

2016-09-19

A Regression Study of Salary Determinants in Indian Job Markets for Entry Level Engineering Graduates

Rajveer Singh
Technological University Dublin

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomdis>



Part of the [Computer Engineering Commons](#)

Recommended Citation

Singh, R. (2016). A Regression Study of Salary Determinants in Indian Job Markets for Entry Level Engineering Graduates. Masters Dissertation. Technological University Dublin.

This Dissertation is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.



This work is licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 3.0 License](#)

**A Regression Study of Salary
Determinants in Indian Job Markets for
Entry Level Engineering Graduates**

Rajveer Singh

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

August 2016

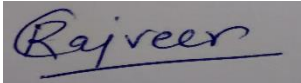
Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

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

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed:

A handwritten signature in blue ink that reads "Rajveer" is centered within a grey rectangular box. The signature is written in a cursive style.

Date:

31 August 2016

ABSTRACT

The economic liberalisation of Indian markets in early 90s boosted the economic growth of the nation in various sectors over the next two decades. One such sector that has seen a massive growth in this time is Information Technology (IT). The IT industry has played a very crucial role in transforming India from a slow moving economy to one of the largest exporters of IT services. This growth created a huge demand in the labour markets for skilled labour, which in turn made engineering one of the top choices of study after high school over the years. In addition, the earning potential and an opportunity to contribute to technology advancements after engineering, makes it a popular choice of study.

These growth dynamics along with the diversified education and labour markets demands gives insight into the factors affecting the employment outcomes of engineering students. This research study focuses on studying the key salary determinants for entry-level engineering graduates in India Labour Markets. The research examined the impact of demographics, academic performance, personality traits and standardised test scores on the starting salary.

The research findings indicated that the academic performance in school and college, college reputation, school affiliation and engineering major are key predictors for starting salary. The findings also revealed that Cognitive skills English and Quantitative ability along with a desire to do a task well are significant contributors to the starting salary of engineering graduates in Indian Labour Markets.

Key words: *Salary Predictors, Regression, Hypothesis Testing, Support Vector Machines, Feature Selection, Salary Prediction*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks my project supervisor, Dr. Basel Magableh and dissertation coordinator Dr. Luca Longo whose support and advice have been invaluable throughout this dissertation right from the start.

I would like to thank all of the staff at DIT, particularly the school of computing, whose dedication and passion for the subject have been an ongoing inspiration over the course of my Masters' programme. Also, a big thank you to my classmates who have made this past year unforgettable.

I would like to say a special thank you to my family and friends; in particular, my mother, Sajju, father, Pokharmal and brother Ranveer, for all the support they have given me, particularly over the last number of months. In addition, I would also say thank you to my dear friend Siddharth Kumar for his support and confidence in me throughout the course. I would also like to extend my gratitude to Vikas and Dimpi for their support through the process of enrolment for this programme.

TABLE OF CONTENTS

Contents

ABSTRACT.....	II
1 INTRODUCTION.....	1
OVERVIEW OF RESEARCH PROJECT	1
1.1 RESEARCH BACKGROUND	3
1.1.1 <i>A brief overview of Indian Education System in context to the research study</i>	4
1.2 RESEARCH PROJECT	5
1.3 RESEARCH OBJECTIVES AND HYPOTHESES.....	5
1.4 RESEARCH METHODOLOGY.....	6
1.5 SCOPE AND LIMITATIONS.....	7
1.6 ORGANISATION OF RESEARCH.....	7
2 LITERATURE REVIEW	9
2.1 INTRODUCTION.....	9
2.2 BACKGROUND AND CONTEXT OF RESEARCH.....	9
2.3 CORRELATION AND REGRESSION ANALYSIS.....	16
2.4 PREDICTIVE MODELLING FOR SALARY	20
2.4.1 <i>Multivariate Linear Regression with Stepwise Selection</i>	21
2.4.2 <i>Ridge regression</i>	22
2.4.3 <i>Lasso Regression</i>	23
2.4.4 <i>Support Vector Regression</i>	25
2.4.5 <i>Conclusion</i>	26
3 DESIGN AND METHODOLOGY.....	27
3.1 INTRODUCTION.....	27
3.2 RESEARCH DESIGN	27
3.3 DATA.....	28
3.4 DATA PREPARATION.....	30
3.5 ASSUMPTIONS	30

3.6	DATA INVESTIGATION AND EMPIRICAL MODEL	31
3.7	EVALUATION AND DIAGNOSTICS.....	31
3.7.1	<i>Residual Analysis</i>	32
3.7.2	<i>Evaluate homoscedasticity</i>	32
3.7.3	<i>Outlier / High Leverage points</i>	32
3.7.4	<i>Box-Cox Transform</i>	32
3.7.5	<i>Goodness-of-fit</i>	33
3.7.6	<i>Root Mean Square Error (RMSE)</i>	33
3.7.7	<i>Data Split</i>	34
3.8	CONCLUSION.....	34
4	IMPLEMENTATION AND RESULTS.....	35
4.1	INTRODUCTION	35
4.2	DATA PRE-PROCESSING	35
4.3	DATA EXPLORATION AND VISUALISATION	36
4.4	COMPARISON AND REGRESSION ANALYSIS	42
4.4.1	<i>Mean Salary Comparison Based on Gender</i>	42
4.4.2	<i>Mean Salary Comparison Based on Engineering Specialization</i>	43
4.4.3	<i>Mean Salary Comparison Based on College Tier</i>	44
4.4.4	<i>Regression Analysis for Cognitive Skills and Salary</i>	45
4.4.5	<i>Regression Analysis for Cognitive Skills, Gender and Academic Features</i> 48	
4.4.6	<i>Regression Analysis using full set of variables</i>	50
4.5	PREDICTIVE MODELLING.....	52
4.5.1	<i>Baseline Multiple Linear Regression Model</i>	53
4.5.2	<i>Model Selection using Forward Stepwise Selection</i>	54
4.5.3	<i>Model Selection using L1 Regularization (Lasso Regression)</i>	56
4.5.4	<i>Model Selection using L2 Regularization (Ridge Regression)</i>	58
4.5.5	<i>Support Vector Regression with Linear Kernel</i>	60
4.5.6	<i>Support Vector Regression with Non-Linear Kernel</i>	61
4.5.7	<i>Model Comparison Based on RMSE Value</i>	62
4.6	CONCLUSION.....	63
5	DISCUSSION AND ANALYSIS	64

5.1	INTRODUCTION	64
5.2	EXPLORATORY DATA ANALYSIS	64
5.3	HYPOTHESIS TESTING.....	64
5.4	REGRESSION ANALYSIS	65
5.5	PREDICTIVE MODELLING.....	66
5.6	AWARENESS OF STRENGTHS AND WEAKNESSES.....	66
5.7	CONCLUSION.....	67
6	CONCLUSION	68
6.1	INTRODUCTION.....	68
6.2	RESEARCH DEFINITION & RESEARCH OVERVIEW	68
6.3	CONTRIBUTIONS TO THE BODY OF KNOWLEDGE	68
6.4	EXPERIMENTATION, EVALUATION, AND LIMITATION	69
6.5	FUTURE WORK & RESEARCH	70
	BIBLIOGRAPHY	72
	APPENDIX A.....	85

TABLE OF FIGURES

FIGURE 1.1 ENGINEERING STUDENT INTAKE SINCE 1947 IN INDIA.....	2
FIGURE 2.1 OECD PROJECTIONS ON TERTIARY DEGREE HOLDERS	16
FIGURE 2.2 FITTED LINE FOR SIMPLE LINEAR REGRESSION	19
FIGURE 2.3 MULTIPLE LINES FITTED TO SIMPLE LINEAR REGRESSION	19
FIGURE 2.4 COEFFICIENTS VS LAMBDA - RIDGE.....	23
FIGURE 2.5 COEFFICIENTS VS SHRINKAGE PARAMETER – LASSO	24
FIGURE 2.6 SUPPORT VECTOR REGRESSION – GENERAL ARCHITECTURE (SMOLA AND SCHÖLKOPF, 2004).....	25
FIGURE 4.1 DISTRIBUTION OF SALARY	36
FIGURE 4.2 SCATTER PLOTS, DENSITY PLOTS AND CORRELATION COEFFICIENTS FOR SALARY AND ACADEMIC PERFORMANCE VARIABLES	38
FIGURE 4.3 SCATTER PLOTS, DENSITY PLOTS AND CORRELATION COEFFICIENTS FOR SALARY AND STANDARDISED AMCAT SCORES.	40
FIGURE 4.4 SALARY DISTRIBUTION BY JOB LOCATION.....	41
FIGURE 4.5 SALARY TREND BASED ON YEAR OF GRADUATION	41
FIGURE 4.6 BOX PLOT FOR SALARIES FOR MALE AND FEMALE GROUPS	42
FIGURE 4.7 BOX PLOT FOR SALARIES BY ENGINEERING SPECIALIZATION.....	44
FIGURE 4.8 RESIDUAL PLOTS OF THE REGRESSION MODEL FOR COGNITIVE SKILLS AND SALARY	46
FIGURE 4.9 RESIDUAL PLOTS FOR THE BASELINE REGRESSION MODEL	53
FIGURE 4.10 PREDICTED VS ACTUAL SALARY ON TEST SET	54
FIGURE 4.11 MODELS TRAINED BY STEPWISE METHOD AGAINST THE CP- STATISTICS..	55
FIGURE 4.12 VARIABLE SELECTION IN THE SELECTED MODEL.....	56
FIGURE 4.13 RMSE FOR TRAINING AND TEST SET	56
FIGURE 4.14 VARIABLE PATH CREATED BY LASSO REGRESSION	57
FIGURE 4.15 DEVIANCE PLOT FOR LASSO REGRESSION	57
FIGURE 4.16 MEAN – SQUARED ERROR VS LOG(LAMBDA)	58
FIGURE 4.17 VARIABLE PATH CREATED BY RIDGE REGRESSION	59
FIGURE 4.18 MEAN-SQUARED ERROR FOR ALL LAMBDA VALUES FOR RIDGE REGRESSION	59
FIGURE 4.19 PARAMETERS FOR SUPPORT VECTOR REGRESSION MODEL WITH LINEAR KERNEL.....	60

FIGURE 4.20 SVR – LINEAR KERNEL HYPER PARAMETER OPTIMISATION	61
FIGURE 4.21 PARAMETERS FOR FINAL SVR MODEL WITH LINEAR KERNEL AFTER OPTIMISATION	61
FIGURE 4.22 PARAMETERS FOR FINAL SVR MODEL WITH RADIAL BASIS KERNEL AFTER OPTIMISATION	62
FIGURE 4.23 RMSE SUMMARY FOR PREDICTIVE MODELLING	62

TABLE OF TABLES

TABLE 3.1 SUMMARY TABLE FOR DATASET.....	28
TABLE 4.1 DESCRIPTIVE STATICS SUMMARY FOR SALARY.....	37
TABLE 4.2 SUMMARY STATISTICS FOR ACADEMIC PERFORMANCE VARIABLES	38
TABLE 4.3 SCATTER PLOTS, DENSITY PLOTS AND CORRELATION COEFFICIENTS FOR SALARY AND COGNITIVE SKILL VARIABLES	39
TABLE 4.4 SUMMARY STATISTICS FOR COGNITIVE SKILLS	39
TABLE 4.5 MEAN AND STANDARD DEVIATION OF SALARIES FOR BOTH GROUPS.....	42
TABLE 4.6 MEAN SALARIES BY CHOICE OF ENGINEERING SPECIALIZATION	43
TABLE 4.7 SUMMARY FOR MEAN SALARY AND STANDARD DEVIATION BASED ON COLLEGE TIER.	44
TABLE 4.8 CORRELATION HEAT MAP FOR COGNITIVE SKILLS AND SALARY	45
TABLE 4.9: REGRESSION SUMMARY FOR COGNITIVE SKILL AND SALARY	47
TABLE 4.10 REGRESSION COEFFICIENTS FOR COGNITIVE SKILL AND SALARY MODEL..	47
TABLE 4.11 REGRESSION SUMMARY FOR COGNITIVE MODEL WITH INTERACTION TERMS	48
TABLE 4.12 REGRESSION COEFFICIENTS OF COGNITIVE SKILLS CONTROLLING FOR GENDER	49
TABLE 4.13 REGRESSION COEFFICIENTS FOR COGNITIVE SKILLS AND ACADEMIC VARIABLES MODEL.....	49
TABLE 4.14 REGRESSION COEFFICIENTS FROM THE COMPLETE MODEL (ABSOLUTE VALUES).....	51
TABLE 4.15 STEPS TO BUILD REGRESSION MODELS.....	52

TABLE OF EQUATIONS

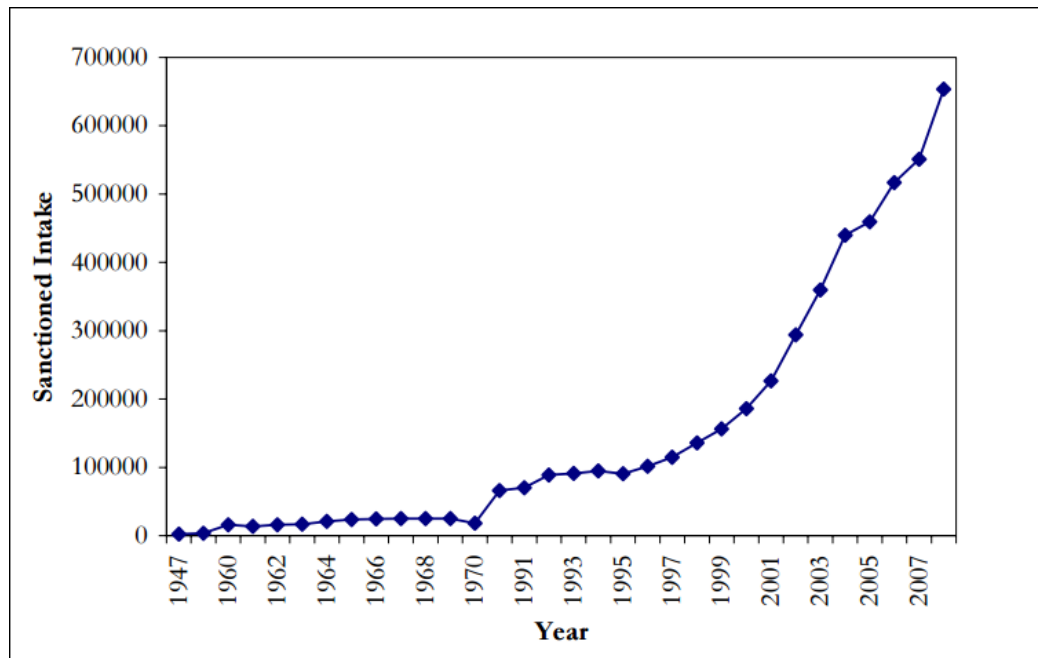
EQUATION 2.1 SIMPLE LINEAR REGRESSION	17
EQUATION 2.2 MULTIPLE LINEAR REGRESSION	18
EQUATION 2.3 ESTIMATING TARGET VARIABLE IN REGRESSION.....	18
EQUATION 2.4 LEAST SQUARES EQUATION	19
EQUATION 2.5 EQUATION FOR RIDGE REGRESSION COEFFICIENTS	22
EQUATION 2.6 EQUATION FOR RIDGE REGRESSION COEFFICIENTS	23
EQUATION 2.7 EQUATION FOR SUPPORT VECTOR MACHINES	25
EQUATION 2.8 REPRESENTATION OF LINEAR KERNEL.....	26
EQUATION 2.9 REPRESENTATION OF RADIAL BASIS KERNEL	26
EQUATION 3.1 BOX-COX TRANSFORMATION.....	33
EQUATION 3.2 FORMULA TO CALCULATE R- SQUARED MEASURE	33
EQUATION 3.3 FORMULA TO CALCULATE MEAN SQUARED ERROR	34

1 INTRODUCTION

Overview of Research Project

The new economic policies introduced in the early 1990s enhanced foreign investment portfolios into the domestic Indian Market¹. Since then, the increasing globalisation has integrated Indian labour markets to global markets. Indian labour markets have seen a tremendous growth in the last two decades in the private sector. This growth has manifested itself with a significant growth in employment opportunities in Labour Markets.

India, with one of the fastest growing economies in the world, also has one of the largest Information Technology (IT) industries in the world, hence generating a huge demand for skilled labour. In order to cater this demand, several interventions were made to encourage the engineering curriculum in Indian education sectors. The figure² below shows the increasing trend in the All India Council for Technical Education (AICTE) sanctioned intake of engineering students.



¹<http://www.ilo.org/newdelhi/lang--en/index.htm>

²<https://www.gedcouncil.org/sites/default/files/Engineering%2BEducation%2Bin%2BIndia%2BDec1608-1.pdf>

Figure 1.1 Engineering Student Intake since 1947 in India

Every year a massive number of engineers are entering into Indian Markets. In recent times, the rapidly emerging economy and increasing technology sector have a significant impact on demand and supply for specific skills, practices, and employability of engineers. Even though this demand is diversified across industrial sectors, a large number of these engineers are employed within the IT sector. The major segments within the IT industry employing these engineers are IT services, IT product development, and various associated ITeS Operations.

In recent years the supply of engineers has surpassed the demand in the Indian IT sectors. Recent survey studies have indicated that this quantity surge has also degraded the quality and employability of engineering graduates. According to National Employability Report 2011³, the increase in the number of engineers has a significant impact on the quality of engineers. The survey also revealed that there is a severe decline in the quality of education for engineers with increasing number of engineering colleges, which in turn lead to the low employability of engineering graduates. Another recent survey by Aspiring Minds in 2014⁴, states that out of almost 600,000 engineers graduating every year, less than 20% are employable within IT sector. One of the biggest challenges for the human resources policy makers, has been the employability of this massive workforce of engineers. Hence, the dynamics surrounding employability of engineering graduates have been a focus of research in recent times.

In spite of the recent shift in supply and demand of engineering graduates, engineering is still one of the top choices of undergraduate course for students after high school in India. Among the variety of reasons for these choices, such as peer pressure, awareness, society dynamics etc. salary is considered one of the major reasons for pursuing engineering as college studies.

The aim of this study is to determine the various factors that determine the starting salaries of engineering graduates in Indian Labour Markets.

³<http://www.aspiringminds.in/docs>

⁴<http://www.aspiringminds.in>

1.1 Research Background

The elements around employment outcomes after graduate studies have been an important focus of research studies over the years. The significance of employment prospects after obtaining an undergraduate degree is of critical importance. In a study by Bureau of Labour Statistics (2009)⁵ in the United States of America (USA), it was indicated that by 2018 more than half of the new jobs will require an undergraduate degree as a prerequisite. Not having an undergraduate degree will result in no access to the jobs within a number of designated employment sectors. In a social study by (Chengwen and Guiying, 2008), the authors observed that getting a job has been closely related to realising one's self and social worth. The growing importance of undergraduate studies for securing employment makes it crucial to understand the dynamics of employment post studies.

In spite of a huge amount research into the career development and graduate employment, there is still a lack of literature examining the factors determining the earning potential and starting salaries of undergraduate and graduate students (Sagen, Dallam, & Laverty, 1999).

The earning prospects and future career status of an individual are significantly determined by their first job and starting salary after an undergraduate degree (Steffy, Shaw, and Noe, 1989). In addition, Rosenbaum (1979) measured the earning benefits indicating that the starting salaries of undergraduate students have a significant impact on one's level of achievement and future wage increments.

“The starting salary of a fresh graduate is considered a potential indicator of career advancement” (Ge, Kankanhalli, and Huang, 2015).

According to (Ge, Kankanhalli, and Huang, 2015) the striking importance of starting salary of a graduate makes it very important to explore its key determinants. Undergraduate students make numerous decision during their academic years which influence the course of their career. There are a number of factors which are in play

⁵ Bureau of Labor Statistics (2009). Employment projections: 2008-2018 summary. Economic News Release. Retrieved from <http://www.bls.gov/news.release/ecopro.nr0.htm>.

such as choice of university, study majors, internships etc. and which may have a significant impact on the starting salaries and career options.

This research study will focus on determining the impact of academic performance, cognitive skills, personality traits, standardised test scores and demographics on the starting salaries of undergraduate students specific to Indian Labour Markets.

1.1.1 A brief overview of Indian Education System in context to the research study

India has one of the most diverse education systems in developing economies. There are a number of aspects leading to this diversification. India has total 36 states and union territories⁶ with 22 different recognised languages⁷. The school education is divided into two primary sections after primary education as following:

- 1) Secondary School - 10th Standard
- 2) Senior School - 12th Standard

There are more than 50 education boards in India including both Central (Government of India) and State Boards (State Government). All these schools are affiliated either to a Central Board or to one of the State Boards. The first language in central board schools is generally Hindi or English. For the State Boards, the first language can be any of the recognised languages. A student can opt to study a second and third language in school depending on the availability. The curriculum for each affiliation is not identical and varies accordingly.

For engineering studies, the student generally joins an engineering college through an entrance exam or merit based criteria depending on the type of college, which then in turn has an affiliation to one of the Central, State or Deemed Universities. The curriculum and structure of study again differ depending on the universities.

These diversities in education ecosystem make it difficult for the employer to evaluate students based on standard merit. So, in order to standardise the evaluation criteria more than 3500 organisation refers to AMCAT (Aspiring Minds) scores – A

⁶ *Library of Congress Country Studies* (5th ed.), Library of Congress Federal Research Division, December 2004, retrieved 30 September 2011

⁷ *The Constitution Of India*. Ministry of Law & Justice. Retrieved 13 April 2011. <http://lawmin.nic.in/coi/coiason29july08.pdf>

standardised test students take after the undergraduate course⁸. The dataset used for this study is released by Aspiring Minds, the organisation which facilitates this test.

1.2 Research Project

The main purpose of the research study is to examine the academic factors, cognitive skills and personality factors, which best predicts the salary of a recently graduated engineer in Indian Markets. The sample of study focuses on engineering graduates in India. Within this study, independent variables will include: Personal Information, Pre-University Information, Standardised Test Scores and Demographics Information of candidates. The dependent variables in the dataset are Starting Salary, Job location, and Job Title. Regression analysis will be performed to study the relationship between these variables. The study also performs a comparative analysis of various salary prediction models based on prediction accuracy to find an optimal salary predictor. The study aims to answer the following Research Question:

RQ: What are the primary factors, in determining the starting salary of a recently graduated engineer in Indian Labour Markets?

The results of the study would allow an engineering graduate to best navigate through various choices to achieve higher salaries. The results will also help the leaders in the education system to develop programs and resources to align with the requirements of higher wages into the Indian Labour Markets. The outcomes of the study can significantly inform the students and education administrators in terms of choices and focus on achieving a higher return for both parties.

1.3 Research Objectives and Hypotheses

In order to answer the research question, a quantitative study will be conducted using the AMEO dataset (Aggarwal, Srikant, and Nisar, 2016).

The research objectives of the research study are:

⁸ <https://www.myamcat.com/about-amcat>

- To explore the existing knowledge base by measuring or evaluating the employability and salary dynamics of undergraduate and graduate students.
- To understand the impact of different factors such as cognitive skills, academic choices, academic performance, demographics, and personality traits on the salary of fresh engineering graduates.
- To examine which of the cognitive skills is contributing the most to Salary.
- To build and select the best prediction model evaluated on the basis of Root Mean Square Error (RMSE) as the performance measuring evaluation, in predicting the salary of recent engineering graduate.

In addition, there are a few hypotheses which are established from the general understanding of the Indian education systems and Indian labour markets. These hypotheses will be tested under the research study.

H1: Male candidates are paid higher starting salaries than their Female counterparts.

H2: The engineering graduates from Tier A colleges are paid higher starting salaries than the graduates from Tier B college.

H3: The computer science graduates are paid a higher salary than the other engineering domains.

H4: English is the strongest predictor of salary compared to Logical and Quantitative ability.

1.4 Research methodology

An exploratory research method is used to address the research question using secondary data. A quantitative research approach is used to conduct an investigation into the data to understand the quantitative properties and underlying relationships within the data.

The research objectives defined earlier will be achieved using the course of action outlined below:

- An extensive literature review is conducted, which is used to summarise the existing research studies in the context to the research question. Additionally, the literature review will allow to objectively shape the course of the research project.

- Statistical test and Regression are used for hypothesis testing
- The relationships between the various factors on determining the starting salary are explored using regression analysis.
- Predictive regression modelling with various techniques is used to build an accurate salary predictor and the best performing model is selected using model accuracy based on Root Mean Squared Error.

1.5 Scope and limitations

The scope of the research study is targeted on recently graduated engineers within Indian Labour Markets. There is a significant amount of research literature available that is focused on examining the various factors that predict job seeking behaviour and re-employment of experienced professionals (Wanberg, Watt, and Rumsey, 1996). This study will not focus on experienced professional but rather on newly graduated engineers. The study is concentrated around the first job placement of engineering students after graduation and not the successive job offers.

Another important limitation might be the generalisability of few aspects of the study. The primary reason behind this would be the sample size, as there is not much information available on how the data was collected. Considering the diversification involved in the Indian population due to different education systems affiliation, language differences, demographics, etc. the sample size might be a relatively small to generalise some results of the study.

1.6 Organisation of Research

The dissertation report is organised in the following sections:

Chapter 1: This chapter is a detailed introduction to the research project. It describes the background and the objectives of the study. Also, it provides a brief overview of the research methodology, along with the scope and limitations of the research. It also outlines the organisation of the research.

Chapter 2: This chapter will address the review of the literature relating to the various labour market studies pertaining to employability factors and salary determinants. The literature will be the underlying guide for the experimental design for this research

project. The chapter will also include the detailed review of literature for the state of art techniques used for the research project.

Chapter 3: This chapter contains the detailed experimental design and research methodology for the research project. A complete overview of the dataset used is outlined. The section also covers the data semantics for the research along with the technical approach and methods employed during the course of this research.

Chapter 4: This chapter contains the implementation of the experiments, and their results. It examines and explores the dataset to address the research question. It presents the detailed experimentation and results.

Chapter 5: This chapter discusses the results from the experiments in context to research question along with strengths and weaknesses of research.

Chapter 6: This chapter provides an overview of the complete research study, briefs the contributions and limitation of the research study. It also outlines the future directions for research.

2 LITERATURE REVIEW

2.1 Introduction

This chapter reviews the existing body of literature in the context to employment outcomes for undergraduate and graduate students. The various regression techniques are assessed and discussed in detail. The literature review will provide a deep understanding of the existing work in the field, which will help to define the course of this research.

2.2 Background and context of research

Employability of undergraduate and graduate students has been a key area of research over the years. Busse (1992) stated in one of his research study that highly skilled candidates are required to fill the necessities for the rapidly changing skilled job markets to increase value to organizations.

In spite of the availability of a substantial amount of literature examining the various employability and career development aspects, the research area relevant to the prediction of salaries of graduates and the factors determining it, is fairly new (Sagen et.al (1999).

The majority of the focus in older studies has been towards the recruiting aspects and career mobility of general population (Rosenbaum, 1979, Wanberg et al., 1996). Studies also concentrated on examining the individual differences in different pieces of training and their impact on employability and income (Rosenbaum, 1979, Wanberg et al., 1996). A variety of research studies by (Chengwen and Guiying, 2008; Sagen et al., 1999; Saks and Ashforth, 1999) all investigated the various factors which predict the employment outcome of students.

Previous studies in this field have examined the effect of university scores and university status on the salary of university graduates. Boissiere, Knight and Sabot (1985) concluded that that university scores are used as a selection criteria to filter through the competition among job applicants. In addition, the students with good academic records are viewed as being better prepared for their first job (Jones and Jackson (1990). The relationship between academic performance and starting salary,

has been examined by a number of researchers in various experimental settings. Few of the earlier studies in context are (*e.g.* James *et al.*, 1989; Weisbrod and Karpoff, 1968; Wise, 1975; Murnane, Willett and Levy (1995).

Tchibozo (2007) observed that the participation in various types of extra-curricular activities during graduation plays a significant role in the employability of an individual. The required grades in core subjects are highly associated across subjects and along with micro and macro grades are significant predictors of student job placement (Athey *et al.*, (2007). Athey *et al.* (2007) also examined the contribution of academic results such as first-year grades, GRE scores, grades in core modules to the prediction of an employment conditional on Ph.D. completion.

Similar to the employability factors, in context to the salaries of undergraduate and graduate students there have been studies investigating the impact of grades, graduate majors and extra- curricular activities on to the salaries.

In an another seminal study by Hamermesh and Donald (2008), the authors determined that more than half of the variation in income is explained by factors such as ability, high school performance, parent's economic status and student's demographic characteristic. The salary difference also significantly differs for different study majors (Hamermesh and Donald, 2008; Rumberger and Thomas, 2003).

Rumberger and Thomas (1993) stated that the three different types of qualitative factors labelled as individual and institutional factors namely college major, school quality and educational performance, have an impact on starting salaries. Furthermore, the impact of school quality and educational performance is not uniform across different majors (Rumberger and Thomas (1993).

Jones and Jackson (1990) investigated the role of college GPA on the salary after the five years of graduation into the job and observed that there is an 8.9% increase in salary per unit change in GPA. However, these findings are within the limited scope of experimental design scenario.

In another study conducted by Godofsky *et al.* (2011), authors examined the impact of internships and industrial training on the transition between the education and employment. The study showed that there was a significant difference between the starting salaries of the graduates who took up internships or industrial training against the students who did not participate in any internships during the course of their graduate program (Godofsky *et al.*, 2011). Additionally, the graduates with previous

work experience related to their choice of college major have a higher starting salary compared to the ones who don't.

The findings from the study by (Godofsky et al.,2011) support another research study by Gault, Leach and Duey (2010) which observed that even the student who performs averagely in their internships are compensated more than the ones who don't participate in internships. Similarly, in another study carried out by Callanan and Benzing (2004) conducted at a mid-Atlantic university in the U.S.A, within the business degree students reflected upon the significance of completion of an internship assignment during the course of their graduate program.

The researchers have also studied the impact of academic variables in the presence of other non-academic factors. In one of the earlier works by (Fuller and Schoenberger, 1991) the academic variables were combined with the co-curricular factors to observe their impact on the starting salaries of graduate students. The study also observed a significant difference between the impact of these variables at the time of graduation to several years into the job after graduation for business students (Fuller and Schoenberger, 1991). Fuller and Schoenberger (1991) determined a significant difference between the starting salary depending on the co-curricular variables and academic performance. However the study dismissed any evidence of salary difference between students after few years into the job due to these factors.

Factors such as choice of study majors and specialisation, are also examined by researchers in relation to salary outcomes. Arcidiacono, P. (2004) observed that there is a substantial difference in the salaries corresponding to the choice of college majors. One of the interesting aspects of the study conducted by Arcidiacono, P. (2004) is that the preference for particular majors causes most of the ability sorting which establishes that the selection of a major is not based on the income outcomes but based on the interest of the student. It creates a possibility of students choosing different majors if provided with the income outcomes attached to the college majors. Apart from this, there are numerous other studies documented showing that there is a substantial difference in the salaries pertaining to the selection of some of the majors (Daymont and Andrisani,1984; Grogger and Eide,1995; James, Alsalam, Conaty, and To, 1989; Loury ,1997; Loury and Garman, 1995).

Another study examining the effects of study majors on salaries by (Scholz,1996) indicated that the most specialised fields of study fetch higher average wages and higher variance, and this is because of the theorised risk factor associated with the

majors. Scholz (1996) theorised this risk associated with the major using cobweb model of equilibrium with the argument that the demand for higher risk majors will cause higher changes in the respective income.

In another significant study by Chia and Miller (2008) using the data from University of Western Australia Graduate Destination Survey, the authors documented that the most important determinant of starting salary of graduates is the weighted average marks they achieve at the university. The study showed that the choice of major also has a significant impact but not as much as the overall academic performance at the university (Chia and Miller, 2008). Chia and Miller (2008) also observed that the difference of the salary due to the higher marks in Australia is relatively higher than the labour markets of the United States and the United Kingdom within the contextual scope of the study.

In another study by Jagacinski et al. (1985), authors identified the difference in the factors influencing the choice of engineering as a career for students from different demographics. A comparative analysis of the factors for choosing engineering streams was reported in the study by (Jagacinski et al. (1985).

Wise (1975) argued that other than the academic performance, soft skills such as leadership skills and interpersonal skills are not measured by the academic weight marks but play an important role in determining the employment outcomes and salaries.

Boissiere, Knight, and Sabot (1985) reported the effect of natural ability (cognitive skill) on the income and employment outcome in underdeveloped economies based on the two micro datasets from Kenya and Tanzania. The study also stated that the students with higher math skills are paid more than the ones with their inferior counterparts in this context (Boissiere, Knight, and Sabot, 1985). Another interesting aspect of the study (Boissiere, Knight, and Sabot, 1985), was that the development and economic factors are very similar which was reflected in the results and it allowed the assessment of the replication of study under similar external factors such as country's literacy rates, economic factors, development index, low university enrolments etc.

Research studies also reported that college reputation and ranking also have a significant impact on the earnings of graduate students. Studies conducted in U.S.A have indicated that students from high-ranking private institutions earn a higher salary than the students who attended mid-level ranking institutions (Brewer, Eide, and Ehrenberg, 1999). On the contrary, similar studies in the United Kingdom indicated

that the effect of institution's ranking is relatively less important on the salaries of graduate students (Belfield and Fielding, 2001).

Another key research topic which has been in focus over the years, is specific studies to inspect the impact of demographics factors on salaries. Studies targeted at examining the gender bias on salaries have presented mixed results. The literature shows that these mixed results seemed to be relatively more evident in IT industry.

In one of the earlier works by (Becker,1985), the author documented that due to the prioritization and responsibilities of women towards child care and housework there is a significant impact on the salaries between men and women. Berts's (1993) research work also observed similar results from a survey study of demographics, salary and job satisfaction for information system jobs.

In an earlier research work by Gerhart, (1988), the author used data from the youth cohort (ages16-19 in 1979; 19-24 in 1982) of National Longitudinal Surveys of labour markets to study the impact of academic performance on the salary difference between the gender groups. The study showed that the majority of the difference between men and women is explained by the choice of study majors of the students (Gerhart, 1988).

In another interesting work by (Fortin, 2008) based on two single group longitudinal surveys investigated the role of four non-cognitive traits – self-esteem, external locus of control (beliefs such as luck), the importance of money/work and the importance of people/family in the salary differences among men and women. The study determined that these non-cognitive traits have a reserved yet significant impact on wage difference (Fortin, 2008).

Tan and Igbaria (1994) observed that there is a variation in salaries for the gender variable specific to the information technology industry in Singapore. A research study by (Truman and Baroudi,1994) indicated that in information systems job profiles specific to managerial positions, female candidates receive relatively lower compensation compared to their male counterparts. These results were consistent even if the experiment was controlled for variables such as job level, age, education and work experience (Truman and Baroudi, 1994). This gender bias in salaries is found to be consistent by another research study on the salary outcomes from 1991 through 2008 for information system positions (Mikita, Dehondt, and Nezlek, 2012).

On the contrary, another research based in Singapore, specific to local IT markets showed that there is no difference in salaries based on gender when controlled for other academic and demographic variables (Ang, Van Dyne, and Begley, 2003). Ang,

Van Dyne and Begley (2003) used a sample of archival salary data for 1,576 IT professionals across 39 institutions in Singapore for the study.

Quan et al. (2007) documented in his research, that there was no significant difference between the salaries of male and female professionals while controlling for IT certification. In a survey study of 213 management information system (MIS) graduates by Fang (2004), the author observed no salary difference based on gender, though the results were not controlled for on any other variable.

Blau and DeVaro (2007) also reported that women have lower chances of getting promoted compared to men but there is no significant difference between the salary growth with or without promotion.

Another study by (Lyness and Heilman, 2006) investigating the relationship between position types and gender by evaluating 448 upper-level managers using individual performance metrics, stated that the organisations set a higher standard for promotions for females compared to their male counterparts and also, in general, women received lower performance ratings.

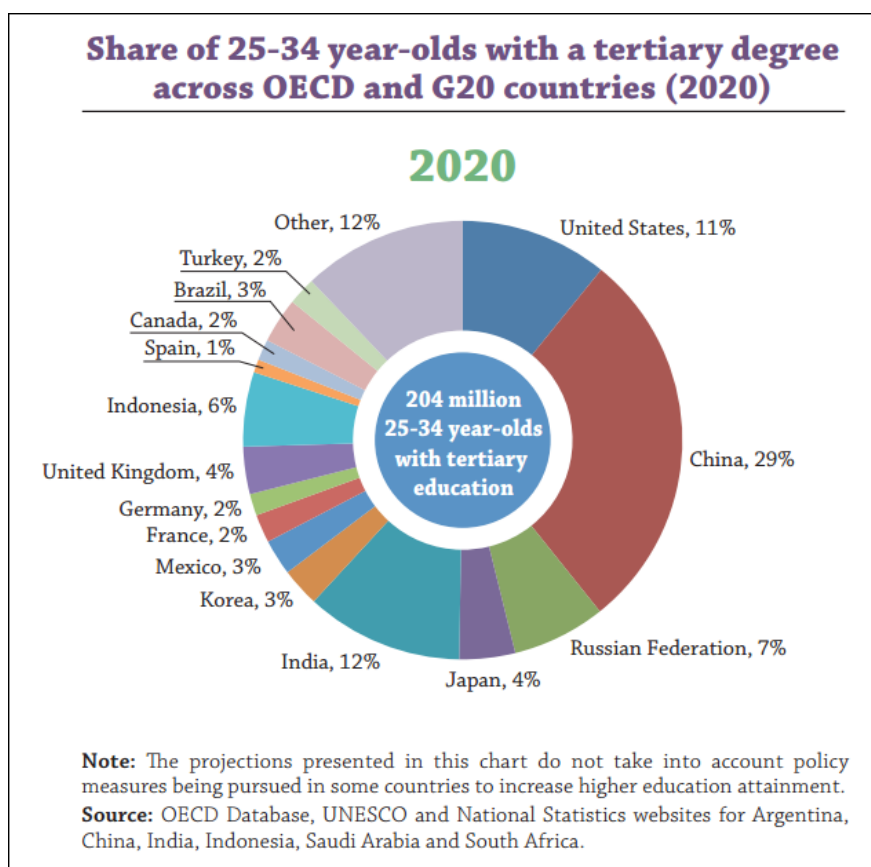
In an influential research by (Sandvig, Tyran, and Ross, 2005), the authors closely examined the effects of internship experience, grade point average (GPA) and job market on starting salaries of management information systems students from Western Washington University. Sandvig, Tyran, and Ross (2005), used demographic variables such as age, gender, and country economic factor as control variables for the analysis. The study reported internship as one of the strongest predictors of salary in the experimental model (Sandvig, Tyran, and Ross, 2005,)

In one of the most recent research studies in the field conducted by (Ge, Kankanhalli and Huang 2015), the authors theorised that the impact of demographic determinants such as foreigner status and gender have direct effects on the starting salary for entry level IT graduates in Singapore job markets. Ge, Kankanhalli, and Huang (2015) observed that females are less likely to land up into an IT job after a degree in IT compared to men. Researchers also outlined that foreign IT graduates are more likely to join an IT job as compared to local graduates but are offered lower starting salaries than the local graduates (Ge, Kankanhalli, and Huang, 2015).

The majority of the influential research literature available in context to the education, employability factor, and salary prediction is geographically centred to the studies within developed countries such as United States of America, Australia, Singapore etc. Relatively, there is less research to study the similar dynamics in underdeveloped and

developing economies. Some interesting studies from such socio-economic status nations are (Boissiere, M., Knight, & Sabot, 1985), where the authors studied the dataset from Tanzania and Kenya, (Tripney et al.,2013) documented a system review outlining the various employability factor and interventions to improve for basic technical and vocational education in low and middle-income countries. Other similar significant studies concentrating on the Chinese labour markets are (Bai, L. 2006); Xiangquan,2009).

That said, there has been a paradigm shift in the focus of such research groups majorly towards the fastest growing economies such as China and India. According to Organisation for Economic Co-operation and Development (OECD)⁹ report for education indicators in focus, the massive growth in higher education in fastest growing G20 economies has led a decrease in the share of Japan, Europe, and U.S.A in the global talent pool. With this continuous growth rate, OECD projections state that China and India alone will account for 40% of the total personal with tertiary graduates.



⁹ <https://www.oecd.org/edu/50495363.pdf>

¹⁰ **Figure 2.1 OECD Projections on tertiary degree holders**

These growth projections are believed to influence the research focus towards developing economies and to understand the underlying factors in context. One of the most important research areas for the socio-economic growth has been the education, employability factors and their impact on salaries.

In a recent study by (Gokuladas, 2011), the author studied the various factors which decide on the employability of undergraduate engineers in India based on a sample of 559 engineering graduates from a reputed engineering college in Southern India. Ge, Kankanhalli and Huang (2015) used linear modelling to understand the effects of demographics factors on the starting salary of IT graduates.

The authors (Ge, Kankanhalli, and Huang, 2015) suggested for future work to utilise a sample from multiple universities combined with academic data, to study the effects of salary. The data of the study in line was limited to demographic variables in their study, which could be extended to academic and other external factors.

In order to bring transparency to the one of most diversified educational ecosystem towards the employability outcome, it is very critical to understand the underlying factors that determine the starting salaries of graduates. The AMEO-2015 (Aggarwal, Srikant, and Nisar, 2016) dataset provides a unique opportunity to study the effects of demographic variables, along with academic performance with standardized test scores for cognitive and personality scores. This study will focus on understanding the various salary determinants for entry-level engineering graduates in Indian Labour Markets.

It is evident from earlier research in the field that academic factors, demographic variables, and natural ability have a great influence on the salaries of graduate and undergraduate students. There are various research techniques used by the researchers to study these effects. The next section provides a review of literature relating to the techniques used by researchers in the past.

2.3 Correlation and Regression Analysis

The most common way to investigate the relationship between two variables is correlation and regression. A correlational research is used to identify and quantify

¹⁰ <https://www.oecd.org/edu/50495363.pdf>

relationships between two or more variables within a population (Leedy and Ormrod, 2010; Curtis, Comiskey, and Dempsey, 2016). The initial usage of correlation and linear regression techniques for research purposes can be traced back to the work of Sir Francis Galton (Miller and Millar 1996; Curtis, Comiskey, and Dempsey, 2016). In his work, Galton primarily concentrated on studying inheritance, which eventually resulted in the development of regression.

Correlation can be defined as the degree of linear relationship between two variables. It is one of the most widely used tools to establish the strength and direction of the relationship between two variables. For example, a researcher might be interested in identifying the relationship between the intake of nutrients (N) with the growth in height (H) for children over a period of 10 years. It would be expected that the children with higher intake of nutrients would grow more compared to those who have a lower intake of nutrients.

Correlation defines how the rate of increase or decrease in one variable corresponds to increase and decrease in another variable. That said, this relationship is not causal in nature i.e correlation doesn't infer that the change in one variable is causing the change in another. In general, a causation effect cannot be inferred from a correlation study.

Regression, in general, can be defined as a family of techniques for estimating relationships. The simplest, yet most powerful form of regression is called Linear Regression.

Linear regression assumes the relationship between the target or dependent variable 'y' and features or independent variables 'x₁, x₂, x₃ x_n' is linear in nature. Linear regression takes the relationship between dependent and independent variables and fits a line to the distribution, which can be used to predict the target or dependent variable using the feature or independent variables (Han, Kamber, and Pei, 2011). For example, a simple linear regression with one feature variable can be represented as a simple equation of a line as:

$$y = w_0 + w_1 x + e$$

Equation 2.1 Simple Linear Regression

Where;

w₀ → Intercept

y → Response/Target/Dependent Variable

$x \rightarrow$ Input/Feature/Independent Variable

$w_1 \rightarrow$ slope

$e \rightarrow$ Error Term or 'noise' term that represents the other variables to have an influence on target variable.

In real application scenarios there is typically more than one independent or explanatory variable, so in order to handle the problem of omitted bias, multiple regressions are applied in which there are 'n' independent variables. A multiple linear regression is represented as below:

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n + e$$

Equation 2.2 Multiple Linear Regression

The value of w_i indicates the measure of a relationship, a value closer to 0 indicates a weak relationship and a value farther from 0 represents a strong relation (positive or negative). The noise term 'e' states that our model will not fit the model perfectly. Here the model is created for 'e' being Gaussian. The target value y_p for a given point can be predicted using the equation :

$$y_p = w_0 + w_1 x_1 + e$$

Equation 2.3 Estimating Target variable in Regression

The difference between y and y_p is called residuals which are equal to $y - (w_0 + w_1 x_1)$ (Myers et al., 2012). The residual sum of squares(RSS) is the total error over all the data points.

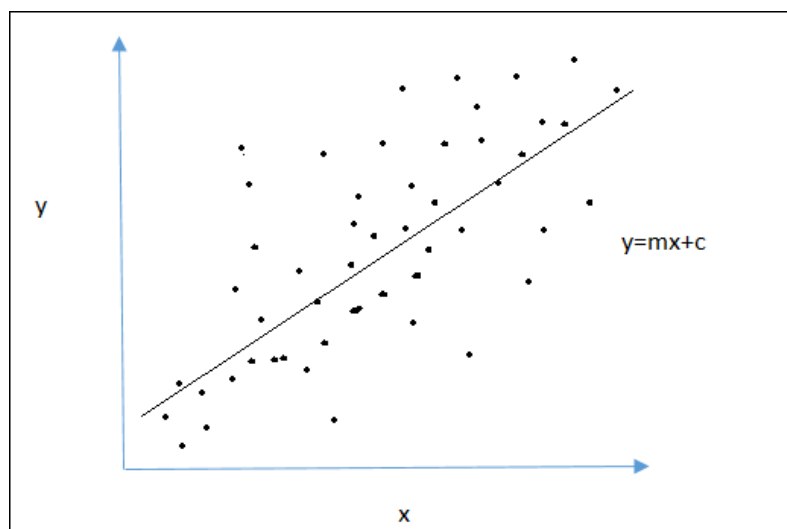


Figure 2.2 Fitted Line for Simple Linear Regression

The figure (above) shows the fitted line of regression for a single input variable **x**. This line/model can be used to predict the value of ‘y’ if the value of **x** is known. But as the figure below represent there can be multiple fitted lines on the given data point.

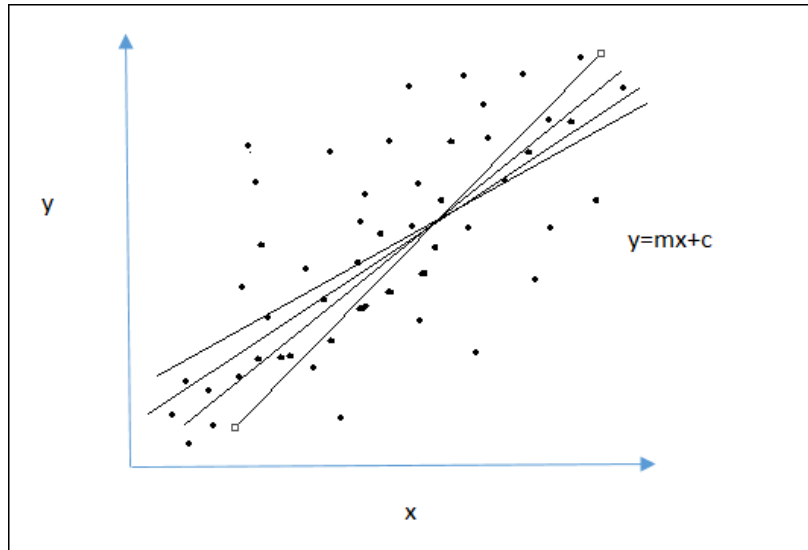


Figure 2.3 Multiple Lines fitted to Simple Linear Regression

To find the best line of fit for the observed data, we solve the optimization condition for the line where the probability of data is highest (James et al.,2013).

$$\min_{w_0, w_1} \sum (y - (w_0 + w_1x_1))^2$$

Equation 2.4 Least Squares Equation

where **min** w_0, w_1 means “minimize over w_0, w_1 ”. This is also called a Least square linear regression problem (James et al.,2013). Also, the assumption here of Gaussian Noise provided the need for the squared error to be the minimization criteria. There could be other assumptions for the cost function for a different distribution.

The significance and validity of the model is evaluated by p-value and R-Squared value for the model¹¹.

Howell (1969) used regression analysis for the very first time in one of the early research works in the field where the weighted factor is used to best describe the relations for the factors affecting the average wages (Howell, Gorfinkel, and

¹¹ <https://onlinecourses.science.psu.edu/stat501/node/311>

Bent,1966). In (Jagacinski et al., 1985), the authors used statistical test to describe the factors in influencing the career choice for engineers.

In another seminal study by Ge, Kankanhalli and Huang (2015) considering the advantage of its fewer distribution requirement to covariance-based structural analysis to a test multistage stage deterministic model (Gefen, Straub, and Rigdon, 2011) used Partial Least Squares to investigate the determinants of starting salary (Götz, Liehr-Gobbers and Krafft, 2010).

Some of the other methodological studies to understand the relationship among variables by (Guyon and Elisseeff, 2003) and (Karagiannopoulos et al., 2007), provides an overview of multivariate feature selection criteria's and methods for different regression models.

2.4 Predictive Modelling for Salary

“The increasing awareness and concern with equity issues in higher education, along with the escalating litigation, has prompted institutions to undertake salary prediction studies” (Johnson, Riggs, and Downey (1987).

Johnson, Riggs, and Downey (1987) performed a comparative salary predictive modelling study using predicted rank, tenure and objective variables for equity studies. The research outcome helped institutions to derive appropriate analytical strategies for predictive modelling for salaries. Carter et al. (1984) proposed alternative techniques using canonical analysis and multiple discriminant analysis to identify and define the new evaluations of magnitudes for the salary decisions. Ramsay (1979) used generalized linear regression models to predict the salaries of college faculties.

Prediction of salaries has been a very active field of research in the field of sports. In one of the most recent studies by (Magel and Hoffman, 2015), the authors used a number of stepwise multivariate linear regression models to predict the average salaries of baseball players.

In a unique study in terms of the dataset, in Finland, researchers build a penal data model to predict individual income with a third degree polynomial of age, duration of employment and GDP as independent variables (Koskinen, Nummi, and Salonen, 2005).

According to researchers, the salary prediction is a very old problem in pension insurance and a numerous number of models have been proposed over the years for the same (Koskinen, Nummi, & Salonen, 2005).

Carriere and Shand (1998) defined salary as an accumulation function based on inflation and merit and built a parametric model to determine the salaries with a comparative modelling of age and service based modelling.

In another interesting salary prediction application, recently a Kaggle competition was targeted at predicted the salaries based on the job advertisements on job portals.

The participants employed a number of machine learning algorithms to build predictive models. Among others, the popularly applied models were Lasso and Ridge regression¹².

As part of this study, a salary predictive model will be built and the best performing model will be selected using various model selection techniques based on the minimised error function. The following section provides a review of techniques used for this study.

2.4.1 Multivariate Linear Regression with Stepwise Selection

Multivariate Linear Regression is one of the most widely used predictive tools to estimate a continuous target using multiple predictors.

Stepwise methods are often used in education and psychological research to find a subset of predictor variables based on their relative importance (Huberty, 1989; Thompson, 1995). Stepwise Regression builds a number of linear models sequentially, by entering one best predictor at a time Snyder (1991). Stepwise Selection has three different variations called– 1) Forward 2) Backward and 3) Stepwise.

A forward stepwise selection starts from a NULL model and then adds one predictor on every step which best improves the error for the model. A backward stepwise selection, on the other hand, starts from a FULL model with all predictors as input and then on every step removes one predictor which adds maximum error to model. A stepwise variation is a hybrid of forward and backward which adds a predictor at each step and also considers the removal of predictors which no longer contribute to the

¹² <http://www.cs.ubc.ca/~nando/540-2013/projects/p58.pdf>

prediction power when considered in combination with the newly added predictors (Thompson,1989).

2.4.2 Ridge regression

Ridge Regression is a technique to estimate the coefficients of regression by adding a bias to the regression estimates (Hoerl and Kennard, 1970). In the ordinary least squares regression technique, the regression coefficients are estimated using the sum of squares error.

From earlier Equation (2.1)

$$y_p = w_0 + w_1 x_1 + e$$

Where 'e' is the associated error.

The error is the difference between the Observed and Predicted values. In a linear equation, this error can be divided into two components, namely the 'error due to bias' and 'error due to variance'. The regression parameters such as w_1 are estimated by minimizing the error term.

From Equation (2.4)

$$\min_{w_0, w_1}: \sum (y - (w_0 + w_1 x_1))^2$$

Ridge regression uses an additional term to the above equation in proportion to the weighted sum of the squared parameter to penalize the very large values of parameters and to control the variance (Hoerl and Kennard, 1970). This is also known as regularizing the coefficients(L2).

$$\min_{w_0, w_1}: \sum (y - (w_0 + w_1 x_1))^2 + \lambda \sum (w_i)^2$$

Equation 2.5 Equation for Ridge Regression Coefficients

$\lambda \rightarrow$ Shrinkage parameter.

Ridge regression reduces the model complexity and reduces variance by introducing a penalty to regression coefficients. The penalty is called shrinkage penalty as it encourages the coefficients to shrink toward zero. The shrinkage parameter lambda decides the amount by which they are encouraged. So if lambda is zero then the regression is equivalent to simple least squares. So, Ridge regression creates a number

of models with a range of values of lambda. The figure¹³ below shows a ridge regression, where the coefficients values are plotted against a range of lambda (Ridge Parameter) values.

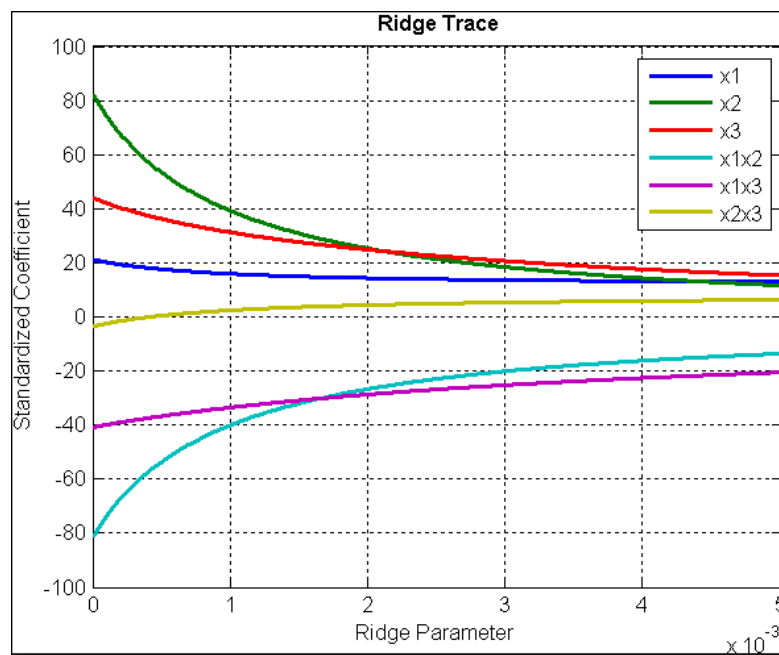


Figure 2.4 Coefficients Vs Lambda - Ridge

The optimal value of lambda is chosen by considering the overall model performance. In general, cross-validation is used to determine the value of lambda¹⁴. Ridge regression shrinks the value of coefficients towards zero but never actually replaces them with zero.

2.4.3 Lasso Regression

In contrast to Ridge, Lasso minimizes or puts a penalty (L1 Regularization) on the coefficients with the absolute values of the coefficients.

So the equation can be represented as:

$$\min_{w_0, w_1} \sum (y - (w_0 + w_1 x_1))^2 + \lambda \sum |w_i|$$

Equation 2.6 Equation for Ridge Regression Coefficients

¹³ <https://sites.google.com/site/bantimeena/software-link/regression-and-optimization>

¹⁴ <https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/courseware/8878fb6f600042fe98d774e0db26f87a/b91ee2b82a6d49eb91e1dc6641cf5efe/>

$\lambda \rightarrow$ Shrinkage Parameter

Along with shrinkage, Lasso also performs subset selection on a variable by pushing the value of the coefficient to zero. Lasso regression has been of great interest in recent years due to its ability to shrink the coefficients to exactly zero to provide a sparser solution (Tibshirani,1996). The general approach of Lasso towards bias-variance is¹⁵:

- The bias increases as λ (amount of shrinkage) increase
- The variance decreases as λ (amount of shrinkage) increases

Similar to ridge regression, lasso also provides a range of models with a corresponding value of shrinkage parameter lambda. The optimal value lambda can be chosen by any optimisation technique. The figure below gives an example of lasso for a path of coefficient created, with a corresponding value of lambda¹⁶.

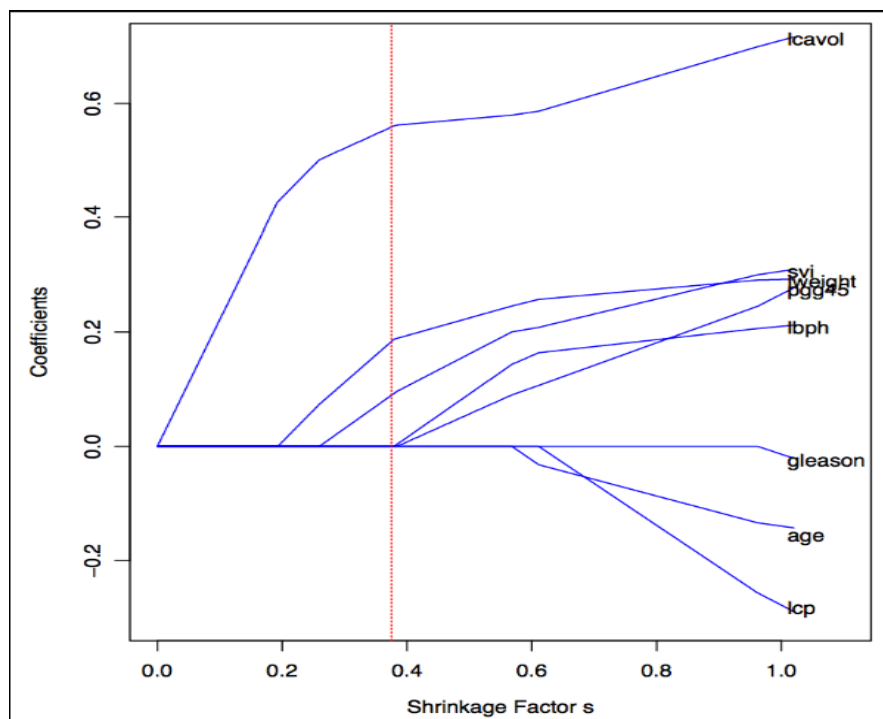


Figure 2.5 Coefficients Vs Shrinkage parameter – Lasso

¹⁵ <http://www.stat.cmu.edu/~ryantibs/datamining/lectures/17-modr2.pdf>

¹⁶ <http://andrewgelman.com/2013/03/18/tibshirani-announces-new-research-result-a-significance-test-for-the-lasso/>

2.4.4 Support Vector Regression

Although the origin of Support Vectors algorithm goes back to sixties (Vapnik and Lerner,1963; Vapnik and Chervonenkis, 1964;), Support Vector Regression (SVR) technique was proposed by (Vapnik, Steven Golowich, and Alex Smola, 1997). Support vector regression is a widely used application of the Support Vector Machine family.

The basic idea behind the SVR technique is to establish an input pattern mapped into a feature space. A dot product of input vectors is calculated to evaluate a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ under a mapping Φ . These dot products of input vectors are then added by introducing a weight criterion (α). The input patterns are then used to predict a target (y_i). The goal here is to find a function $F(x)$ that has at most a deviation of (ϵ) for the predicted value. Deviations below the value (ϵ) are ignored and the values exceeding it are not accepted (Smola and Schölkopf, 2004).

So, the function for the output will be:

$$y_i = \sum \alpha (x_i, x_j) + b \text{ with } x \in \Phi$$

Equation 2.7 Equation for Support Vector Machines

Where (x_i, x_j) is dot product in Φ .

$b \rightarrow$ constant

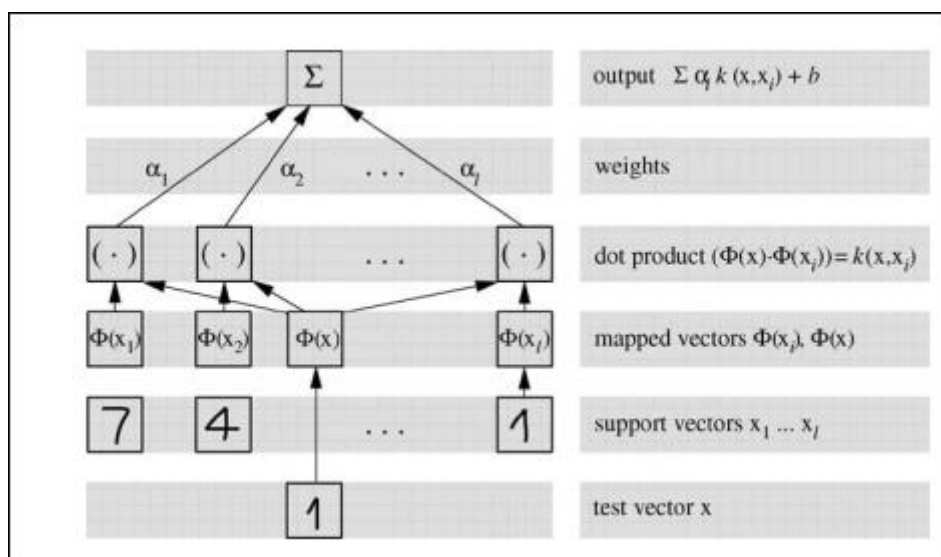


Figure 2.6 Support Vector Regression – General Architecture (Smola and Schölkopf, 2004)

In contrast to general regression techniques, instead of minimising the training error, Support Vector Regression aims to minimise the generalised bound error (ϵ) to achieve robust generalised performance . This generalised error bound is calculated using a combination of training error and a regularization term which controls the complexity of hypothesis space (Basak, Pal, and Patranabis, 2007). In order to minimise the generalised bound error, the estimated kernel function used can be both linear and non-linear in nature. The two popular types of kernel used are: (Durgesh and Lekha, 2010):

1) Linear Kernel

A linear kernel function maps the data into a linearly separable feature space and is represented in the below form:

$$K(x_i, x_j) = x_i^T x_j$$

Equation 2.8 Representation of Linear kernel

2) Radial Basis Kernel

Radial Basis Kernel function is of non-linear class. The radial basis kernel samples a high dimensional feature space with fewer hyper-parameters and less numerical difficulties (Durgesh and Lekha, 2010). These characteristics make radial basis one of the most widely used non-linear kernel function and is represented as below.

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$

Equation 2.9 Representation of Radial Basis Kernel

Where γ is kernel parameter.

2.4.5 Conclusion

This chapter has summarized a complete review of the existing literature in the context of the research study. The review comprised of the domain-specific material along with a detailed assessment of the techniques along with the research gaps and the significance of the study are coherently described. Finally, techniques which were employed in the existing research have been reviewed

3 DESIGN AND METHODOLOGY

3.1 Introduction

The research design is one of the most critical tasks in order to objectively carry out a successful research project. A design methodology lays out an objective platform that guides a research project. This chapter will provide a detailed overview of the complete experimental design about the research study. It provides an overview of the data being used for the research study. It outlines the methods and tools taken for implementation of the research study. Finally, it discussed the techniques used to address the research question in this context and the various methods to evaluate those techniques.

3.2 Research Design

The majority of the research studies can be categorised into three types: Explanatory, Descriptive, and Exploratory (Saunders, Lewis, and Thornhill, 2000).

An explanatory research is one where the researcher attempts to connect different ideas to study causation and effects¹⁷. A descriptive research is one where the researcher attempts to examine and explain a rather more complex idea or phenomenon. An exploratory research is one where the researcher seeks to understand and explore a theoretical idea. An exploratory research determines whether what is observed can be explained by a theory. Considering the nature of this study is to explore and understand the underlying theoretical ideas, it would be exploratory research.

For the course of this study, quantitative research methods will be employed. The idea of quantitative research is to estimate if a predictive theory holds true or not. Quantitative research, in general, is used to explore the quantitative properties and underlying relationships within data.

This research study can also be categorised as correlation research design. A correlation research study is used as a research design strategy in order to examine the relationships in data (Fraenkel and Wallen, 2009). The correlation analysis will be

¹⁷<http://study.com/academy/lesson/purposes-of-research-exploratory-descriptive-explanatory.html>

performed based on the experimental design using a deductive (top-down) approach where a pre-established hypothesis will be tested. According to (Saunders, Lewis, and Thornhill, 2000), deductive research is one in which the hypothesis is tested based on the existing theory.

There are two types of data that can be obtained for use in any research study: Primary and Secondary. Primary data is one where the data collection or data generation is part of the research study and the researcher creates the data himself/herself. On the contrary, in secondary data, the data used for the research is already available for use.

Since the data collection is not part of this research study and existing data will be used for the purpose of the study, the research is secondary data analysis. The data was acquired as per AMEO-2015(Aggarwal, Srikant, and Nisar, 2016).

The further sections will discuss the details of data semantics and methodology used for the research study in detail.

3.3 Data

The dataset was downloaded through the ACM dataset released as per (Aggarwal, Srikant, and Nisar 2016). The dataset was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO). The study is primarily limited only to students with engineering disciplines. The dataset contains the employment outcomes of engineering graduates as dependent variables (Salary, Job Titles, and Job Locations) along with the standardized scores from three different areas – cognitive skills, technical skills and personality skills (Aggarwal, Srikant, and Nisar 2016). The dataset also contains demographic features. The dataset contains around 40 independent variables and 4000 data points. The independent variables are both continuous and categorical in nature. The dataset contains a unique identifier for each candidate. Table 3.2.1. contains the details for the original dataset. The next section will outline the detailed data preparation and data refining steps carried out for the research work.

Table 3.1 Summary Table for dataset

VARIABLES	TYPE	Description
ID	UID	A unique ID to identify a candidate
Salary	Continuous	Annual CTC offered to the candidate (in INR)
DOJ	Date	Date of joining the company

DOL	Date	Date of leaving the company
Designation	Categorical	Designation offered in the job
JobCity	Categorical	Location of the job (city)
Gender	Categorical	Candidate's gender
DOB	Date	Date of birth of candidate
10percentage	Continuous	Overall marks obtained in grade 10 examinations
10board	Continuous	The school board whose curriculum the candidate followed in grade 10
12graduation	Date	Year of graduation - senior year high school
12percentage	Continuous	Overall marks obtained in grade 12 examinations
12board	Date	The school board whose curriculum the candidate followed in grade 12
CollegeID	NA/ID	Unique ID identifying the college which the candidate attended
CollegeTier	Categorical	Tier of college
Degree	Categorical	Degree obtained/pursued by the candidate
Specialization	Categorical	Specialization pursued by the candidate
CollegeGPA	Continuous	Aggregate GPA at graduation
CollegeCityID	NA/ID	A unique ID to identify the city in which the college is located in
CollegeCityTier	Categorical	The tier of the city in which the college is located
CollegeState	Categorical	Name of States
GraduationYear	Date	Year of graduation (Bachelor's degree)
English	Continuous	Scores in AMCAT English section
Logical	Continuous	Scores in AMCAT Logical section
Quant	Continuous	Scores in AMCAT Quantitative section
Domain	Continuous/ Standardized	Scores in AMCAT's domain module
ComputerProgramming	Continuous	Score in AMCAT's Computer programming section
ElectronicsAndSemicon	Continuous	Score in AMCAT's Electronics & Semiconductor Engineering section
ComputerScience	Continuous	Score in AMCAT's Computer Science section
MechanicalEngg	Continuous	Score in AMCAT's Mechanical Engineering section
ElectricalEngg	Continuous	Score in AMCAT's Electrical Engineering section
TelecomEngg	Continuous	Score in AMCAT's Telecommunication Engineering section
CivilEngg	Continuous	Score in AMCAT's Civil Engineering section
conscientiousness	Continuous/ Standardized	Scores in one of the sections of AMCAT's personality test

agreeableness	Continuous/ Standardized	Scores in one of the sections of AMCAT's personality test
extraversion	Continuous/ Standardized	Scores in one of the sections of AMCAT's personality test
neuroticism	Continuous/ Standardized	Scores in one of the sections of AMCAT's personality test
openness_to_experience	Continuous/ Standardized	Scores in one of the sections of AMCAT's personality test

3.4 Data Preparation

After a detailed and careful examination of the original dataset, a number of data manipulation steps were carried out to prepare the data for investigation and predictive modelling. Each of the dependent and independent variables is analysed. Non-relevant variables were discarded and removed from the dataset. The probability distributions for the features were examined using histograms and density curves to understand the variance and outliers. Data transformations are performed if the distributions are highly skewed from normal to meet the underlying statistical assumptions. Scatter plots are used to visualize the relationship between feature variables and response variables. The significance and strength of the relationships are determined using the correlation coefficients and p values. Similarly, the relationships between features are also examined.

After a detailed investigation into the semantics of Indian Education System and existential diversities, a number of data manipulation tasks were performed to prepare the data for this research. For regression modelling, the categorical features are re-coded to continuous features and new features were created. The data was partitioned into train and test sets for predictive modelling.

3.5 Assumptions

The research study is expected to make a contribution to the existing body of research in this field, therefore it is of utmost importance that the underlying assumptions for a research study must be true. As the data used is of secondary type, one of the assumptions held true would be the consistency during the data collection. Also, it is assumed that the features were collected through a random survey where the participants answered the questionnaire honestly and truthfully. Another assumption about the data is that the dataset is free from selection bias.

3.6 Data Investigation and Empirical Model

The data investigation is performed using SPSS¹⁸ and R (R Studio Team, 2015). Initial exploratory data analysis is performed using Tableau and SPSS. Tableau¹⁹ is used for data visualisation. Statistical tests (Pearson correlation) are used to test the correlation between the feature variables in the dataset for the study.

A regression analysis is a powerful statistical tool that allows for establishing relationships and characterisation within data. In brief, a regression analysis is used for:

- A statistical description of variables.
- Estimation of a response variable provided a given set of input variables.
- To determine the risk factors which can influence the response variable

The empirical model will be of multivariate linear regression:

$$\text{Salary} = w_0 + w_1(\text{Feature}_1) + w_2(\text{Feature}_2) + w_3(\text{Feature}_3) + \dots + w_n(\text{Feature}_n) + e$$

Linear regression will allow us to estimate the effect of each variable on salary. The coefficient of variables will provide us respective impact on the response variable.

Linear models are by far the most widely used technique on the subject and have provided successful results (Gerhart, 1988; Rumberger, 1993; Scholz, 1996).

The predictive regression models are build using R. Multiple regression models are built using various techniques. All the models are compared based on their accuracy using Root mean squared Error.

3.7 Evaluation and Diagnostics

In order to establish relationships, it is very important to critically evaluate a regression model structure. The aptness of a regression model is critical to derive effective inference from the model. A regression model is susceptible to misguided inference if the underlying assumptions are not met. There a number of methods available to perform diagnostics and evaluation of regression models - in one of the studies by (Alff,1984; Lommele and Sturgis, 1974), the author discusses a few standard criteria to

¹⁸ IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

¹⁹ <http://www.tableau.com/products/desktop>

evaluate and diagnose regression models. The various evaluation and diagnostics measures used for the purpose of this study are discussed below.

3.7.1 Residual Analysis

Residual analysis is by far one of the most common methods employed in research to advocate the aptness of a regression model. The following statistical assumptions have been tested using residual analyses:

- Regression function is a linear function in terms of parameters.
- The associated error/noise has a constant variance term
- The residuals are normally distributed.

Residual plots from the fit have been used to perform the analysis.

3.7.2 Evaluate homoscedasticity

One of the assumptions for regression models is homoscedasticity of the data. This can be evaluated using studentized Breusch-Pagan test or by examining the residual plots (Koenker, 1981). The distribution of residual terms is also examined to check for homoscedasticity.

3.7.3 Outlier / High Leverage points

An outlier is a data point whose response variable doesn't follow the standard distribution of the data. In contrast to the outlier, a point is a high leverage point if the data point has extreme values for the feature or input variable. Outlier and high leverage data points have a tendency to influence the regression model. Outlier tests and Leverage plots are used to examine such data points within the model.

3.7.4 Box-Cox Transform

In a seminal study by (Box and Cox,1964), the authors proposed box-cox transformation methods to ensure the usual hold for linear model assumptions. The proposed box-cox transformation holds the below form (Box and Cox,1964): ⁽²⁰⁾

²⁰ <http://www.ime.usp.br/~abe/lista/pdfm9cJKUmFZp.pdf>

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Equation 3.1 Box-Cox Transformation

The box-cox transformation tests are employed to handle the distribution assumptions.

3.7.5 Goodness-of-fit

A goodness-of-fit determines how well the selected model fits the underlying data. One of the widely adopted measures for determining the goodness-of-fit is the R-squared coefficient of determination. The coefficient is calculated as the square of the correlation between observed response values and predicted response values.

(²¹)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2},$$

Equation 3.2 Formula to calculate R- squared measure

The R^2 (R-squared) value has been used to analyse the variance explained towards the response variable (Salary of a candidate) by the input features within a model. Also, a p-value for the model fit is used to determine the significance of the model. A *p-value* lower than (<0.05) implied that the coefficients are statistically significant.

3.7.6 Root Mean Square Error (RMSE)

There are multiple measures to model performance for the prediction of outcome variable values to the actual values in regression. One such commonly used performance measure for the prediction performance of regression models is Root Mean Square Error (RMSE) (Willmott, 1981). RMSE can be calculated by taking the square root of the Mean squared error. Mean Squared error is computed using the below formula:

²¹ <https://www.otexts.org/fpp/4/4>

$$\text{MSE} = 1/n \sum (y_p - y_o)^2$$

Equation 3.3 Formula to calculate Mean Squared Error

Where n → Number of data points

y_p → Predicted Value

y_o → Observed Value

3.7.7 Data Split

In data mining applications, the source dataset is generally split into two or three parts for multiple purposes. The train set is used to train a predictive model and then a test set is used to measure performance on unseen data. The test is used to measure the accuracy of the model. Sometimes a third set, a validation set, is used for the optimisation of models (Dobbin, and Simon, 2011). A 70/30 (70% for Training and 30% hold out the sample as Test set) split will be used for this study.

3.8 Conclusion

This chapter discussed the overall design methodology for the research study. The section also provided the overview of the data and the variables used for the research. Furthermore, the chapter also outlined the assumptions and evaluation techniques employed in the study. The next chapter will provide a detailed implementation of the research experiment and the results from the experiment.

4 IMPLEMENTATION AND RESULTS

4.1 Introduction

This chapter provides a detailed description of all the experiments and tasks performed during this research study. The implementation of each experiment is discussed in detail along with the results.

4.2 Data Pre-Processing

The initial data analysis indicated that the dataset is fairly clean. The data manipulation is done using Excel and R. There are a few data manipulation steps carried out to make the data feasible for the research.

- There were a few extreme outliers in the data based on target variable Salary. Even though the number of these outliers was very low, they were causing a heavy skew into the distribution. These data points with outliers were removed from the data based on an outlier test in R. Furthermore, any high leverage points were evaluated and removed later based on residual analysis for regression models.
- The two variables ‘10board’ and ‘12board’ which represent the affiliation of school which the candidate attended, had more than 50 different levels. These variables were transformed to just two levels with values – ‘centre board’ and ‘state board’.
- The engineering domains such as aerospace engineering, biotechnology etc. with less than 20 observations were combined together as ‘Other Engineering domains’.
- There were less than 5% missing values in the dataset with no values missing for target variable Salary. The missing continuous variables were imputed using the mean value and categorical variables were labelled as Unknown.

4.3 Data Exploration and Visualisation

After data preparation, the dataset for the project contains 27 variables, both continuous and categorical.

All the dependent and independent variables in the dataset were examined individually and with respect to target variables in order to establish a better understanding of the data. Some of the key observation from the data exploration are described below.

Salary - Salary is the target variable for the research experiment. The unit of Salary is INR (Indian Rupee). The histogram below shows the distribution of Salary. The data is slightly skewed on the right.

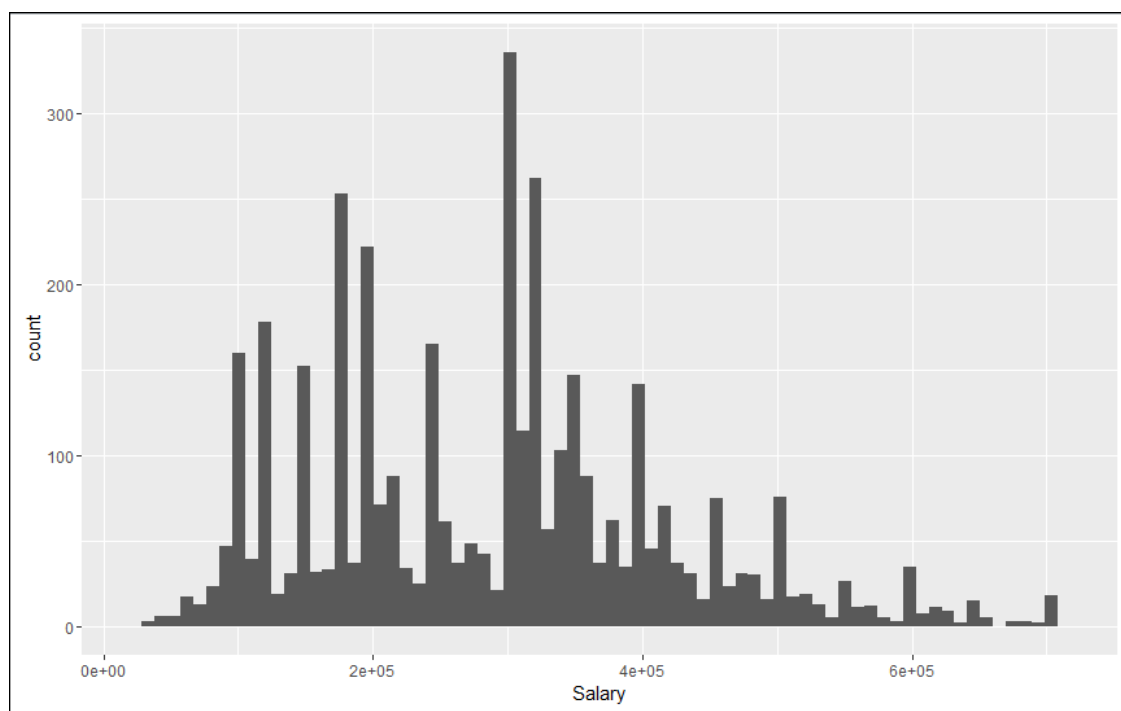
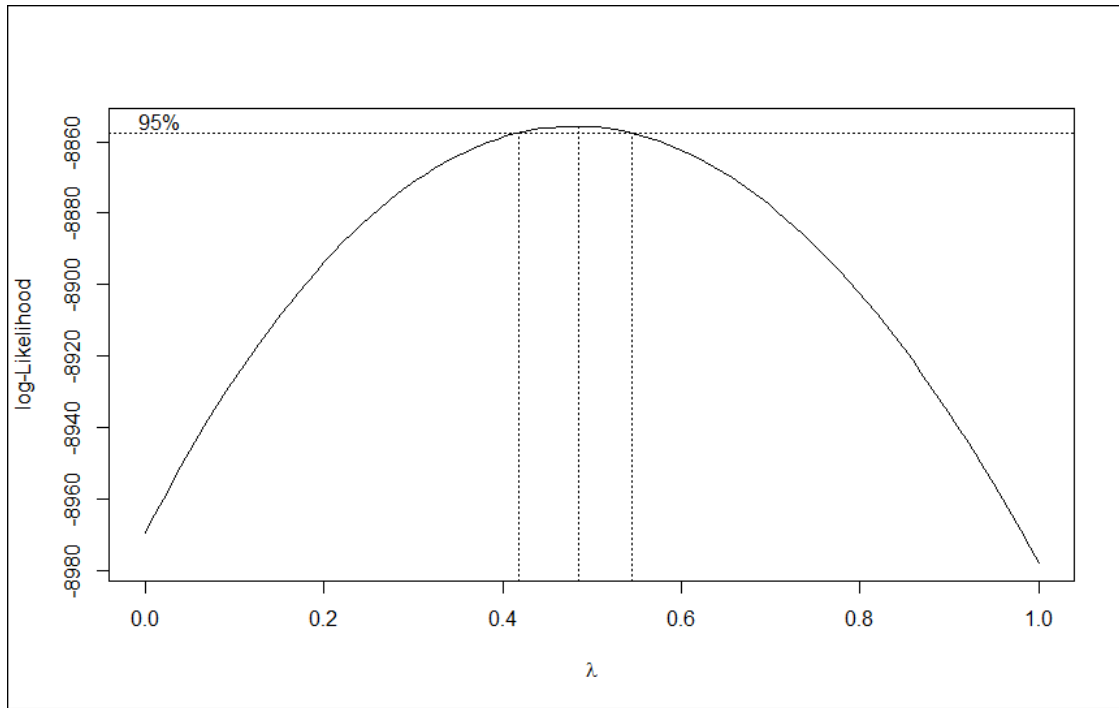


Figure 4.1 Distribution of Salary

In order to deal with skewness a box-cox transformation test is done, taking Salary as a target. The output of the box-cox suggested a transformation with a λ value of 0.5 (Figure below) for linear modelling.



The descriptive Statistics for Salary are summarised in the below table.

Table 4.1 Descriptive Statics Summary for Salary

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Salary	35000	180000	300000	283237	360000	705000

Salary and Academic Variables:

Scatter plots between the Salary and academic variables are examined. The Pearson correlation coefficients show a weak positive correlation between the Salary and academic variables: 10percentage,12percentage, and collegeGPA.

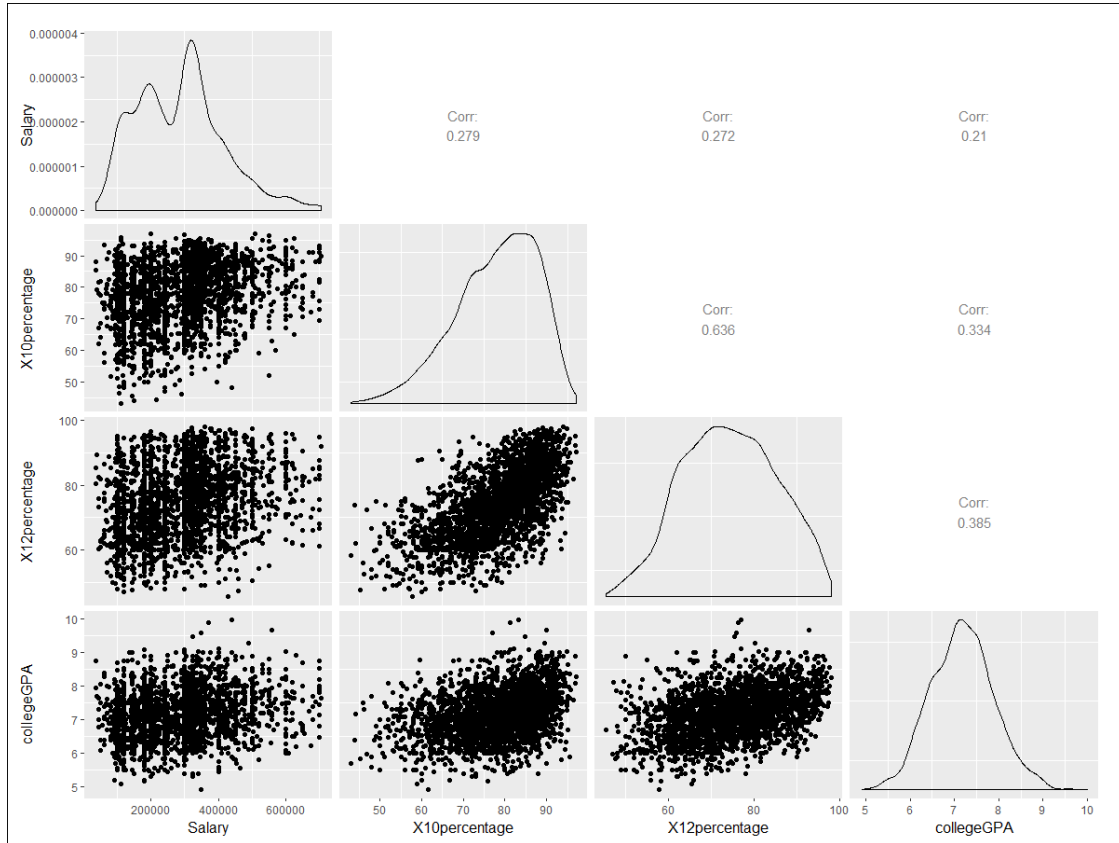


Figure 4.2 Scatter plots, Density plots and Correlation Coefficients for Salary and Academic Performance Variables

The below tables provides the summary statistics for academic variables.

Table 4.2 Summary Statistics for Academic Performance Variables

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
10percentage	43.00	71.60	79.00	77.88	85.60	97.12
12percentage	43.42	66.00	74.14	74.41	82.40	98.20
collegeGPA	4.907	6.665	7.172	7.166	7.627	9.993

Salary and Cognitive Skills

Scatter plots between the Salary and academic variables are examined (Below Figure). The Pearson correlation coefficients show a positive correlation between the Salary and Cognitive Skill variables: English, Logical, and Quant. It is also evident that there is a correlation between the cognitive variables.

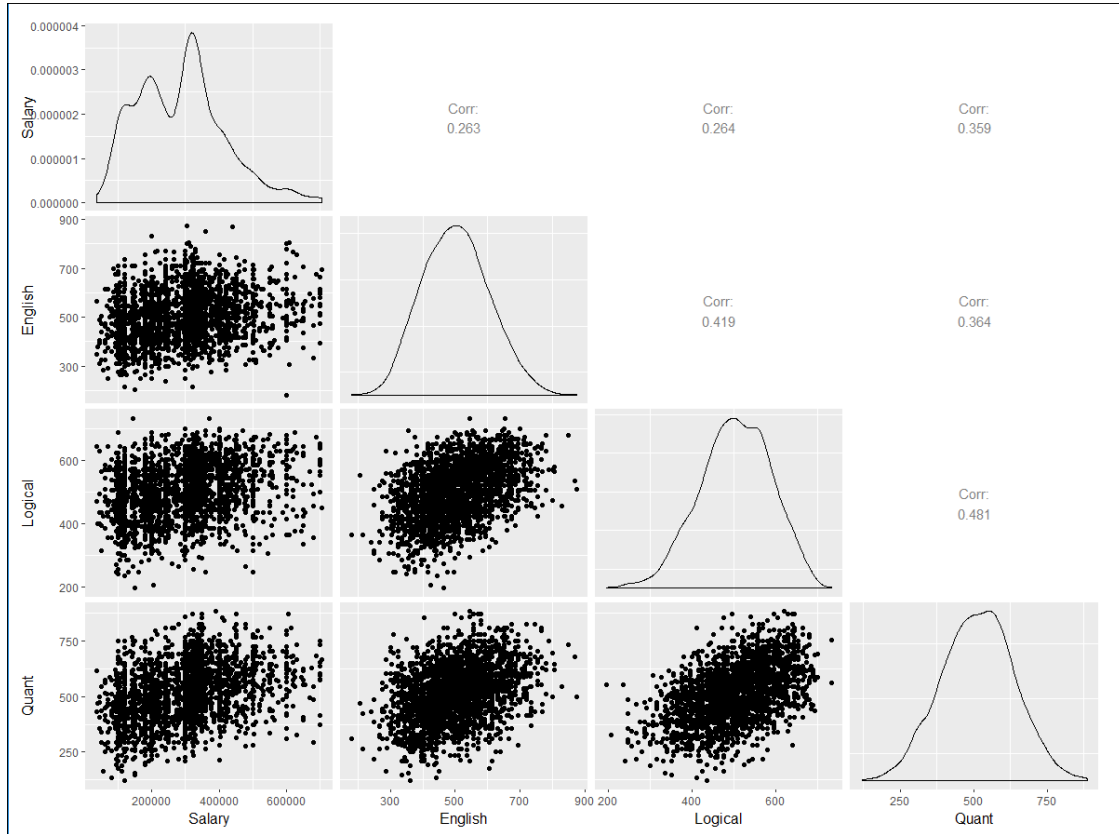


Table 4.3 Scatter plots, Density plots and Correlation Coefficients for Salary and Cognitive Skill Variables

Summary statistics for the Cognitive skill scores are provided in the below table:

Table 4.4 Summary Statistics for Cognitive Skills

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
English	43.00	71.60	79.00	77.88	85.60	97.12
Logical	43.42	66.00	74.14	74.41	82.40	98.20
Quant	4.907	6.665	7.172	7.166	7.627	9.993

Salary and Standardized Test Scores

The standardised test scores from the AMCAT test scores are examined against salary using scatter plots and Pearson correlation coefficient. Interestingly, the engineering domain scores have a weak positive correlation with Salary. The personality scores also seem to have a very low positive correlation with Salary. In addition, a few of the personality scores are correlated to each other such as ‘openness_to_experience’,

which has a moderate correlation with agreeableness, extraversion, and conscientiousness. Conscientiousness has a moderate correlation with agreeableness.

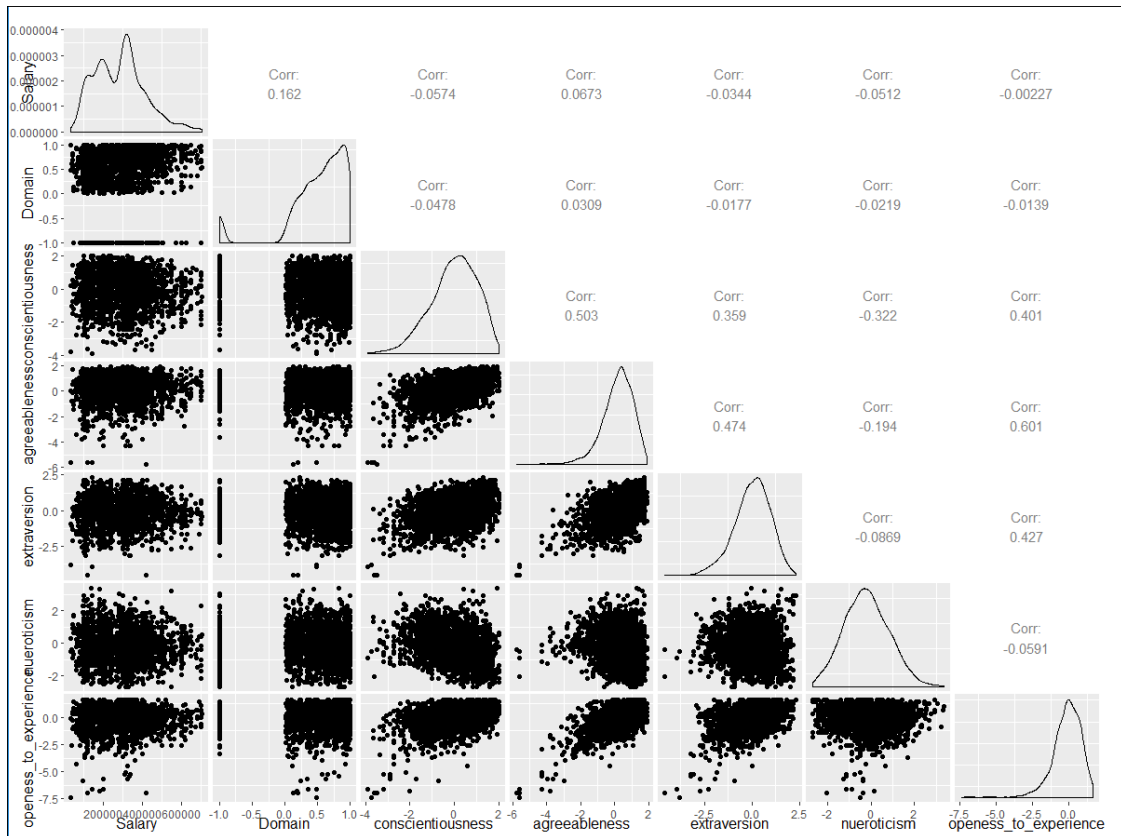


Figure 4.3 Scatter plots, Density plots and Correlation Coefficients for Salary and Standardised AMCAT scores.

Salary and Job Location

A bar plot of Salary distribution with respect to the city of job location shows that the top job destinations Mumbai, Bangalore and Pune, have a higher mean Salary than others do. Another key observation here is Kolkata, despite being one of the four Metro cities in India it has the lowest average Salary.

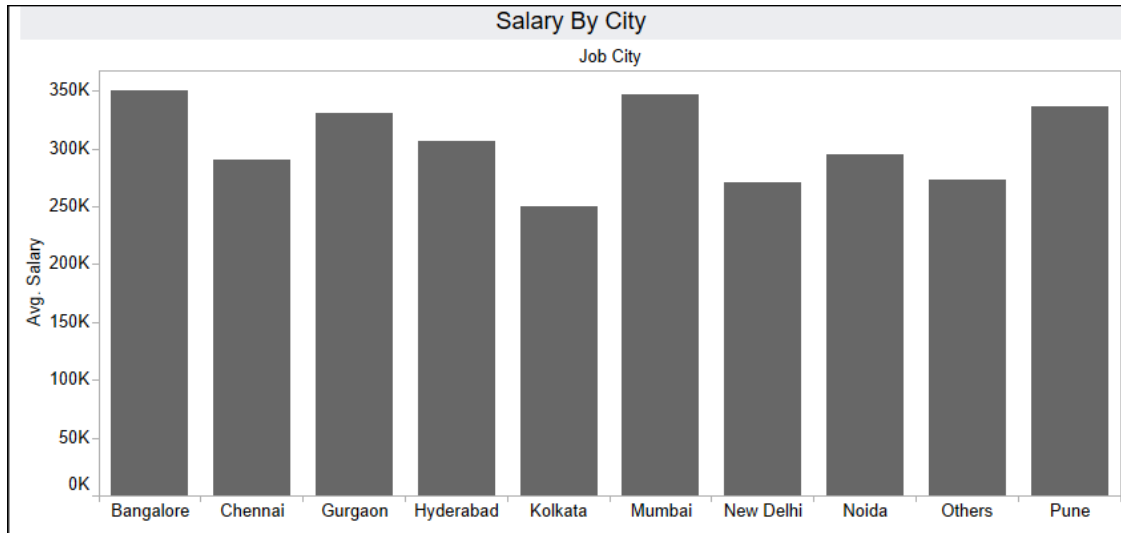


Figure 4.4 Salary Distribution by Job Location

Salary and Graduation Year

On examining the year of graduation with the Salary there are some interesting trends in data. The data shows that there is a continuous increase in the salary for the students graduating from the year 2007 until 2010. Then there is a decrease in the salary up to 2014 and an increase towards the year 2015. On a breakdown of Salary by specialization, the trend is approximately the same.

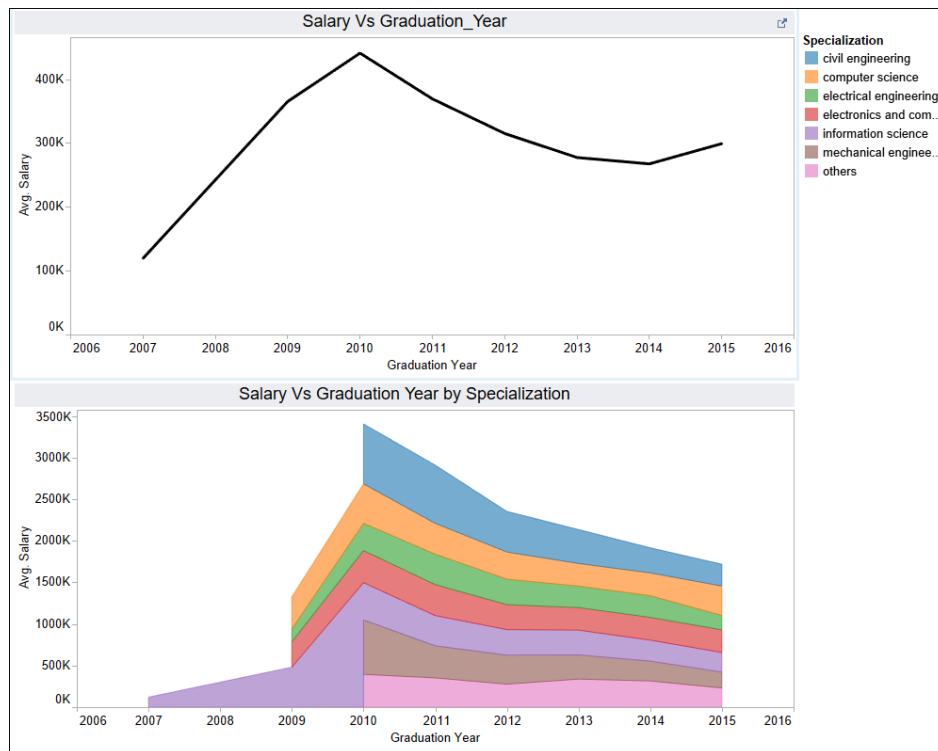


Figure 4.5 Salary Trend based on Year of Graduation

4.4 Comparison and Regression Analysis

4.4.1 Mean Salary Comparison Based on Gender

The salary variable is examined with respect to the gender variable.

The below table shows the average salaries and the respective standard deviations.

Table 4.5 Mean and Standard Deviation of salaries for both groups

Salary (INR)	Male	Female
Mean	290548.12	281439.15
Standard Deviation	133020.597	122613.199

A t-test for equality of means holds the null hypothesis with a p-value of 0.052 (> 0.01). Hence, the data shows no statistical evidence of gender bias in the salaries of entry level engineering graduate's salaries.

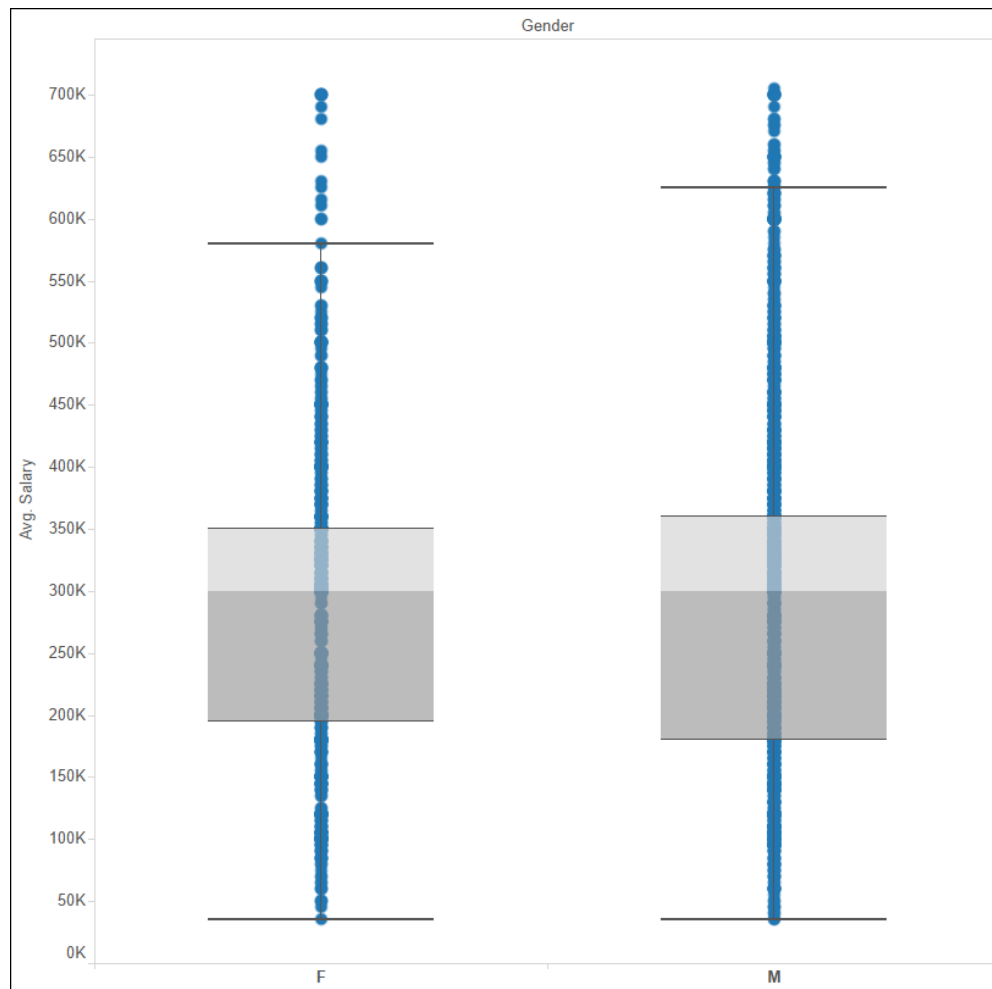


Figure 4.6 Box plot for Salaries for Male and Female groups

Conversely, the starting salary for engineering graduates did not differ by gender. There is the only small insignificant difference from a mean value of INR 281439 to INR 290548.

4.4.2 Mean Salary Comparison Based on Engineering Specialization

The choice of engineering major is generally considered to be one of the important factors for the employability of a graduate. The table below shows the individuals mean and standard deviation for salary each of the specialization.

Table 4.6 Mean Salaries by choice of engineering Specialization

Specialization	Mean (Salary INR)	Standard Deviation (Salary INR)
civil engineering	339038.5	25591.64
computer science	290128	3300.694
electrical engineering	270809.4	7826.406
Electronics and communication	285562.7	3963.42
information science	294284.7	5011.525
mechanical engineering	279198.1	9069.84
Others	309012.3	14499.14

The majority of engineering graduates join the huge IT (Information Technology) markets after graduating due to the huge demand in the sector. As such, it is expected that the graduates majoring from computer science and information science will be paid relatively more than those from other majors. That said, a pairwise comparison of

all specialization groups does not show any significant difference between the average salaries with a p-value greater than 0.01.

Even though there is no statistically significant difference between the average salaries between different engineering groups, it can be observed in the box plot (Figure 4) comparison below that civil engineering graduates have a high lower quartile range compared to other groups.

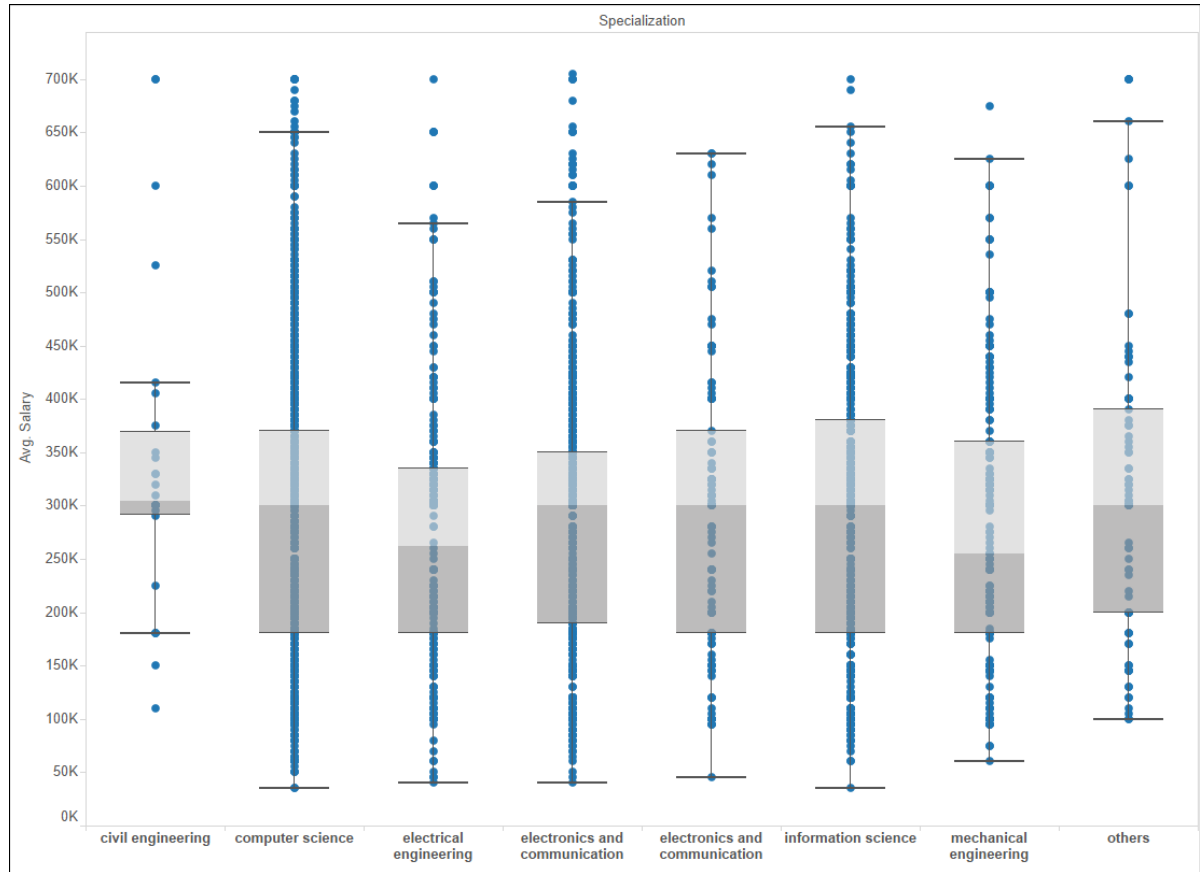


Figure 4.7 Box plot for Salaries by Engineering Specialization

4.4.3 Mean Salary Comparison Based on College Tier

The table below shows the average salaries of a student based on the college tier they graduated from. Also, a t-test for equality of means supported the evidence to reject the null hypothesis with a p-value of .000 (< 0.01) with t-statistic of 12.113 with a degree of freedom of 3915.

Table 4.7 Summary for Mean Salary and Standard Deviation based on College Tier.

College Tier	Mean Salary (INR)	Standard Deviation
A	379227.94	129248.293
B	281569.00	128190.135

The statistical test supports the general notion that the students graduating from Tier A college are paid more than the student's graduating from Tier B (Appendix B).

4.4.4 Regression Analysis for Cognitive Skills and Salary

A multiple regression was conducted to examine the relationship between the cognitive skill test scores towards the Starting Salary. The cognitive skill scores consist of three variables – English, Logical, and Quant.

The correlation tests (Pearson – Heat Map Below) between Salary and cognitive test scores showed a weak but statistically significant correlation.

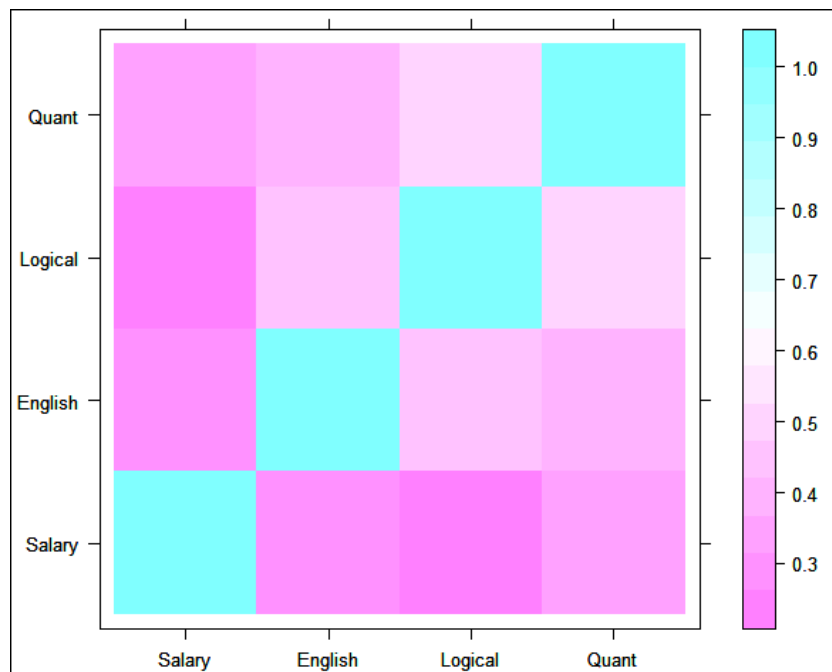


Table 4.8 Correlation heat map for Cognitive skills and Salary

A multivariate regression model was built and the residual plots were examined to verify the validity of the model. The figure below shows all the residual plots for the regression model.

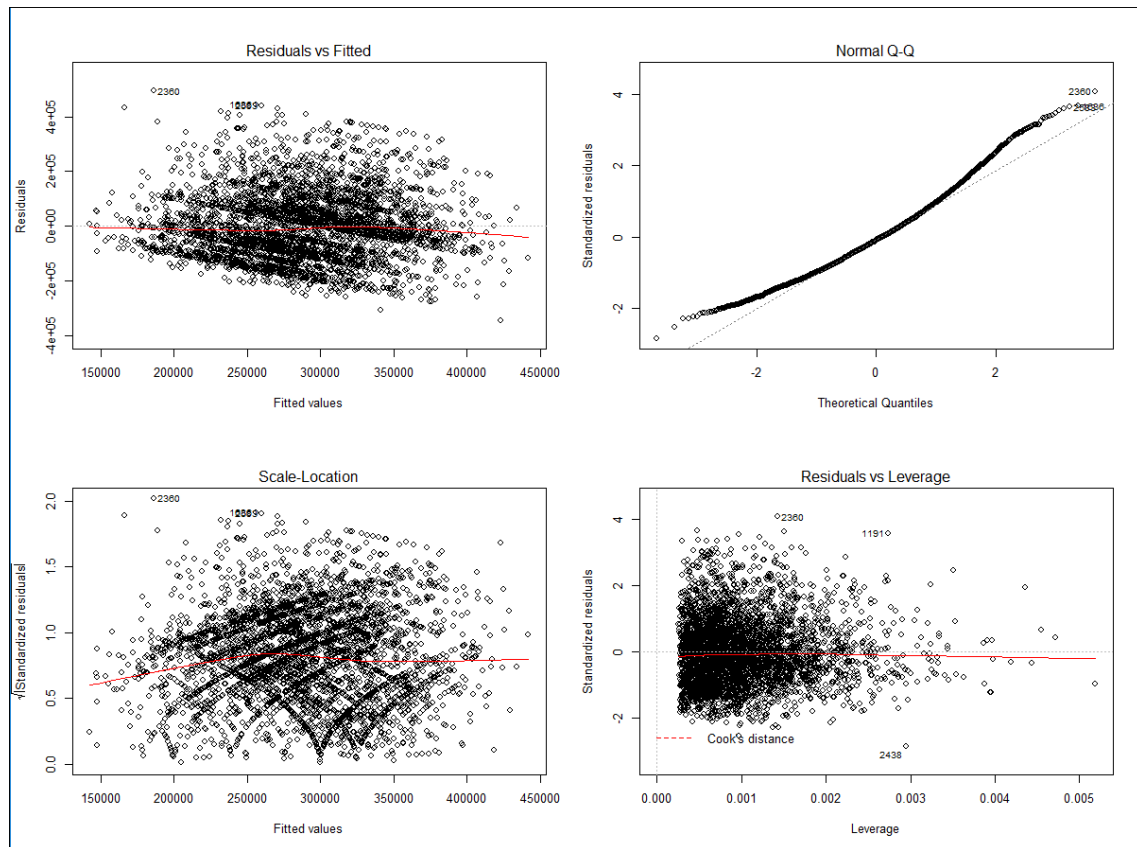


Figure 4.8 Residual plots of the Regression Model for Cognitive skills and Salary

Residual Plot Analysis:

- The top-left (Figure 5) plot for the ‘Residual Vs Fitted’ values shows an even distribution around the zero mean following the regression assumptions.
- The top-right (Figure 5) QQ plot follows the normal distribution assumption for residuals for the model. There are data points on both ends which reflect a little deviation from normal which is because of the infrequent width of the salary data points on both ends.
- The bottom-left (Figure 5) plot is another plot of ‘Fitted Vs Residual’, except that the residuals have been standardized here. The plot shows that the model follows the assumption of homoscedasticity.
- The bottom-right (Figure 5) plot shows that there are three high leverage points in the model. The model was re-created without these data points and the same results were produced, so in the final model these data points were not removed.

A regression model with three predictors indicated that Cognitive Skill scores are statistically significant (p-value < 0.001) predictors of Salary. The table below shows the regression results supporting the same (Appendix B).

Table 4.9: Regression Summary for Cognitive Skill and Salary

Adjusted R-Squared (R²)	F - statistic	p-value
0.1389	211.6	2.2e-16

The adjusted R-squared value states the total variance explained by the independent variable towards Salary. The absolute t-statistic (|t|) value indicates that Quant is a stronger predictor of Salary in comparison to English and Logical ability scores. The table below summarizes the Regression coefficients summary for the model.

Table 4.10 Regression Coefficients for Cognitive Skill and Salary Model

Variable Name	Regression Coefficient	Std. Error	T - statistic	p- value
English	159.94	21.01	7.614	3.31e-14 ***
Logical	91.56	27.26	3.358	0.000791***
Quant	286.65	18.75	15.285	< 2e-16***

* Indicates Significance at the .05 level

** Indicates Significance at the .01 level

*** Indicates Significance at the 0.001 level

The regression models with the interaction terms within cognitive variables are also evaluated to examine the interaction effects. The residual plots for each of the regression models are evaluated for model validation (APPENDIX A). The results from the regression models with interaction terms are summarised in the below table.

Table 4.11 Regression Summary for Cognitive Model with Interaction Terms

Variable	Adjusted R-Squared	Regression Coefficients	p- value
Model -1			
English	0.1493	1.673e+02	6.94e-15 ***
Logical		4.897e+01	0.6066
Quant		2.362e+02	0.0117 *
Logical:Quant		9.525e-02	0.5993
Model -2			
English	0.1496	48.4490	0.667
Logical		-17.6822	0.872
Quant		284.3922	<2e-16 ***
English:Logical		0.2329	0.283
Model -3			
English	0.1493	1.959e+02	0.015518 *
Logical		9.621e+01	0.000634 ***
Quant		3.119e+02	5.84e-05 ***
English:Quant		-5.475e-02	0.714051

* Indicates Significance at the .05 level

** Indicates Significance at the .01 level

*** Indicates Significance at the 0.001 level

In each of the individual regression models with interaction terms, none of the interaction terms were found to be statistically significant.

4.4.5 Regression Analysis for Cognitive Skills, Gender and Academic Features

The cognitive skills are also analysed using regression with Gender as a control variable. The regression model is significant with a p-value of < 2.2e-16 (< 0.001) with an Adjusted R-squared value of 0.1494. The coefficients of from the regression are summarised in the below table.

Table 4.12 Regression Coefficients of Cognitive Skills controlling for Gender

Variable Name	Regression Coefficient	Std. Error	T - statistic	p- value
English	167.68	21.40	7.835	6.15e-15 ***
Logical	97.16	28.12	3.456	0.000556 ***
Quant	283.26	19.03	14.884	< 2e-16 ***
GenderM	2564.45	4689.54	0.547	0.584519

* Indicates Significance at the .05 level

** Indicates Significance at the .01 level

*** Indicates Significance at the 0.001 level

The results from the regression indicate that the control variable Gender is statistically insignificant in the model with cognitive skill features.

Another regression model is built for cognitive skills including the academic variables. The regression results are significant with a p-value of < 2.2e-16 (<0.001) with an Adjusted R-Squared value of 0.1922. The regression coefficients from the model are summarised in the below table.

Table 4.13 Regression Coefficients for Cognitive skills and Academic variables model

Variable Name	Regression Coefficient	Std. Error	T - statistic	p- value
English	112.62	22.24	5.064	4.33e-07 ***
Logical	45.39	27.73	1.637	0.1017
Quant	212.71	19.20	11.081	< 2e-16 ***
X10boardstate board	-10277.58	4190.45	-2.453	0.0142 *
X12boardstate board	19738.18	23732.47	0.832	0.4056
X10percentage	925.52	269.47	3.435	0.0006 ***
X12percentage	1302.69	243.14	5.358	8.96e-08 ***
collegeGPA	1151.83	263.66	4.369	1.29e-05 ***

CollegeTierB	-46981.59	7753.08	-6.060	1.51e-09 ***
---------------------	-----------	---------	--------	-----------------

* Indicates Significance at the .05 level

** Indicates Significance at the .01 level

*** Indicates Significance at the 0.001 level

The results from the regression indicated that the Logical ability and ‘X12 Board’ are insignificant when the cognitive skills and academic variables are combined. The cognitive skills ‘English’ and ‘Quant’ are still significant in the model towards predicting Salary. The academic variables ‘X10board’, ‘X10percentage’, ‘X12Percentage’, ‘collegeGPA’ and ‘CollegeTier’ are significant contributors in predicting Salary. In addition, the students from state board affiliation (X10board) and Tier B (CollegeTier) colleges have a negative contribution in predicting the Salary.

4.4.6 Regression Analysis using full set of variables

A complete multiple linear regression was examined to understand the salary predictors using demographic variables, cognitive skills scores, academic performance and personality scores. The categorical variables were re-coded by creating additional (k-1) variables for k levels of each variable. Such variables are: Gender, Specialization, 10board, 12board and CollegeTier.

The regression results were statistically significant with a p-value of 2.2e-16 (<0.001) and an adjusted R-squared value of 0.243. The F-statistic value for the model is 35.71. Many of the variables were highly significant and as were many control variables. This suggests that there are several aspects for students that significantly affect the starting salary of engineering graduates. The below table shows the regression coefficients and t-statistics summary from the regression model (Only significant variables).

Table 4.14 Regression Coefficients from the complete model (Absolute Values)

Variables	Coefficients	T- Statistics
GenderM	20436.84	4.364***
X10percentage	1183.05	4.415***
X10boardstate board	10234.41	2.448*
X12percentage	1358.13	5.528***
CollegeTierB	42058.37	5.475***
Specializationcomputer science	121408.21	2.823**
Specializationelectrical engineering	100701.34	2.241*
Specializationelectronics and communication	110726.42	2.428*
Specializationinformation science	122268.05	2.547*
Specializationothers	120799.73	2.620**
collegeGPA	1382.69	5.195***
English	123.34	5.611***
Quant	175.50	9.137***
Domain	12180.60	2.310*
ComputerProgramming	33.69	2.659**
ComputerScience	122.50	10.784***
ElectricalEngg	94.15	3.250**
CivilEngg	336.67	3.297***
conscientiousness	7672.51	3.381***

* Indicates Significance at the .05 level; ** Indicates Significance at the .01 level

*** Indicates Significance at the 0.001 level

Cognitive Skill Scores variables (English, Logical, and Quant):

The results of regression found that English and Quant's scores are significant predictors for the starting salary for engineering graduates. The model also revealed that the variable 'Logical' is not significant anymore in the complete model when controlled for with other variables. Also, the t-statistic values state that the Quant Score is still relatively a strong predictor compared to English score.

Academic Variables:

The regression results indicated that the academic performance variables such as 10percentage, 12percentage and collegeGPA have a significant impact on the salary of an engineering graduate. In addition, the choice of 10board, CollegeTier and Engineering Specialisations also presented to have a significant contribution towards the Salary.

Standardised Test scores for Domain and Personality:

The results from the regression indicated that even the scores in domain specific tests for Computer programming, Computer Science, Electrical Engineering and Civil Engineering contributed significantly towards Salary. On the other hand, the only personality score (of BIG5 Personality Test) which had a significant effect on Salary is conscientiousness. The rest of the Personality Test scores are insignificant in the model.

4.5 Predictive Modelling

In order to build an accurate salary prediction model, multiple regression models are created using feature selection and regularization techniques. The regression models are built using R Statistical Package. All the models are compared based on the Root Mean Square Error (RMSE) on the test set. The dataset was split into (70:30) as training and test set. In this study, the below steps are used to build models.

Table 4.15 Steps to build Regression Models

Steps	Description
1.	<ul style="list-style-type: none">• Import Data and split (70:30) as Training and Test Set.
2.	<ul style="list-style-type: none">• Train the model on the training set.
3.	<ul style="list-style-type: none">• Select Best Model using Feature Selection/ Regularization
4.	<ul style="list-style-type: none">• Apply parameter engineering to improve performance

5.	<ul style="list-style-type: none"> • Validate Model and Calculate RMSE
----	---

4.5.1 Baseline Multiple Linear Regression Model

A baseline model is build using all of the predictor variables. The model is trained on the training split. The residual plots are then examined to validate the model assumptions. The Figure below contains the residual plot of the baseline regression model.

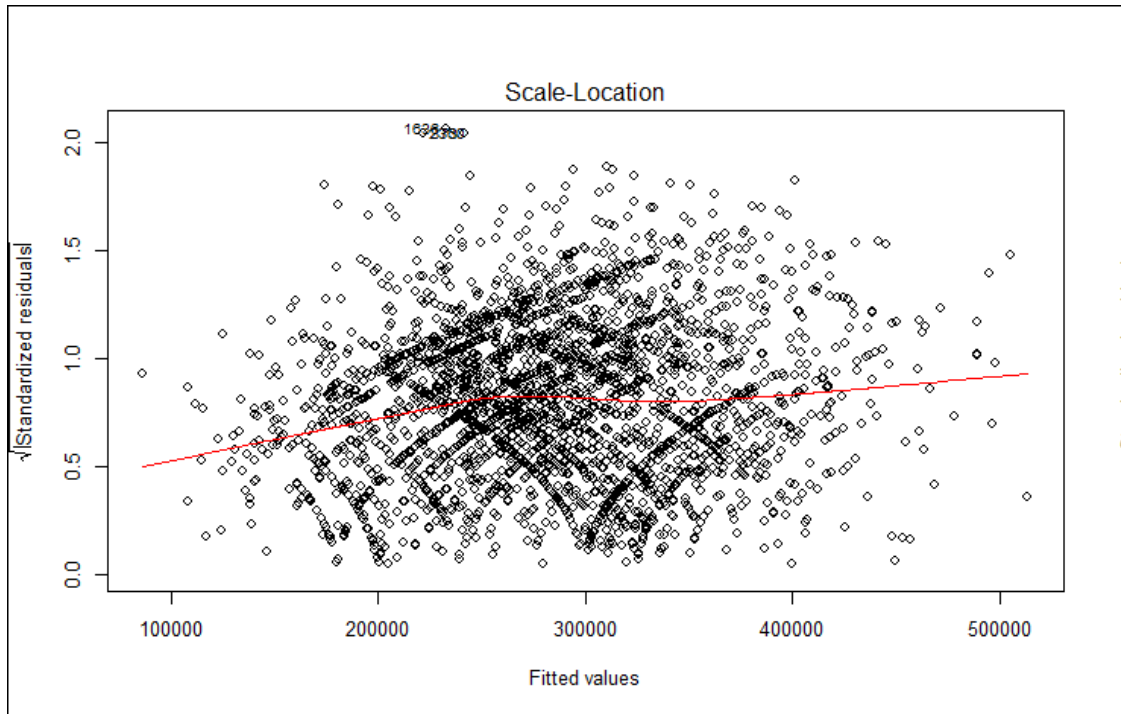


Figure 4.9 Residual plots for the Baseline Regression Model

The ‘Standardized Residual Vs Fitted’ plot holds the regression model assumption of constant variance.

The trained baseline model is then used to make predictions on the test set and Root Mean Square Error is calculated. The RMSE of the baseline model on the test set by 70/30 method is 144194.3.

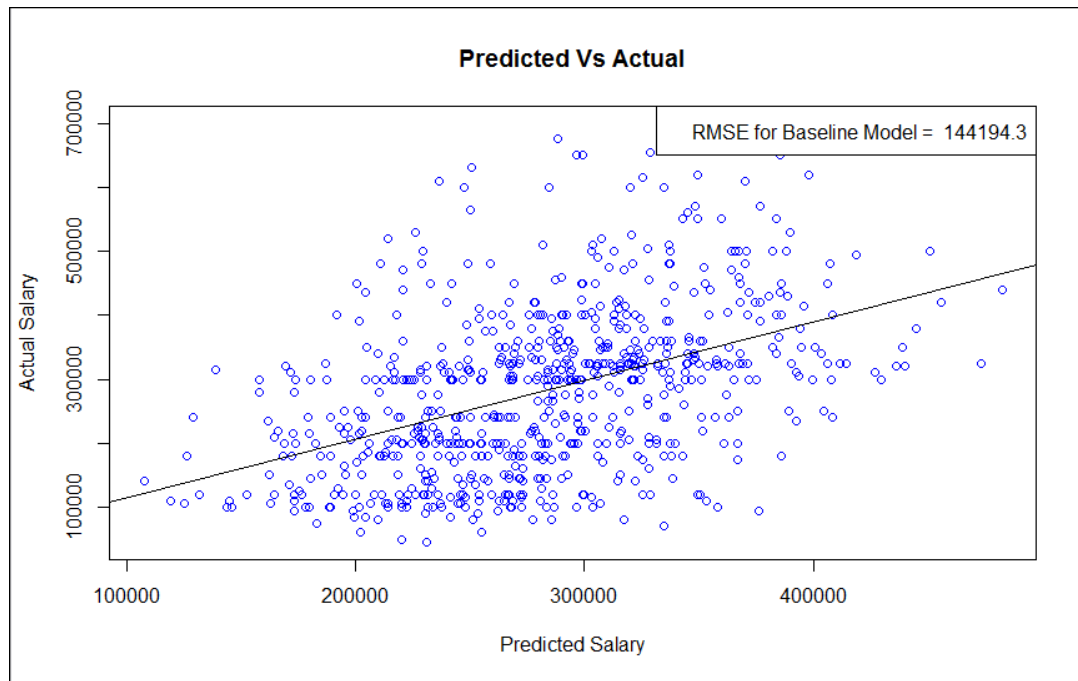


Figure 4.10 Predicted Vs Actual Salary on Test Set

4.5.2 Model Selection using Forward Stepwise Selection

A Forward Stepwise Subset feature selection is employed to improve on the baseline model. Forward Stepwise selection in each iteration includes the next best variable for the model. It creates a nested sequence of models by including the variable that improves the model most at each step.

A stepwise selection method is used to train a total of 32 models on the training data using 70/30 method. The best model is selected with the minimum value of Cp-statistic value. Mallows' Cp-statistic is one of the most commonly used measures to compare all possible regressions and select the best model among them (Gilmour,1996).

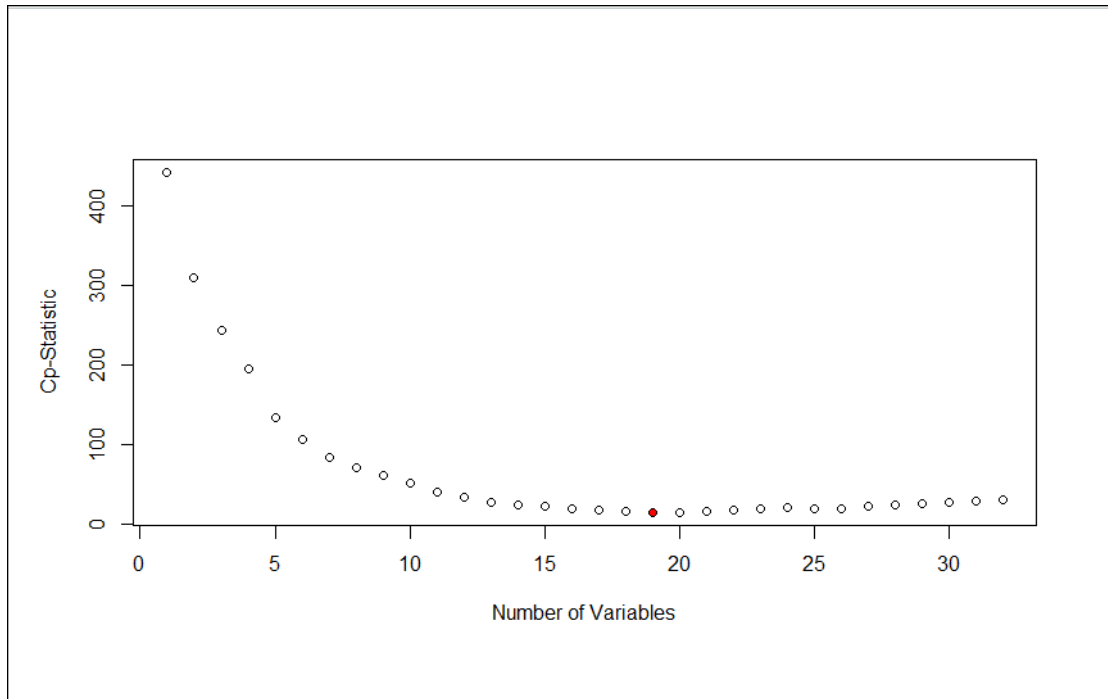


Figure 4.11 Models trained by Stepwise Method against the Cp- Statistics.

A final model with 19 variables is selected with the lowest Cp-statistic (Figure Above) from stepwise forward selection method. The below figure shows the list of variables included in the final model.

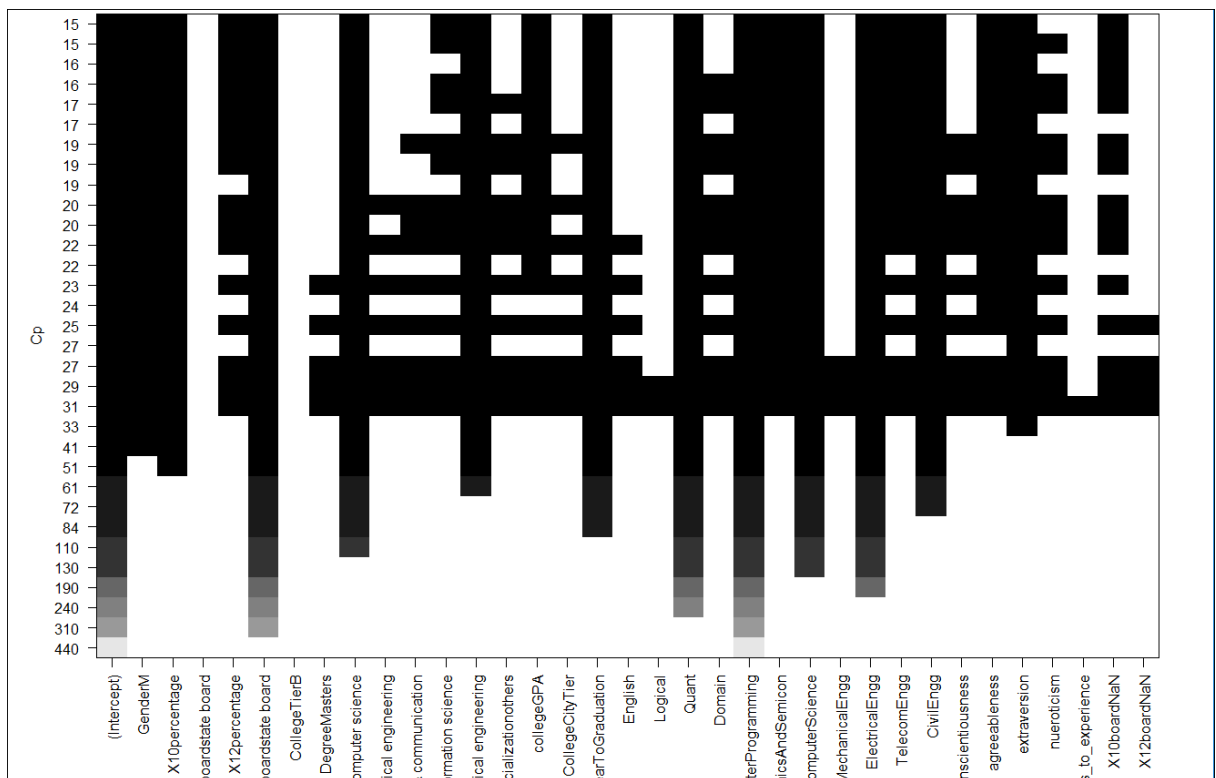


Figure 4.12 Variable selection in the selected model

The model is then used to make predictions on the test set (From 70/30 split method) and RMSE is calculated for the same. The final RMSE from the Stepwise Regression Model on the test set is 126320.9. The RMSE for training and test is plotted in the below figure.

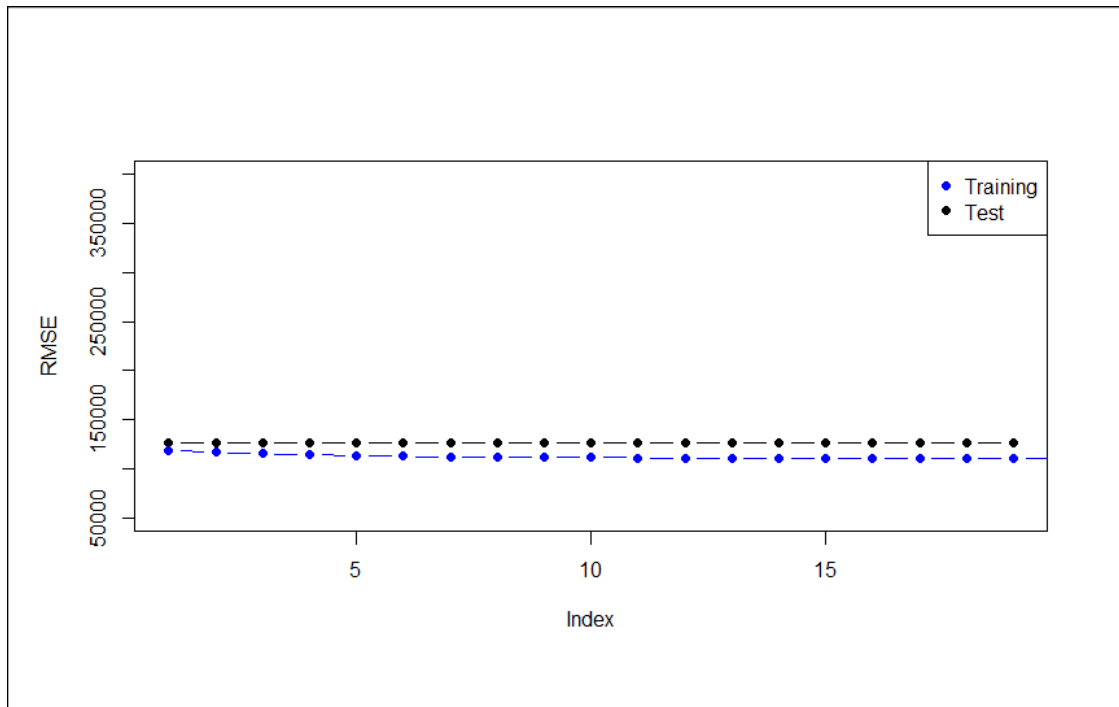


Figure 4.13 RMSE for Training and Test Set

4.5.3 Model Selection using L1 Regularization (Lasso Regression)

In order to perform least absolute shrinkage and variable selection, a Lasso Regression Model is fitted using ‘glmnet’ package (Friedman, Hastie, and Tibshirani, 2010) of R. Lasso Regression performs L1 regularization. It adds the absolute value of coefficients to the optimisation function for the model.

As the package does not use formal language for R, so an input matrix of predictors and a response vector is created in order to build the model. Lasso Regression Model is fitted using ‘glmnet’ function with $\alpha=0$. The below plot shows the variable coefficients and associated value of lambda.

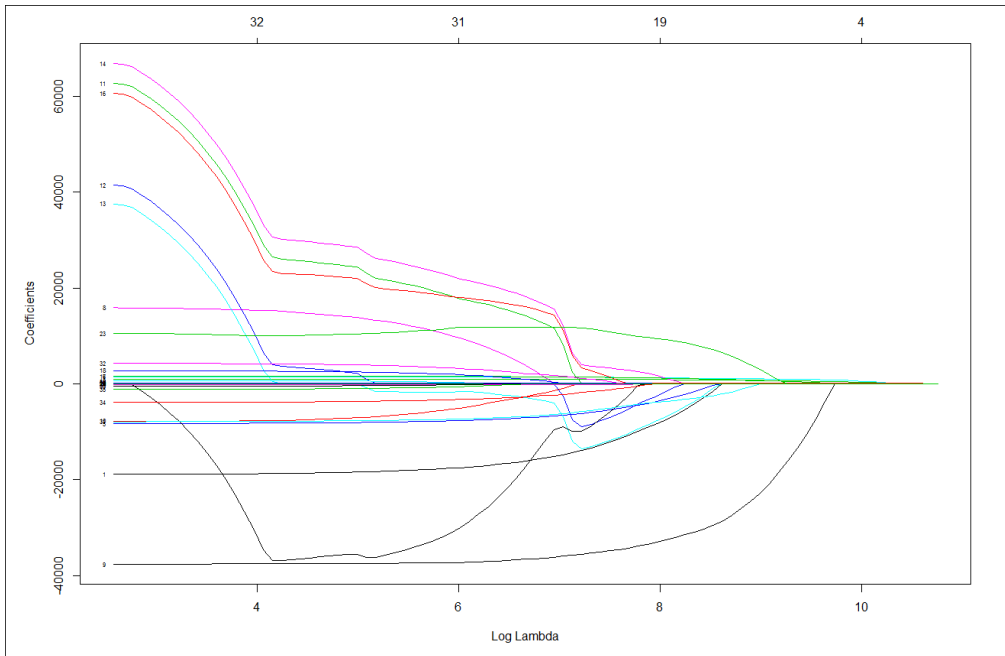


Figure 4.14 Variable Path created by Lasso Regression

Also, the deviance plot (figure below) indicates that coefficients grow very large with a small increase of 0.23 to 0.26 of the total deviance explained (similar to R-squared in Linear Regression).

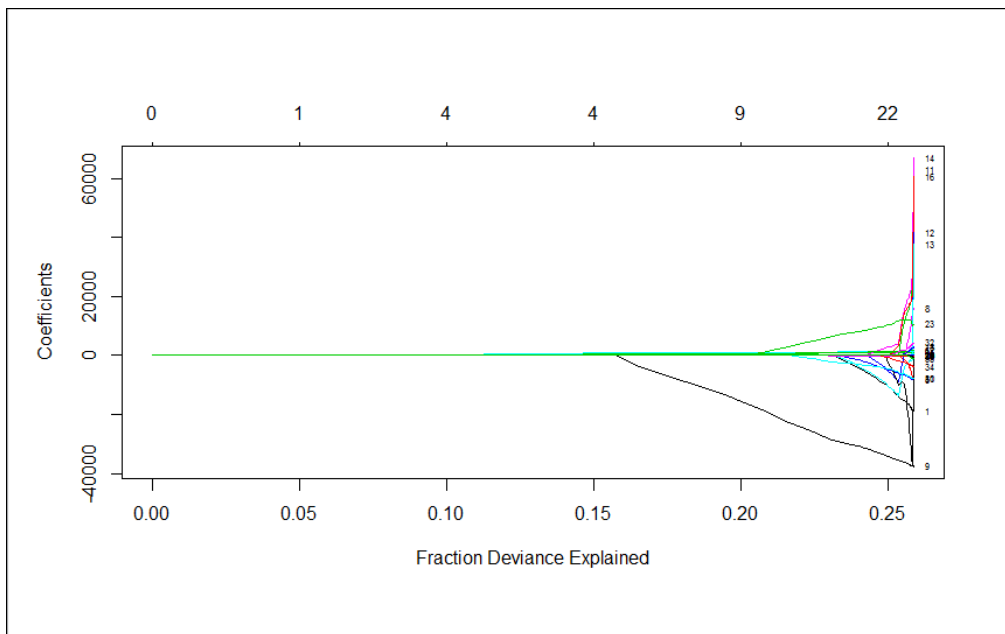


Figure 4.15 Deviance plot for Lasso Regression

Lasso will generate a wide range of possible values of coefficients indexed by different values of lambda. The best set of coefficients will be selected by choosing the corresponding lambda value.

The best value for lambda is selected using a 10-fold cross validation with a minimum mean squared error. In the below plot (Mean Squared Error Vs Log (Lambda)), it can be observed that Mean Squared Error is high in the starting (Left) and then it gradually decreases and levels off. It can be observed that after a while there are no significant decreases in the error even if the coefficients keep increasing.

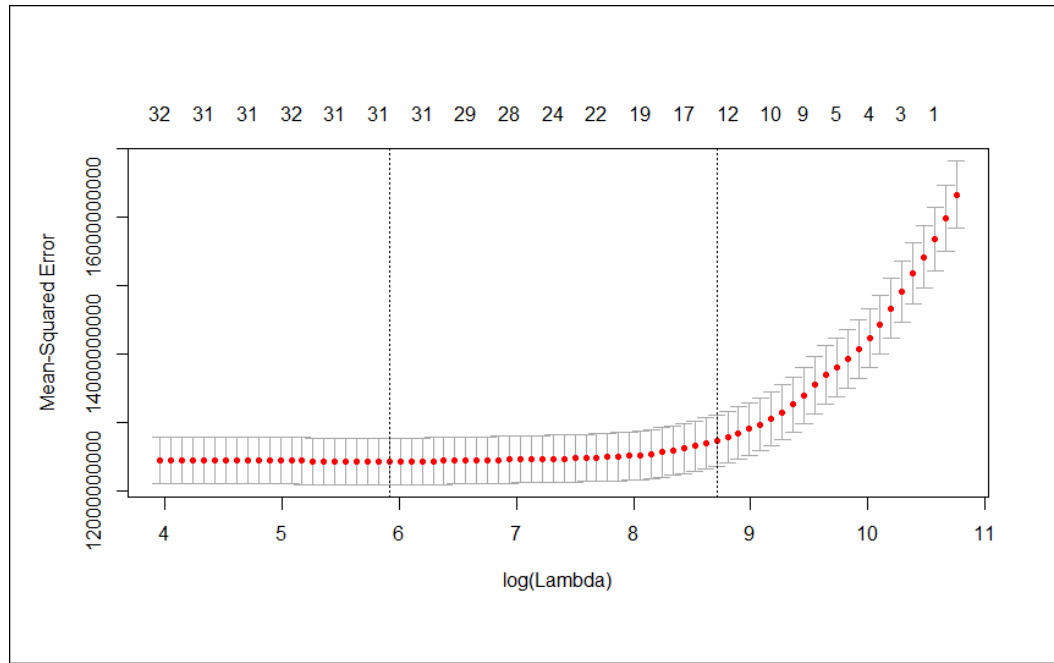


Figure 4.16 Mean – Squared Error Vs Log(Lambda)

The lambda value is selected based on the minimum value of Mean Squared Error and the simplest model. The selected model is then used to make a prediction on the test set and Root Mean Square is calculated for the same. The final RMSE from the best selected Lasso regression model on the test set is 114492.5. The final model selected had 11 variables.

4.5.4 Model Selection using L2 Regularization (Ridge Regression)

Similar to Lasso, a ridge regression model is fitted using the ‘glmnet’ package. The function for ridge regression is residual sum of squares plus lambda times the sum of squares of coefficients. The ridge-regression is fitted by calling the ‘glmnet’ function with $\alpha=0$. ‘Glmnet’ package sprays over a range of values of lambda and creates a path of variables (Figure below).

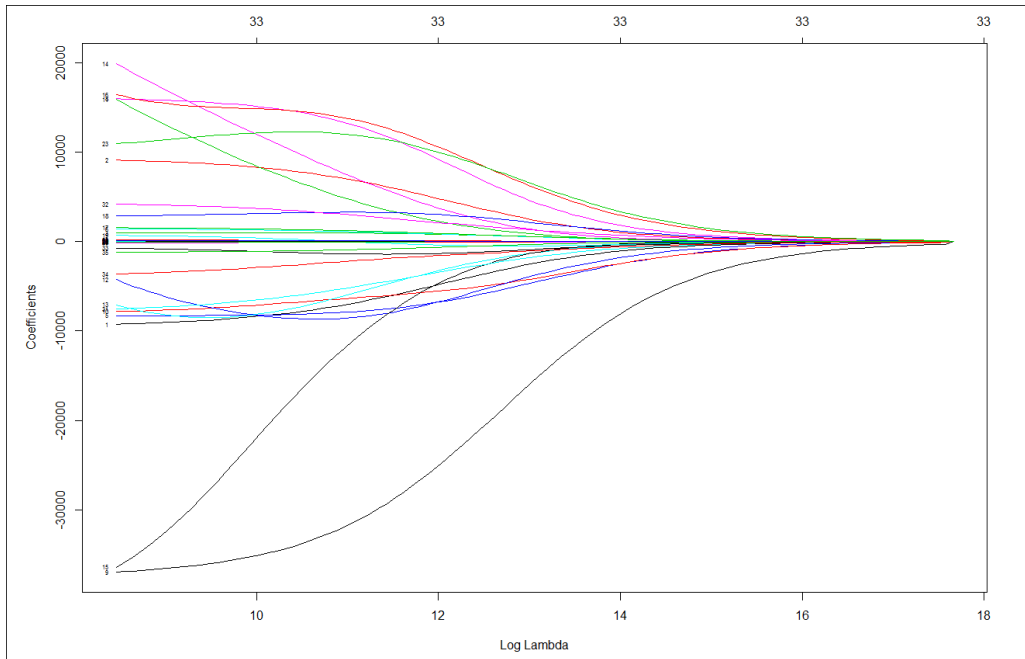


Figure 4.17 Variable Path created by Ridge Regression

The best value of the lambda is chosen using a 10-fold cross-validation on the training set. The selection of Lambda is done based on the minimum value of mean squared error. The complete set of lambda is plotted with the associated mean squared error (figure below).

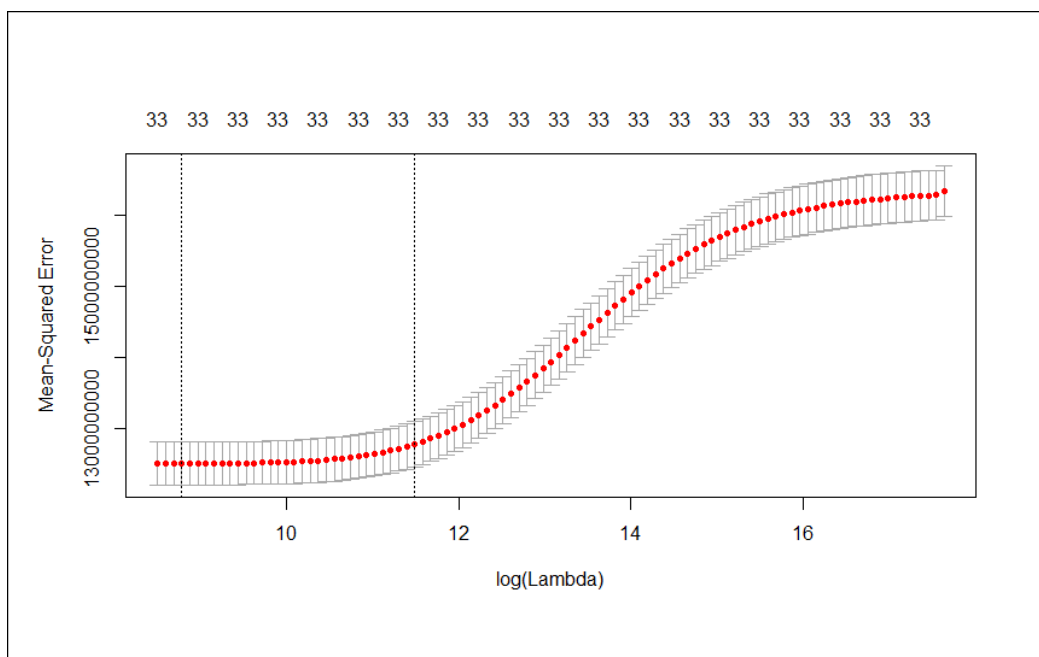


Figure 4.18 Mean-squared Error for all Lambda values for Ridge Regression

The selected model is used to make predictions on the test set (From 70/30 split method) and RMSE is calculated for the same. The final RMSE from the best selected Ridge regression model on the test set is 114459.1.

4.5.5 Support Vector Regression with Linear Kernel

As a part of the initial experimental design, a support vector regression model was also proposed to predict salary. A support vector regression model is built using 'e1071' R package (Dimitriadou et al.,2009).

The baseline SVR model is built with a linear kernel with the following parameter values:

```
Parameters:
  SVM-Type:  eps-regression
  SVM-Kernel: linear
  cost:      1
  gamma:    0.02777778
  epsilon:  0.1
```

Figure 4.19 Parameters for Support Vector Regression Model with Linear Kernel

The model is used to make predictions on the test set and RMSE is calculated: 115709. The model is then optimised for performance using a grid search hyper-parameter optimisation. The method uses a 10-fold cross-validation to tune a number of models by adjusting the values of epsilon and the cost parameters. The figure below shows the improved error with the darker shaded region.

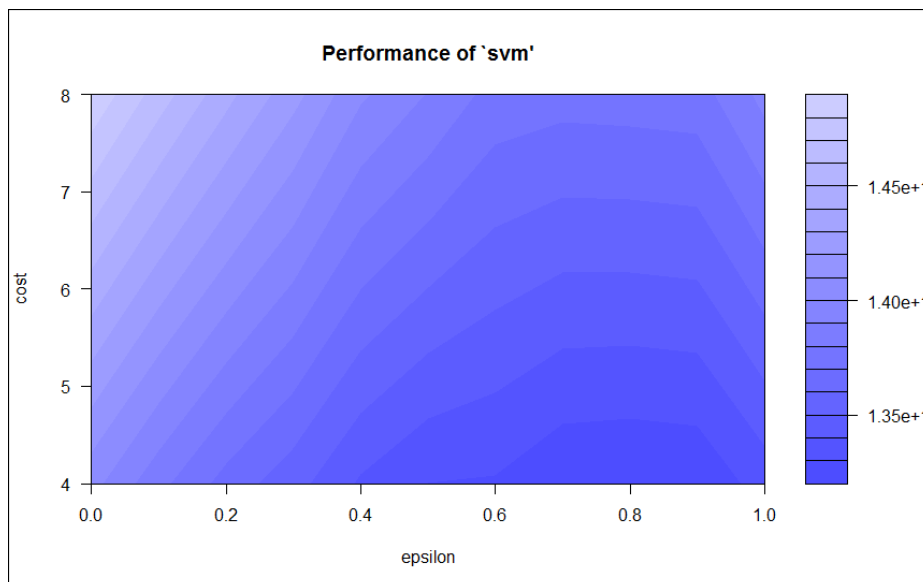


Figure 4.20 SVR – Linear Kernel Hyper-Parameter Optimisation

The model with the below parameters is selected after the grid search for the optimal parameter values.

```
Parameters:  
SVM-Type: eps-regression  
SVM-kernel: linear  
cost: 4  
gamma: 0.02777778  
epsilon: 0.7
```

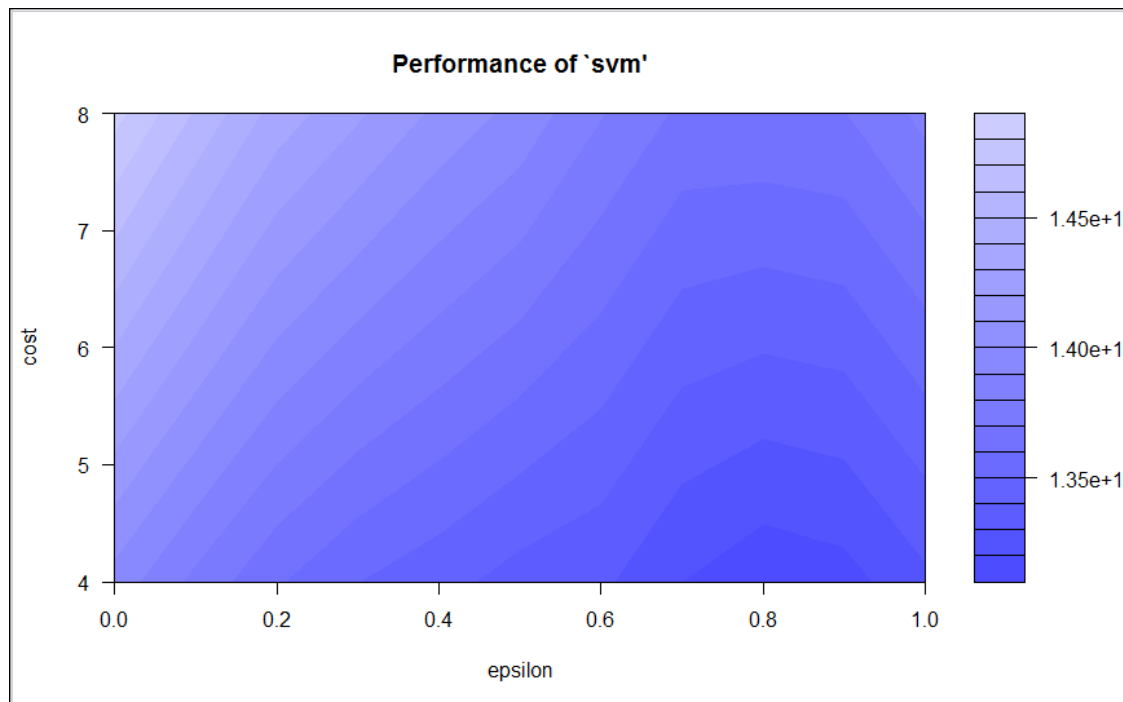
Figure 4.21 Parameters for Final SVR Model with Linear Kernel after optimisation

The model is used to make predictions on the test set and RMSE is calculated: 114107. The RMSE of the model has shown improvement from 115709 to 114107.

4.5.6 Support Vector Regression with Non-Linear Kernel

Another Support Vector Regression model is built using a non-linear kernel. Radial Basis kernel is used to train the model. The initial SVR Model with Radial kernel has the same RMSE as initial SVR with a linear kernel of 115709.

The model is then optimized using a grid search (Figure below) to find the optimal parameter values. The darker shades reflect the parameter values with minimum error.



The final optimized model is selected with the below parameter values:

```
Parameters:  
SVM-Type: eps-regression  
SVM-kernel: radial  
cost: 4  
gamma: 0.02777778  
epsilon: 0.9
```

Figure 4.22 Parameters for final SVR Model with Radial Basis Kernel after optimisation

The selected model is then used for prediction on the test set and RMSE is calculated as 117435. The RMSE results show that the SVM with Linear kernel is performing better than the SVR with Radial basis kernel.

4.5.7 Model Comparison Based on RMSE Value

The below table shows the Root Mean Squared error from all the regression models.

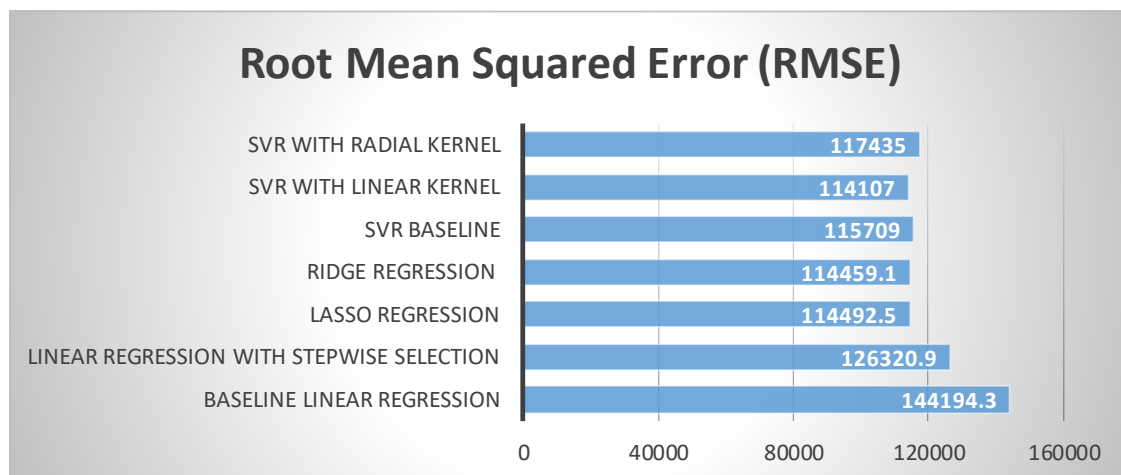


Figure 4.23 RMSE Summary for Predictive Modelling

The Support Vector Regression with a linear kernel after parameter engineering has the minimum RMSE value on the test data in comparison to the other regression models. It is evident from the results that Support Vector Regression outperforms the Multiple Linear Regression Model based on RMSE. Also, interestingly, the SVR

model with linear kernel has a lower value of RMSE as compared to the SVR with Radial basis kernel.

Another key observation in the predictive modelling is that Ridge Regression and Lasso Regression have a very small difference in RMSE, but due to the variable selection capability of Lasso, it provided a more interpretable model compared to Ridge Regression.

4.6 Conclusion

This chapter has outlined the detailed experiments conducted for the study. A quantitative approach based on the research framework was used to determine the impact of all the factors under study. The underlying assumptions for the models were validated and the corresponding results were reported. The results of the study will be discussed in detail in the next chapter.

5 DISCUSSION AND ANALYSIS

5.1 Introduction

This chapter discusses the experiment results from exploratory data, Hypothesis testing, Regression Analysis and Predictive modelling in the context to the research objectives. The key strengths and weaknesses of the study are also discussed.

5.2 Exploratory Data Analysis

Exploratory data analysis led to some key insights into the data. The Salary difference by Job Location revealed that Mumbai, Bangalore, and Pune are top-paying cities. This trend is expected as these are the fastest growing information technology hubs in India, which may attribute to the higher salaries in these cities. Another interesting observation in location context is, in spite of being third largest metropolitan city in India, the entry level salaries are the lowest in Kolkata. The exploratory analysis also showed differences in the salary based on the choice of specialisation in engineering. It indicated that civil engineers have a high lower quartile for salaries. Another key trend which was indicated by the data, is the increase in salaries by graduation year from 2007 until 2010 which is a little unexpected since the world economy was recovering from the 2008 economic crisis.

5.3 Hypothesis Testing

The results from hypothesis testing are outlined below in relation to the initial hypotheses.

H1: Male candidates are paid higher starting salaries than their female counterparts.

- Statistical Tests showed no evidence to reject the null hypothesis. There was no statistical evidence found to support that there is a gender bias in the starting salaries of engineering graduates.

H2: The engineering graduates from Tier A colleges are paid higher starting salaries than the graduates from Tier B college.

- A t-test for equality rejected the null hypothesis. Statistical tests supported the initial hypothesis stating that engineering graduates from Tier A college receive a higher salary than students from Tier B. The initial hypothesis was supported.

H3: The computer science graduates are paid a higher salary than the other engineering domains.

- A pairwise t-test provided no evidence to reject the null hypothesis. There was no statistical evidence found to support the claim that computer science graduates are paid more than those from other engineering specialisations. The initial hypothesis was not supported.

H4: English is the strongest predictor of salary compared to Logical and Quantitative ability.

- Regression Analysis revealed that Quantitative skills are a relatively stronger salary predictor than English. Individual regression on cognitive scores revealed that 'Quant' scores explain twice the variance in Salary as compared to English. The initial hypothesis proved to be wrong in this case.

5.4 Regression Analysis

The regression analysis indicated that the cognitive skills 'English' and 'Quant' are both significant determinants of Salary in all of the regression models. The variable 'Logical ability' is a significant predictor when only cognitive skills were included in the regression but, in the presence of other features, it was found to be statistically insignificant.

The results from this study supported the general notion of earlier research by (Hamermesh and Donald, 2008; Jones and Jackson, 1990; Chia and Miller, 2008), that academic performance from school to college is a significant contributor to the salary of engineering graduates, even though these earlier studies were not specifically targeted at engineering graduates. The regression results illustrated that 'X10percentage', 'X12percentage' and 'collegeGPA' are major predictors of Salary.

The choice of engineering major is also found to be a significant contributor in determining the salary which is in sync with the results from existing research focused

on the impact of choice of study major on salary by (Hamermesh and Donald, 2008; Rumberger and Thomas, 2003; Arcidiacono, 2004).

The unique properties of the dataset also provided an opportunity to study the relationship of personality traits with salary, and the regression analysis reflected that conscientiousness (i.e. a desire to do a task well) is a significant predictor of salary in the presence of all the other feature variables under study.

5.5 Predictive Modelling

A comparative analysis of Salary prediction models indicated that a Support Vector Regression (SVR) model with a linear kernel after parameter optimisation outperformed other models based on RMSE.

Moreover, the performance results from Ridge and Lasso regression showed the very close performance of these models to the SVR model. Between Ridge and Lasso, the difference of RMSE was extremely low, but Lasso, due to its ability to shrink the coefficients to exactly zero, provides a more interpretable model. The final selected regression model from Lasso had only 11 variables whereas the Ridge Regression had all 33 variables.

5.6 Awareness of Strengths and Weaknesses

The strengths of the research study are:

- 1) The combination of academic variables, along with demographics, cognitive skills and personality traits provided the study more robustness in terms of controlling effects.
- 2) Even though the initial hypothesis of English scores being a relatively stronger predictor of Salary compared to Logical and Quant scores was rejected, the English language is still a statistically significant predictor of Salary in all the models which is very critical in Indian Markets.
- 3) The findings of the results hold true to one of the hypotheses that students from Tier A college are paid more than those from Tier B colleges. These results are significant as the data was collected from 1350 different colleges after elimination of the elite government colleges.
- 4) The educational parameters in Indian Markets are referenced for data preparation which further strengthen the research outcomes.

The weaknesses of the research study are:

- 1) Even though the dataset is unique, the sample could be biased for generalisation of some results considering the various diversities in Indian Education system and Indian Labour Markets. For example, the data contains more candidates from computer science background compared to mechanical and civil engineering.
- 2) The examination of gender bias from the data might not be very reliable considering the fact that the data does not have an equally distributed sample of male and female candidates.
- 3) The recruitment process in itself is a very subjective process and in some cases private companies conduct their own written test to score students in campus placement drives. There is no information regarding whether the candidate score for AMCAT test has been used for recruitment or not.
- 4) There is no statistical evidence to support that salaries differ by specialisation but there can be other factors in effect that the majority of candidates in the study might be getting recruited only by Information Technology companies.

5.7 Conclusion

This chapter discussed the results from the experiments in details. The results were used to evaluate the initial hypotheses and to contextualise them with earlier studies. The chapter also outlined the key strengths and weaknesses of the study. In the next chapter, we will summarise the major findings in terms of contribution and recommendations for future work.

6 CONCLUSION

6.1 Introduction

This chapter will provide a brief summary of the research study. It provides an overview of the course of action carried out for the research. It also reflects on the contribution made to the existing literature body for the research area in question. Finally, the future directions and options for research are discussed.

6.2 Research Definition & Research Overview

The project is aimed at understanding the primary salary determinants of entry level engineering graduates in Indian Labour Markets. The primary factors under examination against the salary were: academic features, cognitive skills, standardised test scores, and personality traits. In addition, another objective of the study was to select a best performing salary prediction model. The research study allowed us to achieve the following objectives:

- To perform an extensive literature review on the employability factors and salary determinants of graduate students in diverse geographies and study majors, hence enabling a profound platform for the research study.
- To explore the salary differences and salary trends with various underlying factors for engineering graduates in Indian Labour Markets.
- To identify the primary factors determining the starting salary of engineering graduates using regression analysis.
- To build and compare accurate salary prediction models based on Root Mean Squared Error.

6.3 Contributions to the Body of Knowledge

The research study examined two aspects in the context of Indian Labour markets: To determine the best salary predictors and to select a most accurate salary predictor based on an accuracy comparison. The idea was to explore the semantics of employment

outcome for engineering graduates in Indian Markets and add value to the existing body of knowledge. The findings of this thesis can bring valuable insights to researchers in the field. The novelty of research is driven by the data and its demographics. The study looked at a number of factors such as cognitive skills, academic variables, demographics and personality traits, as salary determinants for recent engineering graduates. Even though this is an ongoing research area, this combination of factors is considered for the first time for a developing socio-economic nation like India.

Researchers can use this research in further examining these factors within other fields of study and in addition to other external factors. The predictive modelling results can be used as a benchmark model for further research in applications of predictive salary modelling with a wider range of other techniques and ensemble models.

The findings of regression analysis could be fruitful for students to tailor their choices for the maximum financial return on jobs. This is because few of the factors such as academic grades and choice of major are under the student's control. In addition, the data indicated that the choice of school from an affiliation perspective is also a significant determinant of starting salary. The results from the study can well inform these choices with respect to maximise the salary outcomes. The findings can be helpful for education administrators to bring interventions to improve certain skills, for example, English language. With India being a multilingual country, this can be a critical factor.

6.4 Experimentation, Evaluation, and Limitation

The research project used a dataset released by Aspiring Minds (Aggarwal, Srikant, and Nisar, 2016). The study followed a general course of secondary research with the below experiment steps:

- In-depth examination of the existing research works in the context of employability factors and salary determinants of graduate and undergraduate students.
- Exploratory data Analysis to develop a deeper understanding of the data.
- Data preparation steps based on the research of external factors of education systems in India.

- Hypothesis testing examining the salary inequalities among entry-level engineering graduates.
- Regression Analysis of the factors under study towards the target variable - Salary of an engineering graduate.
- Linear and Non – Linear Predictive Modelling to find the best salary prediction model based on an Error function (RMSE).

The scope of the study was only limited to engineering graduates and the dataset did not include student data from other study domains. In addition, the generalisation of results might be a little subjective to the effect of external environmental factors such as National economic factors, inflation rates, policy changes etc. Considering these limitations of the study, the future research directions and options are discussed in the next section.

6.5 Future Work & Research

The study examined a larger number of factors in combination, affecting the salaries for engineering graduates in India Labour Markets, than existing literature. Although many of these factors are examined in existing literature, the unique combination of cognitive skills, standardised test score along with academic and demographic factors perceived to add to the body of research.

There are several aspects of the research which could be perused to contribute to the research body. Based on the literature review, it was observed that internship or industrial training is a key factor for starting salaries and employability. Although this study examined a varied combination of variables, it is recommended to adjust the data collection/survey instrument to possibly capture that information. In addition, information such as exchange programs, experience studying abroad, part-time work experience and volunteering experience could be examined, provided the data is captured for these factors.

This thesis has presented an inductive, data-driven approach for the prediction of salaries in the Indian job market for entry level engineering graduates. Because of the dynamism of the features involved in such a prediction, this study could be tackled from a different perspective by, for instance, employing deductive reasoning techniques for inference including (Longo, 2012; Longo, 2013; Longo, 2014;

Longo,2015; Rizzo,2016). Additionally, the feature set could be extended by mining social data from the cloud or online social networks (Dondio, 2011), and extract relevant determinants for predicting salaries (Longo, 2009; Longo 2010).

Another area for future research in this context would be to try to employ other predictive modelling techniques to improve the salary prediction performance.

External factors such as a country's economic growth metrics, inflation rates, and other environmental changes might have an impact on salaries. The data for these could be factored into the study to delve deeper into the subject area. In addition to this, the scope of the research study could be extended to graduates from other study domains such as business, arts etc.

BIBLIOGRAPHY

Aggarwal, V., Srikant, S. & Nisar, H. (2016). AMEO 2015: A dataset comprising AMCAT test scores, biodata details and employment outcomes of job seekers. In M. Marathe, M. K. Mohania, Mausam & P. Jain (eds.), *CODS* (p./pp. 27),: ACM. ISBN: 978-1-4503-4217-9.

Ajit, V., & Deshmukh, P. B. (2013). Factors Impacting Employability Skills of Engineers.

Alff, G. N. (1984). A Note Regarding Evaluation of Multiple Regression Models. *PCAS LXXI*, 84-95.

Ang, S., Slaughter, S., & Yee Ng, K. (2002). Human capital and institutional determinants of information technology compensation: Modelling multilevel and cross-level interactions. *Management Science*, 48(11), 1427-1445.

Ang, S., Van Dyne, L., & Begley, T. M. (2003). The employment relationships of foreign workers versus local employees: A field study of organizational justice, job satisfaction, performance, and OCB. *Journal of organizational behaviour*, 24(5), 561-583.

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1), 343-375.

Athey, S., Katz, L. F., Krueger, A. B., Levitt, S., & Poterba, J. (2007). What does performance in graduate school predict? Graduate economics education and student outcomes. *The American economic review*, 97(2), 512-518.

Bai, L. (2006). Graduate unemployment: Dilemmas and challenges in China's move to mass higher education. *The China Quarterly*, 185, 128-144.

Banerjee, R., & Muley, V. P. (2007). Engineering education in India. Report to Energy Systems Engineering, IIT Bombay, *sponsored by Observer Research Foundation*, September, 14.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203-224.

Becker, G. S. (1985). Human capital, effort, and the sexual division of labor. *Journal of labor economics*, S33-S58.

Becker, G. S. (1985). Human capital, effort, and the sexual division of labor. *Journal of labor economics*, S33-S58.

Becker, G. S. 1993 Human Capital: A Theoretical and Empirical Analysis with Specific Reference to Education (3rd Ed.) *The University of Chicago Press Chicago*.

Belfield, C. R., & Fielding, A. (2001). Measuring the relationship between resources and outcomes in higher education in the UK. *Economics of Education Review*, 20(6), 589-602.

Blau, F. D., & DeVaro, J. (2007). New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires. *Industrial Relations: A Journal of Economy and Society*, 46(3), 511-550.

Boissiere, M., Knight, J. B., & Sabot, R. H. (1985). Earnings, schooling, ability, and cognitive skills. *The American Economic Review*, 75(5), 1016-1030.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Bretts, B. 1993. "She Shall Overcome," *in: Computerworld*. pp. 67-70.

Brewer, D. J., Eide, E. R., & Ehrenberg, R. G. (1999). Does it pay to attend an elite private college? Cross-cohort evidence on the effects of college type on earnings. *Journal of Human resources*, 104-123.

Busse, R. (1992). The New Basics: Today's Employers Want the. *Vocational Education Journal*, 67(5), 24.

Callanan, G., & Benzing, C. (2004). Assessing the role of internships in the career-oriented employment of graduating college students. *Education+ Training*, 46(2), 82-89.

Carriere, J. F., & Shand, K. J. (1998). New salary functions for pension valuations. *North American Actuarial Journal*, 2(3), 18-26.

Carter, R. D., Das, R. S., Garnello, A. H., & Charboneau, R. C. (1984). Multivariate alternatives to regression analysis in the evaluation of salary equity-parity. *Research in Higher Education*, 20(2), 167-179.

Chen, D., Aghdam, A. R., Kamalpour, M., & Sim, A. T. H. (2013, November). The impact of College English Test (CET) on graduates' salaries using data mining techniques. In *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on* (pp. 559-563). IEEE.

Chia, G., & Miller, P. W. (2008). Tertiary performance, field of study and graduate starting salaries. *Australian Economic Review*, 41(1), 15-31.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Curtis, E. A., Comiskey, C., & Dempsey, O. (2016). Importance and use of correlational research. *Nurse Researcher*, 23(6), 20-25.

Daymont, T. N., & Andrisani, P. J. (1984). Job preferences, college major, and the gender gap in earnings. *Journal of Human Resources*, 408-428.

- Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F. (2009). Package 'e1071'. *R Software package, available at <http://cran.r-project.org/web/packages/e1071/index.html>*.
- Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC medical genomics, 4*(1), 1.
- Dondio, P., and Longo, L. (2011) Trust-based techniques for collective intelligence in social search systems. *In Next Generation Data Technologies for Collective Computational Intelligence, pages 113–135. Springer.*
- Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology, 12*(1), 1-7.
- Fang, X., Lee, S., Lee, T. E., & Huang, W. (2004). Critical factors affecting job offers for new MIS graduates. *Journal of Information Systems Education, 15*(2), 189.
- Fortin, N. M. (2008). The gender wage gap among young adults in the united states the importance of money versus people. *Journal of Human Resources, 43*(4), 884-918.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (1993). How to design and evaluate research in education (Vol. 7). New York: McGraw-Hill.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.
- Fuller, R., & Schoenberger, R. (1991). The Gender Salary Gap: Do Academic Achievement, Internship Experience, and College Major Make a Difference?. *Social Science Quarterly, 72*(4), 715-26.
- Gault, J., Leach, E., & Duey, M. (2010). Effects of business internships on job marketability: the employers' perspective. *Education+ Training, 52*(1), 76-88.

Gault, J., Redington, J., & Schlager, T. (2000). Undergraduate business internships and career success: are they related?. *Journal of marketing education*, 22(1), 45-53.

Ge, C., Kankanhalli, A., & Huang, K. W. (2015). Investigating the Determinants of Starting Salary of IT Graduates. *ACM SIGMIS Database*, 46(4), 9-25.

Gefen, D., Straub, D. W., & Rigdon, E. E. (2011). An update and extension to SEM guidelines for administrative and social science research. *Management Information Systems Quarterly*, 35(2), iii-xiv.

Gerhart, B. (1988). Sources of variance in incumbent perceptions of job complexity. *Journal of Applied Psychology*, 73(2), 154.

Gilmour, S. G. (1996). The Interpretation of Mallows's Cp-Statistic. *The Statistician*, 49-56.

Godofsky, J., Zukin, C., & Van Horn, C. (2011). Unfulfilled expectations: Recent college graduates struggle in a troubled economy. *Heldrich Center for Workforce Development, Rutgers University*.

Gokuladas, V. K. (2010). Technical and non-technical education and the employability of engineering graduates: an Indian case study. *International Journal of Training and Development*, 14(2), 130-143.

Gokuladas, V. K. (2011). Predictors of employability of engineering graduates in campus recruitment drives of Indian software services companies. *International Journal of Selection and Assessment*, 19(3), 313-319.

Götz, O., Liehr-Gobbers, K., & Krafft, M. (2010). Evaluation of structural equation models using the partial least squares (PLS) approach. In *Handbook of partial least squares* (pp. 691-711). Springer Berlin Heidelberg.

Grogger, J., & Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *Journal of Human Resources*, 280-310.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3, 1157-1182.

Hamermesh, D. S., & Donald, S. G. (2008). The effect of college curriculum on earnings: An affinity identifier for non-ignorable non-response bias. *Journal of Econometrics*, 144(2), 479-491.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Howell, R. P. (1969). On professional salaries. *Spectrum, IEEE*, 6(2), 22-29.

Howell, R. P., Gorfinkel, M., & Bent, D. (1966). INDIVIDUAL CHARACTERISTICS SIGNIFICANT TO SALARY LEVELS OF ENGINEERS AND SCIENTISTS. STANFORD RESEARCH INST MENLO PARK CA.

Huberty, C. J (1989). Problems with stepwise methods-better alternatives. In B. Thompson (Ed.).

Jagacinski, C. M., Lebold, W. K., Linden, K. W., & Shell, K. D. (1985). Factors influencing the choice of an engineering career. *Education, IEEE Transactions on*, 28(1), 36-42.

James, E., Alsalam, N., Conaty, J. and To, D., (1989). "College Quality and Future Earnings: Where Should You Send Your Child to College?", *American Economic Review, Papers and Proceedings*, Vol. 79, No. 2, pp. 247-252

James, E., Alsalam, N., Conaty, J. C., & To, D. L. (1989). College quality and future earnings: where should you send your child to college?. *The American Economic Review*, 79(2), 247-252.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Johnson, C. B., Riggs, M. L., & Downey, R. G. (1987). Fun with numbers: Alternative models for predicting salary levels. *Research in Higher Education*, 27(4), 349-362.

Jones, E. B., & Jackson, J. D. (1990). College grades and labor market rewards. *The Journal of Human Resources*, 25(2), 253-266.

Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2007). Feature selection for regression problems. *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece, 2022*.

Kidwell, J. S., & Brown, L. H. (1982). Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and the Family*, 287-299.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17(1), 107-112.

Koskinen, L., Nummi, T., & Salonen, J. (2005). Modelling and Predicting Individual Salaries: A Study of Finland's Unique Dataset. Finnish Centre for Pensions.

Koskinen, L., Nummi, T., & Salonen, J. (2005). *Modelling and Predicting Individual Salaries: A Study of Finland's Unique Dataset*. Finnish Centre for Pensions.

Leedy, P. D., & Ormrod, J. E. (2010). Planning and design.

Liu, H., Dougherty, E., Dy, J. G., Torkkola, K., Tuv, E., Peng, H., ... & Zhao, Z. (2005). Evolving feature selection. *Intelligent systems, IEEE*, 20(6), 64-76.

Lommele, J. A. & Sturgis, R. W. (1974). "An Econometric Model of Workmen's Compensation," *PCAS*, Volume LXI, 1974. p. 170.

Longo, L. (2015). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour and Information Technology*, 34(8):758–786.

Longo, L. and Dondio, P. (2014). Defeasible reasoning and argument based medical systems: an informal overview. *In 27th International Symposium on Computer-Based Medical Systems, New York, USA, pages 376–381. IEEE.*

Longo, L. and Hederman, L. (2013). Argumentation theory for decision support in health-care: A comparison with machine learning. *In Brain and Health Informatics - International Conference, BHI 2013, Maebashi, Japan, October 29-31, pages 168–180.*

Longo, L., Kane, B. and Hederman, L., (2012). Argumentation theory in health care. *In Proceedings of CBMS 2012, The 25th IEEE International Symposium on Computer-Based Medical Systems, June 20-22, Rome, Italy, pages 1–6.*

Longo, L., Dondio, P., and Barrett, S., (2010). Enhancing social search: A computational collective intelligence model of behavioural traits, trust and time. *Transactions on Computational Collective Intelligence, 2:46–69.*

Longo, L., Barrett, S., and Dondio, P. (2009). Information foraging theory as a form of collective intelligence for social search. *In Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems, First International Conference, ICCCI, Wroclaw, Poland, October 5-7, pages 63–74.*

Loury, L. D. (1997). The gender earnings gap among college-educated workers. *Industrial & Labor Relations Review, 50(4), 580-593.*

Loury, L. D., & Garman, D. (1995). College selectivity and earnings. *Journal of Labor Economics, 289-308.*

Lyness, K. S., & Heilman, M. E. (2006). When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology, 91(4), 777.*

Magel, R., & Hoffman, M. (2015). Predicting Salaries of Major League Baseball Players. *International Journal of Sports Science, 5(2), 51-58.*

Mikita, D., Dehondt, G., & Nezlek, G. S. (2012). The Deployment Pipeline. In *Proceedings of the Conference on Information Systems Applied Research ISSN* (Vol. 2167, p. 1508).

Miller D.A. & Millar I. (1996) *The Cambridge Dictionary of Scientists*. Cambridge University Press, Cambridge.

Murnane, R. J., Willett, J. B., & Levy, F. (1995). *The growing importance of cognitive skills in wage determination* (No. w5076). National Bureau of Economic Research

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences* (Vol. 791). John Wiley & Sons.

Nezlek, G., & DeHondt, G. (2012). Gender Wage Differentials in Information Systems: 1991–2008 A Quantitative Analysis. *Integrations of Technology Utilization and Social Dynamics in Organizations*, 31.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), 673-690.

Pezzullo, T. R., & Brittingham, B. E. (1979). Salary Equity: Detecting Sex Bias in Salaries among College and University Professors.

Quan, J. J., Dattero, R., & Galup, S. D. (2007). Information technology wages and the value of certifications: A human capital perspective. *Communications of the Association for Information Systems*, 19(1), 6.

Quan, J. J., Dattero, R., & Galup, S. D. (2007). Information technology wages and the value of certifications: A human capital perspective. *Communications of the Association for Information Systems*, 19(1), 6.

R Studio Team (2015). RStudio: Integrated Development for R. R Studio, Inc., Boston, MA URL <http://www.rstudio.com/>.

Ramsay, G. A. (1979). A generalized multiple regression model for predicting college faculty salaries and estimating sex bias. In T. Pezzullo and B. Brittingham (eds.), *Salary Equity: Detecting Sex Bias Among College and University Professors* pp. 37–53. *Lexington, MA: Lexington Books*.

Rizzo L., Dondio, P., Delany, S. J. and Longo, L. (2016). Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile, pages 215–229. *Springer International Publishing, Cham*.

Rosenbaum, J. E. (1979). Tournament mobility: Career patterns in a corporation. *Administrative science quarterly*, 220-241.

Rumberger, R. W., & Thomas, S. L. (1993). The economic returns to college major, quality and performance: A multilevel analysis of recent graduates. *Economics of Education Review*, 12(1), 1-19.

Sagen, H. B., Dallam, J. W., & Laverty, J. R. (1999). Job search techniques as employment channels: Differential effects on the initial employment success of college graduates. *The Career Development Quarterly*, 48(1), 74-85.

Sandvig, J. C., Tyran, C. K., & Ross, S. C. (2005). Determinants of graduating MIS students starting salary in boom and bust job markets. *Communications of the Association for Information Systems*, 16(1), 29.

Saunders Mark, N. K., Lewis, P., & Thornhill, A. (2000). *Research Methods for Business Students*.

- Scholz, D. (1996). Risk Associated With Different College Majors. *Illinois Wesleyan University: Senior Honors Project, 1*, 996.
- Smart, J. C. (1988). College influences on graduates' income levels. *Research in Higher Education, 29*(1), 41-59.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing, 14*(3), 199-222.
- Steffy, B. D., Shaw, K. N., & Noe, A. W. (1989). Antecedents and consequences of job search behaviors. *Journal of Vocational Behavior, 35*(3), 254-269.
- Tan, M., & Igarria, M. (1994). Turnover and remuneration of information technology professionals in Singapore. *Information & Management, 26*(4), 219-229.
- Tchibozo, G. (2007). Extra-Curricular Activity and the Transition from Higher Education to Work: A Survey of Graduates in the United Kingdom. *Higher Education Quarterly, 61*(1), 37-56.
- Thomas, S. L. (2000). Deferred costs and economic returns to college major, quality, and performance. *Research in Higher Education, 41*(3), 281-313.
- Thomas, S. L., & Zhang, L. (2005). Post-baccalaureate wage growth within four years of graduation: The effects of college quality and college major. *Research in Higher Education, 46*(4), 437-459.
- Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development, 21*(4), 146-148.
- Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*, 525-534.

Thorson, A. (2005). The effect of college major on wages. *Illinois Wesleyan University: Senior Honors Project*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tripney, J., Hombrados, J. G., Newman, M., Hovish, K., Brown, C., Steinka-Fry, K. T., & Wilkey, E. (2013). Post-Basic Technical and Vocational Education and Training (TVET) Interventions to Improve Employability and Employment of TVET Graduates in Low-and Middle-Income Countries: A Systematic Review. *Campbell Systematic Reviews*, 9(9).

Truman, G. E., & Baroudi, J. J. (1994). Gender differences in the information systems managerial ranks: An assessment of potential discriminatory practices. *MIS quarterly*, 129-142.

V. Vapnik and A. Lerner, (1963), "Pattern recognition using generalized portrait method", *Automation and Remote Control*, 24.

Wanberg, C. R., Watt, J. D., & Rumsey, D. J. (1996). Individuals without jobs: An empirical study of job-seeking behavior and reemployment. *Journal of Applied Psychology*, 81(1), 76.

Weisbrod, B. A., & Karpoff, P. (1968). Monetary returns to college education, student ability, and college quality. *The Review of Economics and Statistics*, 50(4), 491-497.

Wise, D. A. (1975). Academic achievement and job performance. *The American Economic Review*, 65(3), 350-366.

Xiangquan, Q. S. Z. (2009). Employability, Internship and Graduate Employment: Based on Shandong Survey Data. *Chinese Journal of Population Science*, 6, 013.

Yi, Z., Chunguang, Z., Lan, H., Yan, W., & Bin, Y. (2009, December). Support vector regression for prediction of housing values. In *Computational Intelligence and Security, 2009. CIS'09. International Conference on* (Vol. 2, pp. 61-65). IEEE.

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3), 331-350.

Zeng, X. Q., & Chen, Q. S. (2014, July). A comparative study of redundant feature detection based feature selection methods. In *Computer, Information and Telecommunication Systems (CITS), 2014 International Conference on* (pp. 1-5). IEEE.

APPENDIX A

Summary for Salary Mean comparison by gender

Group Statistics					
	Gender	N	Mean	Std. Deviation	Std. Error Mean
Salary	M	2972	290548.12	133020.597	2440.026
	F	945	281439.15	122613.199	3988.608

Independent Samples Test										
		Levene's Test for Equality of Variances			t-test for Equality of Means			95% Confidence Interval of the Difference		
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Salary	Equal variances assumed	8.639	.003	1.868	3915	.062	9108.962	4876.818	-452.381	18670.306
	Equal variances not assumed			1.948	1706.815	.052	9108.962	4675.759	-61.860	18279.784

Summary for Salary Mean comparison by collegetier

Group Statistics					
	CollegeTier	N	Mean	Std. Deviation	Std. Error Mean
Salary	A	272	379227.94	129248.293	7836.829
	B	3645	281569.00	128190.135	2123.273

Independent Samples Test										
		Levene's Test for Equality of Variances			t-test for Equality of Means			95% Confidence Interval of the Difference		
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Salary	Equal variances assumed	.985	.321	12.113	3915	.000	97658.943	8062.082	81852.666	113465.219
	Equal variances not assumed			12.028	312.121	.000	97658.943	8119.370	81683.322	113634.563

Regression Summary for the Cognitive Skill Score Regression Model

```
##
## Call:
## lm(formula = Salary ~ English + Logical + Quant, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -343316  -87793  -11565   70595  493785
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15571.85   12533.94   1.242 0.214173
## English      159.94     21.01   7.614 3.31e-14 ***
## Logical       91.56     27.26   3.358 0.000791 ***
## Quant        286.65     18.75  15.285 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121200 on 3913 degrees of freedom
## Multiple R-squared:  0.1396, Adjusted R-squared:  0.1389
## F-statistic: 211.6 on 3 and 3913 DF,  p-value: < 2.2e-16
```

Regression Summary for Cognitive Skills with Interaction Terms – Model 1

Residuals:

Min	1Q	Median	3Q	Max
-301877	-84383	-10656	66871	499194

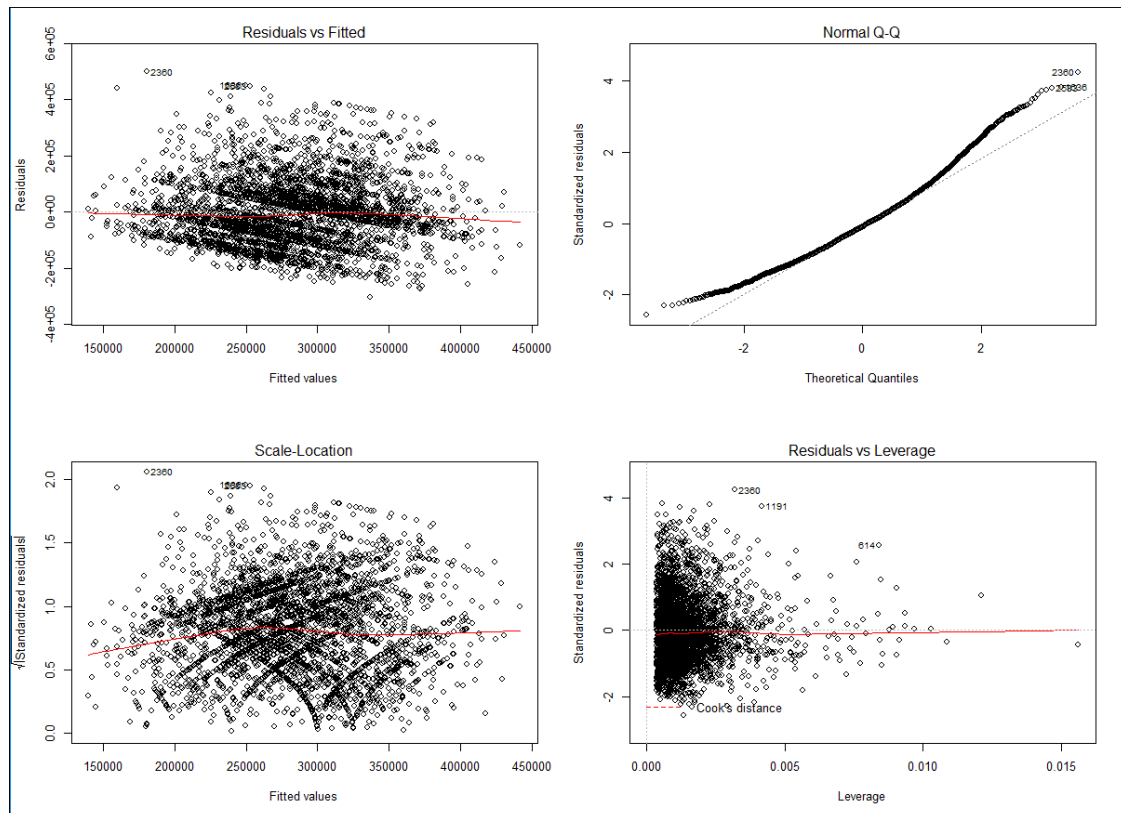
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.773e+04	4.685e+04	0.592	0.5540
English	1.673e+02	2.139e+01	7.819	6.94e-15 ***
Logical	4.897e+01	9.509e+01	0.515	0.6066
Quant	2.362e+02	9.360e+01	2.524	0.0117 *
Logical:Quant	9.525e-02	1.813e-01	0.525	0.5993

 signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 117900 on 3564 degrees of freedom
 Multiple R-squared: 0.1503, Adjusted R-squared: 0.1493
 F-statistic: 157.6 on 4 and 3564 DF, p-value: < 2.2e-16

Residual Plots for Cognitive Skills with Interaction Terms – Model 1

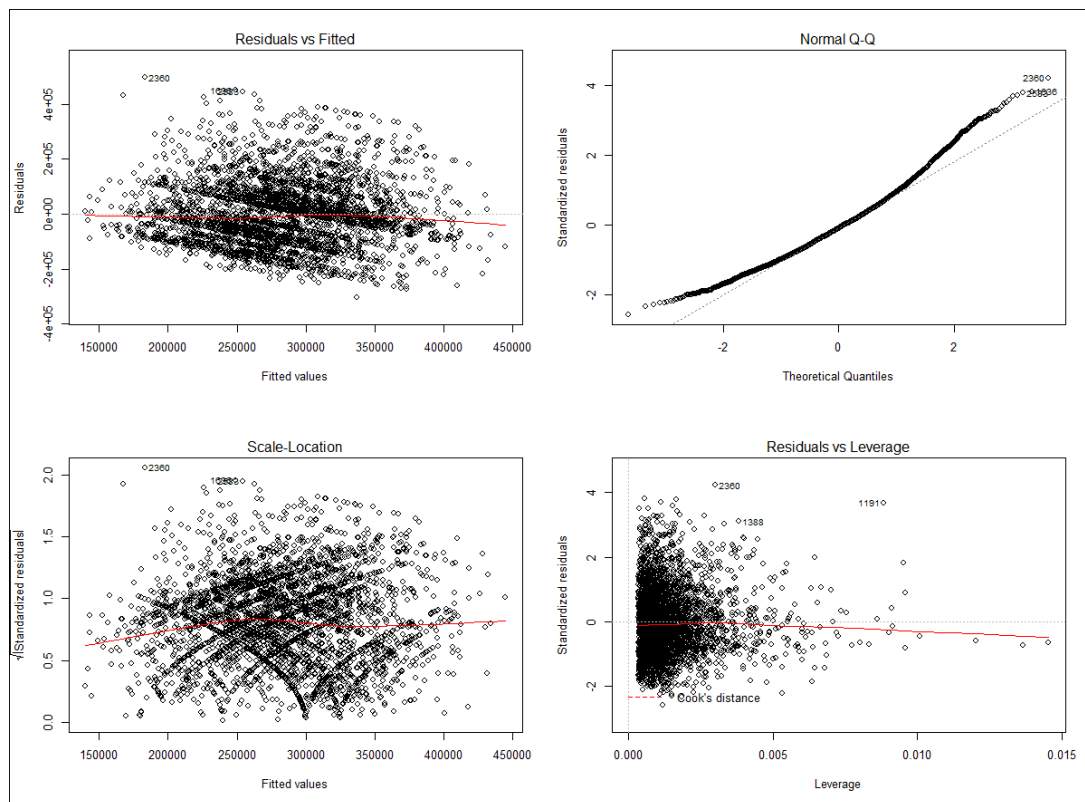


Regression Summary for Cognitive Skills with Interaction Terms – Model 2

Residuals:				
Min	1Q	Median	3Q	Max
-302288	-84932	-10631	67158	496921
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	61526.7883	55046.7333	1.118	0.264
English	48.4490	112.7291	0.430	0.667
Logical	-17.6822	110.1574	-0.161	0.872
Quant	284.3922	18.9183	15.033	<2e-16 ***
English:Logical	0.2329	0.2169	1.074	0.283

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 117900 on 3564 degrees of freedom				
Multiple R-squared: 0.1505, Adjusted R-squared: 0.1496				
F-statistic: 157.9 on 4 and 3564 DF, p-value: < 2.2e-16				

Residual Plots for Cognitive Skills with Interaction Terms – Model 2

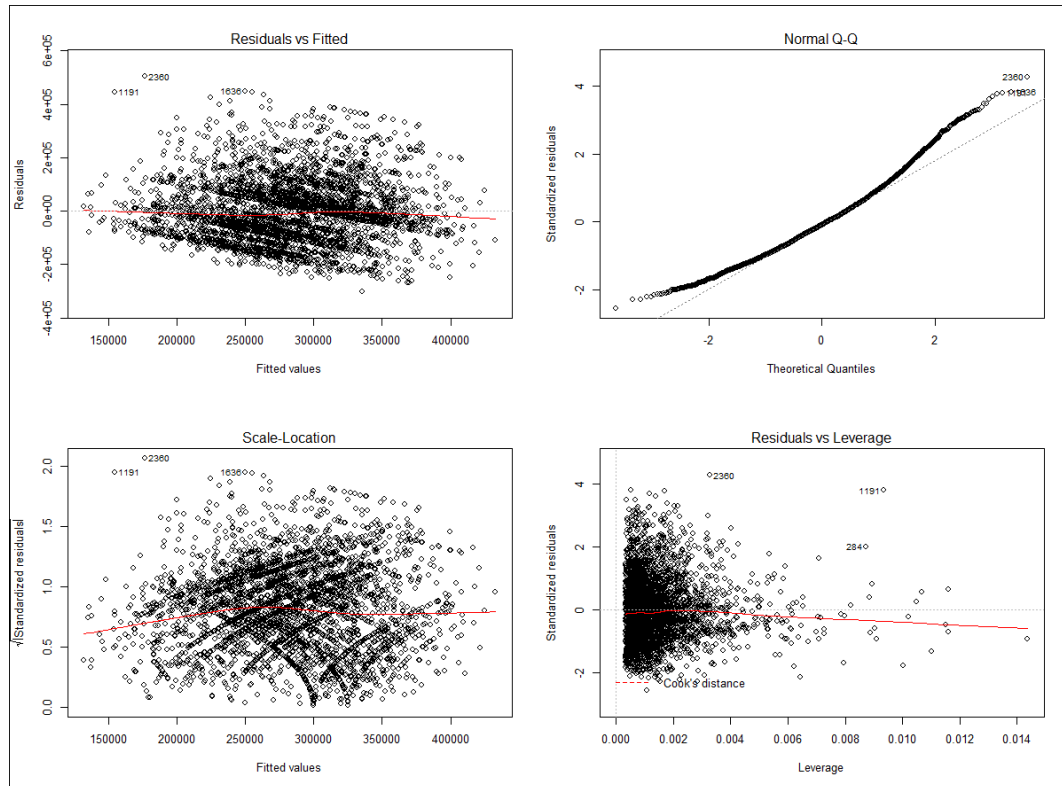


Regression Summary for Cognitive Skills with Interaction Terms – Model 3

Residuals:					
Min	1Q	Median	3Q	Max	
-300736	-84182	-10738	66569	503425	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-9.814e+03	4.003e+04	-0.245	0.806329	
English	1.959e+02	8.091e+01	2.421	0.015518	*
Quant	3.119e+02	7.751e+01	4.024	5.84e-05	***
Logical	9.621e+01	2.814e+01	3.419	0.000634	***
English:Quant	-5.475e-02	1.494e-01	-0.366	0.714051	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 117900 on 3564 degrees of freedom					
Multiple R-squared: 0.1503, Adjusted R-squared: 0.1493					
F-statistic: 157.6 on 4 and 3564 DF, p-value: < 2.2e-16					

Residual Plots for Cognitive Skills with Interaction Terms – Model 3



Regression Summary for Cognitive Skills with Gender

Residuals:

Min	1Q	Median	3Q	Max
-301434	-84088	-10422	66638	500953

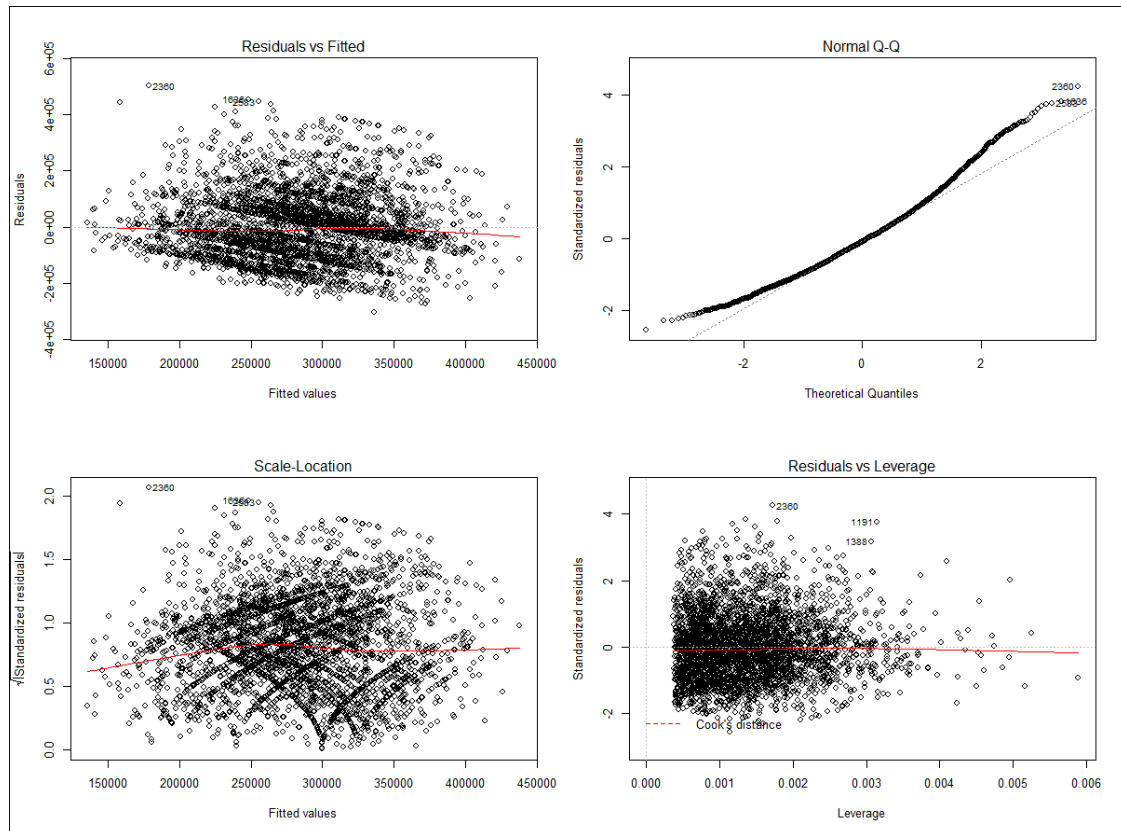
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2254.95	13349.35	0.169	0.865870	
English	167.68	21.40	7.835	6.15e-15	***
Logical	97.16	28.12	3.456	0.000556	***
Quant	283.26	19.03	14.884	< 2e-16	***
GenderM	2564.45	4689.54	0.547	0.584519	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 117900 on 3564 degrees of freedom
 Multiple R-squared: 0.1503, Adjusted R-squared: 0.1494
 F-statistic: 157.6 on 4 and 3564 DF, p-value: < 2.2e-16

Residual Plots for Cognitive Skills with Gender



Regression Summary for Cognitive Skills with Academic variables

Residuals:

Min	1Q	Median	3Q	Max
-320676	-80420	-11852	62943	491005

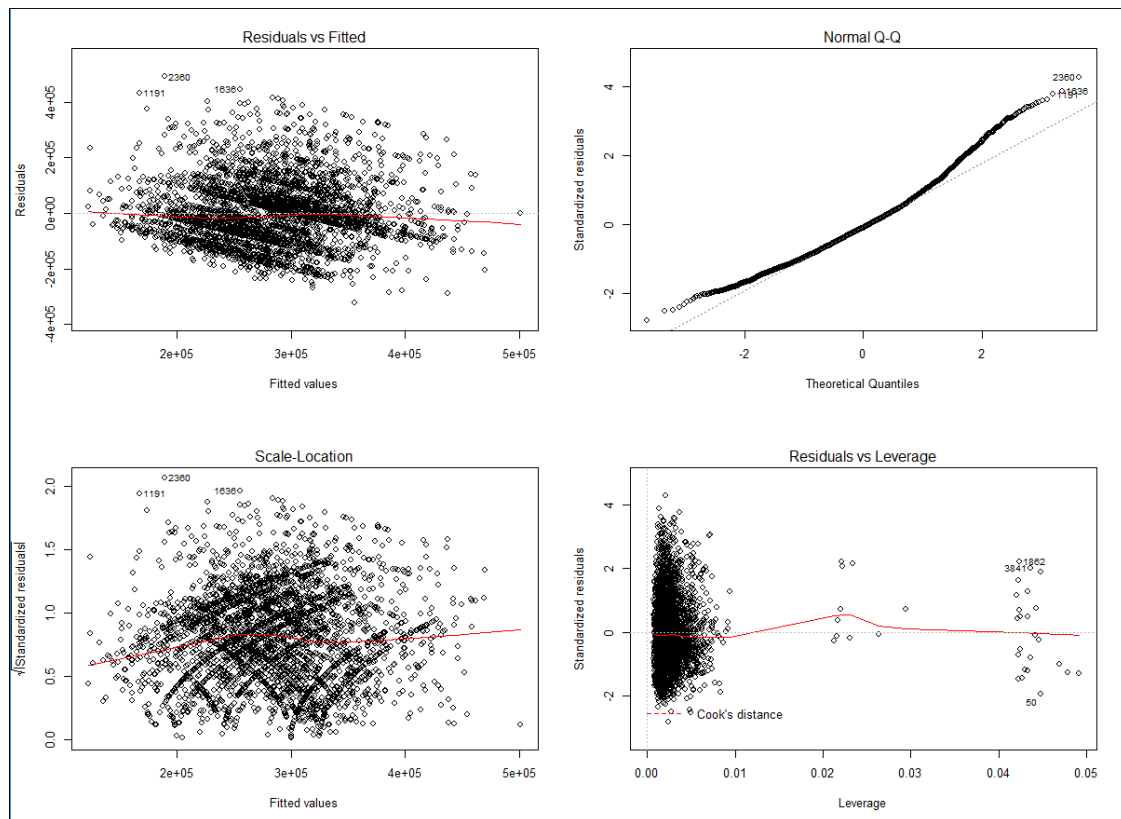
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-108209.14	24126.93	-4.485	7.52e-06	***
English	112.62	22.24	5.064	4.33e-07	***
Logical	45.39	27.73	1.637	0.1017	
Quant	212.71	19.20	11.081	< 2e-16	***
X10boardstate board	-10277.58	4190.45	-2.453	0.0142	*
X12boardstate board	19738.18	23732.47	0.832	0.4056	
X10percentage	925.52	269.47	3.435	0.0006	***
X12percentage	1302.69	243.14	5.358	8.96e-08	***
collegeGPA	1151.83	263.66	4.369	1.29e-05	***
CollegTierB	-46981.59	7753.08	-6.060	1.51e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114900 on 3559 degrees of freedom
 Multiple R-squared: 0.1942, Adjusted R-squared: 0.1922
 F-statistic: 95.32 on 9 and 3559 DF, p-value: < 2.2e-16

Residual Plots for Cognitive Skills and Academic variables



Regression Summary for the full set of variables

Residuals:				
Min	1Q	Median	3Q	Max
-365154	-77753	-6067	63969	469726

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-276385.12	66850.03	-4.134	3.64e-05 ***
GenderM	20436.84	4683.37	4.364	1.32e-05 ***
X10percentage	1183.05	267.95	4.415	1.04e-05 ***
X10boardstate board	-10234.41	4181.25	-2.448	0.014426 *
X12percentage	1358.13	245.66	5.528	3.46e-08 ***
X12boardstate board	7135.97	23133.79	0.308	0.757746
CollegeTierB	-42058.37	7682.46	-5.475	4.69e-08 ***
DegreeMasters	-9838.30	8940.21	-1.100	0.271209
Specializationcomputer science	121408.21	43011.06	2.823	0.004788 **
Specializationelectrical engineering	100701.34	44931.87	2.241	0.025075 *
Specializationelectronics and communication	90654.83	44242.74	2.049	0.040533 *
Specializationelectronics and communication	110726.42	45597.07	2.428	0.015217 *
Specializationinformation science	122268.05	43059.86	2.839	0.004544 **
Specializationmechanical engineering	51329.82	50686.87	1.013	0.311280
Specializationothers	120799.73	46107.07	2.620	0.008831 **
collegeGPA	1382.69	266.16	5.195	2.16e-07 ***
CollegeCityTier	309.44	4213.22	0.073	0.941456
NoYearToGraduation	463.64	1897.59	0.244	0.806989
English	123.34	21.98	5.611	2.17e-08 ***
Logical	34.63	27.38	1.265	0.206079
Quant	175.50	19.21	9.137	< 2e-16 ***
Domain	12180.60	5272.23	2.310	0.020927 *
ComputerProgramming	33.69	12.67	2.659	0.007873 **
ElectronicsAndSemicon	17.01	22.77	0.747	0.455150
ComputerScience	-122.50	11.36	-10.784	< 2e-16 ***
MechanicalEngg	85.30	45.94	1.857	0.063456 .
ElectricalEngg	-94.15	28.97	-3.250	0.001164 **
TelecomEngg	-23.60	20.97	-1.126	0.260374
CivilEngg	336.67	102.11	3.297	0.000987 ***
conscientiousness	-7672.51	2269.44	-3.381	0.000731 ***
agreeableness	3228.92	2813.95	1.147	0.251267
extraversion	1072.56	2330.09	0.460	0.645323
neroticism	-3319.97	2030.81	-1.635	0.102180
openess_to_experience	-1258.07	2365.86	-0.532	0.594924

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111200 on 3535 degrees of freedom
 Multiple R-squared: 0.25, Adjusted R-squared: 0.243
 F-statistic: 35.71 on 33 and 3535 DF, p-value: < 2.2e-16

Residual Plot for Regression Model for full set of variables

