Impact of Imbalanced Data on Landslide Susceptibility Prediction

Fong Bao Xian, Loh Jiahui, Sherinah Rashid

SINGAPORE MANAGEMENT UNIVERSITY

Introduction

Landslides occur more frequently than any other geological event, and can happen anywhere in the world. WHO reported that between 1998 & 2017 worldwide:

18,000 deaths

4.8 million people affected

Objectives

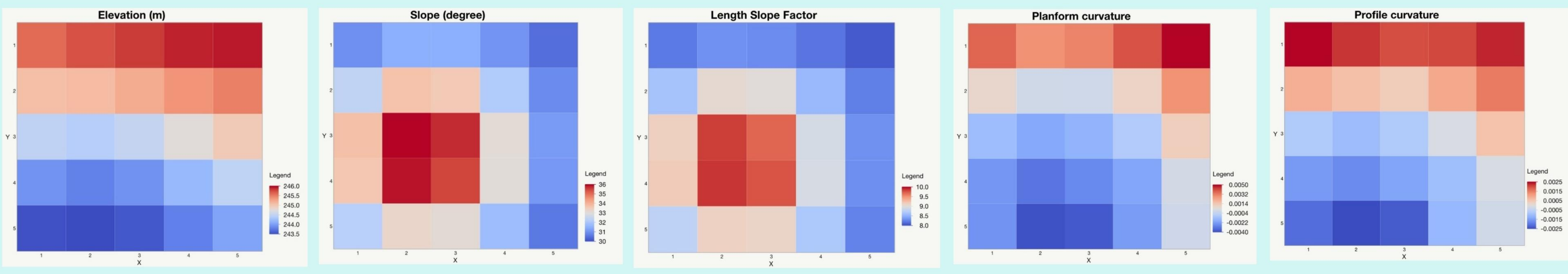
To perform Landslide Susceptibility Prediction for the effective prevention and management of landslide risks

1. Create & identify **key variables** for prediction
2. Investigate **impact of imbalanced data** on prediction performance
3. Assess efficacy of **different classifier methods:**
   - Statistical Method: Logistic Regression
   - Recursive Partitioning: Decision Tree, Bootstrap Forest, Boosted Tree

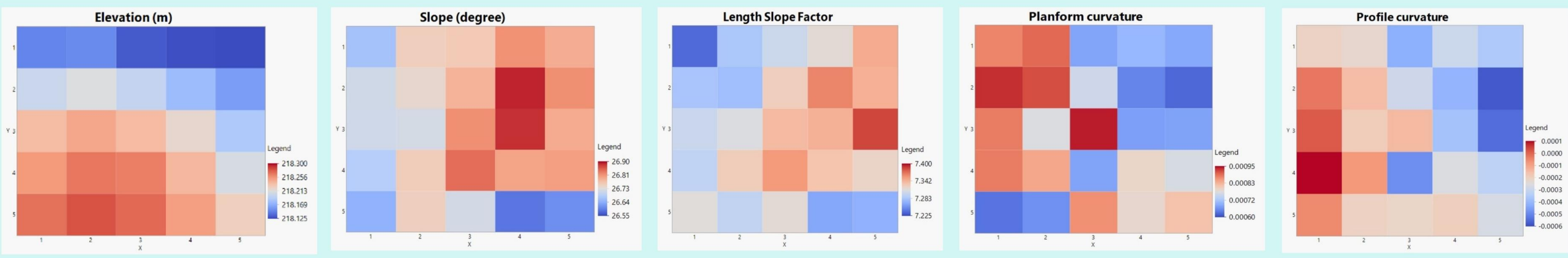
Heat Maps of Predictors

Heatmaps of selected continuous predictions highlighted that landslide cases had distinct patterns vis-à-vis non-landslide cases.

Landslide



Non-Landslide



Data Preparation

Data Source

- Contains 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
- Cell 13 is the location of landslide

Data Preparation

- Retained Cell 13 for aspect, geology, topographic wetness index & step duration orographic intensification factor
- Derived new variables for elevation, slope, length-slope factor, and planform & profile curvature
- SMOTE was utilized to expand our minority samples.

Logistic Regression

- Removed 2 derived variables to avoid multi-collinearity.

Recursive Partitioning

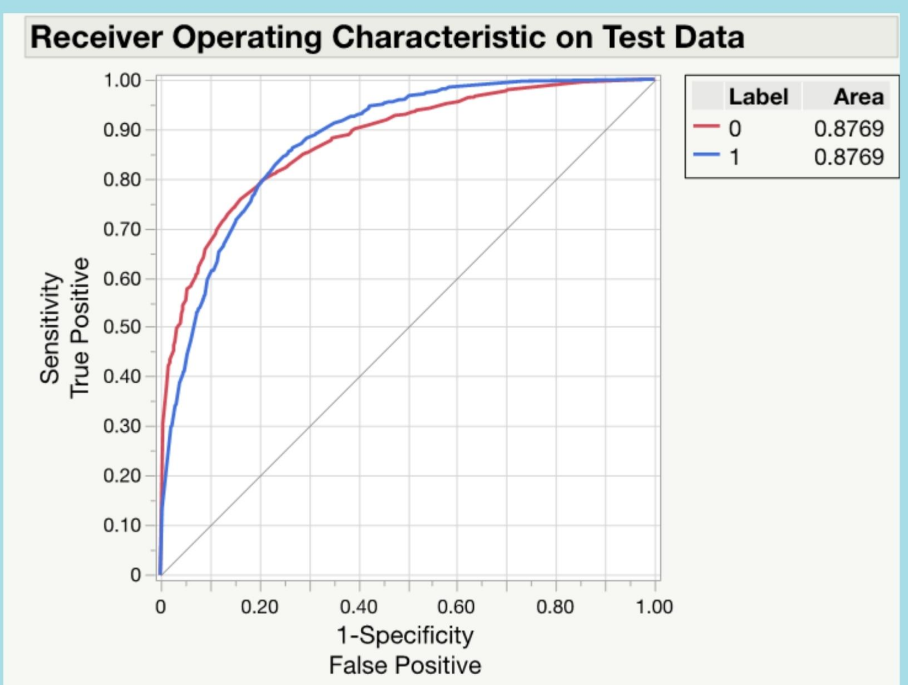
- Retained all variables

4 PREDICTIVE MODELS

Model Comparison

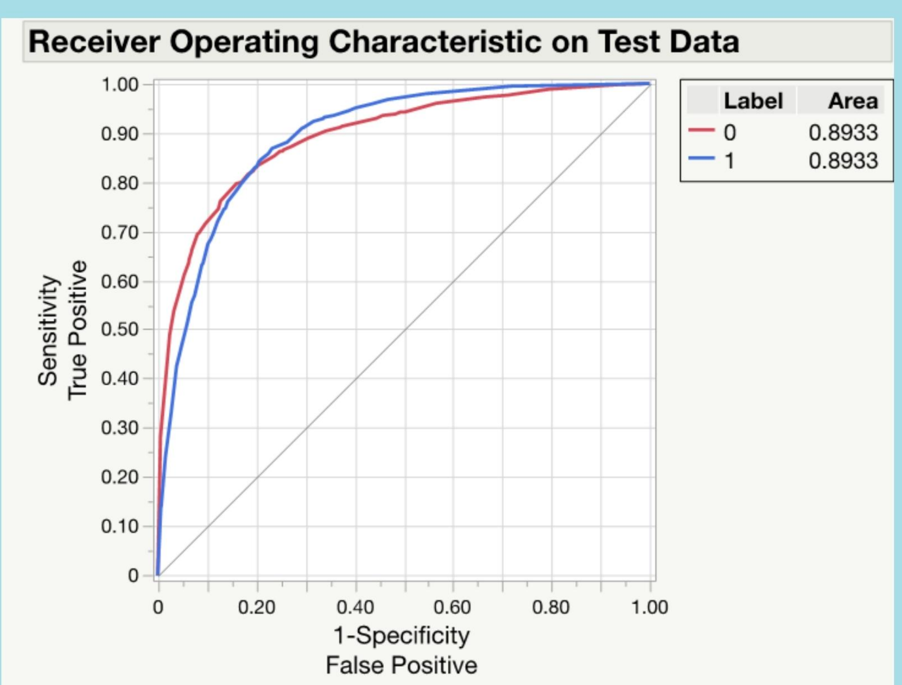
The best model was chosen based on its performance across various evaluation metrics.

Original



	Logistic Regression	Decision Tree	Bootstrap Forest	Boosted Tree
TP	0.441	0.357	0.513	0.450
TN	0.923	0.949	0.929	0.925
Accuracy	80%	80%	82%	81%
Misclassification	20%	20%	18%	19%
Precision	66%	70%	71%	67%
Sensitivity	44%	36%	51%	45%
Specificity	92%	95%	93%	92%

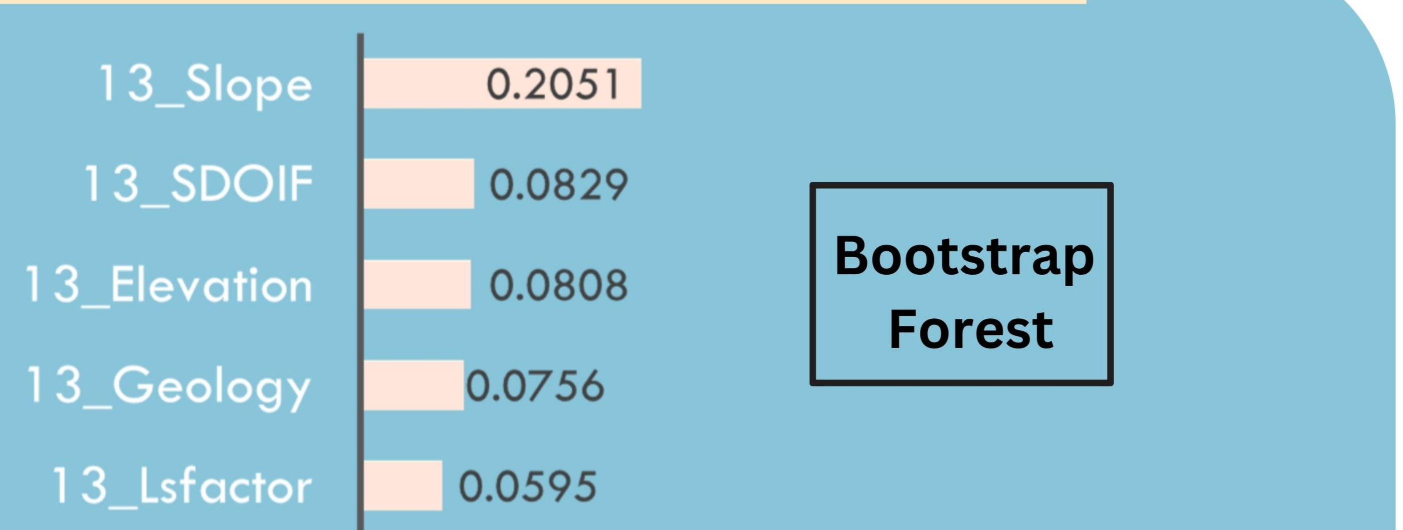
Post-SMOTE



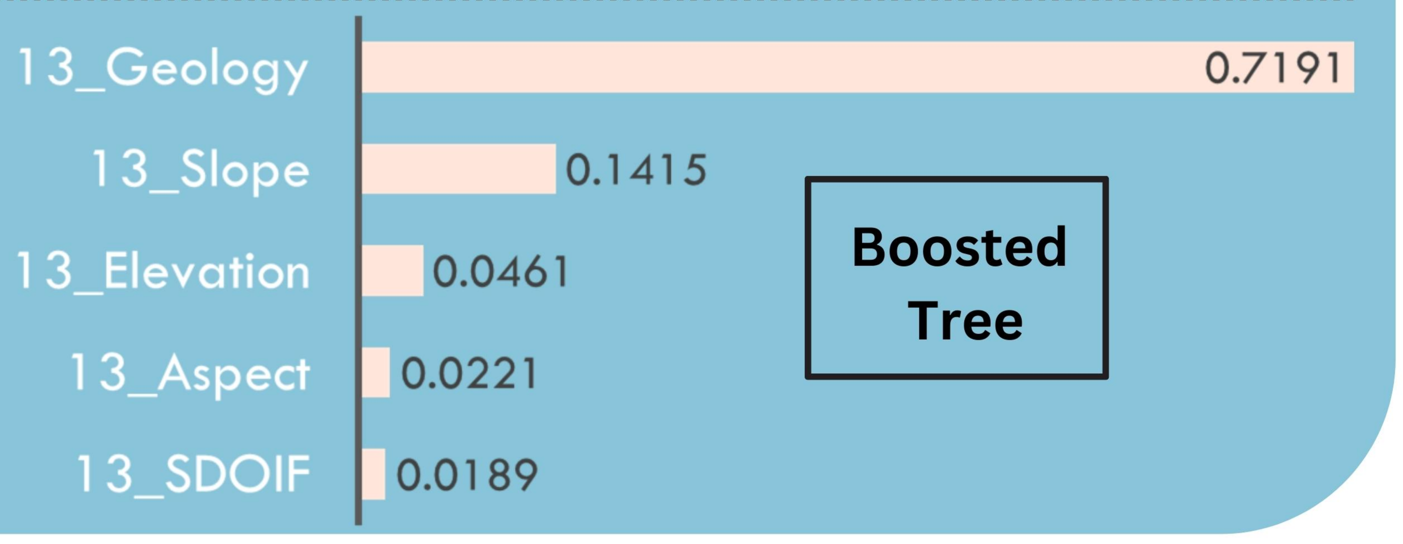
	Logistic Regression	Decision Tree	Bootstrap Forest	Boosted Tree
TP	0.801	0.897	0.860	0.864
TN	0.757	0.656	0.790	0.772
Accuracy	78%	78%	82%	82%
Misclassification	22%	22%	18%	18%
Precision	77%	72%	80%	86%
Sensitivity	80%	90%	86%	79%
Specificity	76%	66%	79%	85%

Top 5 Contributing Factors

Original — **Bootstrap Forest**



Post-SMOTE — **Boosted Tree**



Conclusion & Future Work

1. Landslide cell variables were better predictors than created variables
2. Usage of balanced data led to improved prediction outcomes across all models
3. Recursive partitioning methods yield better outcomes than Logistic Regression

1. Replication of study across other landslide sites to finetune predictor variables and understand model applicability
2. Experiment with alternative sampling methods, e.g., SMOTE with Tomek
3. Explore other classifier methods, e.g., Artificial Neural Networks and Frequency Ration Models