Manufacturing Data Science Final Project

Bridge Deterioration Factor Analysis and Deterioration Rates Prediction

Yi-Ting Lin, Tsung-Han Tsai, Po-Yi Wang, Ting-En Tsai,

Abstract

Following the 2019 Nanfang'ao Bridge collapse in Taiwan, this project investigates key factors influencing bridge deterioration in northern Taiwan and develops a predictive model for deterioration rates. Collaborating with the Northern Region Maintenance Engineering Office, we optimized bridge inspection processes using historical data.

Our methodology involved data cleaning and feature selection using OLS, Elastic Net, and Random Forest. Significant variables ($p\text{-value} < 0.05$) were selected, including interaction terms. Elastic Net identified key standardized features, and Random Forest highlighted top variables by importance. Bridges were labeled as rapidly deteriorating based on a threshold, resulting in imbalanced data. We applied SMOTE to balance the dataset, yielding nine significant variables. Models—including Logistic Regression, Random Forest, XGBoost, and SVM—were trained using cross-validation, stratified splitting, and grid search, focusing on maximizing recall.

We developed a Streamlit-based webpage for highway bureau employees to analyze bridge data and highlight key deterioration factors. This data-driven approach enhances safety management and resource allocation, offering a framework for improving bridge safety nationwide.

1. Background and Motivation

1.1 Research Motivation

The 2019 Nanfang'ao Bridge collapse highlighted the critical importance of bridge safety in Taiwan. Investigations revealed non-compliant construction, inadequate supervision, and neglected maintenance as primary causes, emphasizing the need for effective inspections.

Bridges are essential for daily commuting; their deterioration can cause public concern and force the government to invest more in monitoring, leading to resource waste. We aim to optimize inspection processes by identifying key deterioration factors and predicting rates, providing data-driven evidence for efficient resource utilization.

1.2 Research Background

Collaborating with the Northern Region Maintenance Engineering Office, we utilized their database for bridge inspections in northern Taiwan. Moving beyond traditional "experience-based" two-year cycles, we applied data science to assist in maintenance decisions and predict optimal schedules, enhancing efficiency and safety.

As of 2023, Taiwan has 2,492 highway bridges, with 883 in the northern region. Inspections are categorized into routine, regular, special, detailed, and monitoring.

Northern Taiwan's high population density and humid climate result in bridges bearing heavy traffic, frequent rain, and exposure to corrosive environments, necessitating more frequent inspections. Currently, inspection planning relies on inspectors' experience and intuition, following a two-year cycle and assigning DER & U values to prioritize maintenance. Despite available data, it's underutilized for analysis and prediction.

Therefore, we applied data science methods to analyze bridge data, identify key deterioration factors, and predict rates, providing evidence to optimize schedules and processes.

1.3. Problem Definition

We will use data science techniques to analyze key factors causing bridge deterioration, employing Ordinary Least Squares (OLS), Random Forest, and Elastic Net. After balancing dataset labels ("normal" vs. "rapid" deterioration) using SMOTE, we will apply classification methods like Logistic Regression, Random Forest, XGBoost, SVM, and LSTM to predict rapid deterioration. Finally, we will develop a web interface displaying results of nine variables and classification model predictions, allowing users to select bridge information and deterioration levels via drop-down menus. Real-time analyses will output important variable charts through regression, Random Forest, and Elastic Net, providing insights into the causes of bridge deterioration for inspection companies and staff.

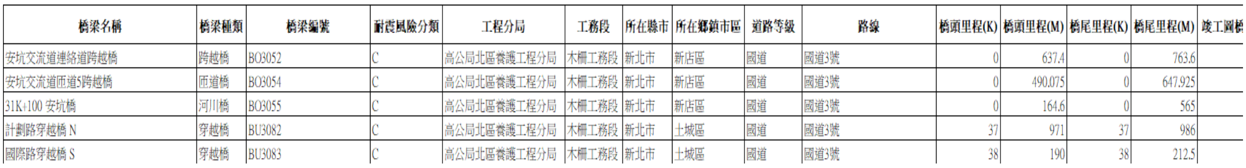
2. Data Sources and Data Preprocessing

2.1 Data Sources

We utilized data authorized by the Northern Region Maintenance Engineering Office of the Ministry of Transportation and Communications (MOTC), specifically from the National Highway Bridge Management System. The data includes:

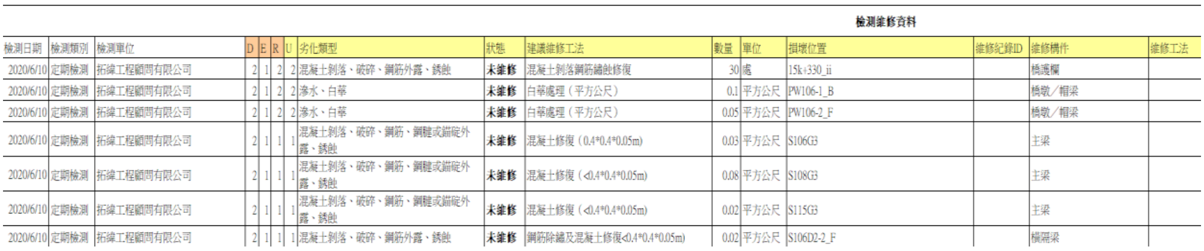
- **Bridge Basic Information**: 883 records with 105 variables, each representing a bridge in the northern region.
- **Bridge Maintenance Records (2018–2023)**: Includes fields such as inspection categories, DERU values, damage locations, and maintenance components.

Due to the high dimensionality of the dataset, inputting all variables into the model could lead to the "curse of dimensionality." Therefore, after a meeting on November 14 with the Northern Region Maintenance Engineering Office and their bridge inspection contractor, Highway Engineering Consultants Co., Ltd., we identified important variables based on domain expertise. We ultimately selected 35 independent variables $\boldsymbol{X}$and calculated the bridge urgency value $\boldsymbol{U}$ based on past inspection results as the dependent variable $\boldsymbol{Y}$ for analysis.



橋梁名稱	橋梁種類	橋梁編號	耐震風險分類	工程分局	工務段	所在縣市	所在鄉鎮市區	道路等級	路線	橋頭里程(K)	橋頭里程(M)	橋尾里程(K)	橋尾里程(M)	竣工圖橋
安坑交流道連絡道跨越橋	跨越橋	BO3052	C	高公局北區養護工程分局	木柵工務段	新北市	新店區	國道	國道3號	0	637.4	0	763.6	
安坑交流道匝道5跨越橋	匝道橋	BO3054	C	高公局北區養護工程分局	木柵工務段	新北市	新店區	國道	國道3號	0	490.075	0	647.925	
31K+100 安坑橋	河川橋	BO3055	C	高公局北區養護工程分局	木柵工務段	新北市	新店區	國道	國道3號	0	164.6	0	565	
計劃路穿越橋 N	穿越橋	BU3082	C	高公局北區養護工程分局	木柵工務段	新北市	土城區	國道	國道3號	37	971	37	986	
國際路穿越橋 S	穿越橋	BU3083	C	高公局北區養護工程分局	木柵工務段	新北市	土城區	國道	國道3號	38	190	38	212.5	

Table 1: Northern Region Bridge Data from the National Highway Bridge Management System



												檢測維修資料			
檢測日期	檢測類別	檢測單位	D	E	R	U	劣化類型	狀態	建議維修工法	數量	單位	損壞位置	維修紀錄ID	維修構件	維修工法
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	2	2	混凝土剝落、破碎、鋼筋外露、銹蝕	未維修	混凝土剝落鋼筋鏽蝕修復	30	處	15k+330_ii		橋護欄	
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	2	2	滲水、白華	未維修	白華處理（平方公尺）	0.1	平方公尺	PW106-1_B		橋墩／帽梁	
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	2	2	滲水、白華	未維修	白華處理（平方公尺）	0.05	平方公尺	PW106-2_F		橋墩／帽梁	
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	1	1	混凝土剝落、破碎、鋼筋、鋼腱或錨碇外露、銹蝕	未維修	混凝土修復（0.4*0.4*0.05m)	0.03	平方公尺	S106G3		主梁	
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	1	1	混凝土剝落、破碎、鋼筋、鋼腱或錨碇外露、銹蝕	未維修	混凝土修復（<0.4*0.4*0.05m)	0.08	平方公尺	S108G3		主梁	
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	1	1	混凝土剝落、破碎、鋼筋、鋼腱或錨碇外露、銹蝕	未維修	混凝土修復（<0.4*0.4*0.05m)	0.02	平方公尺	S115G3		主梁	
2020/6/10	定期檢測	拓緯工程顧問有限公司	2	1	1	1	混凝土剝落、破碎、鋼筋外露、銹蝕	未維修	鋼筋除鏽及混凝土修復<0.4*0.4*0.05m)	0.02	平方公尺	S106D2-2_F		橫隔梁	

Table 2: Sample of Inspection and Maintenance Data

2.2 Data Preprocessing

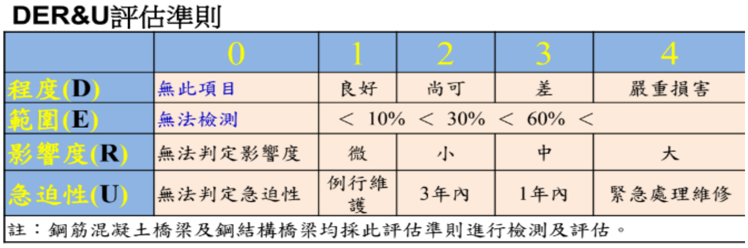
2.2.1 Dependent Variable

The target variable ***Y*** is the urgency value ***U*** of bridge maintenance, evaluated based on each inspection's D (degree), E (extent), and R (influence) values. Following the "Highway Steel Bridge Inspection and Reinforcement Specifications" issued by the Northern Region Maintenance Engineering Office, we calculated weighted scores for each component based on their importance to the bridge (refer to Table 3).

For each inspection:

- Maintained components were assigned a ***U*** value of 1; others remained unchanged.
- We computed the weighted average ***U*** value for each bridge's components.
- The deterioration rate ***U_change_ave*** was calculated as the difference between current and previous periods divided by the number of days.
- Averaging these rates over 2018–2023 yielded each bridge's average deterioration rate.
- Weighting scores are detailed in Table 4.

We observed some bridges with negative deterioration rates, indicating improved conditions without maintenance, which is implausible. Therefore, we used the IQR*1.5 method to remove lower outliers. After removing outliers, 836 bridge records remained. Detailed data is presented in Figure 1.



DER&U評估準則

	0	1	2	3	4
程度(D)	無此項目	良好	尚可	差	嚴重損害
範圍(E)	無法檢測	＜ 10% ＜ 30% ＜ 60% ＜			
影響度(R)	無法判定影響度	微	小	中	大
急迫性(U)	無法判定急迫性	例行維護	3年內	1年內	緊急處理維修
註：鋼筋混凝土橋梁及鋼結構橋梁均採此評估準則進行檢測及評估。					

Table 3: Evaluation Standards for Bridge DERU Values

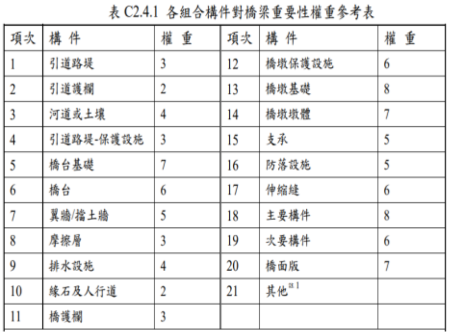


表 C2.4.1 各組合構件對橋梁重要性權重參考表

項次	構 件	權 重	項次	構 件	權 重
1	引道路堤	3	12	橋墩保護設施	6
2	引道護欄	2	13	橋墩基礎	8
3	河道或土壤	4	14	橋墩墩體	7
4	引道路堤-保護設施	3	15	支承	5
5	橋台基礎	7	16	防落設施	5
6	橋台	6	17	伸縮縫	6
7	翼牆/擋土牆	5	18	主要構件	8
8	摩擦層	3	19	次要構件	6
9	排水設施	4	20	橋面版	7
10	緣石及人行道	2	21	其他[註1]	
11	橋護欄	3			

Table 4: Weighted Importance of Bridge Components

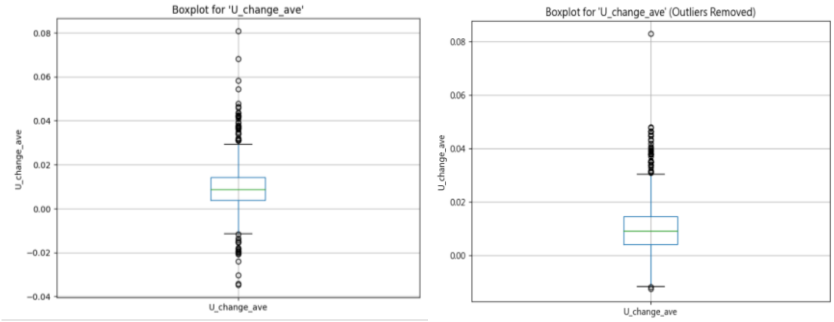


Figure 1: Removal of Extreme Values

2.2.2 Independent Variables

We processed a comprehensive set of independent variables, carefully sourced from multiple government datasets and systems, to ensure accurate analysis of bridge deterioration rates. The table below provides a detailed description of each variable, the corresponding data processing method, and how missing values were handled.

Variable Name	Data Processing Method	Missing Value Handling
Monthly Average Temperature	Scraped data from 29 stations of the Central Weather Bureau (CWB) matching each bridge's location; calculated average monthly temperatures (2018–2023).	No missing value
Average Monthly Relative Humidity	Scraped CWB data matching bridge locations; calculated average monthly relative humidity (2018–2023).	No missing value
Average Total Sunshine Hours	Collected CWB data; calculated average annual total sunshine hours (2018–2023).	No missing value
Average Annual Zinc Corrosion Rate (μm/yr)	Used 'Taiwan Zinc Metal Annual Corrosion Rate Data' from the Institute of Transportation (IOT), MOTC; matched to bridge locations; calculated averages (2019–2023).	No missing value
Average Annual Carbon Steel Corrosion Rate (μm/yr)	Used 'Taiwan Carbon Steel Metal Annual Corrosion Rate Data' from IOT; matched to bridge locations; calculated averages (2019–2023).	No missing value
Annual Average Daily Traffic Volume	Used 'Electronic Toll Collection Traffic Statistics' from the Northern Region Maintenance Engineering Office; matched	No missing value

	interchange and bridge locations; calculated averages (2018–2023).	
Annual Average Precipitation	Scraped CWB data; calculated average annual precipitation (2018–2023).	No missing value
Is the Bridge a Water Crossing	Based on National Highway Bridge Management System data; converted 'Yes'/'No' to binary (1/0).	No missing value
Soil Liquefaction Potential	Used Soil Liquefaction Potential Query System; assigned values 3 (high), 2 (medium), 1 (low), and 0 (no data) based on bridge locations.	Missing values were categorized as 'no data' (0).
Total Bridge Length (m)	Obtained **from** National Highway Bridge Management **System data.**	Missing values were filled with the mode (most common length for similar bridges).
Maximum Net Width (m)	From the system data; missing values filled with the mode.	No missing value
Minimum Net Width (m)	From the system data.	No missing value
Total Number of Lanes	From the system data.	No missing value
Number of Mainline Lanes	From the system data.	No missing value
Under-Bridge Leasing	From the system data; converted 'Yes'/'No' to binary (1/0).	No missing value
Bridge Site Conditions	From the system data (used for post-earthquake s)pecial inspection classification).	No missing value
Distance from Coastline (m)	From the system data.	No missing value
Nearest Fault Distance	From the system data.	No missing value
Is the Bridge Monitored	From the system data; converted 'Yes'/'No' to binary (1/0).	No missing value
Main Girder Material	From the system data; categories include 8 material combinations, such as Prestressed Concrete and Reinforced Concrete. One-Hot Encoding was used to represent each material as a binary variable.	No missing value.
Design Seismic Intensity	From the system data; missing values filled with the mode.	Missing values were filled using the mode.
Design Live Load	From the system data; missing values filled with the mode. Encoded 'HS20-44' as 1, 'HS20-44+10%' as 1.1, etc., based on proportion. Other categories filled with the mode.	Missing values were filled with the mode for comparable bridge types.
Design Horizontal Ground Acceleration	From the system data	Missing values were filled using the mode.
Maximum Considered Ground Acceleration	From the system data	Missing values were filled using the mode.

Design Horizontal Seismic Coefficient	From the system data	Missing values were filled using the mode.
Design Vertical Seismic Coefficient	From the system data	Missing values were filled with the mode.

3. Methodology

3.1 Feature Selection

We employed three methods to analyze key factors influencing bridge deterioration and selected the most significant variables based on the combined results.

(1) Ordinary Least Squares (OLS)

- Observed the p-values of variables, considering those with p-values below 0.05 as significant.
- For significant variables inconsistent with assumptions from the Northern Region Maintenance Engineering Office, we created interaction terms with other important variables and re-ran the regression to confirm their interactions and causal relationships.

(2) Elastic Net

- Standardized all variables.
- Adjusted penalty terms and Lasso ratios to limit the output to 10 variables.

(3) Random Forest

- Identified the top 10 variables based on feature importance.

3.2 Regression Model Prediction

3.2.1 Prediction Objective

- To assist the Northern Region Maintenance Engineering Office in classifying bridges that may require more frequent inspections, preventing sudden collapses or damages that could endanger lives and increase repair costs.

3.2.2 Labeling Method

- Bridges exceeding the below threshold were labeled as "1" (rapid deterioration); others were labeled as "0".

$$\begin{aligned} &\mathit{Rapid\ deterioration\ threshold} \\ &\quad = median(U\_change\_ave) + 2\sigma(U\_change\_ave) \end{aligned}$$

- After labeling, the data distribution became highly imbalanced (see Figure 2). Initial predictions on unbalanced data had high accuracy but failed to identify rapidly deteriorating bridges. Therefore, we applied data balancing methods to the training set.

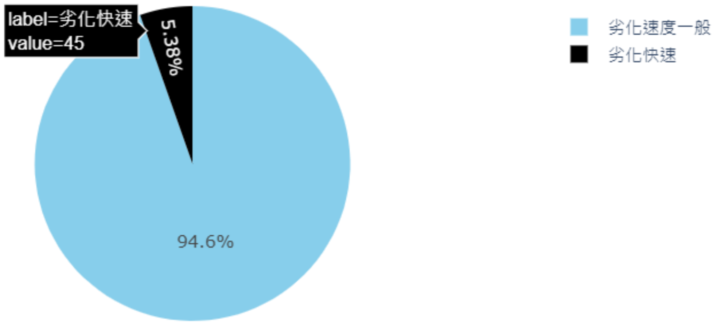


Figure 2: Data Distribution After Rapid Deterioration Labeling

3.2.3 Data Balancing

- Split the dataset into 80% training and 20% testing sets using *random_state = 42.*
- Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the training set by generating synthetic samples near minority class samples, achieving a 1:1 sample size ratio.

$$x_{new} = x_{chosen} + (x_{nearest} - x_{chosen}) * \delta;\ \delta \in [0,1]$$

3.2.4 Feature Selection

- After re-labeling and adjusting the data distribution with SMOTE, we re-applied feature selection methods to identify important features in the training set.
- Used Random Forest to select the top 10 important variables, as it provided higher recall scores before data balancing compared to other models and the variables were easier to input into the interface.
- Performed correlation analysis (see Figure 3) and removed variables with high correlation (coef. = 0.81), specifically the "Average Annual Carbon Steel Corrosion Rate (μm/yr)", since it is directly influenced by relative humidity and adds little explanatory power beyond that.

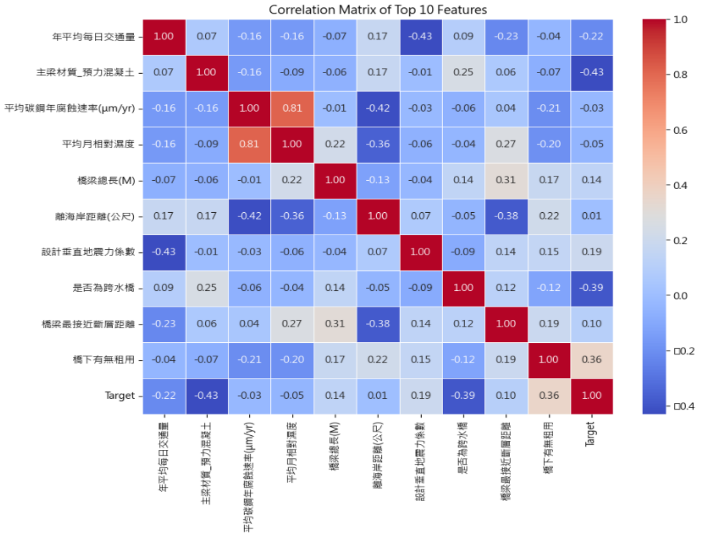


Figure 3: Correlation analysis for feature selection

- The final nine variables were:
    1. Annual Average Daily Traffic Volume
    2. Main Girder Material: Prestressed Concrete
    3. Average Monthly Relative Humidity
    4. Total Bridge Length (m)
    5. Distance from Coastline (m)
    6. Design Vertical Seismic Coefficient
    7. Is the Bridge a Water Crossing
    8. Nearest Fault Distance
    9. Under-Bridge Leasing

3.2.5 Model Training

- Employed four models: **Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM)**, using the nine selected variables for training.
- Implemented the following methods to optimize model parameters (training results are shown in Table 5):
    - i. **Cross-validation**: Applied SMOTE to balance the training set within cross-validation and used unbalanced data as the validation set.
    - ii. **Stratified Splitting**: Ensured each split's training and validation sets had similar label proportions.
    - iii. **Grid Search**: Used grid search to find the optimal parameters.
    - iv. **Recall Optimization**: Aimed to maximize the recall score to avoid missing bridges that may rapidly deteriorate.
- Due to limited expertise in bridge engineering, we also ensured the training and testing sets belonged to the same data distribution. We used PCA to select 14 variables that explained 80% of the data variance for model training. The training results are shown in Table 7.

Model	**Best Parameters**	**Best Recall**
Logistic Regression	C: 0.05	0.6143
Random Forest	max_depth: 3, n_estimators: 30	0.6464
XGBoost	learning_rate: 0.01, max_depth: 5, n_estimators: 50	0.7000
SVM	C: 0.01, kernel: 'linear', max_iter: 100,000	0.6500

Table 6: Training Results of Models

Model	**Best Parameters**	**Best Recall**
Logistic Regression	C: 0.3	0.6821
Random Forest	max_depth: 3, n_estimators: 30	0.6781
XGBoost	learning_rate: 0.01, max_depth: 3, n_estimators: 40	0.6536

Model	Best Parameters	Best Recall
SVM	C: 0.1, kernel: 'linear', max_iter: 30,000	0.7107

Table 7: PCA Training Results

- To ensure that deep learning models do not miss specific information in the training data, all variables were included during model training. We used SMOTE to balance 64% of the training data. The data was split into 64% for training, 16% for validation, and 20% for testing in the LSTM model.

Model	Best Parameters	Best Recall
LSTM	best_epoch=700, hidden_size=25, num_layers=1, optimizer = Adam	0.6250

Table 8: Training Results of LSTM

3.2.6 Model Selection

- Our primary goal is to avoid failing to predict bridges that may rapidly deteriorate, as their collapse or damage could result in significant personnel and repair costs.
- Since rapidly deteriorating bridges are rare in the test set, mispredicting even one can significantly lower the recall score.
- Therefore, we quantified each model's prediction performance using the formula:

$$\text{Performance Score} = 0.5 \times \text{Recall} + 0.5 \times \text{Accuracy}$$

- Based on this metric, we compared the models' prediction results and selected the best-performing model to deploy on our web interface for use by the Northern Region Maintenance Engineering Office.

4. Analysis Results

4.1 Analysis of Key Factors Influencing Deterioration

We identified the top 10 significant features using OLS, interaction term analysis, Elastic Net, and Random Forest methods:

4.1.1 Ordinary Least Squares (OLS)

- **Monthly Average Temperature** (p-value = 0.002)
  - **Coefficient**: Positive (+)
  - **Insight**: Higher monthly average temperatures are associated with faster deterioration rates.
- **Average Total Sunshine Hours** (p-value = 0.043)
  - **Coefficient**: Negative (−)
  - **Insight**: More sunshine hours correspond to slower deterioration rates.
  - **Discussion**: In northern regions during winter, the northeast monsoon leads to more

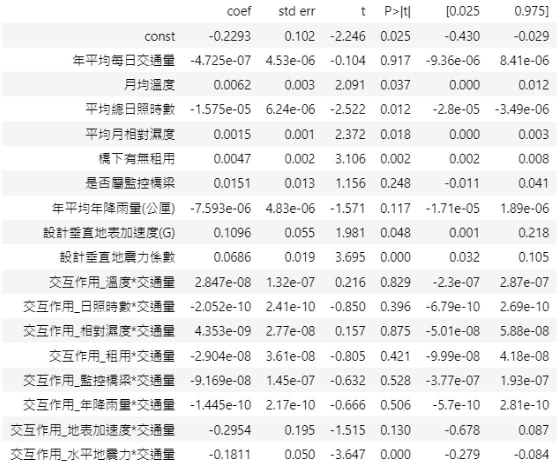
rainy days and fewer sunshine hours. Fewer sunshine hours imply more rainy days, which accelerates deterioration.

- **Average Monthly Relative Humidity** (p-value = 0.035)
    - **Coefficient**: Positive (+)
    - **Insight**: Higher humidity levels accelerate deterioration.
    - **Discussion**: This aligns with expectations, as the high number of rainy days and humidity in northern regions facilitate corrosion of bridge components.
- **Annual Average Daily Traffic Volume** (p-value = 0.044)
    - **Coefficient**: Negative (−)
    - **Insight**: Higher traffic volumes are associated with slower deterioration rates.
    - **Discussion**: Contrary to expectations that more vehicles would accelerate deterioration. This may be due to bridges designed to handle higher traffic loads being built more robustly from the outset.
- **Under-Bridge Leasing** (p-value = 0.001)
    - **Coefficient**: Positive (+)
    - **Insight**: Bridges with under-bridge leasing deteriorate faster.
    - **Discussion**: Leasing may introduce additional human activities that contribute to erosion and accelerate deterioration.
- **Is the Bridge Monitored** (p-value = 0.004)
    - **Coefficient**: Positive (+)
    - **Insight**: Monitored bridges are more likely to exhibit faster deterioration.
    - **Discussion**: This meets expectations, as bridges under special attention by the engineering bureau are likely those with faster deterioration rates.
- **Annual Average Precipitation (mm)** (p-value = 0.004)
    - **Coefficient**: Negative (−)
    - **Insight**: Higher precipitation correlates with slower deterioration rates.
    - **Discussion**: Contrary to initial assumptions that more rainfall accelerates deterioration. This might be because southern regions receive more rainfall than northern regions, suggesting humidity might be a more significant factor.
- **Design Vertical Ground Acceleration (G) / Design Horizontal Seismic Coefficient / Design Vertical Seismic Coefficient** (p-values = 0.023 / 0.014 / 0.011)
    - **Coefficient**: Positive (+)
    - **Insight**: Higher design values are associated with faster deterioration rates.
    - **Discussion**: Bridges designed to be more robust or to withstand higher seismic forces may deteriorate faster, indicating they experience greater stress in reality.

4.1.2 Further Analysis Using Interaction Terms

- **Annual Average Daily Traffic Volume** (Refer to Table 9)
    - After adding interaction terms between annual average daily traffic volume and other

significant variables, we found that its interaction with the horizontal seismic coefficient significantly reduces deterioration rates, while the variable itself becomes insignificant. This suggests that the effect of traffic volume on deterioration is moderated by the horizontal seismic coefficient, allowing bridges to handle higher traffic volumes with reduced deterioration.

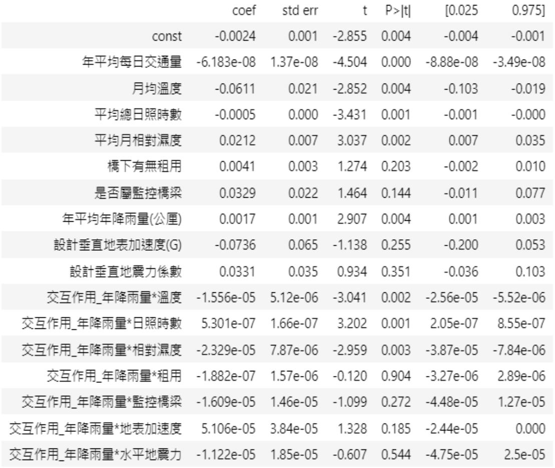


	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2293	0.102	-2.246	0.025	-0.430	-0.029
年平均每日交通量	-4.725e-07	4.53e-06	-0.104	0.917	-9.36e-06	8.41e-06
月均溫度	0.0062	0.003	2.091	0.037	0.000	0.012
平均總日照時數	-1.575e-05	6.24e-06	-2.522	0.012	-2.8e-05	-3.49e-06
平均月相對濕度	0.0015	0.001	2.372	0.018	0.000	0.003
橋下有無租用	0.0047	0.002	3.106	0.002	0.002	0.008
是否屬監控橋梁	0.0151	0.013	1.156	0.248	-0.011	0.041
年平均年降雨量(公厘)	-7.593e-06	4.83e-06	-1.571	0.117	-1.71e-05	1.89e-06
設計垂直地表加速度(G)	0.1096	0.055	1.981	0.048	0.001	0.218
設計垂直地震力係數	0.0686	0.019	3.695	0.000	0.032	0.105
交互作用_溫度*交通量	2.847e-08	1.32e-07	0.216	0.829	-2.3e-07	2.87e-07
交互作用_日照時數*交通量	-2.052e-10	2.41e-10	-0.850	0.396	-6.79e-10	2.69e-10
交互作用_相對濕度*交通量	4.353e-09	2.77e-08	0.157	0.875	-5.01e-08	5.88e-08
交互作用_租用*交通量	-2.904e-08	3.61e-08	-0.805	0.421	-9.99e-08	4.18e-08
交互作用_監控橋梁*交通量	-9.169e-08	1.45e-07	-0.632	0.528	-3.77e-07	1.93e-07
交互作用_年降雨量*交通量	-1.445e-10	2.17e-10	-0.666	0.506	-5.7e-10	2.81e-10
交互作用_地表加速度*交通量	-0.2954	0.195	-1.515	0.130	-0.678	0.087
交互作用_水平地震力*交通量	-0.1811	0.050	-3.647	0.000	-0.279	-0.084

Table 9: Interaction Term Analysis (Annual Average Daily Traffic Volume)

  - **Speculation**: Bridges with higher traffic volumes may be better designed (higher seismic coefficients) due to higher construction costs to prevent accidents. However, confirming this causal relationship would require a Difference-in-Differences (DiD) design with experimental and control groups.

- **Annual Average Precipitation** (Refer to Table 10)
  - After adding interaction terms between annual average precipitation and other significant variables, factors like sunshine hours, relative humidity, and temperature significantly impact deterioration rates. This indicates that the effect of precipitation on deterioration is not fixed and is influenced by these variables, necessitating their inclusion in further analyses.



	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0024	0.001	-2.855	0.004	-0.004	-0.001
年平均每日交通量	-6.183e-08	1.37e-08	-4.504	0.000	-8.88e-08	-3.49e-08
月均溫度	-0.0611	0.021	-2.852	0.004	-0.103	-0.019
平均總日照時數	-0.0005	0.000	-3.431	0.001	-0.001	-0.000
平均月相對濕度	0.0212	0.007	3.037	0.002	0.007	0.035
橋下有無租用	0.0041	0.003	1.274	0.203	-0.002	0.010
是否屬監控橋梁	0.0329	0.022	1.464	0.144	-0.011	0.077
年平均年降雨量(公厘)	0.0017	0.001	2.907	0.004	0.001	0.003
設計垂直地表加速度(G)	-0.0736	0.065	-1.138	0.255	-0.200	0.053
設計垂直地震力係數	0.0331	0.035	0.934	0.351	-0.036	0.103
交互作用_年降雨量*溫度	-1.556e-05	5.12e-06	-3.041	0.002	-2.56e-05	-5.52e-06
交互作用_年降雨量*日照時數	5.301e-07	1.66e-07	3.202	0.001	2.05e-07	8.55e-07
交互作用_年降雨量*相對濕度	-2.329e-05	7.87e-06	-2.959	0.003	-3.87e-05	-7.84e-06
交互作用_年降雨量*租用	-1.882e-07	1.57e-06	-0.120	0.904	-3.27e-06	2.89e-06
交互作用_年降雨量*監控橋梁	-1.609e-05	1.46e-05	-1.099	0.272	-4.48e-05	1.27e-05
交互作用_年降雨量*地表加速度	5.106e-05	3.84e-05	1.328	0.185	-2.44e-05	0.000
交互作用_年降雨量*水平地震力	-1.122e-05	1.85e-05	-0.607	0.544	-4.75e-05	2.5e-05

Table 10: Interaction Term Analysis (Annual Average Precipitation)

4.1.3 Elastic Net

Top variables identified:

1. Under-Bridge Leasing
2. Nearest Fault Distance
3. Monthly Average Temperature
4. Design Vertical Seismic Coefficient
5. Main Girder Material—Prestressed Concrete, Reinforced Concrete
6. Average Total Sunshine Hours
7. Average Annual Zinc Corrosion Rate
8. Annual Average Daily Traffic Volume
9. Average Total Sunshine Hours (duplicate, may need verification)
10. Minimum Net Width

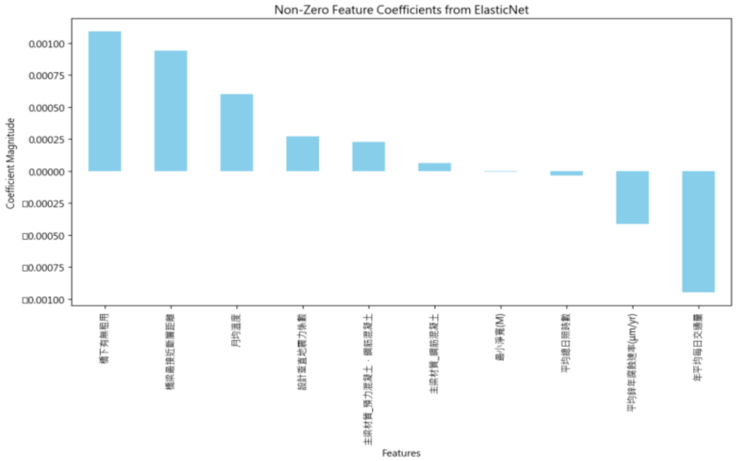


Figure 4: Key Factors identified by Elastic Net

4.1.4 Random Forest

Top variables identified:

1. Nearest Fault Distance
2. Distance from Coastline
3. Total Bridge Length
4. Design Vertical Seismic Coefficient
5. Annual Average Daily Traffic Volume
6. Maximum Net Width
7. Under-Bridge Leasing
8. Minimum Net Width
9. Main Girder Material—Reinforced Concrete
10. Design Live Load

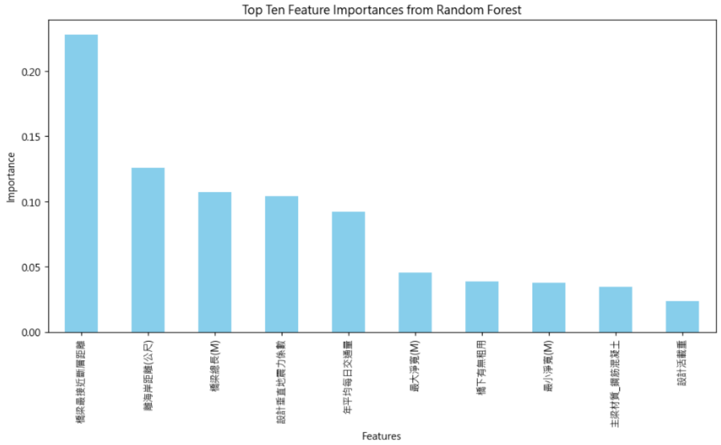


Figure 4: Key Factors identified by Random Forest

4.2 Prediction Results for Rapidly Deteriorating Bridges

Weighted Score = (Recall × 0.5) + (Accuracy × 0.5)

Model	Recall	Accuracy	Weighted Score
Logistic Regression	0.75	0.8095	0.78
Random Forest	0.875	0.8810	0.88
XGBoost	0.75	0.8631	0.81
SVM	0.625	0.8036	0.71
Logistic Regression (PCA)	1	0.7381	0.87
Random Forest (PCA)	0.75	0.8988	0.82
XGBoost (PCA)	1	0.8274	0.91
SVM (PCA)	1	0.7143	0.86
LSTM	0.5	0.78	0.64

Table 7: PCA Training Results

4.2.1 Findings:

- Models trained and predicted using 14 variables selected via PCA yielded more accurate results compared to feature selection using Random Forest.
- Deep learning methods performed poorly, possibly due to insufficient sample size leading to overfitting, or because we focused solely on improving recall scores during training. Additionally, LSTM models are better suited for time-series data. Although deterioration rates consider time, most variables are not time-dependent, leading to poorer predictions.
- XGBoost under PCA and Random Forest under feature selection were the best-performing models, accurately predicting bridges with potential rapid deterioration in the test set.

4.2.2 Final Model Selection:

We selected **Logistic Regression, Random Forest, and XGBoost** for our web platform to predict potential rapid bridge deterioration, using a voting method where the majority prediction determines the output. Reasons for selecting these models include:

- **User-Friendly Variable Input:** Using variables selected through feature selection reduces the number of inputs required by staff, enhancing usability.
- **Interpretability:** Helps staff understand which variables significantly affect bridge deterioration.
- **Avoiding Bias:** The voting method mitigates one-sided predictions.
- **Performance:** These models had the highest weighted scores after feature selection.

5. Conclusions

5.1 Key Factors

Factors Identified by Professionals During Interviews:

- **Average Annual Precipitation:** Influenced by the northeast monsoon in northern regions.
- **Annual Average Daily Traffic Volume:** Higher volumes potentially lead to faster deterioration.
- **Main Girder Material—Reinforced Concrete:** More prone to deterioration.

Factors Identified in Our Research:

- Under-Bridge Leasing
- Design Vertical Seismic Coefficient
- Interaction Between Annual Average Daily Traffic Volume and Horizontal Seismic Coefficient
- Monthly Average Temperature
- Average Total Sunshine Hours
- Main Girder Material—Reinforced Concrete

5.2 Comparison of Prediction Results with Key Monitored Bridges

To assess alignment with the engineering bureau's key monitored bridges, we checked whether the following bridges were predicted as "rapidly deteriorating" by our models:

- **Yuanshan Bridge:** 1 vote
- **Wuyang Elevated Northbound:** 0 votes
- **Wuyang Elevated Southbound:** 0 votes
- **Nankan River Bridge:** 0 votes

Outcome: The results were inconsistent with the bureau's key monitored bridges. This suggests that the criteria used by professionals (domain knowledge) are not fully captured in our data.

6. Discussion

- **Variability in Inspection Assessments:** During data preprocessing and deterioration rate calculations, we found discrepancies due to different inspectors assessing the deterioration degree of specific components differently, affecting the calculated deterioration rates.

- **Outdated Weighting for U Value:** The weighting used for the U value is currently outdated. However, we believe that using a weighted average is more convincing than a simple average of all components.
- **Scope of Data:** Using data from bridges across Taiwan might result in more accurate predictions.

Reference:

https://www.kaggle.com/code/marcinrutecki/best-techniques-and-metrics-for-imbalanced-dataset