YourNextGeocache

A Recommendation App for Avid Geocachers.

Christina Holland

Sort by: **Distance (Near – Far)**

GCTF5W ♡ 0
Just the prescription
0.51mi

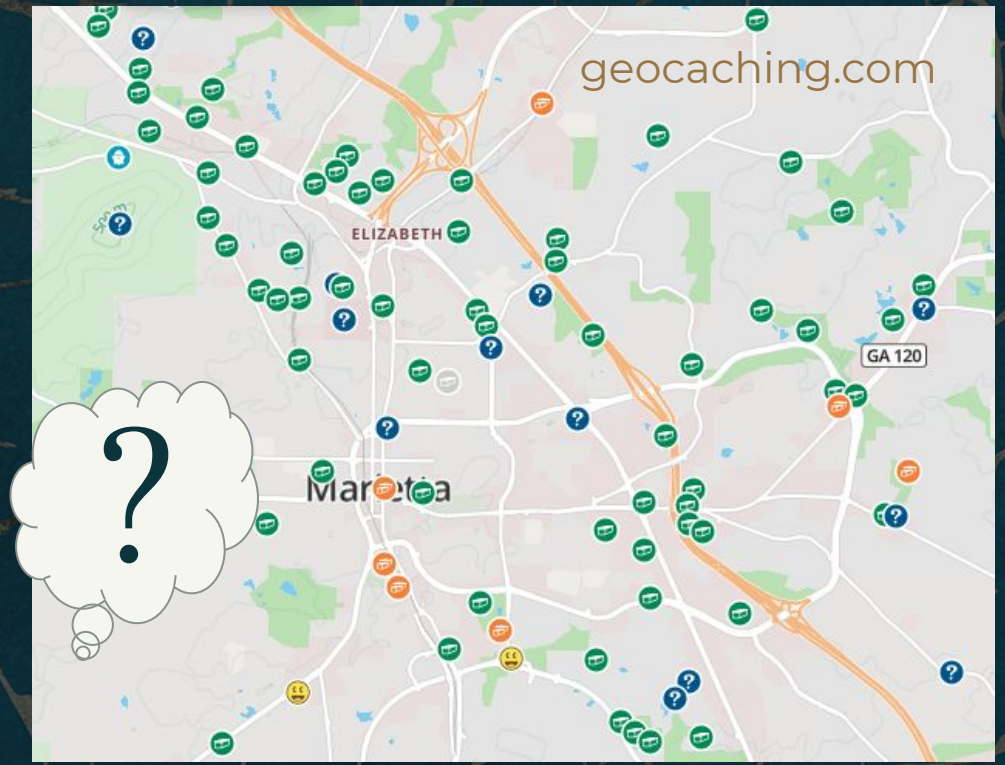
GC8FXA8 ♡ 0 PREMIUM
Ghost Town
1.10mi

GC92EA2 ♡ 6
Sprayberry Rock
1.14mi

GC8BH4X ♡ 0
The King has left the area
1.18mi

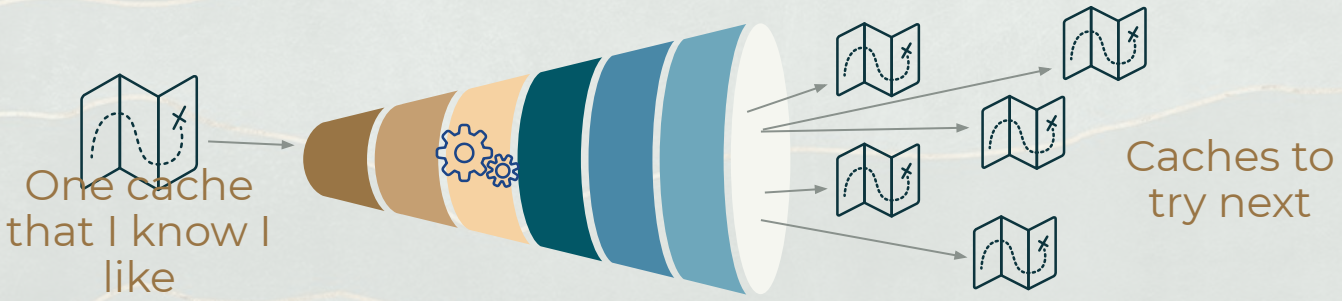
GC1TRHC ♡ 6
Down the Wabbit Hole
1.21mi

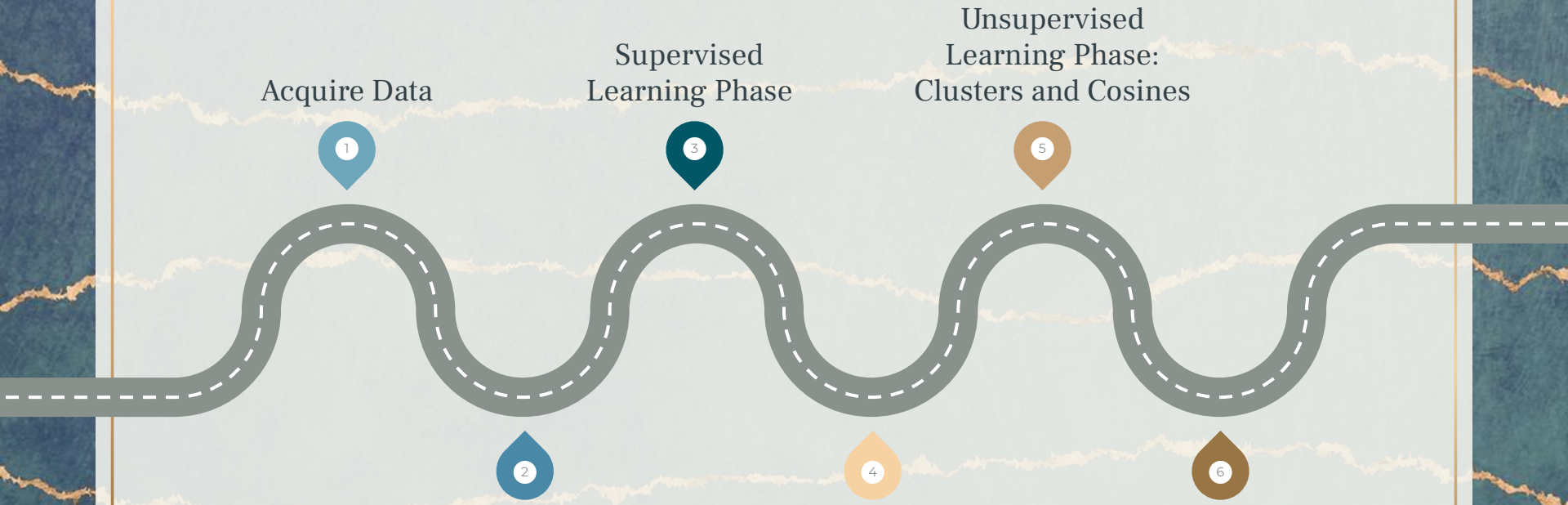
GC89CFG ♡ 0 PREMIUM
Muggle Challenge #9 - Chicken ...
1.27mi



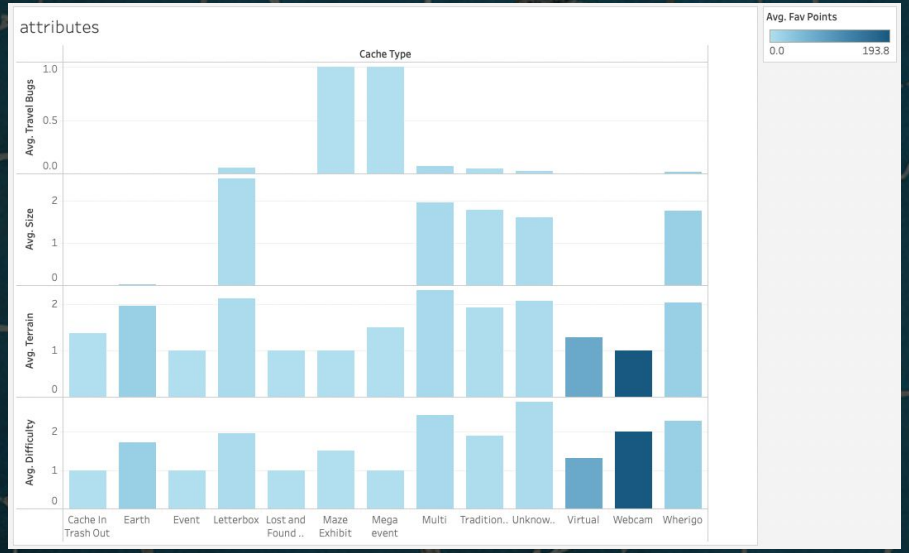
PROBLEM STATEMENT

My objective is to create a "content-based" recommender system for geocaches. For a geocacher using my app, they will get results something like: "If you liked GC47X42 'That Really Big Tree', then you'll probably enjoy …."



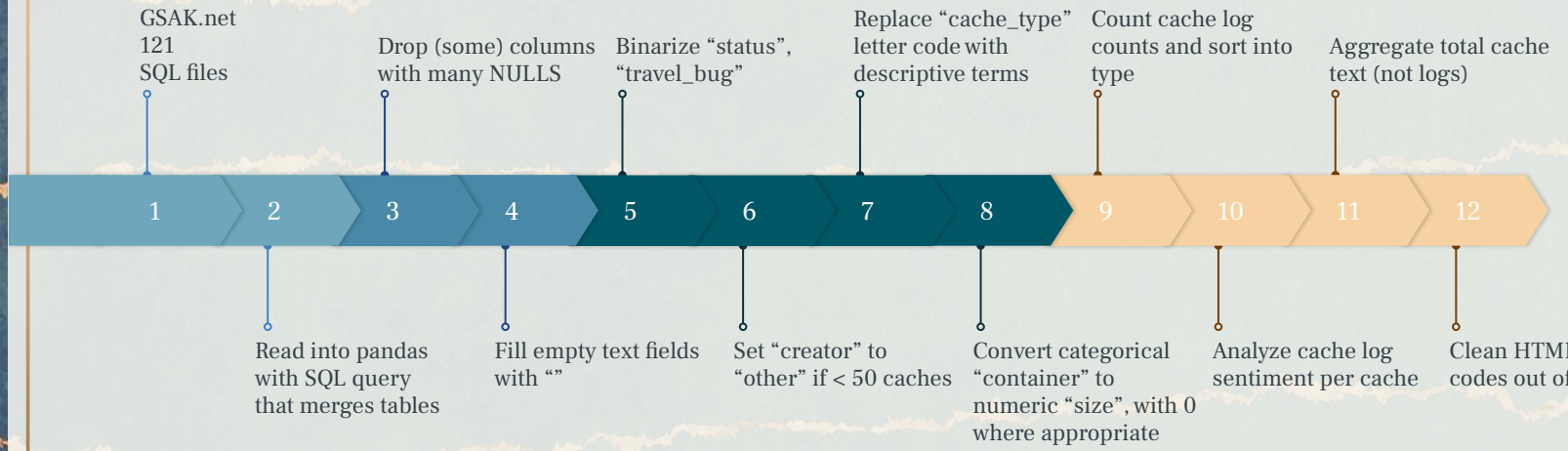

Acquire Data
1
Supervised
Learning Phase
3
Unsupervised
Learning Phase:
Clusters and Cosines
5
2
Data Cleaning
and EDA
4
Popularity
Predictor App
6
Cache
Recommender App

1.
The Data



Geocaching Swiss Army Knife
GSAK.net

Data Acquisition & Cleaning



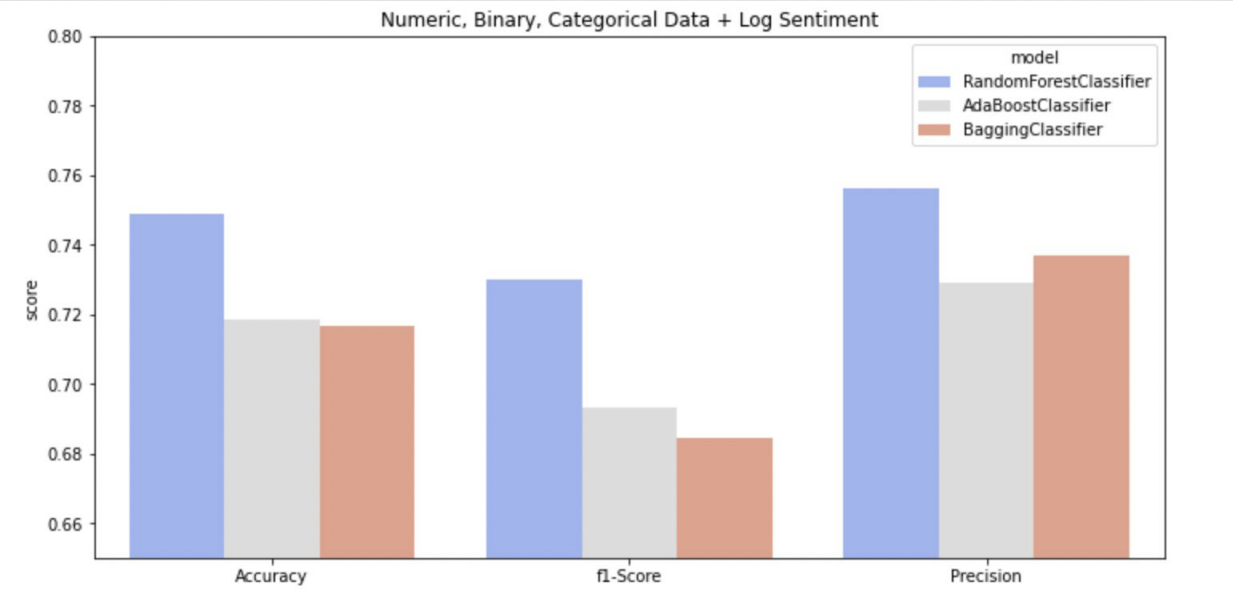
Feature Types

Numerical	Binary	Text	Identifiers	Categorical	Datetime
good_logs_num	status	good_logs_txt	code	creator	placed
neutral_logs_num	is_premium	neutral_logs_txt	name	cache_type	
bad_logs_num	short_description	bad_logs_txt			
difficulty	long_description	cache_text			
terrain	hints				
latitude	travel_bugs				
longitude					
fav_points					
size					

2.
Supervised Learning

Models Attempted

- Log Regression
- KNeighbors
- Decision Tree
- Random Forest
- Bagging
- AdaBoost
- Support Vectors
- Feed-Forward Neural Net



"Kitchen Sink" model

Target = "FavPoints", binarized: 0-1 ⟶ 0, 2+ ⟶ 1

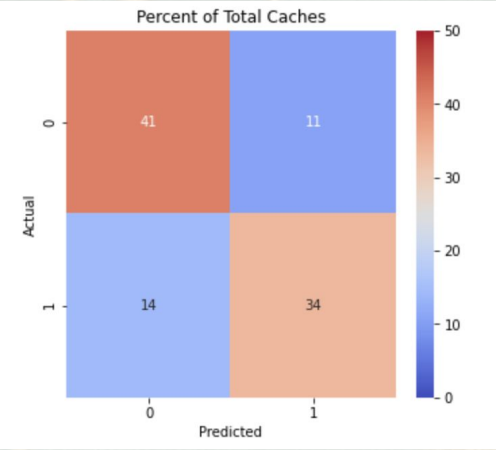
NULL model: FavPoints = 0, Accuracy = 0.519

MODEL SELECTION

Feature Selection

Goal: To help cachers placing a NEW cache predict its likely popularity

Action: Exclude creator name, location, and log info



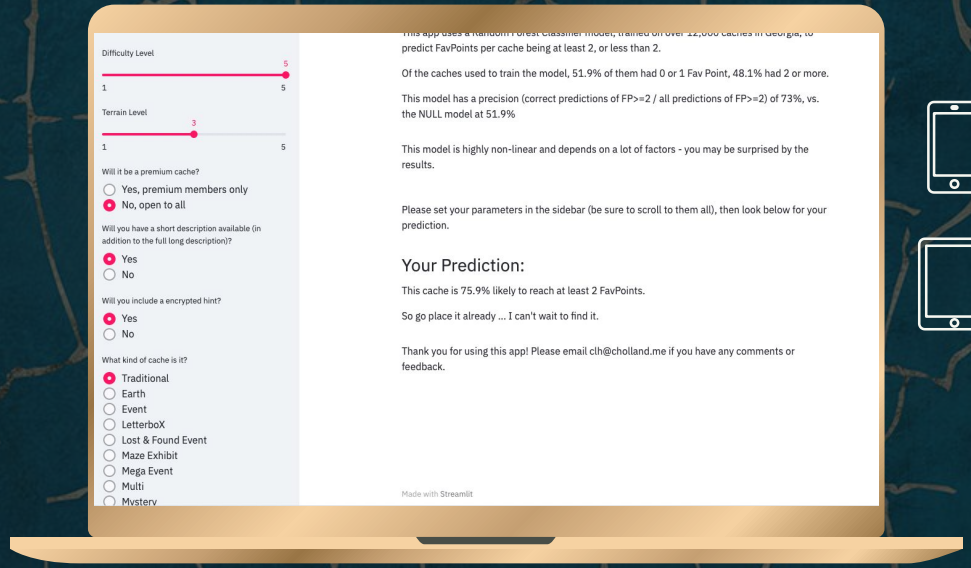
Best Model (so far)

Random Forest
- 50 estimators
- Max depth = 20

Scores
- Accuracy: 0.75
- f1-Score: 0.73
- Recall: 0.70
- Precision: 0.76

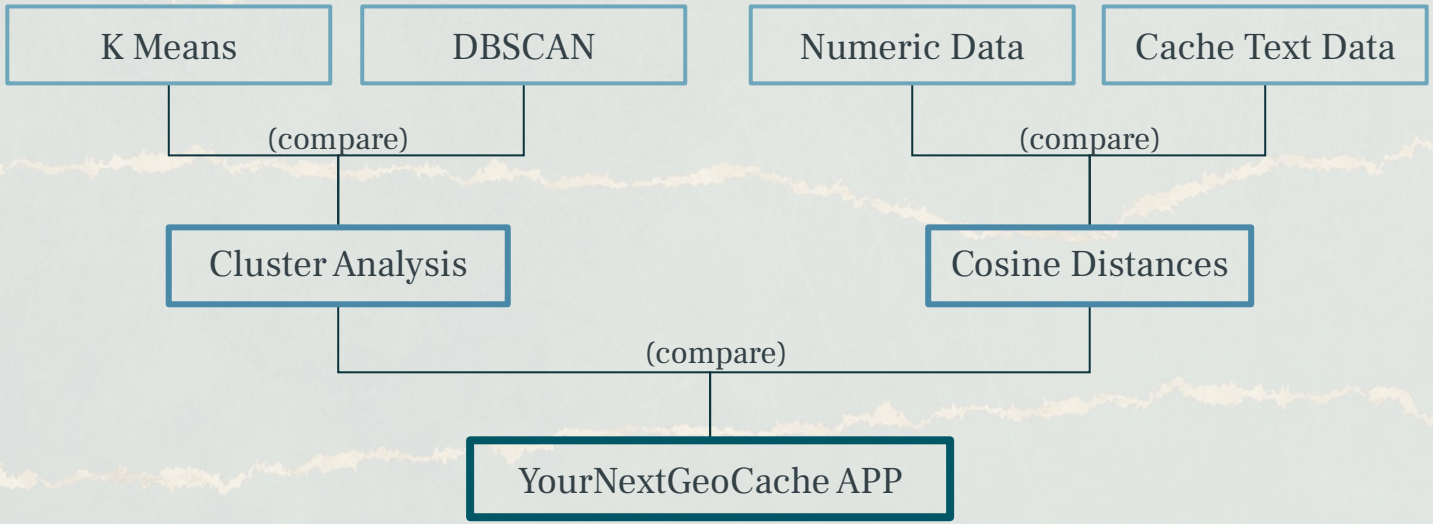
Cache Popularity Predictor App

https://tinyurl.com/cache popularity



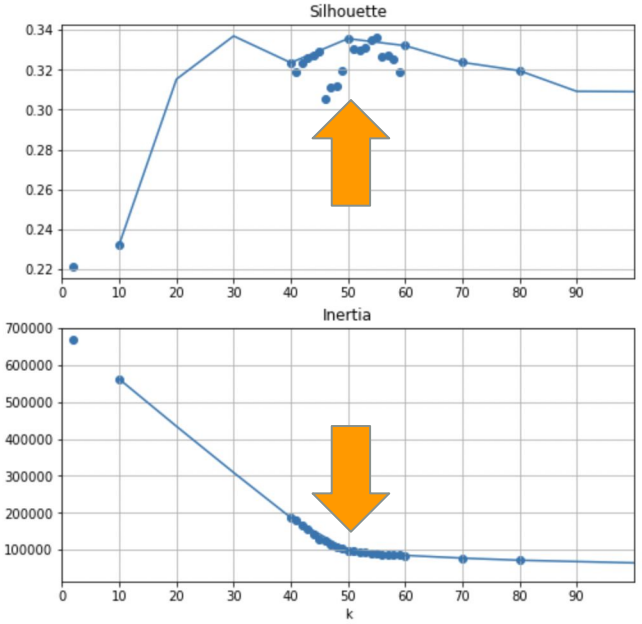
3.
Unsupervised Learning

APPROACH


K Means
DBSCAN
Numeric Data
Cache Text Data
(compare)
(compare)
Cluster Analysis
Cosine Distances
(compare)
YourNextGeoCache APP

K MEANS CLUSTERS: K = 50

cluster #	number of caches	noted characteristics
0	2908	
1	821	Much more likely to be premium access only
2	965	Somewhat more difficult than normal
...		
7	134	Consistently size=0 (more likely to have no container)
...		
10	93	Fairly consistently size of 2-3 (small-regular)
...		
17	60	All difficulty = 2.5; All terrain = 2.5; Likely to be size = 3 (regular)
...		
29	4	100% of the logs are neutral type
...		
46	467	Much more likely to acquire travel bugs
...		



DBSCAN Clusters, eps = 5, min samples = 2 → 58 Clusters

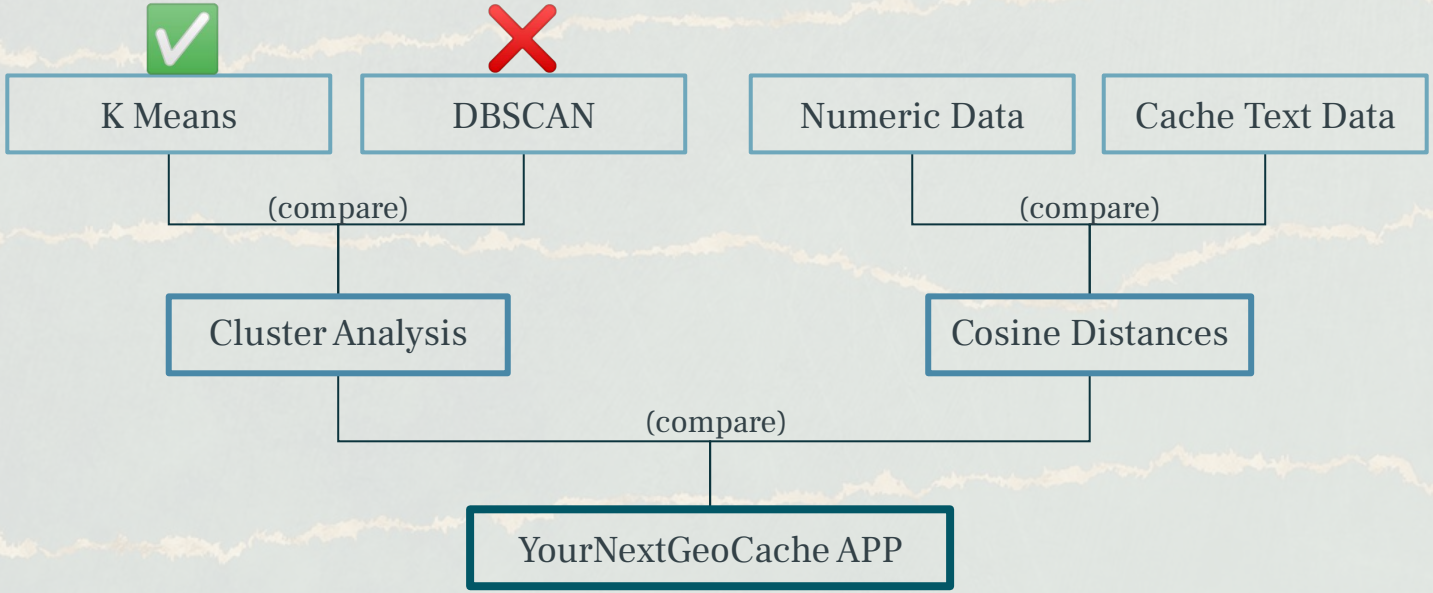
Good

- 0.66% “noise”
- Silhouette score = 0.424, better than K Means (0.336)
- On average clusters were 98.45% homogeneous with a single K Means cluster

Not So Good

- Several K Means clusters have been split into 2 or more clusters here
- Cluster #0 has 99.9% of K Means #0, plus another 5575 caches from 8 other K Means clusters, for 69.8% of the data

Keep K Means Clustering for the App



Cosine Distance

Cache #1
Difficulty 4,
Terrain 3,
Small,
Traditional, ...

Cache #2
Difficulty 5,
Terrain 1,
Micro,
Traditional, ...

Cosine Similarity = $\cos(\Theta)$

Cosine Distance = 1 - cos. sim.

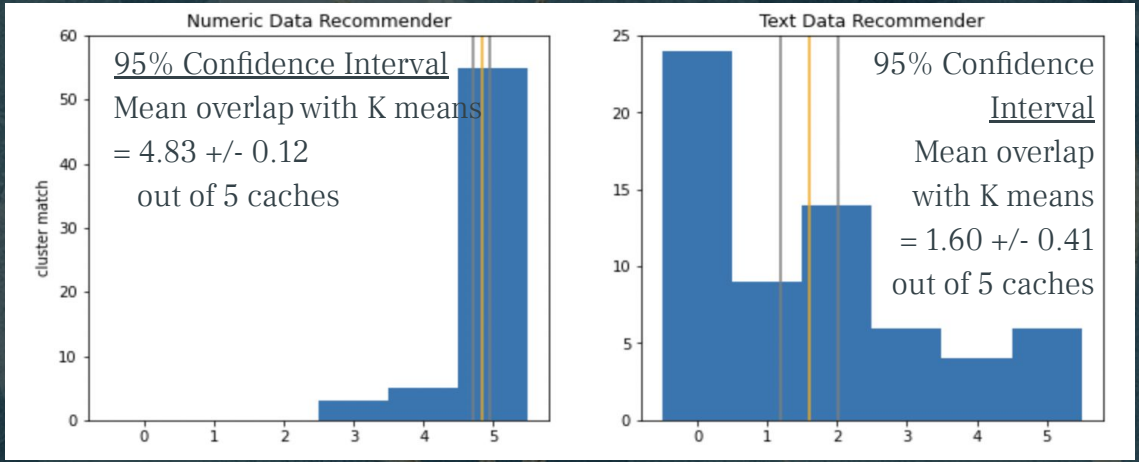
cos simularity	cos distance
-1	2
0	1
1	0

Recommend these

Cos. dist. Compared to K Means clusters

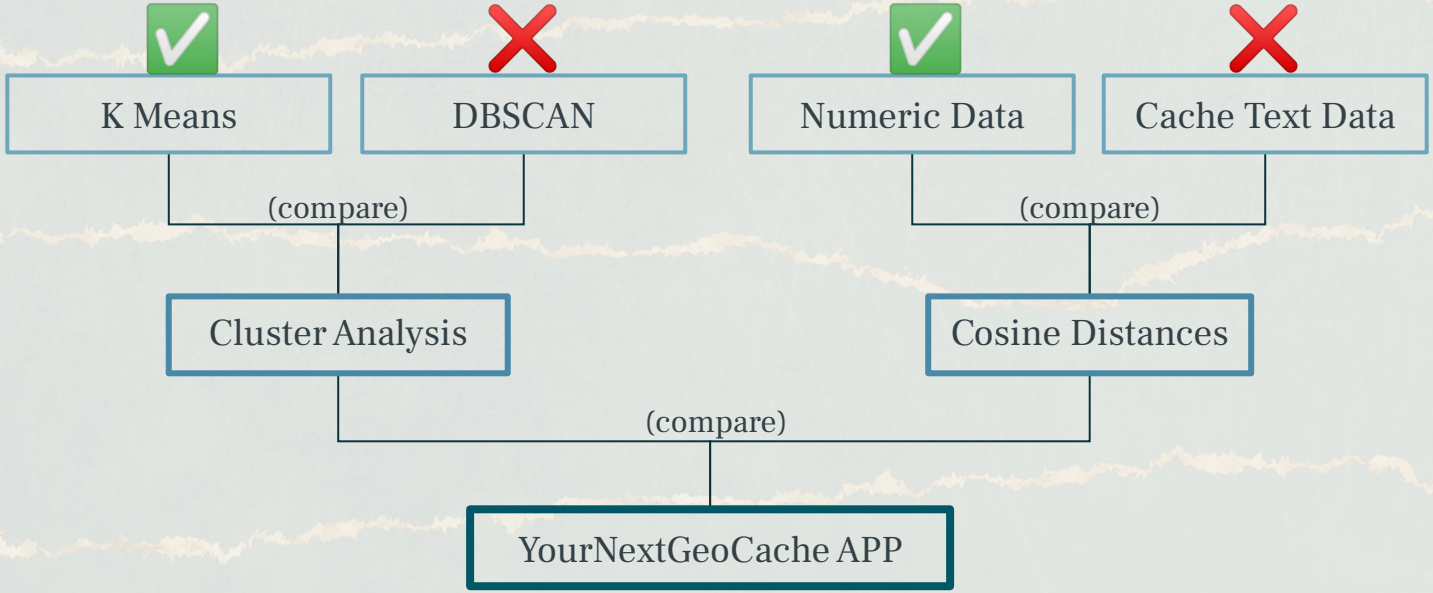
(1) Numeric, binary and categorical features, & log sentiment

(2) Vectorization of any word that appears in the descriptive text of at least 10% of the caches



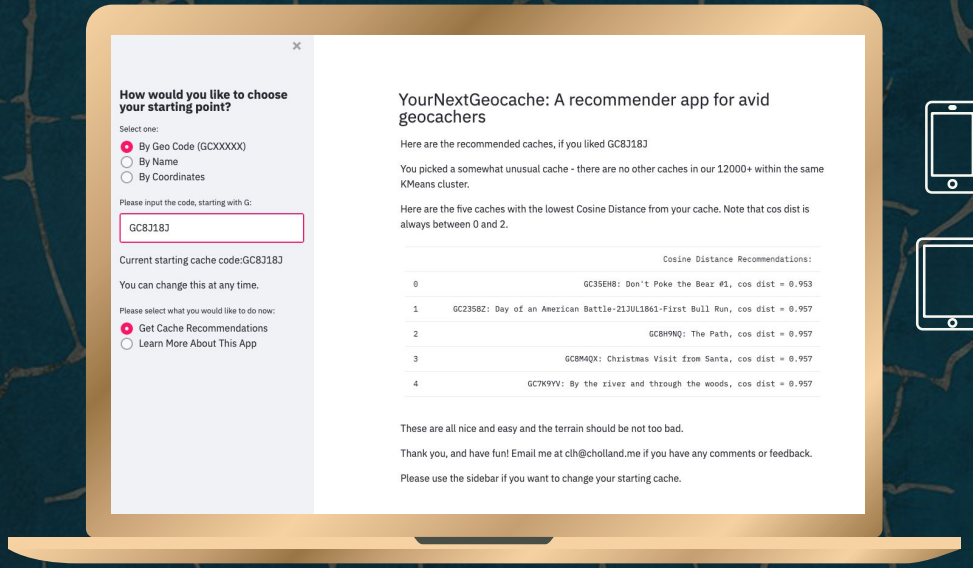
(based on a systematic selection of caches: every 200th)

Keep K Means Clustering and Cosine Distance for the App

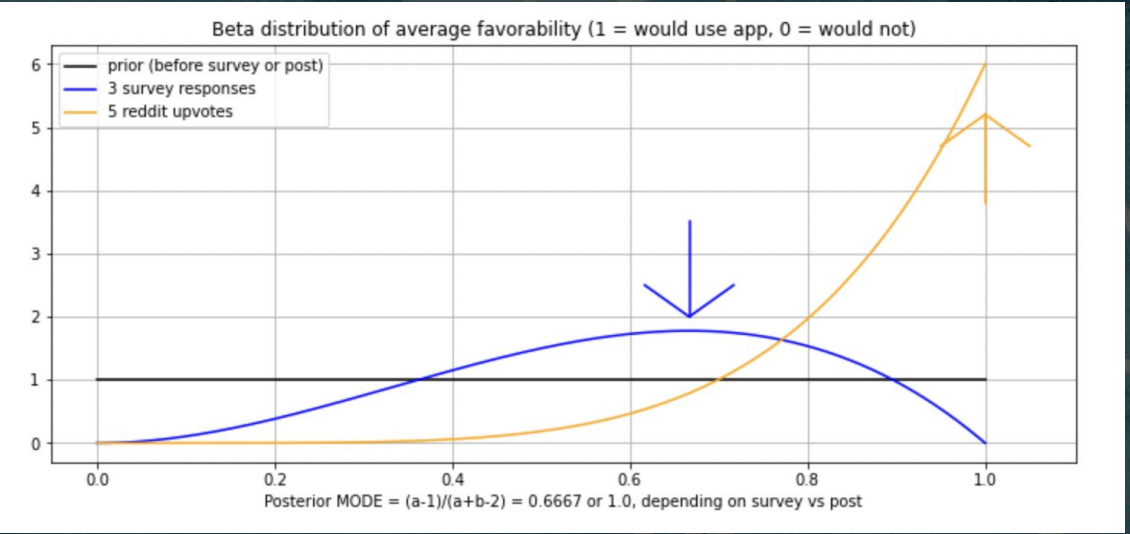


Cache Recommender App

https://tinyurl.com/YourNextGeoCache



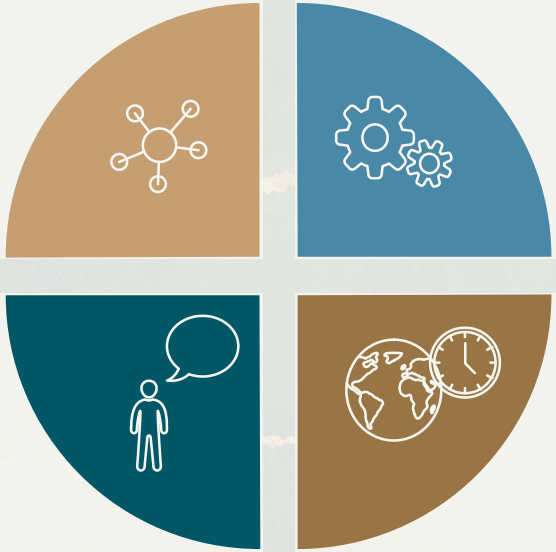
Feedback


Beta distribution of average favorability (1 = would use app, 0 = would not)
prior (before survey or post)
3 survey responses
5 reddit upvotes
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
Posterior MODE = (a-1)/(a+b-2) = 0.6667 or 1.0, depending on survey vs post

AREAS OF POTENTIAL FUTURE WORK

STRENGTH

Multifaceted approach with solid agreement between the K Means clusters, the DBSCAN clusters, and the cosine distance computation.



(1)

Continue working with neural net models to improve the predictive power of the cache popularity predictor.

Continue to assess the app feedback, and work on the app user interface.

(2)

Expand to include more data, other states, and to have it access up to date data.

(3)

THANK YOU!

Any questions?

clh@cholland.me
christinaholland.github.io

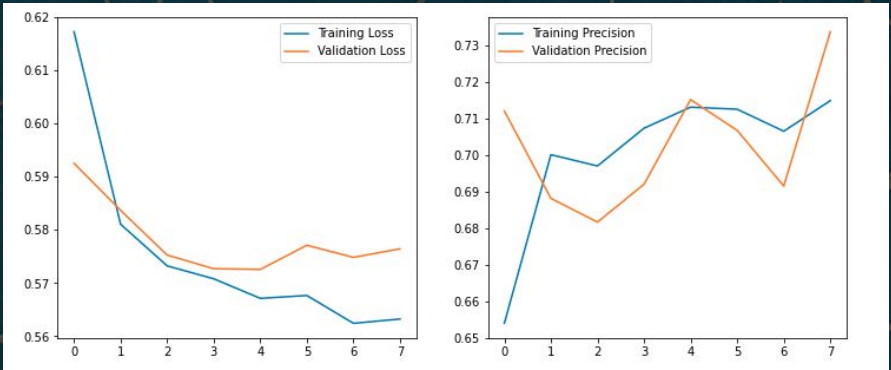
Appendices

Feature Importance

	feature	importance
0	difficulty	0.263414
1	terrain	0.204609
2	size	0.159353
3	status	0.009626
4	is_premium	0.047004
5	short_description	0.044982
6	long_description	0.019873
7	hints	0.041675
8	travel_bugs	0.024998
9	cache_type_Earth	0.007654
10	cache_type_Event	0.001618
11	cache_type_Letterbox	0.004029
12	cache_type_Lost and Found Event	0.000384
13	cache_type_Maze Exhibit	0.000244
14	cache_type_Mega event	0.000320
15	cache_type_Multi	0.015960
16	cache_type_Traditional	0.099497
17	cache_type_Unknown/Mystery	0.042899
18	cache_type_Virtual	0.008585
19	cache_type_Webcam	0.000302
20	cache_type_Wherigo	0.002974

Neural Net Development

```
Model: "sequential_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_92 (Dense)             (None, 256)               5632
_________________________________________________________________
dense_93 (Dense)             (None, 256)               65792
_________________________________________________________________
dense_94 (Dense)             (None, 128)               32896
_________________________________________________________________
dense_95 (Dense)             (None, 128)               16512
_________________________________________________________________
dense_96 (Dense)             (None, 64)                8256
_________________________________________________________________
dense_97 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_98 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_99 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_100 (Dense)            (None, 32)                2080
_________________________________________________________________
dense_101 (Dense)            (None, 32)                1056
_________________________________________________________________
dense_102 (Dense)            (None, 16)                528
_________________________________________________________________
dense_103 (Dense)            (None, 8)                 136
_________________________________________________________________
dense_104 (Dense)            (None, 1)                 9
=================================================================
Total params: 145,377
Trainable params: 145,377
Non-trainable params: 0
_________________________________________________________________
```



	precision	recall	f1-score	support
0	0.68	0.80	0.73	1290
1	0.73	0.59	0.65	1197
accuracy			0.70	2487
macro avg	0.71	0.70	0.69	2487
weighted avg	0.70	0.70	0.70	2487

HTML Cleaning Function

```
clean_text = [clean_html(txt) for txt in cache_texts]
```

```
cache_texts[0]
```

```
'Nickajack Two for OneThis doesn\'t qualify as one of my river/railroad crossing caches, but it is clos
e.<br />\r\n<br />\r\nHere is an interesting spot on Nickajack Lake with a <b>Two for One</b> for your
statistical pleasure. Of course that isn\'t all this pretty spot provides. You can see the Tennessee Ri
ver with all it has to offer - folks fishing and boating, birds looking for a meal, Nickajack
commerical barge traffic. You can see trains go by just across the highway and the sunset is
too. I saw all of this during the 20 minutes I was there hiding the cache, shooting the coord
d photographing the benchmark. The benchmark disk EE1511 can be found on the southeast corner
idge further to the west of the listed coordinates.<br />\r\n<br />\r\nSo come on out, leave
t the parking coordinates, take a short walk, and enjoy the area.<br />\r\n<br />\r\nTennesse
ghway 156 is between the railroad and both the cache and parking spot, so it isn\'t an issue.
n the south side of the road at all times.</i> There is plenty of room to walk from parking b
bank of the lake and the highway guardrail. However, I would advise keeping the small childre
y short leash.<br />\r\n<br />\r\nSome of you might recognize this as an area of contention f
ter rights discussion during the recent droughts. If the State of Georgia had its way, this c
not be listed in Tennessee. <img src="http://www.geocaching.com/images/icons/icon_smile_wink.
r="0" align="middle" /><br />\r\n<br />\r\n<center><b>Please return the cache container to th
cation and position in which it was found and email me with any problems. Good luck and enjoy
\n<br />\r\n<font color="black"><b>Why did I place this cache here? I wanted you to see and e
spot and give you a Two for One.</b></font></b></center>\r\n<br />\r\n<br />'
```

```
clean_text[0]
```

```
"Nickajack Two for OneThis doesn't qualify as one of my river/railroad crossing caches, but i
e. \r\n \r\nHere is an interesting spot on Nickajack Lake with a Two for One for your statist
ure. Of course that isn't all this pretty spot provides. You can see the Tennessee River with
s to offer - folks fishing and boating, birds looking for a meal, Nickajack Dam, and commeric
raffic. You can see trains go by just across the highway and the sunset is quite nice too. I
this during the 20 minutes I was there hiding the cache, shooting the coordinates, and photog
e benchmark. The benchmark disk EE1511 can be found on the southeast corner of the bridge fur
e west of the listed coordinates. \r\n \r\nSo come on out, leave your car at the parking coor
ake a short walk, and enjoy the area. \r\n \r\nTennessee State Highway 156 is between the rai
both the cache and parking spot, so it isn't an issue. Stay on the south side of the road at
There is plenty of room to walk from parking between the bank of the lake and the highway gua
wever, I would advise keeping the small children on a very short leash. \r\n \r\nSome of you
gnize this as an area of contention from the water rights discussion during the recent drough
State of Georgia had its way, this cache might not be listed in Tennessee. \r\n \r\n Please return the
cache container to the exact location and position in which it was found and email me with any problem
s. Good luck and enjoy! \r\n \r\n Why did I place this cache here? I wanted you to see and enjoy this s
pot and give you a Two for One."
```

```python
import mechanize
import nltk
from bs4 import BeautifulSoup
from html2text import html2text
import re

def clean_html(html):
    """
    Copied from NLTK package.
    Remove HTML markup from the given string.

    :param html: the HTML string to be cleaned
    :type html: str
    :rtype: str
    """

    # First we remove inline JavaScript/CSS:
    cleaned = re.sub(r"(?is)<(script|style).*?>.*?(</\1>)", "", html.strip())
    # Then we remove html comments. This has to be done before removing regular
    # tags since comments can contain '>' characters.
    cleaned = re.sub(r"(?s)<!--(.*?)-->[\n]?", "", cleaned)
    # Next we can remove the remaining tags:
    cleaned = re.sub(r"(?s)<.*?>", " ", cleaned)
    # Finally, we deal with whitespace
    cleaned = re.sub(r" ", " ", cleaned)
    cleaned = re.sub(r"  ", " ", cleaned)
    cleaned = re.sub(r"  ", " ", cleaned)
    return cleaned.strip()

# https://stackoverflow.com/questions/26002076/python-nltk-clean-html-not-implemented
```

Cosine Distance Examples

CD ~ 0.00

27 HSC Shine Run

A cache by horseshoechamp & Shine Run Crew Message this owner Hidden : 02/05/2018

Difficulty: ★★½☆☆
Terrain: ★★½☆☆
Size: (regular)
1 Favorites

This is a Premium Member Only cache.

18 HSC Shine Run

A cache by horseshoechamp & Shine Run Crew Message this owner Hidden : 02/07/2018

Difficulty: ★★½☆☆
Terrain: ★★½☆☆
Size: (regular)
2 Favorites

This is a Premium Member Only cache.

CD ~ 1.79

Down By The River

A cache by Busterkeeper Message this owner Hidden : 10/07/2007

Difficulty: ★☆☆☆☆
Terrain: ★☆☆☆☆
Size: (small)
1 Favorites

Related Web Page

Geocache Description:

This is an easy drive-by cache out in apple country on the Cartecay River in Ellijay, GA.

Nickajack Dam: Deep Woods

A cache by NewAgeOutlaw Message this owner Hidden : 03/21/2020

Difficulty: ★★½☆☆
Terrain: ★½☆☆☆
Size: (small)
0 Favorites

Geocache Description:

Placed with permission of TVA. Daytime Only. This will take Yu deep into the Flora & fauna of this area. Great hike for destressing. Enjoy the solitude. Looking for camo container with swag & log(BYOP) Watch for critters both walking & crawling during warmer weather. Enjoy yur hunt.

Geocaching 101