I began this project with doing some exploratory analysis. This map is colored and sized based on the count of deceased or injured in the report. And you might be asking yourself, what the heck is going on over here? Well, I’m glad you asked!

The majority of landslides take place near mountain ranges! Something fascinating about this section, is that these two red dots represent the 2 biggest events in the dataset: India with 5000 deaths, and Afghanistan 2100 deaths. And this is where we start to see the drawbacks of this type of data. The counts are highly suspect since it’s a very even number and no number reported for injured/not deceased.

Then, we can see where the other major drawback to reported data is: not every report is going to be present, so results need to be looked at while keeping these two things in mind.

Data pre-processing:

The original dataset had 31 features. After removing unnecessary features such as links, IDs, or repetitious features, I was left with 13 features. For the NAN data, I tried to fill in wherever I could. I even used a reverse lookup with the lat and long to get the city and country names. Others were filled in using “unknown” or 0s.

4 new variables were created: month, year, season, biome was created by using a dataset from the World Wildlife federation, and risk was created with the sum of fatalities and injured, then classified into 4 class levels.

Feature Selection:

Here’s where I started to get really excited, then really panicked.

I used SelectKBest to filter the dataset for the best features, and removed correlation of anything above 0.8. The first test of SVC is showing promising results. And I was excited to see that my biome variable was getting used so heavily.

I was about halfway through model selection before I realized I had a big problem! I did not account for class imbalance! So, everything was upsampled, then SelectKBest was run again and we can see here why it was so important. The coefficients indicate that the variables became so much more influential.

Model selection:

Data was split 70/30 to training and test sets and gridSearchCV was run for every model to select the best possible hyper parameters. Here we see the final models chosen and the accuracy scores.

Validation:

All models were validated using precision and recall scores as well as confusion matrices.

Feature importance:

Permutation importance was used for the Random Forest and CART models while the coefficients were used to look at the 2 SVC models.

Recurring features between all 4 models include size, trigger\_downpour , and setting. The Random Forest and CART models predominately used population while the two SVC models used either size or a country