MULTIPLE REGRESSION ANALYSIS SYSTEM IN MACHINE LEARNING AND ESTIMATING EFFECTS OF DATA TRANSFORMATION & MIN-MAX NORMALIZATION

Ayla Saylıa, Ceyda Akbulutb*, Kemal Kosutac

Abstract

Machine learning area is a recent topic in data analysis and a researcher or worker of the area is called “Data Scientist” which nowadays has been a highly preferred job title in computing. In this study, we have two aims that the first is to implement a multiple regression analysis system which is developed in Ubuntu operating system on the Anaconda platform using Python3 in order to construct models of each attribute to make their estimations for future decisions taking less risk in advance of past experiences hided in cumulated data and the second aim is to find out effects of data transformation and min-max normalization in the data preparation before building models. After the system implementation, we test the system to determine the best estimation model of each attribute of the vehicles sold in the five European countries between 1970 and 1999. We have constructed six versions of the original dataset and these versions are used to construct regression models for further estimations. Finally, we compute the regression criterion value of R-Squared for each constructed-model and we compare the models according to these values. Computational results are very promising that the system can be used efficiently and the effects of the data transformation and min-max normalization are significant for some attributes.

Keywords: Data Preparation, multiple regression, machine learning, python, r-squared criterion
1. Introduction

In the digital age we live on, huge amounts of structural and non-structural data have been formed as a result of the activities generated on the internet. The size of data stored in the world in 2000 was 800,000 petabytes and it is expected to be up to 35 zettabytes by 2020 [1]. Developing and changing environmental conditions, globalization of the internet, competition with different research and development activities, marketing methods and difficulties in customers’ satisfaction are increasing the importance of information obtained from data day by day. Database management systems are used to collect and manage the data by multi-users for their queries in real-time systems. Nowadays the size of the data is very big and it can be used to build worthwhile models by engineers in order to make better estimations. For this purpose, the area of machine learning can be useful and helpful. Machine learning has three main topics; Supervised Learning, Unsupervised Learning and Reinforcement Learning [2]. Supervised learning has two subtopics; Classification and Regression. Classification is used for detection problems and the regression for prediction problems like success rates of student, forecasting, population growth and sales amount. It has been recently used in many sector data analysis such as education, health, business, bioinformatics and many others. Therefore the works on the regression analysis and modelling is quite up-to-date and important by researchers, especially using big datasets [3-11].

Our goals of this paper in supervised learning are to implement a multiple regression analysis system to construct models of each attribute to make their estimations for future decisions taking less risk in advance of past experiences hided in cumulated data and to find out effects of data transformations and min-max normalization during the data preparation before building models. Before the regression analysis, the dataset is prepared in Section 2; null values of the related attributes are cleaned, outliers of the numeric type of the attributes are detected and removed then dummy values of the categorical attributes are assigned. After the min-max normalization and two different transformations of logarithmic and square-rooted, we have six versioned datasets named as “Prepared”, “Prepared-Logarithmic”, “Prepared-Square-rooted”, “Normalized”, “Normalized-Logarithmic and “Normalized-Square-rooted”. In Section 3, we introduce our system and give the background information about the regression analysis in machine learning. In Section 4, the computational results of constructed models based on our datasets and our computed regression criterion value of R-Squared for each constructed-model to compare the models are given. Finally, in Section 5, we present our conclusion and future work.

2. Data preparation

In the machine learning, before any analysis, the data should be prepared. The main purpose at this point is to pass through a set of transformation operations to ensure that the information content of the datasets is in the best form for learning tools [1, 2, 7 and 11]. The preparation must be formatted appropriately according to the software tool used. Also, there should be enough data under your hand according to each method. In theory, everything seems to be perfect, but in practice the data is usually unstructured.

Therefore the starting work of our study is the preprocessing of the data. The dataset we work with has vehicle sales information in five European countries between 1970 and 1999. These countries are Belgium, France, Germany, Italy and United Kingdom [12]. The dataset contains 11550 instances and 15 attributes. There are 11 numeric and 4 categorical attributes.
In Section 2.1, outliers are removed. In Section 2.2, the dummy value assignments are given. In Section 2.3, the min-max normalization process is shown in detail.

### 2.1 Outlier detection

A graph of each attribute is drawn to see the data intensity to determine the outliers. The outliers are deleted with the set threshold values to focus on the dataset's concentrated range [13]. The threshold values for the types of attributes and peaks are given in Table 1 below.

**Table 1. Data Description**

| Attribute Name (abbreviation) | Data Type | Threshold Value |
|-------------------------------|-----------|-----------------|
| year                          | categorical | -               |
| brand                         | categorical | -               |
| model                         | categorical | -               |
| home                          | categorical | -               |
| quantity (qu)                 | integer    | 100.000         |
| cylinder radius (cy)          | double     | 2.500           |
| weight (we)                   | double     | 1.500           |
| height (he)                   | double     | 150             |
| width (wi)                    | double     | -               |
| horse power (hp)              | double     | -               |
| length (le)                   | double     | -               |
| speed (sp)                    | double     | 200             |
| tax                           | double     | 0.275           |
| price (pr)                    | integer    | 40,000.000      |
| acceleration (ac)             | double     | 20              |

In Figure 1 below, the acceleration (ac) attribute is given before (a) and after the outliers are cleared (b) for an example.

![Figure 1. Before and after cleaning the outliers of acceleration attribute](image-url)
The data density that occurs when cutting is performed according to the threshold value determined for the quantity (qu) attribute is shown in Figure 2 below.

![Figure 2](image)

**Figure 2.** Before and after cleaning the outliers of the quantity attribute

The following table specifies the threshold values for numeric attributes and how many rows have been deleted.

**Table 2. Outliers Detail**

| Attribute abbreviation | Threshold Value | Number of Cleaned Records |
|------------------------|-----------------|---------------------------|
| qu                     | 50.000          | 1200                      |
| cy                     | 2000            | 600                       |
| we                     | 1500            | 17                        |
| he                     | 150             | 367                       |
| wi                     | -               | -                         |
| hp                     | 100             | 300                       |
| le                     | -               | -                         |
| sp                     | 200             | 188                       |
| tax                    | -               | -                         |
| pr                     | 20.000.000      | 409                       |
| ac                     | 20              | 793                       |

**2.2 Dummy value assignment**

After clearing the dataset outlier’s values, the dummy value assignment is processed for categorical attributes of year, brand, model and home. After this process, we name this version of the original dataset as the “Prepared” dataset. We then apply the min-max normalization between 0 and 1 to ignore different size problems on numeric attributes in the following subsection.
2.3 Min-Max normalization

After the assignment of dummy values, the min-max normalization is processed that the min-max normalization method is applied to convert the data to numeric values between 0 and 1[14]. This method is based on determining the largest and smallest numerical values of each numeric attribute and transforming the others accordingly. The commonly used formula is shown below:

\[
X^* = \frac{X - X_{min}}{X_{max} - X_{min}}
\]  

(1)

Where \(X^*\) is the transformed value, \(X\) is the observation value, \(X_{min}\) is the smallest observation value, and \(X_{max}\) is the largest observation value. The values in the dataset are reduced to \(\{0, 1\}\).

Table 3. Min and max values of attributes

| Attribute abbreviation | Minimum Value | Maximum Value |
|------------------------|---------------|---------------|
| qu                     | 53,00         | 49988,00      |
| cy                     | 499,00        | 1999,00       |
| we                     | 520,00        | 1460,00       |
| he                     | 117,50        | 149,00        |
| wi                     | 129,50        | 182,00        |
| hp                     | 13,00         | 99,50         |
| le                     | 297,00        | 493,00        |
| sp                     | 95,00         | 199,00        |
| tax                    | 0,12          | 0,33          |
| pr                     | 498,00        | 19.986.000,00 |
| ac                     | 8,30          | 19,70         |

After the min-max normalization, we name this version of “Prepared” dataset as “Normalized” dataset. We then apply two different transformations; logarithmic in Section 2.4 and square rooted in Section 2.5 to both datasets.

2.4 Logarithmic transformation

In the logarithmic transformation, the logarithm value of each numeric attribute value is calculated and the logarithm values are taken into the multiple regression analysis that it is applied after each target attribute was determined. Then the inverse function is applied to estimate the values of the target attribute. The logarithmic is taken and the results are achieved. We name these versions of the datasets as “Prepared-Logarithmic” and “Normalized-Logarithmic”.

2.5 Square rooted transformation

In the square root transformation, values of each attribute are square rooted and then the multiple regression analysis is applied after each target attribute was determined. Then the inverse function is applied to estimate the values of the target attribute. The square root is
taken and the results are achieved. We name these versions of the datasets as “Prepared-Square-rooted” and “Normalized-Square-rooted”.

After the transformations, we have six versions of the original dataset: “Prepared”, “Prepared-Logarithmic”, “Prepared-Square-rooted”, “Normalized”, “Normalized-Logarithmic” and “Normalized-Square-rooted”. These datasets are used to construct multiple regression models in order to find out the effects of the transformations and min-max normalization.

3. Multiple regression analysis system in machine learning

Our multiple regression analysis system is self-coded on the Anaconda platform using Python3 for scientists, engineers and data analysts. Regression analysis is summarized in Section 3.1 below and then the used criterion of R-Squared is described briefly in Section 3.2.

3.1 Regression analysis in machine learning

Regression analysis is a method used to examine the relationship between attributes. When a correlation between attributes is found, this relation can be expressed in a model. It is used to construct a linear or non-linear model based on a single or multiple independent attributes to estimate values of a dependent attribute. “Single linear regression” model assumes that the relationship between the dependent attribute $y_i$ and the independent attribute $x_i$ is linear. The model of the regression can be formed with $y_i = a + bx_i$ where $a$ is the offset and $b$ is the slope of the linear relationship [3]. If the regression model includes a dependent attribute based on multiple independent attributes and is called as "Multiple Linear Regression". In this paper, we are focused on this model which is shown in Formula (5) as follows:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

(5)

$\hat{Y}$ is the estimated $Y$ value which is the dependent attribute, $b_0$ is the estimated regression cut-off point, $b_1, b_2, \ldots, b_k$ are estimated slope coefficients and $X_1, X_2, \ldots, X_k$ are independent attributes. In this paper, the coefficients of each model for every attribute are calculated based on the six datasets.

3.2 R-Squared criterion

After making the data preparation, it is necessary to calculate the erroneous estimation rates to compare successes of models to choose which is better or optimum to use for future decisions. R-squared criterion in Formula (6) is a statistical measure of how close the data are to the fitted regression line. 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

(6)

SSE is the sum of squared errors of the model shown in Formula (7).
\[ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

(7)

SST is the sum of squared errors of our baseline model shown in Formula (8).

\[ SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \]  

(8)

\( \hat{y}_i \) is the predicted value of \( y_i \) which is the real value and \( \bar{y}_i \) is the average value of all \( y_i \).

4. Computational results

During our experiments, the system is executed many times to construct the models based on taking each attribute as a dependent attribute and the others as independent attributes in our six versioned datasets. In the following sections from 4.1 to 4.6, we give the results of the ac attribute for each dataset in detail for an example. There are 6 models for each attribute based on six versioned datasets and the number of the constructed models is 66 in total for 11 attributes. We could not give all the models detailed due to the page restriction but we compared all models according to the regression criterion of R-Squared represented in Section 4.7.
4.1 Regression model for ac attribute based on “prepared” dataset

The regression model of ac attribute is shown in Figure 3.

Figure 3. Regression model for ac attribute based on “prepared” dataset

4.2 Regression model for ac attribute based on “prepared–logarithmic” Dataset

The regression model of ac attribute is shown in Figure 4.
**Figure 4.** Regression model for ac attribute based on "prepared–logarithmic" dataset
4.3 Regression model for ac attribute based on “prepared–square-rooted” dataset

The regression model of ac attribute is shown in Figure 5.

![OLS Regression Results](image)

**Figure 5.** Regression model for ac attribute based on “prepared–square-rooted” dataset
4.4 Regression model for ac attribute based on “normalised” dataset

The regression model of ac attribute is shown in Figure 6.

Figure 6. Regression model for ac attribute based on “normalised” dataset
4.5 Regression model for ac attribute based on “Normalised–Logarithmic” dataset

The regression model of ac attribute is shown in Figure 7.

![Regression model results](image)

---

**Figure 7.** Regression model for ac attribute based on “normalised–logarithmic” dataset
4.6 Regression model for ac attribute based on “normalised–square-rooted” dataset

The regression model of ac attribute is shown in Figure 8.

![OLS Regression Results](image)

**Figure 8.** Regression model for ac attribute based on “normalised–square-rooted” dataset
4.7 R-Squared criterion

R-squared values of 6 different regression models for each attribute were obtained and are shown in Figure 9 (a) and (b). From this figure, the best model of each attribute can be determined and then it can be used to make better estimations.

Figure 9 (a). R-squared values of qu, sp, tax, we, wi and cy

Figure 9 (b). R-squared values of ac, he, hp, le and pr

From Figure 9, it can be seen that the highest R-squared value for the qu attribute is obtained when the Ln transformation applied and the largest R-squared value for ac attribute is obtained after the min-max normalization and the Ln transformation applied. This figure can be taken into further consideration to choose which version of the original dataset should be used for each attribute to make the multiple regression model.
5. Conclusion and future work

The results show that the multiple regression analysis can be used for predictions and the transformations can be used to reach better results for some attributes such as ac and hp than using the original dataset. Therefore, in addition to our system, the following improvements can be worked in the future. The first one can be a mixed method in which the conversion of the others can be done in a mixed way so that each attribute can be estimated in its best way. For example, for the qu attribute, the best R-squared results were reached by logarithmic transformation, whereas the other attributes should be applied whichever yield their best results. The second works may take a long term to have new attributes which may affect the sales of automobiles and various analyzes can be made further like exchange rates, per capita national incomes. Last but not the least important one can be to have a real data to analyze which attributes are more effective than the others in order to estimate the sales amount of each vehicle.

References

[1] Zikopoulos, P.C., Eaton, C., deRoos, D., Deutsch, T., Lapis, G., Understanding Big Data, McGrawHill, New York, 2012.
[2] Witten, Ian H., et al., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.
[3] Friedman, J., Trevor H., and Tibshirani R., The Elements of Statistical Learning, Vol. 1. Springer series in statistics, New York, 2001.
[4] Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, Bauerschlag DO, Jo¨ckel KH, Erbel R, Mu¨hleisen TW, Zenke M, Bru¨mmendorf TH, Wagner W., “Aging of Blood Can Be Tracked by DNA Methylation Changes at Just Three CpG Sites”, Genome Biol 15.2 (2014):1–11.
[5] Gareth J., Witten D., Hastie T., Tibshirani R., An Introduction to Statistical Learning, Springer, New York, ISBN 978-1-4614-7137-0, 2015.
[6] Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A., “Deep Biomarkers of Human Aging : Application of Deep Neural Networks to Biomarker Development”, Aging 8.5 (2016):1–13.
[7] Hox, Joop J., Mirjam M., and Rens Van de Schoot, Multilevel Analysis: Techniques and Applications, Routledge, 2017.
[8] Hu, Rui, et al., "A Short-term Power Load Forecasting Model based on the Generalized Regression Neural Network with Decreasing Step Fruit Fly Optimization Algorithm", Neurocomputing, 221 (2017): 24-31.
[9] Kristof De W. and López-Torres L., "Efficiency in Education: a Review of Literature and a Way Forward", Journal of the Operational Research Society, 68.4 (2017): 339-363.
[10] Gunasekaran M. and Lopez D., "Health Data Analytics Using Scalable Logistic Regression with Stochastic Gradient Descent", International Journal of Advanced Intelligence Paradigms, 10.1-2 (2018): 118-132.
[11] Markus H., et al., "Economic Development Matters: A Meta-Regression Analysis on the Relation between Environmental Management and Financial Performance", Journal of Industrial Ecology, 22.4 (2018): 720-744.

[12] https://sites.google.com/site/frankverbo/data-and-software/data-set-on-the-european-car-market.

[13] Aggarwal, C. C., An introduction to outlier analysis. In Outlier analysis, New York NY: Springer, (2013): 1-40.

[14] Ilango, V., Subramanian, R., & Vasudevan, V., “A five step procedure for outlier analysis in data mining”, European Journal of Scientific Research, 75(3) (2012): 327-339.