Washington Trails Rating Project

By Marcel Colvin

Introduction

- Can we create a regression model to find which features of hikes relate to user ratings?
- A web designer may find this analysis interesting to create a recommendation system for hikes that users may enjoy.

Methodology

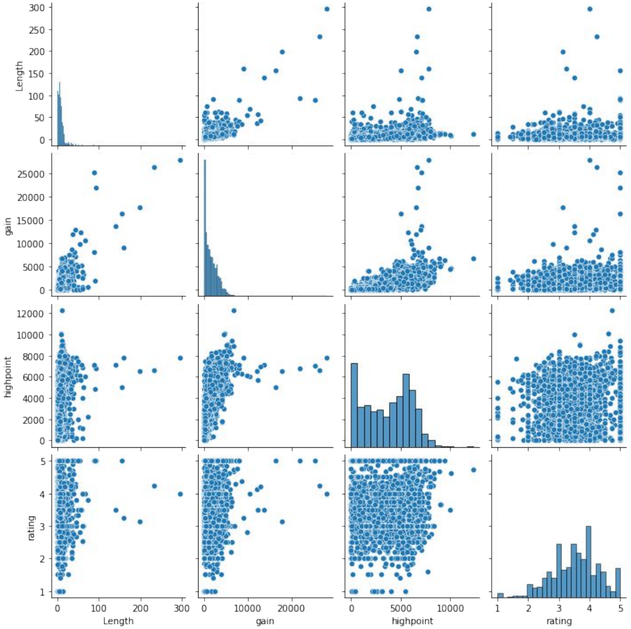
- Data: The dataset was scraped from the Washington Trails Association (WTA.org), which included about 4000 hikes with numerical and categorical data descriptors.
- Tools: Pandas, SKLearn, NLTK VADER, Selenium

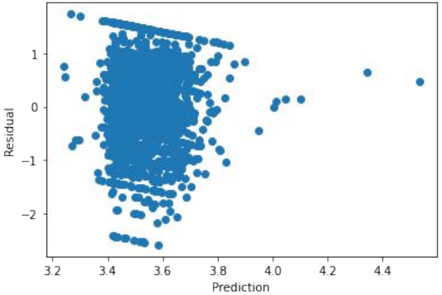
Methodology

- Metrics:
  - Length, Gain, Highpoint
  - Lakes, Old Growth, Coast, etc.
  - VADER summary sentiment score
  - Latitude and Longitude
- Models:
  - Linear Regression
  - Random Forest Regression

Results (Linear Regression)

- Was poor with just the Numerical Data R^2 ~0.016
- Residuals distributed around the average rating
- Categorical Data added made R^2 ~ 0.1
- Log and Polynomial transforms made R^2 worse





Results (Linear Regression)

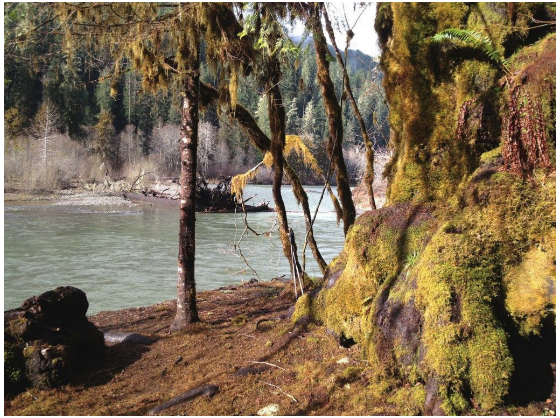
- VADER Sentiment analysis scores of written summary made almost no difference in Linear Regression model scoring.
- Latitude and Longitude also made no difference.
- Subsetting by number of ratings > 10 made $R^2 \sim 0.2$, but this makes the dataset only around 400 points.

Results (Random Forest Regression)

- Performed decently on Train Data R^2 ~0.85, but terribly on Validation and Test data (R^2 ~0.06) implying a high bias model.
- Subset by rating count > 10, subset by highest performing features, and performed a Grid Search CV for the best parameters.
- Final model had R^2 ~0.58 Train and R^2 ~0.57 Test.

Conclusions

- Linear Regression was not able to perform well due to the lack of linearity between the numerical data and the rating.
- Random Forest Regression performed better because the multiple decision trees make the model more robust to the categorical and numerical data. With the tuning it was able to create a better model, but with less interpretability.



Future Work

- Scrape exact Latitude and Longitude from the WTA website to get more exact locations for all hikes.
- Scrape the entirety of the “Trip Reports” and perform sentiment analysis about those data points for each hike(~200k)