Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 451–454

DOI: 10.15439/2022F292
ISSN 2300-5963 ACSIS, Vol. 30

An Approach for Predicting the Costs of Forwarding Contracts using Gradient Boosting

Haitao Xiao[1,2], Yuling Liu[1,2*], Dan Du[1], Zhigang Lu [1,2]
[1]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China*
[2]*School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China*
{xiaohaitao, liuyuling, dudan, luzhigang}@iie.ac.cn

***Abstract*—Predicting the cost of forwarding contract is a severe challenge to road transport management system. The transportation cost of a forwarding contract often depends on many factors. It is hard for humans to evaluate the various factors in transportation and calculate the cost of forwarding contract. In this paper, we propose an approach to address such a problem by following the sequence of machine learning steps which consist of data analysis, feature engineering and model construction. First, we conduct a detailed analysis of the given data. Then, we generate effective features to characterize the cost of forwarding contract and eliminate redundant features. Finally, in the model construction phase, we propose a gradient boosting decision tree based method to train and predict the cost of forwarding contract. The proposed approach achieves RMSE scores of 0.1391 on the test set, which is the 2^{nd} final score in the competition.**

***Index Terms*—cost prediction, gradient boosting, model ensemble**

I. Introduction

COST PREDICTION is widely used in various fields, such as transportation[1], cybersecurity[2], construction[3], and healthcare[4]. Cost prediction is generally a method of studying historical data and predicting future costs. Effective cost prediction can help businesses better control costs and adjust management strategies for the future in a timely manner, allowing them to gain a competitive advantage. Transportation cost prediction is one of the aspects of cost prediction. Anitha and Patil [1] predicted the transportation costs using a regression algorithm, which assisted the retail sector predict the cost incurred for logistics. In the freight company, predicting the costs of freight forwarding contracts by using historical data can help freight companies better understand the causes of costs and select profitable contracts. Thus, from both academia and industry, predicting transportation costs has drawn a lot of attention. Accurate transportation cost prediction is still a challenging problem.

Based on the same background, the FedCSIS 2022 Challenge: Predicting the Costs of Forwarding Contracts[5] released the task to develop a predictive model that assesses the actual costs of individual orders as accurately as possible. In this challenge, six years of history of contract data and planned routes are provided. The key to the problem is to find the factors related to the cost.

In this paper, we propose an approach for predicting the cost of forwarding contracts using the gradient boosting method. First, we analyse the data given in this challenge and have a clear understanding of the meaning of each feature. The data analysis step also provides guidance for the following feature engineering. Then, in feature engineering, we generate the features that can effectively characterise the costs of forwarding contracts from contract data, planned routes, and historical wholesale fuel prices. We also remove redundant features after feature generation. The redundant feature elimination can reduce training time and the impact of noise on the training model. Finally, in the model construction phase, we propose a gradient boosting based method to train and predict the costs related to the execution of forwarding contracts. We introduce the model stacking mechanism as an ensemble method to enhance generalisation performance, which is a frequently used strategy in machine learning competitions. The experiments and competition results have both shown the effectiveness of the proposed approach.

In summary, this paper makes the following contributions:

- We analyse the given data and provide guidance for the following feature engineering. The guidance helps to generate effective features in feature engineering.
- We generate effective features from contract data, planned routes, and historical wholesale fuel prices for cost prediction. The generated features can improve the prediction performance significantly. And we also remove redundant features to reduce training time and the impact of noise on the training model.
- We propose an effective stacking approach using a gradient boosting based method to train and predict the costs of forwarding contracts, which achieves RMSE scores of 0.1402 on the preliminary testing subset and 0.1391 on the complete testing set. And we get the 2^{nd} final score in the competition.

The remainder of this paper is structured as follows: Section II introduces the FedCSIS'22 challenge. Section III provides the analysis of the data and the details of our proposed approach. Section IV shows the results of the experiments. Finally, conclusions are drawn in Section V.

[*] Corresponding author

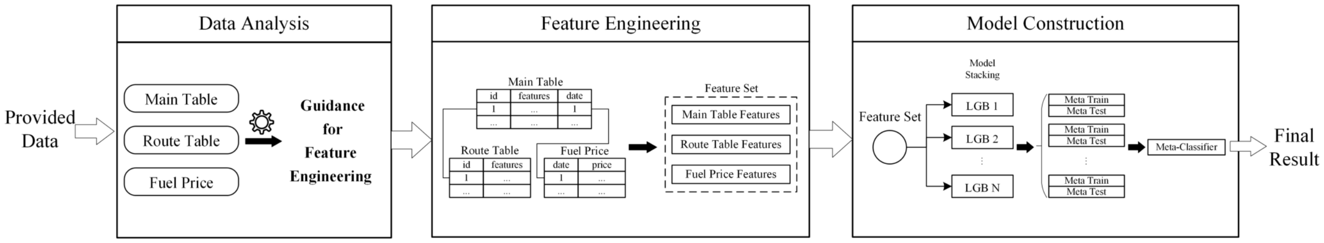

Fig. 1. Overview of proposed approach

II. FedCSIS 2022 Challenge

In this section, we will briefly introduce the FedCSIS 2022 Challenge, titled as Predicting the Costs of Forwarding Contracts.

The task in this challenge is to predict the costs related to the execution of forwarding contracts based on contract data and planned routes. Accurate cost prediction of forwarding contracts can support freight forwarders in selecting profitable contracts. The data provided in the challenge is collected by Control System Software, which is a software company that has been delivering solutions for the transportation, spedition, and logistics industry for 20 years[6].

The data set provided in this challenge contains a six-year history of orders appearing on the transport exchange. Details of the dataset are presented in the Section III. The aim of the competition is to develop a predictive model using the training set that assesses the actual costs of individual orders as accurately as possible.

The preliminary scores of submitted solutions are computed on a small subset of the testing data and published on the public leaderboard. The final evaluation is published using all of the test records after the competition ends.

III. Methodology

In this section, we describe in detail our proposed approach for predicting the cost of forwarding contracts. An overall framework of our proposed approach is provided in Figure 1. The approach includes data analysis, feature engineering, and model construction.

A. Data Analysis

TABLE I
Brief view of the competition data

Data	Format	Size (train/test)
Main Table	csv	69.9 MB / 14.2 MB
Routes Table	csv	204 MB / 58.2 MB
Fuel Prices	csv	68.1 KB

There are two data tables and an additional set of data provided in this competition. A brief view of the data is shown in Table I. The main table contains basic information about the contracts, and the routes table describes the main sections of the planned routes associated with each contract. The main table and routes table are linked by “id_contract”. One contract usually consists of several route sections. The additional set of data is fuel prices, which contains historical wholesale fuel prices for the period of training and test data.

The main table records the details of individual contracts. The main table contains unique contract ids and their related information, as well as the expenses of each contract. The associated information of a contract id includes 1) the general information of contract such as payer, currency, direction, contract type, service type, duty count, planned time. 2) the basic characteristics of the shipped goods, such as refrigeration, temperature, and maximum weight. 3) expected route information, such as longitude and latitude of loading and unloading positions, longitude and latitude of route start and end positions, kilometers to be covered according to the route plan, ferries and trains usage. As these data are important to predict the cost of forwarding contract, they form the core set of features used to train our model.

The route table records the main sections of each contract’s planned route. The route table has more detailed information about the planned route, which contains all route steps of each contract. The associated information of a route step includes 1) general information of route step such as the sequence number, step type, latitude and longitude of the end route step point, city, address, and estimated time. 2) information of ferry and train usage. 3) the vehicle and trailer status information. These data contain information about all of the route steps, which can help our model understand the specific route composition in a contract and assess the cost of a contract at the granularity of route step.

The fuel price table records historical wholesale fuel prices from 2016-01-01 to 2021-11-30, which covers the period of training and test data. This table describes three different types of fuel price information. The cost of transportation can be directly affected by the price of fuel. Thus, it is also an important component of freight forwarding contract costs.

Among these data tables, the main table provides the base information for the competition task. We can explore the properties for further feature engineering when we have a clear understanding of the meaning of the data and features. First, we analyse the expenses column, which is a continuous value to indicate the cost of forwarding contract. The min value of expenses is 2.879139, and the max value is 9.598065. It can be regarded as a regression task. Then, we find that the route table contains information about all route steps, and the fuel price table can indicate the fuel price at the time of the contract. The route table and fuel price table can provide more specific information for predicting the cost of forwarding contract. Thus, it is necessary to extract more information related to the cost from the route table and fuel price table. This makes for more targeted feature engineering and more representative extracted features.

B. Feature Engineering

Following the data analysis phase described above, we generate three types of new features from the main table, route table, and fuel price table. We remove redundant features after all new features have been generated. In the following subsections, three types of new features and the process feature selection are comprehensively presented.

1) Main Table Features: Main table features are mainly generated from the main data table summarizing contract's definition and can indicate the transportation costs. As the temperature feature is quite messy, we perform a simple data clean to correct the temperature in different formats. Then, we generate a series of new features to characterise the cost of contract. The generated features of main table can be divided into three parts: 1) basic features such as duration time, the haversine distance from route start to end location and load to unload location, the ratio of kilometers to be covered with empty trailer and so on. 2) cross features such as "direction×contract_type", "first_load_country×last_unload_country, and so on. The cross features can characterize the links between different category features. 3) time features such as the year of route start time, the quarter of route start time, and so on. The time features can characterize the impact on costs at different times.

2) Route Table Features: Route table features are mainly generated from the route data table based on "id_contract". Each contract has a different number of route sections. We aggregate the section attributes of each contract in different ways. For numerical attributes, we aggregate them by count, sum, and ratio operations. We also use statistical methods like mean, max, min, and median to summarise the numerical features of each contract. For category attributes, we aggregate them by counting their occurrences. Moreover, we count top 1000 cities and address to characterize the geographical situation of each contract.

3) Fuel Price Features: Fuel price features are mainly extracted from the fuel price data table. As it is generally to assess the cost before the contract starts, we directly merge the fuel price into each contract that matches the route start time. The merged fuel prices can represent current fuel price levels at the beginning of the contract.

4) Feature Selection: The feature set has over 2600 features after all the new features generated. We remove any features that are redundant or duplicate. A redundant feature is defined as the percentage of the value of a particular feature that is greater than 99.9%. Finally, we get 1092 features after simple feature selection.

C. Model Construction

Gradient boosting decision tree[7] is an model which uses decision tree as weak learner and improves model quality with a boosting strategy[8]. The gradient boosting based method has been shown to achieve superior performance in various machine learning tasks, such as prediction[9] and ranking[10]. Due to its excellent performance and high accuracy, we choose the gradient boosting based method as our base model. There are multiple implementations of gradient boosting based method, like XGBoost[11], LightGBM[12], and CatBoost[13].

In model construction, we try various gradient boosting based methods to train the selected features and select the best method as the base model of our final solution. We finally choose LightGBM as our base model for its ability to handle the high dimensions of features and high efficiency. We also propose an ensemble approach, which introduces the model stacking mechanism, to improve the generalisation performance. The generalisation ability of an ensemble approach is usually much stronger than that of base learners[14].

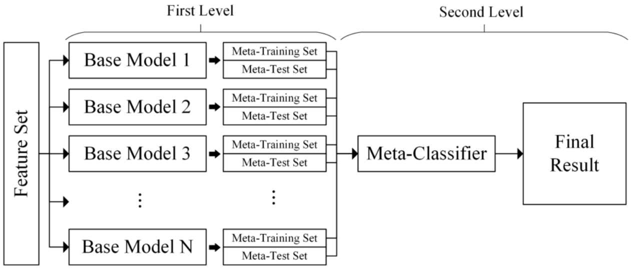

Fig. 2. The framework of proposed ensemble approach

The proposed ensemble approach can be separated into two levels, as shown in Figure 2. First, the training set is split into two parts, one part is used for first-level model training which is LightGBM, the other part, along with the test set, is used for prediction by the trained first-level model. Then, we use the predicted results from the first-level model to train the meta classifier. Typically, the meta classifier is a simple linear model. Therefore, we choose ridge regression as the meta classifier. The predicted result of the meta-classifier is our final result.

IV. Experiment and Result

A. Experiment Setup

1) Environment: The operating environment is Ubuntu 21.10, memory at 128GB, an Intel Xeon Silver 4210 CPU @ 2.20GHz, with 40 physical processors.

2) Toolkit: For feature engineering implementation, we use Pandas 1.2.4, Numpy 1.20.1 to generate new features. We also use LightGBM 3.3.2 as the implementation of our base model.

3) Evaluation Metric: We take the root mean square error (RMSE) as the evaluation measurement, which is exactly the same as the evaluation criterion used for the competition.

B. Experiment Result

To ensure the scores between the local validation and the actual testing results remain within a certain range, it is important to have a good validation strategy. Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. In typical cross-validation, the training and validation sets must cross over in successive rounds such that each data point has a chance of being validated against[15]. Since the labels are continuous values, we use the standard k-fold cross-validation as our local validation strategy. To determine the value of k, we tried different values of k and eventually found that the gap between local validation and online validation was smallest when $k = 3$. Thus, we set $k = 3$ and the gap between our local validation score and the public leaderboard score was less than 0.014.

In the experiment, we use different feature sets to evaluate each set's performance and the improvement of the model stacking. The experimental results are shown in Table II. In this table, "$Baseline$" represents the feature set contained in the original main data table, "$FeatureSet_1$" represents the merging feature set of the original main data table and generated main table features, "$FeatureSet_2$" represents the merging feature set of $FeatureSet_1$ and generated route table features, "$FeatureSet_3$" represents the merging feature set of $FeatureSet_2$ and generated fuel price features, and "$Stacked$" represents the model stacking based on the $FeatureSet_3$.

TABLE II
THE EXPERIMENTAL RESULTS OF DIFFERENT FEATURE SETS

Feature Set	RMSE Score(Local Validation)
$Baseline$	0.1463
$FeatureSet_1$	0.1398
$FeatureSet_2$	0.1276
$FeatureSet_3$	0.1275
$\boldsymbol{Stacked}$	**0.1267**

From Table II, it can be derived that: 1) From the result of training each feature set, the RMSE score get smaller and the performance gets better as the feature increase. The local validation result of $FeatureSet_3$ is 0.1275, which improves 0.0188 RMSE scores compared to the $Baseline$. The experimental results prove the effectiveness of our proposed feature engineering. 2) Comparing the base model and stacked model, the stacked model "$Stacked$" has 0.1267 RMSE scores, which improves by 0.0008 RMSE scores compared to the base model $FeatureSet_3$. The results of the comparison show that the model stacking strategy can improve the generalisation performance of the base model. Both the results of local validation and the public leaderboard can prove the effectiveness of our proposed approach.

V. Conclusion

In this work, we propose a gradient boosting based approach to predict the cost of forwarding contracts. We first analyse the data related to the freight forwarding contracts and provide the gudiance for the feature engineering. Then, we focus on the feature engineering step to generate new features from the given data which can effectively characterise the cost of contracts. Finally, we present an ensemble approach that introduces the model stacking mechanism to improve the generalisation performance of base models. Both the results of our self-validation and the competition have shown that our proposed approach is competitive. Future work can focus on trying a deep learning model as the base model for modelling the contract data and considering time as a trend factor for the features.

Acknowledgment

This research is supported by National Key Research and Development Program of China(No.2021YFF0307203, No.2019QY1300), and Youth Innovation Promotion Association CAS (No.2021156), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02040100). This work is also supported by the Program of Key Laboratory of Network Assessment Technology, the Chinese Academy of Sciences, Program of Beijing Key Laboratory of Network Security and Protection Technology. We thank the anonymous reviewers for their feedbacks on the paper and volunteers of the FedCSIS 2022 Challenge.

References

[1] P. Anitha and M. M. Patil, "Forecasting of transportation cost for logistics data," in *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. IEEE, 2021. doi: 10.1109/CONECCT52877.2021.9622576 pp. 01–06.

[2] R. Leszczyna, "Cost of cybersecurity management," in *Cybersecurity in the Electricity Sector*. Springer, 2019, pp. 127–147.

[3] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, "A novel construction cost prediction model using hybrid natural and light gradient boosting," *Advanced Engineering Informatics*, vol. 46, p. 101201, 2020. doi: 10.1016/j.aei.2020.101201

[4] M. A. Morid, O. R. L. Sheng, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Healthcare cost prediction: Leveraging fine-grain temporal patterns," *Journal of biomedical informatics*, vol. 91, p. 103113, 2019. doi: 10.1016/j.jbi.2019.103113

[5] (2022, Jun.) Fedcsis 2022 challenge: Predicting the costs of forwarding contracts. [Online]. Available: https://knowledgepit.ml/fedcsis-2022-challenge/

[6] 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.

[7] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001. doi: 10.1214/aos/1013203451

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000. doi: 10.1214/aos/1016218223

[9] H. Xiao, Y. Liu, D. Du, and Z. Lu, "Wp-gbdt: An approach for winner prediction using gradient boosting decision tree," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021. doi: 10.1109/BigData52589.2021.9671688 pp. 5691–5698.

[10] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin, "Parallel boosted regression trees for web search ranking," in *Proceedings of the 20th international conference on World wide web*, 2011. doi: 10.1145/1963405.1963461 pp. 387–396.

[11] 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*, 2016. doi: 10.1145/2939672.2939785 pp. 785–794.

[12] 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," *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.

[13] 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*, 2018, pp. 6639–6649.

[14] Z.-H. Zhou, *Ensemble learning*. Springer, 2021, pp. 181–210.

[15] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.