Study on Cost Estimation of the External Fleet Full Truckload Contracts

Jan Kaniuka[1][a], Jakub Ostrysz[1][b], Maciej Groszyk[1][c], Krzysztof Bieniek[1][d], Szymon Cyperski[2][e] and Paweł D. Domański[1,2][f]

[1] *Warsaw University of Technology, Institute of Control and Computation Engineering, Nowowiejska 15/19, 00-665 Warsaw, Poland*

[2] *Control System Software Sp. z o.o., ul. Rzemieślnicza 7, 81-855 Sopot, Poland*

Keywords: Cost Estimation, Full Truck Loads, Machine Learning, Regression, kNN, Decision Trees, Gaussian Processes.

Abstract: Goods shipping supports the operation and the development of the global economy. As there are thousands of logistics companies, there exists a big need for solutions for their daily operation. The shipment can be carried out in many ways. This work focuses on the road transportation in form of the Full Truck Load (FTL). Once the service is supported by the third party, there is a need to have a tool that compares various offers and allows to estimate the cost. Generally, FTLs are used in the long range routes and the estimation of such contracts can be handled in many ways starting from the simple calculators up to data based machine learning solutions. Nonetheless, the need for the cost estimation appears for the short routes, which often support long range ones. Their pricing rules differs from the long range ones and the required approaches should differ as well. This work presents the wide comparison of 35 regression and machine learning approaches applied to the task. The assessment is performed using real contract data of several companies operating in Europe.

1 INTRODUCTION

Full truckload (FTL) is a common transportation way, where the goods fill an entire truck. It ideally suits for large volume of goods where a load covers the whole truck space. There is an alternative approach called less than truckload (LTL), in which a truck takes partial loads to different contract load/unload locations within a single travel. This paper focuses on the FTL approach, however from a rare perspective.

First of all, the case of external fleet contract pricing is considered. It is assumed the contractor may use some custom dynamic pricing model, which is associated with serious challenges (Stasiński, 2020). These issues become even more significant in the case of short routes, when common relationships with the fuel costs and driver time start to matter much less. External fleet long range contracts pricing can be solved with the use of popular fright cost calculators or using the artificial intelligence (AI) and machine learning (ML) (Tsolaki et al., 2022).

The task of the short range external fleet FTL shipment cost estimation, is not specifically addressed in the literature. Actually, this subject is hidden behind the general FTL estimation and people practically do not distinguish short routes cost prediction as separate task. The findings that appeared during the investigation of this project repeat despite the method used. The biggest challenge in the FTL cost prediction is that the highest estimation residua appear for the short routes and low costs. The shorter the route, the more difficult it is to be estimated. The general absolute performance measures are low, while the relative ones appear to be suspiciously high. This happens due to the possible high share of the low cost routes. Thus, we have decided to take a closer look at this subject decomposing the problem into two subproblems. This work copes with the problem of the short routes, while the longer ones are already covered in (Cyperski et al., 2023). Concluding, our work aims to fill this gap. General FTL cost estimation approaches are introduced in Section 2. The case study and used data are presented in Section 3. This work compares 35 various estimators described in Section

[a] https://orcid.org/0009-0009-9379-4906
[b] https://orcid.org/0009-0006-8178-0134
[c] https://orcid.org/0009-0006-4050-6585
[d] https://orcid.org/0009-0007-5232-3628
[e] https://orcid.org/0009-0009-5655-7143
[f] https://orcid.org/0000-0003-4053-3330

Kaniuka, J., Ostrysz, J., Groszyk, M., Bieniek, K., Cyperski, S. and Domański, P.
Study on Cost Estimation of the External Fleet Full Truckload Contracts.
DOI: 10.5220/0012251000003543
In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2023) - Volume 2*, pages 316-323
ISBN: 978-989-758-670-5; ISSN: 2184-2809

4, which are applied to the task. Section 5 analyzes the results, while the work is concluded in Section 6.

2 DYNAMIC FTL PRICING

The FTL freight cost estimation model is needed as external fleets use dynamic pricing strategies (Stasiński, 2020). Any information about cost influential factors helps, especially in its structure evaluation and features selection. One should consider a combination of general and custom market and non-market factors (Vu et al., 2022). Proper selection of the influencing features improves the estimation.

Contract dependent factors are included in the order limiting the solution. They may address the type of the truck and required specific equipment, ADR (*l'Accord européen relatif au transport international des marchandises Dangereuses par Route* – transport with hazardous materials) or drivers' certificates. Shipping load and unload locations determine the route. The location and the contract timeframe has to be matched with the drivers availability. The contract also defines the payment terms. The literature focuses on the blind machine learning approaches (Pioroński and Górecki, 2022; Tsolaki et al., 2022) or more complex hybrid ones (Cyperski et al., 2023).

3 ESTIMATION CASE STUDY

The data used to evaluate and test the method originated from the databases for selected Polish transportation companies (Janusz et al., 2022). Original orders database consists of approximately 414,000 positions. The data considered are limited only to the short range contracts, which limits the number of data to 20,239 records from the time period from January 1^{st}, 2016 to April 30^{th}, 2022. These record are considered as the training data. Records from May 1^{st}, 2022 till August 1^{st}, 2022 are considered as the validating dataset. Therefore, we obtain 703 records in the validating dataset (see Table 1).

Table 1: Size of datasets used in experiments.

	Raw data	Preprocessed data
Train set	414 404	20 239
Test set	14 968	703

3.1 Data Preprocessing

Each registered contract is described by 22 variables from the production databases. While modeling the transportation cost, we limit this number to 12 the most important features: date of payment, min and max transport time, time interval, date of transport, lead time, total and total empty distances, number of pickups and unloadings, demand for cold storage and fuel-cost. The following descriptors are excluded:

- ID number → it is just a sequence number,
- maximum weight and tonne-kilometres → these data are frequently incomplete, and cannot be practically used,
- location cluster number (Cyperski et al., 2023), the latitudes and longitudes of the loading and unloading site → are used only in the selection of the short range contracts.

Python programming language (*scikit_learn* and *torch* libraries) and MATLAB (*Statistics and Machine Learning Toolbox*) are used during data processing and the estimation process.

4 ESTIMATION APPROACHES

This section describes the machine learning approaches taken into account during the study. The choice of black box approaches is motivated by the unknown pattern behind the data due to the dynamic geopolitical situation and the specificity of the very short shipping. Among other factors, rising inflation, Brexit and the COVID-19 pandemic are impacting truck cargo transit prices. The following regression estimation methods are taken into account:

– regression approaches:

LMS: Least Mean Squares,

R-LMS: Robust Linear Regression, which is robust against outliers (Holland and Welsch, 2007),

SLR: Stepwise Linear Regression (Yamashita et al., 2006),

TS-LR: Theil-Sen Regression (Wang et al., 2009),

H-LR: Huber Regressor (Huber and Ronchetti, 2011),

– support vector machines (Wang and Hu, 2005):

LSVM: Linear Support Vector Machines,

KSVM: Kernel Support Vector Machines,

QSVM: Quadratic Support Vector Machines,

CGSVM: Coarse Gaussian Support Vector Machines,

MGSVM: Medium Gaussian Support Vector Machines,

FGSVM: Fine Gaussian Support Vector Machines,

- Gaussian processes (Schulz et al., 2018):

  EGPR: Exponential Gaussian Process,

  SEGPR: Squared Exponential Gaussian Process Regression,

  MGPR: Matern 5/2 Gaussian Process Regression,

  RGPR: Rational Quadratic Gaussian Process Regression,

- **k-NN:** k-Nearest Neighbors Regressor (Yao and Ruzzo, 2006),
- **OMP** Orthogonal Matching Pursuit (Tropp, 2004),
- ridge regressions (Li et al., 2020):

  RR: Ridge Regression,

  ARD-RR: Automatic Relevance Determination,

  B-RR: Bayesian Ridge Regression,

- decision trees (Breiman et al., 1984):

  DTR: Decision Tree Regressor (Breiman, 2017),

  BoostRT: Boosted Regression Trees (Bergstra et al., 2012),

  GBoostRT: Gradient Boosting Regression (Friedman, 2001),

  HGBoostRT: Histogram Gradient Boosting Regression (Tiwari and Kumar, 2021),

  ERTR: Extremely Randomized Trees (Geurts et al., 2006),

  BRTR: Bagged Regression Trees (Sutton, 2005),

  F-DTR: Fine Regression Tree,

  M-DTR: Medium Regression Tree,

  C-DTR: Coarse Regression Tree,

  RFR: Random Forest Regression (Ho, 1995),

- regularization techniques (Tibshirani, 1996):

  LASSO-R: LASSO Regression,

  LARS-R: LARS Lasso (Efron et al., 2004),

  ENR: Elastic Net Regression (Zou and Hastie, 2005),

  LAR: Least Angle Regression,

- **ANN:** Artificial Neural Network optimized according to the following hyperparameters:
  - number of hidden neurons: 12-2048,
  - number of hidden layers: 1-8,
  - activation fun.: RELu, Tanh, identity, logistic,
  - optimizer: Adaptive Moment Estimation, Stochastic Gradient Descent.

5 RESULTS

The comparison of the methods uses the residuum analysis. Three performance measures are used: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Square Error (MSE). MAE error is defined as (1), while the MSE as (2), where y is the actual and $\hat{y}$ is the predicted value.

$$\mathrm{MAE} = \frac{1}{n}\sum|y-\hat{y}| \tag{1}$$

$$\mathrm{MSE} = \frac{1}{n}\sum(y-\hat{y})^2 \tag{2}$$

MAPE defines the accuracy of a forecasting method that is given by the formula (3)

$$\mathrm{MAPE} = \frac{100\%}{n}\sum\left|\frac{y-\hat{y}}{y}\right|. \tag{3}$$

Colin David Lewis proposed in (Lewis, 1982) the table (see Table 2) containing interpretation of typical MAPE values. MAPE as a relative error allows us to more naturally interpret of how accurate the model is.

Table 2: Interpretations of MAPE values.

MAPE [%]	Interpretation
<10	Highly accurate forecasting
10–20	Good forecasting
20–50	Reasonable forecasting
>50	Inaccurate forecasting

5.1 Comparison of the Performance

At first, the models are simply compared by their performance measures. Table 3 shows the respective values. Even the draft review brings interesting observations. First of all, each measure indicates different models. The MSE highly penalizes residua with large values in opposition to the small ones. It is shown (Seborg et al., 2010) that the MSE punishes large deviations and is sensitive to the outlying observations. The MAE is less conservative. It has the closest relationship to economic considerations (Shinskey, 2002).

The difference between MAE and MAPE also requires for some discussion. The fact that both indexes indicate different regressors is due to the fact that the errors lie in different parts of data. Minimized MAE is due to the residua of the higher cost sacrificing the

Table 3: Comparison of the regression models. Grey color highlights the extreme values of the models: the worst (red) and the best (green). Bold numbers indicate the worst and the best one.

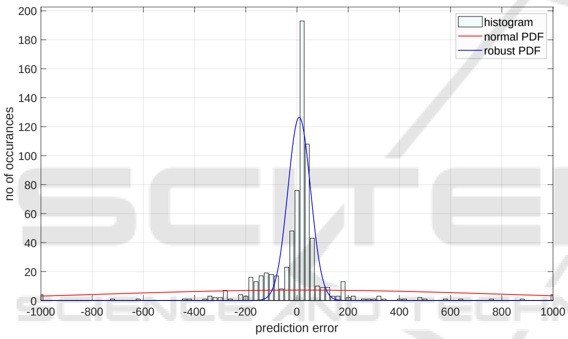
No	Descriptor	Method Name	MAE	MSE	MAPE [%]
1	LMS	Least Squares	205.8	704903	66.24
2	R-LMS	Robust Linear Regression	189.2	837335	37.69
3	SLR	Stepwise Linear Regression	226.8	1664193	75.34
4	TS-LR	Theil-Sen Regressor	188.4	851928	46.61
5	H-LR	Huber Regressor	177.6	763898	41.4
6	LSVM	Linear Support Vector Machines	175.6	764256	36.11
7	KSVM	Kernel Support Vector Machines	298.9	**1943548**	60.79
8	QSVM	Quadratic Support Vector Machines	187.1	877200	45.61
9	CGSVM	Coarse Gaussian Support Vector Machines	195.6	1088107	50.5
10	MGSVM	Medium Gaussian Support Vector Machines	235.0	1479565	56.79
11	FGSVM	Fine Gaussian Support Vector Machines	**304.8**	1746572	**79.59**
12	EGPR	Exponential Gaussian Process Regression	155.5	748644	42.58
13	SEGPR	Squared Exponential Gaussian Process Regression	202.4	1193684	42.76
14	MGPR	Matern 5/2 Gaussian Process Regression	182.1	967428	42.37
15	RGPR	Rational Quadratic Gaussian Process Regression	159.7	718013	38.31
16	k-NN	k-Nearest Neighbors Regressor	200.9	824671	57.08
17	OMP	Orthogonal Matching Pursuit	198.0	858053	40.65
18	RR	Ridge Regression	205.8	704902	66.24
19	ARD-RR	Automatic Relevance Determination	205.4	704991	65.99
20	B-RR	Bayesian Ridge Regression	205.7	704758	66.31
21	DTR	Decision Tree Regressor	167.9	614842	35.45
22	BoostRT	Boosted Regression Trees	164.7	719251	33.58
23	GBoostRT	Gradient Boosting Regression	151.8	695577	38.04
24	HGBoostRT	Histogram Gradient Boosting Regression	140.0	**490936**	34.01
25	ERTR	Extremely Randomized Trees	131.2	712589	27.68
26	BRTR	Bagged Regression Trees	**128.5**	626157	26.83
27	F-DTR	Fine Regression Tree	161.6	688117	29.23
28	M-DTR	Medium Regression Tree	140.6	675587	27.12
29	C-DTR	Coarse Regression Tree	130.5	609023	**26.25**
30	RFR	Random Forest Regression	136.1	625584	30.77
31	LASSO-R	LASSO Regression	205.9	705301	66.54
32	LARS-R	LARS Lasso	205.9	705301	66.54
33	ENR	Elastic Net Regression	204.9	702740	66.39
34	LAR	Least Angle Regression	205.8	704903	66.24
35	ANN	Artificial Neural Network	134.0	651000	27.82

contracts of low cost. In contrary, minimization of the relative error (MAPE) does not distinguish between the contract absolute value. This observation may be used to design estimation performance index.

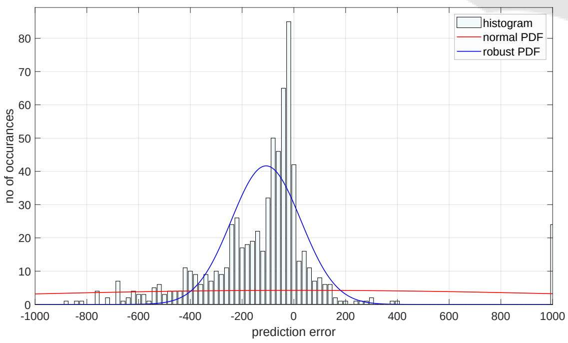
Furthermore, the reference between obtained numbers and the MAPE errors interpretation sketched in Table 2 shows that none of models can be considered as useful for the *good* or *highly accurate forecasting*. The majority of them are categorized as *reasonable*. Few of these reasonable models, such as Coarse, Medium or Bagged Regression Trees are positioned the closest to the good forecasting category. In contrary there are some models, like the Least Squares, Stepwise Linear Regression, the majority of Support Vector Machines, k-Nearest Neighbors, ridge regression variants and regularized models that are clearly unacceptable.

Thus, we notice that some categories of the approaches perform better, while the other ones, despite the version applied are not able to cope with the problem. We may clearly see, that despite the applied variant decision trees deliver consistently good results, in contrary to the Support Vector Machines. Linear Regression models are characterized by high variability in their performance. Robust Linear Regression is twice better than the Stepwise Linear version. It is probably due to the frequent outlying observations in data. Artificial Neural Network model is relatively good, however it looses against decision trees.

Further residuum analysis might be performed using the comparison of the obtained residua histograms. Fig.1 shows sample histograms for the best Coarse Regression Tree and the worst performing Fine Gaussian Support Vector Machines model. We clearly notice the non-Gaussian properties of the residua probably due to the original process data or the relatively small sample size. Furthermore, we observe that the results are skewed towards negative errors, i.e. overestimation. We also see a lot of outlying observations, which is reflected by the large difference between normal and robust version of the fitted Gaussian probabilistic density function (PDF).



(a) Coarse Regression Tree model.



(b) Fine Gaussian Support Vector Machines model.

Figure 1: Sample histogram plots for the best and the worst models with the fitted normal and robust Gaussian PDFs.

Next plot shows the predicted versus the real costs plot analyzing the hypothesis that the error and the quality of estimation may depend on the cost of the shipping. Fig. 2 shows this relationship for two selected models: C-DTR and FGSVM. We observe the largest difference in the decision tree model improvement is achieved for low and medium cost contracts.

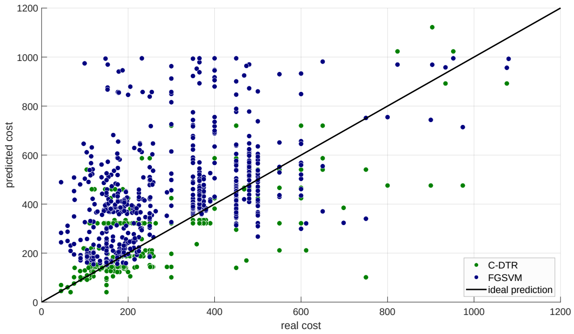


Figure 2: The predicted versus real cost for selected models.

It can be observed in Fig. 3, which shows the relationship between prediction error and the real cost. This dependence is well seen through the polynomial fitting the error and the real cost. To discard the effect of outliers robust regression is used for fitting.

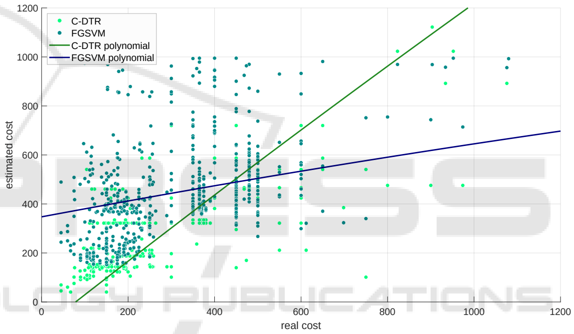


Figure 3: The relationship between the prediction error (3^{rd} order polynomial) and the contract real cost.

Residuum analysis is not a simple comparison between two numbers, as each measure, their relations matter, which can be further assessed with the statistical analysis or dedicated multi-criteria plots. Clearly, this is multi-criteria assessment. As one can see, absolute and relative errors may indicate different prediction models as good ones. There is a need for an aggregating approach and the according measure. We propose the two-dimensional plot of the relative (MAPE) versus the absolute (MAE) measure, denoted as Index Ratio Diagrams (IRD). The best model (predictor) is the one, which is the closest to the origin. The distance measure is called the Aggregated Distance Measure (ADiMe). Fig. 4 presents IRD plot for the considered FTL cost estimation models.

The general formulation of the aggregated ADiMe index is sketched in Eq. (4). Generally, the measure might be scaled with coefficients w_1 and w_2, which set the ratio between the relative and the absolute index. In the considered case both coefficients are equal and

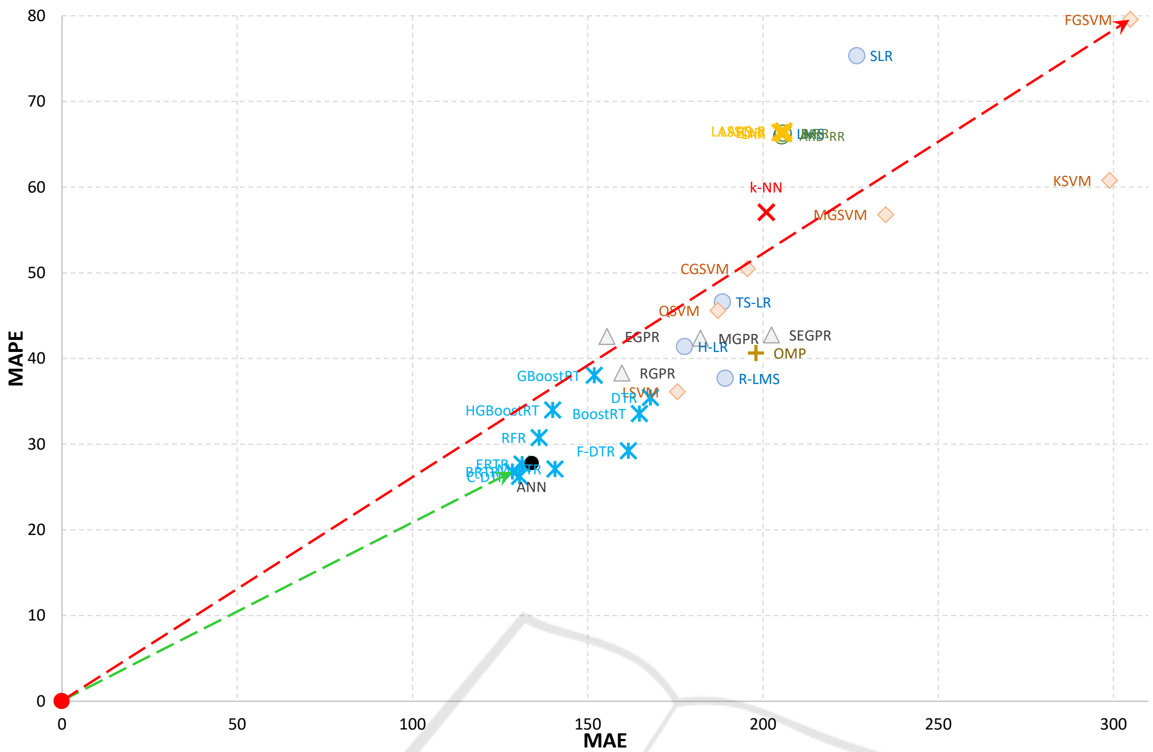


Figure 4: The IRD plot comparing the models: red line shows the worst model (the longest distance), green the best one (the shortest distance).

$w_1 = w_2 = 1$.

$$\text{ADiMe} = \sqrt{w_1 \cdot \text{MAE}^2 + w_2 \cdot \text{MAPE}^2} \tag{4}$$

The colors on the plot indicates various classes of the models. In this way, the assessment of the models is much easier, as each model is represented by a single point in the two-dimensional plot. We clearly see that the class of the decision tree models is the most homogeneous and all of them achieve relatively good performance. In contrary, the support vector machines models represent the most scattered class, with the biggest difference in their performance. We also see that the ridge regression variant and the selection of the regularization technique make no difference for the model property. Gaussian regression schemes are also quite homogeneous, however their performance is worse comparing to the decision trees.

Table 4 compares the models. It presents ordered first five the best models and five the worst ones as well. We see that the difference between the methods is significant, as poor estimators might be three times worse than the best one. This observation once more confirms the well-known fact that the model selection is crucial to achieve satisfactory estimators.

Comparison of the best models is not clearly visible in the non scaled plot. Therefore, the Fig. 5 shows the same plot, but with the magnified region

Table 4: Comparison of five the best and five the worst models according to the ADiMe.

rank	method	ADiMe	MAE	MAPE [%]	MSE
1	BRTR	131.3	128.5	26.83	626157
2	C-DTR	133.2	130.5	26.25	609023
3	ERTR	134.1	131.2	27.68	712589
4	ANN	136.9	134.0	27.82	651000
5	RFR	139.6	136.1	30.77	625584
	⋮	⋮	⋮	⋮	⋮
30	LASSO-R	216.4	205.9	66.54	705301
31	LARS-R	216.4	205.9	66.54	705301
32	SLR	239.0	226.8	75.34	1664193
33	MGSVM	241.7	235.0	56.79	1479565
34	KSVM	305.0	298.9	60.79	1943548
35	FGSVM	315.0	304.8	79.59	1746572

of the best models. It's worth to notice that the artificial neural network model (ANN) behaves similarly to the best predictors, being ranked as the fourth best.

6 CONCLUSIONS AND FURTHER OPPORTUNITIES

The presented work has three dimensions and conclusion types. Firstly, it analyzes the issue of the short routes external fleet FTL contracts cost assessment, which actually is hardly analysed in the literature. Subject seems to be underrated, despite of its huge

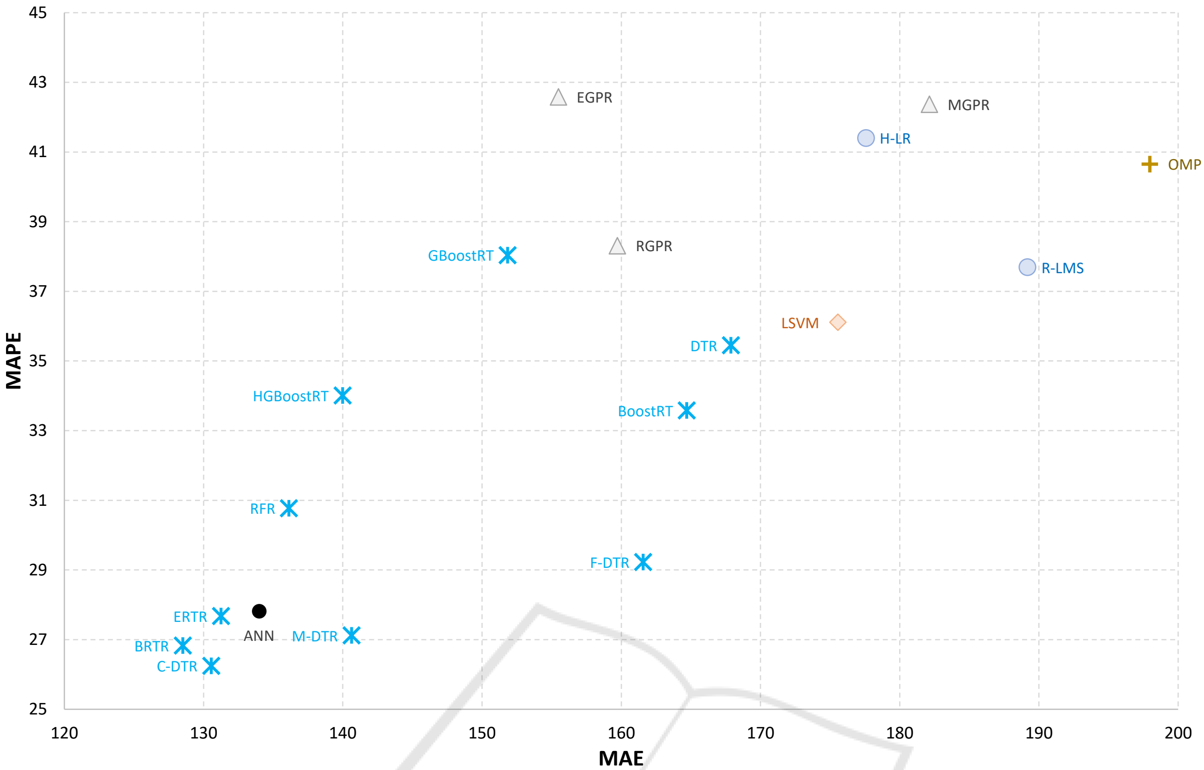


Figure 5: The scaled IRD plot comparing the best prediction models.

practical value. Probably it is hidden behind the general FTL pricing, or the results are not so impressive though being practically acceptable.

The second contribution is in fact that we were able to find out approaches, originating from the decision trees that allow to get satisfactory models. Especially, the Coarse, Medium or Bagged Regression Trees deliver reasonable prediction accuracy.

Finally, the obtained results are used to perform deeper residuum analysis. It shows that *an error is not equal to the error*. The specific use of the measure may favor one method against the other and the interpretation of the results can be easily biased. Therefore, the estimation model analysis should not be limited to the simple comparison of one measure numbers, but further investigation using statistical methods, or just different residua presentation can help.

It is also proposed the aggregated approach to the multi-criteria residuum analysis using the visual approach through the Index Ratio Diagrams (IRD) and resulting Aggregated Distance Measure (ADiMe).

The analysis of the FTL shipping is still not over. Several subjects remain open. How to improve obtained models and optimize their hyperparameters? How to improve the residuum analysis and how to get it simpler? There is still a work to be done.

ACKNOWLEDGEMENTS

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

Bergstra, J., Pinto, N., and Cox, D. (2012). Machine learning for predictive auto-tuning with boosted regression trees. In *2012 Innovative Parallel Computing (InPar)*, pages 1–9. IEEE.

Breiman, L. (2017). *Classification and Regression Trees (1st ed.)*, chapter 8. Routledge.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Cyperski, S., Domański, P. D., and Okulewicz, M. (2023). Hybrid approach to the cost estimation of external-fleet full truckload contracts. *Algorithms*, 16(8).

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63:3–42.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Holland, P. W. and Welsch, R. E. (2007). *Robust regression using iteratively reweighted least-squares*. Taylor & Francis.

Huber, P. and Ronchetti, E. (2011). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.

Janusz, A., Jamiołkowski, A., and Okulewicz, M. (2022). Predicting the costs of forwarding contracts: Analysis of data mining competition results. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 399–402.

Lewis, C. (1982). Industrial and business forecasting methods. *London: Butterworths*.

Li, D., Ge, Q., Zhang, P., Xing, Y., Yang, Z., and Nai, W. (2020). Ridge regression with high order truncated gradient descent method. In *12th International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 1, pages 252–255.

Pioroński, S. and Górecki, T. (2022). Using gradient boosting trees to predict the costs of forwarding contracts. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 421–424. IEEE.

Schulz, E., Speekenbrink, M., and Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16.

Seborg, D. E., Mellichamp, D. A., Edgar, T. F., and Doyle, F. J. (2010). *Process dynamics and control*. Wiley.

Shinskey, F. G. (2002). Process control: As taught vs as practiced. *Industrial and Engineering Chemistry Research*, 41:3745–3750.

Stasiński, K. (2020). A literature review on dynamic pricing - state of current research and new directions. In Hernes, M., Wojtkiewicz, K., and Szczerbicki, E., editors, *Advances in Computational Collective Intelligence*, pages 465–477, Cham. Springer International Publishing.

Sutton, C. D. (2005). 11 - classification and regression trees, bagging, and boosting. In Rao, C., Wegman, E., and Solka, J., editors, *Data Mining and Data Visualization*, volume 24 of *Handbook of Statistics*, pages 303–329. Elsevier.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tiwari, H. and Kumar, S. (2021). Link prediction in social networks using histogram based gradient boosting regression tree. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–5.

Tropp, J. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242.

Tsolaki, K., Vafeiadis, T., Nizamis, A., Ioannidis, D., and Tzovaras, D. (2022). Utilizing machine learning on freight transportation and logistics applications: A review. *ICT Express*.

Vu, Q. H., Cen, L., Ruta, D., and Liu, M. (2022). Key factors to consider when predicting the costs of forwarding contracts. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 447–450. IEEE.

Wang, H. and Hu, D. (2005). Comparison of SVM and LS-SVM for regression. In *2005 International conference on neural networks and brain*, volume 1, pages 279–283. IEEE.

Wang, X., Dang, X., Peng, H., and Zhang, H. (2009). The Theil-Sen estimators in a multiple l;inear regression model. accessed on 18 August 2023.

Yamashita, T., Yamashita, K., and Kamimura, R. (2006). *A Stepwise AIC Method for Variable Selection in Linear Regression*. Taylor & Francis.

Yao, Z. and Ruzzo, W. (2006). A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC bioinformatics*, 7 Suppl 1:S11.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.