UFO Sighting Analysis based on Weather and Geo Information

Jin Yan (jyan16), Lei Tian (ltian1), Yao Yao (yyao17), Jing Qian (jqian7)
Dan Potter, Carsten Binnig, Eli Upfal CS1951-A, Spring 2017

INTRODUCTION & HYPOTHESIS

- UFO reports have been around us for decades, and people tell their UFO sightings in different ways. Predicting anything with UFO is intriguing since these tiny fickle light dots leave us little trace behind.
- We will try to investigate this topic in three steps. Firstly, we will 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 will develop models to detect fake UFO reports. Finally, we will build a web application to visualize our analysis results, and enable users to report their own sightings.

DATA ETL

- Data source:
  UFO reports data: NUFORC(National UFO Report Center)
  GIS data: Google's API
  Weather data: DarkSky
  U.S. population data: U.S. Census Bureau

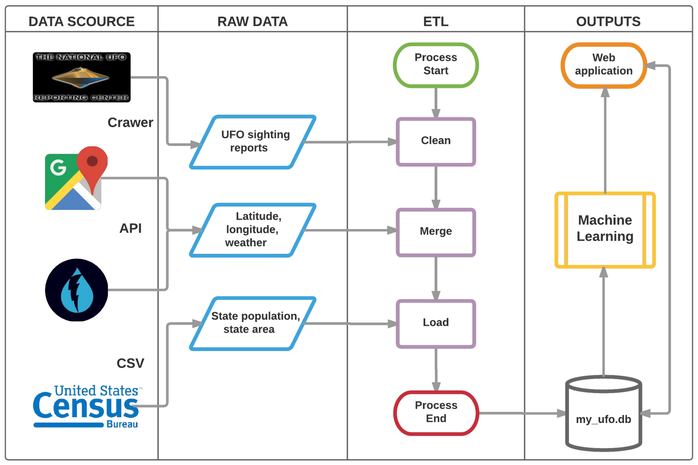

Figure 1: Whole Data Pipeline Architecture

- Data collection:
  We write python scripts to scrap UFO reports data from NUFORC online database, use Google's API to get latitude and longitude data and use Darksky API to get weather data and download CSV files from U.S. Census Bureau to get population data.
- Data integration:
  To clean UFO reports data, we use regular expressions to remove useless HTML tags, remove rows containing columns of "Unknown" or "Unspecified" and transform date format to yyyy-mm-dd, hh:MM:ss. We use event-id in UFO reports as a key to merge UFO data, location data and weather data. Then we create a sqlite3 database and load them from CSV files into it.

METHODOLOGY

- Fake detection:
  We use four models to do fake detection: decision tree and logistic regression to train text features, decision tree and SVM to train numeric features. The average of the probabilities generated by the four models can be used to predict the reliability of the UFO reports.
- Sighting number Analysis:
  We draw a heatmap to show the distribution of accumulate UFO sightings, and polynomial regression to analyze it. Some statistic methodology are used.
- Website building:
  We build a web application for users to report their UFO sightings, view our checking results of their reports, and the statistic conclusion of this topic. The structure of our web application are shown in figure 1.

STATISTIC RESULT

Figure 2: Statistic Result of Sighting Reports

- Most sightings occurs during clear night, as well as partly cloudy night. Weekends have a little more sightings than weekdays. The number of sightings is increasing according to year, while a little bit decreasing recently.
- For the sighting distribution around the U.S, California reports more than any states, and with a much higher estimated marginal means of duration. The abundance of reports might due to its geography condition: on the coast and with desert. With the same condition, Texas and Florida and Washington state also have many reports than others. We need more data perhaps on the night time activities and air traffic in order to explain the spike of duration for California.

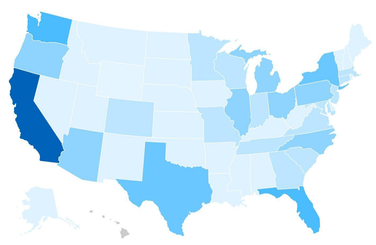

Figure 3: Cumulative Distribution around U.S.

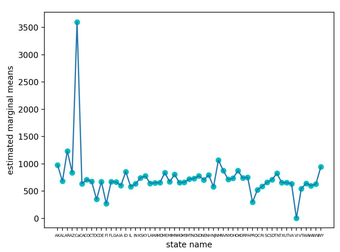

Figure 4: Estimated Marginal Means

MACHINE LEARNING

- Our machine learning work focuses on supervised learning algorithms to detect fake reports, and use regression method for distribution analysis.
- We use logistic regression and decision tree to process summary data, as well as rbf kernel SVM and decision tree to check numeric data. All these classifiers vote in average to give out the final grade of user reports. The words in top eight weights for summary are shown in figure 6. The classifier performance is listed in table 1. We choose class weight based on judge score: **cross_validation_mean * 0.7 + recall * 0.3**

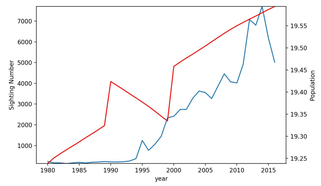

Figure 5: Population and Sighting Number

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

Table 1: Classifier Performance

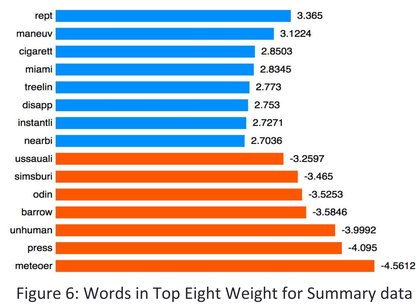

Figure 6: Words in Top Eight Weight for Summary data

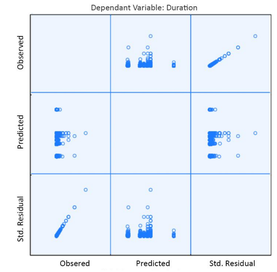

Figure 7: Intercept + weather

- We use 4 degree polynomial regression to fit UFO sighting number of a state with year and state population. The total number of UFO sightings have higher correlation to year and log(population), as illustrated in figure 5, while not much with area. Our regression score is **0.96 out of 1**.

CHALLENGES & CONCLUSIONS

- Challenges: 1) UFO data has two kinds of features: text features and numeric features, which are difficult to analyze at the same time. 2) Distribution of data with different labels is imbalanced. It caused difficulties in training our models. 3) Correlation relationships are hard to find, as figure 7 indicates the weak correlation among the features with duration of UFO appearances.
- Our work: we tried multiple classifiers to train our data, assigned different class weight to data, compared the cross validation scores and averaged the output of different models and found a feasible fake detector of UFO reports.
- Conclusions: 1) UFO sightings has close relationship to weather and time of a day. Also, the number of UFO reports shows a trend of increment by year. 2) Words in UFO reports can be used as indicators to detect fake reports. An interesting thing is that "reptile" give the most positive influence. 3) Sightings distributed around the U.S. are imbalanced. States that near lakes, deserts and oceans tend to report more sightings.