POLITECNICO
MILANO 1863

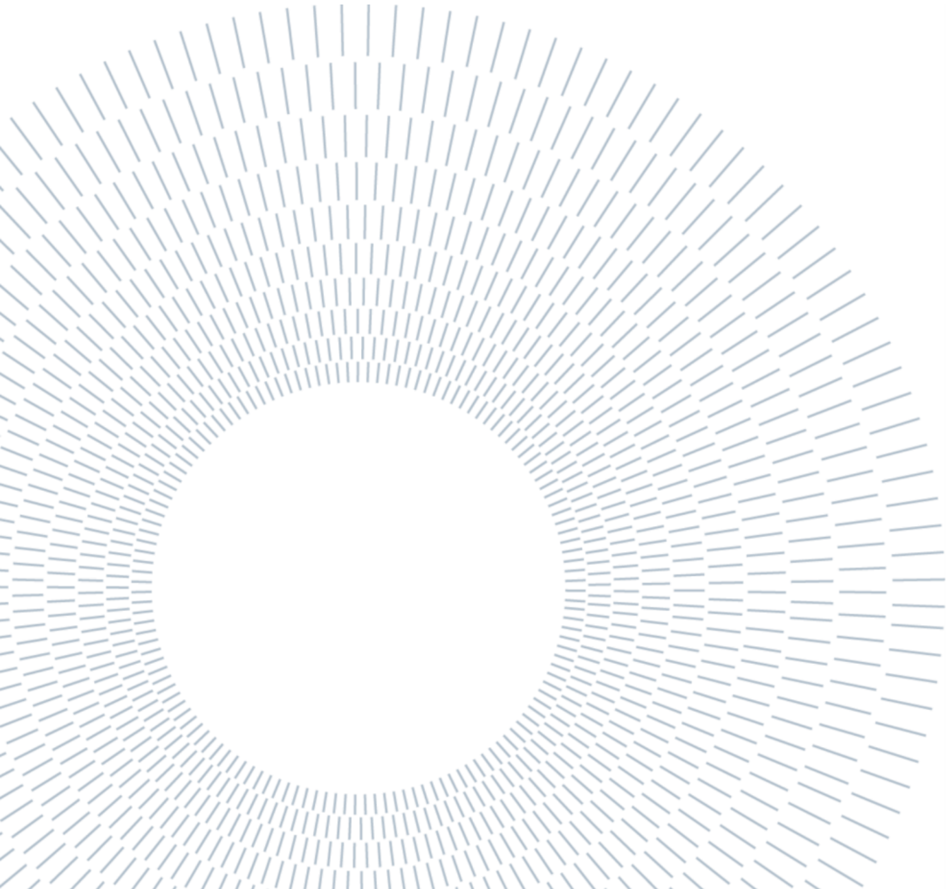
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Risk evaluation of ground motion using machine learning models for the Lombardy infrastructure

TESI DI LAUREA MAGISTRALE IN
CIVIL ENGINEERING FOR RISK MITIGATION

Author: **Ziyang Wang**

Student ID:	952432
Advisor:	Scaioni Marco
Academic Year:	2021-2022



Abstract

The issue of damage to infrastructure, including highways and railways, caused by ground motion in Lombardy is a problem which cannot be disregarded. Based on the ground motion data from European Ground Motion Service (EGMS), the reason of this process could be analyzed. The identification of relevant condition factors associated with ground motion, and their relationship with the latter, can be accomplished through the application of AutoML function in ArcGIS Pro. Through the file derived from Machine Learning algorithms, the ground motion risk map of infrastructure can be generated. The comparison between the risk map and the EGMS data indicates a high level of agreement for areas deemed to be at risk. However, analysis of highway ground motion is overestimated in areas where ground motion is concentrated. The conclusion illustrates the characteristics of infrastructure which is prone to ground motion and it can help designers to implement precautionary measures in infrastructure ground motion risk management.

Key-words: ground motion, EGMS, machine learning.

Abstract in lingua italiana

La questione dei danni alle infrastrutture, comprese autostrade e ferrovie, causati dal movimento del suolo in Lombardia è un problema che non può essere trascurato. Sulla base dei dati di movimento del suolo dell'European Ground Motion Service (EGMS), è stato possibile approfindire il motivo di questi spostamenti. L'identificazione dei fattori di condizione rilevanti associati al movimento del suolo e la loro relazione con quest'ultimo può essere realizzata attraverso l'applicazione della funzione AutoML in ArcGIS Pro. Attraverso le informazioni derivate dagli algoritmi di apprendimento automatico, è possibile generare la mappa del rischio di movimento del suolo in prossimità di una infrastruttura. Il confronto tra la mappa del rischio ei dati EGMS indica un alto livello di accordo per le aree ritenute a rischio. Tuttavia, l'analisi del movimento del suolo in corrispondenza delle autostrade risulta essere sovrastimata nelle aree di maggiore movimento del suolo. La tesi illustra le caratteristiche delle infrastrutture soggette a movimenti del suolo e può aiutare i progettisti a implementare misure precauzionali nella gestione del rischio legato al movimento del suolo.

Parole chiave: moto del suolo, EGMS, apprendimento automatico.

Contents

Introduction

Ground motion is a term used to describe the movement of the ground surface that can result from various natural and man-made events such as earthquakes, volcanic eruptions, heavy rain and human activities like construction (Hill et al., 2002; Grünthal et al). It is a complex phenomenon that is influenced by a number of factors, including geology, topography, soil conditions, and environmental factors (Wang and Xie, 2010; Ni and Wu, 2021). By analyzing these factors, the importance can be known. Based on importance and characteristics of regions, corresponding measures could be taken.

The analysis of ground motion is a complex and multifaceted process, requiring the integration of multiple data sources and the application of advanced technologies. In recent years, the integration of Geographic Information Systems (GIS) and machine learning techniques has shown great promise in this area, providing new and innovative methods for data analysis and model building (Karimzadeh et al.,2014;. Khosravikiaet al., 2021). More studies highlighted the application of other machine learning techniques in ground motion prediction and any machine learning models such as Naïve Bayesian, Naïve Bayesian, Logistic model tree and Random Forest are used in ground motion analysis (Trugman and Shearer, 2018; Kong et al., 2019; Li and Zhang, 2023). They suggested that machine learning approaches to ground motion prediction could be a new powerful tool in the next generation of seismic hazard assessments.

The ground motion data is provided by European Ground Motion Service (EGMS) which provides consistent and reliable information regarding natural and anthropogenic ground motion over the Copernicus Participating States and across national borders, with millimeter accuracy (Crosetto et al, 2020). EGMS provides service in European Union, including the Lombardy region of Italy. The data collected by EGMS allows us to study the factors that influence ground motion and develop more accurate models for forecasting and mitigating its effects.

In this article, the study of ground motion in Lombardy provides valuable insights into the factors that contribute to this phenomenon and highlights the importance of considering a range of factors in the analysis of ground motion. Seven factors are considered in this research including elevation, slope angel, slope aspect, rainfall, curvature, solar radiation and normalized difference vegetation index (NDVI). Machine learning models such as Decision Tree (DT), Linear regression (LR), Light GBM (LG), XGBoost (XG), Random Forest (RF) and Extra Trees (ET) are used in analysis and all the processes are based on ArcGIS Pro. The research area mainly focuses on infrastructure like highway and railway. With different range of area which is near to the infrastructure, the data trained separately. According to the result of machine learning, the best model of infrastructure will be selected and we will use this model to make an infrastructure risk map in Lombardy. In addition, based on the importance of factors, we also need to provide corresponding prevention and control strategies for locations with high importance factors.

1. Study area

Lombardy, located in northern Italy, is an area of significant seismic activity, with a history of ground motion events caused by both natural and man-made processes (Brunelli et al., 2023). The analysis of ground motion in Lombardy is therefore of great importance, both for improving our understanding of the underlying causes of these events and for informing mitigation efforts to reduce the risk to human life and property.

The orography of Lombardy is characterised by three distinct belts; a northern mountainous belt constituted by the Alpine relief, a central piedmont area of mostly alluvial pebbly soils, and the Lombard section of the Padan Plain in the south of the region. The most important mountainous area is the Alpine zone, which includes the Lepontine and Rhaetian Alps—Piz Bernina (4,020 m), the Bergamo Alps, the Ortler Alps and the Adamello massif. It is followed by the Alpine foothills zone Prealpi, the main peaks of which are the Grigna Group (2,410 m), Resegone (1,875 m), and Presolana (2,521m).

This study focuses on the infrastructure including highway and railway. The Lombardy Highway and Railway, is a major transportation located in the northern region of Italy. Spanning approximately 160 0km and 2115km, it connects Milan to several smaller towns and cities along the shores of Lake Maggiore and Lake Como. However, there are some geological risks associated with ground motion of the Lombardy infrastructure including landslides, seismic activity and ground settlement.

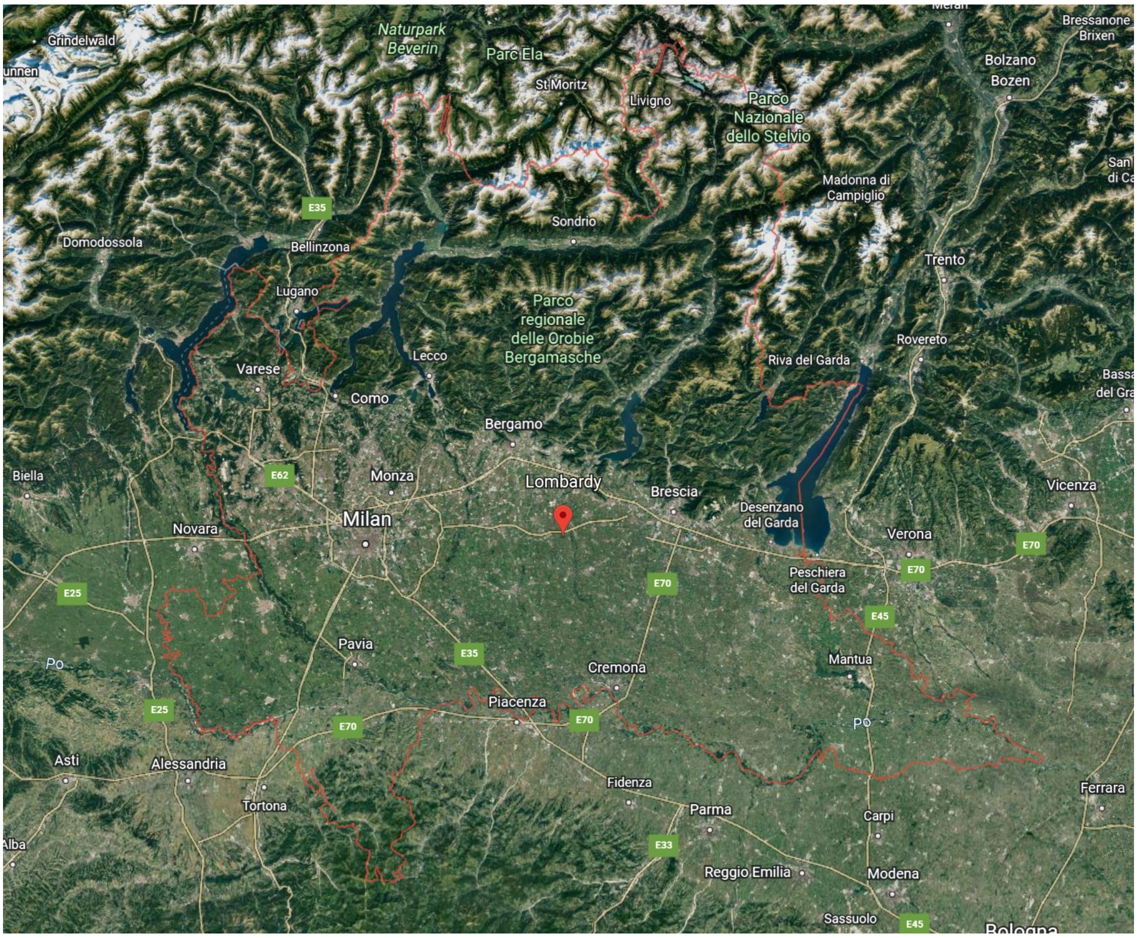


Figure 1.1: Lombardy region

The Lombardy region is prone to landslides, particularly during heavy rains and snow melts, which can cause damage to the infrastructure (Antonielli et al.,2019). Lombardy is located in a seismically active area, and the highway passes through several zones that are at risk of earthquakes (Garbin et al.,2013). This can pose a risk to the stability of the infrastructure such as bridges and tunnels. High elevation of groundwater and soil instability is an ordinary condition in Lombardy (Gattinoni et al.,2017), which can lead to ground settlements (Ikuemonisan et al.,2021). These conditions would result in ground motion to the infrastructure, particularly in areas where the soil is composed of soft or poorly compacted material.

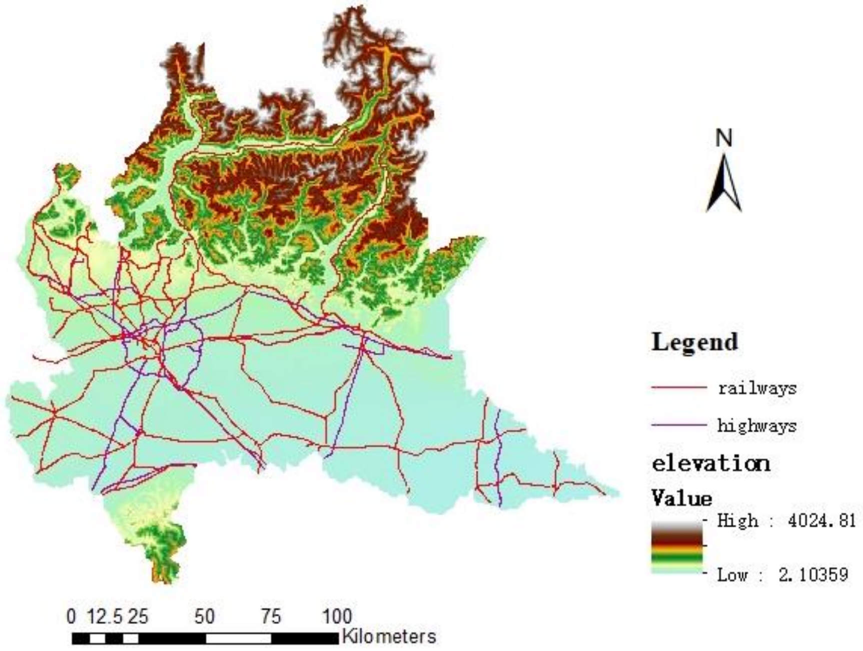


Figure 1.2: The infrastructure in Lombardy

2. Methodology

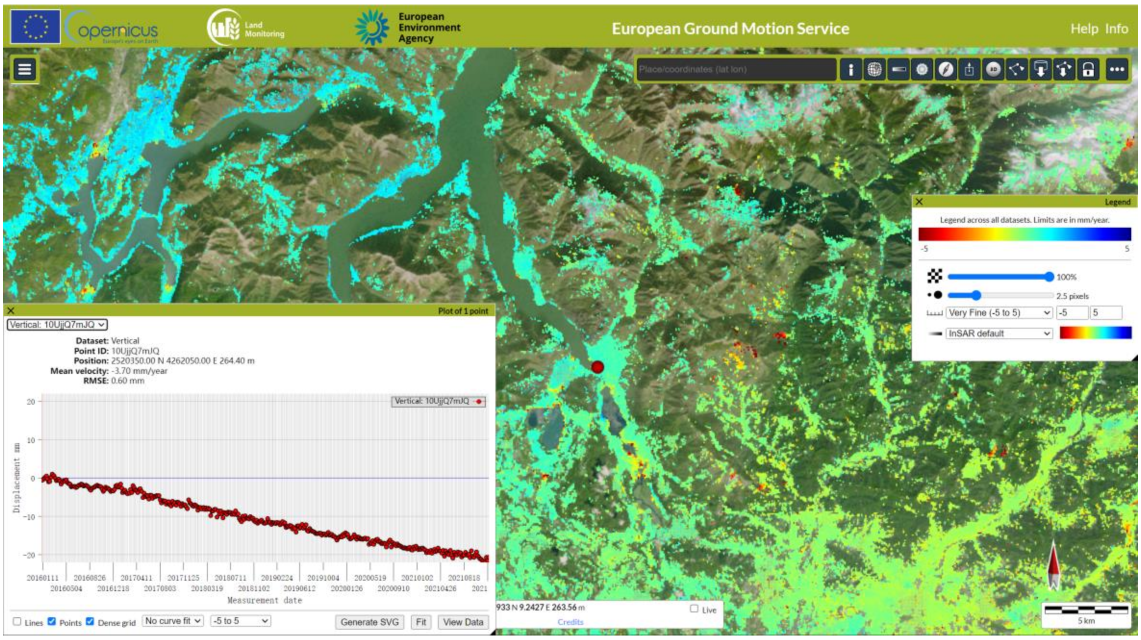
A final chapter containing the main conclusions of your research/study and possible future developments of your work have to be inserted in this chapter.

2.1 ArcGIS Pro

ArcGIS Pro is a professional desktop GIS application designed for advanced geospatial analysis, visualization, and collaboration. It provides access to 2D and 3D mapping, advanced analysis tools, data management, and real-time collaboration capabilities. It is a part of the Esri ArcGIS platform and is used by various industries for tasks such as spatial analysis, data management, and collaboration. In this study, the AutoML function in Toolboxes will be used for machine learning.

2.2 EGMS data

The data of ground motion is download from EGMS and we can view and download EGMS data via the EGMS Explorer. Because interferometric processing of a time series of acquisitions from synthetic aperture radar (SAR) satellites, it is possible to detect and measure ground motion phenomena, typically caused by landslides, subsidence, earthquakes or volcanic activity, with millimeter-scale precision. This enables, for example, monitoring of the stability of slopes, mining areas, buildings and infrastructures. In Fig, every point on the map has the ground motion data from 2016 to 2021 and it includes vertical motion data and horizontal motion data.



Dataset: Vertical
Point ID: 10UjjQ7mJQ
Position: 2520350.00 N 4262050.00 E 264.40 m
Mean velocity: -3.70 mm/year
RMSE: 0.60 mm

□ Lines ☑ Points ☑ Dense grid | No curve fit | -5 to 5 | Generate SVG | Fit | View Data

933 N 9.2427 E 263.56 m □ Live

Legend across all datasets. Limits are in mm/year.

-5 5

100%

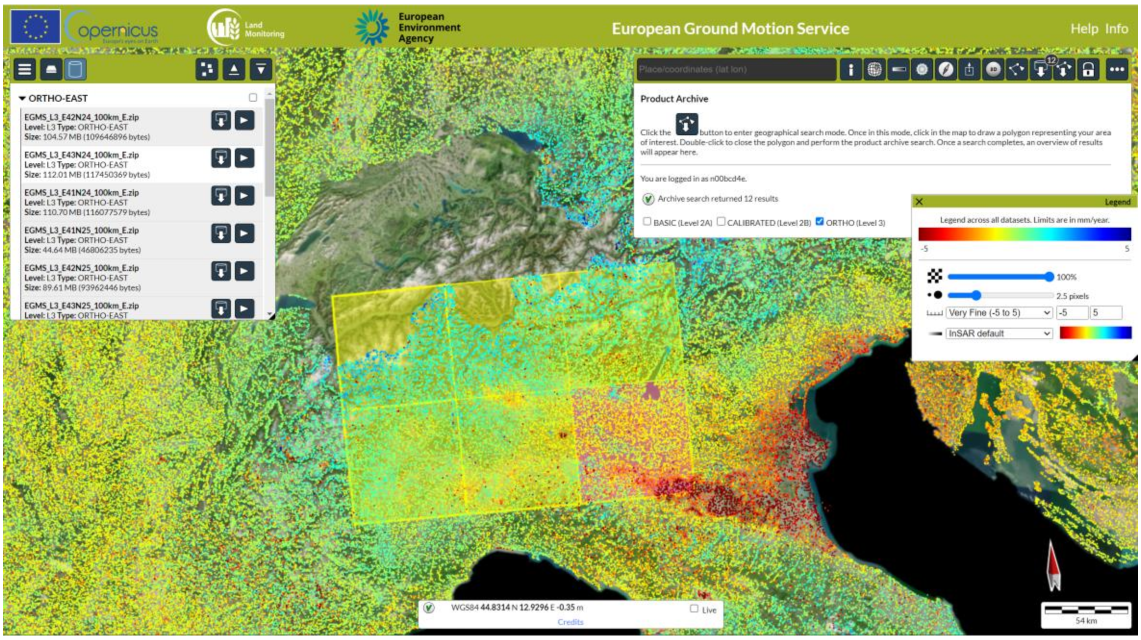
2.5 pixels

Very Fine (-5 to 5) | -5 | 5

InSAR default

Figure 2.1: Interaction with InSAR data in the EGMS Explorer interface

Functionality for downloading EGMS products is completely and seamlessly integrated into the EGMS Explorersystem (Fig 2.1). We can select the area we are interested in. Then, the right toolbar will show the information about download and the range of area which will be download is 100km ×100km (Fig 2.2).



▾ ORTHO-EAST □

EGMS_L3_E42N24_100km_E.zip
Level: L3 Type: ORTHO-EAST
Size: 104.57 MB (109646896 bytes)

EGMS_L3_E43N24_100km_E.zip
Level: L3 Type: ORTHO-EAST
Size: 112.01 MB (117450369 bytes)

EGMS_L3_E41N24_100km_E.zip
Level: L3 Type: ORTHO-EAST
Size: 110.70 MB (116077579 bytes)

EGMS_L3_E41N25_100km_E.zip
Level: L3 Type: ORTHO-EAST
Size: 44.64 MB (46806235 bytes)

EGMS_L3_E42N25_100km_E.zip
Level: L3 Type: ORTHO-EAST
Size: 89.61 MB (93962446 bytes)

EGMS_L3_E43N25_100km_E.zip
Level: L3 Type: ORTHO-EAST

Product Archive

Click the button to enter geographical search mode. Once in this mode, click in the map to draw a polygon representing your area of interest. Double-click to close the polygon and perform the product archive search. Once a search completes, an overview of results will appear here.

You are logged in as n00bcd4e.

✓ Archive search returned 12 results

□ BASIC (Level 2A) □ CALIBRATED (Level 2B) ☑ ORTHO (Level 3)

Legend across all datasets. Limits are in mm/year.

-5 5

100%

2.5 pixels

Very Fine (-5 to 5) | -5 | 5

InSAR default

WGS84 44.8314 N 12.9296 E -0.35 m □ Live

Figure 2.2: Overview of the workflow to search and download EGMS products.

The document includes vertical motion data and horizontal motion data. In this study, we only analyze the vertical ground motion. Two types of ground motion data are provided including .csv and .tiff. The .csv document includes latitude, longitude, mean velocity, acceleration, seasonality and ground motion data each month. As for .tiff data, it can be input in to ArcGIS Pro and its format is raster map. The value of each pixel means the millimeter per year of vertical ground motion (Fig 2.3).

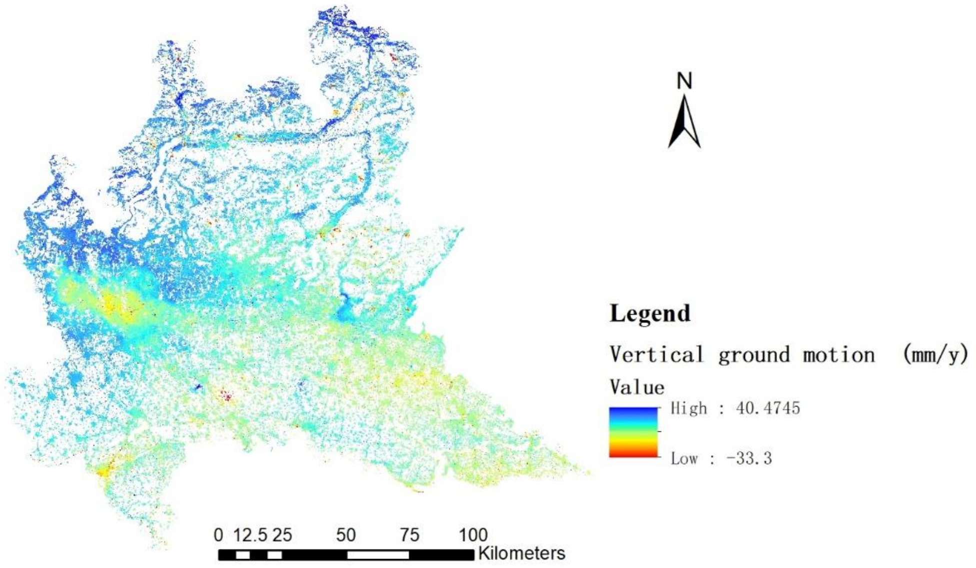


Figure 2.3: Vertical ground motion data from EGMS

2.3 Condition factors data

The ground motion is caused by natural and manmade reason. Its level is controlled by geographical and geological factors. Therefore, 7 ground motion condition factors are introduced in this research, including elevation, slope angel, slope aspect, rainfall, curvature, solar radiation and normalized difference vegetation index (NDVI). Data for some conditioning factors in 2020 are selected. The Digital Terrain Model (DTM) with a resolution of 20 m × 20 m is downloaded from the Geoportale della Lombardia (https://www.geoportale.regione.lombardia.it/), and used to generate the maps of topographic factors in ArcGIS (Fig. 4a-e). The NDVI map is provided by the vegetation part of the Copernicus Global Land Service (CGLS), which is a component of the Land Monitoring Core Service (LMCS) of Copernicus, the European flagship program on Earth Observation (https://land.copernicus.eu/global/). The rainfall map is received from GISgeography (https://gisgeography.com/gis-weather-data-sources/).

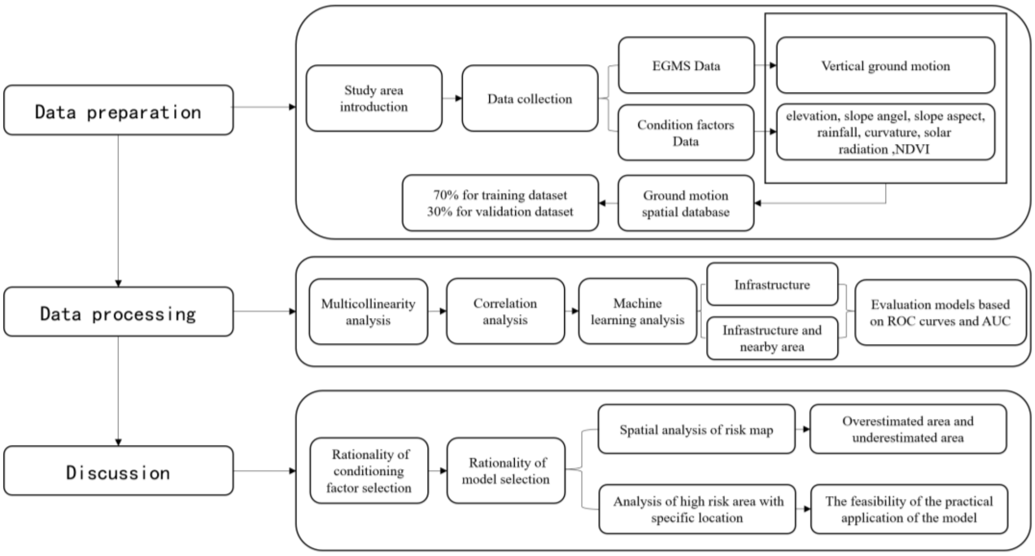
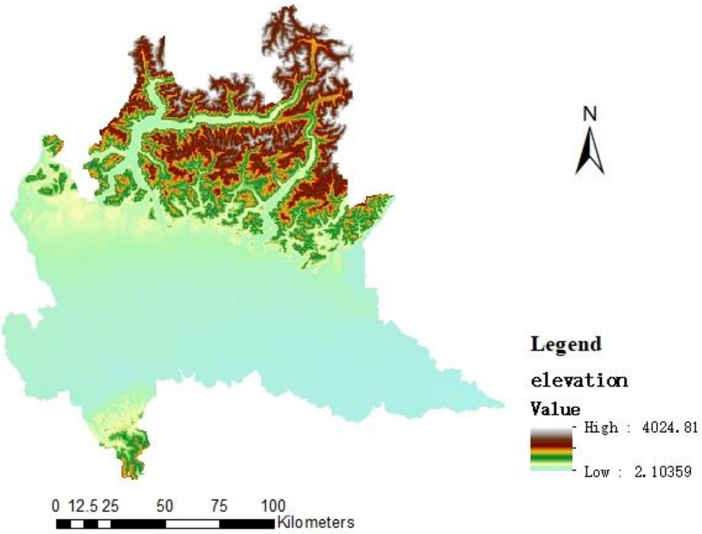
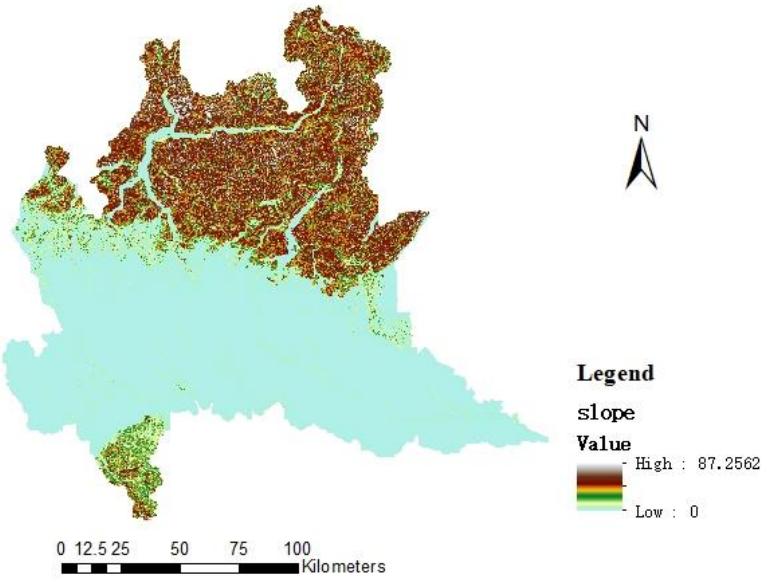

Data preparation
Study area introduction
Data collection
EGMS Data
Condition factors Data
Vertical ground motion
elevation, slope angel, slope aspect, rainfall, curvature, solar radiation ,NDVI
70% for training dataset 30% for validation dataset
Ground motion spatial database
Data processing
Multicollinearity analysis
Correlation analysis
Machine learning analysis
Infrastructure
Infrastructure and nearby area
Evaluation models based on ROC curves and AUC
Discussion
Rationality of conditioning factor selection
Rationality of model selection
Spatial analysis of risk map
Overestimated area and underestimated area
Analysis of high risk area with specific location
The feasibility of the practical application of the model

Figure 2.4: The flowchart of ground motion risk research.

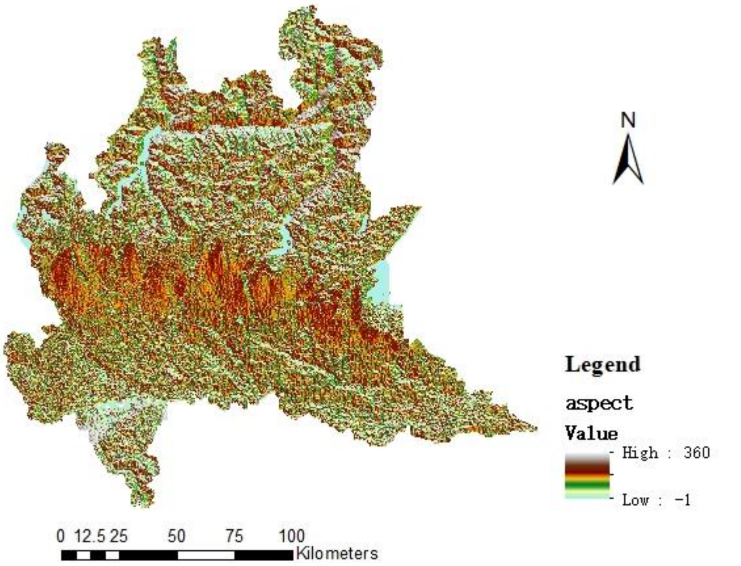
In order to build ground motion risk models, the spatial database is randomly divided into training and validation datasets with the ratio 7:3, respectively (Wang et al., 2022; Youssef and Pourghasemi, 2021). Although this ratio can be adjusted, it shows a better performance for the ML model, and adopted by many researches (Chen et al., 2020; Nguyen et al., 2021; Tien Bui et al., 2016). Therefore, we use this ratio to divide the training and validation data. The flowchart of thaw settlement risk evaluation was shown in Fig 2.4. The thematic map of condition factors and their types, resolutions and sources are shown in Fig 2.5 and Table 2.1.



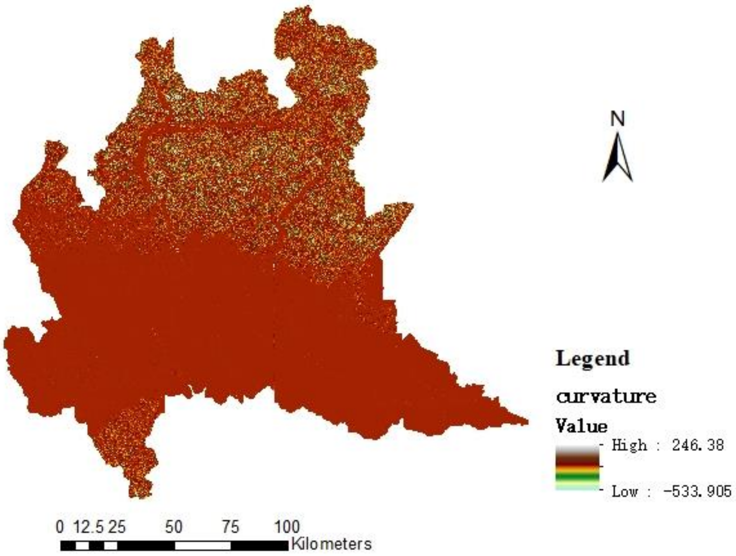
(a) Elevation



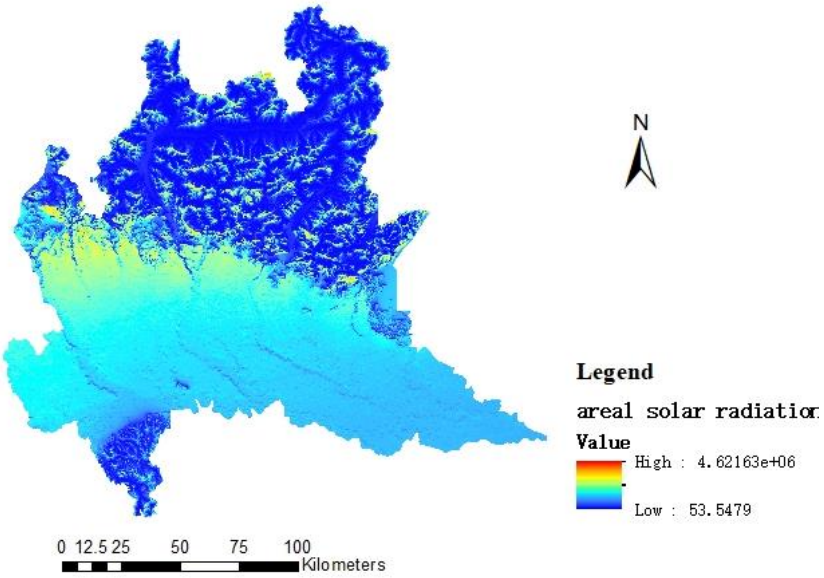
(b) Slope angel



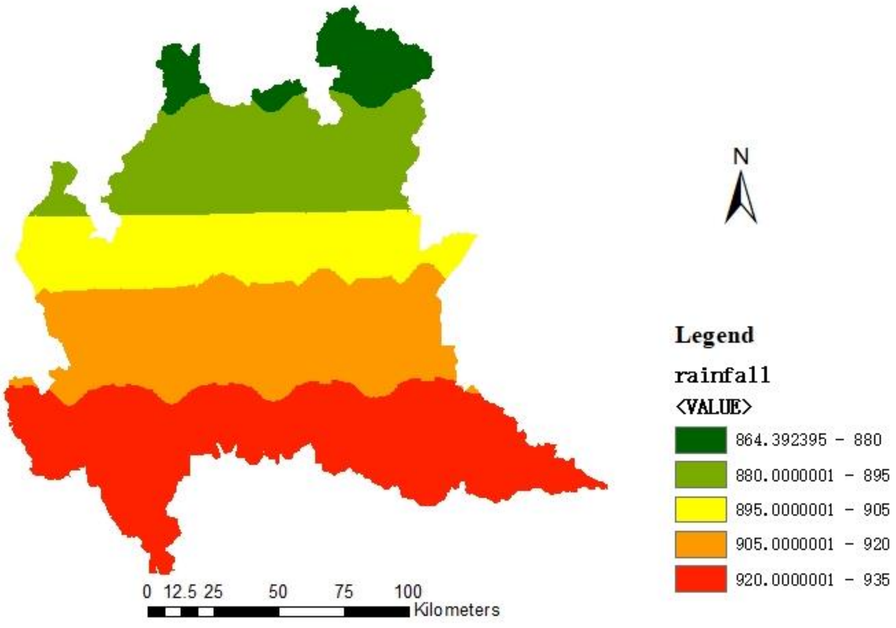
(c) Slope aspect



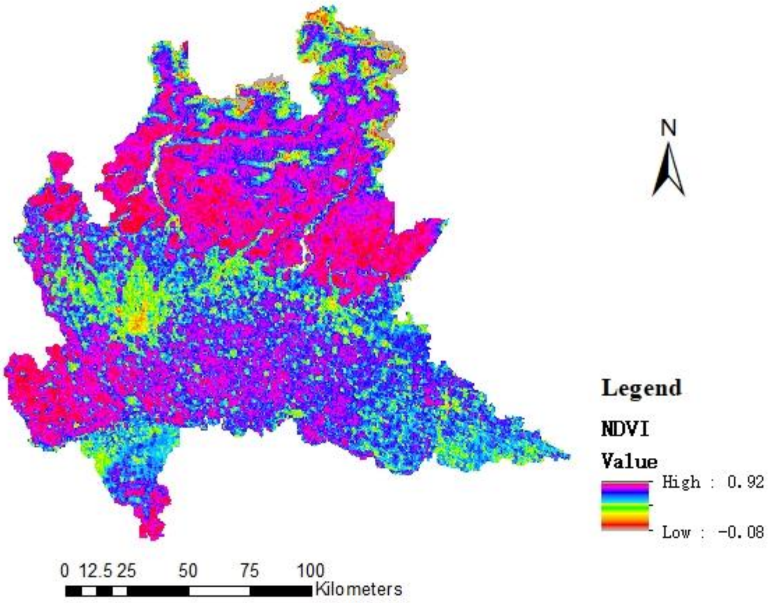
(d) Curvature



(e) Areal solar radiation



(f)Rainfall



(g)NDVI

Figure 2.5: Thaw ground motion conditioning factors.

Factors	Types	Resolutions	Sources
Elevation	Continuous	20 m × 20 m	Geoportale
Slope angel	Continuous	20 m × 20 m	ArcGIS
Slope aspect	Continuous	20 m × 20 m	ArcGIS
Curvature	Continuous	20 m × 20 m	ArcGIS
Areal solar radiation	Continuous	20 m × 20 m	ArcGIS
Rainfall	Continuous	1km× 1km	GISgeography
NDVI	Continuous	300m× 300 m	CGLS

Table 2.1: Ground motion conditioning factors and their types, resolutions and sources.

2.4 Description of condition factors

The main causes of ground motion in Lombardy are rainfall and earthquakes (Luino et al.,2005: Peresan et al.,2009). The mechanism of ground motion is related to the intensity and duration of the rainfall, the terrain conditions and the vegetation cover (Take et al., 2004; Lourenço et al., 2006; An and Zheng, 2012). In this study, factors can be divided into gestation factor and triggering factor. The conditioning factors in the study consist of elevation, slope angle, slope aspect, curvature, and the triggering factors include areal solar radiation, rainfall, and NDVI.

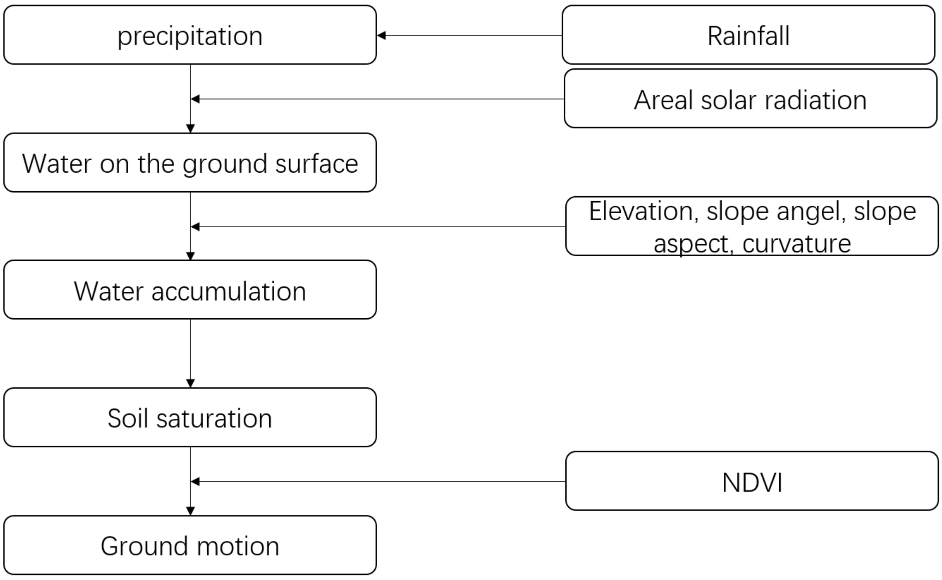


Figure 2.6: Physical process of ground motion caused by rainfall

Elevation plays a crucial role in shaping the distribution of ground motion in the study area. It causes variations in topography and temperature, which in turn affects the thermal stability of permafrost (Qin et al., 2020). Slope angle is closely associated with the hydraulic connection within the slope (Ohlmacher, 2007). The slope aspect determines the distribution of sunny and shady slopes and rainfall, resulting in temperature and vegetation variability on the slope (Beullens et al., 2014; Wu and Chau, 2013). Curvature is frequently utilized to reflect the shape of the slope surface, which profoundly impacts the evolution of the landscape and changes the direction of surface water flow (Li and Wang, 2019; Ohlmacher, 2007). Vegetation coverage significantly affects land subsidence (Rahmati et al., 2019). Solar radiation has a close relationship with the growth and types of vegetation (Naumburg et al., 2005).

2.5 Machine learning models

The machine learning prat is based on the function AutoML in ArcGIS pro. There is a general process or workflow associated with a machine learning (ML) project. A typical ML workflow begins with identifying the business problem and formulating the problem statement or question. This is followed by a series of steps, including: data preparation (or preprocessing), feature engineering, selecting a suitable algorithm and model training, hyperparameter tuning, and model evaluation. This is an iterative process and the optimal model is often only reached after multiple iterations and experiments (Caruana et al.,2004). The work flow is shown in Fig 2.7.

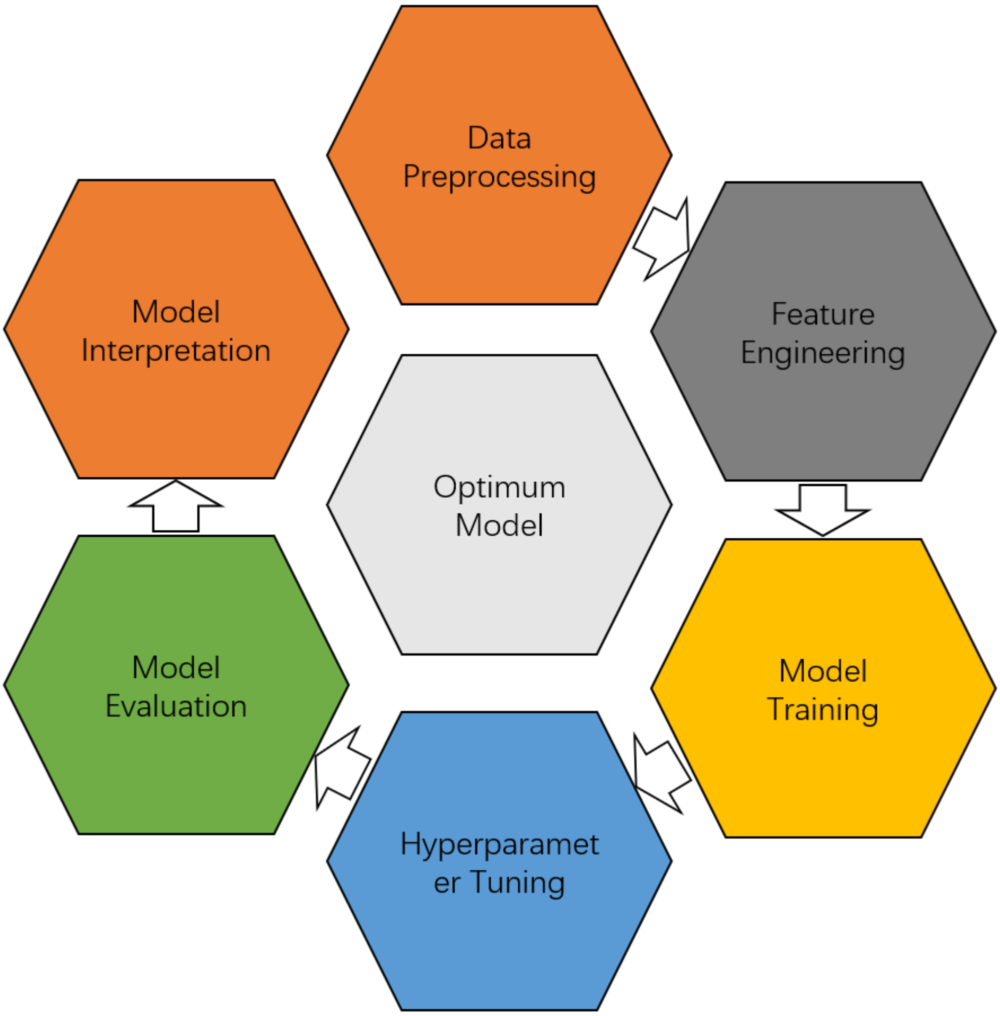


Figure 2.7: AutoML tool workflow

2.5.1 Linear regression

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots \tag{2.1}$$

In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

In Fig 2.8, a linear regression model is described by the regression line y = 153.21 + 900.39x. The model describes the relationship between the dependent variable, Diabetes progression, and the explanatory variable, Serum triglycerides level. A positive correlation is shown. This example demonstrates a linear regression model with two variables. Although it is not possible to visualize models with more than three variables, practically, a model can have any number of variables.

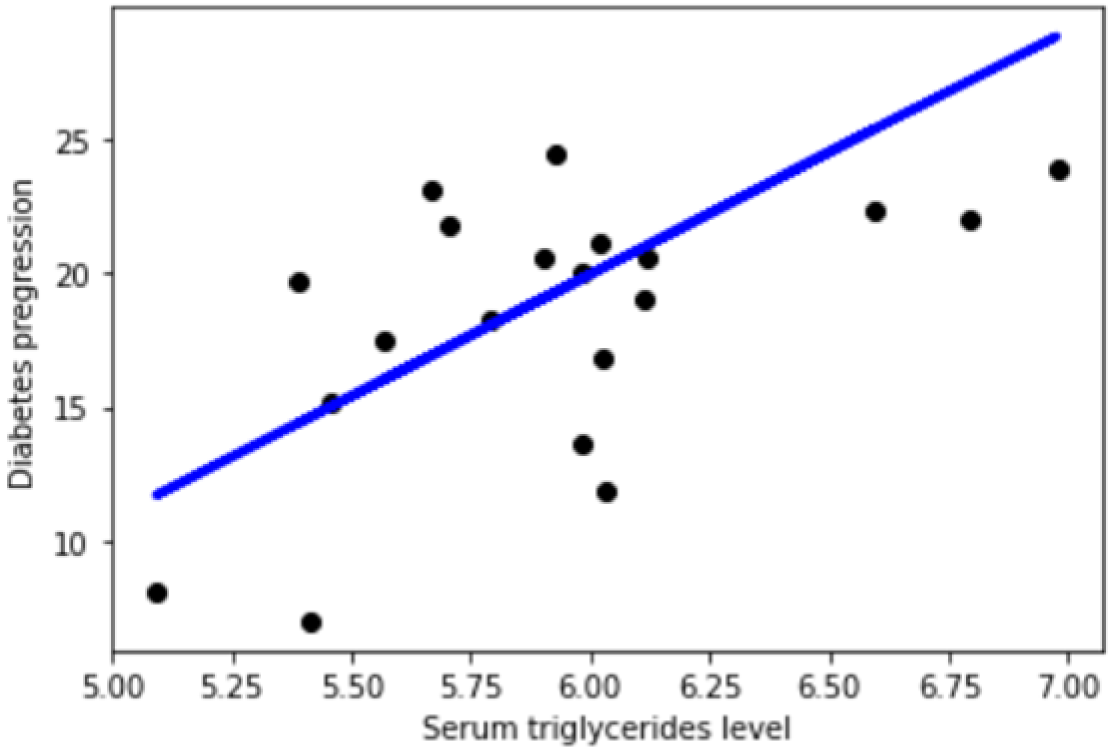


Figure 2.8: A linear regression model description.

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

2.5.2 Decision tree

Decision trees is a type of supervised machine learning algorithm that is used by the Train Using AutoML tool and classifies or regresses the data using true or false answers to certain questions. The resulting structure, when visualized, is in the form of a tree with different types of nodes—root, internal, and leaf. The root node is the starting place for the decision tree, which then branches to internal nodes and leaf nodes. The leaf nodes are the final classification categories or real values. Decision trees are easy to understand and are explainable.

To construct a decision tree, start by specifying a feature that will become the root node. Typically, no single feature can perfectly predict the final classes; this is called impurity. Methods such as Gini, entropy, and information gain are used to measure this impurity and identify how well a feature classifies the given data. The feature with the least impurity is selected as the node at any level. To calculate Gini impurity for a feature with numerical values, first sort the data in ascending order and calculate the averages of the adjoining values. Then, calculate the Gini impurity at each selected average value by arranging the data points based on whether the feature values are less than or greater than the selected value and whether that selection correctly classifies the data. The Gini impurity is then calculated using the equation below, where K is the number of classification categories and p is the proportion of instances of those categories.

$$Gini\ Impurity = 1 - \sum_{i=1}^{k} p_i^2 \qquad (2.2)$$

The weighted average of the Gini impurities for the leaves at each value is calculated. The value with the least impurity is selected for that feature. The process is repeated for different features to select the feature and value that will become the node. This process is iterated at every node at each depth level until all the data is classified. Once the tree is constructed, to make a prediction for a data point, go down the tree using the conditions at each node to arrive at the final value or classification. When using decision trees for regression, the sum of squared residuals or variance is used to measure the impurity instead of Gini. The rest of the method follows similar steps.

2.5.3 Random forest

A decision tree is overly sensitive to training data. In this method, many decision trees are created that are used for prediction. Each tree generates its own prediction and is used as part of a majority vote to make final predictions. The final predictions are not based on a single tree but on the entire forest of decision trees (Fig 2.9). The use of the entire forest helps avoid overfitting the model to the training dataset, as does the use of both a random subset of the training data and a random subset of explanatory variables in each tree that constitutes the forest (Ho et al., 1995).

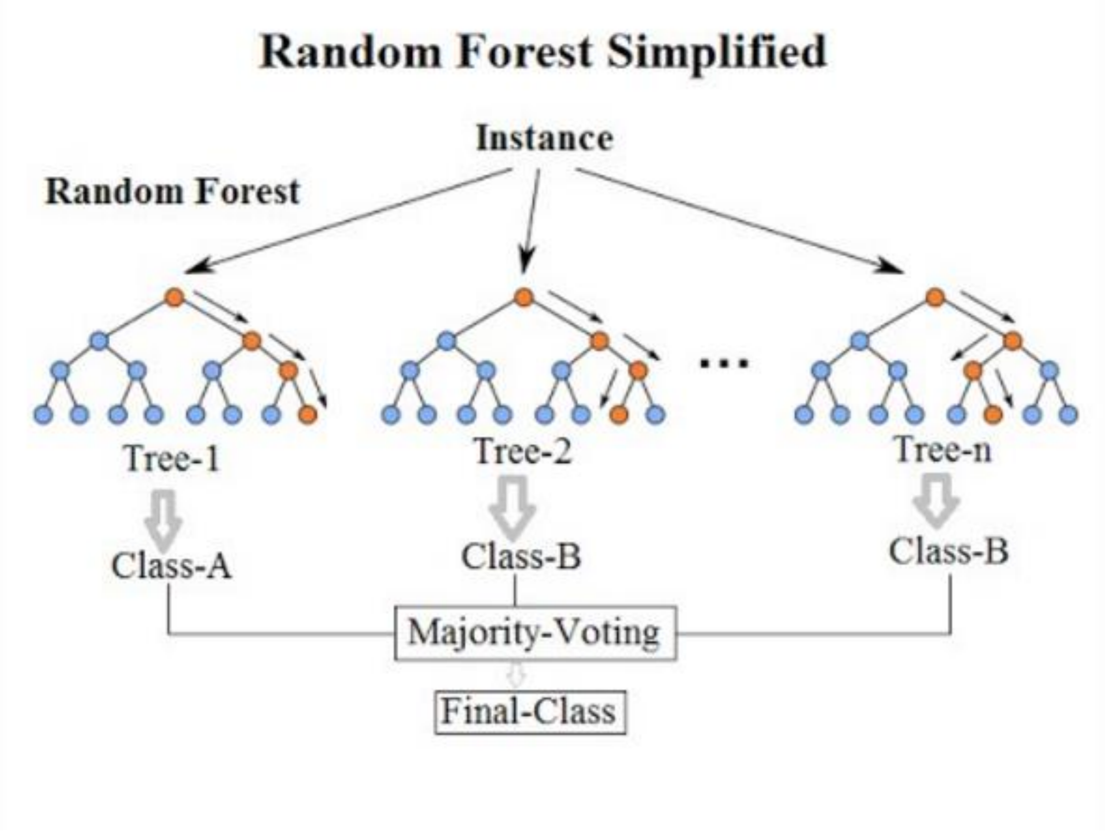


Figure 2.9: Random Forest concept.

Bootstrapping is used to create a random subset of the training data. The subset is the same size as the original training data since the data is selected randomly with repetition. This makes the model less sensitive to the original training data. The random selection of explanatory variables reduces the correlation between trees and causes less variance. This level of variance makes random forest more effective than decision trees. Using bootstrapping and the aggregation of results together is called bagging. To test the accuracy of a tree, the subset of data that is not selected (out-of-bag) is used. The method iterates different settings to find the forest with the least out-of-bag error.

2.5.4 Extra trees

Extra trees (short for extremely randomized trees) is an ensemble supervised machine learning method that uses decision trees and is used by the Train Using AutoML tool. This method is similar to random forests but can be faster.

The extra trees algorithm, like the random forests algorithm, creates many decision trees, but the sampling for each tree is random, without replacement. This creates a dataset for each tree with unique samples. A specific number of features, from the total set of features, are also selected randomly for each tree. The most important and unique characteristic of extra trees is the random selection of a splitting value for a feature. Instead of calculating a locally optimal value using Gini or entropy to split the data, the algorithm randomly selects a split value. This makes the trees diversified and uncorrelated (Geurts et al.,2006).

2.5.5 XGBoost

XGBoost is short for extreme gradient boosting. This method is based on decision trees and improves on other methods such as random forest and gradient boost. It works well with large, complicated datasets by using various optimization methods.

To fit a training dataset using XGBoost, an initial prediction is made. Residuals are computed based on the predicted value and the observed values. A decision tree is created with the residuals using a similarity score for residuals. The similarity of the data in a leaf is calculated, as well as the gain in similarity in the subsequent split. The gains are compared to determine a feature and a threshold for a node. The output value for each leaf is also calculated using the residuals. For classification, the values are typically calculated using the log of odds and probabilities. The output of the tree becomes the new residual for the dataset, which is used to construct another tree. This process is repeated until the residuals stop reducing or for a specified number of times. Each subsequent tree learns from the previous trees and is not assigned equal weight, unlike how Random Forest works.

To use this model for prediction, the output from each tree multiplied by a learning rate is added to the initial prediction to arrive at a final value or classification.

XGBoost uses the following parameters and methods to optimize the algorithm and provide better results and performance:

Regularization—A Regularization parameter (lambda) is used while calculating the similarity scores to reduce the sensitivity to individual data and avoid overfitting.

Pruning—A Tree Complexity Parameter (gamma) is selected to compare the gains. The branch where the gain is smaller than the gamma value is removed. This prevents overfitting by trimming unnecessary branches and reducing the depth of the trees.

Weighted quantile sketch—Instead of testing every possible value as the threshold for splitting the data, only weighted quantiles are used. The selection of quantiles is done using a sketch algorithm, which estimates a distribution on multiple systems over a network.

Parallel Learning—This method divides the data into blocks that can be used in parallel to create the trees or for other computations.

Sparsity-aware split finding—XGBoost handles sparsity in data by trying both directions in a split and finding a default direction by calculating the gain.

Cache-aware Access—This method uses the cache memory of the system to calculate the similarity scores and output values. The cache memory is a faster access memory compared to the main memory and improves the overall performance of the model.

Blocks for Out-of-core Computation—This method works with large datasets that cannot fit in the cache or the main memory and that must be kept in hard drives. The dataset is divided into blocks and compressed. Uncompressing the data in the main memory is faster than reading from the hard drive. Another technique called sharding is used when the data must be kept on multiple hard drives.

2.5.6 LightGBM

LightGBM is a gradient boosting ensemble method that is used by the Train Using AutoML tool and is based on decision trees. As with other decision tree-based methods, LightGBM can be used for both classification and regression. LightGBM is optimized for high performance with distributed systems.

LightGBM creates decision trees that grow leaf wise, which means that given a condition, only a single leaf is split, depending on the gain. Leaf-wise trees can sometimes overfit especially with smaller datasets. Limiting the tree depth can help to avoid overfitting.

LightGBM uses a histogram-based method in which data is bucketed into bins using a histogram of the distribution. The bins, instead of each data point, are used to iterate, calculate the gain, and split the data. This method can be optimized for a sparse dataset as well. Another characteristic of LightGBM is exclusive feature bundling in which the algorithm combines exclusive features to reduce dimensionality, making it faster and more efficient.

Gradient-based One Side Sampling (GOSS) is used for sampling the dataset in LightGBM. GOSS weights data points with larger gradients higher while calculating the gain. In this method, instances that have not been used well for training contribute more. Data points with smaller gradients are randomly removed and some are retained to maintain accuracy. This method is typically better than random sampling given the same sampling rate.

3. Results

3.1 Multicollinearity analysis

In the context of studying ground motion, multicollinearity analysis serves as a crucial step in data preprocessing. When a strong relationship exists among the conditioning factors, it is advised to avoid using Decision Tree and Extra Tree models. To assess the relationship between the factors, this study employs the use of tolerance (TOL) and variance inflation factor (VIF), which are reciprocals and are commonly used in this type of analysis. A TOL value greater than 0.1 typically indicates independence of the factor under examination from the other factors (Chen et al.,2019). The results of the multicollinearity analysis, presented in Table 3.1, demonstrate that the selected factors are appropriate and that the Decision Tree and Extra Tree models are suitable for this study.

$$TOL = \frac{1}{VIF} \tag{3.1}$$

$$VIF = \frac{1}{1-R_J^2} \tag{3.2}$$

where R_J^2 is the determination coefficient for regression analysis of other conditioning factors.

Factors	highway		railway	
	TOL	VIF	TOL	VIF
Elevation	0.281	3.556	0.272	3.678
Slope	0.645	1.551	0.311	3.219
Aspect	0.994	1.006	0.993	1.008
Curvature	0.983	1.017	0.997	1.003
Areal solar radiation	0.651	1.537	0.744	1.344
Rainfall	0.293	3.417	0.570	1.754
NDVI	0.978	1.023	0.899	1.113

Table 3.1: Multicollinearity analysis results.

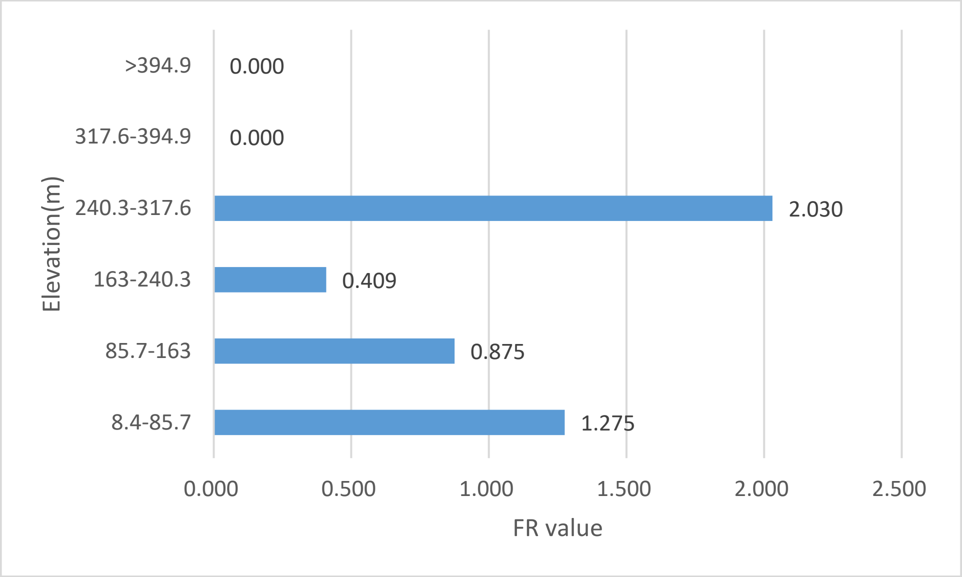
3.2 Correlation analysis

The study employed the frequency ratio (FR) method to investigate the relationship between conditioning factors and ground motion. The FR method involves reclassifying the conditioning factors and determining the intervals that promote the occurrence of ground motion. FR values less than 1 indicate that the corresponding conditioning factor is not conducive to ground motion, while values equal to 1 indicate a critical relationship between the factors. FR values greater than 1 imply that the conditioning factor is favorable to the occurrence of ground motion (Aditian et al., 2018).

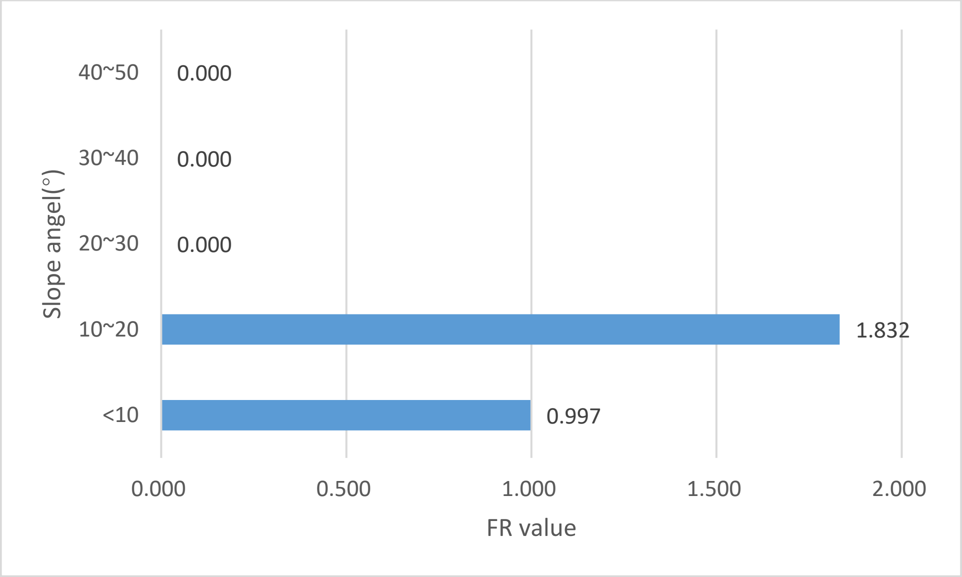
$$FR = \frac{S'/S}{C'/C} \tag{3.3}$$

where S' is the number of ground motion in a factor's class, S is the total number of ground motion, C' is the number of pixels in a factor's class and C is the total number of pixels in the study area.

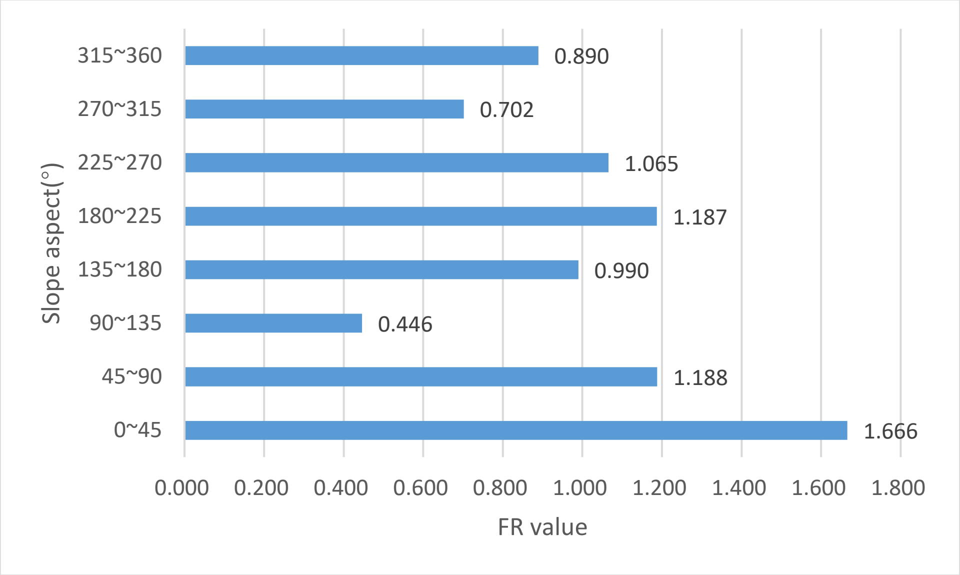
In the highway study, the ground motion occurs in the elevation range of 240.3m to 317.6m with an FR value of 2.030 (Fig 3.1a). The area with a slope lower from 10 to 20 is closely related to ground motion. Ground motion is concentrated in the area with a curvature from -1 to 1 (Fig 3.1e), but it do not show the tendency to curvature. This may be because highways are generally located in the plains and the aspect also does not show an obvious contribution to ground motion (Fig 3.1b and Fig 3.1c). Higher areal solar radiation is favorable to the occurrence of ground motion, with an FR value of 1.527 in the range of 0.6 to 0.8 (Fig 3.1d). Areas with less precipitation are more prone to ground motion. NDVI in the range of 0.2 to 0.4 shows the highest FR value of 1.372 (Fig 3.1g). This indicates that ground motions are mainly concentrated in areas with low vegetation coverage. As for rainfall, the higher value will promote the ground motion (Fig 3.1f).



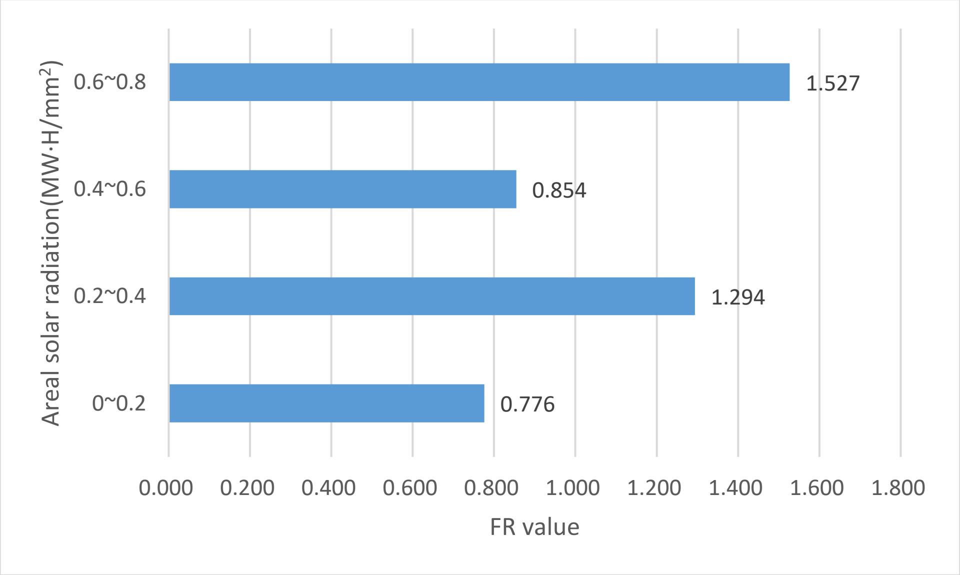
(a) Elevation



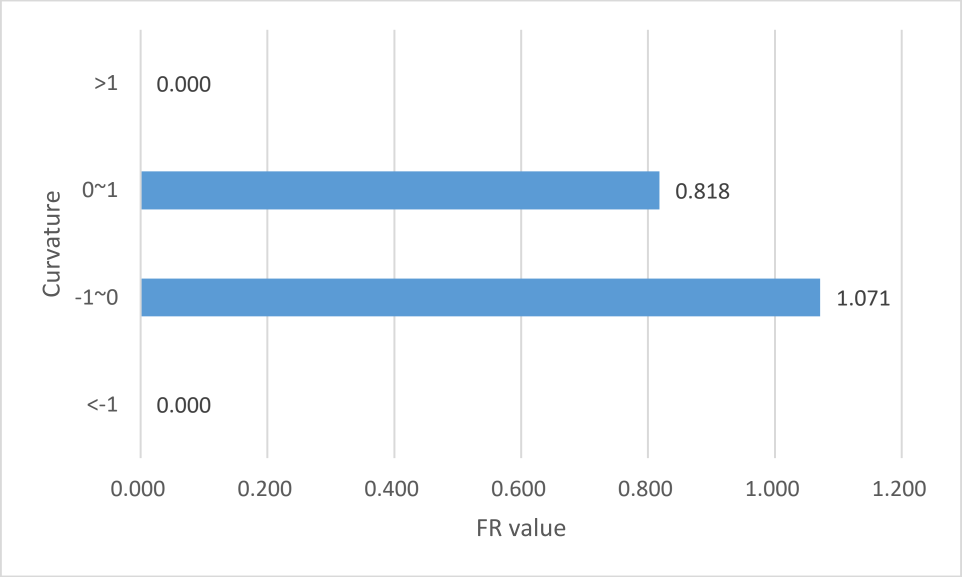
(b) Slope angle



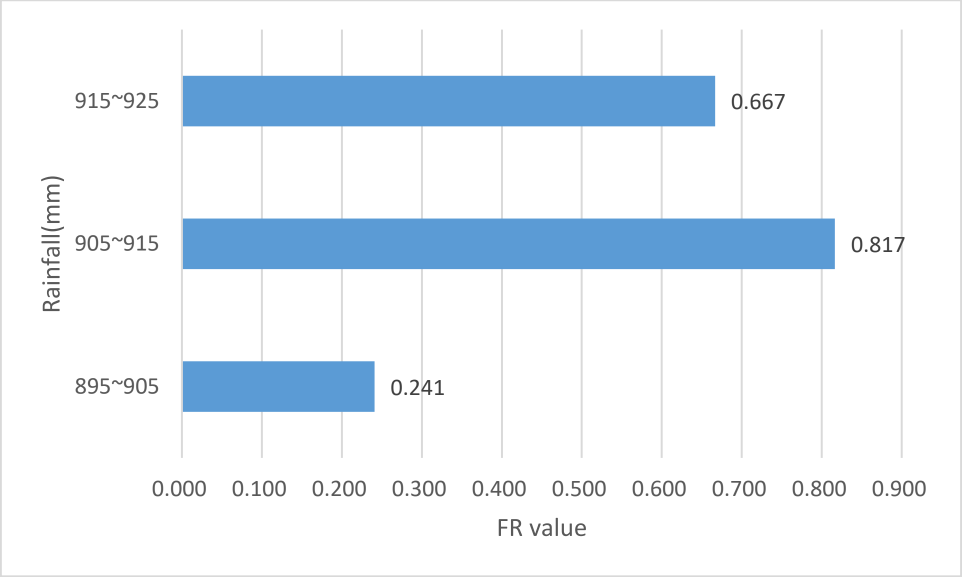
(c) Slope aspect



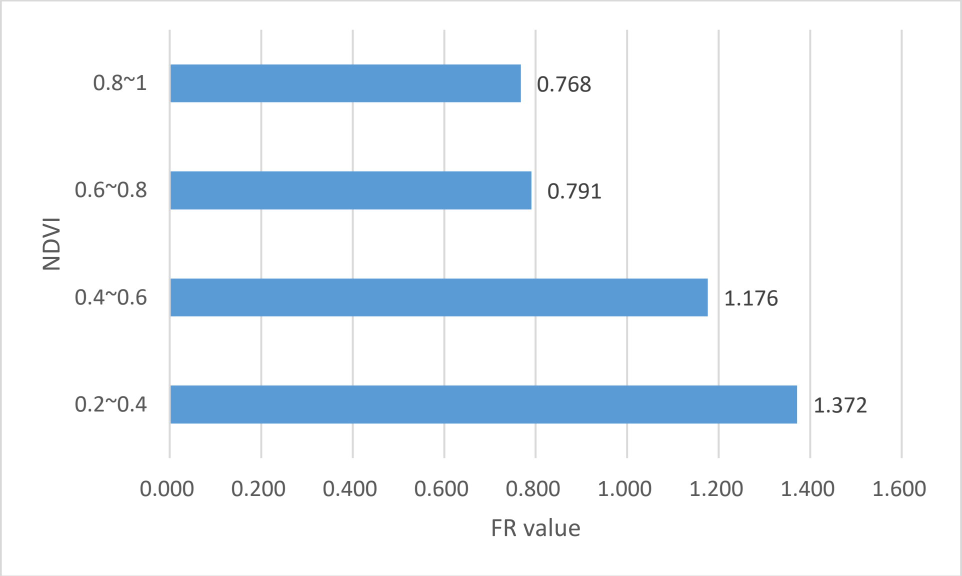
(d) Areal solar radiation



(e) Curvature



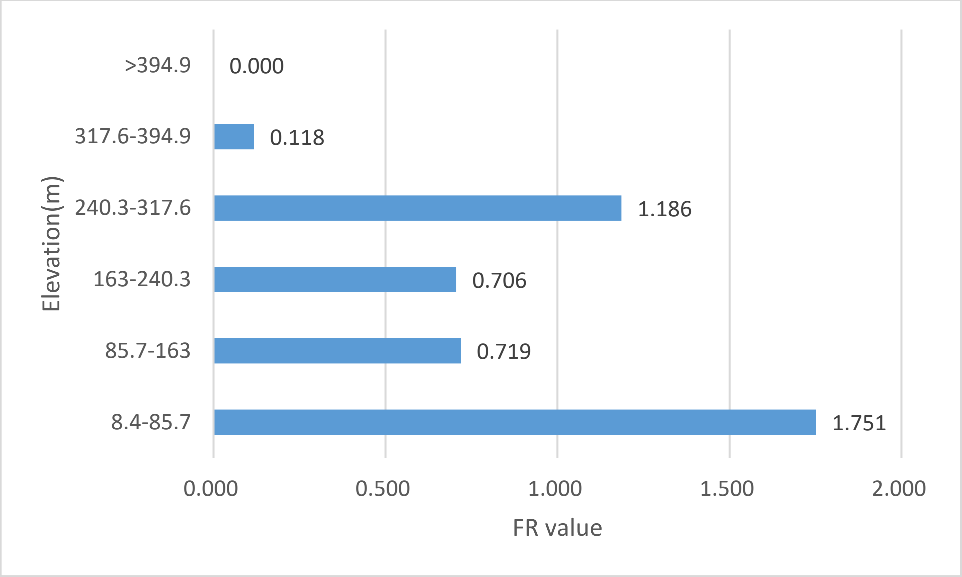
(f) Rainfall



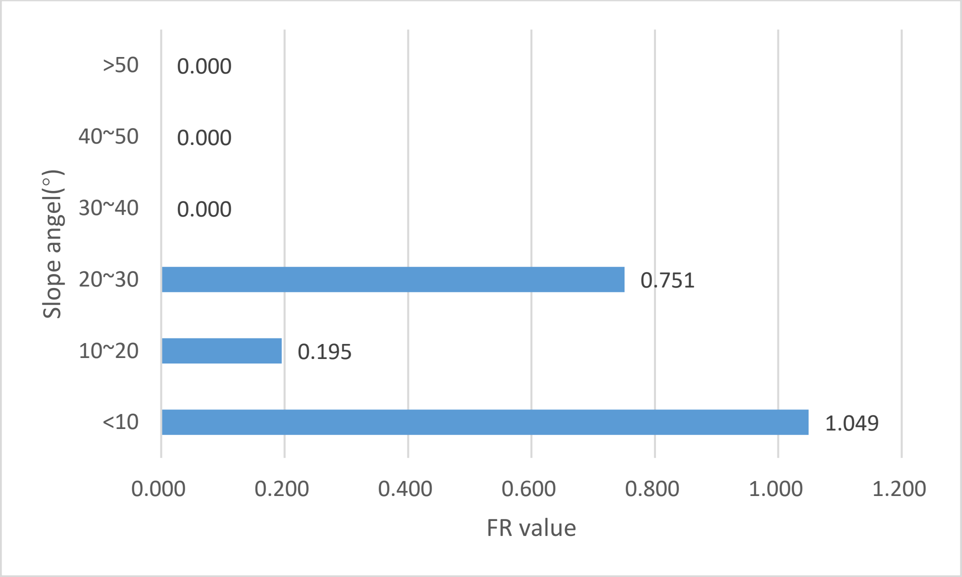
(g) NDVI

Figure 3.1: Correlation analysis of highway

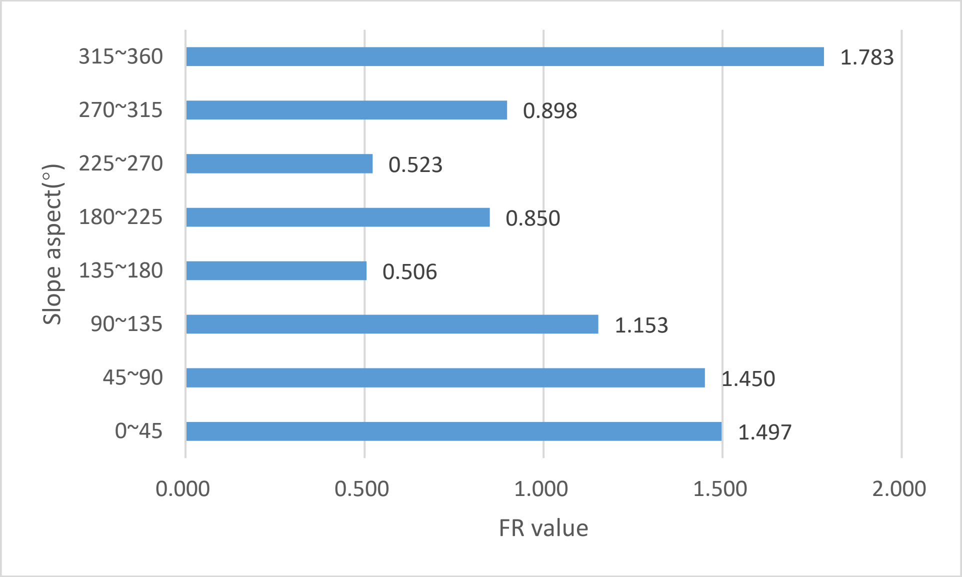
In the railway study, ground motion easily occurs at elevations below 317.6m (Fig 3.1a), and areas with a slope angel less than 10 degrees are more likely to experience ground motion (Fig 3.1b). Different with highway research, the higher curvature of the land is also associated with ground motion about railway (Fig 3.1e). The aspect does not seem to have a significant effect on the occurrence of ground motion(Fig 3.1c). However, areal solar radiation is closely associated with ground motion in the range of 0.2 to 0.4 (Fig 3.1d). Areas with higher levels of precipitation are also more prone to ground motion (Fig 3.1f). Because areas with a higher NDVI value, particularly in the range of 0.8 to 1.0, tend to have higher frequency ratio values and are more likely to experience ground motion (Fig 3.1g). This suggests that ground motion tends to occur more frequently in areas with less vegetation coverage.



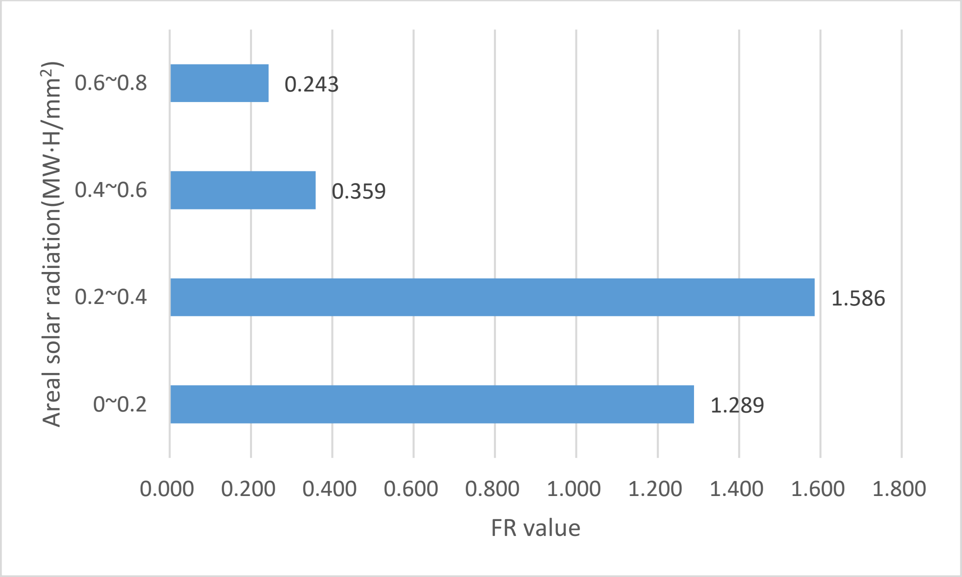
(a) Elevation



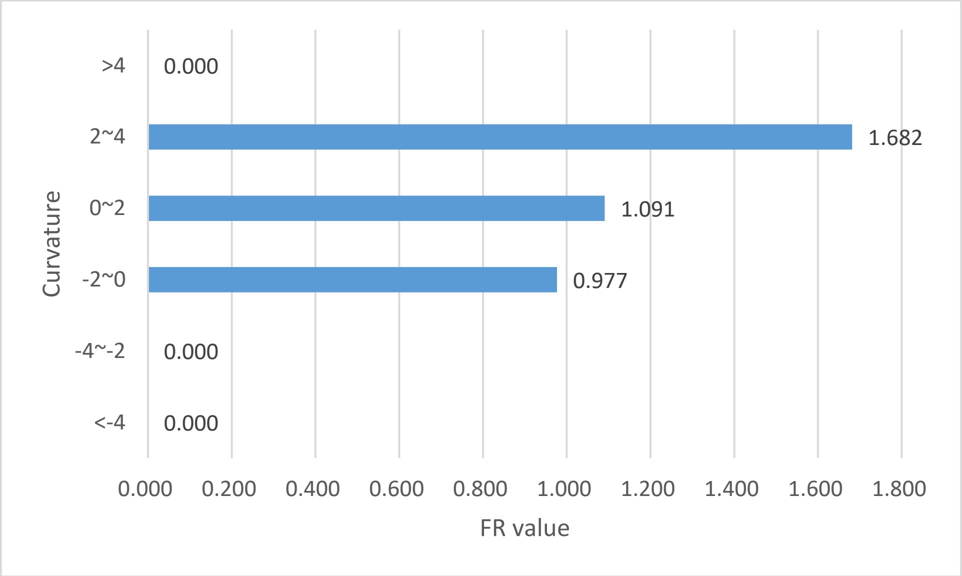
(b) Slope angle



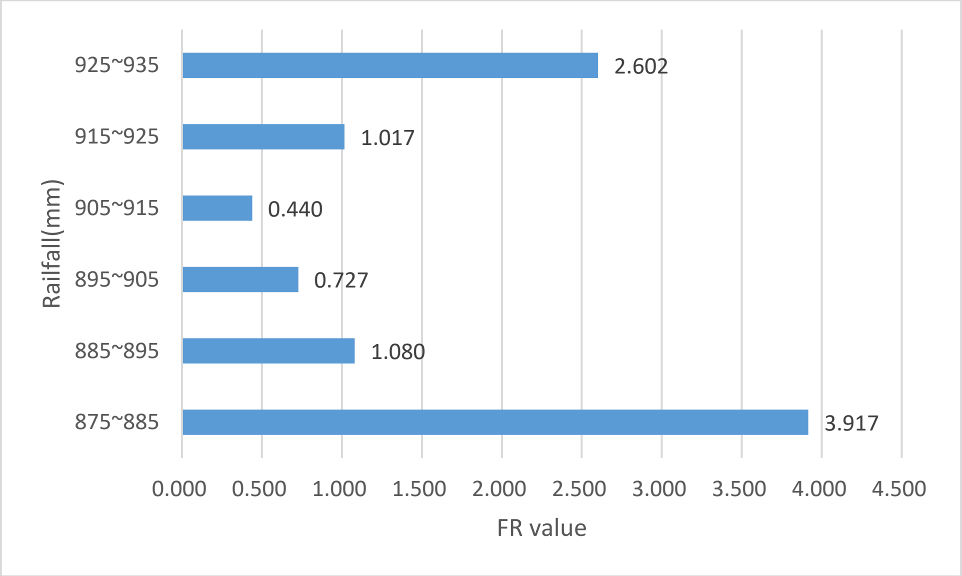
(c) Slope aspect



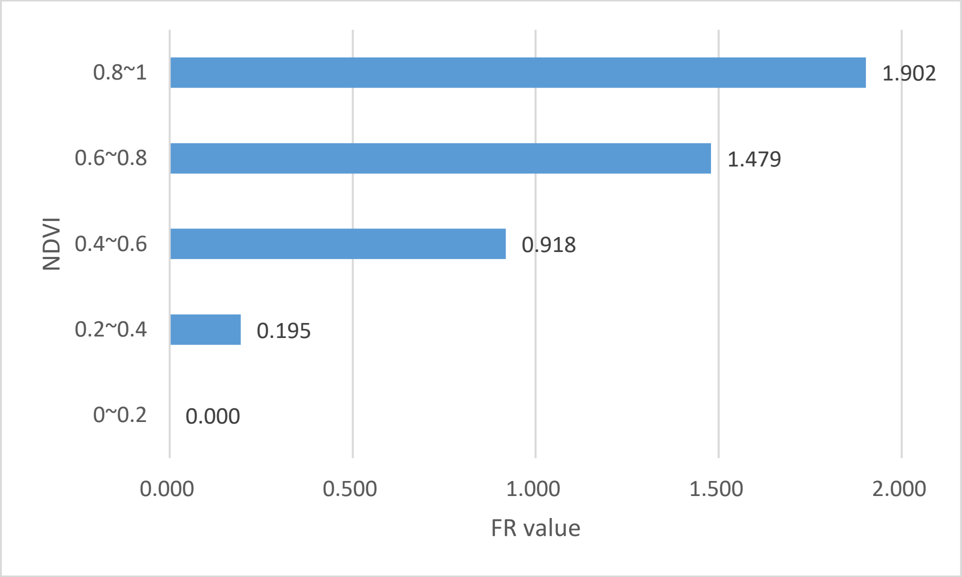
(d) Areal solar radiation



(e) Curvature



(f) Rainfall

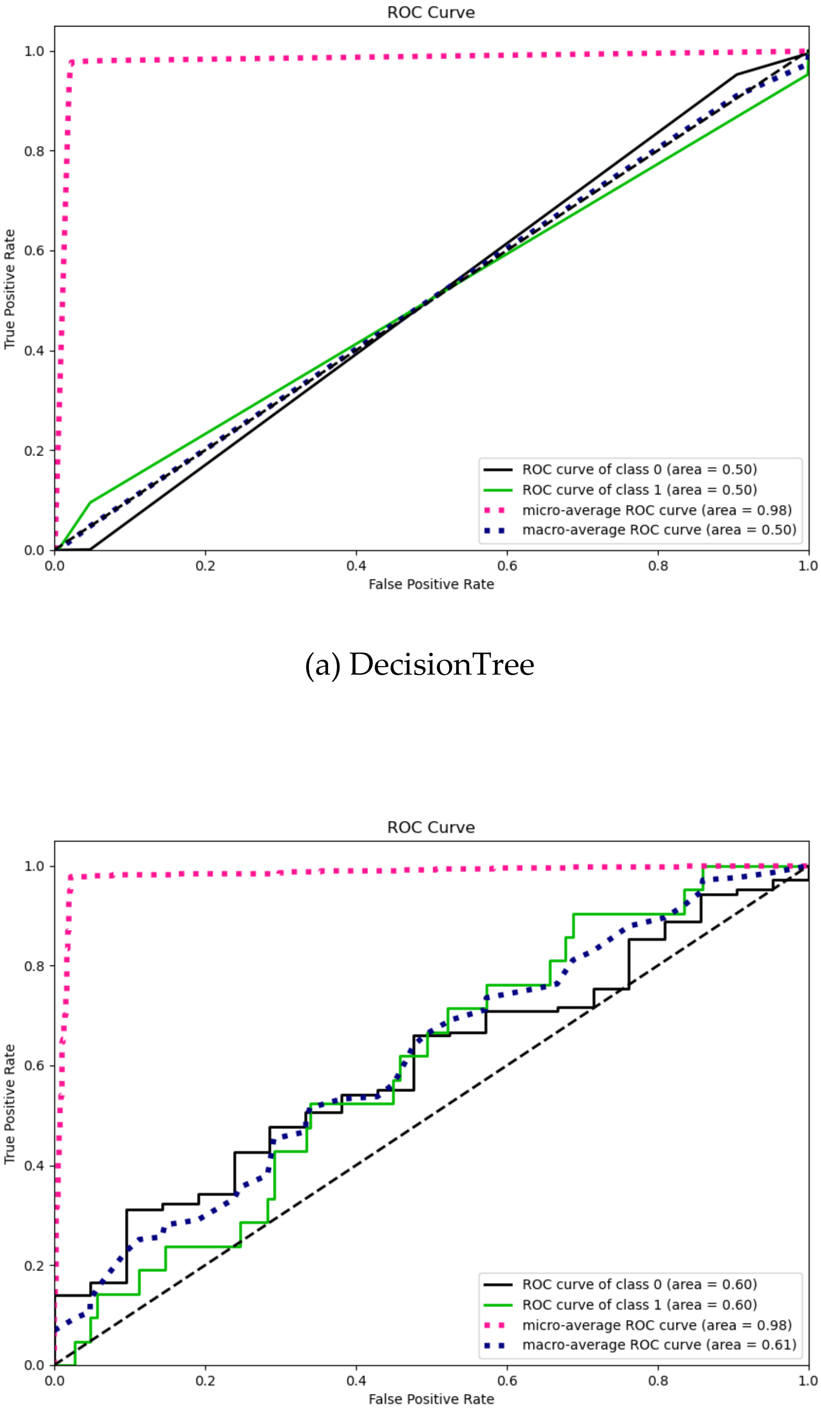


(g) NDVI

Figure 3.2: Correlation analysis of railway

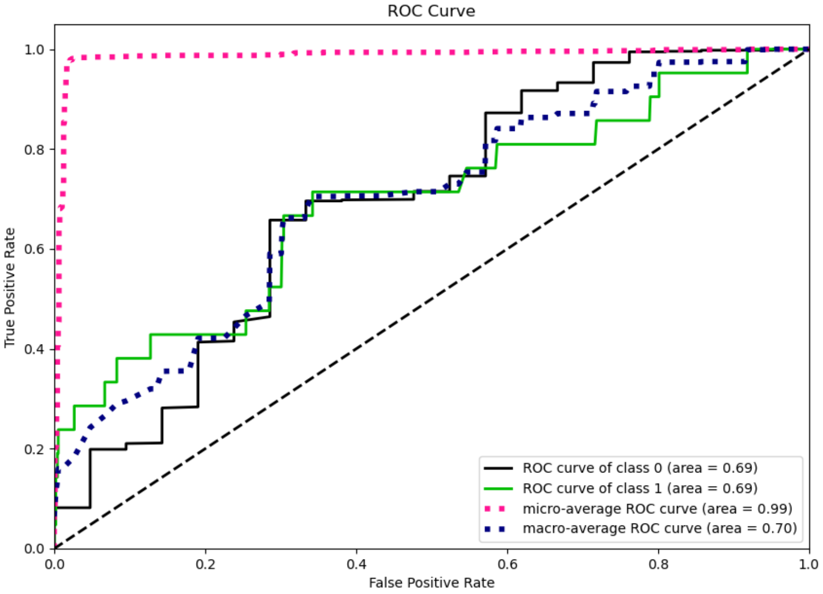
3.3 Evaluation of machine learning models

In machine learning, the ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information for us, called AUC. AUC stands for "Area under the ROC Curve." In general, an AUC of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding. Theoretically, the ROC curve and AUC based on validation dataset is known as the predictive rate curve (Hong et al., 2015; Tacconi Stefanelli et al., 2020).

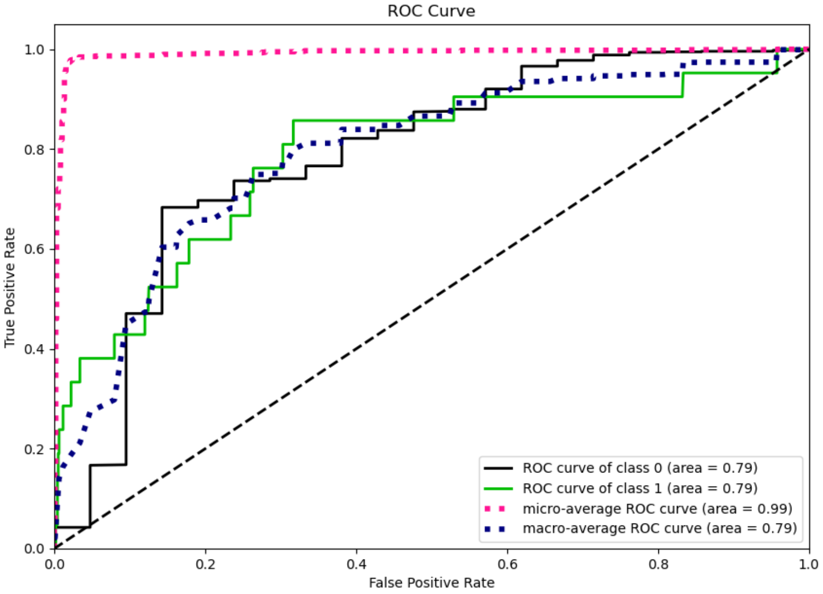


(a) DecisionTree

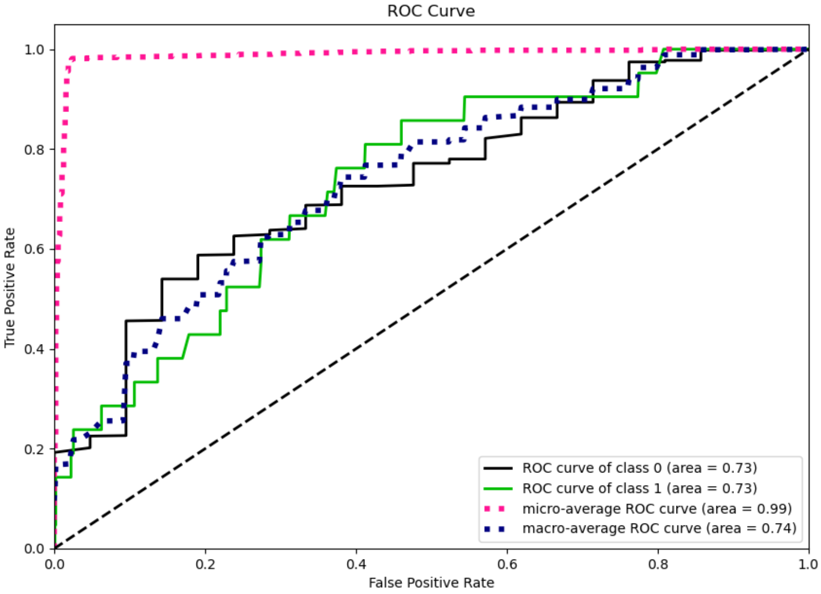
(b) LinearRegression



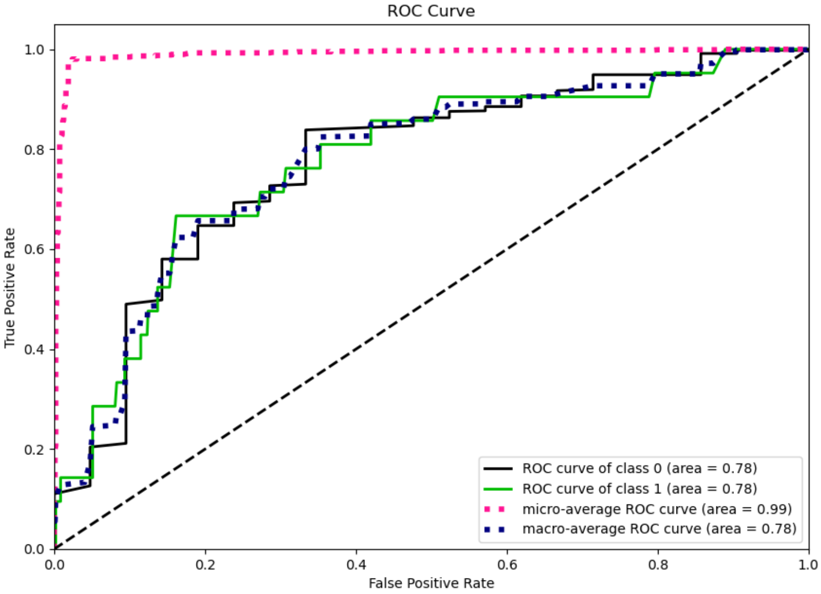
(c) LightGBM



(d) Xgboost



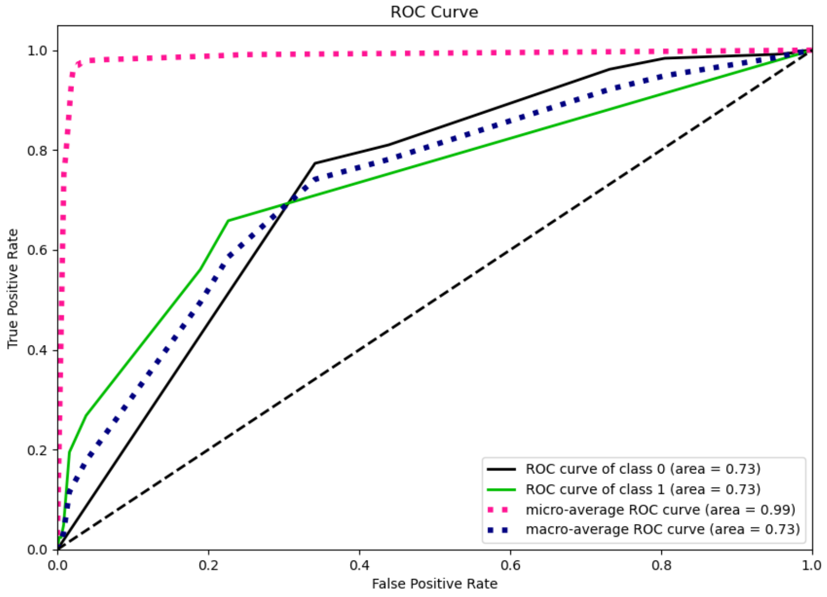
(e) RandomForest



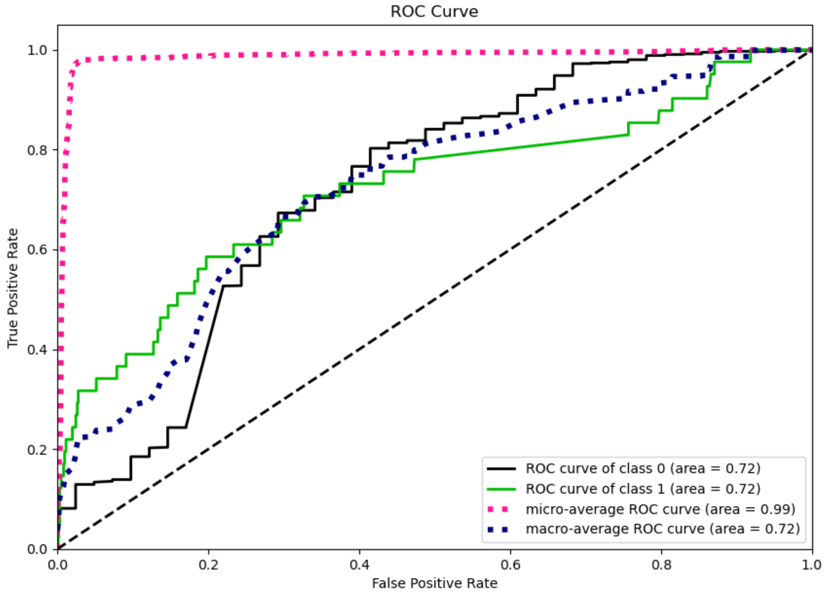
(f) ExtraTrees

Figure 3.3: ROC curves of highway study

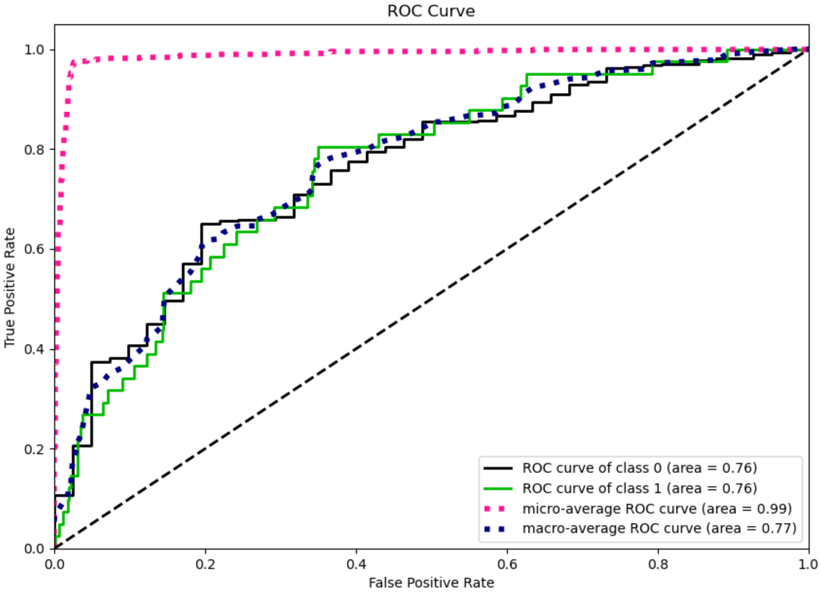
In highway study, the ROC of DecisionTree and LinearRegression do not show an obvious curve, which means the performance of these two models are not suitable for this research (Fig 3.3). Similarly, in the railway study, the ROC curve also shows this condition (Fig 3.4).



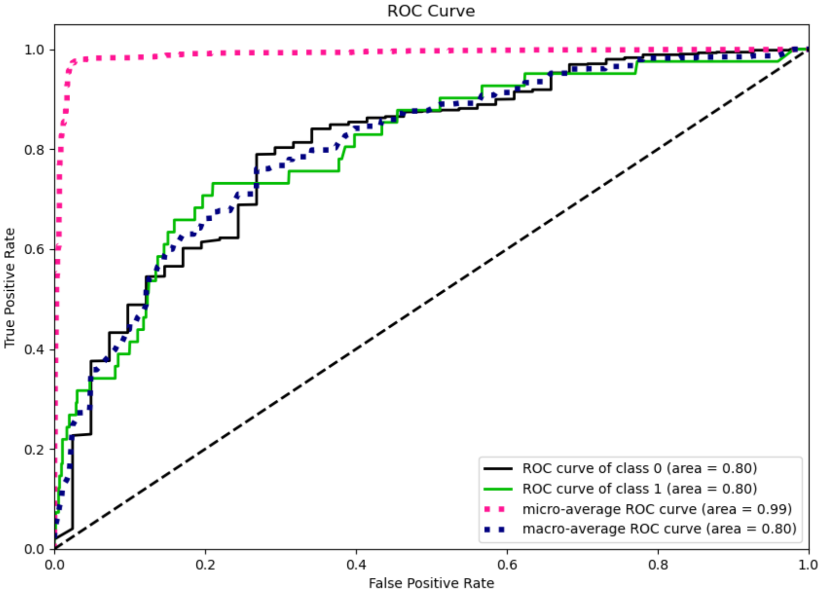
(a) DecisionTree



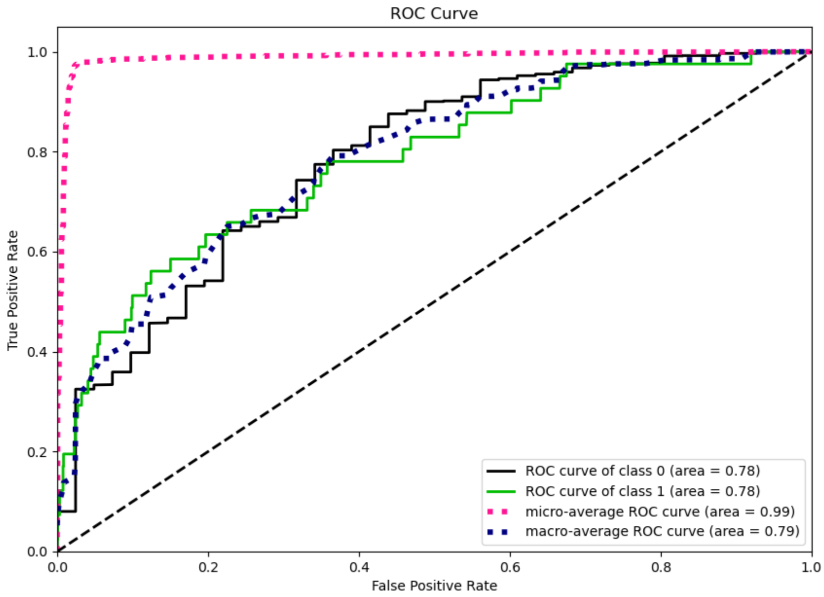
(b) LinearRegression



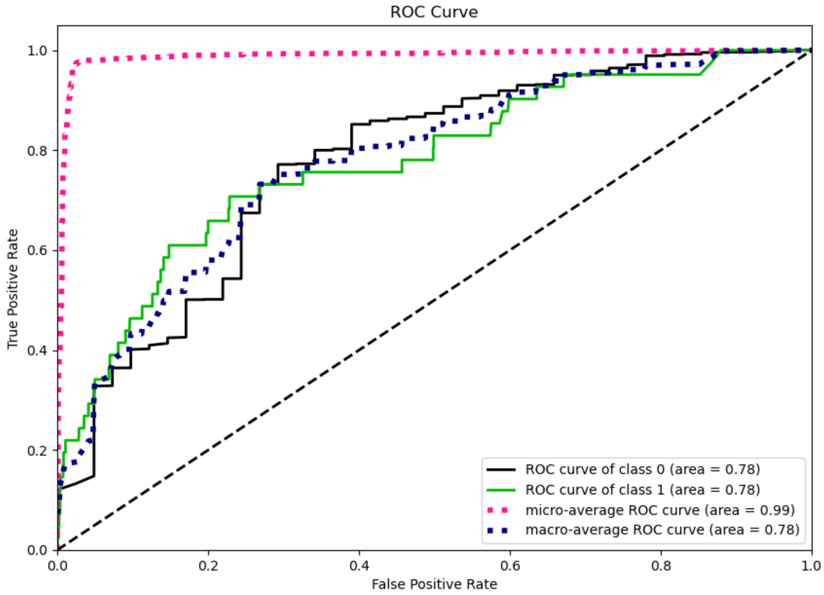
(c) LightGBM



(d) Xgboost



(e) RandomForest



(f) ExtraTrees

Figure 3.4: ROC curves of railway study

Models	highway	railway
1DecisionTree	0.50	0.73
2LinearRegression	0.60	0.76
3LightGBM	0.70	0.72
4Xgboost	0.78	0.80
5RandomForest	0.77	0.78
6ExtraTrees	0.73	0.78

Table 3.2: AUC value of different machine learning models

Based on the ROC curves and AUC value in Fig 3.2, it shows that Xgboost, and RandomForest exhibit a higher AUC value compared to the other models. In the highway study, the AUC value of LightGBM, Xgboost, RandomForest and ExtraTrees are found to be over 0.7, which indicates that these models perform well in terms of accuracy. The AUC value of Xgboost for these models exceeded 0.8 in the railway study, further emphasizing their suitability for this task. In contrast, the AUC values of other models were found to be less than 0.8 in the railway study. In terms of performance comparison between the six models, it was observed that RandomForest and Xgboost exhibit better results. The performance of RandomForest and ExtraTrees are slightly inferior compared to the other two models. This implies that these two models are more suitable for ground motion prediction.

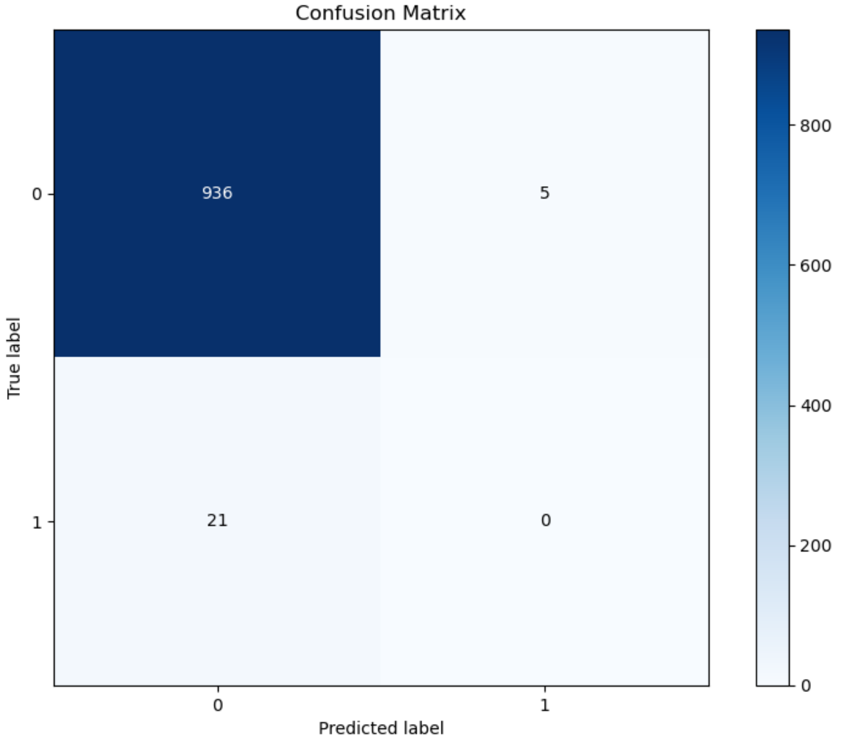
A confusion matrix is a table that is used to evaluate the performance of a machine learning or classification model by comparing the predicted values of the model with the actual values. It is a matrix of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) that are calculated based on the classification results.

In a binary classification problem, the confusion matrix has two rows and two columns, as follows in Table 3.3:

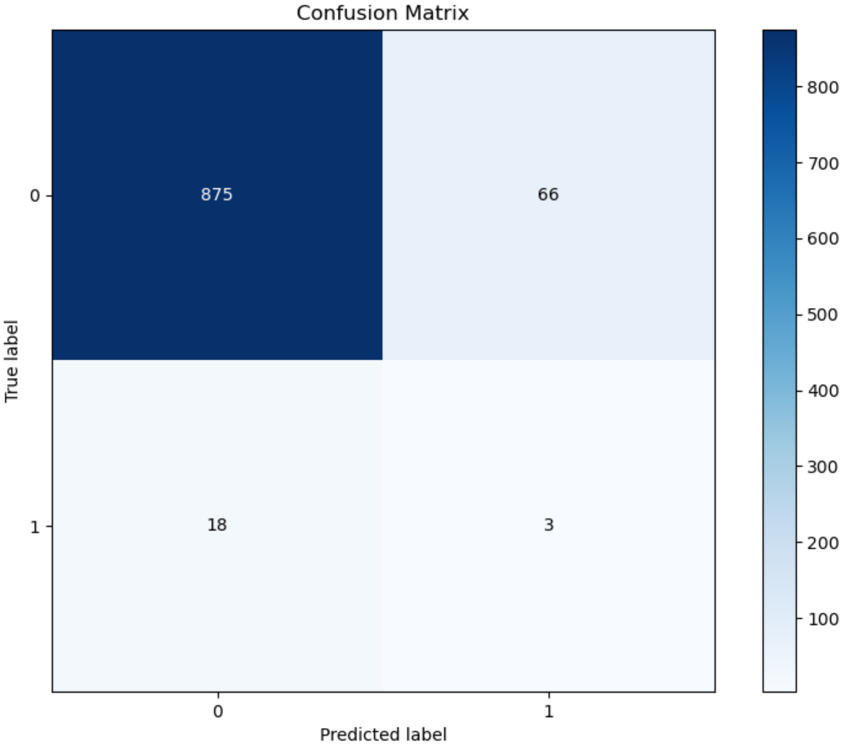
	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3.3: Element in confusion matrix

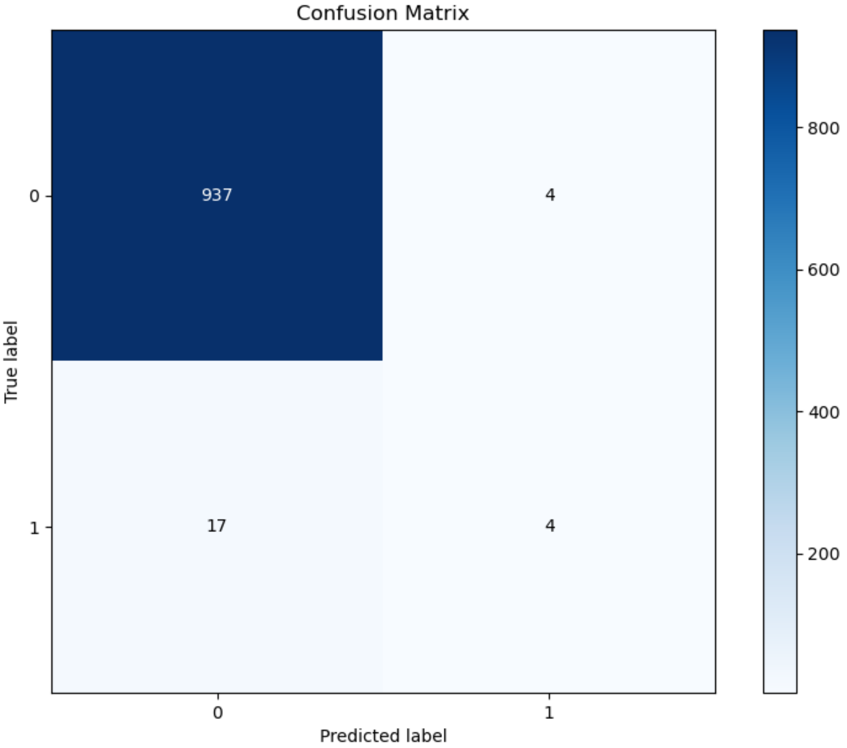
By analyzing the values in the confusion matrix, we can calculate various performance metrics of the model, such as accuracy, precision, recall, F1 score, and others. These metrics provide insight into how well the model is performing and can help in fine-tuning the model to achieve better results.



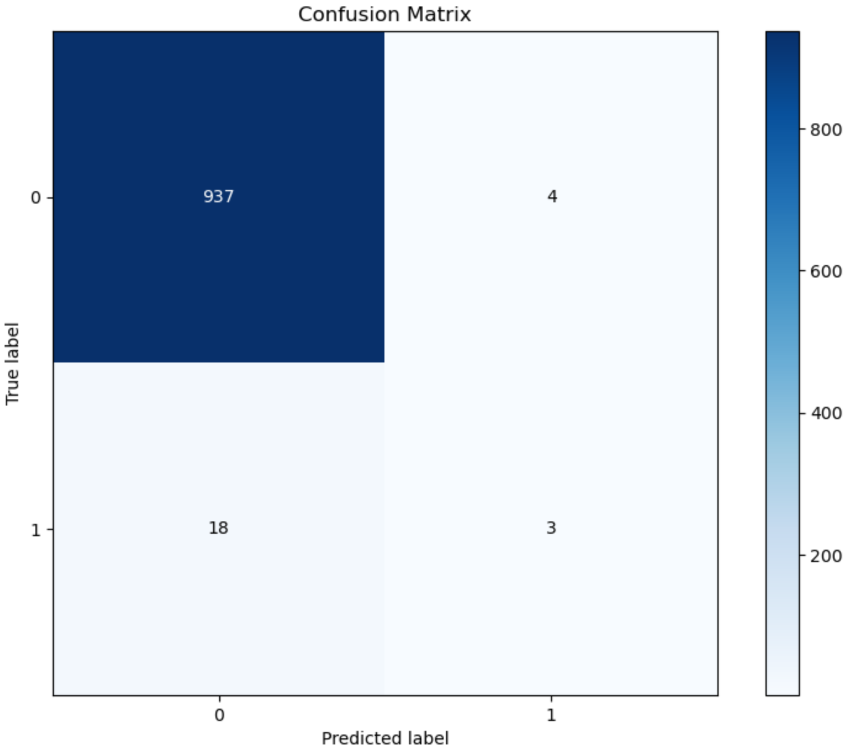
(a) DecisionTree



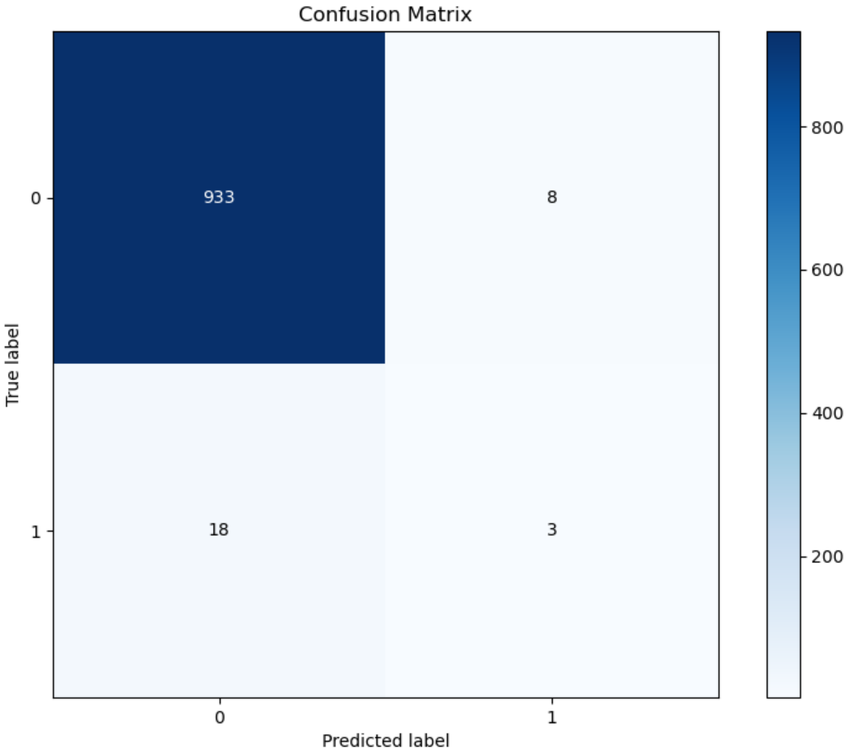
(b) LinearRegression



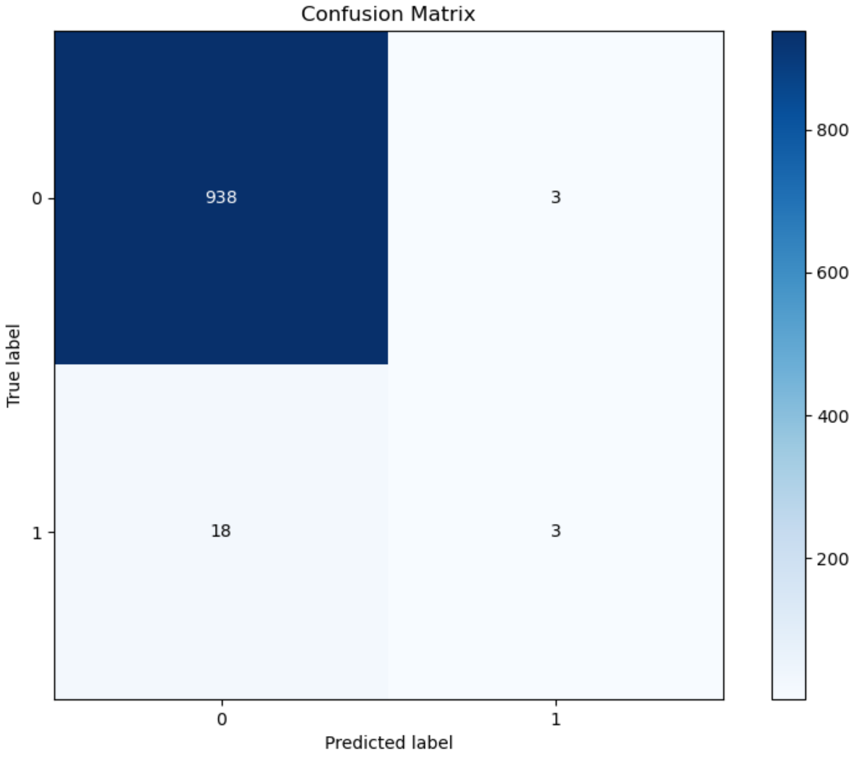
(c) LightGBM



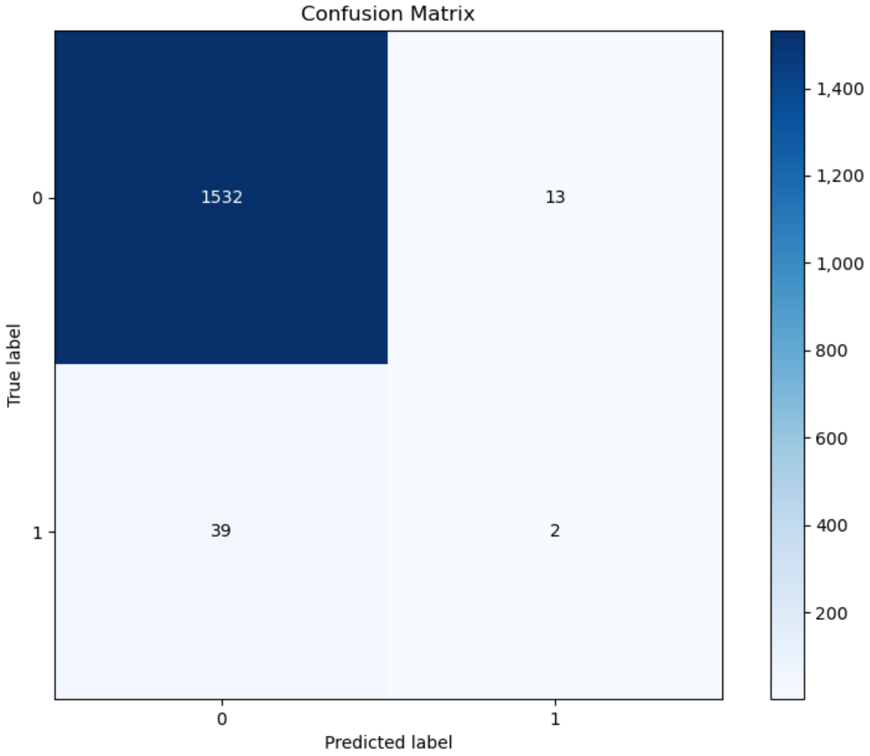
(d) Xgboost



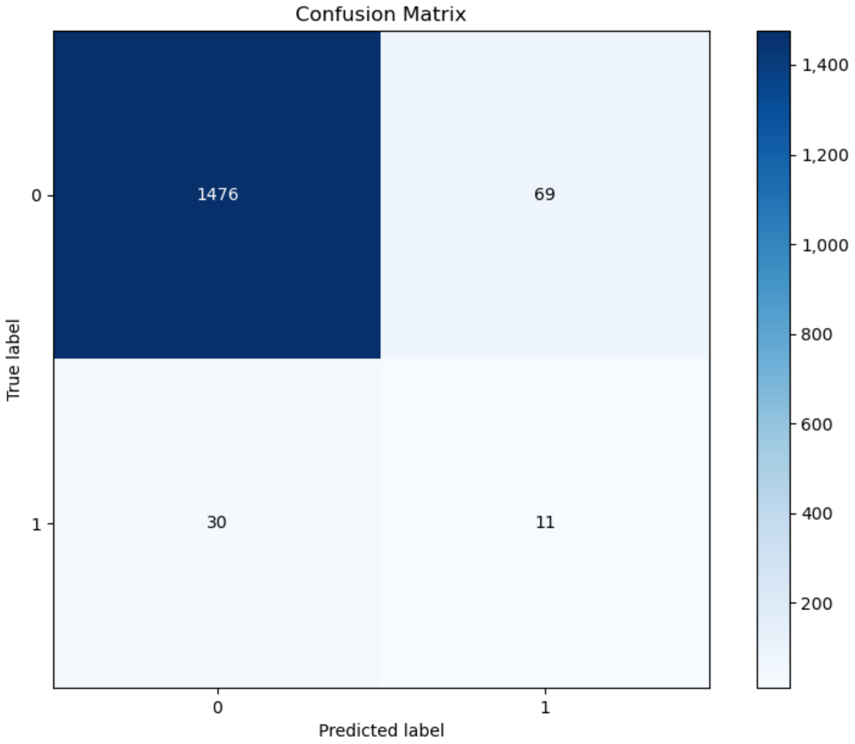
(e) RandomForest



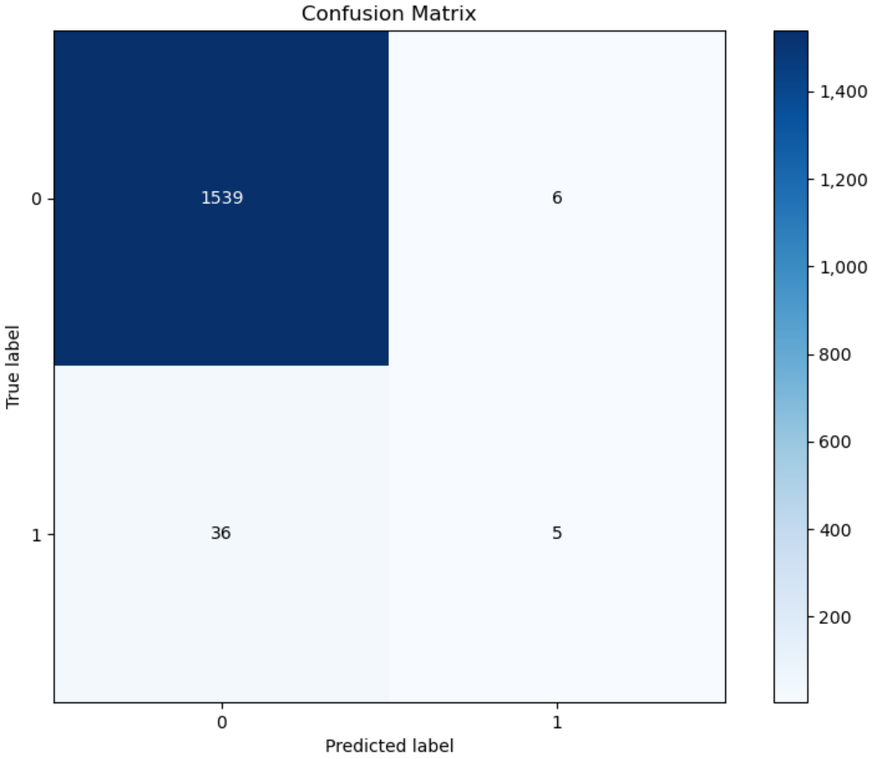
(f) ExtraTrees



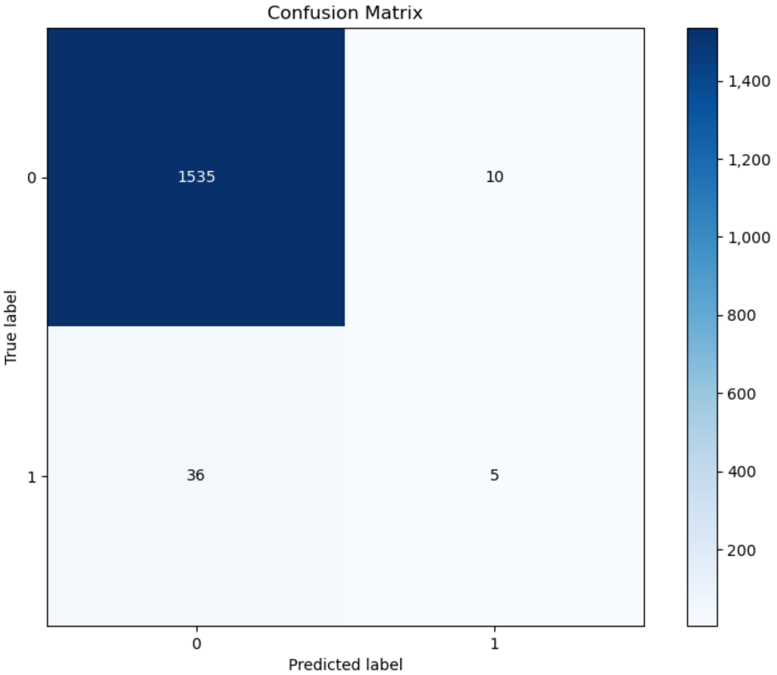
(g) DecisionTree



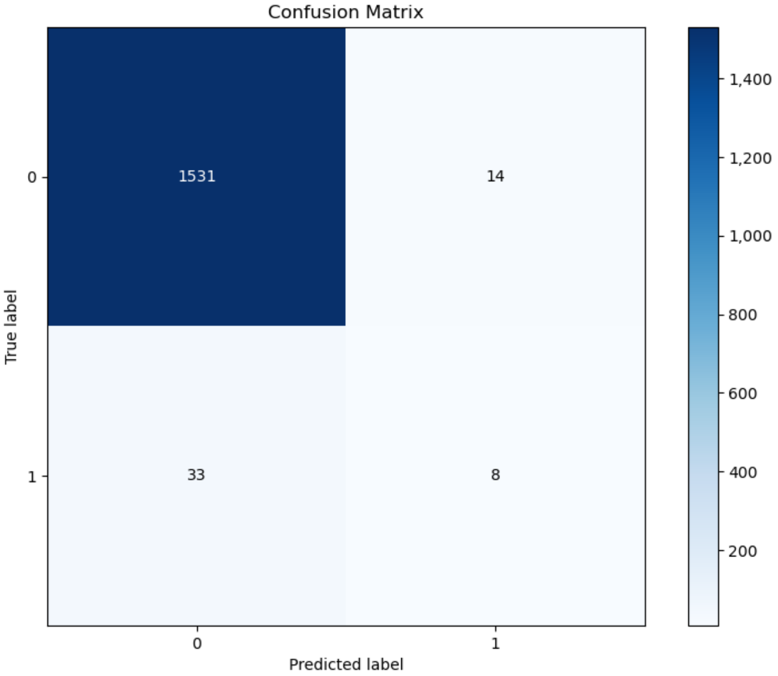
(h) LinearRegression



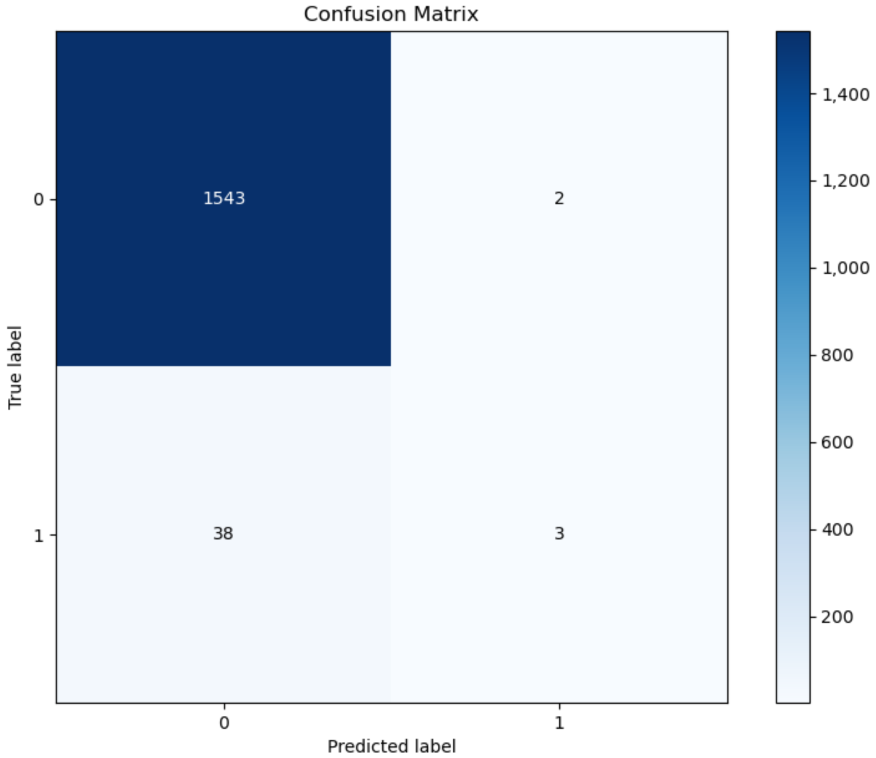
(i) LightGBM



(j) Xgboost



(k) RandomForest

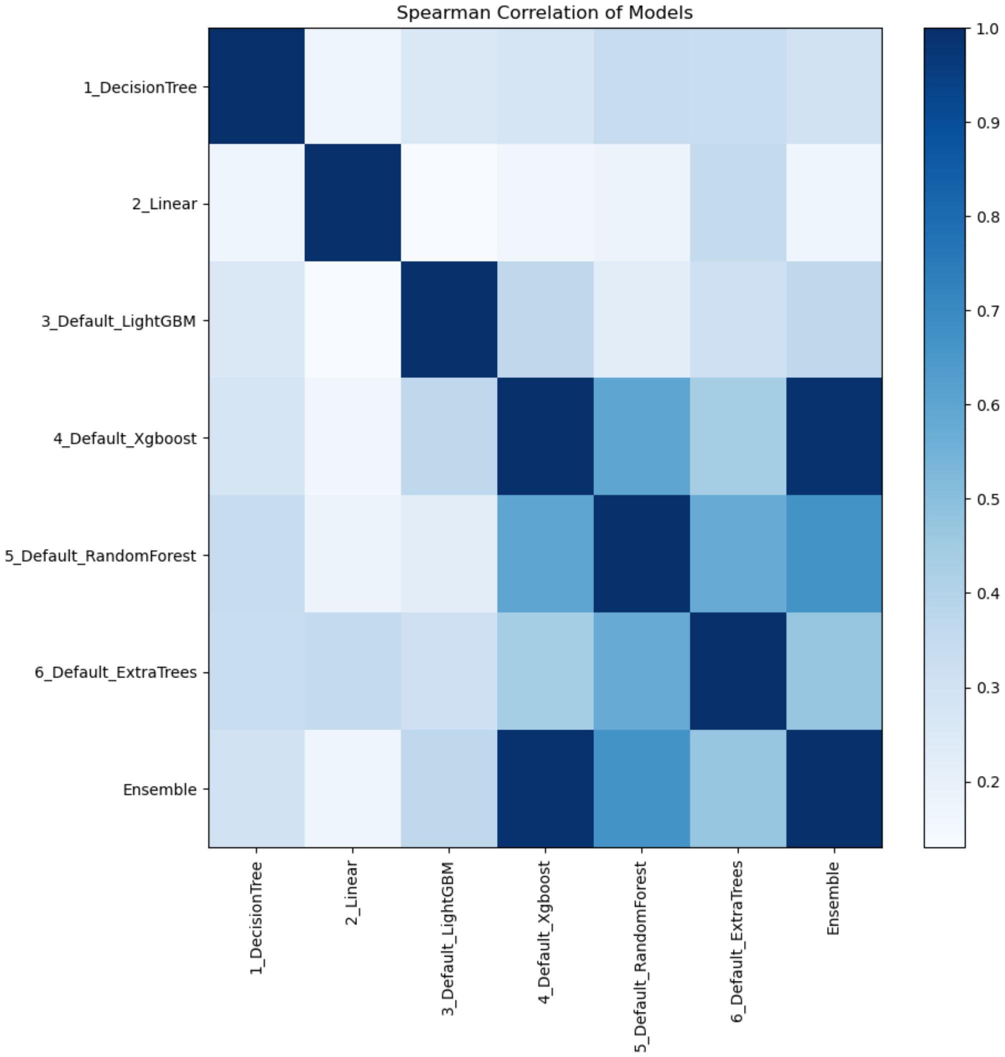


(l) ExtraTrees

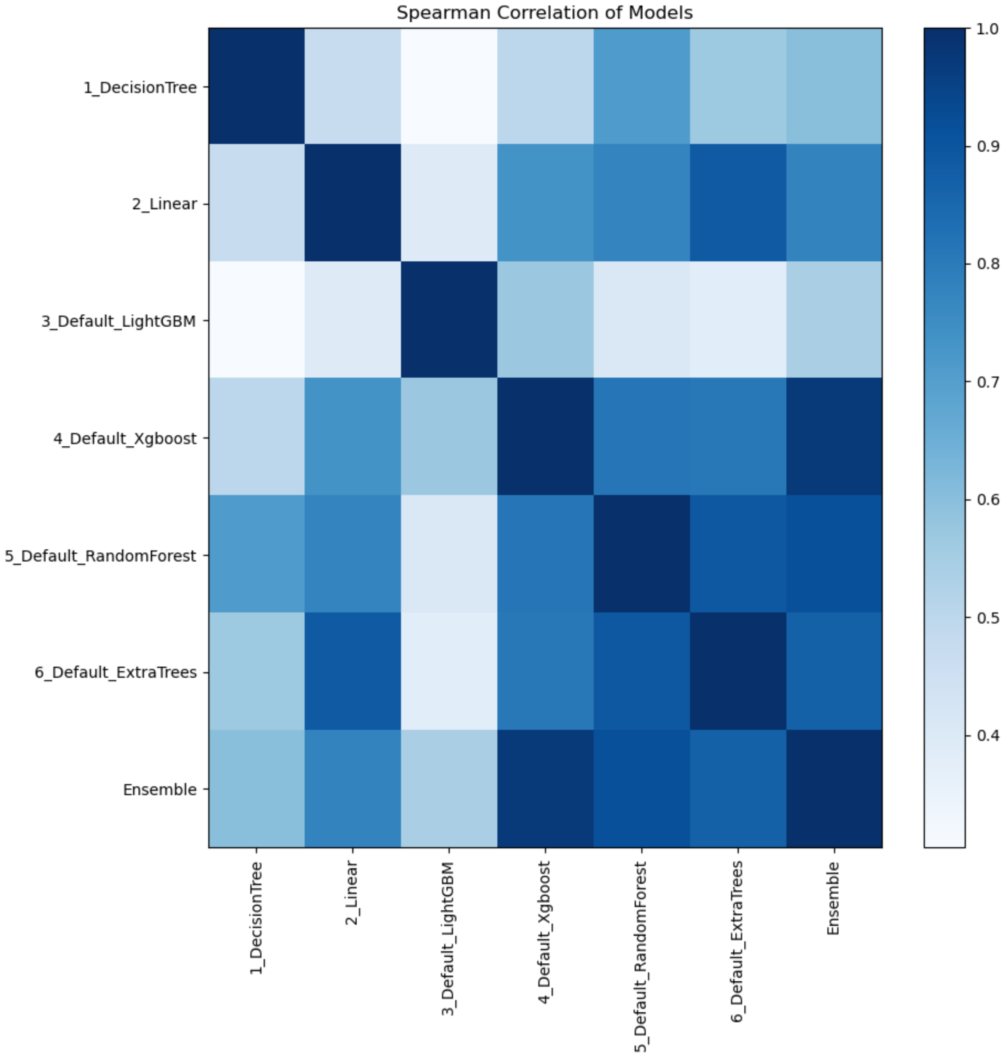
Figure 3.5: Confusion matrix of railway study: (a) (b) (c) (d) (e) (f). Confusion matrix of railway study: (g) (h) (i) (j) (k) (l).

Based on Fig 3.5, it can be concluded that the models' performance is good. This method is suitable to this study. The accuracy metric provides the percentage of correctly classified instances out of the total number of instances, indicating that the model has a high level of predictive power. This result suggests that the model can effectively classify instances into their correct categories, as measured by the true positive and true negative values in the confusion matrix. However, it is important to consider other performance metrics such as precision, recall, and F1 score to gain a more comprehensive understanding of the model's performance and its ability to generalize to new data. Overall, a high accuracy score is a positive sign of the model's effectiveness and its potential value in practical applications.

Spearman correlation is a statistical measure used to evaluate the strength and direction of the monotonic relationship between two variables. It is often used to compare the performance of machine learning models by calculating the correlation between the predicted values and the actual values. A higher correlation indicates a better fit between the predicted and actual values and thus a more accurate model. The Spearman Correlation of Models of six machine learning is shown in Fig 3.6.


Spearman Correlation of Models
1_DecisionTree
2_Linear
3_Default_LightGBM
4_Default_Xgboost
5_Default_RandomForest
6_Default_ExtraTrees
Ensemble
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

(a)highway



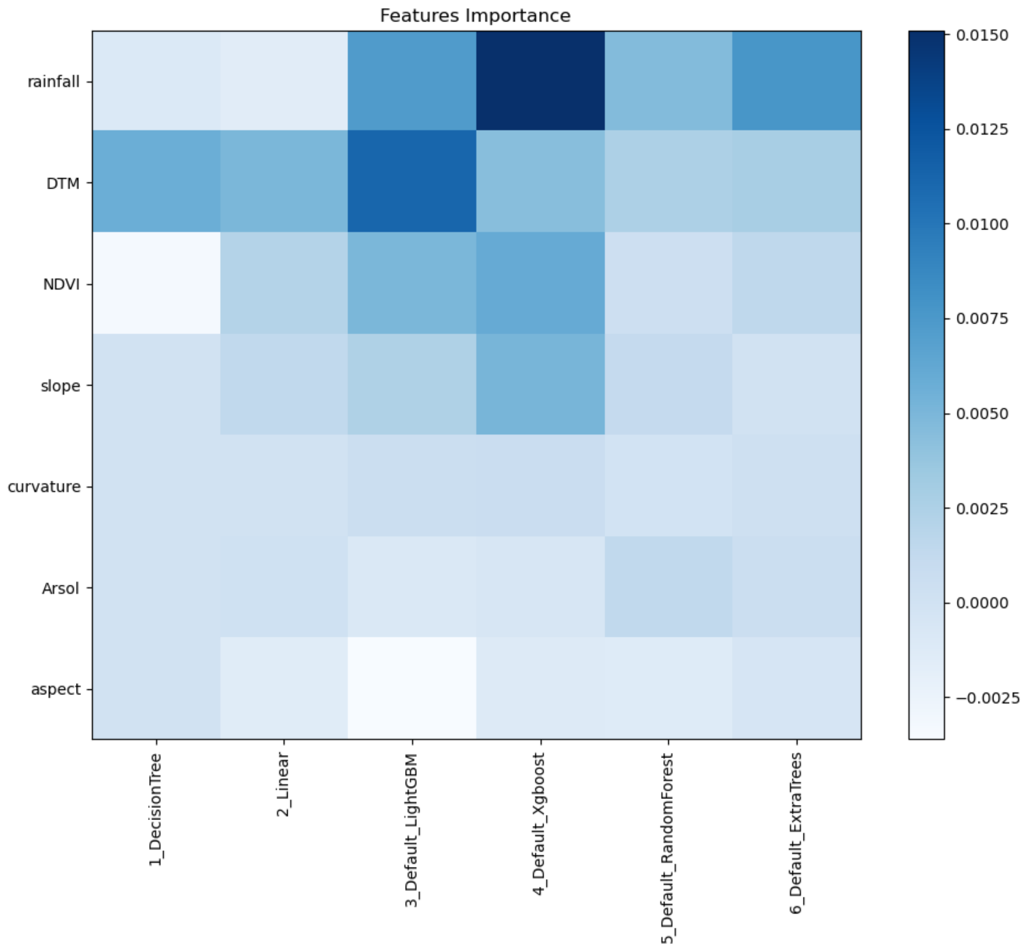
(b)railway

Figure 3.6: Spearman Correlation of Models

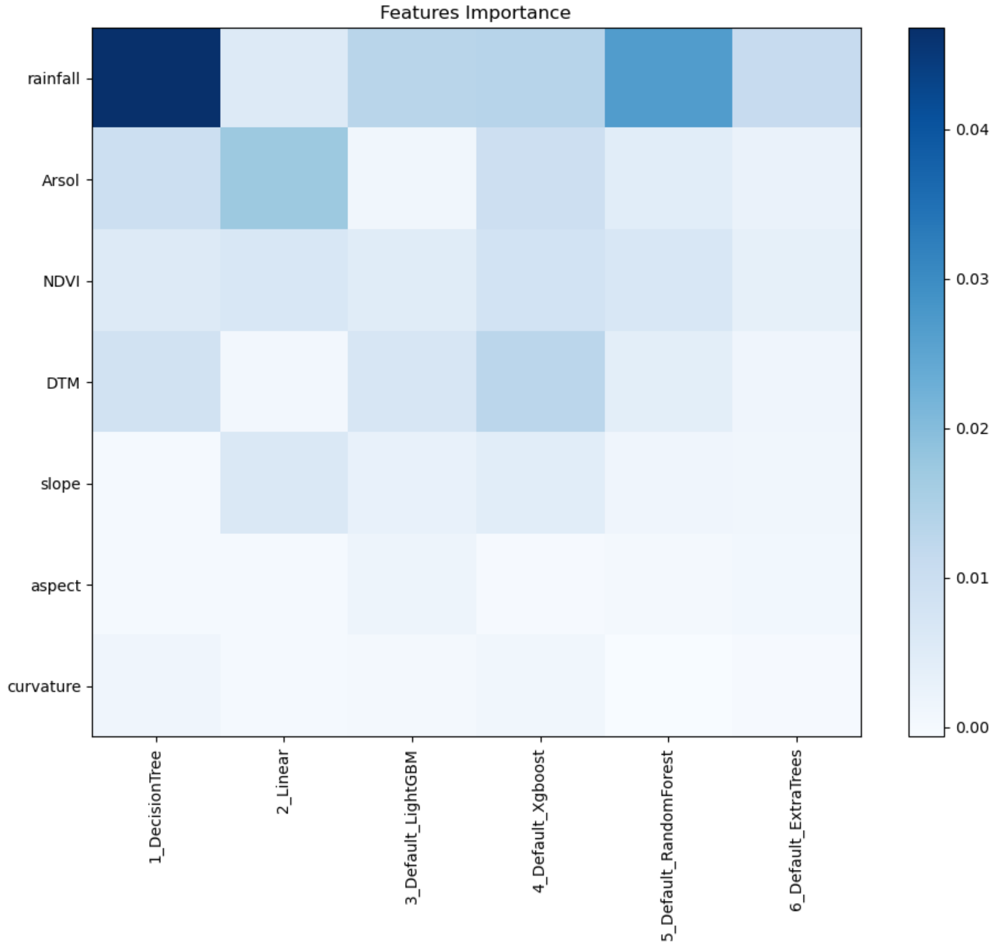
In Spearman Correlation of Models, it is obvious that the results of last three models (Xgboost, RandomForest and ExtraTrees) are more consistent with each other. The ensemble model which is automatically generated by the function Auto ML also select these three models as component. For highway study, the ensemble model includes Xgboost and RandomForest with weight 5 and 1 resprecively. As for railway study, the ensemble model includes Xgboost and RandomForest with weight 2 and 1 resprecively. The ensemble model will be used to make risk map.

3.4 Factor analysis in machine learning

Importance analysis is a crucial step in the interpretation and understanding of machine learning results. It involves identifying the relative importance of each feature used in a predictive model, in terms of its contribution to the overall accuracy of the model. By analyzing feature importance, we can determine which features are the most relevant for predicting the target variable, and gain insights into the underlying relationships and patterns in the data. This can be especially useful for understanding complex models and for identifying areas where the model may be overfitting or underperforming. There are various methods for calculating feature importance, ranging from simple methods like correlation analysis to more complex techniques like permutation feature importance and SHAP (Shapley Additive Explanations) values. Importance analysis is essential for ensuring that machine learning models are reliable and transparent. The Fig 3.7 is the importance of factors about the relationship between ground motion and infrastructure.



(a)highway importance



(b) railway importance

Figure 3.7: Features Importance

By Features Importance, it is obvious that precipitation is the main factor affecting ground motion whether for railways or highways. For the highway, elevation will influence ground motion more, but for railway, areal solar radiation is the second reason. Meanwhile, NDVI also cannot be ignored in ground motion. As for slope, it influences more highway than railway. Because the structure of highway and railway is different. The railways always build on the plain area and in the mountain area, the bridge will be built for railway.

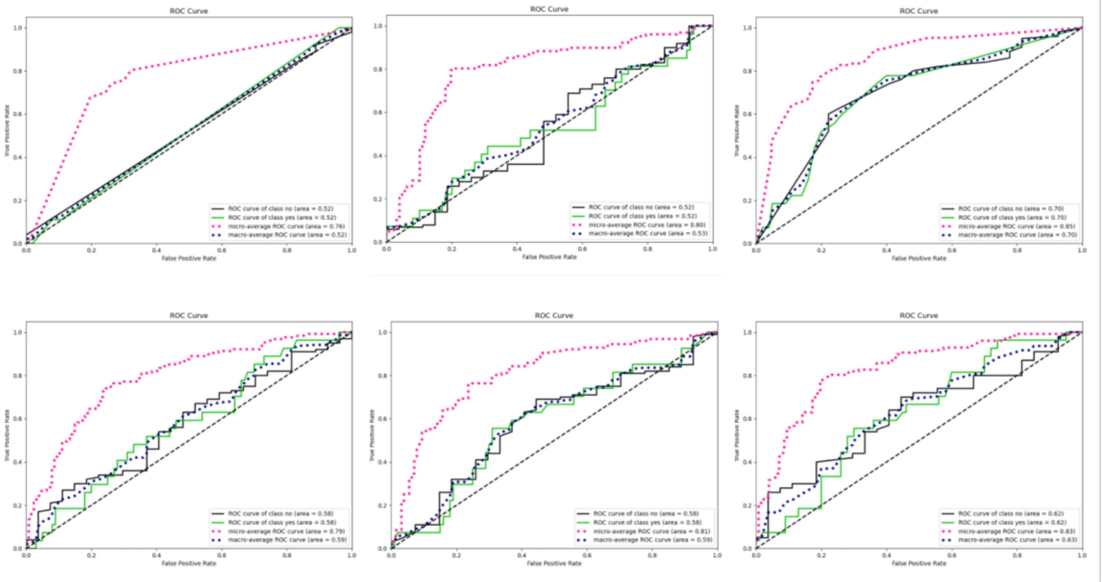
Permutation-based Importance and SHAP Importance are two popular techniques used for feature importance analysis in machine learning. Permutation-based Importance involves shuffling the values of each feature in the dataset and measuring the effect on the model's performance. The higher the drop in performance after shuffling a feature, the more important that feature is considered to be. On the other hand, SHAP (SHapley Additive exPlanations) Importance is a game theory-based method that measures the contribution of each feature to the model's prediction for a particular data point. It computes the average impact of each feature across all possible combinations of features and assigns a score to each feature based on its contribution to the model's output. Both techniques are useful for identifying the most important features in a machine learning model and can help in feature selection and model optimization.

Importance type	Permutation-based Importance		SHAP Importance	
Range(km)	highway	railway	highway	railway
DecisionTree	elevation	rainfall	rainfall	rainfall
LinearRegression	elevation	solar radiation	rainfall	solar radiation
LightGBM	elevation	rainfall	rainfall	NDVI
Xgboost	rainfall	rainfall	rainfall	solar radiation
RandomForest	rainfall	rainfall	rainfall	rainfall
ExtraTrees	rainfall	rainfall	rainfall	rainfall

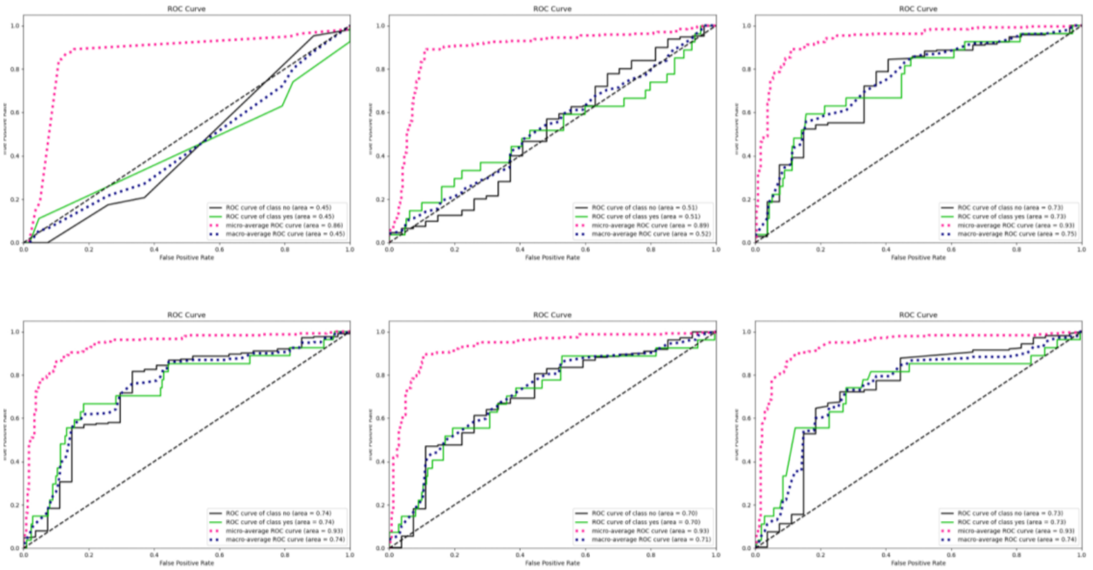
Table 3.4: The main reason of ground motion

3.5 Ground motion on the infrastructure and nearby area

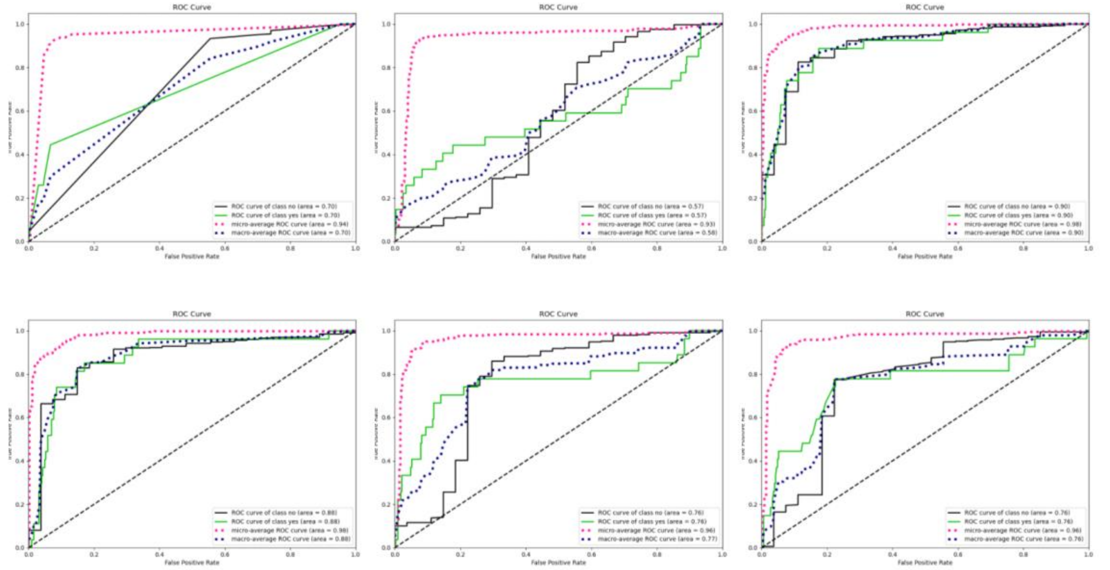
In the above part, we discussed the factors that cause ground motion on infrastructure. The ground motion tends to be concentrated in a certain area. Therefore, in order to further explore the relationship between conditional factors and ground motion, the ground motion happen on the infrastructure and its nearby area should be analyzed. In this part, we set different range with 1km, 3km, 5km and 10km to analyzes difference between these two objects about ground motion, and machine learning method also is used in this part.



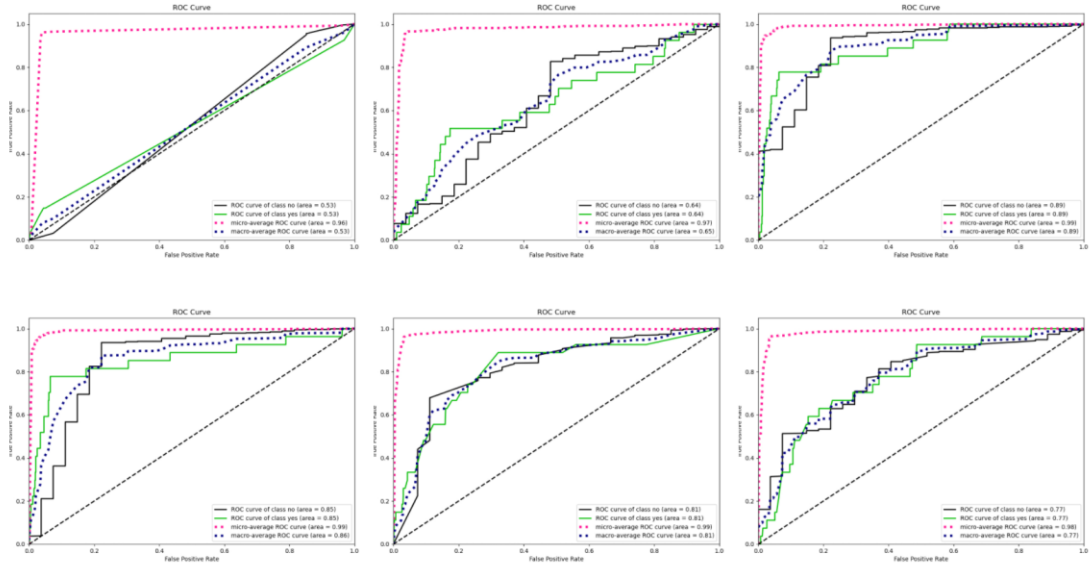
(a)highway analysis in 1km range



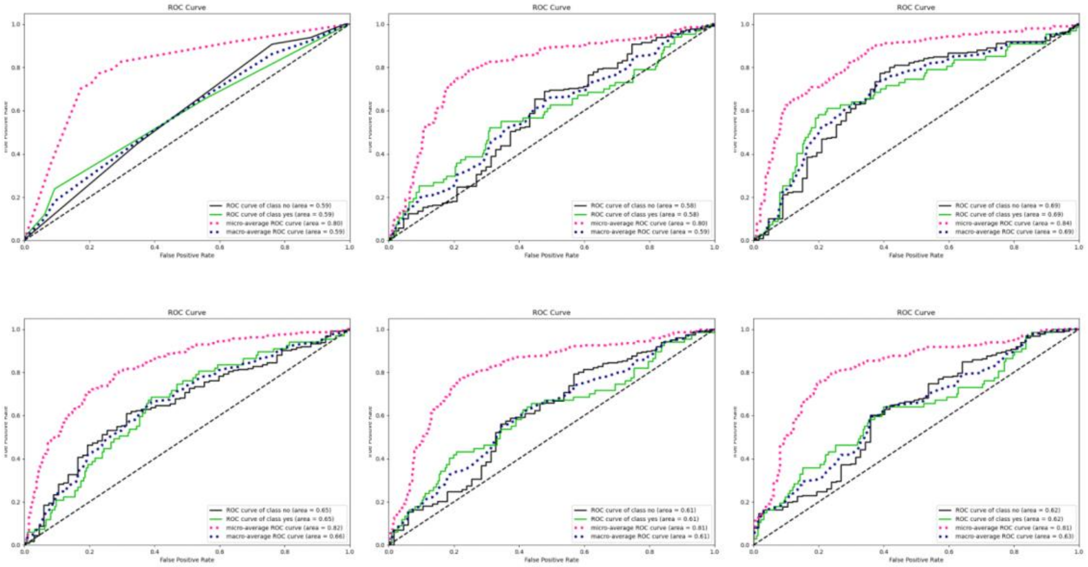
(b)highway analysis in 3km range



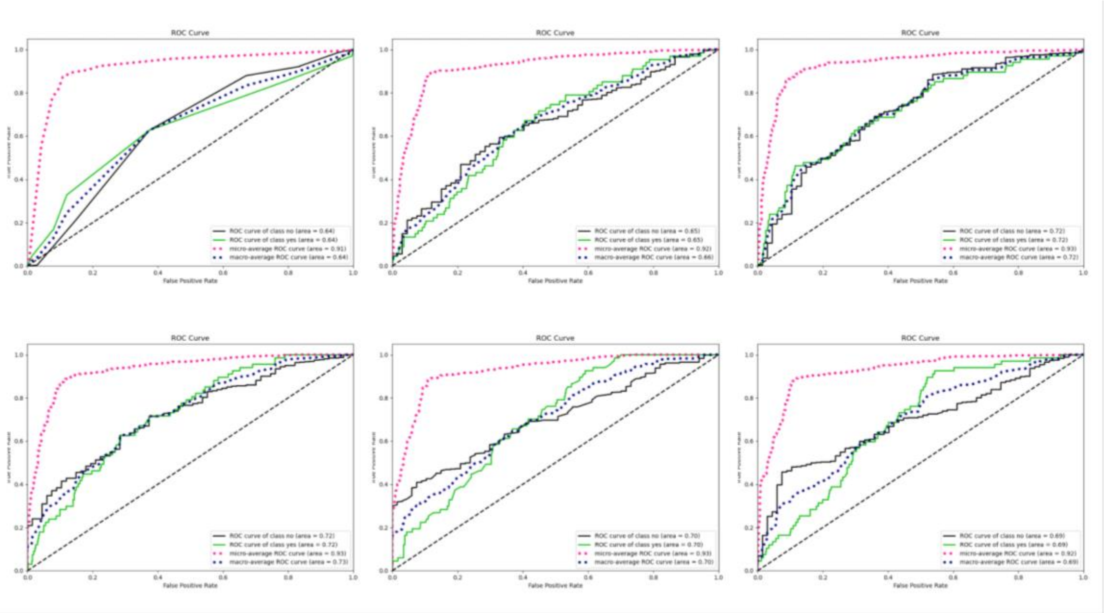
(c)highway analysis in 5km range



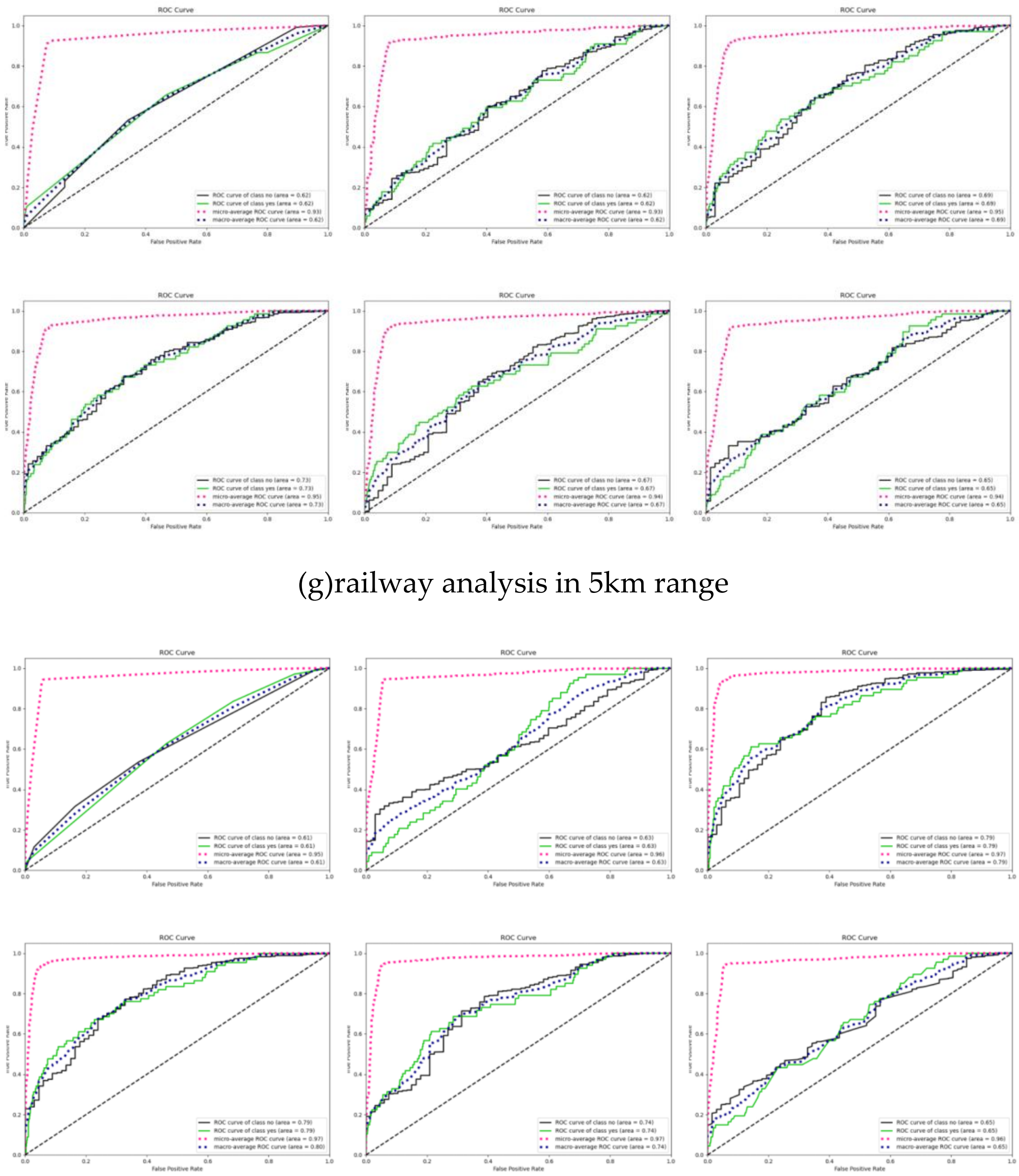
(d)railway analysis in 10km range



(e)railway analysis in 1km range



(f)railway analysis in 3km range



(g)railway analysis in 5km range

(h)railway analysis in 10km range

Figure 3.8: ROC curves and AUC value

In Fig 3.8, with the increasement of range, the ROC curves are more tend to the left and up part, which means that there are more data are used to training and validation. This result is corresponded that the performance of learners can benefit significantly from much larger training sets (Banko and brill, 2001). Meanwhile, the result of AUC also certificates it in Fig 3.4 and Table 3.2.

Range	1km	3km	5km	10km
1DecisionTree	0.58	0.63	0.62	0.61
2LinearRegression	0.58	0.65	0.62	0.63
3LightGBM	0.68	0.72	0.69	0.79
4Xgboost	0.65	0.72	0.73	0.80
5RandomForest	0.61	0.70	0.66	0.73
6ExtraTrees	0.62	0.69	0.65	0.65

Table 3.4: AUC in machine learning of railway

Range	1km	3km	5km	10km
1DecisionTree	0.52	0.44	0.7	0.53
2LinearRegression	0.52	0.51	0.57	0.64
3LightGBM	0.70	0.73	0.90	0.89
4Xgboost	0.58	0.74	0.88	0.85
5RandomForest	0.58	0.70	0.76	0.81
6ExtraTrees	0.62	0.73	0.76	0.77

Table 3.5: AUC in machine learning of highway

As for the permutation Permutation-based Importance and SHAP Importance of ground motion in infrastructure and nearby area, we can know the main reason which would cause ground motion in certain area. The following table are importance of highway and railway (Table 3.6 and Table 3.7).

Importance type	Permutation-based Importance				SHAP Importance			
Range(km)	1	3	5	10	1	3	5	10
DecisionTree	rainfall	elevation	elevation	slope	elevation	elevation	elevation	elevation
Linear	elevation	slope	slope	slope	elevation	elevation	slope	elevation
LightGBM	elevation	elevation	elevation	elevation	elevation	elevation	elevation	elevation
Xgboost	NDVI	elevation	elevation	elevation	slope	elevation	elevation	elevation
RandomForest	NDVI	elevation	elevation	elevation	rainfall	elevation	elevation	elevation
ExtraTrees	NDVI	elevation	elevation	rainfall	elevation	NDVI	elevation	rainfall

Table 3.6: Main reason in highway study

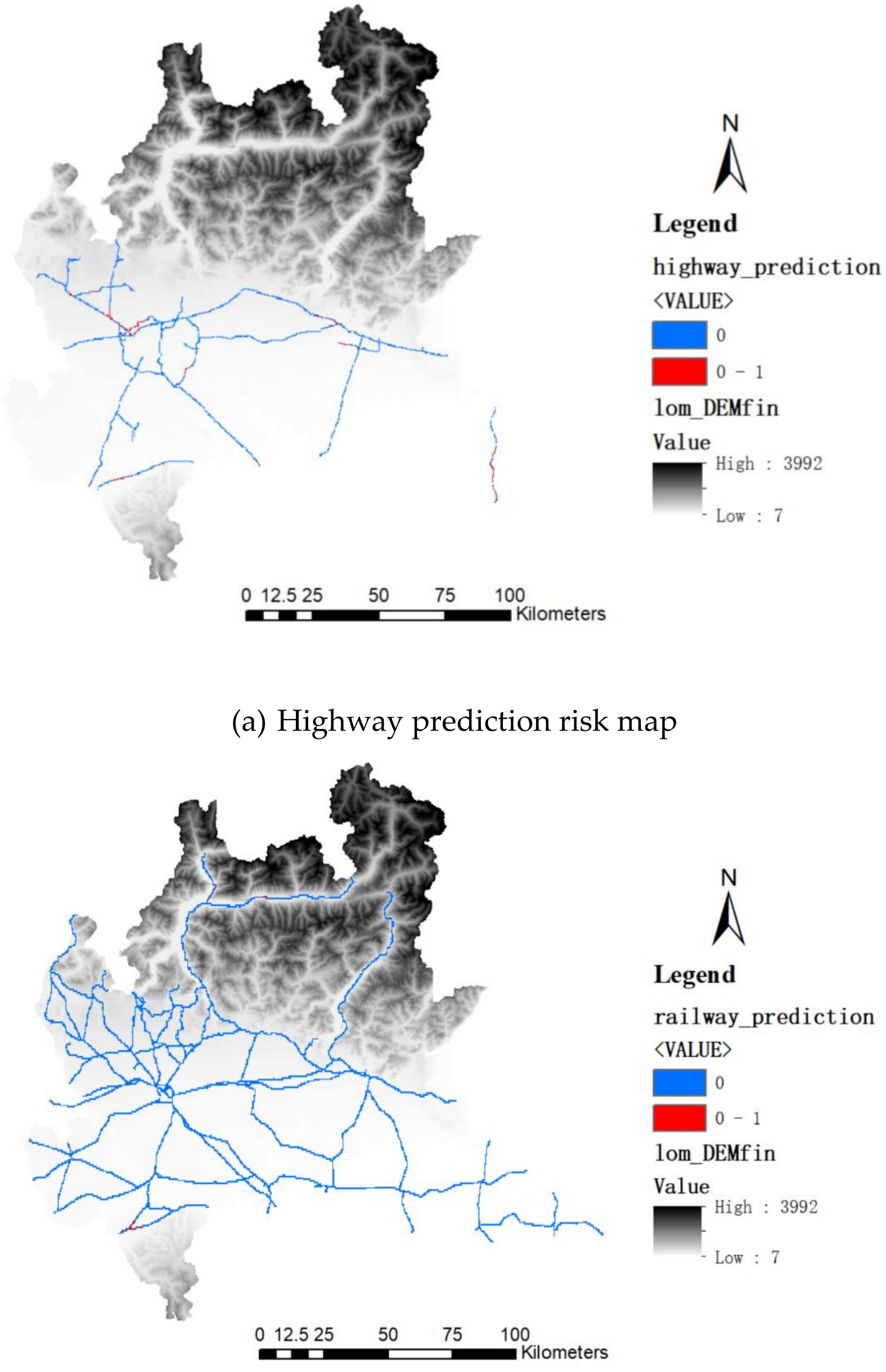
Importance type	Permutation-based Importance				SHAP Importance			
Range(km)	1	3	5	10	1	3	5	10
DecisionTree	rainfall	slope	elevation	elevation	rainfall	elevation	elevation	elevation
Linear	curvature	elevation	elevation	elevation	elevation	elevation	elevation	elevation
LightGBM	rainfall	elevation	elevation	rainfall	rainfall	elevation	elevation	elevation
Xgboost	NDVI	solar radiation	rainfall	rainfall	solar radiation	elevation	elevation	elevation
RandomForest	NDVI	elevation	elevation	elevation	elevation	elevation	elevation	elevation
ExtraTrees	NDVI	rainfall	rainfall	slope	rainfall	rainfall	rainfall	slope

Table 3.7: Main reason in railway study

By the result of importance analysis, in highway study, there are four factors (elevation, NDVI, rainfall and slope), which is the main reason in each machine learning model. It is obvious that elevation appear the most times, which means in the high elevation area, the ground motion prefers to happen in the infrastructure. As for railway study, the factors are elevation, slope, NDVI, rainfall and areal solar radiation. Different from the highway study, which is only concentrate on elevation, the frequency of elevation and rainfall is similar. Therefore, for the railway, we need to give more attention on the area which have high elevation and more rainfall.

3.6 Risk maps

After above process, the next step is to generate the ground motion risk map. Risk mapping is an important part of ground motion analysis that helps to identify and assess the potential risks associated with ground motion events, such as earthquakes or landslides. The goal of risk mapping is to provide a visual representation of the potential hazards, vulnerabilities, and exposure to various types of ground motion events in a given area. Risk mapping can help decision-makers and stakeholders to understand the potential consequences of ground motion events and develop appropriate mitigation and preparedness measures to reduce the risks. In this study, the risk maps are based on the result of ensemble models in machine learning (Fig 3.9).



(a) Highway prediction risk map

(b) Railway prediction risk map

Figure 3.9: Prediction risk map from ensemble machine learning mode

4. Discussion

4.1 Rationality of conditioning factor selection

In ground motion study, there seven condition factors are introduced including elevation, slope angel, slope aspect, curvature, areal solar radiation, rainfall and NDVI. At first, for the study which is related to ground motion on highways and railways, the FR method is used to analyze the relationship. In this part, for the highway, the main distribution areas of ground motion are characterized by elevation (240.3-317.6m), slope angle (10-20°), slope aspect (0-45°), areal solar radiation (0.6-0.8 MW·H/m2), curvature (-1-0°), rainfall (905-915 mm), NDVI (0.2-0.4). The FR value of ground motion of these conditioning factors (classes) are 2.030,1.832, 1.666, 1.527, 1.071, 0.817 and 1.372, respectively. Thus, a more concise conclusion can be drawn from the data analysis, that is, regions with elevation (240.3-317.6 m), slope angle (10-20°) and slope aspect (0-45°).

As for railway study, the main distribution areas of ground motion are characterized by elevation (8.4-85.7m), slope angle (<10°), slope aspect (315-360°), areal solar radiation (0.2-0.4MW·H/m2), curvature (2-4°), rainfall (875-885 mm), NDVI (0.8-1). The FR value of ground motion of these conditioning factors (classes) are 1.751,1.049, 1.783, 1.586, 1.682, 3.917 and 1.902, respectively. Based on the results of the data analysis, it can be inferred that areas exhibiting a slope aspect within the range of 315-360°, rainfall between 875-885 mm, and NDVI values in the range of 0.8-1, are more likely to be associated with ground motion. It is worth noting that the FR value is more significant than other condition factors for highway in FR analysis.

Then the machine learning method is used in this study. For the highway, rainfall, elevation and NDVI are the most closely related to the ground motion. As for railway, these three condition factors are rainfall, areal solar radiation and NDVI. Compared with FR method, the importance of condition factors of highway is different. However, in railway study, the rainfall is the most important condition factor for these two methods. Meanwhile, the importance of NDVI is also reflected in both methods. In Permutation-based Importance and SHAP Importance, elevation, rainfall and areal solar radiation appear in highway research, and NDVI, rainfall and areal solar radiation are closely related to ground motion of railway.

According to the importance analysis of the ground motion on the infrastructure and its nearby area, the location with higher elevation is easily to happen ground motion on the highway. For the railway, the area with more rainfall and higher elevation has higher chance of ground motion on the railway.

In a word, for the highway and railway, we need to respond accordingly in the location with specific features. For the ground motion location, which is on the high elevation area or with more rainfall, the highway and railway are more prone to ground motion compared with its nearby area.

4.2 Rationality of model selection

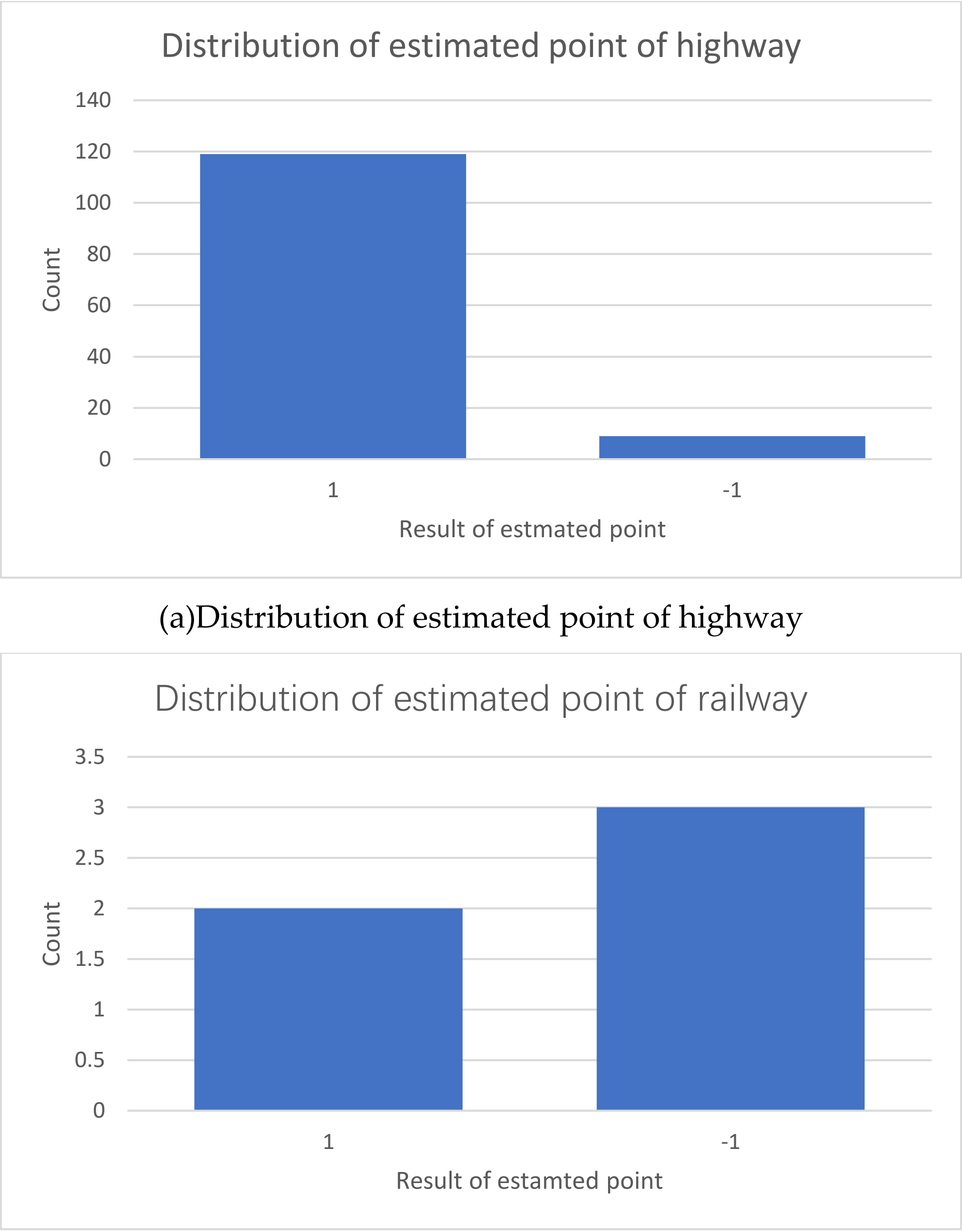
Model selection played an important role in this research. However, there are no widely accepted criterions to guide model selection. As this study is the first to evaluate ground motion risk using ML models, and there is a lack of similar studies which can be referred in model selection. Therefore, six well-performed and typical ML models are adopted to conduct this study, i.e. DecisionTree, Linear, LightGBM, Xgboost, RandomForest and ExtraTree models.

As shown in Table 3.2, in terms of highway, Xgboost, RandomForest and ExtraTree show a better performance with AUC value 0.78, 0.77 and 0.73. Conversely, the value of DecisionTree, Linear, LightGBM is 0.50, 0.60 and 0.70. Similarly, the better machine learning model of railway are also Xgboost, RandomForest and ExtraTree with 0.80, 0.78 and 0.78 AUC value. In the Table 3.4 and Table 3.5, in the study of infrastructure and its nearby area, especially in a large range, the Xgboost also demonstrate a superior performance with 0.80 and 0.85 of AUC value for highway and railway respectively.

Despite achieving reasonably high accuracies, it is important to acknowledge that the models utilized in this study were conventional. Future research efforts should focus on the following areas: (1) employing advanced models, such as state-of-the-art hybrid machine learning and deep learning models; (2) predicting the spatial distribution of ground motion risk under anticipated climate scenarios and geological phenomena, such as earthquakes; and (3) utilizing optimization algorithms to refine the model parameters, thereby enhancing the overall predictive accuracy.

4.3 Analysis of ground motion risk maps and real case

In this part, we compare the prediction result and the real ground motion condition. The result of estimation can be divided into two types, 1 and -1, which means overestimated point and underestimated point. Overestimated point means that the point without ground is the ground motion point in the prediction. Conversely, underestimated point is the point with ground motion but not be predicted. The Fig 4.1 is the distribution of estimated point.



(a)Distribution of estimated point of highway

(b) Distribution of estimated point of railway

Figure 4.1: Distribution of estimated point

According to the Fig 4.1, the accuracy of railway forecasting is extremely high. The wrong predicted point is only 5. Compared with the number of total points of railway, the number of mispredictions is almost negligible. However, the estimation result of highway is worthy of discussion. The overestimated points are nearly 120, but the number of underestimated points is less than 10. The reason why the overestimated point is much more than underestimated point need to be discussed.

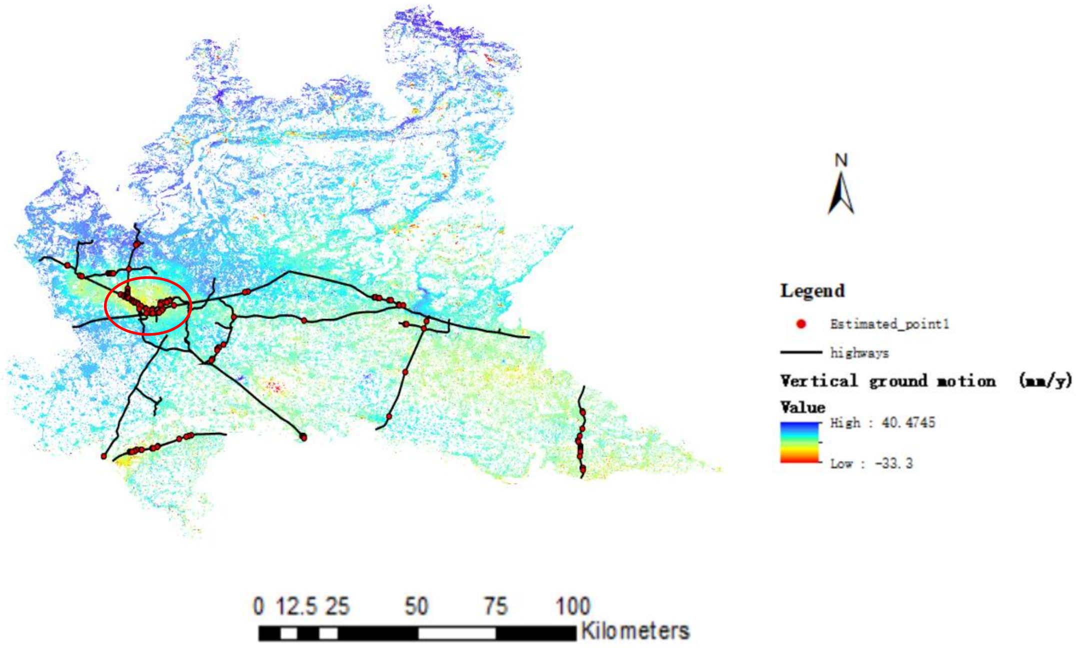


Figure 4.2: The location of overestimated point

In the Fig 4.2, we can know the overestimated points are near to the real ground motion point. Especially, in the select area in Fig 4.2, this concentration is more significant. By analyzing the distribution of features of this typical position, the main reason rainfall is almost unchanged in these points (Fig 4.3).

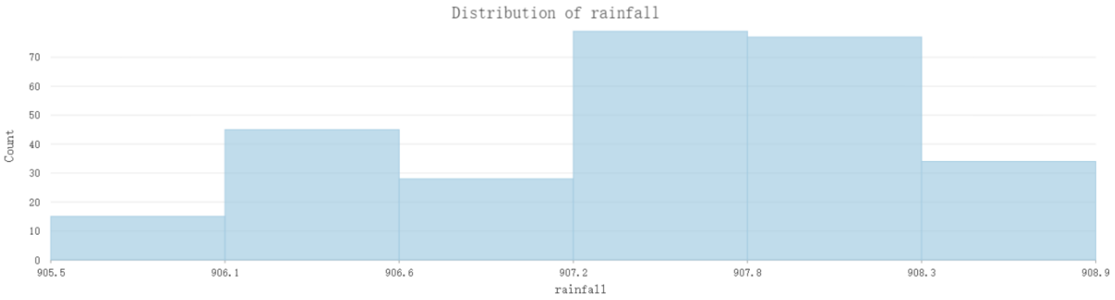


Figure 4.3: Distribution of rainfall

In the Fig 4.4, the red points are overestimated points and the green points are ground motion points. It is obvious that two types of points are very near, which means the difference of features of points are similar. So, in machine learning models, for the areas where ground movement is relatively concentrated, ground motion on the highway will be overestimated. In terms of this condition, more accurate and closely related feature should be input as training dataset, so as to provide more judgment methods to improve the accuracy of machine learning.

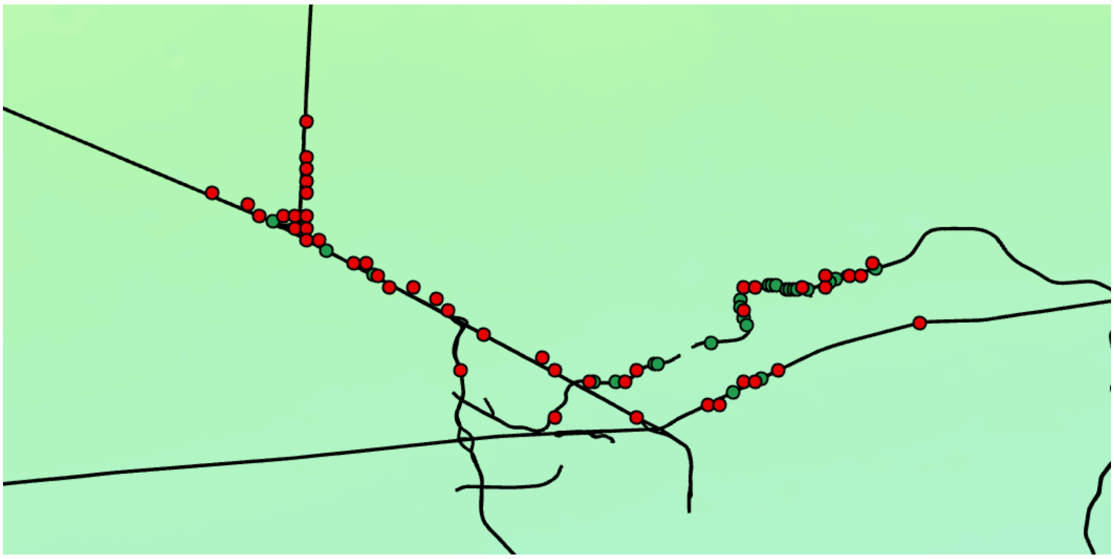


Figure 4.4: Overestimated points and ground motion points

5. Conclusion

In this study, we presented a ground motion risk evaluation for the Lombardy region using six machine learning models, i.e. DecisionTree, Linear, LightGBM, Xgboost, RandomForest and ExtraTree models, and the following conclusions can be drawn from this study:

(1) The correlation analysis between ground motion and condition factors shows that highway with elevation (240.3m-317.6m), NDVI (0.2-0.4) and rainfall (905m-915 mm) are the main distribution areas of ground motion. As for railway, the main reasons are rainfall (875mm-885mm), NDVI (0.8-1.0) and areal solar radiation (0.2-0.4).

(2) The results of the machine learning analysis demonstrate that, in the context of highway analysis, the Xgboost model exhibits the highest level of performance as measured by the AUC metric, with a value of 0.78. The RandomForest model follows closely behind with an AUC value of 0.77, while the ExtraTrees model has an AUC of 0.73. Through the evaluation of the models, the elevation, rainfall, and NDVI factors are found to have the greatest impact on ground motion. In contrast, the railway study reveals that the Xgboost model performed best with an AUC value of 0.80, followed by the ExtraTrees (0.78) and RandomForest (0.78) models. The most significant factors contributing to ground motion are identified as elevation, rainfall, and areal solar radiation.

(3) In the ground motion area, which has higher elevation, highway and railway are easy to happen ground motion. Compared with highway, ground motion of railway is also prone to occur in areas with a lot of precipitation.

(4) As for the railway, machine learning model provide an accurate ground motion risk map. However, in the prediction of highway ground motion, the risk in areas of concentrated ground motion is overestimated.

(5) It is important to note that this study serves as a preliminary examination into the relationship between factors and ground motion. While it has provided valuable insights, there are several limitations that must be addressed in future research, because the factors causing ground motion are complex and numerous. Further investigations may benefit from incorporating additional relevant factors into the

models. Additionally, future studies should also focus on exploring the potential adverse impacts of ground motion on infrastructure availability.

Bibliography

Hill D P, Pollitz F, Newhall C. Earthquake-volcano interactions[J]. Physics Today, 2002, 55(11): 41-47.

Grünthal G. Induced seismicity related to geothermal projects versus natural tectonic earthquakes and other types of induced seismic events in Central Europe[J]. Geothermics, 2014, 52: 22-35.

Wang H Y, Xie L L. Effects of topography on ground motion in the Xishan park, Zigong city[J]. Chinese Journal of Geophysics, 2010, 53(7): 1631-1638.

Karimzadeh S, Miyajima M, Hassanzadeh R, et al. A GIS-based seismic hazard, building vulnerability and human loss assessment for the earthquake scenario in Tabriz[J]. Soil Dynamics and Earthquake Engineering, 2014, 66: 263-280.

Khosravikia F, Clayton P. Machine learning in ground motion prediction[J]. Computers & Geosciences, 2021, 148: 104700.

Ni J, Wu T, Zhu X, et al. Risk assessment of potential thaw settlement hazard in the permafrost regions of Qinghai-Tibet Plateau[J]. Science of the Total Environment, 2021, 776: 145855.

Li R, Zhang M, Pei W, et al. Risk evaluation of thaw settlement using machine learning models for the Wudaoliang-Tuotuohe region, Qinghai-Tibet Plateau[J]. Catena, 2023, 220: 106700.

Khosravikia F, Clayton P. Machine learning in ground motion prediction[J]. Computers & Geosciences, 2021, 148: 104700.

Kong Q, Trugman D T, Ross Z E, et al. Machine learning in seismology: Turning data into insights[J]. Seismological Research Letters, 2019, 90(1): 3-14.

Trugman D T, Shearer P M. Strong correlation between stress drop and peak ground acceleration for recent M 1–4 earthquakes in the San Francisco Bay area[J]. Bulletin of the Seismological Society of America, 2018, 108(2): 929-945.

Crosetto M, Solari L, Mróz M, et al. The evolution of wide-area DInSAR: From regional and national services to the European Ground Motion Service[J]. Remote Sensing, 2020, 12(12): 2043.

Brunelli G, Lanzano G, Luzi L, et al. Data-driven zonations for modelling the regional source and propagation effects into a Ground Motion Models in Italy[J]. Soil Dynamics and Earthquake Engineering, 2023, 166: 107775.

Antonielli B, Mazzanti P, Rocca A, et al. A-DInSAR performance for updating landslide inventory in mountain areas: An example from lombardy region (Italy)[J]. Geosciences, 2019, 9(9): 364.

Garbin M, Priolo E. Seismic event recognition in the Trentino area (Italy): Performance analysis of a new semiautomatic system[J]. Seismological Research Letters, 2013, 84(1): 65-74.

Ikuemonisan F E, Ozebo V C, Olatinsu O B. Investigating and modelling ground settlement response to groundwater dynamic variation in parts of Lagos using space-based retrievals[J]. Solid Earth Sciences, 2021, 6(2): 95-110.

Gattinoni P, Scesi L. The groundwater rise in the urban area of Milan (Italy) and its interactions with underground structures and infrastructures[J]. Tunnelling and Underground Space Technology, 2017, 62: 103-114.

Wang, C., Lin, Q., Wang, L., et al., 2022. The influences of the spatial extent selection for non-landslide samples on statistical-based landslide susceptibility modelling: a case study of Anhui Province in China. Nat. Hazards 112, 1967–1988.

Youssef, A.M., Pourghasemi, H.R., 2021. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. Geosci. Front. 12, 639–655.

Chen, W., Li, Y., 2020. GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models. CATENA 195, 104777.

Nguyen, Q.H., Ly, H.-B., Ho, L.S., et al., 2021. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Mathematical Problems in Engineering 2021, 1–15

Tien Bui, D., Tuan, T.A., Klempe, H., et al., 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides 13, 361–378.

Caruana, Rich et al. "Ensemble Selection from Libraries of Models." Proceedings of the 21st International Conference on Machine Learning. Banff, Canada (2004).

Hong, H., Pradhan, B., Jebur, M.N., et al., 2015. Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines. Environm. Earth Sci. 75, 40.

Tacconi Stefanelli, C., Casagli, N., Catani, F., 2020. Landslide damming hazard susceptibility maps: a new GIS-based procedure for risk management. Landslides 17, 1635–1648.

Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation[C]//Proceedings of the 39th annual meeting of the Association for Computational Linguistics. 2001: 26-33.

Chen, W., Yan, X., Zhao, Z., et al., 2019. Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China). Bull. Eng. Geol. Environ. 78, 247–266.

Aditian, A., Kubota, T., Shinohara, Y., 2018. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. Geomorphology 318, 101–111.

Ho T K. Random decision forests[C]//Proceedings of 3rd international conference on document analysis and recognition. IEEE, 1995, 1: 278-282.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. Machine learning, 2006, 63: 3-42.

Luino F. Sequence of instability processes triggered by heavy rainfall in the northern Italy[J]. Geomorphology, 2005, 66(1-4): 13-39.

Peresan A, Zuccolo E, Vaccari F, et al. Neo-deterministic seismic hazard scenarios for North-Eastern Italy[J]. Bollettino della Società geologica italiana, 2009, 128(1): 229-238.

Take W A, Bolton M D, Wong P C P, et al. Evaluation of landslide triggering mechanisms in model fill slopes[J]. Landslides, 2004, 1: 173-184.

Lourenço S D N, Sassa K, Fukuoka H. Failure process and hydrologic response of a two layer physical model: implications for rainfall-induced landslides[J]. Geomorphology, 2006, 73(1-2): 115-130.

An J, Zheng F, Lu J, et al. Investigating the role of raindrop impact on hydrodynamic mechanism of soil erosion under simulated rainfall conditions[J]. Soil Science, 2012, 177(8): 517-526.

Beullens, J., Van de Velde, D., Nyssen, J., 2014. Impact of slope aspect on hydrological rainfall and on the magnitude of rill erosion in Belgium and northern France. CATENA 114, 129–139.

Wu, C.L., Chau, K.W., 2013. Prediction of rainfall time series using modular soft computingmethods. Eng. Appl. Artif. Intell. 26, 997–1007.

Li, R., Wang, N., 2019. Landslide Susceptibility Mapping for the Muchuan County (China): A Comparison Between Bivariate Statistical Models (WoE, EBF, and IoE) and Their Ensembles with Logistic Regression. Symmetry 11, 762.

Ohlmacher, G.C., 2007. Plan curvature and landslide probability in regions dominated by earth flows and earth slides. Eng. Geol. 91, 117–134

Rahmati O, Golkarian A, Biggs T, et al. Land subsidence hazard modeling: Machine learning to identify predictors and the role of human activities[J]. Journal of environmental management, 2019, 236: 466-480.

Naumburg E, Mata-Gonzalez R, Hunter R G, et al. Phreatophytic vegetation and groundwater fluctuations: a review of current research and application of ecosystem response modeling with an emphasis on Great Basin vegetation[J]. Environmental Management, 2005, 35(6): 726-740.