Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 431–436
DOI: 10.15439/2022F291
ISSN 2300-5963 ACSIS, Vol. 30

Diversified gradient boosting ensembles for prediction of the cost of forwarding contracts

Dymitr Ruta, Ming Liu, Ling Cen
EBTIC, Khalifa University, UAE
{dymitr.ruta,liu.ming,cen.ling}@ku.ac.ae

Quang Hieu Vu
ZaloPay, VNG Corporation, Vietnam
hieuvq@vng.com.vn

***Abstract*—A common business practice for transportation forwarders is to bid for shipping contracts at the transport or freight exchanges. Based on the detailed contract requirements they try to estimate the total expected cost of its execution and accordingly bid with the fixed price in advance for delivering such shipping service at the prescribed specification and schedule. The capability to accurately predict the cost of contract execution is the critical factor deciding about the profitability of offered shipping services as well as the amount of business drawn from freight exchanges. However, given highly volatile nature of the transport services ecosystem, it is difficult to simultaneously account for countless dynamically changing factors like fuel prices, currency exchange rates, temporal and spatial multitude of routing and implied traffic risks, the properties of cargo and shipping vehicles etc., which leads to big cost under- or over-estimation resulting with loss-making contracts or equally painful missed revenue opportunities. In the context of FedCSIS 2022 data mining competition we propose an accurate and robust predictor of the cost of forwarding contracts built upon the detailed contract data using the ensemble of the state-of-the-art gradient boosting-based regression models. Our established feature engineering framework combined with deep parametric optimization of the individual models and multi-faceted diversification techniques guiding hybrid final model ensembles were instrumental to outperform all the competitive predictors and win the FedCSIS 2022 contest.**

***Index Terms*—Cost Prediction of Forwarding Contracts, Gradient Boosting Trees, CatBoost, XGBoost, LightGBM, Stacking, Diversity, Model Diversification, Ensemble Learning.**

I. Introduction

With the development of IoT (Internet of things), e-commerce and continuous globalization, the business providing logistics service and involving supply chain for supply chain planning functions or transport management has become increasingly important. Big data analytics for intelligent transportation and prediction analysis with machine learning techniques have boosted management of transportation and logistics by providing intelligent solutions aiming for more efficient and safer transportation at cheaper cost. In the recent years, the technologies of data mining and machine learning have been applied to investigate a range of issues in international freight transportation, supply chain and logistics management, e.g. driver behavior analysis [2], [3], origin-destination parameter estimation [4], pavement maintenance [5], traffic control and forecasting [6], [7], [8], freight logistics [9], air traffic management [10], vehicle classification [11], travel time prediction [12], traffic pattern analysis [13], freight demand prediction [14], traffic volume forecasting [15], transportation cost forecasting [16], etc. A good literature review regarding utilizing machine learning on freight transportation and logistics applications has been published in [18].

The objective of the FedCSIS 2022 challenge [1][1], which is in cooperation with PTI and QED Software and sponsored by Control System Software[2], is to forecast the costs of forwarding contracts, which are, although rather useful in the business providing logistics service or involving supply chains, quite challenging since it can be affected by many static/dynamic and internal/external factors. Besides contract nature and transportation arrangement, the actual transportation cost is constrained by the factors such like fuel prices, currency exchange, drivers behavior, weather, traffic, market demand, etc [16]. There is quite little work published in the literature on cost prediction of forwarding contracts. In [17], AI based models were developed to predict the long-term cost of the logistics service, and attempted to construct a risk-aware interval for the prices to be offered in the bid, aiming to boost competitiveness in the application for tenders. In addition, historical data was used to develop statistical learning models for predicting the success likelihood of a tender based on the actual data and predicted service prices achieved from previous stage. The work proposed in [16] identified the most significant predictive criteria by a trapezoidal neutrosophic fuzzy analytical hierarchy process (TNF-AHP) and based on the criteria found the transportation cost was predicted with an artificial neural network (ANN) model, which, claimed by the authors, can also be employed in supply chain management and inventory control management.

In this paper, an ensemble learning model based on gradient boosting decision trees together with efficient feature engineering and model hyper-parameter optimization has been developed for predicting the costs of forwarding contracts to complete the task given in FedCSIS 2022 challenge. Gradient boosting decision trees (GBDT), developed in the late nineties, is a commonly used boosting methods for solving regression and classification problems in the form of an ensemble of decision trees as weak prediction model, which achieves state-of-the-art results for many commercial and academic applications [19], [20]. In the GBDT, each new model correlates

[1]https://knowledgepit.ai/fedcsis-2022-challenge/
[2]https://controlsystem.com.pl/

to the negative gradient of the system's loss function that is minimized by using the gradient descent method, which is successively fitted to delivery better estimation of dependent variables via training, resulting in gradual improvement of prediction accuracy. Three efficient GBDT implementations, i.e. XGBoost, CatBoost, and lightGBM, which have shown their powerful learning capabilities by many winning teams in a number of machine learning competitions, are employed to construct the ensemble learning model for forecasting the cost of contract forwarding in our method.

The remainder of the paper is organized as follows. The FedCSIS 2022 Challenge is briefly described in Section II. Data transformation and feature engineering is presented in Section III, followed with the description of the gradient boosting models, model diversity, and ensemble learning in in Sections IV, V and VI, respectively. The experimental results in Section VII. Concluding remarks are provided in Section VIII.

II. FedCSIS 2022 challenge

The FedCSIS 2022 data mining competition focused on the prediction of the costs of forwarding contracts' execution based on 6 years of detailed history of orders on the European transport exchange. The data contained both the general information about the contracts as well as detailed data of planned routes' segments including geo-located and timed path, specification of shipping vehicles and cargo and even financial details including daily currency rates and wholesale fuel prices. The objective of the competition was to develop a prediction model to accurately estimate the total cost of the contract execution based on all available data. The competitors were provided with the training data from 330055 contracts along with the true realized cost, as well a the testing data from 72452 contracts but without the realized cost. The knowledgepit.ai platform [3], on which the competition was hosted operated a leaderboard, which provided the feedback to the competitive model prediction submissions in a form of the preliminary RMSE score [4] computed over the unknown 10% of the testing set, while the final RMSE score for the complete testing set - constituting the final results, were provided after the submissions' closure.

III. Data transformation and feature engineering

The data provided by the competition organizers included already a well curated, cleaned and carefully selected set of features, however only main dataset providing general contract details was organized in a tabular format of one raw (record) per forwarding contract. The extended route data, on the other hand, contained detailed records of between 1 and 31 subsequent steps of the planned route segments of the same contract and hence it became immediately clear that in order to build a competitive cost prediction model all the individual steps data would have to be incorporated hence eventually somewhat aggregated per each contract. We have developed a generic aggregation filter and applied it all useful columns of the detailed route segments dataset to achieve per-contract aggregates. For numerical columns eleven self-explanative aggregators were applied: *'first', 'last', 'min', 'max', 'argmin', 'argmax', 'mean', 'mode', 'sum', 'range', 'std'*. For categorical columns the aggregation treatment was made dependent on the number of unique values. For more than 100 unique values the occurrence of each value was considered sparse enough to limit the aggregation to just the four operators of *'first', 'last', 'mode', 'nuq'*, where *'nuq'* simply denotes the number of unique elements. For categorical columns with fewer than 100 unique values aside of the above-mentioned 4 categorical aggregators we have also applied one-hot-encoding on the original feature and with thereby up to 100 new numerical columns we have applied again all the 11 above-mentioned numerical aggregators to receive quite a large number of final features in the order of thousands. To avoid redundancy and wasteful poor quality features we have automatically eliminated duplicate features and removed features with at most one unique value different than nan/null, which typically resulted in the final set of up to 2000 features.

We have also included features that measured country disagreement between the origin and destination, extracted days of the week, various expected segment duration and the prices of different type of fuel during the trip segment days. Among the alternative but less successful data preprocessing techniques we have explored flattening all trip segments along the single contract record of up to 31 possible segments as well as organizing the data as sequences of consistent route step segments.

IV. Gradient boosting models

Preliminary experiments on the main dataset very clearly revealed that gradient boosting models performed by far the best in terms of the reference predictive accuracy and actually quite well in terms of the computational cost, even comparing to simple linear regression and by far comparing to deep networks. Among gradient boosting models XGBoost, LightGBM, CatBoost were used and subsequently optimized throughout the competition. Their variants trained on different parameters were utilized for second level ensembles both executed by simple aggregation and by stacked retrained ensemble of gradient boosting model versions.

A. Individual models' parametric optimization

Current state of the art Machine Learning models are highly customized and flexible to accommodate a very wide range of different options, versions and parametric settings during the model build. Gradient boosting models are good examples of such models with tens of algorithmic, representational, modelling and statistical parameters available to tune in to best fit or represent the data and ultimately to learn robust regression function between the inputs and the continuous output that generalizes well on the previously unseen data.

Given a set of distinct models each with a large numbers of parameters to tune we have decided to apply fast greedy,

[3] https://knowledgepit.ai/

[4] https://en.wikipedia.org/wiki/Root-mean-square_deviation

rotational grid search for each of the gradient boosting models: XGBoost, CatBoost and LightGBM. Each optimizable parameter, whether numerical or categorical is assigned up to 5 unique values comprehensively covering the domain of this parameter. Contrary to the exhaustive parametric grid search, which given the numbers of parameters in our case would prove intractable, our method incrementally finds local optimum of a specific parameter with remaining staying fixed, before rotationally progressing to the next until no improvement can be found from any local change. To further boost the reliability of the best found configurations of parameters we have applied 5-fold cross-validation to rule out accidentally high performance, yet as a consequence to limit an additional cost of cross-validation the process of local optima search for each parameter was reduced to just a pair of neighbouring checks: above and below the current value per turn and shifting the current optimal to the value for which the maximum performance improvement was reported.

This parameters optimization process is terminated when no improvement in cross-validated RMSE performance was found from any local changes of parameters.

V. Model diversification techniques

Well performing but diverse models produce diverse outputs which after aggregation produce significant reduction of both variance and bias error components. The critical challenge here is how to develop diverse but well performing models and also to which degree worse performing but quite diverse models are still worth combining to achieve the performance gain. We have developed two generic diversification methods applicable to gradient boosting models. The first method focuses on maximizing the number and magnitude of differences between as many parameters as possible of the same model. The second method takes specific categorical focal feature with a few unique values and proceeds with training an array of models specific to each of the unique value of the focal feature. Both method yield good results with parametric-diversity achieving lower levels of output decorrelation but higher individual performances, while decompositional-diversity achieving higher diversity but lower performance mainly due to smaller number of training examples to train on.

A. Parametric model diversification

Parametric model diversification method was developed in conjunction with the parametric model optimization discussed above. The method simply retains model parametric configurations and the corresponding regression accuracy throughout the optimization process and tries to establish the the population of the best performing model versions with the most diverse configurations of the parameters. To assess the level of diversity among model versions' parametric configurations we developed a simple disagreement measure adding up the differences in grid positions of all parameters conceptually similar to the City Block (Manhattan) distance metric. Once all parameter configurations encountered throughout the optimization are evaluated in terms of their performance P and the diversity measure D, the final step involves selecting k best model versions from the performance-diversity profile which could be associated with the normalized ratio of $D/RMSE$ assuming our performance measure is $P = 1/RMSE$. Selected models versions outputs are then subsequently aggregated using the simple average operator.

B. Decompositional model diversification

It is known that certain level of model diversity, seen as the level of disagreement among model outputs, could be achieved by the training on mutually exclusive data sets. This effect can be further reinforced if the instead of training on random partitions of the training set, model versions are trained on partitions associated with the different values of certain categorical variables as they typically represent significantly different subsets of the available data. It can be justifiable argued, though, that any limitation or reduction of the training set size exposed to the model is likely to reduce its predictive performance. While it indeed could be the case, particularly for the small training data sets, we argue and have experimentally verified that when the data set is large enough and the categorical variable has only very few unique values, the benefits from combining the outputs from that way diversified model versions outweigh the negligible reduction in performance of a single model trained on the whole set. Given all our models utilize the decision tree construction mechanism in the back end, such guided data partitioning could be considered as a forced first splits that branch out into several categorical-variable-value specific tree-based model versions. Given 300k+ size of the training set we have identified several suitable categorical variables with only a few unique values and trained boosting models on subsets corresponding to specific values of these variables obviously without this variable included in the training process. Trained model versions were then applied separately to the testing sets to generate the outputs which were finally combined with the simple mean operator to produce a single output. Figure 1 comparatively illustrates how the decompositional model diversification differs from the traditional individual model build along the training and testing processes.

VI. Ensemble model

In the construction of the final ensemble, we have utilized 3 baseline gradient boosting models: XGBoost (XGB), LightGBM (LGBM), and CatBoost (CatB) subjected to both parameteric (DivP) and decompositional (DivD) diversification filters. Diversification techniques are designed to improve the classifier generalization performance by first expanding it into a number of diverse versions, train them on a whole or subsets of the training set and apply them to the testing set before merging the model versions' outputs back together by a simple aggregation. To further boost diversity but also in a search for better complementary predictive performance we have trained all baseline regression models with their DivP/D filters on two different subsets of features generated by our feature engineering engine. The only difference between these two

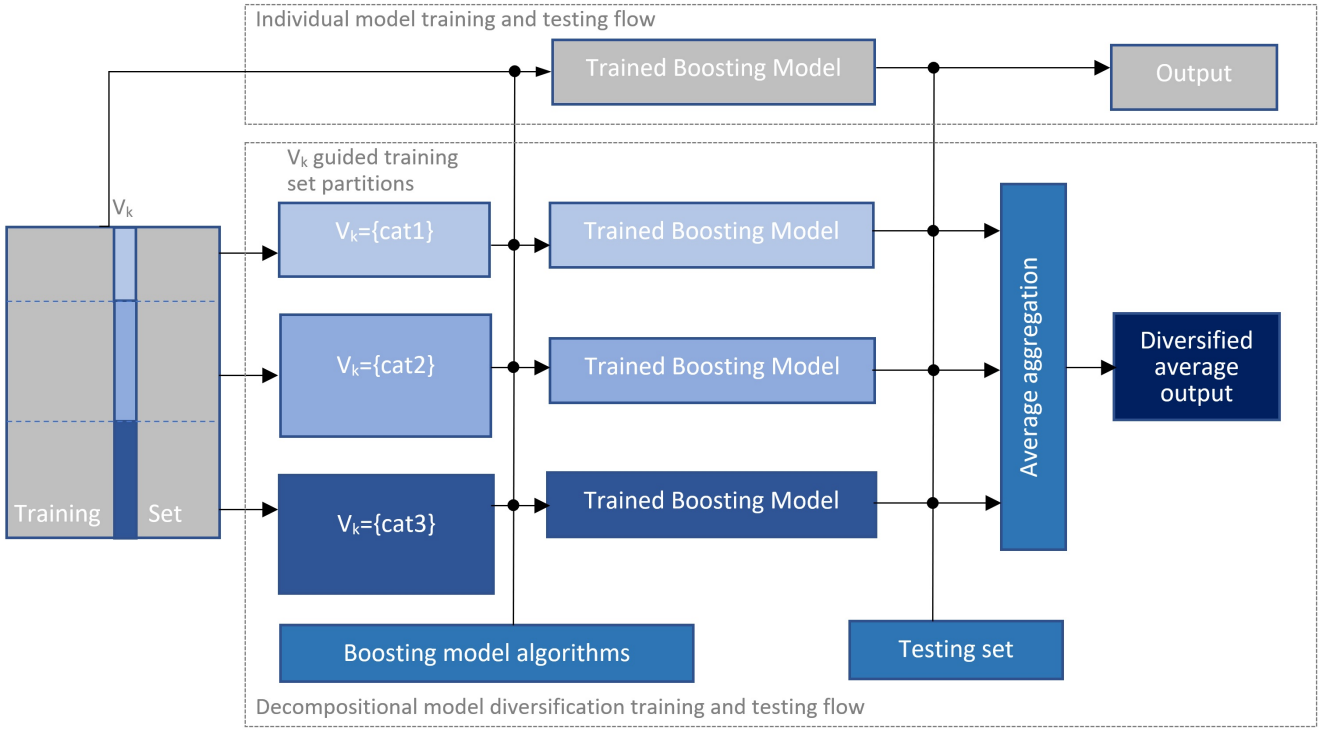

Figure 1. Decompositional model diversification compared to traditional individual model training and testing flow chart.

feature subsets were that the second set included many more sparse columns extracted from much more prolific application of one-hot-encoding to categorical features.

Moreover, in a search for further performance gains we have added another stacked layer of simple linear regression trained on the outputs from diversified baseline models. To properly accommodate stacking layer the training data were split into two parts, one used for building the baseline models and their diversified versions, while the other for learning the parameters of the linear regression in the stacking layer.

Eventually all diversified individual model outputs along with the outputs from linear regression based stacking were averaged together. The architecture or rather flow chart of the final ensemble is depicted in Figure 2.

VII. Experimental results

To establish a baseline predictability for the presented problem of forwarding cost prediction we first optimized individual gradient boosting models: XGB, LGBM and CatB on two above-mentioned subsets of extracted features and received the following RMSE results along with the optimal parameters returned by the optimization process.

1) Feature set 1 with 894 features:
   - CatB (learning rate 0.05, depth 8, iterations 2000): 0.1442
   - LGBM (learning rate 0.02, depth 8, iterations 2000): 0.1444
   - XGB (learning rate 0.02, depth 1000, iterations 1000): 0.1496

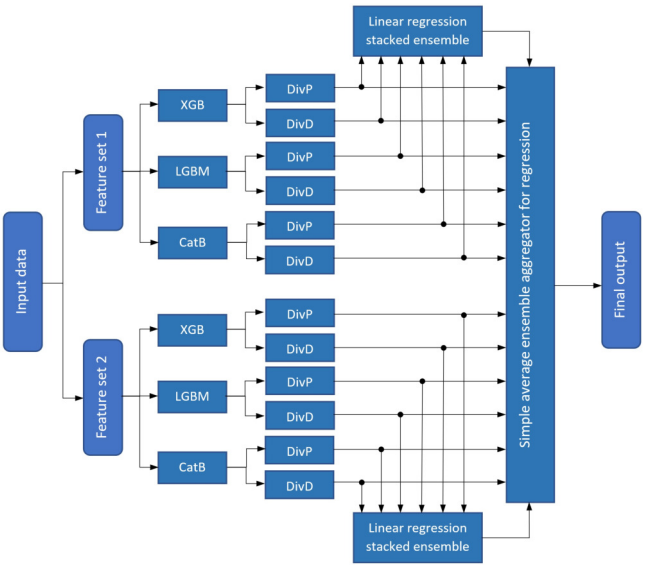

Figure 2. Flowchart of the final ensemble model.

2) Feature set 2 with 3431 features:
   - CatB (learning rate 0.05, depth 8, iterations 2000): 0.1513
   - LGBM (learning rate 0.02, depth 8, iterations 2000): 0.1494
   - XGB (learning rate 0.02, depth 1000, iterations 1000): 0.1640

Secondly, linear regression stacking models were trained on the outputs from the diversified individual models and reported the preliminary RMSE performance of:

- Stacking model with 894 features: 0.1441
- Stacking model with 3431 features: 0.1465

The final predictions are achieved by ensemble averaging of the outcomes from both stacking models along with diversified individual baseline models, yielding the RMSE around 0.141.

A. Visual efficient performance-diversity horizon method for a robust ensemble composition

Even though the ensemble model illustrated in Figure 2 resulted with the best preliminary performance, the journey of incremental model build and performance improvement led to hundreds of model output submissions as potential solutions each with different performance and mutual dependency with the other solutions. Given the submissions represented the outputs from multiple different models with many different versions of the training data, parametric setups and model design choices we have considered them as a final-stage resource for potential further performance improvement through simple aggregation. The question posed in this stage was: given n model outputs with preliminary RMSE scores is it possible to improve the best model result and if so how to achieve the biggest possible improvement.

The intuition backed by the bias-variance error decomposition theorem suggested that the best results should be achieved by combining the outputs from the best performing (model) solutions that differ the most from each other, inline with the diversification techniques discussed in section V. At the final stage when only model outputs and their preliminary scores were available the natural disagreement or resultant diversity measure that could be computed upon the model outputs was an average from the correlation coefficients between the specific model output and all the rest in the considered pool of solutions. After computing such outputs' diversity c_i for all solutions with the preliminary RMSE scores $e_i < 0.152$ we have plotted all such best solutions from our submissions as points (c_i, e_i) on the 2-dimensional diagram depicting dependency $e = f(c)$, as shown in Figure 3.

The points stretching along the bottom horizon approximately marked by the dashed line represent the continuum of the best performing and at the same time the most diverse solutions and are expected to be the most promising choices for final stage combination to achieve the performance improvement. Rather than arbitrarily take some solution for combination along such horizon we have developed a simple greedy algorithm to choose the best solution candidates from around the the horizon for final combination. The greedy sequence in our case is starting from the best performing model in the bottom right corner and then adding the next best model but only out of the more diverse solutions, i.e. from the top solution as a pivot the next solution added is represented by the lowest point to the left from the current pivot. Then the pivot is shifted to the newly added point and process repeats until no more points can be added. Such greedy sequential addition leads to the staircase connected set op points marked in black along the diversity horizon. The final stage is testing at what point such greedy sequence does not improve the overall ensemble performance any more.

Such method of output-level ensemble combination is particularly effective when a large number of black-box models are suddenly at the disposal and a quick decision is needed on which models' outputs are best to aggregate to maximally reduce the overall predictive error. In our cases the team merger pose an exact situation as described above and following a quick testing a final ensemble output has been generated by aggregation of the outputs from the first 11 models along the diversity horizon, and yielded the top predictive RMSE error below 0.14 and even better result on the full testing set, thereby securing the first place in the FedCSIS 2022 competition.

VIII. Conclusions

We have attempted to improve predictive performance of the already highly robust regression models from the gradient boosting family: XGBoost, LGBM, CatBoost. To achieve that we have proposed a range of model diversification methods coupled with various ensemble combination schemes. Decompositional diversity forced by training on significantly different input data subsets, combined with actively encouraged parametric diversity led to an improvement in performance achieved from aggregation of the expanded diverse model versions, additionally boosted with linear regression based stacking and output level selection of the most efficient ensemble candidates in terms of the performance-diversity trade-off. The proposed ensemble has been applied to the complex problem of advance prediction of the total realized cost of forwarding contracts' based on a variety of data coming in different forms and types, in the competitive setup of the FedCSIS 2022 data mining challenge. Our proposed solution scored the first place in this challenge producing the lowest (RMSE) error below 0.14, which corresponds to only 2% in relative cost in monetary units. The proposed solution can enable forward contractors to better estimate their expected shipment cost, further reducing their business risks and boosting the efficiency across transport services domain.

References

[1] A. Janusz, A. Jamiołkowski, M. Okulewicz, Predicting the Costs of Forwarding Contracts: Analysis of Data Mining Competition Results, *Proceedings of the 17th Conference on Computer Science and Intelligent Systems (FedCSIS)*, 2022.

[2] Z. Li, K. Zhang, B. Chen, Y. Dong and L. Zhang, Driver identification in intelligent vehicle systems using machine learning algorithms, *IET Intell. Transp. Syst.*, vol. 13, no. 1, pp. 40-47, 2019.

[3] S. Bhattacharya. Novel approach for Ai based driver behavior analysis model using visual and cognitive data, 2019.

[4] S. Kikuchi, R. Nanda and V. Perincherry. A method to estimate trip O-D patterns using a neural network approach. *Transp. Planning Technol.* 17(1): 51-65, 1993.

[5] A. Pozarycki. Pavement diagnosis accuracy with controlled application of artificial neural network, *The Baltic Journal of Road and Bridge Engineering* 10(4): 355-364, 2015.

[6] M. Bielli, G. Ambrosino, M. Boero and M. Mastretta. Artificial intelligence techniques for urban traffic control. *Transportation Research Part A: General* 25(5):319-325, 1991.

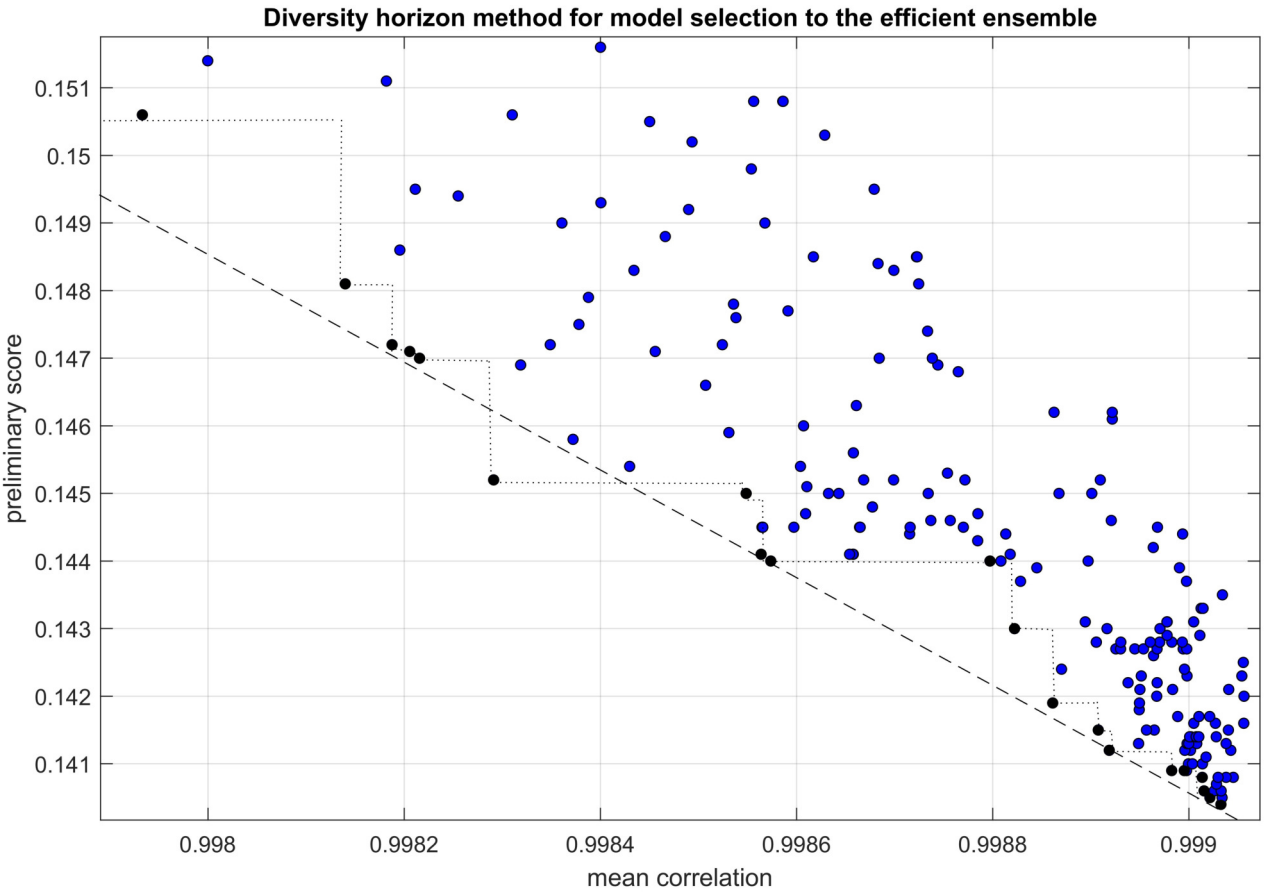

Figure 3. Diversity horizon method for model selections.

[7] M. Ghanin and G. Abu-Lebdeh. Projected state-wide traffic forecast parameters using artificial neural networks. *IET Intel. Transp. Syst.* 13(4):661-669, 2019.

[8] J. Lu, L. Feng, J. Yang, M. Hassan, A. Alelaiwi and I. Humar, Artificial agent: The fusion of artificial intelligence and a mobile agent for energy-efficient traffic control in wireless sensor networks, *Future Gener. Comput. Syst.* 95:45-51, 2019.

[9] Y. Kayikci. A conceptual model for intermodal freight logistics centre location decisions. *Proc. Soc. Behav. Sci.* 2(3): 6297-6311, 2010.

[10] F. Saadaoui, H. Saadaoui and H. Rabbouch. Hybrid feedforward ANN with NLS-based regression curve fitting for US air traffic forecasting. *Neural Computing and Appl.* 32:10073-10085, 2019.

[11] J. George, A. Cyril, B. Koshy and L. Mary. Exploring sound signature for vehicle detection and classification using ANN. *Int. J. Soft Comput.* 4(2):29-36, 2013.

[12] H. Lin, R. Zito and M. Taylor. A review of travel-time prediction in transport and logistics. *Proc. Eastern Asia Soc. Transp. Stud.* 5:1433-1448, 2005.

[13] H. Kirby and G. Parker. The development of traffic and transport applications of artificial intelligence: An overview. *Artificial Intelligence Applications to Traffic Engineering*, The Netherlands:VSP, pp. 3-27, 1994

[14] I.C. Bilegan, T.G. Crainic and M. Gendreau. Forecasting freight demand at intermodal terminals using neural networks–an integrated framework. *Eur. J. Oper. Res* 13(1):22-36, 2008.

[15] K. Kumar, M. Parida and V. Katiyar. Short term traffic flow prediction for a non urban highway using artificial neural network. *Proc. Social Behav. Sci.* 104:755-764, 2013.

[16] 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* 9:103497-103512, 2021.

[17] S. Nataraj, C. Alvareza, L. Sadaa, A. Juana, J. Panaderoa and C. Bayliss. Applying Statistical Learning Methods for Forecasting Prices and Enhancing the Probability of Success in Logistics Tenders. *Transportation Research Procedia (Elsevier)* 47:529–536, 2020.

[18] K. Tsolaki, T. Vafeiadis, A.N. Dimosthenis and I.D. Tzovaras. Utilizing machine learning on freight transportation and logistics applications: A review. *ICT Express*, 2022.

[19] L. Mason, J. Baxter, P.L. Bartlett, and M. Frean. Boosting Algorithms as Gradient Descent In S.A. Solla and T.K. Leen and K. Müller. Advances in Neural Information Processing Systems 12: 512–518, MIT Press, 1999.

[20] J.H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29(5): 1189-1232, 2001.