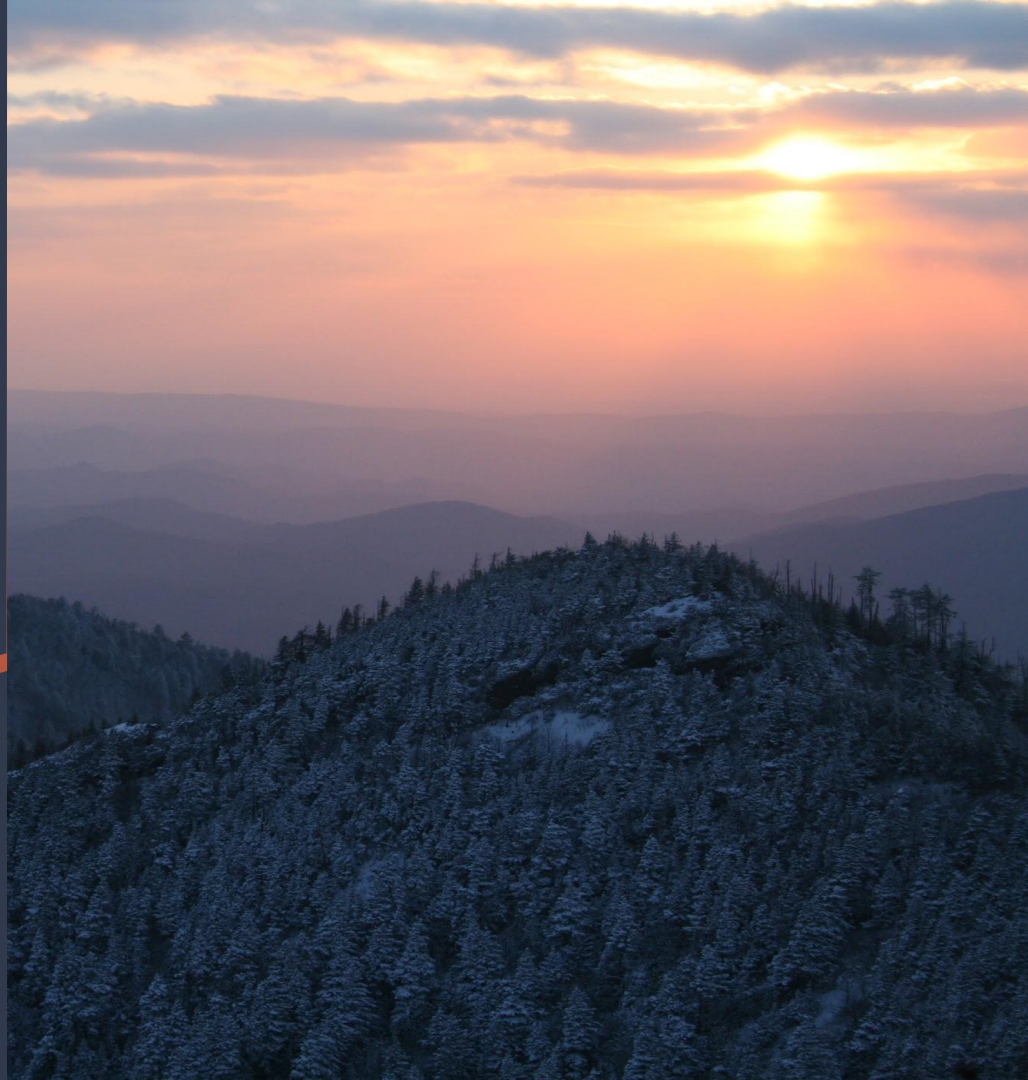
National Park Trails

What do we love most?

Contents



Motivation

Hiking in the United States has been steadily rising for years; in 2018 there was a recorded 47.87 million hikers.

This project is focused on the National Park Trails and understanding the biggest draws to trails and an attempt to predict which trails will have have the highest ratings based on the features and activities it offers.

Source:
https://www.statista.com/statistics/191240/participants-in-hiking-in-the-us-since-2006/

Overview

The data originally included 3313 rows (entries) and 18 columns which include trail name, state, trail activities & features such as backpacking, fishing, if dogs are allowed, birding, waterfalls, views, caves, if the trails are paved, and 21 more.

All data (csv file) was loaded from kaggle datasets (user: planejane) which was acquired from AllTrails.com

Source:
https://www.kaggle.com/planejane/national-park-trails

Cleaning and Organizing

- Removed rows with missing values
- Gave each specific activity & feature its own column
- Corrected false values
  - Maui being listed as its own state
  - Mislabeled units
- Georgia has no national parks but had one listed which was removed
- Various minor issues

California	633
Washington	297
Wyoming	290
Utah	282
Colorado	242
Maine	175
Arizona	164
Virginia	163
Tennessee	162
Montana	144
North Carolina	109
Texas	91
Ohio	50
Hawaii	42
Florida	29
Alaska	28
Nevada	25
Kentucky	22
Oregon	19
North Dakota	19
Arkansas	16
Indiana	15
South Dakota	15
South Carolina	8
New Mexico	8
Minnesota	7
Missouri	2
Michigan	2

Exploratory Data Analysis

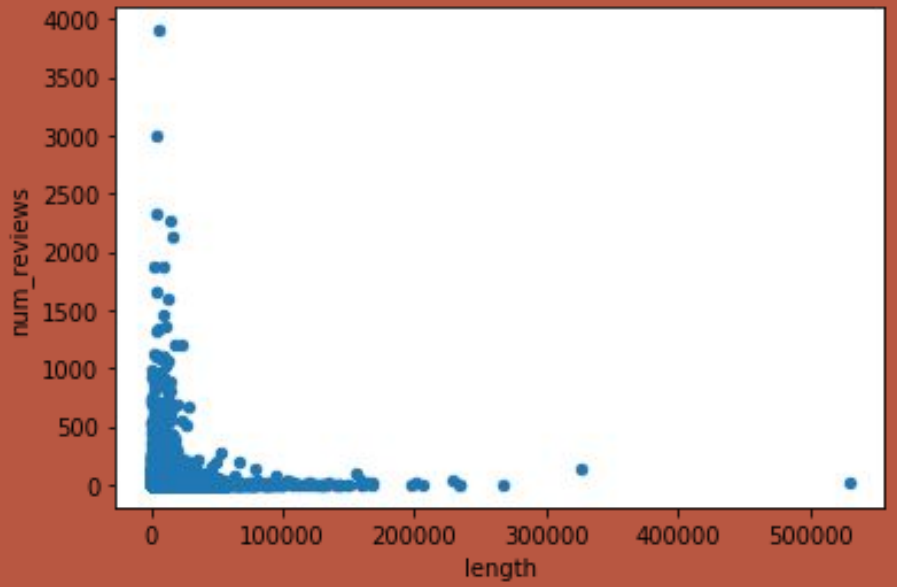
This list is the distribution of trails throughout states.

California holds >20% of all national park trails, while Missouri and Michigan have <.1%

This distribution will have a large effect when comparing other factors.

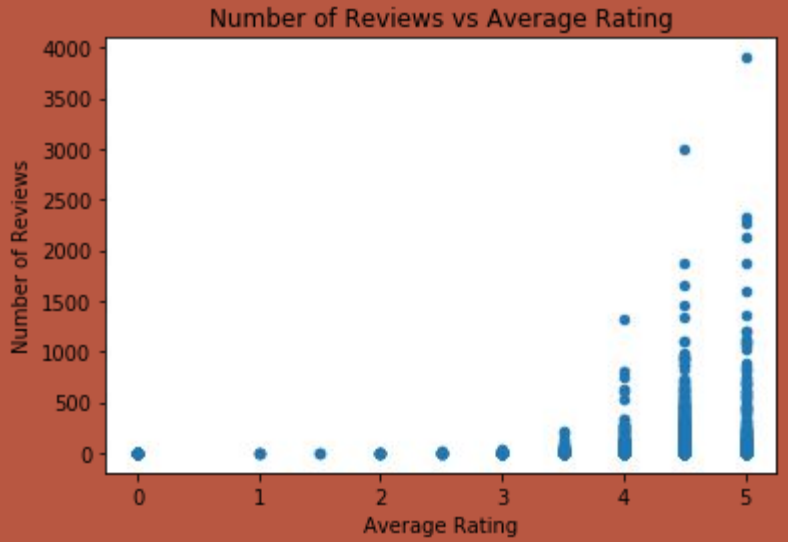
Number of reviews vs trail length

- Large outliers for both length and number of reviews
- Length has a negative relationship with number of reviews
- ⅓ of trails are <15,000 meters (9 miles)



Number of reviews vs average rating

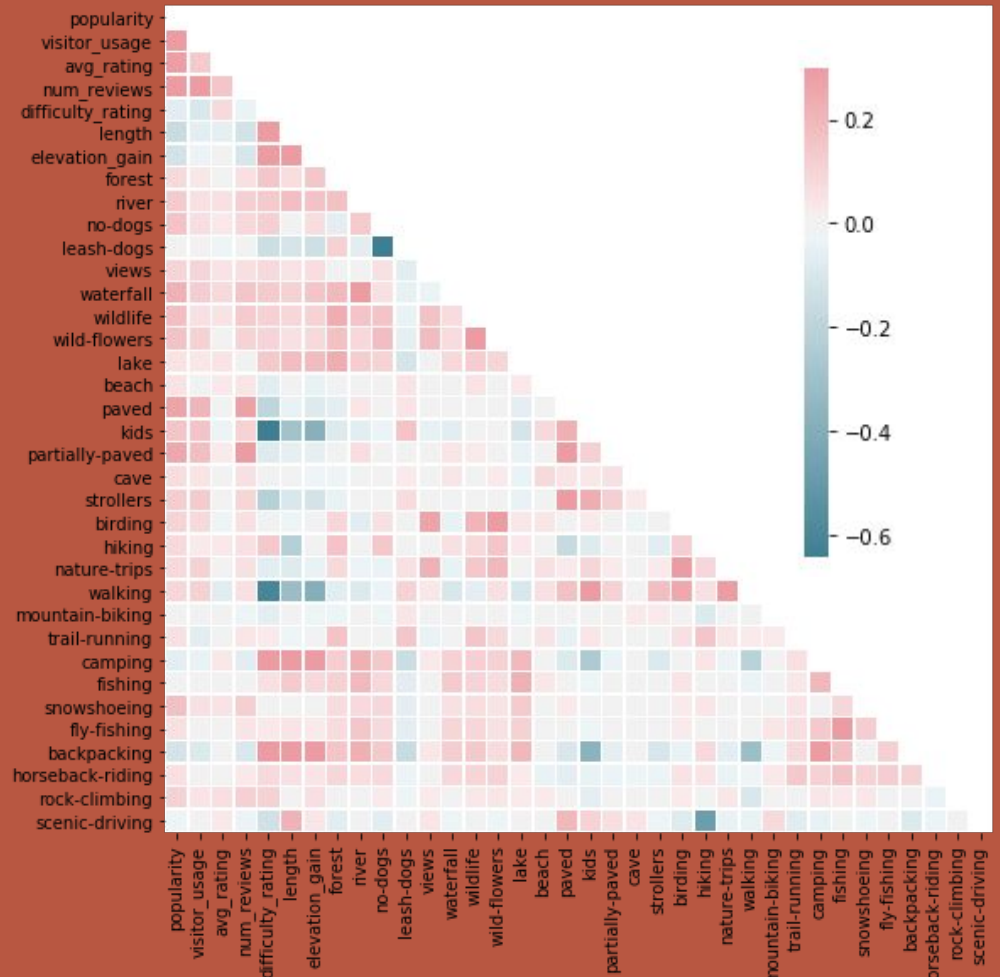
- Reviews increase with average rating
- Better reviews = more visitors = more reviews
- Very few low to mid ranged reviews



Inferential Statistical Analysis

Conclusions:

- Average rating for California's trails are not equal to those of Washington's
- Elevation gain has a strong effect on trail popularity as well as difficulty rating
- Review other correlations in heatmap



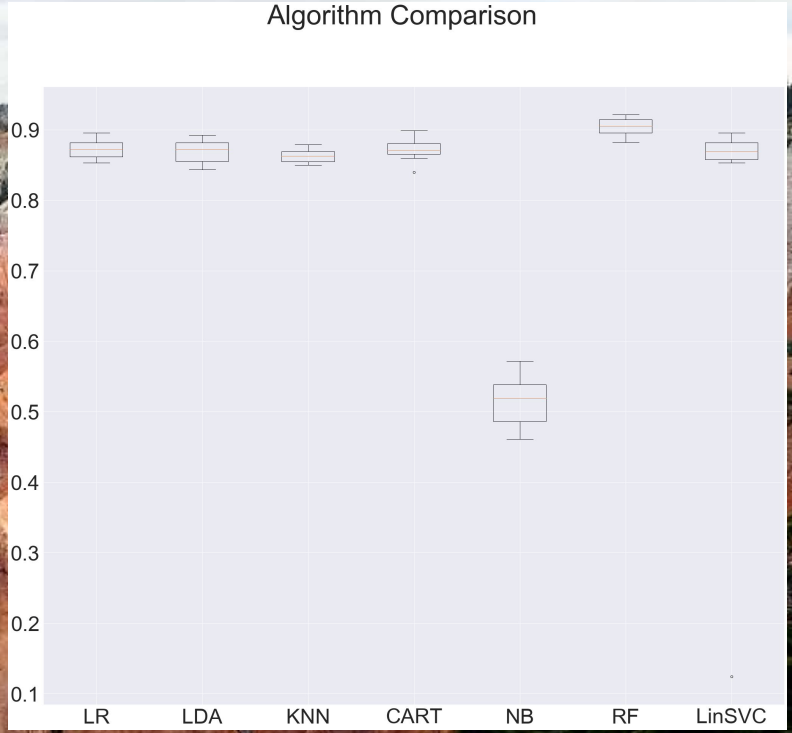
Correlation Heatmap

First columns are where we're most interested. Dark blue indicates a strong negative correlation while dark pink is a strong positive correlation

- No strong relationships with average rating
- Walking and kids have a large negative effect on difficulty while length, elevation gain, camping and backpacking increase difficulty ratings

Machine Learning Models

This graph of accuracies (y-axis, %) was created using a Kfold cross validation function for a series of out of box classifiers. Logistic regression (LR), support vector machine (LinSVC), and random forest (RF) models were chosen to move on with.

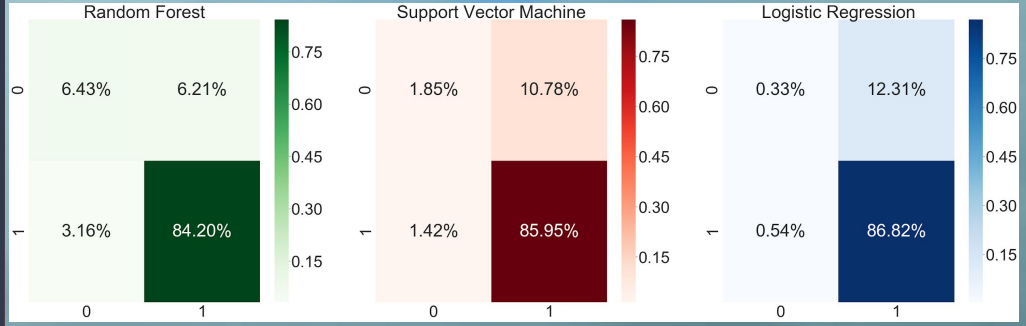


Confusion Matrices

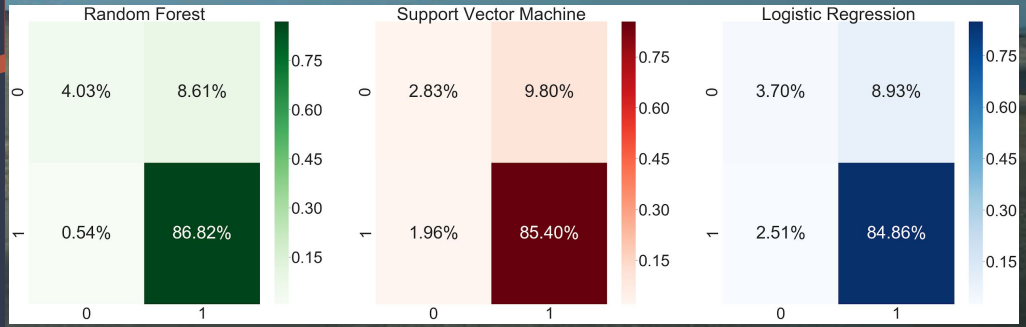
After training and testing on split data sets these are the resulting confusion matrix for each model, both before and after optimizing for accuracy.

Within each 2x2 matrix:
Bottom left: false negatives
Top right: false positives
Bottom right: true positives
Top left: true negatives

Out of box:



GridSearch CV:



Algorithm Metrics

The overall metrics were improved after using GridSearch CV to tune the parameters for each model.

Out of box:

```
Random Forest:
Accuracy  = 90.63
Precision = 93.13
Recall    = 96.38
F1 Score  = 94.73

Support Vector Machine:
Accuracy  = 87.8
Precision = 88.85
Recall    = 98.38
F1 Score  = 93.37

Logistic Regression:
Accuracy  = 87.15
Precision = 87.58
Recall    = 99.38
F1 Score  = 93.11
```

GridSearch CV:

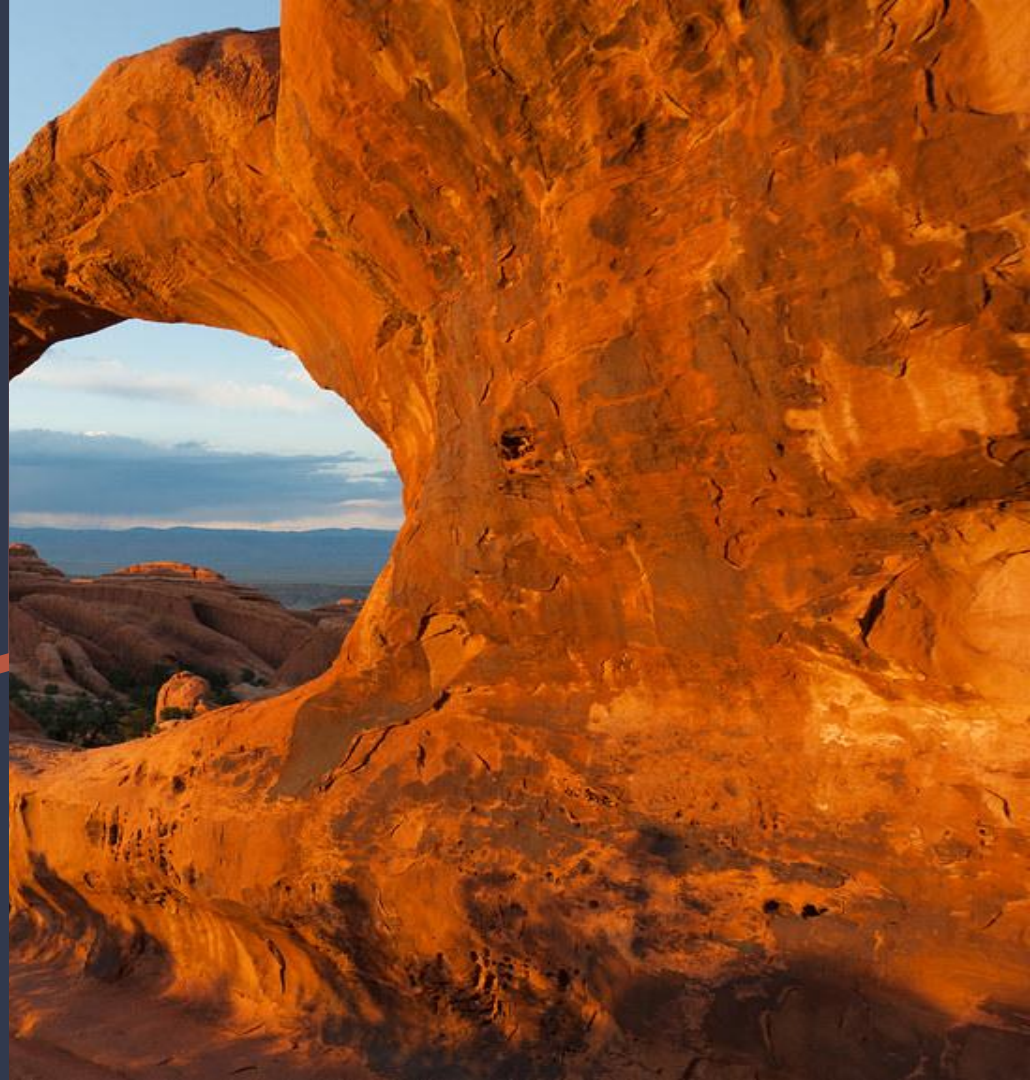
```
Random Forest:
Accuracy  = 90.85
Precision = 90.98
Recall    = 99.38
F1 Score  = 94.99

Support Vector Machine:
Accuracy  = 88.24
Precision = 89.7
Recall    = 97.76
F1 Score  = 93.56

Logistic Regression:
Accuracy  = 88.56
Precision = 90.48
Recall    = 97.13
F1 Score  = 93.69
```

Conclusions

Biggest impacts on average rating:

Negative: walking & length
Positive: number of reviews and waterfalls

The random forest classifier algorithm was the most accurate (90.85%), both before and after parameter tuning.



Future Work

- Train models to predict popularity, visitor usage, and number of reviews
- Include areas and states as independent variables
- Gather data on elevation levels at the trailhead (beginning of trails)
- Gather data on non-national Park trails to compare


Thank you for
your time!