Spencer Tollefson

October 3, 2018

Proposal for Project Benson

Predict hiking trail rating on wta.org

The Washington Trail Association is a non-profit advocacy group that focuses efforts on promoting hiking in the state of Washington. The organization, beginning as *Signpost* magazine in 1966, currently pursues goals including reducing barriers to hiking, volunteer trail maintenance, organizing hiking activities, encouraging people to access the outdoors, and providing information to the public.

WTA has created and provided, free-of-charge, a database of over 3,500 hikes in Washington state. These hikes differ from short strolls to multi-day backpacking extravaganzas. The database web GUI allows visitors to filter from the list of all hikes by selecting from over 30 features, such as location, length, mountain views, beaches, and more. Additionally, visitors may write a "trip report", essentially a forum comment, and rate each hike on a 5-star scale.

Each "hike" has its own web page which displays all hike features in an aesthetically pleasing manner, as well as a hike description, directions to the trail, and Google Map view.

Here is an example page

Question/Need:

Predict the rating (on a 0-5 scale) of any hike given its set of features.

This prediction model, while mostly for my fun and learning, could potentially help WTA and other trail authorities decide which features to invest in when maintaining or creating new trails.

Description of data:

Each hike can contain up to 30 features. Included are a number of continuous features, such as the length of the hike or number of trip reports. Additionally, about half of the features are binary-categorical features. They indicate the presence of something like a "coast" on a hike, and is indicated as present or not present.

For some pages, there is extensive information on the authority responsible for maintaining the trail head or on the author of the website description. Because of this I have made multiple fields (author 1, author 2) to capture this information for pages in which it exists. It will fill as None for pages it does not exist.

Each hike page contains the following features.

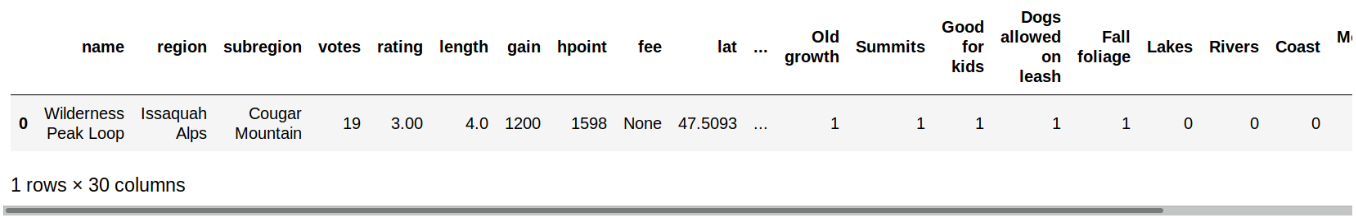
Features	Type	Description	Use in Model	Typically Available for All Hikes?
name	string	Name of hike	Y	Y

Features	Type	Description	Use in Model	Typically Available for All Hikes?
region	string	Name of greater geographic region, limited to 11 areas	Y	Y
subregion	string	Name of one of the multiple subregions within each region	Y	Y
total votes	int	Total number of votes cast to calculate rating	Y	Y
rating	float	Rating on a scale of 0-5 (out to 2 decimals) of the hike by visitors to wta.org	Y	
length (miles)	float	Number of miles (to one decimal) of the hike	Y	Y
length is one-way or round trip	string	Defines if length is determined "one-way" or "round trip"	Y	Y
gain (feet)	int	Number of change in vertical feet from low point to high point of hike	Y	Y
highest point (feet)	int	Elevation of highest point of hike measured in feet	Y	Y
parking/entrance fee	string	Either "None", a $ value, or name of pass required	Y	N
latitude	float	Latitude coordinates of the trailhead	Y	Y
longitude	float	Longitude coordinates of the trailhead	Y	Y
trailhead information 1	string	Text information describing certain trails or features (such as cliffs) on trail	Y	N
trailhead information 2	string	Text information describing authority responsible for trailhead information	Y	N
author 1	string	Name of group responsible for writing web page description	Y	N
author 2	string	Name of individual responsible for writing web page description	N	
count of trip reports	int	Number of total trip reports written for hike	Y	Y
Wildflowers/Meadows	binary	Presence on hike (Y/N)	Y	Y
Ridges/passes	binary	Presence on hike (Y/N)	Y	Y

Features	Type	Description	Use in Model	Typically Available for All Hikes?
Wildlife	binary	Presence on hike (Y/N)	Y	Y
Waterfalls	binary	Presence on hike (Y/N)	Y	Y
Old growth	binary	Presence on hike (Y/N)	Y	Y
Summits	binary	Presence on hike (Y/N)	Y	Y
Good for kids	binary	Presence on hike (Y/N)	Y	Y
Dogs allowed on leash	binary	Presence on hike (Y/N)	Y	Y
Fall foliage	binary	Presence on hike (Y/N)	Y	Y
Lakes	binary	Presence on hike (Y/N)	Y	Y
Rivers	binary	Presence on hike (Y/N)	Y	Y
Coast	binary	Presence on hike (Y/N)	Y	Y
Mountain views	binary	Presence on hike (Y/N)	Y	Y
Established campsites	binary	Presence on hike (Y/N)	Y	Y

Characteristics of each row of data:

Each row of data will be a record for a particular hike. There are currently 3,555 hikes in the database according to the WTA website. Thus I expect that many rows, with a field for each of the features listed above.



	name	region	subregion	votes	rating	length	gain	hpoint	fee	lat	...	Old growth	Summits	Good for kids	Dogs allowed on leash	Fall foliage	Lakes	Rivers	Coast	M
0	Wilderness Peak Loop	Issaquah Alps	Cougar Mountain	19	3.00	4.0	1200	1598	None	47.5093	...	1	1	1	1	1	0	0	0	

1 rows × 30 columns

Known Unknowns

- Hikes with low number of total votes are more susceptible to delineating from a model. A single or few voters can heavily influence rating.
- Unknown how many hikes have substantial number of votes (arbitrarily say, greater than 10)
- As with many rating systems, often people who take the time to rate have a highly positive or experience which can lead to fluctuations in the model.
- Weather conditions, interaction with other hikers on the trail during the day of the hike, and other likely important factors are not captured by this model.
- 1,822 of the 3,555 hikes have "incomplete information". Unknown how much information is missing until scraping. A large amount of missing information may make modeling difficult.