European Research Studies Journal
Volume XXVII, Special Issue 2, 2024

pp. 179-189

Construction of Regression Models Predicting Lead Times and Classification Models

Submitted 18/02/24, 1st revision 16/03/24, 2nd revision 20/04/24, accepted 16/05/24

Paweł Olszewski[1], Leszek Gil[2], Natalia Rak[3], Tomasz Wołowiec[4], Michał Jasieński[5]

Abstract:

Purpose: *This article presents the process of building and applying regression models to predict lead time and classification models in supply chain management.*
Design/Methodology/Approach: *The article presents the construction of regression models predicting lead times and classification models for partial orders and complete orders*
Findings: *Using classification and regression models in the furniture industry increases customer satisfaction through timely order fulfillment, reduced costs associated with delays, and effective management of company resources.*
Practical Implications: *Using regression models to determine forecast delivery times for delayed orders allows you to manage customer expectations better and minimize delays' impact on the entire supply chain. With accurate lead time forecasts, the company can make informed decisions about resource allocation, production planning, and logistics, contributing to operational efficiency.*
Originality/Value: *Using predictive models in the procurement management process allows for continuous improvement of logistics processes by analyzing historical data and identifying trends.*

Keywords: *Regression, classification, XGBoost, knn.*

JEL codes: *C45, C01, C53, L74, M11.*

Paper type: *Research article.*

[1]*Corresponding Author: Netrix S.A/ WSEI University, Lublin, Poland, e-mail: pawel.olszewski@netrix.com.pl;*
[2]*WSEI University, Lublin, Poland, e-mail: Leszek.Gil@wsei.lublin.pl;*
[3]*WSEI University, Lublin, Poland, e-mail: Michal.Jasienski@wsei.lublin.pl;*
[4]*WSEI University, Lublin, Poland, e-mail: Natalia.Rak@wsei.lublin.pl;*
[5]*Wyższa Szkoła Biznesu - National Louis University, e-mail: mjasienski@wsb-nlu.edu.pl;*

1. Introduction

In today's rapidly growing business competition, effective supply chain management has become a critical success factor for many companies. One of the essential elements of this management is the precise prediction of order lead times (Kadłubek *et al.*, 2022). In this context, constructing regression and classification models becomes extremely important (Tyagi *et al.*, 2023).

Regression modeling allows you to forecast numerical values, such as order lead time. With these models, companies can analyze the impact of various factors on delivery times and make decisions to optimize logistics processes (Zampeta and Chondrokoukis, 2023). On the other hand, classification models allow new cases to be assigned to specific categories, which can be particularly useful in identifying potential delays in order fulfillment.

We will examine various modeling techniques, key factors influencing delivery times, and opportunities to optimize logistics processes using these models. In addition, we will provide practical examples of how these techniques can be applied to various industries to illustrate their potential benefits and challenges.

Knowledge of building regression and classification models to predict order lead time is essential for companies striving to manage their supply chains effectively. I hope that this article will be helpful to anyone interested in data analysis and logistics.

CLASSIFICATION (Loh, 2011):
The goal is to predict whether a customer will buy a particular commodity, taking into account the customer's information, average monthly spend, and available data about the commodity's buyers. This helps businesses target their marketing efforts and increase profits.

REGRESSION (Sarstedt, 2014):
The goal is to predict average monthly customer spending based on information about the customer, the buyer of a given item, and available data on average monthly expenditure. First, we will perform the necessary cleaning (removing duplicates and checking for missing values) and data visualization.

2. Building Regression Models

A dataset containing 1,595,196 observations was prepared for the study. The following variables are present in the set:

- numberOrderDoc – number of the purchase order document
- Merchandise – item identifier
- OrderDate

- issue date
- SaleDate
- issued number – the number of the goods issued document
- SalesDoc number – the number of the sales document
- Recipient Country
-Quantity
- PLN value – the value of the transaction expressed in PLN
- CommodityGroup – the commodity group to which the commodity belongs
- CountrySuppliers

The table presents selected observations of the set (Table 1).

Table 1. *A few observations of the set*

No.	numberOrderDoc	Merchandise	OrderDate	issue date	DateSales
0	ZS-64225/18/ICK	83096	2018-07-19	2018-07-25	2018-07-25
1	ZS-74797/18/ICK	132394	2018-08-16	2018-09-26	2018-09-26
2	ZS-1815/17/FE	121748	2017-05-25	2017-06-05	2017-06-05
3	ZS-53571/17/ICK	1083	2017-07-04	2017-07-12	2017-07-12
4	ZS-67208/17/ICK	118822	2017-08-16	2017-08-17	2017-08-17

No.	issue number	Recipient Country	Quantity	Value PLN	Commodity Group	Country Suppliers
0	FS-10638/18/FH	PL	3	179.34	10 Wieszaki	CN
1	FS-10378/18/EL	PL	1	326.60	09 Meble	CN
2	FS-1322/17/FE	CZ	1	75.00	06 Ławy	CN
3	FS-16428/17/FH	PL	2	250.80	17 Fotele	CN
4	FS-19586/17/FH	Pl	1	110.70	06 Ławy	CN

Source: *Own creation.*

The data set does not contain missing data or duplicates. However, the same order document number may appear in the database several times, with differences in goods, values, etc. This occurs when the available part of the goods is released first, and the other goods are released later after delivery.

Based on the repetitive numbers of the order documents, a partial binary variable was created containing information about the partial execution of the order:

- 1 – a given order is partially fulfilled (in several suborders)
- 0 – the order is fully processed (in full)

A variable has also been created to specify the lead time of an order called the number of lead days (as the difference between the Sale Date and the Order Date).

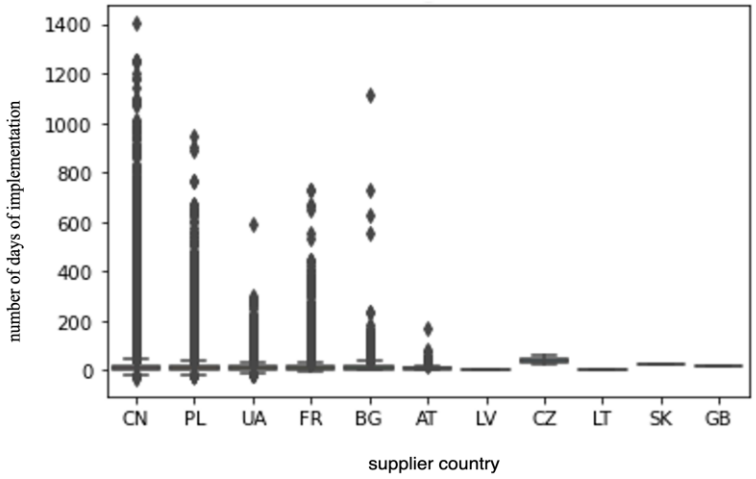
Table 2. *Added variables named partial and number of days of implementation.*

No.	numberOrderDoc	partial	DateSales	OrderDate	number of days of implementation
0	ZS-86185/20/ICK	1	2020-05-28	2020-05-20	8
1	ZS-94898/20/ICK	1	2020-08-28	2020-06-02	78
2	ZS-46995/20/ICK	1	2020-09-02	2020-08-31	2
3	ZS-466705/20/ICK	1	2020-09-04	2020-08-27	8
4	ZS-108975/20/ICK	0	2020-09-03	2020-06-25	70

Source: *Own creation.*

Observing the box plot of the number of lead days broken down by the supplier's country (Figure 1), you can see outliers in the number of days of order fulfillment (there are values even above 1000 days and negative ones). For further analysis, observations for which the lead time was non-negative or for which the lead time exceeded 365 days were selected. This results in a table of size (1591471, 16).

Figure 1. *Box plots of the number of lead days by Supplier Country*



Source: *Own creation.*

Table 3. *Statistics for lead times (after removing outliers)*

Name	Number of days of implementation
Mean	18.235016
Std	30.586949
Min	0.00
25%	3.00
50%	7.00
75%	18.00
max	365.00

Source: *Own creation.*

We have 81 commodity groups.
Regression – predicting the number of days an order will be fulfilled.
The predictors are the following features: categorical – Recipient Country, Product Group, Supplier Country, and numerical – Quantity, PLN Value.

Table 4. *Predictors of regression models*

No.	Quantity	ValuePLN	Recipient Country	Commodity Group	Country Suppliers
0	3	179.34	PL	10 Wieszaki	CN
1	1	326.60	PL	09 Meble	CN
2	1	75.00	CZ	06 Ławy	CN
3	2	250.80	PL	17 Fotele	CN
4	1	110.70	Pl	06 Ławy	CN

Source: *Own creation.*

For categorical variables (Rachwał, 2023), it is necessary to perform a one-hot encoding transformation.

Figure 2. *Dataset after One-hot encoding*

	Ilość	WartoscPLN	KrajOdbiorcy_AT	KrajOdbiorcy_BA	KrajOdbiorcy_BE	KrajOdbiorcy_BG	KrajOdbiorcy_BY	KrajOdbiorcy_CA	KrajOdbiorcy_CH	K
0	3.0	179.34	0	0	0	0	0	0	0	
1	1.0	326.60	0	0	0	0	0	0	0	
2	1.0	75.00	0	0	0	0	0	0	0	
3	2.0	250.80	0	0	0	0	0	0	0	
4	1.0	110.70	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...
1595191	1.0	359.47	0	0	0	0	0	0	0	
1595192	2.0	682.26	0	0	0	0	0	0	0	
1595193	3.0	493.92	0	0	0	0	0	0	0	
1595194	1.0	149.94	0	0	0	0	0	0	0	
1595195	6.0	513.24	0	0	0	0	0	0	0	

1591471 rows × 136 columns

Source: *Own creation.*

The dummy variables collection has 136 columns.

Table 5. *Results of Regression Models*

Model	RMSE	R^2
Linear Regression	29.97597	0.03817
Random Forest Regressor	25.054138	0.328091
K Nearest Neighbors Regressor(Yuan 2019)	28.969907	0.3036981
XGBoost Regressor	25.3531875	0.31195531

Source: *Own creation.*

The best results were achieved by a random forest model(Schonlau 2020, Cutler 2012), with an RMSE of about 25, where the execution days ranged from 0 to 365).
Classification is the order broken down into subcontracts.
The predictors have the same characteristics as for regression.
The XGBoost (Kozłowski 2021) classifier was used for classification.
The model was also built on data from the last two years (2020-2021).
Grouping of items according to the time of order fulfilment.

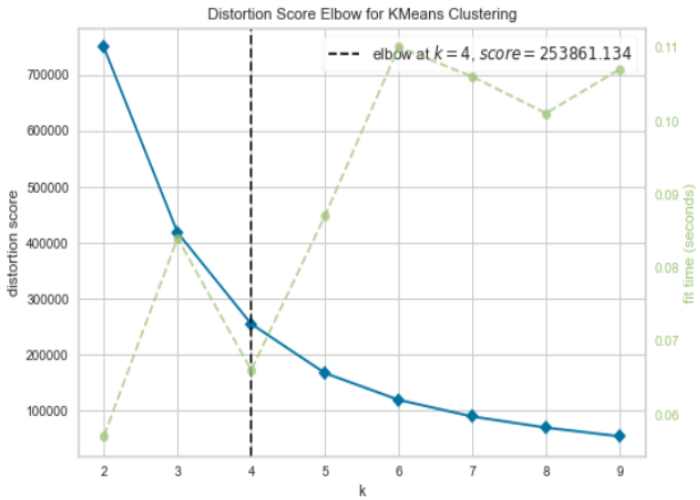
Table 6. *Average lead time for individual items*

No.	product	number of days of implementation
0	715	13.356157
1	743	10.175573
2	778	17.375
3	793	10.55
4	795	11.947977
…		…
5844	144314	46
5845	144315	46
5846	144389	14
5847	144397	7.145455
5848	144426	12

Source: *Own creation.*

The data set has been aggregated: for each item ID, the average lead time has been calculated. With the help of the K Means method, the goods were grouped into clusters with similar lead times.

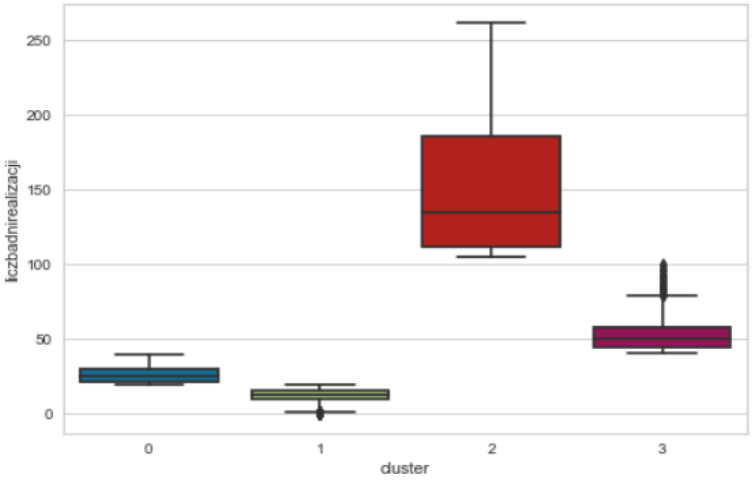
Figure 3. *Selection of the optimal number of clusters (elbow criterion)*



Source: *Own creation.*

Based on the elbow method, the optimal number of clusters is 4.
Sizes of designated clusters:
Cluster No. 1 is the group with the lowest lead time. Next up is cluster 0. Cluster 3 has a lead time of about 50 days (on average). Cluster No. 2 is characterized by the most incredible diversity of lead times. For this cluster, they exceed 100, and the median is close to 150.

Figure 4. *Box plots of order lead times by group*



Source: *Own creation.*

The exact statistics are presented in the table below:

Table 7. *Descriptive statistics in groups*

	number of days of implementation				
cluster	min	mean	median	std	max
0	19.4167	26.1824	24.8370	5.4161	40.0000
1	0.0000	12.5611	13.000	4.3505	19.4146
2	105.0000	150.2290	134.1850	45.7217	261.0000
3	40.0706	53.7360	50.1667	12.8620	100.1078

Source: *Own creation.*

Table 8. *Results of XGBoost regression models implemented separately in each cluster*

Nr klastra	Zakres liczby dni	RMSE	R^2
0	19.42 – 40	30.2641346593569	0.29031599427756805
1	0 – 19.41	19.959052320469777	0.34683860608959216
2	105 – 261	60.67251480476456	0. 4384349629301073
3	40.07 – 100.11	45.20584970616358	0.25682177148802054

Source: *Own creation.*

3. Regression Models of the Number of Lead Days by Different Countries, Suppliers, and Customers

Volume and weight are attached to the predictors, and quantity was removed. Thus, the independent variables include Commodity Group, Volume, and Weight.

Variables have been standardized. Regression models were performed on subsets selected based on the recipient's country and the supplier's country. The following groups of recipients have been distinguished according to their geographical location.

Audience:
1. Poland
2. Northern European countries

'LV' (Latvia), 'LT' (Lithuania), 'SE' (Sweden), 'NO' (Norway), 'EE' (Estonia), 'FI' (Finland), 'GB' (Great Britannia), 'DK' (Denmark), 'IE' (Ireland),

3. Western European countries

'NL' (Netherlands),'BE' (Belgium),'DE' (Germany),'FR' (France) ,'CH' (Sweden),'AT' (Austria)

4. Southern Europe

'HU','RO','BG','CZ','SK','SI','GR','MK','HR','MD','XK','BA','ME','AL','IT','ES','ML'

5. Eastern countries 'BY' (Belarus),'RU' (Russia),'UA' (Ukrainian),'KZ' (Kazakhstan)
6. America (but only one observation) 'CA'
7. Australia 'NC' (New Caledonia)
8. Africa ‘MU’ (Mauritius), ‘GM’ (Gambia), ‘GA’ (Gabon)

Supplier Country is China

***Table* 9.** *Results for: CountrySuppliers CN-China, CountryReceive: PL-Poland*

Model	RMSE	R2
Regression	33.25469318950208	0.017530631009539444
Random Forest	27.05337157926618	0.3497865879158082
KNN	28.006894240147346	0.30314402019596953
XGBoost	27.487197392746545	0.3287657457902394

Source: *Own creation.*

Table 10. *Results for: CN-China Supplier Country, Recipient Country: 'FI','SE','NO','DK','EE','LT','LV','IE','GB' Recipients are Northern European countries.*

Model	RMSE	R2
Random Forest	24.497520495953513	0.4400822253863216
XGBoost	24.48421421253843	0.4406903573340982

Source: *Own creation.*

Table 11. *Results for: CountrySuppliers PL, CountryRecipient: PL-Poland*

Model	RMSE	R2
Random Forest	18.442530335633492	0.4189205778939028
XGBoost	18.40840693272455	0.42106884815166556

Source: *Own creation.*

Table 12. *Results for: CN Supplier Country, Customer Country 'FR', 'BE', 'NL', 'AT', 'DK', 'CH' Western European countries.*

Model	RMSE	R2
Random Forest	23.6624988739506	0.25187186095857816
XGBoost	23.68957285002003	0.2501589296496408

Source: *Own creation.*

Table 13. *Results for: CN Supplier Country, Recipient Country: 'HU','RO','BG','CZ','SK','SI','GR','MK','HR','MD','XK','BA','ME','AL.','IT','ES','ML' Southern European countries.*

Model	RMSE	R2
Random Forest	24.6665535521045	0.3239139621762003
XGBoost	24.774328123773106	0.31799304973129716

Source: *Own creation.*

Table 14. *Supplier Country: CN, Recipient Country: Canada There is only one observation. Results for Supplier Country: CN, Recipient Country: NC Countries from Australia*

Model	RMSE	R2
Regression	0.0	1.0
Random Forest	0.0	1.0

Source: *Own creation.*

Table 15. *Results for Supplier Country: CN, Recipient Country: 'BY','RO','UA','KZ' To East*

Model	RMSE	R2
Random Forest	29.684972948038023	0.15888846525783062
XGBoost	29.465688633402497	0.17126922015857626

Source: *Own creation.*

Table 16. *Results for Supplier Country: CN, Recipient Country: GM, GA, MU To Africa*

Model	RMSE	R2
Random Forest	30.183723056976365	0.17968627094744638
XGBoost	34.18249006367331	-0.052063137004950244

Source: *Own creation.*

Regression and classification models were also built on the subset before 2020 (before the pandemic). Unfortunately, this did not significantly improve the models' fit to the data.

4. Conclusions, Proposals, Recommendations

Using classification and regression models to predict order fulfillment times and identify order delays in the furniture industry brings numerous benefits to companies. Firstly, determining the delivery delay threshold based on company policy allows quick response to potential logistics problems, increasing operational process efficiency.

Thanks to classification models, it is possible to quickly and automatically recognize orders at risk of delay, allowing the company to focus on preventive actions. Using regression models to determine the forecast delivery time for delayed orders allows you to better manage customer expectations and minimize the impact of delays on the entire supply chain.

Thanks to precise forecasts of order fulfillment times, the company can make informed decisions regarding resource allocation, production planning, and logistics, contributing to increased operational efficiency.

Moreover, using predictive models in the order management process allows for continuous improvement of logistics processes through analyzing historical data and identifying trends. Thanks to this, the company can better understand the causes of delays and make appropriate corrections to minimize the risk of their occurrence in the future.

As a result, the use of classification and regression models in the furniture industry contributes to increasing customer satisfaction through timely execution of orders, reduction of costs related to delays, and effective management of company resources. The company can achieve a competitive advantage in the market by offering high-quality delivery services and building lasting relationships with customers.

References:

Cutler, A., Cutler, D.R., Stevens, J.R. 2012. Random Forests. In: Zhang, C., Ma, Y. (eds). Ensemble Machine Learning. Springer, New York, NY.

Grima, S., Spiteri, V.J., Thalassinos, E.I. 2020. Risk management models and theories. Frontiers in Applied Mathematics and Statistics.

Kadłubek, M., Thalassinos, E.I., Domagała, J., Grabowska, S., Saniuk, S. 2022. Intelligent transportation system applications and logistics resources for logistics customer service in road freight transport enterprises. Energies, 15(13), 4668.

Kozłowski, E., Borucka A., Świderski A., Skoczyński A. 2021. Classification Trees in the Assessment of the Road–Railway Accidents Mortality. Energies 14, 12, 3462. https://doi.org/10.3390/en14123462.

Loh, W. 2011. Classification and regression trees. WIREs Data Min. Knowl. Discov., t. 1, nr 1, s. 14-23.

Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., Rymarczyk, T. 2023. Determining the Quality of a Dataset in Clustering Terms. Applied Sciences, vol. 13, nr 5, 1-20.

Schonlau, M., Zou, R.Y. 2020. The random forest algorithm for statistical learning. Stata J., 20, 3-29.

Sarstedt, M., Mooi, E. 2014 Regression Analysis: A Concise Guide to Market Research. Berlin, Heidelberg: Springer Berlin Heidelberg, 193-233.

Tyagi, P., Grima, S., Sood, K., Balamurugan, B., Özen, E., Thalassinos, E.I. (Eds.). 2023. Smart analytics, artificial intelligence and sustainable performance management in a global digitalised economy. Emerald Publishing Limited.

Yuan, C., Yang, H. 2019. Research on K-Value Selection Method of K-Means. Clustering Algorithm J, t.2, nr 2, 226-235.

Zampeta, V., Chondrokoukis, G. 2023. A Comprehensive Approach through Robust Regression and Gaussian/Mixed-Markov Graphical Models on the Example of Maritime Transportation Accidents: Evidence from a Listed-in-NYSE Shipping Company. Journal of Risk and Financial Management, 16(3), 183.