Navdeep Raj

Capstone 2 Project Proposal

Mining and Predicting UFO Sightings

Problem: We often wonder if we are alone in this universe. But lot of stories have been heard for the last century but the solid evidence is still lacking, this project is to analyze and mine the documented data available to us based on UFO sightings and find interesting trends and predict some behavior in terms of their shape, size, color, etc. Most of the analysis is based on visualizing maps to understand the geographies of UFO sightings

Context – Data has been collected from Kaggle where the first sighting dates back to 1949 and covers up to 2013. Dataset has around 80000 rows and 11 columns. It will be interesting to map various attributes together and analyze the sightings in terms of region, shape and time.

Criteria for Success – This analysis is to perform the following:

- Predict the duration of the sighting
- Analyze which country has the maximum encounters
- Analyze what time of the day has more sightings
- Text Analysis of the description to find the color and sighter and cluster & classify them(NLTK)
- Maximum shape size reported
- Which season and hemisphere had more sightings

Scope – The data is based on NUFORC reports till 2013 and has been scraped from their website.

Constraints –

- Although the dataset has 80000 rows, but still scientists are debating the existence of extraterrestrial, and all the observations may be subjective to the individual making the claim.
- Lot of text Analysis required for better understanding of the observers

Shareholders : Space Research organization team dealing in UFOs,

NUFORC director

Data & Sources -

Dataset link - https://www.kaggle.com/NUFORC/ufo-sightings

Approach –

- Cleaning Dataset
- Formatting various columns such as city and country and datatime
- Filling in the missing values
- Univariate & Bivariate Analysis
- Visualizing relationships
- Splitting dataset
- Choosing appropriate algorithm for model development
- Evaluating models and choosing best prediction model

Text Analysis –

- Cleaning data, removing digits, non-letters, unicode
- Spellcheck, removing stop words
- Shape classification
- Observer clustering
- Visualizing trends

Deliverables:

- Github Repo including:
- Code
- Powerpoint Presentation
- Prediction results summary