Empirical avalanche prediction in Colorado:

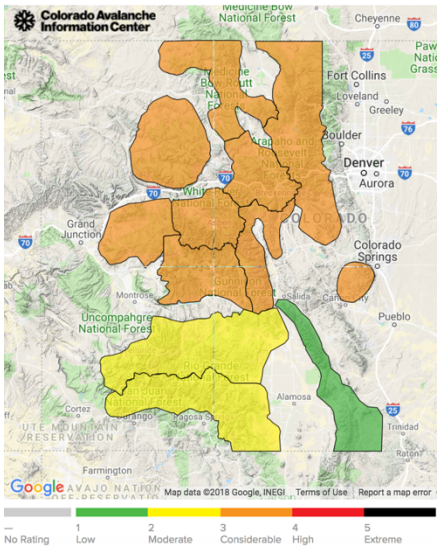
Can a machine-learning model trained on historical climatic and avalanche data augment prediction of avalanche risk?

Drew Thayer

Project summary:

Big Picture: Avalanches are a prime example of a natural phenomena caused by a combined sequence of non-linear processes aggregated over time. Accordingly, they be very hard to predict, however the physical processes that cause avalanches are well-identified, and can be measured. Experts use probabilistic models combined with human assessment to predict risk of avalanches.

Colorado Avalanche Information Center data:
The CAIC issues **daily avalanche risk forecasts** for 10 geographic zones across Colorado. These forecasts are highly influenced by human interpretation of recent events, and are quite accurate. Experts consider every nuance of weather and climate history, and their primary goal is issuing forecasts that will help people make safe decisions.



The CAIC also meticulously documents **avalanche observations**. Each record includes zone location, type of avalanche, start zone elevation, aspect, type (storm slab, persistent slab, wet, etc.) and destructiveness on an ordinal scale (D1 to D5). These data contain 10,128 observations over 18 years.

The NRDC's **SNOTEL** network has sensors distributed throughout the high country (near avalanche zones). There are ~50 sensors; each records 6 numeric fields related to precip, air temp, and snow-water equivalent.

Question:

Using incidence of avalanches as targets, can I train a machine-learning model on climatic data to predict the risk of an avalanche occurring?

Project Progression:

Minimum Viable Product: predictive model
A machine-learning model trained on daily climatic data that can predict the probability of an avalanche occurring, given conditions on a certain day. The model will be trained on all SNOTEL data and sectored into CAIC's 'Backcountry Zones'.

As proof of concept, I trained the following models to predict statewide incidents of avalanches from the Berthoud Pass SNOTEL weather data for years 2000-2018 (a spatially ill-posed problem):

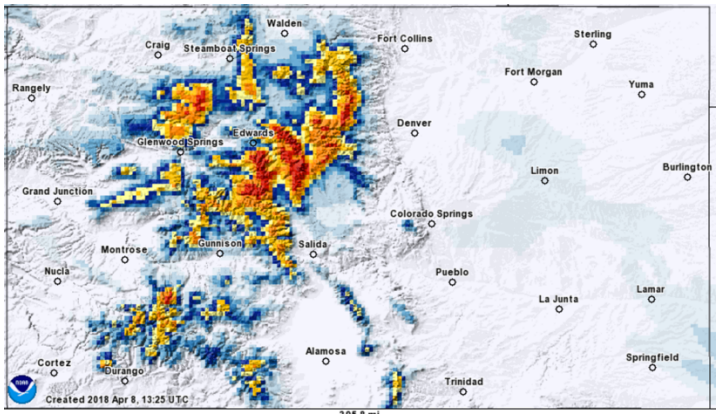
- *logistic regression*: target = (1,0) for occurrence of avalanche. Must address class imbalance.
- *linear regression*: target = # of avalanches on a given day (range 0 to 16). Very poor performance (not surprising given non-linear nature of physical processes). Training score = 0.126, test RMSE = 484.124.
- *gradient boosting regression*: performed much better. With non-optimized parameters, training score = 0.965, test RMSE = 64. Feature importances highlight precipitation and snow-water-equivalent on given day as most useful in training the model (makes sense physically).

These modest results from the gradient boosting regression give me confidence that a model trained on data and targets within the same spatial/climatic domains is worth pursuing.

Improvement 1: Comprehensive weather data
SNOTEL data is limited to SWE, precip, and air temp. **Wind speed** should be a very useful feature in predicting avalanches, as most avalanches before spring warming are caused by wind-slabs. More comprehensive weather data will likely improve predictability.

Improvement 2: **Learning from interpolated snowfall intensity models**
Every day, the NSIDC publishes an interpolated model of snowfall in the previous 24 hours. These models are raster images colored by snowfall amount.

example: the recent big storm on April 7, 2018. The central mountains received 18-24 inches in 24 hours! And there were MANY natural avalanches.



While this in an interpolated product, it still represents a *geographically distributed* data set, as opposed to point data (e.g. SNOTEL/weather stations). In many natural science fields, an active area of research involves incorporating new spatially distributed data into modeling techniques which have relied for decades on point measurements.

Hypothesis: A **Convolutional Neural Network** trained on these spatially-distributed representations of snowfall information can out-perform the model trained on weather data. *(Caveat: I expect this to be true in spring, but wind-speed may be more important in deep winter.)*

Improvement 4: **time-sequence of events matters**
The conditions that create avalanches are almost always time-aggregated *(e.g. wind-loading over several days, or a big storm followed by rapid warming, or snow followed by rain.)* Can a **Recurrent Neural Network** that considers past events perform even better?

Ideas for insights from unsupervised learning:
The CAIC dataset contains features such as aspect, start zone elevation, and type (storm slab, persistent slab, wet...) that could be appropriate for unsupervised analysis, e.g. clustering or PCA.

Potential topics of investigation:

- *What is the most common type of avalanche in March? What aspect is most likely to slide?*
- *Do avalanches below tree-line actually become more likely after nightly temperatures rise above freezing? Or daily highs consistently above 50 F?*