Applications of Predictive Modeling on the Appalachian Trail

A Presentation by: Chris Campell of Appalachian State University

Mentored by: Dr. Mitch Parry

Presented at: NCUR 2017

Inspiration and Background

Research Goals and Applications

- Goals:
  - Create a machine learning prediction algorithm to assist backpackers hiking the Appalachian Trail
    - The prediction model estimates how far a backpacker will go (in miles-per-day) along the trail
    - This distance prediction enables backpackers to plan in advance where they will spend the night
- Applications:
  - The creation of an Appalachian Trail Guide mobile application

Existing Solutions

- Mobile applications such as BACKPACKER GPS Trails [1] which provides the user with the following features:
  - The ability to save a series of GPS points for offline use
  - The ability to save a topographical map for offline use
  - The ability to locate the users location via GPS
- Existing backpacking mobile applications do not provide:
  - Distance predictions for hikers
  - Estimation of arrival time at shelters

Methodology: Model Formulation

Model Formulation

- It was hypothesized that the distance in miles-per-day of a hiker could be calculated as:
  - $\widehat{miles\_per\_day} \sim$ user_bias + location_direction + intercept_constant
    - Where user_bias is a measure of the user's miles-per-day in relation to the average hiker along a given segment of the Appalachian Trail.
    - Where location_direction is a unique identifier for a specific shelter and an associated direction of travel along the trail (Northbound or Southbound).
    - Where intercept_constant is the y-intercept calculated by the Ordinary Least Squares regression as the average miles-per-day of hikers in the dataset.

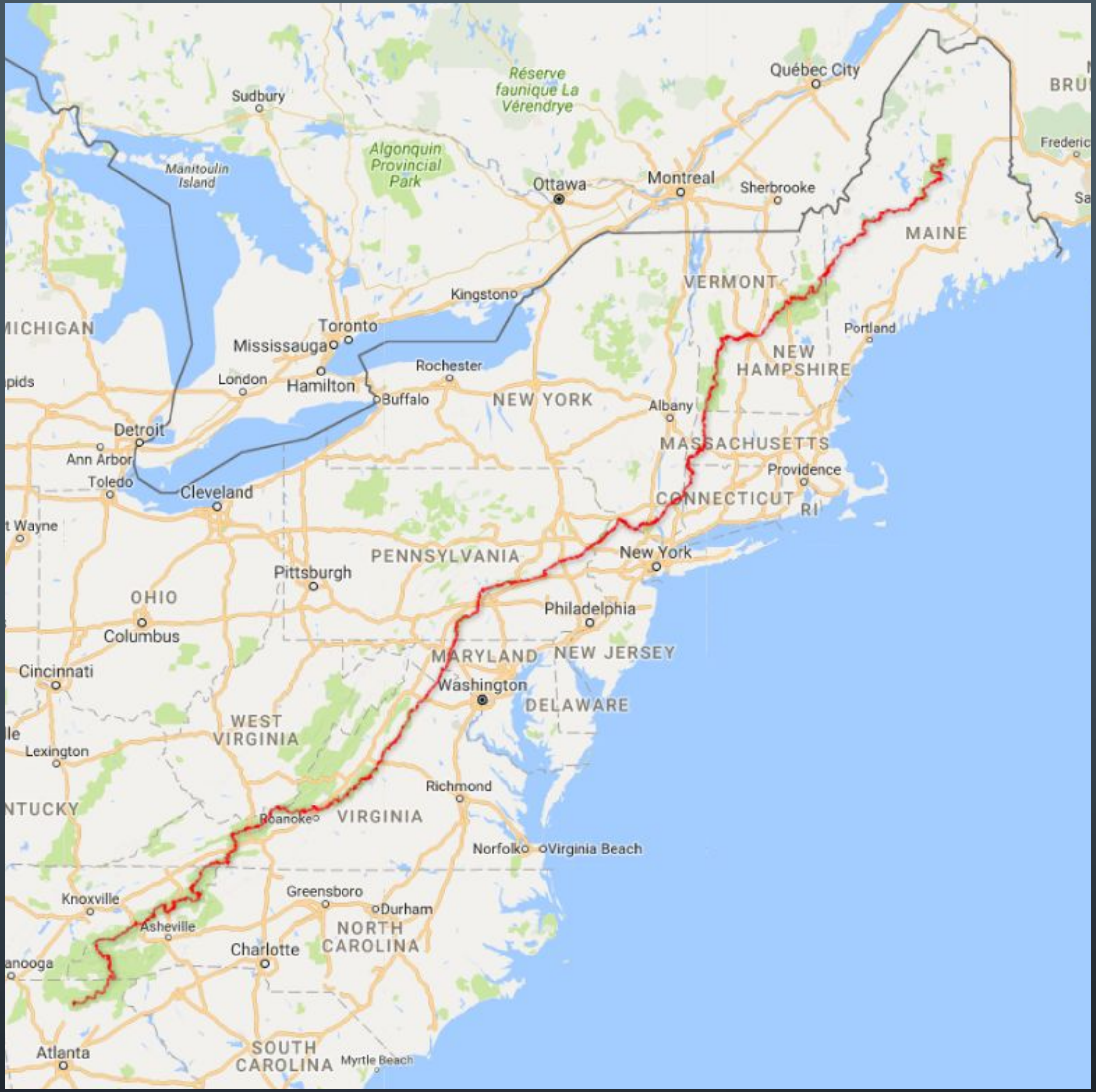
Training the Model

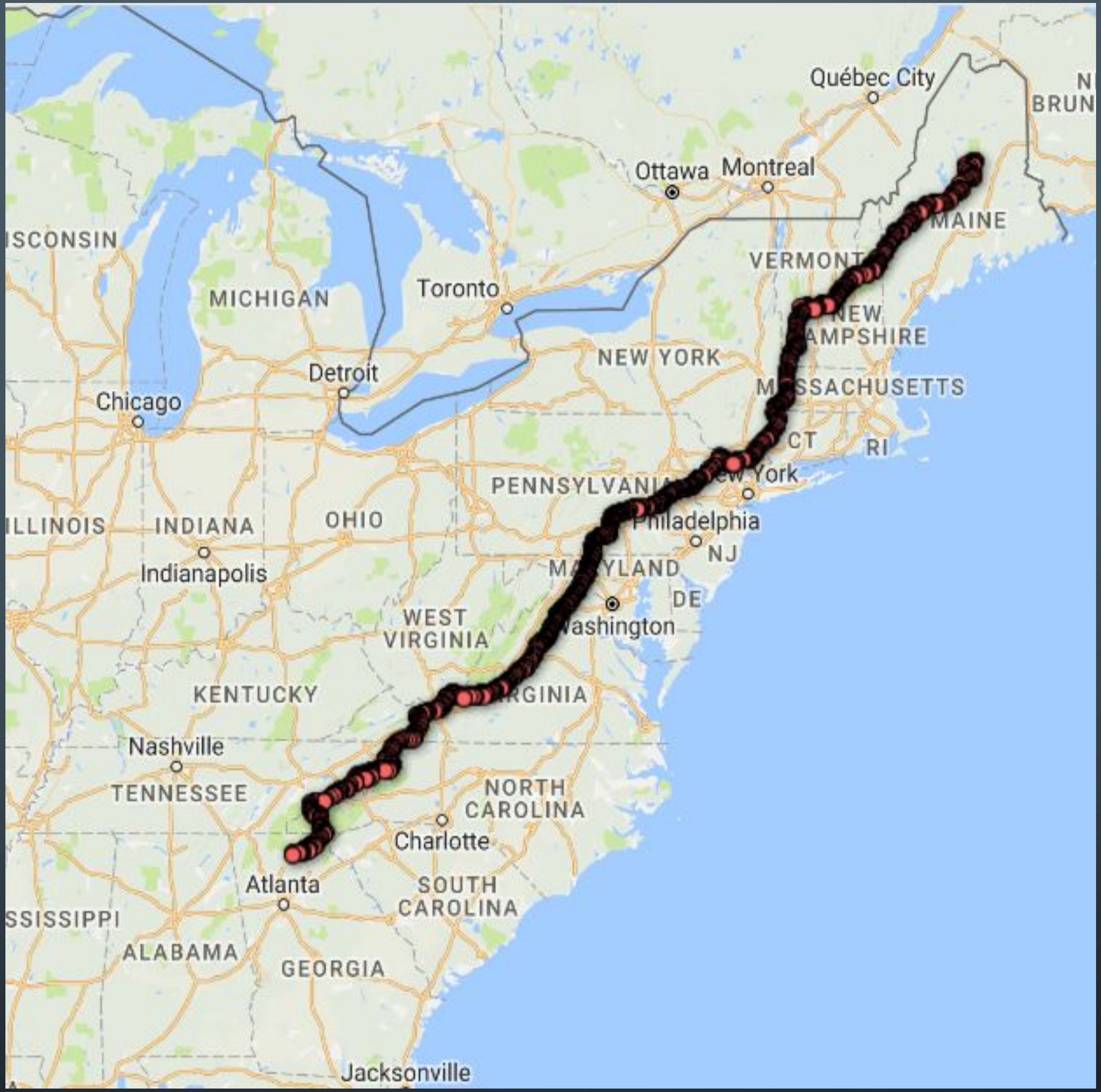
- The model was trained using the StatsModels module [2] in conjunction with Numpy [3]
- There are three main datasets which in conjunction form the training dataset utilized by the Hiker Distance Prediction Model:
  - The Appalachian Trail Profile (ATP) Dataset
  - The Appalachian Trail Shelters (ATS) Dataset
  - The Appalachian Trail Hiker (ATH) Dataset
- The required data was stored in CSV form, then loaded into memory using the Pandas [4] module in Python.

Methodology: Data Mining

The Appalachian Trail Profile (ATP) Dataset

- The Appalachian Trail Profile (ATP) Dataset is comprised of the 279,406 GPS coordinates which make up the Appalachian Trail Centerline
- GPS data was obtained from the Appalachian Trail Conservancy [5] in Keyhole Markup Language (KML) form


Sudbury
Réserve faunique La Vérendrye
Québec City
BRU
Frederic
Sa
Manitoulin Island
Algonquin Provincial Park
Ottawa
Montreal
Sherbrooke
MAINE
Kingston
VERMONT
Portland
MICHIGAN
Toronto
Mississauga
NEW HAMPSHIRE
London
Hamilton
Rochester
pids
Buffalo
NEW YORK
Albany
Detroit
MASSACHUSETTS
Ann Arbor
Toledo
Providence
Cleveland
CONNECTICUT
RI
t Wayne
PENNSYLVANIA
New York
Pittsburgh
OHIO
Philadelphia
Columbus
Cincinnati
MARYLAND
NEW JERSEY
Washington
DELAWARE
WEST VIRGINIA
le
Lexington
Richmond
NTUCKY
Roanoke
VIRGINIA
Norfolk
Virginia Beach
Greensboro
Knoxville
Durham
Asheville
NORTH CAROLINA
Charlotte
anooga
Atlanta
SOUTH CAROLINA
Myrtle Beach



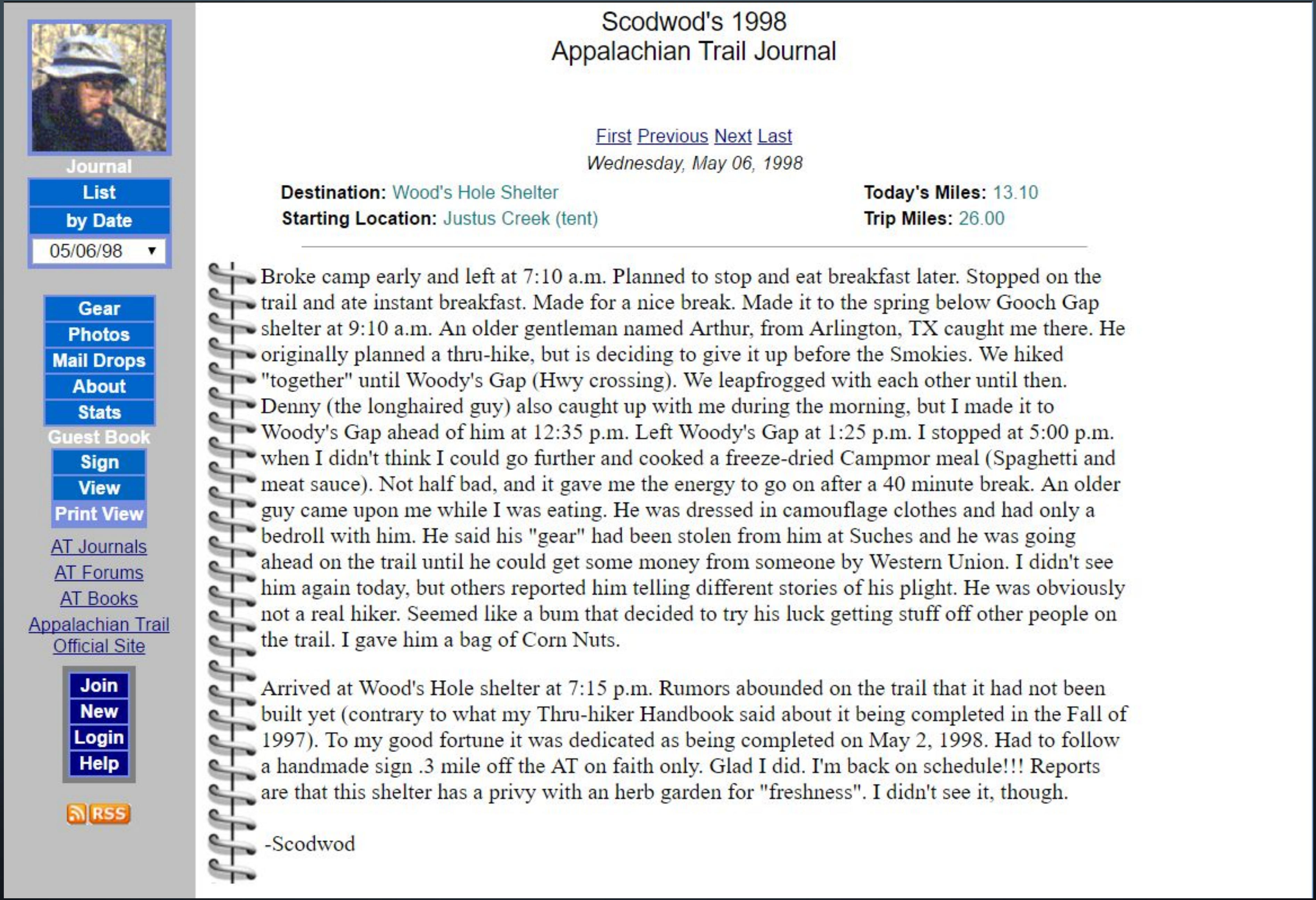
Québec City
Ottawa
Montreal
MAINE
VERMONT
NEW HAMPSHIRE
MASSACHUSETTS
CT
RI
NEW YORK
New York
PENNSYLVANIA
Philadelphia
NJ
MARYLAND
DE
Washington
VIRGINIA
WEST VIRGINIA
KENTUCKY
NORTH CAROLINA
Charlotte
SOUTH CAROLINA
Atlanta
GEORGIA
ALABAMA
TENNESSEE
Nashville
MISSISSIPPI
OHIO
INDIANA
Indianapolis
ILLINOIS
Chicago
MICHIGAN
Detroit
Toronto
WISCONSIN
Jacksonville

The Appalachian Trail Shelters (ATS) Dataset

- Contains information about shelters along the Appalachian Trail
- The Appalachian Trail Shelters (ATS) dataset was constructed by manually combining the Appalachian Trail Conservancy’s shelter data [5], with the Tennessee Landforms’ shelter dataset [6]
- The resulting ATS dataset contains 287 shelters along the Appalachian Trail
- Each shelter has an associated: GPS coordinate, unique identifier, shelter name and shelter type (e.g. Log, Lean-To, Stone, etc…)

The Appalachian Trail Hiker (ATH) Dataset

- The Appalachian Trail Hiker (ATH) dataset was the most challenging to obtain.
- Data was mined via web scraper from the Trail Journals website [7]
- Each hiker has several key pieces of information:
  - Hiker Name
  - Hiker Trail-Name
  - Hiker Trail Start Date
  - **Hiker Trail Journal**
- No validation is performed on user-entered data by the Trail Journals'



Scodwod's 1998
Appalachian Trail Journal

First Previous Next Last

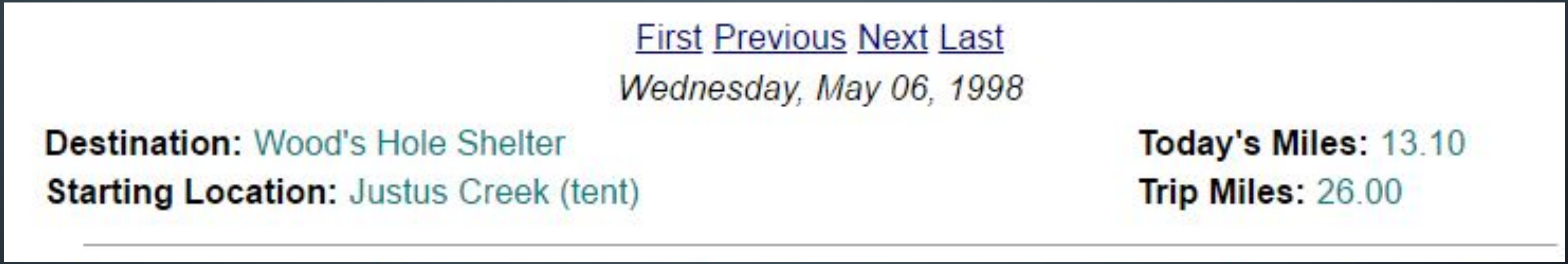
Wednesday, May 06, 1998

Destination: Wood's Hole Shelter	**Today's Miles:** 13.10
Starting Location: Justus Creek (tent)	**Trip Miles:** 26.00

Broke camp early and left at 7:10 a.m. Planned to stop and eat breakfast later. Stopped on the trail and ate instant breakfast. Made for a nice break. Made it to the spring below Gooch Gap shelter at 9:10 a.m. An older gentleman named Arthur, from Arlington, TX caught me there. He originally planned a thru-hike, but is deciding to give it up before the Smokies. We hiked "together" until Woody's Gap (Hwy crossing). We leapfrogged with each other until then. Denny (the longhaired guy) also caught up with me during the morning, but I made it to Woody's Gap ahead of him at 12:35 p.m. Left Woody's Gap at 1:25 p.m. I stopped at 5:00 p.m. when I didn't think I could go further and cooked a freeze-dried Campmor meal (Spaghetti and meat sauce). Not half bad, and it gave me the energy to go on after a 40 minute break. An older guy came upon me while I was eating. He was dressed in camouflage clothes and had only a bedroll with him. He said his "gear" had been stolen from him at Suches and he was going ahead on the trail until he could get some money from someone by Western Union. I didn't see him again today, but others reported him telling different stories of his plight. He was obviously not a real hiker. Seemed like a bum that decided to try his luck getting stuff off other people on the trail. I gave him a bag of Corn Nuts.

Arrived at Wood's Hole shelter at 7:15 p.m. Rumors abounded on the trail that it had not been built yet (contrary to what my Thru-hiker Handbook said about it being completed in the Fall of 1997). To my good fortune it was dedicated as being completed on May 2, 1998. Had to follow a handmade sign .3 mile off the AT on faith only. Glad I did. I'm back on schedule!!! Reports are that this shelter has a privy with an herb garden for "freshness". I didn't see it, though.

-Scodwod

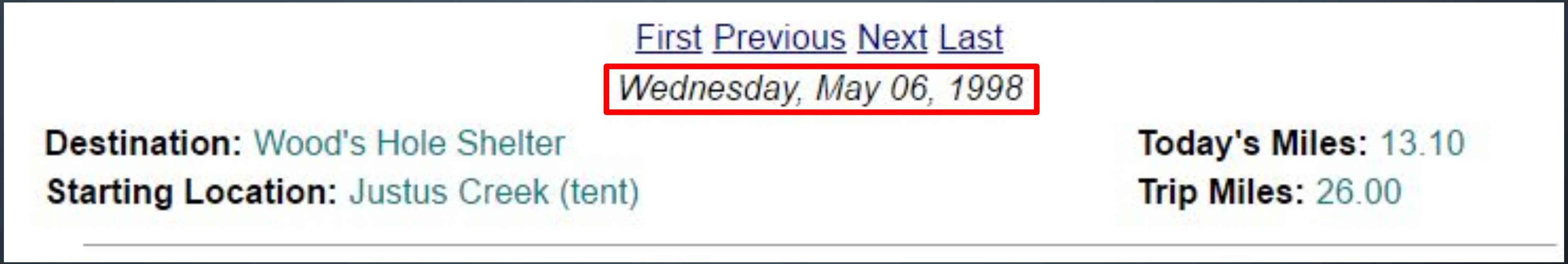


First Previous Next Last

Wednesday, May 06, 1998

Destination: Wood's Hole Shelter **Today's Miles:** 13.10

Starting Location: Justus Creek (tent) **Trip Miles:** 26.00



First Previous Next Last

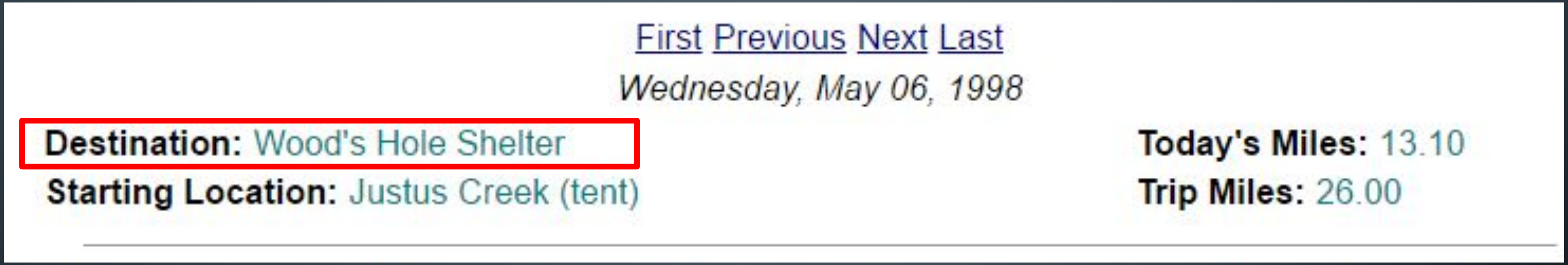
Wednesday, May 06, 1998

Destination: Wood's Hole Shelter

Starting Location: Justus Creek (tent)

Today's Miles: 13.10

Trip Miles: 26.00



First Previous Next Last

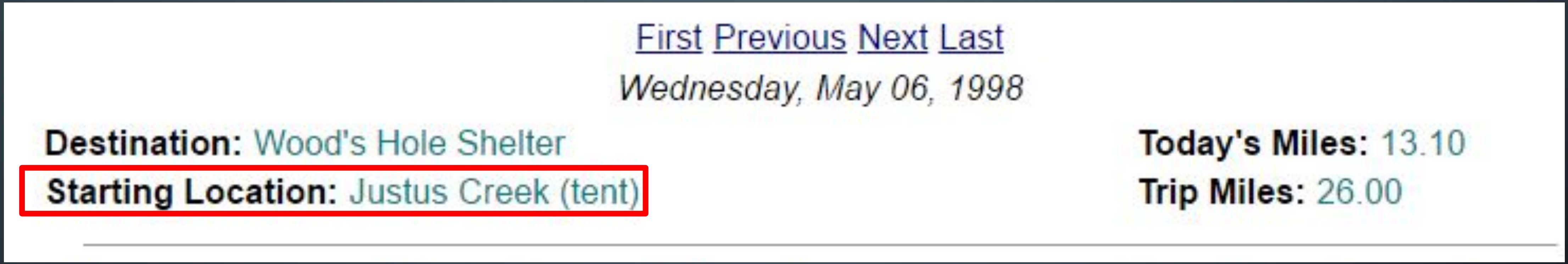
Wednesday, May 06, 1998

Destination: Wood's Hole Shelter

Starting Location: Justus Creek (tent)

Today's Miles: 13.10

Trip Miles: 26.00



First Previous Next Last

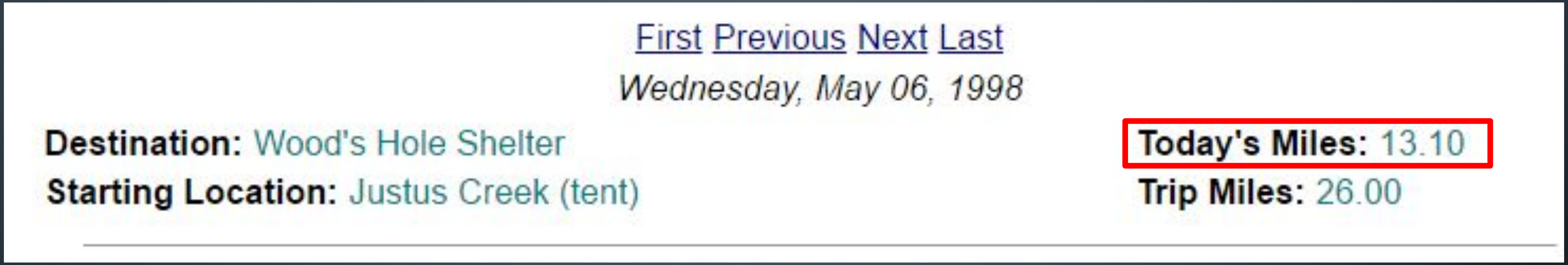
Wednesday, May 06, 1998

Destination: Wood's Hole Shelter

Today's Miles: 13.10

Starting Location: Justus Creek (tent)

Trip Miles: 26.00



First Previous Next Last

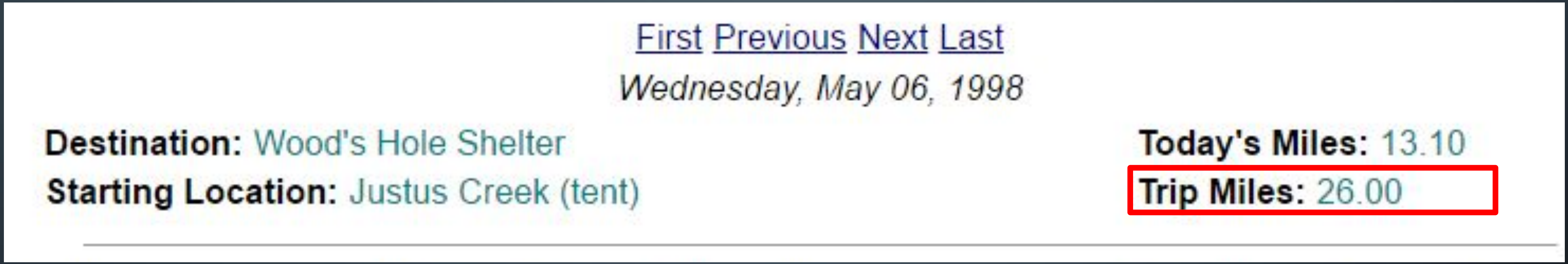
Wednesday, May 06, 1998

Destination: Wood's Hole Shelter

Starting Location: Justus Creek (tent)

Today's Miles: 13.10

Trip Miles: 26.00



Wednesday, May 06, 1998

Destination: Wood's Hole Shelter

Starting Location: Justus Creek (tent)

Today's Miles: 13.10

Trip Miles: 26.00

The Appalachian Trail Hiker (ATH) Dataset

- A web scraper was utilized in order to obtain the hikers' information

- This web scraper functioned in three distinct phases:

  1. Obtained hiker identifiers and hiker URL
  2. Obtained hiker information and trail journal URL
  3. Obtained hiker trail journal entries

- The web scraper was created using the Scrapy module's Selector functionality [8] in conjunction with Xpath [9] and Python's native urllib module [10]

The Appalachian Trail Hiker (ATH) Web Scraper

http://www.trailjournals.com/entry.cfm?trailname=X

The Appalachian Trail Hiker (ATH) Web Scraper

The Appalachian Trail Hiker (ATH) Web Scraper

Location Mapping the ATH Dataset

- User-entered location data was mapped to existing shelters in the Appalachian Trail Shelters dataset
- This was accomplished via a comparison algorithm known as Fuzzy String Matching [11] utilizing the Python package: FuzzyWuzzy [12]
    - Fuzzy String Comparison allowed for multiple entries to be interpreted the same:
        - “Black Rock Shelter” and “Black Rock” will be mapped in the same way
    - Through experimentation, a comparison threshold of 95% was determined

Location Mapping the ATH Dataset

The Appalachian Trail Hiker (ATH) Dataset

- Results of the Web Scraping Process:
  - 2,260 registered users retrieved during stage one of the web scraper
  - 1,882 users/hikers with an existing trail journal for the Appalachian Trail
  - 1,817 hikers with trail journals after removing entries with unmapped starting locations
- Results of the Location Mapping Process:
  - 118,200 trail journal entries with mapped starting locations

Methodology: Noise Removal

Noise in the ATH Dataset

- Every hiker had miles-per-day between consecutive mapped starting locations calculated and stored
  - Elapsed number of days between entries was calculated using Python’s native datetime module [13]

Outlier Removal

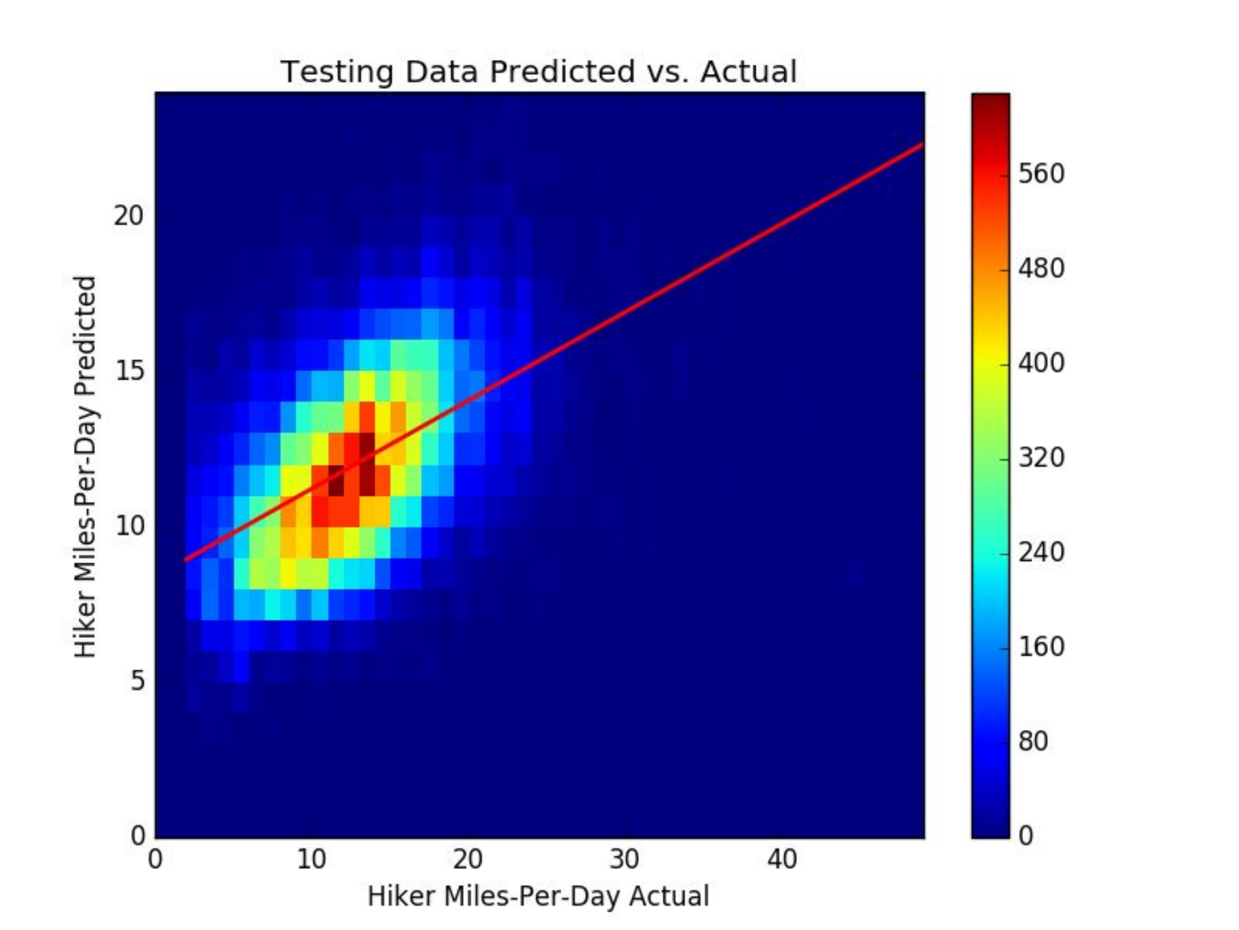
- Journal entries were removed from the training dataset when:
  - Elapsed mileage calculations resulted in an entry with negative miles-per-day
  - Calculated distance between consecutive mapped starting locations was less than 5 miles-per-day.
  - Calculated distance between consecutive mapped starting locations was greater than 25 miles-per-day
- A hiker with negative miles per day is usually indicative of a user who did not insert journal entries chronologically.
- A hiker with greater than 25 miles-per-day or less than 5 miles-per-day was a significant outlier which negatively impacted model performance.

Results: Model Evaluation

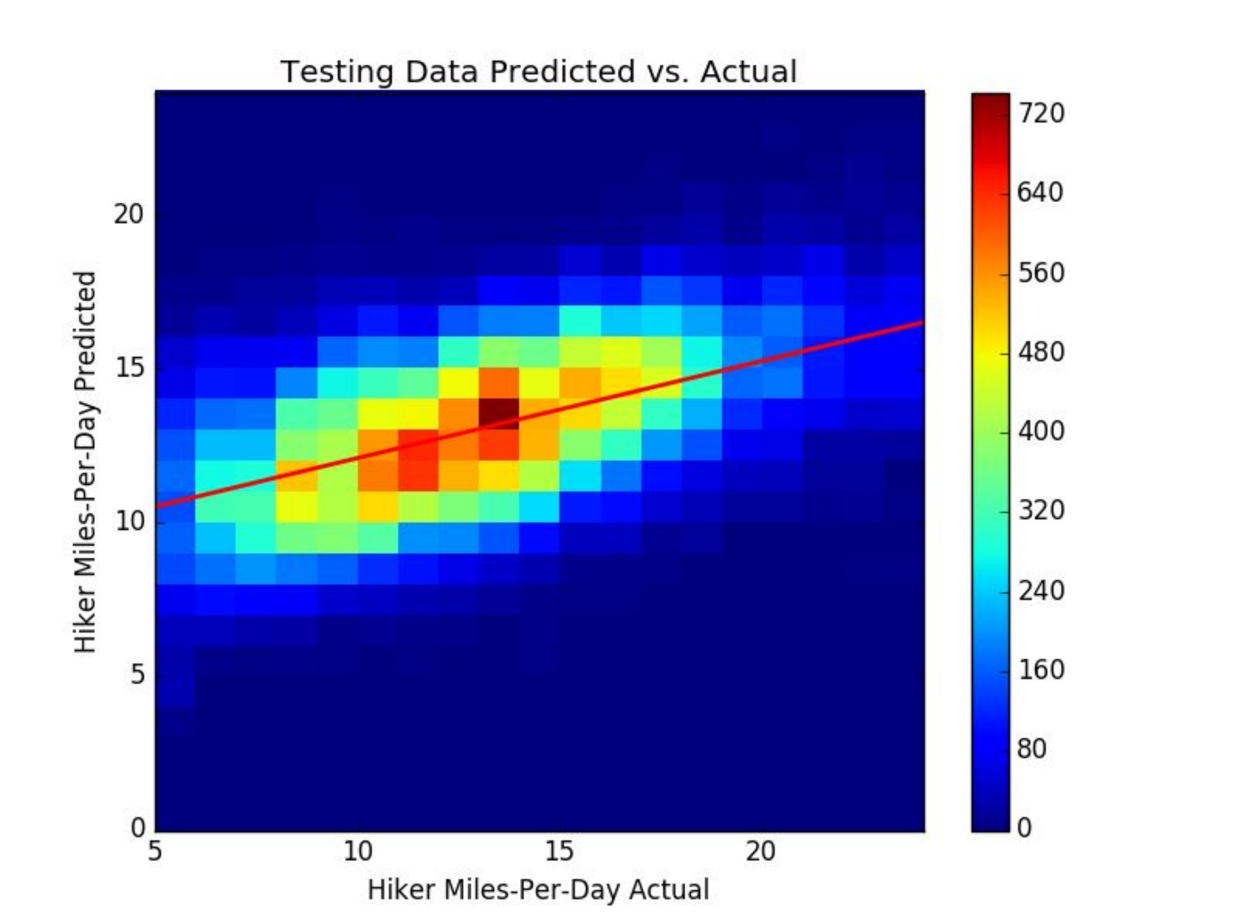
Model Evaluation

- The standard practice of cross validation was followed to avoid overfitting the model to the training data
  - Cross validation was performed using the Sklearn [14] module in Python
  - Sixty percent of the data was retained as a test dataset
  - Forty percent of the data was utilized in the training dataset
- The Root Mean Squared Error (RMSE) of the hiker distance prediction model is 3.65 miles-per-day
  - This means that the difference between the distance predicted by the model and the distance actually traveled by hikers in the test dataset is on average ± 3.65 miles-per-day.

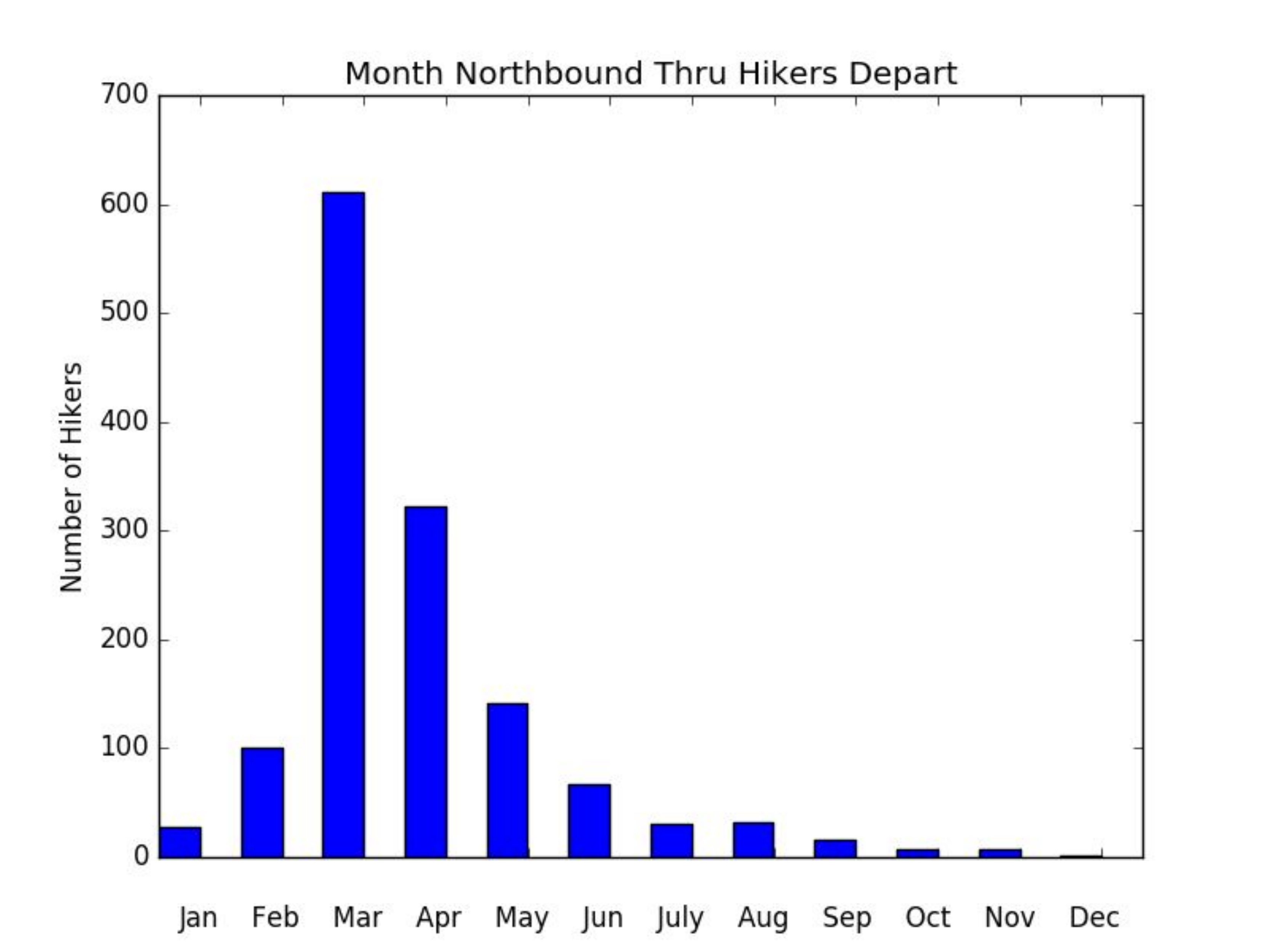
Model Evaluation

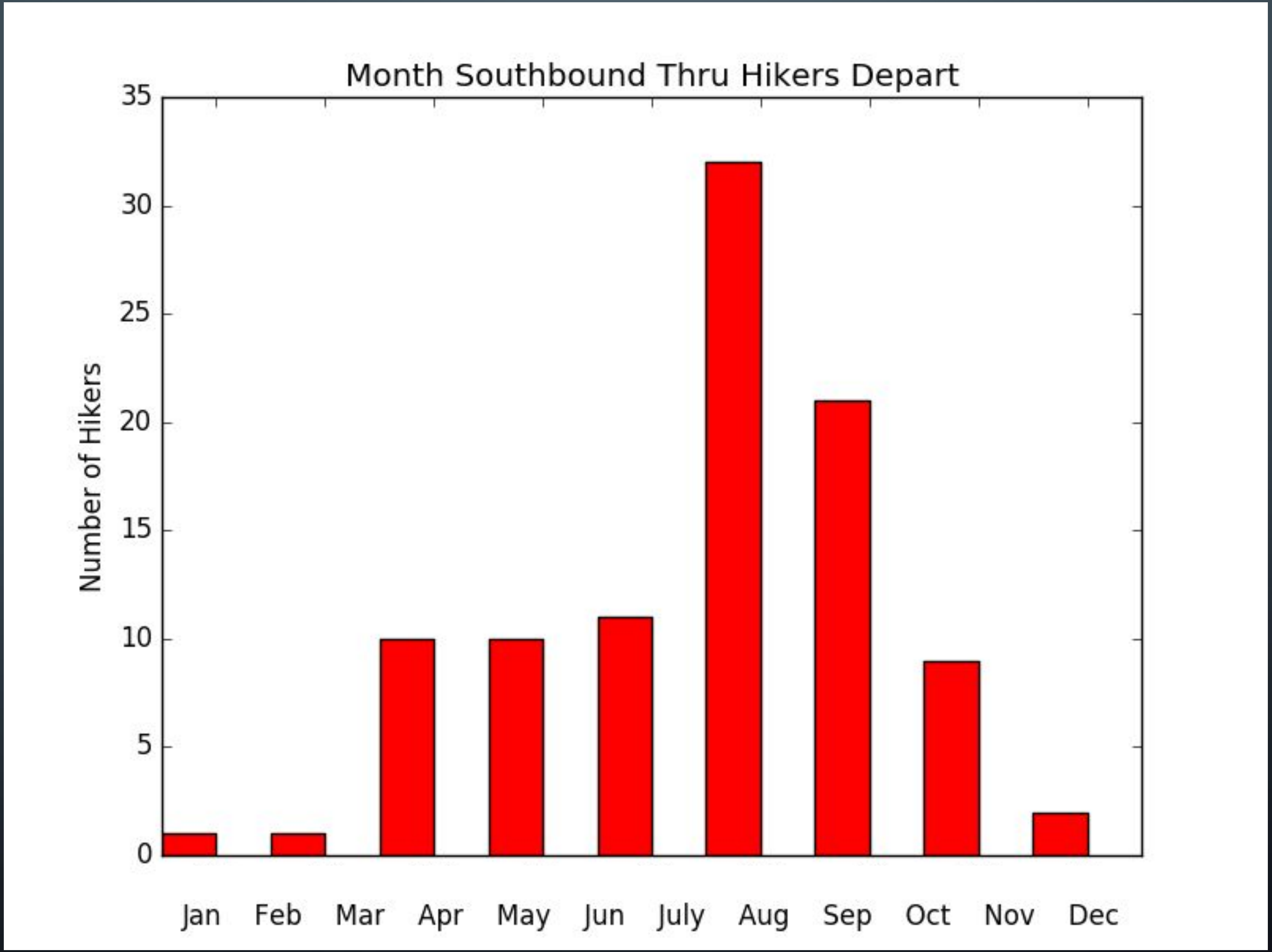


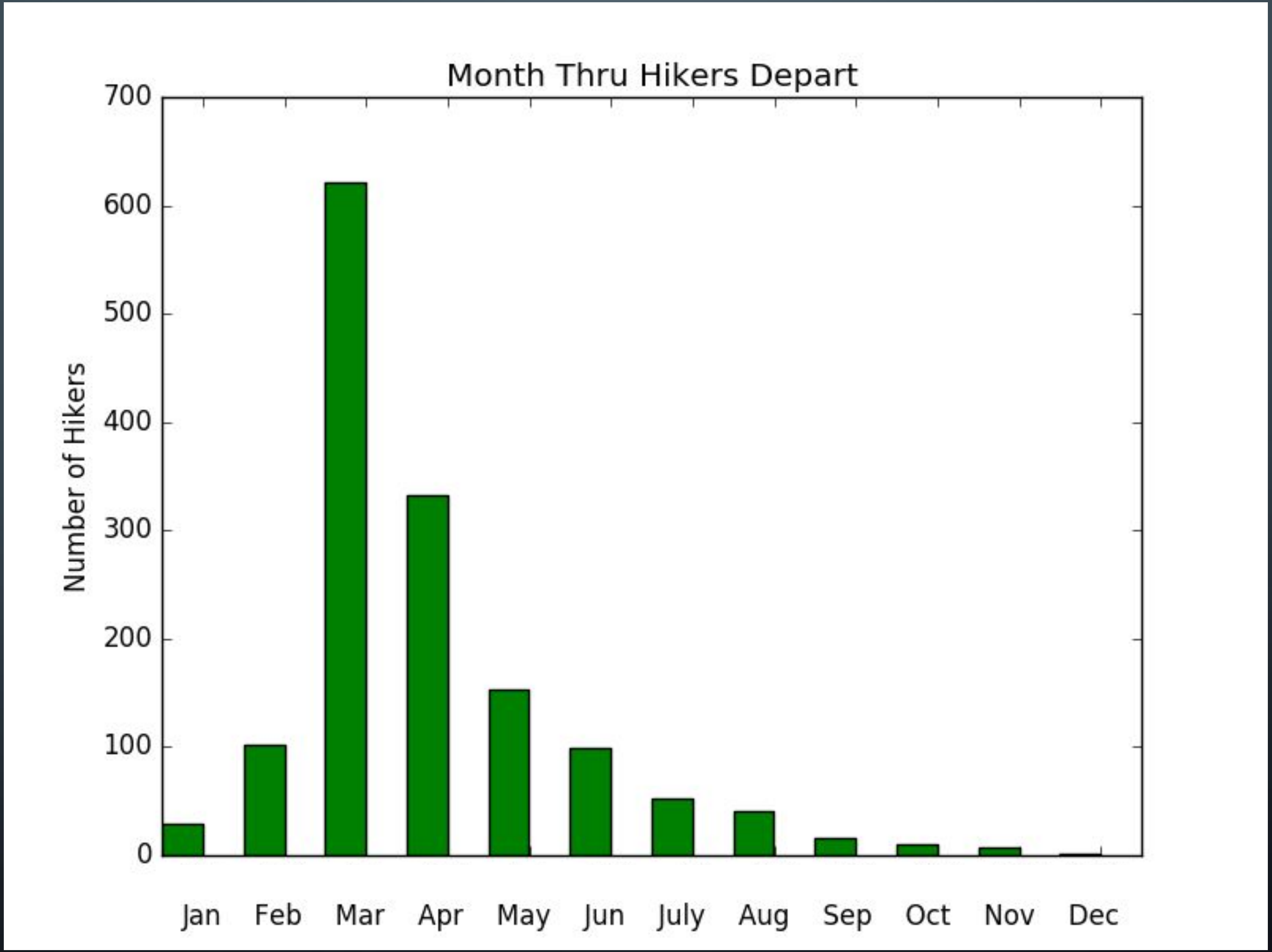
Model Evaluation

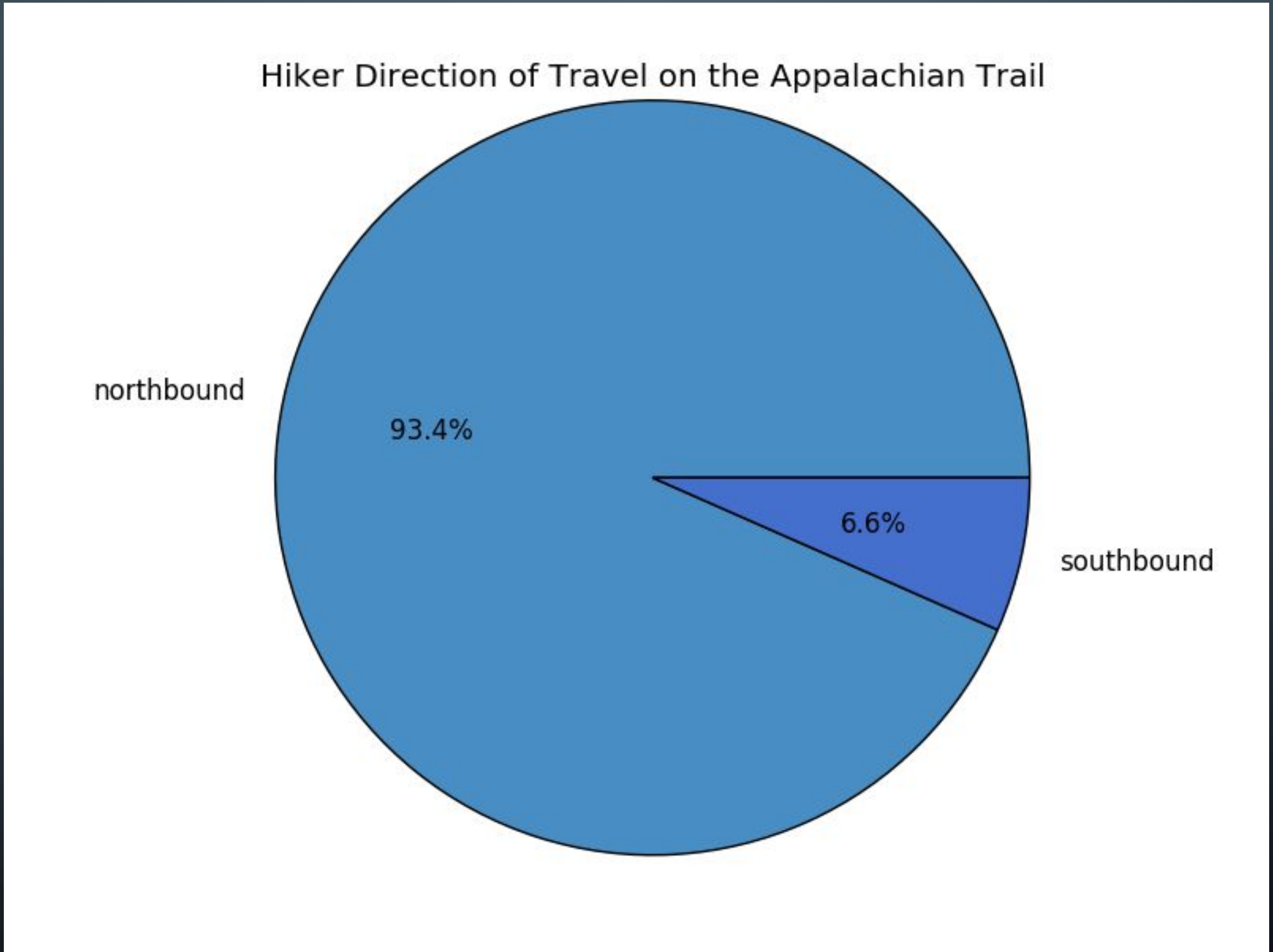


Results: Interesting Findings


Month Northbound Thru Hikers Depart
Number of Hikers
0
100
200
300
400
500
600
700
Jan
Feb
Mar
Apr
May
Jun
July
Aug
Sep
Oct
Nov
Dec


Month Southbound Thru Hikers Depart
Number of Hikers
0
5
10
15
20
25
30
35
Jan
Feb
Mar
Apr
May
Jun
July
Aug
Sep
Oct
Nov
Dec


Month Thru Hikers Depart
Number of Hikers
0
100
200
300
400
500
600
700
Jan
Feb
Mar
Apr
May
Jun
July
Aug
Sep
Oct
Nov
Dec


Hiker Direction of Travel on the Appalachian Trail
northbound
93.4%
6.6%
southbound

Results: Conclusions and Future Work

Conclusions and Future Work

- Conclusions:
  - Although an RMSE of 3.65 miles-per-day is better than the alternative of having no distance prediction, there is certainly room for improvement
  - Regardless, the creation of such a model proves it is possible to estimate hiker distance using only hiker journal text (shelter locations and daily mileage) as a predictor.
- Future Work:
  - Utilize the RANSAC algorithm for better outlier removal
  - Sort hiker trail journal entries chronologically instead of by order-of-insertion
  - Create a mobile application to display the prediction model superimposed onto a topographic map of the trail

Works Cited

1. Cruz Bay Publishing Inc. (2015). *Backpacker GPS Trails for Android* [Website]. Available: http://www.backpacker.com/gear/backpacker-gps-trails-for-android
2. (2017, March 5). *Stats Models* [Online Python Code Module]. Available: http://www.statsmodels.org/stable/index.html
3. NUMFocus (2017 April 2). *Numpy* [Online Code Repository]. Available: http://www.numpy.org/
4. NUMFocus (2016, December) *Pandas: Python Data Analysis Library* [Online Python Code Module]. Available: http://pandas.pydata.org/
5. The Appalachian Trail Conservancy and The National Park Service Appalachian Trail Park Office. (2002-2014). *Appalachian Trail GIS Data* [Online Dataset]. Available: http://www.appalachiantrail.org/explore-the-trail/gis-data
6. T. Dunigan. (2017, January 6). *Appalachian Trail Shelters* [Online Dataset]. Available: https://tnlandforms.us/at/
7. Trailjournals LLC. (2017, March 26). *Appalachian Trail Backpacking Journals* [Online Forum]. Available: http://www.trailjournals.com/journals/appalachian_trail/
8. Scrapinghub (2017, April 5). *Scrapy* [Online Code Repository]. Available: https://scrapy.org/
9. W3schools (2017, April 5). *Xpath* [Online]. Available: https://www.w3schools.com/xml/xml_xpath.asp
10. Python (2017, April 5). *Urllib* [Online Code Repository]. Available: https://docs.python.org/3/library/urllib.request.html
11. J. B. Moreno. (2015, February 26). *Fuzzy String Matching – a survival skill to tackle unstructured information* [Online Blog]. Available: http://bigdata-doctor.com/fuzzy-string-matching-survival-skill-tackle-unstructured-information-r/
12. SeatGeek Inc. (2017, March 19). *FuzzyWuzzy* [Online Code Repository]. Available: https://github.com/seatgeek/fuzzywuzzy
13. Python 3.0 (2017, April 5). *Datetime* [Online Python Module]. Available: https://docs.python.org/3/library/urllib.request.html
14. Scikit-Learn Developers (2017, March 27). *Scikit-Learn* [Online Python Code Module]. Available: http://scikit-learn.org/stable/index.html