



Predictive Modeling: Data Mining Regression Technique Applied in a Prototype

¹Festim Halili, ²Avni Rustemi

^{1,2}Department of Informatics State University of Tetovo, SUT Tetovo, Macedonia

¹ festim.halili@unite.edu.mk, ² avnirustemi11@gmail.com

Abstract - Increasingly with the rapid development of technology also there are various sophisticated software which enable us to solve problems in various spheres of our life. Today, thanks to various software and methods that are discovered a long time ago, we manage to solve various problems and dangers that may threaten us, so today thanks to these software, different methods we can predict any danger that our work may be threatened in the future. With the introduction of sophisticated software, there is a need also for new terms where it will be stored these data because we know that software cannot function if it does not have the most significant part and that is database. Firstly, we will introduce some terms that are related with our topic such as, some brief description of Big Data, Data Warehouses, Data Mining and their classification and finally also will do the particular analysis for Regression technique (linear and multiple regression) and our approach regarding a prototype and how can be used and for what reasons regression techniques giving explanations with concrete examples. Regression as technique although is predictive technique, but based on analyzes conducted to reach the conclusion most scientists, they have concluded that the reliability percentage is around 95%. Through our paper we will try to demonstrate this scale of reliability through concrete examples.

Keywords: Predictive modeling, Data warehouses, Simple and Multiple Regression, Data Mining.

I. INTRODUCTION

Predictive modeling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or “dependent” variable and various predictor or “independent” variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable. Because these relationships are never perfect in practice, it is desirable to give some measure of uncertainty for the predictions, typically a prediction interval that has some assigned level of confidence like 95%.

Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors. Regression, as a data mining technique, is supervised learning. Supervised learning partitions the database into training and validation data. The techniques used in this research were simple linear regression and multiple linear regression. Some distinctions between the use of regression in statistics versus data mining are: in statistics The data is a sample from a population, but in Data Mining The data is taken from a large database (e.g. 1 million records). Also in statistics The regression model is constructed from a sample, but in Data Mining the regression model is constructed from a portion of the data (training data). Predictive analytics encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events. The variety of techniques is usually divided in three categories: predictive models, descriptive models and decision models.

Predictive models look for certain relationships and patterns that usually lead to a certain behavior, point to fraud, predict system failures, assess credit worthiness, and so forth. By determining the explanatory variables, you can predict outcomes in the dependent variables.

Descriptive models aim at creating segmentations, most often used to classify customers based on for instance socio-demographic characteristics, life cycle, profitability, product preferences and so forth. Where predictive models focus on a specific event or behavior, descriptive models identify as many different relationships as possible. Lastly, there are decision models that use optimization techniques to predict results of decisions. This branch of predictive analytics leans particularly heavily on operations research, including areas such as resource optimization, route planning and so forth.

II. DATA MINING

It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. To refer to the mining of gold from rocks or sand, we say gold mining instead of rock or sand mining. Analogously, data mining should have been more appropriately named “knowledge mining from data,” which is unfortunately somewhat long. However, the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material (Figure 1.3). Thus, such a misnomer carrying both “data” and “mining” became a popular choice. In addition, many other terms have a similar meaning to data mining—for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is shown in Figure 1 as an iterative sequence of the following steps:

- Data cleaning (to remove noise and inconsistent data)
- Data integration (where multiple data sources may be combined)
- Data selection (where data relevant to the analysis task are retrieved from the database)
- Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)⁴
- Data mining (an essential process where intelligent methods are applied to extract data patterns)
- Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
- Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users).

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data). Therefore, we adopt a broad view of data mining functionality: Data mining is the process of discovering

interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data/pattern analysis.

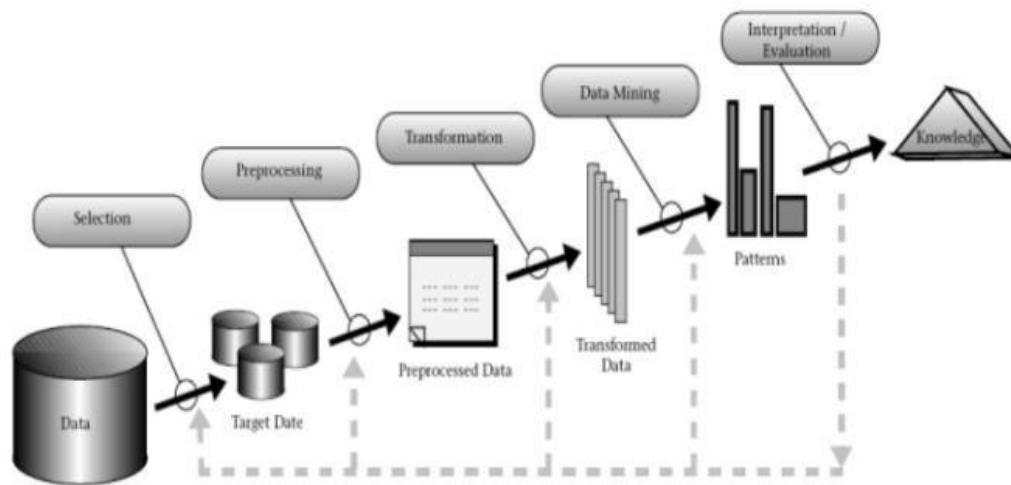


Figure 1. Process of knowledge discovery process

A. DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

A.1. CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

A.2. CLUSTERING

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

A.3. PREDICATION

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models. Types of regression methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

A.4. ASSOCIATION RULE

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. Types of association rule:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

A.5. NEURAL NETWORKS

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Types of neural networks: Back Propagation.

III. DATA WAREHOUSES

Suppose that A-Electronics is a successful international company with branches around the world. Each branch has its own set of databases. The president of A-Electronics has asked you to provide an analysis of the company’s sales per item type per branch for the third quarter. This is a difficult task, particularly since the relevant data are spread out over several databases physically located at numerous sites. If A-Electronics had a data warehouse, this task would be easy. A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. Figure 2 shows the typical framework for construction and use of a data warehouse for A-Electronics. To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity). The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region. A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount).

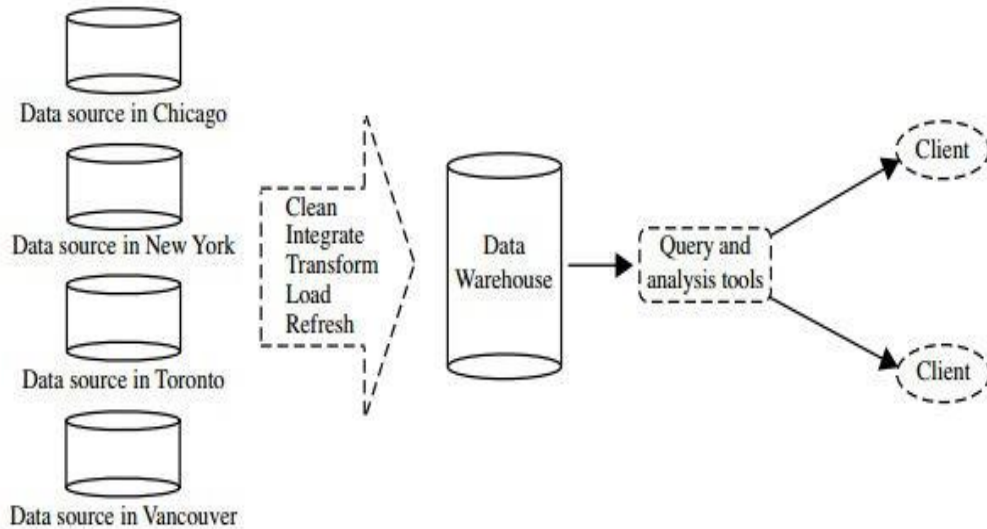


Figure 2. Typical framework of a data warehouse for A-Electronics

A data cube provides a multidimensional view of data and allows the pre-computation and fast access of summarized data. By providing multidimensional data views and the pre-computation of summarized data, data warehouse systems can provide inherent support for OLAP. Online analytical processing operations make use of background knowledge regarding the domain of the data being studied to allow the presentation of data at different levels of abstraction. Such operations accommodate different user viewpoints. Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization. For instance, we can drill down on sales data summarized by quarter to see data summarized by month. Similarly, we can roll up on sales data summarized by city to view data summarized by country. Although data warehouse tools help support data analysis, additional tools for data mining are often needed for in-depth analysis. Multidimensional data mining (also called exploratory multidimensional data mining) performs data mining in multidimensional space in an OLAP style. That is, it allows the exploration of multiple combinations of dimensions at varying levels of granularity in data mining, and thus has greater potential for discovering interesting patterns representing knowledge.

IV. LINEAR AND MULTIPLE REGRESSION AND OUR APPROACH

In statistics prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction. The simplest form of regression is simple linear regression that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model). Graphically this would look as it does in Figure 1.3. The simplest form of regression seeks to build a predictive model that is a line that maps between each predictor value to a prediction value. Of the many possible lines that could be drawn through the data the one that minimizes the distance between the line and the data points is the one that is chosen for the predictive model. On average if you guess the value on the line it should represent an acceptable compromise amongst all the data at that point giving conflicting answers. Likewise if there is no data available for a particular input value the line will provide the best guess at a reasonable answer based on similar data.

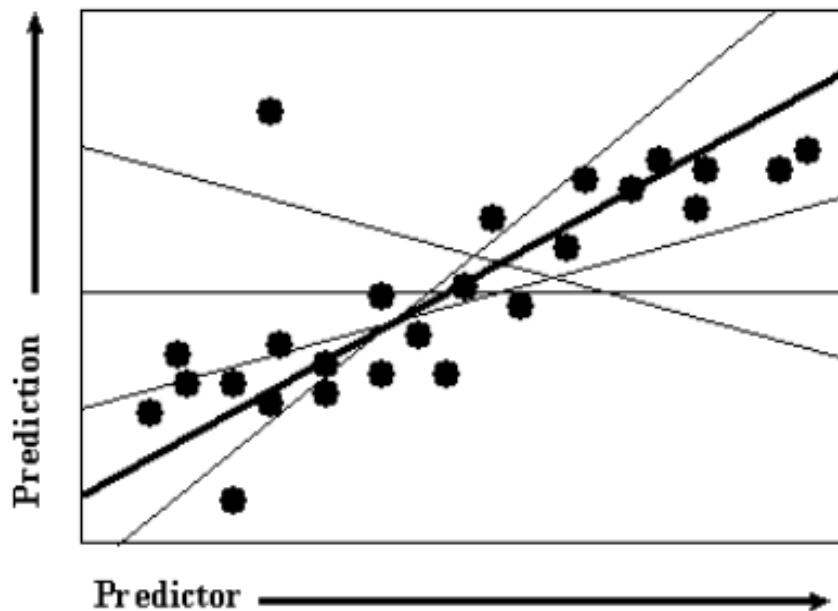


Figure 3. Linear regression is similar to the task of finding the line that minimizes the total distance to a set of data.

The predictive model is the line shown in Figure 3. The line will take a given value for a predictor and map it into a given value for a prediction. The actual equation would look something like: $\text{Prediction} = a + b * \text{Predictor}$. Which is just the equation for a line $Y = a + bX$. As an example for a bank the predicted average consumer bank balance might equal $\$1,000 + 0.01 * \text{customer's annual income}$. The trick, as always with predictive modeling, is to find the model that best minimizes the error. The most common way to calculate the error is the square of the difference between the predicted value and the actual value. Calculated this way points that are very far from the line will have a great effect on moving the choice of line towards themselves in order to reduce the error. The values of a and b in the regression equation that minimize this error can be calculated directly from the data relatively quickly.

A statistical tool that allows you to examine how multiple independent variables are related to a dependent variable. Once you have identified how these multiple variables relate to your dependent variable, you can take information about all of the independent variables and use it to make much more powerful and accurate predictions about why things are the way they are. This latter process is called “Multiple Regression”.

In this paper we will present to you ,some analysis made for a specific prototype, respectively analysis of profit in a particular shop where initially we have provided information regarding the number of sales of certain products (in this case we receive only 4 types of products), and then doing the regression analysis and when we have regression table with specific information we can much easier predict how much profit we will have in specific day with certain number of sales of given products. Below we have given table of products and profits from the analysis made in a designated enterprise over a week and the data are as follows:

Table1. Tables with data with the number of sold products and profit

Day	Profit	Prod. 1	Prod. 2	Prod. 3	Prod. 4
1	\$ 7,378.40	356	432	356	456
2	\$ 7,284.00	324	456	324	456
3	\$ 6,395.80	432	356	326	344
4	\$ 3,070.70	563	106	108	108
5	\$ 7,280.00	500	400	300	450
6	\$ 7,493.60	356	450	360	456
7	\$ 6,378.00	308	308	456	338

From the table above we see that analyzes are made with one week (seven day) , and there are given only 4 different products, and the profit from the same products . Almost we already have the data table, with information of given products, now we can find regression of these product in comparison with profit. Why we need to find a regression technique? Regression technique is predictive technique, but with high level of reliability (95 %) , so if we would like to know one day how will be profit from a certain number of sold products must use this method . To find the regression method is not easy, because we need to make various mathematical calculations, but such a thing thanks to the different software can find much easier so we will not lose time in finding different relations to achieve regression but will give tables from found multiple regression and thereafter shall give the relation of multiple regression and how we can use it in practice for finding in this case a profit from sold different products. Regression informations in form of table found by the data we have in Table 1, are:

Table 2 . Regression statistics

<i>Regression Statistics</i>		ANOVA					
Multi. R	1		<i>d</i> <i>f</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sign. F</i>
R Sq.	1	Reg.	4	14 76 08 41	369 021 0.19	2.1 04 75 E+ 30	4.751 15E- 31
Adj. R Square	1	Res.	2	3.5 1E- 24	1.75 33E- 24		
Std. Er	1.32 E- 12	Tot.	6	14 76 08 41			
Obser.	7						

	<i>Coefficients</i>		<i>Standard Error</i>		<i>t Stat</i>
Intercept	2.73E-12		9.37E-12		0.2911726
Product 1	2.5		1.37E-14		1.8278E+14
Product 2	5.4		3.29E-14		1.6402E+14
Product 3	4.5		1.28E-14		3.5036E+14
Product 4	5.6		3.18E-14		1.7584E+14
<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>	
0.79834	- 3.8E- 11	4.3E- 11	-3.8E-11	4.3E-11	
2.99E-29	2.5	2.5	2.5	2.5	
3.72E-29	5.4	5.4	5.4	5.4	
8.15E-30	4.5	4.5	4.5	4.5	
3.23E-29	5.6	5.6	5.6	5.6	

From table 2 already we have all information that we need to calculate regression to predict a certain number of products sold in a particular period but before we give a concrete example we must initially give the overall equation for finding multiple regression. The equation that equation which enables us regression calculation, respectively, respectively $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \epsilon$, where ϵ , the “noise” variable, is a Normally distributed random variable with mean equal to zero and standard deviation σ whose value we do not know.

We also do not know the values of the coefficients $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.

We estimate all these $(p + 2)$ unknown values from the available data. The data consist of n rows of observations also called cases, which give us values $y_i, x_{i1}, x_{i2}, \dots, x_{ip}; i = 1, 2, \dots, n$. The estimates for the β coefficients are computed so as to minimize the sum of squares of differences between the fitted (predicted) values at the observed values in the data.

Now the question is why we need all these calculations. All these calculations are related to each other i.e. all calculation are found sequentially But thanks to the software we are not entered at all in mathematical calculations but will use them ready.

So if the we want to know how much will be profit from our products ,i.e. if we want to know how much will be profit if we sell for example 500 of product A , 450 of product B , 356 of product C and 452 of product D . So with the equation above these numbers are x values, so $x_1=500$, $x_2=450$, $x_3=356$ and $x_4=452$. And coefficient β_0 , β_1 , β_2 , β_3 and β_4 we take from table 2. So the number $\beta_0= 2.73E-12$, $\beta_1=2.5$, $\beta_2=5.4$, $\beta_3=4.5$ and $\beta_4=5.6$. So after already we have all the necessary variables we can calculate profitability based on the formula and we have : $Y = 2.73E-12 + 2.5 * 500 + 5.4 * 450 + 356*4.5 + 452*5.6 = 7813.2$. So from the result we can conclude that , if one day we sell so products as we describe , the profit will be 7,813.20 \$ where if we make a more precise analysis we will conclude that these result even they are prediction but are 99% sure. So we can say that regression is one of the predictive methods that enable to predict a result under some certain parameters but with a high degree of reliability (around 95% are credible results) .

V. CONCLUSION

Although there are not just one or two types of regression but we have more types, we have illustrated only two of them which are associated with the analysis of our prototype. We tried to present briefly and clearly before you regression analysis with aim how to use and for what purpose to use regression techniques in the future. To be successful in various businesses we should do a lot of analyses to be sure that our business will go properly in the future. Among these analyses and techniques is also regression, through which we manage to predict some phenomenon but based on some other phenomenon. This means that for the regression techniques we must have independent variables to find the dependent variables. This is related with that of example which we illustrate before you, where to find how can be profit one day, we must have the number of sold products. Our next job is to make other analysis related with predictive models, where will provide analysis and concrete examples of how these types of models used in practical forecasting and what benefit we have of them.

REFERENCES

- [1] John P. Hoffmann, "Linear Regression Analysis: Applications and Assumptions Second Edition", 2010, USA.
- [2] Alan O. Sykes, " An Introduction to Regression Analysis" The Inaugural Coase Lecture.
- [3] P. Bastos, I. Lopes , L. Pires : "Application of data mining in a maintenance system for failure prediction ", 2014 Taylor & Francis Group, London. [4] David A. Dickey, N. Carolina State U., Raleigh, NC, "Introduction to Predictive Modeling with Examples", 2012.
- [5] John O. Rawlings Sastry G. Pantula David A. Dickey, "Applied Regression Analysis: A Research Tool, Second Edition", 1998.
- [6] Britney Robinson and Joi Officer Advisor: Dr. Fred Bowers: " Data Mining: Predicting Laptop Retail Price Using Regression".
- [7] Data Mining for Business Intelligence, Galit Shmueli, John Wiley, Nitin Patel, and Peter Bruce, 2007.
- [8] Berk, Richard A. "Statistical Learning from a Regression Perspective", Springer Series in Statistics. New York: Springer-Verlag. (2008).
- [9] Wilhelmiina H'am'ala'inen, "Descriptive and Predictive Modelling Techniques for Educational Technology", Licentiate thesis August 10, 2006 Department of Computer Science University of Joensuu.
- [10] Dan Campbell, Sherlock Campbell, "Introduction to Regression and Data Analysis", October 28, 2008.