Prediction of dwell time of railway freight cars at the terminal based on Gradient Boosting Regression Tree

Anqi Shi[1, a], Baotian Dong[1, *], Fangcan Zhao[1, b], Yang Wang[2, c]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China.

[2]Institute of Planning and Standards, National Railway Administration of the People's Republic of China, Beijing 100891, China.

*Corresponding author:
btdong@bjtu.edu.cn, [a]17120976@bjtu.edu.cn, [b]15114244@bjtu.edu.cn, [c]13661380301@139.com

Keywords: Gradient Boosting Regression Tree, Railway Freight, Dwell Time.

Abstract: Railway transportation is an important part of China's transportation system. Due to the original business model of railway, its punctuality is poor. In addition, the current calculation method of dwell time of railway freight cars is still be the general old-fashioned algorithm. In order to improve the punctuality of dwell time of cars at the station, this paper takes the terminal as an example, and puts forward a prediction model based on Gradient Boosting Regression Tree. The influencing factors of cars' dwell time in terminal are characterized. Six influencing factors are selected, i.e. car number, goods type, car type, start station, end station and train number. The six discrete characteristics are quantified. The six characteristics are combined into input vectors, and the data in 2016 and July 2017 are selected for prediction. The experimental results show that the model has a good prediction effect.

1. Introduction

With the rapid development of national economy, the express delivery industry has developed rapidly. Due to the strong punctuality of express delivery industry, some transportation goods, such as LTL cargo, are more in favor of using express delivery. In addition, with the rapid development of information technology in recent years, express delivery can be monitored in real time in the transportation process, observing the distribution progress and geographical location, which makes the market of express delivery has been further developed. Due to the low punctuality of cars, railway transportation which is an important part of China's transportation system, has caused a lot of waste of labor power, material and financial resources, but also limited the development of railway transportation. Due to the original model of railway transportation, the dwell time of railway freight cars in each station cannot be accurately calculate, which is also one of important reason for the low punctuality rate of cars. In order to solve the problem of poor punctuality of cars, if we can predict the dwell time of cars in each station step by step and improve the accuracy of prediction, it will be very helpful to improve the car punctuality. Secondly, China's railway has a long history of development. With the rapid development of computer storage technology and database technology, the railway has accumulated a lot of historical data information. If we analyze and mining useful data information, it will improve the accuracy of the prediction of the dwell time of cars at the station. Because different railway stations often have different situation, the prediction model and results are often different, and different date (National Day and other major holidays) will also affect the prediction results. Therefore, this paper will take the terminal as an example and select the historical data in July 2016 and 2017 to predict the arrival time of cars.

At present, the calculation of dwell time of railway freight cars is still by traditional way, which mainly consists of three parts: special operation time, distribution period and transportation time. If the result is less than 3 days, it will be calculated as 3 days. This calculation method is very general, in fact, for different date (such as holidays or ordinary working days), weather and other factors will also affect the prediction result. Therefore, compared with the traditional way, machine learning will be very helpful to improve the punctuality rate. There are many kinds of machine learning algorithms used in regression prediction, such as neural network, K-means, SVR, etc., but according to the characteristics of railway data, there are two aspects to consider: 1) if you choose too complex model, it is easier to make the results over fitting, so as to reduce the accuracy of prediction; 2) many black box algorithms are poor interpret the model. Based on the above considerations, this paper proposes a strong learning algorithm- Gradient Boosting Regression Tree (GBRT), which is composed of many weak learning algorithms. This machine learning algorithm can effectively prevent over fitting and has good generalization. GBRT has good prediction ability.

2. The Theory of Gradient Boosting Regression Tree

GBRT mainly combines the ideas of Regression Decision Tree and Boosting Decision tree, and proposes to use residual gradient to optimize the integration process of regression tree.

2.1. Regression Decision Tree

Regression Decision Tree is a binary decision tree constructed recursively according to the principle of minimizing square error. X and Y are input and output variables respectively. X is composed of multiple eigenvectors. Y is a continuous variable. Given the training data set, i.e. cars' dwell time.

$$\mathrm{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \tag{1}$$

Regression Decision Tree divides the input space (i.e. the feature space) into M units $\{R_1, R_2, \dots, R_M\}$. Each node of the Regression Decision Tree corresponds to a unit, which has a fixed output value of c_m. When the input characteristic is x, the Regression Decision Tree will determine it to a node, and take the output value c_m corresponding to the node as the output of the Regression Decision Tree. In this way, the Regression Decision Tree model can be expressed as

$$\mathrm{T}(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{2}$$

$I(x \in R_m)$is the indicator function. When the Regression Decision Tree determines that x belongs to R_m, its value is 1, otherwise it is 0.

The Regression Decision Tree is aim to select the appropriate spatial division method (i.e. the generation method of decision tree) and corresponding output value for data set D to minimize the square error.

$$\sum_{x_i \in D} \big(y_i - T(x_i)\big)^2 \tag{3}$$

First of all, choose the appropriate way of space division. According to the way of constructing decision tree, select the jth dimension feature of variable x ($x[j]$ represents the value of jth dimension of x) and corresponding threshold s at each decision node as the segmentation feature and threshold, then the node divides the space into two regions:

$$R_1(j, s) = \{x | x[j] \leq s\} \text{ 和 } R_2(j, s) = \{x | x[j] > s\} \tag{4}$$

The optimal segmentation feature j and segmentation threshold s at this node are found as follows

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right] \tag{5}$$

This node divides the sample set into two sub sample sets according to the segmentation feature j and the segmentation threshold s. The specific purpose of Eq. (5) is to minimize the variance of the two sub sample sets. In formula (5), the values of c_1 and c_2 are the mean values of each sample, and the selection of j and s is determined by traversal. The segmentation threshold s is a continuous variable, but its value can be selected according to the actual distribution of samples, without continuous traversal.

Then, determine the output value of the regression tree. For each sub region R_m (the node of the tree), the corresponding output value c_m can directly select the mean value of category, that is

$$c_m = arg\min_c \sum_{x_i \in R_m} (y_i - c)^2 = aver(y_i | x_i \in R_m) \tag{6}$$

$aver$ is the mean value.

2.2. Boosting Decision Tree

Boosting Decision Tree is actually a simple superposition of multiple decision trees, which can be expressed as

$$f_M(x) = \sum_{m=1}^{M} T(x; \theta_m) \tag{7}$$

$T(x; \theta_m)$ represents the Decision Tree, and θ_m represents the parameters of the Decision Tree; M is the number of trees.

For sample $\mathrm{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, the training of Boosting Decision Tree is to select the parameter $\theta = \{\theta_1, \theta_2, \dots, \theta_M\}$ to minimize the loss function $\sum L(y_i, f_M(x_i))$.

$$\arg\min_\theta \sum_{i=1}^{N} L(y_i, f_M(x_i)) = arg\min_\theta \sum_{i=1}^{N} L(y_i, \sum_{m=1}^{M} T(x; \theta_m)) \tag{8}$$

The loss function is used to reflect the difference between the output $f_M(x_i)$ of the Boosting Decision Tree and the sample label y_i. Here, the square error loss function can be selected:

$$\mathrm{L}(y, f(x)) = (y - f(x))^2 \tag{9}$$

According to Eq. (7), the Boosting Decision Tree can also be expressed as an iterative process

$$f_m(x) = f_{m-1}(x) + T(x; \theta_m), m = 1,2, \dots, M \tag{10}$$

Therefore, the training of the Boosting Decision Tree can also be completed according to the iterative process. Through m times of iteration, a new decision tree $T(x; \theta_m)$ is generated.

Specifically, initialize the boosting tree $f_0(x) = 0$, and then determine the mth decision tree $T(x; \theta_m)$, i.e. select the appropriate parameter θ_m of decision tree to minimize the loss function.

$$\hat{\theta}_m = arg\min_{\theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x; \theta_m)) \tag{11}$$

The solution of Eq. (11) is the key to the Boosting Decision Tree. If the square loss function is used, then

$$L(y_i, f_{m-1}(x_i) + T(x; \theta_m)) = [y_i - f_{m-1}(x_i) - T(x; \theta_m)]^2 = [r_{m,i} - T(x; \theta_m)]^2 \tag{12}$$

Here, $r_{m,i} = y_i - f_{m-1}(x_i)$ represents the residual of model $f_{m-1}(x)$ fitting data (x_i, y_i).

In this way, the solution of equation (11) becomes to select the appropriate parameter θ_m of decision tree, so that the error between the output $T(x; \theta_m)$ and the residual $r_{m,i}$ of the decision tree is as small as possible. Therefore, you can use $\{(x_i, r_{m,i})\}_{i=1,2,\dots,N}$ as the sample set of decision tree $T(x; \theta_m)$, the optimal value of parameter θ̂m is obtained according to the conventional decision tree generation process.

2.3. Gradient Boosting Regression Tree

The Gradient Boosting Regression Tree combines the idea of Regression Decision Tree and Boosting Decision Tree, and generalizes them to more general cases. The calculation of the residual in the Boosting Decision Tree is done when the loss function is a square loss function. If the loss function is a logarithmic function, it is not very convenient to calculate the residual $r_{m,i}$. When training the regression tree, it is not easy to calculate the output value of the node C_m.

The Gradient Boosting Regression Tree uses the approximate method to calculate the approximate value of the residual.

$$r_{m,i} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{13}$$

When calculating the output value, if the loss function is the square loss function, the solution is consistent with the regression decision tree, and the average value can be taken directly. If it is other loss function, it needs to be solved specifically. Specifically, we take the derivative as zero to solve the equation

3. Data preprocessing

3.1. Calculation of dwell time of railway freight cars

At present, the data about cars is the packet data which is the operation types of railway freight cars in each station of the whole country, so it is necessary to calculate the dwell time of cars at the terminal.

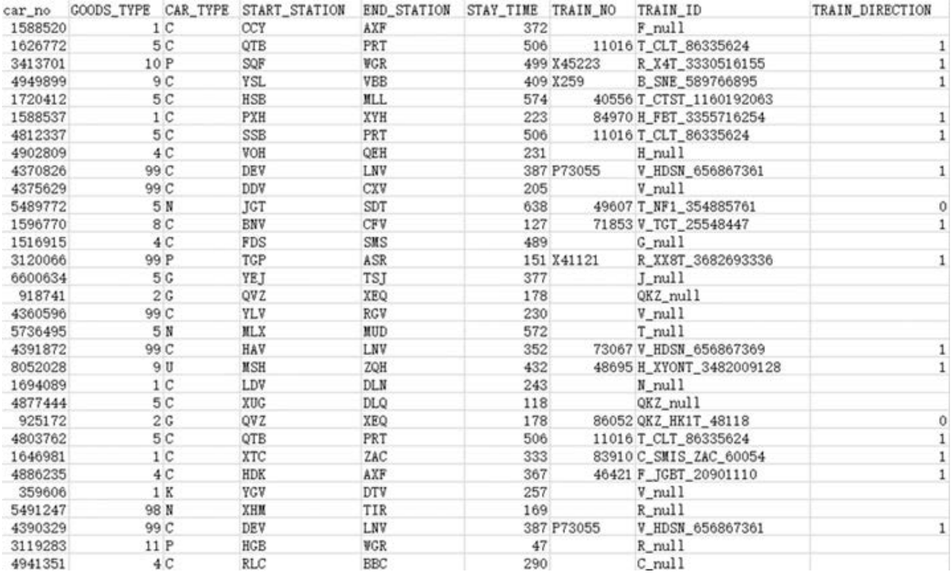
To get the dwell time of the cars at the terminal, first determine it by current station and end station. When the current value of the car at the terminal is the same as that of the target terminal, it indicates that the car is carrying out relevant information at the terminal. Therefore, after finding all the cars' information at the terminal, find the correct "LCDD" (train arrival) and corresponding to each car's "XCBG" (unloading report), the time corresponding to "XCBG" (unloading report) minus the time corresponding to "LCDD" (train arrival) is used to get the dwell time of the cars at the terminal station. The specific steps are as follows:

Step 1: query all the field information, and combine the car no, car type, start station, current station, end station, goods type, train no, train id and other fields into the new filed (od info), and arrange the queried data according to the car no, report type fields.

Step 2: take out the data whose report type field is "LCDD" (train arrival) in the query, take out its od info field information and put it into the map, take out the data in the next row in turn, if the field information of od info is consistent, and the report type field is "XCBG" (unloading report), put it into the corresponding map; otherwise, look up the data whose report type field is "LCDD" (train arrival), Continue to cycle.

Step3: take the value from the map. If there are both "LCDD" (train arrival) and "XCBG" (unloading report) in the map, and the corresponding time of "XCBG" (unloading report) is greater than the corresponding time of "LCDD" (train arrival), subtract the corresponding time of "LCDD" (train arrival) from the corresponding time of "XCBG" (unloading report).

Based on the preliminary calculation of the original data, we get the preliminary cleaning data including the cars' dwell time, which is shown in Figure 1.



car_no	GOODS_TYPE	CAR_TYPE	START_STATION	END_STATION	STAY_TIME	TRAIN_NO	TRAIN_ID	TRAIN_DIRECTION
1588520	1	C	CCY	AXF	372		F_null	
1626772	5	C	QTB	PRT	506	11016	T_CLT_86335624	1
3413701	10	P	SQF	WGR	499	X45223	R_X4T_3330516155	1
4949899	9	C	YSL	VBB	409	X259	B_SNE_589766895	1
1720412	5	C	HSB	MLL	574	40556	T_CTST_1160192063	
1588537	1	C	PXH	XYH	223	84970	H_FBT_3355716254	1
4812337	5	C	SSB	PRT	506	11016	T_CLT_86335624	1
4902809	4	C	VOH	QEH	231		H_null	
4370826	99	C	DEV	LNV	387	P73055	V_HDSN_656867361	1
4375629	99	C	DDV	CXV	205		V_null	
5489772	5	N	JGT	SDT	638	49607	T_NF1_354885761	0
1596770	8	C	BNV	CFV	127	71853	V_TGT_25548447	1
1516915	4	C	FDS	SMS	489		G_null	
3120066	99	P	TGP	ASR	151	X41121	R_XX8T_3682693336	1
6600634	5	G	YEJ	TSJ	377		J_null	
918741	2	G	QVZ	XEQ	178		QKZ_null	
4360596	99	C	YLV	RGV	230		V_null	
5736495	5	N	MLX	MUD	572		T_null	
4391872	99	C	HAV	LNV	352	73067	V_HDSN_656867369	1
8052028	9	U	MSH	ZQH	432	48695	H_XYONT_3482009128	1
1694089	1	C	LDV	DLN	243		N_null	
4877444	5	C	XUG	DLQ	118		QKZ_null	
925172	2	G	QVZ	XEQ	178	86052	QKZ_HK1T_48118	0
4803762	5	C	QTB	PRT	506	11016	T_CLT_86335624	1
1646981	1	C	XTC	ZAC	333	83910	C_SMIS_ZAC_60054	1
4886235	4	C	HDK	AXF	367	46421	F_JGBT_20901110	1
359606	1	K	YGV	DTV	257		V_null	
5491247	98	N	XHM	TIR	169		R_null	
4390329	99	C	DEV	LNV	387	P73055	V_HDSN_656867361	1
3119283	11	P	HGB	WGR	47		R_null	
4941351	4	C	RLC	BBC	290		C_null	

Figure 1. Information form of railway freight cars dwell time.

3.2. Characteristic quantization and selection

First of all, cargo type, car type and start station are all likely to be the influencing factors of cars' dwell time. The influencing features are basically discrete features, in order to mine more factors that affect cars' dwell time, we must first quantify all discrete features.

Discrete data is commonly known as categorical data. Machine learning cannot directly process the categorical data, so we need to process and transform the data. Categorical data that can be classified is a discrete value, which means that they belong to a limited category. In the response variables predicted by the model, they are often called categories / tags. These discrete values can be text or numbers (even unstructured data such as images). There are two kinds of classification data: ordinal and nominal. The values of ordinal classification have certain meaning or concept of order. There is no concept of order between the values in nominal. Generally speaking, there is no general module or function that can automatically transform and map these features to numerical representation according to these sequences. You can use a custom coding / mapping scheme.

Since our discrete features belong to the nominal data, we can see the basic performance of each feature by using the customized coding scheme after quantifying the discrete features, as shown in Figure 3. From left to right, we can see the goods_type, car_type, start_station, end_station and stay_time.

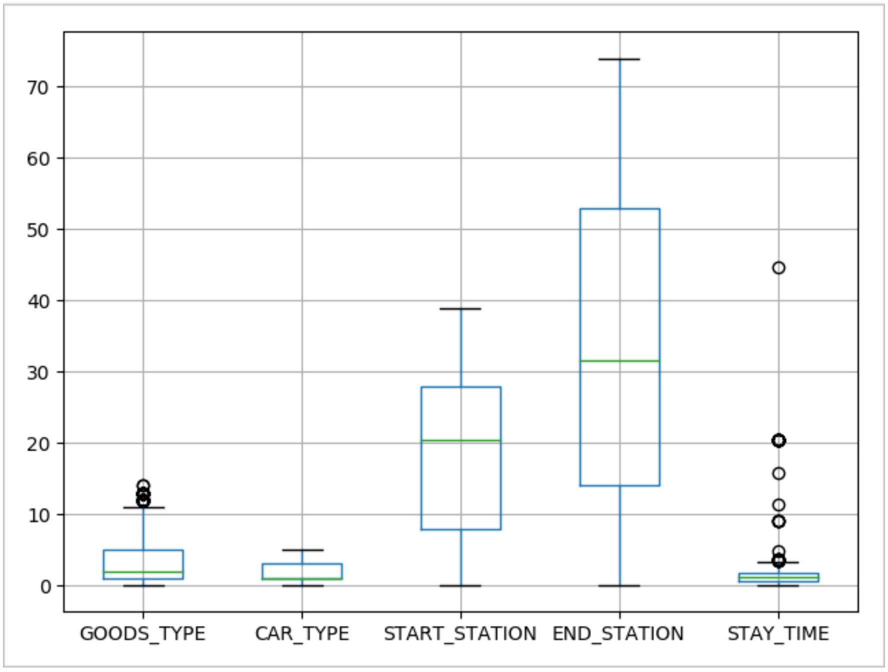


Figure 2. The box plot of characteristic.

Through the box plot, you can directly know some basic conditions of each feature, as shown in Figure 2. You can clearly see the outlier, upper quartile value, lower quartile value, upper limit value, lower limit value and median of each feature value. There is no outlier for car type, start station and end station, and the data distribution is very even, especially for start station and end station.

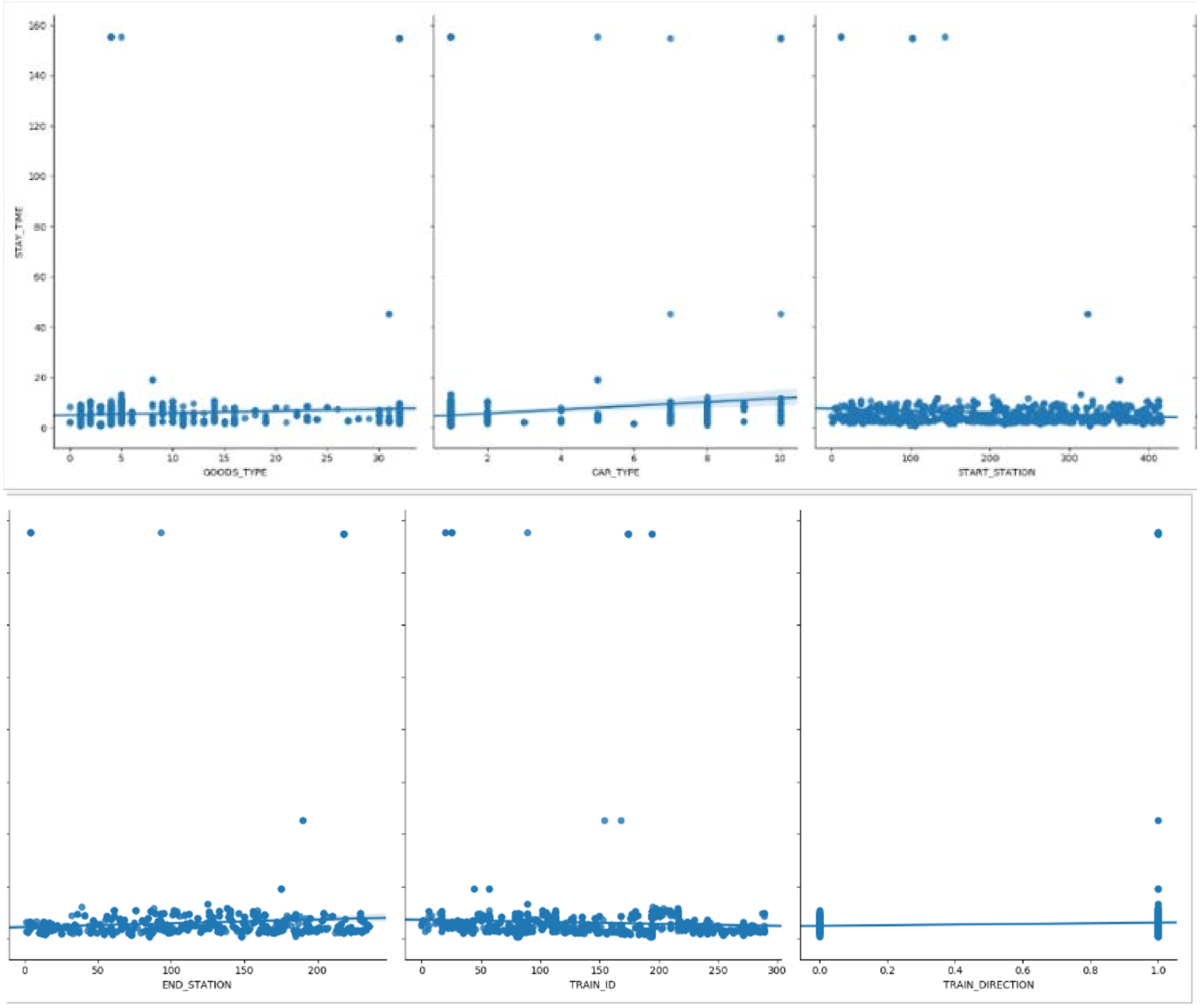


Figure 3. Scatter plots and fitting lines with 95% confidence interval.

After comparing the box plot of important features, the relationship between each feature and dwell time is compared. Set at 95% confidence interval, and add a fitting line, as shown in Figure 3. From left to right, from top to bottom, it is good_type, car_type, start_station, end_station, train_id and train_direction. We can see that the characteristics of goods_type, start_station, end_station and train_id have a higher correlation with the dwell time.

Based on the above analysis, car no, good type, car type, start station, end station and train id are selected as the eigenvectors of model input.

4. Model prediction results and analysis

80% of the data are training data and 20% of the data are testing data. In addition, in order to highlight the advantages of GBRT algorithm, we also compare four other machine learning algorithms, namely Bayesian Ridge Regression, Linear Regression, Elastic Net regression and Support Vector Regression. in Figure 4, It shows the comparison between the predicted value and the real value of various machine learning algorithms, it can be seen that the prediction result of The Gradient Boosting Regression Tree is very close to the real value.

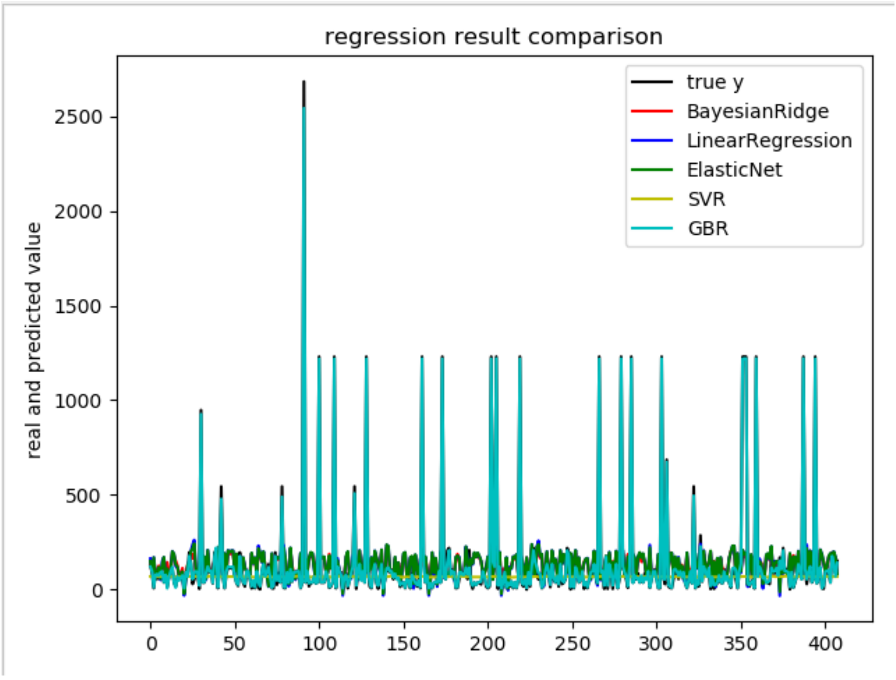


Figure.4. A comparison between the predicted value and the real value of five machine learning models.

Four indexes are selected to evaluate the model. They are explained_variance_score: interpretation the variance score of the regression model, whose value range is [0,1]. The closer to 1, the independent variable can better explain the variance change of the dependent variable. The smaller the value, the worse the effect. Mean absolute error is used to evaluate the degree of proximity between the predicted result and the real data set. The smaller the value is, the fitting effect will be better. Mean squared error: the indicator calculates the mean value and the sum of squares of the errors of the corresponding sample points of the fitting data and the original data. The smaller the value is, the better the fitting effect. r2_score: determination coefficient, which also means interpretation the variance score of the regression model. The value range is [0,1]. The closer to 1, the result same as the index of explained_variance_score. Table 1 shows the evaluation index values of the predicted values of various machine learning algorithms. In order to enhance the accuracy of the prediction results, the number of cross tests is set as 6. From the table, we can see that the explained_variance_score of GBRT and the judgment coefficient are basically close to 1, and the average absolute error and mean square error are relatively small, so the fitting degree of the model is the best.

Table.1. Four Evaluation Indexes

Machine Learning	explained_variance_score	Mean absolute error	Mean squared error	r2_score
Bayesian Ridge	0.022174	220.004376	654792.362123	0.022174
Linear Regression	0.034779	235.343710	646351.152328	0.034779
Elastic Net	0.034726	234.142831	646386.692425	0.034726
SVR	0.000665	181.127998	680156.788608	-0.015704
GBRT	0.911824	101.724085	59046.547023	0.911824

5. Conclusion

Based on the GBRT, this paper forecasts the dwell time of railway freight cars at the terminal, taking the car type, goods type, start station, end station, car id and train id as the dwell time that affects the cars at the terminal, and quantifies the characteristics. Since there is no dwell time data in the original data, the cars' dwell time is calculated according to the characteristics of the original data. Taking feature group as input variable and cars' dwell time at terminal as output value, different models are fitted. By comparing GBRT with Bayes Ridge Regression model, Linear Regression model, Elastic Net regression model and Support Vector Regression model in the prediction error between the predicted results and the real one, it is concluded that GBRT has the highest accuracy in calculation. Therefore, GBRT proposed in this paper improves the accuracy of the prediction model for the dwell time of railway freight cars at the terminal. However, the current railway models are very complex, and the influencing factors are also very complex. Based on the analysis of the existing historical data, this paper obtains some influencing factors, which improves the accuracy of the prediction. However, if more accurate prediction of the dwell time is needed, more influencing factors need to be mined.

Acknowledgments

This paper is supported by the project railway big data application technology research - Key Technology Research of big data for railway freight marketing and safety management (2017X004-A) and the project passenger and freight transportation organization optimization technology research - Key Technology Research of railway freight arrival time monitoring and warning (2017X009-C).

References

[1] Carey M., Carvile S. Scheduling and platforming trains at busy complex stations [J]. Transportation Research Part A Policy & Practice, 2003, 37 (3):195-224.

[2] Rodriguez J. A constraint programming model for real-time train scheduling at junction [J]. Transportation Research Part B Methodological,2007,41 (2):231-245.

[3] Carey M. A model and strategy for train pathing with choice of lines, platforms, and routes [J]. Transportation Research Part B Methodological,2008,28 (5):333-353.

[4] Cebon P., Samson D. Using real time information for transport effectiveness in cities [J]. City, Culture and Society,2011, 2 (4):201-210.

[5] Necula E. Analyzing Traffic Patterns on Street Segments Based on GPS Data Using R [J]. Transportation Research Procedia, 2015, 10:276-285.

[6] Junior O. S. S., Lopes D., Silva A. C., et al. Developing Software Systems to Big Data Platform based on MapReduce model: an Approach based on Model Driven Engineering [J]. Information & Intelligence, 2017, 8 (4): 31-44.