ISyE 6740 - Spring 2021

Bayesian Backcountry Boarder - Final Report

Evan Anderson (eanderson42)

Colorado Backcountry Avalanche Prediction

Problem Statement

In the American West, particularly in the state of Colorado with easy access to the rocky mountain range, winter gravity sports are an incredibly popular activity. In particular, backcountry skiing, snowboarding, and snowmobiling have been increasing in popularity of the past decade. With this rise in popularity, and with a lack of knowledge by these new sport participants, the inherent dangers of being in a backcountry environment in the winter are not adequately observed, in part because the information is not easily accessible to those who are untrained in interpreting the weather and avalanche safety reports being published. This has culminated in a significant increase in deaths in the Colorado Front Range region, and over the entire United States over the past decade (CAIC, 2021).

The goal of this project is to take the vast amount of information around weather observations and avalanche reporting, and provide a simple to interpret avalanche risk level for anyone wanting to recreate in backcountry, stating whether or not an avalanche is likely for a given day in a given area.

Data Source

For data collection and model learning, two major sources will be used, with supplemental sources as necessary.

The first data source comes primarily from a snow fall reporting service, Snow Telemetry, via National Water and Climate Center (SNOTEL). This information is collected from multiple of stations across the rocky mountain range (and others) and was expected to include data such as daily snow fall, wind direction and speed, and high, low, median, and average temperatures, all on a daily basis. After scraping the data of interest, it was found that wind speed and directional data was largely unavailable, which would impact prediction ability. SNOTEL data was specifically selected because it is the standard that the CAIC prefers to use for all avalanche forecasting, and its data is readily accessible for large historical sets.

The second data source supplies the pre-classified values came from the Colorado Avalanche Information Center (CAIC). This organization collects avalanche observation information and forecasts avalanche types for large scale areas across Colorado. These forecasts are manually intensive, with human reported observations and expert forecasts analyzing data each evening for the given forecast the next day. The data utilized from the CAIC will confirm avalanche events for the past 10-20 years, with the events filtered only to “large” scale avalanches, meaning the snow slide has the potential to bury or harm humans in its path. These avalanche reports include information on location, face aspect, avalanche size, avalanche type, and trigger events.

Merging these two data sets required significant data cleaning and transformation. The first step was determining what backcountry zones in the historic CAIC data would be of most use. From there, the SNOTEL data collection station that most accurately represented that zone need to be researched and mapped. Deciding on zones of interest, a simple count of the number of large avalanches in each zone helped narrow down what zones provided the most data. From there using the landmark descriptions for the avalanche observations helped to narrow down which SNOTEL stations are most applicable for each zone.

The data was then mined for the daily weather data for 10 years of data by parsing the URL for the CSV file, and iterating through time periods and SNOTEL stations. The final data collected was as shown below.

Backcountry Zone	SNOTEL Station	# of Large Avs in past 10 Years
Northern San Juan	Idarado	741
Front Range	Loveland Basin	453
Southern San Juan	Red Mountain Pass	391
Gunnison	Schofield Pass	385
Vail & Summit County	Vail Mountain	356
Aspen	Independence Pass	333

Once the data was collected, the final step was the in-line aggregations to account for the times series aspect of the data when building and testing models. To do this, for each SNOTEL station, we aggregated precipitation, minimum temperature, maximum temperature, and attempted wind speed and direction for the past 1, 3, 7, and 14 days. During this stage was when it was discovered the data collection at each SNOTEL station for wind related data was not reliable, with sparse collection at best, and no collection of data at all for some of the stations. With this info, it became apparent that predicting the aspect (direction the mountain is facing) for a given avalanche would be near impossible, as wind direction plays a critical role in determining the more dangerous aspects for an area.

The final data preparation aspect was taking the avalanche data, and mapping a binary flag to each row on the SNOTEL weather data. This was accomplished with a mapping from backcountry zone to SNOTEL station, joining on these two and the data of the avalanche observation.

Methodology and Evaluation

The goal of avalanche prediction utilizing the data described above was to split into two distinct steps. The first is a binary classification of "avalanche likely" or "avalanche unlikely" given the current weather conditions. The second step was planned to further specify what aspects are most at risk for avalanche danger given that avalanche conditions are presently and likely from the initial classification. These two pieces of information will provided the most critical and most useful data points for backcountry recreations when making the decision of if and where to ski in avalanche terrain. Unfortunately, because of the missing wind speed and direction data, as well as classification issue that will be discussed further below, the aspect classification was not possible.

For the initial binary classification, multiple models where attempted. The included; Naive Bayes, KNN, One Class SVM, Random Forest, CART Models, and Linear Regression. Each will be discussed in detail below. All models were produced by shuffling data and splitting into 80% train and 20% test data sets.

Naive Bayes - The bayes classifier was used without doing any principal component analysis. This classifier produced predictions exclusive to no avalanche days. This gave a reasonable accuracy because of the nature of rare event data.

Naive Bayes	Predict No Av	Predict Avalanche
True - No Av	2545	0
True - Avalanche	238	0

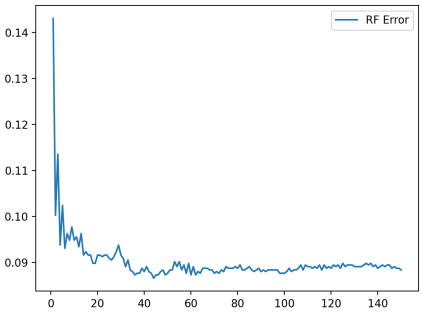
K-Nearest Neighbors - For the KNN classifier, we tested K in a range from 2 to 20. The ideal K when as far as both accuracy and falsely classifying Avalanche days as No Avalanche days was k=2. While the KNN model resulted in an over all worse accuracy, at 86.38%, there was significantly less false negatives compared to the other models, which is the outcome we want to seek from a safety perspective in avalanche prediction

KNN	Predict No Av	Predict Avalanche
True - No Av	2360	176
True - Avalanche	203	44

One Class SVM - The one class support vector machine model was the most time intensive model, as its requires both significant machine time, and it the nodel also benefits from reduced dimensionality, so principal component analysis was utilized. For this model, we tested from 5 to 24 components via PCA, and iterated over a gamma values, 1/n, with n in range 2 to 500. After running these many models, they ideal combination came to 11 components, selected via PCA, and a gamma of 1/47. This model resulted in an accuracy of 91%.

One Class SVM	Predict No Av	Predict Avalanche
True - No Av	2531	14
True - Avalanche	236	442

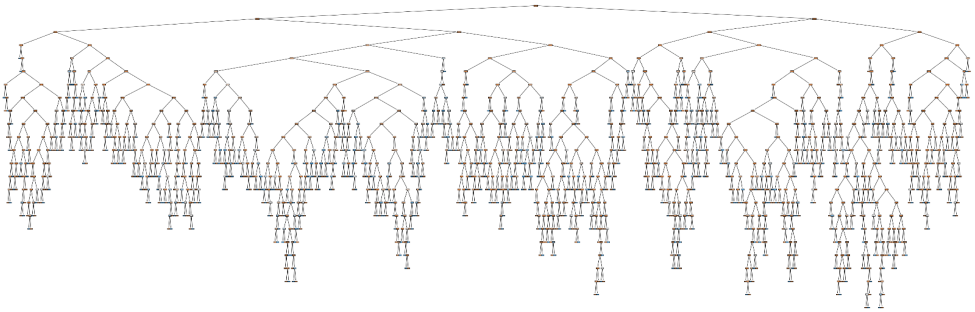
Random Forest - Multiple forests were generated over a range of maximum estimators, and the error was plotted to determine the optimum number of estimators. The optimum number was determined to by around 37 estimators.



From here, the model was built with the 37 estimator forest. The confusion matrix below showed that while accuracy was good, at around 91.23% this was largely because most instances where predicted to be no-avvy days. Again, largely due to over-prediciting no avalanche days.

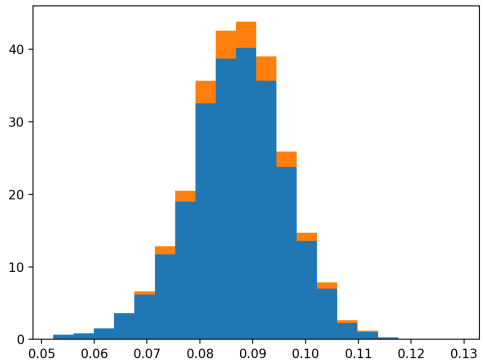
Random Forest	Predict No Av	Predict Avalanche
True - No Av	2519	17
True - Avalanche	227	20

CART - The CART decision tree model was produced with a final max depth of 22. As seen with previous models, the tree is optimizing for accuracy, which results in poor performance when it comes to false negatives. The overall accuracy was 86.52%



CART Model	Predict No Av	Predict Avalanche
True - No Av	2519	17
True - Avalanche	227	20

After running these models, it became apparent that binary classification was not the ideal solution. This is because it is difficult to optimize for the “safe” approach when it comes to predicting backcountry, which is minimizing the false negatives, or saying that the conditions are safe when in reality they may not be. This lead to thinking that a likelihood prediction would come into play. Instead of the binary No Avalanche vs Avalanche day, we could output a likelihood of an avalanche occurring on a given day. This ideal lends itself perfectly to **Linear Regression**. When creating the linear regression models, models were created using all features, as well as attempted feature selection using methods such as LASSO (with alphas tested from 1 to 10k) and PCA. In the end, no specific linear regression model produced a valuable likelihood estimation for the an avalanche on a given day. Below is an example of linear regression with PCA’d components = 20.



The orange represents avalanche days in there associated probability ranges, and the blue represents non avalanche days. Both are roughly normally distributed, indicating little prediction power over random chance.

Final Results

After all of the models built, the conclusion was reached that the weather data alone is not sufficient to classify whether an avalanche will or will not occur, or even the likelihood of an avalanche occurring based on a probabilistic output. Possible reason for this being the case include issues with hyper localized geography playing a factor. Avalanches typically occur on slopes over 30 degrees, and depend largely on the wind direction for snow accumulation at the top of the 30 degree slope. The SNOTEL weather stations, however, collect data typically at a

relatively neutral slope, with wind protection to ensure accurate accumulation totals. This discrepancy could explain the inability to associate the weather data with avalanche causes.

A further issue could come from the fact that large scale avalanches are “rare events”. With a ratio of about 20:1 avalanche days to no avalanche for days in the winter season. These models explored may not be the best for rare events such as this, and further models may need to be explored.

With more time, or in further exploration, a potential approach to this problem could be to reverse engineer the CAIC prediction values, an example prediction shown below. Instead of the binary prediction produced in this project, instead a “danger level” for a given aspect that was produced by experts in the field could be used in conjunction with weather data for certain SNOTEL sites. This reverse engineering could then help forecasters with an initial prediction, and they could then use their knowledge and data collected by the backcountry community to make a more timely and informed decision. The main issue with this approach (and why it wasn’t taken) is that this data is not made publicly available in an easily consumable format. It would need to be manually collected from the public internet, or the CAIC would need to expose the data for analysis.

Front Range

Backcountry Avalanche Forecast | Forecast Discussion | Observations & Weather Data

Print Share

Mon, Dec 13, 2021 at 7:36 AM
Issued by: Austin DiVesta

	Monday	Tuesday
Above Treeline	**Considerable (3)** Dangerous avalanche conditions. Cautious route-finding and conservative decision-making essential.	**Moderate (2)** Heightened avalanche conditions on specific terrain features. Evaluate snow and terrain carefully.
Near Treeline	**Moderate (2)** Heightened avalanche conditions on specific terrain features. Evaluate snow and terrain carefully.	**Moderate (2)** Heightened avalanche conditions on specific terrain features. Evaluate snow and terrain carefully.
Below Treeline	**Moderate (2)** Heightened avalanche conditions on specific terrain features. Evaluate snow and terrain carefully.	**Low (1)** Generally safe avalanche conditions. Watch for unstable snow on isolated terrain features.

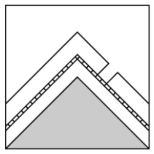
Danger Scale

—	1	2	3	4	5
No Rating	Low	Moderate	Considerable	High	Extreme

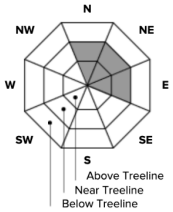
Summary

You can trigger dangerous avalanches on north, through northeast, to east-facings slopes in the alpine. This is especially true in areas near Rocky Mountain National Park and north through Cameron Pass, that received the most snow late last week. Avalanches will be larger and easier to trigger below ridgelines and in cross-loaded terrain features that have the most wind-drifted snow. You can identify these areas by looking for smooth pillows and cornices. If you see signs of instabilities like recent avalanches, cracking and collapsing, or hear whumpfing, consider moving to lower-angle, wind-sheltered slopes.

Avalanche Problem Persistent Slab

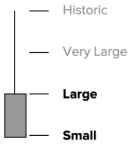


Avalanche Character



Aspect/Elevation

Likelihood



Size