Sasquatch Classification Model

Project Overview

Section: 2021FA_MSDS_434-DL_SEC55 Analytics Application Engineering
Author: Mark Stockwell, MarkStockwell2021@u.northwestern.edu
Updated: November 28, 2021

Introduction

The North American Wood Ape (aka "Sasquatch", "Bigfoot") may exist and many legitimate scientists are actively researching the possibility[1]. The Bigfoot Field Researchers Organization[2] has been collecting and analyzing reports of sightings from credible witnesses for several decades. Historians note that Native American peoples have centuries old oral traditions of large hominids with multiple descriptions depending on geographic region[3]. It is likely that distinct populations exist[4] and can be classified based on geographic features.

Goals & Objectives

The goal of this project is to build a Machine Learning model using the Google Cloud Platform that classifies bigfoot sightings based on location, elevation, climate, and population. Analyze the predictive power of the model and develop visualizations of data using open source tools.

Data Sources

- Bigfoot sightings database - This is a curated collection of sightings with location information and will be the primary source. This data has been standardized and enhanced at data.world/timothyrenner/bfro-sightings-data
- Precipitation data by county, available from the National Oceanic Atmospheric Agency.
- Population data by zip code and county from the US Census Bureau.
- Elevation data via the Google Elevation API.
- Plant Hardiness Zones are used as a proxy for temperature, code available at waldoj/frostline: A dataset, API, and parser for USDA plant hardiness zones.

Data Architecture

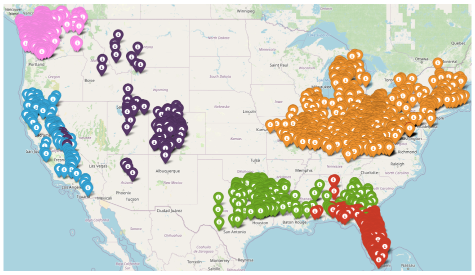
Data is ingested from original sources or derived from an API, see below:

[1] Meldrum, 2016 Sasquatch & Other Wildmen: The Search for Relict Hominoids
[2] See https://www.bfro.net/
[3] The RELICT HOMINOID INQUIRY 1:1-12 (2012) , see https://www.isu.edu/rhi/research-papers/
[4] See https://en.wikipedia.org/wiki/Skunk_ape

- BFRO data is loaded to Cloud storage via python load.py script from data.world.
- Precipitation data is loaded directly to a one column BigQuery table in raw/space delimited format via the load.py script and then converted to structured data using a view
- Population data is accessed live as needed from bigquery-public-data:census_bureau_usa.population_by_zip_2010
- Zip code data is from bigquery-public-data:geo_us_boundaries.zip_codes
- Elevation data is loaded from a csv into the storage bucket via a Jupyter notebook utility Elevation_API.ipynb

Once raw data is imported into the project, a DDL script creates the bfro_reports_geocoded_final table which includes yearly precipitation, elevation, location, zip code, county, population density, and all other attributes in a single wide table. This table is then used as a source for a BigQueryML model. The ML.PREDICT function is used to create the bf_centroids table, which has a cluster identifier for each Bigfoot sighting and the distance to the centroid for all centroids. The distance of the best match centroid is assigned a fit_rating percentage based on the distance to centroid vs total distance of all centroids. The fit_rating can be used in predictions to determine the best centroid match and quality for a random point.

Visualizations are done using a folium map hosted on Google App Engine (main.py). Each sighting with associated data is color coded to the centroid, with data filtered by fit_rating.



System Maintenance & Operations

Source data changes slowly, on the order of days for base BFRO data. The data however is sourced from data.world, which is updated monthly. Suggested maintenance window is daily check for updated data, and if found then rerun the data load, model creation, and final analytical table creation. Additional alerts should be set up and monitored for app engine latency and errors. Over time updates to the python version and related packages may be needed similar to other systems.

Future Enhancements

The current architecture relies heavily on the data.world dataset. This could be bypassed by going directly to bfro.org data. Some sightings have coarse location (i.e. county) but are missing lat/long. These could be enhanced to use the county center location. County to zip code translation has some errors due to zip codes overlapping county boundaries. This could be corrected by using the center of the zip code polygon to determine the county. Zip code and county data also contain percentage of coverage by water; this could be added to the list of features.