Joey Chen
Catherine McNabb
Anuraag Mohile
Preetika Srivastava

How to See a UFO

Description of Project Goals

Humans have always been fascinated by the idea of extraterrestrial life, and our dataset explores potential alien sightings from the past 35 years. We used data from National UFO Report Center (NUFORC) to explain the likelihood that a given UFO sighting is explainable, as well as what a typical sighting might be like. The dataset predictors include when and where the sighting occurred, as well as the shape and description of the object. We added variables such as proximity to air force base, extreme weather presence and the moon phase on the day of sighting to learn more about when and where a UFO may be seen. These variables were just educated guesses by our team as to what could be correlated with a sighting. Furthermore, some of the UFO sightings have been labeled “explainable” by NUFORC. NUFORC compared satellites, aircraft and comet data to identify whether given UFO sightings might be explainable.

Though gaining insight into UFOs will not have an impact on the community, it is of interest to many, many people. In the United States alone, where we filtered our data by location, there are 8 active UFO organizations, dedicated to sharing information about unidentified flying objects[1]. Many movies, books and other pop culture are influenced by the idea of extraterrestrial life, such as E.T., Men in Black, Star Wars and others. Due to this fascination with UFO sightings, we explored both explainable and unexplainable UFO sightings, so we can know when and where they occurred and hopefully see the next one ourselves.

Exploratory Analysis

First, we began exploring our data by looking at variables one by one. What is the most likely day to see a UFO? Where would it be? These are some of the initial questions we sought to examine in the exploratory phase.

Looking at the graphs in the appendix, we can see that certain days and times of year correspond to the most UFO sightings. Saturdays and the middle of the night are some of the most common times to see UFOs. You are also more likely to see them in July through September. July 4th and New Year’s Day are the most common days of the year. These are all logical times for sightings to peak. They are times when more people are out, and, for the holidays and weekends, many of those people probably had been imbibing all day before spotting a UFO. Moreover, the two holidays have an additional explanation - these are the most common holidays for fireworks.

But, we can also assume that sightings occur in some places more than others, just as they occur at some times more than others. Logically, California, Florida, Texas and New York are four of the top five states with the most UFO sightings. These states have high populations, so proportionally we would expect this disproportionate number of sightings. However, Washington is the state with the second most UFO sightings and Arizona the sixth, and their populations are not near that of the other four states.

Seattle is also the city with the most of UFO sightings, even though its population is not near one of the largest of U.S. cities. The other U.S. cities included in the top five for UFO sightings are Phoenix, Las Vegas, Portland, and Los Angeles, all west coast cities.

However, the biggest location factor that determines whether you will see a UFO is your proximity to an air force base. Most sightings occur within 100 miles of a base.

Solution and Insights

Using the variables stated earlier in the report, we ran several models on the data to try to predict whether sightings were easily explained. If a sighting could be categorized as "explained", then we do not need to waste time investigating that sighting, and we can spend more time analyzing the "true" UFO sightings.

First, we tried logistic regression to see if we could explain some of these unidentified flying objects. The "explainable" UFOs accounted for only about 3.6% of total, so the data was a bit off balance before we even began to analyze. However, we used a "stratify" function to make sure that equal proportions of the explained events would be in the test and train set, and we also lowered the threshold of what counted as "explainable" down to 0.3, rather than the 0.5 that logistic regression might typically use.

The prediction percentage of the regression was very good, but mostly because the majority of the observations were "unexplainable" and we predicted them as such. The problem with our prediction was the false negatives. We correctly predicted 114 events as explainable, but we incorrectly predicted 941 events as unexplainable when they could be explained. Even though our total correct was almost 97%, we did not do a good job predicting the explainable events[2].

However, we did gain some insight from the regression. By looking at coefficient weights, we found which shapes were the most unexplainable - chevron, cigar, triangle and egg. And which was the most explainable - a light. Putting this together with other models can help us know what a "real" UFO will look like.

Second, we tried a random forest model to our data to build a classifier which can identify which sightings might be explainable. Like we said in the logistic regression model, only 3.6% UFO events are labeled as explainable. As a result, we could expect that random forest classifier would not perform well. However, we could still try to answer these questions: How to help National UFO Report Center to label these UFO events? If we want to find a place to see UFOs, which sightings should we avoid?

The accuracy of the random forest classifier was nearly 97%. Just like what happened in logistic regression, the classifier would guess "unexplainable" most of the time due to the unbalanced data. If we want to automate the labeling process for NUFORC, this classifier could not do a decent job. For 844 labeled test data, random forest classifier could only identify 88 events "explainable". The True Positive rate is only 10%, so we do not recommend using this random forest model to automate the labeling process[3].

However, if we want to see UFOs, this classifier might come in handy. For all 23,873 test data, how should we pick where to visit? We all know it is hard to find a UFO, so we should believe most of the UFO events could be explained. We changed the random forest classifier threshold to 0.95, and tried to find the sightings that not likely to see UFOs. In other words, we wanted to have a high True Positive rate regardless of the total accuracy. The result was that we could successfully identified 372 UFO events are "explainable" and the True Positive rate was 44%. We believed that using the random forest classifier could help UFO lovers avoid visiting places not likely to have a UFO.

Although random forest was not a useful model for our data, we could get some insights from random forest feature importance form. We could see that "Time", "Proximity to USAFB," and "Date" were the top 3 important features in classifying.

We also applied K-means clustering to obtain discernible patterns, but, most of the clusters were some combination of dominant variables like Texas, California, Washington, Florida, etc. for states; July, August, etc. for months; weekends for day of the week and light for shape.

We did get a couple of differentiated clusters; for example, one with only Alaska and Hawaii as the states, and another with “Sphere” as the dominant shape and January as the dominant month. The former could be explained by both those states being on similar longitudes, towards the west of the mainland USA, as well as having really high proximities to Air Force bases. However, there were no other clusters with significant patterns.

Conclusion

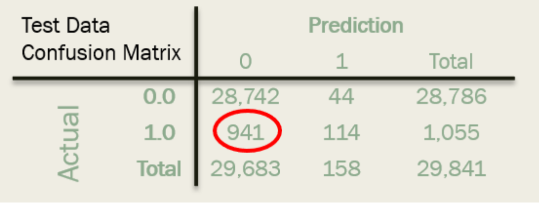
In conclusion, using data analysis and insights, we would recommend going to Seattle, Washington during a weekend in July or August to find a “true” UFO. Also, we recommend searching the skies at night rather than during the day for a “triangle” or “cigar” shape. This recommendation comes from careful analysis of all variables in the dataset, as well as the ones we collected separately.

The regression and random forest helped us to find which events may be the most “unexplainable”, which will in turn lead us to the “true” UFOs, rather than the ones that can be explained by other phenomena.

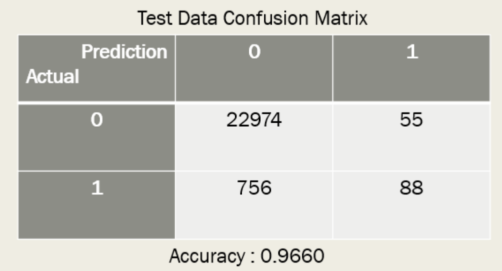
We have included in the appendix a plot that shows the proportions of explainable and unexplainable variables for states[4]. Here, we can see that Washington has a higher proportion of unexplained sightings, so that’s one of the best places to go to see a “true” UFO. Also, within Washington, Seattle has the highest number of unexplained or “true” sightings, making it the best place to spot an actual UFO.

Appendix

1. https://en.wikipedia.org/wiki/List_of_UFO_organizations#United_States

2. 

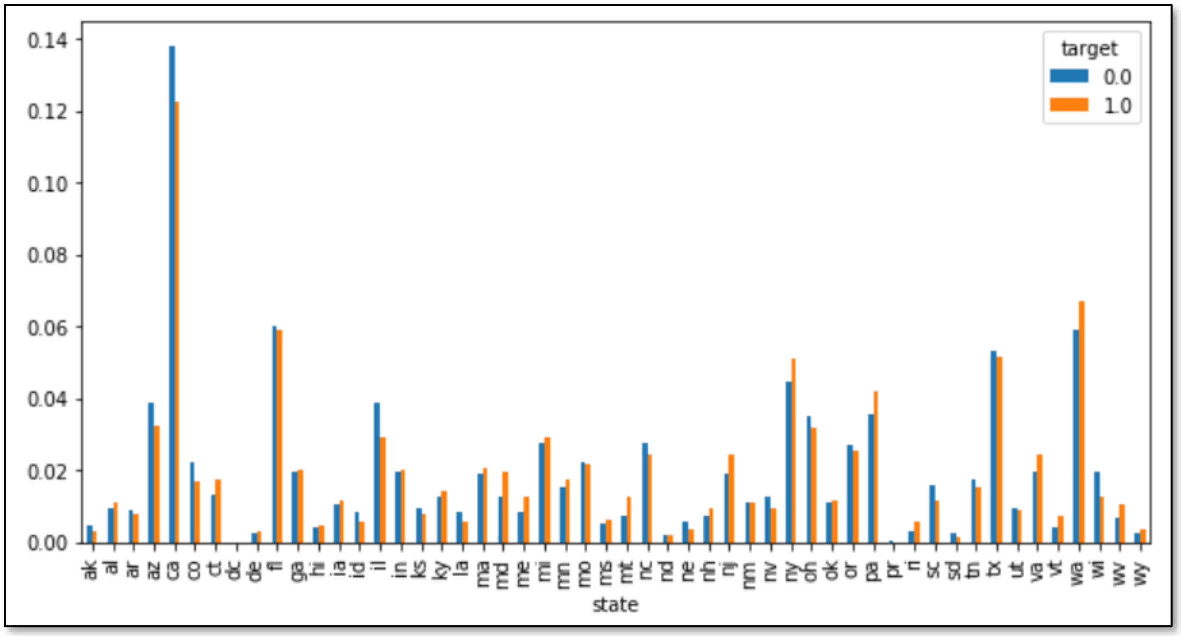
Test Data Confusion Matrix		Prediction		
		0	1	Total
Actual	0.0	28,742	44	28,786
	1.0	941	114	1,055
	Total	29,683	158	29,841

3. 

Test Data Confusion Matrix

Prediction Actual	0	1
0	22974	55
1	756	88

Accuracy : 0.9660

4. 
target
ak al ar az ca co ct dc de fl ga hi ia id il in ks ky la ma md me mi mn mo ms mt nc nd ne nh nj nm nv ny oh ok or pa pr ri sc sd tn tx ut va vt wa wi wv wy
state