Brown University

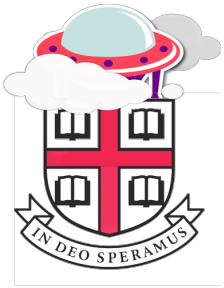
cs1951-A Final Report

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 7, 2017



1 Introduction

People are crazy about exploring habitable exoplanet and extraterrestrial intelligence these days. Meanwhile, as an important event with aliens, UFO reports have been around us for decades, and people tell their UFO encounters in thousands of different ways. Predicting anything with 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. Firstly, we figure out correlations between UFO sightings and other source of information such as geometry, weather, population and area data. Secondly, based on statistic analysis and machine learning techniques, we develop models to detect fake UFO reports. Finally, we build a web application to visualize our analysis results, as well as enable users to report their own sightings.

The whole data pipeline architecture of our project is illustrated in figure 1. In section 2, we discuss our processing about data. In section 5, we demonstrate how our system works, and will focus on the web application we achieved. Statistic results will be shown in section 14(a) and machine learning methods will be shown in section 4. Section 6 discusses the challenges we meet when analyzing data. Finally, section 7 gives out project conclusions, as well as future work. Section 8 gives our kind appreciate to anyone helps us.

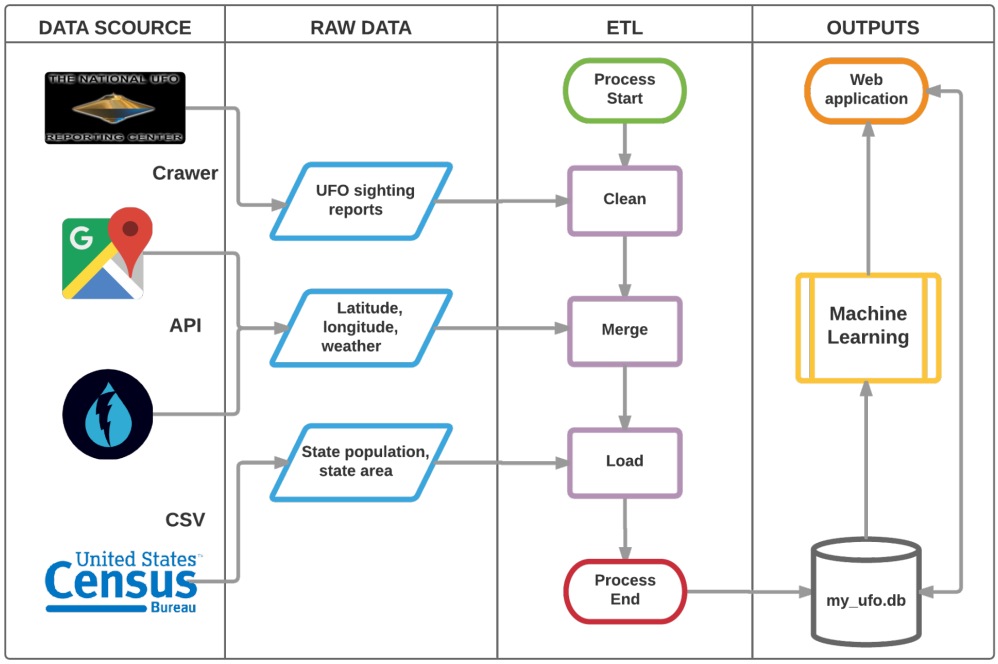


Figure 1: data pipeline architecture

2 Data

2.1 Data Source & Collection

Our data is collected from four different data sources:

UFO Sighting Data is from National UFO Report Center [1]. It is the main part of our data set. For each sighting, it contains time (year, month, day, hour), city, state, shape, duration, summary and posted date. This website doesn't provide API, so we write a web crawler to collect data from it. There are about 96000 sightings.

Geo Information Data is collected from Google Map [2]. Google Map provides convenient API for us. To get the best efficiency, we first statistic cities that have sightings, and only require longitude and latitude information of those cities. About 19000 city information needs to be collected.

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

U.S. Area/Population data is from U.S. Census Bureau [4]. Data is stored in separate .CSV file. We just manually download those .CSV files. It really takes us some time to do so.

2.2 Data ETL

Within our raw data, only NUFORC data needs to be cleaned. A possible raw UFO report may look like this:

```
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 need to separate each item, unify unit (like for duration), clean the summary, and drop reports that contain blank items. Each duration is unified to seconds. Also, to make further data process easier, we do the following steps for each summary:

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

Note, we use NUFORC's comments on each summary to label report as true (1) or fake (0).

In order to organize database, We assign an event_id to each report, along with the same event_id for each weather correlated to that sighting report. Each city is also assigned a location_id at first. However, considering it's size and how we use location information, city latitude and longitude are finally integrated into each report, as two new columns. Because of the size of reports data and weather data, we store them in to different tables, although they share the same index event_id. As for population and area data, since they are independent from UFO sightings, we build two other tables, which don't use index. Figure 3 shows the data schema of our database.

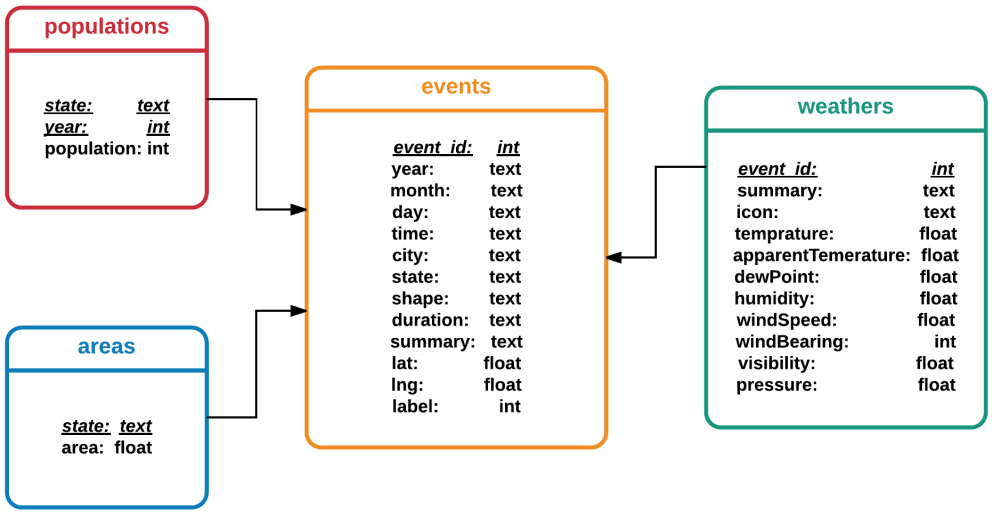
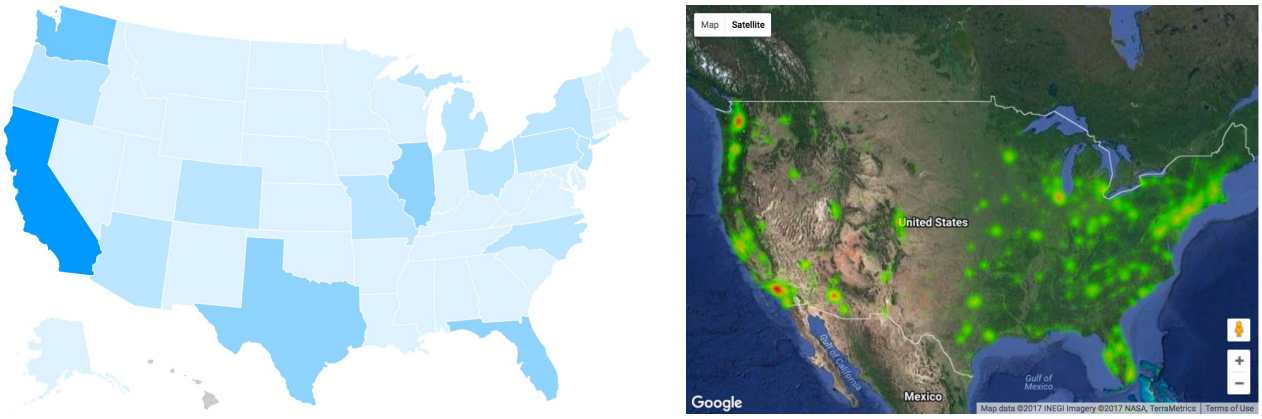


Figure 3: data schema of my_ufo.db

3 Statistic Analysis

In this section, we discuss some results based on our statistic analysis of UFO dataset. Figure 4(a) and 4(b) demonstrates the cumulative sighting numbers from 1950 to 2016 around U.S. The UFO sightings distributed around 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 conditions, Texas, Florida and Washington also report more sightings than others. Another kind of geographical condition, near the Great Lakes, also results of more sighting numbers, like Illinois, Michigan, Ohio, Pennsylvania and New York. As shown in figure 5, we calculate estimated marginal means of duration for each state. Just like what we expected, California also has an extremely high mean than other states.



(a) state distribution (b) heat map

Figure 4: cumulative distribution around U.S.

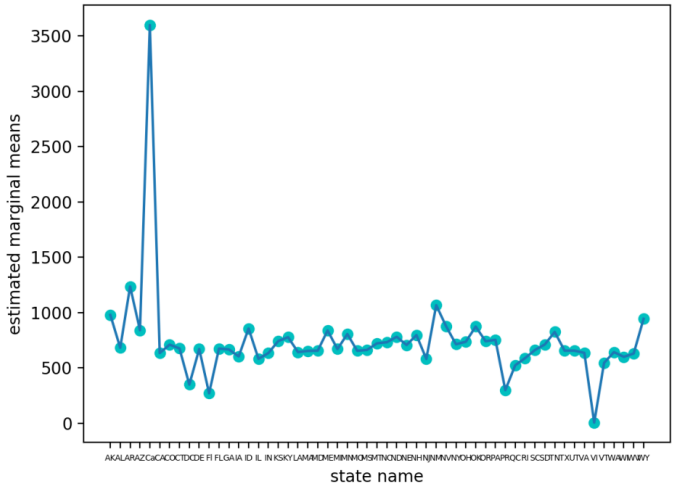


Figure 5: estimated marginal means of duration for each state

Next, we analyze 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 shape.

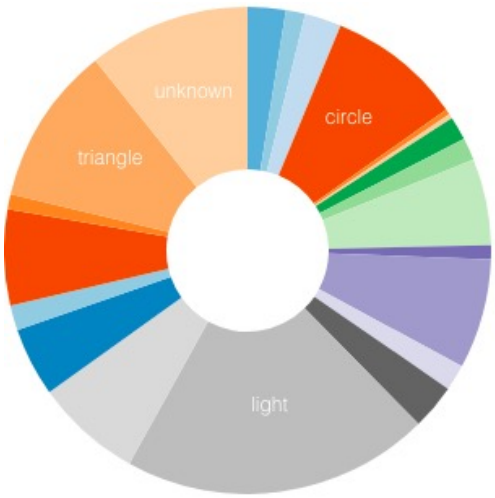
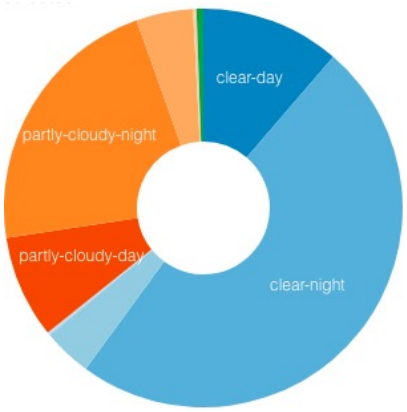
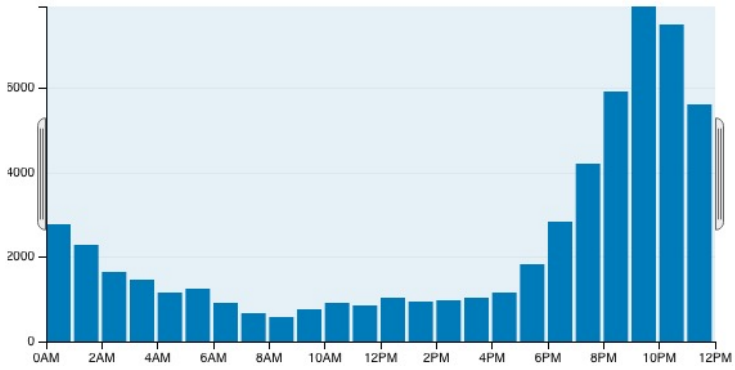


Figure 6: shape distribution

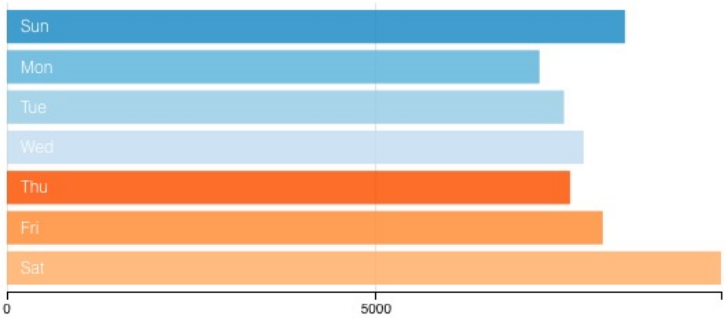
On the one hand, most UFO sightings occurred at night, no matter clear or cloudy the weather is, as shown in figure 7(a), which agrees with figure 7(b), as most sightings occur between 8PM and 12PM. This reflects a basic rule of UFO appearance — at night. On the other hand, sighting number on Sunday and Saturday is a little more than weekdays, although not much. We don't feel this make sense to be a rule for UFO sightings. It is just because of most people are busy at working, and come home early at night during weekday.



(a) weather



(b) hour



(c) weekday

Figure 7: weather and time distribution

4 Machine Learning

4.1 Sighting Number Regression

In order to predict how many sightings may occur in the future, we need to find the relationship between number of sightings and other macro factors. Such as area, year and population. Figure 8 shows detailed distribution of sighting number across year, while figure 9 shows how the number of sightings and log(population) varies relatively. It is clear that the number of sightings increases as a whole during the last 60 years, although slightly decreased recently. However, by analyzing states' area and their own sighting report number, there is little correlation between them.

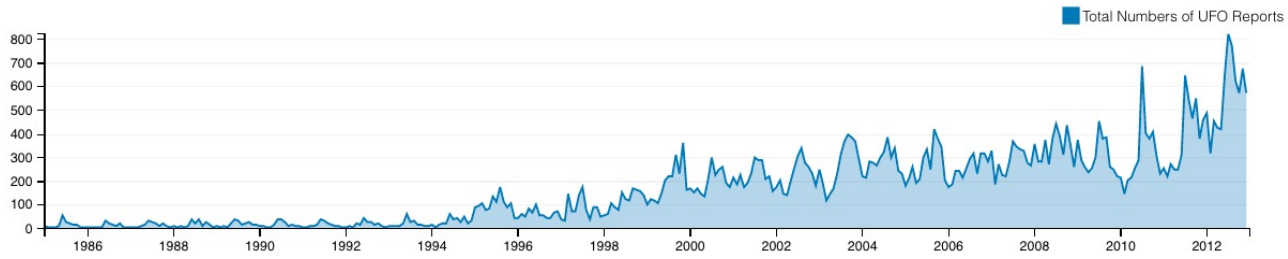


Figure 8: sighting number

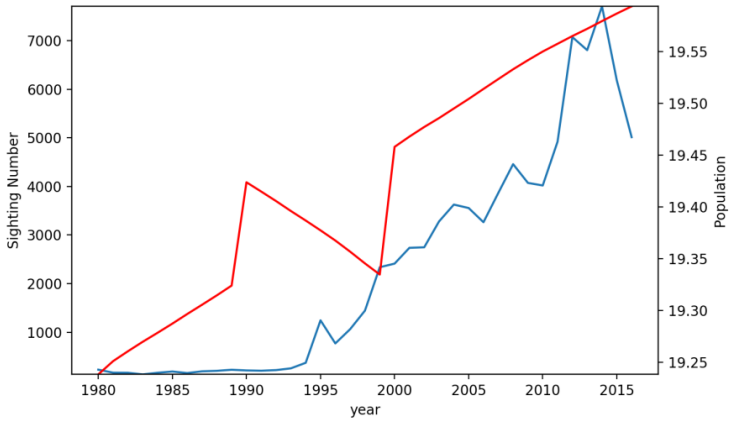


Figure 9: population and sighting number

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

4.2 Fake Detection

We encounter two main problems when doing fake detection. For each sample, there are text feature, say report summary, and numeric features. It is difficult to analyze them at the same time, or within the same classifier. Another difficulty is that as described in section 2, each report is labeled by NUFORC's comment. Since only a small set (about 5000) are labeled as fake, our data is extremely imbalanced.

For the first problem, we use different classifiers for different features, and then fuse all results together. After testing, we choose decision tree and svm with rbf kernel to train numeric features, and logistic regression and decision tree to train summary feature. Some features, like weather condition and shape, 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. For the second problem, we do many experiments to find out the best class weight for different classifiers. The judge score we use is:

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

Figure 10 shows the classifiers we use in project. Note that except for decision tree model of description 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 of fake samples. Also note that the best judge score is from decision tree model, no matter what kind of features.

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 words with top weight in svm-rbf model. Since we apply Porter stemming algorithm before load data into database, here it only shows stem of words. Reptile has the biggest positive score, while meteor has the biggest negtive score. What surprises us is that Miami, as a state name, is regarded as one of the top positive word. So we can assume a situation that mostly to be true: a reptile liked alien gets of of a cigarette like ship in forest, and all of these are near witness. While a mostly fake report maybe: a meteor like ship with an alien that is unhuman.

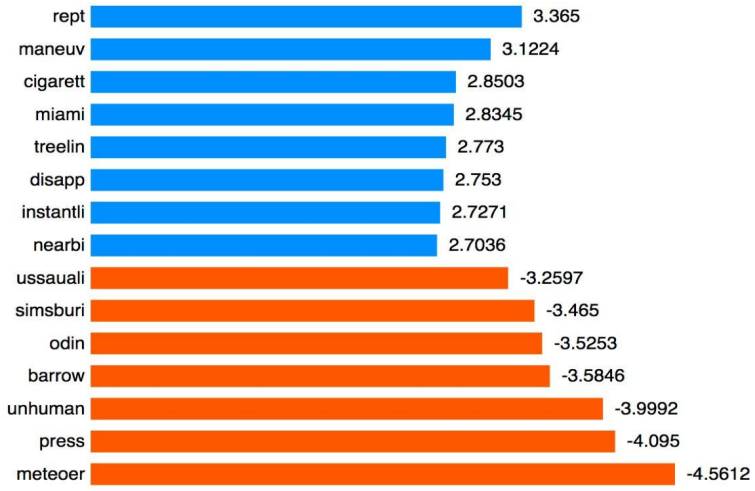


Figure 11: word weight of svm-rbf

5 Web Application

We also build a web application to interact with users. The home page is shown in figure 12. There are three major functions of our website — reporting UFO sighting, viewing statistic result, viewing database.

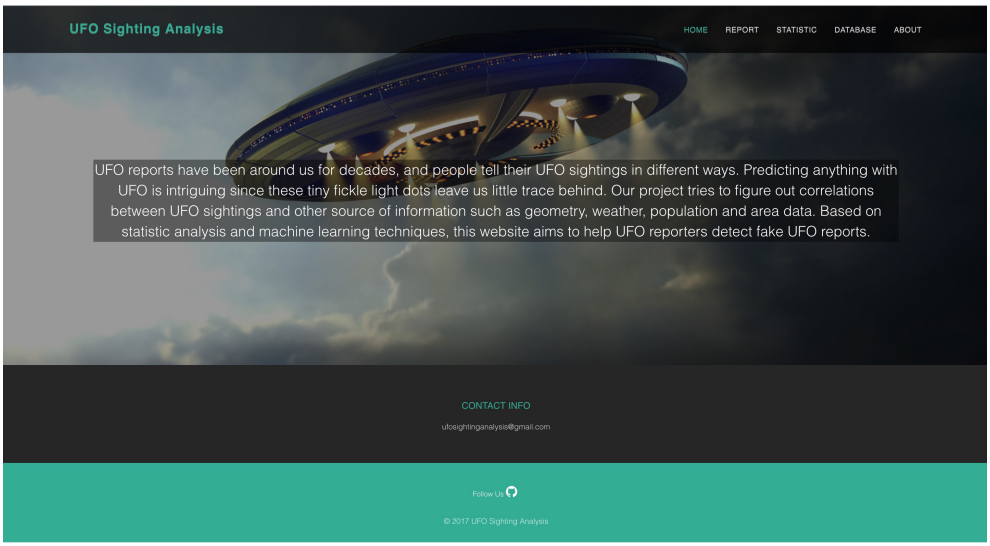
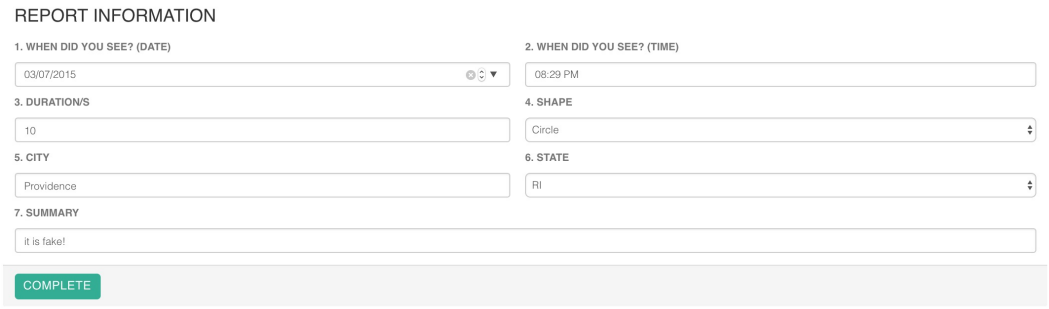
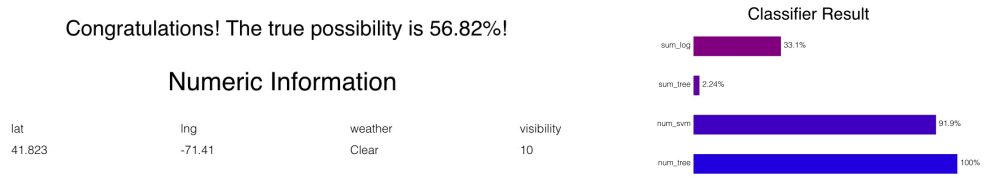


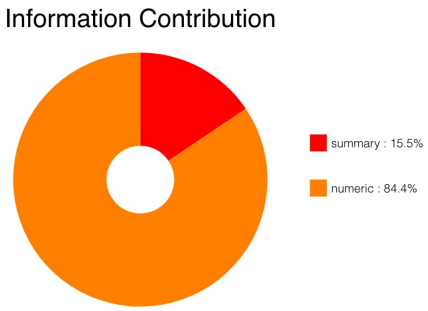
Figure 12: home page

Report page displayed in figure 13(a) requires user to fill in some information about their sighting: where, when, how long, what shape and their own summary. By clicking complete, we will gather his report and calculate the credibility of it by machine learning model illustrated in section 4. Figure 13(b) and 13(c) are 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

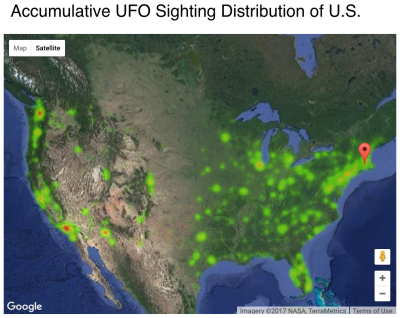


(a) report page





(b) feedback-1



(c) feedback-2

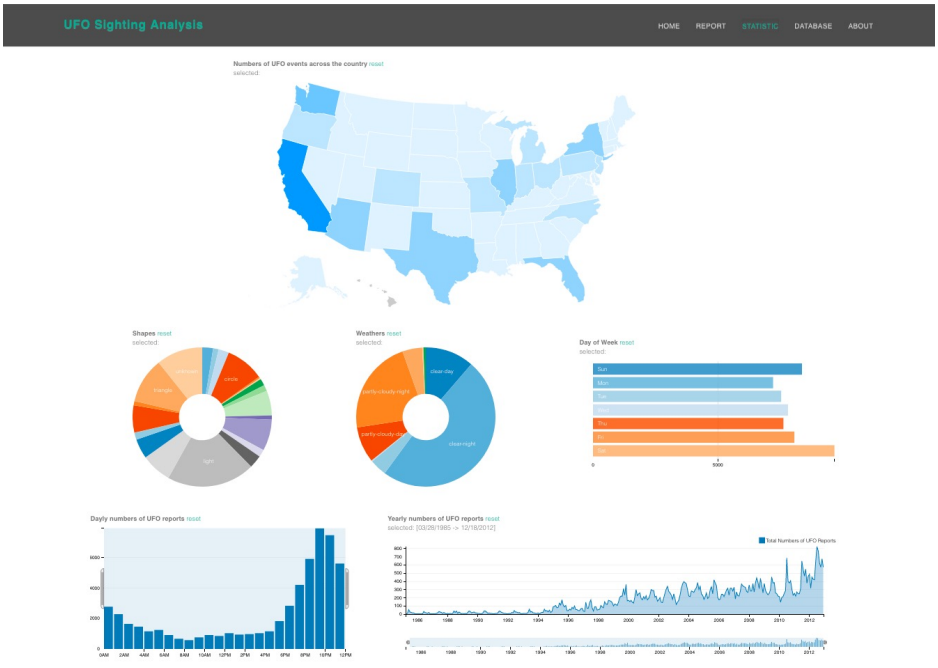
Figure 13: report page

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:

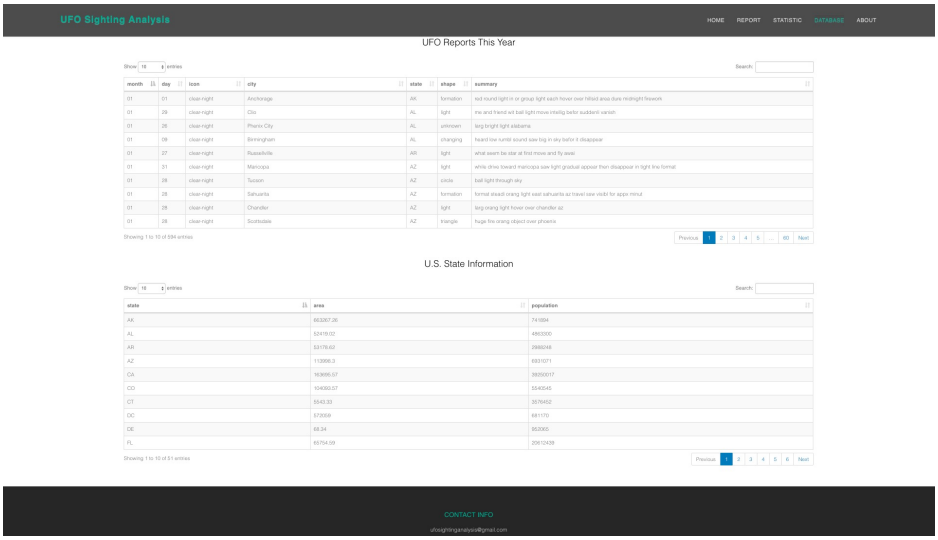
$$sum\_log + sum\_tree : num\_svm + num\_tree$$

Finally, we mark out user sighting against the accumulative UFO sighting distribution around U.S.

Statistic page in figure 14(a) shows our results described in section 14(a). We use crossfilter [5] to provide instant feedback to user interaction. Database view exhibit in figure 14(b) 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.



(a) statistic page



(b) database page

Figure 14: statistic and database page

6 Challenges

To be honest, UFO data is really difficult to handle and analyze. Not only because of limited number of reports, but also the difficulty to find out suitable models along with suitable labels to predict. Actually we have tried many many other ways, all of which finally failed to achieve our destination.

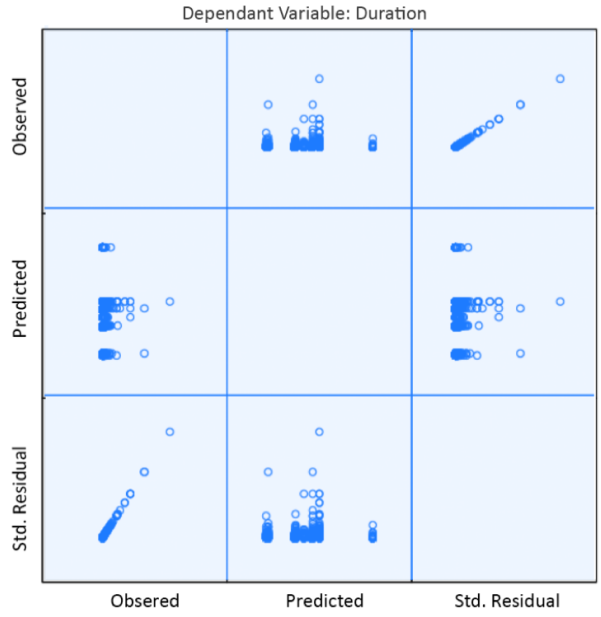


Figure 15: intercept + weather

At first beginning, we try to predict duration, shape or location based on other features. However, our statistic analysis has tested with a range of meteorological features (temperature, wind speed, humidity, visibility, etc.,), geographical features along with the UFO reports. Most of them do not have significant correlations with any features of UFO we have. For example, figure 15 indicates the weak correlation among features with duration of UFO appearances.

We have tried a lot of machine learning algorithms to detect fake reports. One approach is unsupervised learning — cluster 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. Another way is to convert a 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 bigger than 50%, which is absolutely unacceptable.

7 Conclusion & Future Work

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

- Many UFO sightings occur at night, between 9PM and 12PM, no matter it is clear or partly-cloudy.
- Sighting distribution around U.S. are imbalanced. States that near lakes, deserts and oceans tend to report more sightings. California reports the most of them, which also gets highest estimated marginal means of duration.
- Whether an UFO sighting report is true or just a hoax has close relationship to numeric features we collect. By using location information, time of a day, UFO shape, weather

condition, our models can detect fake report with a relatively high accuracy.

- For description feature, reptile has the biggest positive score, while meteor has the biggest negative score.
- Total UFO sighting number increases according to year, while slightly decreased recently. By using population and year as independent variable, our regression model predicts that there maybe 5589 UFO sightings in 2017.

8 Acknowledgements

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.