Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 413–416

DOI: 10.15439/2022F298
ISSN 2300-5963 ACSIS, Vol. 30

Prediction of the Costs of Forwarding Contracts with Machine Learning Methods

Stanisław Kaźmierczak
Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland
Email: stanislaw.kazmierczak@pw.edu.pl

***Abstract*—This paper summarizes experiments conducted and findings related to FedCSIS 2022 Challenge that we participated in. The task was to develop a predictive model that estimated costs pertained to the execution of forwarding contracts (FC). We thoroughly analyze the dataset and present steps performed in the data preprocessing stage. Then we describe our approach to building a predictive model, which placed us eighth out of 135 teams. In the end, a wide range of ideas for further research is provided.**

***Index Terms*—forwarding contracts, data preprocessing, machine learning**

I. Introduction

DUE TO the specialization of work, the economy of production scale, and mass consumption, places where products are manufactured do not coincide with places where the demand for them is reported. Therefore, transportation is essential to bridge the gap between the buyer and the seller. It is a part of a logistic chain and plays a crucial role in attaining its primary goal – resource optimization.

Freight forwarding is an activity consisting in organizing the transport of goods. The forwarding company performs various activities for the client related to the organization of transportation, starting from adjusting the means of transportation through consulting in the field of cargo transport and ending with unloading the goods. Forwarder operates on the freight exchange, a place of information exchange between carriers and forwarding companies. Its purpose is to facilitate communication and accelerate the conclusion of transactions in the economic sector of transport.

Forwarder's work consists in searching orders on freight exchange, evaluating them, and selecting profitable ones. The ultimate goal is to sign an FC. It is an agreement whereby the freight forwarder makes a commitment for the sender or recipient of the goods to transport them to the place of delivery, not conducting the carriage himself but finding the carrier who will carry the goods. The forwarder signs a carriage contract on his behalf but for the account of the sender or recipient.

This study describes the model created to predict the costs related to the execution of FCs. Such a model aims to support freight forwarders in selecting profitable contracts.

II. Related literature

Increased interest in predicting the cost of transport has been observed among researchers since the late 1990s. Various means of transport, parts of the world as well as predictive methods were analyzed. This section summarizes more significant, in our opinion, studies.

In [1], the authors investigated factors that affected transportation costs. They applied the tobit model to find that infrastructure is its most important determinant in the considered area. Other crucial factors included details of geography, administrative barriers, and the structure of the shipping industry.

The authors of [2] used regression-based methods to examine the determinants of shipping costs to the US. It was found that distance, containerization, and efficiency of a port were significant factors influencing freight. The study provided some examples of how private involvement in port management along with labor reform and reduced monopoly power led to efficiency and lower costs.

Reference [3] studied the impact of port characteristics on international maritime transport costs. It considered 16 Latin American countries and maritime trade transactions in containerizable goods. The authors employed a regression model and proved that doubling port efficiency in a pair of ports would have the same impact on international transport costs as halving the distance between them.

The authors of [4] created a microeconomic model of interregional freight transportation. They utilized an ordinary least squared regression model and showed that besides determinants of transport cost incorporated in the model, the degree of competition also played a significant role in freight charge prediction.

Reference [5] is another paper focused on the prediction of costs of maritime transport, more precisely, logistics costs in container ports. The authors applied transaction cost economics (TCE) to support and explain empirical findings. They found that the quality of port infrastructure, port services, and port connectivity are among the most important determinants of logistics costs in container ports.

In [6], the authors summarized crucial findings from previous studies related to the estimation of transport costs. Most of the approaches concentrated only on statistical analysis or employed regression-based methods.

To our knowledge, there is no study that elaborates on more sophisticated machine learning (ML) methods and considers various means of transport. We hope that our study contributes to filling this gap.

IEEE Catalog Number: CFP2285N-ART ©2022, PTI

III. Datasets

The dataset contains orders that appeared in the freight exchange and were accepted by a large Polish company. Training samples come from the period between January 2016 and November 2020, while test instances from the period between September 2020 and November 2021. It means that train orders are generally followed by test ones. There are two main reasons why the test set partially overlaps the train set: complex orders from the training part that require a long time to complete, as well as some reversed start and end times (the end time is earlier than the start time) in 452 orders. Order details such as its type, basic characteristics of the shipped goods, along with the expected route that a driver will have to cover are provided. Input columns are the same for both training and test sets except for the target variable – costs of individual orders – which is given in the case of the former one and needs to be predicted in the case of the latter.

A. Dataset description

In more detail, the training set consists of two tables: *css_main_training.csv* and *css_routes_training.csv*. The former contains fundamental information about the contracts. It has 330 055 rows and 36 columns. The latter describes the main sections of the planned routes associated with each contract. It consists of 1 189 654 rows and 60 columns. The first column of each table contains contract identifiers that allow matching records from both tables. *css_main_training.csv* contains a column with the prediction target. Analogously, the test set consists of two tables: *css_main_test.csv* and *css_routes_test.csv*. Their structure is the same as corresponding training tables, but the column with the prediction target is empty. The first table consists of 72 452 records, while the second – 325 222 records. Additionally, *fuel_prices.csv* contains wholesale prices of three different types of fuel for the period of training and test data. More details about data and the competition itself are provided in [7].

B. Data analysis

In the case of the analyzed dataset, there are three types of data: numerical, categorical (including binary), and text (only one column of this type – requirement related to the temperature). In terms of most features, we can observe significant skewness (numerical columns) or noticeable imbalance (binary columns), which is generally not positive from a machine learning perspective.

Each order in the main table is timestamped. Thus, if orders are grouped, columns may be viewed as time series. We stick to the most important column, *expenses*, which is the target variable. Fig. 1 depicts expenses from orders grouped by different time periods. Several conclusions can be drawn. First, orders that are planned to start on Friday are the most expensive, and those beginning during a weekend – the cheapest. Conversely, orders scheduled to be finished on a weekend are high-priced. Second, throughout a year, there are peaks in cost irrespective of whether we consider the beginning or end of the order. It concerns both fixed and floating holidays. Finally, one may observe that contracts planned to be accomplished in the summer months are cheaper than those in other parts of the year.

The results obtained in standard cross-validation in which training and test instances are mixed in terms of timestamp were approximately 10% better than those registered on the competition platform in which training samples are followed by test ones. It suggests that data or/and concept drift occurs. The difference in the distribution of some features between the training and test data is not a sufficient explanation of this phenomenon. There is no statistically significant difference in the distribution of the target variable. As the last step, we applied TSNE to map training and test instances to the 2D plane. We did not observe any significant dissimilarities between both types of samples. Fig. 2 depicts the results of the algorithm. Data/concept drift needs further investigation we did not manage to perform before the end of the competition.

IV. Data preprocessing

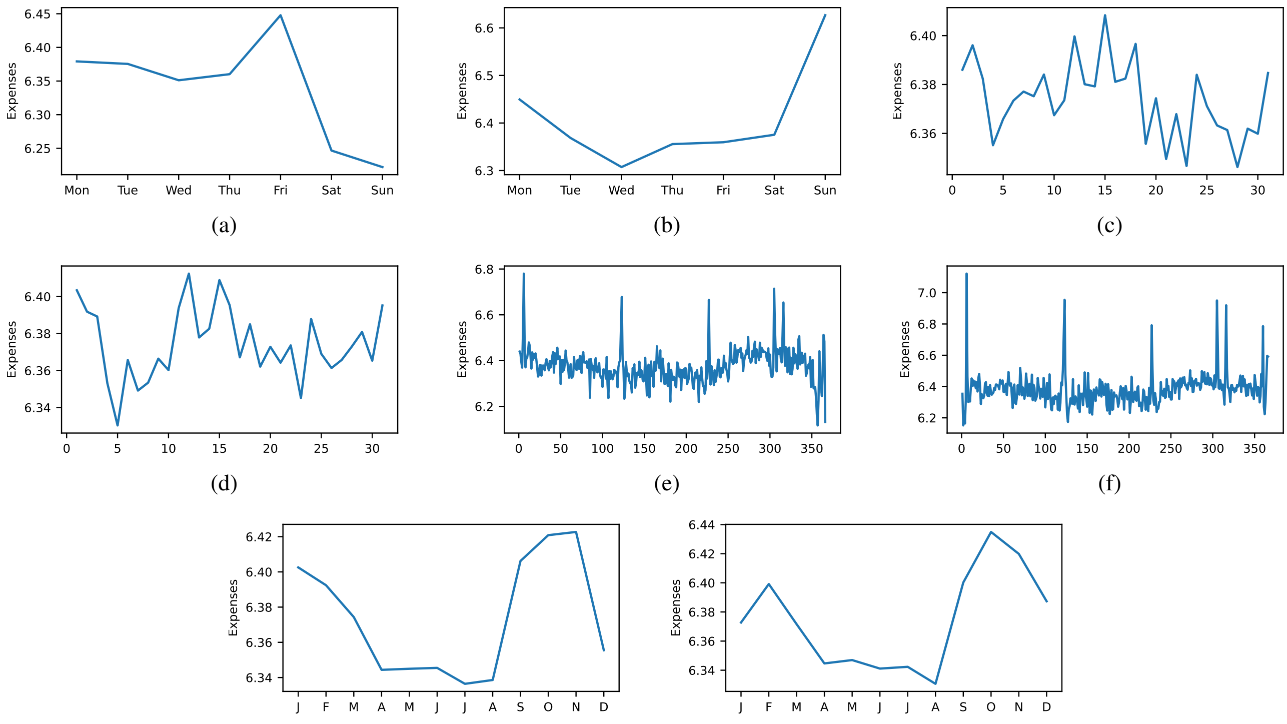
Quality data is necessary for machine learning models to operate efficiently. In general, data preparation requires more time and effort than actual modeling [8]. In this section, we present preprocessing steps that made our data prepared to build a predictive model.

A. Main tables

1) All data except the following columns were loaded: *temperature*, *first_load_lat*, *first_load_lon*, *last_unload_lat*, *last_unload_lon*, *route_start_lat*, *route_start_lon*, *route_end_lat*, *route_end_lon*.
2) The following columns were one-hot encoded: *direction*, *id_service_type*, *contract_type*, *id_payer*, *first_load_country*, *last_unload_country*, *route_start_country*, *route_end_country*, *id_currency*, *prim_train_line*, *load_size_type*, *prim_ferry_line*.
    a) *id_payer* was limited to 50 payers with the most numerous contracts within both train and test set.
    b) In terms of *prim_train_line* and *prim_ferry_line*, the additional category representing missing values was created (in other one-hot encoded columns, there were no missing values).

B. Routes tables

1) The following columns were loaded: *id_contract*, *external_fleet*, *id_vehicle*, *id_trailer*, *if_empty*, *ferry*, *train*, *step_type*, *country_code*, *id_vehicle_model*, *id_vehicle_type*, , *vehicle_type*, *vehicle_capacity_type*, *trailer_generator*, *id_trailer_model*, *id_trailer_type*, *ferry_line*, *train_line*.
2) The following columns were one-hot encoded: *country_code*, *id_vehicle_model*, *id_vehicle_type*, *vehicle_capacity_type*, *step_type*, *trailer_generator*, *id_trailer_model*, *id_trailer_type*, *ferry_line*, *train_line*, *vehicle_type*.
    a) For all columns other than *step_type*, the additional category representing missing values was created



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 1: Expenses (target variable) as a function of time. In the consecutive subfigures, orders are grouped by: (a) and (b) – day of a week, (c) and (d) – day of a month, (e) and (f) – day of a year, (g) and (h) – month of a year. Subfigures (a), (c), (e), and (g) relate to the planned time of the beginning of a route, subfigures (b), (d), (f), and (h) – to the planned time of the end of a route.

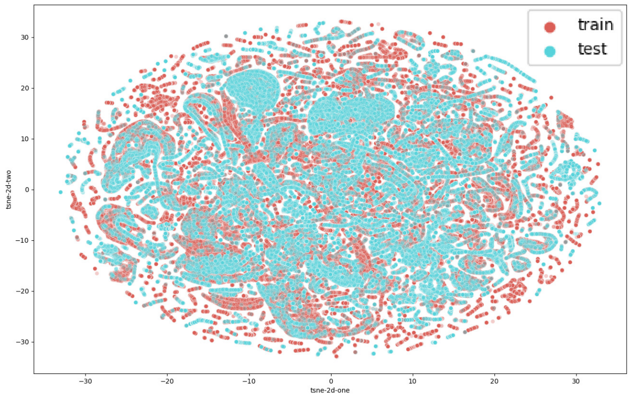


Fig. 2: TSNE applied to compare training and test instances.

(in the *step_type* column, there were no missing values).

3) Aggregation: instances were grouped by *id_contract*, and values from all other columns were summed for each contract. One more column reflecting the number of routes per contract was added.
4) Above data frames with routes were merged with the main data frames on the *id_contract* column which was then removed.

C. Fuel prices

1) *disel_type2_price* column was added to the above merged data frames based on date.
   a) *disel_type1_price* and *disel_type3_price* were not utilized since their Pearson correlation with *disel_type2_price* ranges between 0.98 — 0.99.

D. Data transformations

1) *ferry_intervals* was the only column with missing values. They were imputed on the basis of *ferry_duration*. If *ferry_duration* equals 0, then *ferry_intervals* is also 0. Otherwise, *ferry_intervals* is imputed with 1.
2) Time correction: if *route_end_datetime* was earlier than *route_start_datetime*, the values were swapped.
3) New features being the product of *disel_type2_price* and features related to distance (*km_empty*, *km_nonempty*, *km_total*, *km_train*, *ferry_duration*) were created. They should reflect the total cost of fuel related to particular contracts.
4) Additional features created:
   a) *duration_h* — the difference expressed in hours between *route_end_datetime* and *route_start_datetime*.

b) *start_day_of_week* and *end_day_of_week* — days of week related to *route_start_datetime* and *route_end_datetime*, respectively.

c) *day_of_year* — day of a year, ranges from 1 to 365 (in the case of the leap year, 1 was subtracted from all days after 29th February to be consistent with common years; it turned out that the predicted prices are highly correlated with some dates, especially related to festivals); one-hot encoded.

5) After all the aforementioned transformations, *route_start_datetime* and *route_end_datetime* columns were ultimately removed from the data frame.

6) Eventually, the data frame contained over 1000 columns. Features were assessed using XGBoost's *feature_importance* property. For final modeling, different number of the most valuable features were left (more details are provided in Section V).

V. Prediction results

The task of the created models was to predict the actual costs of individual orders as accurately as possible. Such models aim to assist freight forwarders in picking beneficial contracts. The quality of algorithms is evaluated using the RMSE measure.

The whole code was written in Python 3.7. Neural networks were created in Keras 2.3.1. Gradient boosting was implemented with the xgboost 1.0.2 package. In terms of all other machine learning algorithms, scikit-learn 0.24.2 was applied. If not mentioned otherwise, hyperparameters were left at their default settings.

The best results were obtained by XGBoost built on the most valuable 200 or 500 features mentioned in Subsection IV-D. It is not a surprise in terms of the tabular data since XGBoost is the top choice on the Kaggle platform in such cases as well. In terms of hyperparameter tuning, we selected hyperparameters and their considered value range as suggested in [9] and [10]. Due to time constraints, we optimized them one by one, assuming fixed (default) values for the others. It turned out that the following values brought the best results: *subsample* – 1, *max_depth* – 6, and *eta* – 0.3.

The final solution was constituted by the averaging ensemble of three XGBoost models with *n_estimators* set to 205 and built on all, 500, and 200 most valuable features, respectively. The RMSE obtained amounted to 0.1529, which placed us eight out of 135 teams.

It is worth mentioning that many algorithms other than XGBoost, were analyzed (we submitted 108 valid solutions). Before focusing on XGBoost, we tested a wider range of algorithms – linear regression, random forest, and different neural architectures. All of them were more than 10% worse than the final solution.

VI. Conclusions and further ideas

In this paper, we present our approach to building a model able to predict costs related to FC. Despite many conducted experiments and the reasonable score achieved, we still see a lot of room for improvement.

First, concept/data drift was detected but not addressed successfully. It requires further investigation. We believe that the application of some dedicated methods (please refer to [11]) may lead to prediction enhancement.

Second, data aggregation requires more experiments. In the current approach described in subsection IV-B, values from the routes table are summed for each corresponding contract. Such a method may cause the loss of some valuable information.

Third, we believe that there is still some uncovered potential in neural networks. Neural architectures are relatively hard to tune and prone to overfitting due to their complexity. Even if they do not outperform XGBoost, they can constitute a valuable element of the ensemble model by increasing its diversity.

Next, it may be worth looking one more time at encoding categorical features with a large number of values, e.g., *id_payer*. On the one hand, we should not expand a feature space massively. On the other, we must not allow valuable information to be lost.

Last but not least, it is worth taking a closer look at the feature selection. We applied a simple approach based on XGBoost's *feature_importance* property. However, this method does not take into account feature correlation. We strongly believe that the application of some more sophisticated feature selection algorithms along with other aforementioned ideas will further boost the prediction quality.

References

[1] N. Limao and A. J. Venables, "Infrastructure, geographical disadvantage, transport costs, and trade," *The world bank economic review*, vol. 15, no. 3, pp. 451–479, 2001.

[2] A. Micco and N. Pérez, "Determinants of maritime transport costs," *Inter-American Development Bank*, 2002.

[3] G. Wilmsmeier, J. Hoffmann, and R. J. Sanchez, "The impact of port characteristics on international maritime transport costs," *Research in transportation economics*, vol. 16, pp. 117–140, 2006.

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

[5] H.-s. Cho, "Determinants and effects of logistics costs in container ports: The transaction cost economics perspective," *The Asian Journal of Shipping and Logistics*, vol. 30, no. 2, pp. 193–215, 2014.

[6] 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.

[7] A. Janusz, A. Jamiołkowski, and M. Okulewicz, "Predicting the costs of forwarding contracts: Analysis of data mining competition results," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022, Sofia, Bulgaria, September 4-7, 2022*. IEEE, 2022.

[8] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015, vol. 72.

[9] A. Jain, "Complete Guide to Parameter Tuning in XGBoost with codes in Python," 2016, online; accessed 23-Jul-2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

[10] D. Martins, "XGBoost: A Complete Guide to Fine-Tune and Optimize your Model," 2021, online; accessed 23-Jul-2022. [Online]. Available: https://towardsdatascience.com/xgboost-fine-tune-and-optimize-your-model-23d996fab663

[11] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.