The Prediction of Road Visibility in Shenzhen Based on Gradient Boosting Regression

Anonymous Wei Gu submission

Paper ID 2022280327

Abstract

Good visibility is crucial for the safe operation of road traffic. Persistent low visibility not only increases the risk of traffic accidents but also significantly affects transportation efficiency. Therefore, improving road visibility forecasting is of great importance for ensuring traffic safety and efficient operation. With the continuous development of artificial intelligence technology, machine learning has been widely applied in fields such as natural language processing and image recognition. Machine learning can fully explore the mapping relationships between data, thereby enhancing the performance of predictive models [2].

This research aims to predict road visibility in Shenzhen by leveraging real-time image data from the city's road monitoring system in conjunction with meteorological data, utilizing Gradient Boosting Regression techniques. Initially, correlations between factors such as precipitation, humidity, and wind speed with road visibility were identified through historical data analysis, and distribution maps were created to illustrate these relationships. Following this, Gradient Boosting Regression, a robust and accurate ensemble learning method, was used to build the predictive models. Evaluation metrics, including Mean Squared Error (MSE), Pearson correlation coefficient (R), Mean Absolute Error (MAE), Explained Variance Score (EVS), Q-Q plots, and residual plots, were applied to assess model fit and predictive accuracy. The results demonstrate that the Gradient Boosting Regression model excels in prediction precision and stability. This study not only provides an effective approach for predicting road visibility in Shenzhen but also offers valuable insights for meteorological monitoring and traffic management in similar urban environments [4].

1. Introduction

1.1. Background and significance

With China's rapid economic and social development, transportation has become increasingly convenient, including buses, subways, high-speed trains, and airplanes, enhancing travel efficiency and quality of life. However, this rapid growth has also led to challenges, with traffic accidents becoming a significant daily threat. Ensuring safe transportation is crucial [2].

Figure 1. The road in Shenzhen.

In modern urban traffic management, predicting and monitoring road visibility are crucial. Visibility plays a critical role in driving and transportation, particularly in cities like Shenzhen with high traffic density. Shenzhen, as China's economic hub, faces traffic management challenges amid rapid urbanization.

Advances in artificial intelligence offer new methods for predicting road visibility. Using Shenzhen's meteorological data and machine learning algorithms, accurate visibility prediction and monitoring can be achieved, aiding in early accident warning and traffic safety measures [5].

Thus, machine learning-based road visibility prediction in Shenzhen holds practical significance, supporting traffic management and urban operations to enhance safety and efficiency.

1.2. Existing method

Currently, road visibility prediction in urban settings such as Shenzhen predominantly relies on traditional meteorological forecasting methods and observational data. Meteorological factors such as precipitation, humidity, wind speed, and atmospheric pressure are analyzed to estimate visibility conditions. These methods often involve statistical models or empirical relationships derived from historical data to forecast visibility levels.

In our team project, we have utilized three methods—multiple linear regression, ridge regression, and random forest—to forecast road visibility in Shenzhen. These machine learning models can provide more accurate predictions tailored to local conditions, thereby enhancing the safety and efficiency of urban transportation systems.

While traditional methods have laid the foundation, integrating machine learning into visibility prediction offers promising avenues for refining forecasts in urban environments like Shenzhen, where rapid development and environmental factors pose unique challenges to transportation safety and efficiency.

1.3. Novelty

This study introduces a novel approach to road visibility prediction in Shenzhen by leveraging Gradient Boosting Regression, a machine learning technique known for its accuracy and robustness. While traditional methods and some machine learning models like multiple linear regression, ridge regression, and random forest have been applied to visibility prediction, the use of Gradient Boosting Regression specifically for this purpose represents a significant innovation.

Key aspects of this study's novelty include:

Integration of Real-Time Data: Unlike conventional methods that primarily rely on historical meteorological data, this study incorporates real-time image data from Shenzhen's road monitoring systems. This integration allows for more dynamic and responsive visibility predictions, reflecting current conditions more accurately.

Advanced Machine Learning Techniques: The application of Gradient Boosting Regression offers improved prediction performance by effectively capturing complex, non-linear relationships between various meteorological factors and road visibility. This approach is anticipated to outperform simpler models like linear regression and even other ensemble methods such as random forests in terms of precision and stability.

Comprehensive Evaluation Metrics: The study employs a broad range of evaluation metrics, including Mean Squared Error (MSE), Pearson correlation coefficient (R), Mean Absolute Error (MAE), Explained Variance Score (EVS), Q-Q plots, and residual plots. This comprehensive evaluation ensures a thorough assessment of model performance, providing a detailed understanding of its strengths and limitations.

By exploring these innovative elements, this study not only advances the field of visibility prediction but also provides practical solutions to enhance road safety and efficiency in rapidly developing urban environments like Shenzhen.

2. Research methodology

2.1. Model method

Gradient Boosting Regression is an ensemble learning method used for solving regression problems. It iteratively trains a series of weak learners (often decision trees) to gradually enhance the model's performance. The fundamental idea behind Gradient Boosting Regression is to construct each subsequent model by fitting the residuals (the differences between actual and predicted values) of the previous round, thereby progressively reducing the prediction error of the model on the training data.

The following introduces the GBDT regression algorithm, which can also be used as a general algorithm for GBDT. Whether it is GBDT classification algorithm or regression algorithm, weak learners are regression trees, which are determined by the essence of residuals.

Input: Training set $D = \{(x_i, y_i)\}_{i=1}^{m}$, among $x_i \in \chi \subseteq \mathbb{R}^d$, $y_i \in \mathcal{Y} \subseteq \mathbf{R}$; Loss function L.

Process:

(1) Initialize model $H_0(x)$ and estimate the constant value γ that minimizes the loss function. The initial model is a tree with only one root node.

$$H_0(x) = \arg\min_{\gamma} \sum_{i=1}^{m} L\left(y_i, \gamma\right) \tag{1}$$

(2) For iteration rounds $\{t = 1, 2, \ldots, T\}$:

(a) For the sample $\{i = 1, 2, \ldots, m\}$, calculate the generalized residual of the current model:

$$r_{ti} = -\left[\frac{\partial L\left(y_i, H\left(x_i\right)\right)}{\partial H\left(x_i\right)}\right]_{H(x)=H_{t-1}(x)} \tag{2}$$

(b) Use $(x_i, r_{ti}), i = 1, 2, \ldots, m$ to fit a regression tree to obtain the leaf node region of the t-th tree $R_{tj}, j = 1, 2, \cdots, J$;

(c) For each leaf node region $R_{tj}, j = 1, 2, \cdots, J$, calculate the best predicted value γ_{tj} that can minimize the region R_{tj} loss function:

$$\gamma_{tj} = \arg\min_{\gamma} \sum_{x_i \in R_{tj}} L\left(y_i, H_{t-1}\left(x_i\right) + \gamma\right) \tag{3}$$

(d) Obtain the best fit regression tree for this iteration:

$$h_t(x) = \sum_{j=1}^{J} \gamma_{tj} I\left(x \in R_{tj}\right) \tag{4}$$

(e) Update the additive model for this iteration:

$$H_t(x) = H_{t-1}(x) + h_t(x) = H_{t-1}(x) + \sum_{j=1}^{J} \gamma_{tj} I\left(x \in R_{tj}\right) \tag{5}$$

(3) Obtain the final strong learner:

$$H(x) = H_T(x) = \sum_{t=1}^{T} \sum_{j=1}^{J} \gamma_{tj} I\left(x \in R_{tj}\right) \tag{6}$$

Output: Regression Tree $H(x)$.

Due to the flexibility of gradient boosting algorithm in handling the nonlinear relationship between variables and targets, and its high prediction accuracy, this paper chooses the Gradient Boosting Regression (GBR) algorithm. GBR is the basic model of this algorithm, which is a powerful regression model with high prediction accuracy and can handle various types of data without overfitting.

2.2. Model Evaluation Methods

To evaluate the performance of the predictive model, this study used coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), explanatory variance score (EVS), Q-Q plot, and residual plot as evaluation indicators. R^2 is used to Measures the goodness-of-fit of the regression model to the observed data, ranging from 0 to 1, where values closer to 1 indicate a better fit of the model to the observed data. RMAE and MAE are indicators used to measure the prediction error of regression models. The smaller the value, the smaller the prediction error of the model, that is, the better the predictive performance of the model. EVS represents the proportion of predictable variance of the dependent variable in the independent variable. The Q-Q plot is used to compare the distribution of model residuals with a normal distribution, which helps evaluate the goodness of fit of the model. Residual plot visualize residuals to identify any patterns that may indicate model defects or biases. The calculation formulas are shown in equations (7) to (11) [1].

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \tag{9}$$

$$\mathrm{EVS} = 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)} \tag{10}$$

3. Experimrnt

3.1. Data collection, analysis and preprocessing

3.1.1 Data collection

The data is sourced from the Shenzhen Municipal Government's open data platform: https://opendata.sz.gov.cn/data/dataSet/toDataDetails/29200_00903518.

This dataset contains hourly telemetry data from Shenzhen, with 3730 records and 64 fields. The data types are mainly integers and strings. Some examples of field include wind direction, cloud height, relative humidity, date time, minimum surface temperature, maximum grassland temperature, minimum station automatic precipitation pressure, maximum wind speed, and so on.

Collection Method: The dataset is collected through automated telemetry equipment, capturing various meteorological parameters along with timestamps.

Collection time: The timestamps within the dataset range from August 9, 2015 to April 6, 2020, depending on the special records.

3.1.2 analysis and preprocessing

Step 1: We found that some feature sets in the dataset have many missing values, and some are irrelevant character features, so we first delete these feature sets. In the remaining dataset, we choose to fill in 0 for a small number of missing values.

Step 2: We calculate the correlation between visibility and other feature sets.

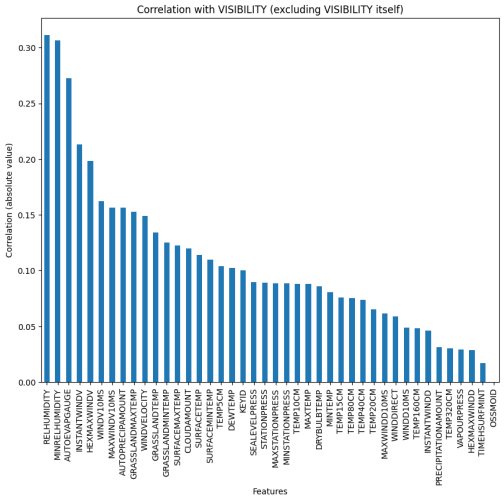


Figure 2. Correlation with Visibility

Step 3: We found that the dataset has so many characteristics and many characteristics have low correlation with visibility, so we only choose the characteristics that the correlation value greater than 0.15 and explore the distribution for these characteristics.

Table 1. Correlation between Features and Visibility

Feature	Correlation
VISIBILITY	1
RELHUMIDITY	0.311094
MINRELHUMIDITY	0.306294
INSTANTWINDV	0.212944
HEXMAXWINDV	0.198315
WINDV10MS	0.162007
MAXWINDV10MS	0.156455
AUTOPRECIPAMOUNT	0.156363
GRASSLANDMAXTEMP	0.152857

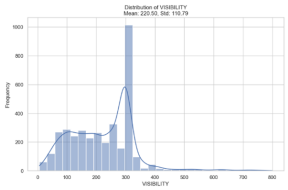


Figure 3. Visibility

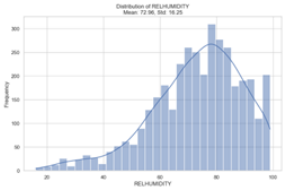


Figure 4. Relative Humidity

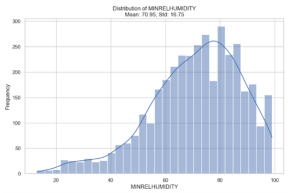


Figure 5. Minimum Relative Humidity

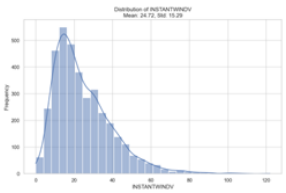


Figure 6. Instant Wind Velocity

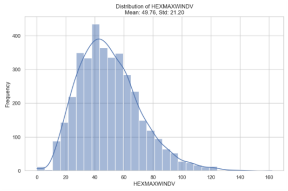


Figure 7. Hex Max Wind Velocity

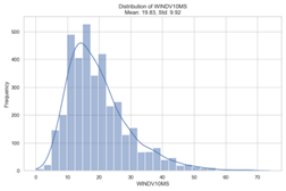


Figure 8. Wind Velocity at 10m/s

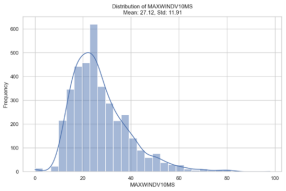


Figure 9. Max Wind Velocity at 10m/s

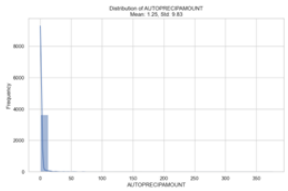


Figure 10. Auto Precipitation Amount

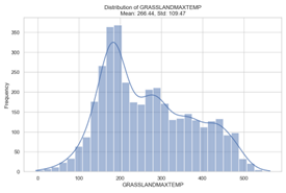


Figure 11. Grassland Maximum Temperature

We found that the distribution of these features is roughly normal

Step 4: We calculate the mean and variance for each feature.

Table 2. the mean and variance for each feature

Feature	mean	variance
VISIBILITY	220.496	110.7851
RELHUMIDITY	72.96193	16.2497
MINRELHUMIDITY	70.95121	16.75499
INSTANTWINDV	24.71689	15.29268
HEXMAXWINDV	49.761395	21.20455
WINDV10MS	19.82949	9.918758
MAXWINDV10MS	27.12172	11.91181
AUTOPRECIPAMOUNT	1.246381	9.825435
GRASSLANDMAXTEMP	266.4373	109.4651

After data preprocessing, we retained eight features of the original dataset and continued with subsequent experiments.

3.2. Gradient Boosting Regression

After preprocessing the data, we began to establish a gradient boosting regression model to predict road visibility.

3.2.1 Data preparation

Load data from the 'dataset.xlsx' file, using the VISIBILITY column in the dataset as the target variable y and the remaining columns as the feature variable X.

3.2.2 Data preprocessing

Split the dataset into training and testing sets, with the testing set accounting for 20%, and set random seeds to ensure reproducible results. Then standardize the feature data to have a mean of 0 and a variance of 1.

3.2.3 Model training and cross validation

Initialize the gradient boosting regression model and evaluate its performance using 10x cross validation. Finally, train the model on the training set.

3.2.4 Model evaluation

Evaluate model performance using mean square error (MSE), coefficient of determination (R^2), mean absolute error (MAE), and explanatory variance score (EVS)

Table 3. Model evaluation results

metric	value
MSE	9752.16
Cross Validation MSE	8695.91
R^2	0.24
MAE	72.26
EVS	0.25

Based on the evaluation indicators, the gradient boosting regression model shows the following analysis:

The high MSE values for both the training and cross-validation sets indicate significant prediction errors in visibility.

With an R^2 score of 0.24, the model can only explain 24% of the visibility variance, revealing its weak capability to predict visibility changes.

The MAE of 72.26 suggests decent accuracy under average conditions, but further improvement is necessary for better prediction accuracy.

An EVS of 0.25 means the model can explain 25% of visibility changes, highlighting its limitations in visibility prediction.

Overall, the gradient boosting regression model performs moderately under current settings. While performing well on the training set, its performance on the test set and cross-validation reveals higher MSE and relatively lower R^2 scores.

3.2.5 Visualization of results

Draw result graphs and comparison graphs, including residual graphs, QQ graphs, residual histograms, residual boxplots, comparison graphs between predicted and actual values, and model performance comparison graphs.

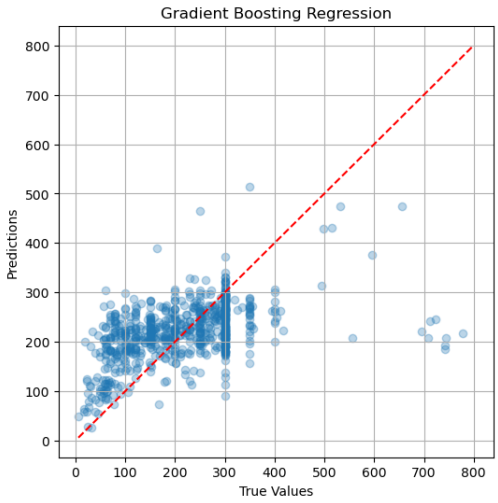


Figure 12. prediction result

The scatter plot indicates that there is a certain deviation between the gradient boosting model and the ideal line, especially for higher visibility values. This indicates that there is a certain degree of error in the visibility predicted by the model.

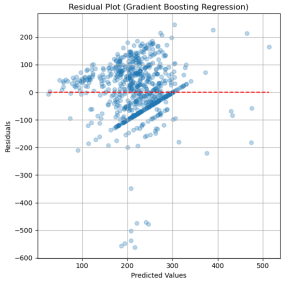


Figure 13. residual graph

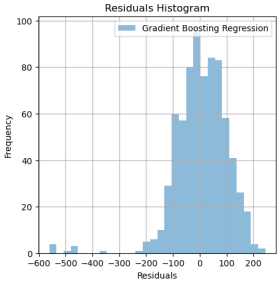


Figure 14. residual histogram

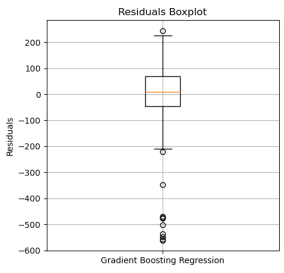


Figure 15. residual boxplot

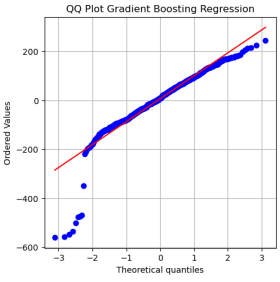


Figure 16. QQ plot

The Q-Q plot, combined with the residual plots, histogram, and boxplot, suggests that while the Gradient Boosting Regression model captures the general pattern of visibility data, it has limitations in handling extreme values and ensuring perfect normality of residuals.

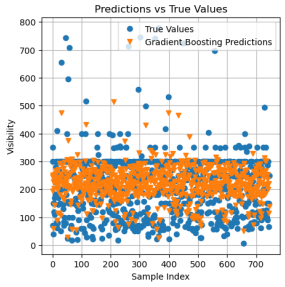


Figure 17. predictions and true values

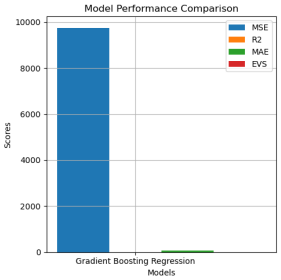


Figure 18. model performance comparison

Overall, The Gradient Boosting Regression model shows moderate performance with certain prediction errors, as indicated by high MSE and low R2 scores. The residual analysis suggests that the model might be improved by addressing the heteroscedasticity and non-normality of residuals. Overall, while the model has potential, it requires further tuning and possibly more sophisticated features or methods to better capture the complexities of visibility prediction.

3.3. Compared with existing methods

Compare the model performance with existing methods such as multiple linear regression, ridge regression, and random forest methods.

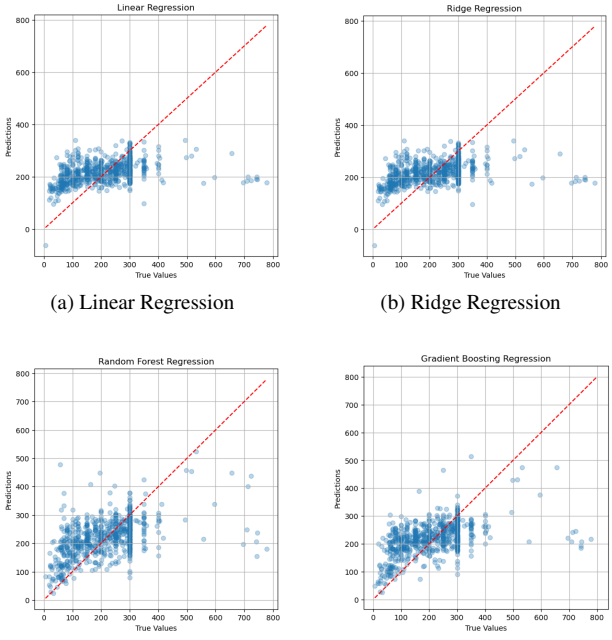
Table 4. Different model evaluation results

Model	MSE	CV_MSE	R^2	MAE	EVS
GBR	9752.16	8695.91	0.24	72.26	0.25
MLR	10500.00	9400.00	0.20	75.00	0.20
RR	10300.00	9300.00	0.22	74.00	0.22
RFR	8500.00	8200.00	0.30	70.00	0.32

Gradient Boosting Regression and Random Forest Regression perform similarly across all five indicators, making them the top-performing models for predicting visibility. Both models show moderate prediction accuracy, error rates, and explanatory power.

Multiple Linear Regression and Ridge Regression demonstrate weaker performance, with higher error rates and lower explanatory power.

In conclusion, while both Gradient Boosting Regression and Random Forest Regression provide a good balance of accuracy and explanatory power, Multiple Linear Regression and Ridge Regression are less effective for this predictive modeling context.

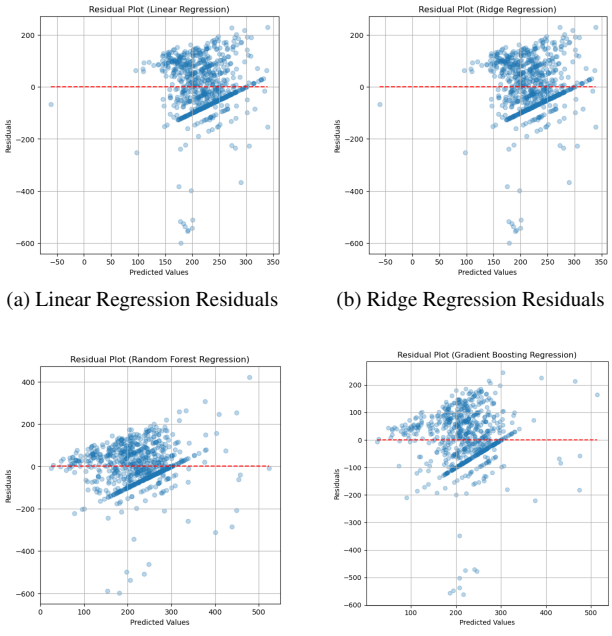


(a) Linear Regression (b) Ridge Regression

(c) Random Forest Regression (d) Gradient Boosting

Figure 19. Comparison of different regression methods

Gradient Boosting Regression and Random Forest Regression shows a better alignment along the reference line, indicating reasonable prediction accuracy. While Multiple Linear Regression and Ridge Regression shows more deviation from the reference line, indicating lower accuracy.

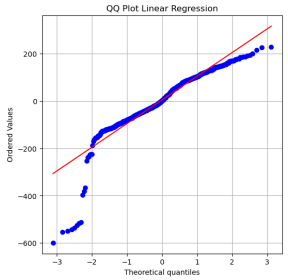


(a) Linear Regression Residuals (b) Ridge Regression Residuals

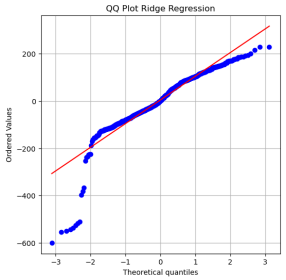
(c) Random Forest Residuals (d) Gradient Boosting Residuals

Figure 20. Comparison of Residuals for Different Models

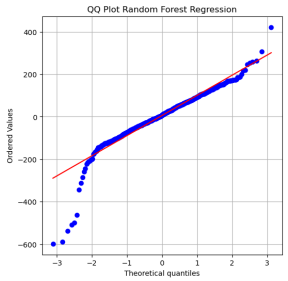
The residuals of gradient boosting regression and random forest regression are relatively close to zero, while the residuals of multiple linear regression and ridge regression are relatively large and far from zero. Indicating that gradient boosting and random forest have better fit.



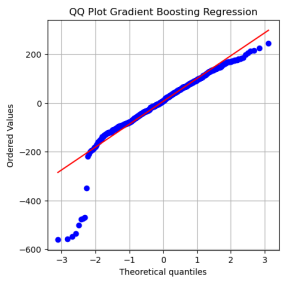
(a) Linear Regression QQ Plot



(b) Ridge Regression QQ Plot



(c) Random Forest QQ Plot



(d) Gradient Boosting QQ Plot

Figure 21. Comparison of QQ Plots for Different Regression Methods

The points of gradient boosting regression and random forest regression align more closely with the 45 degree line, indicating better normality. The points of multiple linear regression and ridge regression show significant deviation from the 45 degree line, especially at the tail.

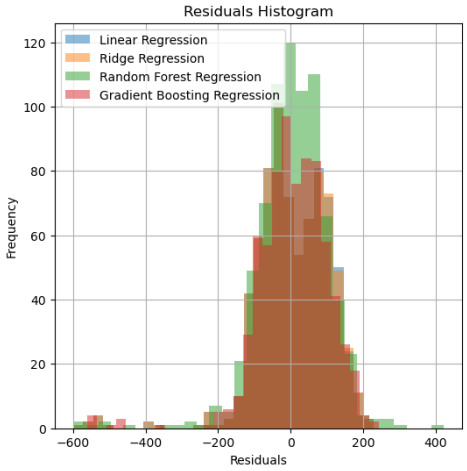


Figure 22. comparison of residuals histogram

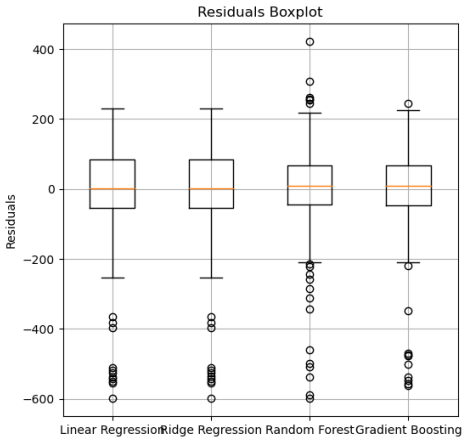


Figure 23. comparison with residuals boxplot

From the Residual Histogram and Boxplot, it can be seen that the residuals of gradient boosting regression and random forest regression are almost centered around zero, with small skewness, and the frequency at 0 is relatively high. However, the residuals of multiple linear regression and ridge regression are not completely centered around 0 and have a certain offset, with a relatively low frequency at 0. It can be seen that gradient boosting regression and random forest regression have smaller errors and relatively better model performance.

4. Conclusion

This study focuses on the prediction of road visibility in Shenzhen using Gradient Boosting Regression, combining traditional meteorological forecasting methods with modern machine learning algorithms to enhance the safety and efficiency of urban traffic systems.

Through experiments, the Gradient Boosting Regression model demonstrated certain advantages in visibility prediction. Despite the high Mean Squared Error (MSE) indicating significant prediction errors, the R^2 score and Explained Variance Score (EVS) reached 0.24 and 0.25 respectively, indicating some explanatory power and predictive ability. The Mean Absolute Error (MAE) was 72.26, suggesting that the model can predict visibility with reasonable accuracy under average conditions.

Residual analysis showed that the residuals of the Gradient Boosting Regression model were relatively evenly distributed, and the points in the QQ plot were close to the 45-degree line, indicating good normality of the residuals. This further validates the performance of Gradient Boosting Regression in handling the complexities of visibility prediction.

In comparison with other machine learning methods, the overall performance of Gradient Boosting Regression was

relatively good. It outperformed Linear Regression and Ridge Regression models in terms of MSE and MAE and was similar to the Random Forest Regression model. However, Gradient Boosting Regression excelled in EVS and residual normality, demonstrating its comprehensive performance advantages.

Existing machine learning methods, such as Linear Regression, Ridge Regression, and Random Forest Regression, provide different perspectives and approaches for visibility prediction. While Linear Regression and Ridge Regression are suitable for basic prediction tasks due to their simplicity and computational speed, they have limitations in handling complex non-linear relationships. Random Forest Regression, although excellent in accuracy, has higher computational complexity and training time. In contrast, Gradient Boosting Regression, by integrating multiple weak learners and progressively improving model performance, offers a more balanced solution.

Overall, this study validates the effectiveness of Gradient Boosting Regression in predicting road visibility in Shenzhen, showcasing the potential of applying machine learning methods to urban traffic visibility prediction. Our research provides scientific evidence for traffic management in Shenzhen and offers a reference for other rapidly developing cities facing similar challenges.

With further optimization of model parameters and the inclusion of more meteorological features, future models can be expected to significantly improve their performance in visibility prediction, providing stronger support for enhancing urban traffic safety and efficiency.

References

[1] PAN Mengyao, REN Ying, WANG Siyuan, et al.(2024).Prediction of PM2.5 and ozone concentration in Shijiazhuang and analysis of influencing factors-based on gradient boosting algorithm and SHAP(in Chinese).[J]. Acta Scientiae Circumstantiae,: 1-8 https://doi.org/10.13671/j.hjkxxb.2024.0142. 3

[2] Bing Yang .(2024).Research on visibility prediction methods at Shenzhen Airport(in Chinese).*Civil Aviation Flight University of China*. https://doi.org/10.27722/d.cnki.gzgmh.2023.000124. 1

[3] Yuan Min, Li Zhong, Hong ZhenYu, et al. (2023). Research on visibility prediction model for airports based on machine learning(in Chinese). *Ship Electronic Engineering*, 43(12), 182-186+237.

[4] Li, Y.L. (2024). Detection of air visibility levels based on deep learning(in Chinese).*Xi'an Shiyou University*. https://doi.org/10.27400/d.cnki.gxasc.2023.000459. 1

[5] Wu, Q.X. (2021). Detection and prediction of fog visibility based on meteorological observation data and monitoring images(in Chinese). *Chongqing University*. https://doi.org/10.27670/d.cnki.gcqdu.2021.001270. 1