Journal of Engineering Technology and Applied Sciences, 2018 *e-ISSN:* 2548-0391
Received 26 October 2019, Accepted 23 December 2018 *Vol. 3, No. 3, 189-204*
Published online 30 December 2018 *doi: 10.30931/jetas.475215*

Citation: Saylı, A., Akbulut, C., Kosuta, K., "Multiple Regression Analysis System in Machine Learning and Estimating Effects of Data Transformation&Min-Max Normalization". Journal of Engineering Technology and Applied Sciences 3 (3) 2018 : 189-204.

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

Ayla Saylı[a], Ceyda Akbulut[b*], Kemal Kosuta[c]

[a]*Department of Mathematical Engineering, Faculty of Chemistry and Metallurgical, University of Yildiz Technical, Istanbul, Turkey*
sayli@yildiz.edu.tr

[b*]*Department of Mathematical Engineering, Faculty of Chemistry and Metallurgical, University of Yildiz Technical, Istanbul, Turkey (corresponding author)*
cey.akbulut@gmail.com

[c]*Department of Mathematical Engineering, Faculty of Chemistry and Metallurgical, University of Yildiz Technical, Istanbul, Turkey*
kosuta@yildiz.edu.tr

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.

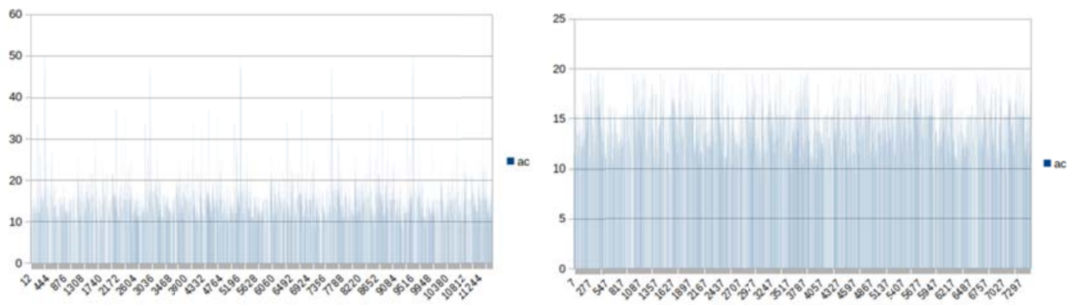
2.1 Outlier detection

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

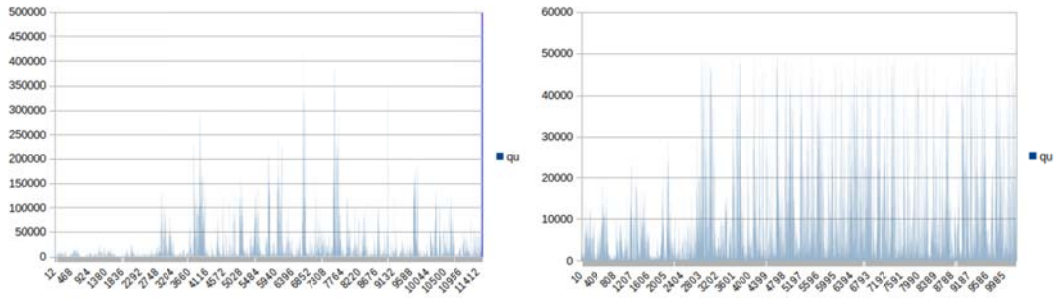
In Figure 1 below, the acceleration (ac) attribute is given before (a) and after the outliers are cleared (b) for an example.



(a) Before (b) After

Figure 1. Before and after cleaning the outliers of acceleration attribute

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.



(a) Before (b) After

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 normalisation 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:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

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:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{5}$$

$\hat{Y}$ is the estimated Y value which is the dependent attribute, b_0 is the estimated regression cut-off point, $b_1, b_2 \ldots\ldots, b_k$ are estimated slope coefficients and $X_1, X_2, \ldots\ldots, X_k$ 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.

$$R^2 = 1 - \frac{SSE}{SST} \tag{6}$$

SSE is the sum of squared errors of the model shown in Formula (7).

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

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

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{8}$$

$\hat{y}_i$ 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.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.380
Method:	Least Squares	F-statistic:	14.25
Date:	Sun, 21 Jan 2018	Prob (F-statistic):	0.00
Time:	22:10:59	Log-Likelihood:	-77118.
No. Observations:	7453	AIC:	1.549e+05
Df Residuals:	7108	BIC:	1.573e+05
Df Model:	344		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.178e+04	7779.872	-1.514	0.130	-2.7e+04	3471.497
x1	-5844.0190	3901.044	-1.498	0.134	-1.35e+04	1803.190
x2	-5935.3505	3885.870	-1.527	0.127	-1.36e+04	1682.113
x3	6771.0925	6960.219	0.973	0.331	-6873.009	2.04e+04
x4	-3090.0931	6953.499	-0.444	0.657	-1.67e+04	1.05e+04
x5	-3592.8156	7263.501	-0.495	0.621	-1.78e+04	1.06e+04
x6	1844.2740	7398.384	0.249	0.803	-1.27e+04	1.63e+04
x7	-1191.2182	7415.769	-0.161	0.872	-1.57e+04	1.33e+04
x8	-6037.7680	7255.349	-0.832	0.405	-2.03e+04	8184.876
x9	-5699.7052	7213.242	-0.790	0.429	-1.98e+04	8440.397
x10	-6281.2509	7282.724	-0.862	0.388	-2.06e+04	7995.057
x11	-8082.1098	1.05e+04	-0.767	0.443	-2.87e+04	1.26e+04
x12	-5836.6107	7282.053	-0.802	0.423	-2.01e+04	8438.381
x13	7322.7005	7246.159	1.011	0.312	-6881.929	2.15e+04
x14	-7726.4701	7419.599	-1.041	0.298	-2.23e+04	6818.153
x15	-7734.0368	7307.675	-1.058	0.290	-2.21e+04	6591.182
x16	-8105.6919	7446.659	-1.089	0.276	-2.27e+04	6491.977
x17	-550.8464	3433.478	-0.160	0.873	-7281.486	6179.793
x18	1914.8614	3370.507	0.568	0.570	-4692.335	8522.058
x19	-595.3690	4799.770	-0.124	0.901	-1e+04	8813.609
x20	-2020.7381	3423.413	-0.590	0.555	-8731.647	4690.171
x21	-2051.3453	4647.194	-0.441	0.659	-1.12e+04	7058.540
x22	923.3504	7153.050	0.129	0.897	-1.31e+04	1.49e+04
x23	1957.3939	3447.255	0.568	0.570	-4800.253	8715.041
x24	-1062.3825	1654.843	-0.642	0.521	-4306.368	2181.603
x25	-585.2615	1808.224	-0.324	0.746	-4129.919	2959.396
x26	2471.1705	1497.496	1.650	0.099	-464.368	5406.709
x27	-3178.7539	2258.598	-1.407	0.159	-7606.278	1248.770
x28	-3791.5672	2909.505	-1.303	0.193	-9495.064	1911.929
x29	1.394e+04	1925.483	7.241	0.000	1.02e+04	1.77e+04
x30	1.773e+04	2007.155	8.836	0.000	1.38e+04	2.17e+04
x31	-2158.6690	4034.743	-0.535	0.593	-1.01e+04	5750.629
x32	706.5170	4778.792	0.148	0.882	-8661.338	1.01e+04
x33	-61.5373	4768.120	-0.013	0.990	-9408.472	9285.397
x34	-2182.5539	3908.078	-0.558	0.577	-9843.551	5478.444
x35	-326.7756	7942.648	-0.041	0.967	-1.59e+04	1.52e+04
x36	7106.3023	3648.287	1.948	0.051	-45.426	1.43e+04
x37	6846.7233	3056.495	2.240	0.025	855.082	1.28e+04
x38	-2332.4184	1528.424	-1.526	0.127	-5328.584	663.747
x39	-1424.0300	1468.850	-0.969	0.332	-4303.413	1455.353
x40	632.5004	5223.618	0.121	0.904	-9607.347	1.09e+04

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.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.380
Method:	Least Squares	F-statistic:	14.25
Date:	Sun, 24 Dec 2017	Prob (F-statistic):	0.00
Time:	23:32:15	Log-Likelihood:	-77118.
No. Observations:	7453	AIC:	1.549e+05
Df Residuals:	7108	BIC:	1.573e+05
Df Model:	344		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.316e+04	3.94e+04	-1.603	0.109	-1.4e+05	1.41e+04
x1	-3.15e+04	1.97e+04	-1.598	0.110	-7.01e+04	7139.444
x2	-3.166e+04	1.97e+04	-1.607	0.108	-7.03e+04	6949.810
x3	6169.1015	6960.361	0.886	0.375	-7475.280	1.98e+04
x4	-3679.7521	6957.712	-0.529	0.597	-1.73e+04	9959.435
x5	-4263.0734	7283.052	-0.585	0.558	-1.85e+04	1e+04
x6	989.9136	7424.710	0.133	0.894	-1.36e+04	1.55e+04
x7	-2046.3421	7440.392	-0.275	0.783	-1.66e+04	1.25e+04
x8	-6858.1248	7288.322	-0.941	0.347	-2.11e+04	7429.156
x9	-6551.8255	7243.238	-0.905	0.366	-2.08e+04	7647.079
x10	-7253.2000	7308.669	-0.992	0.321	-2.16e+04	7073.968
x11	-9197.6554	1.06e+04	-0.870	0.384	-2.99e+04	1.15e+04
x12	-6759.3314	7315.375	-0.924	0.356	-2.11e+04	7580.981
x13	6383.6837	7270.789	0.878	0.380	-7869.229	2.06e+04
x14	-8760.0128	7457.410	-1.175	0.240	-2.34e+04	5858.731
x15	-8663.4566	7337.215	-1.181	0.238	-2.3e+04	5719.669
x16	-9088.6694	7470.920	-1.217	0.224	-2.37e+04	5556.558
x17	-691.4182	3441.239	-0.201	0.841	-7437.271	6054.434
x18	2210.0463	3373.629	0.655	0.512	-4403.272	8823.364
x19	-968.0335	4815.705	-0.201	0.841	-1.04e+04	8472.182
x20	-2125.7329	3435.789	-0.619	0.536	-8860.903	4609.437
x21	-2199.8477	4655.698	-0.473	0.637	-1.13e+04	6926.707
x22	1793.4947	7179.664	0.250	0.803	-1.23e+04	1.59e+04
x23	2027.0578	3430.838	0.591	0.555	-4698.406	8752.521
x24	-964.2440	1649.393	-0.585	0.559	-4197.546	2269.058
x25	-661.6203	1811.051	-0.365	0.715	-4211.820	2888.579
x26	2566.0061	1508.164	1.701	0.089	-390.444	5522.456
x27	-3331.8915	2299.986	-1.449	0.147	-7840.548	1176.765
x28	-4144.8940	2972.276	-1.395	0.163	-9971.440	1681.652
x29	1.371e+04	1938.796	7.070	0.000	9905.733	1.75e+04
x30	1.762e+04	2045.653	8.611	0.000	1.36e+04	2.16e+04
x31	-2896.9408	4052.099	-0.715	0.475	-1.08e+04	5046.380
x32	883.7458	4794.339	0.184	0.854	-8514.587	1.03e+04
x33	80.9520	4783.725	0.017	0.986	-9296.574	9458.478
x34	-2076.5426	3919.416	-0.530	0.596	-9759.766	5606.680
x35	-519.7190	7935.751	-0.065	0.948	-1.61e+04	1.5e+04
x36	6942.2544	3583.618	1.937	0.053	-82.704	1.4e+04
x37	6717.3202	3056.516	2.198	0.028	725.638	1.27e+04
x38	-2385.4276	1524.187	-1.565	0.118	-5373.288	602.433
x39	-1446.4423	1448.965	-0.998	0.318	-4286.845	1393.960
x40	557.8951	5217.101	0.107	0.915	-9669.177	1.08e+04

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

Dep. Variable:	y	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.380
Method:	Least Squares	F-statistic:	14.25
Date:	Mon, 25 Dec 2017	Prob (F-statistic):	0.00
Time:	00:25:31	Log-Likelihood:	-77118.
No. Observations:	7453	AIC:	1.549e+05
Df Residuals:	7108	BIC:	1.573e+05
Df Model:	344		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.315e+04	1.55e+04	-1.491	0.136	-5.36e+04	7277.058
x1	-1.152e+04	7771.495	-1.482	0.138	-2.68e+04	3718.249
x2	-1.163e+04	7752.852	-1.500	0.134	-2.68e+04	3565.762
x3	6570.1467	6961.934	0.944	0.345	-7077.318	2.02e+04
x4	-3288.2959	6954.940	-0.473	0.636	-1.69e+04	1.03e+04
x5	-3829.3325	7271.129	-0.527	0.598	-1.81e+04	1.04e+04
x6	1533.9576	7407.324	0.207	0.836	-1.3e+04	1.61e+04
x7	-1503.8005	7424.299	-0.203	0.839	-1.61e+04	1.31e+04
x8	-6342.4428	7267.153	-0.873	0.383	-2.06e+04	7903.342
x9	-6006.2136	7225.867	-0.831	0.406	-2.02e+04	8158.638
x10	-6666.5011	7294.042	-0.914	0.361	-2.1e+04	7631.993
x11	-8514.8210	1.06e+04	-0.807	0.420	-2.92e+04	1.22e+04
x12	-6182.2185	7294.933	-0.847	0.397	-2.05e+04	8118.022
x13	6976.4452	7254.517	0.962	0.336	-7244.568	2.12e+04
x14	-8111.6135	7441.820	-1.090	0.276	-2.27e+04	6476.570
x15	-8075.5850	7317.260	-1.104	0.270	-2.24e+04	6268.424
x16	-8473.3342	7457.488	-1.136	0.256	-2.31e+04	6145.564
x17	-563.7575	3439.660	-0.164	0.870	-7306.516	6179.001
x18	2105.4816	3378.397	0.623	0.533	-4517.183	8728.146
x19	-787.1538	4830.942	-0.163	0.871	-1.03e+04	8682.931
x20	-2019.9391	3424.109	-0.590	0.555	-8732.213	4692.335
x21	-2064.2512	4646.320	-0.444	0.657	-1.12e+04	7043.920
x22	1419.1578	7165.162	0.198	0.843	-1.26e+04	1.55e+04
x23	2046.3986	3448.544	0.593	0.553	-4713.774	8806.571
x24	-986.7618	1658.146	-0.595	0.552	-4237.223	2263.699
x25	-600.8968	1823.755	-0.329	0.742	-4176.000	2974.207
x26	2591.7652	1518.136	1.707	0.088	-384.234	5567.764
x27	-3208.0238	2274.807	-1.410	0.159	-7667.322	1251.274
x28	-3911.5175	2947.075	-1.327	0.184	-9688.663	1865.628
x29	1.386e+04	1932.497	7.175	0.000	1.01e+04	1.77e+04
x30	1.773e+04	2023.673	8.761	0.000	1.38e+04	2.17e+04
x31	-2519.9860	4051.803	-0.622	0.534	-1.05e+04	5422.755
x32	842.1466	4783.934	0.176	0.860	-8535.788	1.02e+04
x33	56.1365	4772.739	0.012	0.991	-9299.853	9412.126
x34	-2084.5836	3906.080	-0.534	0.594	-9741.663	5572.496
x35	-339.2206	7940.377	-0.043	0.966	-1.59e+04	1.52e+04
x36	7060.5475	3621.365	1.950	0.051	-38.406	1.42e+04
x37	6829.4221	3055.099	2.235	0.025	840.518	1.28e+04
x38	-2330.6976	1526.999	-1.526	0.127	-5324.071	662.676
x39	-1419.2502	1468.222	-0.967	0.334	-4297.402	1458.902
x40	604.4892	5222.353	0.116	0.908	-9632.877	1.08e+04
x41	-3808.9685	3875.008	-0.983	0.326	-1.14e+04	3787.201

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.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	14.25
Date:	Sun, 24 Dec 2017	Prob (F-statistic):	0.00
Time:	22:21:15	Log-Likelihood:	-77118.
No. Observations:	7453	AIC:	1.549e+05
Df Residuals:	7108	BIC:	1.573e+05
Df Model:	344		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3079.1980	1656.714	-1.859	0.063	-6326.851	168.455
x1	-1494.9363	847.579	-1.764	0.078	-3156.445	166.572
x2	-1584.2617	841.951	-1.882	0.060	-3234.736	66.213
x3	6821.9049	6961.560	0.980	0.327	-6824.826	2.05e+04
x4	-3026.6352	6954.200	-0.435	0.663	-1.67e+04	1.06e+04
x5	-3475.7672	7264.392	-0.478	0.632	-1.77e+04	1.08e+04
x6	1935.9821	7397.435	0.262	0.794	-1.26e+04	1.64e+04
x7	-1108.3852	7414.357	-0.149	0.881	-1.56e+04	1.34e+04
x8	-5903.7420	7253.703	-0.814	0.416	-2.01e+04	8315.676
x9	-5591.7169	7214.462	-0.775	0.438	-1.97e+04	8550.778
x10	-6180.7596	7283.583	-0.849	0.396	-2.05e+04	8097.233
x11	-7950.1518	1.05e+04	-0.754	0.451	-2.86e+04	1.27e+04
x12	-5713.3481	7281.098	-0.785	0.433	-2e+04	8559.771
x13	7399.9526	7244.005	1.022	0.307	-6800.453	2.16e+04
x14	-7575.3261	7428.890	-1.020	0.308	-2.21e+04	6987.511
x15	-7610.2425	7304.288	-1.042	0.297	-2.19e+04	6708.336
x16	-7945.9865	7448.879	-1.067	0.286	-2.25e+04	6656.034
x17	-544.4616	3441.085	-0.158	0.874	-7290.012	6201.089
x18	1895.0503	3374.427	0.562	0.574	-4719.831	8509.932
x19	-665.8996	4838.516	-0.138	0.891	-1.02e+04	8819.033
x20	-1956.9808	3421.635	-0.572	0.567	-8664.404	4750.443
x21	-1963.3153	4645.728	-0.423	0.673	-1.11e+04	7143.694
x22	893.0590	7153.231	0.125	0.901	-1.31e+04	1.49e+04
x23	1944.4361	3453.101	0.563	0.573	-4824.670	8713.543
x24	-1082.4623	1658.099	-0.653	0.514	-4332.829	2167.905
x25	-630.0883	1820.222	-0.346	0.729	-4198.265	2938.088
x26	2449.0554	1514.173	1.617	0.106	-519.174	5417.285
x27	-3116.0757	2258.522	-1.380	0.168	-7543.451	1311.300
x28	-3731.8983	2925.596	-1.276	0.202	-9466.937	2003.140
x29	1.397e+04	1927.142	7.249	0.000	1.02e+04	1.77e+04
x30	1.779e+04	2008.838	8.857	0.000	1.39e+04	2.17e+04
x31	-2174.6213	4053.561	-0.536	0.592	-1.01e+04	5771.565
x32	714.4394	4778.602	0.150	0.881	-8653.043	1.01e+04
x33	-40.9598	4767.954	-0.009	0.993	-9387.570	9305.650
x34	-2135.7152	3904.652	-0.547	0.584	-9789.996	5518.566
x35	-336.5386	7944.034	-0.042	0.966	-1.59e+04	1.52e+04
x36	7163.3636	3653.007	1.961	0.050	2.382	1.43e+04
x37	6909.3647	3055.239	2.261	0.024	920.185	1.29e+04
x38	-2315.7876	1529.124	-1.514	0.130	-5313.326	681.751
x39	-1497.7557	1469.609	-1.019	0.308	-4378.628	1383.116
x40	617.8119	5224.737	0.118	0.906	-9624.229	1.09e+04

Figure 6. Regression model for ac attribute based on "*normalised*" dataset

4.5 Regression model for ac mttribute based on *“Normalised–Logarithmic”* dataset

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

OLS Regression Results

Dep. Variable:	y	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.851
Method:	Least Squares	F-statistic:	124.8
Date:	Sat, 02 Dec 2017	Prob (F-statistic):	0.00
Time:	23:27:13	Log-Likelihood:	8654.8
No. Observations:	7453	AIC:	-1.662e+04
Df Residuals:	7108	BIC:	-1.423e+04
Df Model:	344		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6225	0.017	37.411	0.000	0.590	0.655
x1	0.3098	0.009	36.389	0.000	0.293	0.326
x2	0.3127	0.008	36.981	0.000	0.296	0.329
x3	0.1389	0.070	1.986	0.047	0.002	0.276
x4	0.1226	0.070	1.755	0.079	-0.014	0.260
x5	-0.0014	0.073	-0.019	0.985	-0.144	0.142
x6	0.1816	0.074	2.444	0.015	0.036	0.327
x7	0.1991	0.074	2.673	0.008	0.053	0.345
x8	0.2589	0.073	3.553	0.000	0.116	0.402
x9	0.2215	0.072	3.057	0.002	0.079	0.364
x10	0.1652	0.073	2.258	0.024	0.022	0.309
x11	0.4091	0.106	3.862	0.000	0.201	0.617
x12	0.2581	0.073	3.530	0.000	0.115	0.402
x13	0.2200	0.073	3.024	0.003	0.077	0.363
x14	0.1286	0.075	1.723	0.085	-0.018	0.275
x15	0.0823	0.073	1.122	0.262	-0.061	0.226
x16	0.1373	0.075	1.835	0.066	-0.009	0.284
x17	-0.4041	0.035	-11.692	0.000	-0.472	-0.336
x18	0.0764	0.034	2.254	0.024	0.010	0.143
x19	0.1603	0.049	3.299	0.001	0.065	0.256
x20	-0.3445	0.034	-10.023	0.000	-0.412	-0.277
x21	0.2555	0.047	5.476	0.000	0.164	0.347
x22	-0.6076	0.072	-8.456	0.000	-0.748	-0.467
x23	-0.1050	0.035	-3.027	0.002	-0.173	-0.037
x24	0.1517	0.017	9.111	0.000	0.119	0.184
x25	0.1420	0.018	7.769	0.000	0.106	0.178
x26	0.0021	0.015	0.139	0.890	-0.028	0.032
x27	0.3767	0.023	16.605	0.000	0.332	0.421
x28	0.0665	0.029	2.263	0.024	0.009	0.124
x29	0.2140	0.019	11.055	0.000	0.176	0.252
x30	0.1769	0.020	8.765	0.000	0.137	0.216
x31	0.2375	0.041	5.833	0.000	0.158	0.317
x32	-0.3239	0.048	-6.747	0.000	-0.418	-0.230
x33	-0.3223	0.048	-6.730	0.000	-0.416	-0.228
x34	-0.3092	0.039	-7.884	0.000	-0.386	-0.232
x35	-0.0332	0.080	-0.416	0.678	-0.190	0.123
x36	-0.2709	0.037	-7.382	0.000	-0.343	-0.199
x37	-0.2717	0.031	-8.854	0.000	-0.332	-0.212
x38	-0.1362	0.015	-8.870	0.000	-0.166	-0.106
x39	0.0621	0.015	4.207	0.000	0.033	0.091

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

Dep. Variable:	y	R-squared:	0.852
Model:	OLS	Adj. R-squared:	0.845
Method:	Least Squares	F-statistic:	119.3
Date:	Mon, 22 Jan 2018	Prob (F-statistic):	0.00
Time:	21:24:28	Log-Likelihood:	8509.2
No. Observations:	7453	AIC:	-1.633e+04
Df Residuals:	7108	BIC:	-1.394e+04
Df Model:	344		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8476	0.030	27.799	0.000	0.788	0.907
x1	0.4223	0.015	27.426	0.000	0.392	0.452
x2	0.4253	0.015	27.830	0.000	0.395	0.455
x3	0.1332	0.071	1.869	0.062	-0.007	0.273
x4	0.1424	0.071	2.001	0.045	0.003	0.282
x5	0.0018	0.074	0.025	0.980	-0.144	0.148
x6	0.2381	0.076	3.141	0.002	0.089	0.387
x7	0.2522	0.076	3.318	0.001	0.103	0.401
x8	0.3238	0.074	4.361	0.000	0.178	0.469
x9	0.2861	0.074	3.868	0.000	0.141	0.431
x10	0.2554	0.075	3.424	0.001	0.109	0.402
x11	0.4991	0.108	4.619	0.000	0.287	0.711
x12	0.3299	0.075	4.424	0.000	0.184	0.476
x13	0.2736	0.074	3.685	0.000	0.128	0.419
x14	0.2236	0.076	2.947	0.003	0.075	0.372
x15	0.1338	0.075	1.789	0.074	-0.013	0.280
x16	0.1904	0.076	2.505	0.012	0.041	0.339
x17	-0.3794	0.035	-10.778	0.000	-0.448	-0.310
x18	0.0590	0.035	1.702	0.089	-0.009	0.127
x19	0.1561	0.050	3.132	0.002	0.058	0.254
x20	-0.3248	0.035	-9.303	0.000	-0.393	-0.256
x21	0.2499	0.047	5.271	0.000	0.157	0.343
x22	-0.7471	0.074	-10.099	0.000	-0.892	-0.602
x23	-0.1229	0.035	-3.491	0.000	-0.192	-0.054
x24	0.1331	0.017	7.862	0.000	0.100	0.166
x25	0.1199	0.019	6.462	0.000	0.084	0.156
x26	-0.0185	0.015	-1.214	0.225	-0.048	0.011
x27	0.4136	0.023	17.636	0.000	0.368	0.460
x28	0.1172	0.030	3.872	0.000	0.058	0.176
x29	0.2313	0.020	11.626	0.000	0.192	0.270
x30	0.1986	0.021	9.511	0.000	0.158	0.239
x31	0.2839	0.042	6.754	0.000	0.201	0.366
x32	-0.2922	0.049	-5.978	0.000	-0.388	-0.196
x33	-0.2851	0.049	-5.848	0.000	-0.381	-0.190
x34	-0.2837	0.040	-7.140	0.000	-0.362	-0.206
x35	-0.0295	0.081	-0.363	0.716	-0.189	0.130
x36	-0.3236	0.037	-8.817	0.000	-0.396	-0.252
x37	-0.2832	0.031	-9.081	0.000	-0.344	-0.222
x38	-0.1361	0.016	-8.701	0.000	-0.167	-0.105
x39	0.0411	0.015	2.772	0.006	0.012	0.070
x40	0.0924	0.053	1.729	0.084	-0.012	0.197

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.

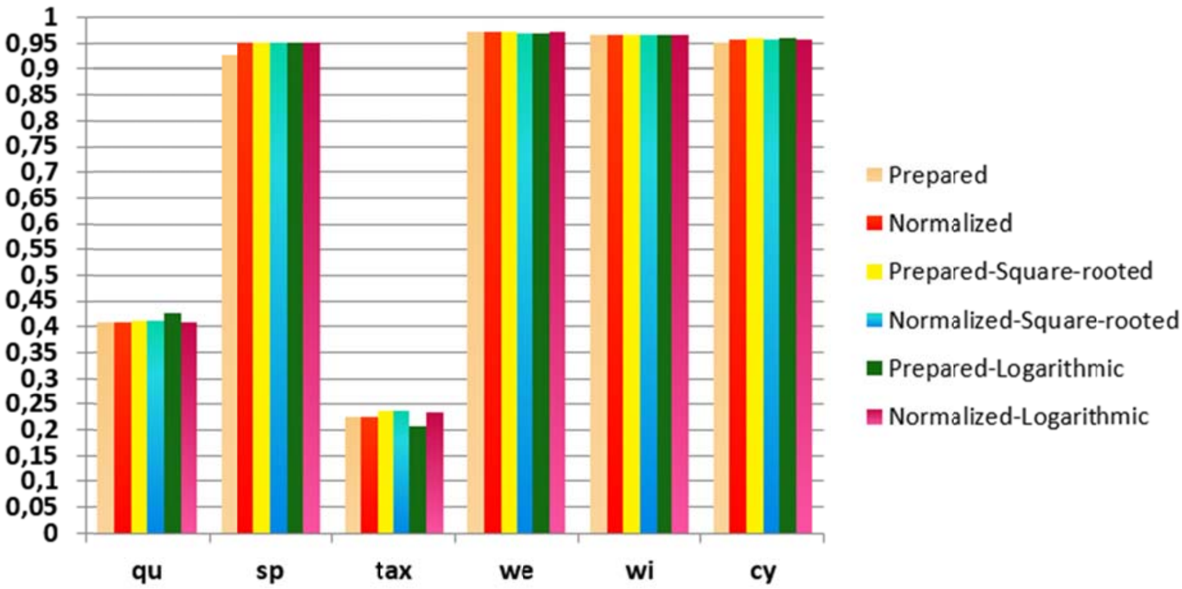


Figure 9 (a). R-squared values of qu, sp, tax, we, wi and cy

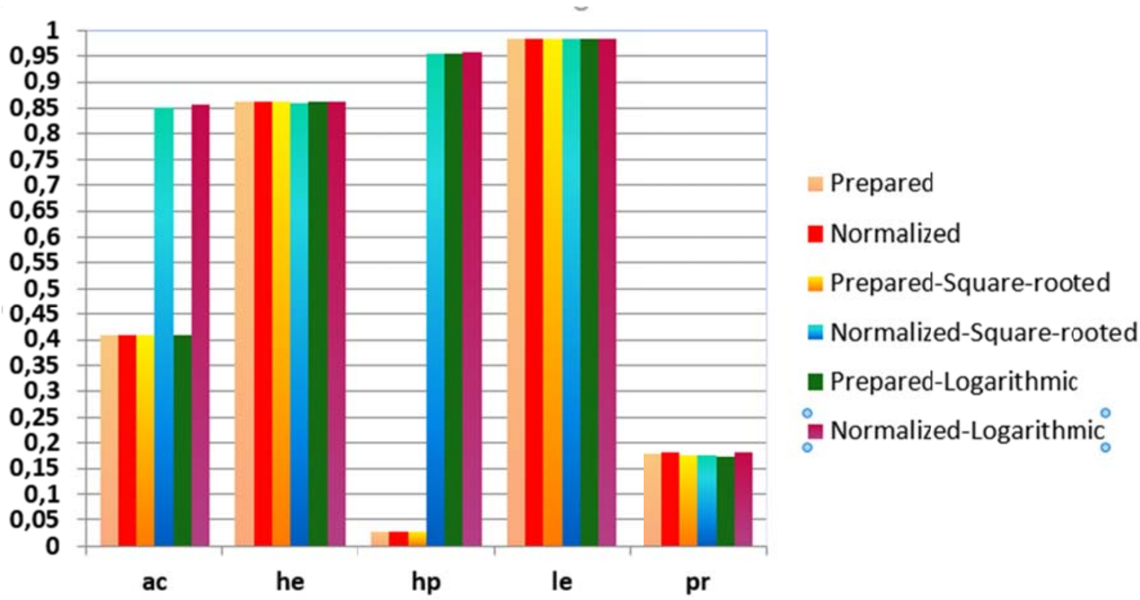


Figure 9 (b). R-squared values of ac, he, hp, le and pr

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.