ISSS602 – Data Analytics Lab Group Project

Impact of Imbalanced Data on Landslide Susceptibility Prediction

Fong Bao Xian, Loh Jiahui, Sherinah Binte Rashid, Singapore Management University

ABSTRACT

One common encounter in landslide susceptibility prediction is the lack of landslide samples to train the models. The main objective of this study is to investigate the impact of imbalanced data on landslide susceptibility prediction and compare the performance of models using imbalanced (original) and balanced data. Terrain information were obtained from samples of land with landslides and without landslides. Using exploratory data analysis, the characteristics of the variables in landslide and non-landslide cells in relation to their surrounding cells were identified and new independent variables were created to augment the existing dataset. Statistical learning method like logistic regression and recursive partitioning approaches including Decision Tree, Bootstrap Forest and Boosted Tree were used for landslide classification. Then, synthetic minority oversampling technique (SMOTE) was applied to expand the quantity of landslide samples and the same models were ran again. Results indicated that across all models, the usage of balanced data and increase in minority samples have led to improved outcomes, with true positive rates increasing from around 50% or less, to over 80% in all models. Recursive partitioning approaches like Bootstrap Forest and Boosted Tree generally performed better compared to logistic regression, giving higher true positive rates and a balance of performance among other evaluation metrics.

INTRODUCTION

According to the World Health Organisation (2018), landslides occur more frequently than any other geological event, and can happen anywhere in the world. Between 1998 and 2017, landslides caused 18,000 fatalities, and affected an estimated 4.8 million people worldwide. In Italy, Austria, Switzerland and France, the mean annual costs of landslides were estimated between USA 1 to 5 billion for each country (Strumpf & Kerle, 2011). With growing occurrences, landslide identification plays a significant role in landslide risk assessment and management (Wang et al., 2019).

The use of statistically based models and machine learning techniques to understand landslide susceptibility is not an uncommon practice. A meta-analysis conducted by Korup and Stolle (2014) across 674 scientific papers published, found that most machine learning techniques achieved overall success rates of 75 to 95 percent and added that logistic regression was the most commonly adopted approach (33 percent). This was followed by Artificial Neural Networks (31 percent) and Frequency Ration Models (18 percent).

Despite preference and high utility for certain approaches, there is no agreed upon best method for empirical susceptibility modelling (Goetz et al., 2015). A point of interest that came up in recent literature, however, was the issue of class imbalance in landslide susceptibility data. Studies have shown that a balanced dataset improves overall classification performance compared to an imbalanced dataset in several classifier algorithms. While this does not imply classifiers cannot learn from imbalanced data, the application of sampling methods does indeed aid in improved classifier accuracy for most imbalanced datasets (Haibo & Garcia, 2009). Specific to landslide susceptibility, in a study conducted by Stumpf and Kerle (2011), test runs using Random Forests with naturally imbalanced training sets resulted in serious underestimation of the landslide class. Such biases were undesirable as an over or underestimation of affected areas would lead to over or underestimation of the associated risks. However, treating imbalanced samples is not commonly practiced in the field.

In our study, we will be applying synthetic minority oversampling technique (SMOTE) to expand the quantity of landslide samples and doing comparisons of the results pre- and post-SMOTE. Proposed by Chawla in 2022, SMOTE is an oversampling technique that performs k nearest neighbours on the minority class and interpolates between them to generate new data observations. This method has its advantages over random undersampling, which may throw out potentially useful data; and random oversampling, which may be susceptible to overfitting since it simply replicates existing examples in the minority class (Singh & Sharma, 2019). SMOTE is also generalised to handle datasets with both continuous and nominal features, which will be appropriate for the features available in our study. Several other landslide susceptibility studies (Wang et al., 2018 and Gao et al., 2020) have also used SMOTE method to augment the minority class samples and reported favourable prediction results.

DATA

This project will utilise the dataset obtained from the Landslide Prevention and Innovation Challenge on Zindi Africa Platform that is provided by the Hong Kong University of Science and Technology (2022). The dataset contains information on terrain information taken from plots of land samples. Each sample is composed of data from 25 cells, covering an area of 625 m^2. Each cell represents an area of 5 x 5 m^2. For cases with a landslide, the middle cell, cell 13, is the location of the landslide. Cell orientation and independent variables available in the dataset are presented in the figure and table below.

Figure 1. Dataset Cell ID allocation

1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

Table 1. Independent variables for landslide identification

Feature name	Data type	Description
CELLID_elevation	Continuous	Digital elevation of the terrain surface in meter
CELLID_slope	Continuous	Angle of the slope inclination in degree
CELLID_aspect	Continuous	Exposition of the slope in degree
CELLID_placurv	Continuous	Planform curvature, curvature perpendicular to the direction of the maximum slope
CELLID_procurv	Continuous	Profile curvature, curvature parallel to the slope, indicating the direction of maximum slope
CELLID_lsfactor	Continuous	Length-slope factor that accounts for the effects of topography on erosion
CELLID_twi	Continuous	Topographic wetness index, an index to quantify the topographic control on hydrological process
CELLID_geology	Categorical	Lithology of the surface material 1: Weathered Cretaceous granitic rocks 2: Weathered Jurassic granite rocks 3: Weathered Jurassic tuff and lava 4: Weathered Cretaceous tuff and lava 5: Quaternary deposits 6: Fill 7: Weathered Jurassic sandstone, siltstone and mudstone
CELLID_sdoif	Continuous	Step duration orographic intensification factor: an index to quantify the amplification of orography on rainfall
Label	Categorical	1: Landslide 0: Non-landslide

DATA PREPARATION

Heatmaps of Predictors

Prior to data preparation, heatmaps of the continuous predictions for the landslide and non-landslide cases were developed to better visualize and understand the average values of the factors across the 25 cells.

Comparing Figures 2 and 3, the heatmaps for selected factors differed for landslide vis-à-vis non-landslide cases. In landslide cases for the elevation factor, the top two rows had higher values of up to 246m compared to the bottom two rows, whereas for the non-landslide cases, the bottom two rows of cells had higher elevation of up to 218m. This pattern was repeated for planform and profile curvature – the landslide cases showed a distinct pattern where the first two rows of cells had the highest values vs the bottom two rows which had the lowest values. Contrarily, the non-landslide cases had the highest values to the left of the plot of land, and the lowest values to the right. The range of values for these three factors were also wider for landslide cases compared to non-landslide cases.

For TWI, the lowest value of 4 was at cell 13 for landslide cases, while the lowest TWI values for non-landslide cases were more dispersed and higher in magnitude, at 50.10, and was located at cell 5. Comparing both figures for the slope factor, cells 8, 9, 13, and 14 had the highest values of up to 36 for landslide cases, while it was cells 17 and 18 for the non-landslide cases, with the highest values of only up to 26.9. Similarly for the length slope factor, the same cells had the highest values for the landslide cases of up to 10, while the non-landslide cases had the highest value in cell 23, and only up to 7.4.

For slope and length slope factor, the landslides cases displayed a pattern where higher values were concentrated in the middle cell and its immediate neighbours, whereas the dispersion of values in non-landslide cases were more random.

Figure 2. Heatmaps of continuous predictors (Landslide)

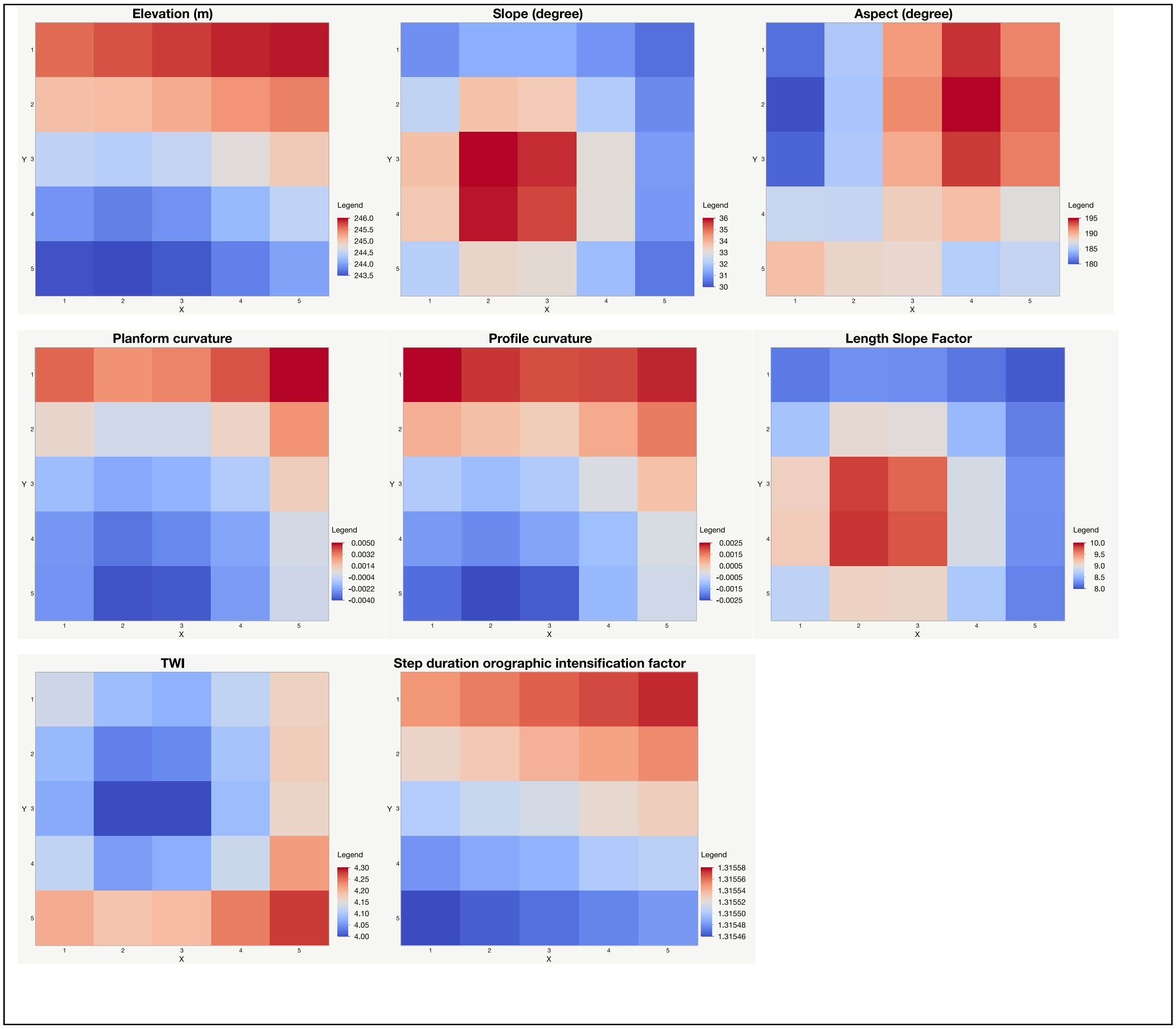
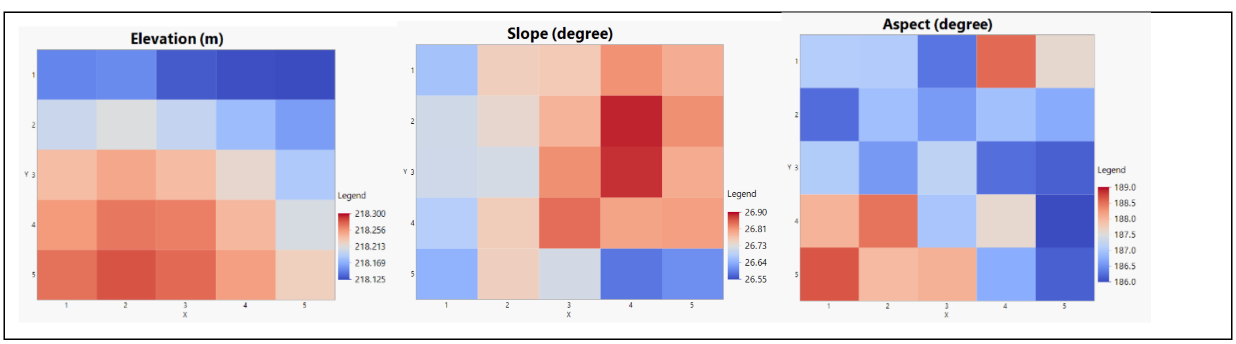
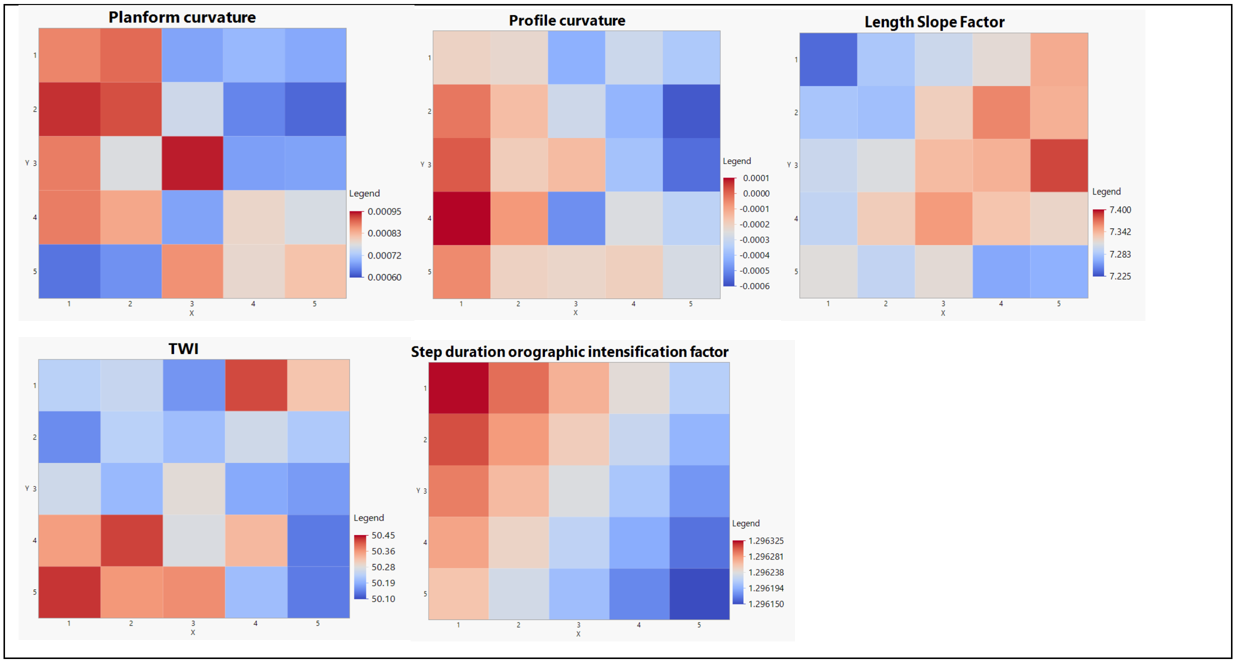


Figure 3. Heatmaps of continuous predictors (Non-Landslide)

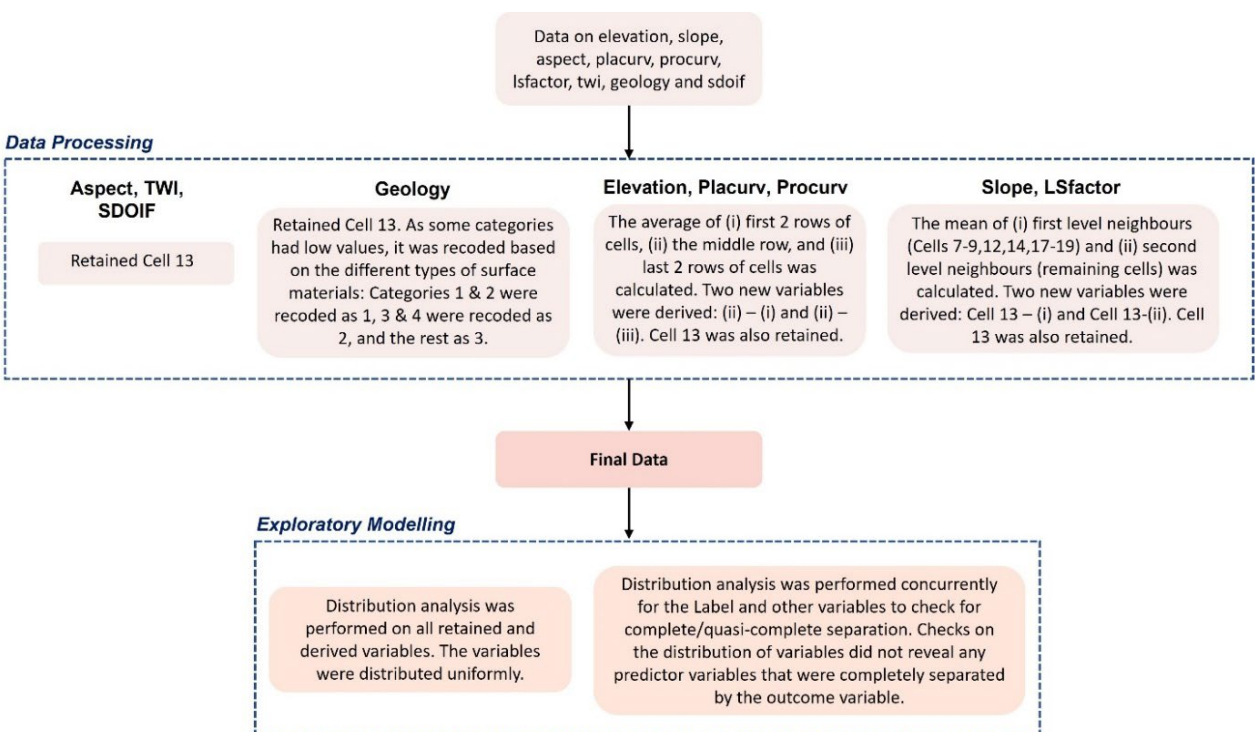




Data Preparation

As cell 13 is the location of the landslide, data for cell 13 was retained for all factors. In line with the interpretation of the heat maps, new features were also created to capture the differences and patterns between cell 13 and its neighbours. The data preparation process is summarized in Figure 4 below.

Figure 4. Data Preparation Process



ANALYTICAL APPROACH

There is no consensus on an optimal machine learning (ML) algorithm as the performance and predictive ability of ML models rely on many factors such as the fundamental quality of the algorithms and the quality of the landslide inventory. As such, literature suggests choosing the best-performing ML model after building several (Liu et al., 2021, Merghadi et al., 2020). Since landslide prediction is a classification problem with the binary outcome of presence/absence of a landslide, and training data has been provided, we will build 4 models using supervised ML classification algorithms. The 4 methods are summarised in the table below (Merghadi et al., 2020, Wang et al., 2020, Hurley, 2012).

Table 2. Summary of ML methods Utlised

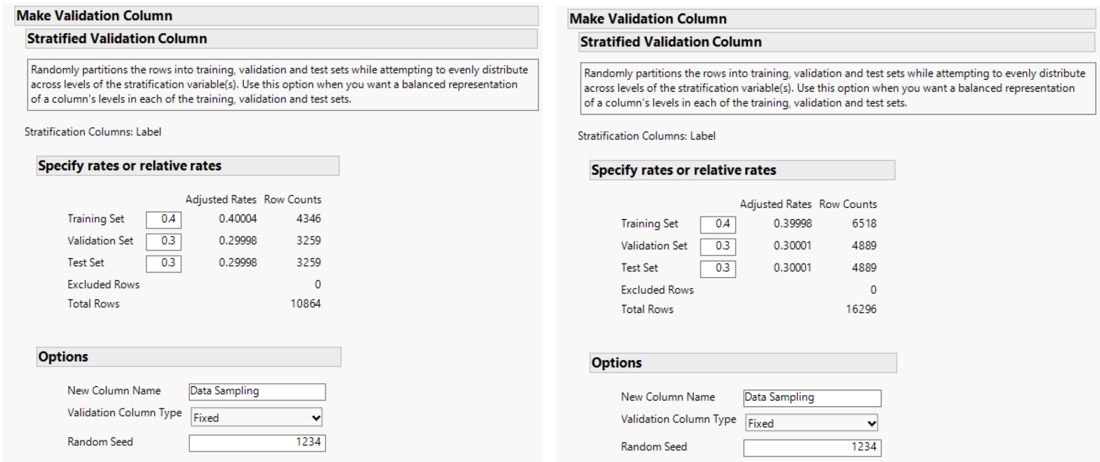
Models	Goal of ML Method	Strengths	Limitations
Logistic Regression	Predicts the probability of the occurrence of an event.	Has been shown to have high accuracy in predicting landslides.	Assumptions to be met: - Little or no multicollinearity between factors - Dependent variable has to be in binary form - Large sample size
Decision Tree	Splits the data based on independent factors in the input dataset and generates decision nodes inferring a predictor value.	Performance may sometimes be superior when compared to linear models such as logistic regression.	It may return a biased solution if one class label dominates the dataset, thus it is necessary to use a balanced dataset.
Boosted Tree	A sequence of weak learners (such as decision tree) is fitted to weighted versions of the training data.	Combines the performance of a number of weak classifiers to produce a powerful “committee”, so it is regarded as a strong classifier.	It may overfit the data and thus predictions on new data may not be accurate.
Bootstrap Forest	Utilises multiple decision tree type classification models to determine an optimal model.	The resulting model is usually more powerful than the initial decision tree, and reduces overfitting and helps improve accuracy.	It often produces the best results when there is tuning of hyper-parameters, such as the number of trees to be combined, or the maximum number of features considered at each split.

ANALYSIS PROCESS AND RESULTS

Sampling for data validation – Original and SMOTE dataset

After preparation of the predictors, the sample was split into 3 sub-groups for training (40%), validation (30%) and testing (30%). To address the main objective of the study, a separate sample was also prepared using SMOTE techniques to address the issues of class imbalance, using the same split in proportion. Consequently, 5432 additional rows of observations with *label* of value 1, i.e., landslide observations were generated. The figures below show the parameters used for the sampling for both Original and SMOTE dataset.

Figure 5. Sampling parameters for Original Sample (left) and SMOTE Sample (right)



Logistic Regression Approach

While logistic regression is not limited by several key assumptions of linear regression and general linear models that are based on ordinary least squares algorithms – linearity, normality, homoscedasticity, and measurement level, it still shares some assumptions with linear regression. One main assumptions of logistic regression is for there to be little or no multicollinearity among the independent variables (Schreiber-Gregory & Bader, 2018). Hence, before performing logistic regression, we first checked for multicollinearity among the continuous independent variables.

Multivariate Analysis of variables

Among the 19 independent variables, two pairs were strongly correlated with correlation values above 0.8. Hence, we remove one variable from each pair, specifically *middle_second_layer_diff_slope* and *middle_second_layer_diff_ls and* retained 17 variables for logistic regression.

Figure 6. Pairwise Correlations sorted in descending order of Correlation (Top 5)

Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
middle_second_layer_diff_slope	middle_first_layer_diff_slope	0.9058	10864	0.9023	0.9091	<.0001*	
middle_second_layer_diff_ls	middle_first_layer_diff_ls	0.8691	10864	0.8645	0.8737	<.0001*	
13_lsfactor	13_slope	0.7905	10864	0.7833	0.7974	<.0001*	
middle_second_layer_diff_slope	middle_second_layer_diff_ls	0.6874	10864	0.6773	0.6972	<.0001*	
middle_second_layer_diff_slope	13_slope	0.6708	10864	0.6603	0.6810	<.0001*	

Logistic Regression – Original sample

For logistic regression using the original data, the Whole Model Test shows p-value < 0.0001, lower than significance level of 0.05. Hence, we can reject the null hypothesis and conclude that the logistic model is useful to explain the *label* (landslide or no landslide).

Figure 7. Whole Model Test for Logistic Regression – Original Sample

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	625.7424	18	1251.485	<.0001*
Full	1817.6167			
Reduced	2443.3591			

RSquare (U)	0.2561
AICc	3673.41
BIC	3794.4
Observations (or Sum Wgts)	4346

For the Lack of Fit (Goodness of Fit) test, Prob>Chisq is 1 and we do not reject the null hypothesis at significance level of 0.05. This supports the conclusion that the model is adequate and there is little to be gained by introducing additional variables.

Figure 8. Lack of Fit Test for Logistic Regression – Original Sample

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4321	1817.6167	3635.233
Saturated	4339	0.0000	**Prob>ChiSq**
Fitted	18	1817.6167	1.0000

Both the Effect Likelihood Ratio Tests and Parameter Estimates show the same 12 independent variables as significant given that Prob>ChiSq of these 12 variables are less than significance level of 0.05. Five variables, including the middle cell for length-slope factor and the newly-created variables for planform curvature and profile curvature were found to be not significant.

Among the significant variables, the middle cell (i.e., cell 13) for step duration orographic intensification factor (sdoif), planform curvature (placurv) and profile curvature (procurv) were the strongest indicators. Specifically, if a specific plot of land cell had higher values of sdoif or procurv; or lower value of placurv, it had a higher landslide susceptibility.

Figure 9. Parameter Estimates (left) and Effect Likelihood Ratio Tests (right) for Logistic Regression – Original Sample

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
13_geology 2	2	2	249.657016	<.0001*
13_elevation	1	1	106.344569	<.0001*
13_sdoif	1	1	75.9969025	<.0001*
13_twi	1	1	51.9702884	<.0001*
middle_first_layer_diff_slope	1	1	43.5232645	<.0001*
13_slope	1	1	37.4255029	<.0001*
middle_first_layer_diff_ls	1	1	36.5104448	<.0001*
middle_top_diff_elevation	1	1	35.1865417	<.0001*
13_placurv	1	1	33.6771537	<.0001*
middle_bottom_diff_elevation	1	1	13.4173891	0.0002*
13_aspect	1	1	4.1973646	0.0405*
13_procurv	1	1	4.05065535	0.0442*
13_lsfactor	1	1	1.78115175	0.1820
middle_top_diff_procurve	1	1	1.03044383	0.3101
middle_bottom_diff_placurve	1	1	0.94938538	0.3299
middle_top_diff_placurve	1	1	0.61242623	0.4339
middle_bottom_diff_procurve	1	1	0.01158183	0.9143

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
13_sdoif	10.4500744	1.3023543	64.38	<.0001*
13_procurv	5.89583935	2.9456749	4.01	0.0453*
middle_bottom_diff_placurve	4.22162618	4.331776	0.95	0.3298
13_geology 2[2]	0.98177385	0.0688279	203.47	<.0001*
middle_first_layer_diff_ls	0.27385354	0.0462203	35.11	<.0001*
13_slope	0.10766614	0.0167489	41.32	<.0001*
13_aspect	0.00089602	0.0004385	4.18	0.0410*
13_elevation	-0.003142	0.0003175	97.90	<.0001*
13_lsfactor	-0.0615323	0.045236	1.85	0.1738
middle_bottom_diff_elevation	-0.1197211	0.0328009	13.32	0.0003*
middle_first_layer_diff_slope	-0.1276769	0.0196262	42.32	<.0001*
middle_top_diff_elevation	-0.2000799	0.0340643	34.50	<.0001*
middle_bottom_diff_procurve	-0.4862333	4.5184466	0.01	0.9143
13_twi	-0.4972746	0.083024	35.87	<.0001*
13_geology 2[1]	-0.7672033	0.0892296	73.93	<.0001*
middle_top_diff_placurve	-3.4328945	4.3889165	0.61	0.4341
middle_top_diff_procurve	-4.5195066	4.4597589	1.03	0.3109
Intercept	-15.402907	1.7673598	75.95	<.0001*
13_placurv	-16.633313	2.8967521	32.97	<.0001*

For log odds of 1/0

Looking at results on the test dataset, a low true positive rate of 44%, and a misclassification rate of approximately 20% was observed.

Figure 10. Summary results for Logistic Regression – Original Sample

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2561	0.2509	0.2580	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3706	0.3640	0.3731	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4182	0.4211	0.4176	∑ -Log(ρ[j])/n
RASE	0.3678	0.3693	0.3685	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2704	0.2724	0.2714	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1993	0.2001	0.1979	∑ (ρ[j]≠ρMax)/n
N	4346	3259	3259	n

Confusion Matrix

Training

Actual Label	Predicted Count 1	Predicted Count 0
1	468	618
0	248	3012

Actual Label	Predicted Rate 1	Predicted Rate 0
1	0.431	0.569
0	0.076	0.924

Validation

Actual Label	Predicted Count 1	Predicted Count 0
1	367	447
0	205	2240

Actual Label	Predicted Rate 1	Predicted Rate 0
1	0.451	0.549
0	0.084	0.916

Test

Actual Label	Predicted Count 1	Predicted Count 0
1	360	456
0	189	2254

Actual Label	Predicted Rate 1	Predicted Rate 0
1	0.441	0.559
0	0.077	0.923

Logistic Regression – SMOTE sample

Performing logistic regression on the SMOTE-treated data, the Whole Model Test shows p-value < 0.0001. Hence, we can reject the null hypothesis at significance level of 0.05 and conclude that this logistic model is useful to explain the *label* (landslide or no landslide).

Figure 11. Whole Model Test for Logistic Regression – SMOTE Sample

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	1353.0122	18	2706.024	<.0001*
Full	3164.9199			
Reduced	4517.9321			

RSquare (U)	0.2995
AICc	6367.96
BIC	6496.7
Observations (or Sum Wgts)	6518

Lack of Fit test similarly shows Prob>Chisq less than significance level of 0.05. Hence, we can conclude that the model is adequate and there is little to be gained by introducing additional variables.

Figure 12. Lack of Fit Test for Logistic Regression – SMOTE Sample

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	6489	3164.9199	6329.84
Saturated	6507	0.0000	Prob>ChiSq
Fitted	18	3164.9199	0.9197

Compared with the original sample with 12 significant independent variables, logistic regression on the SMOTE sample has 13 significant independent variables as shown by the Parameter Estimates and Effect Likelihood Ration Tests. Four variables, including the middle cell for aspect and most of the newly-created variables for planform curvature and profile curvature were found to be not significant.

The strongest indicators for this model were similarly the middle cell for sdoif, placurv and procurv. Higher values of sdoif or procurv; or lower value of placurv, indicated a higher likelihood of the land cell having landslide.

Figure 13. Parameter Estimates (left) and Effect Likelihood Ratio Tests (right) for Logistic Regression – SMOTE Sample

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
13_geology 2	2	2	398.524177	<.0001*
13_elevation	1	1	176.201798	<.0001*
middle_first_layer_diff_slope	1	1	132.053291	<.0001*
13_sdoif	1	1	129.239322	<.0001*
13_slope	1	1	116.520764	<.0001*
13_twi	1	1	97.855444	<.0001*
middle_first_layer_diff_ls	1	1	63.3574095	<.0001*
13_placurv	1	1	52.5826603	<.0001*
middle_top_diff_elevation	1	1	47.8787706	<.0001*
middle_bottom_diff_elevation	1	1	17.497825	<.0001*
middle_top_diff_placurve	1	1	7.77173594	0.0053*
13_procurv	1	1	5.29022939	0.0214*
13_lsfactor	1	1	4.96599825	0.0259*
13_aspect	1	1	3.29768689	0.0694
middle_top_diff_procurve	1	1	1.47358322	0.2248
middle_bottom_diff_placurve	1	1	0.04352958	0.8347
middle_bottom_diff_procurve	1	1	0.01002845	0.9202

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
13_sdoif	9.52483649	0.8961531	112.97	<.0001*
13_procurv	5.51180384	2.4042358	5.26	0.0219*
13_geology 2[2]	0.9045612	0.0498685	329.02	<.0001*
middle_first_layer_diff_ls	0.30411769	0.0387412	61.62	<.0001*
13_slope	0.14267494	0.0126363	127.48	<.0001*
13_aspect	0.00059032	0.0003251	3.30	0.0694
13_elevation	-0.0032485	0.0002519	166.35	<.0001*
13_lsfactor	-0.0755328	0.033306	5.14	0.0233*
middle_bottom_diff_elevation	-0.1125077	0.0270103	17.35	<.0001*
middle_first_layer_diff_slope	-0.1820306	0.0160283	128.98	<.0001*
middle_top_diff_elevation	-0.1931957	0.0281335	47.16	<.0001*
middle_bottom_diff_procurve	-0.3577074	3.5717322	0.01	0.9202
13_twi	-0.5316051	0.0603205	77.67	<.0001*
middle_bottom_diff_placurve	-0.7293011	3.4953929	0.04	0.8347
13_geology 2[1]	-0.7773112	0.0627789	153.31	<.0001*
middle_top_diff_procurve	-4.3727559	3.6022723	1.47	0.2248
middle_top_diff_placurve	-9.7784371	3.515762	7.74	0.0054*
Intercept	-13.798862	1.2224	127.43	<.0001*
13_placurv	-16.277292	2.2647878	51.65	<.0001*

For log odds of 1/0

True positive rate for the test dataset also improved significantly from 44% to 80%. The misclassification rate, however, increased slightly to 22%.

Figure 14. Summary results for Logistic Regression – SMOTE Sample

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2995	0.2865	0.2966	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4530	0.4370	0.4495	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4856	0.4946	0.4876	∑ -Log(ρ[j])/n
RASE	0.3968	0.4005	0.3973	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3175	0.3220	0.3205	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2252	0.2252	0.2209	∑ (ρ[j]≠ρMax)/n
N	6518	4889	4889	n

Confusion Matrix

	Training		Validation		Test	
Actual Label	Predicted Count 1	0	Predicted Count 1	0	Predicted Count 1	0
1	2606	651	1990	457	1958	486
0	817	2444	644	1798	594	1851

	Training		Validation		Test	
Actual Label	Predicted Rate 1	0	Predicted Rate 1	0	Predicted Rate 1	0
1	0.800	0.200	0.813	0.187	0.801	0.199
0	0.251	0.749	0.264	0.736	0.243	0.757

Recursive Partitioning Approaches

As prediction techniques under this approach are nonparametric, these methods do not rely on any assumption about the type of dependence of the dependent variable on the predictors (Landau & Barthel, 2010). Therefore, all 19 independent variables were included in the analysis for all six models. To ensure reproducibility of the data, a seed of '1234' was set for both the Bootstrap Forest and Boosted Tree analyses.

Decision Tree – Original sample

For Decision Tree using the original data, the resultant model with 14 splits yielded 15 terminal leaf nodes with a response count ranging from 10 to 274 for landslide observed rows and 25 to 1131 for non-landslide rows. Eight variables were identified to have contributed to the model, with the middle cells (i.e., cell 13) for slope, geology and elevation being identified as the top 3 indicators for landslide susceptibility. Looking at results on the test dataset, a low true positive rate of 36%, and a misclassification rate of approximately 20% was observed.

Figure 15. Column Contributions for Decision Tree – Original Sample

Column Contributions

Term	Number of Splits	G^2	Portion
13_slope	4	669.983709	0.5101
13_geology 2	2	257.467785	0.1960
13_elevation	2	123.119388	0.0937
middle_second_layer_diff_ls	1	94.3830903	0.0719
middle_top_diff_elevation	2	56.7809542	0.0432
13_twi	1	45.1987828	0.0344
middle_bottom_diff_elevation	1	35.9432772	0.0274
13_sdoif	1	30.493046	0.0232
13_lsfactor	0	0	0.0000
middle_first_layer_diff_ls	0	0	0.0000
13_procurv	0	0	0.0000
middle_top_diff_procurve	0	0	0.0000
middle_bottom_diff_procurve	0	0	0.0000
13_placurv	0	0	0.0000
middle_top_diff_placurve	0	0	0.0000
middle_bottom_diff_placurve	0	0	0.0000
13_aspect	0	0	0.0000
middle_first_layer_diff_slope	0	0	0.0000
middle_second_layer_diff_slope	0	0	0.0000

Figure 16. Summary results for Decision Tree – Original Sample

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2688	0.2535	0.2475	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3863	0.3673	0.3599	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4111	0.4196	0.4235	∑ -Log(ρ[j])/n
RASE	0.3634	0.3681	0.3709	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2643	0.2697	0.2663	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1967	0.1976	0.1991	∑ (ρ[j]≠ρMax)/n
N	4346	3259	3259	n

Confusion Matrix

Training

Actual Label	Predicted Count 0	1
0	3113	147
1	708	378

Actual Label	Predicted Rate 0	1
0	0.955	0.045
1	0.652	0.348

Validation

Actual Label	Predicted Count 0	1
0	2317	128
1	516	298

Actual Label	Predicted Rate 0	1
0	0.948	0.052
1	0.634	0.366

Test

Actual Label	Predicted Count 0	1
0	2319	124
1	525	291

Actual Label	Predicted Rate 0	1
0	0.949	0.051
1	0.643	0.357

Decision Tree – SMOTE Sample

Using the SMOTE-treated data, the resultant decision tree model with 13 splits yielded 14 terminal leaf nodes with a response count ranging from 3 to 1444 for landslide observed rows and 45 to 878 for non-landslide rows. Six variables were identified to have contributed to the model, with the middle cells (i.e., cell 13) for slope, geology and sdoif being identified as the top 3 indicators for landslide susceptibility. Looking at results on the test dataset, the true positive rate noted significant improvement from 36% to close to 90%. The misclassification rate, however, remained stable at 22%.

Figure 17. Column Contributions for Decision Tree – SMOTE Sample

Column Contributions

Term	Number of Splits	G^2	Portion
13_slope	3	1806.65772	0.6345
13_geology 2	3	379.110951	0.1331
13_sdoif	2	232.431151	0.0816
13_elevation	3	231.457778	0.0813
middle_second_layer_diff_ls	1	150.265143	0.0528
middle_bottom_diff_elevation	1	47.5670765	0.0167
13_lsfactor	0	0	0.0000
middle_first_layer_diff_ls	0	0	0.0000
13_procurv	0	0	0.0000
middle_top_diff_procurve	0	0	0.0000
middle_bottom_diff_procurve	0	0	0.0000
13_placurv	0	0	0.0000
middle_top_diff_placurve	0	0	0.0000
middle_bottom_diff_placurve	0	0	0.0000
13_twi	0	0	0.0000
13_aspect	0	0	0.0000
middle_first_layer_diff_slope	0	0	0.0000
middle_second_layer_diff_slope	0	0	0.0000
middle_top_diff_elevation	0	0	0.0000

Figure 18. Summary results for Decision Tree – SMOTE Sample

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3151	0.3038	0.3016	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4719	0.4583	0.4557	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4747	0.4826	0.4841	∑ -Log(ρ[j])/n
RASE	0.3923	0.3961	0.3966	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3082	0.3135	0.3134	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2195	0.2277	0.2232	∑ (ρ[j]≠ρMax)/n
N	6518	4889	4889	n

Confusion Matrix

Training

Actual Label	Predicted Count 0	1
0	2149	1112
1	319	2938

Actual Label	Predicted Rate 0	1
0	0.659	0.341
1	0.098	0.902

Validation

Actual Label	Predicted Count 0	1
0	1553	889
1	224	2223

Actual Label	Predicted Rate 0	1
0	0.636	0.364
1	0.092	0.908

Test

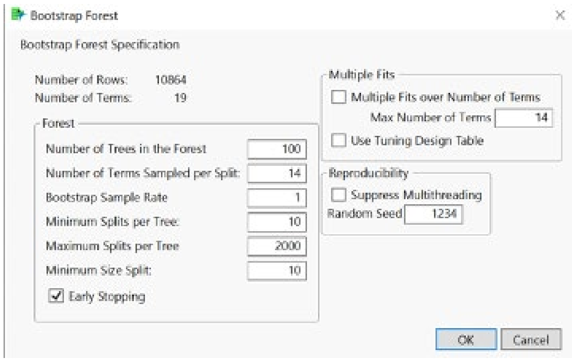
Actual Label	Predicted Count 0	1
0	1605	840
1	251	2193

Actual Label	Predicted Rate 0	1
0	0.656	0.344
1	0.103	0.897

Bootstrap Forest – Original Sample

Bootstrap Forest approach was also conducted on the same 19 variables. Figure 19 shows the parameters used for the analysis.

Figure 19. Bootstrap Forest analysis parameters



The resultant model had 62 trees, 19 terms, and 14 terms sampled per split. The true positive rate of the bootstrap forest model test set was 51%, an improvement over 36%, which was the test accuracy of the decision tree model. The 3 most important predictors were the middle cell (i.e., cell 13) for slope, sdoif and elevation.

Figure 20. Summary results for Bootstrap Forest – Original Sample

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5819	0.3288	0.3431	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7113	0.4577	0.4743	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2350	0.3773	0.3697	∑ -Log(ρ[j])/n
RASE	0.2552	0.3455	0.3451	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1851	0.2521	0.2493	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0716	0.1675	0.1752	∑ (ρ[j]≠ρMax)/n
N	4346	3259	3259	n

Confusion Matrix

Training

Actual Label	Predicted Count 0	Predicted Count 1
0	3215	45
1	266	820

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.986	0.014
1	0.245	0.755

Validation

Actual Label	Predicted Count 0	Predicted Count 1
0	2286	159
1	387	427

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.935	0.065
1	0.475	0.525

Test

Actual Label	Predicted Count 0	Predicted Count 1
0	2269	174
1	397	419

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.929	0.071
1	0.487	0.513

Figure 21. Column Contributions for Bootstrap Forest – Original Sample

Column Contributions

Term	Number of Splits	G^2	Portion
13_slope	913	415.620961	0.2051
13_sdoif	862	168.05266	0.0829
13_elevation	882	163.641637	0.0808
13_geology 2	478	153.262822	0.0756
13_lsfactor	595	120.645178	0.0595
13_twi	691	110.299345	0.0544
middle_second_layer_diff_ls	611	100.834486	0.0498
middle_bottom_diff_elevation	666	97.0388873	0.0479
middle_second_layer_diff_slope	467	83.7009583	0.0413
middle_top_diff_elevation	600	76.5916489	0.0378
13_placurv	591	73.3142415	0.0362
middle_top_diff_placurve	537	66.2129702	0.0327
13_aspect	673	61.912127	0.0306
13_procurv	541	57.6606254	0.0285
middle_top_diff_procurve	540	57.3152784	0.0283
middle_first_layer_diff_ls	527	56.3395064	0.0278
middle_first_layer_diff_slope	466	56.0661788	0.0277
middle_bottom_diff_placurve	481	54.8622217	0.0271
middle_bottom_diff_procurve	499	52.9559326	0.0261

Bootstrap Forest – SMOTE Sample

Using the SMOTE treated data, Bootstrap Forest approach was also conducted on the same 19 variables using the same analysis parameters. The resultant model had 100 trees, 19 terms, and 14 terms sampled per split. The true positive rate of the bootstrap forest model was 86%, a significant improvement over 51%, which was the test accuracy of the same model using the original data sample. The 3 most important predictors were the middle cell (i.e., cell 13) for slope, elevation and sdoif.

Figure 22. Summary results for Bootstrap Forest – SMOTE Sample

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.6044	0.4391	0.4366	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7565	0.6080	0.6054	$(1-(L(0)/L(model))^{(2/n)})/(1-L(0)^{(2/n)})$
Mean -Log p	0.2742	0.3888	0.3905	$\sum -Log(\rho[j])/n$
RASE	0.2786	0.3509	0.3511	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2141	0.2704	0.2722	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0875	0.1702	0.1751	$\sum (\rho[j]\neq\rho Max)/n$
N	6518	4889	4889	n

Confusion Matrix

Training

Actual Label	Predicted Count 0	Predicted Count 1
0	2906	355
1	215	3042

Validation

Actual Label	Predicted Count 0	Predicted Count 1
0	1920	522
1	310	2137

Test

Actual Label	Predicted Count 0	Predicted Count 1
0	1931	514
1	342	2102

Training

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.891	0.109
1	0.066	0.934

Validation

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.786	0.214
1	0.127	0.873

Test

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.790	0.210
1	0.140	0.860

Figure 23. Column Contributions for Bootstrap Forest – SMOTE Sample

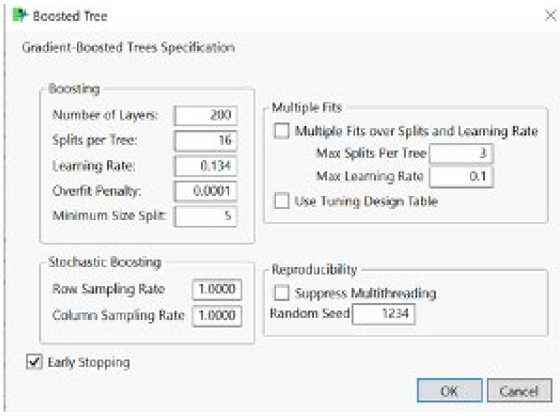
Column Contributions

Term	Number of Splits	G^2	Portion
13_slope	1827	1060.3575	0.2943
13_elevation	1923	317.284026	0.0881
13_sdoif	1720	309.710734	0.0860
13_geology 2	861	271.060829	0.0752
13_lsfactor	1034	218.501858	0.0606
middle_second_layer_diff_ls	1269	195.290267	0.0542
13_twi	1190	173.905903	0.0483
middle_bottom_diff_elevation	1094	115.336122	0.0320
middle_top_diff_elevation	1046	108.970235	0.0302
middle_first_layer_diff_slope	1024	99.2468353	0.0275
middle_second_layer_diff_slope	797	96.0055064	0.0266
13_aspect	1253	90.6720796	0.0252
13_placurv	1027	89.6208799	0.0249
middle_top_diff_placurve	967	86.4599752	0.0240
middle_first_layer_diff_ls	905	78.9931736	0.0219
middle_bottom_diff_placurve	921	78.5827817	0.0218
middle_top_diff_procurve	944	78.3584626	0.0217
13_procurv	950	72.5826354	0.0201
middle_bottom_diff_procurve	813	62.2688904	0.0173

Boosted Tree – Original Sample

Lastly, the same 19 variables were also used to implement a boosted tree analysis. Figure 24 shows the parameters used for the analysis.

Figure 24. Boosted tree analysis parameters



The resultant model had 44 layers and 16 splits per tree. The true positive rate of the boosted tree model was 45%, and a misclassification rate of 19%. All 19 variables contributed to the model, with the middle cell (i.e., cell 13) for geology, slope and elevation identified as the top 3 strongest indicators.

Figure 25. Summary results for Boosted Tree – Original Sample

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4397	0.2658	0.2794	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5777	0.3826	0.3994	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3150	0.4127	0.4055	∑ -Log(ρ[j])/n
RASE	0.3114	0.3599	0.3592	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2236	0.2587	0.2562	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1247	0.1844	0.1942	∑ (ρ[j]≠ρMax)/n
N	4346	3259	3259	n

Confusion Matrix

Training

Actual Label	Predicted Count 0	Predicted Count 1
0	3152	108
1	434	652

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.967	0.033
1	0.400	0.600

Validation

Actual Label	Predicted Count 0	Predicted Count 1
0	2258	187
1	414	400

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.924	0.076
1	0.509	0.491

Test

Actual Label	Predicted Count 0	Predicted Count 1
0	2259	184
1	449	367

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.925	0.075
1	0.550	0.450

Figure 26. Column Contributions for Boosted Tree – Original Sample

Column Contributions

Term	Number of Splits	G^2	Portion
13_geology 2	79	200820.945	0.5771
13_slope	200	56132.6103	0.1613
13_elevation	83	26091.8257	0.0750
13_aspect	67	13155.0841	0.0378
middle_top_diff_elevation	28	10468.4421	0.0301
middle_bottom_diff_elevation	30	10082.3177	0.0290
13_sdoif	25	7707.38145	0.0221
13_twi	28	5386.43168	0.0155
middle_second_layer_diff_ls	15	3849.92531	0.0111
middle_top_diff_placurve	27	2673.12798	0.0077
13_lsfactor	11	2307.50659	0.0066
middle_second_layer_diff_slope	12	1884.82537	0.0054
13_placurv	18	1834.67672	0.0053
middle_top_diff_procurve	13	1373.31219	0.0039
13_procurv	17	1134.29995	0.0033
middle_bottom_diff_procurve	16	1018.68904	0.0029
middle_bottom_diff_placurve	15	848.086628	0.0024
middle_first_layer_diff_slope	12	626.466082	0.0018
middle_first_layer_diff_ls	8	604.956139	0.0017

Boosted Tree – SMOTE Sample

The Boosted Tree approach was also implemented on the SMOTE treated data using the same variables and analysis parameters. The resultant model had 78 layers and 17 splits per tree. The true positive rate of the boosted tree model was 86%, and a misclassification rate of 18%. All 19 variables contributed to the model, however, only the top 2 contributed to more than 10% of the model. The middle cells (i.e., cell 13) for geology, slope and elevation were identified as the top 3 indicators.

Figure 27. Summary results for Boosted Tree – SMOTE Sample

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5289	0.3878	0.3989	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6929	0.5545	0.5663	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3265	0.4243	0.4167	∑ -Log(ρ[j])/n
RASE	0.3153	0.3685	0.3635	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2411	0.2835	0.2801	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1329	0.1933	0.1820	∑ (ρ[j]≠ρMax)/n
N	6518	4889	4889	n

Confusion Matrix

Training

Actual Label	Predicted Count 0	Predicted Count 1
0	2615	646
1	220	3037

Validation

Actual Label	Predicted Count 0	Predicted Count 1
0	1827	615
1	330	2117

Test

Actual Label	Predicted Count 0	Predicted Count 1
0	1887	558
1	332	2112

Training

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.802	0.198
1	0.068	0.932

Validation

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.748	0.252
1	0.135	0.865

Test

Actual Label	Predicted Rate 0	Predicted Rate 1
0	0.772	0.228
1	0.136	0.864

Figure 28. Column Contributions for Boosted Tree – SMOTE Sample

Column Contributions

Term	Number of Splits	G^2	Portion
13_geology 2	148	779768.133	0.7191
13_slope	451	153399.491	0.1415
13_elevation	72	49979.5921	0.0461
13_aspect	255	23992.5975	0.0221
13_sdoif	31	20475.1356	0.0189
13_twi	31	14161.1391	0.0131
middle_top_diff_elevation	38	11196.8462	0.0103
middle_second_layer_diff_ls	23	9657.69282	0.0089
middle_bottom_diff_elevation	42	7032.84354	0.0065
13_placurv	20	2616.16897	0.0024
middle_bottom_diff_placurve	29	2495.82218	0.0023
middle_first_layer_diff_slope	25	1976.80618	0.0018
middle_top_diff_placurve	32	1967.47657	0.0018
middle_second_layer_diff_slope	20	1267.61214	0.0012
middle_first_layer_diff_ls	17	1237.39159	0.0011
13_lsfactor	24	901.965302	0.0008
13_procurv	22	846.021922	0.0008
middle_bottom_diff_procurve	28	784.14739	0.0007
middle_top_diff_procurve	18	610.066177	0.0006

MODEL COMPARISON & EVALUATION

FACTORS OF IMPORTANCE

Each model generated a unique set of variables and varies in contribution proportion to the model. It was observed that the trend for Logistic Regression was distinctive from the recursive partitioning models.

For Logistic Regression, the planform curvature, profile curvature and step duration orographic intensification factor were the variables with highest magnitude of estimates for log odds. These same variables however, featured much lower importance in the recursive partitioning models. Instead, across Decision Tree, Bootstrap Forest and Boosted Tree, slope, elevation and geology were noted as important variables. Notably, new variables that were created to calculate the differences in parameters between the landslide cell (i.e., cell 13) and its neighbouring cells did not perform as well as untreated variables in all models.

Table 3. Summary of variable importance on test datasets

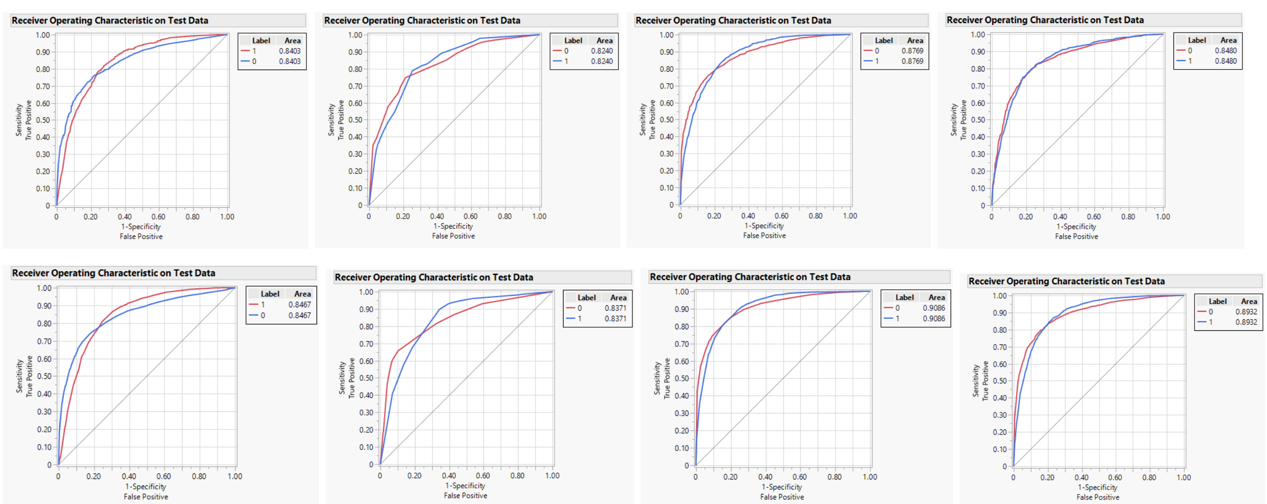
	Logistic Regression		Decision Tree		Bootstrap Forest		Boosted Tree	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
13_elevation	-0.0031	-0.0032	0.0937	0.0813	0.0808	0.0881	0.075	0.0461
middle_top_diff_elevation	-0.2001	-0.1932	0.0432		0.0378	0.0302	0.0301	0.0103
middle_bottom_diff_elevation	-0.1197	-0.1125	0.0274	0.0167	0.0479	0.032	0.029	0.0065
13_slope	0.1077	0.1427	0.5101	0.6345	0.2051	0.2943	0.1613	0.1415
middle_first_layer_diff_slope	-0.1277	-0.1820			0.0277	0.0275	0.0018	0.0018
middle_second_later_diff_slope					0.0413	0.0266	0.0054	0.0012
13_lsfactor		-0.0755			0.0595	0.0606	0.0066	0.0008
middle_first_layer_diff_ls	0.2739	0.3041			0.0278	0.0219	0.0017	0.0011
middle_second_layer_diff_ls			0.0719	0.0528	0.0498	0.0542	0.0111	0.0089
13_procurve	5.8958	5.5118			0.0285	0.0201	0.0033	0.0008
middle_top_diff_procurve					0.0283	0.0217	0.0039	0.0006
middle_bottom_diff_procurve					0.0261	0.0173	0.0029	0.0007
13_placurve	-16.6333	-16.2773			0.0362	0.0249	0.0053	0.0024
middle_top_diff_placurve		-9.778			0.0327	0.024	0.0077	0.0018
middle_bottom_diff_placurve					0.0271	0.0218	0.0024	0.0023
13_geology_2	-0.7672 [1] 0.9818 [2]	-0.7773 [1] 0.9046 [2]	0.196	0.1331	0.0759	0.0752	0.5771	0.7191
13_sdoif	10.4501	9.5248	0.0232	0.0816	0.0829	0.086	0.0221	0.0189
13_aspect	0.0009				0.0306	0.0252	0.0378	0.0221
13_twi	-0.4973	-0.5316	0.0344		0.0544	0.0483	0.0155	0.0131

Note: The heatmap is visualised per data column. Measure used for Logistic Regression are the Parameter Estimates for Log Odds of 1/0. Measure used for Decision Tree, Bootstrap Forest and Boosted Tree are the portion of column contributions.

OVERALL MODEL PERFORMANCE

Comparing the receiver operating characteristic curve (ROC) of the models, the differences among the models were not visibly significant. All models were noted to have good predictive power, with the curve bowing above the diagonal. The areas under the curve, an indicator of how good the classifier performs, were also similar across the models. There were also no major fluctuations in the curve, indicating that the models are stable.

Figure 29. ROC Comparison for all models on Original Sample (top row) and SMOTE Sample (bottom row) (from left to right: Logistic Regression, Decision Tree, Bootstrap Forest, Boosted Tree)



For more detailed comparison among the models, we investigate further into the key performance indicators. The table below summarises the key indicators between the training and test datasets.

Table 4. Summary of Model Performance on Training Datasets

	Logistic Regression		**Decision Tree**		**Bootstrap Forest**		**Boosted Tree**	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
TP	0.431	0.8	0.348	0.659	0.755	0.934	0.6	0.932
TN	0.924	0.749	0.955	0.098	0.986	0.891	0.967	0.802
Accuracy	80%	77%	80%	78%	93%	91%	88%	87%
Misclassification	20%	23%	20%	22%	7%	9%	12%	13%
Precision	43%	80%	72%	73%	95%	90%	86%	82%
Sensitivity	65%	76%	35%	90%	76%	93%	60%	93%
Specificity	83%	79%	95%	66%	99%	90%	97%	80%

Table 5. Summary of Model Performance on Test Datasets

	Logistic Regression		**Decision Tree**		**Bootstrap Forest**		**Boosted Tree**	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
TP	0.441	0.801	0.357	**0.897**	0.513	0.860	0.450	0.864
TN	0.923	0.757	**0.949**	0.656	0.929	0.790	0.925	0.772
Accuracy	80%	78%	80%	78%	**82%**	**82%**	81%	**82%**
Misclassification	20%	22%	20%	22%	**18%**	**18%**	19%	**18%**
Precision	66%	77%	70%	72%	71%	80%	67%	**86%**
Sensitivity	44%	80%	36%	**90%**	51%	86%	45%	79%
Specificity	92%	76%	95%	66%	93%	79%	92%	**85%**

OVERFITTING ASSESSMENT

The misclassification rate between training and test datasets was used as an indicator to assess if there are signs of overfitting across the eight models. For both models implemented using logistic regression and decision tree, misclassification remained stable between the training and test datasets, while a bump in misclassification rate was observed for bootstrap forest and boosted tree. The increase for bootstrap forest was the largest, increasing by approximately 10%; suggesting that this model may be slightly overfitted.

OTHER ASSESSMENT METRICS

Models would be assessed on two levels, first models developed using the original dataset, and at the overall level across all eight models. Aside from comparing the true positive and misclassification proportions, the accuracy, precision, sensitivity and specificity of the test dataset of each model are tabulated in Table 5. Among the models implemented on the original dataset, the model using the Bootstrap Forest approach resulted in better outcomes, with the highest true positive rate of 51%. While sensitivity rate for this model was at a modest 51%, it was also highest compared to all other models. However, as mentioned in earlier paragraphs, a sizable increase in misclassification rate was observed for this model between the training and test dataset, a potential sign of overfitting. Therefore, depending on the purpose of the model, there may be value in considering the logistic regression or boosted tree method, with slightly poorer results, but lower likelihood of overfitting.

Across all 4 methods, the test dataset with the SMOTE treated sample recorded a significant increase in true positive cases detected, compared to when the methods were applied to the original test dataset. Notably, overall accuracy of the model did not improve despite higher true positive scores for models implemented on the SMOTE treated test samples. This lack of improvement may be due to the class imbalance, with the successful number of non-landslide cases predicted on the original dataset accounting for the high accuracy proportions.

Two important measurements for consideration are the Precision and Sensitivity levels of the models. Precision quantifies the number of positive landslide predictions that belong in the landslide class; a sharper tool compared to accuracy that is specific to the landslide class. Models implemented on the SMOTE treated sample performed better, when compared to the original sample, regardless of method. Between methods, Boosted Tree appeared to have recorded the highest improvement, increasing its precision rate by 19%. Sensitivity quantifies the proportion of observed landslide cases that were predicted as such. This is an important measure as a model with low sensitivity would mean that a sizable proportion of cases go undetected, and in the case of landslide detection, it may be disastrous. Like precision, models implemented on the SMOTE treated sample recorded higher sensitivity than those implemented on the original sample. Among the methods, Decision Tree recorded the sharpest improvement of 34%.

Lastly, specificity quantifies the proportion of observed non-landslide cases that were accurately predicted as such. It was interesting to note that across all models, specificity went down for models that were implemented on the SMOTE treated sample. As such, it is important to seek a balance between the measures when assessing model performance. For example, the decision tree model on the SMOTE sample noted the highest sensitivity, but lowest specificity.

CONCLUSION AND FUTURE RECOMMENDATIONS

This study set out to understand the efficacy of different classifier methods on detecting landslide susceptibility and if addressing the issue of class imbalance using SMOTE would improve overall classification performance compared to an imbalanced dataset, and our findings showed recursive partitioning methods such as Bootstrap Forest and Boosted Tree tended to yield better outcomes, compared to logistic regression. These methods, however, were also more likely to show signs of overfitting and should be monitored closely if chosen for future studies. On addressing class imbalance, results improved across all 4 classifier methods, to varying extents. Decision Tree applied on the SMOTE sample led to a model with the highest sensitivity, however Boosted Tree on the SMOTE sample yielded the most balanced results across multiple measures. Therefore, there is value in exploring sampling techniques such as SMOTE when doing similar landslide susceptibility studies in the future. It would also be useful to test multiple classifier methods and evaluate them based on the decision threshold required on a case-by-case basis, in order to decide on the best approach.

With these results in mind, there may be value in expanding the study on three fronts. First, use classifier methods such as Artificial Neural Networks and Frequency Ration Models that were not explored in this study. Second, to experiment with other sampling methods such as SMOTE with Tomek to see if results would improve beyond just using SMOTE. And lastly, to replicate the study across other landslide sites, to understand if improvements to the model is uniform across all sites, or due to unique features observed in this specific Hong Kong case.

REFERENCES

Chawla, N. V. et al. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Gao, H., Fam, P.S., Tay, L.T. et al. (2020). Three oversampling methods applied in a comparative landslide spatial research in Penang Island, Malaysia. SN Appl. Sci. 2, 1512. https://doi.org/10.1007/s42452-020-03307-8

Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & Geosciences, 81, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007

Haibo, He., E.A., Garcia. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9):1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hong Kong University of Science and Technology. (2022). *Landslide Prevention and Innovation Challenge* [Data set]. Zindi Africa. https://zindi.africa/competitions/landslide-prevention-and-innovation-challenge

Hurley, G. J. (2012). *JMP® Pro Bootstrap Forest*. JMP. https://www.mwsug.org/proceedings/2012/JM/MWSUG-2012-JM04.pdf

Korup, O., & Stolle, A. (2014). Landslide prediction from machine learning. Geology Today, 30(1), 26–33. https://doi.org/10.1111/gto.12034

Landau, S, & Barthel, S. (2010). International Encyclopedia of Education 3rd Edition.

Liu, Z., Gilbert, G., Cepeda, J. M., Lysdahl, A. O. K., Piciullo, L., Hefre, H., & Lacasse, S. (2021). Modelling of shallow landslides with machine learning algorithms. *Geoscience Frontiers*, *12*(1), 385–393. https://doi.org/10.1016/j.gsf.2020.04.014

Ma, Z., Mei, G. & Piccialli, F. (2021). Machine learning for landslides prevention: a survey. Neural Comput & Applic 33, 10881–10907. https://doi.org/10.1007/s00521-020-05529-8

Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., Avtar, R., & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, *207*, 103225. https://doi.org/10.1016/j.earscirev.2020.103225

Schreiber-Gregory, Deanna & Bader, Karlen. (2018). Logistic and Linear Regression Assumptions: Violation Recognition and Control.

Singh, Priyanka & Sharma, Prof. (2019). Analysis of Imbalanced Classification Algorithms: A Perspective View. International Journal of Trend in Scientific Research and Development. Volume-3. 974-978. https://doi.org/10.31142/ijtsrd21574

Stumpf, A., & Kerle, N. (2011). Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment, 115(10), 2564–2577. https://doi.org/10.1016/j.rse.2011.05.013

Wang, H., Zhang, L., Yin, K.S., Luo, H., & Li, J. (2021). Landslide identification using machine learning. Geoscience frontiers, 12, 351-364. https://doi.org/10.1016/j.gsf.2020.02.012

Wang, Y., Wu, X., Chen, Z., Ren, F., Feng, L., & Du, Q. (2019). Optimizing the Predictive Ability of Machine Learning Methods for Landslide Susceptibility Mapping Using SMOTE for Lishui City in Zhejiang Province, China. International journal of environmental research and public health, 16(3), 368. https://doi.org/10.3390/ijerph16030368

World Health Organization. (2018). *Landslides*. https://www.who.int/health-topics/landslides#tab=tab_3

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Fong Bao Xian
E-mail: bxfong.2022@mitb.smu.edu.sg

Name: Loh Jiahui
E-mail: jiahui.loh.2022@mitb.smu.edu.sg

Name: Sherinah Binte Rashid
E-mail: sherinahr.2022@mitb.smu.edu.sg