UFO Sightings in the United States

Jennifer Ruiz

DSC 550 Data Mining

Bellevue University

August 3, 2020

Abstract

The focus of the project at hand is to examine trends in sightings of Unidentified Flying Objects (UFOs) with a focus on sightings in the United States. Exploratory data analysis was performed on the data with a focus on identifying trends and points for further analysis. Several machine learning techniques were applied to the data, including clustering, text mining, and time series analysis. These analyses were used to make predictions regarding the frequency of sightings.

UFO Sightings in the United States

I. Introduction

Motivation: Sightings of Unidentified Flying Objects (UFO) have been occurring in the United States since the 1940s (Wikipedia, 2020). With more than 100,000 sightings reported to date, it cannot be denied that some phenomenon is occurring that is not yet understood by humans. As an aviation enthusiast, I found this data to be fascinating and wanted to gain experience applying data mining concepts in a practical manner. The prevalence of data related to this topic and the practical experience to be gained made this a good candidate for my data mining project.

Significance: Unidentified Flying Objects are defined by the United States Airforce as "any apparent aerial phenomenon whose cause cannot be easily or immediately identified by the observer" (USAF, 2020). There is a strong element of mystery and very little historical analysis on this topic. For this project, I chose to focus on trends in sightings, commonality in eyewitness accounts, and understanding sighting frequency over time.

Working Structure:

Figure 1 shows the steps that were kept in mind during this project. I began with data collection from Kaggle. Data was then cleaned and transformed prior to exploratory analysis. During EDA, various visualizations were used to better understand the data and identify trends for follow up analysis. Cluster analysis was performed, followed by text mining. The final step in the workflow was to train a model to predict frequency of sightings based on historical data.

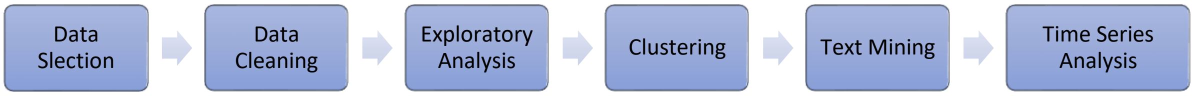

Figure 1 Project Workflow

II. Proposed Approaches

Due to the variety and complexity of the data, several different methods needed to be employed for effective analysis. EDA indicated that cluster analysis was an ideal basis to start from. Time Series analysis was also needed, as was a natural language processing feature. The following analyses were selected based on their fit for the data:

1. <u>DBSCAN Clustering:</u> This is an unsupervised learning method which performs density-based clustering. Given a set of points, the algorithm creates clusters based on Euclidean distance (Salton, 2017). This method is useful in handling data which contains valid outliers and groups them in low density areas. It is superior to k-means clustering in this regard and in the fact that it determines the optimal *k* automatically.

2. <u>Word Cloud:</u> This is a word visualization technique for frequency and importance of words in a text. The size of the word indicates its corresponding frequency in the text (Luvsandorj, 2020). The technique used natural language processing principles to measure word frequency.

3. Amira Model: This is a time series forecasting algorithm which performs univariate time series forecasting using past values (Prabhakaran, n.d.). The algorithm can be fitted to both seasonal and non-seasonal data.

III. Dataset

Data Collection: The data for this project was sourced via Kaggle from the National UFO Research Center (NUFORC). The data consists of more than 80,000 individual reports spanning from the early 1950s through 2014 and contains sightings from around the world. The dataset contained 11 variables and this project primarily focused on 5 of those. Both a scrubbed version of the dataset and a complete version were downloaded from Kaggle and used during the analysis.

Data Cleaning: Since the data was collected and aggregated over a long period of time, there were inconsistencies in formatting that needed to be handled. The biggest challenge was dealing with the number of missing values within the dataset. The first step in the data cleaning process was fill all missing values, apart from the datetime columns, with a value of ‘unknown’. The rows with missing datetime columns were dropped from the dataset, removing 243 observations. Data types for several columns were changed from object to numeric for future usage and datetime information was formatted and split into separate columns.

Data Exploration: Once data collection and cleaning was completed, exploratory analysis began. The first step in the analysis was examining trends in sightings based on year through a bar plot visualization. This analysis reveals that sightings began increasing in the early 1990s and have continued to climb in the years since. Next, a pie chart was used to visualize sightings by country of origin. The United States was shown to have the highest number of reported sightings in the dataset with 65114 observations.

Based on the preliminary EDA, I chose to shift the focus of the project to sightings in the United States only. The next step in analysis was examining sightings by state. This analysis revealed that California had the most reported sightings. I also examined sightings by state based on population. This revealed that Washington state had the highest number of sightings based on population.

The final phase of exploratory analysis focused on sighting locations and shape of object seen. Based on data visualization, the highest number of sightings occurs in June, July, and August in both the Northern and Southern Hemispheres. A bar plot visualization examined shape of the object seen, with the four most prominent shapes were lights, triangles, circles, and fireballs. The final visualization for exploratory analysis was a bar plot to examine the duration of sightings with 0-15 seconds being the most common sighting length.

IV. Implementation & Results

DBSCAN Clustering:

For the implementation of DBSCAN Clustering, the cluster method from Scikit-Learn Python library was used. DBSCAN takes core samples within the dataset and builds dense clusters around them (Harris, 2015). Outliers within the data are represented in low density areas and show no clear center. For this analysis, the longitude and latitude of each reported sighting were isolated, and the dataset was reduced to 70000 rows to save computational time. Epsilon was set to a value of 10 with a minimum number of samples per cluster being set at 30. Figure 2 shows the algorithm output which produced 7 distinct clusters, clustering outliers (as noise) in black throughout the plot.

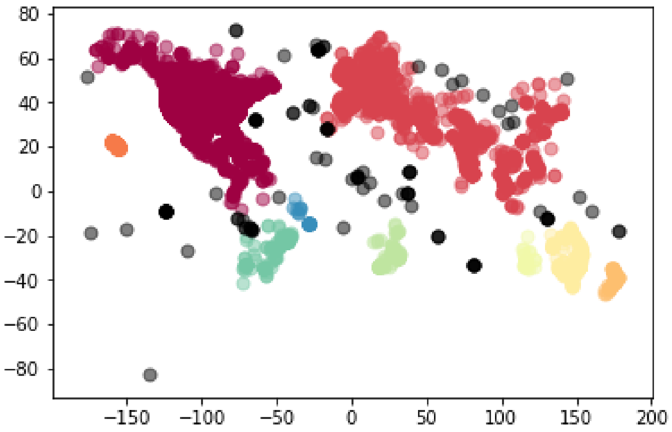

Figure 2 DBSCAN Clustering of Longitude and Latitude of Sightings

Text Mining:

For the implementation of a text mining algorithm, the Word Cloud package for Python was utilized. The package analyzes the frequency or importance of words used in a text and automatically excludes common stop words for the English language (Gurucharan, 2020). The method was used to examine the descriptions and comments used by eyewitnesses when reporting sightings. The goal was to identify trends or keywords frequently used to characterize sightings.

Figure 3 shows the results of the Word Cloud mining. The results indicate that the most common phrases or words associated with sightings are “bright lights”, “in the sky”, “white or orange light”, and references to “WA”. This last association is intriguing when correlated with the exploratory analysis performed, which highlighted Washington State as having the highest number of sightings per capita.

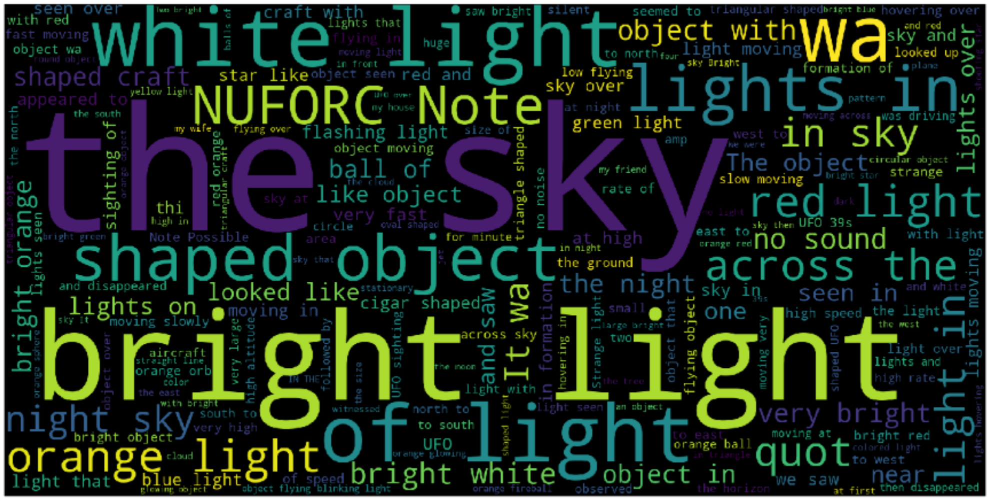

Figure 3 Common Words Used to Describe Sightings

Amira Model:

For time series analysis, the time series analysis library of Stat Models package for Python was used to implement the AMIRA model. The model functions as linear regression model which uses historical values only to forecast future values. To begin the analysis, the datetime columns were converted to datetime format and any missing values changed to zero. Sightings were then grouped by month and annual totals calculated. Time series decomposition was performed on the resulting data. The AMIRA algorithm for forecasting was then fitted to the data and predictions regarding the forecast of future sightings were made.

Figure 4 shows the visual results of the model forecasting. The model did not perform as well as expected in predicting future sightings. Upon further analysis, it was discovered that the data was highly positively autocorrelated and a seasonal auto regression should have been applied to the data. Though the original model was unsuccessful in forecasting, with the appropriate adjustments a new model could have different results.

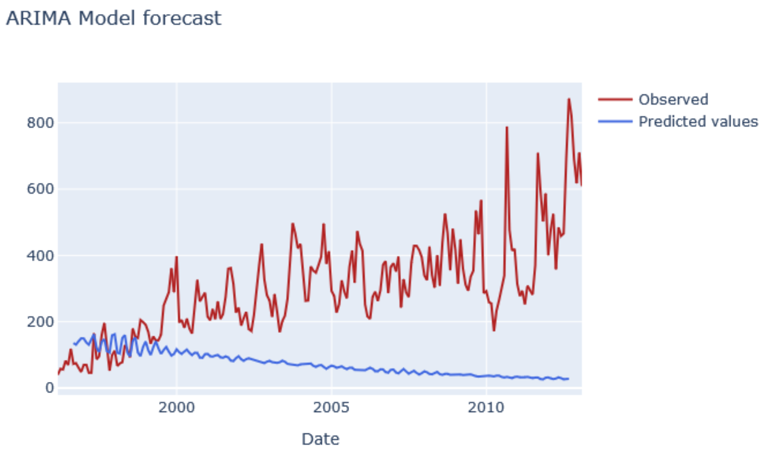

Figure 4 AMIRA Forecasting of Frequency of Future Sightings

V. Conclusions

Cluster and text mining analyses were both successfully implemented and produced actional insight. Though the time series analysis did not perform as predicted, valuable data regarding the seasonality of sightings was gained. This project can be enhanced through predicting additional characteristics of sightings and exploring additional methods to predict frequency and location of UFO sightings.

VI. Future Analyses

The focus of future analyses will center upon connecting the insights gained during this project to data collected by international organizations regarding sightings of UFOs. Additional natural language processing techniques should be applied to eyewitness descriptions to better understand trends in sighting data. Finally, a seasonal auto regression should be applied to the AMIRA algorithm used in this analysis to determine differences in predictions.

VII. Acknowledgements

This project is to fulfill a requirement of DSC550 for Bellevue University. I would like to thank Professor Deitenbeck for her feedback and guidance throughout the semester. I would also like to thank my fellow classmates for their input on potential analyses and code feedback.

VIII. References

Gurucharan, M. K. (2020). Create your own Word Cloud. Retrieved on August 4, 2020 from https://towardsdatascience.com/create-your-own-word-cloud-705798556574

Harris, N. (2015). Visualizing DBSCAN Clustering. Retrieved on August 3, 2020 from https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Luvsandorj, Z. (2020). Simple Word Cloud in Python. Retrieved on July 26, 2020 from https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5

Prabhakaran, S. (n.d.) ARIMA Model – Complete Guide to Time Series Forecasting in Python. Retrieved on July 30, 2020 from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Salton, K. (2017). How DBSCAN works and why we should use it. Retrieved on July 26, 2020 from https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

USAF. (2020). Air Force Declassification Office: UFO. Retrieved on July 23, 2020 from https://www.secretsdeclassified.af.mil/Home/Top-Flight-Documents/Unidentified-Flying-Objects/

Wikipedia. (2020). UFO Sightings in the United States. Retrieved on July 23, 2020 from https://en.wikipedia.org/wiki/UFO_sightings_in_the_United_States