Exploratory Machine Learning with UFOs

Linear Regression, Neural Networks, Decision Trees & Random Forests

May 7th, 2025

DS 3001 Professor Terrence Johnson

Avery Anderson, Zoe Gates, Katrina Garcia

Abstract

Using a dataset of self-reported UFO sightings, this project aims to discover patterns in those sightings to serve as an indicator of UFO existence and predictability. Our research narrows in on the shape, duration, country and dates of the respective sightings and uses those variables to attempt to create models of prediction. For our project we used linear regression, decision trees and neural networks to examine our data for patterns. In the end we found very little, if any, predictable patterns that point to the reality of UFOs. However, this investigation did lead us to see an uptick of sightings on certain holidays and in certain seasons, which lead us to believe that outside variables might play more of a role in these sightings than the true existence of UFOs.

Introduction

Unidentified Flying Objects (UFOs)— the most mysterious phenomena of our era. Are they real? Do aliens exist? Could we still have hope that we are not alone, and our dating pools are larger than we think? Well, we needed to know, so we sought out the most comprehensive UFO dataset on the world wide web. It was our belief that if we could find a pattern in these sightings, we may be able to not just work to confirm the existence of our extraterrestrial friends, but also predict when we might see them next. Unfortunately, either our friends are very type B and don't have any sort of pattern to their flyovers or, more likely, these sightings are misinterpretations.

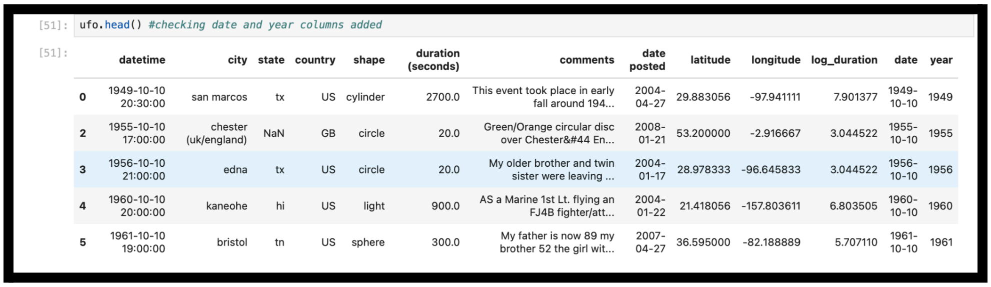
Prediction Question

Can we predict UFO sightings (where they will happen, how long they will be, where they will happen) based on observed characteristics? Is there consistency amongst responses and can we find any meaningful relationships?

a. **If yes,** we propose there might be some truth to the sightings that should be investigated. There may be multiple similar observations of UFOs coming from different people, and perhaps the consistent reports point to something worth investigating further.

b. **If not,** we propose the data is built on misinterpretations and inconsistencies. We cannot go as far as to claim complete fraudulence, but it would point to the fact that UFOs are probably not real. If we cannot successfully apply machine learning models to model the predictive power of different characteristics and measure consistency, we think that the sightings are a collective of misinterpreted sky events.

Data

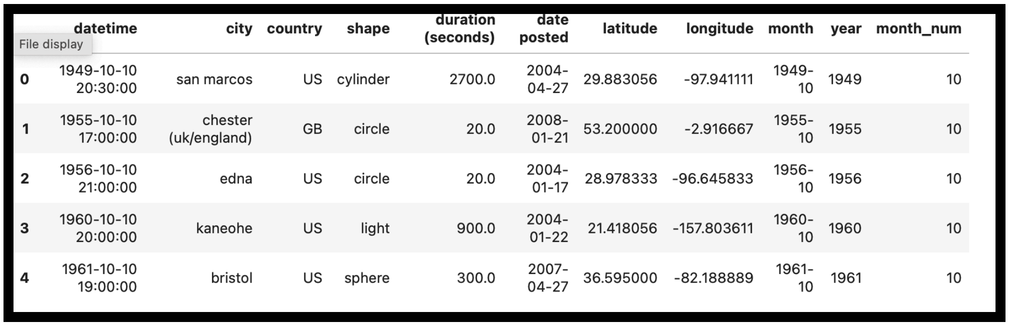


```
[51]: ufo.head() #checking date and year columns added
```

[51]:

	datetime	city	state	country	shape	duration (seconds)	comments	date posted	latitude	longitude	log_duration	date	year
0	1949-10-10 20:30:00	san marcos	tx	US	cylinder	2700.0	This event took place in early fall around 194...	2004-04-27	29.883056	-97.941111	7.901377	1949-10-10	1949
2	1955-10-10 17:00:00	chester (uk/england)	NaN	GB	circle	20.0	Green/Orange circular disc over Chester&#44 En...	2008-01-21	53.200000	-2.916667	3.044522	1955-10-10	1955
3	1956-10-10 21:00:00	edna	tx	US	circle	20.0	My older brother and twin sister were leaving ...	2004-01-17	28.978333	-96.645833	3.044522	1956-10-10	1956
4	1960-10-10 20:00:00	kaneohe	hi	US	light	900.0	AS a Marine 1st Lt. flying an FJ4B fighter/att...	2004-01-22	21.418056	-157.803611	6.803505	1960-10-10	1960
5	1961-10-10 19:00:00	bristol	tn	US	sphere	300.0	My father is now 89 my brother 52 the girl wit...	2007-04-27	36.595000	-82.188889	5.707110	1961-10-10	1961

This dataset is a collection of self-reported UFO sightings including columns like duration of sighting, shape seen, date observed, and location. To clean this dataset, we removed unnecessary variables like comments about each sighting. We also split up the datetime value in order to look at just dates, months and years. We did some overall cleaning including dropping NAs, making consistent capitalization, and coercing values to integers if needed for easier use of the data during modeling. Additionally, because we knew we would use the shape variable for predicting, we limited the potential options for shape down to only 16 possible shapes that were reported.



	datetime	city	country	shape	duration (seconds)	date posted	latitude	longitude	month	year	month_num
0	1949-10-10 20:30:00	san marcos	US	cylinder	2700.0	2004-04-27	29.883056	-97.941111	1949-10	1949	10
1	1955-10-10 17:00:00	chester (uk/england)	GB	circle	20.0	2008-01-21	53.200000	-2.916667	1955-10	1955	10
2	1956-10-10 21:00:00	edna	US	circle	20.0	2004-01-17	28.978333	-96.645833	1956-10	1956	10
3	1960-10-10 20:00:00	kaneohe	US	light	900.0	2004-01-22	21.418056	-157.803611	1960-10	1960	10
4	1961-10-10 19:00:00	bristol	US	sphere	300.0	2007-04-27	36.595000	-82.188889	1961-10	1961	10

Methods

Types of Learning

We conducted supervised learning since our data was clearly labeled, and we aimed to analyze if and which variables could predict sightings and their characteristics. We trained our models using train-test-split on data that was already collected. We had quantitative, continuous and qualitative, string data, so we decided to use regression and classification to thoroughly explore the data. By building regression models, we tried to identify significant predictors and see how they influenced duration, latitude, and longitude. Using classification helped us investigate shapes and seasons, and see if there was consistency in the observations.

Types of Models

We planned to use linear regression, decision trees/random forests, and neural networks in our analysis to examine the ability of various factors to model UFO sightings.

For linear regression, we will predict the duration of UFO sightings using country, month, and UFO shape. Our exploratory data analysis (EDA) revealed significant variation in the kernel density plots of log-transformed duration across different countries, suggesting potential patterns. If we obtain a high $R2$, it will indicate a strong correlation between these factors and sighting duration.

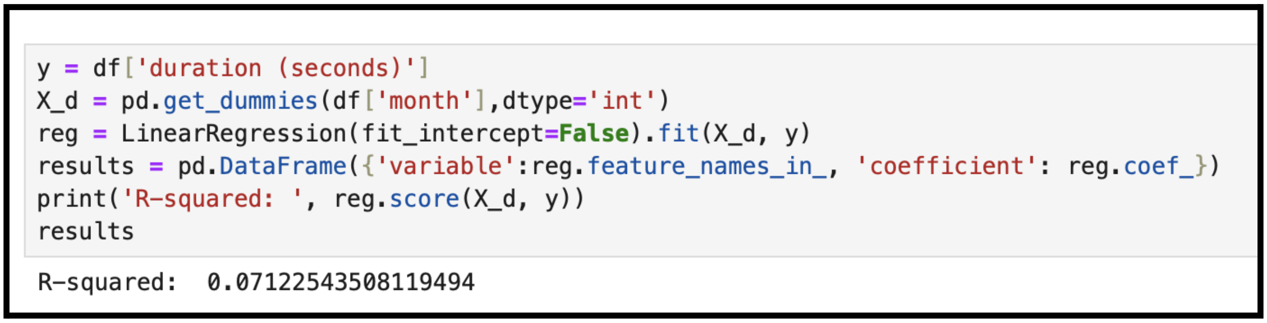
We will use decision trees to capture more complex, non-linear, and sensitive relationships. We plan to use this because of its powerful modeling capabilities. We will also use a random forest to implement bootstrapping and see if we can improve upon our results. Our predictor variables include location (country), shape, duration, latitude, longitude, year, and date.

Finally, we will use Neural Networks to look for patterns in seasonal sightings. Using variables such as location, duration, shape, and country we hope to create a neural network model that can predict the season (Summer, Fall, Winter or Spring) that a sighting will occur.

Results

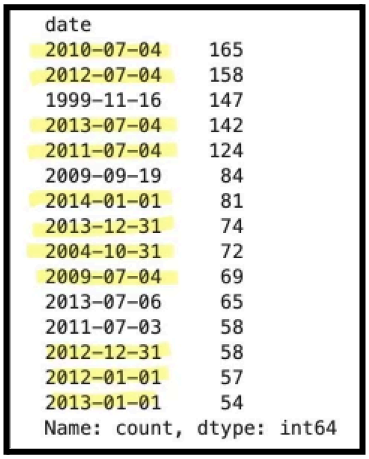
Linear Regression

We tried multiple types of linear regression in order to look for potential patterns. As we were using linear regression, the numerical value we attempted to predict was duration of sighting. We wanted to see if there were any variables that could be used as a predictor for how long a UFO would be visible to people. After trying a few different variables, we were getting extremely low r values. The highest r-value we received would come from using the month of the year that the sighting took place. I want to note, when I say highest r-squared, it was still very low. However, this r-squared value was about 100 times the size of previous run models, which led us to want to take a second look. It was this increase in r-squared that led us to taking a closer look into our date columns.



```
y = df['duration (seconds)']
X_d = pd.get_dummies(df['month'],dtype='int')
reg = LinearRegression(fit_intercept=False).fit(X_d, y)
results = pd.DataFrame({'variable':reg.feature_names_in_, 'coefficient': reg.coef_})
print('R-squared: ', reg.score(X_d, y))
results
```

```
R-squared:  0.07122543508119494
```



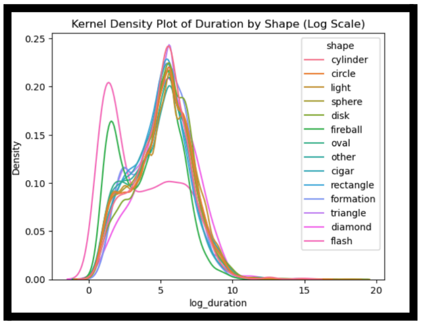
```
date
2010-07-04    165
2012-07-04    158
1999-11-16    147
2013-07-04    142
2011-07-04    124
2009-09-19     84
2014-01-01     81
2013-12-31     74
2004-10-31     72
2009-07-04     69
2013-07-06     65
2011-07-03     58
2012-12-31     58
2012-01-01     57
2013-01-01     54
Name: count, dtype: int64
```

While running a linear regression using day of the year yielded another significantly low r-squared, we did notice something interesting about the dates category. By sorting the dates by number of sightings on that date, we noticed a similarity in what days people were seeing UFOs. Of the top 15 days where the most UFOs were sighted, 11 were on holidays, 5 of which were 4th of July. This may indicate that there could be another unknown factor causing these sightings, like fireworks for example.

Decision Tree & Random Forest

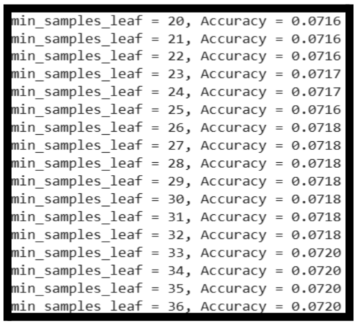
The decision tree model showed that predictability is generally low. Variables such as year, duration, month, year, shape, latitude, longitude, and country often don't have the predictive power to explain each other. We used a decision tree regressor and classifier to find a more sensitive model and capture underlying relationships, since we were unable to discover anything with linear and logistic regression.

Using a decision tree classifier, we attempted to predict the observed shape of the UFO. We ended up classifying



shapes on latitude, longitude, month, and year. The latitude and longitude were meant to account for location, since in our heatmap we saw that there may be a relationship between shape and country. Initially, the classifier was performing under 3% accuracy, which was worse than random.

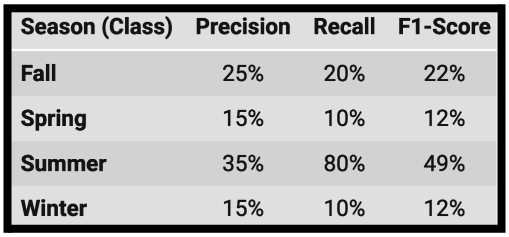
Initial exploratory data analysis (EDA) showed that duration didn't vary across shapes, except for the fireball and flash shapes. This would make sense, for the nature of these observed forms implies the duration of no more than a moment. This meant that excluding duration (seconds), even in its log-transformed form, was more beneficial in trying to model meaningful relationships. By recognizing this, and trying to make the data less noisy, we got it up to 7.2% accuracy. However, iterating across different numbers of min_sample leaves and adjusting the depth didn't change the performance of the model much. This means that the model is underfitting and unable to find any meaningful relationships in the data. We proceeded to create a more complicated and powerful model — a random forest — and we witnessed no improvement in the accuracy of results, which were just above random. This shows that the observations are widely inconsistent, and there exists little to no learnable pattern amongst them.



```
min_samples_leaf = 20, Accuracy = 0.0716
min_samples_leaf = 21, Accuracy = 0.0716
min_samples_leaf = 22, Accuracy = 0.0716
min_samples_leaf = 23, Accuracy = 0.0717
min_samples_leaf = 24, Accuracy = 0.0717
min_samples_leaf = 25, Accuracy = 0.0716
min_samples_leaf = 26, Accuracy = 0.0718
min_samples_leaf = 27, Accuracy = 0.0718
min_samples_leaf = 28, Accuracy = 0.0718
min_samples_leaf = 29, Accuracy = 0.0718
min_samples_leaf = 30, Accuracy = 0.0718
min_samples_leaf = 31, Accuracy = 0.0718
min_samples_leaf = 32, Accuracy = 0.0718
min_samples_leaf = 33, Accuracy = 0.0720
min_samples_leaf = 34, Accuracy = 0.0720
min_samples_leaf = 35, Accuracy = 0.0720
min samples leaf = 36, Accuracy = 0.0720
```

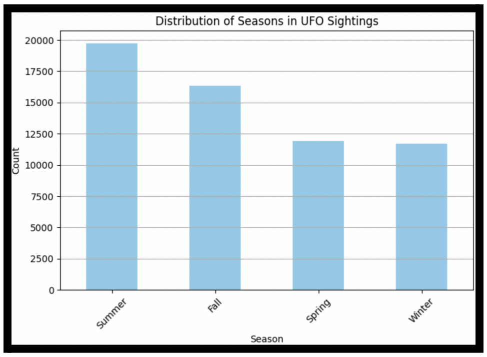
Neural Network Model

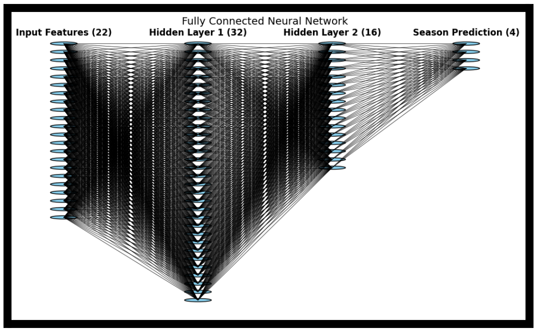
The model's performance was poor, achieving an overall validation accuracy of approximately 32%, slightly better than random guessing, and comparable to always predicting the most frequent season (~33%). Precision, recall, and F1-scores



Season (Class)	Precision	Recall	F1-Score
Fall	25%	20%	22%
Spring	15%	10%	12%
Summer	35%	80%	49%
Winter	15%	10%	12%

were especially low for Fall, Spring, and Winter. While the model recalled about 80% of Summer sightings, it frequently mislabeled Spring, Fall, and Winter as Summer, demonstrating strong bias toward the majority class. Overall macro-averaged precision, recall, and F1-score were around 22–27%, reflecting generally weak classification across seasons.





Several factors likely contributed to the poor results. The model heavily favored Summer due to class imbalance (shown in the graph) and showed little ability to differentiate the other seasons. Additionally, the input features: location, duration, country, and shape, had little direct relationship with the season of a sighting, making the task inherently difficult. The lack of explicit information like the actual date or month limited predictive power.

In conclusion, the model struggled to accurately predict the season of UFO sightings because the explanatory variables used, such as location, duration, shape, and country, lacked strong predictive power. Although the dataset included the exact sighting date, we chose not to use it, as it would have directly indicated the season and introduced multicollinearity. This outcome highlights that, without direct seasonal indicators, the remaining features provide limited signal for distinguishing between seasons. As a result, the classification task remains difficult, likely due to both data noise and weak associations between the chosen variables and the target.

Conclusion

While we may not have uncovered definitive proof of alien life, our exploration into UFO sightings revealed important insights about the limitations of self-reported data and the challenges of making predictions without strong feature relationships. Across linear regression, decision trees, random forests, and neural networks, we consistently found low accuracy and weak patterns, especially in models attempting to predict shape, duration or season. This suggests that UFO sightings, at least as captured in this dataset, are largely unpredictable and likely influenced by external factors such as holidays, human perception, or environmental noise. Our most interesting finding was the spike in sightings on holidays like the Fourth of July, hinting that social or cultural events may drive reports more than actual extraterrestrial visits. Ultimately, while our models failed to prove the existence of UFOs, they succeeded in teaching us about data limitations, model selection, and the importance of critically examining both inputs and outputs.