Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 399–402

DOI: 10.15439/2022F303
ISSN 2300-5963 ACSIS, Vol. 30

Predicting the Costs of Forwarding Contracts: Analysis of Data Mining Competition Results

Andrzej Janusz[*†], Antoni Jamiołkowski[†], Michał Okulewicz[‡§]
[*]Institute of Informatics, University of Warsaw, Warsaw, Poland
[†]QED Software, Warsaw, Poland
[‡]Control System Software, Sopot, Poland
[§]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

***Abstract*—We discuss the international competition *FedCSIS 2022 Challenge: Predicting the Costs of Forwarding Contracts* that was organized in association with the FedCSIS conference series at the KnowledgePit platform. We explain the scope and outline the results obtained by the most successful teams.**

***Index Terms*—Data mining competitions; costs of forwarding contracts; KnowledgePit platform**

I. Introduction

TRANSPORTATION and logistics are among the most influential sectors in the global economy. It is their capacity that determines whether the commodities will reach the end customers. Moreover, the cost of transportation has a direct impact on the prices of essential goods. According to the European Union's Science Hub[1], production expenses among European Union companies consist of up to 15% of transportation and warehousing fees. Hence, appropriate decisions of freight forwarders – individuals involved in the overall arrangement of transportation services – remain vital. The *FedCSIS 2022 Challenge* is an attempt to address this issue.

Freight forwarders commonly use their expert intuition to decide whether to accept or reject a contract. It allows them to arrive at accurate choices, though underneath it is a complex process. The contract cost is affected by, but not limited to, the size of transported goods, their weight, fuel prices and, most importantly, the contract route. Various countries are likely to have significantly different transit-related costs. Furthermore, outlier contracts such as hazardous freight or those requiring additional safeguarding are particularly hard to handle. The motivation for the competition was the presumption that machine learning (ML) can make efficient use of this data.

The paper is organized as follows: Section II reviews the literature on ML in freight forwarding tasks. Section III outlines the objective of the competition and gives some details about the data. Section IV describes the baseline solution that we prepared for the competition purposes. Section V reports the winning solutions. Section VI concludes the paper.

II. Analysis of Related Literature

The income of the freight forwarder is largely based on a commission from the profit of successfully finding transportation services for the cargo that needed moving. Freight forwards operate on online freight exchanges, such as Timocom[2], Trans.eu[3] or Teleroute[4], seeking the most profitable contracts for which they can find a transportation service. The accurate transportation cost prediction is one of the key problems that need to be solved by freight forwarders to be successful. The importance of managing forwarders' information is discussed in [1], while [2] analyzes other factors impacting the financial effectiveness of managing transportation logistics.

An interesting topic is the prediction of the future freight demand [3], which would enable freight forwarders to balance risk and expected income. Other ML-based approaches focus on finding estimated time of arrival (ETA) or predict fuel consumption. In [4] a random forest is used to predict ETA for intermodal transportation (i.e. including sea and/or railway transport). In [5], [6] random forests and support vector machines are also used to predict fuel consumption in order to monitor and prevent fuel fraud. A recent review [7] summarizes the aspects of freight forwarding and transportation, whereby ML approaches have been utilized up to now.

Meanwhile, factors other than time and fuel consumption influencing transportation costs, especially in data-driven ML approaches, have not been thoroughly studied. However, [8], [9] propose expert models to calculate such transportation costs. Moreover, [10] proposes a statistical model and analyzes the impact of various factors on the estimated cost. While [11] also takes into account risk factors for the ocean transportation costs. Finally, [12] tried to solve the problem of cost prediction using artificial neural networks. However, as in most of the mentioned studies, these results were based only on small data sets or expert surveys. We believe that providing research community with a more comprehensive data will prove crucial for finding new factors that impact the accuracy of predicting transportation cost for forwarding contracts.

III. FedCSIS 2022 Challenge Outline

The challenge was launched at the KnowledgePit platform[5] on March 1, 2022, and the submission system was opened until

This research was partially 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.

[1] https://joint-research-centre.ec.europa.eu/scientific-activities-z/transport-sector-economic-analysis_en

[2] https://www.timocom.co.uk/smart-logistics-system/freight-exchange

[3] https://www.trans.eu/en/carriers/

[4] https://teleroute.com/en-en/

[5] https://knowledgepit.ai/fedcsis-2022-challenge/

May 27, 2022. We refer to [13] for more information about KnowledgePit, as well as about the previous KnowledgePit competitions that have quite a long tradition at FedCSIS.

The data used this year was provided by the competition sponsor, i.e. Control System Software – a Polish software company that is specializing in solutions for the Transportation, Spedition, and Logistics industry. The task for participants was to predict execution costs of forwarding contracts described in the available test data. An accurate prediction model for this task could be used in future to support freight forwarders.

A. Data preparation

The data sets that were made available in our competition describe an over six-year history of contracts accepted by a large Polish transportation company. The main data was composed of two separate tables. The first one contained basic information about the contracts, and the second one described the main sections of the planned routes associated with each contract. The first column in both tables, i.e. *id_contract*, stored identifiers that allow matching records between them. Additionally, the second column in the first table (the main data file), i.e. *expenses*, contained information about the actual prediction target values. A short description of the remaining data attributes from both tables was also made available in separate files. Finally, an additional data table containing historical wholesale prices of fuel was provided.

Since the data came from a real transportation company, all sensitive information had to be scrambled prior to publishing. All identifiers were removed or encoded by random strings. Geo-location data related to key points on the routes was modified. Instead of original values, we used the Nominatim service[6] to generate coordinates of the central points in the corresponding post code areas. In the published data, the original geographical coordinates are changed into the generated ones. Some of the characteristics of trucks and trailers were transformed into indicators. Finally, the fuel prices and the target values (the contract execution costs) were rescaled.

For the purpose of the evaluation, the data was divided into separate training and test data tables. The training data contained approximately five-year history of the accepted contracts, and the test set was composed of the data collected in the last year (between Nov. 1, 2020 and Nov. 23, 2021). In total, training data stored information about 330,055 contracts described by 36 attributes, and the total number of route parts was 1,189,654 (the route data table had 60 attributes). The empirical distribution of the target expenses looked like a mixture of a few Gaussians, with the mean value 6.3735 and standard deviation 1.059. The histogram of target values is presented on Figure 1. The test data contained 72,452 contracts, and the corresponding route data consisted of 325,222 entries.

B. Evaluation procedure

The evaluation procedure for our competition was typical to challenges held at KnowledgePit [14], [15]. Competitors submitted solutions as text files with each line containing a single prediction for the corresponding test instance. The quality of submissions was evaluated online. We used RMSE as the error measure. The preliminary score of each submission was computed on a small subset of the test records. Approximately 10% of test data was used for the preliminary evaluation. The best preliminary result for each team was published on the public leaderboard. The final evaluation was performed after the competition's completion using the remaining part of the test instances. Those results were then published online too. Only the teams which submitted a short report describing their approach were qualified for the final evaluation.

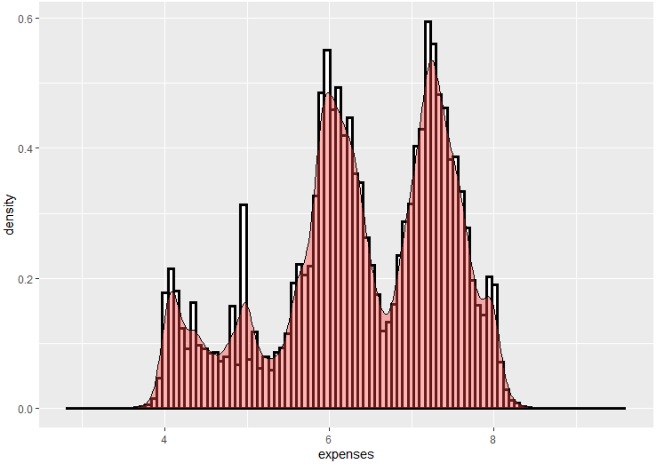

Fig. 1: The target expenses values in the training data set.

IV. Competition Baseline

To give a reference to competitors, we prepared and submitted at the very beginning predictions of our own baseline model. To construct it, we first analyzed the available data and identified features that could be useful. We divided the data preprocessing task into stages. At each stage, we extracted different types of features describing the contracts from the available data tables. This part of the model preparation process (i.e. feature extraction [16], [17]) proved to be crucial to the performance of the resulting prediction model.

Firstly, we processed the main data table. After consulting with domain experts, we identified categorical features that could have predictive value. For each of such features, we narrowed its set of possible values to those which appeared in at least 1% of training data. All other non-missing values were changed to *other*. After this transformation, we used two types of encoding. The selected features were one-hot encoded. Additionally, we created a numeric version of the categorical features by transforming each value into the mean *expenses* of contracts from the training data with that value. Overall, we applied this transformation to features *id_payer*, *id_currency*, *direction*, *load_size_type*, *contract_type*, *id_service_type*, *first_load_country*, *last_unload_country*, *route_start_country*, *route_end_country*, *prim_train_line*, *prim_ferry_line*, *route_start_month*, and *route_end_month*.

In the second stage, data from the route tables was processed. Again, the filtration of rare categorical values was performed. After that, for each contract we performed projections of aggregated *km*, *km_haversine*, and *kg_current* values on the values of selected categorical features. We added those

[6] https://nominatim.org/

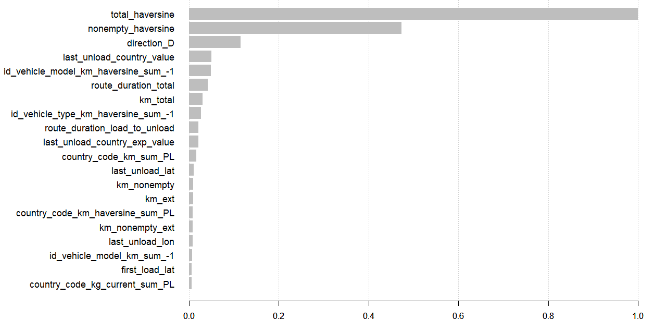

Fig. 2: Estimated feature importance for the baseline model.

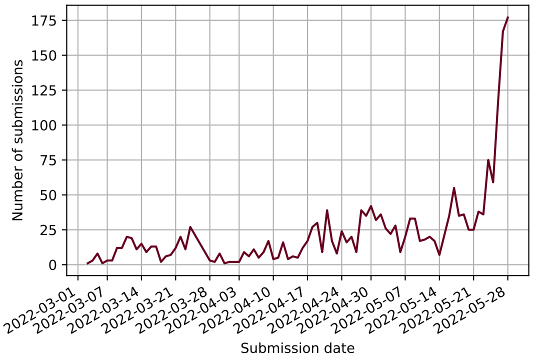

Fig. 3: Activity of participants by means of daily submissions.

projections as new features describing individual contracts. We also constructed a few dozen of other auxiliary features, such as the total number of steps without a load, the number of steps performed by external contractors, the average ratio between the current weight and the length of the route part, etc.

Finally, we added time features (e.g. month number, quarter, whether the contract will be executed through the weekend), fuel prices at the route beginning, and truck features (e.g. average size of the required trucks during the route). In total, we defined 585 numeric features describing the contracts.

We constructed the prediction model using the XGBoost library [18]. We did not focus on the hyperparameter tuning. We used a small portion of the training data as a validation set and experimentally checked several settings. The final model was heavily regularized, i.e. the learning rate was set to 0.01, $\alpha = \lambda = 1$, and subsampling was used on both instances and columns. The total number of used trees was 2,500, and the maximum depth of trees was set to 8. Figure 2 shows the estimation of feature importance in the resulting model. In the competition, our model had the fifth score with the preliminary RMSE value 0.1491 and the final result 0.1475.

V. Competition Results

The challenge was taken up by 130 teams from 24 countries. The teams came e.g. from Poland (76), India (14), and the USA (4). There were 1,927 solutions submitted. Figure 3 shows activity of competitors expresses in terms of the number of daily submissions. It shows that the number of daily submissions remained stable in the first half of the competition. Teams increased their activity in the second half, reaching the peak in the last week. Solutions submitted during the last 3 days of the competition account for nearly 24% of all solutions. This shows that the competition between participants continued until the last moment. In Table I, we present the final ranks, scores, and the number of submissions of the best performing teams. Figure 4 follows with more details about the teams that eventually managed to exceed our baseline solution.

TABLE I: Top 10 final results of the FedCSIS 2022 Challenge.

Rank	Team name	Preliminary	Final score	#subs
1	Dymitr	0.1398	0.1383	619
2	Cyan	0.1402	0.1391	181
3	hieuvq	0.1396	0.1407	159
4	Lord of the ML	0.1434	0.1420	147
5	baseline	0.1491	0.1475	-
6	kubapok	0.1502	0.1494	32
7	DeepIf	0.1500	0.1498	28
8	Stan	0.1529	0.1519	131
9	Artur Budzyński	0.1549	0.1520	45
10	Nindza Zhelki	0.1567	0.1573	36
. . .	. . .	. . .	. . .	. . .

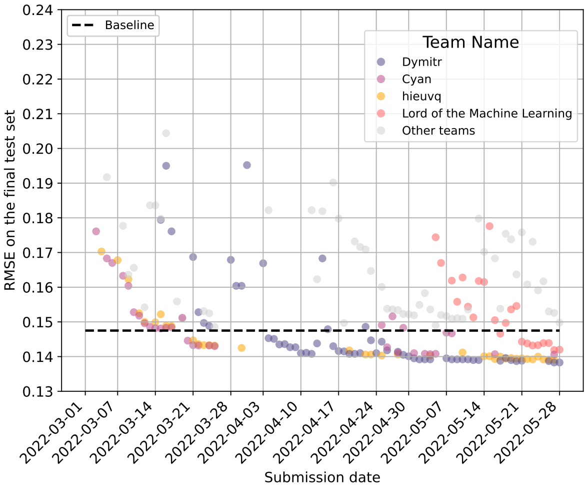

Fig. 4: RMSE (final test set) of solutions of the best teams.

As in previous KnowledgePit competitions, feature extraction was a crucial step. The solutions of the best-performing teams were preceded by in-depth data analysis and processing. The well-established approach of a feature selection preceded by a feature generation was the most common. Feature generation methods ranged from simple statistics to a manual selection of feature combinations and regex-based information extraction from text describing temperature requirements.

Every team that exceeded the baseline used gradient boosting methods (XGBoost [18], LightGBM [19] or CatBoost [20]) and model ensemble techniques. The winners' solution puts emphasis on choosing suitable ensemble methods and model diversification, whereby their two proposed approaches focus on the diversity of the models' hyperparameters and the level of disagreement of the models' output. Both approaches were used in the final solution. The decision of their final model ensemble was the average of the models' predictions on two subsets with different features. The runner-up team used the model stacking approach (i.e. ridge regression trained on top of gradient boosted trees' outputs). On the other hand, the team that finished third conducted a forecast post-processing that aimed to predict a trend in the contract costs. That forecast was used to adjust the predictions of the final model.

VI. Conclusions

We presented an international data mining competition related to a vital problem in the transportation and logistic industry, i.e. predicting the execution costs of forwarding contracts accepted by a freight company. We described the competition scope and available data sets, and we proposed a baseline model for the task. We also discussed the most successful solutions proposed by the participants.

The competition was a successful event, with 130 registered teams from 24 countries. The most accurate solutions were largely dominated by gradient boosting models implemented in popular libraries, such as XGBoost, and LightGBM. They were typically combined with feature extraction techniques in the data preprocessing phase. Moreover, a few teams decided to mix several models trained on different parts of data, and their final solutions were generated using an ensemble.

Reducing transportation expenses by selecting optimal contracts, or by identifying the most costly factors can decrease the price of production of many goods including those purchased on a daily basis. We believe that our competition contributed to the discussion on the estimation of forwarding contract costs. The solutions developed by participants, and outcomes of future research may pronouncedly influence the decision-making process of transportation and logistics companies. By providing the research community with a large-scale data set, we hope to accelerate the advances in this area.

References

[1] E.-S. Lee and D.-W. Song, "Knowledge Management in Freight Forwarding as a Logistics Intermediator: Model and Effectiveness," *Knowledge Management Research & Practice*, vol. 16, no. 4, pp. 488–497, 2018. [Online]. Available: https://doi.org/10.1080/14778238.2018.1475848

[2] R. Burkovskis, "Efficiency of Freight Forwarder's Participation in the Process of Transportation," *Transport*, vol. 23, no. 3, pp. 208–213, 2008. [Online]. Available: https://doi.org/10.3846/1648-4142.2008.23.208-213

[3] J.-A. Moscoso-López, I. T. Turias, M. Come, J. Ruiz-Aguilar, and M. Cerbán, "Short-Term Forecasting of Intermodal Freight Using ANNs and SVR: Case of the Port of Algeciras Bay," *Transportation Research Procedia*, vol. 18, pp. 108–114, 2016. [Online]. Available: https://doi.org/10.1016/j.trpro.2016.12.015

[4] A. Balster, O. Hansen, H. Friedrich, and A. Ludwig, "An ETA Prediction Model for Intermodal Transport Networks Based on Machine Learning," *Business & Information Systems Engineering*, vol. 62, no. 5, pp. 403–416, 2020. [Online]. Available: https://doi.org/10.1007/s12599-020-00653-0

[5] S. Wickramanayake and H. D. Bandara, "Fuel Consumption Prediction of Fleet Vehicles Using Machine Learning: A Comparative Study," in *2016 Moratuwa Engineering Research Conference, MERCon 2016*, 2016, pp. 90–95. [Online]. Available: https://doi.org/10.1109/MERCon.2016.7480121

[6] M. A. Hamed, M. H. Khafagy, and R. M. Badry, "Fuel Consumption Prediction Model Using Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021. [Online]. Available: https://doi.org/10.14569/IJACSA.2021.0121146

[7] K. Tsolaki, T. Vafeiadis, A. Nizamis, D. Ioannidis, and D. Tzovaras, "Utilizing Machine Learning on Freight Transportation and Logistics Applications: A Review," *ICT Express*, 2022. [Online]. Available: https://doi.org/10.1016/j.icte.2022.02.001

[8] Y. Konishi, S.-i. Mun, Y. Nishiyama, and J. E. Sung, *Determinants of Transport Costs for Inter-regional Trade*. Research Institute of Economy, Trade and Industry, 2012.

[9] B. Kordnejad, "Intermodal Transport Cost Model and Intermodal Distribution in Urban Freight," *Procedia – Social and Behavioral Sciences*, vol. 125, pp. 358–372, 2014. [Online]. Available: https://doi.org/10.1016/j.sbspro.2014.01.1480

[10] S. Camisón-Haba and J. A. Clemente, "A Global Model for the Estimation of Transport Costs," *Economic Research – Ekonomska Istraživanja*, vol. 33, no. 1, pp. 2075–2100, 2020. [Online]. Available: https://doi.org/10.1080/1331677X.2019.1584044

[11] S. Nataraj, C. Alvarez, L. Sada, A. Juan, J. Panadero, and C. Bayliss, "Applying Statistical Learning Methods for Forecasting Prices and Enhancing the Probability of Success in Logistics Tenders," *Transportation Research Procedia*, vol. 47, pp. 529–536, 2020. [Online]. Available: https://doi.org/10.1016/j.trpro.2020.03.128

[12] A. Singh, A. Das, U. K. Bera, and G. M. Lee, "Prediction of Transportation Costs Using Trapezoidal Neutrosophic Fuzzy Analytic Hierarchy Process and Artificial Neural Networks," *IEEE Access*, vol. 9, pp. 103 497–103 512, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3098657

[13] A. Janusz and D. Ślęzak, "KnowledgePit Meets BrightBox: A Step Toward Insightful Investigation of the Results of Data Science Competitions," in *Proceedings of the 2022 Federated Conference on Computer Science and Intelligence Systems, Sofia, Bulgaria, September 4-7, 2022*, ser. Annals of Computer Science and Information Systems, M. Ganzha, M. Paprzycki, and D. Ślęzak, Eds., vol. 30, 2022.

[14] A. Janusz, T. Tajmajer, M. Świechowski, Ł. Grad, J. Puczniewski, and D. Ślęzak, "Toward an Intelligent HS Deck Advisor: Lessons Learned from AAIA'18 Data Mining Competition," in *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, Poznań, Poland, September 9-12, 2018*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 15, 2018, pp. 189–192. [Online]. Available: https://doi.org/10.15439/2018F386

[15] A. Janusz, M. Przyborowski, P. Biczyk, and D. Ślęzak, "Network Device Workload Prediction: A Data Mining Challenge at Knowledge Pit," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 21, 2020, pp. 77–80. [Online]. Available: https://doi.org/10.15439/2020F159

[16] D. Ślęzak, M. Grzegorowski, A. Janusz, M. Kozielski, S. H. Nguyen, M. Sikora, S. Stawicki, and Ł. Wróbel, "A Framework for Learning and Embedding Multi-Sensor Forecasting Models into a Decision Support System: A Case Study of Methane Concentration in Coal Mines," *Information Sciences*, vol. 451-452, pp. 112–133, 2018. [Online]. Available: https://doi.org/10.1016/j.ins.2018.04.026

[17] H.-M. Wong, X. Chen, H.-H. Tam, J. Lin, S. Zhang, S. Yan, X. Li, and K.-C. Wong, "Feature Selection and Feature Extraction: Highlights," in *2021 5th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, ser. ISMSI 2021, New York, NY, USA, 2021, pp. 49–53. [Online]. Available: https://doi.org/10.1145/3461598.3461606

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, New York, NY, USA, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA, 2017, pp. 3149–3157.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, Red Hook, NY, USA, 2018, pp. 6639–6649.