National Parks Trail Capstone
Full report, final
Evan Hintz

This project is focused around understanding national park trail data (U.S.), specifically what trail statistics, features, and activities have the strongest and weakest effects on the trail's popularity, rating, and usage. The purpose is to help learn how and where future trails can best be designed to create the most interest. The process of trail-making is a long and difficult one so having a better idea of what the demand is like may help National Park Services improve overall trail systems.

All of the data was collected by an application that helps users find a fitting trail. The following paper will review how the data was cleaned, organized, explored, and analyzed.

The data was imported and converted into a pandas dataframe. Each of the activities (hiking, camping, biking, etc) and features (waterfalls, wildflowers, viewpoints, etc) were grouped under a single variable. Since we want to see the effects of each feature and activity implementing one hot-encoding to the dataframe will split each one up into fifteen and fourteen separate categorical (boolean) variables of features and activities respectively. Now between these two and the general trail statistics there are a total of thirty four independent variables not including location (area, city, and state).

The data was explored for any missing values using pandas info function; it was found that visitor usage was lacking 253 values. Out of 3313 this is not too significant and while it may change some specifics of how the trails are distributed through the states etc. there will still be plenty of data left to compare the three main dependent variables of interest.

Some data is mislabeled at well; Maui is listed as its own state (it's a county/island within Hawaii) so this has been fixed. The country name for both also needed to be changed from 'Hawaii' to 'United States'.

Digging a bit deeper into the data it was found that Georgia only had one trail listed and it was actually a local park/events space and not within a national park, so Georgia was removed entirely from the data.

The data appears to be clean and organized for the purposes of this project; moving on to exploratory data analysis and storytelling to get a better understanding of the data itself.

Only twenty-nine states contain any national park trails at all, with California having about 20% of them. Location and accessibility will be very important in trail popularity and how often they're used. This makes sense as a large portion of the state (which is quite large) is national park land. California has a great location which also gives it generally nice weather and is already a highly-populated area. Some guesses can be made about variables like elevation gain. For example Florida, North Dakota, and Minnesota are very flat states but summed they only have 55/3313 trails (3060 in the cleaned set). Meanwhile California contains huge areas of mountains and hills.

Another huge factor will be trail length, people have varying levels of skills and physical conditioning when it comes to hiking. Some people go out to push their limits for weeks on end, while others are simply out for a casual sight-seeing stroll. Below is a box plot of the trail lengths in approximate yards. The trails have quite a large range of length, from hundreds of miles to one that is so short it is listed at a length of 0. The biggest outlier is a 329 mile driving loop around the Olympic Peninsula. The second and third are ~200 mile backpacking trail through Yosemite Valley. The fourth is another driving trail (offroad) through Death Valley.

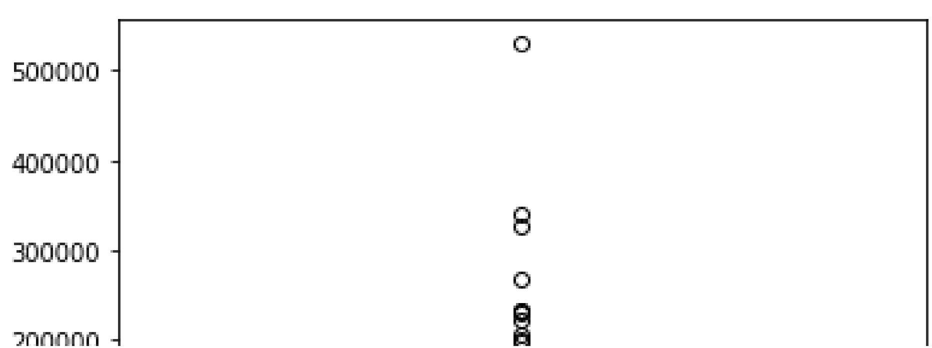
On the other end of the spectrum is the trail with 0 length mentioned earlier. This is the Newspaper Rock Trail in Utah. This ‘trail’ really is more like a large rock that doesn’t require any real hiking at all and is more of a drive-up viewpoint that is covered in Indian Petroglyphs which are a type of ancient carving. This is a fantastic trail option because it is accessible to almost everyone and has a unique feature that could attract people of all ages and interests in outdoor activities. Since having such a particular and uncommon feature, this also needs to be kept in mind when comparing trail popularity as it is not something that is included directly in the data set.

There are a few other trails as well with incredibly short lengths that are designed more as quick road-side stopping points for drivers, for example “Sunset View Trail” which is less than a tenth of a mile long.

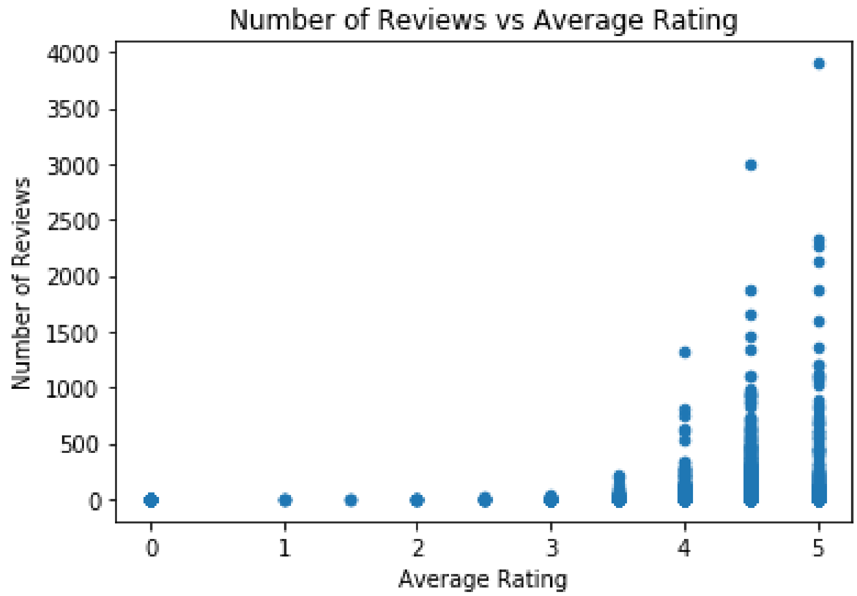
While these are major outliers to the data they are still important points when we’re looking at national park trails. They help demonstrate the wide variety of not only the type of trails but also how they’re used. These data points won’t be taken out of the set but they will be accounted for and potentially excluded when looking at some statistics such as averages.

Note: While there are plenty of trails above the 50,000 and 100,000 mark, a large majority of the trails are found below 50,000 (~28miles). 3108 trails are 28 miles or shorter and 205 trails are longer.

A similar box plot can be seen for change in elevation with similar conclusions. It has a large factor in trail difficulty and in turn can affect consumer’s trail choice. A few very very large outliers and a high density through the lower half of the range (2000 to



5000 ft). This could be interpreted as the demand of trail type; more people want a 'do-able' trail that doesn't require a lot of exertion but it is also restricted by the topography within the parks.

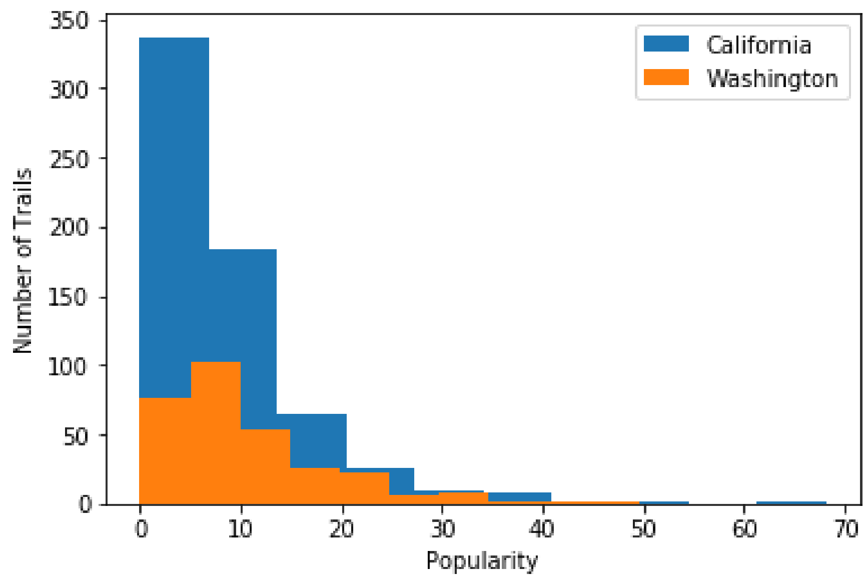


When comparing the number of reviews versus average rating, it's not entirely surprising that there is a positive relationship between the two. When a trail receives good ratings it increases the chances that more people are going to use that trail, increasing both the trail's overall rating as well as usage.

After getting a bit more acquainted with the data it's time to do some actual analysis. First using inferential statistics to look at some other relationships within the data.

Null hypothesis: The average rating for California's trails are equal to those of Washington's
Alternative hypothesis: The average ratings are not equal
Using an alpha of 5%, this two-tailed test will have a z-score of 1.96

After the necessary calculations the Z-test value was -11.29; which is far to the left of the distribution and strongly recommends the null hypothesis to be rejected. This concludes that the average rating for California's trails are not equal to those of Washington's. Similar results were found when comparing popularity and usage.



Above is a histogram of CA and WA's trail popularity. It is obvious here that CA has many more trails, however it's worth noting that both state's distribution of trail popularity are very similar. (Histograms of the other variables look very similar as well)

While CA has more than twice as many trails as WA it might be a good idea to run the same tests with Wyoming and Utah, states that only have a difference of 8 trails. CA also has a much higher population as well as tourist traffic than the other two states. After running through the hypothesis testing program with new inputs the end conclusion is similar (Z-test = -5.33), the rating along with the popularity of the two states' trails are not equal. Oddly enough, the visitor usage does fall within the 1.96 Z-score distribution at .098 for these two states, so we can state with significance that visitor usage between Wyoming and Utah are equal.

Looking at a few more variables, elevation gain, popularity, and the difficulty rating:

Null hypothesis: Elevation gain has no effect on trail popularity
Alternative hypothesis: Elevation gain has a positive or negative effect on popularity

A two tailed test yielded that the t stat is 39.68 indicating that elevation gain has a strong effect on trail popularity. The p-value is far below the alpha (1.36e-306) supporting a rejection of the null hypothesis. A similar test and results can be run with elevation gain and difficulty as well - as expected elevation gain has a significant effect on a trail's difficulty rating.

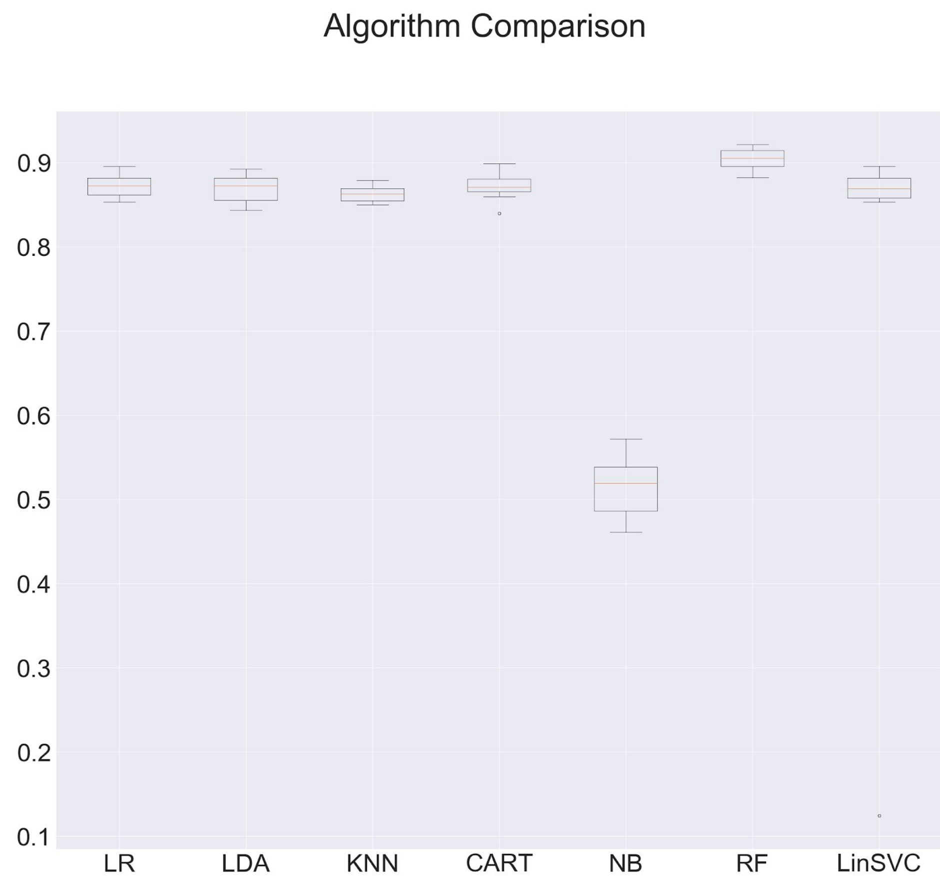
The remainder of this report will be focusing on the machine learning aspect of the National Park Trail analysis project.

The scale for average rating was originally set between 0 and 5 (0 being the worst and 5 being the best possible rating) however it was converted to a binary (low vs high) classification problem where 0-3.5 ratings now became a 'low' rating and 4-5 ratings became a 'high' rating. This new boolean column was used as the dependent variable (y); to be predicted. All trail features and activities were used as the independent variables (X); the predictors. Some columns were excluded including the area names and state names. These could be included in future analysis but would need to be one-hot encoded (each value would need a separate column to be used as predictors).

This project mainly utilizes the SciKit Learn python package for data analysis. In order to test the predictions of the models the data must be split into training and testing data. Using the train_test_split function 70% of the data was used for training and the remaining 30% used for testing. Since some of the features have various scales with their numeric values (for example trail length will have much higher values in general versus trail popularity) a standard MinMaxScaler was used on the independent variables (X) for both the training and testing data so that all data has a normal distribution.

To see how well the average ratings were predicted the model is ran against the testing data (which it did not see at all during the training process) and is evaluated on how well values were predicted. A lot of information is gained on the model's ability, the metrics in the analysis include accuracy, precision, recall, and F1 score. These values are based off of the confusion matrix which includes all values predicted compared to the actual values and whether they are true positive, true negative, false positive, or false negative.

The following graph takes a quick look at the range of accuracies (y-axis) the out of box algorithms (meaning no modifications were made) can achieve on the data set. The graph includes the following models in order from left to right: logistic regression, linear discriminant analysis, K neighbors, decision tree, GaussianNB, random forest, and linear SVC.



The random forest, linear SVC, and logistic regression look like they are doing pretty well. A KFold model was used to produce these results which uses cross-validation so the accuracies are not based on predictions of the X_test data set vs the actual values. We'll take a closer look at the three algorithms by testing them against the split data sets.

Metrics explained further:

- <u>Accuracy</u>: the percentage of all test values that were properly predicted out of all values in the data set
- <u>Precision:</u> the percentage when the model predicts positive (a high rating) and was correct
- <u>Recall:</u> the percentage when the value is positive and the model predicts positive

- <u>F1 Score:</u> the percentage of a weighted average of the true positive rate (recall) and precision

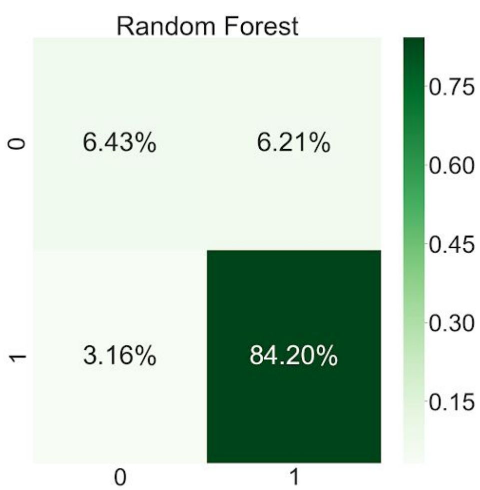
```
Random Forest:
Accuracy  = 90.63
Precision = 93.13
Recall    = 96.38
F1 Score  = 94.73
```

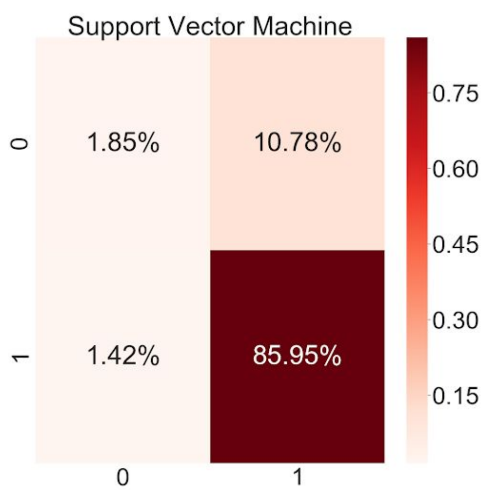
```
Support Vector Machine:
Accuracy  = 87.8
Precision = 88.85
Recall    = 98.38
F1 Score  = 93.37
```

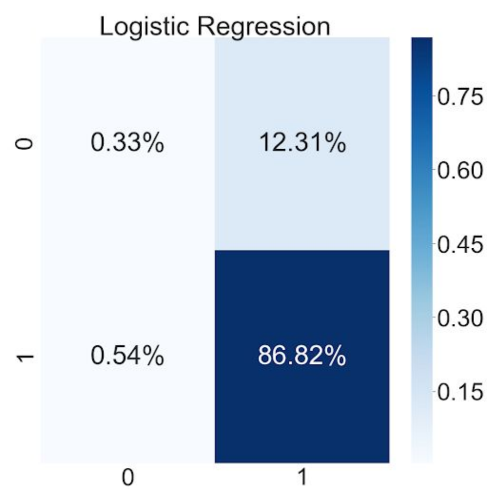
```
Logistic Regression:
Accuracy  = 87.15
Precision = 87.58
Recall    = 99.38
F1 Score  = 93.11
```

These metrics are based on the out of box models for each algorithm. Overall the random forest is doing a significantly better job predicting.

The following graphs are the confusion matrices for each model (where the above metrics are calculated from). The top left square represents true negatives (low rates that were properly predicted), the bottom right true positives (high rates that were properly predicted), these two percentages add up to the accuracy of the model. The top right and bottom left represent false positives and false negatives respectively, together they're the total percentage of incorrectly predicted values.







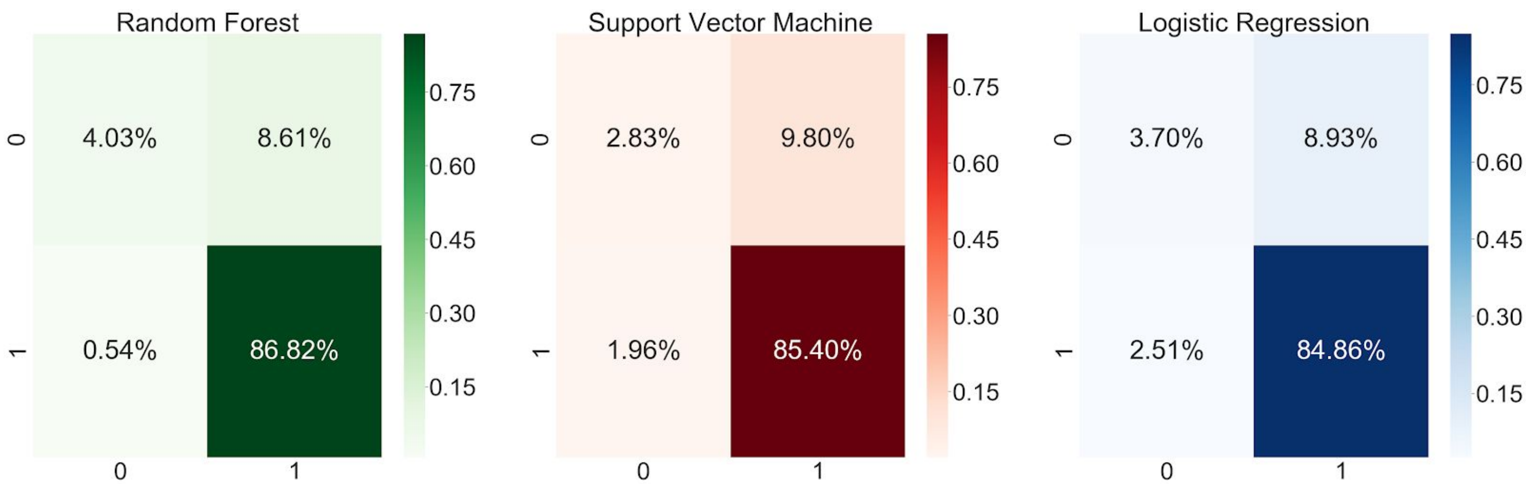
The next step is to tune the parameters for each model to see which can become the best predictor for average trail ratings. To do this without having to manually tune each parameter (each model can contain a multitude of relevant parameters) another SciKit Learn function is employed; GridSearch CV. This function goes through fitting the training data to the model with each possible parameter that it's given and each combination of them in order to find the most accurate.

The following values and confusion matrices are based on the optimized models using the GridSearch CV method.

```
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
```



As stated, these models are tuned to make the most accurate predictions, it's important to note some of the other metrics can fall as a result. It is also interesting to note that for the random forest algorithm it improved the accuracy by only .22% and it did so by doing a better job with true positives, however its percentage of true negatives predicted actually fell. The opposite was true for the support vector machine and the logistic regression model.

The random forest remains as the best overall predictive model and does a fairly good job, being able to correctly predict 90.85% of the data as low or high. While this is the clear winner between these three models it may be worth it to look into some of the other classifiers and their parameters for possible improvements on the random forest classifier.

This dataset invites a multitude of different analyses all with many insights and possible interpretations. I would like to continue exploring these by looking at the effects of the variables on the number of reviews, trail rating, visitor usage, and maybe most importantly the popularity. I think future work should include designing models around these other variables as dependents to see how they compare with rating. It may also benefit the accuracy of the models to include the areas and state names as independent predictor variables. It may be interesting as well to see this work expanded to non-national park trails both to compare any differences but most likely simply to increase and improve the data size.