Matthew Thompson & Lowell Deschenes

10/09/20

CS 287

Project Proposal

**Forecasting Next Day Probability of**

**Avalanche Occurrence for Tuckerman Ravine**

Motivation and Problem Statement:

Every year—on average—someone is either caught, carried, or buried by an avalanche in Tuckerman Ravine. The Mount Washington Avalanche Center (MWAC) forecasters do an exceptional job at giving up-to-date information on snow stability; however, each forecast is released with snow data from the early morning, when the snowpack is at its most stable. What about during the day when the snow is impacted by sun and warmer temperatures? While forecasters do account for how the snowpack will develop throughout the day, there is a lack of precision to this estimation. The goal of this project is to achieve the first step in graduating from 24-hour interval avalanche forecasts to more frequent information distribution. By using snow, weather, and avalanche data from MWAC and Mount Washington Observatory (MWOBS), the goal of this project is to train a machine learning model to predict the probability of an avalanche occurring the next day. If successful, forecasters and recreational users will have a numerical probability to use in their decision-making process. Finally, with the potential addition of snowpack data collected during the afternoons, prediction models such as these can begin to increase the frequency with which avalanche forecasts are release to the general public for recreational safety purposes.

Related Work:

The field of machine-learning based avalanche forecasting has not been widely researched, but there have been a few studies that have displayed high-accuracy models. These projects use high-dimensionality terrain characteristic data sets in conjunction with snow, weather, and avalanche information for effective predictions. Researchers at the University of Iran published a paper in 2019 attempting to predict snow avalanche hazard paths using support vector machines (SVM) and spatial data were able to produce a model with an area under the curve (AUC) receiver operating characteristics (ROC) curve value of 90 [Choubin 1]. AUC provides an aggregate measure of model performance across all possible classification thresholds, so this score is incredibly impressive. Unfortunately, MWAC does not publish specific terrain and avalanche location data necessary to create a model of this caliber; however, studies done without the use of spatio-temporal information have been shown to predict the occurrence of an avalanche with a decent effectiveness of >70% accuracy ([Dyer 1], [Gauthier 1], [Pozdnoukhov 1]). These papers used a variety of models, but SVM was shown to be the most successful. Others used include logistic regression (LR) and nearest neighbors (NN). For feature selection, all reports agreed upon the importance of new snow, temperature variation, max wind speed, liquid precipitation, wind direction, cloud cover, and recent avalanche activity as variables important to model outcome, with some including data from the previous 1 to 2 days. Approaches such as sensitivity analysis, recursive feature elimination, and a backward feature search algorithm were used to select these given variables.

Data Collection and Cleaning Path:

The first step of the data collection process was to open communication with representatives at both MWAC and MWOBS. From MWAC, this project requires snow and weather data from Hermit Lake, located just a few hundred vertical feet below the proper “bowl” of Tuckerman Ravine. At the summit of Mount Washington, approximately 2,000 vertical feet above the bowl of Tuckerman Ravine, MWOBS collects specific weather data and has access to information from as far back as 1934. Individuals from both groups have been contacted and the data should be retrieved within the next couple weeks. While that information is being acquired, the next step of scraping avalanche forecast and occurrence data from MWAC’s archives can be pursued. This data is posted in a blog-style update format, so it would not be useful to retrieve this raw data from the representative. Instead, variables like danger level, primary/secondary avalanche concerns, and avalanche occurrence can be collected via this method.

Once the datasets are acquired, they can begin to be cleaned. There are two data sets that need to be cleaned. One is the contains the snow and the other contains the weather data. Matt can complete the cleaning of the snow data and Lowell can complete the cleaning of the weather data. We must then merge the two data sets together. The data cleaning plan is as follows:

1. Check for missing values
   * The Hermit Lake snow data has already been identified as missing rows, in which case a row will be added with the difference of values of season total ground snow from the days before and after the lacking row. This represents the new, unrecorded snowfall, as long as that value is positive, although the snow water equivalent (SWE) and snow density percentage values will be unable to be filled. This data is also missing a very large number of inputs for snow density percentage. Luckily, it is easy to calculate this for daily intervals as long as there is data for 24-hour snowfall accumulation and 24-hour SWE:

Snow Accumulation \* SWE \* 100 = Density Percentage

1. Check for missing values that were filled in
2. Check for human error inputting data
   * Spelling errors
   * Abnormally small/large values
   * Unique values
3. Convert datasets to daily temporal space
   * The weather data from MWOBS is provided in hourly intervals, whereas both the snow data from Hermit Lake and the avalanche forecast/occurrence data are in daily intervals. The weather dataset values will have to be averaged and added to new columns, then this averaged data can be relocated into its own file.
4. Check for data faithfulness
5. Join snow, weather, and avalanche forecast/occurrence datasets
6. Remove data from outside the ski season
   * Based on the date of the final avalanche forecast for each season, drop all rows until beginning of next season
   * Add column containing which ski season the data is from (ex. 2014-2015 Season)

Data Exploration, Modeling, and Analysis Plan:

Exploring:

* Explore data types
* Feature selection (one of: Univariate Selection, Feature Importance, or Correlation Matric with Heatmap, Lasso Regression)
* Small multiple chart for exploring correlations, skewness, and outliers

Modeling:

* Stratified K-Fold Cross-Validation (K = 5)
  + There is large imbalance of the target value in the dataset, Stratified K-Fold Cross-Validation is designed to be effective in these cases
* Support Vector Machine (SVM) - Gaussian Kernel
  + SVM shows the most promise after its consistent success in similar projects, also allows for scalability to higher-dimensionality data for the future (spatial data) ([Choubin 1], [Dyer 1], [Pozdnoukhov 1], [Pozdnoukhov 2])
  + The SVM Gaussian Kernel works the best for datasets with a small feature size and a medium number of training examples
* Nearest-Neighbors (NN)
  + Showed similar results to SVM in other papers, worth investigating and could be included in Ensemble Learning suite ([Dyer 1], [Pozdnoukhov 1])
  + Lacks high-dimensionality data scalability that SVM boasts
* Logistic Regression (LR)
  + Showed similar results to SVM in other papers, worth investigating and could be included in Ensemble Learning suite ([Gauthier 1], [Dyer 1])
* Mini-batch Gradient Descent
* Ensemble Learning ([Pozdnoukhov 2])
  + Due to lacking spatial data the prediction accuracy increases gained from using Ensemble Learning are intriguing and could increase model practicality

Visualizing:

* Avalanche occurrence by liquid precipitation, season snow total, average wind speed, and temperature variance
* Probability of avalanche occurrence by season snow total, liquid precipitation, average wind speed, and temperature variance
* Probability of avalanche occurrence by avalanche occurrence and day of season as line plot

Analyzing:

* Confusion Matrix and Precision/Recall Plot
  + To get concise picture of the performance of the model due to dataset being unbalanced
* ROC Curve (for each model)
  + To compare the performance of each individual model
* Diagnose Bias and Variance
  + To understand the path to improving model accuracy based on whether bias or variance is too high
* Learning Curves (for each model)
  + To understand the differences in data fitting for each model

Needs Assessment:

Areas of Uncertainty:

* Obtaining data on avalanche likelihood, size, and information on specific aspects and elevations from the Mount Washington Avalanche Center Snow Rangers may or may not be possible because it is displayed to the public via JPG file. Unfortunately, without this the models will have inaccuracies introduced because Tuckerman Ravine does not include any West, South West, or North West facing aspects, whereas the avalanche forecasts include these aspects.
* It would be preferred to use Stratified K-Fold Cross-Validation, however, it may make the most sense to keep data from the same ski season together and not be divided by the data splitting process.

Contingency Plans:

* First plan:
  + Predict probability of avalanche occurrence
* Second Plan:
  + Predict probability of avalanche occurrence
  + Ensemble learning model
* Third Plan:
  + Predict probability of avalanche occurrence for each aspect in each elevation classification
  + Ensemble learning model

Proposed Timeline:

October 25 – Check-In #1:

* Completed scraping data from MWAC archives
* Acquired all data

November 8 – Check-In #2:

* Completed data cleaning process

November 15 – Check-In #3:

* Completed data exploration
* Completed model implementation

November 22 – Check-In #4:

* Completed data and model analysis
* Completed visualizing results

November 29 – Project Due:

References:

* [Choubin 1] Choubin, Bahram; et al. (2019, October). “Snow avalanche hazard prediction using machine learning methods”. <https://www-sciencedirect-com.ezproxy.uvm.edu/science/article/pii/S0022169419306493>
* [Dyer 1] Dyer, Wes. (2010). “Forecasting Avalanches in the Pacific Northwest”. <http://cs229.stanford.edu/proj2010/Dyer-ForecastingAvalanchesInThePacificNorthwest.pdf>
* [Gauthier 1] Gauthier, F.. (2017, June 15). “Logistic models as a forecasting tool for snow avalanches in a cold maritime climate: northern Gaspésie, Québec, Canada”. <https://link.springer.com/article/10.1007/s11069-017-2959-3>
* [Pozdnoukhov 1] Pozdnoukhov, A.; Purves, R.S.; Kanevski, M.. (2008). “Applying machine learning methods to avalanche forecasting”. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.687.6577&rep=rep1&type=pdf>
* [Pozdnoukhov 2] Pozdnoukhov, A.; Matasci, G.; Purves, R.S.; Kanevski, M.. (2011). “Spatio-temporal avalanche forecasting with Support Vector Machines”. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.671.3966&rep=rep1&type=pdf>