Web Scrapping and Linear Regression Project Writeup

*Create a database from the web, and find the correlation among the data*

By: Scott Turransky

**Abstract**

The goal of this project is to develop a database from data scraped from a website (not an already established database). The database must have 1000+ points and 10+ features. From there, perform feature engineering and modify the model to get the best regression model possible. Once these improvements were implemented in the train and validation subsections, they were tested against the test subsection.

This project used data on trails in National Parks. It seeks to identify how different features affect a trail’s attendance.

**Design**

The original design of this project was to identify factors that affect number of hikers on trails, so when new parks are added to the National Park system staff can identify which trails will be most popular and need additional maintenance. Originally the design of this project was based on the website AllTrails, which has a metric “Completed” that was to be used as our attendance target. AllTrails identified I was trying to scrape their website early on, so all scrapping had to be from a similar site Hikingproject.com. While less popular than AllTrails, it had many of the features needed to build a database. However, since there was no “completed” category, the number of raters of a particular trail served as our target.

**Data**

Information on 2938 trails among all the USA national parks were pulled from Hiking project.

For each trail, the following 16 data points were found and identified:

1. Park Name
2. Trail Website
3. Trail difficulty
4. Average Rating of the trail
5. Number of people who rated the trail
6. Distance in miles
7. High elevation in feet
8. Low elevation in feet
9. Distance hiker goes uphill in feet
10. Distance hiker goes downhill
11. Trail type (loop, out and back etc.)
12. Average grade (%)
13. Max grade (%)
14. Check-ins
15. State
16. Dog policy

After the data was cleaned up, the model was also assessed on the following additional features derived from the features above:

* Number of National Parks per state
* Region of USA where the park is located

**Algorithms**

**Web Scrapping:** A list of the national parks was found at <https://www.hikingproject.com/search?q=national%20park>

From there, the websites for the US National Parks were compiled in a list using Beautiful Soup and HTML headers. From each National Park page, a dictionary was created with the park as a key, and all the trails at that park combined into a list serving as the value for the key.

The next step was to take the dictionary of trail website and scrape the necessary data from each page. Each step of the scrap was tested on a pair of websites to ensure it worked before applying it to every site. This was saved as a database and saved to Excel.

**Regression:** in a separate Jupyter Notebook (so the scrapping did not constantly have to occur) the database was cleaned.

Then it was split into train, validation and test data. Each change to the model was tested on the train dataset and the validation dataset. Once the model was optimized, the train and validation dataset were combined and used to assess the test dataset.

**Tools**

The following tools were used:

* Selenium- This was used to control the dynamic website to get the data needed
* Beautiful Soup- This was used to parse the HTML code
* Pandas- this was used to manage the database
* Statmodels- This was used to test R\_squared on the models
* SKlearn- This was used to split the model into train- validate-test subsets as well as check the R\_squared score for the predicted Y values.

**Results**

|  |  |  |
| --- | --- | --- |
| Dataset | Initial R\_Squared | Final R\_Squared |
| Train | 0.081 | 0.114 (+0.033) |
| Validation | 0.070 | 0.097 (+0.027) |

The score for training increased by about 40%, the validation set increased by 39%.

Those datasets were combined and used against the test dataset:

|  |  |  |
| --- | --- | --- |
|  | Train + Validation datasets | Test dataset |
| R\_squared | 0.119 | 0.124 |