MODELING PROBLEMS IN A REGIONAL LABOR MARKET - BY MARS AND ARTIFICIAL INTELLIGENCE - POLAND CASE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SELMA GÜTMEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ACTUARIAL SCIENCES

SEPTEMBER 2019

Approval of the thesis:

MODELING PROBLEMS IN A REGIONAL LABOR MARKET - BY MARS AND ARTIFICIAL INTELLIGENCE - POLAND CASE

submitted by **SELMA GÜTMEN** in partial fulfillment of the requirements for the degree of **Master of Science in Actuarial Sciences Department, Middle East Technical University** by,

Prof. Dr. Ömür Uğur
Director, Graduate School of **Applied Mathematics** _____

Prof. Dr. Sevtap Kestel
Head of Department, **Actuarial Sciences** _____

Prof. Dr. Sevtap Kestel
Supervisor, **Actuarial Sciences, IAM, METU** _____

Prof. Dr. Gerhard Wilhelm Weber
Co-supervisor, **Chair of Marketing and Economic Engineering, Poznan University of Technology** _____

Examining Committee Members:

Assoc. Prof. Dr. Ali Devin Sezer
Financial Mathematics, IAM, METU _____

Prof. Dr. Sevtap Kestel
Actuarial Sciences, IAM, METU _____

Prof. Dr. Gerhard Wilhelm Weber
Chair of Marketing and Economic Engineering,
Poznan University of Technology _____

Prof. Dr. Vilda Purutcuoğlu
Department of Statistics, METU _____

Assoc. Prof. Dr. Könül Bayramoglu
Department of Actuarial Sciences, Hacettepe University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: SELMA GÜTMEN

Signature :

ABSTRACT

MODELING PROBLEMS IN A REGIONAL LABOR MARKET - BY MARS AND ARTIFICIAL INTELLIGENCE - POLAND CASE

Gütmen, Selma
M.S., Department of Actuarial Sciences
Supervisor : Prof. Dr. Sevtap Kestel
Co-Supervisor : Prof. Dr. Gerhard Wilhelm Weber

September 2019, 57 pages

Laborers' skills are critical for advancement in the labor market in the economy and, eventually, in the areas of health, personal and social security, fulfillment, life quality and expectation. In this respect, it is essential to monitor needed knowledge and available core skills in the market, as well as to make this knowledge accessible to decision makers from economy, business and educational sectors. In Wielkopolskie Voivodeship (Greater Poland region, Poland), a Professional System has been implemented for many years as a vast and diverse dataset, to inspect requests for professional skills by employer, and to accelerate the flow of information between educational system and different areas of the labor market. The first step of the thesis was to preprocess this big dataset to make it suitable for our studies. The main aim of this thesis study is to mathematically model students' possible contributions ("promise") to jobs in terms of professionals skills (Z), with particular interest in the dependence of these contributions on common skills (O), general skills (W), any other economic or social circumstances, as well as time variables. Hence, in our model, the response variable is (Z) and all other aspects are implemented as input variables. Each row vector of our dataset is a pair (i,j) where student i is a student and a j is a job offer i applies to. Our aim is to figure out the relationship between the response variable and input variables by applying Linear Regression (LR), Multivariate Adaptive Regression Splines (MARS) and Artificial Neutral Networks (ANNs), as an AI-kind of methodology.

We compare the results of MARS model and ANNs model, by the help of statistical performance criteria and statistical graphs, herewith demonstrating the high competitiveness of our MARS approach. Through this analysis we also comment on how to determine the fulfillment between the demands of employers and decision markers of government and university leadership, especially, in the field of education.

Keywords: Data Mining, Modeling, Linear Regression, Multivariate Adaptive Regression Splines, Artificial Neutral Networks, Professional Skills, Education System, Labor Market, Applied Mathematics, Interdisciplinary Approach, Health and Safety.

ÖZ

BÖLGESEL İŞGÜCÜ PİYASASINDA MARS VE YAPAY ZEKA İLE MODELLEME SORUNLARI - POLONYA ÖRNEĞİ

Gütmen, Selma
Yüksek Lisans, Aktüerya Bilimleri Bölümü
Tez Yöneticisi : Prof. Dr. Sevtap Kestel
Ortak Tez Yöneticisi : Prof. Dr. Gerhard Wilhelm Weber

Eylül 2019 , 57 sayfa

Çalışanların belirli becerilere (mesleki veya teknik) yeterliliği, işgücü piyasasında, ekonomide ve nihayetinde sağlık, kişisel ve sosyal güvenlik, icra etme, yaşam kalitesi ve beklentiler alanında kaydedilen ilerlemeler için kritik bir konudur. Bu bakımdan, piyasadaki ihtiyaç duyulan bilgileri ve mevcut temel becerileri izlemenin yanı sıra bu bilgiyi ekonomi, işletme ve eğitim sektörlerinden yetkililere erişilebilir kılmak esastır. Wielkopolskie Voyvodalığı'nda (Büyük Polonya Bölgesi, Polonya), işveren tarafından mesleki beceri taleplerini denetlemek ve eğitim sistemi ile bilgi sistemi arasındaki bilgi akışını hızlandırmak amacıyla uzun yıllardır geniş ve çeşitli veri toplama platformu olarak Profesyonel Sistem incelenmiştir. Tezimizdeki ilk adım, bu büyük veri setini önceden hazırlayıp çalışmalarımız için uygun hale getirmektir. Tezdeki temel amaç; öğrencilerin mesleki beceriler (Z)'ye göre muhtemel katkılarını ("söz"), bu katkıların ortak becerilere (O), genel becerilere (W), diğer herhangi bir ekonomik veya sosyal koşullara ve zamana bağlı değişkenlere göre gösterdiği farklılıkları özellikle matematiksel olarak modellemektir. Dolayısıyla, modelimizde, yanıt değişkeni (Z)' ye bağlı olarak belirlenmiş ve diğer tüm koşullar girdi değişkeni olarak uygulanmıştır. Her satır vektör veri setimizin bir çifti (i, j) olarak belirlenmiş ve burada her öğrenci i, her öğrencinin başvuru yaptığı iş teklifi ise j olarak kullanılmıştır. Bu çalışmada Doğrusal Regresyon (LR), Çok Değişkenli Adaptif Regresyon Spline'ları (MARS) ve Yapay Nötr Ağları (ANNs) kullanarak yanıt değişkeni ve girdi değişken-

leri arasındaki ilişki bulunmaya çalışılmıştır. MARS modelinin ve ANNs modelinin sonuçlarını, istatistiksel performans kriterleri ve istatistiksel grafikler yardımıyla, ve MARS yaklaşımımızın rekabet edebilirliğinin yüksek olduğunu kanıtlayarak karşılaştırıyoruz. Bu analizle, özellikle eğitim alanında, işverenlerin talepleri ile devletin ve üniversite liderliğinin karar mercileri arasındaki taleplerin nasıl karşılanacağı ve nasıl belirleneceği üzerine yorum yapılmaktadır.

Anahtar Kelimeler: Veri Madenciliği, Modelleme, Doğrusal Regresyon, Çok Değişkenli Uyarlamalı Regresyonlu Şemaları, Yapay Nötr Ağ, Mesleki Beceriler, Eğitim Sistemi, İşgücü Piyasası, Uygulamalı Matematik, Disiplinlerarası Yaklaşım, Sağlık ve Güven.

ACKNOWLEDGMENTS

During my study in Poland, my supervisor Prof. Dr. Sevtap Kestel always helped me and answered my questions patiently. Even though she was in Turkey, she was a great source of help for following all my technical problems for my thesis process. Thanks to her existence, I had a chance to be a student abroad and have excellent Erasmus mobility experience. I would like to thank her for all of her technical support for this part of my life.

I would also like to express my deepest gratitude wholeheartedly to my co-adviser Prof. Dr. Gerhard-Wilhelm Weber. He is the scientific head of this research and inspiration of all my work during this thesis as well. I do not know how I can repay and thank for his support, guidance, enthusiastic encouragement and his immense patience that he showed towards me during this unique experience. I would like to say thanks for sharing his fountain of knowledge with me and his protective attitude towards me. His continues sharing of his experience and sparing his valuable time steered me towards to real life and helped me to find right path. I am honored to work side by side with such an excellent of character, to learn and to improve under his supervision. Working together with him made me realize to feel myself a lucky person for the first time in my life. He is one of my real best friends and *an actual guardian angel* for my vita.

I had this great chance to study for this thesis with the help of my colleagues, namely, Dr. Maciej Szafrański, Dr. Magdalena Graczyk-Kucharska, Dr. Marek Goliński, Dr. Maĺgorzata Spychala, and others. They guided me and were closely collaborating with me for the project, inspring and the closest study to my thesis, entitled *"Modeling Problems in a Regional Labor Market in Poland with MARS"* [17]. I sincerely thank them to share their work and give an opportunity to work together with such great people. It was a great experience to be in Poland thanks to their warm friendship.

Besides my colleagues and advisors, I would like to thank to Dr. Ayşe Özmen for her help in modeling part, her valuable explanations to make me understand about my work and to share her technical knowledge during whole of my study. None can deny that her great advices guided me successfully.

I would also like to show my appreciation to Dr. Alper Çevik for all of his help in analyzing dataset. He always kept his help for any technical problem during my thesis study. He has an important role to start for this study and understand dataset structure. He was a pioneering member of our work because of his remarkable talent

for data-mining.

With his undeniable amount of help, Dr. Semih Kuter was also an important part of this work. I would like to show my glad to him for sharing his knowledge.

I am also grateful to Dr. Erik Kropat for all his support and to my friends MSc. Ece Köksal and MSc. Mehmet Alp Üreten for all their understanding and great help for the technical problems of my study.

I also want to thank to my family,especially my older brother Faruk Gütmen, all of my relatives and all of my friends, especially Serap Hınıslı and Arzu Tekin, for their support and their understanding during my education years.

Most importantly, I would like to point out that my mother, Ülker Gütmen, is the most important person who I would like to show and mention about this work. I found my power within me to succeed this study thanks to *her endless love and existence deep in my heart*. With all my hearth full of her love, I want to thank her for giving me a chance in this world to become a person who is able to success such a great work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

α_i^k	Value of each Vocational Skill k by Student i
β_j^k	Value of each Vocational Skill k from the view of particular Job j
θ_j^k	Weight Factor
ε	Noise Term
σ^2	Variance (Error Variance)
AAE	Average Absolute Error
Adjusted R^2	Multiple Coefficient of Determination
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
AU	Administration and Service
AWT	Technology Knowledge Accelerator
BEP	Back-Error-Propagation
BF	Basis Function
EE	Electro-electronic
GCV	Generalized Cross Validation
GUI	Graphical User Interface
IT	Information Technology
LOF	Lack of Fit
LR	Linear Regression
LS	Least Squares
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Splines
ME	Mean Error
MG	Mechanical and Mining-Metallurgical
MS	Medical and Social
MSE	Mean Square Error
NNET	Neutral Networks Toolbox

O	Common Skills
R	Correlation Coefficient
R^2	Coefficient of Determination
RL	Agriculture and Forestry with Environmental Protection
RMSE	Root Mean Square Error
SPM	Salford Prediction Modeler
ST	Artistic
TG	Tourist and Gastronomic
W	General Skills
X_O	Student's Total Value of Common Skills for particular Offer
X_W	Student's Total Value of General Skills for particular Offer
Y_Z	Student's Total Value of Professional Skills for particular Offer
Z	Specific Skills

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

It has become a widely except-able truth that having a "good education" could bring a "good position". What is more, it may be said that "good employment" leads to "good work" and increases the chances to get fulfillment and happiness in human's lives. In this respect, the main points eventually are (i) a better "matching" on the labor market, (ii) a better Education System - young people becoming better prepared for their professional life and, finally, for their personal future, and (iii) a better Human Resource Management by companies and institutions. This thesis is a scientific and operational step that will be able to become a road map for better understanding of undergraduate and graduate students, i.e., it facilitates and encourages advanced decision making which leads to improved working conditions and superior advancements in instruction [8]. Organizations are required to rapidly identify and respond to changes, to be imaginative and to adapt fast to new circumstances [14, 16]. As a scientific support to fulfill this requirement, a first study [17] has been done based on *General skills, W*, in Wielkopolskie which is one of the regions in Poland populated by about 3.6 million people [11]. From another point of view, in this thesis, *Specific skills, Z*, have been proposed to study in the development of Educational and Economic Network in this region.

In the model, student's total evaluation value of professional skills (X_Z) is determined as *response variable*, Y and all other aspects are implemented as input variables, X_i's. We have 15 inputs composing the input vector $\mathbf{X}$. The name of all input variables are

listed in Table 2.1. We aim to figure out the relationship regarding how much the response variable and the inputs match each other in accordance with the demands of job offers and the skills which students hold by employing Linear Regression (LR), Multivariate Adaptive Regression Splines (MARS) and Artificial Neutral Networks (ANNs). Here, a MARS model has been created and settled as our core approach of applied mathematics. In fact, MARS abbreviates Multivariate Adaptive Regression Splines. While LR is used for a basic description and discussion of regression, ANNs from Artificial Intelligence have been developed by us, statistically compared with the MARS model, herewith disclosing that our chosen approach by MARS is very much competitive and powerful, useful and promising. According to comparison results, MARS model gives the best results in our aim for this study. Additionally, for each student *i* and for each job offer *j* where student *i* applies to, we have the pair (i, j), representing a row vector of data $\boldsymbol{x}$ in our application.

The flow of changes within the labor market asks for a steady cooperation between science, industries and educational institutions [11]. Emerging Insurance Industries, sectors of Finance and Retirement Fund departments can also be among companies and institutions. Apart from these organizations, there will be many other upcoming industries and their different branches which have a vast influence on the feelings about, e.g., certainty, security and safety, of young people, on having a perspective in their working and private life, on realizing, developing and using their skills, on contributing to the economy, the society and to humanity, on feeling safe and being secure – now and in future. Eventually, all of the points above affect life standard and life quality in today's society and in future young families. In this context, one important issue and important purpose is that insurance companies aim at decreasing the number of "cases" through research studies, measurements and campaigns in the fields of Actuarial Sciences, Management of Risk, Education and Human Resource Management. Furthermore, although the employees represent "main goods" of a labor market, the number of skilled workers in the market is less, and even, not enough to satisfy the demands by the companies [11]. In this regard, young workers lack the attitude of "I want to". In other words, "will power" required by companies as well as skills and competences are lacking [15]. Yet, companies' high demands along with these deficiencies creates an essential challenge. This leads to a need by

the companies for an identification of current motivational systems built upon their foundations, which would be possible in a well developed and refined economy. It is also essential to know the needs and motivations of young workers to provide them to work more effectively and in a more encouraged way - with further will to stay in the company for a longer time in their life. This thesis is a pioneering project in this sense, indeed, fostered by modern applied mathematics.

On the other hand, when employees are hired as, e.g., beginners or workers, it can be considered that they are possible candidates in the future for becoming managers - upcoming leaders. This makes it very important that employees and, eventually, employers, are to be selected with a high care so that they are sufficient in social skills as well as their technical knowledge [10] since the managers are not only leading and kinds of technical works but also managing a social network among workers, contract partners, and coworkers. Along with all these skills, other information also exists in the system. Therefore, in our system, there are also available some clusters of Competences and Qualifications which have been arranged according to the Skills. For a company, determining and predicting further demands for competencies is a responsible and strategic management task as development built on competencies ensures some benefits in the market [9]. Consequently, to scientifically clarify the aforementioned needs and goals, we focus on two modeling approaches which set up a frame and delivered the response data: a General Model and a Reduced (or Simplified) Model. We may call this step the "forward problem". The following core part of the study consist of the modeling work – the "inverse problem" - which gives us our estimation results.

Since our final aim is to improve the Networking between companies, universities, and students, to foundations of better solutions, better occupation and fulfillment for all of them, we proposed to contribute to more happiness and health - being of high importance also for modern emerging companies of insurance and institutions of pension fund systems. Moreover, we contribute to new scientific advances in the design of respected technologies of Data Mining, Machine Learning, Deep Learning and Artificial Intelligence.

This thesis is a pioneering contribution to the **Networking** *and related* **Modeling**

sides of **Human Resource Management** *by Applied Mathematics and Artificial Intelligence that become carefully compared for the problem posed and solved. As the* **Human Factor** *is richly involved in this study, eventually aiming at a best matching between the humans' offers and the jobs' needs, according to the motto: "The right person at the right place (and at the right time)", this thesis naturally serves to personal happiness, fulfillment, safety and health and, herewith, to emerging modern industries such as those where* **Actuarial Sciences** *are applied. Hence, and for the various aspects presented and discussed, this thesis can be regarded as the beginning of a research agenda of the future.*

1.2 Survey on Skills, Comptences and Qualifications

Many associations require to identify the particular and discrete abilities and information related with employments. Naturally, this is required for extremely specialized employments, such as those in Information Technology or Engineering. Companies address the requirement for quite special knowledge and abilities by adding a certain list or a so-called inventory to supplement the competencies. This inventory, with the combinations of competences, gives increasingly complete data required by the organization to select the competitors with the best fit, to recognize preparing needs, and to get ready for progress. A **competence** has mainly expresses itself by two types: ***soft competences*** or ***behavioral***, and ***hard competences*** or ***technical*** ones. Behavioral competences articulate personality characteristics or traits of employees, whereas technical competences include gained expertness, e.g., knowledge and skills. Behavioral competences (life skills) comprise the *abilities* as one branch of competence, such as: facilitation, initiative, impact and influence or resilience, and many others [3]. The other elements which are learned through experiences from study or practice, i.e., *skills* and *knowledge*, are needed for success in specialized fields by considering them as technical competences, like marketing research, information technology, business operations or human resources management [4]. Necessarily, each job offer must contain these three types of competencies: skills, knowledge and ability [14]. To succeed, candidates need to show the right *portfolio* of knowledge, ability, and skill. A student, candidate or job-seeker cannot be compelling without

the combination of:

- ***Knowledge***: "*know-how*", "*know-what*", "*know-who*" and "*know-who*", expertise or specifies areas, such as: accounting, engineering, history or branches or specific techniques for areas, etc. Knowledge is what one gets from the course (in educational institutions and outside) which is related to carry out one's task.

- ***Ability***: to fulfill or satisfy complicated demands by activating and performing *natural* or *inherited* attitudes or behaviors. Ability is the set of the human skills acquired with learning or directly obtained from daily-life experiences, to deal with problems and handle questions. Abilities are used properly and accountably for handling routine works. For abilities, examples are: professionalism, resourcefulness, collegiality or analytically thinking.

- ***Skill***: a learning *capacity* or *potential* to realize the outcomes you need with greatest proficiency and certainty. Skill is the application part of one's knowledge in work or profession. Skills are something comprehended to be able to perform job functions including cognitive, technical or interpersonal issues.

Altogether, these elements form *competence*. To make clear what a competence is and to give a simple and lively example, let us think about hair dressing. During the time when a hair dresser is cutting hair, keeping hand steadily is a ability. The ways or techniques which one can learn from a hair dressing course are a knowledge, and to be able to do something which one has learned in an attractive or proper way is a skill.

As a summary, competences are more than skills. They must consist of knowledge, behavior or abilities of humans. Definitely, competencies should always embody with various skills. Technical competences, social competences and key competences are some core types of competences [27].

Qualifications are formal outcomes of the ways of assessments and approval from situations, when a competent body decides that a person has accomplished learning results which are agreeing with given conditions. Such a body may be a teacher, an expert, a certification or supervision institution, etc. In order to be able to say that one

has a qualification, one must have got the essential certificate, diploma, degree etc. with the *approved* required skills and competences from a competent body, its associated authority or institution. We note that often the employing company, when it also serves for education and examination, can be an institution of qualification as well. However, as an additional point, when one has got qualification, it does not mean for all, the skills and competences under this qualification could be used efficiently for related task. Sometimes, that could happen because of a lack of competent bodies' attendance, and also candidates' less practice in time. At the end, qualifications should not be only criterion to decide about candidates. That is why, employers should focus on employees' skills more during the hiring process.

The triple of Skills, Competences and Qualifications are of high importance for the all: students and job-seekers, companies and related agents, and the local and whole economies. They are main keys to get human capital in a best possible way to the labor market, to corporations and institutions and, eventually, to the economic future prospects. In this way, Data on Skills, Competences and Qualification, and their Data Mining strongly contribute to fulfillment, health, perspectives and social peace, regarding individuals, their relations, and entire nations.

In our project [17] and this study, we address many competences and, under them, we have numerous Skills. We partition these Skills into three classes: ***Specific skills***, Z, ***Common skills***, O, and ***General skills***, W.

1.3 Outline of the Thesis

In Chapter 2, we explain how we prepared our Regression Dataset from the so-called system's dataset of the project team. Here, we discuss a calculation method for some of the variables and the mathematical model a frame that results from this method, the ways which were developed to get a quantitative dataset from qualitative data information, and the treatment of missing data points. Moreover, weight factors and features of students and so-called the reference time issue are discussed. Chapter 3 is given as a methodology part. In this chapter, Linear Regression is presented, also to summarize what is behind regression basically. Then, MARS method is presented

as our main model, and ANNs has been chosen by us a method for comparison. Chapter 4 contains three models, calculated by LR, MARS and ANNs, respectively. The statistical criteria to compare the performances of these models are introduced in Chapter 5. Finally, we summarize all our results and discussions in Chapter 6. In a compact form, we state our findings and we submit ideas about our further studies as an outlook.

CHAPTER 2

DATASET AND ITS PREPROCESSING

2.1 Preparation of Data for Regression

At the beginning of the study, with the dataset which we received from the system "Zawodowcy" [2], we have the set that includes both quantitative and qualitative information. After some preparation and calculation, we described all of our qualitative variables as quantitative variables. These preparations were done for introducing categorical data into our statistical model by several ways explained below.

The dataset represents the information for both users (students) and offers (companies' job offers). In the system, we have 3346 students of IT - information technology department from between October 2012 and November 2015 and also other students with at least three IT professional skills from other departments (more information about students' departments will be given below). In the system, 566 of these students are active. They were contacted and interviewed from technical high schools during the years 2012-2015. Additionally, there were 423 companies in different sizes: small, medium and large, respectively. From these companies, our system included 619 job offers, from which 318 are active offers.

From both students and offers, we gathered lots of information about several kinds of skills that are related to questions about which skills students have and which skills are needed by the companies' job offers. These skills are divided into 3 groups: *common skills, O, specific (technical or professional) skills, Z* and *general skills, W*. Group O is the cluster of common skills for professions in one area. In the educational system, there are eight fields characterized as follows: (1) administration and

service (AU); (2) construction (BD); (3) electro-electronic (EE); (4) mechanical and mining-metallurgical (MG); (5) agriculture and forestry with environmental protection (RL); (6) tourist and gastronomic (TG); (7) medical and social (MS); (8) artistic (ST). Common skills for all areas are in the group W. The last group Z includes the characteristics of the each professions in the system. Totally, we have 3807 skills collected from both students and companies' job offers. 3324 of them are from professional skills, 335 of them are from common skills, and the rest part (148) is from soft skills, respectively.

Furthermore, we introduce all further information as quantitative variable to our model. Hence, we obtained 16 different variables. In fact, 15 of them are used as *input variables (predictors)* and the one (X_Z) (we shall call it as Y_Z) is the output variable that is related to professional skills. Here, the response variable Y may be viewed from numerous perspectives, such as:

- Strong dependence between vocational skills and the other skills which appear in the model.
- There are deficiencies in the educational system: a situation that students tend to always being passive, just repeating and having not much creatively. In addition to these drawbacks, companies often have a quite limited picture, few ideas only about students and the possible fulfillment levels which students could provide and receive through a task, respectively.
- The description and content of the job offer can be too narrow in order to include the variety or diversity of skills.
- Our model also bases on the choice of a maximal number of basis functions, the number of the input variables,and, a maximal number of interactions (multiplications) between the input variables.

For 16 variables, more details (names and definitions of variables) are stated in Table 2.1.

Some of these variables which are given above were collected directly from the system. However, sometimes, we needed to use other ways to introduce qualitative vari-

Table 2.1: All used variables for developing Regression models.

Names of all variables	
Student's total evaluation value of professional skills for particular offer:	Y_Z
Student's gender:	X_1
Student's birthday:	X_2
Student's profile creation date:	X_3
Job offer visible from:	X_4
Job offer visible to:	X_5
Job offer's date of work start:	X_6
Job offer's date of creation:	X_7
Job offer's type of employment:	X_8
Job offer's time work:	X_9
Job offer's no. of work station:	X_{10}
Job offer's shift work:	X_{11}
Job offer's non stationary work:	X_{12}
Job offer and student are in the same county:	X_{13}
Student's total evaluation value of common skills for particular offer:	X_{14} or X_O
Student's total evaluation value of general skills for particular offer:	X_{15} or X_W

ables to our MARS model. To give an example, one time, we consider the county of the students where they live and which county the job offer comes from. If the county is not the same for both student and job offer, we used "1" to represent the situation, otherwise, we wrote "2". Similar to this way, we introduced gender of the students as well. Moreover, we have several kinds of offer-shift style, offer non-stationary work style, etc. Similar ways are also followed to introduce these variables.

Furthermore, for the variables which depend on any time issue, we used *reference time point* (for closer information, please cf. Subsection 2.1.2).

In addition to these preparations, to get students' total evaluations value of any kind of skill for the companies' particular job offers: X_O, X_W, X_Z ($Y_Z = Y$) we generated our own calculation method by considering the skills that students have and are required by companies' job offers. (More information about our calculation method for these variables will be given in Subsection 2.1.3.1.) [1]

[1] Regarding from X_2 to X_7, we refer to *reference time* or *absolute time* as own input variable for all variables which depend on time. By this, we are in one line with the use of reference time like in other parts of management, e.g., in accounting.

Why do we also evaluate the students with at least 3 IT skills?

We should offer diversity and add variety of the labor market which be gathered and represented in the system to our MARS modeling. In fact, we include IT-related data of students who wish to join the IT branch, but also enrich this particular sub-dataset with the further data of sufficiently IT-skilled students who prefer other branches of the labor market, i.e., of the economy. This united dataset will serve with our Regression Dataset and, eventually, for our MARS model. In order to specify the meaning of being sufficiently IT-skilled by a value, called as a threshold, we chose the discrete values of 3 IT skills. Actually, in future studies, this threshold value could be assessed statistically by the help of Model Selection in order to ensure a high or even maximal information content of our MARS model. In other words, the set of those additional students gets smaller when the threshold becomes increased, whereas this set grows when the threshold is diminished.

With the help of these thresholds, we are tuning and giving diversity and variety to the MARS model. This is very important, as we include into this model students with all of their skills and not just with their intentions to join a special branch of the industry, such as the emerging IT sector. That is, a higher variety of students or their skills leads to a higher diversity and property of the sector.

An additional aspect of the aforementioned model preparation is related with the overall interests of the companies. In fact, we can not imagine a successful company which is full of persons that are only interested in computer science. Every company also needs other staff members in order to have a good, a pleasant and fruitful mixture. The recruiters can give the opportunity to possible job applicants to enter with other skills to enrich the company through such further and often quite different skills. Providing the labor market with a broad spectrum of skills also helps to not exclude or discriminate such different people in the labor market competition, and to support a peaceful and creative atmosphere in companies.

Each company should give a basic chance to all other people who are interested in working in a company to show their skills which are related to companies' offers in a wider sense, and to contribute to company, economy and society by them. This group of persons can here compensate to some extent their deficiencies in IT skills

with other skills, e.g., social skills or further skills. Through relevant students, having at least 3 IT skills, we can obtain an atmosphere in the firm that turns their interests into skills. Actually, in the process of time their IT skills can enhance and mature through learning, identifying with the company, its goals and personal, and through joint successes in the company. By this way, we can experience (and enjoy) whether and how the distribution of different kinds of skills matters in the context of labor market.

2.1.1 Missing Data Points

The issue of missing has always been quite important for data analyses. Missing information is a typical real-life phenomenon and can significantly affect the ends that may be drawn from the information. In statistical areas, missing data points, or missing values, arise when no information has been received for an observed variable. For our study, we computed the median of each column to replace missing data.

We used the ***median*** instead of ***mean***, since all our values are integers, not decimals. In fact, a main reason for preferring the median (from ℓ^1 regression) over the arithmetic mean (from ℓ^2 regression) consists in its ***robustness***, which is so important in our real-world situations characterized by noise and further uncertainty. This advantage becomes very clear in the presence of outliers [12]. That we did not prefer to use some polynomial interpolation here to overcome missing data is due to the fact that our data are not enumerated along the time axis or any other underlying and connecting variable; there is not yet any evidence about a *functional* dependence on the enumeration underlying. The use of multi-dimensional polynomial interpolation is intended as an alternative approach which could give a chance to benefit from the *power* of the dataset, so that we could regress a variable with missing data on further variables without missing data. That sub-project of this study would require at least three steps: (*i*) statistical investigation about which variables could be regressed on meaningfully and with a sufficient correlation (or functional relationship), including model selection and dimensional reduction; (*ii*) setting up and solving a suitable multidimensional interpolation model; related to this: (*iii*) investigating and improving (regularizing) any ill-possessedness related with the interpolation polynomial, as it

would also influence the stability of our MARS model eventually. As a matter of fact, we could even address a *piecewise* multidimensional interpolation polynomial, e.g., a multivariate spline to more accurately and in a more stable manner fulfill the interpolation. In any way, such a sub-project should remain in the right *size* in relation to our entire study and always remain sufficiently *handy*.

Indeed, the treatment of *missing data* is one of the modules within our entire methodology and program, which is situated between the given (*robustness*) data and our intended MARS model, and is evaluation and use. The more stable that *missing data* module will be, the less can noise or later changes in the data impact the MARS model and make it behave in an unstable way.

In contrast, the weight can be reduced for any student who says, or pretends, that he/she has *all the skills* with very many and high values of self-assessment.

2.1.2 Reference-Time Issue

We need to address *time* as a variable, mainly *absolute time*. In fact, we wish to know and refer to *absolute time*, e.g., which time points on the time axis have to be included in our data in order to build up the final MARS model. At those discrete time points the given measurements, observations, etc., are considered as valid. But even a bit earlier or a bit later, the model might change with new data points and be different. This reflects that we are addressing a *changing world*; it needs to have *time* as an input variable. Later on, we will show that our MARS model with time variable(s) included has a better statistical performance values than MARS without time variable(s). In fact, time needs to be *absolute*, in other words, *historical* or *chronological* time, since *relative* or *incremental time*, such as time differences, is hiding the time at the time axis, i.e., when in the flow of time the data were taken and when, eventually, the model is meant and understood. This hiding of absolute time would mean a great *modeling loss* and loss of statistical performance of our MARS model. According to the terminology from the mathematics of differential equations and dynamical systems, we shall call MARS model with time variable as *non autonomous*, and MARS model without time variable as *autonomous*.

Moreover, if we have new dates (time data), then the whole dataset changes and, eventually, our final MARS model will become different. In data mining, it is preferred that a dataset is stable, we also say *regular* (what we sometimes achieve after some process of regularization) and reference times should be constant times, i.e., very certain times. For this reason, the reference time should ideally not be a particular, e.g., personal, company-owned, etc., element or component of the data set. Actually, the values of those elements or components could be affected by *noise*, by wrong information, etc., and they might even have to be corrected later on; such corrections would lead to vast changes in the whole Regression Dataset and, eventually, in the whole MARS model. Thus, we cannot choose any time from the dataset as a reference time. Because of all these reasons, we chose reference time "0" as $31.12.1999$. It is used that the initial and included day January 1, 2000, is the starting time "1" and the last date which belongs to variable is the ending time. As indicated above, there are different time variables. Indeed, we also have information about (a) the time when a job offer is created by a company, (b) the time of work start, (c) the start point for visibility of job offers in the system, (d) the end point for visibility of job offer in the system, and (e) birthdays of the students, (f) the creation date of students' profile. Then, to enhance our model performance, we decided to extend our knowledge about the entire dataset of the system by treating those six time values (see (a)-(f)) as new, supplemental input variables. In our MARS model, these time variables will be among the *input variables*, also called as *predictor variables*. Typically, we would like to see how changes in one or several predictors lead to changes of the response variable. This kind of Sensitivity or Stability Analysis is explained more closely in Section 6.2.

2.1.3 Our Mathematical Model to Prepare the Regression Data

In this section, we are right in between the System's Dataset which we got from our project [17] partners at PUT (Poznan University of Technology), and the Regression Dataset which will be the basis for creating our MARS model. This important preparation or preprocessing step, right between two datasets, requires to closely investigate the links between students' skill and companies' demands in terms of required skills. By this way, we generate new variables according to the 3 different classes of

skills, namely, X_O, X_W, and X_Z, as we mentioned before. We chose X_Z as our response (or output) variable that we aim to understand through all the other variables, including X_O and X_W eventually. Therefore, we need to introduce X_Z through a "Calculation Method" subsequently. We are also going to indicate how this will lead us to the MARS model f for X_Z, as we shall discuss in an *Interpretation* part below.

2.1.3.1 Calculation Method

For each student *i* and for each job offer *j* where student *i* applies to, we have the pair (i, j), representing a row vector of data. We herewith have 2173 different pairs (i, j) in our Regression Dataset.
We have the value α_i^k for each vocational (professional) *Z* skill *k* of *i*. To determine the value of α_i^k, look whether skill *k* which student *i* has is requested by *j*:

if ***yes***: note the given evaluation value for each *i*;
if ***not***: note the value 0.

At the same time, we consider the value of β_j^k for each vocational (*Z*) skill *k* from the view of particular job *j*. Value of β_j^k (to be assigned by the company or practitioner) for any skill *k* says to what an amount (level) company needs the skill for job *j*, independent of any particular student *i*. In addition to these values, we have the weight factors $\theta_{i,j}^k$ which decision makers give us for each *k*, taking into account the entire distribution of the student *i*'s skills. Each weight factor defines the importance of each skill *k* for each pair (*i,j*), so that it can be included into the model: fairness criteria between students' and companies' possible preferences, e.g., in focused (low variance) skill distributions of students.

After all these analyses, for given (i, j),
we multiply these three weighted terms and sum them up:

$$y = y_Z = \sum_k \theta_j^k \alpha_i^k \beta_j^k. \tag{2.1}$$

In this way, we have built the total evaluation value y_Z about professional skills of

pair (i, j). This step can be called as the *Forward Problem*, in distinction from the *Inverse Problem* which shall become the main content of this thesis.

Analogously, we find the total evaluation values x_O, x_W about common (*O*) and soft (*W*) skills of each pair (i, j).

Herewith, along all those pairs (*i,j*), we get the compound (total) skill values for our response variable ($Y = Y_Z$) and for certain input random variables (X_O, X_W).

In fact, the formula of Eq. (2.1) represents the algorithm for our *General Model*. After careful discussion, because of some local and technical problems and lack of information, we could not obtain certain result for the value of weight factors. That is why, we needed to turn our general model into a so-called *Simple Model*. In this simplified version, our formula is the following:

$$y = y_Z = \sum_k \theta_{i,j}^k \alpha_i^k, \tag{2.2}$$

where for the pair (i, j) which consists of the *i*-th student and the *j*-th job offer.

For data point $\boldsymbol{x}_1 = (\, \boldsymbol{x}_{1,1} \,, \boldsymbol{x}_{1,2} \,)$, where $\boldsymbol{x}_{1,1}$ is data vector on the 1-st student and $\boldsymbol{x}_{1,2}$ is the data vector on the 1-st job offer (i.e., $(i, j) = (1, 1)$).

We note that any parameter $\theta_{i,j}^k$ takes the value $0, 1, 2$ instead of the values which are calculated by decision makers in this form. Herewith, these numbers show the required level of each skill by companies. The triplet $0, 1, 2$ means "*no need to have*", "*be nice to have*" and "*have to have*", respectively, for each needed skill by related job offer.

2.1.3.2 Weight Factors

At the stage of processing the knowledge-accelleration (system's) data, preparing our Regression Data, we have two different approaches to calculate weight factors θ_j^k, with respect to job offer *j* and skill *k*. As a matter of fact, there are further approaches, and any formula from real-life practice could be considered and used.

Note: In the special case which was introduced by the practitioners, the weight factor θ_j^k does not depend on *i*, in mathematical term: $\theta_{i,j}^k \equiv \theta_j^k$.

Approach 1. With our first so-called *Averaging Approach*, we determine the value of the weight factors θ_j^k by using an additive form such as the arithmetic mean. In this formula, we calculate our factors by multiplying the average of given importance of all skills which are related to a job offer and the number of the job offers which demand these skills. When computing the average importance of all related skills, in our study, we use the arithmetic mean. Herewith, the importance of each skill in the system is decided by the decision makers in the companies. The following small example illustrates the basic idea.

Example 1. Assume that we have skill *communication or communicativeness* which has been chosen in 100 job offers. In 50 job offers, the assessment was 4 and in another 50, it was 5. So, the average assessment is $4,5$. Then, the weight is counted as:

$$100 \times 4,5 = 450.$$

Moreover, we would like to mention the geometric mean which can also be used to assess weight factors. By this way, our formula would be based on a multiplicative form of goals. Other ways can be the median or further forms of determining an average.

Approach 2. For our second way, called as *General* or *Distributional Approach*, the weight factors θ_j^k can be personalized, e.g., they can also depend on student *i* and on his/her (a) number of skills and (b) distribution of skills. When we evaluate the skills of a student, we determine the weight factor for each given skill and given job offer, based not only on the number of skills which they have, but also the distribution of the skills with their evaluation values. If in his/her supervised self-assessment a modest or particularly ***focused*** student names quite few skills, we increase the weight factor value of his/her skills a bit, herewith raising the assessment curve, in order not to dominate or suppress students with the aforementioned self-assessment behavior and profile with a small number of skills. In contrast, the weight can be reduced for any student who says, or pretends, that he/she has *all the skills* with many and high values of self-assessment. Herewith, at our Regression Dataset preprocessing stage, we offer to our MARS modeling and, eventually, to the decision maker, additional

and important ideas and modeling techniques in order to present fairness. What is more, by that technique and its fine-tuning of increasing or diminishing the weight factors, the company can also represent and conducts its preferences about how much of a ***generalist*** and how much of a ***specialist*** its future employee should list.

Indeed, we tried to create and propose a more fair and more widely operational environment for setting up and analyzing our Regression Dataset and, eventually, our MARS model. In this way, decision makers may give a chance to different and non-mainstream persons, to enter a company, to participate in joint works, to enrich the company by their various and soft skills and, herewith, to even improve and mature their skills, especially, in the typical IT domain. Furthermore, firms are provided with an additional tool to respect and protect people who are disabled or shy, and they can give encouragement to the young people to show and enhance their capacities.

2.1.3.3 *Interpretation*

Let us present our *response variable* as follows: Y_Z:
the introduction of a student's total evaluation value of professional skills requested by a company for a particular job, this means, the potential of a student as he/she would introduce it into the job if the corresponding company will hire him/her.

Similarly, we present and calculate the input variables X_O and X_W.

In fact, $Y(= Y_Z)$ represents the response variable of our model. When the dataset is uploaded into MARS algorithm together with all the other data on further variables, such as demographic, biographic, economic ones, then MARS will give us the model formula f:

$$Y = f(\boldsymbol{X}) + \varepsilon.$$

CHAPTER 3

METHODOLOGY

3.1 Linear Regression

Linear Regression (LR) models are keystones and very valuable tools in several scientific areas for many years. Regression methods are especially used in science and engineering, on statistical and financial aspects and for the assessment of expectation.

The fundamental point of LR is to demonstrate a connection between factors by building linear equations. We utilize linear regression as a common tool to fit a forecasting model and to compare it with other methods. One can find a relation between observations and based on them, a model (between one dependent variable and some independent variable) easily by using a linear method. It is a rather easy way to make a prediction according to the actual or historical values. The dependent variable is represented by Y and it will be directly based on the independent variables (predictors) which are shown by the input vector $\boldsymbol{X}$. Even though LR has several types, we will mention about simple linear regression and multiple linear regression basically. In the event that there exists just a single independent variable, at that point, the model is named as *simple linear regression.* On the other hand, the multiple linear regression covers one dependent variable and more than one independent variables.

The principal thought behind linear regression, as expressed by its simple case, is to fit a solid line through observations points. This line can be represented as:

$$Y = \gamma_0 + \gamma_1 X + \varepsilon, \tag{3.1}$$

where Y is the estimated dependent random variable, X is the independent random

variable, γ_0 is the intercept, γ_1 is the regression coefficient, and ε is the noise term with $N(0, \sigma^2)$. The model which is given above has a single independent variable, that is why it shows a "*Simple Linear Regression Model*".

Simple linear regression can also be written in a different form [20]:

$$\hat{Y} = \hat{\gamma_0} + \hat{\gamma_1} X,$$

where the independent variable is the *X* term, $\hat{\gamma_0}$ is the intercept and the slope, $\hat{\gamma_1}$, is the change in the mean of the distribution of Y by change by one unit in *X*.

In any case, however, not every one of the observations is actually on the plot. At this time, the term of noise, ε, should be considered in the formula. Then, the equation occurs as follows:

$$Y = \gamma_0 + \gamma_1 X + \varepsilon.$$

Herewith, the error term, ε, herewith, is used to represent the real relation between response variable and regressors, and it is supposed to have a normal distribution $\varepsilon \sim N(0, \sigma^2)$.

Moreover, since the response variable is a random variable, it has the two parameters mean and variance. They can be specified respectively in the following way of conditional expectation:

$$E(Y \mid X = x) = \gamma_0 + \gamma_1 x$$

and

$$Var(Y \mid X = x) = Var(\gamma_0 + \gamma_1 x + \varepsilon) = \sigma^2.$$

Generally, the parameters γ_0 and γ_1 have to be estimated. In order to perform that, least-squares method looks for optimizing the function on the sum of squares, $S(\gamma_0, \gamma_1)$, of the vertical differences between observed (response) values which belong to y_i and the estimated line.

Least-Squares (LS) criterion is to minimize $S(\gamma_0, \gamma_1) = \sum_{i=1}^{l} \varepsilon_i^2 = \sum_{i=1}^{l} (y_i - \gamma_0 - \gamma_1 x_i)^2$. The function *S* has to be reduced according to the unknown coefficients and (x_i, y_i), $i = 1, 2, \ldots, l$, are herewith the given data, and the ε_i values are the residuals.

Naturally, the different number of the predictors is the principal distinction between simple linear regression and multiple linear regression. *Multiple linear regression* generally includes more regressors than simple linear regression. Multiple linear regression model with *m* regressors can be written in the following form [7]:

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_m X_m + \varepsilon.$$

In just the same way, the noise term ε is a random variable which is distributed according to $\varepsilon \sim N(0, \sigma^2)$. It usually shows the error between model and data.

In Figure 3.1, the model is represented with $\boldsymbol{X}$ in $\mathbb{R}^2$. However, in our study, we will use 15 different regressors to represent the general random vector $\boldsymbol{X}$.

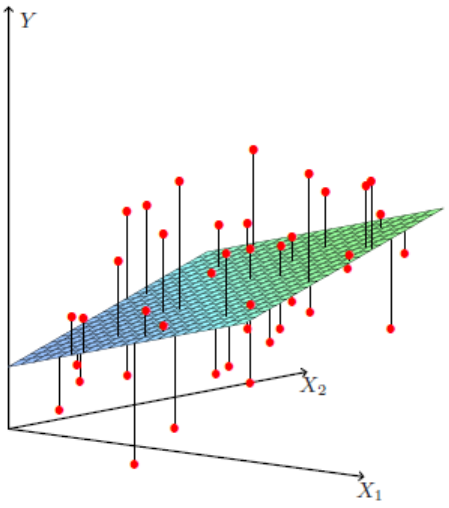

Figure 3.1: Illustration of Linear Regression model [20].

In general, multiple linear regression model is stated as follows:

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \ldots + \gamma_m X_{im} + \varepsilon_i \\ &= \gamma_0 + \sum_{j=1}^{m} \gamma_j X_{ij} + \varepsilon_i, \qquad \text{where} \quad i = 1, 2, \ldots, l. \end{aligned}$$

On the other hand, in simple linear regression, the response variable, *Y*, correlates with only one regressors X_i for each data point. That is why, one can define the expected value of multiple linear regression as:

$$E(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_m x_m,$$

where γ_0 is the intercept value, and the other parameters γ_j are univariate slopes.

Remark: We would like to emphasize that in this chapter, we prefer to use the indexes *i* and *j* according to the usual habits and customs in statistics and data mining, i.e., for enumerating data and input variables, respectively. Herewith, this notation may serve for the readers' convenience, especially, given the already very high number of variables, parameters and indexes in this work.

In fact, we trust that here and in later chapters, there will be no confusion with the meaning of the indexes *i* and *j* in the main research part of this thesis, where they represent students and companies' jobs. Please note the basic and *symbolic* relation between the two meanings: $\boldsymbol{i} = (i, j)$, where $\boldsymbol{i}$ indicates a data point and (*i*, *j*) signifies a pair consisting of a student and a job. ■

In this part of our methodological chapter, the notations, definitions and fundamental properties of simple linear regression and multiple linear regression have been mentioned. Moreover, in the project of this thesis, we employed multiple linear regression model with 15 different input variables.

3.2 Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS) developed and presented by Jerome Friedman in 1991 might be introduced as an augmentation of direct models that *consequently* generate nonlinearities and associations [20]. However, originally, the first version of MARS was carried out in and performed by Salford System Software, called Salford Prediction Modeler (SPM), for the clients' benefits [6]. MARS is an information-driven technique, and not at all like other widely utilized display-driven or managed learning techniques and calculations; it is fundamentally a regression

model. In fact, regression, generally being utilized for forecasting and prediction, is principally built on the strategies of least-squares estimation, and estimation of highest probability.

In modeling of applied mathematics, multivariate analysis may be a set of processes for estimating the relationships among variables. Multivariate analysis is additionally accustomed to perceive that among the independent variables there is an association with a certain variable, and to explore the characteristics of these relationships. Especially, regression analysis enables one to understand how the typical value of the dependent variable changes, i.e., one may see how the response (output or dependent) variable is affected by the other variables (input or independent). It was intended to anticipate continuous numeric results and it is worthwhile to perform variable choice, variable change, connection recognition and self-testing for large numbers of measurements in a very speedy way, and the majority of all of this *naturally* [34].

Basis Functions (BFs), here likewise known as ***splines***, are incorporated as indicators that are originally dependent on data, which allows for a very flexible regression, e.g., local-global distinction. In fact, MARS has the flexibility to approximate the BFs' contributions in order that each additive, but also interactive or multiplicative impact of the predictors are admitted to join the investigation of the response variable. In MARS, all conceivable positions and all indicators are followed and found together with each conceivable connection in the model. After MARS having decided about the ideal amounts of BFs and knots, the least-squares estimator strategy is connected so as to build the last model that gives the best estimation of the dataset with the rest of the BFs. The last MARS model is established with a two-stage process including Forward and Backward Stage.

At the MARS' forward stage, the model is commonly an over fitting model which consists of an extensive arrangement of BFs. The BFs are added into the model quickly and consistently by MARS' algorithm, until achieving the greatest number of BFs in the model. Consequently, the model received contains all possible BFs without considering the contribution of these BFs to the model performance. The forward stage develops a large model that includes a rather vast set of BFs; subsequently, it should be cleaned from excessive BFs, and this requires another stage which is called

as the backward stage.

At MARS' backward stage, the over-fitted approach is modified to lessen the complication of the model, i.e., its complexity. In any case, the model still regulates the general performance with the fit to the information. At this stage, the BFs that lead to the minimal increment in the *Residual Sum of Square (RSS)* are erased from the formula at every step and, accordingly, an ideal model is acquired [23]. BFs are removed from the model so as to obtain the ideal essential amounts by thinking about their minimal contribution to the entire performance of the model. The key criterion at this stage to stop expects to accomplish an ideal balance between bias and variance, between accuracy and stability.

While conducting forward and backward stage, MARS divides the whole dataset into several sub-sections and specifies various mathematical algorithms for every subsection. Eventually, by these algorithms, MARS establishes a relation between some sub-section of input variables and the response variable. In order to achieve that, MARS utilizes linear BFs which are generated from the given dataset. The two one-dimensional BFs, $[x-\phi]_+$ and $[x-\phi]_-$, are described subsequently:

$$[x-\phi]_+ = \begin{cases} x-\phi, & \text{if} \quad x > \phi, \\ 0, & \text{otherwise}, \end{cases}$$

$$[x-\phi]_- = [\phi-x]_+ = \begin{cases} \phi-x, & \text{if} \quad \phi > x, \\ 0, & \text{otherwise}, \end{cases}$$

where x, $\phi \in \mathbb{R}$.

Figure 3.2 represents basis functions to give an example. These truncated power functions or hinge functions are the basic components of MARS models. Both functions are a *reflected pair* together. They show the piece-wise linearity for the model in every dimension.

The general MARS model which is represented subsequently shows a link between

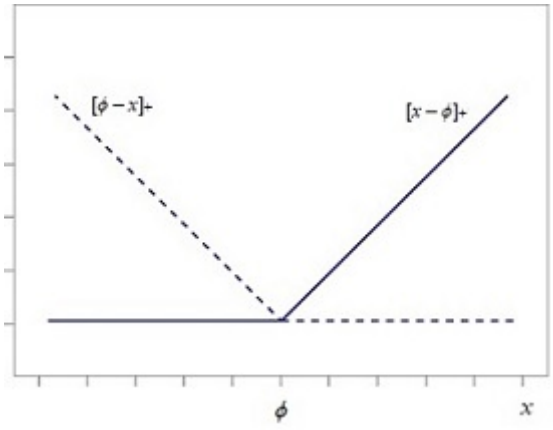

Figure 3.2: Sample or Truncated Functions $[x-\phi]_+$ (solid) and $[\phi-x]_+$ or $[x-\phi]_-$ (dashed) are used by MARS model.

response or output variable, and input variables or predictors:

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{3.2}$$

where the response variable is demonstrated by Y, $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)^T$ is a vector representing each of the input variables, and ε is an added stochastic term called as noise with zero mean and limited variance. Here, X_j $(j = 1, 2, \ldots, m)$ are the corresponding Cartesian coordinates with *m*-dimensional knots $\boldsymbol{\phi}_i = (\phi_{i,1}, \phi_{i,2}, \ldots, \phi_{i,m})^T$ at input data vectors $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$ $(i = 1, 2, \ldots, l)$ with the input data values x_{ij}.

According to this, the set of 1-dimensional basis functions looks as follows:

$$F : \{[x_j - \phi]_+, [x_j - \phi]_- \quad | \quad \phi \in \{x_{1j}, x_{2j}, \ldots, x_{lj}\}, \quad j = 1, 2, \ldots, m\},$$

where *l* is the number of observations.

In the forward part of MARS, the model which continuously adjusts the dataset is developed with BFs of the set F and also products of BFs. Eventually, the model becomes the following:

$$Y = \gamma_0 + \sum_{n=1}^{N} \gamma_n F_n(\boldsymbol{X}) + \varepsilon, \tag{3.3}$$

with the output variable *Y*, the vector of input variables $\boldsymbol{X}$ whose corresponding data vector is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$ $(i = 1, 2, \ldots, l)$ with the input data values x_{ij}, and noise term ε, which herewith is a randomly added noise assumed with zero mean and fixed variance. Moreover, the number of BFs, *N*, is shown in the existing model, $F_n(X)$ is the *n*-th BF introduced in the set F or the multiplication of at least two functions from this set, the constant γ_0 is the intercept and γ_n is the coefficient of the *n*-th BF.

In addition, these current BFs remain in the model together with the recently built BFs in order that spline fitting turns into a spline fitting of higher order. The definition of the *i*-th multivariate BF for all observations represented by $(\boldsymbol{x}_i, y_i)$ $(i = 1, 2, \ldots, l)$ is the following:

$$F_n(\boldsymbol{x}) := \prod_{t=1}^{T_n} [H_{t,n} \cdot (x_{v(t,n)} - \phi_{t,n})]_+,$$

where the multiplication number of truncated linear functions is T_n and the related input variable with *t*-th truncated linear function is $x_{v(t,n)}$ in the *n*-th BF. Furthermore, $H_{t,n} = \pm 1$ and for variable $x_{v(t,n)}$, the knot value is $\phi_{t,n}$.

MARS looks through all combinations of generated terms and all values for each variable at the forward pass. Then, the knots and corresponding pair of BFs that contribute the largest decrease in residual error are selected by the model. After all, at this stage, MARS adds the succeeding products to the algorithm. The product-adding part lasts until the value of maximum number of terms is achieved. By this way, one gets an over-fitted or highly complex model with many incorrect terms.

After constructing the over-fitted model in the forward step, the backward step starts to built a pruned model by removing redundant basis functions to decrease the complexity of the model. This removing lasts until arriving at the best sub-model, $\hat{f}_b$, with the optimal number of terms *b*. While the forward step includes terms to the model as pairs, some terms which contributed least into the model are excluded from the model. MARS utilizes Lack-Of-Fit (LOF) criterion which is described by Generalized Cross Validation (GCV) to get an optimum number of terms, *b*, that presents the best predictive fit. The GCV formula was developed by Friedman in 1991 in the

subsequent way [20]:

$$\text{LOF}(\hat{f}_b) = \text{GCV}(b) := \frac{\sum_{i=1}^{l}(y_i - \hat{f}_b(\boldsymbol{x}_i))^2}{(1 - P(b)/l)^2},$$

with sample observations number, l, the estimated best model, $\hat{f}_b$, and the efficient number of parameters, $P(b)$, whose formula is

$$P(b) := v + dK,$$

where v is the number of linearly independent functions, K is the number of knots, and d shows a "price" of penalty for one unit of complexity. Actually, d is generally designated as 3. In fact, $d = 2$ is preferred when we have a merely additive model [22].

Eventually, the best-fit model is chosen which reduces GCV along with the help of the lack-of-fit criterion in the backward step.

Moreover, one can analyze both discrete and continuous datasets by using MARS. This is rather convenient to realize and implement. Another fundamental favorable position of the model originates from the utilization of certain piecewise, truncated linear one-dimensional functions. Moreover, MARS technique is able to catch some critical points, interactions and nonlinearities in the model [32]. Essentially, MARS is a non-parametric procedure and it might be considered as a linear model augmentation. The principle and favorable position of MARS is that it is not in need of any forecasting, and it is more flexibly adaptable than linear models.

In Figure 3.3, non-parametric MARS model is compared with parametric LR model.

3.3 Artificial Neutral Networks

Artificial Neutral Networks (ANNs) comes from a family of techniques that forms efficient computational models. An ANN consists of several basic artificial neurons

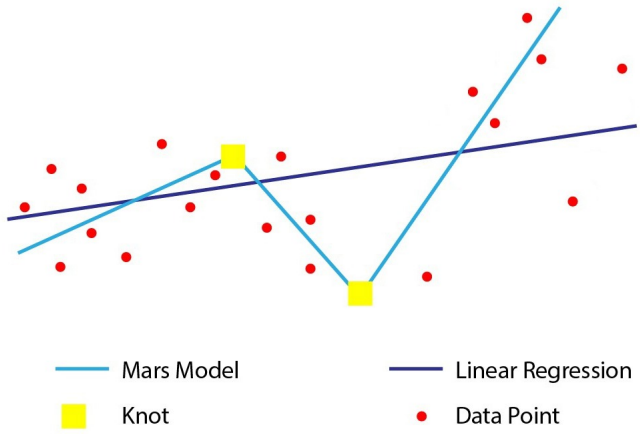

Figure 3.3: Visualization of a MARS model.

in order to be used as processing elements. These processing units are structured like neurons of the human brain [13]. Neurons have been put to use for assignments such as image recognition, grasping linked memory and object detection, etc. [19]. ANN is able to learn from given examples and, thus, making it successfully doable to produce complex nonlinear models easier [34]. It has benefits like working through incomplete datasets, and making generalization for unfamiliar cases too [26].

Input variables which are the components of input vector $\boldsymbol{X}$ and output variable Y are also in existence for ANNs. Input variables are entered into a neural network by a user while the output variables sets as outputs. This is a main reason of why ANNs can be employed when conclusions are done about specific unknown information based on already known informations. *Artificial neuron* is the most important part of any ANNs. Given quantities of input signals should arrive at neurons and, then, via some relation of certain weights, a signal amount is presented to the neuron. A threshold limit that triggers the neuron is specified for a certain neuron. At the end, an overall stimulation of the neuron is turned into the whole of an ANNs.

An ANNs consists of 3 layers: *input layer*, one or more *hidden layers* and *output layer*. By the input layer, it is possible to introduce a given dataset to the network, and the hidden layer processes information. After that, in the output layer, a result or output value y has occurred and a conclusion is drawn. The layers have a certain number of processing units. All the units are core elements and constitute artifi-

cial neuron models. Units from different consecutive layers are connected with each other. Weights are determined according to given connections to specify strengths of connection. All the associated weights are comprised by the weight vector $\boldsymbol{w}$.

A neuron sends output signals with values which are calculated in two steps. ***Step 1*** is about multiplication of the weights *w* with signals *x*; their numbers is denoted by n. These weights are liable to a specified function. This is called as an *internal processing* function which is actually a summation practice. The outcome of an internal *e* processing depends on the given input-output function. This is named as Φ an *activation function*; overall, it constitutes our ***Step 2***. In the end, the output signal resulted from the subsequent formula for each neuron:

$$y := \Phi(\sum_{i=1}^{n} w_i x_i) = \Phi(\boldsymbol{W} \cdot \boldsymbol{X}), \tag{3.4}$$

where $\boldsymbol{X}$ is the input data vector, $\boldsymbol{W}$ is the weight vector, Φ is the activation function and y is the output signal. With Eq. (3.4), we follow classical notation in literature on ANNs (cf. Section 3.3).

Another type of network is the *multilayer* network based on topology. Through such a network, it is possible to define input, hidden and output layer separately. The connections between these layers have a particular structure. However, these are often complete, full connections, with all neurons and their layers being connected. This is a network that follows one direction, whereby a finite number of response variables can be permitted. Those finitely many responses are well-known in statistical learning, e.g., in *classification*. For such a network, the formula can be stated as follows [18]:

$$\boldsymbol{Y} := \Phi_{wy}(\boldsymbol{W}_{wy} \cdot \Phi_{ukr}[\boldsymbol{W}_{ukr} \cdot \Phi_{ws}(\boldsymbol{W}_{ws} \cdot \boldsymbol{X})]), \tag{3.5}$$

where

$\boldsymbol{X}$ is the input data vector, $\boldsymbol{W}$ is the weight matrix for the appropriate layer, Φ is the activation function for the appropriate layer, and $\boldsymbol{Y}$ is the output data vector.

The main purpose of this network is to perform a certain function in order to get the expected response values with given inputs. This aim is obtained by the way of a learning process network, i.e., the proper change in the neurons' weight values.

We would like to mention that ANNs often benefits from a "library" of problems that are already solved "optimally". Given a new problem submitted, many of these (library or benchmark) problems are tried out in the hope for resemblances - to fit into the new model, i.e., for an optimal and efficient solution of the new problems posed. For all these reasons and preparedness of an ANNs, it has become a highly efficient and successful modeling technology whose rise has contributed very strongly to our modern industries and societies, e.g., in commodity markets, especially, the energy sector, and markets of natural resources, in medicine and, very naturally, in neuroscience. In fact, *Artificial Intelligence* (*AI*) is so much in the public discussion and future decision making in both our economies and our societies. Along with Machine Learning, mostly ANNs is meant whenever a discussion is about modern AI.

Herewith, we recognize the high importance of ANNs in our modern time, as a computational model and very helpful technique. However, ANNs certainly is much less of an analytical technique that could explore and disclose structural or theoretical connections and analytic properties. Just at this regard, the qualities of the more mathematical tool of MARS come into the comparison and unfold their remarkable usefulness, so that, in our study, MARS turns out to be very much competitive to ANNs.

CHAPTER 4

MODEL SELECTION AND ANALYSIS

In our study, the main aim is building a model to understand the relation between input variables and output variable. After collecting information from our practitioners' system, we made many additions as being mentioned in Chapter 2 for rearranging variables in order to get the Regression Dataset. Later on, we separated the dataset of 32595 real-life observations into two groups: $\%70$ of the whole dataset for training with 22816 data points, and $\%30$ of the dataset for testing with 9778 data points. In this thesis, *Linear Regression* will be used as an informative tool as a modeling method, while *Artificial Neutral Networks* will be employed as an emerging reference method of comparing with our MARS model.

4.1 Linear Regression Model

Linear Regression is used as a rather primitive statistical modeling method in our study. We just used LR in order to give some basic information about regression. What is more, later on we shall add a few words of interpretation and comparison with our MARS model.

Moreover, the residuals of Linear model which come from our dataset do not follow the generally assumed normal distribution. That is why, our dataset is not suitable to apply Linear Regression method. For the application of Linear Regression, the preconditions which which can be studied in [12, 20] and are addressed in our discussions below should be fulfilled by the dataset. However, in the case of our dataset, as Figures 4.1-4.4 show, the required condition of Normal Distribution is not satisfied.

In fact, the figures also demonstrate many high (in terms of absolute value) residual values and their distribution.

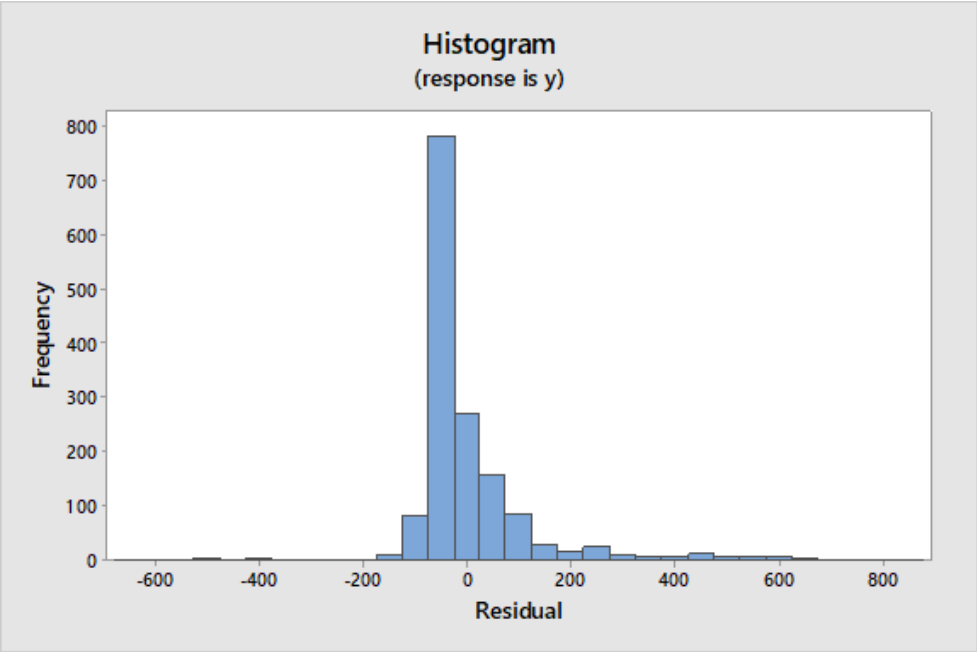

Figure 4.1: Histogram of the residuals for Linear model.

In Figure 4.1, the *Histogram* (Residual-Frequency) shows to us that residuals, as samples of the noise term in the model, do not obey a normal distribution around mean 0. Mostly, the residuals are accumulated between -200 and 0, rather than immediately around 0. What is more, the histogram of our set of residuals is *right-skewed*, not symmetrical.

Another graph, Figure 4.2 on the *Normal Probability Plot* (Residual-Percentage) reveals whether the distribution of residuals is normal or not, and; therefore, it represents the case of normal distribution of the residuals, too. If two certain "curves" are about the same, then we consider the normal-distribution assumption as being fulfilled. In closer detail: the "dotted curve" which is comprised of the data-based residuals of our Linear Model is plotted against the "ideal" linear curve, representing coincidence between practice (data) and theory (probability). If the "dotted curve" closely enough approximates the line, then we view the normal distribution as fulfilled; otherwise, it is seen as not satisfied sufficiently. In our thesis, at this additional agenda point on Linear Model, we confine with the user-friendly GUI (graphical user

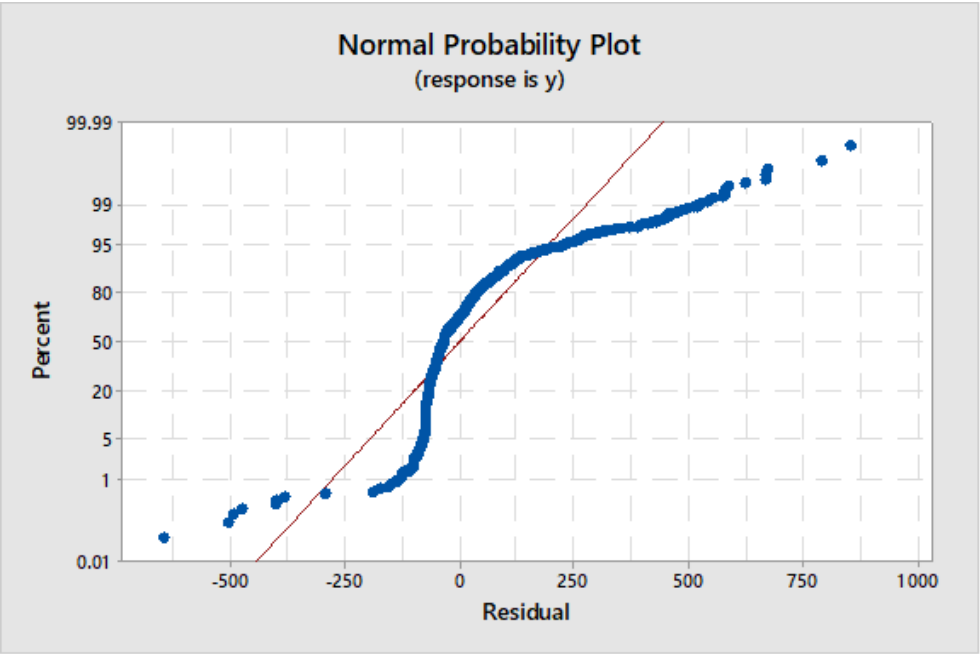

Figure 4.2: Normal Probability Plot of the residuals for Linear model.

interface) [1], given by Figures 4.1-4.4, rather than sophisticated techniques of testing or validating.

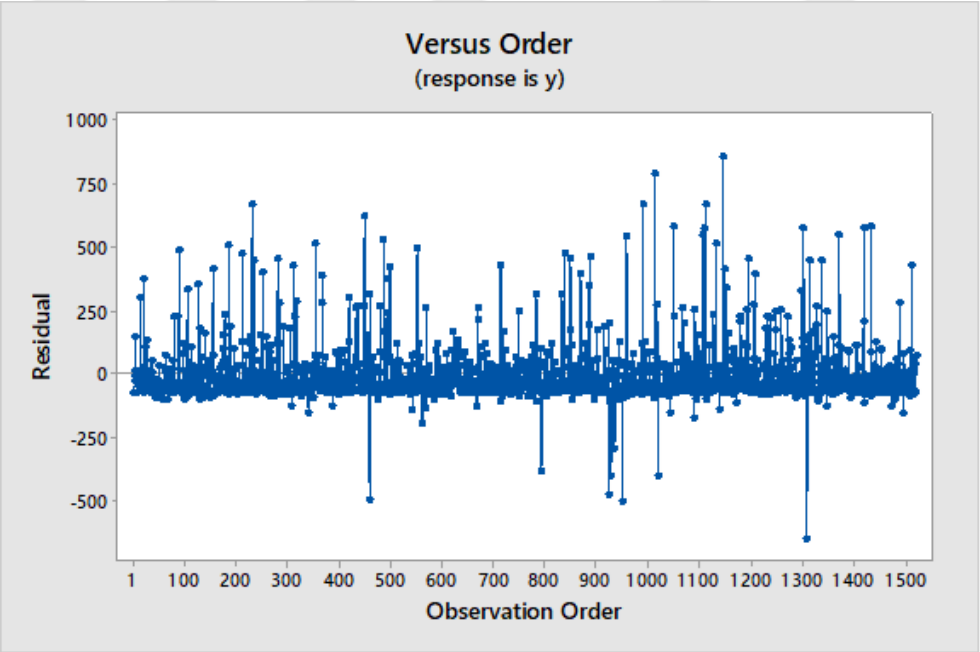

Figure 4.3: Residual Versus Order plot for Linear model.

Additionally, in Figure 4.4, entitled as *Versus Order* (Residual-Observation Order),

the residuals for data of the training set observations are plotted according to the observations' order. It also shows that we have more (larger) positive residual values than negative values. That shows another difference of our dataset compared to normal distribution.

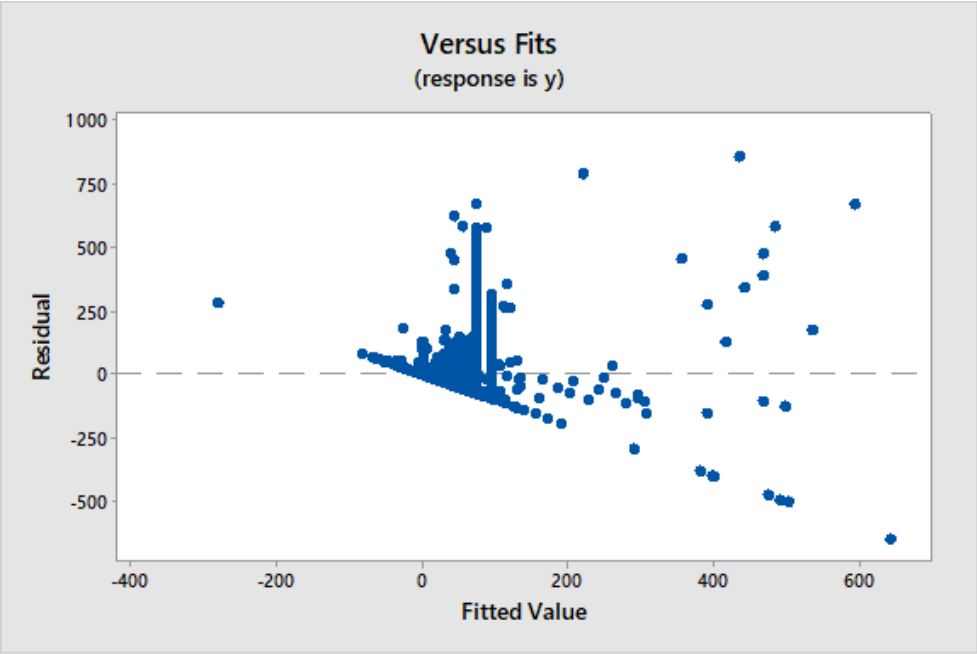

Figure 4.4: Residual Versus Fits Plot for Linear model.

Similarly, in Figure 4.4 on the *Versus Fits* (Residual-Fitted Value), the residuals are plotted for each fitted value.

Eventually, our Linear Model is calculated and found to be of the following form:

$$Y = 83.8 - 29.98X_8 - 5.88X_{10} + 25.97X_{11} + 4.237X_{14} - 0.4420X_{15} + \varepsilon. \quad (4.1)$$

In Table 4.1, all *p*-values, *t*-test values and coefficients of our Linear model are displayed.

Although all *p*-values in Table 4.1 are significant, we could not evaluate the LR model since the *pre-conditions* of LR are not satisfied.

Table 4.1: *p*-values, *t*-test values and coefficients for Linear model.

Variables	Coefficients	*p*-value	*t*-test
Constant Term	83.8	0.000	6.32
X_8	-29.98	0.000	-4.31
X_{10}	-5.88	0.000	-5.73
X_{11}	25.97	0.003	3.02
X_{14}	4.237	0.000	17.91
X_{15}	-0.4420	0.000	-5.52

4.2 MARS Model

Aiming at our intended MARS methodology model, the optimally determined model implying a reduced number of BFs, and given an upper bound of M_{max} many BFs, are obtained. For this purpose, we have concluded both MARS' forward step and MARS' backward step by its software (SPM 2018), addressing generalized cross-validation (GCV) given in Section 3.2, and the model providing the optimized predictive fit is selected. In our research, the maximal degree of interaction is 3, whereas the bound M_{max} is preassigned as 35. After MARS' Backward Stage, the number of BFs went down to 21. In consequence, the BFs given at the end of the backward stage are displayed subsequently:

$$
\begin{aligned}
BF_1 &= \max\{0, X_{14} - 2\},\\
BF_2 &= \max\{0, 2 - X_{14}\},\\
BF_3 &= \max\{0, X_8 - 1\} \cdot BF_1,\\
BF_5 &= \max\{0, 1 - X_{15}\} \cdot BF_2,\\
BF_6 &= \max\{0, X_4 - 5281\} \cdot BF_5,\\
BF_7 &= \max\{0, 5281 - X_4\} \cdot BF_5,\\
BF_8 &= \max\{0, X_4 - 5269\} \cdot BF_5,\\
BF_{10} &= \max\{0, X_4 - 5297\} \cdot BF_5,\\
BF_{12} &= \max\{0, X_4 - 5163\} \cdot BF_5,\\
BF_{14} &= \max\{0, X_2 + 1695\} \cdot BF_3,
\end{aligned}
$$

$$
\begin{aligned}
BF_{16} &= \max\{0, X_{15} - 31\} \cdot BF_1, \\
BF_{17} &= \max\{0, 31 - X_{15}\} \cdot BF_1, \\
BF_{19} &= \max\{0, 4864 - X_3\} \cdot BF_{16}, \\
BF_{20} &= \max\{0, X_5 - 5426\} \cdot BF_5, \\
BF_{21} &= \max\{0, 5426 - X_5\} \cdot BF_5, \\
BF_{22} &= \max\{0, X_4 - 5115\} \cdot BF_5, \\
BF_{24} &= \max\{0, X_4 - 5178\} \cdot BF_5, \\
BF_{26} &= \max\{0, X_4 - 5233\} \cdot BF_5, \\
BF_{28} &= \max\{0, X_5 - 5426\} \cdot BF_3, \\
BF_{29} &= \max\{0, 5426 - X_5\} \cdot BF_3, \\
BF_{30} &= \max\{0, X_3 - 5373\} \cdot BF_5, \\
BF_{33} &= \max\{0, 32 - X_{15}\} \cdot BF_3, \\
BF_{34} &= \max\{0, X_5 - 6575\} \cdot BF_5,
\end{aligned}
$$

The best MARS model with the BFs shown above is also presented in the following form:

$$
\begin{aligned}
Y = {} & 12.41 + 4.21 \cdot BF_1 + 10.26 \cdot BF_3 - 81.92 \cdot BF_5 \\
& - 29.62 \cdot BF_6 + 0.369 \cdot BF_7 + 20.81 \cdot BF_8 \\
& + 11.22 \cdot BF_{10} - 9.461 \cdot BF_{12} - 0.005 \cdot BF_{14} \\
& - 0.1405 \cdot BF_{17} - 0.00172 \cdot BF_{19} - 0.0531 \cdot BF_{20} \\
& - 0.317 \cdot BF_{21} + 3.040 \cdot BF_{22} + 8.216 \cdot BF_{24} \\
& - 4.307 \cdot BF_{26} - 0.0315 \cdot BF_{28} + 0.098 \cdot BF_{29} \\
& + 0.263 \cdot BF_{30} - 0.374 \cdot BF_{33} + 0.116 \cdot BF_{34} \\
& + \varepsilon.
\end{aligned}
\tag{4.2}
$$

In our MARS model, 7 input variables turned out to be included, we may say: important, among 15 variables. (Some of the other input variables may be important as

well but statistically "covered" through dependence by the help of the 7 included variables already.) Variables that are affecting the response variable through the MARS model are these ones: student's birthday (X_2), student's profile creation date (X_3), job offer visible from (X_4), job offer visible to (X_5), job offer's type of employment (X_8), student's total evaluation value of common skills for particular offer (X_{14} or X_O), and student's total evaluation value of general skills for particular offer (X_{15} or X_W). In fact, 4 of these 7 input variables are related to *time*. That means: 4 of our 6 time variables (cf. Subsection 2.1.2 and, in particular, Table 2.1) are visible in our MARS model. This shows a high time dependence of our model for the particular application of human resources. We have, such to say, a *4 times*-dependent MARS model, a 4-times dimensional model.

In Table 4.2, all estimated parameter or coefficient values, γ_n, corresponding to the selected BFs of our MARS model are displayed.

Table 4.2: Coefficient values of MARS model.

γ_0	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}
12.41	4.21	10.26	-81.92	-29.62	0.369	20.81	11.22	-9.461	-0.005	- 0.1405
γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{17}	γ_{18}	γ_{19}	γ_{20}	γ_{21}	γ_{22}
-0.00172	-0.0531	-0.317	3.040	8.216	-4.307	-0.0315	0.098	0.263	-0.374	0.116

Moreover, there is an interesting and positive point of comparison: when we look at all the included variables in our Linear model and MARS model, respectively, then we have 3 common input variables for both models: job offer's type of employment (X_8), student's total evaluation value of common skills for particular offer (X_{14} or X_O), and student's total evaluation value of general skills for particular offer (X_{15} or X_W) for both models. Let us recall that Linear model just contains 5 input variables (see Eq. (4.1)) and MARS model implies 7 input variables only.

4.3 Artifical Neutral Networks Model

In this investigation, the feed forward and multi-layered Artificial Neural Networks (ANNs) meta-models were induced by the help of Back-Error-Propagation (BEP) learning algorithm in order to provide an entrepreneurship prediction. For building up the proposed NN model, we applied MATLAB software (NNET toolbox). As seen

in Figure 4.5, we have 1 input layer, 1 hidden layer and, as usually, 1 output layer. We generated an ANNs with 14 inputs, 1 neuron in the output layer and 20 neurons in the hidden-layer for our ANNs model. Here, w is representing weight, applying on the b stands for bias and tansig function is employed, a so-called an activation function, in applying, on the hidden-layer neurons.

Additionally, it has already been appeared that one hidden layer regularly tends to generalize well and can demonstrate any multi-dimensional work with higher accuracy [24, 25].

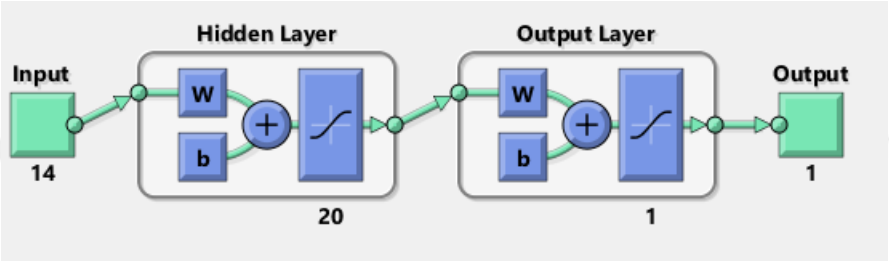

Figure 4.5: ANNs model for Professional Skill approximation.

Note: In Table 2.1, we have shown 15 input variables of our work. The correlation coefficients between all these variables were checked by us and we noticed that the correlation value between *job offer visible from* (X_4) and *job offer's date of creation* (X_7) is so high. As the variables with high correlation value are taken off easily by our refined MARS methodology, we just involved all 15 variables together in our MARS approach. However, since Artifical Neutral Networks cannot tolerate any such a kind of situation, before modeling with ANNs, here we reduced our input variables number to 14 by deleting the variable job offer's date of creation (X_7).

CHAPTER 5

STATISTICAL EVALUATION

5.1 Statistical Performance Criteria

In total, we analyze 32595 data points from all observations for our 3 different models: LR, MARS and ANNs. In order to get results more precisely, the dataset is split into two groups as train dataset and test dataset, as mentioned before in Chapter 4. While the aim was to receive a potential predictive model with train dataset, the test dataset was used to measure the performance of the models [21]. Furthermore, before comparing between the prediction models, LR, MARS and ANNs, for each of the three approaches, we have to choose our best model among other possible models which shows a better accuracy performance in the course of statistical evaluation [29]. In order to understand which model has a superior predictive ability than others, we addressed Average Absolute Error (AAE), Root Mean Square Error (RMSE), Multiple Coefficient of Determination (Adjusted R^2) and Correlation Coefficient (R) as accuracy measures. The accuracy criterion tries to assess the predictive capability of the model developed [30]. AAE measures the average magnitude of the errors. It shows the differences between actual response value and predicted response value. The following equation shows the formulation of AAE:

$$\text{AAE} := \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| = \frac{1}{N}\sum_{i=1}^{N}|e_i|,$$

where y_i is the *i*-th actual response value, $\hat{y}_i$ is the *i*-th predicted (fitted) response value, e_i is the error term, and N the is number of observations.

RMSE indicates the results for the square root of mean-squared residuals. RMSE is calculated with the formula stated in the following:

$$\text{RMSE} := \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}};$$

here, again for y_i and $\hat{y}_i$ we used the *i*-th actual response value and the *i*-th predicted (fitted) response value, respectively, and N is the number of observations.

Additionally, R expresses the relation of linearity between predicted response and observed response. Namely, R is represented as given below:

$$R := \frac{1}{1-N}\sum_{i=1}^{N}\Big(\frac{y_i - \bar{y}}{s_y}\Big)\Big(\frac{\hat{y}_i - \bar{\hat{y}}}{s_{\bar{y}}}\Big),$$

where $s_y = \sqrt{\frac{1}{1-N}(\sum_{i=1}^{N} y_i - \bar{y})^2}$, $\quad s_{\bar{y}} = \sqrt{\frac{1}{1-N}(\sum_{i=1}^{N} \hat{y}_i - \bar{\hat{y}})^2}$ and

$$R^2 := \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of the actual values, $\hat{y}$ is the predicted response variable, $\bar{\hat{y}}$ is the mean of the predicted response variable, $s(y)^2$ is the standard deviation of the actual response variable; $s(\hat{y})^2$ is the standard deviation of the predicted response variable and N is the number of observations.

Adjusted R^2 indicates the variation percentage in the response variable generated by the model. The fomula is as follows:

$$\text{Adjusted } R^2 := 1 - \left(\frac{(N-1)(1-R^2)}{N-K-1}\right),$$

where N is the number of observations, K is the number of independent variables in the regression equation.

Table 5.1: Accuracy performance criteria: results based on train and test dataset for LR, MARS and ANNs.

	Regression and Estimation Models					
	LR		MARS		ANNs	
	Train	Test	Train	Test	Train	Test
AAE	73.914	71.097	37.321	38.719	38.822	47.599
RMSE	119.092	111.324	69.262	73.050	73.331	94.752
Adjusted R^2	0.197	-0.005	0.725	0.556	0.693	0.256
R	0.447	0.311	0.854	0.767	0.837	0.626

As a conclusion, in any model, smaller values of AAE and RMSE closer to 0 indicate better results, whereas the values of R and R^2_{adj} closer to 1 are preferred [28].

5.1.1 Results and Comparison

As mentioned before in previous Section 5.1, models were evaluated with regard to performance criteria by implementing related formulas. For both train dataset and test dataset of LR, MARS and ANNs, the performance criteria results for accuracy are shown in Table 5.1.

Now, we can infer from Table 5.1 that the AAE value for both LR and ANNs is higher than the AAE value of MARS. This shows that any prediction with MARS has a high reliability [33]. The RMSE value of MARS is lower than the other two ones' RMSE values, which means that MARS ensures more accurate results. For MARS model, the value of Adjusted R^2, 0.725, is closer to 1; it is better than the other two. The R criterion has the highest value for MARS model comparing with the ones of LR and ANNs. A higher correlation coefficient value shows a stronger relationship, i.e., this lets us get more accurate results as well.

In the aforementioned comparison, the results obtained from the train dataset were used. It can be easily seen that test dataset results for the models always affirm all our main statements about train dataset's results.

Likewise, the values listed in Table 5.2 are compared with the stability performance for each comparison criterion on both train and test dataset. The values closer to 1 indicate a more stable model. Stable methods are the ones that perform equally well

Table 5.2: Stability results of performance criteria for LR, MARS and ANNs.

	Regression and Estimation Models		
	LR	MARS	ANNs
AAE	1.0396	1.0374	1.2260
RMSE	1.0697	1.0546	1.2921
Adjusted R^2	-0.0253	1.3039	0.3694
R	0.6957	0.8981	0.7479

on both training and test datasets [31].

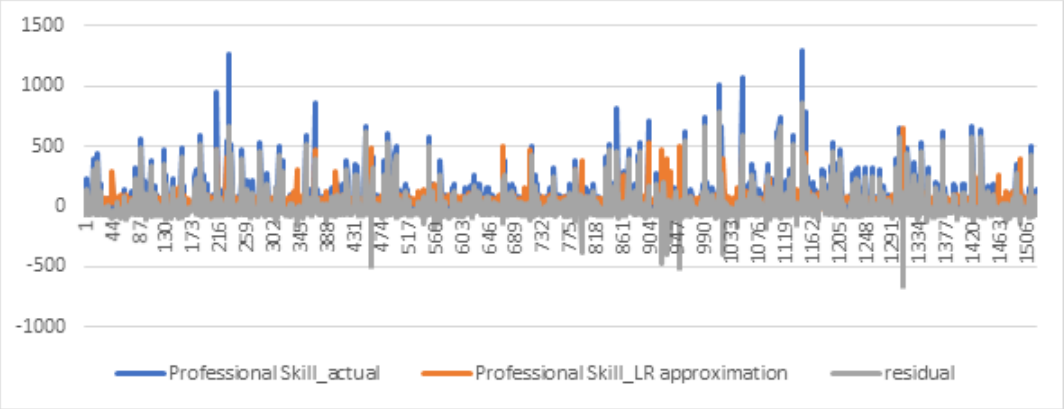

Figure 5.1: Actual and predicted values with LR model for training data.

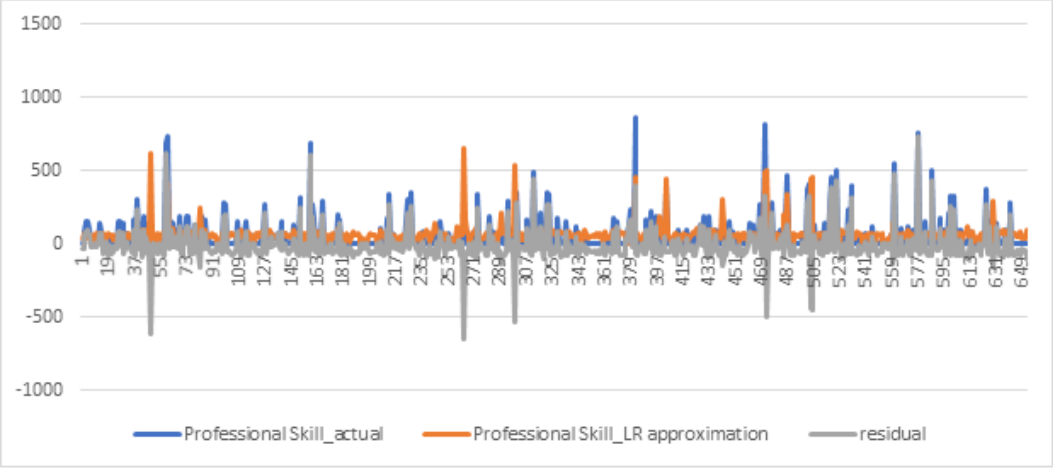

Figure 5.2: Actual and predicted values with LR model for test data.

Under normal conditions, the residuals could be bigger for the test dataset. However, in LR model, the values of residuals for the test dataset are smaller than the ones for the train dataset. In fact, among the training dataset points there are more high values, slightly bigger than among the test data points, and the famous *proportionality effect*

of LR applies [12].

In addition to this, the preconditions are not all fullfiled for LR. In fact, a proper foundation of comparison between training set- and test set-based training and test based residuals, in terms of as preconditions, is not satisfied in our LR model.

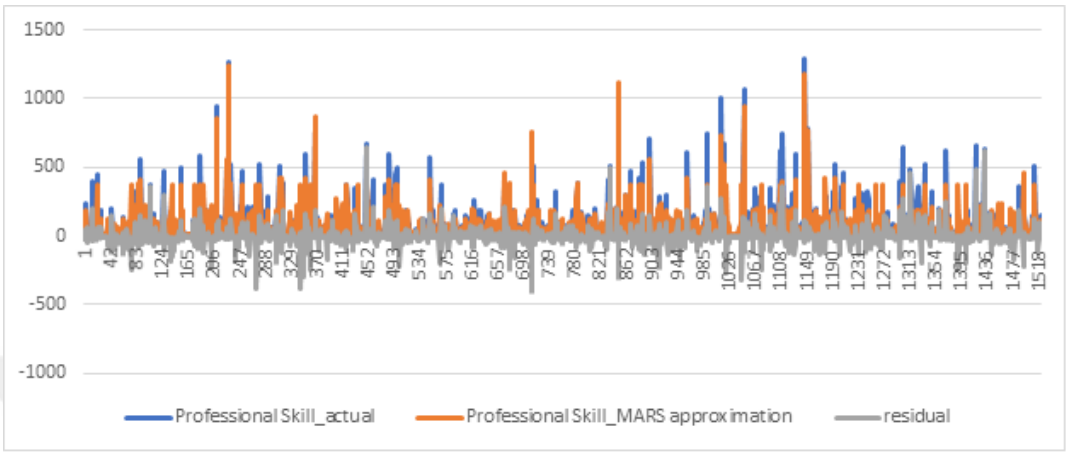

Figure 5.3: Actual and predicted values with MARS model for training data.

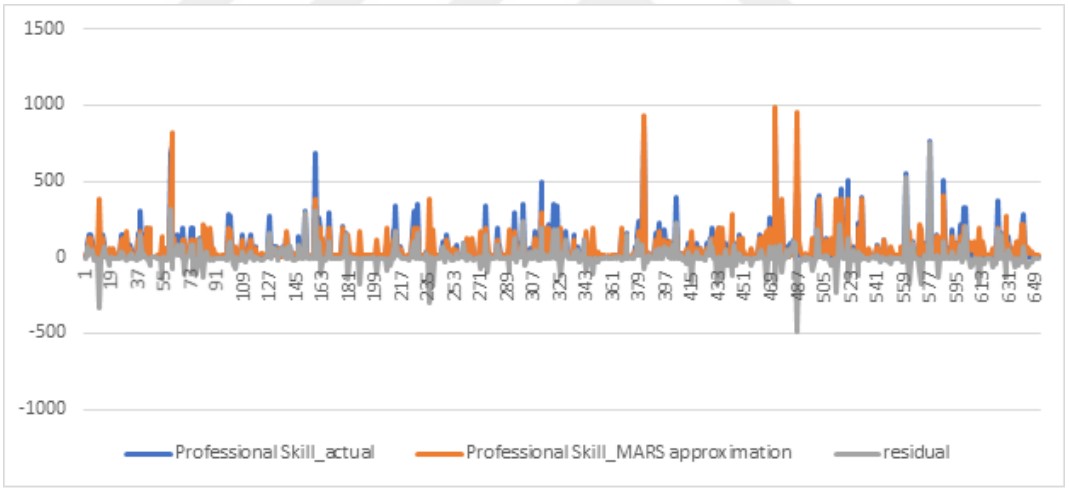

Figure 5.4: Actual and predicted values with MARS model for test data.

Although, while the residuals of train dataset become enlarged in MARS' backward step compared with forward step, the residuals for the test data remains small because of the excellent prediction capability of MARS model. MARS has to give up accuracy goal a bit after the forward step in order to satisfy the stability goal, too. This means that it takes off some basis functions. Then, the "cost" of this can be a slightly reduced accuracy and, thus, slightly increased residuals. We would like to emphasize that

MARS does all of this in a very well-controlled analytic way.

Let us underline this excellent and highly competitive property of MARS as we see it based on the test data, as one of our premium findings among all our results.

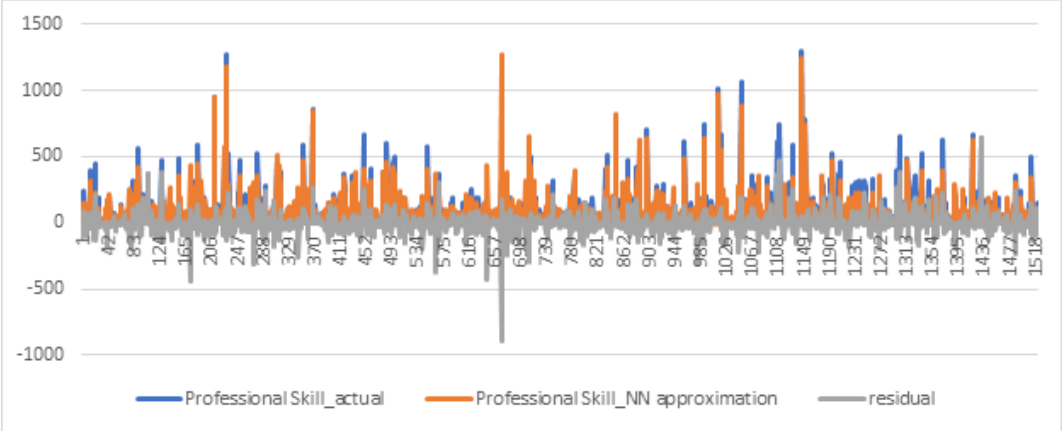

Figure 5.5: Actual and predicted values with ANNs model for training data.

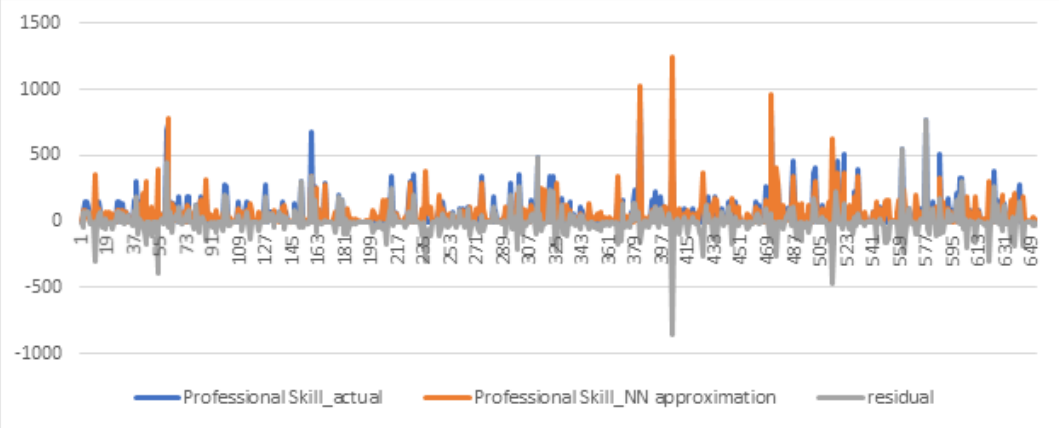

Figure 5.6: Actual and predicted values with ANNs model for test data.

In addition to these reflections, Figure 5.7 and Figure 5.8 summarize our findings for the 3 methods. Especially, the figure related to the test set demonstrates the superiority of our MARS model, when we compare it with LR and ANNs.

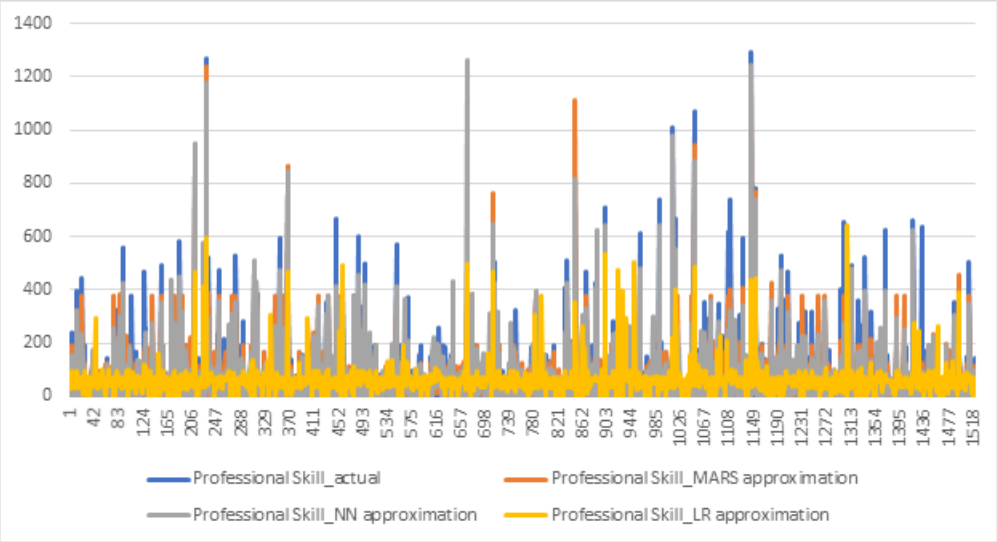

Figure 5.7: Actual and predicted values with LR, MARS, ANNs models for training data.

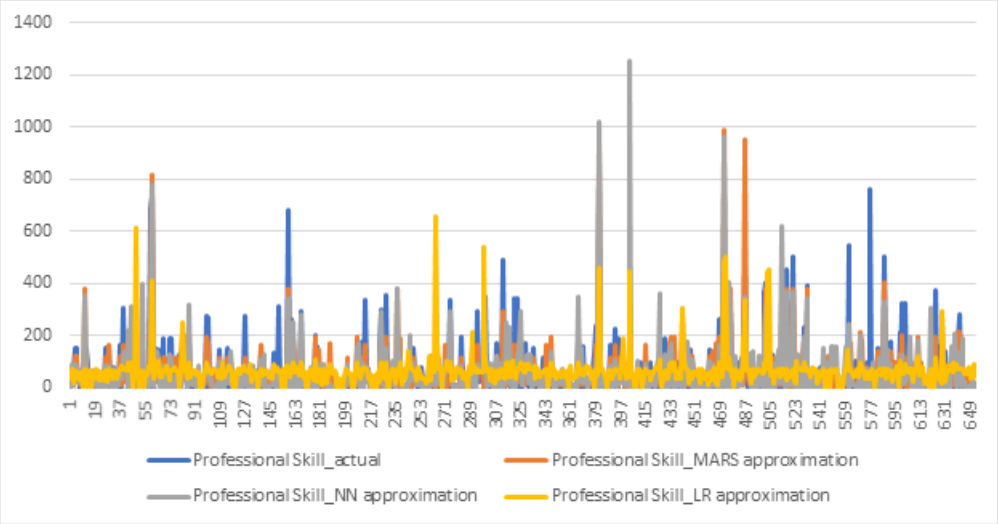

Figure 5.8: Actual and predicted values with LR, MARS, ANNs models for test data.

CHAPTER 6

CONCLUSION AND OUTLOOK

6.1 Conclusion

In this study, it is aimed to discuss, model and analyze challenges in human-resource management modeling problems in labor market stemming from education systems by applying Linear Regression, Artificial Neutral Networks, and Multivariate Adaptive Regression Splines. The dataset used in the thesis includes information in web page called "Zawodowcy" during the years 2012-2015. Before using the dataset for regression, the response variable had to be decided about, response data had to be calculated, and some certain modifications have been made to transform the inputs including qualitative information into quantitative variables. Here we also used the name of a Forward Problem. After that, three models have been built: while LR is used to introduce the main idea of regression, MARS model is provided as our fundamental model which we particularly compare with our ANNs. Indeed, when we look at these models, we can see that there are three common input variables (X_8: Job offer's type of employment: X_{14} (or X_O): Student's total evaluation value of common skills for particular offer and X_{15} (or X_W): Student's total evaluation value of general skills for particular offer) between LR model and MARS model. This shows that our models have a good and reasonable relation between each other, even though LR model is not further evaluated, neither as a main model nor as a comparing model.

When we view our graphs generated, at some figures and values from comparison criteria for LR, MARS and ANNs models, the results show a higher reliability, more accurate and stronger relationship between variables for our MARS model. In LR

model, 5 variables are presented. Therefore, our LR model is quite simple when we compare it with MARS model. In our MARS model, several BFs are in correlation and have some multi-linearity among them. It leads to various interactions between input variables obtained from different BFs. That is why, our MARS model is rather sophisticated and has an additive-multiplicative structure; it allows for analytical investigations, on mathematical subjects ranging from structure, dependences and sensitivity, to simulation. Because of all these reasons, we can say that MARS model is the most interesting and important model for us. ANNs model has some performance values close to the ones of MARS model. This shows that our ANNs approach, chosen also for comparison, is an applicable one, and our MARS model is good enough in order to make valuable comments on the results found. In contrast, ANNs methodology provides more of a computational system or input-output automation and it is, to some high degree, a "black box" approach.

Furthermore, for our triplet model (LR, MARS and ANNs), the results coming from performance criteria for training and test dataset are always supporting each other. This point is another privilege of our work. A further finding point is that the stability values of the performance criteria for MARS model show numbers which are the ones closest to 1. Hence, MARS model is our best approach with respect to statistical stability criteria.

6.2 Outlook to Future Studies

For our future projects, we have plans to collect new datasets, reflecting more recently information, in order to implement the General Model (cf. Chapter 2) (as "forward problem") as well, to use again MARS and ANNs methods, but also further data mining, deep learning, machine learning and AI (Artificial Intelligence) methods, such as RMARS (Robust MARS), CMARS (Conic MARS), RCMARS, GPLMs (Generalized Partial Linear Models), CGPLMs and RCGPLMs. When we apply and compare them for understanding and presenting specific skills (Z) and general skills (W), but also common skills (O) with the help of statistical performance criteria, several computational systems, statistical graphs, all kinds of models, and Sensitivity Analyses, as we shall indicate subsequently.

Whenever we hear or use the word *Sensitivity* in daily life about any object, person or system, it can have both a more positive and a more negative meaning. In fact, on the one hand, sensitivity stands for the flexibility of responding well and sufficiently to stimuli or other changes within a system or from the environment. Herewith, sensitivity stands for the readiness to adapt, to adjust and to survive, whereas a too high "stability" (insensitivity) or entropy could imply to become left behind, to lose competitiveness, and even to disappear or "die". On the other hand, *Insensitivity* (stability) can be another word for the needed persistence, the conservative element, easiness, lack of complexity, but structure and order which life in all its forms needs, tradition, comfort or easiness. This can be accomplished by saving energy, time and cost of any form. Herewith, insensitivity of stability can also mean reliability, survival, identity and protection. In the following, we shall give a clue on Sensitivity to the reader if he/she is interested to employ it as an analytical instrument, or if he/she is a practitioner or decision maker in order to evaluate the MARS model found, according to his/her particular questions and interests, and to draw by it managerial or educational conclusions.

In fact, MARS provides, in every input variable which becomes involved into it, a piecewise linear model; herewith, by an easy case study (comparison of 2 values for maximum) it can be reduced to a linear model of a regarded input variable. Our project partners at Poznan University of Technology (Poland) have a particular interest in the "impact" or "importance" of the different predictors on the response variable. This impact (or importance) can be an *absolute* one, regarding the sizes of the values or the "orders" of the factors (sensitivities); or it can be a *relative* one, comparing the relative impacts of the different input variables within the context of the entire additive-multiplicative MARS model.

Another plan is to perform similar implementations and scientific investigations accordingly with not fully academic labor markets, e.g., migrant communities in various countries, e.g., in Turkey [5]. Through all of these we shall serve to strengthen and refresh both the PUT project on Technical Knowledge Acceleration Program [17] and the variety of modern and "charismatic" scientific mathematics-supported subjects at METU and PUT as Centers of Excellence. Future projects will be located in areas of (i) Engineering Management, especially related with Quality Management

and Marketing, (ii) Applied Mathematics and Statistics, (iii) Education, (iv) Actuarial Sciences, especially related to insurance companies and pension fund systems, and (v) Cognitive Science and Neuroscience, too. In fact, we do use the notion of ***Human Resource Management*** in order to represent a big part of the content and final aim of both this thesis and our intended research. As also Informatics is involved in the form of the development of a Graphical User Interface (GUI) and as the "Human Factor" is so richly represented in our investigation. In this work, we can also see "Artistic Problems" addressed through these scientific efforts. Our future research intention also includes the preparation of a ***Graphical User Interface*** (in short: ***GUI***) to conveniently visualize, see and understand the inner dependences, effects and impacts of our MARS models, both on microscopic level and macroscopic level. Then, we would have at hand a very helpful scientific product, a tool of greatest value for all the aforementioned groups interested and, in fact, for our whole societies concerned with our modern industries and advanced education, in Poland, Turkey and all over the world.

Another point for further studies, a ***Simulation*** analysis, an imitation of a real-world process or system over time, can be considered, and conducted with the help of our MARS model f. Here, we aim at different levels, eventually: at the maximal level, of the different levels of contribution to emerging working places and industries, based on different realizations of skills, as illustrated in Figure 6.1. For several (here: 3) fictive datasets, skills are arranged according to different policies in the educational policy, at technical high school level or at the governmental level. One of these paradigmatic policies could be strengthening of mathematical and computational skills through university or school education.

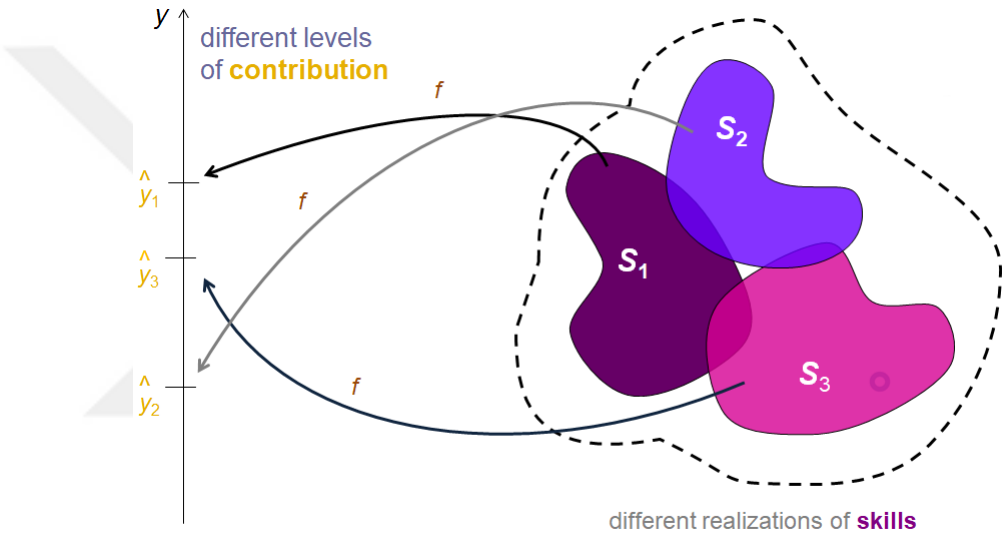

Figure 6.1: Simulation with the help of MARS model f, scheme based on 3 different educational policies [17].

REFERENCES

[1] Computer hope, https://www.computerhope.com/jargon/g/gui.htm, accessed: 2019-03-08.

[2] Czas zawodowcow, https://system.zawodowcy.org/, accessed: 2018-12-20.

[3] Human resource systems group, https://www.hrsg.ca/behavioral-competencies, accessed: 2019-01-05.

[4] Human resource systems group, https://www.hrsg.ca/technical-competencies, accessed: 2019-01-05.

[5] Liberated social entrepreneur using business metrics: Migport refugee big data analytics, http://www.ifors.org/wp-content/uploads/2019/02/Willi-Berat-Refugees-and-Labor-Market.pdf, accessed: 2019-08-05.

[6] Salford predictive modeller, https://www.salford-systems.com/products/spm, accessed: 2019-02-15.

[7] Technical knowledge accelerator programme, http://www.awt.org.pl/t-iii5/?lang=ens, accessed: 2018-12-11.

[8] *Monitoring Demand for Professional Skills in SMEs of the Wielkopolska Region*, 2015.

[9] *Competency Management as the Direction of the Development of Enterprises—Based on Research*, 2016.

[10] *From the Research on Social Competencies of Future Managers*, 2017.

[11] *Models of Businesses' Support for Technical Knowledge Development in Wielkopolska Region - a Qualitology Approach*, 2017.

[12] R. Aster, B. Borchers, and C. Thurber, *Parameter Estimation and Inverse Problems*, Academic Press, 2003.

[13] M. Bernard, What are artificial neural networks - a simple explanation for absolutely anyone, https://www.forbes.com/sites/bernardmarr/2018/09/24/what-are-artificial-neural-networks-a-simple-explanation-for-absolutely-anyone/#6a6617191245, accessed: 2019-06-10.

[14] A. Branowska, P. Siemieniak, and M. Spychala, *Workers' occupational competencies in a modern enterprise*, Publishing House of Poznan University of Technology, 2011.

[15] M. Graczyk-Kucharska, M. Szafrański, M. Goliński, and M. Spychala, Competencies of the future as an impulse for innovation in the management of smart organizations, in *Proceedings of the 6th Central European Conference in Regional Science – CERS, 2017*, 2017.

[16] M. Graczyk-Kucharska, M. Szafrański, M. Goliński, M. Spychala, and K. Borsekova, *Model of Competency Management in the Network of Production Enterprises in Industry 4.0-Assumptions*, 2018.

[17] M. Graczyk-Kucharska, M. Szafrański, S. Gütmen, A. Çevik, G. W. Weber, Z. Wĺodarczyk, M. Goliński, and A. Özmen, Modeling problems in a regional labor market in Poland with MARS, https://www.researchgate.net/publication/334249870_Modeling_Problems_in_a_Regional_Labor_Market_in_Poland_with_MARS, accessed: 2019-06-23.

[18] M. Graczyk-Kucharska, A. Özmen, M. Szafrański, G. W. Weber, M. Goliński, and M. Spychała, Knowledge accelerator by transversal competences and multivariate adaptive regression splines, pp. 1–25, 2019.

[19] V. Gudivada, *Cognitive Computing: Theory and Applications*, North Holland, 2016.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2008.

[21] R. J. Hydman and A. B. Koehler, Another look at measures forecast accuracy, International Journal of Forecasting, vol. 22, pp. 679–688, 2006.

[22] E. Köksal, *Modeling of Exchange Rates by Multivariate Adaptive Regression Splines and Comparison with Classical Statistical Methods*, Master's thesis, Institute of Applied Mathematics, Middle East Technical University, 2017.

[23] S. Kuter, Z. Akyürek, and G. W. Weber, Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines, Remote Sensing of Environment, vol. 205, 2018.

[24] R. Lippmann, An introduction to computing with neural nets, IEEE ASSP Magazine, vol. 4, pp. 4–22, 1987.

[25] J. Mass and J. Flores, The application of artificial neural networks to the analysis of remotely sensed data, International Journal of RemoteSensing, vol. 29, pp. 617–663, 2008.

[26] V. Sharma, S. Rai, and A. Dev, A comprehensive study of artificial neutral networks, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, pp. 278–284, 2012.

[27] M. Szafrański and M. Goliński, System for professionals-monitoring employers' demands for key competences in wielkopolska, in *Recent Advances in Computer Science; Proceedings of the 19th International Conference on Computers*, pp. 184–191, 2015.

[28] P. Taylan, F. Özkurt Yerlikaya, and G. W. Weber, Precipitation modeling by polyhedral RCMARS and comparison with MARS and CMARS, Environmental Modeling & Assessment, vol. 19, pp. 425–435, 2014.

[29] G. W. Weber, İ. Batmaz, G. Köksal, P. Taylan, and F. Özkurt Yerlikaya, CMARS: A new contribution to non - parametric regression with multivariate adaptive regression splines supported by continuous optimization, Inverse Problems in Science and Engineering, vol. 20, pp. 371–400.

[30] A. Özmen, *Robust Conic Quadratic Programming Applied to Quality Improvement - A Robustificaiton of CMARS*, Master's thesis, Institute of Applied Mathematics, Middle East Technical University, 2010.

[31] A. Özmen, İ. Batmaz, and G. W. Weber, An approach to the mean shift outlier model by tikhonov regularization and conic programming, Intelligent Data Analysis - Business Analytics and Intelligent Optimization, vol. 18, pp. 79–94, 2014.

[32] A. Özmen and G. W. Weber, RMARS robustification of multivariate adaptive regression spline under polyhedral uncertainty, Journal of Computational and Applied Mathematics, vol. 259, 2014.

[33] A. Özmen, G. W. Weber, and G. Nalcacı, Long-term load forecasting: Models based on MARS, ANN and LR methods, Central European Journal of Operations Research, vol. 24, pp. 1033–1049, 2018.

[34] A. Özmen, Y. Yılmaz, and G. W. Weber, Natural gas consumption forecast with MARS and CMARS models for residential users, Energy Economics, vol. 70, pp. 357–381, 2018.