On Cost Estimation of the Full Truckload Contracts

Szymon Cyperski [1], Michał Okulewicz [1], and Paweł D. Domański [1,2,]*

[1] Control System Software Sp. z o.o., ul. Rzemieślnicza 7, 81-855 Sopot, Poland
[2] Institute of Control and Computation Engineering, Warsaw University of Technology, Nowowiejska, Warsaw, Poland
Email: scyperski@betacom.com.pl (S.C.); mokulewicz@betacom.com.pl (M.O.); pawel.domanski@pw.edu.pl (P.D.D.)
*Corresponding author

***Abstract*—Today's global economy can operate efficiently due to timely and cost-effective delivery of goods. There are a huge number of shipping companies on the market, whose sole purpose is to transport the ordered goods. Road transportation can be carried out using the company's own fleet of trucks or using third-party companies. This work focuses on Full Truckload (FTL) transportation. The shipper must be aware of the potential costs of a given service during the process of selecting the right service provider. Our solution analyzes the cost estimation of the FTL shipping. Market offers many approaches, from detailed calculators to solutions using various Artificial Intelligence (AI) solutions. This study compares hybrid solutions that combine different Machine Learning (ML) techniques. The solution is tested on real data covering multi-year contracts of several freight forwarding companies operating in the European FTL market. The results obtained are implemented in a commercial solution used by freight forwarding companies daily.**

***Keywords*—regression, clustering, cost estimation, Full Truckload (FTL), k-Nearest Neighbors (k-NN), eXtreme Gradient Boosting (XGBoost), Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**

I. Introduction

The Full Truck Load (FTL) shipping services constitute the main variant of the road long range transportation. It well works for large volumes, where a load fills the whole truck space. There is an alternative approach, i.e., the Less than Truckload (LTL), in which a truck takes partial loads to/from different locations during one travel. Presented paper focuses on the FTL approach.

The FTL external fleet cost estimation is very important, as a lot of shipment is done in this way and an operator should know if a given order can be approved using some objective measures, not only his experience.

The FTL shipping is characterized by lower damage risk, as goods stay on the same truck all the time from loading to the final unloading. It is easier to keep the load secure and the whole shipment is more reliable. Due to the lack of the loading stops the delivery process is faster minimizing the unexpected delays probability. Thus, it's easier to guarantee the delivery. The FTL shipping is realized using the own fleet, the trucks from regular suppliers, or can utilize the truck services offered by the external companies. Each approaches applies different shipping cost model. The realization by the own fleet is the simplest in the cost estimation, as the owner just knows exactly all the costs, which include the shipping time, the fuel cost, according road and transportation fees (intermodal costs, ferries, road fees, etc.), driver rates, truck fixed costs, taxation, company overheads and others. Regular partners, which are often called as the leased carriers are directly connected and operate according to known contract rules defining the aggregate transportation cost on a given route. Leased carriers might use the dynamic pricing, but with a different model. External third party subcontractors operate according to the current market situation, their own policies, habits and specialization. The price knowledge that can be put on the table for a given shipment allows to make decision to accept or to reject the contract. The FTL cost estimation determines crucial part of the business.

Despite its importance, this is not an easy task. The process data do not consist independent records. They are seriously affected by the external economic environment, and what is much more challenging, by human interventions and erroneous entries. Thus, the methodology should incorporate the knowledge about any affecting impacts customizing the model accordingly.

This paper continues works that use hybrid approach mixing Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with regression tree eXtreme Gradient Boosting (XGBoost) and the k-Nearest Neighbors (k-NN) estimation [1]. Proposed approach uses the process knowledge about the shipping and is not just another black-box model.

We assume that the offer of the contract price depends on some unknown dynamic pricing model [2] from an external shipping service provider. External fleet long route pricing is generally solved with the use of popular cost calculators [3], or using estimation. Simple linear estimation originating from general least squares assumptions does not applies as the data are highly contaminated with outlying observations. Robust regressions estimators are then required [4]. Researchers

Manuscript received January 16, 2024; revised February 6, 2024; accepted March 1, 2024; published May 16, 2024.

doi: 10.12720/jait.15.5.614-621

simply overcome this issue using Artificial Intelligence (AI) and Machine Learning (ML) methods [5–7].

One has to be aware that ML solutions are not just out-of-the-shelf universal estimators that can be fed with any data to produce a brilliant result. Such an approach often ends up with a simple result: *garbage in, garbage out*.

The general research in other contexts meets the same challenge [8, 9]. Input data must be carefully assessed. The methods has to be cautiously used and parametrized. Only then one may expect to get reasonable results.

As the FTL external fleet shipping cost estimation brings money to the shipment operator, and simultaneously mitigates his risk, it is worth to be considered. Moreover, it is always worth to embed the process knowledge into the process to minimize blind methods. This paper addresses this issue bringing decision support to the shipment operator.

The paper introduces the novel, hybrid methodology to estimate the cost of the external fleet FTL shipping. The method, as any, has some limitations. In this case these are the short routes, for which the dynamic pricing is less dependent on objective rules. Section II presents the problem and the utilized methods. Next, the real case study is presented in Section III with the results and their analysis. It is followed by the conclusions and the opportunities for further research described in Section IV.

II. Full Truckload Shipment

Proposed algorithm uses hybrid approach to incorporate into the solution as much as possible of the available process knowledge. It is an intentional approach to exclude fully black-box solutions, which are nontransparent, inflexible and their use requires enormous calculation power. The hybrid approach allows estimation procedure decomposition to smaller tasks, which can be separately assessed and maintained.

Developed algorithm uses three ML approaches: a two-dimensional DBSCAN clustering, XGBoost and k-Nearest Neighbors estimations. Two dimensional grouping aims at diminishing the number of pickup and unloading locations (exchanged by clusters). The DBSCAN algorithm is selected as it delivers reasonable clusters. The k-NN brings flexibility to incorporate the knowledge and it allows self-adaptation. The XGBoost increases the efficiency in case of short range routes.

To confront the clustering based approach with other options, there are performed two alternative tests: manual clusters and the non-clustered solution. These tests are applied as in practical implementation validation and confirmation of the DBSCAN clustering may be hard for the end-user and may strongly depend on the existing contracts. Manual clustering or dynamic assignment does not require any specific knowledge from the end user.

The paper compares three FTL cost estimation methods differing in the clustering: DBSCAN, manual and dynamic assignment. Each algorithm is tested in its original form and with an improvement by the XGBoost. The algorithm workflow is shown in Fig. 1.

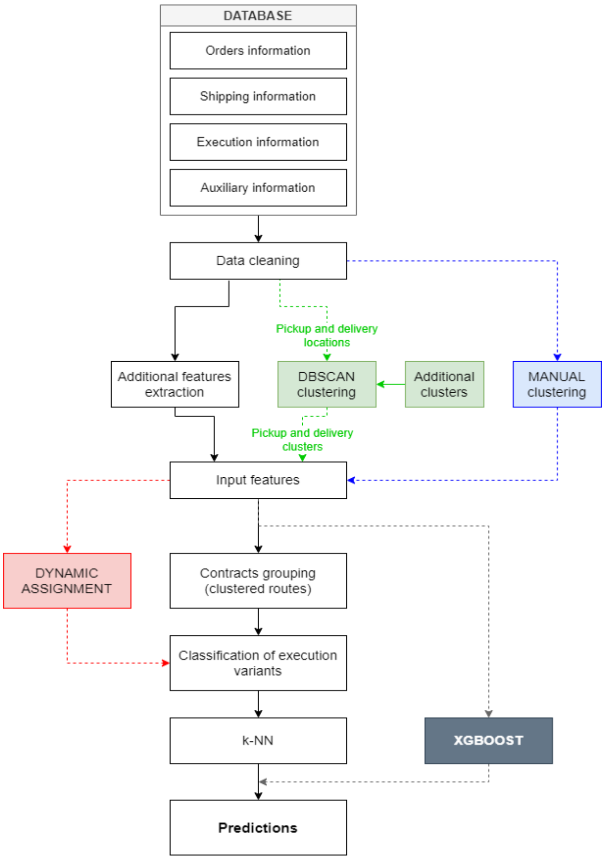

Fig. 1. The FTL estimation algorithm workflow.

A. *The DBSCAN Clustering*

There are many clustering algorithms, which differ by data processing, application, and the cluster concept definition [10]. The DBSCAN belongs to the class of density algorithms [11]. It enables to find groups of complex shapes, not only spherical or to find clusters surrounded by the others. It works well with noisy data and the algorithm finds the number of clusters itself.

DBSCAN has two input parameters. The first is the minimum number of points required by a cluster n and the maximum radius of the neighborhood ε.

The algorithm reviews all the unclassified points. For each of them, the method checks the number of all points in its neighborhood. If the number of such points is less than the n parameter, the point is temporarily marked as the noise and the algorithm takes another one. Otherwise, a totally new group, which initially consists all points from the surroundings of the current center is created.

For each point belonging to the seed, the number of neighbors is checked. If this number is greater than or equal to n all previously not visited points are added to the seed and the cluster expands. If in a neighborhood there are noise points they are added to the cluster. Its expansion ends when all points are examined. The detailed description can be found in [11]. In the described implementation it is used in a spatial context. It also takes into account geographical borders into the algorithm [12].

B. Manual Clustering

Manual clustering is done in a simple way using manual determination of the geographical centers, mostly of the European administrative regions.

C. The k-NN Estimation

The k-NN algorithm is an example of a memory-based approach, which, unlike other statistical methods, does not require learning as such. The algorithm uses the idea of prototypes, which assumes that similar objects are in the same class. Prediction of class membership of a new object is therefore based on a comparison with a set of prototype objects. In classification, the voting of the nearest *k* neighbors is used, while in regression (considered case), average estimators are used.

The k-NN estimation and prediction methods does not require any learning as such, which increases its attractiveness and popularity [13]. It is interesting that this approach has been already adopted into the shipping context [14], what supports the selection.

D. The XGBoost Regression

The Extreme Gradient Boosting or the gradient boosting algorithm seems to be one of the most popular data mining approaches [15]. In XGBoost, we deal with a set of classifiers, like the decision trees. The final decision depends on all the trees used by the algorithm. The algorithm uses an incremental strategy, as it is simpler than training all the trees simultaneously.

The change is in the use of a regularization element. Regularization is a kind of penalty put on the model for having too many final leaves in the decision tree, which controls model complexity. Thus, the general formulation of the algorithm has two parts. The first component is responsible the error minimization and is called the loss (cost) function. The regularization as the second element, prevent overtraining and holds the complexity.

The algorithm can be used for the classification task and for the regression. The last version is applied in the considered solution.

III. Estimation Case Study

Proposed algorithm uses custom hybrid approach to take into account as possible of the process custom knowledge, pricing best practices and all available data. The hybrid approach allows estimation procedure to be decomposed into smaller tasks, which can be assessed and maintained separately.

The data used during the project originate from the databases of the selected Polish transportation companies [16]. Original single contract record extracted from the production database includes many fields and features. Proposed approach tries to minimize the features' set considering only the most reasonable the knowledge about the process. Contract features are filtered to remove unimportant variables [1].

The final list of used features is presented below:

- TimeTillNow (the measure of a contract time),
- isCooler (classifier for cargo cooler),
- FuelPrice (diesel fuel price),
- RouteLength,
- PaymentTerm,
- MinShippingTime and MaxShippingTime (shipping time window),
- LoadNo (number of pick-ups),
- UnloadNo (number of unloadings),
- CargoWeight,
- EmpyDist (distance to be travelled empty to pick up the cargo),
- TonnesKM,
- TimeDiff.

Apart from the above model inputs the following classifiers are used:

- /ANY → any existing contract for the current route
    - /INTERMODAL → intermodal shipping
    - /ROAD → road truck only shipping
        - /LEASED → leased carriers
            - /SPECIFIC → specific leased carrier
            - /OTHER → other leased carrier
        - /EXTERNAL → external carriers
            - /SPECIFIC → specific external carrier
            - /OTHER → other external carrier

The resulting estimation of the price **RouteEstimCost** is calculated as the weighted mean out of $k = 5$ nearest neighbors. All costs are re-scaled to the current moment using inflation rate coefficient. Moreover, the algorithm returns the minimum and maximum cost, as the cheapest and the most expensive Nearest Neighbor. The travel times are returned as the trimmed (trimming equal to 1) time out of the already chosen $k = 5$ Nearest Neighbors. Detailed description of the algorithm can be found in [1].

Original orders database consists of approximately 583,000 orders. After the pre-processing the 414,000 records remain, as the own fleet contracts and the erroneous data are removed. The data considered covers the time period from January 1st, 2016 to April 30th, 2022. These records are considered as the training data. Records from May 1st, 2022 till August 1st (15,000 orders), 2022 are used for validation.

A. Clustering

There are two reasons for the loading/unloading points clustering: one comes out of the process specifics, while the second origins from the estimation algorithm. The shipping cost depends on the route. The database reveals many close to each other pickup/unloading locations. Thus, the difference between two routes: from Bonn to Warsaw or from Legionowo to Cologne is neglectable. The exact locations differ, while the cost difference is almost none. It is just inefficient to keep exact locations, and it is reasonable to merge similar ones.

The second reason derives from the estimation method. Its efficiency depends on the number of neighbors. The more similar orders we have, it is easier to select the most relevant one. We observe dozens of the routes, starting and ending close to each other. Their clustering merges different orders and then the estimation has better examples to find appropriate neighbors for a route. It also allows to find the cost for a new locations, as long as they are located close to the already known ones.

Data clustering is performed in two steps. At first, main clusters are evaluated using classical DBSCAN procedure with the following parameters: $\varepsilon = 7$ km and $n = 100$. These clusters reflect real road infrastructure and locations follow the road patterns. Such clustering does not fill the entire map leaving uncovered. This subject is solved with final manual clustering. Uncovered regions are filled with the manually set, which address regions without order history, as shown in Fig. 2.

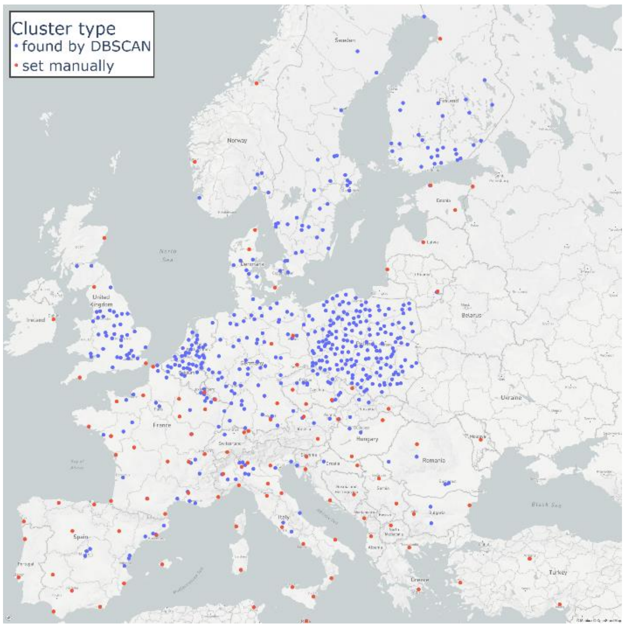

Fig. 2. Evaluated cluster centers locations: DBSCAN and manual ones.

In such a way the clustering is complete. Each estimated location can be assigned to the nearest cluster, as shown in Fig. 3.

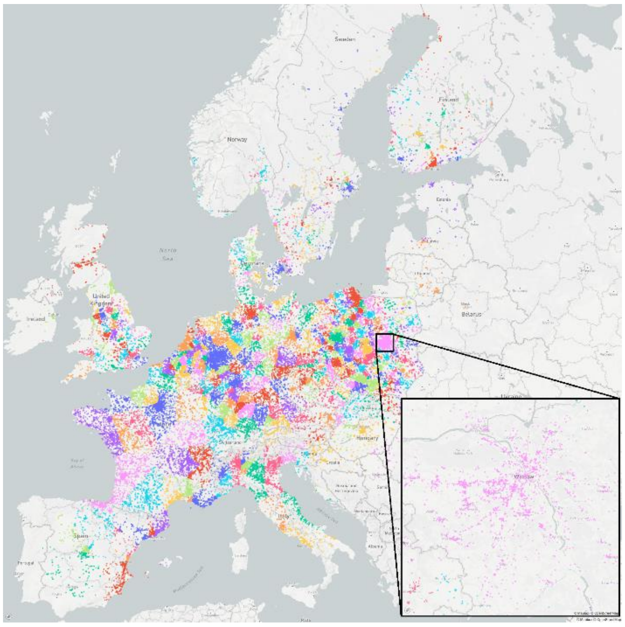

Fig. 3. Final clustering result with enlarged Warsaw region.

The DBSCAN clustering has an advantage, which during commercial operation might be considered as a shortcut. It is highly customized and requires good and long history of data. The DBSCAN clustering, though simple for researchers might cause hesitation for a shipping company personnel resulting in the abandoning the use of the solution. Such the result is observed in case of the advanced control solutions in process industry [17]. Short records history can significantly bias clustering, resulting in only the manual ones being left.

Above consideration lead to the simplification and the application just of the manual clustering. The results of the manual clustering are shown below. Fig. 4 shows the centers location while Fig. 5 presents the result of the manual clustering on the training data.

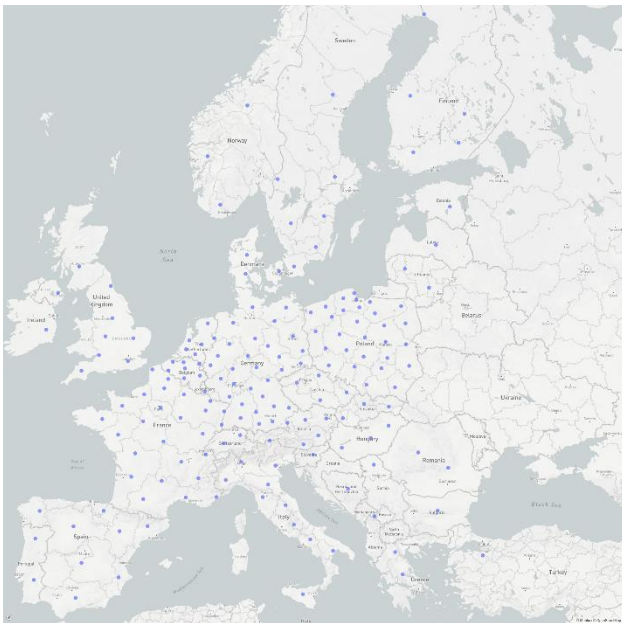

Fig. 4. Manual clustering: The centers.

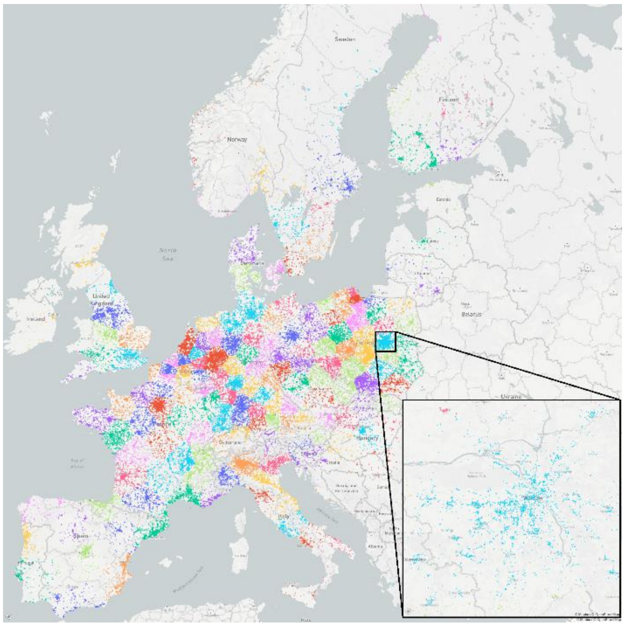

Fig. 5. Manual clustering: the location assignment with enlarged Warsaw region.

The resulting manual clustering assignment differs from the DBSCAN one. Especially, the number of clusters are minimized. Such an approach can be used for a fresh customer, without a significant history of contracts in.

Finally, the clustering step itself can be omitted and the neighboring location might be assigned dynamically. Such a procedure can be defined simply. For a current location we take some perimeter and search for the known locations in the database. If we find none, we increase the radius.

During the proposed procedure we use the Haversine formula [18] to calculate the distance. We use the following limits: {10,25,50,100,250,400} km. It has to be noted that the algorithm takes into consideration the natural geographical borders as for instance the Alps ridge or the English channel.

A. XGBoost Integration

The integration with the XGBoost follows the observation done during the initial analysis [1]. The k-NN estimation generally works better than the XGBoost alone. However, we observe that the XGBoost model favors low costs, while the higher costs routes are better estimated with the k-NN.

This observation leads to the integration concept done in a fuzzy-like way. Once there are not enough solutions, i.e., the neighbors for the k-NN the XGBoost is taken into account. We check the number of the historical contracts on the same route and for the same shipping variant. Once it is low, the XGBoost is introduced with its weight opposite to the k-NN one (see Figs. 6 and 7).

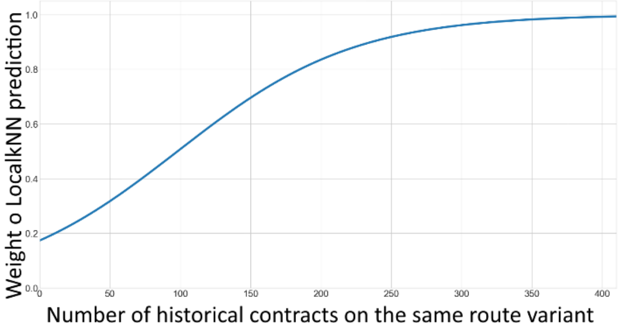

Fig. 6. The k-NN membership function for the same routes in the same variant.

Once the variant does not exists in data we take general relation, which is presented in Fig. 7.

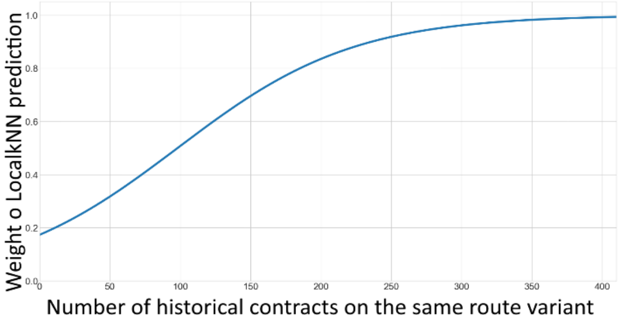

Fig. 7. The k-NN membership function for the same routes (independent of the variant).

B. Estimation Results

Once the clustering and XGBoost integration are properly set we may continue with the price estimation. To compare the solutions proper residuum analysis has to be applied, as its selection may influence the conclusions.

Three main integral indexes: Mean Square Error (MSE), mean Integral Absolute Error (IAE) and Mean Percentage Integral Absolute Error (MAPE) are used [19]. Apart of standard integral indexes, two statistical measures are also considered: normal standard deviation (stDev) and the robust estimator of standard deviation in form of the logistic psi-function estimator (RstDev) [4].

Nine models are evaluated, as explained in previous paragraphs:

1. kNN: pure k-NN model with DBSCAN clustering,
2. XG: pure XGBoost model,
3. kNN-XG: hybrid: DBSCAN and XGBoost,
4. fNN: k-NN model with fixed clusters,
5. fNN-XG: hybrid: fixed clusters and XGBoost,
6. dNN: k-NN model with dynamic clusters,
7. dNN-XG: hybrid: dynamic clusters and XGBoost,
8. meanModel: mean prediction of k-NN models,
9. meanModel-XG: mean prediction of hybrid models.

The residuum analysis of the obtained models starts from the calculation of the above performance indicators. They are shown in Table I. Red color indicates the worst models, while the green the best performing ones.

TABLE I. RESIDUUM ANALYSIS OF THE CONSIDERED MODELS

Models	MSE	IAE	MAPE	stDev	Rst
kNN	1111817	492.3	14.94	1053	385.9
XG	902040	456.2	16.18	944	400.1
kNN-XG	850713	425.3	14.42	921	358.2
fNN	1375971	516.1	16.57	1172	388.2
fNN-XG	762451	392.0	13.92	872	342.2
dNN	1333483	538.5	18.39	1153	385.0
dNN-XG	807891	409.0	13.25	896	344.0
meanModel	1169071	505.3	15.90	1080	391.4
meanModel-XG	903630	450.5	13.87	947	372.4

The analysis of the results evaluated using various performance indices answers several questions. We see that each index favors different model. Thus, it matters which one we select, as it may bias further decisions.

We observe that XGBoost model is never selected as the best one and even in case of one index (robust standard deviation) it is indicated as the worst one. The difference between normal standard deviation and its robust version depends on their perception of outliers, Normal standard deviation is highly affected by even the small number of outlying observations [20, 21]. Its robust estimator is robust against outliers. Therefore, we may say that XGBoost can be considered as the better in case of outlying observations failing in case of the main process representatives. However, once we combine the XGBoost estimator with any k-NN we obtain improved results. Such smart integrations improves k-NNs, which are quite good in case of the most common samples (from the perspective of the fundamental process behind the data) with the estimation of the outlying samples. These outlying samples are associated with the roots with a small number of historical contracts.

It's interesting to observe that generally the worst predictions are associated with the fixed clusters based k-NN, while introduction of the XGBoost to this model results in the best predictions. The fact that dynamic routes assignment integrated with the XGBoost (dNN-XG) gives the best results with the relative residuum measure (MAPE) means that this model behaves better with the short routes (lower absolute costs).

We should remember that simple index oriented analysis does not explain the nature of the model and the causes for the poor performance. The analysis of the residua histograms and their properties delivers further insight. Sample histograms for two models are shown below. Fig. 8 presents the histogram with fitted Probabilistic Density Functions (PDF) of normal and robust distributions for the fNN model. Fig. 9 shows similar diagrams for the fNN-XG model.

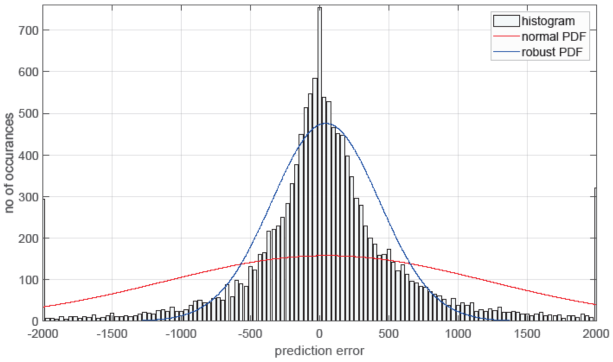

Fig. 8. Sample histogram plots with the fitted normal and robust Gaussian distributions for the fNN model.

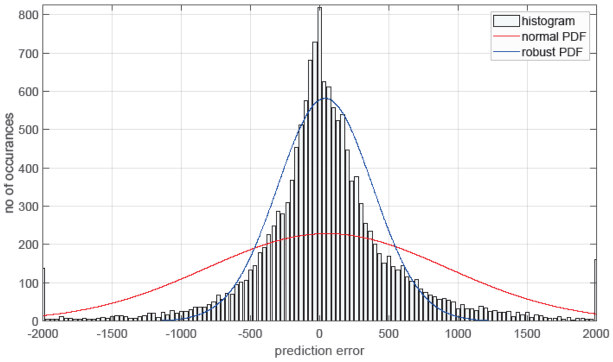

Fig. 9. Sample histogram plots with the fitted normal and robust Gaussian distributions for the fNN-XG model.

They present the best and the worst models, respectively. We observe that normal distribution is not appropriate approximation of the error stochastic properties, as it is significantly disturbed by the histogram tails (outliers). Robust estimator behaves more reliable. Therefore, normal mean and standard deviations should not be used. Following the above, the MSE should not be used, as it is equivalent to the standard deviation (the feature shown by Gauss). The IAE measure (absolute or percentage) and robust standard deviation are acceptable alternative measures of model quality. The mindless use of the MSE index is misleading and, hance should be used with the caution.

The comparison of all nine models is shown in Fig. 10, which presents all fitted robust Gaussian distributions in a single plot. We can better observe how the prediction error improves. The observation about non-Gaussian error properties, distribution fat tails and the related outlying results suggest further analysis.

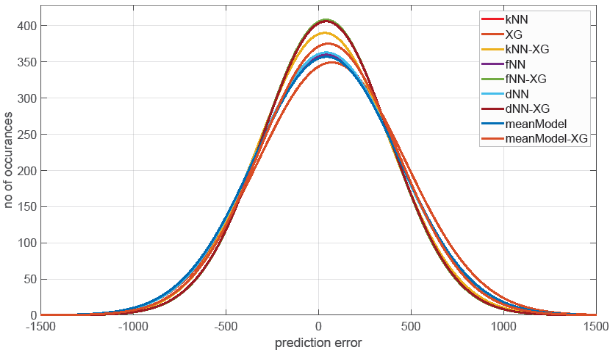

Fig. 10. Comparison of the robust Gaussian PDFs.

Fig. 11 shows the box-plot representation of the errors. In this way we can compare the models.

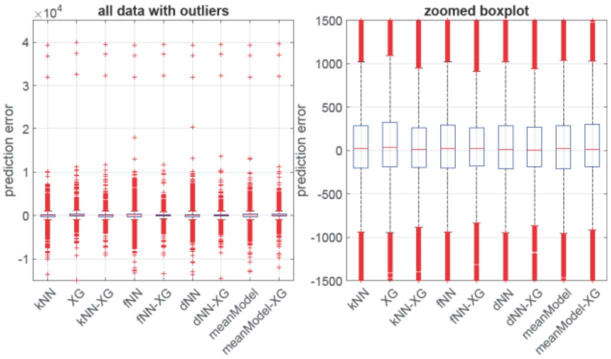

Fig. 11. Comparison of the models with the box plot.

The comparison of the IAE and MAPE suggest that the estimation quality may be related to the shipping cost. One of the ways to detect it is to analyze the predicted versus the real costs relationship. Fig. 12 shows this plot for four models: kNN-XG, fNN-XG, dNN-XG, and meanModel-XG.

This feature can be better visualized with the diagram shown in Fig. 13, which presents the relation between prediction error and the real shipping cost. This dependence is well visualized with the 3rd order polynomial fitting of the residuum to the real cost.

We observe the difference in performance between the kNN-XG, fNN-XG, dNN-XG, and meanModel-XG models. We can bring the hypothesis that the prediction quality depends on the number of historical examples along the route. Fig. 14 shows this relationship. The figure shows that the more often a given connection is used the more accurately we can estimate its cost.

All the above demands further understanding of the fundamental process behind the data and the task, deep understanding of human influence, and an attempt to develop a solution for routes that are short, infrequent or have not been used recently. The case of the short and/or infrequent routes is still challenging and requires further attention with dedicated approaches.

Observation of the results enables not only to indicate prediction effectiveness. The fact that a particular estimators is better or worse according to some index is insufficient. One should try to detect the reasons for such results and to perform a kind of root-cause analysis [22].

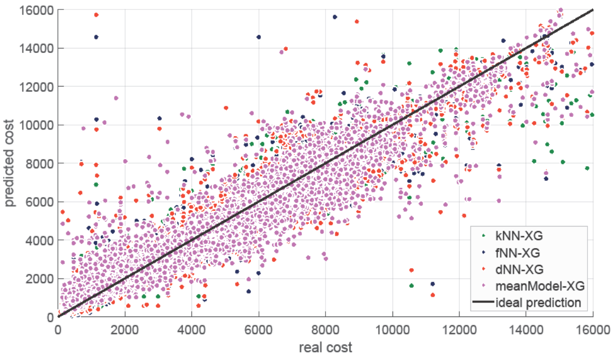

Fig. 12. Predicted versus real cost for selected models: kNN-XG, fNN-XG, dNN-XG and meanModel-XG.

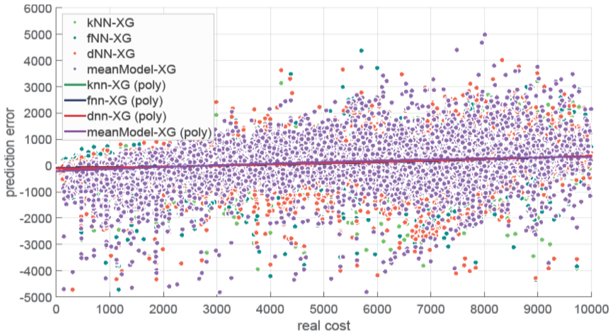

Fig. 13. The relationship between the model quality (prediction error) and the route shipping cost.

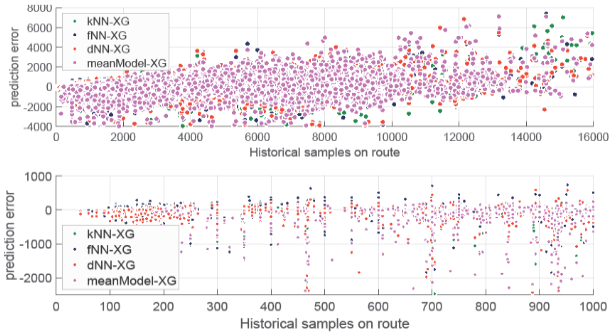

Fig. 14. The relationship between the number of the historical samples on route and its actual shipping cost.

IV. Conclusions

This paper presents a proposal for the novel cost modeling approach for the FTL contracts in case of the dynamic pricing of the third-party transportation companies. This approach uses a combination of various clustering approaches, the k-Nearest Neighbors modeling and the Extreme Gradient Boosting estimation.

In the presented example, the non-Gaussian nature of the phenomenon and the distribution of errors is demonstrated, which should preclude the use of the MSE index or the Gaussian normal PDF. Considered logistics processes are associated with a lot of outliers resulting in fat tails.

It is shown that the quality of the model depends on the cost of the route. The short routes cost estimation is much more demanding and requires customized specific solutions. Apart of that, the cost estimation of infrequently used routes is a much more challenging than the analogous task for frequent connections.

Presented models prove their quality not only in paper, but are practically used in commercial solutions for freight forwarding companies.

Realization of this project suggests open issues. One is to better understand real shipping properties behind the data. The second is to find a way of fitting the appropriate estimation, as the universal method does not exist.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

SC conducted the research; SC and PDD analyzed the data; PDD wrote the paper; MO supervised the work; all authors had approved the final version.

Funding

Research is supported by the Polish National Centre for Research and Development, grant no. POIR.01.01.01-00-2050/20, application track 6/1.1.1/2020-2nd round.

References

[1] S. Cyperski, P. D. Domański, and M. Okulewicz, "Hybrid approach to the cost estimation of external-fleet full truckload contracts," *Algorithms*, vol. 16, no. 8, 360, 2023.

[2] K. Stasiński, "A literature review on dynamic pricing—State of current research and new directions," in *Advances in Computational Collective Intelligence*, M. Hernes, K. Wojtkiewicz, E. Szczerbicki, Eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 465–477.

[3] Freightfinders GmbH. Freight Cost Calculator. (2023). [Online]. Available: https://freightfinders.com/calculating-transport-costs/

[4] P. Huber and E. Ronchetti, *Robust Statistics*, Wiley Series in Probability and Statistics, 2011.

[5] K. Tsolaki, T. Vafeiadis, A. Nizamis, D. Ioannidis, and D. Tzovaras, "Utilizing machine learning on freight transportation and logistics applications: A review," *ICT Express*, vol. 9, pp. 284–295, 2022.

[6] Ł. Podlodowski and M. Kozłowski, "Predicting the costs of forwarding contracts using XGBoost and a deep neural network," in *Proc. the 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Sofia, Bulgaria, 2022, pp. 425–429.

[7] S. Kaźmierczak, "Prediction of the costs of forwarding contracts with machine learning methods," in *Proc. the 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Sofia, Bulgaria, 2022, pp. 413–416.

[8] A. Alshammari, R. Almalki, and R. Alshammari, "Developing a predictive model of predicting appointment no show by using machine learning algorithms," *Journal of Advances in Information Technology*, vol. 12, no. 3, pp. 234–239, 2021.

[9] W. Kusonkhum, K. Srinavin, N. Leungbootnak, P. Aksorn, and T. Chaitongrat, "Government construction project budget prediction

using machine learning," *Journal of Advances in Information Technology*, vol. 13, no. 1, pp. 29–35, 2022.

[10] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 100, 104743, 2022.

[11] M. Ester, H. P Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. the Second International Conference on Knowledge Discovery and Data Mining*, Palo Alto, CA, USA: AAAI Press, 1996, pp. 226–231.

[12] Q. Du, Z. Dong, C. Huang, and F. Ren, "Density-based clustering with geographical background constraints using a semantic expression model," *ISPRS International Journal of Geo-Information*, vol. 5, 72, 2016.

[13] P. D. Domański and M. Więcławski, "Memory-based prediction of district heating temperature using GPGPU," in *Progress in Automation, Robotics and Measuring Techniques*, R. Szewczyk, C. Zieliński, M. Kaliczyńska, Eds. Cham, Switzerland: Springer International Publishing, vol. 350, 2015, pp. 33–42.

[14] M. A. Mohammed, M. K. A. Ghani, R. I. Hamed, S. A. Mostafa, D. A. Ibrahim, H. K. Jameel, and A. H. Alallah, "Solving vehicle routing problem by using improved k-nearest neighbor algorithm for best solution," *Journal of Computational Science*, vol. 21, pp. 232–240, 2017.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794.

[16] A. Janusz, A. Jamiołkowski, and M. Okulewicz, "Predicting the costs of forwarding contracts: Analysis of data mining competition results," in *Proc. the 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Sofia, Bulgaria, 2022, pp. 399–402.

[17] J. F. Smuts and A. Hussey, "Requirements for successfully implementing and sustaining advanced control applications," in *Proc. the 54th ISA POWID Symposium*, 2011, pp. 89–105.

[18] H. B. Goodwin, "The haversine in nautical astronomy," *Naval Institute Proceedings*, vol. 36, no. 3, pp. 735–746, 1910.

[19] P. D. Domański, *Control Performance Assessment: Theoretical Analyses and Industrial Practice*, Springer International Publishing, Cham, Switzerland, 2020.

[20] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York, NY, USA: John Wiley & Sons, Inc., 1987.

[21] P. D. Domański, Y. Chen, and M. Ławryńczuk, *Outliers in Control Engineering: Fractional Calculus Perspective*, De Gruyter, 2022.

[22] M. J. Falkowski and P. D. Domański, "Causality analysis with different probabilistic distributions using transfer entropy," *Applied Sciences*, vol. 13, no. 10, 5849, 2023