Assignment 6: Geospatial Machine Learning

(60 Points Total)

Data available under Resources>Assignment Data.

You have been provided with the following files:

Landslides>lsm_data.csv: set of example locations of "slope failures" and "not slope failures" point locations across the Valley and Ridge physiographic region of West Virginia.

Landslides>stack.tif: stack of raster predictor variables for a small subset of the Valley and Ridge physiographic region of West Virginia.

These data are from the following publication, which can be accessed for free:

Maxwell, A.E., M. Sharma, J.S. Kite, K.A. Donaldson, J.A. Thompson, M.L. Bell, and S.M. Maynard, 2020. Slope failure prediction using random forest machine learning and LiDAR in an eroded folded mountain belt, *Remote Sensing*, 12(3): 486. https://doi.org/10.3390/rs12030486.

Your Tasks

The goal of this assignment is for you to train and validate random forest-based machine learning models to predict the probability of slope failure occurrence using a variety of terrain, lithology/soils, and distance variables. The predictor variables are described in the tables below. Note that I am not asking you to optimize the hyperparameters or predict to the raster grid. However, you can predict to the raster grid if you want to experiment with this process.

T1: Read in the **lsm_data.csv** table. Randomly split the data into **training** and **testing** sets. Roughly 66.6% should be used for training while 33.3% should be withheld for testing. This should be stratified by the "class" attribute.

T2: Split the data into **Y** and **X** components. The Y variable is the "class" attribute. All other attributes are the predictor or X variables.

T3: Create subsets of the training and testing predictor variables as follows:

- All (All indexes)
- Just Terrain ([:,0:32])
- Just Lithology/Soils ([:,40:])
- Just Distance ([:,32:40])
- All Not Terrain ([:,32:])

T4: Train separate **random forests** models using each of the five predictor variable sets and the **training** samples. You do not need to optimize the hyperparameters.

T5: Generate **confusion matrices** and **classification reports** for each predictor variable set using the predictions on the **test** data and correct classifications.

T6: Calculate the **AUC metric** from the **ROC curve** for each predictor variable set using the probabilistic predictions on the **test** data and the correct classifications.

T7: Generate a **ROC curve plot** that includes the ROC curves for all five models. Each curve should use a different color and a legend should be provided.

T8: Provide a short write up focused on a comparison of the models based on the binary assessments and probabilistic assessments.

Optional T9: Use your all variables model to predict to the raster stack (**stack.tif**). You will need to rename the bands in the following order:

- “slp","sp21","sp11","sp7","rph21","rph11","rph7","diss21","diss11","diss7","slpmn21","slpmn11","slpmn7","sei","hli","asp_lin","sar","ssr21","ssr11","ssr7","crossc21","crossc11","crossc7","planc21","planc11","planc7","proc21","proc11","proc7","longc21","longc11","longc7","us_dist","state_dist","local_dist","strm_dist","strm_cost","us_cost","state_cost","local_cost", “steve" “dspm","drain"

Deliverables

- Jupyter Notebook. Use Markdown to generate your short write up within the Notebook.

Table 1. Terrain predictor variables.

Variable	Abbreviation	Description	Window Sizes (cells)
Slope	slp	Gradient or rate of maximum change in Z as degrees of rise	1
Mean Slope	slpmn	Slope average over a local window	7, 11, 21
Linear Aspect	asp_lin	Transform of topographic aspect to linear variable	1
Profile Curvature	proc	Curvature parallel to direction of maximum slope	7, 11, 21
Plan Curvature	planc	Curvature perpendicular to direction of maximum slope	7, 11, 21
Longitudinal Curvature	longc	Profile curvature intersecting with the plane defined by the surface normal and maximum gradient direction	7, 11, 21
Cross-Sectional Curvature	crossc	Tangential curvature intersecting with the plane defined by the surface normal and a tangent to the contour - perpendicular to maximum gradient direction	7, 11, 21
Slope Position	sp	Z – Mean Z	7, 11, 21
Topographic Roughness	rph	Square root of standard deviation of slope in local window	7, 11, 21
Topographic Dissection	diss	$\frac{Z - Min\ Z}{Max\ Z - Min\ Z}$	7, 11, 21
Surface Area Ratio	sar	$\frac{Cell\ Area}{Cosine(slope * \pi * 180)}$	1
Surface Relief Ratio	ssr	$\frac{Mean\ Z - Min\ Z}{Max\ Z - Min\ Z}$	7, 11, 21
Site Exposure Index	sei	Measure of exposure based on slope and aspect	1
Heat Load Index	hli	Measure of solar insolation based on slope, aspect, and latitude	1

Table 2. Additional predictor variables.

Variable	Abbreviation	Description
Distance to Roads	us_dist, state_dist, local_dist	Euclidean distance to nearest US, state, and local road
Cost Distance to US Roads (US, state, and local)	us_cost, state_cost, local_cost	Euclidean distance to nearest US, state, and local road weighted by slope
Distance from Streams	strm_dist	Distance from mapped streams
Cost Distance from Streams	strm_cost	Distance from
Geomorphic Presentation	Steve	Classification of rock formations based on geomorphic presentation
Dominant Soil Parent Material	dspm	Dominant parent material of soil
Soil Drainage Class	drain	Drainage class of soil