Traffic flow prediction based on machine learning

Wenqi Zhang [1], Yixin Li [2, *]

[1] College of Transportation Engineering, Shandong Jianzhu University, Yantai, China, 264000

[2] College of Mechanical and Automotive Engineering, Anhui Polytechnic University, Anhui, China, 241000

* Corresponding Author Email: baekyqi1@163.com

Abstract. The rapid development of intelligent transportation systems requires accurate prediction of short-term traffic flow. However, the nonlinearity, uncertainty, and spatiotemporal variability of traffic flow make it difficult for traditional traffic flow prediction methods to achieve ideal results. This study uses machine learning methods to improve the accuracy of prediction and expand features by evaluating the clustering effect of the k-means algorithm through data dimensionality reduction, interpolation of missing data, and the silhouette coefficient method. The 5-fold cross-validation and grid search methods were used to optimize the hyperparameters, and the Lasso regression model, ridge regression model and random forest regression model were compared. It was found that the feature expansion of the random forest had the highest fitting coefficient and the smallest error. This improves prediction accuracy, provides a valuable reference for intelligent traffic flow prediction, and has the potential for further optimization in feature selection, model design, and traffic control strategies.

Keywords: Traffic Flow, K-means Algorithm, The Random Forest Regression Model.

1. Introduction

With the acceleration of urbanization around the world, the number of vehicles continues to rise, causing traffic congestion to become more and more serious. Forecasting traffic flow is an important measure to alleviate traffic congestion. At present, cities are developing towards intelligence. Intelligent Transportation Systems (ITS) are an important part of smart cities. Traffic flow prediction can provide strong support for ITS. Traffic flow forecasting refers to using historical and real-time data to predict the number of vehicles passing through a certain road in the future. For different time scales, traffic flow forecasting can be divided into short-term traffic flow forecasting and long-term traffic flow forecasting. Short-term traffic flow forecasts usually focus on changes in traffic flow from minutes to hours, which can help traffic management departments adjust plans in a timely manner and provide travelers with real-time road conditions. Short-term traffic flow data usually exhibits the characteristics of repeatability, randomness and uncertainty, which brings challenges to accurate forecasting of urban traffic flow. Therefore, in order to alleviate traffic congestion and improve road traffic control strategies, relevant prediction aspects must be carried out. Scientific research.

In recent years, many scholars have used machine learning methods to conduct research on predicting short-term traffic flow. Based on several machine learning algorithms, Liu Di [1] focused on designing accurate traffic light management strategies for traffic congestion and different levels of traffic flow, and proposed a reliable real-time traffic flow prediction algorithm for dynamically changing traffic flow. Xu Jianfeng et al [2]. Established a multi-dimensional data model of traffic flow for multiple related routes in the same domain, and based on this data model provided a traffic flow prediction method based on multi-machine learning competition strategy. Jiang Xiaofeng et al [3]. Used support vector machine (SVM) to build a short-term traffic flow prediction model, and used a genetic algorithm (GA) to optimize the penalty factor C and kernel parameters of SVM to improve the accuracy of predicting traffic flow. Du Jinbiao [4] also provided a random forest regression prediction model (KNN-CEEMD-RFR) that can be decomposed through an improved fusion complementary lumped empirical mode filtered from KNN monitoring sites.

Although the above research has made significant progress in short-term traffic flow prediction, there are still some shortcomings. First of all, traditional research focuses on the application of a specific model or algorithm and lacks systematic comparison of multiple machine learning models. In addition, in terms of feature processing, existing research still needs to be improved in outlier processing and feature expansion.

In summary, this paper innovatively conducts research on the traffic flow prediction problem. First, this paper reduces the dimensionality of high-dimensional and single-attribute traffic flow data based on the length of a day, then processes outliers, deletes data with missing values greater than 0.3 in each day, and adopts the method for data with missing values less than 0.3. Completion by mean interpolation. Next, the k-means algorithm was used to cluster the data, and the clustering quality was evaluated by the silhouette coefficient method. At the same time, this paper expands the clustering results as new features to improve the accuracy of the prediction model. Finally, this article established the Lasso, Ridge and Random Forest Regression models, optimized the hyper parameters of the models through 5-fold cross-validation and grid search, and compared and evaluated the three models with evaluation indicators (MAE, MSE,R^2). The results showed that Random Forest The regression model has the best fitting effect and the smallest error. The technical roadmap of this article is shown in Figure 1 below.

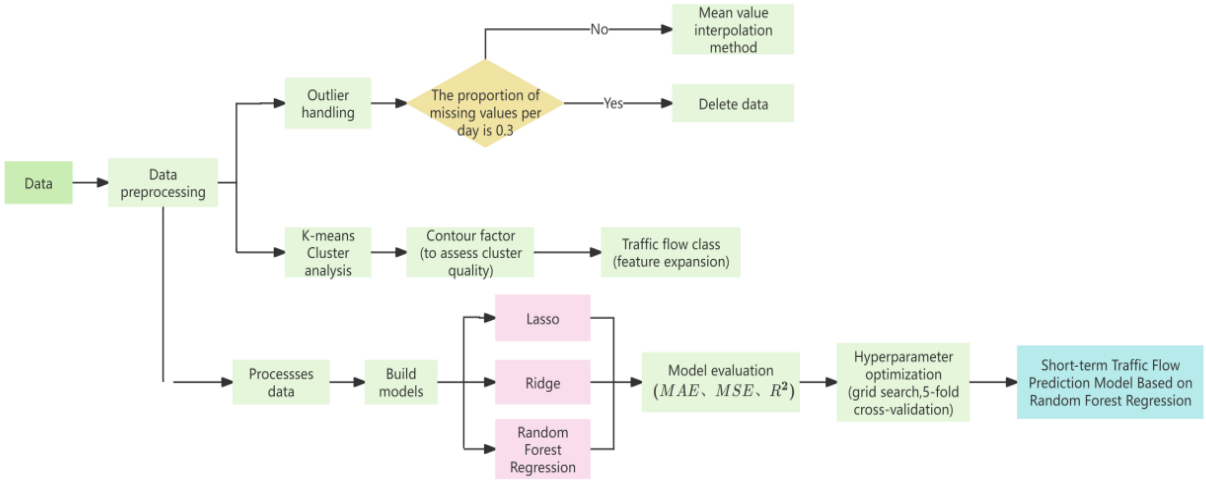


Figure 1. Technology Route.

The innovations of this article are mainly reflected in two aspects: first, the data quality is improved through dimensionality reduction and outlier processing, and second, the prediction accuracy of the model is significantly improved by using k-means feature expansion. This research not only effectively improves the accuracy of traffic flow prediction, but also provides new ideas for short-term traffic flow prediction, improving traffic management efficiency for future intelligent transportation systems.

2. Data Exploration

2.1. Data Source

The traffic flow data package used in this article comes from: https://www.kaggle.com. The traffic flow data records the traffic flow of each road within 3 months. The data sampling interval is 5 minutes, so a total of 26,496 observations were obtained. This data has the characteristics of high dimensionality and single attributes, so the data needs to be dimensionally reduced. Divide these traffic flow data into a series of samples according to the length of the day. Each sample starts at 0:00 and ends at 23:55, with intervals of 5 minutes. Therefore, there are 288 times in one day. Point. The traffic flow data after dimensionality reduction is shown in Figure 2 below.

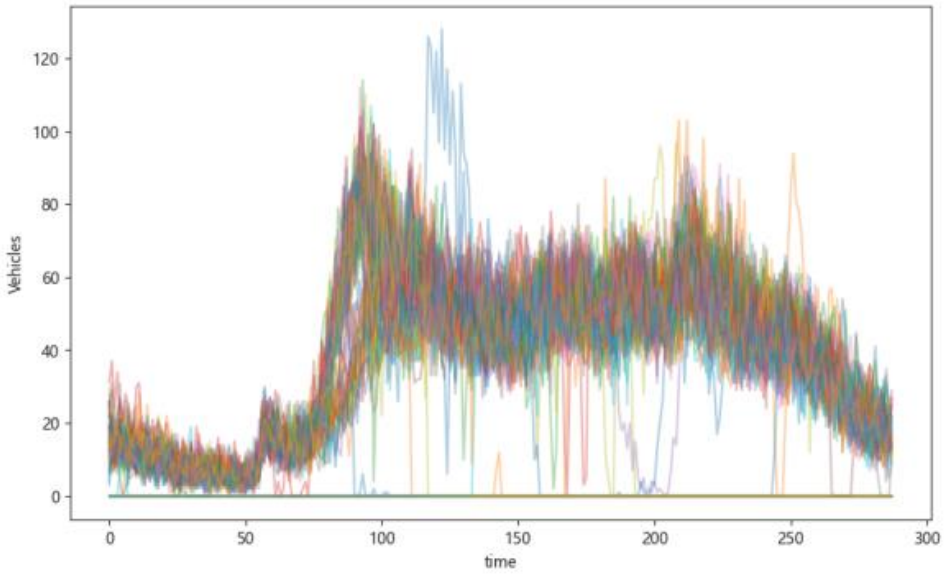


Figure 2. Dimensionality Reduction Traffic Flow Diagram.

2.2. Data Cleaning

In the process of information collection, due to uncertain reasons such as collection equipment hardware problems, vehicle accidents, human factors, etc., the collected information may be omitted or abnormal. In order to better predict traffic flow, the original data needs to be processed and the missing data needs to be filled in.

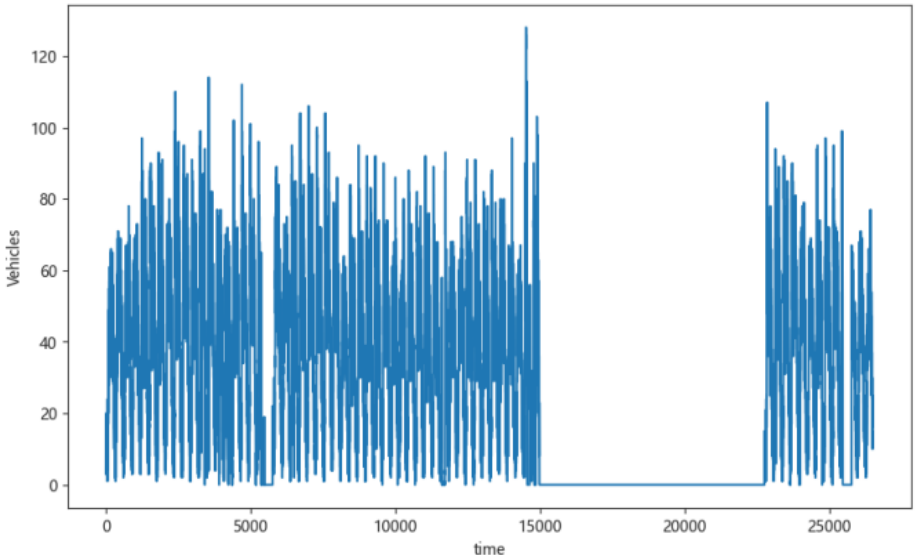


Figure 3. Original Traffic Flow Time Series Diagram.

As can be seen from Figure 3, the traffic flow data has abnormal 0 values within a certain time range, which seriously affects the prediction of traffic flow, so data cleaning is required. Next, the traffic flow data is grouped by day. If the proportion of missing values in each day is greater than 0.3, it is seriously missing, so these data will be deleted; for the rest containing missing values, this article will use the average interpolation method to process the vacant data. The principle of the mean imputation method is to use the average value of the adjacent values of the missing variable to fill in the missing values of the variable. The formula is shown in (1).

$$\bar{x} = \left[x(t-1) + x(t+1)\right] / 2 \tag{1}$$

Among them, is the average value of adjacent values of the missing value, and t is a certain time.

2.3. K-means Analysis

Traffic flow is affected by time, weather, holidays and other factors and shows a non-linear relationship. It usually changes with time and may vibrate in a short period of time. Due to the above characteristics, short-term traffic flow prediction has the above characteristics of uncertainty. In order to reduce the prediction overhead, it is necessary to use additional analysis tools to conduct precise analysis of traffic flow data in order to accurately and effectively predict traffic flow. Therefore, this article will use the k-means algorithm to conduct noise analysis on traffic flow data [5].

Each sample represents a traffic volume category and has the same starting time point and ending time point. Using each time point as a feature, k-means cluster analysis is performed on the traffic flow data to analyze the traffic flow category. And interpret the clustering results based on specific dates. Figure 4 below is a flow chart of traffic flow cluster analysis using k-means algorithm:

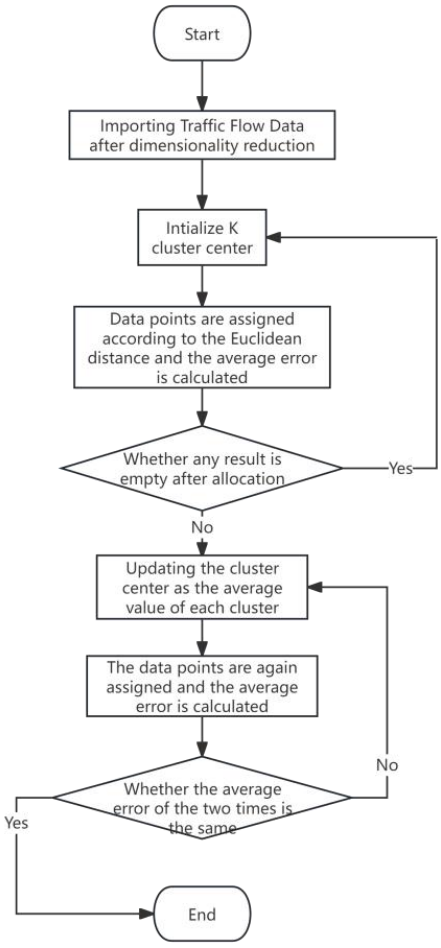


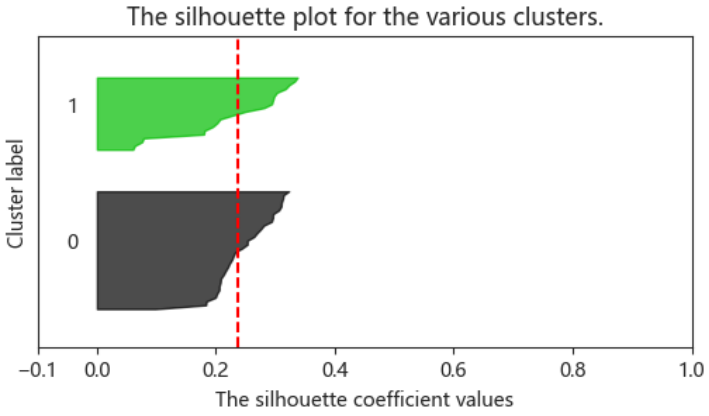
Figure 4. Traffic Flow Cluster Analysis Flow Chart.

This article will evaluate the clustering quality through the silhouette coefficient. The value range of the silhouette coefficient is [-1, 1]. If it is close to 1, it means that the sample point matches its cluster well, and the clustering effect is good; if it is close to -1, it means that its cluster is well matched. The matching effect is poor and the clustering effect is poor. The contour coefficient formula is shown in (2).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(1))} \tag{2}$$

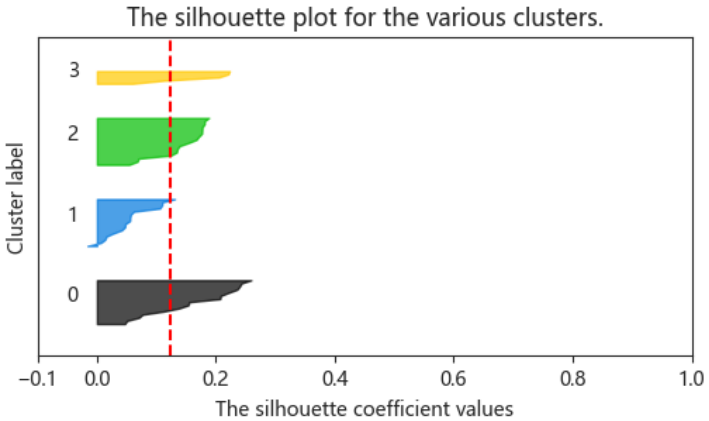
Figures (a), (b), and (c) show the silhouette coefficient values when k is 2, 3, and 4 respectively. The abscissa represents the silhouette coefficient value of each sample, and the ordinate represents k categories of samples. It can be seen from (a), (b) and (c) in Figure 3 that when k=2, the silhouette

coefficient values of the samples are all greater than 0, but as the k value increases, the silhouette coefficient values of each sample begin to gradually decrease. Small, the sample contour values of some categories even have negative values, and the clustering effect is not ideal. Therefore, when k=2, the clustering effect is optimal.



(a) Contour Coefficient Diagram When k=2

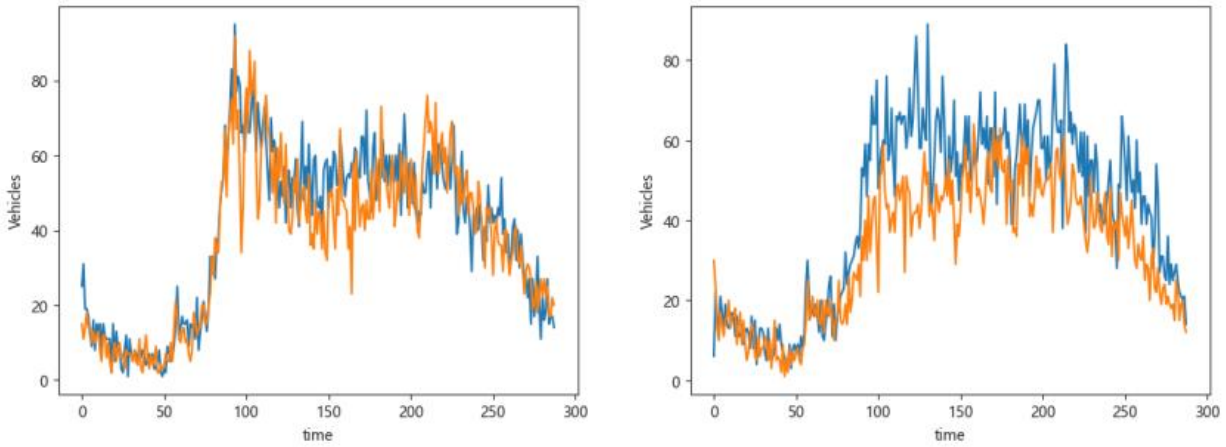
(b) Contour Coefficient Diagram When k=3



(c) Contour Coefficient Diagram When k=4

Figure 5. Silhouette Coefficient Diagram Under Different k Values.

k=2 means that the traffic flow data is divided into two categories. In order to specifically analyze the traffic flow categories, these two types of data are visualized. Figure 6(a) and (b) are the traffic flow trend charts of category 0 and category 1 respectively. The horizontal axis represents 288 time points in a day, and the vertical axis represents the traffic flow at that time point.



(a) Category 0 Traffic Flow Trend Chart. (b) Category 2 Traffic Flow Trend Chart.

Figure 6. Traffic Flow Trend Chart of Two Categories.

3. Model Establishment

3.1. Introduction To Model Principles

In order to predict traffic flow more accurately, this article will establish the following regression model and select the optimal regression model based on three model evaluation indicators.

The Lasso (Least Absolute Shrinkage And Selection Operator) model is a linear regression model that improves the prediction performance of the model and enhances the interpretability of the model by introducing the L1 regularization term based on the traditional least squares method[6]. L1 regularization usually adds an L1 norm penalty term to the loss function, and its expression is;

$$L(\beta)=\sum_{i=1}^{n}\left(y_i-X_i\beta\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{3}$$

Among them, is the response variable, is the characteristic variable, and is the regression coefficient. Is a regularization parameter that controls the strength of the penalty term? It is precisely because of L1 regularization that some regression coefficients can be effectively shrunk to zero, making the model sparse, so the multicollinearity problem between independent variables can be dealt with, and the model has strong interpretability. By eliminating unimportant independent variables, Lasso can simplify the model and improve the model's predictive accuracy. However, Lasso is more sensitive to the choice of pairs, and in some cases, Lasso's solution may not be the only solution.

The Ridge (ridge regression) model is built on the basis of traditional linear regression. The L2 regularization term is introduced to limit the complexity of the model and reduce overfitting, thereby improving the prediction performance of the model. The L2 regularization term is the penalty for the sum of squares of the model parameters. After adding the L2 norm penalty term to the loss function, the objective function becomes:

$$L(\beta)=\sum_{i=1}^{n}\left(y_i-X_i\beta\right)^2+\lambda\sum_{j=1}^{p}\beta_j^{\,2} \tag{4}$$

Ridge regression stabilizes the estimation of regression coefficients by adding penalty terms and can effectively handle multicollinearity problems; by limiting the size of regression coefficients, it can improve the generalization ability of the model and reduce overfitting. Even when the number of features is greater than the number of samples, the solution of Ridge regression is the only solution. However, Ridge does not perform feature selection, the model has poor interpretability, and there are also cases where it is more sensitive to feature selection.

The Random Forest Regression model improves the accuracy of regression analysis by constructing multiple decision trees and combining their prediction results. It is an integrated learning method. The basic principle is:

Randomly select multiple samples from the original data to form a sub-sample set;

For each node of the decision tree, during the splitting process, a part of the features is randomly selected for splitting;

Repeat the above steps to build multiple decision trees and finally form a decision tree forest;

For newly input data, random forest averages or weights the average of the prediction results of multiple decision trees to obtain the final regression result [7].

Because Random Forest randomly samples and randomly selects features to build trees, it has strong resistance to overfitting; by randomly selecting features to reduce calculations, Random Forest can effectively handle high-dimensional data. In addition, a random forest composed of multiple decision trees can usually achieve higher prediction accuracy than a single decision tree.

3.2. Model evaluation indicators

In order to better predict traffic flow, this article will use three evaluation indicators: mean absolute error (MAE), mean square error (MSE) and fitting coefficient (R^2). MAE is the average of the absolute value of the difference between the predicted value and the true value, which can directly reflect the accuracy of the prediction; MSE is the average of the squared errors, which is beneficial to the optimization algorithm and has high theoretical properties when used as a loss function. The smaller the values of MAE and MSE are, the smaller the model error is and the better the model effect is.R^2 Reflects the degree to which the model explains the variability of the data. The closer its absolute value is to 1, the stronger the model's explanation ability. The expressions are as shown in equations (5) to (7).

In the formula, represents the actual value of traffic flow at the i-th time, represents the traffic flow prediction value of the model at the i-th time, and represents the sample mean.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{6}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{7}$$

3.3. Model Hyperparameters

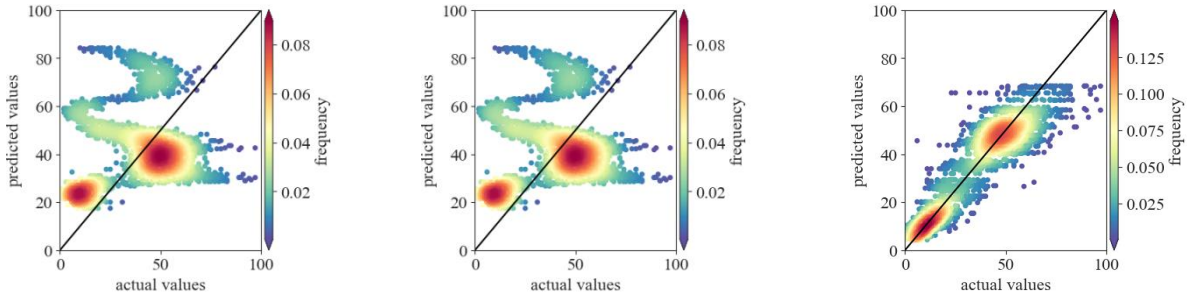
Model hyper parameters are configurations external to the model. Their values cannot be estimated from the data, and their optimal values can be obtained through trial and error [8]. In order to select the best model and parameter combination and improve the accuracy and reliability of traffic flow prediction, this article will use the 5-fold cross-validation method to evaluate the model performance and find the optimal model combination through grid search.5-fold cross-validation refers to dividing the data set into 5 equal-sized subsets. In each validation process, 4 subsets are used as training sets, and the remaining 1 subset is used as test sets. The model is Perform training and prediction on the test set [9], and record the evaluation indicators. Repeat 5 times, and finally get 5 evaluation results. We can use the average evaluation indicator as the performance indicator of the model. Grid search is an exhaustive search method that specifies parameter values. By optimizing the parameters of the estimated function through 5-fold cross-validation, the optimal parameter configuration is finally found [10]. The optimal hyper parameter combinations of Lasso, Ridge and Random Forerst Regression models are shown in Table 1.

Table 1. Display of Optimal Hyper parameter Combinations for Each Model.

Model	Hyperparameter
Lasso	λ=0.0
Ridge	λ=1.0
Random Forerst Regression	min_samples_leaf=7

3.4. Model Results

(a), (b), and (c) in Figure 7 show the density scatter plots between the real values and predicted values of the Lasso, Ridge, and Random Forerst Regression models respectively. The horizontal axis represents the real value of traffic flow, and the vertical axis represents Traffic flow prediction value under the model. The red area is an area with dense data points, indicating that the predicted value is highly consistent with the true value. It can be seen from Figure 7 that the Random Forerst Regression model can predict traffic flow data more accurately.



(a) Density scatter plot of Lasso's true and predicted values

(b) Density scatter plot of Ridge's true and predicted values

(c) Density scatter plot of Random Forerst Regression's true and predicted values

Figure 7. Density Scatter Plot between the Actual Values and Predicted Values of Each Model.

Comparing the model evaluation indicators in Table 2, the MAE of the Lasso regression model is 19.436 and the fitting coefficient is 0.410; the MAE of the Ridge regression model is 19.432 and the fitting coefficient is 0.409; the MAE of the Random Forerst Regression model is 6.444 and the fitting coefficient is 0.823 .It can be seen that the random forest regression model has better fitting effect.

Table 2. Evaluation Indicators of Each Model.

Model	Lasso	Ridge	Random Forerst Regression
MAE	19.436	19.432	6.444
MSE	577.589	577.316	72.709
R^2	-0.410	-0.409	0.823

Then, a comparative experiment was conducted on the Random Forerst Regression model using k-means for feature expansion and not using feature expansion. The comparison results are shown in Table 3 below. The MAE without k-means feature expansion is 6.516, and the MSE is 75.66. The errors are both higher than those with k-means feature expansion. Its fitting coefficient is 0.815, which is smaller than the fitting coefficient of the model with k-means feature expansion., therefore, the Random Forerst Regression model with k-means feature expansion has better fitting effect and more accurate prediction.

Table 3. Is There Any Model Evaluation Index for k-means Feature Expansion?

Model	Random Forerst Regression (k-means feature augmentation)	Random Forerst Regression (no feature augmentation)
MAE	6.444	6.516
MSE	72.709	75.66
R^2	0.823	0.815

4. Conclusion

This paper performs dimensionality reduction and outlier processing on traffic flow data, uses the k-means algorithm to cluster the data, and evaluates the clustering effect through the silhouette coefficient. Adding traffic flow category as a feature enhances model accuracy. Lasso, Ridge, and Random Forest models are optimized via cross-validation and grid search. Random Forest outperforms others in fitting and prediction. Feature expansion via k-means further reduces model error and raises fitting coefficient. The study shows combining traffic flow time series with clustering improves model performance, especially for nonlinear, complex data. Random Forest's predictive power aids short-term traffic flow prediction, aiding traffic management in optimizing congestion and facilitating travel.

However, the limitation of this study is that it did not consider the impact of external factors such as weather on traffic flow. In the future, more comprehensive data should be collected to fully consider the impact of various factors. In addition, with the rise of deep learning technology, future research will try to introduce more complex deep learning models to further improve prediction accuracy and process larger-scale traffic flow data.

References

[1] Liu Di. Research on safe operation of complex transportation networks based on several machine learning algorithms [D]. Southeast University, 2021.

[2] Xu Jianfeng, Tang Tao, Yan Junfeng, et al. Short-term traffic flow prediction based on multi-machine learning competition strategy [J]. Transportation System Engineering and Information, 2016, 16(04): 185-190+198.

[3] Jiang Xiaofeng, Xu Lunhui, Zhu Yue. Short-term traffic flow prediction based on SVM. "Journal of Guangxi Normal University" (Natural Science Edition), 2018, 30.4: 13-17.

[4] Du Jinbiao. Short-term traffic flow prediction based on CEEMD-RFR with improved KNN [D]. Chang'an University, 2018.

[5] Liu Yanli, Zhao Zhuofeng, Ding Weilong, et al. Short-term traffic flow prediction method based on high-speed toll big data [J]. Computers and Digital Engineering, 2019, 47(5):7.

[6] Wang Dexian, He Xianbo, He Chunlin, et al. Research on latent semantic prediction model combining L_1 and L_2 regularization constraints [J]. Computer Engineering and Applications, 2019, 55(19): 121-127.

[7] Wang Jiaqi, Hu Jingxin, Zhang Hongyan. Desert locust remote sensing monitoring based on machine learning [J]. Journal of Anhui Agricultural University, 2023, 50(04): 738-744.

[8] Jiao Pengpeng, An Yu, Bai Zixiu, et al. Research on short-term traffic flow prediction based on XGBoost [J]. Journal of Chongqing Jiaotong University (Natural Science Edition), 2022(008):041.

[9] Li Kunming, Gu Yijun, Wang An. Malicious PDF file detection technology under evasion attacks [J]. Journal of People's Public Security University of China: Natural Science Edition, 2019, 25(3):5.

[10] Liu Wei, Xu Wenfeng. Application of intelligent recommendation technology based on machine learning in transformer selection [J]. Electric Power Information and Communication Technology, 2019, 17(5): 19-24.