Brown University

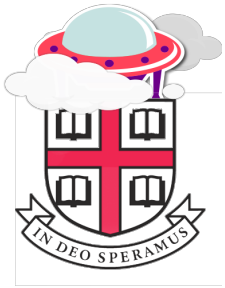
cs1951-A Final Report

UFO Sighting Analysis based on Weather and Geo Information

Author:
Jin Yan
Lei Tian
Yao Yao
Jing Qian

Supervisor:
Dan Potter
Carsten Binnig
Eli Upfal

May 9, 2017



1 Introduction

These days, people are crazy about exploring habitable planets and extraterrestrial intelligence. Meanwhile, UFO events have been happening around us for decades, and people tell their UFO encounters in thousands of ways. Predicting any aspects about UFO is intriguing since these tiny fickle light dots leave us little trace behind.

In this project, we try to investigate this topic in three steps. First, we figure out correlations between UFO sightings and data from other sources, such as geometry, weather, population and territory area. Second, based on statistic analysis and machine learning techniques, we develop models to detect fake UFO reports. Finally, we build up a web application to visualize our analysis results and to provide users a way to interact with our project, such as enabling them to report their own sightings and get access to fake detection result.

The whole data pipeline of our project is illustrated in figure 1. For the rest part of our project report, we will first discuss about data processing in section 2, and statistic analysis on UFO data in section 3. Machine learning methods will be indicated in section 4. As for web application, we will demonstrate the data pipeline and how our system works in section 5. In section 6, we will discuss the challenges we met during working on this project. Finally, in section 7, we will claim the conclusions as well as future work. In section 8, we give our profound acknowledgement to all people who help us with the project.

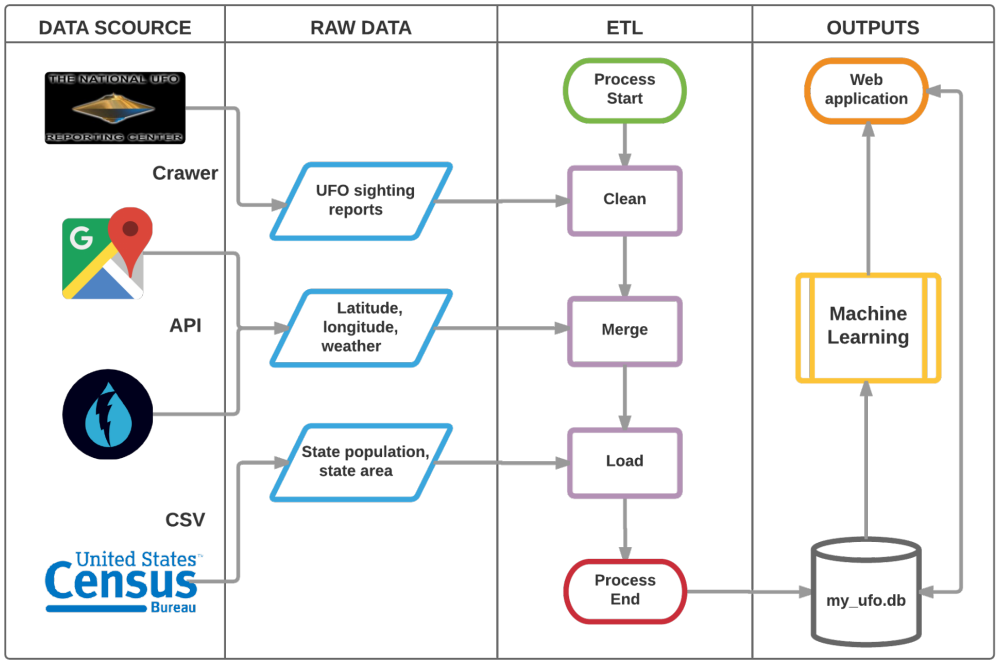


Figure 1: data pipeline architecture

2 Data

2.1 Data Source & Collection

Our data is collected from four different sources:

UFO Sighting Data, the majority of our data set, is from National UFO Report Center [1]. It is composed by 96000 UFO sighting reports and each report contains information of event date (year, month, day, hour), city, state, shape, duration, sighting description and posted date. Since this website doesn't provide an API to access the data, we implement a web crawler to collect data from its web pages.

Geo Information Data is collected from Google Map [2] which provides well-developed APIs for developers to obtain comprehensive information. We first filter out cities where UFO events happened, and then obtain longitude and latitude information of these cities. In total, about 19000 records of geo information have been collected.

Weather Data is collected from DarkSky website [3]. We collect weather conditions (icon, temperature, apparent temperature, dew point, humidity, wind speed, wind bearing, visibility and pressure) when UFO events happened. Weather data has the same size with UFO sighting data.

U.S. Area/population Data is from U.S. Census Bureau [4]. The data is downloaded manually and is stored in separate CSV files sorted by year.

2.2 Data ETL

A snapshot of raw UFO report data is as following:

b'\\r\\n2/22/17 19:30\\r\\nSwanville\\r\\nME\\r\\nLight\\r\\n45 minutes\\r\\nBright pulsating white light brighter than Venus in vicinity of Swanville, ME. ((anonymous report)) ((NUFORC Note: Venus. PD))\\r\\n2/22/17\\r\\n\\r\\n'

Figure 2: raw UFO sighting data

We separate each fields, unify units of measurement (i.e. use seconds to measure any event duration), clean the summary, and remove records that contain invalid data elements. Also, in order to make further data processing easier, we complete the following steps for each summary field:

- lowercase all characters
- strip punctuation
- remove all items in brackets
- apply Porter Stemming algorithm

And what's worth noting is that we use NUFORC's comments on each record's summary field to label report as true (1) or fake (0).

In order to distinguish each data record in database, we assign an event_id to each report data along with the weather data correlated to the sighting report. Each city is also assigned a location_id. However, considering the growing data size and the way we use location information, city's latitude and longitude data are finally integrated into each report as two additional columns. Also, due to the size of reports and weather data, we store them in different data tables, although they share the same index event_id. As for population and area data, since they are independent from UFO sightings, we create two other tables that do not have ids as indexes. Figure 3 shows the data schema of our database.

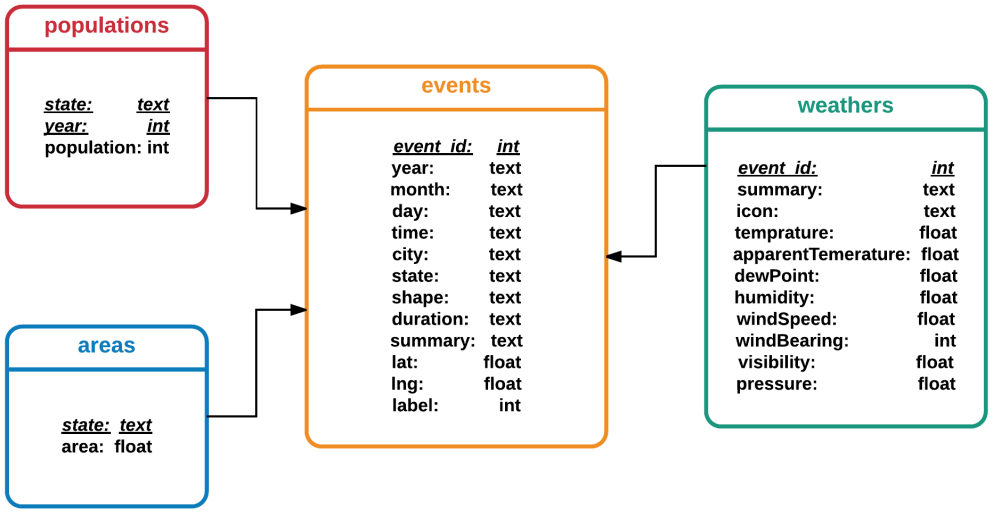
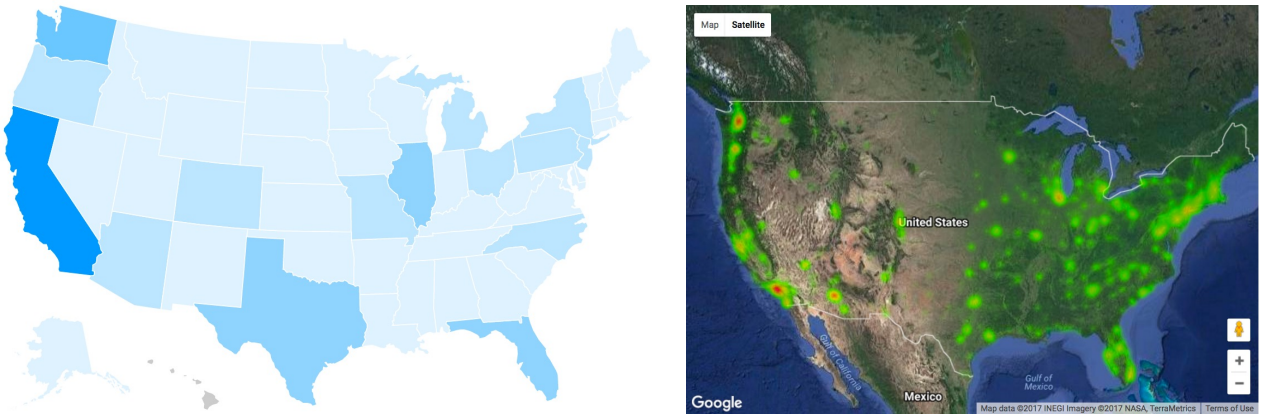


Figure 3: data schema of my_ufo.db

3 Statistic Analysis

In this section, we will discuss some statistic analysis results of UFO dataset. Figure 4(a) and 4(b) demonstrate the cumulative sighting numbers from 1950 to 2016 across U.S. The UFO sightings distributed across U.S. unevenly — California reports more UFO sightings than the other states. We believe this is because of its unique geographical conditions — on the coast and with desert. Having the same geographical conditions, Texas, Florida and Washington also report more sightings than other states. Another kind of geographical condition, near the Great Lakes, also results in more sighting numbers, for example in Illinois, Michigan, Ohio, Pennsylvania and New York. As shown in figure 5, we calculate estimated marginal means of duration of UFO sightings for each state. Just like what we expected, California also has an extremely high mean value than other states.



(a) state distribution (b) heat map

Figure 4: cumulative distribution around U.S.

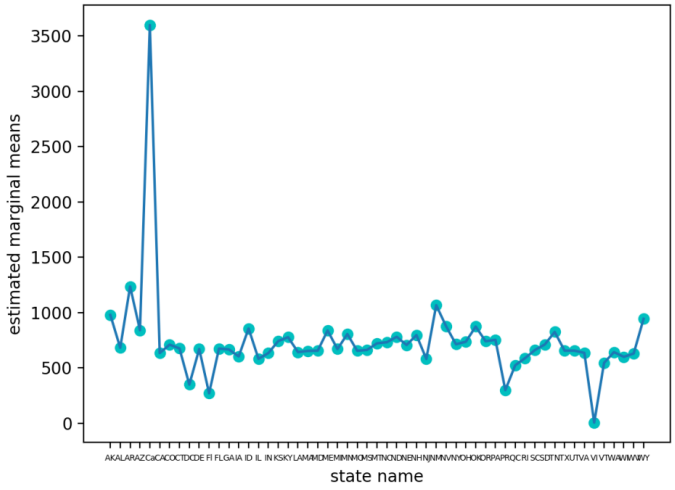


Figure 5: estimated marginal means of duration for each state

Next, we analyze relations among different fields of UFO sighting reports themselves. Pie chart 6 shows that many UFO looks like light (21%), circle (10%) and triangle (10%), with about 10% witnesses are not sure about shapes.

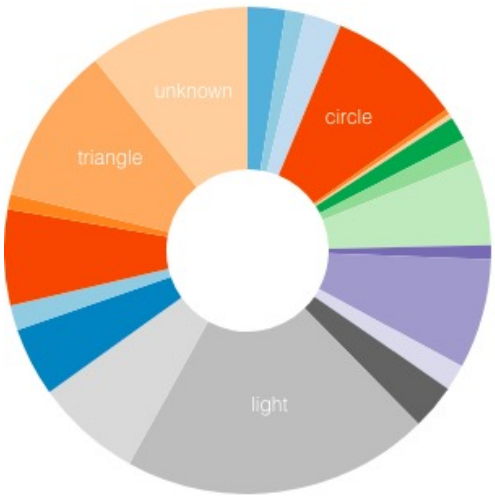
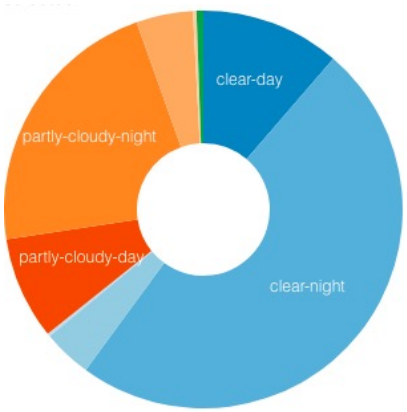
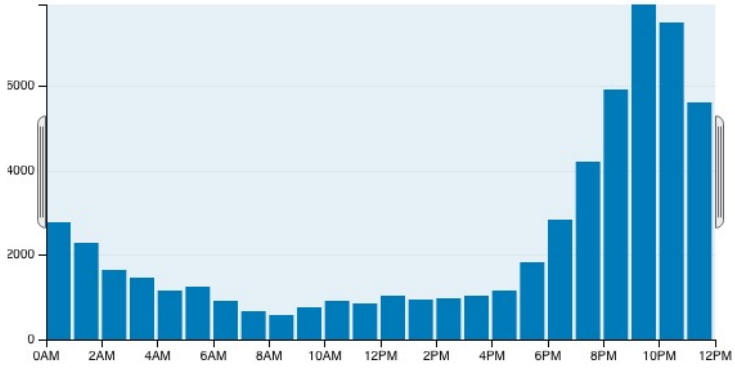


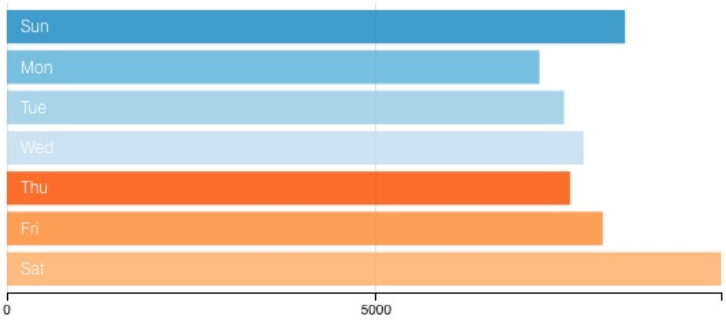
Figure 6: shape distribution



(a) weather



(b) hour



(c) weekday

Figure 7: weather and time distribution

For one thing, most UFO sightings occurred at night, regardless of how the weather is (as shown in figure 7(a)), which is consistent with figure 7(b) where most sightings occurred between 8 PM and 12 PM during a day. This observation reflects a basic rule about when UFO appears — mostly at night. For another thing, the numbers of sightings on weekends are slightly larger than those on weekdays. However, we do not accept this observation as a rule for UFO sightings. Such

phenomenon happens probably because most people are busy at work on work days and they go back home early at night for rest, while they attend outdoor activities during weekend nights. Given that UFO tends to appear during the night, people have less chance to sight one during weekdays.

4 Machine Learning

4.1 Sighting Number Regression

In order to predict how many sightings may occur in the future, we try to find the relations between the number of sightings and other macro factors, such as area, year and population data. Figure 8 shows detailed distribution of numbers of sightings from 1986 to 2016, while figure 9 shows how the number of sightings and $ln(population)$ varies relatively. It is clear that the overall number of sightings has been increasing during the last 60 years, although slightly decreased recently. However, we find there is little correlation between states' area and the numbers of their sighting reports after analyzing them thoroughly.

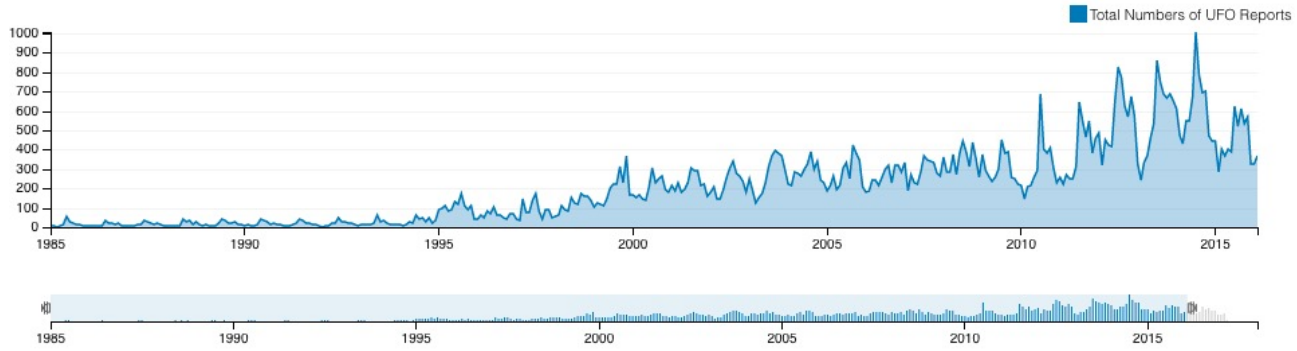


Figure 8: sighting number

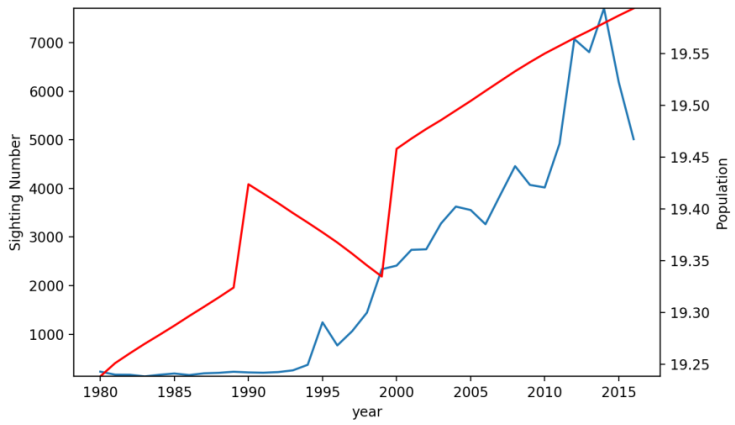


Figure 9: population and sighting number

We use four degree polynomial regression to fit year, total population of U.S. and total number of sightings. Generally, the result is perfect as our regression score is 0.96 out of 1. Given that the population of U.S. is 326474013 this year, we predict that there will be about 5589 UFO sightings across U.S. in 2017.

4.2 Fake Detection

We have encountered two main problems when working on fake detection. First, for each observation, there are two types of features — text and numeric features, which is difficult to analyze at the same time or within the same classifier. Second, as described in section 2, each report is labeled by NUFORC's comment. However, since only a small proportion (about 5000) are labeled as fake, our data is extremely imbalanced.

To tackle the first problem, we use different classifiers for different features, and then combine all results together. After testing, we choose Decision Tree and SVM with RBF kernel to train numeric features, while using Logistic Regression and Decision Tree to train summary features. Other features, such as weather condition and shape of UFO, are quantified to numbers in order to train classifiers. For summary data, we use Porter Stemming algorithm to reduce data dimension, and then vectorize all words.

To solve the second problem, we design experiments to find out the best class weight for different classifiers. Changing class weight during training is equivalent to changing loss function: if a fake report is classified as true, a much bigger penalty will be added to the objective function. For the same reason, we also create a novel judge score, rather than only the cross validation score, to choose the best class weight, which is

$$judge\_score = 0.7 * cross\_valid\_score + 0.3 * recall$$

Figure 10 shows the classifiers we use in our project. Note that except that Decision Tree model is based on text feature, the best class weight for all the other three models are around 10. This might because the number of true samples is about one order of magnitude higher than that of fake samples. Also, we find that the best judge score is from Decision Tree model regardless of the types of features we feed into classifiers.

Features	Model	Best Weight	Judge Score
Numeric	SVM(RBF)	12	0.849
	Decision Tree	10	0.922
Description	Logistic Regreesion	10	0.796
	Decision Tree	1	0.931

Figure 10: classifier performance

Figure 11 illustrates the words with top weights in SVM-RBF model. Here it only shows stems of words, since we apply Porter Stemming algorithm before load data into database. We can conclude that "Reptile" has the biggest positive score, while "meteor" has the biggest negative score. What surprises us is that "Miami", a state name, is regarded as one of the top positive words. Thus, we are able to make up an UFO event description that is true with high possibility: "a reptile-like alien gets off a cigarette-like spaceship in forest, and all of these are near me." While a fake description probably is: "a meteor-like ship with an alien that is unhuman."

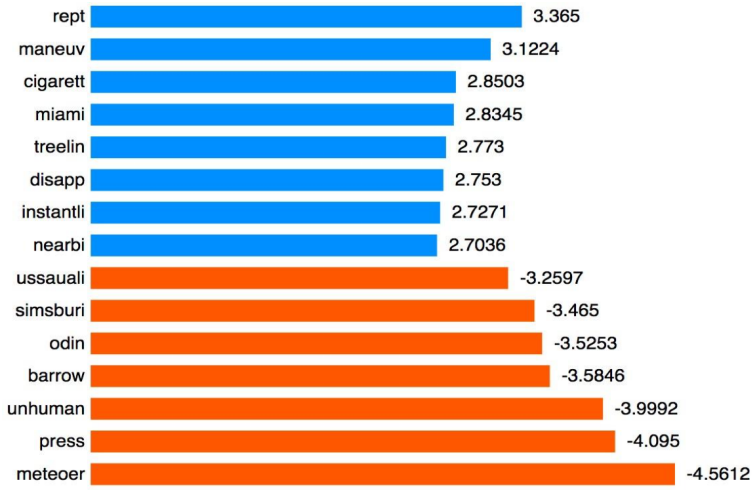


Figure 11: word weight of SVM-RBF

5 Web Application

We also build a web application to interact with users. Our web application structure is shown in figure 12. We use Node JS as web framework. For each new UFO sighting report data, we will call detection program to check it, produce truth possibility. Based on grading, we will assign true (1) or fake (0) to the report, and update my_ufo.db. The system will periodically redo model training process based on all report data.

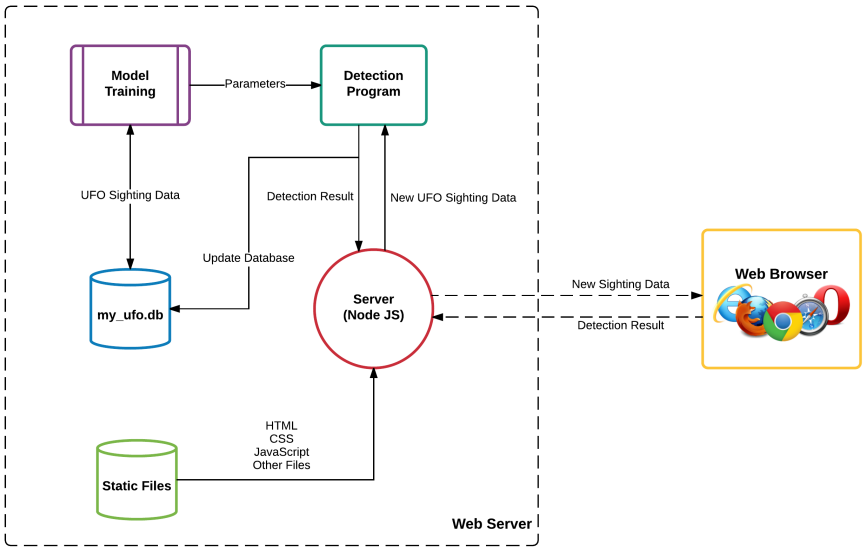
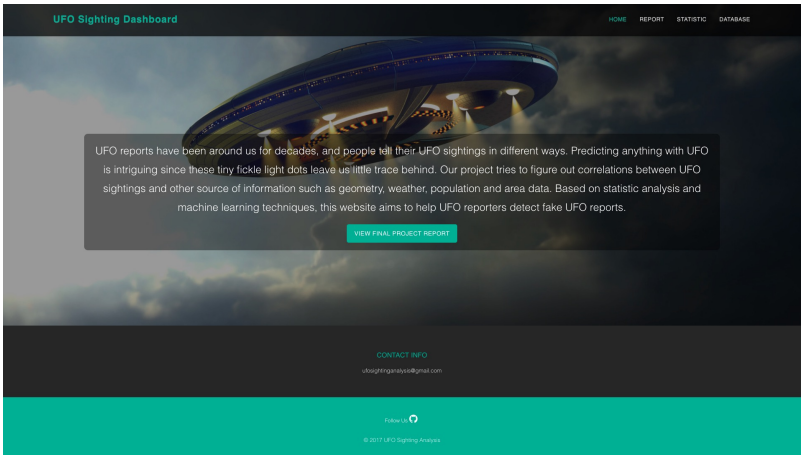


Figure 12: web structure

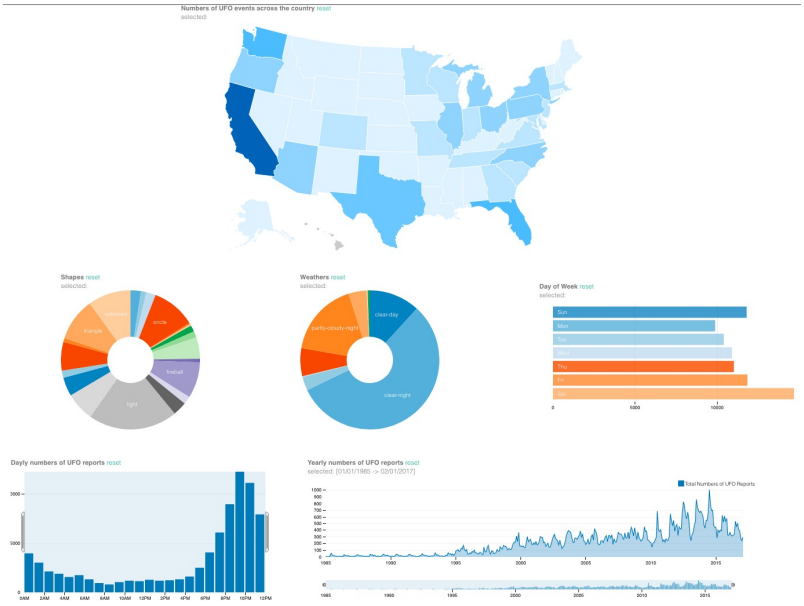
The home page of our website is shown in figure 13(a). There are three major functions of our website — reporting UFO sighting, viewing statistic results, viewing database. Report page displayed in figure 14(a) requires users to enter some information about their sightings: where, when, how long, what shape and their own summary. By clicking complete button, we will gather their reports' data and calculate their credibility by machine learning model illustrated in section 4. Figure 14(b) is a possible feedback to users. The classifier result is the possibility calculated by four classifiers separately. They vote in average to give out the final grade of user report. Numeric information lists other numeric features we gained from their report information. Information contribution is the proportion of grades between numeric and summary classifiers:

$$sum\_log + sum\_tree : num\_svm + num\_tree$$

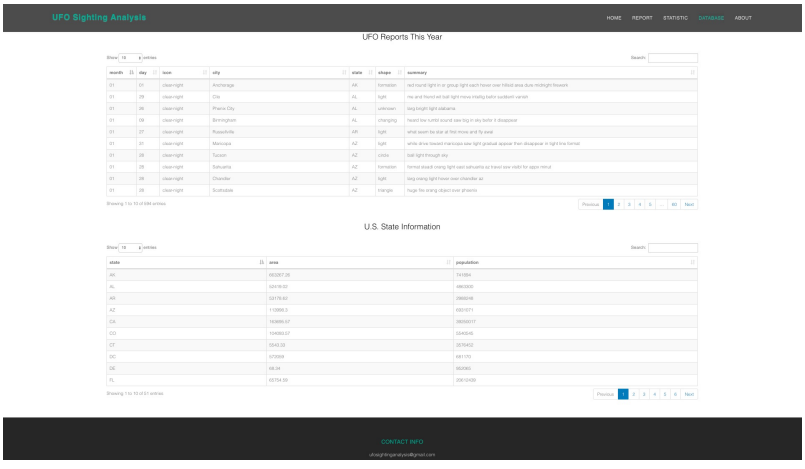
Finally, we mark out user sighting against the accumulative UFO sighting distribution around U.S in a heat map. One thing we need to emphasize is that although Decision Tree and Logistic Regression produce truth possibilities, our SVM model can only produce labels. Consequently, the classifier result of SVM will be either 0% or 100%.



(a) home page

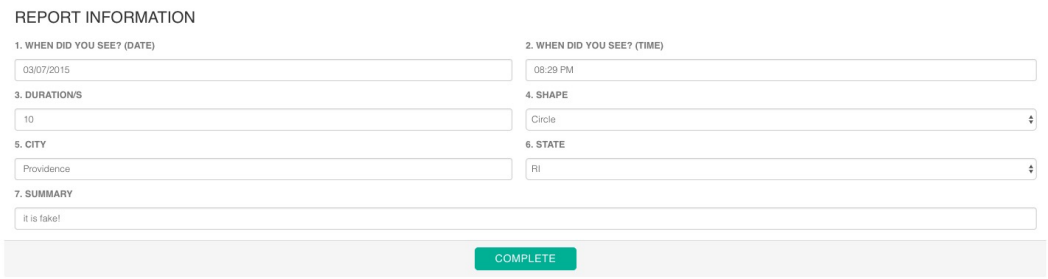


(b) statistic page

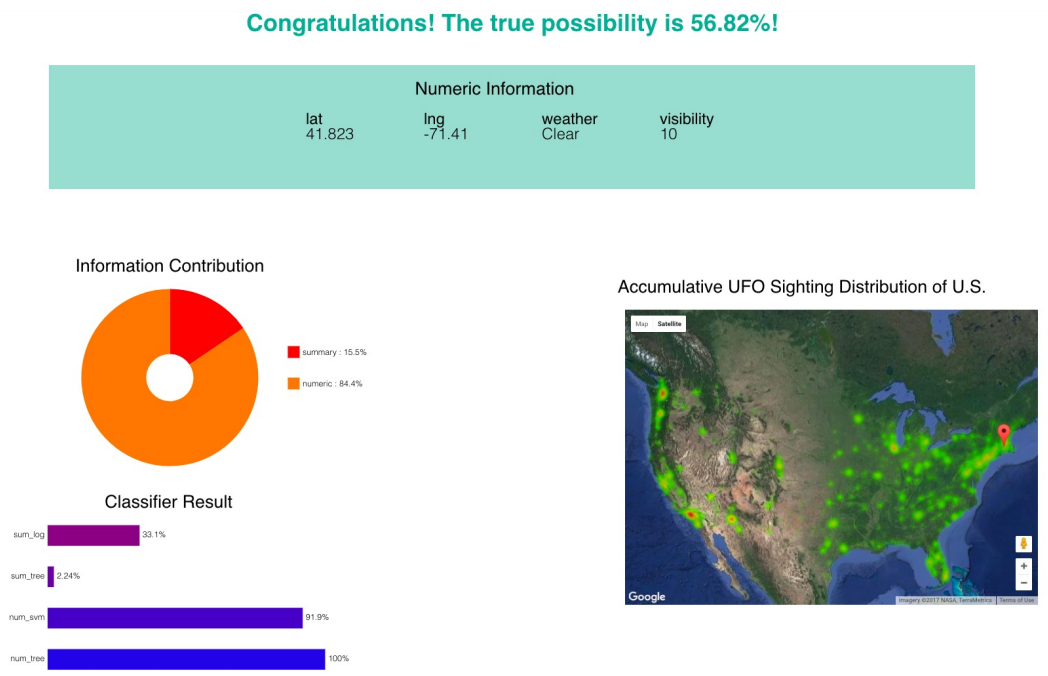


(c) database page

Figure 13: home, statistic and database page



(a) report page



(b) feedback

Figure 14: report page

Statistic page in figure 13(b) shows our analysis results described in section 3. We use crossfilter [5] to provide instant feedback to user interaction. Database view exhibits in figure 13(c) uses DataTable [6] To show UFO reports in current year, as well as population and area information of U.S. that we used for machine learning.

6 Challenges

To be honest, UFO data is really difficult to process and analyze. There are limited number of reports, and it's difficult to find out suitable models along with correct labels to predict. Actually we have tried many other ways, all of which finally failed to achieve satisfactory results.

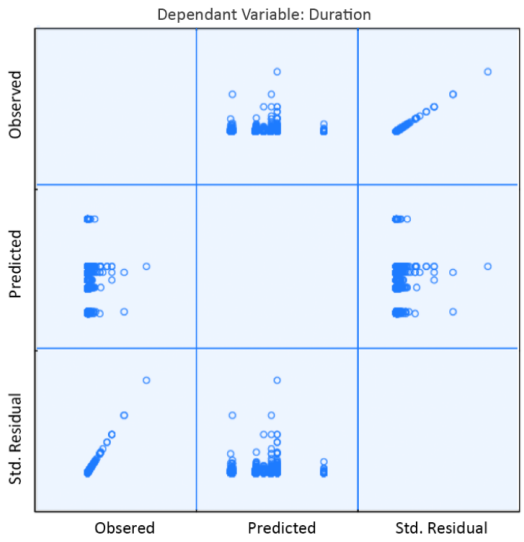


Figure 15: intercept + weather

At first, we try to predict duration, shape or location based on other features. However, when we do statistic analysis on a variety of meteorological features (temperature, wind speed, humidity, visibility, etc.,), geographical features along with the UFO reports data, most of them do not have significant correlations with other features of UFO reports data. For example, figure 15 indicates the weak correlation between other features and duration of UFO appearances.

We have tried a lot of machine learning algorithms to detect fake reports. One approach is unsupervised learning — clustering true and fake reports. However, as our training result shows, the internal cluster measures [7] are not good, actually only 0.2 out of 1. [1] Another way is to convert a classification problem on imbalanced data to an outlier detection problem only on positive data. We use one-class SVM to model true reports only. Based on true report boundary gained above, we use this model to detect which report is outlier,i.e., out of boundary. However, this method only generates a detecting accuracy slightly greater than 50%, which is absolutely unacceptable.

7 Conclusion & Future Work

According to our discussion in previous sections, we can draw following conclusions:

- Many UFO sightings occur at night, between 9 PM and 12 PM during a day, no matter whether it is clear or partly-cloudy.
- Sighting distribution across U.S. are imbalanced. States that near lakes, deserts and oceans tend to report more UFO sightings. California reports the most sightings, which also gets highest estimated marginal means of sighting duration.
- Whether an UFO sighting report is true or just a hoax has close relationship to numeric features we extract. By using location information, time of a day, UFO shapes, weather conditions, our models can detect fake reports with a relatively high accuracy.

[1] $1 - \frac{compactness}{separation}$

- For text features, the word "reptile" has the biggest positive score, while the word "meteor" has the biggest negative score.
- Total UFO sighting number increases as year goes by, although slightly decreased recently. By using population and year as independent variable, our regression model predicts that there maybe 5589 UFO sightings in 2017.

Many further work can be done on UFO sighting data research. First and foremost is to find a better way to label data. NUFORC labeled data manually based on other information. For example, if a sighting's description and environment conditions are similar to satellite launching, they may label it as a fake report. It is possible to build a machine learning model based on related news analysis, which could label UFO reports automatically and more accurately.

Another thing is that we believe UFO sightings are related to many other different factors, like economy and vegetation. Consider what we derived in section 3. We conclude that California has more UFO sightings than the other states because of topographical features of desert and coast. But is this the truth? Is it possible this conclusion results from it's great economy, or any other conditions that is unique for the Golden State? We don't know. Many other data need to be collected, some of which even need some field research.

After all, we believe UFOs, or aliens, do exist. But finding out rules of their occurrence on Earth is really difficult. Hope our work on analyzing sightings across U.S. could help further studies on this field.

8 Acknowledgements

We wish to show our profound gratitude to professors teach us data science. Most of our methods are learned from their great lectures. Meanwhile, we are grateful to our TAs', especially Michael Xu, our mentor TA, selfless help.

References

[1] The National UFO Reporting Center

[2] Google Map API

[3] Dark Sky API

[4] United States Census Bureau

[5] dc.js - Dimensional Charting Javascript Library

[6] DataTable

[7] Liu, Yanchi, et al. "Understanding of internal clustering validation measures." Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.