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:


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:


**Deliverables**

- Jupyter Notebook. Use Markdown to generate your short write up within the Notebook.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Description</th>
<th>Window Sizes (cells)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td>slp</td>
<td>Gradient or rate of maximum change in Z as degrees of rise</td>
<td>1</td>
</tr>
<tr>
<td>Mean Slope</td>
<td>slpmn</td>
<td>Slope average over a local window</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td>Linear Aspect</td>
<td>asp_lin</td>
<td>Transform of topographic aspect to linear variable</td>
<td>1</td>
</tr>
<tr>
<td>Profile Curvature</td>
<td>proc</td>
<td>Curvature parallel to direction of maximum slope</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td>Plan Curvature</td>
<td>planc</td>
<td>Curvature perpendicular to direction of maximum slope</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Profile curvature intersecting with the plane defined by the surface normal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>and maximum gradient direction</td>
<td></td>
</tr>
<tr>
<td>Longitudinal Curvature</td>
<td>longc</td>
<td>Tangential curvature intersecting with the plane defined by the surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>normal and a tangent to the contour - perpendicular to maximum gradient</td>
<td></td>
</tr>
<tr>
<td>Cross-Sectional Curvature</td>
<td>cosscc</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z – Mean Z</td>
<td></td>
</tr>
<tr>
<td>Slope Position</td>
<td>sp</td>
<td>Z – Mean Z</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td>Topographic Roughness</td>
<td>rph</td>
<td>Square root of standard deviation of slope in local window</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td>Topographic Dissection</td>
<td>diss</td>
<td>$\frac{Max Z - Min Z}{Cell Area}$</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td>Surface Area Ratio</td>
<td>sar</td>
<td>$\frac{Cosine(slope * \pi * 180)}{Mean Z - Min Z}$</td>
<td>1</td>
</tr>
<tr>
<td>Surface Relief Ratio</td>
<td>ssr</td>
<td>$\frac{Max Z - Min Z}{Max Z - Min Z}$</td>
<td>7, 11, 21</td>
</tr>
<tr>
<td>Site Exposure Index</td>
<td>sei</td>
<td>Measure of exposure based on slope and aspect</td>
<td>1</td>
</tr>
<tr>
<td>Heat Load Index</td>
<td>hli</td>
<td>Measure of solar insolation based on slope, aspect, and latitude</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2. Additional predictor variables.

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