Project Luther - Predicting a Hike's Rating

Spencer Tollefson

October 12, 2018

Project Design

WTA.org is a website that has a database of 3,555 hikes (at the time of scraping) which includes hike details, trip reports, and a voting system allowing users to rate a scale on a scale of 1-5. The idea of this project was to see if the given information in this database would allow one to predict how highly a hike scored on the voting system. A good model potentially could aid "trail planners" with the foresight to predict how well-liked a newly opened hike would do before they invested in creating the trail.

Although I originally hoped to use nearly all the features I scraped as well as create some new ones, time constraints did not allow in time for project completion. Chiefly, the lat/long coordinates and sub-regions of the hikes both offer unique insights and were not incorporated in my models. Additionally, using other sources of data such as weather patterns/history and alltrails.com were skipped due to time constraints.

Initially, I hoped to create the model without using features that a trail planner would not have access to: such as the use of how many times people had rated a trail and the number of trip reports that were left for the trail. After some simple linear regressions built with those features excluded, it became clear the model had little value. Thus, I decided to use all data available to me.

This project, at different points, used the following models: OLS, Polynomial Regression, Lasso, Ridge Regression, Elastic net, and a taste of Random Forest Regression (with an instructor post-presentation). Cross validation with numerous folds were used on training data for each of these models and later compared with the test data to see how well the model performed on out-of-sample data. All of the best performing models adopted a Polynomial Tranformation of the 2nd degree. The best performing model was an Elastic Net model tuned to 0.1 L1 and 0.9 L2, measured by an R^2 value of .183. Additionally, a pure Lasso regularization returned a value of .18 and a pure Ridge Regression of .16. However, the Elastic Net returned the lowest difference between training data and the hold-out test set of data with only a .04 difference in R^2. This slight difference in R^2 between test and training indicates it likely had a better fit to the data than the other models.

The used model suggested that hikes on the coast, with summit points, and a sizable length were indiciative of people giving a high rating for the hike. However, of equal interest was how much the rest of individual features did **not* correlate to a high or low score. This leads to the belief that perhaps different features would have more predictive value than only the ones evaluated. The number of trip reports for a given hike is the strongest factor indicating a positive rating, however given this is intuitive and nearly a proxy for popularity I tried to stray away from focusing on it.

Tools

- Python: *Web scraping: requests, beautifulsoup *Data storage: pandas, pickle *Data analysis: numpy, pandas, scikit-learn, Jupyter *Presentation: matplotlib, seaborn
- Libreoffice Impress

Data

The data was obtained by scraping WTA.org. The website had pages for 3,555 unique hikes at the time of scraping on October 4, 2018. Each page containd the name of the hike, an embedded Google Maps window, the hiker trip reports for the hike, as well as a number of characteristics such as hike length, elevation gain, if the hike contained features such as coast/old growth forest/campsites/wildflowers, and even coordinates of the trailhead among other features.

All the features can be found in the Appendix. Notably, the Region and Fee categories required preprocessing to manipulate from one column into many one-hot encoded columns. This increased the number of features a total of 37 analyzed by the models. Scikit-learn techniques made it simple with little overhead to parse out the unique features in these columns and build the model outward.

The number of vote and number of trip report features were highly correlated. In order to prevent confusing my model, I chose to eliminate the number of votes feature due to this concern.

What I Would Do Differently

To give my model real value, I would have liked to remove any features which are only available post-hike creation. Ideally, this model could help "hike creators" plan where to create the trail for another hike. They would not have access to features such as the number of trip reports of a hike. However, this feature aided the model so much I chose to keep it for the sake of this project.

I would have taken a closer look at the raw distribution of "rating" pre-modeling. It turns out the rating is heavily left-skewed with a higher percentage of hikes receiving scores in the 3-5 range and fewer in the 1-3 range. By applying a transformation before modeling, it could help tease out and better understand the features that lead to sub-3 hike scores.

More charts during the modeling and EDA phase of residuals, Q-Q, and individual feature histograms would aid in understanding the underlying interaction between feature and target better.

Post-presentation I was introduced to Random Forest Regression and saw markable improvements on my model. I am interested in exploring this technique farther.

Appendix

Features	Type	Description	Use in Model	Typically Available for All Hikes?

Features	Type	Description	Use in Model	Typically Available for All Hikes?
rating	float	Rating on a scale of 0-5 (out to 2 decimals) of the hike by visitors to wta.org	Y	
name	string	Name of hike	N	Y
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	N	Y
total votes	int	Total number of votes cast to calculate rating	N	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"	N/A	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 rame of pass required	Y	N
latitude	float	Latitude coordinates of the trailhead	N	Y
longitude	float	Longitude coordinates of the trailhead	N	Y
trailhead information 1	string	Text information describing certain trails or features (such as cliffs) on trail	N	N
trailhead information 2	string	Text information describing authority responsible for trailhead information	N	N
author 1	string	Name of group responsible for writing web page description	N	N
author 2	string	Name of individual responsible for writing web page description	N	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

Features	Type	Description	Use in Model	Typically Available for All Hikes?
Ridges/passes	binary	Presence on hike (Y/N)	Y	Y
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