Close Encounters: Script

SLIDE 1

Good Evening, everyone! I hope you all had a wonderful day! My teammate, Tifannie, and I will be talking about UFO close encounters.

SLIDE 2

The thought of life beyond our planet has fascinated humankind. It even found its way into every facet of popular culture. From books to movies to magazines and more, the influence of the possibility of alien life and even first contact has captivated generations of people.

We found a data set on Kaggle which included reports of UFOs in the last century. A gentleman named Sigmund Axel took it upon himself to collect and clean data from the NUFORC website which contained data from the early 1900s to 2014. The complete original data set contained entries where the location of the sighting was found and not found or blank. The time for each sighting was converted and normalized to seconds. We used the scrubbed data set with only resolved locations and normalized time. This was a total of 81185 total records that contained information about the description of the sighting including the shape people saw, the date of sighting, the duration, any additional comments, the city, and the state.

We also collected movie data from The Movie Database API on Science Fiction movie releases related to UFOs. Data was collected as well as from various websites that documented state populations, locations of US airports, locations of US military bases, weather conditions each day, and binge drinking prevalence in the US.

With all of this data, we would like to look at areas of the united states that are most likely to have UFO sightings, see if there any trends in UFO sightings over time, look to see if there is any correlation between clusters of UFO and landmarks, such as airports or military bases. We would also like to create a UFO sighting classifier to predict if a UFO sighting is credible, and look at algorithms that give us a high prediction accuracy.

SLIDE 3

To begin, we cleaned the UFO data to pertain only to sightings within the United States. As you can see from this plot, UFO sightings have increased over time. The year 2013 had the most UFO sightings with a total of 2135. Followed by 2012 with 2119.

SLIDE 4

We wanted to know which states had the most reported UFO sightings. We created the following bar chart which reveals the top 15 states for reported UFO sightings. We immediately noticed that California has significantly more reported UFO sightings than any other state. It has approximately two times more reported sightings than the next highest state, Washington. The next three states include Florida, Texas, and New York. Four out of the top five states in terms of reported UFO sightings are also the largest four states in the United States by population. This bar chart confirms to us that a state’s population is a significant factor in reported UFO sightings per state.

SLIDE 5

Then we looked at what days and months UFO sightings are likely to occur. From the barplot on the left, you can see UFOs like to go out on Saturdays. Aliens are just like you and me, in that sense. They still are busy during the week, though. From the barplot on the right, we see that UFOs like the Summer months, especially the month of July. This makes sense since people are out and about during the Summer than the Winter months when people stay at home more.

SLIDE 6

When we looked at our descriptive statistics, we saw that 3 quarters of the UFO sightings lasted for about 3 minutes. We did see that some cases lasted up to 3 years. In fact, we searched through the comment attribute and saw there were 20 abduction cases included in the data.

SHOW MAP: We made a map to show where UFO reports lasted over 24 hours.

SLIDE 7

The most commonly reported shape for UFOs is “light.” We see on the plot from the right that this was not the case in the past. We see in the mid-1990s a rise of the “unknown shape”.

SHOW MAP: We made a map to show where UFO reports with egg shape happened. We also made one for the shape of light, but as you can see, it occurs all over the nation.

SLIDE 8

We calculated the distance between a UFO sighting and a US military base. As you can see in this histogram, there is no correlation. UFOs are nowhere near military bases. A reason for this is that just a small number of people are allowed to be around these places.

SLIDE 9

The same does not apply to airports. Per our exploratory analysis, it is more likely that a sighting will occur the closer the person is to an airport.

SHOW MAP: In this map, we only marked only large airports. As you can see, our observation is emphasized.

SLIDE 10 and 11

There does not seem to be a correlation between UFO sightings and alcohol consumption in the US up until 1994. Once the alcohol consumption hit an all-time low, UFO sightings began to increase sharply which also led to an increase in alcohol consumption.

ADD QUOTE TO PRESENTATION: Add the comment about the guy who saw a UFO on thanksgiving when we drank a Heineken.

Tiffannie continues from here

SLIDE 12

Figuring out how to use the TMDB API was tricky. But, after much trial and error, it was easy to read in various features about the films. The API had a search feature where we searched based on various terms like UFO, alien, outer space, etc. That only returned about 70 movies. Then we found a list of films featuring extraterrestrials on Wikipedia and searched for their details on the API based on their title. This returned over 400 movies. We ended up with 580 films altogether. The earliest film came out in 1902! There was a toon of different details you could use, but we stuck to the Title, Release Date which is the day the movie premiered, and Overview of the movie for this analysis. The overview contains a brief description of the movie and the plot. We thought it would be interesting to compare to the descriptions of the UFO sightings to see if there was any interesting overlap.

SLIDE 13 and 14

Off the bat, it seemed as though space-related movies were being released as increasingly as the ufo sightings were being made. The trend seemed similar.

The movie and present UFO data seemed to follow a similar pattern over time.

SLIDE 15

We did a little bit of text mining on the description given by UFO seers and the overview of each film. We produced two word clouds, however, there does not seem to be very many words in common.

SLIDE 16

The trends seem less likely to align when plotted on top of each other.

We still wanted to see what kind of correlation could exist, if any

MODELING ----------------------------------------------------------

SLIDE 17

Other people have analyzed the UFO data from Kaggle and they use the UFO’s shape as their dependent variable. If a shape was clearly described, then that meant there was an “outcome” or a UFO sighting, otherwise, it was not counted as such. Many of the projects were all the same, so we wanted to do something different.

After collecting airport and military, ufo movie, weather, addiction, and population information and analyzing it in comparison to the UFO data, we wanted to see if we could predict a “credible” UFO sighting.

“Credible” was the variable created. Certain criteria had to be made in order for a sighting to be labeled as credible.

We took into account various weather events like tornados, fog, snow, rain, thunder, and hail. If the distance to the nearest airport or military base was less than 1.6 kilometers then not credible. If it was on new years or the fourth of July, then not credible. We really tried to think of events that would invalidate the UFO sighting or be easily confused with the event that was going on during the same day.

The variables we chose to focus on were

0 city,

1 State,

2 Shape,

3 Year,

4 Month, day, rain, thunder, snow

5 Day,

6 season,

7 Military\_base\_distance to sighting,

8 Airport\_distance to sighting,

9 Temperature that day,

10 Rain,

11 Visibility that day,

12 Thunder,

13 Fog,

14 Tornado,

15 Snow,

16 movies\_released (which were the number of UFO-related movies released within 2 years before the ufo sighting,

17 2015PrescriptionRate,

18 2015AlcoholPercentage,

And 19 population

Since credibility is a binary outcome, we chose a few classifiers and a neural network to predict whether our set of variables could indicate a UFO sighting.

SLIDE 18

The dataset was split into train and test sets.

We will review the models beginning with logistic regression

SLIDE 19

**Logistic regression**. It’s classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable.

It yielded a 65 % accuracy which isn’t the best.

The next model we used a **random forest classifier**, this performed well and got a 100 percent accuracy (we may have been doing something wrong…. haha)

For **XGBoost Classification**. XGBoost is a Scalable Tree Boosting System. It is an ensemble of decision tree algorithms where new trees fix errors of those trees that are already part of the model. Trees are added until no further improvements can be made to the model.

This model did very well and also predicted with an accuracy of 100 percent. Again, kind of fishy, but excellent, nonetheless.

SLIDE 20

For the logistic regression, random forest classifier, and xgboost classifier, we were able to see which features were the most important in predicting the outcome. For logistic regression it was military base distance. For the random forest classifier, airport distance was the most telling with 0, 4, 5, and 10 being not nearly as significant, but still making the chart. For XGBoost, the airport distance was the most telling, with 4,5,10,12,13, and 15 adding a tiny bit of explanation.

This is xgboost classifier importance test output (model.feature\_importances\_)

This goes to prove that the distance from the airport was the most telling of a UFO sighting.

SLIDE 21

For the **neural network**, we went with Tensorflow’s sequential model. The categorical attributes were appropriately encoded and the remaining features were scaled to a range between 0 and 1. We added 3 dense layers. (A Dense layer feeds all outputs from the previous layer to all its neurons, each neuron providing one output to the next layer. It's the most basic layer in neural networks.) We used 2 layers with relu activation which is the rectified linear activation function or ReLU for short is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. The last layer used Softmax activation. Softmax assigns decimal probabilities to each class in a multi-class problem. Those decimal probabilities must add up to 1.0. This additional constraint helps training converge more quickly than it otherwise would. Softmax is implemented just before the output layer.

After running the model a number of times, we were able to predict the credibility of a UFO sighting with an accuracy of 99 percent!

SLIDE 22

Another classification method we tried was to use KNN to predict the city and state of the UFO sighting. The model was able to predict the city with 35% accuracy. The model (when trained and tested for the credibility) predicted with an accuracy of 85%

Finding optimal k. Blue is training and orange is test score. We went with a k of 21 in this instance for 85%. A K of 31 yielded an 83 % accuracy

**In the future** we would have liked to use PCA to train and test the models.

We ran a PCA on our data and concluded that most of the variance (we shot for 95%) could be captured within about 16 components.

