Predicting Landslides in Oregon

Reed Laverack, February 2016

Executive Summary

Landslides in Oregon were studied to find which characteristics lead to an increased risk in an unstable slope. By finding these slopes, preventative measures can be put in place in order to prevent heavy repair costs and the risk of loss of life. The following steps were taken to achieve this goal.

- Landslides over the past 15 years in Oregon were extracted from the SLIDO database. Along with the historic landslides, an equivalent number of stable slopes were selected for comparison.
- This point data was joined with land coverage, digital elevation maps, geologic information, soil information and weather data to find which characteristics we can use to predict landslides.
- The accuracy of the random forest classifier was 85%. After the application of a factor of safety to the analysis, the number of false negatives can be reduced to save both lives and repair costs due to a missed landslide.

The final model should be built to minimize the false negative so the model picks up as many potential landslides as possible. The model can be used when planning transportation and commercial projects throughout Oregon to predict the likelihood of failure of the slopes.

Project Description

On March 22, 2014, 43 people in Oso, Washington were suddenly killed when the slope above the Stillaguamish River failed covering most of the small town in mud and trees. Efforts to find people in the rubble were sent out, but to little avail. The slope dammed the river causing flooding upstream, and threatening people downstream if the earth was to fail causing flash floods.

This tragedy was caused by a number of factor, included slope, precipitation and geology. By tracking these factors, we can make maps of the dangerous slopes, track the weather conditions, and alert areas when the potential for a slide is high.

Landslides can account for 25-50 deaths per year and an estimated one billion dollars in damage. Many of these deaths are attributed to rock falls. Debris flows cause the greater civil damage to areas often blocking or destroying transportation routes, housing and commercial structures. This study will focus on the debris flows.

Slope stability can be calculated by the factor of safety,

$$\text{Factor of Safety} = \frac{\text{Resisting Force}}{\text{Driving Force}}$$

As long as the resisting forces are greater than the driving forces, the slope is stable. The driving forces in the case of a landslide are the weight on the soil; the resisting forces are the frictional force within the soil body. In this study we will look for causes for the frictional forces to be low to begin with, and what factors cause them to be less than the weight of the soil mass. Several factors can lead to the reduction of these resisting forces including rainfall increasing the moisture in the soil, forestry or forest fires changing the vegetation of the slopes, or the disturbance of the ground due to earthquakes.

This study will look into landslides in the past 15 years; collect data on soil, slope and weather factors for each landslide slopes and selected stable slopes and use machine learning techniques to determine relationships between the factors to give more accurate alerts to future events.

This landslides studied will be located in Oregon where damages from landslides are typically in the tens of millions every year. Oregon also has low earthquake susceptibility when compared to states such as California, so the factors we will study will not include seismic events. The results of this study can prove the technique useful for other states in the US and abroad.

Data Sources

Historic landslides have been compiled based on maps from the SLIDO databases released by the Oregon Department of Transportation. For this study, the dates for the landslide need to be known in order to match weather and soil conditions. Also, the land coverage maps are available for dates after 2001. This leaves 1418 landslides in the SLIDO database. The database coordinates were transformed into the WGS 84 reference system. This reference will be in latitude and longitude and can be used in other databases such as weather and soil data.

In addition to the slopes that have slid in the past, slopes were defined that have not experienced landslides in order to have both stable and unstable slopes in the database. These slopes were chosen in proximity to the landslides in the complete SLIDO

database. The purpose of this study is to analyze what characteristics cause a slope to become unstable, which requires the comparison of nearby slopes to reduce the variation in testing features. The dates for the stable slopes were defined from the closest landslide occurrence in order to reduce variation in the data. So the database will compare nearby slopes under the same weather conditions to see what factors caused the unstable slopes.

Weather Underground was queried by location and date to show weather events on the day of the landslides. A sample query for 120W and 45N on January 1, 2001 is shown below,

http://api.wunderground.com/api/{api key}/history_20010101/geolookup/q/-120,45.json

These parameters included dew point, barometric pressure, wind speed, temperature and precipitation.

The USGS releases Land Use Land Coverage Maps every five years, 2001, 2006 and 2011 are available online for download. These maps are a raster data set comprised of 104424 x 161190 pixels each having defined ground type (e.g. Developed Land, Mixed Forest, etc.). The land coverage for the location of each landslide was taken from the image corresponding to a year before the landslide.

USGS also releases DEM (Digital Elevation Map) images for Oregon. These can be converted to .tiff files whose pixels represent the slope of land at that point. Geologic data was included in the study to find what kinds of rock exist below the slides in question.

The geologic raster data was imported into QGIS. All the raster images were converted to the WGS 84 reference system to be able to work with the landslide data. Data at the points of each landslide was extracted from each of the raster images. This data was collected in a .csv format and merged with the point and weather data.

Statistics and Visualization

In order to find which factors will be included in the final model we will compare the factors of the landslide site to stable slopes in the vicinity of the landslides.

Geology

The geology that underlies the slope often defines what the soil will be comprised of and how competent the substrates will be. Looking at the age of rocks in the data, unstable slopes are mostly underlain by rocks in the Oligocene epoch, where stable slopes are either in the more recent Miocene or older Eocene epochs, as shown in Figure 1.The geology of the rocks will have to be combined with the epochs to make any statements on the stability, in Figure 2 the unstable slopes are mostly comprised of terrestrial sedimentary rocks, which tend to be lightly cemented. This can lead to the rocks being weathered into loose sands. Loose sands have very little cohesion and are susceptible to sliding. The stable slopes tend to be comprised of the more competent rock types, volcanic and marine sedimentary.

Vegetation

Vegetation on the slopes helps hold the slope together through its root structure. The denser and deeper the root structure, the more stable a slope should be. In Figure 3, the unstable slopes are mostly grassland, which has very light and shallow roots in the soil, making the slopes more susceptible to sliding. The stable slopes tend to be comprised of forests whose trees will have strong and deep roots. Shrubs will provide

some stability to the slopes but not as much as trees, which is why shrubs account similar percentages of both stable and unstable slopes.

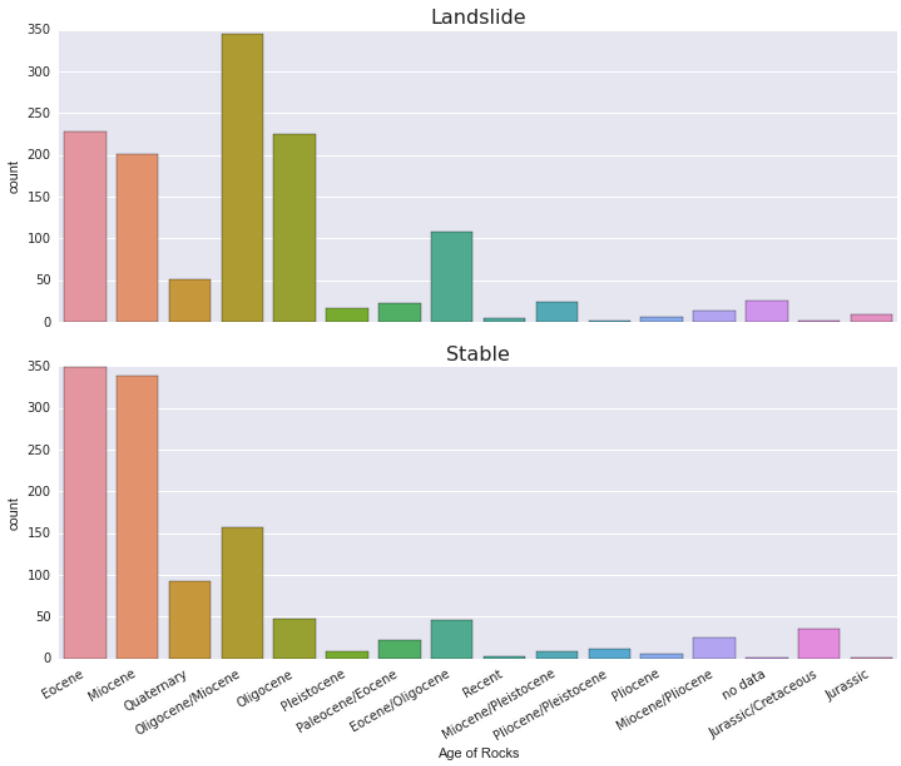

Figure 1 - Age of the rock underlying the stable and unstable slopes. The stable slopes tend to be in the Eocene of Miocene epochs while the unstable slopes are typically from the Oligocene epoch.

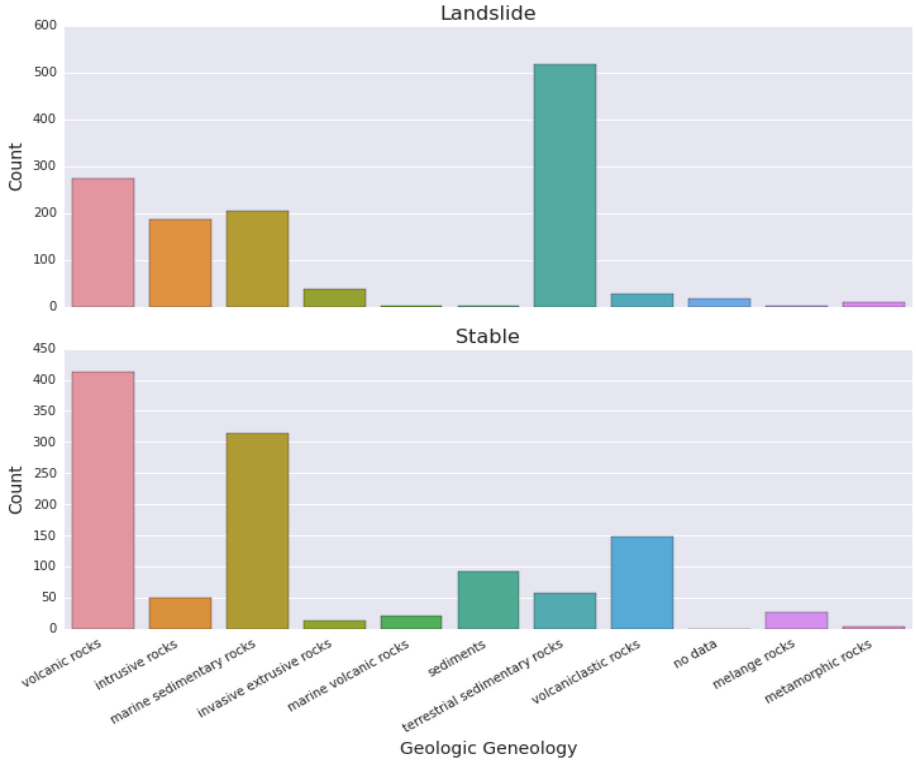

Figure 2 - General rock types for the slopes studied. The unstable slopes have a high proportion of underlying rocks comprised of the less competent terrestrial sedimentary rocks.

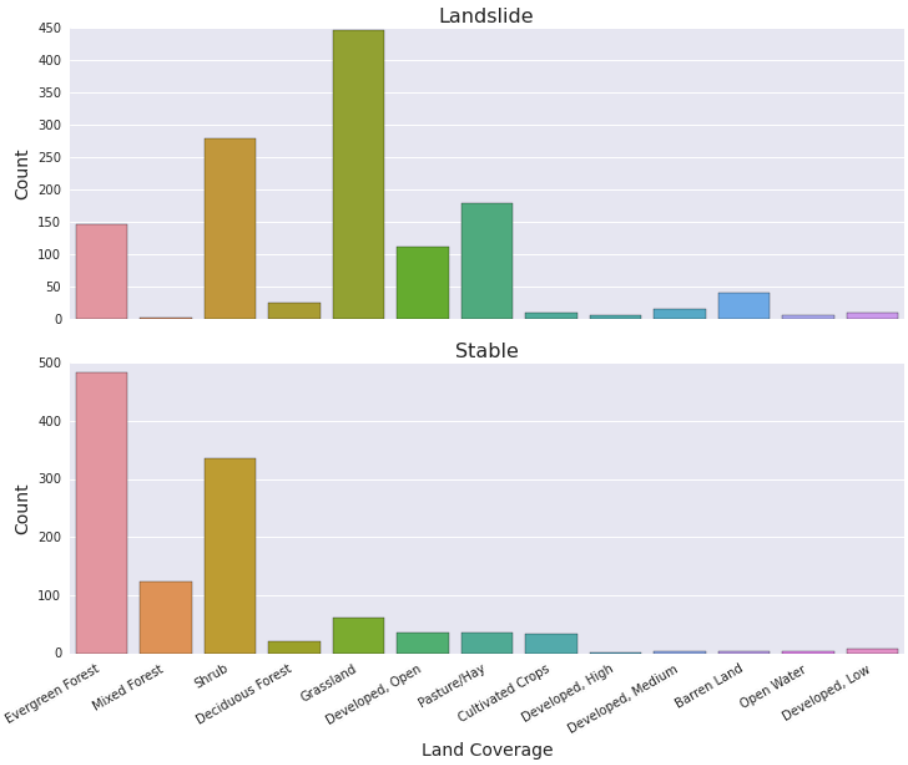

Figure 3 - Land Coverage for each slope studied. The unstable slopes tend to be overlain by vegetation with shallow and weak rot structure such as grass and shrubs.

Slope

The slope of the area has a large effect on the driving force of the factor of safety equation, increasing the component of the weight in the direction of the slide. Figure 4 shows that the unstable slopes tend to have a higher slope than the unstable slopes. Note that for this study we are comparing similar slopes in similar areas to landslide locations, if we were to include all the stable slopes, the mean for the stable slope would be much lower.

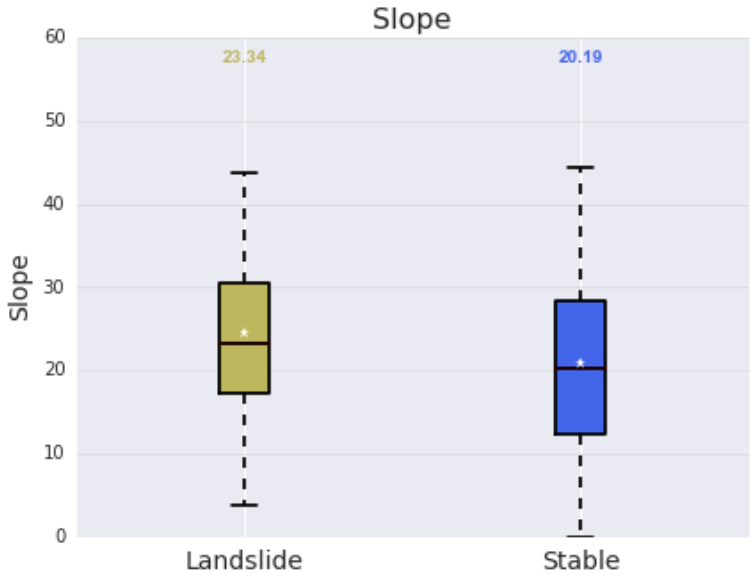

Figure 4 – Slope of the unstable and stable slopes in the study. For this study stable slope were selected to be similar to the unstable slopes, which greatly increases the slope of the stable points.

These comprise the permanent factors for the landslides. Though vegetation can change, it's typically of the scale of multiple months or years, barring human interaction, so for this study it will be considered permanent. The transient factors comprise moistly of weather factors, rainfall, soil moisture, wind, etc.

Weather

Moisture can reduce the cohesion of soils. Figure 5 and Figure 6 show that the daily and month precipitation totals are greater for the unstable slopes. Many of the slopes had no rain on the day of the slide, showing that soil moisture has a much greater effect than the surficial water. Other transient factors such as wind speed (Figure 7) and temperature are fairly consistent between the stable and unstable slopes. Differences may arise when they are taken in combination of the permanent features.

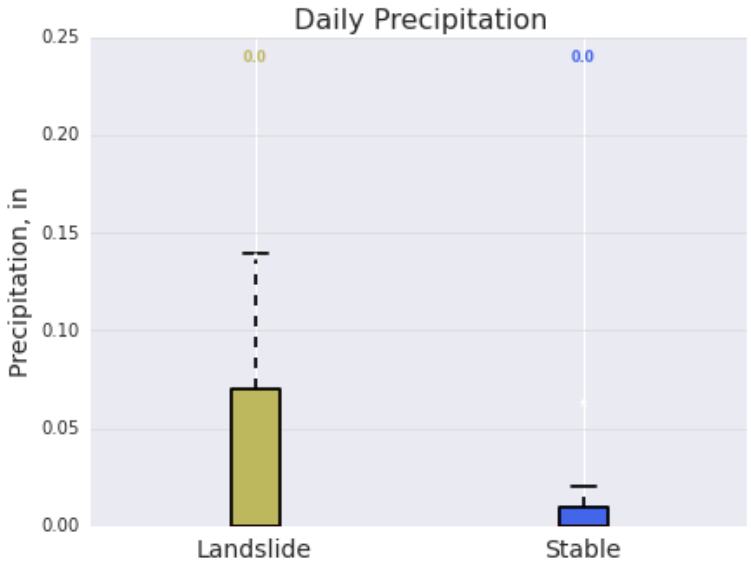

Figure 5 – Precipitation amounts for stable and unstable slopes on the day on the landslide. Many of these days had no precipitation accumulation, accounting for the average and median to be zero.

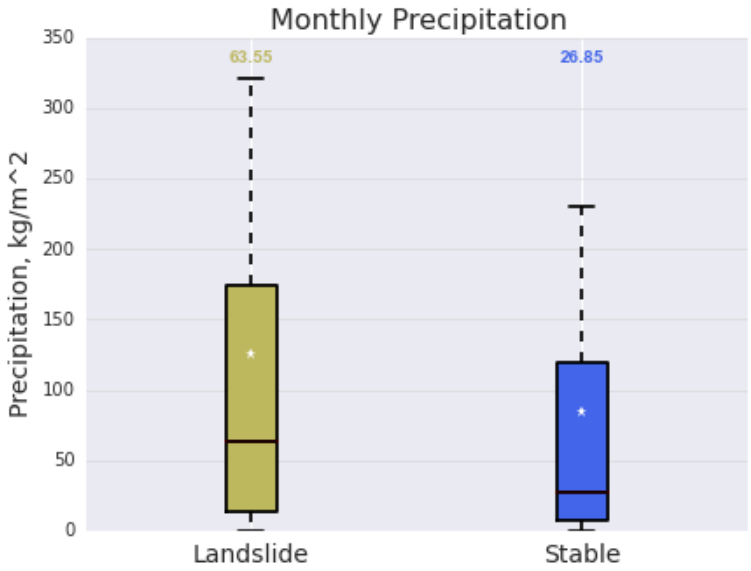

Figure 6 – Precipitation amounts for the month that the slopes failed. The unstable slopes tended to have higher amounts of rain that would lead to a higher saturation in the soil.

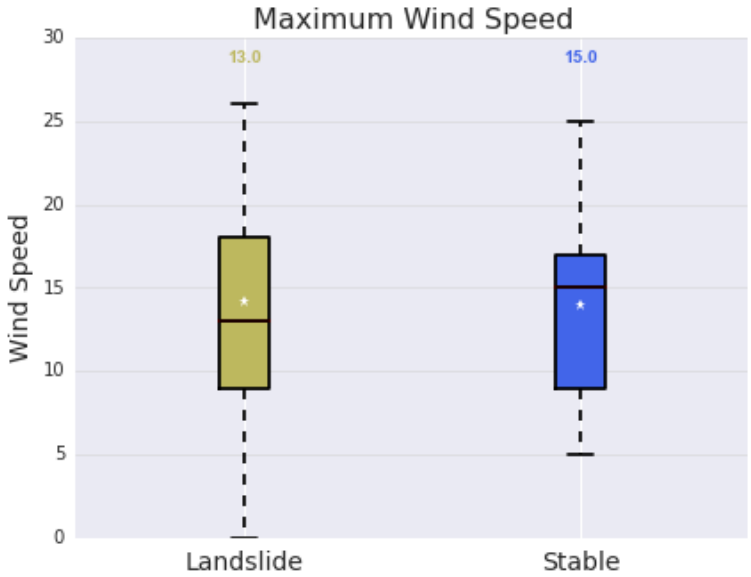

Figure 7 – The maximum wind speed at the locations of stable and unstable slopes. The wind speed does not show a bias towards the stable or unstable slopes but may become a factor when combined with the other features of the model.

Data Preparation

From these data sets the following features were extracted to be put into a model: Slope, Precipitation, Monthly Precipitation, Soil Moisture, Soil Moisture at 3 feet, Vegetation Coverage, Soil Temperature, Max Wind Speed, Max Temperature, Min Temperature, Max Humidity, Max Pressure, Geologic Age Name, Geologic Genealogy, Coverage.

There were multiple categorical features in the dataset including the geologic age name and genealogy and the coverage; these had to be assigned dummy variables in order to be used with the Random Forest Classifier. This assigns multiple Boolean columns to represent each unique value in the categorical column.

Analysis

A Random Forest Classifier was used on the data. Logistic Regression and SVM was also applied to the data but did not predict the data as well as random forest. The decision tree was found to have an accuracy of 85% on the testing set.

The feature importance was extracted from the decision tree; the 5 most important features were as follows:

```
Feature ranking:
1. feature 0 : Slope : (0.133970)
2. feature 48 : COVERAGE_Developed, Open : (0.094626)
3. feature 2 : Precip_Mon : (0.066917)
4. feature 5 : Vegetation : (0.061164)
5. feature 3 : Soil_Moist : (0.057708)
```

This follows our understanding of landslides that slope, soil, vegetation and moisture leads to failures in the soil. The Open feature for the land coverage refers to land that has human impact on it, but is primarily low vegetation land such as grasslands. The slope was predicted coming into the study not to be a predominant feature due to how the stable slopes for comparison had been chosen, but turned out to still be a great indication of landslides in the decision tree.

Looking at the incorrect answers in the decision tree, it is important to remember that a false positive (landslide on a stable slope) is acceptable, where a false negative (stable on an unstable slope) is not. The goal of these trees should be to encompass as many of the unstable slopes as possible. The predictions from our random forest classifier are shown below in a confusion matrix.

Table 1 - The results of the random forest classifier on the 730 slopes in the test dataset. 61 slopes were misclassified as stable. The goal of the study is to be able to classify as many landslides as possible.

	Landslide	Stable
Predicted Landslide	391	48
Predicted Stable	61	230

By printing the probabilities for the random forest classifier we can manually change answers that the probability of a landslide is greater than 33% rather than 50% which random forest does by default. This lowers the accuracy to 79%, but does the following to the confusion matrix.

Table 2 - The results of the random forest classifier for the probably of a landslide being greater than 33%. 37 landslides that were misclassified as stable were picked up using this analysis; the cost was 80 more stable slopes that were classified as landslides.

	Landslide	Stable
Predicted Landslide	415	128
Predicted Stable	24	163

For the sake of this paper, we will assume a landslide costs $500,000 to fix on average and the stabilization of a slope we classify as unstable costs $100,000. By reducing landslide criteria to a probability of 33% we correctly predicted an extra 37 landslides, but we also incorrectly predicted 80 stable slopes. With this method we have added an additional $8,000,000 in slope stabilization costs, but have potentially saved

$18,500,000 in repair costs to these. This method can be applied at different probabilities to whichever cost-benefit suits the needs of a department.

An ROC curve (**Error! Reference source not found.**) was created from the results of the analysis. An ROC curve typically works on the true and false positive rates on the data. Due to the nature of landslides, this has been reversed for this analysis to show the true negative and false negative rates of the prediction. Using the same prices as with the confusion matrix. A slope of 5 on the ROC curve would be the break-even point for cost. This states that to gain one more landslide in our prediction, we would need to misclassify 5 stable slopes. This slope occurs approximately at a false negative rate between 15% and 20%. So we can eliminate 80% of the false negatives while still keeping a positive cost-benefit.

These numbers are an oversimplification, and monetary cost is not the only risk during a landslide, but it serves as an example for how this model can be used in the allocation of resources within a department.

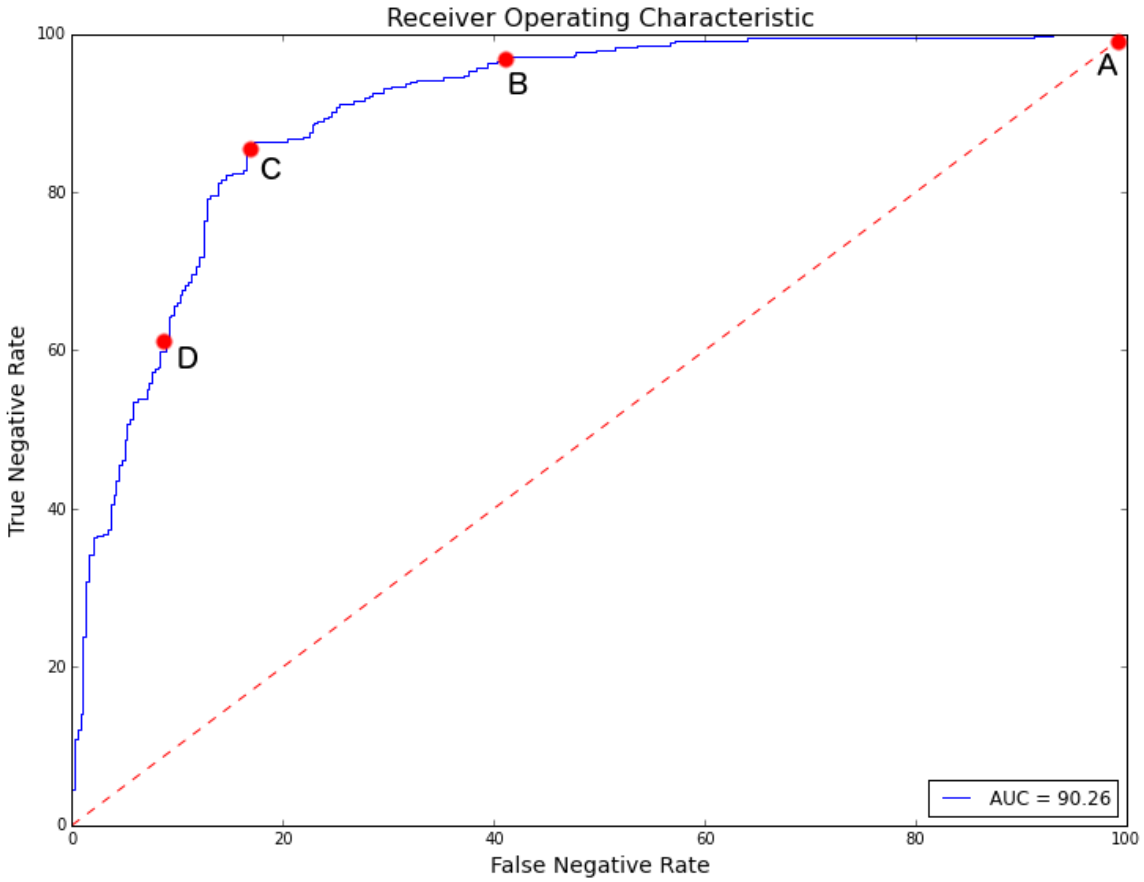

Figure 8 - ROC curve showing the true negative rate. Point (A) shows the model after the first run, where a landslide is labeled when the probability is greater than 50%. Moving down the curve to (B) the model is picking up false negatives faster than it loses true negatives. At point (C) the false negatives become more costly and point D it takes allowing a lot of misclassifications of true negative to get rid of false negative. For the most cost effective plan, the model should operated between point (B) and (C) in order for the best cost benefit.

Conclusions

Landslides occurring in the past 15 years have been compared to stable slopes in order to find characteristics leading to the failure of slopes. SLIDO landslide points were

combined with land coverage, geologic, topographic and weather data to accurately describe the conditions under which a landslide occurred. This data was trained in a random forest classifier to define the probability of a landslide occurring in our test data set.

The model performed well on multiple training and testing sets. By dropping the condition of the classifier for the probability of a landslide from 50% to 33% more of the landslides in the data were being captured by the decision tree. The cost-benefit analysis would need to be suited to whichever department is using the data, landslides occurring closer to populated areas have a higher risk of death, which is not included in the simple example provided in this paper. The takeaways from this model are as follows,

- The model predicts important features inline with what is generally accepted in the field as causes of landslides, including slope and moisture.
- By allowing more slopes to be classified as potential landslides we can minimize the unknown landslides and repair costs caused by those landslides.
- Using this classifier slopes can be rated on the potential danger for current and future residents and government agencies looking to provide civil works to areas.

The study presented in this paper shows a good proof of product that these types of studies could provide government and commercial industries with useful and actionable information about landslides to proactively avoid disasters such as the Oso Landslide. Using a cost benefit analysis alongside this model as discussed here, the potential savings are in the tens of millions in the long run.

Next Steps

The data going into this model could be improved upon. A landslide is comprised on the entire slope, the shape, change in slope, the toe composition, etc. a better approach would be to select the landslides as polygons instead of points in the GIS files. Another large piece of information would be the soil composition, coarse grained and fine grained soils can have very different effects on landslides and should be taken into account in future studies.

The data can be broken into permanent and transient features. A model could be set up to find susceptible slopes that match the permanent features of landslides. Then by tracking transient features, early warning systems could be issued to areas when the moisture data causes a dangerous condition.

Bibliography

SLIDO Database

Oregon Department of Geology and Mineral Industries (2016) SLIDO-3.2 Statewide Landslide Information Database for Oregon, http://www.oregongeology.org/sub/slido/

National Land Cover Database (2001, 2006, 2011)

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K., 2015, Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information.*Photogrammetric Engineering and Remote Sensing*, v. 81, no. 5, p. 345-354

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J., 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States, *PE&RS*, Vol. 77(9):858-864.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., and Wickham, J. 2007. Completion of the 2001 National Land Cover Database for the Conterminous United States. *Photogrammetric Engineering and Remote Sensing*, Vol. 73, No. 4, pp 337-341.

NASA Land Data Assimilation Systems

Xia, Y., et al. (2012), Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, J. Geophys. Res., 117, D03109, doi:10.1029/2011JD016048.

Xia, Y., et al. (2012), Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow, J. Geophys. Res., 117, D03110, doi:10.1029/2011JD016051.

Oregon Digital Elevation Maps

Oregon Geospatial Enterprise Office, 2016, Digital Elevation Models, http://www.oregon.gov/DAS/CIO/GEO/pages/data/dems.aspx

Weather Underground

http://api.wunderground.com/api