深 圳 大 学 实 验 报 告

课程编号：2801000049

课程名称：机器学习

实验项目名称：基于机器学习的深圳道路能见度预测

学院：电子与信息工程学院

专业：电子信息工程

指导教师：麦晓春

报告人：古炜
林凡超
卫宏林 学号：古炜
林凡超
卫宏林 2022300013 班级：文华班

实验时间：2024.5.28 - 2024.6.28

实验报告提交时间：2024.6.28

教务部制

基于机器学习的深圳道路能见度预测

古炜 林凡超 卫宏林

Contributing percentage: 33.33% + 33.33% + 33.33%

Abstract

(250-300 words)

This study focuses on predicting road visibility in Shenzhen by harnessing real-time image data from the city's road monitoring system alongside meteorological data, employing machine learning methodologies. Initially, through an analysis of historical data, correlations between factors such as precipitation, humidity, and wind speed with road visibility were identified, and distribution maps illustrating these relationships were generated. Subsequently, various machine learning algorithms including multiple linear regression, ridge regression, and random forest were employed to construct predictive models. Evaluation metrics such as Mean Squared Error (MSE), Pearson correlation coefficient (R), Mean Absolute Error (MAE), Explained Variance Score (EVS), Q-Q plots, and residual plots were utilized to assess model fit and predictive accuracy. The results indicate that the random forest model outperforms other methodologies in terms of both prediction precision and stability. This research not only offers an effective method for predicting road visibility in Shenzhen but also provides valuable insights for meteorological monitoring and traffic management in similar urban settings.

1. Introduction

Background

In modern urban traffic management, the prediction and monitoring of road visibility are of paramount importance. Visibility, as a common indicator in daily life, plays a crucial role in driving and transportation. Especially in rapidly developing cities like Shenzhen, with high traffic density and a large number of vehicles, timely and accurate prediction of road visibility is particularly crucial.

Figure 1 haze weather in shenzhen

Shenzhen, as China's economic center and a modernized city, faces challenges in traffic management alongside its rapid economic growth. Road visibility directly impacts the safety of citizens' travel and their quality of life.

The development of machine learning technology provides new approaches and methods for solving the problem of road visibility prediction. By leveraging meteorological data from Shenzhen and integrating machine learning algorithms, it is possible to achieve accurate prediction and monitoring of road visibility. This not only helps in early warning of traffic accidents but also provides scientific basis for traffic management authorities to take effective measures to ensure road traffic safety.

Therefore, research on machine learning-based road visibility prediction in Shenzhen holds significant practical significance and application prospects. By delving into the formation laws and influencing factors of road visibility and combining advanced technological means, it can provide strong support for traffic management and urban operations in Shenzhen, promoting the safety and efficiency of urban traffic.

Motivation

In rapidly developing cities like Shenzhen, the prediction and monitoring of road visibility are crucial for ensuring traffic safety and enhancing urban operational efficiency. However, traditional prediction methods may be limited by the timeliness and accuracy of data collection, leading to unreliable or untimely predictions. Hence, there is an urgent motivation and purpose for research into machine learning-based road visibility prediction.

Firstly, by leveraging machine learning algorithms combined with meteorological data from Shenzhen, it is possible to improve the accuracy and precision of predicting road visibility changes. This aids in early warning of potential traffic accidents, reducing traffic congestion, and minimizing casualties and economic losses caused by accidents.

Secondly, machine learning technology can effectively uncover patterns and influencing factors in road visibility changes, providing scientific basis and decision support for traffic management authorities. By delving into the mechanisms and related factors of visibility formation, targeted traffic management measures can be devised to enhance urban traffic safety and operational efficiency.

Additionally, with the continuous development and application of machine learning technology, its application in the field of road visibility prediction holds significant practical significance. This not only promotes the intelligent and informatization of traffic management but also provides valuable experience and reference for traffic management in similar cities.

Therefore, this study aims to utilize machine learning technology, coupled with meteorological data from Shenzhen, to achieve accurate prediction and real-time monitoring of road visibility. This provides scientific basis and decision support for traffic management authorities, fostering traffic safety and operational efficiency in Shenzhen.

The problem our solve

The problem addressed in this study is the prediction of road visibility in Shenzhen. This involves understanding the correlation between various meteorological factors such as precipitation, humidity, and wind speed, and their impact on road visibility. The goal is to develop predictive models that can accurately forecast road visibility based on real-time image data from the city's road monitoring system and meteorological data.

Method

The methodology employed in this study involves several steps. Firstly, historical data analysis is conducted to identify correlations between meteorological factors and road visibility. Distribution maps are then generated to visualize these relationships. Subsequently, multiple machine learning algorithms including multiple linear regression, ridge regression, and random forest are utilized to construct predictive models. These models are trained using the historical data and evaluated using various metrics such as Mean Squared Error (MSE), Pearson correlation coefficient (R), Mean Absolute Error (MAE), Explained Variance Score (EVS), as well as graphical analyses such as Q-Q plots and residual plots.

Contributions and Novelty

The main contributions of this research lie in providing an effective method for predicting road visibility in Shenzhen. The study demonstrates the superiority of the random forest model in terms of both prediction precision and stability compared to other methodologies. Additionally, the research provides valuable insights for meteorological monitoring and traffic management in similar urban settings by uncovering the correlations between meteorological factors and road visibility and offering predictive models to enhance decision-making processes.

2. Method (Use at least two pages to illustrate the methodology)

2.1 Multiple linear regression

Multiple linear regression models:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon \tag{1}$$

Asume

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \tag{2}$$

Minimize the function

$$Q(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 \tag{3}$$

Finding the partial derivative for each parameter that needs to be estimated, we can get a series of equations as follows

$$\begin{aligned} &\sum (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}) = 0 \\ &\sum (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}) x_{i1} = 0 \\ &... \\ &\sum (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}) x_{ip} = 0 \end{aligned} \tag{4}$$

This is equivalent to the sum of the residuals of each element being 0

$$e^T X = 0 \tag{5}$$

Because $e^T = y - X\hat{\beta}$

$$\begin{aligned} &(y - X\hat{\beta})^T X = 0 \\ &y^T X = \hat{\beta}^T X^T X \\ &X^T y = X^T X \hat{\beta} \\ &\hat{\beta} = (X^T X)^{-1} X^T y \end{aligned} \tag{6}$$

Define $H = (X^T X)^{-1} X^T$, so

$$\hat{\beta} = Hy \tag{7}$$

MSE

$$\frac{Q(\beta_0, \beta_1, ..., \beta_p)}{n} = \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2}{n} \tag{8}$$

2.2 Ridge regression

On the basis of multiple linear regression, a penalty for coefficient values is added

$$\hat{\beta}_{\text{ridge}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I}_{\mathrm{n}})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \tag{9}$$

I_n is the identity matrix, and the diagonal is all 1, similar to "mountain."

λ is the ridge coefficient, the value of which can be changed by changing the value of the diagonal of the identity matrix

The cost function becomes:

$$\begin{aligned} J_{\beta}(\beta) &= \sum_{\mathrm{i}=1}^{\mathrm{p}} (y_i - X_i\beta)^2 + \lambda \sum_{\mathrm{j}=0}^{\mathrm{n}} \beta_{\mathrm{j}}^2 \\ &= \sum_{\mathrm{i}=1}^{\mathrm{p}} (\mathrm{y}_1 - \mathrm{X}_1\beta) + \lambda ||\beta||^2 \end{aligned} \tag{10}$$

Ridge regression is an improved least squares estimation method, which obtains the regression coefficient at the cost of losing part of the information and reducing the accuracy by giving up the unbiased nature of the least squares method, which is a more practical and reliable regression method, and the fitting of the data with outliers is stronger than that of the least squares method.

2.3 Random Forest Model

Basic Principles of the Random Forest Model

Random forest is an ensemble learning method that constructs multiple decision trees and combines their results through voting or averaging to produce the final prediction. The primary advantage of this method is that by aggregating multiple models, it effectively addresses overfitting issues and enhances the model's generalization ability.

Modeling Process and Key Parameters of the Random Forest Model

Modeling Process

1. Bootstrap Sampling: From the original dataset, draw multiple subsets using the bootstrap sampling method (sampling with replacement).
2. Decision Tree Construction: For each subset, construct a decision tree. At each node, randomly select a subset of features to determine the best split.
3. Repetition: Repeat the above steps until the specified number of decision trees is generated.

Key Parameters

1. Number of Trees: Generally, the more decision trees in the forest, the better the model's performance, although this increases computational cost.
2. Number of Randomly Selected Features**: At each node split, a subset of features is randomly selected. Typically, the number of selected features is the square root or the logarithm of the total number of features. The choice of feature subset size affects the model's bias and variance.

Inner Mechanism of the Random Forest Model

The inner mechanism of the random forest primarily revolves around its randomness and ensemble characteristics:

1. Randomness: This arises from both sample randomness and feature randomness, leading to different decision trees and increasing model diversity.
2. Ensemble: By combining the predictions of multiple decision trees through voting (for classification) or averaging (for regression), the model effectively reduces variance, thereby enhancing stability and accuracy.

To ensure accurate predictions from the random forest, it is essential to provide useful information and ensure that each tree can independently offer its opinion. Consequently, when combined, the collective decision is both accurate and reliable.

Additionally, random forest models offer some level of interpretability. They can indicate feature importance, helping to explain the model's predictions. These characteristics make random forests highly effective for a wide range of practical applications.

3. Experiments and Results

3.1 Data Collection, Preprocessing and Analysis

Data Collection from the website:
https://opendata.sz.gov.cn/data/dataSet/toDataDetails/29200_00903518
This dataset contains hourly telemetry data from Shenzhen, with 3,730 records and 64 fields. The data types are primarily integers and strings. Some example fields include wind direction, cloud height, relative humidity, datetime, surface minimum temperature, grassland maximum temperature, automatic precipitation amount, minimum station pressure, maximum wind speed, and more.
Collection Method: The dataset appears to be collected through automated telemetry equipment, capturing various meteorological parameters along with timestamps.

Collection Time: The timestamps in the dataset range from August 9, 2015, to April 6, 2020, depending on the specific record.

Preprocessing
Step1: we find the dataset that some characteristic sets have lots of missing value, so we first delete the sets.

Step2: because the dataset that has been processed still have some missing value, we chose to change the missing value to 0.

Step3: we calculate the relationship value for the visibility and other characteristic sets.
The result below:

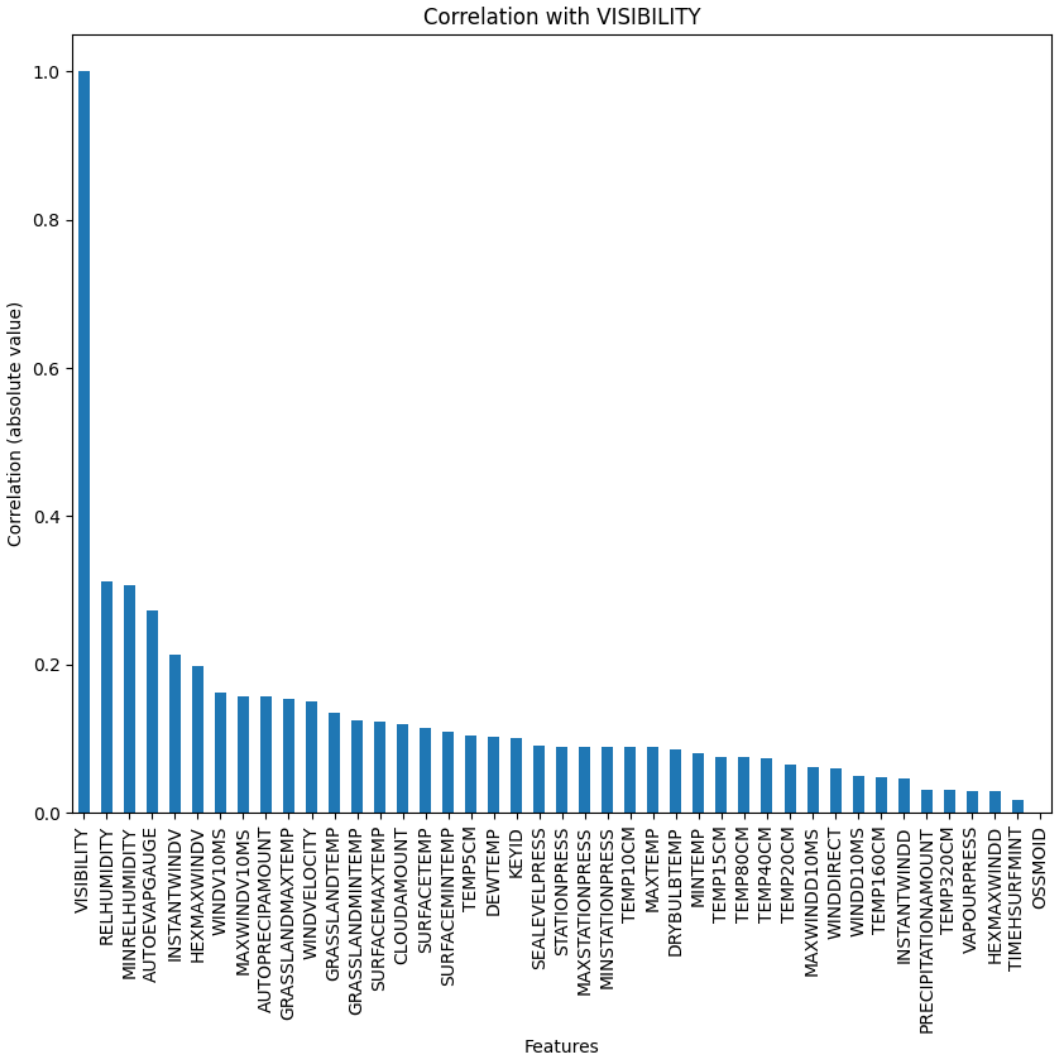


Figure 2 correlation with VISIBILITY

Step4: we found that the dataset have so many characteristic and many charateritic have low correlation with visibility, so we only choose the characteristic that the correlation value greater than 0.15.

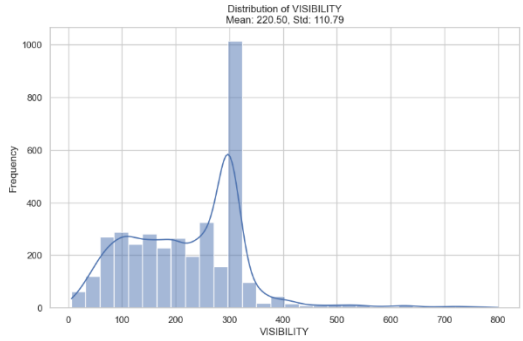
The result below:

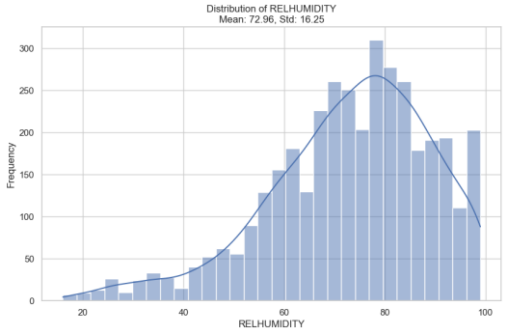
Feature	Correlation
VISIBILITY	1
RELHUMIDITY	0.311094
MINRELHUMIDITY	0.306294
INSTANTWINDV	0.212944
HEXMAXWINDV	0.198315
WINDV10MS	0.162007
MAXWINDV10MS	0.156455
AUTOPRECIPAMOUNT	0.156363
GRASSLANDMAXTEMP	0.152857

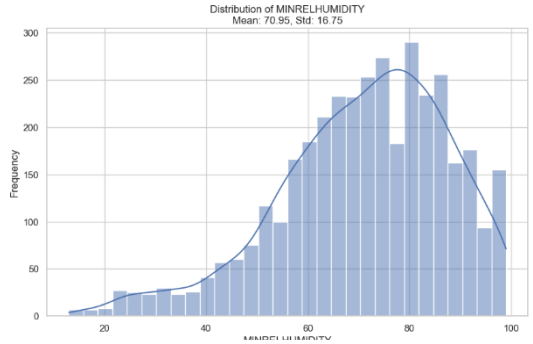
We also delete the AUTOEVAPGAUGE, because the sets is have so many missing values.

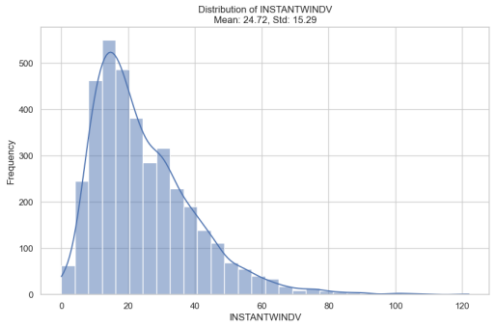
Step5: we explore the distribution for every characteristic.

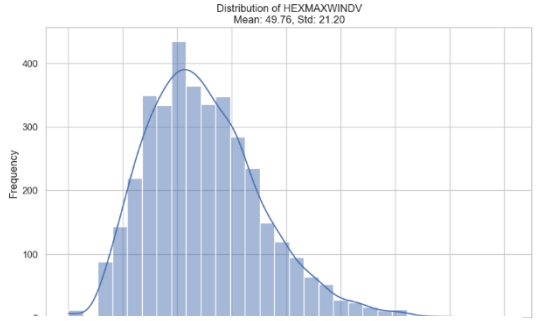
The result below:

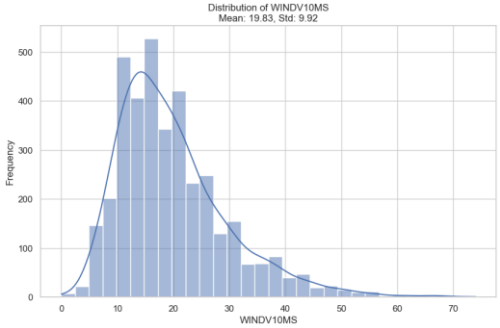


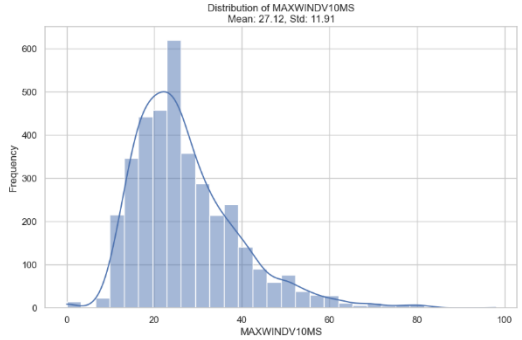


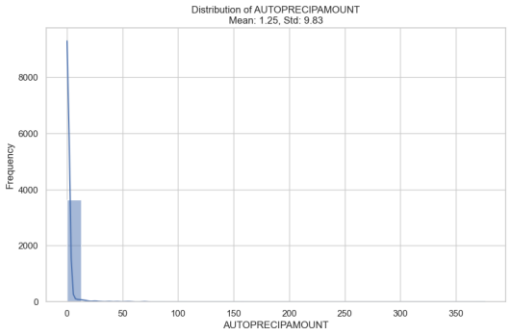


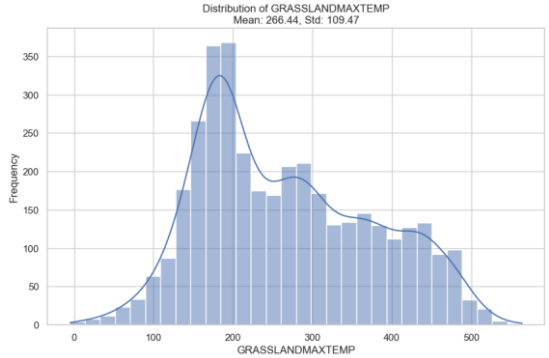












Expect the autoprecipamount, other characterstic similar adapt to normal distribution.

Step6:
We calculate the mean and variance for each characteristic.
The result below.

	mean	std
VISIBILITY	220.496	110.7851
RELHUMIDITY	72.96193	16.2497
MINRELHUMIDITY	70.95121	16.75499
INSTANTWINDV	24.71689	15.29268
HEXMAXWINDV	49.76139	21.20455
WINDV10MS	19.82949	9.918758
MAXWINDV10MS	27.12172	11.91181
AUTOPRECIPAMOUNT	1.246381	9.825435
GRASSLANDMAXTEMP	266.4373	109.4651

3.2 Evaluation Metrics

Model performance was evaluated, and the mean square error (MSE), coefficient of determination (R2), mean absolute error (MAE), and explanatory variance score (EVS) were calculated. These evaluation metrics are used to measure the prediction accuracy and performance of the model.

MSE: A measure of the mean squared error between the predicted value and the actual value.

R2: The proportion of variation in the explanatory variable, ranging from 0 to 1, with closer to 1 indicating a better model.

MAE: The average absolute error between the predicted value and the actual value.

EVS: Interpretive variance score, which measures how well the model explains the variation in the data.

3.3 Experiments

- **Data Section**

1. **Data Reading and Preprocessing**: Read the data and separate the feature data (X) from the target data (y). The feature data is used for model training, and the target data is what we aim to predict (VISIBILITY).
2. **Dataset Splitting**: Split the dataset into a training set and a test set, with the test set comprising 20% of the data. The training set is used to train the model, while the test set is used to evaluate the model's performance.
3. **Feature Scaling**: Standardize the feature data so that it has a mean of 0 and a standard deviation of 1. Standardization can improve the model's convergence speed and predictive performance.

- **Model Section**

1. **Linear Regression**: Train the model using linear regression and evaluate its mean squared error (MSE) using ten-fold cross-validation. Linear regression is a simple linear model suitable for regression problems with linear relationships.

2. **Ridge Regression**: Train the model using ridge regression and evaluate its mean squared error (MSE) using ten-fold cross-validation. Ridge regression adds L2 regularization to the basic linear regression, which helps reduce overfitting.
3. **Random Forest Regression**: Train the model using random forest regression and evaluate its mean squared error (MSE) using ten-fold cross-validation. A random forest is an ensemble model composed of multiple decision trees, capable of capturing more complex nonlinear relationships.

- **Evaluation Section**

Model Evaluation Function: Evaluate the model's performance by calculating the mean squared error (MSE), coefficient of determination (R2), mean absolute error (MAE), and explained variance score (EVS). These evaluation metrics measure the model's predictive

1. accuracy and performance. MSE measures the average squared error between predicted and actual values; R2 indicates the proportion of variance explained by the model, ranging from 0 to 1, with values closer to 1 indicating a better model; MAE measures the average absolute error between predicted and actual values; EVS measures the degree to which the model explains the variance in the data.

- **Image Section**

1. **Predicted vs. Actual Values Comparison Plot**: Plot a scatter diagram of predicted versus actual values, with a red line indicating the ideal situation where predicted values equal actual values. This plot visually compares the differences between predicted and actual values.
2. **Residual Plot**: Plot a scatter diagram of predicted values versus residuals, with a red dashed line indicating zero residuals. This plot checks whether the residuals are systematically biased.
3. **QQ Plot**: Plot a Quantile-Quantile (QQ) plot of the residuals to check if they follow a normal distribution. This plot evaluates the normality assumption of the residuals and determines if the model is biased.
4. **Residual Histogram and Boxplot**: Plot a histogram and boxplot of the residuals. The histogram shows the shape of the residuals' distribution, while the boxplot compares the distribution and outliers of residuals across different models.
5. **Predicted vs. Actual Values Comparison Graph**: Plot a scatter diagram of predicted and actual values against sample indices. This graph visually compares the differences between predicted and actual values for different models.

Through these data and images, we can comprehensively evaluate the performance of each model, understand its predictive accuracy and residual distribution, and thus select the best model for practical application.

3.3 Experimental Results and Analysis

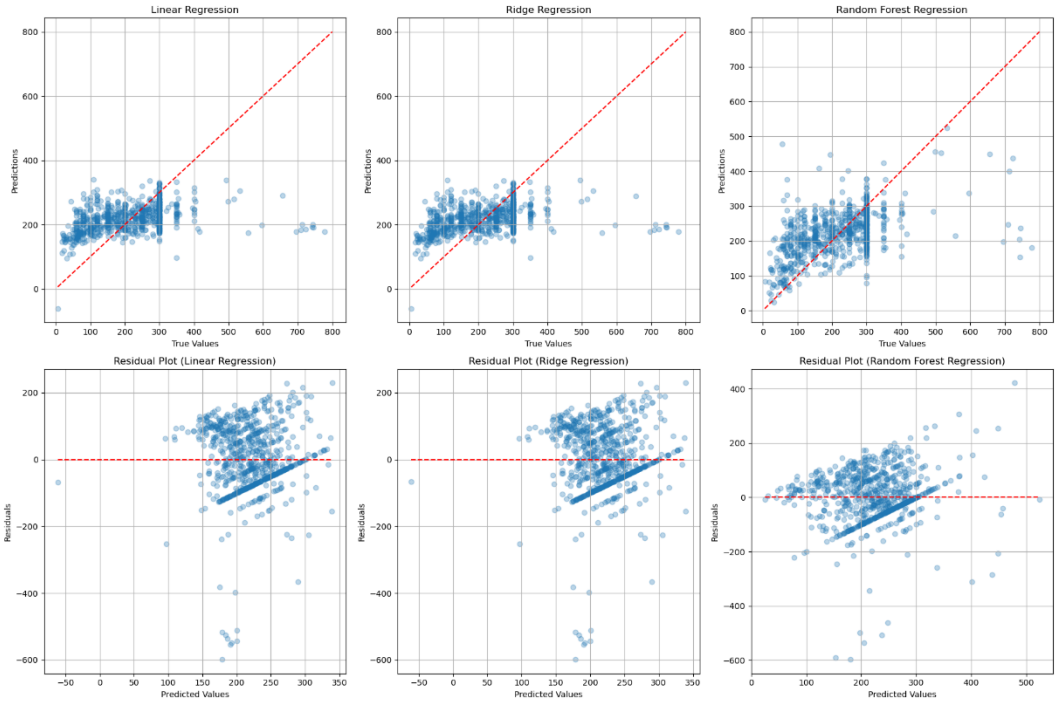
Table 1 Comparison of model evaluation indicators

Model	MSE	CV MSE	R2	MAE	Explained Variance Score
Linear Regression	11221.22	10390.05	0.129734	79.32443	0.131769

Ridge Regression	11220.58	10389.56	0.129784	79.32393	0.131819
Random Forest Regression	9790.59	8806.87	0.240687	71.50635	0.243476

Table 2 Model coefficient importance:

Model	**Coefficient Importance**							
Linear Regression	-27.276	-3.019	17.490	28.036	-8.023	-22.142	-12.064	5.567
Ridge Regression	-26.882	-3.405	17.492	27.961	-8.013	-22.074	-12.068	5.558
Random Forest Regression	0.269	0.109	0.116	0.108	0.081	0.083	0.021	0.212



Linear Regression

Plot Description: The top left plot shows the predictions of the linear regression model against the true values.

Analysis: The points are widely scattered around the red diagonal line (ideal predictions). This indicates that the linear regression model has a significant amount of error and may not be capturing the underlying patterns effectively.

Ridge Regression

Plot Description: The top middle plot shows the predictions of the ridge regression model against the true values.

Analysis: Similar to the linear regression plot, the points are scattered around the diagonal line but appear slightly more clustered, suggesting a modest improvement over plain linear regression.

Random Forest Regression

Plot Description: The top right plot shows the predictions of the random forest regression model against the true values.

Analysis: The points are more closely aligned along the diagonal line compared to the other two models, indicating better performance and a higher accuracy of predictions.

Residual Plots

Linear Regression

Plot Description: The bottom left plot shows the residuals (errors) of the linear regression model against the predicted values.

Analysis: The residuals show a clear pattern, suggesting that the model is not capturing all the underlying trends in the data. This indicates potential issues with model fit and assumptions.

Ridge Regression

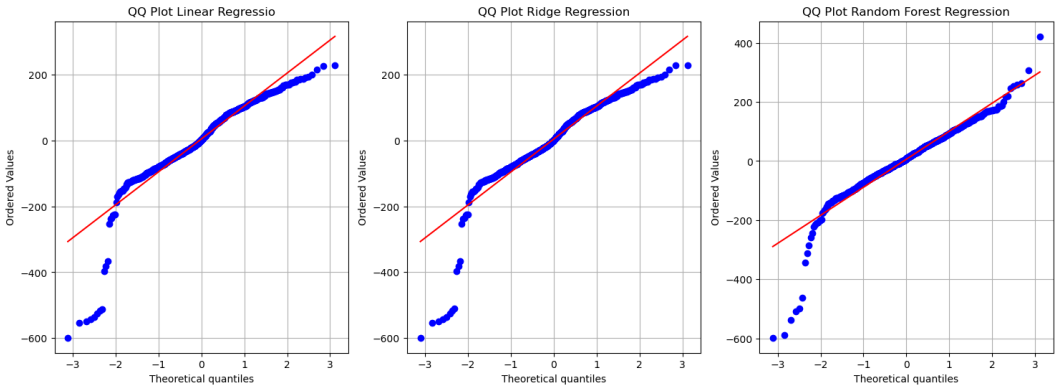
Plot Description: The bottom middle plot shows the residuals of the ridge regression model against the predicted values.

Analysis: The residuals still show a pattern but are more dispersed compared to the linear regression residuals, indicating a slight improvement in model fit but still room for better modeling.

Random Forest Regression

Plot Description: The bottom right plot shows the residuals of the random forest regression model against the predicted values.

Analysis: The residuals appear more randomly distributed, which is a good sign as it suggests the model captures the data patterns more effectively without systematic bias.



Linear Regression

Plot Description: The bottom left QQ plot shows the quantiles of the residuals from the linear regression model against the theoretical quantiles.

Analysis: The points deviate significantly from the red line at both ends, indicating that the residuals are not normally distributed and there may be issues with heteroscedasticity or non-linearity.

Ridge Regression

Plot Description: The bottom middle QQ plot shows the quantiles of the residuals from the ridge regression model against the theoretical quantiles.

Analysis: The points still deviate from the red line, particularly at the tails, but the fit is slightly better than the linear regression QQ plot, suggesting improved but not perfect normality.

Random Forest Regression

Plot Description: The bottom right QQ plot shows the quantiles of the residuals from the random forest regression model against the theoretical quantiles.

Analysis: The points are closer to the red line compared to the other models, indicating that the

residuals are more normally distributed, which suggests a better overall model fit and less bias.

Summary

Linear Regression: Exhibits significant prediction errors, systematic patterns in residuals, and non-normal residuals, indicating poor model fit.

Ridge Regression: Shows slight improvement over linear regression with better clustering of predictions and slightly more dispersed residuals, but still has noticeable issues with residual patterns and normality.

Random Forest Regression: Demonstrates the best performance with predictions closely aligned to true values, randomly distributed residuals, and residuals that approximate normal distribution well.

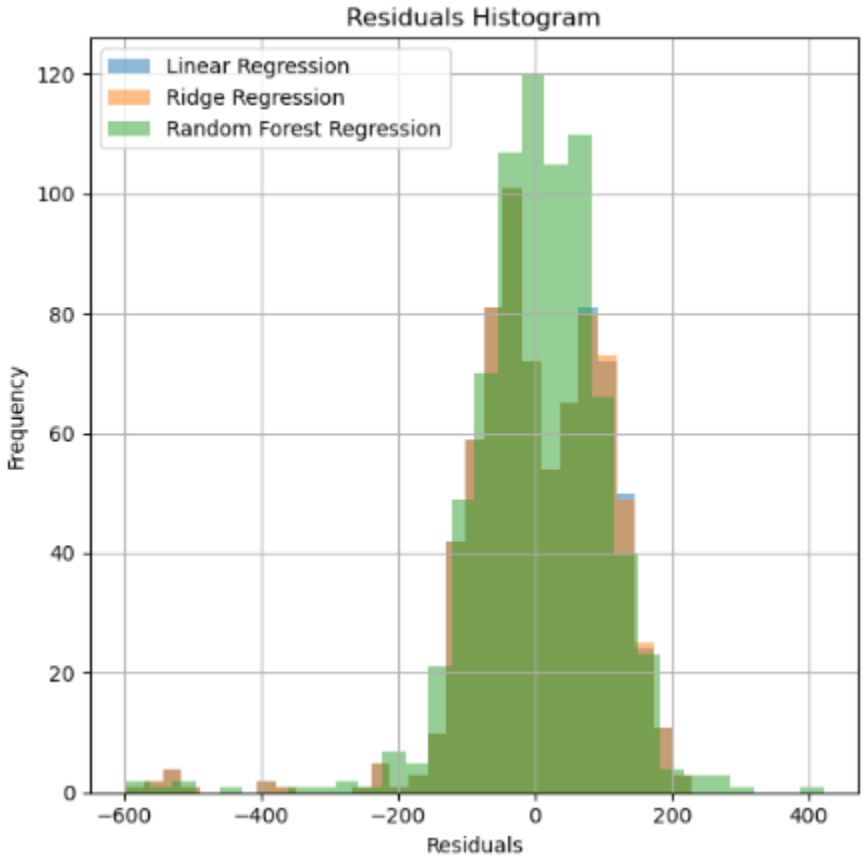


Figure 3 Residuals histogram

The residuals histogram compares the distribution of residuals (errors) for the three models:

Linear Regression: The residuals are roughly centered around zero, with a somewhat normal distribution but a wider spread compared to the other models.

Ridge Regression: The residuals distribution is similar to linear regression but appears slightly tighter, indicating that the regularization in ridge regression has helped reduce some variance in the predictions.

Random Forest Regression: The residuals are also centered around zero and have the tightest distribution among the three models, indicating that this model has the smallest errors and is potentially the most accurate.

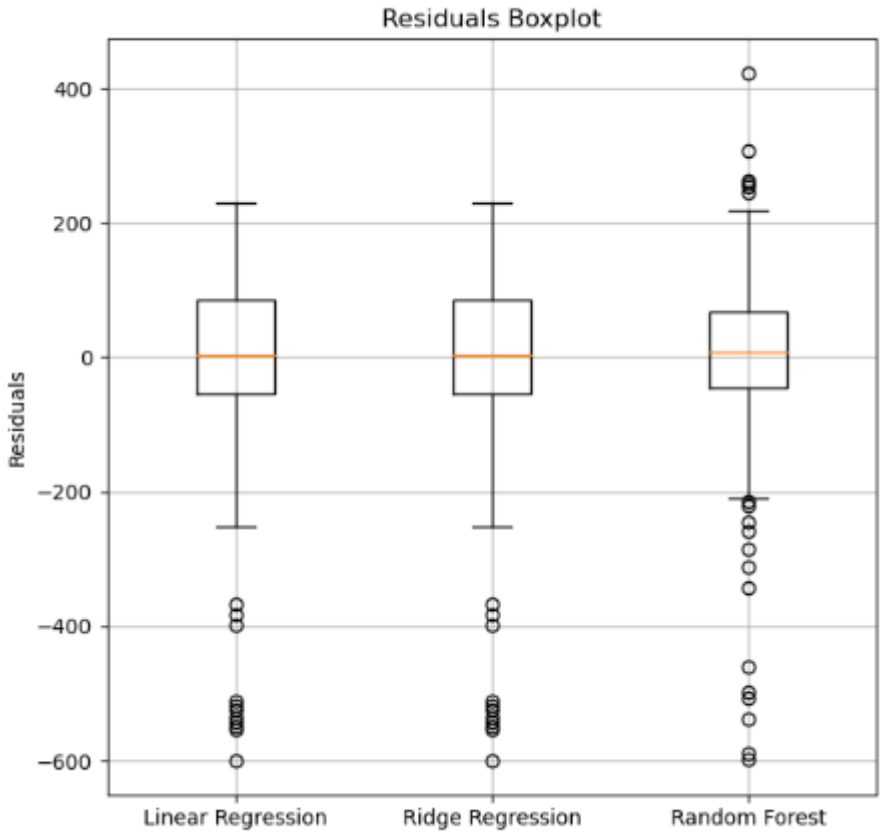


Figure 4 Residuals Boxplot

The residuals boxplot shows the spread and outliers of the residuals for each model:

Linear Regression: The interquartile range (IQR) is wider compared to ridge regression and random forest, with several outliers, especially on the lower end.

Ridge Regression: Similar to linear regression but with a slightly narrower IQR, indicating a reduction in variance due to regularization.

Random Forest Regression: The IQR is the narrowest, with fewer outliers, indicating that the random forest model has the most consistent performance with the least amount of error.

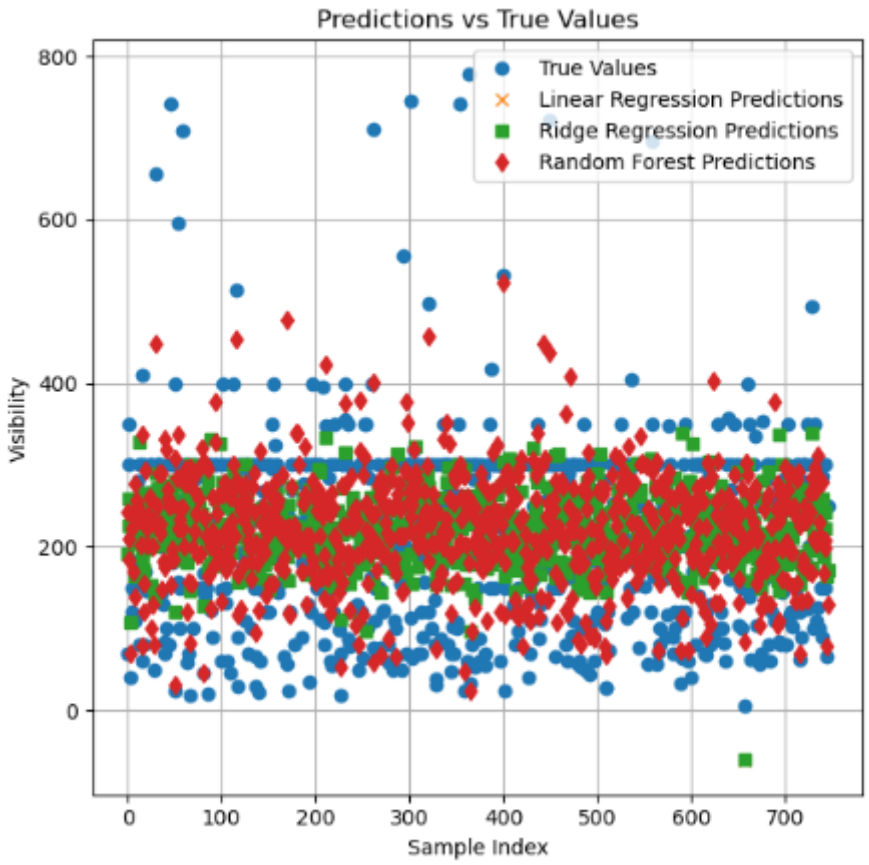


Figure 5 Prediction vs True values

This scatter plot compares the predicted values from each model against the true values:
True Values (blue dots): Represent the actual visibility values.
Linear Regression Predictions (green crosses): Predictions are scattered around the true values but with noticeable deviations, indicating some inaccuracy.
Ridge Regression Predictions (orange squares): Predictions are similar to linear regression but slightly more clustered around the true values, indicating an improvement in accuracy due to regularization.
Random Forest Predictions (red diamonds): Predictions are densely packed around the true values, suggesting the highest accuracy among the three models. The random forest model seems to predict visibility values more accurately and consistently.

4. Discussion

The dataset:

The data set from:

https://opendata.sz.gov.cn/data/dataSet/toDataDetails/29200_00903518

the dataset has lots of loss value, we pre-processed it that eliminated some attributes to make the dataset can be utilized.

The attributes cor-relationship:

only few attributes have high cor-relationship with visibility. We choose eight attributes to regress the visibility.

The linear model discussion:

Residual Distribution: The residuals are widely spread with many outliers, indicating high variance and low predictive accuracy.

Model Complexity: Low, easy to understand and interpret.

Interpretability: High, as each feature's coefficient can be explained as its contribution to the prediction result.

The ridge model discussion:

Residual Distribution: Compared to linear regression, the residuals are more tightly distributed, variance is reduced, and predictive accuracy is somewhat improved, though there are still some outliers.

Model Complexity: Medium, with a slight increase in complexity due to the introduction of regularization terms.

Interpretability: Medium, while the coefficients can still be interpreted as contributions to the prediction result, the interpretability is somewhat reduced compared to linear regression due to regularization

The random forest model discussion:

Residual Distribution: The residuals are the most tightly distributed, with the fewest outliers, indicating the highest predictive accuracy and the most stable performance.

Model Complexity: High, consisting of multiple decision trees, which requires substantial computational resources.

Interpretability: Low, as it is difficult to directly interpret the contribution of each feature. Although feature importance can provide some understanding of the model, it is far less intuitive than linear regression.

The results indicate that the random forest model outperforms other methodologies in terms of both prediction precision and stability. This research not only offers an effective method for predicting road visibility in Shenzhen but also provides valuable insights for meteorological monitoring and traffic management in similar urban settings.

5. Conclusions

The random forest model can effectively predict the Shenzhen’s road visibility, comparing to other model. The MSE of prediction is which indicates the highest accuracy. The MSE

of ridge model is . which can always predict correctly but sometimes make abnormal values. The MSE of the linear model is whose majority is right but compared with the before two , it has more abnormal

Reference

[1]袁敏,李忠堃,洪震宇,等.基于机器学习对机场能见度预测模型研究[J].舰船电子工程,2023,43(12):182-186+237.

[2]吴晴霞. 基于气象观测数据及监控图像的雾景能见度的检测与预测[D].重庆大学,2023.DOI:10.27670/d.cnki.gcqdu.2021.001270.

[3]李元龙. 基于深度学习的空气能见度等级检测[D].西安石油大学,2024.DOI:10.27400/d.cnki.gxasc.2023.000459.

Appendix

Appendix A: xxx

Appendix B: xxx

指导教师批阅意见：   成绩评定：   指导教师签字： 年　　月　　日
备注：

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后 10 日内。