citations, etc.) associated with the international outlook of universities[1]. It is also necessary to study some new indicators such as investigating the influence of using the English language as a primary academic language in the syllabus, exams and all the academic papers on assessing the quality as argued by (Altbach, 2008; Yingqiang & Yongjian, 2016).

[1] The two terms **international outlook** and **international quality** of a university will be used interchangeably throughout this research.

Nevertheless, from the literature review, it is seen that there is no paper or study examined the impact of the University location on the assessment of the outlook. This is the first study that provides a precise investigation of the indicators that have been used for the current ranking systems as briefly described in the literature review chapter. Also, it was noticed that machine learning algorithms are not used for assessing the quality of higher education at international level. Significant contributions to existing literature can also be made in this area.

1.6 Scope and Limitations

The scope of this study, 818 universities from different countries, in particular, 72 countries that are included for assessing the international quality of the universities (names of the universities and countries are listed in the appendix). Also, this study uses five years of ranking the universities starting with 2011 until 2016, and the total number of features is 13, the universities are not the same each year, some new universities are ranked, and other were excluded in different years.

Due to reliance on the secondary research, the dataset available for study is limited to THE dataset and its features that are related to institutions, professors, administrators and students' enrolments.

1.7 Document Outline

The rest of this document is organised in the following chapters:

Chapter 2 - Literature review: This chapter reviews the existing works related to the methodologies in the universities assessments. It also summarises the factors that are included for ranking universities. Also, it reviews the usage of machine learning models in education, such as regression and support vector machines. It also reviews some evaluation metrics such as R-squared and Root Mean Squared Error. It concluded by defining the gaps and the limitations in the previous papers.

Chapter 3 - Design of the implementation: This chapter explains the exact steps, software and packages that will be considered in the implementation chapter. Also, the section of the limitations and strengths of the design is provided at the end of this chapter.

Chapter 4 - Implementation: This chapter explains in detail the processes and results from the experiments.

Chapter 5 - Evaluation: This chapter provides critical assessment and analysis of the results observed in the implementation chapter and concluded by outlining the key strengths and weaknesses of the experiments.

Chapter 6 - Conclusion: This chapter describes briefly all the work that has been done from the beginning. It summarises all the steps in the previous chapters. Lastly, it provides some suggestions for the future research.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

This chapter reviews different methodologies and concepts in the context of Assessment of Quality for International Higher Educational Institutions. Different Ranking metrics and Machine Learning techniques are reviewed. This chapter provides a detailed explanation of already existing relevant work, which will help in understanding the course of this study.

2.2 Research Context and Background

Increasing standards and international character of Higher Education systems around the world has led many universities, students and governments to take an interest in knowing the comparative Quality and Ranking of a University as compared to other Universities and Institutions. Due to a massive increase in the number of universities in each continent, the Analysis of the quality of a University has become of much importance in past few years around the world.

The first work in the universities ranking is "America's Best Colleges" which was published by the Journal U.S. News and World Reports in 1983. Many other countries started following this enterprise by creating their standards for quality measurements with the added purpose of providing information to consumers and using it as an institutional Marketing strategy. Since then, university quality assessment methods have increased rapidly not only from private institutions but also from public entities and professional organisations.

There are three main issues related to the Assessment of Quality of Universities:

- Who assesses the quality?
- Why assess the quality
- The audience for assessment of quality (Merisotis, 2002)

Most of universities quality assessments and then rankings by assessments are done by media-based and private entities, but many governments and professional organisations and institutions are also focusing on this issue. The primary purpose of Quality Assessment is to provide quality related information to consumers as this helps

them make an informed decision when selecting a particular institution, and also works as a marketing strategy. Another purpose of Quality assessment is to promote a sustainable high-quality and hence, to create a competitive environment between different universities. The last purpose is to address the concerned audience of quality assessment. Students are the most concerned audience of quality assessment. Another consumer of assessment is the Parents of children who manage the expenses of higher education of their children. Some other consumers are government institutions and academic entities who are responsible for educational policies (Buela et al., 2006).

The assessment systems entirely depend on the types of features to be used for quality assessment by a particular author and many rules are established for the quality assessment process (Merisitis, 2005). First of all, data is collected by either original source or from some already available sources. After collection of data, specific types of variables are selected to be used for the assessment of quality. Next step includes the standardisation of the attribute variables and then weights are assigned to these variables. In the last step, comparison and calculations are performed to get the results about the quality of institution under review.

Initially, the Quality Assessment and Ranking of institution was limited to particular nations like rankings of Chinese universities (Liu &Liu, 2005), USA universities (Vaughn, 2002), British universities (Eccles, 2002), Russian universities (Filinov & Ruchkina, 2002), Polish universities (Van Dyke, 2005), German universities (Feferkeil, 2002) and Japanese universities (Yonezawa et al., 2002). With the fast increase in technology, mobility of students and expansion in the economy, the concept of quality assessment at the national level has shifted to an international scale and now the assessment of higher education quality is being done on the basis of international comparisons. This concept has become so much international, and it is no longer sufficient for universities to be compared against universities from the same country. Universities are now compared with their global counterparts and compete with each other globally for acquiring resources (Beula-Casal et al., 2006). Worldwide Academic quality assessment and ranking was first done by The Institution of Higher Education of Shanghai Jiao Tong University (Rust & Kim, 2015). After this, other countries also started working on the comparison between universities around the globe. The first step in the Quality assessment is the selection of attribute by which assessment is to be done.

Next step is the selection of an approach to be used for the assessment using already selected attributes. There are two main approaches for this:

2.2.1 Weight and Sum Based Approach

This method involves assigning some specific weights to each attribute on the basis of its importance and then the calculation of final score by calculating the sum of all the attribute values with weights. A brief explanation of different methodologies for Quality Assessment of universities at international level using weight and sum based approach are presented here. These methodologies are presented according to the typology proposed by Professor Jamie Merisotis (Merisotis, 2002). She presented following components of a systematic Assessment Typology:

Assessment Types:

- Unified: In this type of assessment, many different attributes with some weights are combined which provide an overall quality of an institution under review.
- Discipline-Based: This type of assessment is done on the basis of specific programs, subjects and specialisation offered by a university.
- Other: It includes the assessment that cannot be characterised quickly.

Assessment Structures:

- Numerical: Numbers 1, 2, 3, 4 … are assigned to universities on the basis of quality level.
- Grouping: Universities are grouped in the top, middle and bottom groups according to the degree of quality.
- Top Quality: Only a specific number of top quality universities are mentioned according to this type of assessment.

Assessment Frequency: Assessment of universities can be done at some regular intervals like annually or at some irregular intervals (Rust & Kim,2015; Yingqiang &Yongjian, 2016; Zineldin, Akdag &Vasicheva, 2011; Steve, 2010).

Assessment Sorting: University quality assessments can be sorted out in many different ways like geographical distribution, mission, age, public and private institutions, etc.

Assessment Related Data Sources: The data to be used for quality assessment can either be collected from already available data sources or can also be gathered from original sources like students and surveys, etc.

Following are some International level Quality Assessment methodologies including the components mentioned in typology above with some additional details:

a) World University Ranking

Type: It involves discipline-based approach and unified approach for Quality assessment of universities worldwide.

- Sorting: The assessment according to this method is done on the basis of geographical distributions of different universities.
- Structure: It uses a combination of top level (200 top universities) and statistical approach.
- Data Source: Original and already available data.
- Frequency: This assessment is done annually.

Six indicators are used for the assessment of the quality of higher education institutions by this methodology (Steve, 2010). The six attributes and the weight of each attribute are as follows:

- Faculty to Students ratio (20%)
- International Staff percentage (5%)
- Review of Recruiters (10%)
- International Students percentage (5%)
- Peer Review (40%)
- Each Faculty member citations (20%)

After the calculation of total quality score by these six attributes, the universities are ranked on the basis of this quality score values.

- *b) Academic Rankings of World Universities*
    - Type: It involves unified approach for Quality assessment of universities worldwide.

- Sorting: The assessment according to this method is done on the basis of geographical distributions of different universities.
- Structure: It uses a combination of top level (500 top universities) and statistical approach.
- Data Source: Already available data.
- Frequency: This assessment is done annually.

Six indicators are used for the assessment of the quality of higher education institutions by this methodology. The six attributes and the weightage of each attribute are as follows:

- Total number of articles published related to Science and Nature (20%). For the institutions that are specialised in the fields of social sciences and humanities, this attribute is not used and the weight allocated to this attribute is then shifted to other remaining attributes.
- Number of university staff members who have won Medals in Fields and Nobel Prizes (20%).
- Number of University Alumni members who have won Medals in Fields and Nobel Prizes (10%).
- Total number of articles cited in Arts and Humanities Citation Index, Science Citation Index Extended, Social Science Citation Index (20%).
- Total number of researchers highly cited in 21 subjects categories (20%).
- The last attribute is the size of an institution which is calculated by dividing the total score calculated by top 5 attributes with number of full-time academic staff (10%).

After the calculation of total quality score value by these six attributes, the ranking is done by this quality score values.

c) International Champion League of Research Institutions

- Type: It involves Discipline-Based Approach for Quality assessment of universities worldwide like Agriculture, Clinical Medicines, Engineering, Art and Humanities, Technology and Computing, Biology, Environmental Science, Life Science, Earth Science, Chemical Science and Physical Science.

- Sorting: The assessment according to this method is done on the basis of no specific sorting type.
- Structure: Clustering based approach is used.
- Data Source: Already available data.
- Frequency: This assessment is done at irregular intervals.

Attributes used for the quality assessment of universities are chosen from two categories, institution and sub-discipline.

Attributes from institution category involves:

- Total number of publications.
- Specialisation degree of all research publications.
- Attributes from sub-discipline category involves:
- Weight impact of research publications.
- Research publications activity.
- Research publications world share.
- A total number of published Articles in ISI database.

After the calculation of quality score values on the basis of these six attributes, no weight is assigned to 5 of the characteristics and the ranking is done only on the basis of a total number of published Articles in ISI database i.e. 100% weight is assigned to this attribute.

d) Higher Education Evaluation and Accreditation Council of Taiwan

The HEECAT quality assessment and Ranking methodology assess the quality of universities worldwide and then present to 500 universities. This methodology also uses many different attributes with specific weightage for the Assessment and Ranking purpose. This program started in 2007 and since then assessment attributes have been changed many times. Only overall score based assessment was done at the start of this program, but it also started field based assessment and rankings like SOC (Social Science), ENG (Engineering), LIFE (Life Sciences), etc. Eight indicators are used for the assessment of the quality of higher education institutions by this methodology. The eight attributes and the weightage of each attribute are as follows:

- Total number of publication articles in the year of assessment (10%).

- Total number of publication articles in last 11 years from the year of assessment (10%).
- Total number of Highly Cited Research Papers (15%).
- Total number of publication articles in high impact journals in the year of assessment (15%).
- Total h index value of last two years from the year of assessment (20%).
- Total number of citation in last 11 years from the year of assessment (10%).
- Total number of citation in last two years from the year of assessment (10%).
- An average number of citations in last 11 years from the year of assessment (10%).

These attributes can evaluate the quality of a university in both short term and long term as compared to other methodologies.

e) THE (Times Higher Education)-QS (Quacquarelli Symonds) Method

THE-QS assessment program was started by THE using the data gathered and analysed by QS company. They also presented some Asian Universities Rankings in the beginning but later split and became THE and QS. The attributes of assessment used by THE-QS were adopted by QS while THE joined Thomson Reuters for the development of some new attributes. THE-QS used six different attributes for quality assessment including both qualitative and quantitative attributes. The six attributes and the weight of each attribute are as follows:

- Total number of Citations of each Faculty member (20%).
- A total number of Academic Peer Reviews (40%).
- The ratio between the number of Teachers and Students (20%).
- Reviews from Employer (10%).
- The number of International Students (5%).
- The number of International Faculty members (5%).

After the calculation of total quality score value by these six attributes, the ranking is conducted by this quality score values (Huang, 2011).

f) Centre for World University Rankings (CWUR)

The CWUR ranking of universities is one of the most useful rankings regarding determining the quality of education such universities offer (Jajo & Harrison, 2014). The methodology applied in ranking these universities makes them the most effective in determining universities performance using external factors. High-quality graduates produce high-quality content, and that is why a university with the highest published articles in reputable journals reflects a quality education. Also, patents show ownership of some high-quality content, when a university has signed many patents; it indicates that they are producing high quality and original content that others may want to copy and that is why the universities with these features can be assumed to be offering high-quality education. This is one of the best ways in which the quality of education of institutions can be assessed without having any form of bias. This is because the institutions are not involved in the analysis and the parameters used are external.

There is no way an institution can influence the outcome of the research or their performance since third parties are involved in analysing the organisation products in the market such as the performance of their alumni in the job market. The use of data that is available from external sources is significant because it cannot be influenced by the universities in an attempt to show that they offer high-quality education. Parameters such as the number of citations and the employment rates of the graduates are external, and different people can observe them to ascertain their authenticity. Therefore, when applying the method in assessing the quality of education offered by a particular university, it is possible to get the right information that can be used to relate to the university in question. Therefore, this is the method that provided attributes that can be used in analysing the quality of education that is offered by the universities (Garwe, 2015).

2.2.2 The Jackknife Technique

This is another approach used for the quality assessment of universities. This method is different than weights and sum based approach because it does not assign any weights to the attributes. This methodology replaces one linear model with another linear model in which the overall score values are used as an output variable, and all the attributes are used as predictor variables (Marginson, 2007). This method removes each attribute variable one by one. It recalculates the overall score value after the removal of

an attribute and then repeats the process for all the attributes. In this way, numbers of regression models equal to the number of attributes are estimated.

2.3 Analysis of Reviewed Methodologies

Weights and Sum based methods are easy to implement and are used by many Ranking institutions, but there are many problems pointed out by many critics in this approach (Soh, 2015). One of the problems is the selection of weight values for each attribute because it varies with the person selecting the values of weights. This method is well accepted for the quality assessment of products like cars etc., but it has divided opinions for Educational Quality related tasks due to the reason that it is tough to measure and quantify educational components like reputation, etc. Also, it is hard to find the difference between overall score values by using weight and sum method because the overall score values change with the change in attributes or weights being used.

While it is easy to find the difference between qualities of educational institutions, the overall score values stay stable using the Jackknife technique (Clarke, 2002). This reflects that there is a need for more robust and stable approach for the Quality Assessment which can take some good decisions about the Quality of an Institution. Also, there are no precise studies regarding the analysis of the relationships between the indicators that have been considered in the ranking assessment process. The aim of this research is to provide such a methodology which can overcome all these issues like analysis of relationships between different attributes and the addition of some new assessment attributes such as English Language Level and the University location, etc.

2.4 Machine Learning and Data Mining - Educational Applications

Up-to-date information related to the effectiveness of educational institutions is a high priority issue nowadays. The success of students is also considered a responsibility of institutions (Campbell & Oblinger, 2007). One way to deal with these issues is the application of Machine Learning and Data Mining techniques on educational data in new ways. Although Machine Learning and Data Mining techniques are already applied in many different fields and sectors but the use of these techniques in Educational Applications is limited (Ranjan & Malik, 2007). With the emergence of Educational Data Mining, new methods can now be designed and applied to solve many

different educational field related problems and issues. Literature related to Machine Learning and Data Mining in the field of Education is discussed in this section.

The literature includes the application of Machine Learning and Data Mining methodologies in the solution and analysis of education-related data (Baker & Yacef, 2009). These research methodologies range from the use of Machine Learning and Data Mining in improving the learning process of students to the use of Data Mining and Machine Learning in increasing the effectiveness of educational institutions. There is a wide range of applications and methodologies for the educational applications of Machine Learning and Data Mining, but this review will focus on the applications which are closely related to students and institutions like an evaluation of the performance of students in Management Systems, retention and success of students and recommender systems, etc.

Journal of Educational Data Mining was started by researchers who were interested in Educational Machine Learning and Data Mining in 2009 and also started a yearly conference since 2008 at an international level. The literature has drawn from different disciplines involving Learning Theory, Machine Learning, Psychometrics, Data Mining and Data Visualization (Baker & Yacef, 2009). Some of the research methodologies proposed earlier are published in International Journal on Artificial Intelligence in Education and Conference on Artificial Intelligence in Education. Since Machine Learning and Artificial Intelligence are a big part of Data Mining techniques, many Data Mining techniques were published in Artificial Intelligence related publications earlier. Different Machine Learning and Data Mining applications are reviewed in this section. Power and Limitations of these methodologies are also discussed in this section.

There are many different methodologies proposed by many researchers for the analysis of massive amount of data for extraction of useful information and analysis to help in decision-making process (Shockley et al., 2012). CRISP-DM is a life cycle process which helps in the analysis and development of different data analysis models and techniques (Ruggiero, 2016). This process is helpful in the whole process of creating a model i.e. from an understanding of data to the deployment of the final model. This process includes six phases, including an understanding of the area of implementation,

understanding of data, preparation of data, modelling of technique, evaluation of model and deployment of the model (Leventhal, 2010). The advantage of this framework is that it is not a software vendor specific framework and provides templates and guidance in data analysis (Leventhal, 2010). This concept has been used in many educational applications related studies (Wang, & Liao, 2002; Vialardi et al., 2011; Wang & Liao, 2011).

Machine Learning algorithms can help faculty members in becoming more proactive to assess and identify the students who are at risk and then enable them to respond accordingly (Campbell & Oblinger, 2007). There are many key techniques that can be applied to education related data like association rules mining, multivariate statistics, web mining and classification (Calders & Pechenizkiy, 2012). These methods help in forecasting and prediction of improvements required in institutions for quality improvement. These methods also help in pointing out the differences between students and thus appropriate measures can be taken to improve their learning process (Corbett, 2001).

These methodologies also help Educational Institutions in the assessment of the quality of education they are providing to enhance the decision-making the process for quality improvement which as a result provides financial gains and improved competitiveness (Nemati & Barko, 2004).q

The researchers, Wang & Liao (2002) used Machine Learning and Data Mining methodologies for the identification and prediction of the type of students who will drop out of school and who will return to school again.

Regression Trees and Classification based approach was applied for the development of this system and predicted which students will not be coming back to school. Student success factor was calculated by using both qualitative and quantitative techniques in this research. It was a valuable research since it was a tool to help the students in improving their efforts for retention. In another similar research, Lin (2012) applied Machine Learning and Data Mining techniques for the prediction of students which are likely to get benefits from retention of the programs offered by the campus. Some other researchers also developed a system for the improvement and support of retention using different Data Mining techniques (Chacon, Spicer & Valbuena, 2012).

This research implemented a retention aiding system successfully by using these techniques, and this system helped the faculty to predict the students at risk and then provide help to them. A team of researchers (Chacon et al., 2012) designed a similar system for real time for retention support which is being used at Bowie State University to help students in retention efforts.

These techniques can also be applied for Courses Management Systems. A team of researchers have developed a system using Data Mining and Machine Learning techniques which work inside Course Management System and enables users to get the information about their courses. This system also allows faculty members to share students' results and collaborate with each other (Romero et al., 2011). These techniques can also be used in the development of customized activities for the learning of a student according to their behaviour and progress. It was used in an English Language learning course which was able to adopt the learning activities on the basis of progress of student (Wang & Liao, 2011).

Another use of Machine Learning is the analysis of complex behaviours of students during learning. The research was conducted using three weeks programming assignment online (Blikstein, 2011). This assignment included different coding and non-coding related tasks. Different behaviors of students were analyzed at the end of assignment using different Data Mining techniques. These behaviors helped in profiling the behaviors of students into three categories copy-paste category, mixed category and self-sufficient category. Another research involved the analysis of student behavior in a broader way as compared to the programming related behavior analysis (Dringus & Ellis, 2005).

The involvement of a student learner in an online course is of critical importance. This issue is handled by a researcher by using Machine Learning and Data Mining techniques which can analyze the involvement of a learner and tell if there is some uninvolved learners present (Cocea & Weibelzahl, 2009). This research used different parameters like the speed of learner's reading and the time spent on a page during learning, etc.

There are also many different ways in which Machine Learning is being used by Higher Education Systems like Adoptive systems for learning which keep track of

student learning and then recommend next steps accordingly, different grading systems that help in automatic assessments of student assignments and detect plagiarism, etc.

All these applications of Machine Learning and Data Mining have many advantages in Educational field. With all the advantages mentioned above, there are also some limitations which can be faced while developing or using these applications. Some of the limitations involve the limited accuracy of these applications, time consumption, data collection, application at an extended level instead of applying it to a single institution, etc.

Regression is a quantitative research method which involves analysis of models and several variables (Hayes & Rockwood, 2016). The relationship that developed by regression analysis is between the dependent and independent variables. Regression analysis is a method that is used to predict various outcomes with changes in various variables. Therefore, regression analysis is simply a statistical process that involves estimation of the relationship between variables. There are two types of regression models, linear and non-linear models. In a linear regression model, the dependent variable is a linear combination of independent variables. In a non-linear regression model, the parameters may not be linear, and they are supposed to be analyzed critically in order to predict the outcomes effectively (Hung et al., 2015).

In research, regression models are important, and there is a need to incorporate them in different studies like Education related applications. This is because they enhance the prediction of outcomes and decisions can be made on the basis of the trends that are developed by these models. For instance, educational trends can be predicted effectively using these models. It is important to note that trends are effective in predicting the future and there is a need to develop these trends using regression models. There are various benefits of using regression analysis in a study.

First, the model can be used to predict the future. Regression-based forecasting techniques are important in determining what is likely to happen in the future. Educational organizations can use these models in determining and estimating their rankings in the foreseeable future following the trends that have been developed in the history. Secondly, the models can be used to develop supporting decisions. Thirdly, the models can be used to correct errors in thinking. For instance, the management team of

an educational institution may develop an idea of working in a certain way to improve the ranking which may not be according to the ranking institutions. However, if they consider the regression analysis and the forecasts from the models, they may change their thinking and act on the trends that are developed by the regression models. Finally, the regression models can build new insights that originate from the large amount of data that may be available (Gilstrap, 2013)

Regression and correlation can be used in research to come up with a detailed analysis of the study. There are different reasons why the two can be used in a study. First, is to test the hypotheses about cause-and-effect relationships under regression analysis. In this case, the researcher determines the impact of independent variables on dependent variables and sees whether variations in independent variables have an effect on dependent variables (Hayes & Rockwood, 2016). Second, the use of correlation analysis can be used to determine whether two variables have a relationship and in which direction, positive, negative or no relationship.

SVM is one of the Machine Learning algorithms which can be used for extraction of useful knowledge from a set of data (Sonali et al., 2012). It is a type of supervised Machine Learning algorithm which can be used for both classifications and regression purposes. Many researchers (Sonali et al., 2012) recommend SVM as a classifier which is able to provide Minimum Error and Maximum Accuracy. SVM has been used in many Educational and Non-Educational applications. One of the Educational applications includes the use of SVM for the prediction of placement of students using different attributes (Pratiyush & Manu, 2016). SVM decides if the placement of student is to be done or not on the basis of these attributes. Sample data of 200 Graduate Students was used for classification. The results provided much help to both students and institution in making a good decision about future. Another researcher used SVM for the classification of Education Resources (Xia, 2016). Due to these and many other useful applications of SVM in the educational sector, it is going to be used in this research for Quality Prediction.

2.4.1 Evaluation Metrics

Various applications of Machine learning in Education sector has been discussed in the previous section. The effectiveness of each of these applications, in the case of a

continuous dependent variable, is measured using R^2, RMSE and MAE. R^2 is used to assess the fit of a given dataset to a proposed model (Chai & Draxler, 2014). A higher value of R^2 (maximum being 1) generally indicates a strong correlation between the objective function chosen and the driving variable. The mean absolute error is arguably the most organic measure of average error. It is also the simplest approach, as it is simply the average value of error across a number of data points (Willmott & Matsuura, 2005). RMSE, on the other hand, depends on the square root of the number of errors and MAE. While more complicated, it has shown to be a better indicator of average error if the error distribution follows a Gaussian pattern (Chai &Draxler, 2014). In this work, only RMSE and R^2 will be used to analyse the relation between global university ranking and various parameters. This approach combining the two metrics will give an enhanced understanding of the problem at hand.

2.5 Conclusion

This chapter has presented a review of literature pertinent to understanding the application of machine learning in ranking international universities. This review included detailed knowledge about the domain and methodologies already implemented and used with their advantages and limitations. This research builds on the existing body of literature and extends it by exploring universities rankings based on their ***international outlook,*** a score that measures the degree of internationalization a university achieves. In other words, The ability of a university to attract students and faculty members from all over the world, as well as producing research co-authored by international researchers. The importance for such indicator stems from the fact that this ability of attracting foreign element is key to its success on the world stage.

CHAPTER 3 – DESIGN AND METHODOLOGY

3.1 Introduction

The primary goal of this chapter is to design the experiment for answering the research question. Different techniques will be used for solving the research problem and exploring the relationships among variables. This chapter provides information about the statistical methods used for conducting the experiments and interpreting the results, the main phases of the CRISP-DM methodology will be considered in the design such as data understanding, data preparation, modelling and evaluation. Finally, there is a brief discussion about the strengths and weaknesses of the implementation design.

3.2 Data Understanding

3.2.1 Data collection

The data has been collected by kaggle from THER which has ranked 818 universities under five groups of indicators. Kaggle gathered that ranking data from 2011 until 2016 for comparing three ranked systems Time higher education ranking, CWUR and Academic Ranking of World Universities from Shanghai. Universities are required to provide the annual academic reputation survey and some statistical information related to staff and students, some sort of data was not provided because of confidentiality issues. For example, industrial income would be estimated by choosing a value between the lowest value and the average of all the values of these indicators. The research output analysis provided by Sci Val analytical tool and Scopus journal database help to calculate this indicator. About the final evaluation, the standardisation method was chosen based on the data distribution between specific indicators and cumulative probability function that is calculated. Further, an evaluation is made that at which point the indicator of the particular institution is located in that function. In this way, the cumulative probability score resulted as X describes that the university having the random values will be falling below X percent of the time for that indicator[2] using Z-scoring for calculating the cumulative probability values of the functions.

[2] www.timeshighereducation.com

3.1.2 Data Description

Table 3. 1: Description of 13 features under study and their types

Indicator name	Data Type	Description
world_rank	Ordinal or Interval	The world ranks given to the university, some of the values here are ordered from 1 to 200, and other values are ranged from 200 to 800. The following explains the different types used in this column: 1, 2, 3, …, 200, 200-300, 300-400.
university_name	Nominal	University name is the name of the university.
country	Nominal	The country indicates the location of the university.
english_fluent	Dummy variable	1 indicates that the syllabus, books and learning in the university is based on the English language. 0 refers that the university is not using English for teaching the curriculum to the students.
staff_ student_ratio	Ratio	A ratio of students taught by each member of the faculty.
Citations	Number	The score of university for citations (research influence)
Research	Continuous	The score of the University for conducting research including the income, volume, and reputation
Teaching	Continuous	The total university score of the teaching, this indicator is comprised of other features such as: using technology, online materials, teacher awarded (alumni or Nobel or other international prizes).
International	Continuous	International-to-domestic student ratio, international-to-domestic-staff ratio, and International collaboration
Income	Continuous	It indicates the income of the university
total_quality	Continuous	The total score yields from the sum of weighted indicators, the result used for ranking universities.
num_students	Continuous	All number of the students in the university.
female_male_ratio	Ratio	Proportion of male and female students
Year	Date	Period of 2011 to 2016

3.2.2 Data cleaning/ handling the outliers and missing values

The data will be explored by using IBM SPSS software to check missing values and outliers. To address these issues, the data will be initially analysed then some techniques will be applied for resolving the outliers and missing values such as using the mean for filling the values having less than of 20% of missing values and the data having more than 50% missing values will be permanently removed.

In addition to this, data exploration, through descriptive statistics and visualization, is performed to help understand the nature of the relationship between each feature and the response variable. Data exploration is also useful in identifying which set of transformations, if any, should be performed to help machine learning models achieve better performance. Since some variables have shown a significant degree of skewness, Box-Cox transformation has been used to adjust the skewness of some variables, where adjustment is needed. Also, all variables have been standardized. Excel worksheet will be used for converting the actual values for some features to another format like converting the ratio to the percentage.

3.5 Modelling

The goal of this section is to choose the best model from the two popular ML algorithms and check their assumptions. As mentioned in the introduction chapter, the predictive models will be built by using different features related to students' and the institutions for predicting the international quality score for 800 universities which will be ultimately leading to predict the global ranking as well.

3.5.1 Multiple Linear Regression Model

The objective of this part is to model the regression equation:

$Y= a+ b_1 X_1 +b_2X_2 +\dots +b_i X_i$ Where: $i=1\dots N$

Y = the dependent variable (International Quality)

X = the independent variables, i.e., teaching score, research score, the number of students, etc.

b_i = the coefficients of independent variables that indicate how much the dependent variable (international quality) is dependent on a particular independent variable, keeping everything constant

3.5.2 ML Regression Assumptions

3.5.2.1 Independence of Observations

SPSS software will be utilised for assessing the independence of the observations through Durbin-Watson Statistics, and if the value is equal to 2 or close to 2, this indicates that the independence of the observations exists.

3.5.2.2 Linearity

One of the common assumptions that should be studied in the regression is the existence of the straight-line relationship between the predictors (the group of student features and institutional features) and the response variable (international outlook). If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. Also, the prediction accuracy of the model can be significantly reduced. Residual plots are a useful graphical method for identifying non-linearity. In MLR cases, the plot of the residuals versus the predicted (or fitted) values will be performed. In Ideal cases, the residual plot will not show any discernible pattern. The presence of a pattern may indicate a problem with some aspect of the linear model. If non-linear associations were detected by the residual plot, then non-linear transformations of the predictors will be used such as Box Cox.

3.5.2.3 Constant Variance of Error Terms

The data should represent the homoscedasticity or equal variance among the residuals of variables. The scattered plot used for above assumptions will also be utilised in this assumption. The non-constant variances in the errors, heteroscedasticity, can be identified through the presence of a funnel shape in the residual plot.

3.5.2.4 Absence of Multicollinearity

Multicollinearity indicates that two or more predictor variables are highly correlated or related to each other. The presence of this assumption causes some problems in the regression context since it can be difficult to separate out the individual effects of collinear variables on the response. That results in a great deal of uncertainty in the coefficient estimates. Hence, it increases the standard error of the estimates. In this

study, the data will be analysed for ensuring the group features including students and the institutional features are not highly correlated to each other. VIF will be used for checking this assumption. It can be calculated by dividing one by the tolerance (see formula 1); tolerance is used to measure the effect of one independent variable on the other independent variables that used to build a regression model. It can be calculated by subtracting 1 from the residual square (Williams, 1987).

$$VIF = \frac{1}{1 - R^2}$$

Formula 3.1: Formula for calculating VIF

3.5.2.5 Absence of significant level of outliers

Outliers mean the abnormality of the data that is not following the distribution of normalisation. Outliers can be detected by using two techniques: one is graphical such as scatter plots. The second technique is Bonferroni Outlier test, the p-value of this test reports the most extreme observation (Williams, 1987). Such noise data can affect the performance of the regression model and therefore, the outlier should have to be removed in the process of training the model (Chen et al., 2015).

3.5.2.6 Check homoscedasticity

One of the most critical assumptions used for the regression analysis is testing the homoscedasticity which means statistically a sequence of random variables. In this way, the test of Studentized Breusch-Pagan is applied for the evaluation of homoscedasticity or otherwise the residual plots technique can also be used (Koenker, 1981). Additionally, the distribution behaviour of residual terms has also been examined for the purpose of analysing the homoscedasticity.

3.5.2.7 Normality of the Residual

Residual analysis has a crucial importance in describing the suitability of the regression model. It estimates the error by calculating the distance between the predicted value and the actual observation. This assumption can be checked by using residual plots; the plots should be organised in a normal curve. Another way for testing the normality of this assumption is Shapiro test which is a statistical approach, and its p-value can decide whether the residuals follow the normal distribution or not.

3.5.3 Accepting / Rejecting hypotheses

A statistical significance or p-value should be specified to accept or reject the null hypotheses which are clearly defined in the introduction chapter. Also in multiple linear regression models, this threshold should be checked for analysing the coefficients of the correlations and MLR model.

3.5.4 Variable Importance

The absolute values of the t-test should have been checked for the purpose of finding the predictors that have a higher level of the influence in the model proposed. The stepwise forward technique for regression model is examined to find the relevant variable for building the model.

3.6 SVM Model

SVM is a supervised machine learning algorithm which can be used for classification or regression. Since this project is about predicting a continuous variable, the international outlook, the regression flavour of SVM will be utilized. In this case, it is referred to as Support Vector Regression (SVR). It is worth noticing that in this work both terms (SVM) and (SVR) are used interchangeably.

The following types of SVR are deployed in this research:

3.6.1 SVR with Linear Kernel

Usually, linear kernels work better if the number of features is large, typically more than the number of observations because the extra complexity resulting from using radial or polynomial kernel is not necessary.

Although this is not the case in this research, because the number of features is much less than the number of observations, SVR with linear kernel will be deployed nevertheless, as the previously mentioned rule is only a rule of thumb and not an established fact.

Two different options for the SVR with linear kernel will be examined:

a) Default Value of the Cost (C) Parameter:

In this option, caret package will be used to train SVR with a default value of the tuning parameter (C), which identifies the cost of violating the margin around the hyperplane used to separate the observations. A smaller value of the cost parameter means a wider margin, and a larger number of support vectors will violate the margin. On the other hand, a larger value of the cost parameter means a narrower margin and a smaller number

of support vectors will violate the margin. In a nutshell, the larger the value of the cost parameter the more the model will try to accurately fit the training data. This doesn't imply that higher values of C are always better, because although higher values of C increases model performance on training data, i.e. decreases the model bias, it also increases its variance when subject to unseen data. All this will be examined in the implementation chapter in detail.

b) Tuned Value of the Cost (C) Parameter:

In this option, caret package will be used to train a SVR with a user defined set of values for the tuning parameter (C). It is expected that by tuning the cost (C), the model can achieve better performance on the data it has been trained on, the training data. But the true test is to achieve the same performance on unseen data, the test data, which will be examined in the implementation chapter.

3.6.2 SVR with Radial Kernel:

Support vector machine with radial basis function (RBF) kernel will also be examined to see if it could outperform the linear SVR or not. In a radial basis function SVM, there are two parameters that control the behaviour of the fit. The cost parameter, and Sigma. Sigma defines how strong the influence of a single training example is. Low values of sigma mean strong influence, and high values mean weak influence. In terms of model fit, the higher the values of sigma, the more accurately the model will fit the training data. Again, this is not always better, because of the bias-variance trade-off.

Again, the same two options will be deployed:

a) Default Values of the Cost Parameters, Cost (C), and Sigma:

In this option, caret package will be used to train an RBF support vector regression model with the default values of the tuning parameters (C) and sigma. This means the fit will be moderately smooth and not trying to be very accurate.

b) Tuned Values of the Cost Parameters, Cost (C), and Sigma:

In this option, a user defined search grid of the tuning parameters C and sigma will be utilized to try to achieve better performance.

3.7 Validation and Evaluation

3.7.1 Split data

The data was divided into two datasets; training (Cross validation method will be applied on this set for resampling data during training and validating the models) and

test dataset (this set will be held as unseen data for evaluating different models). Further, the split was based on the year feature, all observations before 2016 were used for training, and the rest is used for the test.

3.7.1.2 Model Training K-Fold Cross Validations

Throughout this study, k-fold cross validation will be used in the training phase of each model as a resampling method. This technique randomly divides the data set of observations into K folds of almost equal size. It uses the first fold as a validation set, and the method is fit on the remaining K−1 folds. The evaluation metric such as root mean squared error (RMSE) is computed on the remaining observations in the held-out fold. The process is repeated K times; each time a different set of observations will be chosen for validation. The result will be k different values of the metric, RMSE1, RMSE2,..., RMSEK. Then the average will be taken to achieve an overall estimate of the metric. K-fold CV was chosen instead of LOOCV (Leave-one-out CV) for two reasons:

(i) - Computational Efficiency: In LOOCV, it is required to train n models, where n is the number of observations. This is usually very intense, computation wise, especially if n is very large. While in K-fold CV, it only needs to train the model K times.

(ii) - Better Bias-Variance Trade off:

The LOOCV approach leads to a better bias than the K-fold CV, as in LOOCV, almost all the training set observations are used for training the model, which leads to an approximately unbiased estimate of the test error. K-fold CV, on the other hand, leads to an intermediate level of bias since each training set contains (k - 1) n/ k observations, fewer than those in LOOCV approach. Therefore, from the bias reduction perspective, it is clear that LOOCV is to be preferred to K-fold CV. However, bias is not the only source of concern in an estimating procedure; the procedure's variance should also be considered. As compared to K-fold CV, LOOCV has high variance. The reason is that when LOOCV is performed, the average of the outputs of n fitted models is taken, while each of these models is trained on nearly the exact set of observations. Hence, these results are highly correlated with one another. On the other hand, K-fold CV averages the outputs of K fitted models that are somewhat less correlated with each other, since

the overlap between the training sets in each model is smaller. Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does its K-fold CV counterpart(Kohavi,1995).

To summarise, there is a bias-variance trade-off associated with the choice of k in k-fold cross-validation. It has been shown empirically that K-fold cross-validation with k set equal to 5 or 10 gives an estimate of the test error that is characterised neither by high variance nor high bias. In this experiment, k was chosen to be 10.

3.7.3 Evaluation metrics

3.7.3.1 Goodness of the fit Measure

The value of the R-square indicates the proportion of the variance in the dependent variable (international quality) that is predictable from the independent variables in the MLR or the set of features in the SVR.

$$R^2 = \frac{\sum(\grave{y}i - \acute{Y})^2}{\sum(yi - \acute{Y})^2}$$

Formula 3.2: Equation for calculating R^2

3.7.2.2 Root Mean Square Error (RMSE)

There are many different kinds of measures for assessing model accuracy. RMSE is one of the most commonly used methods to estimate how the models perform when predicting unseen data (Willmott & Matsuura, 2005). This metric can be calculated by squaring the mean of squared errors.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\grave{y}t - yt)^2}{n}}$$

Formula 3.3: Equation for calculating RMSE

3.8 Software

All the previous steps either the visualisation or the statistical investigations are conducted by utilising two powerful softwares: SPSS and R. SPSS tool will be used for exploring the data and generating the descriptive analysis while R tool will be used for

finding the correlation between the variables as well as building, training, validating, evaluation and assessing the two families of models proposed in this study.

3.9 Strengths and Weaknesses of the design of the experiment:

3.9.1 Strengths of the Research

1- Adopting two different families of models (regression and SVM) is believed to be invaluable, as SVM is known for its high predictive power, while Regression usually provides interpretability and insights because of the coefficients that are assigned to each predictive feature.

2- In Regression, many models will be deployed, to try to figure out which set of features are significant in predicting the response variable.

4- A repeated pattern throughout the research using 10-fold cross validation in the training process of each model gives a relatively accurate approximation of the true value of the evaluation metrics R^2 and RMSE because the model has been trained and evaluated 10 times, and the average of these 10 evaluations is taken. Also, it is used for achieving optimal values of the tuning parameters of SVM.

3.9.2 Weaknesses of the Design:

1- The number of institutions is not distributed equally or close to equally across countries. Some countries have more than fifty universities, while others have less than five. This might undermine the reliability of coefficients estimates of some countries, and any change in the data would cause a significant change in the model predictions. The research has not investigated this issue carefully to show how the institutions are distributed among countries.

2- For all the models in SVM, the full set of features will be used to predict the response variable. Trying sub-groups of features, as in the regression case, could provide more insights and information about the interaction between each group of features and the dependent variable.

3- Tuning the SVMs for optimal performance only tried very few values of the tuning parameters (Cost, Sigma), due to insufficient computational powers, as well as time constraints.

4- When splitting the data, the test data was all observations in 2016, while train data was all observations before that. Stratified sampling has not been performed to split the data. This could be seen as a weakness from one point of view because it undermines

the predictive power of the models when subject to test data that is significantly different from the train data. On the other hand, it could be seen as a strength, because the objective of training a model is to use it for a prediction on out-of-sample data. In the real world, out-of-sample data is not always a stratified random sample of the training data. So, by doing that, the models are faced with a real challenge, and if they performed well, this could be a true indicator of the model predictive power.

3.10 Conclusion

This chapter has described the overall methodology and the design of the experiment for achieving the research goals. It considered a CRISP-DM methodology for designing the experiment, starting with understanding different kinds of the variables in the dataset and how to prepare them for the modelling phase. Also, as mentioned before, some evaluation metrics for assessing the accuracy level of the models have been selected.

CHAPTER 4: EXPERIMENT AND VALIDATION

4.1 Introduction

This chapter discusses in detail all the steps outlined in the design chapter. It begins with data exploration using descriptive statistics and visualization, then details the steps taken to process the data before modelling. The final phase of this chapter and the most important one is the modelling phase, where two families of models, namely MLR and SVM have been trained and validated using two different validation approaches, 10-fold cross validation and test data validation. These steps are conducted by two pieces of software, SPSS and R.

4.2 Data Exploration

Descriptive statistics are presented below in Table 4.1, international score is less heterogeneous (coef.var=0.49) than research score (coef.var=0.63) and student_staff_ratio (0.66). Percentage of international students along with the number of students are the most dispersed indicators (coef.var>0.8) among those that are present in the data set. Descriptive statistics also show that there is no data entry error because the ranges of all variables are reasonable.

The Figure 4.1 below shows the relationship between each variable and the response variable:

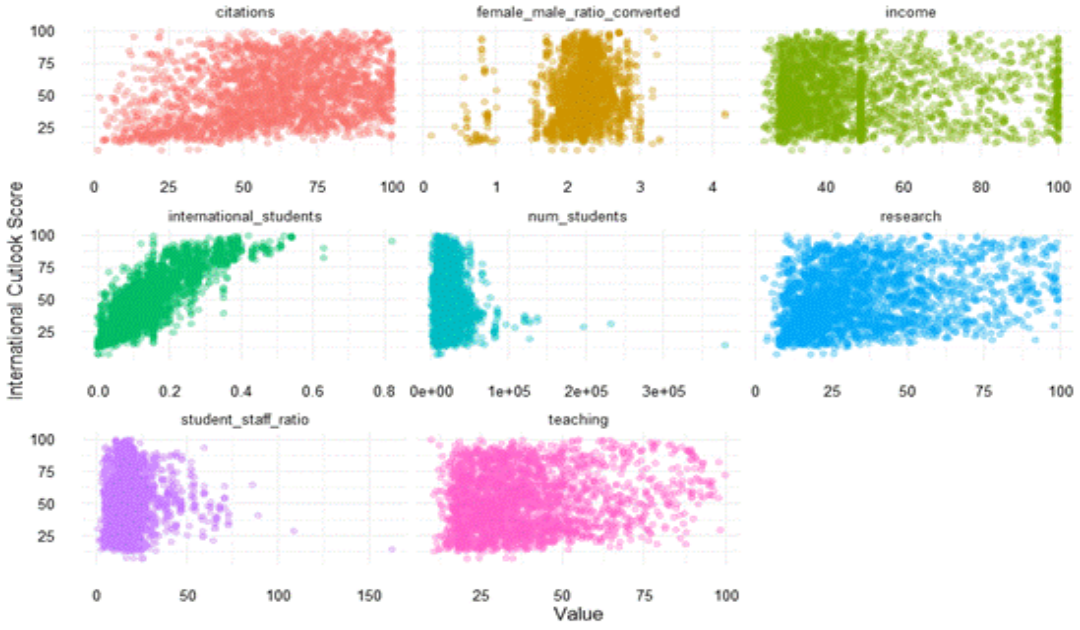


Figure 4. 1: Response variable vs. numerical variables

Table 4. 1: The relationship between variables

	teaching	international	research	citations	income	num_students	student_staff_ratio	num_staff	female_perc	international_students
nbr.val	8603	8603	8603	8603	7763	7793	7793	7793	7736	7790
nbr.null	0	0	0	0	0	0	0	0	0	19
nbr.na	0	0	0	0	37	7	7	7	64	10
Min	9.9	7.1	2.9	1.2	28	462	0.6	28	10	0
Max	95.6	99.9	99	100	100	37923	162.6	11842	78	82
Range	85.7	92.8	96.1	98.8	72	37876	162	11814	68	82
Sum	25267.2	38800.3	22502.9	41047.8	35761.5	19159883	15159.3	1161079.9	37031	10035
Median	27	45.7	22.1	50.3	38.6	20174	16.6	1127	52	10
Mean	31.58	48.5	28.13	51.31	46.87	24161.26	19.12	1464.16	50.31	12.7
SE.mean	0.53	0.84	0.69	0.96	0.74	801.46	0.44	39.51	0.38	0.38
CI.mean.0.95	1.04	1.64	1.36	1.88	1.46	1573.23	0.87	77.55	0.75	0.75
Var	224.56	561.25	381.51	731.46	423.38	509369910	156.8	1237829.88	107.19	113.98
std.dev	14.99	23.69	19.53	27.05	20.58	22569.22	12.52	1112.58	10.35	10.68
coef.var	0.47	0.49	0.69	0.53	0.44	0.93	0.66	0.76	0.21	0.84

It can be noticed that some variables are skewed like num_students and student_staff_ratio. This is an indicator that scaling the data should be considered as an important pre-processing step before building any model.

4.3 DATA PREPARATION

Variables such as world ranking of universities, university name and total_quality were excluded from the analysis, also after 200 top rankings, the ranking was a range (i.e., Universiti Teknologi MARA in Malaysia was ranked as 601-800), as this format was not suitable for the analysis, so they were removed from further analysis.

There are 818 universities listed in the dataset while the numbers of total observation are 8603. As each university is counted once in a year for ranking, this will be a unique variable, so it is not an influential variable.

Additionally, Variable “english_fluent” contained some text observations such as “0Autonomous University of Madrid" apart from 0 and 1. As it was a categorical variable, the text was parsed for the digits and kept the first digit as an observation. So the observation “0Autonomous University of Madrid" was reduced to 0.

Special characters were removed such as “%” from the variable “international_students” and converted it back to ratios. Convert variables include "international", "income" and 'total_quality' from factors to numeric variable. Year variable was converted to factor, as each year will have an individual effect on the dependent variable. Total_quality was removed as it contained more than 50% of NAs observation. Other variables had less than 10% of missing values for the target variable “international outlook”. Missing values were replaced by the mean value (this was done for continuous variables).

After reprocessing the data contained 12 variables in total. There were two categorical variables, and rest were numerical variables. In the following part, some key observations from visualizing the variables are mentioned below:

International: This is the dependent variable. The histogram shows the distribution of international outlook. It can be observed that most of the observations are clustered between 25 and 70 with little observation towards 0. The distribution in Figure 4.2 shows how it closes to normal and does not appear to have much skewness

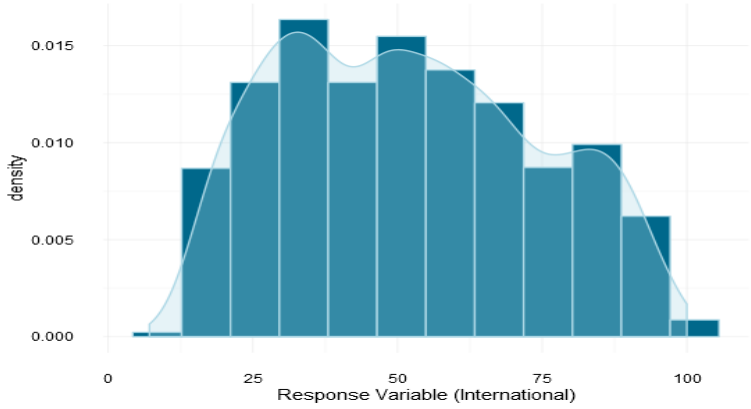


Figure 4. 2: Histogram distribution of international outlook

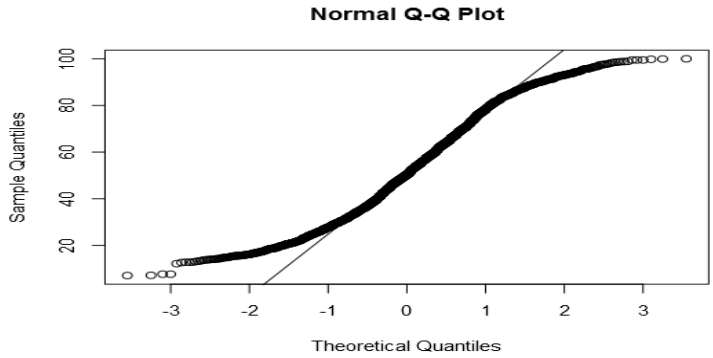


Figure 4. 3: Q-Qnormal plot

Figure 4.3 above is used to check the normality of the data which confirms that the distribution of the international variable is normal with a slight deviation at the end. Also, Shapiro-Wilk was conducted for testing the dependent variable. The test shows that W = 0.96871, p-value < 2.2e-16.

Table 4. 2: Summary descriptive for international variable

Variable name	**Minimum**	**First Quartile**	**Median**	**Mean**	**Third Quartile**	**Maximum**
International	7	33	50	52	69	100

Correlations: There are two groups of features will be used for building the model. International and institution variable: "English_fluent", "teaching", "research", "citations" and "income" variables have been grouped in institution specific variable, and the relation between them has been analysed. Pearson correlation coefficient has been shown in Figure 4.4 below along with the scatter plot between the variables.

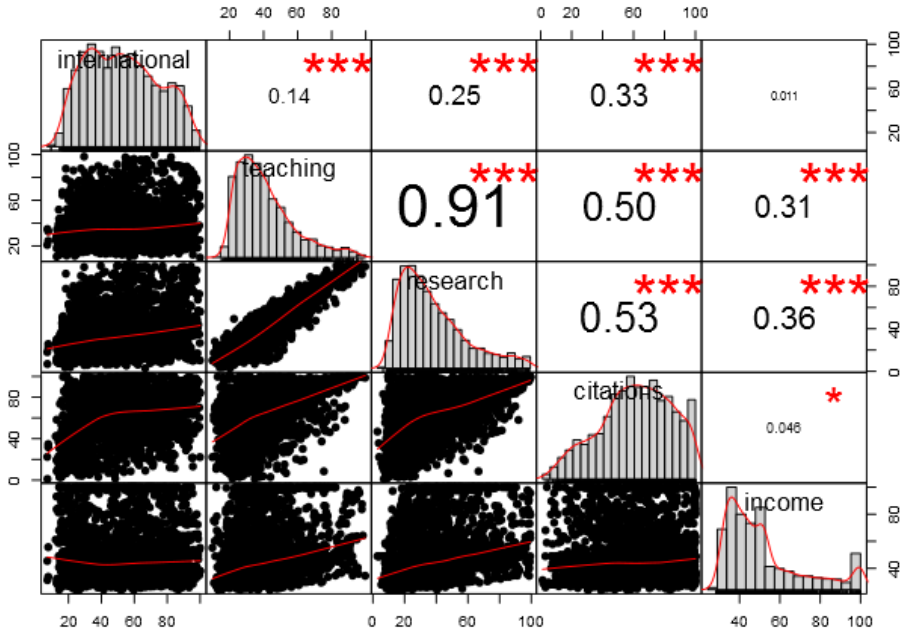


Figure 4. 4: Scatter plot, histogram and correlation plot between different variables.

It can be observed that international variable is not strongly correlated (weak positive correlation) with other institution specific variables. However, teaching and research index are strongly correlated. So it can be said that in the institutions where there is strong research, teaching score is also strong.

International and student specific variables: the interdependencies were explored between international and student specific variables. Student specific variable contained "num_students", "student_staff_ratio", "international_students" and "female_male_ratio_converted". Figure 4.5 shows the scatter plot, histogram and correlation plot between different variables.

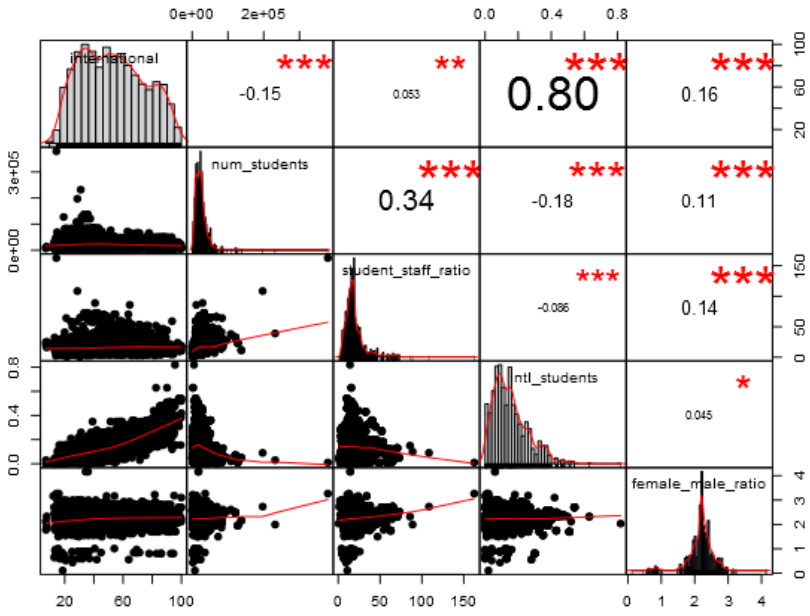


Figure 4. 5: Scatter plot, histogram and correlation plot between international and student specific variables

It can be observed that while the variable "international_students" is strongly correlated with international outlook variable. All other variables have a correlation with the response variable.

4.4 Modelling

In this stage, statistical models were created to predict the value of the international variable using Multivariate Regression and Support Vector Machine Learning (SVM).

4.4.1 Regression Analysis

4.4.1.1 Baseline Model

The mean of international outlook from training data is used as a baseline model prediction. The coefficients of determination (R- square) and the root mean square error have been computed. A value of 0 of R^2 has been observed, suggesting that the baseline model does not explain the variance in the response variable (international outlook score). Similar observations were obtained for test dataset.

Table 4. 3: R-square and RMSE of train dataset using baseline model

Dataset	Train	Test
R-square	0	-0.046
RMSE	21	24

4.4.1.2 The Institutional Model

Multivariate regression was conducted to determine the relation between institution outlook score and all other institution specific variables. The variables teaching, research, citations and income, were included in order to determine the effect of institution related features on institutional outlook score without being affected by the other variables. This was done to eliminate the influence of the other variables to the features related to institution only. The Figure 4.6 below shows all the residual plots to demonstrate the validity of the model.

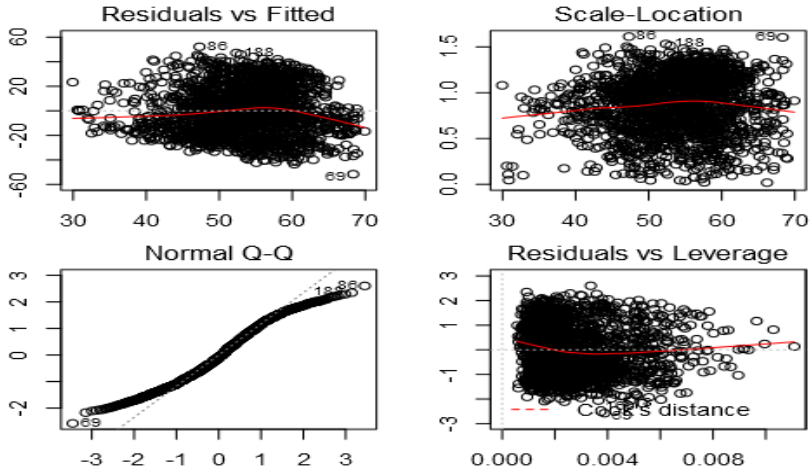


Figure 4. 6: Residual plot of regression model for institutional outlook score and institutional specific variables

Some key observations from above shown residual plots are:

- The first plot on the top left shows the distribution of residuals around zero mean. It also shows if there is any heteroscedasticity in the data. As there is no specific pattern in residual vs. fitted plot, it can be said that there is no heteroscedasticity in the fitted model.

- The second plot on top right shows the QQ plot of the standardized residuals. As regression model assumes that the residual is normally distributed, this plot can be used to check the assumption of the regression model. In this case, residuals deviate from normal towards the end of the curve, which is similar to the distribution of international outlook variable.
- The bottom third plot again shows that there is no heteroscedasticity and bottom right plot can be used to see if there are high leverage points.

The table 4.4 below shows the strength of linear association in relation to explaining the ability of the institution outlook score using features related to institution only. Specifically, the squared value of R (.102) and its adjusted form (.1), measures the percentage of total variation of institution outlook score explained by teaching, research, citations and income. This implies that approximately 10% of the variability of institution outlook score is explained by the features selected.

Table 4. 4. Model Summary for Linear Regression using features related to institution only

MODEL SUMMARY	
Multiple R-Squared	0.102
Adjusted R-Squared	0.1
Residual Standard Error (RMSE)	20.1
F-Statistic	50.8
P-Value	<2e-16

The Figure 4.7 below shows the distribution of RMSE and R-squared across the 10 validation sets that cross validation uses to validate each trained model. The full results of applying 10-fold cross validation on the training data for the institutional model is presented in the appendix.

Figure 4. 7: CV Distribution of RMSE and R-squared for the Institutional Model.

These results show that this model almost fails to explain the variation in the response variable based on the institutional attributes only. To see this more clearly, a scatter plot between the response variable and each of the predictors has been generated below using the training data only, and it shows that there's no clear linear trend between the international outlook of an institution and any of its attributes.

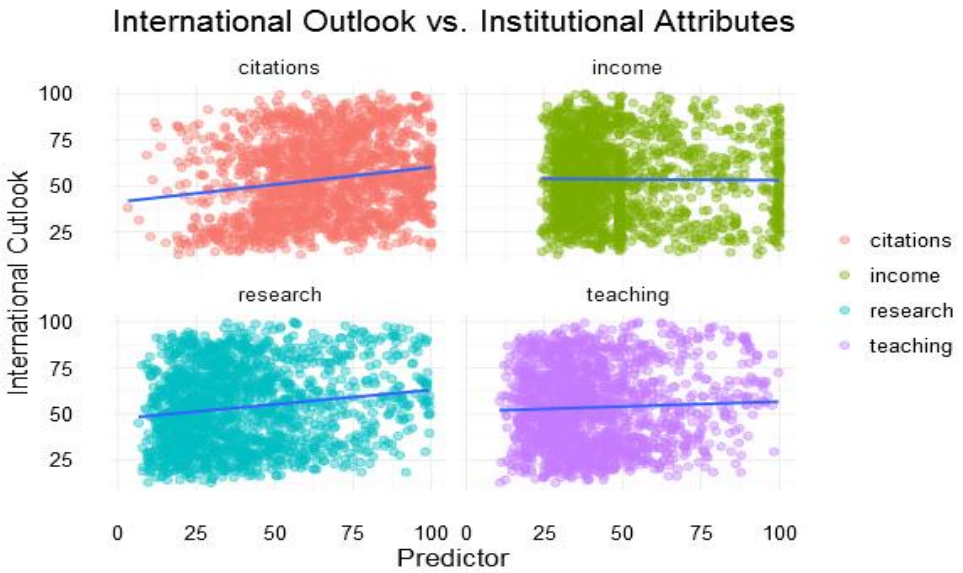


Figure 4. 8: Scatter Plot between the Response Variable and the Institutional Features

Next, the model has been applied to the test data to see how well it performs on out-of-sample data. The results are RMSE = 20.45, almost the same as the CV estimate and R^2 = 26%, which is more than twice that of the CV.

The table 4.5 below shows the model created using features related to an institution with their respective significance statistics. The variables teaching, research and citations are significant at alpha = .05. However, the variable income is only significant at alpha =.10.

Table 4. 5. Coefficients of the model using the significant features related to institution

Variables	Coefficient	Standard Error	T-Value	P-value
(Intercept)	53.571	0.474	112.94	2.00E-16
Teaching	-12.762	1.151	-11.09	2.00E-16
Research	13.754	1.166	11.79	2.00E-16
Citations	3.365	0.542	6.21	6.50E-10
Income	-0.955	0.513	-1.86	0.063

Interpreting the coefficient in the model above:

The intercept suggests that on the average, holding every other variable constant, the predicted value of the institution outlook score is 53.571. Based on the results teaching has a negative impact on the institution outlook score. A one unit increase in teaching causes the institution outlook score to decrease by 12.762.

This seems counter intuitive, and it maybe because of the high collinearity between teaching and research. This issue will be further investigated in this chapter when a model based on the full set of features is developed.

On the flip side, research has a positive effect on the institution outlook score. As the research variable increases by 1 unit, the institution outlook score increases by 13.754. Citations also have a positive effect on the institution outlook score. A one unit

increase in the citation variable causes the institution outlook score to go up by 3.365. Lastly, the variable income negatively affects the institution outlook score. As income increases by 1 unit, the institution outlook score decreases by .955.

4.4.1.3 The Students Model

Same with the previous section, a multivariate regression was conducted to create a model that can predict the value of the institution outlook score. However, instead of including features related to institutions only, features related to students were used in its place. The variables: number of students, student staff ratio, international students, and the converted female to male ratio were included in order to determine the effect of student related features on institutional outlook score without being affected by the other variables. This was done to eliminate the influence of the other variables to the features related to students only. The graph in Figure 4.10 below shows all the residual plots to demonstrate the validity of the model.

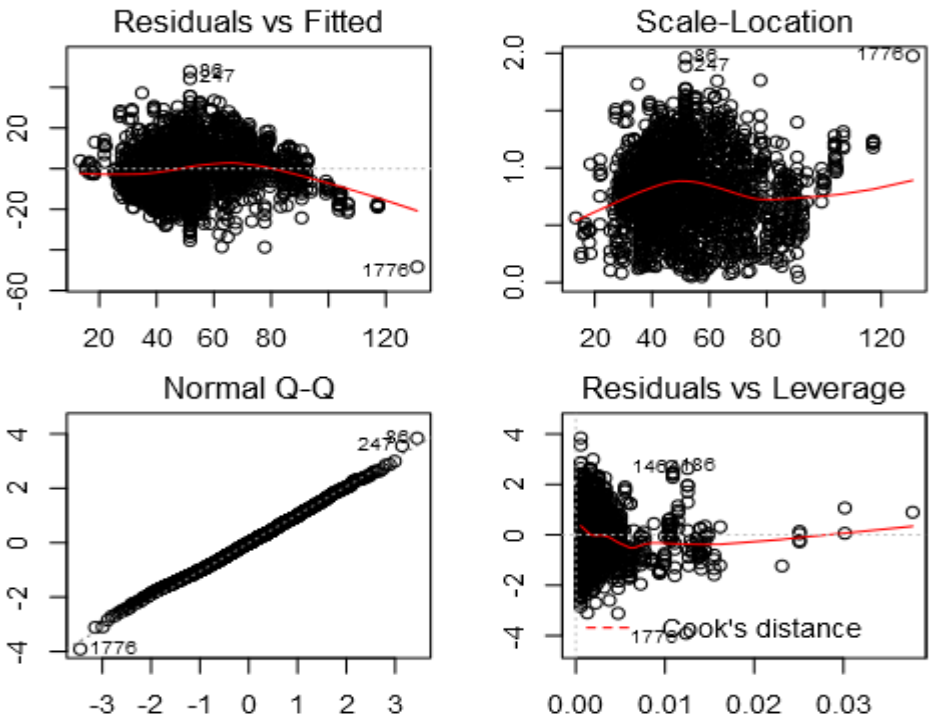


Figure 4. 9: Residual plots

Some key observations from above shown residual plots are:

- The first plot on the top left shows the distribution of the residuals is around zero mean with only a slight deviation on the right side. A pattern cannot be

identified in the residual vs. fitted plot. Therefore, the plot suggests that there is no heteroscedasticity in the model.

- The second plot on the top right shows the QQ plot of the standardized residuals. The QQ plot follows the diagonal line implying that the normality assumption of multivariate regression has been satisfied.
- The bottom third plot again shows that there is no heteroscedasticity and bottom right plot shows that there are no outliers in the model.

The Table 4.6 below shows how much of the variation of the institution outlook score is explained by the independent variables (number of students, student-staff ratio, international students and the converted ratio of male and female). The features related to students only can explain 65.5% of the variation of the institution outlook score

Table 4. 6: Summary model used student Features only

MODEL SUMMARY	
Multiple R-Squared	0.655
Adjusted R-Squared	0.655
Residual Standard Error	12.26
F-Statistic	850
P-Value	<2e-16

The Figure 4. 10 below shows the distribution of RMSE and R-squared resulted from training the model using 10-fold cross validation. The full results of the cross validation procedure can be seen in the appendix:

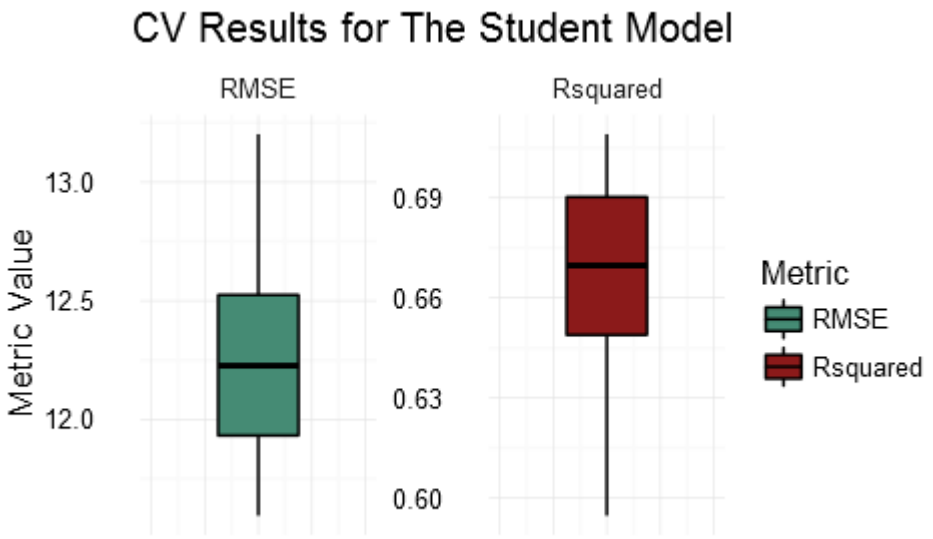


Figure 4. 10: CV Distribution of RMSE and R-squared for the Student Model

From the results summarized above, it can be noticed that this model succeeds in explaining much of the variation in the response variable based on student specific features only. This is more evident in the following plot:

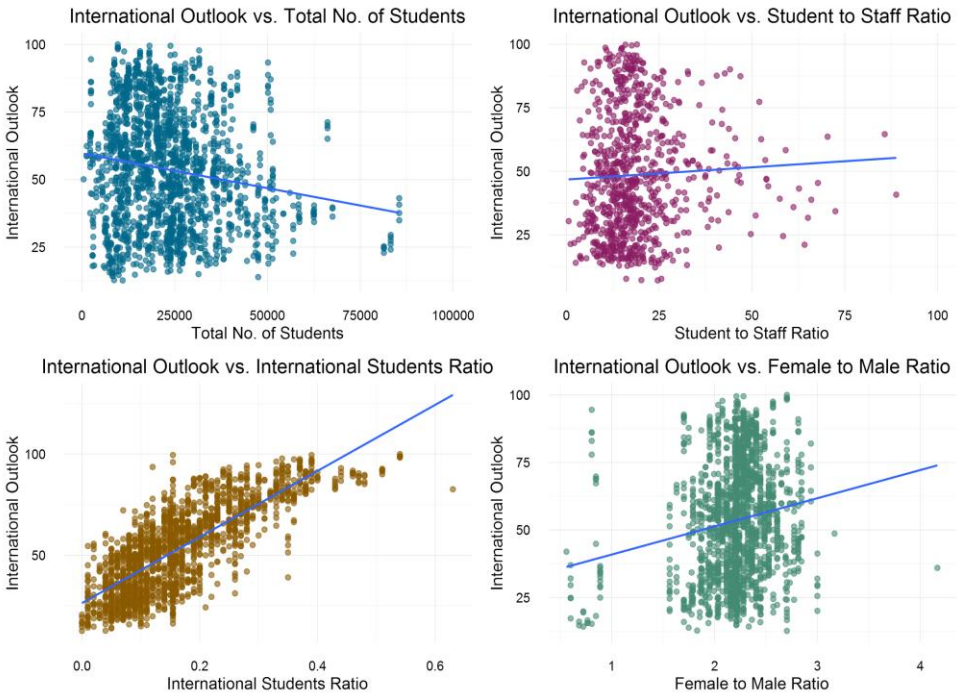


Figure 4. 11: Scatter Plot between the Response Variable and Student Features

The plot shows a strong linear relationship between the international student's ratio and the response variable, which explains the improvement of this model performance compared to the Institutional Model. As for the other three features, they don't seem to have a strong linear relationship with the response.

Next, the model has applied to the test data to assess its performance on out-of-sample data; the results are RMSE = 13.22, almost the same as the CV estimate and R^2 = 69.3%, a 3% increase over the CV estimate

The Table 4.7 below shows the model created using features related to students only with their respective significance statistics. All of the variables (number of students, student-staff ratio, international students and female-male ratio) have p-values that are less than .05. This implies that at .05 level of significance, all of the variables included in the model are significant.

Table 4. 7: Model Coefficients using features related to students only

Variables	**Coefficient**	**Standard Error**	**T-Value**	**P-value**
(Intercept)	53.571	0.294	182.29	< 2e-16
num_students	-1.287	0.31	-4.16	3.40E-05
student_staff_ratio	3.249	0.306	10.63	< 2e-16
international_students	16.55	0.301	54.96	< 2e-16
female_male_ratio_converted	2.585	0.298	8.67	< 2e-16

Interpreting the coefficients in the model above:

The intercept suggests that on the average, holding every other variable constant, the predicted value of the institution outlook score is 53.571. The number of students has a negative effect on the institution outlook score as shown in the model. A one unit increase in the number of students causes a 1.287 decrease in the value of institution outlook score. The rest of the variable included in the model have a positive effect on the institution outlook score. Student-staff ratio increases the value of the institution

outlook score by 3.249 for every 1 unit increase. The international students variable's increase per 1 unit (i.e. 1 percent) causes the institution outlook score to increase by 16.55. Lastly, the value of the institution outlook score increases by 2.585 for every 1 unit increase in the converted female-male ratio.

4.4.1.4 The Country Model

Another multivariate regression using features related to the institution only was conducted to create a model that can predict the value of the institution outlook score. In this section, however, the variable country was added to the list of features in order to take into account the effect of geography on the institution outlook score. The graph below shows all the residual plots to demonstrate the validity of the model.

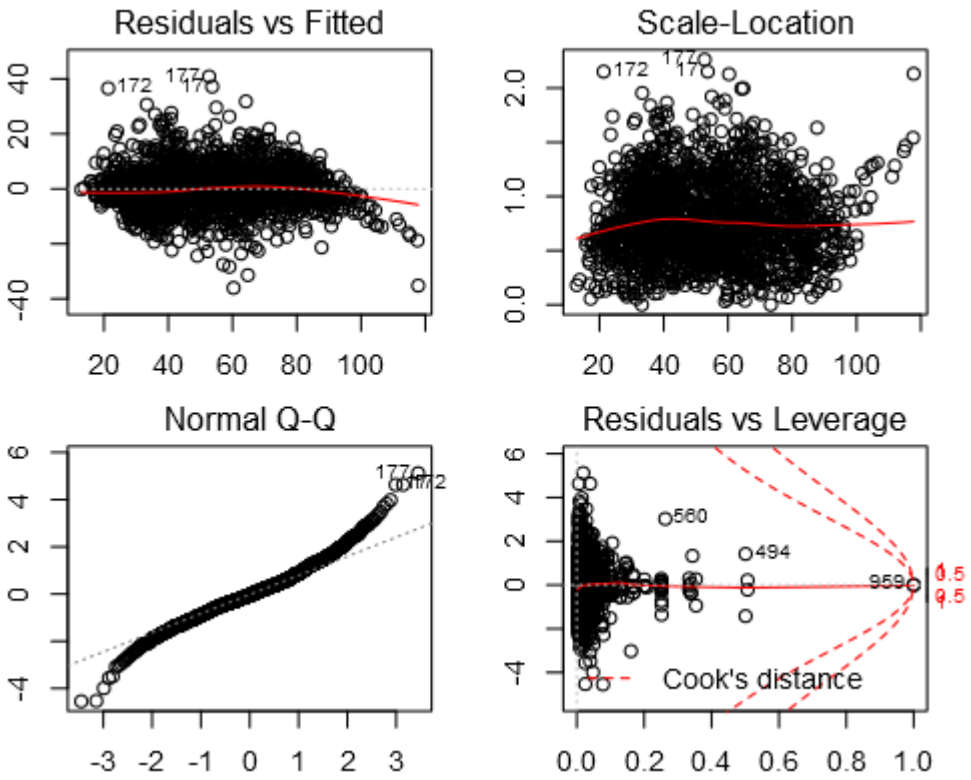


Figure 4. 12. Model used institution features with respect to the locations of the universities

Some key observations from above shown residual plots are:

- The first plot on the top left shows that the data points revolve around zero. The points are distributed in a random manner and since no pattern cannot be identified in our plot, this suggests that there is no heteroscedasticity in

the model. The second plot on top right also shows that there is no heteroscedasticity.

- The first plot on the bottom left shows the QQ plot of the standardized residuals. Although it slightly deviates on the ends, it can be seen that most of the data points follow the diagonal line in the QQ plot implying that the normality assumption of multivariate regression has been satisfied.

The Table 4.8 below shows how much of the variation of the institution outlook score is explained by the independent variables. The features related to geography and institution only can explain 77.2% of the variation of the institution outlook score.

Table 4. 8. Model summary of using features related to institution and locations

MODEL SUMMARY	
Multiple R-Squared	0.775
Adjusted R-Squared	0.768
Residual Standard Error	10.2
F-Statistic	125
P-Value	<2e-16

The box plots below in Figure 4.13 shows the distribution of evaluation metrics resulted from training and validating the model using 10-fold cross validation. Remarkable improvement is noticed, from the results presented above, in predictive power and ability to explain the variation in the response variable compared to the first model (The Institutional Model). R-squared has gone from 10% to 77.5%, and the RMSE has decreased to 10.2, instead of 20. All this improvement has been achieved by adding only one feature to the model, which is country.

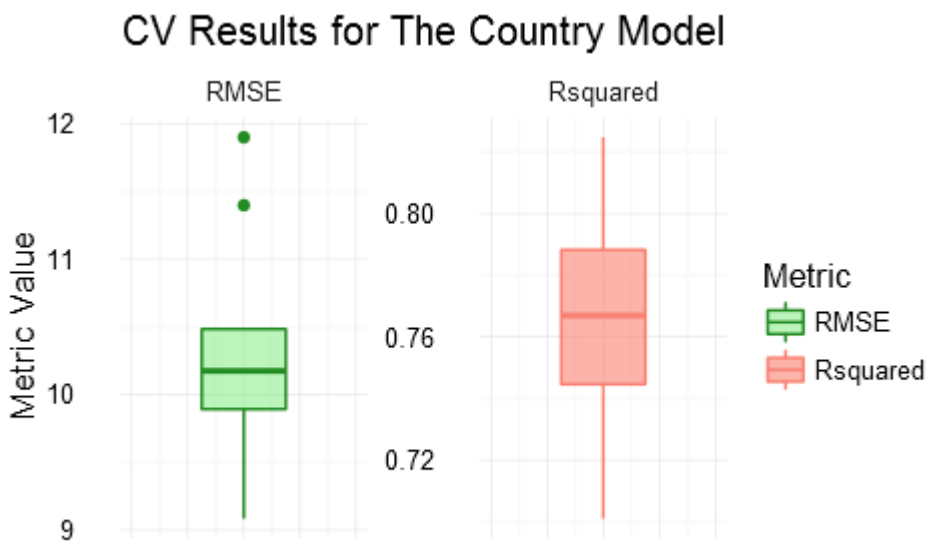


Figure 4. 13. CV Distribution of RMSE and R-squared for the Country Model

The Model then was tested on the test data and results are RMSE = 18.63, almost twice that of the training R^2 = 50%, which is 27% less than its training data counterpart. A significant reduction in performance on the test data is noticed, compared to the training data, which is a strong indicator of a high level of bias. More analysis and insights will be provided to explain this in the next chapter.

The Table 4.9 below shows the model created using features related to geography and institution only with their respective significance statistics. All of the variables included have p-values that are less than .05 except for China, Egypt, Mexico, Morocco, and Thailand. This implies that at .05 level of confidence, all of the variables included in the model are significant except for the mentioned countries. Even though China, Egypt, Mexico, Morocco, and Thailand are insignificant, they were still retained in the model because they are part of one categorical variable: country.

Table 4. 9: Coefficients of the model using institutions and location

Feature	Estimate	Std..Error	t.value	P value
(Intercept)	53.57134894	0.2406411673	222.6192199	0
Teaching	1.903546551	0.6049163375	3.14679309	0.00167859622

Research	1.487258375	0.5929997519	2.508025291	0.01223092905
Citations	2.964355846	0.3129126539	9.473429115	8.47E-21k2
Income	1.036966611	0.3024522026	3.428530532	6.21E-04
Australia	9.377514162	0.275348507	34.05689126	2.33E-195
Austria	4.925502696	0.2485018879	19.82078582	5.09E-79
Belgium	2.934593022	0.2528456371	11.60626324	4.71E-30
Brazil	-0.7315371917	0.2484497135	-2.944407468	0.003278532739
Canada	5.364622262	0.2644154824	20.28860872	2.44E-82
Chile	0.6131457518	0.24241658	2.529306171	0.01151625361
China	-0.3114317927	0.2599437654	-1.198073715	0.2310510176
Colombia	0.9319247105	0.2435459187	3.826484614	1.35E-04
Czech.Republic	1.046559889	0.2433633584	4.300400423	1.80E-05
Denmark	3.583708253	0.2501965913	14.32356946	4.44E-44
Egypt	-0.2793156048	0.2419250666	-1.154554212	0.2484311108
Estonia	0.5970040279	0.2435218441	2.451541997	0.0143219942
Finland	1.346711662	0.248430086	5.420887958	6.76E-08
France	4.098446175	0.2512998795	16.30898584	9.15E-56
Germany	4.444142069	0.2631296804	16.88955067	2.16E-59
Greece	0.684313304	0.2436174612	2.808966568	0.005025460454
Hong.Kong	4.252326304	0.2479567136	17.14947033	4.79E-61
Iceland	1.220779481	0.251139082	4.860969752	1.27E-06

India	-1.220856563	0.2460564563	-4.961692864	7.67E-07
Iran	-0.830813198	0.2461941671	-3.37462584	7.55E-04
Israel	1.26245228	0.248503425	5.080220848	4.17E-07
Italy	1.070699536	0.2610747747	4.101122132	4.30E-05
Japan	-1.768424248	0.2613114054	-6.767497366	1.78E-11
Macau	0.9824908733	0.2416029134	4.066552259	4.98E-05
Mexico	0.09463240037	0.2416474915	0.3916134191	0.695391687
Morocco	0.3576867593	0.2423100957	1.476152936	0.140083262
Netherlands	3.789902617	0.2687305857	14.10298202	7.47E-43
New.Zealand	5.988676714	0.2555972922	23.43012582	8.80E-106
Norway	2.778270511	0.24582709	11.30172639	1.24E-28
Poland	0.6122358466	0.2437883754	2.511341427	0.01211703863
Portugal	1.441115291	0.2469501729	5.835652083	6.37E-09
Republic.of.Ireland	4.842408228	0.2473907902	19.5739228	2.75E-77
Russian.Federation	0.8008902391	0.2452319239	3.265848208	0.001112522914
Saudi.Arabia	1.745157009	0.2429058548	7.18449957	9.95E-13
Singapore	3.903543378	0.2450827303	15.92745182	1.97E-53
South.Africa	2.483732292	0.2513245527	9.882569234	1.90E-22
South.Korea	-0.7797080364	0.2558119144	-3.047973892	0.002338489364
Spain	1.155757602	0.2519206834	4.587783689	4.80E-06
Sweden	3.501942455	0.2594888219	13.49554262	1.49E-39

Switzerland	6.999390375	0.2470936695	28.32687049	1.50E-145
Taiwan	-1.207006568	0.2632186135	-4.585566926	4.85E-06
Thailand	-0.04272569048	0.2441497046	-0.1749979201	0.8611015698
Turkey	0.3968224444	0.247250832	1.604938763	0.1086882298
United.Kingdom	13.30145505	0.28464913	46.72930162	7.94E-310

Interpreting the coefficients in the model above:

The intercept suggest that on the average, holding every other variable constant, the predicted value of the institution outlook score is 53.5713. A 1 unit increase in the university score for teaching leads to 1.9 increase in institution outlook score. As the university score for research increases by 1 unit, the institution outlook score increases by 1.5. Citation and income variable also has positive effect on institution outlook score. This means that as the university score for citation increases by 1 unit, institution outlook score increases by 2.96. As income increases by 1 unit, the institution outlook score increases by 1.03. Finally, it is noticeable that the categorical variable country has the most influential effect on institution outlook score.

4.6.1.5 The Full Model

A model was then created to define the relationship between institution outlook score and all other variables in the data set. Since the other sections already investigated the individual effects of features related to institution and student, this section used all of the variables available in order to see how the interaction of institution and student affects the institution outlook score. The graph below shows all the residual plots to demonstrate the validity of the model.

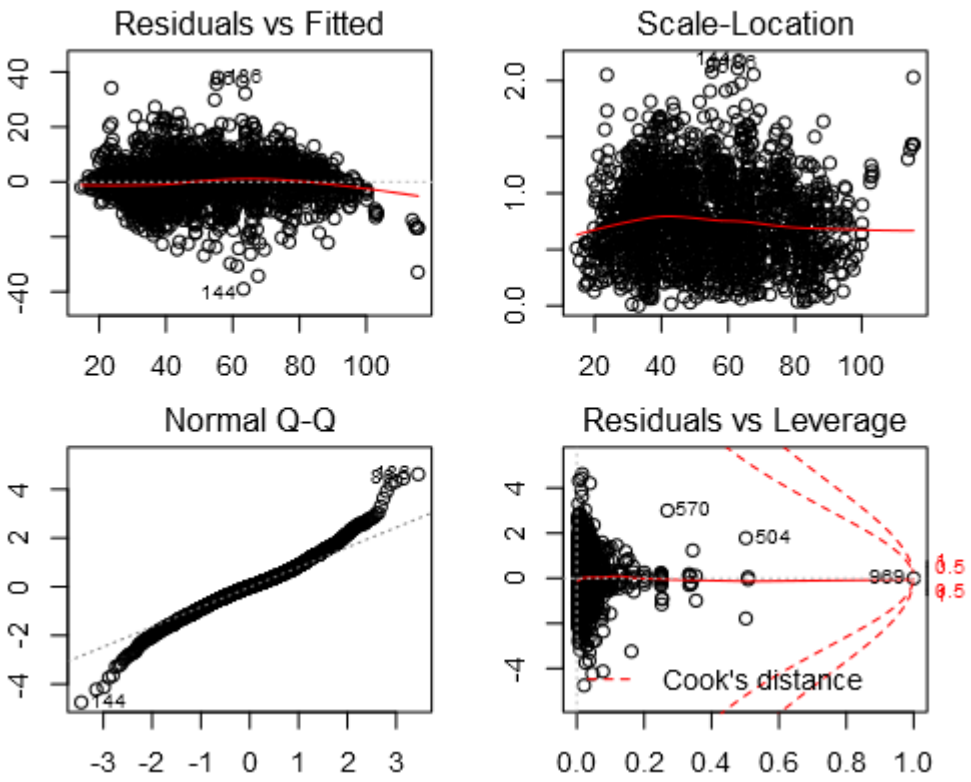


Figure 4. 14: Model Diagnostic Plots using full dataset

Some key observations from above shown residual plots are:

- The first plot on the top left shows that the data points revolve around zero but there a deviation exists on the right most part of the plot. Even though a slight deviation is present, the points are distributed in a random manner, and since no pattern cannot be identified in our plot, this suggests that there is no heteroscedasticity in the model. The second plot on top right also shows that there is no heteroscedasticity.
- The first plot on the bottom left shows the QQ plot of the standardized residuals. Although it slightly deviates on the ends, it can be seen that most of the data points follow the diagonal line in the QQ plot implying that the normality assumption of multivariate regression has been satisfied.

The Bonferroni Outlier test was used to check if there are any potential outliers and influential variables. The Bonferroni outlier test tests the null hypothesis that an observation is not an outlier. The Bonferroni Outlier test p-value is less than 0.05, this means that observation 86 is an outlier. Observation 86 is removed in the next analysis. After removal, the Bonferroni outlier test was checked again to see if there are any more

outlier. According to the result, there are no more Studentized residuals with Bonferroni-p that is less than .05

Table 4. 10: Bonferroni results

Observation #	R- Student	Bonferroni P
86	4.9	.003

The assumption of non-collinearity was also checked. The variance inflation factor (VIF) was computed to identify the severity of multicollinearity in the full model. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. The table below 4.11 shows that multicollinearity is present in the model caused by the teaching and research variable. The next section explains how the violation of non-collinearity was corrected. The standard assumption in linear regression is that the theoretical residuals are independent and normally distributed. The plot in Figure 4.15 below shows the distribution of the normal residuals of the model using student features. Notice that most of the data points revolve around zero and the histogram shows a bell-shaped distribution. From here, it can be concluded that the student residuals are approximately normal. Thus, it can be concluded that the assumption of normality for the full model has been satisfied.

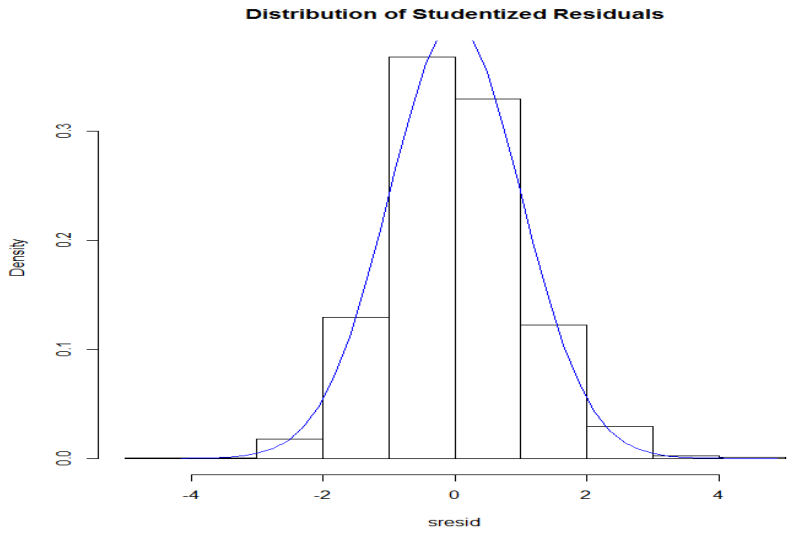


Figure 4. 15: The distribution of the normal residuals of the model

Table 4. 11: Multicollinearity results

Variables	**VIF**	**VIF>2**
english_fluent	1.2	FALSE
Teaching	6.3	TRUE
Research	6.4	TRUE
Citations	1.4	FALSE
Income	1.3	FALSE
num_students	1.2	FALSE
student_staff ratio	1.2	FALSE
international_students	1.3	FALSE
Year	1.1	FALSE
female_male_ratio_converted	1.1	FALSE

The Table 4.12 below shows the summary of the model using full features; it presents how much of the variation of the institution outlook score is explained by the independent variables, it explains 85% of the variation of the institution outlook score.

Table 4. 12: summary of the full model using full features

MODEL SUMMARY	
Multiple R-Squared	0.858
Adjusted R-Squared	0.854
Residual Standard Error	8.08
F-Statistic	195
P-Value	<2e-16

The following Figure 4.15 is for the RMSE and R-squared, distributed across the 10 validation folds:

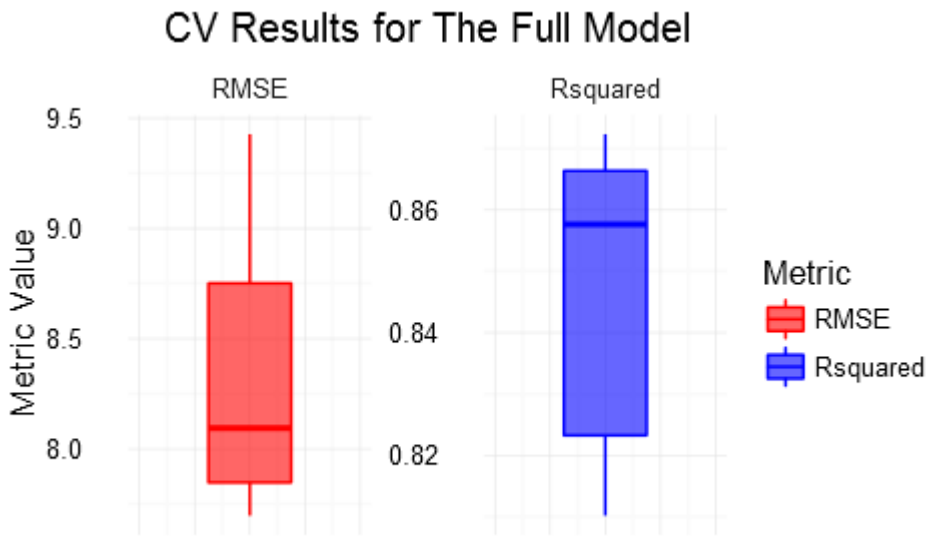


Figure 4. 16: CV Distribution of RMSE and R-squared for the Full Model

The results of the full model are the best so far, with R-squared indicating that almost 86% of the response variance can be explained by the full set of features.

Again, testing on out-of-sample data gives RMSE = 13.69, significantly more than RMSE on training, which is 8 and R^2 = 76.87%, which is less than what it is on the training set, but still by far the best model performed on the test set.

4.6.1.6 The Reduced Model:

The stepwise selection procedure was then utilized to see if the same accuracy achieved in the last model could also be achieved using a smaller set of features. The rationale behind this is that a simple model, holding everything else equal, is better than a complex one. This is because complex models tend to overfit.

Applying the stepwise selection procedure produced the result presented in the following Figure 4.17:

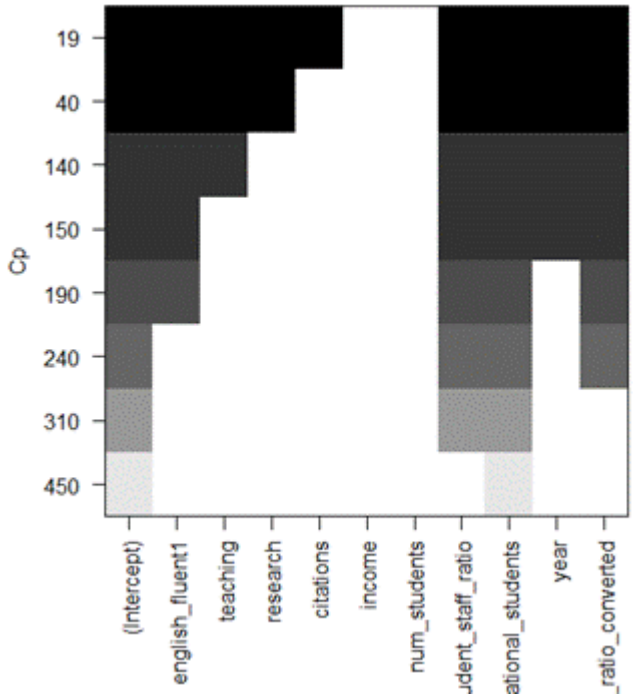


Figure 4. 17: Features important using Stepwise Forward Method

From the plot, it can be seen that the lowest Cp score or best R2 comes with following variables: english_fluent, teaching, research, citations, num_students, student_staff_ratio, international_students, year. Furthermore, it is noticed that teaching and research are correlated, so the research variable has been removed in order to correct the problem of multicollinearity.

The reduced model was created based on the variables retained in the stepwise selection procedure. The residual plot was again examined in order to make sure that the assumption of homoscedasticity is not violated. The plot in Figure 4.17 below shows the residual plot for the reduced model.

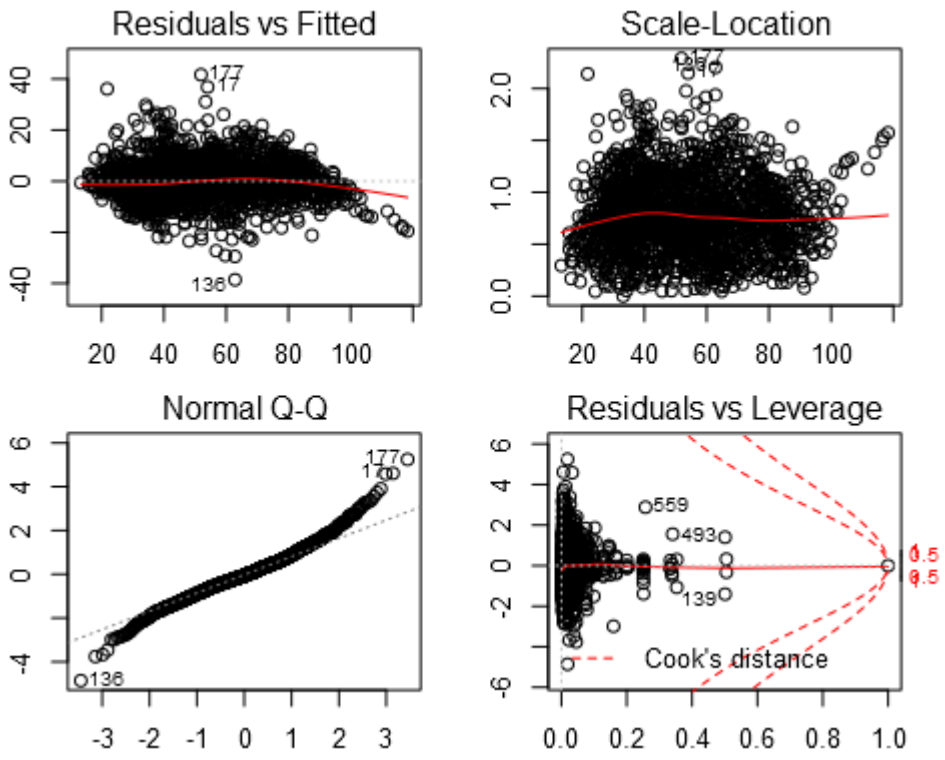


Figure 4. 18: The residual plot for the reduced model.

Some key observations from above shown residual plots are:

- The plots for the reduced model is almost the same with the full model. The first plot on the top left shows a deviation exists on the right most part of the plot but the points are distributed in a random manner, and since no pattern cannot be identified in our plot, this suggests that there is no heteroscedasticity in the model. Also, the plot shows that the residuals are distributed about the zero mean. The second plot on top right also shows that there is no heteroscedasticity.
- The first plot on the bottom left shows the QQ plot of the standardized residuals. Although it slightly deviates on the ends, it can be seen that most of the data points follow the diagonal line in the QQ plot implying that the normality assumption of multivariate regression has been satisfied.

The Bonferroni Outlier test was again used to check if there are any potential outliers and influential variables. The Bonferroni Outlier test p-value is less than .05, this means that observation 237 and 1766 is an outlier. These observations were removed in the next analysis. After removal, the Bonferroni outlier test was checked again to see

if there are any more outlier. According to the result, there are no more Studentized residuals with Bonferroni-p that is less than .05

Table 4. 13: Bonferonni reports the outliers

Observation #	R- Student	Bonferroni P
237	4.4	0.025
1766	-4.2	0.043

The assumption of non-collinearity was again checked. The variance inflation factor (VIF) was computed to identify the severity of multicollinearity in the reduced model. The Table 4.14 below shows that multicollinearity is not present in the model. Thus, verifying the assumption of non-collinearity in the reduced model is satisfied.

Table 4. 14: VIF results to check Multicollinearity

Variables	**Variance Inflation Factor(VIF)**	**VIF>2**
english_fluent	1.2	FALSE
Teaching	1.5	FALSE
Citations	1.4	FALSE
num_students	1.2	FALSE
student_staff ratio	1.2	FALSE
international_students	1.2	FALSE
Year	1.1	FALSE

The assumption of the residuals normality was again checked. The plot below shows the distribution of the student residuals. Notice that most of the data points revolve around zero and the histogram shows a bell-shaped distribution. From here, it

can be concluded that the student residuals are approximately normal. Thus, it can be concluded that the assumption of normality for the reduced model has been satisfied.

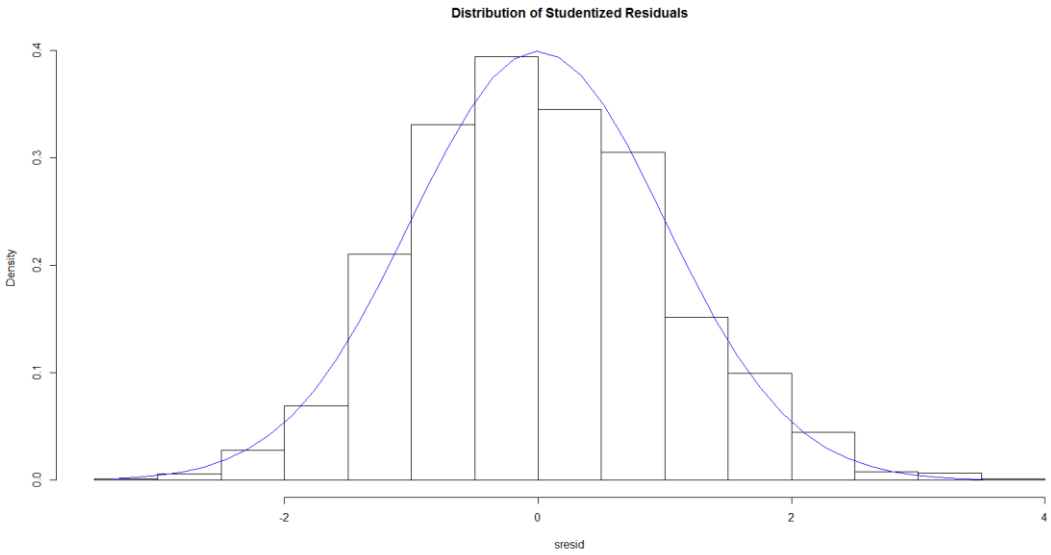


Figure 4. 19: The residuals normality distribution used student's features

The Table 4.15 below shows the summary of the reduced/final model. This shows how much of the variation of the institution outlook score is explained by the features retained by the stepwise selection procedure (english_fluent, teaching, citations, number of students, student-staff ratio, international students and year). The features used can explain 85.5% of the variation of the institution outlook score.

Table 4. 15: Summary of the reduced/final model

MODEL SUMMARY	
Multiple R-Squared	0.86
Adjusted R-Squared	0.855
Residual Standard Error	8.03
F-Statistic	209
P-Value	<2e-16

Figure 4. 20: CV Distribution of RMSE and R-squared for the Full Model

The results presented in the above Table 4.15 and Figure 4.20 show that almost the same level of predictive power could be achieved with fewer features than the Full Model. The advantage of this is that simpler models, holding everything else constant, tend to perform better on out-of-sample data, compared to more complex ones.

The last step is to test on the test data; the results are RMSE = 13, less than the RMSE of the Full Model (13.69) and R^2 = 77.54%, which is also slightly better than the Full Model.

So, this would be the model of choice among all regression models developed so far, and in the next sections, it will be compared to the other machine learning model discussed in this research, namely Support Vector Machines.

The Table 4.17 below shows the final reduced model created using features retained by the stepwise selection procedure with their respective significance statistics. All of the variables except for English fluency and a few countries have p-values that are less than .05. This implies that at .05 level of confidence, all of the variables included in the model, except the above mentioned ones, are significant.

Table 4. 16: Coefficients of the model created using features retained by the stepwise selection

Term	estimate	Std error	statistic	P value
(Intercept)	53.50580681	0.1938556311	276.0085251	0
Australia	5.785762601	0.2565685222	22.55055512	5.37E-99
Austria	3.277922396	0.2729497423	12.00925258	5.62E-32
Belgium	2.158054579	0.3013830726	7.160503608	1.18E-12
Brazil	9.47E-04	0.2307287186	0.004103506435	0.9967263551
Canada	4.667065074	0.2156081062	21.64605569	3.32E-92
Chile	0.6913480239	0.2030895294	3.404153951	6.79E-04
China	0.7706849212	0.3357200882	2.295617535	0.02181620219
Colombia	1.159847004	0.1975245643	5.871912732	5.15E-09
Czech.Republic	0.8777515027	0.2119869469	4.140592218	3.63E-05
Denmark	3.371740428	0.2771327242	12.16651854	9.61E-33
Egypt	0.09013883878	0.2052761651	0.4391101069	0.6606362781
Estonia	0.8285868266	0.2074738751	3.993692344	6.78E-05
Finland	1.697095244	0.2764231966	6.139482017	1.02E-09
France	3.146630848	0.3359653075	9.365939809	2.25E-20
Germany	3.972879228	0.4788244819	8.297151416	2.11E-16
Greece	0.8103507221	0.2119327911	3.823621242	1.36E-04

Hong.Kong	2.289314251	0.2868697796	7.980325618	2.62E-15
Iceland	1.346754471	0.2128335542	6.327735661	3.16E-10
India	-0.4328869005	0.2306408076	-1.876887724	0.06070069891
Iran	-0.1814684747	0.2167463216	-0.8372390053	0.402573213
Israel	1.944337942	0.2465964269	7.884696331	5.52E-15
Italy	1.754681014	0.3751526905	4.677244915	3.13E-06
Japan	-0.6895045064	0.3690211998	-1.868468551	0.06186471972
Macau	0.4691641149	0.1992673595	2.354445385	0.01866095735
Mexico	0.3290226346	0.2000490095	1.644710141	0.1002102324
Morocco	0.4757214872	0.1991094553	2.389246089	0.0169889378
Netherlands	3.488985233	0.3727521467	9.360067442	2.37E-20
New.Zealand	4.626608701	0.2139857492	21.62110664	5.09E-92
Norway	2.710937013	0.2566477295	10.5628716	2.54E-25
Poland	0.8430444295	0.220153294	3.829351877	1.33E-04
Portugal	1.430347228	0.238180717	6.005302387	2.32E-09
Republic.of.Ireland	3.691385798	0.2032713538	18.15989184	1.27E-67
Russian.Federation	0.7233361728	0.2247703653	3.218111835	0.001314123971
Saudi.Arabia	1.54884648	0.2019096133	7.670989285	2.82E-14
Singapore	2.645707997	0.2258475817	11.71457306	1.46E-30

South.Africa	2.632523987	0.1998583072	13.17195179	7.72E-38
South.Korea	-0.1064234676	0.3068605664	-0.3468137626	0.7287731744
Spain	1.169368875	0.3017177339	3.87570482	1.10E-04
Sweden	3.497995336	0.3234194065	10.81566308	1.97E-26
Switzerland	4.834102625	0.306965303	15.74804246	2.44E-52
Taiwan	-0.3655564555	0.3088658813	-1.183544307	0.2367551659
Thailand	0.3084640974	0.211529799	1.45825363	0.1449511157
Turkey	0.752711852	0.2724749396	2.762499381	0.005796411855
United.Kingdom	7.253438455	0.2948460236	24.60076743	6.26E-115
english_fluent.1	0.2849812253	0.8462320214	0.3367648802	0.7363347615
Teaching	1.409155394	0.2905724916	4.849582924	1.35E-06
Citations	2.431278893	0.25480974	9.541546147	4.56E-21
num_students	-0.7861253641	0.2545933575	-3.087768557	0.002048505142
student_staff_ratio	1.236258617	0.314462568	3.931337917	8.78E-05
international_students	10.07465897	0.3354285304	30.03518799	4.95E-160

Interpreting the coefficients in the model above: the intercept suggests that on the average, holding every other variable constant, the predicted value of the institution outlook score is 53.50. Teaching variable has a positive effect; one unit increase in the university score for teaching leads to 1.9 increase in institution outlook score. As the number of student increases by 1 unit, i.e. one standard deviation, as the variables has been normalized, the institution outlook score decreases by 0.85. On the other hand, citations, student-staff ratio, and international students have a positive impact on

institution outlook score. This means that as the university score for citation increases by 1 unit, institution outlook score increases by 1.96. Also, a 1 unit increase in the student-staff ratio leads to a 1.22 increase in the institution outlook score. If the international students variable increases by 1 unit, the institution outlook score increases by 10.

In all of the above, a unit increase or decrease in a predictor means 1 standard deviation above or below, because the data has been normalized.

4.5 SVM

Support Vector Machine or SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges. In this case, it was used to regress the value of institution outlook score on different features. In the following sections, two types of support vector regression (SVR) will be explored: SVR with Linear Kernel, and SVR with Radial Kernel. And, for each one of them, two options will be examined. The first option is to run the model with its default tuning parameters. The second option is to design a custom grid of tuning parameters, and utilize 10-fold cross validation to figure out which tuning parameters give better results.

4.5.1 SVR with Linear Kernel and Default Parameters:

A simple SVR model has been trained, with linear kernel and default parameters. The model has held the cost parameter C at a value of 1. Using 10-fold cross validation, the resampling results are RMSE = 8.476, R^2= 0.8409. So, 84% of the variation in the response variable can be explained by this simple model, which is quite good for a starter. Next, the model has been tested on out-of-sample set, and the results are RMSE = 9.11, R^2= 0.8677. A slight increase is noticed in the test error; RMSE has gone from 8.47 on the training data to 9.11 on the test data. On the other hand, R-squared has increased slightly. Overall, it's noticed that this simple model generalizes well and doesn't seem to suffer from the issue of overfitting, for it didn't experience much change in its evaluation metrics when applied to unseen data.

Expectations are that a better model can be achieved using a better tuning for the cost parameter. This is examined in the next section:

4.5.2 SVR with Linear Kernel and custom designed grid of the Cost Parameter:

The next step in SVR discovery is to train SVR model with the linear kernel as the previous step. But this time, a user-defined grid of the tuning parameter, cost (C), has been created. 10-fold cross validation has been utilized here to achieve two goals:

- Get a better estimate of the generalization error
- Choose a value of the cost parameter that yields better results

The Figure 4.21 below shows the 10-fold CV estimate of the two metrics, across different values of the cost parameter.

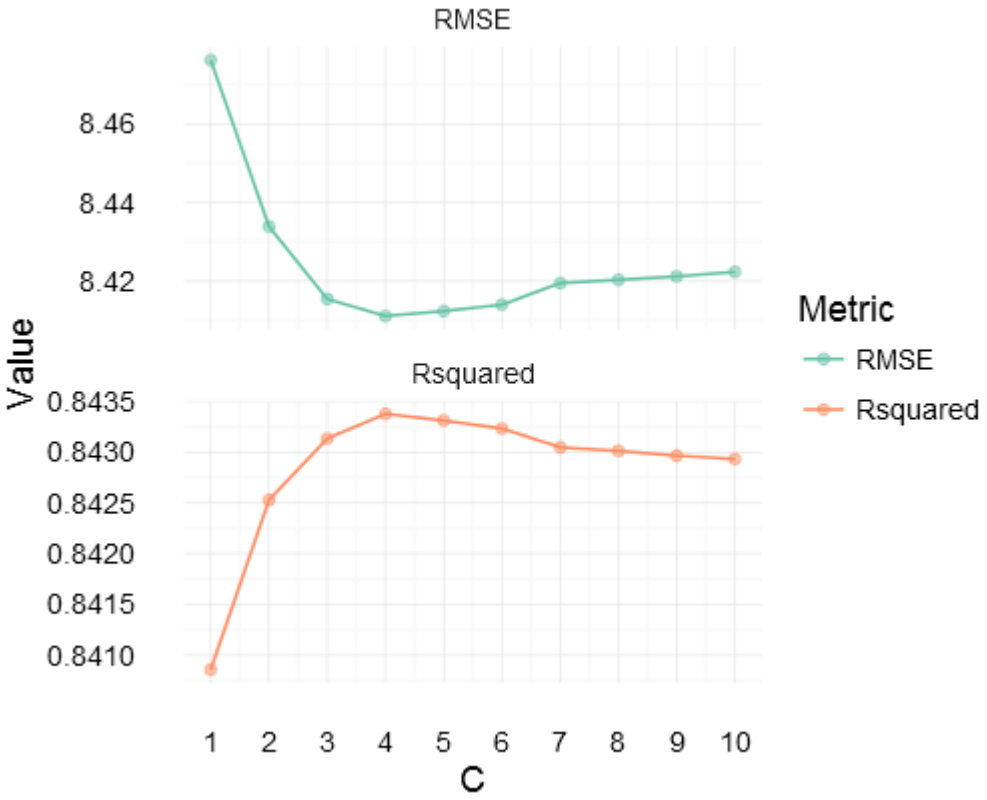


Figure 4. 21. Tuning Results for SVR model with Linear Kernel

RMSE was used to select the optimal model using the smallest value. The final value chosen for the model was C = 4, which results in RMSE = 8.411, R2 = 0.8434. An improvement of nearly 1.5% is noticed in R2.

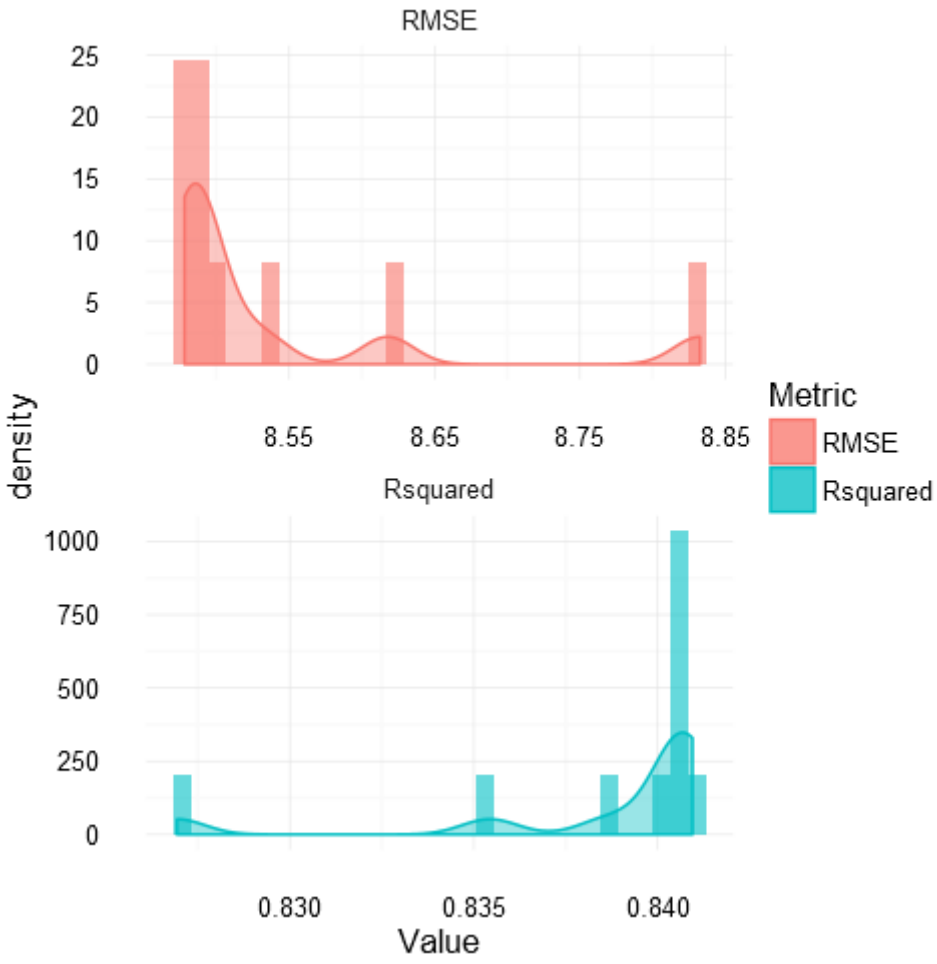


Figure 4. 22. RMSE and R2 Distribution across C

Next, the model has been tested on the unseen test set and yielded an RMSE = 9.11, R^2= 0.8677. A slight increase is observed in the test error; RMSE has gone from 8.41 on the training data to 9.11 on the test data. On the other hand, R-squared has experienced nearly 2.5% improvement. All in all, this model generalizes well and doesn't experience a significant degree of overfitting. And, it's almost identical to the first model with the default tuning parameters, in terms of generalization error.

4.5.3 SVR with Radial Kernel and Default Parameters:

The last SVR model with linear kernel and cost parameter set equal to 4 managed to explain 86% of the variation in the response variable on unseen data. This is a rather good result. But still, better results could be achieved by trying a different kernel with new tuning parameters. So, a new SVR model has been built, this time with Radial kernel and default tuning parameters. Using 10-fold cross validation, sigma was held constant

at 0.02773, and C has been chosen to be 1. At these values of the tuning parameters, RMSE turned out to be 8.556 and R^2equals 0.8383. So, 83.83% of the variation in the response variable can be explained by this simple model. The model has then been tested on out-of-sample set, and the results are RMSE = 9.92 and R^2= 0.8363.

4.5.4 SVR with Radial Kernel and Custom Designed Grid of the Tuning Parameters:

To try to achieve better results, a custom grid of the tuning parameters was designed:

C = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Sigma = 0.025, 0.05, 0.1

Of course, in order to achieve optimal tuning parameters, a wider search grid should've been designed. Nevertheless, due to insufficient computational power and resources, it's been chosen to examine only a few values of the two tuning parameters. The results of training the model, using different combinations of the two tuning parameters, while performing a 10-fold CV, is shown in the Figure 4.23 below:

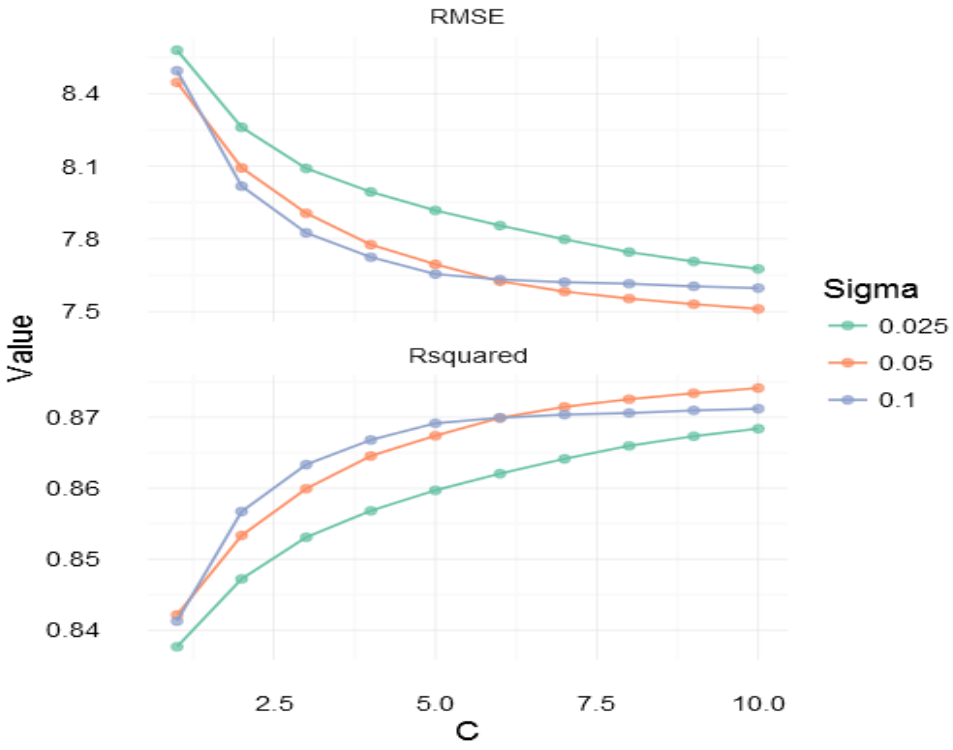


Figure 4.23 Tuning Results for SVR model with Radial Kernel

RMSE was used to select the optimal model using the smallest value. The final values used for the model were sigma = 0.05 and C = 10, which yields RMSE = 7.511, and R^2=0.8741.

The Figure 4.24 below shows the distribution of RMSE and R2 across 10-fold CV resamples:

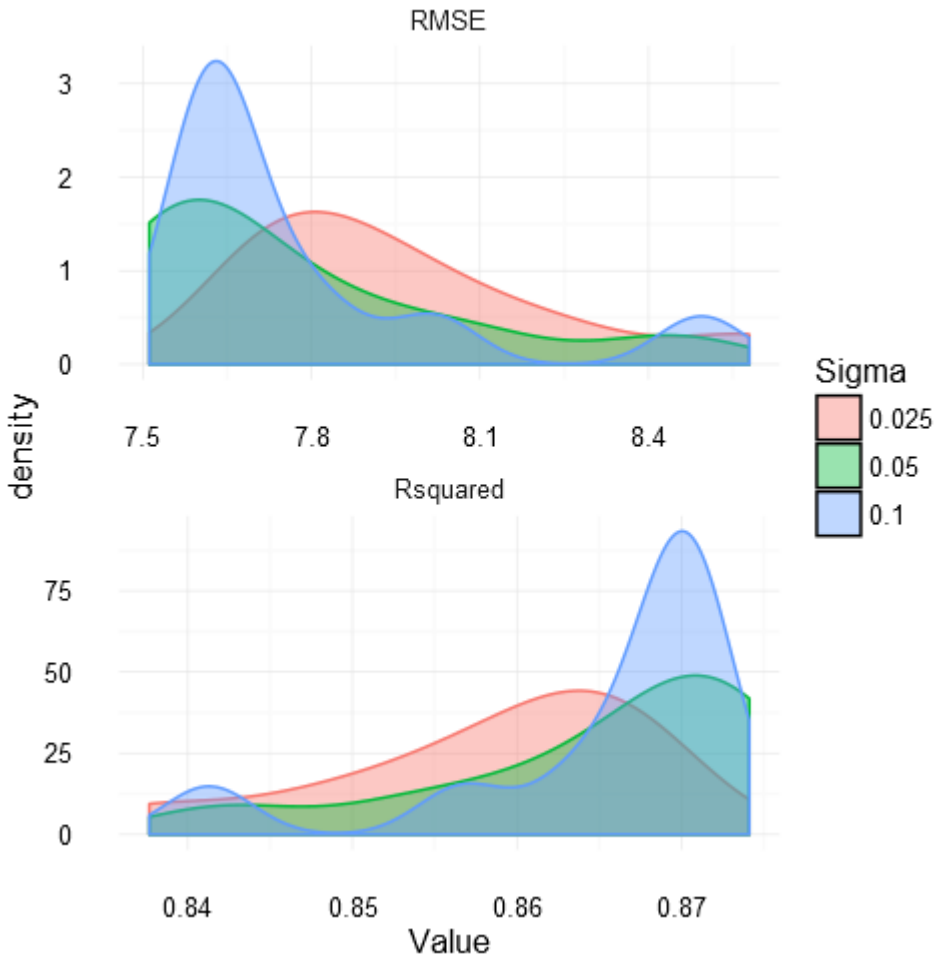


Figure 4. 24: Distribution of RMSE and R2 across Both Tuning Parameters Over the 10-fold CV Resamples

Next, the model has been tested on the unseen test set and yielded an RMSE = 10.82, R^2= 0.7966. A remarkable drop of the model performance is observed on out-of-sample data compared to its performance on the data it was trained on. R^2 has gone from 0.8741 to only 0.79, and from 7.511 RMSE to 10.82. This is a strong indication of

overfitting, and it might be due to the insufficient search for the optimal tuning parameters, due to limited computational powers as mentioned before.

4.6 Conclusion

In this chapter, a total of 9 models were trained and tested; five regression and four SVM models. The 10-fold cross validation technique has been utilized in the training process in order to get a better estimate of the generalization error and to find the optimal tuning parameters. Although Cross Validation succeeded in many cases to achieve a good estimate of the generalization error, this has not always been the case, as some remarkable differences between performance on the training and test set have been noticed for more than one model. Regression Models experienced considerable variations in their performances on in-sample as well as out-of-sample, due to trying different combinations of features. On the other hand, SVM models exhibited very similar performances on the training data, while showed some variation in performance when subject to test data. In the next chapter, critical evaluation, assessment, and analysis of the results shown, will be provided.

CHAPTER 5: EVALUATION AND ANALYSIS

5.1 Introduction

This chapter presents the results of applying two families of models, Regression and SVM, compare them, and highlight the strengths and weaknesses of the study.

5.2 The Regression Family

A summary of the evaluation metrics (RMSE and R2) for the five regression models, and for each method of evaluation: cross validation and out of sample data (Test), is presented in Figure 5.1:

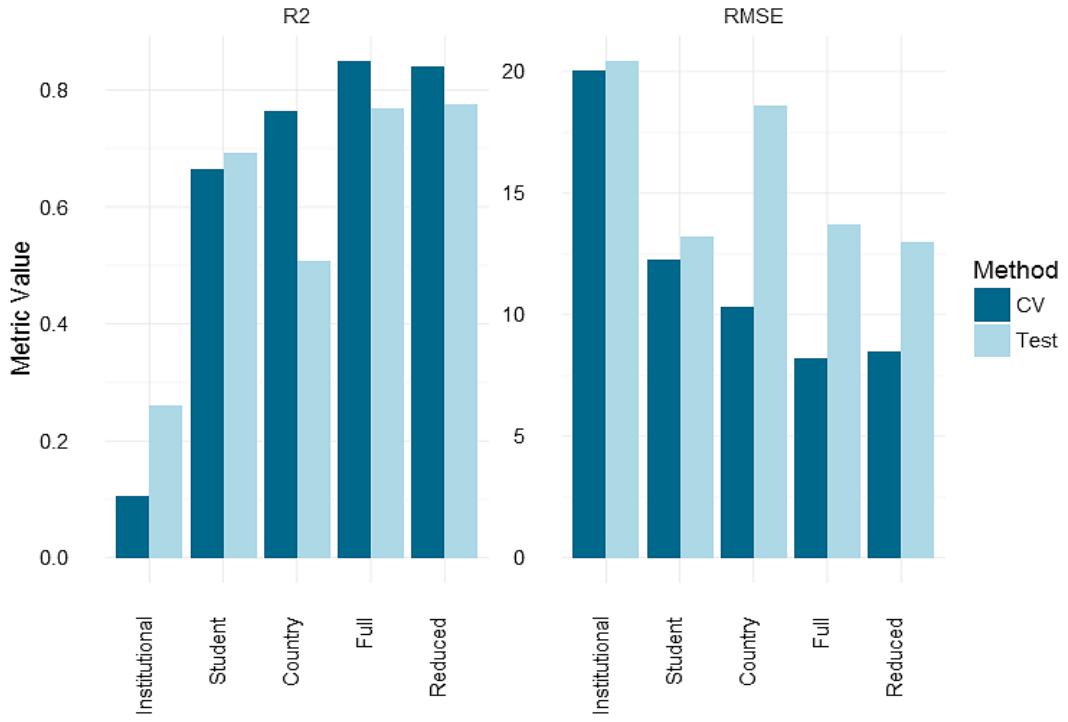


Figure 5. 1: RMSE and R2 for Regression across CV and Test Errors

Model (1): The Institutional Model: The Figure5.1 shows that the Institutional Model experienced very poor performance, and that it is significantly biased and under-fitted. This is also evident in Figure 4.8 that showed there's no clear linear trend between the international outlook score of an institution and any of its attributes.

This actually highlights an interesting phenomenon. That is, the international outlook score of an institution doesn't depend much on how truly the quality of education in this institution is. Because these attributes are chosen to predict the response variable (the international outlook), clearly correlate with the actual quality of education in this

institution, and yet when it comes to the institution outlook, it turned out that people don't put much store on these attributes. The conclusion is that some institutions enjoy a high degree of marketability on the world stage, although not very competitive when it comes to the actual educational quality, while others are totally the opposite; they could be delivering the best education while not being able to market themselves on an international level.

Model (2): The Student Model: The model based on student specific features showed far better results than the Institutional Model on in-sample as well as out-of-sample data. One very interesting observation about this is that there's one specific feature is contributing the most to the Model strength. That is the international students ratio. Figure 4.11 makes this argument crystal clear, as it shows a very strong linear relationship between the international students ratio and the response variable (international outlook).

When subject to out-of-sample data, the Student Model experienced nearly identical results to the ones resulted from the 10-fold cross validation procedure. This shows that the model doesn't suffer from a high level of variance since it produced a similar predictive performance on unseen data.

Model (3): The Country Model: Looking at Figure 5.1, adding just the country feature to the Institutional Model caused a remarkable improvement. Again, this makes a lot of sense, because some countries are very appealing to international students as well as faculty members, while others are not. Being located in the desired country or not clearly affects the ability of an institution to achieve a high level of internationalization. If a student had an opportunity to get his education in one of the countries that are well known for their high educational standards, e.g. USA, Australia, UK, Ireland, and he had the same opportunity to do the same program in a country less known for its high educational standards; holding everything else constant, he would definitely opt for one of the countries in the first group.

Having said that, the model showed far weaker performance on the held out test data, as shown in Figure 5.1. One plausible assumption for this is that splitting data into training and testing has not been randomly stratified. The test data contains all the institutions in 2016, while the training data contains all institutions before 2016. The result of doing that was that 26 countries are present in the test set while they're absent in the training set. This caused a disruption of the distribution of institutions across

countries and led to the significant difference between the training and test set as far as the country feature is concerned, which is the most important feature in the model, giving that it's the one that caused this huge improvement in the model strength. This non-stratified partitioning of the data could be viewed as a weakness on the one hand. On the other hand, it could also be viewed as a strength. More analysis and discussion of this point will be provided at the end of this chapter.

Model (4): The Full Model: This model combined the two groups of features, and as a result exhibited the highest strength on the in-sample-data. When applied to unseen data, the model experienced a reduction in its strength, as shown in Figure 5.1, although the reduction here is less than what happened with the Country Model. This is because the model gained more predictive power from combining the two groups of features. So, although the test data is to some extent different from the training data, the model managed to hold its ground and performed moderately well, as the student specific features, especially the international students ratio backed it up and prevented a strong downfall due to the sudden change in the country feature.

Model (5): The Reduced Model: The Final Model used the strongest features (selected by a stepwise selection procedure) and proceeded to produce slightly better results than the Full Model on the unseen data, and almost identical results on the training data. This improvement in performance could be attributed to the multicollinearity that was present between teaching and research features and then was removed before training the Reduced Model. This Model is considered the best model among the Regression Family, for not only does it outperform the Full Model, it's also a simpler model, and simple models are preferred over complex ones, when the same level of strength is achieved because they tend to be less prone to overfitting.

5.3 The SVM Family

A total of four SVM models have been trained. (1) A SVM with linear kernel using the default value of the parameter (Cost), (2) a SVM with linear kernel, and tuned Cost, (3) a SVM with Radial Kernel using the default values of the parameters (Cost and Sigma), (4) and a SVM with Radial Kernel and tuned Cost and Sigma.

A summary of the evaluation metrics (RMSE and R^2) for the four SVM models, and for each method of evaluation: cross validation (CV) and out of sample data (Test), is presented in Figure 5.2 below:

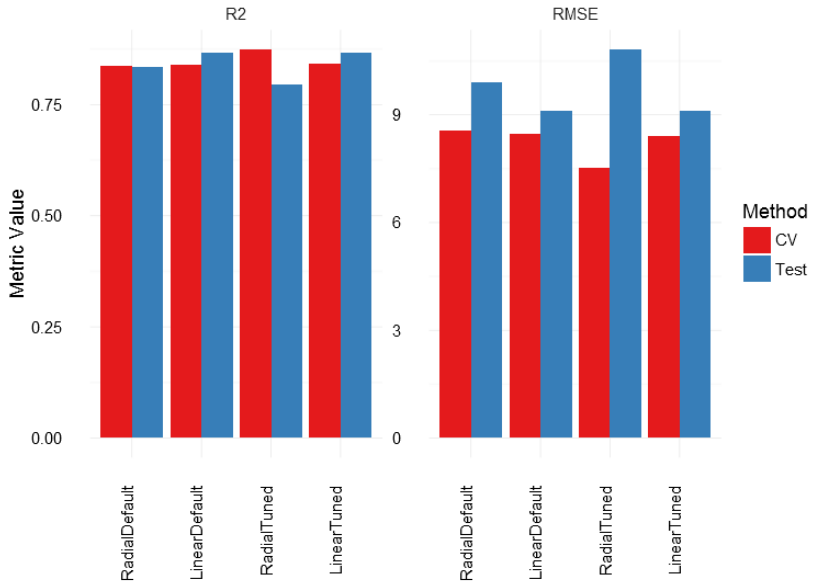


Figure 5. 2. RMSE and R2 for SVM across CV and Test Errors

It is noticed from Fig 5.2 that SVM with Linear kernel (with Default parameters, and with tuned parameters), perform slightly better than their Radial Kernel counterparts, in terms of training data as well as test data, except for the Radial SVM with tuned parameters, which performed best on the training data, due to the tuning process which forced the model to fit the data as accurately as possible. This came at a cost, however. That is, when subject to the test data it suffered a remarkable reduction in its strength, especially in terms of RMSE. Again, that shows that overly complex models are not usually the best ones. Yes, they may perform well on the data they've been trained on, but they usually fail to achieve the same excellent performance on unseen data. That is why the Radial SVM with the default values of the parameters (cost and sigma), i.e. without too much tuning on the training data, performed better on the test data.

Comparison between the two Linear Kernel SVMs is very difficult, however, for they showed almost identical results on training data as well as test data. For that reason, the simple Linear Kernel Model, the one with default Cost parameter is considered to be better. Again, simplicity is the key.

5.4 General Assessment of the Two Families

Figure 5.3 below shows a summary comparison between all models of the two families based on their respective RMSE on the test data:

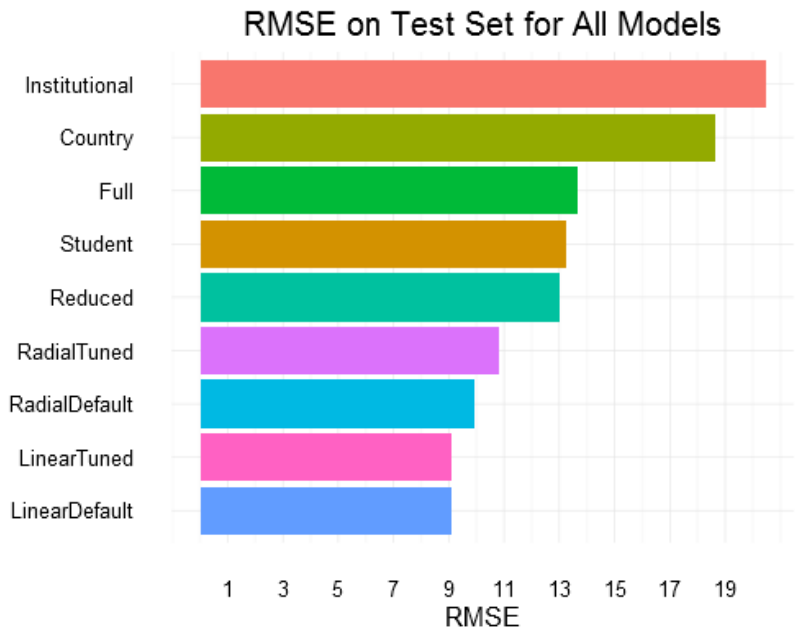


Figure 5. 3. RMSE on Test Data Across all Models

It is clear from the Figure 5.3 that the SVM family are superior to the Regression family in terms of generalization error. This is a very strong indicator of the predictive strength of SVM in general. Because although four different models have been trained using different kernels and tuning parameters, all four models exhibited very strong performance on unseen data. Even the overly fitted SVM model (the Radial Tuned), is still more powerful on test set than all regression models. The comparison between the models with respect to the 10-fold CV estimate of the generalization error is presented in Figure 5.4 below:

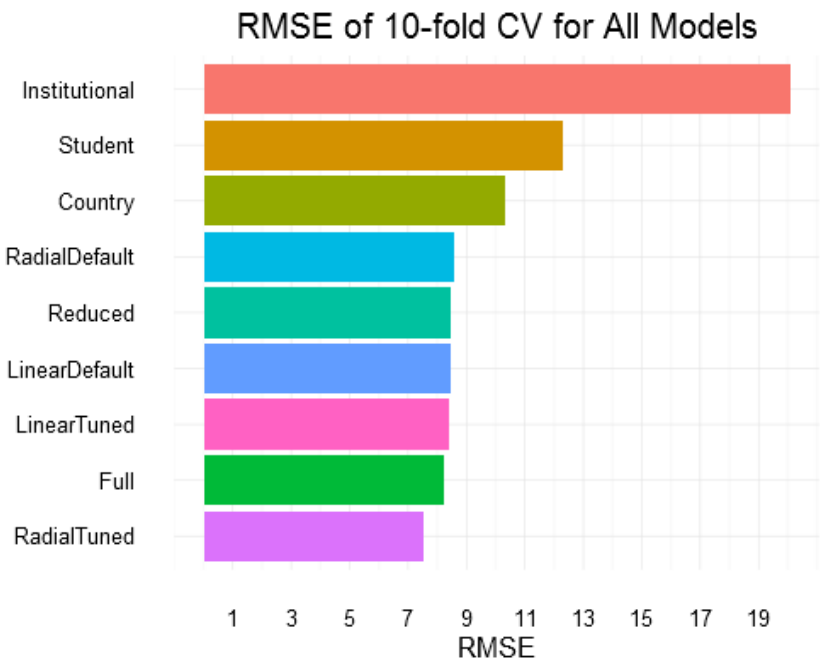


Figure 5. 4. RMSE of the 10-fold CV Across all Models

Still, SVMs exhibit strong performance, although some SVM models were outperformed only slightly by two regression models. On the other hand, the regression family are better than SVM in interpretability, as it provided such insightful remarks as to which features are more important in predicting the response, as well as the coefficients associated with each feature which quantified the relationship between the feature and the response. A virtue that SVM family lacked.

5.5 Strengths of the Research:

1- Adopting two different families of models (regression and SVM) turned out to be invaluable for the research, as one family achieved high predictive power, especially on out of sample data (SVMs), while the Regression family was highly interpretable and provided insights on the data and how each group of features interact with response variable.

2- In Regression, many models have been deployed, to try to figure out which set of features are significant in predicting the response variable. This gives decision makers in those educational institutions and in government as well, a good tool that help them make better decisions when trying to enhance the international outlook of their institutions.

3- The research provided a statistical proof of something that was assumed by common sense, which is the correlation of the country on the international outlook of an institution. Moreover, it quantified this correlation by producing a numeric value associated with each country.
4- A repeated pattern throughout the research was to use 10-fold cross validation in the training process of each model. Gives a relatively accurate approximation of the true value of the evaluation metrics (R^2, RMSE), because the model has been trained and evaluated 10 times, and the average of these 10 evaluations is taken.

5.5 Weaknesses of the Research

Below is the list of major weaknesses of the research:

1- The number of institutions is not distributed equally or close to equally across countries. Some countries have more than fifty universities, while others have less than five. This might undermine the reliability of coefficients estimates, and any change in the data would cause a big change in the model predictive power. The research has not investigated this issue carefully to show how the institutions are distributed among countries.
2- For all the models in SVM, the full set of features have been used to predict the response. Although SVMs have achieved good performance, trying sub-groups of features, as in the regression case, could have provided more insights and information about the interaction between each group of features and the dependent variable.
3- Tuning the SVMs for optimal performance only tried very few values of the tuning parameters (Cost, Sigma), due to insufficient computational powers, as well as time constraints.
4- When splitting the data, the test data was all observations in 2016, while train data was all observations before that. Stratified sampling has not been performed to split the data. This could be seen as a weakness from one point of view because it undermines the predictive power of the models when subject to test data that is significantly different from the train data. On the other hand, it could be seen as a strength, because the objective of training a model is to use it for a prediction on out-of-sample data. In the real world, out-of-sample data is not always a stratified random sample of the training data. So, by doing that, the models are

faced with a real challenge, and if they performed well, this could be a true indicator of the model predictive power.

5.6 Conclusion

This Chapter summarized and discussed the results of the research, including a comparison between each model and its family members, as well as comparing the two families of models as a whole. It also outlined the strengths and weaknesses of the research. The next chapter concludes the research and recommends future work.

CHAPTER 6: CONCLUSION

6.1 Introduction

This chapter provides a brief summary of the whole work starting with the research hypotheses and objectives, going through the CRISP-DM phases such as data pre-processing, modelling and experiments, ending with the evaluation part. It also provides an overview of all steps that have been performed and their results. It also contains the contributions to the body of the knowledge, future options and possibilities are also discussed at the end of this chapter.

6.2 Research Overview and Problem Definition

The aim of this research is to analyse the relationship between the indicators that affect the international outlook of the universities by using both statistical tests and different Machine Learning algorithms (MLR and SVM). Many different features were used for the quality assessment, these features are grouped into two categories: features related to institutions (performance assessment) such as teaching, research, citations, etc. Another group that referred to students.

Another aim was the analysis of using different variables that are not investigated before such as Level of the English spoken and the location of the universities.

The research tried to achieve the following objectives:

- To perform a thorough review of all the available methodologies for the assessment of the universities quality at international level.
- To select and add suitable features to be used for the assessment.
- To analyse the relationships between different features.
- To select the suitable ML algorithms and compare the $R^2$and RMSE

6.3 Research methodology and data understanding

This research is quantitative and experimental in nature that attempts to examine the correlations between variables. The methodology for conducting the experiment is exploratory which utilizes the existing data to construct the research hypotheses. The type of research used in this method is secondary, deductive which means that the hypotheses are tested by utilizing the theories, and it is a reasonable research. Data has been collected by kaggle from Time Higher Education (THER) which has ranked 818 universities on the basis of 13 indicators; Kaggle gathered these ranked data from 2011

until 2016 for comparing three ranked systems Time higher education ranking, Centre World University Rankings (CWUR) and Academic Ranking of World Universities from Shanghai.

Design and Implementation of this research include following steps:

- The addition of the English feature to the dataset, it is presented as dummy variable; 1 indicates that the syllabus, books and learning in the university is based on the English language and 0 referred that the university is not using English for teaching the curriculum to the students.
- The features used in the analysis are country, english_fluent, staff_ student_ratio, citations, research, teaching, international, income, num_students, female_male_ratio and Year.
- Exploration of the data by using IBM SPSS software, to ensure the quality of the data collected, as a result of this step; missing values were found in some features with different percentages.
- Initial analysis of data for determining the missing values and the outliers helped choose the appropriate techniques for solving the missing values related issues like using the mean for filling in variables with less than 20% of missing values, and variables having more than 50% missing values were permanently removed.
- Exclusion of variables such as world ranking and university name because they seemed to be unuseful for the analysis. Also, total_quality was removed from the analysis because it contains more than 50% of missing values.
- Descriptive statistics table was generated for all numerical variables to ensure that the values of variables fall within the acceptable values range, this was achieved using SPSS.
- Correlation analysis between the variables has been investigated by using R packages.
- The result of studying the effect of the university location on the international quality shows that this is a good predictor, because when this feature alone were added to the institutional model, a remarkable improvement in the model strength was noticed.
- Regression analysis model for analysing the relationship between different factors (country, english_fluent, staff_ student_ratio, citations, research,

teaching, international, income, total_quality, num_students, female_male_ratio and Year) and dependent variable (international quality).

- Regression assumptions have been checked like Independence of Observations, Linearity, Constant Variance of Error Terms, Absence of Multicollinearity, Absence of a significant level of outliers, homoscedasticity test and Normality of the Residual, the results show
- Check for the absolute values of the t-test for the purpose of finding the predictors that have a higher level of influence in the model proposed.
- 10-Fold Cross Validation was used to get a better approximation for the generalization error, as well as finding optimal tuning parameters for SVM.
- "english_fluent" variable turned out to be statistically insignificant in assessing the international quality.
- The country variable was statistically significant as a whole, although some countries were not.

6.4 Summary of the evaluation

There are nine models were trained and tested; five regression and four SVM models. 10-fold cross validation technique has been utilized in the training process in order to get a better estimate of the generalization error and to find the optimal tuning parameters. Differences between the performance on the training and test for more than one model have been noticed. Regression Models experienced considerable variations in their performances on in-sample as well as out-of-sample, due to trying different combinations of features. On the other hand, SVM models exhibited very similar performances on the training data as well as test data except for the tuned radial SVM, which was fare stronger on the training data than on the test data, due to overfitting.

6.5 Contribution to the body of the Knowledge

Internationalisation is one of the major forces shaping higher education in the globalized world of the twenty first century. This study explored the rankings of universities based on their international outlook, a score that measures how a university is concerned with the development of a multicultural community of students and staff, and the development of international alliances in research and education.

It used machine learning models to investigate the relationship between different features and the international outlook score, and to predict the value of this score in the future.

The adopted models, especially regression, revealed interesting patterns that could be insightful for academics and researchers:

To begin with, the international outlook score doesn't depend much on the actual quality of education an institution provides, as was made clear by the Institutional Model. This raises a flag to decision makers in any institution that provides high quality of education, while coming short in terms of international outlook score, to try to work more on their marketing strategy.

While working on the marketing strategy, they should focus the most on attracting international students specifically, as the student Model revealed that this is one of the strongest features in terms of predicting the international outlook

The Country Model has provided another insightful finding, for it highlighted that the country of an institution is a very strong determinant of its ability to compete on the world stage. Now, this is intuitive and may arouse a question as to whether or not this model provides any additional knowledge or insights beyond what is already known by common sense? And the answer is definitely yes, for intuition is not always correct, and this has repeatedly been proven in applied sciences. The Country Model has provided a statistical proof that common sense, in this case, is right. And, moreover, it quantified this common sense by calculating how much each country affects, or to be precise, correlates with the outlook score of an institution.

The research provided working predictive models that can be used to predict the international outlook score of universities in the future. Since the models built in this project trained on data prior to 2016 and were capable of predicting the response variable in 2016, the same models could be re-trained on data prior to 2017 and predict the international outlook score in 2017, and so on.

This research also provided some useful nuances regarding applying machine learning in the real world. Some of the key nuances are:

The importance of combining more than one validation technique to assess the quality of any predictive model. It has been repeatedly shown in this project that a model could perform very well on training data, and although 10-fold cross validation has been used to resample the training data to give a better estimate of the error, yet when subject to testing data, more than one model, especially in Regression Family, has experienced a considerable decrease in its strength. This point leads to the next one, which is:

The importance of holding out a test data that is somehow different from the training data. Meaning, the original data is not split using a stratified random sampling technique. This helps achieve a more accurate and true assessment of the strength of any model. And by doing so in this research, the true power of the Support Vector Machine model has been revealed, as most of the SVMs trained in this project performed very well on the test data, and didn't suffer from a significant downfall in their predictive ability.

6.6 Future Work

Although this study provided a thorough analysis for the relationship between the international outlook score and a number of features, more features could be investigated. For example, the age of the institution, student satisfaction, and the GDP per capita in the country of the institution, among many others.

Another important addition to this study is that many of the investigated indicators are engineered using other features. An example of this is the teaching score; it is comprised of multiple features such as: using technology, online materials, teacher awarded (alumni or Nobel prizes), so it may be beneficial to quantify the effect of using each feature of these alone in the analysis.

One possible enhancement is to try different SVMs with different group of features, as was done in regression, and see what kind of knowledge and insights could be extracted from that.

Moreover, adopt more machine learning models, especially ensemble models like Random Forests, and compare the results with the ones achieved.

Also, adding a qualitative element to the research could be invaluable, such as conducting some interviews with international students to investigate which factors they consider most important and test how well these factors work as predictors.

Another important addition is to study the relationship between the international quality and these factors (factors already explored in this research) controlled by time; this needs to apply an advanced statistical analysis such as time series analysis and cross-sectional effect.

6.7 Conclusion

The brief overview of the research problem is mentioned in this chapter, with its limitations and scope. Also, some steps in the implementation and evaluation sections are summarised with their results. At the end of it, there are two sections for the contribution and future work.

REFERENCES

Abdelhalim, A., Traore, I., &Nakkabi, Y. (2016). Creating Decision Trees from Rules using RBDT-1. *Computational Intelligence,* 32(2), 216-239. doi:10.1111/coin.12049.

Abdullah, A., Doucouliagos, H., & Manning, E. (2015). Does Education Reduce Income Inequality? A Meta-Regression Analysis. *Journal of Economic Surveys, 29*(2), 301-316.

Agarwal, S., Pandey, G. N., & Tiwari, M. D. (2012). Data mining in education: data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *2*(2), 140.

Altbach, P. G. (2008). The imperial tongue: English as the dominating academic language. *International Educator*, 17, 56-59.

Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review Future Visions. *Journal of Educational Data Mining, 1*(1).

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance?. *Journal of Science Education and Technology*, *23*(1), 160-182.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*.

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. *In Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 110-116). ACM.

Bookstein, F., Seidler, H., Fieder, M., &Winckler, G. (2010). Too much noise in the Times Higher Education rankings.*Scientometrics*, 85(1), 295-299. doi:10.1007/s11192-010-0189-5.

Breiman, L. (1996). *Out-of-bag estimation* (pp. 1-13). Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708, 1996b. 33, 34.

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M. P., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics, 71*(3), 349-365.

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M. P., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics, 71*(3), 349-365.

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M. P., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics, 71*(3), 349-365.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2016). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, 1-20.

Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter*, *13*(2), 3-6.

Campbell, J., & Oblinger, D. (2007). Academic analytics. *Washington, DC: Educause*, *27*(2) 443-459.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?. *Geoscientific Model Development Discussions*, *7*, 1525-1534.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250.

Chang, C. C., & Chou, S. H. (2015). Tuning of the hyperparameters for L2-loss SVMs with the RBF kernel by the maximum-margin principle and the jackknife technique. *Pattern Recognition*, *48*(12), 3983-3992.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*, *2015*, 12.

CLARKE, M. (2002). Some Guidelines for Academic Quality Rankings. *Higher Education in Europe*, *27*(2) 443-459.

Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Modeling and User - Adapted Interaction*, *19*(4), 341-385. doi:10.1007/s11257-009-9065-5.

Corbett, A. (2001, July). Cognitive computer tutors: Solving the two-sigma problem. *In International Conference on User Modeling* (pp. 137-147). Springer Berlin Heidelberg.

Curto, J. D., & Pinto, J. C. (2011). The corrected vif (cvif). *Journal of Applied Statistics*, *38*(7), 1499-1507.

Dill, D. D. (2006). Convergence and diversity: The role and influence of university rankings. In *Keynote Address presented at the Consortium of Higher Education Researchers (CHER) 19th Annual Research Conference* (Vol. 9).

Dill, D. D., & Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher education*, *49*(4), 495-533.

Douglas, J., Douglas, A., & Barnes, B. (2006). Measuring student satisfaction at a UK university. *Quality assurance in education*, 14(3), 251-267.

Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*(1), 141-160.

Duan, K., Keerthi, S. S., & Poo, A. N. (2003). Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, 51, 41-59.

Fan, J., Guo, S., & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74*(1), 37-65.

FILINOV , N. B., RUCHKINA, S. (2002). The ranking of higher education institutions in Russia: some methodological problems. *Higher Education in Europe, 27*(2) 407–421.

Francis, A., & Taylor, N. (2016). Quality Assurance of Higher Education in the UK: Regulatory Change and Market Competition-the Case of Law. *Revista de Educacion y Derecho,14*.

Franses, P. H. (2016). A note on the Mean Absolute Scaled Error. *International Journal of Forecasting, 32*(1), 20-22.

García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education, 14*(2), 77-88. doi:10.1016/j.iheduc.2010.07.006

Garwe, E. C. (2015). Managing the Quality of Cross-Border Higher Education in Zimbabwe. *Journal of Education and Training Studies*, 3(2), 44-50.

Gilstrap, D. L. (2013). Quantitative Research Methods in Chaos and Complexity: From Probability to Post Hoc Regression Analyses. Complicity: *An International Journal of Complexity and Education*, 10(1-2), 57-70.

Golino, H. F., & Gomes, C. A. (2016). Random forest as an imputation method for education and psychology research: its impact on item fit and difficulty of the Rasch model. *International Journal of Research & Method in Education*, 39(4), 401-421. doi:10.1080/1743727X.2016.1168798

Grömping, U. (2012). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*.

Guerine, M., Rosseti, I., &Plastino, A. (2016). Extending the hybridization of metaheuristics with data mining: Dealing with sequences. *Intelligent Data Analysis*, 20(5), 1133-1156. doi:10.3233/IDA-160860

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics,* 101(2), 1461-1473. doi:10.1007/s11192-013-1228-9

Hayes, A. F., & Rockwood, N. J. (2016). Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour Research and Therapy*, doi:10.1016/j.brat.2016.11.001

Hazelkorn, E. (2007). The impact of league tables and ranking systems on higher education decision making. *Higher education management and policy, 19*(2), 1-24.

Hirata, M., Onodera, H., & Suzuki, M. (2016). Determination of the End of Positioning Phase Using SVM: Kernel Choice and Parameter Tuning. *IFAC-PapersOnLine, 49*(21), 103-108.

Huang, B. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC Bioinformatics*, 171-13. doi:10.1186/s12859-016-1228-x

Huang, M. H. (2011). A comparison of three major academic rankings for world universities: From a research evaluation perspective. *The journal, 9*(1), 1-25.

Hung, B. T., Long, N. P., Hung, L. P., Luan, N. T., Anh, N. H., Nghi, T. D., & ... Hirayama, K. (2015). Research Trends in Evidence-Based Medicine: A Joinpoint Regression Analysis of More than 50 Years of Publication Data. Plos ONE, 10(4), 1-13. doi:10.1371/journal.pone.0121054

Jajo, N. K., & Harrison, J. (2014). World university ranking systems: an alternative approach using partial least squares path modelling. *Journal of Higher Education Policy and Management, 36*(5), 471-482.

Kaba, A. J. (2012). Analyzing the Anglo-American Hegemony in the "Times Higher Education" Rankings. *Education Policy Analysis Archives,* 20(21),

Kaycheng, S. (2015). Multicolinearity and Indicator Redundancy Problem in World University Rankings: An Example Using Times Higher Education World

University Ranking 2013-2014 Data. *Higher Education Quarterly,* 69(2), 158-174. doi:10.1111/hequ.12058

Keerthi, S. S. (2002). Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. *IEEE Transactions on Neural Networks, 13*(5), 1225-1229.

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G.,& Haydock, M. (2012). Analytics: The Widening Divide. *MIT Sloan Management Review, 53*(2), 1-22.

Kiss, M., Kun, A. I., Kapitány, A., & Erdei, P. (2014). Regression Analysis of the Effect of Personality-Career Match on the Academic Performance in Business Higher Education: an Evidence from the University of Debrecen. *Tudás–Tanulás–Szabadság Neveléstudományi Konferencia, Cluj-Napoca, Romania*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Kothari, C. R. (2004). *Research methodology: Methods and techniques*. New Age International.

Kozak, A., & Kozak, R. (2003). Does cross validation provide additional information in the evaluation of regression models?. *Canadian Journal of Forest Research, 33*(6), 976-987.

Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice, 12*(2), 137-153. doi: 10.1057/dddmp.2010.35

Lin, S. H. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges*, *27*(4), 92-99.

Lin, S., Xie, C., Tang, B., Liu, R., & Pan, A. (2016). The data mining application in the power quality monitoring data analysis. *In Industrial Electronics and Applications (ICIEA), 2016 IEEE 11th Conference on* (pp. 338-342). IEEE.

Liu, N. C., & Cheng, Y. (2005). The academic ranking of world universities. *Higher education in Europe*, *30*(2), 127-136.

Liu, N. C., & Liu, L. (2005). University rankings in China. *Higher Education in Europe, 30*(2), 217-227.

Mantovani, R. G., Rossi, A. L., Vanschoren, J., Bischl, B., & de Carvalho, A. C. (2015, July). Effectiveness of Random Search in SVM hyper-parameter tuning. *In 2015 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Marginson, S. (2007). Global University Rankings: Implications in General and for Australia. *Journal of Higher Education Policy and Management,* 29(2), 131-142.

Martínez-Muñoz, Gonzalo, and Alberto Suárez. "Out-of-bag estimation of the optimal sample size in bagging." *Pattern Recognition* 43, no. 1 (2010): 143-152.

Muller, A. (2002). Education, income inequality, and mortality: a multiple regression analysis. *Bmj,324(7328),* 23.

Norouzi, M., Collins, M. D., Fleet, D. J., & Kohli, P. (2015). Co2 forest: Improved random forest by continuous optimization of oblique splits. *arXiv preprint arXiv:1506.06155*.

Pratiyush, G., & Manu, S. (2016). Classifying Educational Data Using Support Vector Machines: A Supervised Data Mining Technique. *Indian Journal of Science and Technology, 9*(34).

Ranjan, J., & Malik, K. (2007). Effective educational process: a data-mining approach. *VINE: The Journal of Information and Knowledge ManagementSystems, 37*(4), 502-515.

Ruggiero Jr., M. A. (2016). DATA MINING FOR PATTERNS. *Modern Trader*, 58.

Ruiyun, Y., Yu, Y., Leyou, Y., Guangjie, H., &Oguti Ann, M. (2016). RAQ--A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors (14248220), 16(1), 1-16. doi:10.3390/s16010086

Rust, V. D., & Kim, S. (2015). Globalization and Global University Rankings. In *Second International Handbook on Globalisation, Education and Policy Research* (pp. 167-180). Springer Netherlands.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR), 34*(1), 1-47.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics.*

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., ... & Ray, T. S. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, *8*(1), 68-74.

Soh, K. (2015). What the Overall doesn't tell about world university rankings: examples from ARWU, QSWUR, and THEWUR in 2013. *Journal of Higher Education Policy and Management*, *37*(3), 295-307.

Stack, M. L. (2013). The Times Higher Education Ranking Product: Visualising Excellence through Media. *Globalisation, Societies and Education*, 11(4), 560-582.

Steve, S. (2010). Constructs of quality in higher education services. *International Journal of Productivity & Performance Management,* 65(8), 1091-1111. doi:10.1108/IJPPM-05-2015-0079

Sun, J., Zheng, C., Li, X., & Zhou, Y. (2010). Analysis of the distance between two classes for tuning SVM hyperparameters. *IEEE Transactions on Neural Networks*, *21*(2), 305-318.

Tondeur, J., van Braak, J., Siddiq, F., & Scherer, R. (2016). Time for a new approach to prepare future teachers for educational technology use: Its meaning and measurement. *Computers & Education, 94,* 134-150.

Van Dyke, N. (2005). Twenty years of university report cards. *Higher Education in Europe, 30*(2), 103-125.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G.,Vinatea, B., Estrella, J., & Ortigosa, Á. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User - Adapted Interaction, 21*(1-2), 217-248. doi: 10.1007/s11257-011-9098-4

Wang, Y., de Gil, P. R., Chen, Y. H., Kromrey, J. D., Kim, E. S., Pham, T., ... & Romano, J. L. (2016). Comparing the Performance of Approaches for Testing the Homogeneity

of Variance Assumption in One-Factor ANOVA Models.*Educational and Psychological Measurement*, 0013164416645162.

Wang, Y.-h., & Liao, H.-C. (2011). Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications, 38*(6), 6480-6485. doi: 10.1016/j.eswa.2010.11.098

Wen, Z., Li, B., Kotagiri, R., Chen, J., Chen, Y., & Zhang, R. (2016). Improving Efficiency of SVM k-fold Cross-validation by Alpha Seeding. *arXiv preprint arXiv:1611.07659*.

Williams, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* **36**, 181–191.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research, 30*(1), 79-82.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research, 30*(1), 79-82.

Yin, S., & Yin, J. (2016). Tuning kernel parameters for SVM based on expected square distance ratio. *Information Sciences, 370*, 92-102.

Yingqiang, Z., &Yongjian, S. (2016). Quality Assurance in Higher Education: Reflection, Criticism, and Change. *Chinese Education & Society*, 49(1/2), 7-19. doi:10.1080/10611932.2016.1192382

Yonezawa, A., Nakatsui, I., & Kobayashi, T. (2002). University rankings in Japan. *Higher Education in Europe, 27*(4), 373-382.

Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2016). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, 1-10.

Zineldin, M., Akdag, H. C., &Vasicheva, V. (2011). Assessing Quality in Higher Education: New Criteria for Evaluating Students' Satisfaction. *Quality in Higher Education*, 17(2), 231-243.

Appendix

Names of the universities

[1] Harvard University

[2] California Institution of Technology

[3] Massachusetts Institution of Technology

[4] Stanford University

[5] Princeton University

[6] University of Cambridge

[7] University of Oxford

[8] University of California, Berkeley

[9] Imperial College London

[10] Yale University

[11] University of California, Los Angeles

[12] University of Chicago

[13] Johns Hopkins University

[14] Cornell University

[15] ETH Zurich ? Swiss Federal Institution of Technology Zurich

[16] University of Michigan

[17] University of Toronto

[18] Columbia University

[19] University of Pennsylvania

[20] Carnegie Mellon University

[21] University of Hong Kong

[22] University College London

[23] University of Washington

[24] Duke University

[25] Northwestern University

[26] University of Tokyo

[27] Georgia Institution of Technology

[28] Pohang University of Science and Technology

[29] University of California, Santa Barbara

[30] University of British Columbia

[31] University of North Carolina at Chapel Hill

[32] University of California, San Diego

[33] University of Illinois at Urbana-Champaign

[34] National University of Singapore

[35] McGill University

[36] University of Melbourne

[37] Peking University

[38] Washington University in St Louis

[39] گcole Polytechnique

[40] University of Edinburgh

[41] Hong Kong University of Science and Technology

[42] گcole Normale Supârieure

[43] Australian National University

[44] Karolinska Institution

[45] University of Gسttingen

[46] University of Wisconsin

[47] Rice University

[48] گcole Polytechnique Fâdârale de Lausanne

[49] University of California, Irvine

[50] University of Science and Technology of China

[51] Vanderbilt University

[52] University of Minnesota

[53] Tufts University

[54] University of California, Davis

[55] Brown University

[56] University of Massachusetts

[57] Kyoto University

[58] Tsinghua University

[59] Boston University

[60] New York University

[61] Emory University

[62] LMU Munich

[63] University of Notre Dame

[64] University of Pittsburgh

[65] Case Western Reserve University

[66] Ohio State University

[67] University of Colorado Boulder

[68] University of Bristol

[69] University of California, Santa Cruz

[70] Yeshiva University

[71] University of Sydney

[72] University of Virginia

[73] University of Adelaide

[74] University of Southern California

[75] William & Mary

[76] Trinity College Dublin

[77] King?s College London

[78] Stony Brook University

[79] Korea Advanced Institution of Science and Technology (KAIST)

[80] University of Sussex

[81] The University of Queensland

[82] University of York

[83] Heidelberg University

[84] University of Utah

[85] Durham University

[86] London School of Economics and Political Science

[87] University of Manchester

[88] Royal Holloway, University of London

[89] Lund University

[90] University of Southampton

[91] University of Zurich

[92] Wake Forest University

[93] McMaster University

[94] University College Dublin

[95] George Washington University

[96] University of Arizona

[97] University of Basel

[98] University of Maryland, College Park

[99] Dartmouth College

[100] ےcole Normale Supârieure de Lyon

[101] Technical University of Munich

[102] University of Helsinki

[103] University of St Andrews

[104] Rensselaer Polytechnic Institution

[105] Rutgers, the State University of New Jersey

[106] Purdue University

[107] National Tsing Hua University

[108] University of Cape Town

[109] Pennsylvania State University

[110] Seoul National University

[111] Hong Kong Baptist University

[112] Bilkent University

[113] Tokyo Institution of Technology

[114] Eindhoven University of Technology

[115] National Taiwan University

[116] University of Hawai?i at M?noa

[117] University of California, Riverside

[118] University of Geneva

[119] KU Leuven

[120] Nanjing University

[121] Queen Mary University of London

[122] Michigan State University

[123] Technical University of Denmark

[124] Ghent University

[125] Lancaster University

[126] Leiden University

[127] University of Alberta

[128] University of Glasgow

[129] Stockholm University

[130] Osaka University

[131] University of Victoria

[132] Tohoku University

[133] University of Freiburg

[134] University of Iowa

[135] University of Bergen

[136] University of Lausanne

[137] University of Sheffield

[138] University of Montreal

[139] VU University Amsterdam

[140] Pierre and Marie Curie University

[141] University of Dundee

[142] University of Barcelona

[143] Utrecht University

[144] Wageningen University and Research Center

[145] University of Auckland

[146] University of Birmingham

[147] Alexandria University

[148] Uppsala University

[149] Hong Kong Polytechnic University

[150] University of Aberdeen

[151] Delft University of Technology

[152] Birkbeck, University of London

[153] Newcastle University

[154] University of New South Wales

[155] Pompeu Fabra University

[156] Indiana University

[157] Iowa State University

[158] Georgia Health Sciences University

[159] Erasmus University Rotterdam

[160] University of Delaware

[161] Arizona State University

[162] Boston College

[163] National Sun Yat-Sen University

[164] Georgetown University

[165] University of Amsterdam

[166] University of Liverpool

[167] Aarhus University

[168] University of Leeds

[169] University of Wپrzburg

[170] University of Groningen

[171] Sun Yat-sen University

[172] Goethe University Frankfurt

[173] Bielefeld University

[174] Nanyang Technological University

[175] University of East Anglia

[176] University of Nottingham

[177] University of Copenhagen

[178] Humboldt University of Berlin

[179] Monash University

[180] University of Bonn

[181] National Chiao Tung University

[182] RWTH Aachen University

[183] Middle East Technical University

[184] University of Exeter

[185] University of Twente

[186] University of Konstanz

[187] Karlsruhe Institution of Technology

[188] University of Innsbruck

[189] University of Tپbingen

[190] Drexel University

[191] University of Cincinnati

[192] Yonsei University

[193] Dalhousie University

[194] KTH Royal Institution of Technology

[195] University of Vienna

[196] Kent State University

[197] University of Illinois at Chicago

[198] Zhejiang University

[199] Simon Fraser University

[200] Swedish University of Agricultural Sciences

[201] University of Wisconsin-Madison

[202] University of Texas at Austin

[203] University of Rochester

[204] University of Bern

[205] Hebrew University of Jerusalem

[206] University of Florida

[207] Brandeis University

[208] Chinese University of Hong Kong

[209] Free University of Berlin

[210] University of Warwick

[211] Radboud University Nijmegen

[212] Medical University of South Carolina

[213] Texas A&M University

[214] University of Reading

[215] Tel Aviv University

[216] Paris Diderot University ? Paris 7

[217] Universitâ Catholique de Louvain

[218] University of Miami

[219] Queen?s University

[220] University of S?o Paulo

[221] University of Oslo

[222] University of Ottawa

[223] University of Western Australia

[224] City University of Hong Kong

[225] Maastricht University

[226] University of Leicester

[227] Autonomous University of Barcelona

[228] Cardiff University

[229] Colorado School of Mines

[230] Nagoya University

[231] Northeastern University

[232] Technion Israel Institution of Technology

[233] Tulane University

[234] Ulm University

[235] Ume University

[236] University at Buffalo

[237] University of Essex

[238] University of Georgia

[239] University of Gothenburg

[240] University of Medicine and Dentistry of New Jersey

[241] University of Otago

[242] University of South Carolina

[243] University of Strasbourg

[244] University of Waterloo

[245] University of Western Ontario

[246] Universitâ Libre de Bruxelles

[247] Carleton University

[248] Chalmers University of Technology

[249] Colorado State University

[250] Creighton University

[251] Fudan University

[252] Korea University

[253] Macquarie University

[254] State University of New York Albany

[255] Tokyo Metropolitan University

[256] University of Bologna

[257] University of Calgary

[258] University of Hamburg

[259] University of Milan

[260] University of Milan-Bicocca

[261] University of Missouri

[262] University of Padua

[263] University of Trieste

[264] Bangor University

[265] Brunel University London

[266] Johannes Kepler University of Linz

[267] Kyushu University

[268] Norwegian University of Science and Technology

[269] Queen?s University Belfast

[270] Ruhr University Bochum

[271] Stellenbosch University

[272] Tilburg University

[273] TU Dresden

[274] University of Bath

[275] University of Graz

[276] University of Kiel

[277] University of Southern Denmark

[278] University of Texas at Dallas

[279] University of the Witwatersrand

[280] University of Tsukuba

[281] University of Wollongong

[282] Victoria University of Wellington

[283] Virginia Polytechnic Institution and State University

[284] Wayne State University

[285] Aberystwyth University

[286] Autonomous University of Madrid

[287] Hokkaido University

[288] Istanbul Technical University

[289] Lomonosov Moscow State University

[290] Montpellier University

[291] Queensland University of Technology

[292] State University of Campinas

[293] Technical University of Darmstadt

[294] Tokyo Medical and Dental University (TMDU)

[295] UiT The Arctic University of Norway

[296] University of Antwerp

[297] University of Crete

[298] University of Guelph

[299] University of Iceland

[300] University of Kansas

[301] University of Kentucky

[302] University of Mپnster

[303] University of Newcastle

[304] University of Texas at San Antonio

[305] University of Trento

[306] York University

[307] Aalborg University

[308] Aalto University

[309] Bar-Ilan University

[310] Binghamton University, State University of New York

[311] Bo?aziài University

[312] Charles Darwin University

[313] Charles University in Prague

[314] George Mason University

[315] Indian Institution of Technology Bombay

[316] Jagiellonian University

[317] Keele University

[318] Keio University

[319] Lehigh University

[320] Linkس ping University

[321] National Taiwan University of Science and Technology (Taiwan Tech)

[322] Plymouth University

[323] Polytechnic University of Milan

[324] Sapienza University of Rome

[325] Shanghai Jiao Tong University

[326] Sharif University of Technology

[327] Sungkyunkwan University (SKKU)

[328] University College Cork

[329] University of Aveiro

[330] University of Canterbury

[331] University of Eastern Finland

[332] University of Ferrara

[333] University of Hertfordshire

[334] University of Houston

[335] University of Hull

[336] University of Li_ge

[337] University of Manitoba

[338] University of Maryland, Baltimore County

[339] University of Modena and Reggio Emilia

[340] University of Oklahoma

[341] University of Pisa

[342] University of Porto

[343] University of South Florida

[344] University of Stirling

[345] University of Surrey

[346] University of Tampere

[347] University of Tasmania

[348] University of Valencia

[349] University of Waikato

[350] University of Warsaw

[351] Vienna University of Technology

[352] Vrije Universiteit Brussel

[353] Washington State University

[354] Aston University

[355] Auburn University

[356] Clemson University

[357] Curtin University

[358] Deakin University

[359] Flinders University

[360] Georgia State University

[361] Griffith University

[362] Harbin Institution of Technology

[363] Heriot-Watt University

[364] Hiroshima University

[365] Kansas State University

[366] Kobe University

[367] Kyung Hee University

[368] La Trobe University

[369] Leibniz University of Hanover

[370] Liverpool John Moores University

[371] Loughborough University

[372] Mahidol University

[373] Massey University

[374] Michigan Technological University

[375] National Central University

[376] National Taiwan Ocean University

[377] National University of Ireland, Galway

[378] National University of Ireland, Maynooth

[379] New Jersey Institution of Technology

[380] New University of Lisbon

[381] Old Dominion University

[382] Polytechnic University of Catalonia

[383] Polytechnic University of Turin

[384] Polytechnic University of Valencia

[385] Pontifical Catholic University of Chile

[386] Saint Petersburg State University

[387] Swansea University

[388] Swinburne University of Technology

[389] Tokyo University of Agriculture and Technology

[390] University of Bari Aldo Moro

[391] University of Coimbra

[392] University of Idaho

[393] University of Kent

[394] University of Paris North ? Paris 13

[395] University of Salento

[396] University of South Australia

[397] University of Strathclyde

[398] University of Tartu

[399] University of Turku

[400] University of Wyoming

[401] University of Zaragoza

[402] Waseda University

[403] Wuhan University

[404] Yuan Ze University

[405] Paris-Sud University

[406] Joseph Fourier University

[407] Johannes Gutenberg University of Mainz

[408] St George?s, University of London

[409] University of Erlangen-Nuremberg

[410] Florida Institution of Technology

[411] Indian Institution of Technology Kharagpur

[412] Koà University

[413] Laval University

[414] Mines ParisTech

[415] National Research Nuclear University MePhI

[416] University of Connecticut

[417] University of Oregon

[418] Bayreuth University

[419] Oregon State University

[420] University of Montana

[421] University of Turin

[422] Claude Bernard University Lyon 1

[423] King Abdulaziz University

[424] Medical University of Vienna

[425] Murdoch University

[426] National Cheng Kung University

[427] North Carolina State University

[428] Renmin University of China

[429] University of Fribourg

[430] University of Pavia

[431] University of Portsmouth

[432] University of Vermont

[433] Indian Institution of Technology Roorkee

[434] King Mongkut?s University of Technology Thonburi

[435] National Autonomous University of Mexico

[436] Paris Dauphine University

[437] Southern Methodist University

[438] Temple University

[439] University of Duisburg-Essen

[440] University of Jyv؋skyl؋

[441] University of KwaZulu-Natal

[442] University of Minho

[443] University of Technology Sydney

[444] University of the Andes, Colombia

[445] University of Vigo

[446] Panjab University

[447] University of Cologne

[448] University of Nebraska-Lincoln

[449] University of Alaska Fairbanks

[450] Wuhan University of Technology

[451] China Medical University, Taiwan

[452] Hanyang University

[453] Indian Institution of Technology Delhi

[454] Indian Institution of Technology Kanpur

[455] King Saud University

[456] San Diego State University

[457] University of Florence

[458] University of Navarra

[459] University of Rovira i Virgili

[460] Scuola Normale Superiore di Pisa

[461] Syracuse University

[462] Sabanc? University

[463] Technical University of Berlin

[464] Federico Santa Mar؊a Technical University

[465] University of Bremen

[466] University of New Mexico

[467] Indian Institution of Science

[468] Lappeenranta University of Technology

[469] University of Macau

[470] Illinois Institution of Technology

[471] Novosibirsk State University

[472] University of Marrakech Cadi Ayyad

[473] University of Nebraska Medical Center

[474] University of Stuttgart

[475] Ewha Womans University

[476] Isfahan University of Technology

[477] Royal College of Surgeons in Ireland

[478] University of Lisbon

[479] University of Rome III

[480] University of Seoul

[481] Western Sydney University

[482] University of Mannheim

[483] Scuola Superiore Sant?Anna

[484] University of Luxembourg

[485] Charitâ - Universit؋tsmedizin Berlin

[486] Copenhagen Business School

[487] Florida State University

[488] Oregon Health and Science University

[489] Paris Descartes University

[490] Peter the Great St Petersburg Polytechnic University

[491] Royal Veterinary College

[492] Rush University

[493] Aix-Marseille University

[494] University of Bordeaux

[495] James Cook University

[496] Justus Liebig University Giessen

[497] Saint Louis University

[498] University of Tennessee, Knoxville

[499] Tomsk Polytechnic University

[500] University of Greifswald

[501] Gwangju Institution of Science and Technology

[502] University of Hohenheim

[503] Kazan Federal University

[504] Medical College of Wisconsin

[505] University of Naples Federico II

[506] ▲rebro University

[507] Technical University of Dortmund

[508] Toulouse 1 Capitole University

[509] V?B - Technical University of Ostrava

[510] University of Cyprus

[511] University of St Gallen

[512] Graz University of Technology

[513] Instituto Superior Tâcnico Lisboa

[514] University of Oulu

[515] Panthâon-Sorbonne University ? Paris 1

[516] University of South Dakota

[517] Lille 2 University ? Health and Law

[518] Verona University

[519] American University

[520] Bournemouth University

[521] University of Brescia

[522] Brno University of Technology

[523] Ca? Foscari University of Venice

[524] University of Cagliari

[525] Catholic University of the Sacred Heart

[526] City University London

[527] Complutense University of Madrid

[528] Concordia University

[529] Dublin City University

[530] East China University of Science and Technology

[531] Florida International University

[532] University of Genoa

[533] Howard University

[534] Indian Institution of Technology Madras

[535] University of Ioannina

[536] Iran University of Science and Technology

[537] University of Kaiserslautern

[538] Louisiana State University

[539] Makerere University

[540] Marche Polytechnic University

[541] University of Nantes

[542] National and Kapodistrian University of Athens

[543] National Institution of Applied Sciences of Lyon (INSA Lyon)

[544] National Yang-Ming University

[545] University of Neuch▪tel

[546] University of Nice Sophia Antipolis

[547] The Open University

[548] Oxford Brookes University

[549] University of Palermo

[550] University of Parma

[551] RMIT University

[552] University of Rome II ? Tor Vergata

[553] University of San Francisco

[554] University of Saskatchewan

[555] University of Siena

[556] Southern Cross University

[557] Tampere University of Technology

[558] University of Ulsan

[559] Ulster University

[560] Universitâ du Quâbec ▪ Montrâal

[561] Universiti Teknologi Malaysia

[562] University of Urbino Carlo Bo

[563] Xiamen University

[564] American University of Beirut

[565] Amirkabir University of Technology

[566] University of Arkansas

[567] Babe?-Bolyai University

[568] University of the Basque Country

[569] Bauman Moscow State Technical University

[570] Ben-Gurion University of the Negev

[571] Blaise Pascal University

[572] University of Burgundy

[573] University of Canberra

[574] University of Catania

[575] Central Queensland University

[576] University of Chile

[577] China Agricultural University

[578] Chung-Ang University

[579] Czech Technical University in Prague

[580] De Montfort University

[581] East China Normal University

[582] Edith Cowan University

[583] Federal University of Rio de Janeiro

[584] University of Granada

[585] University of Haifa

[586] Huazhong University of Science and Technology

[587] Indian Institution of Technology Guwahati

[588] Jadavpur University

[589] Kanazawa University

[590] King Fahd University of Petroleum and Minerals

[591] University of La Laguna

[592] University of Limerick

[593] Manchester Metropolitan University

[594] University of Maribor

[595] Masaryk University

[596] Memorial University of Newfoundland

[597] Missouri University of Science and Technology

[598] Montana State University

[599] Monterrey Institution of Technology and Higher Education

[600] National Taiwan Normal University

[601] National Technical University of Athens

[602] New Mexico State University

[603] University of North Carolina at Greensboro

[604] Oklahoma State University

[605] Osaka City University

[606] Otto von Guericke University of Magdeburg

[607] University of Oviedo

[608] Palack? University in Olomouc

[609] Pontifical Catholic University of Rio de Janeiro (PUC-Rio)

[610] Portland State University

[611] University of Pretoria

[612] Pusan National University

[613] Quaid-i-azam University

[614] University of Regina

[615] University of Rennes 1

[616] University of Salamanca

[617] University of Santiago de Compostela

[618] Semmelweis University

[619] University of Seville

[620] Universitâ de Sherbrooke

[621] Soochow University

[622] South China University of Technology

[623] Tallinn University of Technology

[624] Tehran University of Medical Sciences

[625] University of Texas at Arlington

[626] Tianjin University

[627] University of Toledo

[628] Tongji University

[629] University of Tulsa

[630] United Arab Emirates University

[631] University of Wisconsin-Milwaukee

[632] Xi?an Jiaotong University

[633] University of A Coru_a

[634] Adam Mickiewicz University

[635] AGH University of Science and Technology

[636] Ajou University

[637] University of Alcalت

[638] Alexandru Ioan Cuza University

[639] Aligarh Muslim University

[640] American University of Sharjah

[641] Amrita University

[642] Anadolu University

[643] Andhra University

[644] University of Antioquia

[645] Aristotle University of Thessaloniki

[646] Asia University, Taiwan

[647] Athens University of Economics and Business

[648] Auckland University of Technology

[649] Austral University of Chile

[650] Beijing Institution of Technology

[651] Belarusian State University

[652] University of Belgrade

[653] Birla Institution of Technology and Science, Pilani

[654] University of Bradford

[655] University of Bras؞lia

[656] University of Brighton

[657] University of Bucharest

[658] Budapest University of Technology and Economics

[659] Cairo University

[660] University of Calcutta

[661] California State University, Long Beach

[662] Capital Medical University

[663] University of Castilla-La Mancha

[664] University of Central Lancashire

[665] University of Cergy-Pontoise

[666] Chang Gung University

[667] Carlos III University of Madrid

[668] University of Chemistry and Technology, Prague

[669] Chiang Mai University

[670] Chiba University

[671] China University of Geosciences (Wuhan)

[672] China University of Petroleum (Beijing)

[673] Chonbuk National University

[674] Chongqing University

[675] Chonnam National University

[676] Chulalongkorn University

[677] Chung Yuan Christian University

[678] Chungnam National University

[679] Comenius University in Bratislava

[680] Coventry University

[681] Dalian University of Technology

[682] University of Debrecen

[683] University of Delhi

[684] University of Dhaka

[685] Dublin Institution of Technology

[686] Ehime University

[687] University of Electronic Science and Technology of China

[688] Eسtvس Lorت‌nd University

[689] Erciyes University

[690] Federal University of Bahia

[691] Federal University of Minas Gerais

[692] Federal University of Paranت (UFPR)

[693] Federal University of Rio Grande do Sul

[694] Federal University of Santa Catarina

[695] Federal University of S?o Carlos

[696] Federal University of Viàosa

[697] Federal University of Lavras

[698] Feng Chia University

[699] Fu Jen Catholic University

[700] Gda?sk University of Technology

[701] University of Ghana

[702] Gifu University

[703] Glasgow Caledonian University

[704] University of Greenwich

[705] Hacettepe University

[706] University of Huddersfield

[707] Hunan University

[708] University of Ibadan

[709] University of Indonesia

[710] Inha University

[711] I-Shou University

[712] Istanbul University

[713] Jilin University

[714] University of Jordan

[715] Jordan University of Science and Technology

[716] Juntendo University

[717] K.N. Toosi University of Technology

[718] Kaohsiung Medical University

[719] Khon Kaen University

[720] Kingston University

[721] Kinki University

[722] Konkuk University

[723] Kumamoto University

[724] Kyungpook National University

[725] Kyushu Institution of Technology

[726] University of Latvia

[727] Lille 1 University ? Science and Technology

[728] University of Lincoln

[729] University of Ljubljana

[730] Miami University

[731] Middlesex University

[732] Moscow Institution of Physics and Technology

[733] University of Murcia

[734] Nagasaki University

[735] University of Nairobi

[736] National Chengchi University

[737] National Chung Cheng University

[738] National Chung Hsing University

[739] National Taipei University of Technology

[740] National University of C¢rdoba

[741] National University of Science and Technology (MISiS)

[742] National University of Sciences and Technology

[743] Niigata University

[744] Northumbria University

[745] Northwestern Polytechnical University

[746] Nottingham Trent University

[747] Oakland University

[748] Ocean University of China

[749] Ohio University

[750] Okayama University

[751] Osaka Prefecture University

[752] University of Pardubice

[753] Paris-Sorbonne University ? Paris 4

[754] University of Patras

[755] University of Pâcs

[756] Pontifical Catholic University of Paranت

[757] Pontifical Catholic University of Rio Grande do Sul (PUCRS)

[758] Pontifical Catholic University of Valpara•so

[759] Prince of Songkla University

[760] Qatar University

[761] Rio de Janeiro State University (UERJ)

[762] Rochester Institution of Technology

[763] Saitama University

[764] University of Salford

[765] University of Santiago, Chile (USACH)

[766] S?o Paulo State University (UNESP)

[767] Savitribai Phule Pune University

[768] University of Science and Technology Beijing

[769] Sejong University

[770] Shahid Beheshti University

[771] Shanghai University

[772] Shantou University

[773] Sheffield Hallam University

[774] Shinshu University

[775] Showa University

[776] Sichuan University

[777] University of Silesia in Katowice

[778] Slovak University of Technology in Bratislava

[779] Sogang University

[780] Sophia University

[781] University of South Africa

[782] Southern Federal University

[783] University of Southern Mississippi

[784] University of Southern Queensland

[785] Suez Canal University

[786] Sultan Qaboos University

[787] Suranaree University of Technology

[788] University of Szeged

[789] Taipei Medical University

[790] Taras Shevchenko National University of Kyiv

[791] Technical University of Madrid

[792] University of Tehran

[793] University of Texas at El Paso

[794] Texas Tech University

[795] Tokai University

[796] Tokushima University

[797] Tokyo University of Marine Science and Technology

[798] Tokyo University of Science

[799] Tomsk State University

[800] Tottori University

[801] Toyohashi University of Technology

[802] Universiti Kebangsaan Malaysia

[803] Universiti Putra Malaysia

[804] Universiti Sains Malaysia

[805] Universiti Teknologi MARA

[806] Ural Federal University

[807] V.N. Karazin Kharkiv National University

[808] Vilnius University

[809] Warsaw University of Technology

[810] University of West Bohemia

[811] University of the West of England

[812] West University of Timi?oara

[813] University of Westminster

[814] Xidian University

[815] Yeungnam University

[816] Y?ld?z Technical University

[817] Yokohama City University

[818] Yokohama National University

Name of the countries:

[1] United States of America United Kingdom

[3] Switzerland Canada

[5] Hong Kong Japan

[7] South Korea Singapore

[9] Australia China

[11] France Sweden

[13] Germany Republic of Ireland

[15] Finland Taiwan

[17] South Africa Turkey

[19] Netherlands Belgium

[21] Denmark Norway

[23] Spain New Zealand

[25] Egypt Austria

[27] Israel Brazil

[29] Italy Russian Federation

[31] Greece Iceland

[33] Czech Republic India

[35] Poland Iran

[37] Portugal Thailand

[39] Chile Estonia

[41] Saudi Arabia Mexico

[43] Colombia Macau

[45] Morocco Luxembourg

[47] Unisted States of America Cyprus

[49] Unted Kingdom Uganda

[51] Malaysia Lebanon

[53] Romania Slovenia

[55] Pakistan Hungary

[57] United Arab Emirates Belarus

[59] Serbia Slovakia

[61] Bangladesh Ghana

[63] Nigeria Indonesia

[65] Jordan Latvia

[67] Kenya Argentina

[69] Qatar Oman

[71] Ukraine Lithuania

72 Levels: Argentina Australia Austria Bangladesh Belarus Belgium ... Unted Kingdom

Split data

Figure 1: Distribution of the split in the training set and testing set

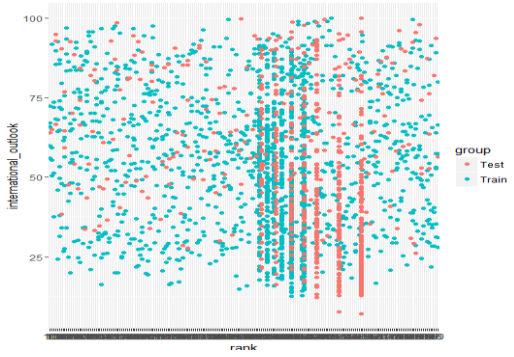
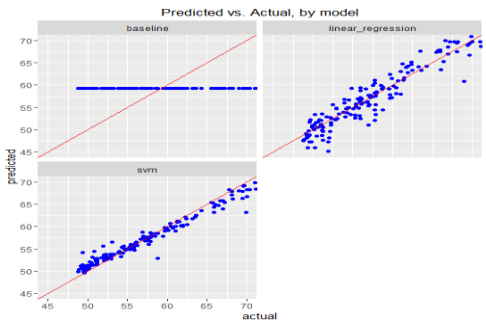


Figure 2: Comparison of the fit between the base line, Linear regression and SVM models



SVM

SVM Linear By Defualt:

C	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
1.0	8.4761741	0.8408600	6.2834763	0.5742392	0.0235611	0.4743523

Table 1: SVM Linear By Defualt Results

SVM Linear Tune:

C	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
1.0	8.476174107376158	0.8408600755970007	6.283476335121427	0.5742392592645387	0.02356115813473072	0.4743523052988671
2.0	8.433873949171277	0.8425298032032803	6.231631452942272	0.5531512774722837	0.02294851734175099	0.4492620572470726
3.0		0.8431344119418892	6.202933876544841	0.5507699689339837	0.02295511641384684 7	0.4448515971253483
4.0	8.411159119296979	0.8433776754367738,	6.188950477894226	0.5536792824230873	0.02292850174058211 7	0.4458857752500319
5.0	8.412380006478115	0.8433125056564722,	6.186861798637 8585	0.5635028028228821	0.02319712838150329 5	0.4542113993692241 7
6.0	8.414010228674837	0.8432332024087574	6.188021473573704	0.5701264283820117	0.02338327831750147	0.4576660408714801 6
7.0	8.419532813591317	0.8430483957955827	6.188397253571874	0.5726097063938295	0.02351692667481319	0.4571728685004644
8.0	8.420342684970594	0.8430142167458748	6.187949227860 23	0.5747683815311972	0.02356510677619852	0.4620939841395 0525

9.0	8.421217115617306	0.8429663929317832	6.18952488592106	0.5758641034359472	0.023649651211865245	0.4628220792241393
10.0	8.422362744597722	0.8429329322494644	6.190167272389565	0.5770075319178892	0.023678311776804923	0.4642279510498106 3

Table 2: Results from tuning SVM Linear Kernel

SVM Radial by Defualt:

sigma	C	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.027725682726442 88	0.25	10.107023299108 85	0.814036337410978 9	7.85173482113523 4	3.195121799947644 4	0.0665854518084033	3.000530307972235
0.027725682726442 88	0.5	9.056497624521734	0.828370730457016	6.82841800517337 2	2.173768468342206	0.0637703409784607 5	2.048622586588945 4
0.027725682726442 88	1.0	8.555614177589018	0.838282487557474 3	6.34036433706539 7	1.648314853958139 1	0.0563484091361727 56	1.602416715143276 2

SVM Radial Tuned:

sigma	C	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.025	1.0	8.579445282273976	0.8376549647170464	6.366507983172338	1.6874815073312956	0.0573837149603059 8	1.6361415093665 03
0.025	2.0	8.260940704065563	0.8472195569594775	6.041081438446459	1.3379789266209978	0.0476006822200 91776	1.3307002714643261
0.025	3.0	8.091055995927885	0.8530842902711757	5.848155834807191	1.1639355449536302	0.0416356502145 6643	1.1766466477966737
0.025	4.0	7.994267404907608	0.8568222917240497	5.743788523701 06	1.0496314460236502	0.0372956034375 3471	1.0574147453553009
0.025	5.0	7.917661443533092	0.8597056315939557	5.668787572802363	0.9808105027391884,	0.0346841594847 99686	0.9638036685742668
0.025	6.0	7.855226850255547	0.8620514489870086	5.609224519440608	0.9419022207406684	0.0329883642435 62246	0.9053492627571368
0.025	7.0	7.798221680245565	0.8641450163390 62	5.558028301054223 5	0.9125609631118512	0.0316607797207 39475	0.8663455832741 18

0.025	8.0	7.745601587069047	0.8659786937630404	5.508131605381680 5	0.8839735324353717	0.030451283654829974	0.8311049627754205
0.025	9.0	7.706602620948133	0.8673345208416 11	5.466135547541696	0.8600715466660828	0.029478383783096147	0.7993952693137615
0.025	10.0	7.676302048148665	0.8683882851231592	5.433844164827 5335	0.8406841202559461	0.028752514084 21457	0.7764565400239514
0.05	1.0	8.446590359913625	0.8421352033869475	6.192559962965072	1.5660336972465427	0.053632833178 19159	1.5197126712747409
0.05	2.0	8.093372149114261	0.8533435985785086	5.825607312172594	1.2009035984965521	0.042218780824 72525	1.1780409219438472
0.05	3.0	7.906170732908506	0.8599327828324 6255	5.634245230274618	1.0657531047685271	0.037213090125 64938	1.0050144629682645
0.05	4.0	7.776285776008098	0.8645450767330527	5.514576160630736	0.9858526104714215	0.034049301272 2609	0.8937077768492391
0.05	5.0	7.6950551917832	0.8674038915443396	5.440566530824 12	0.9346812014181274	0.032016890156 84221	0.8206751257091368
0.05	6.0	7.624994375183068	0.8699135977266459	5.377240055792943	0.8951419598627822	0.030393371508 653715	0.7679385599629159
0.05	7.0	,7.582385608731361	0.8714651012766929	5.327273153576273	0.8540586342135404	0.028765795011 92269	0.7246857268954078
0.05	8.0	7.553772012087358	0.8725454627900282	5.295131083252532	0.8118987078381582	0.027203168365 098	0.6823794859298693
0.05	9.0	7.530406952221977	0.8734006107176983	5.271588470676106	0.7858204724600667	0.026227053127 883083	0.6514282427419349
0.05	10.0	7.511104992975683	0.8741245704929984	5.248817841256609	0.7605709706694153	0.025334052650 43981	0.6179731237119619
0.1	1.0	8.494470874611675	0.8412816417345387	6.169791965135374	1.6908207495700562	0.056814331994 54049	1.6110576324176482
0.1	2.0	8.018108475062668	0.8567172234641772	5.735035009 1977145	1.2484498441364524	0.042129278150 72711	1.1720410527115603
0.1	3.0	7.825018711 1668295	0.8633318004477456	5.537351626391431	1.0419270960 8971	0.035214704400 95724	0.9403739294 03589
0.1	4.0	7.724583323083372	0.8668063267255307	5.424390333369277	0.9011832973681638	0.030545459244 51034	0.7980715501786633

0.1	5.0	7.654637743447404	0.8691539269117816	5.340833213036423	0.8020997396800539	0.027365696567636547	0.6703375122066464
0.1	6.0	7.632444276641013	0.8699392884611069	5.306048422934880 5	0.7432343203573 57	0.025474878990172003	0.5926440994417714
0.1	7.0	7.621472594536502	0.8703855062190056	5.277826235548318	0.6933575424386603	0.023916857585 35143	0.5341562924224028
0.1	8.0	7.615223480375592	0.8706050074122098	5.263815512213118	0.6629759941271722	0.023024992198 040998	0.4920663017441162
0.1	9.0	7.604277008365489	0.8709711550828043	5.247839327 0181085	0.6474206398682713	0.022567264083 057148	0.4664676301 2624115
0.1	10.0	7.597213168432219	0.8712106594103433	5.238647732 1853855	0.6319965421 4588	0.022166613679 79306	0.4441964372178913

Cross Validation Results:

Table (A.1): 10-fold CV Resamples for the Institutional Model

RMSE	Rsquared	Resample
19.3115288	0.1342061721	Fold01
20.66554499	0.05319773899	Fold02
19.82618897	0.144636279	Fold03
20.49878662	0.0844674765	Fold04
19.56090966	0.1598955939	Fold05
20.31715714	0.124987194	Fold06
19.66390402	0.07183731088	Fold07
20.68601679	0.08499722326	Fold08
19.83043714	0.1455217845	Fold09
20.27544831	0.0498769426	Fold10

Table (A.2): 10-fold CV Resamples for the Student Model

RMSE	Rsquared	Resample
11.91391906	0.6741290893	Fold01
12.96911456	0.6273731784	Fold02
12.60189893	0.6544748383	Fold03
12.28893459	0.6712650489	Fold04
12.29104762	0.6681610778	Fold05
11.98488071	0.6956556018	Fold06
12.16336655	0.6470146212	Fold07
11.73033967	0.709098472	Fold08
11.59349106	0.706072049	Fold09
13.20169957	0.5946343718	Fold10

Table (A.3): 10-fold CV Resamples for the Country Model:

RMSE	Rsquared	Resample
11.39985482	0.7095801347	Fold01
9.08177485	0.8247591966	Fold02
9.873500766	0.7884501092	Fold03
9.909384292	0.7875096584	Fold04
10.12527692	0.7729646005	Fold05

11.90213432	0.7009592427	Fold06
10.22312199	0.7507032337	Fold07
9.885972437	0.7912745833	Fold08
10.48519032	0.7608303705	Fold09
10.48248662	0.7425389353	Fold10

Table (A.4): 10-fold CV Resamples for the Full Model:

RMSE	Rsquared	Resample
8.962605779	0.8199484497	Fold01
7.977697652	0.8615908399	Fold02
7.909351313	0.8651453607	Fold03
8.391531961	0.8473917931	Fold04
8.096445036	0.8553859089	Fold05
9.663913036	0.8038835631	Fold06
9.125803523	0.8043251961	Fold07
7.956533952	0.8654648431	Fold08
8.074471704	0.8578443417	Fold09
8.76509451	0.8204090492	Fold10

Table (A.5): 10-fold CV Resamples for the Reduced Model:

RMSE	**Rsquared**	**Resample**
8.811296461	**0.8251054504**	**Fold01**
7.977582656	**0.8610125345**	**Fold02**
7.995334102	**0.8621648133**	**Fold03**
8.302405066	**0.8506772738**	**Fold04**
8.105462567	**0.855201355**	**Fold05**
9.522027194	**0.8093629751**	**Fold06**
9.075155938	**0.806012419**	**Fold07**
8.041375071	**0.8624883776**	**Fold08**
8.161290081	**0.8547292703**	**Fold09**
8.880236443	**0.8158271024**	**Fold10**

Table (A.6): 10-fold CV Resamples for the Default Linear SVM Model:

RMSE	**Rsquared**	**Resample**
8.532746265	**0.8345903269**	**Fold01**
7.523497695	**0.8763362694**	**Fold02**
7.745038305	**0.8688774636**	**Fold03**
8.585978502	**0.8394668975**	**Fold04**
7.696765188	**0.8697106005**	**Fold05**
9.294521482	**0.8186537236**	**Fold06**
9.074356254	**0.809849504**	**Fold07**
8.009770553	**0.8626107085**	**Fold08**
8.097191738	**0.857896789**	**Fold09**
8.653082887	**0.825235481**	**Fold10**

Table (A.7): 10-fold CV Resamples for the Tuned Linear SVM Model:

RMSE	Rsquared	Resample
7.523297833	0.8763550193	Fold02
7.745460234	0.868862367	Fold03
9.074435956	0.8098359232	Fold07
8.292460252	0.8445161835	Fold01
7.692520714	0.8698420406	Fold05
9.2452655	0.8218546146	Fold06
8.655221277	0.8251552961	Fold10
8.231692853	0.8524011474	Fold04
8.011254543	0.8625568507	Fold08
8.096483885	0.8579182189	Fold09

Table (A.8): 10-fold CV Resamples for the Default Radial SVM Model:

RMSE	Rsquared	Resample
10.48048252	0.7637585525	Fold01
7.117069529	0.8921030772	Fold02
6.747431793	0.8997313337	Fold05
11.24171471	0.7568384516	Fold04
6.580384098	0.9060857966	Fold03
10.89238331	0.7778266299	Fold06
7.068167209	0.8907497152	Fold09
7.346658513	0.884464434	Fold08
8.308322994	0.8354159671	Fold07
7.610393546	0.8640608607	Fold10

Table (A.9): 10-fold CV Resamples for the Tuned Radial SVM Model:

RMSE	Rsquared	Resample
7.89739466	0.852830848	Fold07
8.441843837	0.8477180158	Fold04
7.666739908	0.8641807019	Fold01
6.208389061	0.9148876913	Fold05
6.590957221	0.9049720689	Fold09
6.410716059	0.9102618374	Fold03
6.759115633	0.9009825751	Fold02
8.150059003	0.8599304527	Fold06
7.233661508	0.8770832437	Fold10
7.747969371	0.8717287385	Fold08