Regression Models for Lean Production

Ricardo Bragança, Filipe Portela, Manuel Santos

Algoritmi Research Centre, University of Minho, Guimarães, Portugal
A51055@alunos.uminho.pt, cfp@dsi.uminho.pt, mfs@dsi.uminho.pt

Abstract. Data mining models are an excellent tool to help companies that live from the sales of items they produce because it allows the company to optimize its production and reduce costs, for example in storage. When these models are combined with Lean Production, it becomes easier to remove waste and optimize industrial production. This project is based on the phases of the methodology CRISP-DM, and aims to reduce and, if possible, eliminate wastage. The following methods: average, mean and standard deviation, quartiles and Sturges rule regression, were techniques applied to this data to determine which one is the model is less likely to make mistakes, in other words, meaning that the model did correctly predict the target. Most common metrics used at the statistical level, which had already been proven to have good results in similar studies. After performing the tests, the M4 model is what is less likely to make mistakes in terms of regression with a RAE of 21,33%.

Keywords: Data Mining, Regression, CRISP-DM, DSR, Lean Production.

1 Introduction

Companies in industry are increasingly feeling the need to find a way to optimize their production to counteract the adversities of the economic world. One of the best ways to do this is using the data mining models that allow, based on past sales, to get an estimate of how much selling's will occur in the future, for an efficient production and reducing the waste of raw material and labor work. The rating models are a great tool to help businesses achieved success. That said, here were tested several models at the level of regression hoping to find models with Relative Absolute Error (RAE) close to 0%, which means that the model is less likely to fail in target prediction.

This project stems from organizations need to improve its production process. There are several aspects that need special attention, firstly being the necessity to reduce or eliminate the stock completely, as it is equivalent to having money as well as products tied up; Secondly it is necessary to reduce the inventory because in addition to the expenses inherited to the stock of products, there is a product that is not transacted, so it does not bring any kind of profit to the organization; Thirdly, there is a misallocated production in the company for the orders received; and finally, there is a need to reduce or eliminate any type of waste existing in the production process. In this work several common metrics have been used to apply data such as average, mean and standard deviation, quartiles and Sturges rule regression, secondly [1], have proved to obtain good results. This work intends to elaborate forecasting and

decision models in order to optimize the production of raw materials common to several products, as well as intermediate products and to reduce, or if possible, eliminate the waste of raw materials.

This paper started with an overview of the concepts inherent in the project including data mining, Lean Production and Decision Support, and is ending with what exists today. Subsequently the methodologies, CRISP-DM, DSR, and, at the level of the tools, WEKA, are presented. Following up, the phases of the entire project were presented posteriorly to the phases of the CRISP-DM. Finally, a little discussion about the project was created as well as a conclusion containing a presentation of closing arguments.

2 Background

2.1 Data Mining

In the world there is a vast amount of data stored, but these are only examined in a very superficial way, which leads to having a wealth of data and great poverty of knowledge [2]. In the last few decades, data mining has been widely recognized as a powerful and versatile tool of data analysis and has made a significant contribution in the areas of information technology in clinical medicine, sociology, physics, and in the areas of management, economics and finance [3, 4]. Data mining (DM) is the task that seeks to discover patterns in data sets by using methods of artificial intelligence, machine learning, statistics and database systems [5, 6]. Wu [7] suggests a more complete definition for DM; he believes it is the integration of various subjects: (i) databases, (ii) databases technologies, (iii) statistic, (iv) machine learning, (v) math, (vi) neural networks and others. The DM was designed to use two techniques, one predictive and the other descriptive. The predictive, also known as supervised learning, is characterized by obtained examples and input data, predicting future values. In turn, descriptive, also known as unsupervised learning, is characterized by learning through data grouping with identical characteristics [8, 9].

2.2 Lean Production

LP originated from the Toyota Company, and started in the end of World War II with the implementation of the Toyota Production System (TPS) and aimed to increase productivity and reduce the cost of car production by eliminating any waste. The fact that this system allowed for Toyota leadership in the automotive market has led researchers at the Massachusetts Institute of Technology (MIT) dubs it Lean Production [10][10][10] [9] [18]. The LP is an organizational model that brings numerous benefits to the organization that implements it by reducing costs and eliminating waste [11, 12]. There is a great difficulty to find the best setting for LP as the existing definitions and their focus differs from author to author, as well as the practices involved [13]. The Lean term comprises the establishment of a culture of continual improvement and organizational learning [14]. At the operational level, it reduces unnecessary production time, materials and manufacturing efforts, thus

reducing the cost by reducing waste [15]. There are obstacles that must be taken into account when seeking the implementation of this methodology, namely the human factor often brings the rejection of change by employees and that can cast doubt to the implementation of Lean [16]. The implementation of Lean is a lengthy process, thus requiring a well-defined strategy and must involve top management in the process so that there is real knowledge of the organization [10]

2.3 Decision Support

The Decision Support refers to applications involving broad analyses and exploration of current and historical data in organizations supporting the high-level Decision-making. The decision Support requires the integration of two types of management: **Data management**, which includes the organization's databases that are managed by database management systems; **Knowledge management**, which handles tasks related to reasoning. Due to various limitations existing in the processing of human Knowledge it is necessary to use other support [17]. In the Simon's decision-making process there are three main phases [18]: Intelligence, Design, and Choice.

2.4 Present

Ming-Te, Kuo-Chung, & Pan [19], suggest the use of DM, notably Genetic Programming, Artificial Neural Networks and Logistic Regression to improve the LP configuration for improved performance. Groger, Nidermann, & Mitschang [20] presented a paper describing the optimization of manufacturing as a targeted approach to the DM. Based on a huge amount of data, previously defines the use cases of DM that will be applied to identify hidden data patterns, hence optimizing the entire manufacturing process. By using Dashboards and Key Performance Indicators (KPI) it is possible to analyze in depth the reasons for deviations and presented indication in order to improve processes. Unver [21] indicates that there is an opportunity expressed in manufacturing industries for a concept denominated Intelligent manufacturing. In this concept, it is used all the advantages of Lean: cost control; quality improvement and product and waste disposal, combined with business intelligence more precisely the dashboards presenting the processed data and KPI, which allow managers to determine the most appropriate options for applying Lean. Vazan, Tanuska, & Kebisek [22] indicates that there is an opportunity expressed in manufacturing industries for a concept denominated Intelligent manufacturing. In this concept, it is used all the advantages of Lean: cost control; quality improvement and product and waste disposal, combined with business intelligence more precisely the dashboards presenting the processed data and KPI, which allow managers to determine the most appropriate options for applying Lean.

3 Methods and Tools

Throughout the project, we used two methodological approaches: Design Science Research (DSR) and Cross Industry Standard Process for Data Mining (CRISP-DM). As the first methodology is in the field of Information Systems (IS), it will be used in conducting the research process. Given that the second one refers to a project linked to data mining it will be used to carry out the project. In the last years, the DSR regained importance in the paradigm of research specifically in the field of IS [23, 24]. In the IS area, DSR is used in the construction of artifacts in a countless number of areas from systems for decision support, Modeling tools, assessment methods and among others [25]. DSR is a set of techniques and synthetic and analytical perspectives to conduct research processes.

The main objective of this approach is the creation of knowledge through the design of new and innovative artifacts and analysis of their use and performance along with reflection and abstraction [26]. DSR consists of six phases: (i) Identification of the problem and motivation, which defines the specific research problem and where the value of the solution is justified; (ii) Definition of the solution objectives and identification of goals for the problem definition and knowledge of what is possible; (iii) Development and Design of the solution, at this stage artifacts for the solution are created; (iv) Demonstration, is intended at this stage to demonstrate the efficacy of articles developed in the previous stage to solve the problem; (v) Evaluation, at this stage is intended to observe and measure how the artifact can actually support the solution of the problem; (vi) Communication, at this stage communication is made to the researchers and/ or professionals interested, what is the found problem, its importance, which artifacts have been developed, as well as its usefulness and the fact of whether it is a novelty or not [27-29].

One of the main methods of data mining is the CRISP-DM and was developed by a consortium comprised of NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc. and OHRA Verzekering en Bank Groep B.V. [30, 31]. The main objective of this methodology is to define a complete data mining process from Understanding the data for the implementation of developed models through the monitoring of improvements. The main objective of this methodology consists in defining a complete data mining process from Understanding the data to the implementation of developed models through the monitoring of improvements [32]. This approach consists of six main phases: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation and (6) Implementation [2, 33]. The WEKA was the tool chosen to implement this project because despite being widely used in data mining projects is also open source.

4 Industrial Planning

As previously stated the entire project was developed with the CRISP-DM methodology and the following paragraphs will describe the stages of this methodology.

4.1 Business Understanding and Data Understanding

This project is part of the need to eliminate all types of waste inherent in the process of a producer organization of textile labels and their role is to design, produce, manipulate and encode all types of labels for the leading companies of confection and sale of textile products and supplements worldwide. For this purpose, were developed data mining models that will be presented in subsequent points.

The data used to carry out this project concern the quantity sold of a product from 09-11-2012 to 19-02-2016 and contain a total of 4299 records. Table 1 **Error! Reference source not found.** it is found a description of the attributes that initially existed before proceeding to any kind of change. These data are very important because sales can be predicted and after this forecast it is possible to predict which quantities should be produced and this allows to reduce costs in production, storage costs, because only what is sold is produced. After performing a data analysis, it was denoted the existence of periods with no data because there was no data from the beginning to the end of the calendar year analysis.

Table 1. Description of the initial attributes

Attribute	Description	Example	Maximum	Minimum	Average
Article	Article nomenclature	REFXPTA	n.a.	n.a.	n.a.
Data	Day, month and year in which particular item has been sold	2014-01-06	19-02-2016	09-11-2012	n.a.
Quantity	Quantity sold by an article	47.350.692	199.000	-31.800	7.273,498

4.3 Data Preparation

In order to obtain results as reliable as possible the interval was shortened from 2014 to 2015, thereby obtaining 52 weeks in each of the years. Given the introduction of the number of weeks, in addition to changing the attributes, they were also performed calculations in the attribute of quantity sold, particularly sum and medium, thus obtaining the classes to which the quantities belong. In Table 2 **Error! Reference source not found.** is a description of the attributes that will be used later in the models.

Table 2. Description of the attributes that have changed

Attribute	Description	Example
Article	Article nomenclature	REFXPTA
Data	Day, month and year in which particular item has been sold	2014-01-06
Quantity	Quantity sold by a particular article	47.350.692

After completing the grouping, it was necessary to resort for methods able to identify the interval of sold quantities in order to perform predictions using the regression approach. Taking into account the absence of a method able to determine these intervals, it was used: Average, quartiles, medium and standard deviation and Sturges rule. The average was calculated using the mathematical expression

$\bar{x} = \frac{1}{n} \sum_{i=1}^n xi$ where n is the number of weeks and xi is the quantity sold of given article. The intervals are $[min, \bar{x}]$ and $]\bar{x}, max]$. Another important method is the quartiles, a non-central measure of tendency that in face of a sample, is intended to determine a set of four distinct classes. In the first instance, it was necessary to sort the data set in ascending order and then identify the maximum and minimum value. To determine the first class is necessary to identify the minimum quantity sold of the REFEXPTA article sample and the value of the sold amount corresponding to the first quartile to determine this amount was necessary to use the following calculation $xn = \frac{(25\%*n)}{100\%}$. The first defined class is presented as a range between two values, $[min, xn = \frac{25\%*n}{100\%}]$, the following class is determined by the range $]xn = \frac{25\%*n}{100\%}, xn = \frac{50\%*n}{100\%}]$, the third class is defined by the range $]xn = \frac{50\%*n}{100\%}, xn = \frac{75\%*n}{100\%}]$ and finally the fourth grade $]xn = \frac{75\%*n}{100\%}, max]$. The mean and standard deviation methods require the same calculations made in the first describe method, but it needs the value of the standard deviation. Once identified, its value was determined by:

$$s = \sqrt{\frac{\sum_{i=1}^n (xi - \bar{x})^2}{n - 1}}$$

The variable xi is the quantity sold per week and n the number of weeks. From the determined values, the classes were defined. One of the intervals can be set by $]\bar{x}, \bar{x} + s]$, but it is important to note that the upper and lower values cannot exceed the max and min respectively. Finally, we determine the classes using the Sturges rule. The number of classes is determined from the expression $k = 1 + 3,322 * \log_{10} 104$ and the amplitude of each class was determined through reference to $amplitude = (max - min)/k$. The first class corresponds to the range that is determined by the min value calculation and application $h1 = min + amplitude$, The first range is obtained from $k1 = [min, h1]$. The second range will be obtained from $k2 =]h1, h1 + amplitude]$, the remaining intervals are calculated according to this process, sequentially to determine the set of class k . In Table 3 is an example of dataset that corresponds to the average data amount sold, by applying the Sturges rule.

Table 3. Example dataset

Class Number	Class	Quantity
1	[0;112487125,1]	18
2	]112487125,1;224974250,3]	19
3	]224974250,3;337461375,4]	32
4	]337461375,4;449948500,5]	17
5	]449948500,5;562435625,6]	11
6	]562435625,6;674922750,8]	5
7	]674922750,8;787409875,9]	1

4.4 Modeling

In carrying out this project, taking into account that it is intended to build predictive models as a way to predict what has occurred, we have used supervised learning techniques, more precisely regression techniques applied to the models of Table 4.

In the regression, the quantities sold of a product were used as target, applying the methods Average, Mean and Standard Deviation, Quartiles and Sturges Rule. The techniques used in the regression were Regression Trees (RT), Vector Support Machines (SVM) and Linear Regression (LR).

Table 4. Description of models

Model	Description
M1	Average application to the average quantities sold
M2	Application quartile average of quantities sold
M3	Application average and standard deviation of the average quantities sold
M4	Application Sturges rule to the average quantities sold
M5	Average application to the sum of quantities sold
M6	Application of quartiles to the sum of quantities sold
M7	Application the mean and standard deviation to the sum of quantities
M8	Application Sturges rule to the sum of quantities sold

For the implementation of a testing mechanism for the obtained models, the sampling method, 10-folds Cross Validation (10-folds CV) was selected. The techniques used in the regression were LinearRegression, SMOReg, M5Rules and AdditiveRegression. In the LR, the M5 method was used as the attribute selection method. In SVM we used the normalize training data filter type, the PolyKernel kernel and the RegSMOImproved learning algorithm.

4.5 Evaluation

The regression models are evaluated according to several metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE) and Relative Absolute Error Mean (RAEM). The best models are those in which the proportion of SAR (Relative Absolute Error) is closer to 0%. That said, the best model at the RAE level is M1 and M4 with a percentage of 13.74% and 20.34% respectively, which means that these models are less likely to be wrong. The worst models are the M5 and M2 with 60.95% and 39.81% respectively. The algorithms that obtained better classification were LinearRegression followed by SMOReg (Table 5). The average technique applied to the data of the M1 model was the one that obtained the best result, however the same technique applied to the data of the M5 model was the one that obtained the worst result. At MAE level, the models that achieved significantly lower errors were the M4 and M3 models, where the techniques used were the Sturges Rule and Mean and Standard Deviation. The results of the RAEM indicate that the model that obtained the best result was the M4 model followed by the M8 model with 21.33% and 26.66%, to both models was applied the Sturges Rule technique. The worst-rated model was the M5 model with 63.57%.

Table 5. Results of regression models

Model	Best Algorithm	RAE	MAE	MSE	RAE Average
M1	M5Rules	13,74%	26652094,52	38044083,28	33,42%
M2	SMOReg	39,81%	33727613,65	51104716,00	42,22%
M3	Linear Regression	31,83%	26249323,26	32278311,96	32,86%
M4	Linear Regression	20,34%	16971357,77	20917943,31	21,33%
M5	SMOReg	60,95%	85066734,58	108409389,59	63,57%
M6	Linear Regression; M5Rules	37,37%	51532417,16	70292270,46	38,51%
M7	SMOReg	35,66%	49329231,25	76751161,25	36,58%
M8	AdditiveRegression	26,00%	35843672,52	71375015,43	26,66%

5 Discussion

After analyzing the obtained models, it is important to understand which models are more or less susceptible to error, and the models that obtain a percentage of RAEs closer to 0%, the most suitable to be used. Through the obtained data, the algorithms of the SVM were identified, SMOReg and Linear Regression are on average those that obtain better results, with about 37%, but nevertheless, they are in average far from the values that are important to obtain in a model, that is, 10%. Only the algorithm M5Rules, which was applied to the first model, achieves highly acceptable values, very close to the 10% that was intended. This way, the best model at the RAE level is M1 and M4 with a percentage of 13.74% and 20.34% respectively, which means that these models are less likely to be wrong. The worst models are the M5 and M2 with 60.95% and 39.81% respectively. The algorithms that obtained the best classification were LinearRegression followed by SMOReg. The average technique applied to the data of the M1 model was the one that obtained the best result, however the same technique applied to the data of the M5 model was the one that obtained the worst result. At MAE level the models that achieved significantly lower errors were the M4 and M3 models, where the techniques used were the Sturges Rule and Mean and Standard Deviation. The results of the MSAR indicate that the model that obtained the best result was the M4 model followed by the M8 model with 21.33% and 26.66%, to both models was applied the Sturges Rule technique. The worst-ranked model was the M5 model with 63.57%.

6 Conclusion and Future Work

The models obtained allow the organization to predict the number of units to be sold, and thus can adjust the amount of product they have to produce, and when to produce. Combining the Lean methodology and the data mining models obtained, the organization can optimize its service, but it is important to note that in this project are not considered external factors, such as economic recession or meteorology. To conclude, it is important to realize how important data mining models in industrial

planning, together with methodologies such as Lean Production, are a fundamental tool in combating waste and helping to optimize Industrial Production. The model that had the best classification was the M4 model with the lowest RAE percentage with about 20%, which indicates that it is less likely to err. When the aim is to forecast quantities sold, it is not enough to look only at the quantities sold in previous years, it is also necessary to take into account the meteorology to realize the impact that may have existed in production. In relation to future work, it is important to carry out the last phase of CRISP-DM, which is the implementation phase, passing these models into practice, thus helping all organizations to improve their production process. From the obtained models, it is possible to have a solid base for the development and implementation of Decision Support System, able to support not only the industrial planning, but also the actual production and the orders in real time. With this type of systems allows the organization to reduce waste and optimize production.

Acknowledgments

This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

1. I. witten, et al., Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufman, 2011.
2. Y. Alsultanny, "Labor Market Forecasting by Using Data Mining," *Procedia Computer Science*, vol. 18, 2013, pp. 1700-1709.
3. F. Gullo, "From Patterns in Data to Knowledge Discovery: What Data Mining Can Do," *Physics Procedia*, vol. 62, 2015, pp. 18-22.
4. W. Xu, et al., "A Neural Network Based Forecasting Method For the Unemployment Rate: Prediction Using the Search Engine Query Data," ed., IEEE, 2011, pp.
5. S. Ramos, et al., "A data-mining-based methodology to support MV electricity customers characterization," *Procedia Energy and Buildings*, vol. 91, 2015, pp. 16-25.
6. W. Xu, et al., "Data mining for unemployment rate prediction using search engine query data," *Book DEditor ed.^eds., Springer*, 2012, pp. 33-42.
7. W. Yan, "Application Research of Data Mining Technology about Teaching Quality Assessment in Colleges and Universities," *Procedia Eng.*, vol. 15, 2011, pp. 4241-4245.
8. X. Ren, et al., "Data mining of space heating system performance in affordable housing," *Procedia Building and Environment*, vol. 89, 2015, pp. 1-13.
9. L. Martin, et al., "Using data mining techniques to road safety improvement in Spanish roads," *Procedia Social and Behavioral Sciences*, vol. 160, 2014, pp. 607-614.
10. L. Maia, et al., "Metodologias para Implementar Lean Production: Uma Revisão Crítica de Literatura," *Edições INEGI*, 2011, pp.
11. L. Maia, et al., "Definition of a Protocol for Implementing Lean Production Methodology in Textile and Clothing Case Studies," *Editor ed.^eds., 2013*, pp.
12. G. Chauhan and T.P. Singh, "Measuring parameters of lean manufacturing realization," *Measuring Business Excellence*, vol. 16, no. 3, 2012, pp. 57-71.

13. P. Hasle, et al., "Lean and working environment: a review of the literature," *International Journal of Operations & Production Management*, vol. 32, no. 7, 2012, pp. 829-849.
14. Y. Yamamoto and M. Bellgran, "Fundamental mindset that drives improvements towards lean production," *Assembly Automation*, vol. 30, no. 2, 2010, pp. 124-130; DOI 10.1108/01445151011029754].
15. K. Hassan and H. Kajiware, "Application of Pull concept-based Lean production System in the Ship Building Industry," *Journal of Ship Production and Design*, vol. 29, no. 3, 2013, pp. 105-116; DOI 10.5957/JSPD.29.3.120021.
16. E. Balashova and E. Gromova, "Prospects and Specifics of Resource Management in Enterprises Operating in Different Sectors of the Russian Economy," *Economy and Management of the Enterprise*, vol. 216, no. 2, 2015, pp. 102-108; DOI 10.5862/JE.216.12.
17. Z. Chen, *Computational intelligence for decision support*, CRC Press, Inc., 2000.
18. E. Turban, et al., *Decision Support and Business Intelligence Systems*, Prentice Hall, 2011.
19. L. Ming-Te, et al., "Using data mining technique to perform the performance assesment of lean service," *Neural Computing and Applications*, vol. 22, no. 7, 2013, pp. 1433-1444; DOI 10.1007/s00521-012-0848-y.
20. C. Groger, et al., "Data Mining-driven Manufacturing Process Optimization," *Book Data Mining-driven Manufacturing Process Optimization*, Series *Data Mining-driven Manufacturing Process Optimization* 3, ed., Editor ed.^eds., 2012, pp.
21. H. Unver, "An ISA-95-base manufacturing intelligence system in support of lean initiatives," *The International Journal of Advanced Manufacturing Technology*, vol. 65, no. 5, 2013, pp. 853-866; DOI 10.1007/s00170-012-4223-z.
22. P. Vazan, et al., "The Data Mining usage in Production System Management," *International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering*, vol. 5, no. 5, 2011.
23. M. Myers and J. Venable, "A set of ethical principles for design science research in information Systems," *Procedia Information Management*, vol. 51, no. 801-809, 2014.
24. R. Beck, et al., "Theory-generating design science research," *Inf Syst Front*, vol. 15, 2013, pp. 637-651.
25. S. Gregor and A. Hevner, "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly*, vol. 37, no. 2, 2013, pp. 337-355.
26. V. Vaishnavi and B. Kuechler, "Design Science Research in Information Systems," *Book Design Science Research in Information Systems*, Series *Design Science Research in Information Systems*, ed., Editor ed.^eds., 2013, pp.
27. S. Hain and B. Andrea., "Towards a Maturity Model for E-Collaboration - A Design Science Research Approach," *IEEE*, 2011, pp.
28. A. Hevner and S. Chatterjee, *Design Science Research in Information Systems: Theory and Practice*, Springer US, 2010.
29. K. Peffers, et al., "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, 2008, pp. 45-78.
30. O. Erohin, et al., "Intelligent Utilisation of Digital Databases for Assembly Time Determination in Early Phases of Product Emergence," *Procedia CIRP*, vol. 3, 2012, pp. 424-429.
31. A. Nadali, et al., "Evaluating the Sucess Level of Data Mining Projects Based on CRISP-DM Methodology by Fuzzy Expert System," *IEEE*, 2011, pp. 161-165.
32. A. Hoe, et al., "Analyzing Students Records to Identify Patterns of Students Performance," *IEEE*, 2013, pp. 544-547.
33. R. Wallis, et al., "Data Mining - supported Generation of Assembly Process Plans," *Procedia CIRP*, vol. 23, 2014, pp. 178-183.