Empirical avalanche prediction in Colorado:

Can a machine-learning model trained on state-wide weather data augment prediction of avalanche risk?

Project summary:

Big Picture: Backcountry skiing is one of my main passions; it's incredibly rewarding, yet must be taken seriously because of the risk of avalanches. This sport requires a high degree of knowledge, understanding of snow/weather processes, self restraint, and reliable current information. The Colorado Avalanche Information Center provides an invaluable resource for backcountry skiers, publishing daily avalanche risk forecasts and in-depth analysis written by seasoned experts.

CAIC avalanche data: Every winter and spring, experts at the CAIC issue daily avalanche risk forecasts for the state. Risk is predicted for every aspect and for three elevation domains (above, near, and below tree-line) and forecasts are partitioned into 10 geographic zones which experience different weather and climatic patterns.

These avalanche risk forecasts are highly influenced by human interpretation of recent events. They are made by top experts in the field who are considering every nuance of weather and climate history, and who's primary goal is issuing forecasts that will help people make safe decisions.

The CAIC also meticulously documents avalanche observations, and has a record going back to 2000 for thousands of observed avalanches. This data contains consistent records of location, type of avalanche, start zone elevation, aspect, and destructiveness on an ordinal scale (D1 to D5). The data contains 10,128 observations over 18 years.

Question: The CAIC avalanche data is well set up for machine learning. Historical weather data is readily available from SNOTEL sensors, NOAA, WeatherUnderground, and others. Can I train a machine-learning model on climatic data to predict the probability of an avalanche occurring, given climatic conditions?

Project Progression

Minimum Viable Product: A machine-learning model trained on daily climatic data that can predict the probability of an avalanche occurring, given conditions on a certain day.

- As a simple proof of concept, I concatenated a dataset of state-wide incidence of D2 and D3 avalanches (most common) with weather data from the Berthoud Pass SNOTEL station for years 2000-2018. SNOTEL data has 6 numeric features. I trained and tested the following models:
  - **logistic regression**: target = (1,0) for occurence of avalanche. 90% accurate in training, but model useless due to class imbalance (only 8% 1s), need to rectify this...
  - **linear regression**: target = number of avalanches on a given day (range 0 to 16). Very poor performance (not surprising given non-linear nature of physical processes that cause avalanches). Cross-val training score = 0.126, test rmse = 484.124.
  - **gradient boosting regression**: performed much better. With un-optimized parameters, cross-val training score = 0.965, test RMSE = 64. Feature importances highlight precip and snow-water-equivalent on given day as most useful in training the model (makes sense physically).

Improvement 1: *AKA Drew vs Pandas.*

Include all available SNOTEL data sets and sector by 'Backcoutry Zone'. Not only will this improve performance (how often does weather at Cameron Pass have anything to do with the snowpack near Silverton?), but this will allow comparison of my predicted avalanche risk with CAIC's zone-by-zone predictions. It could be really interesting to

explore this comparison.

Improvement 2: *AKA Drew vs the APIs.*

Incorporate more weather data. SNOTEL data is limited to SWE, precip, and air temp data. Wind speed would be a very useful feature in predicting avalanches, as most avalanches before spring warming are related to wind-slabs. More comprehensive weather data from NOAA/ WeatherUnderground will likely improve predictability.

Improvement 3: *AKA Drew vs the NSIDC's API, FTP, and binary -> GeoTIFF conversion tools.*

Every day, the NSIDC publishes an interpolated model of snowfall (inches) in the previous 24 hours. They are raster images with a set # of pixels, 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. <img 'snodas' src='24_snow_20180408.png' width=300>

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. These daily images can be paired with occurence/ # of avalanches.

- **hypothesis:** A Convolutional Neural Net 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.)* Images can be fed in straight-up, no rotations/transformations needed.

Improvement 4: *If everything else works...*

The conditions that create big avalanches are almost always created conditions/events happening over a time-period *(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 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 see a slide in winter? In spring?*
- *Do avalanches below treeline actually become more likely after nightly temperatures rise above freezing? Or daily highs consistently above 50 F?*