**Capstone Project I Report**

**Problem:**

Since 1961, Yuri Gagarin started humanity’s first journey to the universe. Humankind has never stopped exploring the universe. One of the important topics people always discuss is whether we are alone or there are other civilizations in the universe. There are hundreds of reports of sighting UFO every year. However, we still do not have conclusive evidence to prove aliens exist. Therefore, me as a fan of science fiction, I want to do a prediction on the month of UFO appearance so that in the future, base on the weather conditions we can prepare for the UFO sighting, and hopefully we are able to record this moment and help scientists to do further research on it.

**Client:**

NASA, astronomer, physicist and biologist would love to capture vibrant videos or pictures of UFO to do more research and study on it. If we can predict duration of sighting UFO base on the weather information, we will have higher chances of capturing high quality UFO videos or pictures.

**Data:**

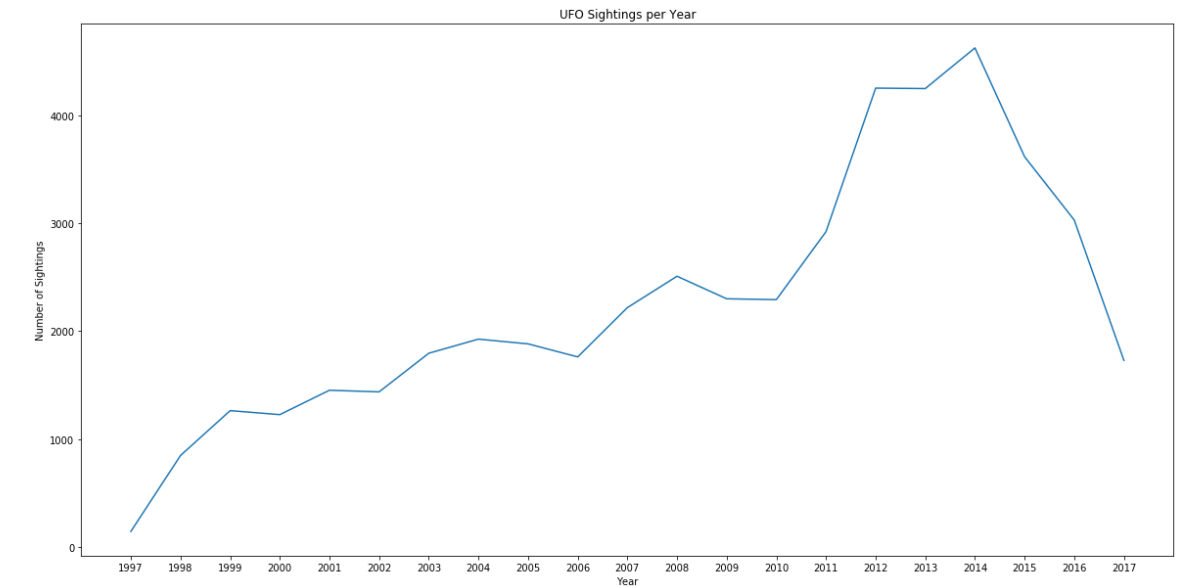
I used the “Consolidated UFO and Weather Data” data from <https://www.kaggle.com/emorelli/consolidated-ufo-weather-data>. This data contains over 51,871 reports of UFO sightings from 1997 to 2017 in 8845 cities. It provides the information such as color, shape, and duration of UFO sighting. Also, it contains historical weather data for every day from 1997 to 2017 such as pressure, temperature, humanity, dew point, wind direction, wind speed, visibility. Furthermore, I also obtained the population and population density data of 50 States from “Census Bureau” by using API, and combined the two data sets together.

**Data Cleaning and Wrangling:**

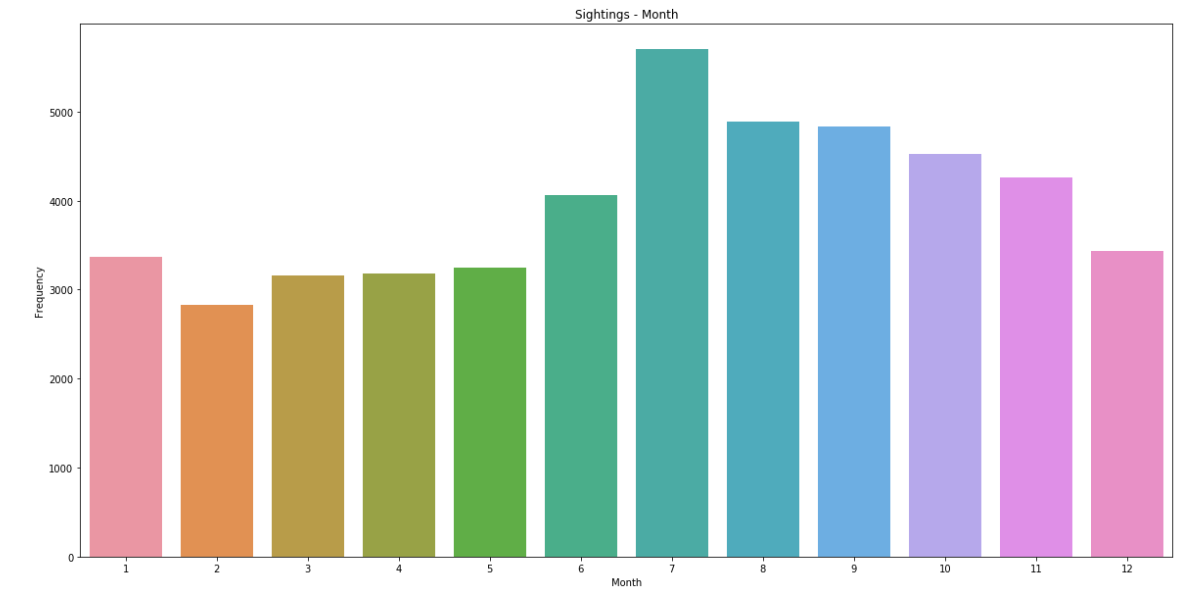
For data cleaning and wrangling, I mainly used “pandas” module in Python. First I dropped the duplicate columns of the data. Next, I checked how many null values of every columns in the data, and I dropped the columns that contain 80% or above null values. Also, in the “duration” column, the zero value rows are also null value so I dropped these rows also. Finally, I merged the population and population density data into the original data.

**Data Story and Inferential Statistics:**

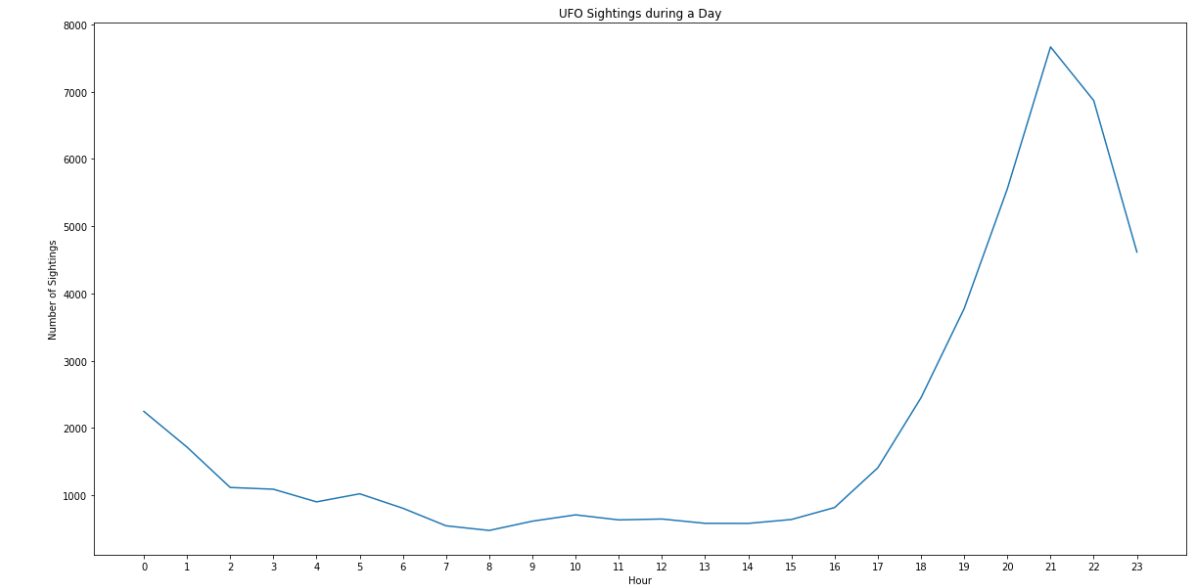
For data storytelling, I used “numpy”, “matplotlib”, and “seaborn” modules in Python. First I plotted the trend of UFO sighting every year. From the plot we can see the number of UFO sighting has an upward trend from 1997 till 2014. 2014 was the year recorded sighting the most UFOs. The number of UFO sighting started decreasing after 2014.



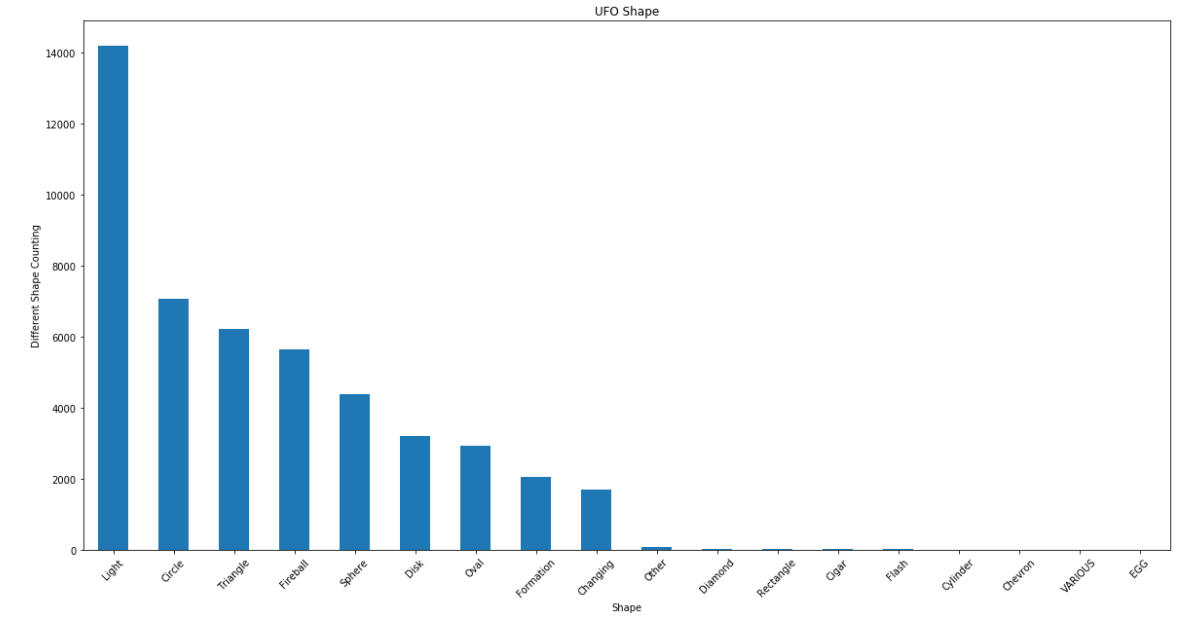
In a year period, summer and fall seasons have higher a frequency of UFO sighting than spring and winter.



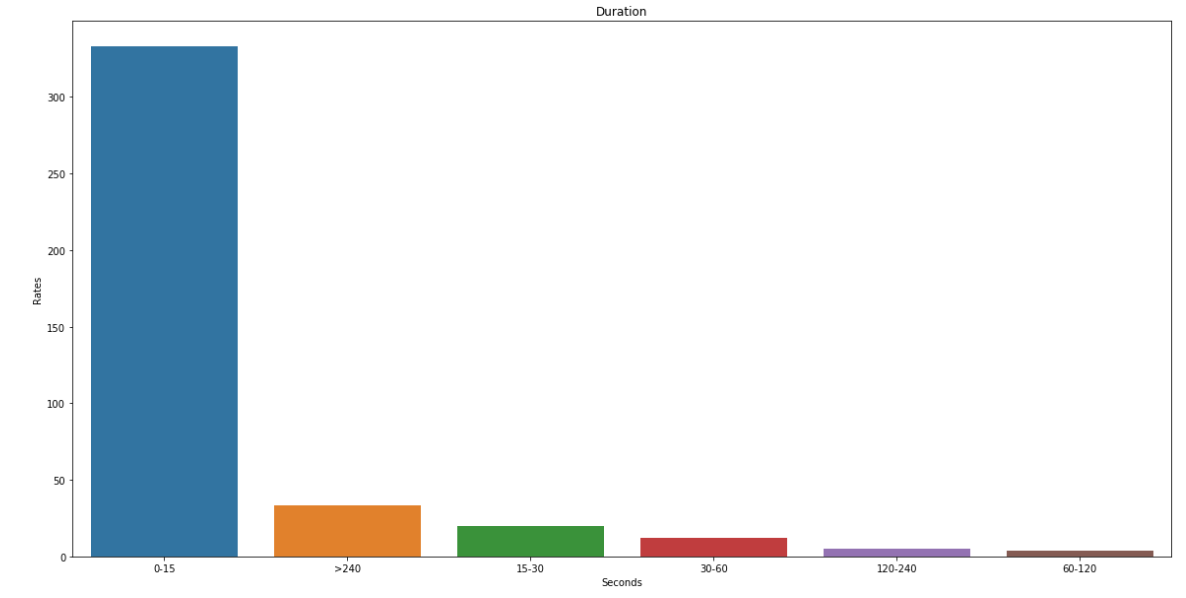
In a day period, UFO sightings mostly happened during night time from 6 p.m. to 12 a.m. The trend starts to drop down after 12 a.m. since people go to sleep.



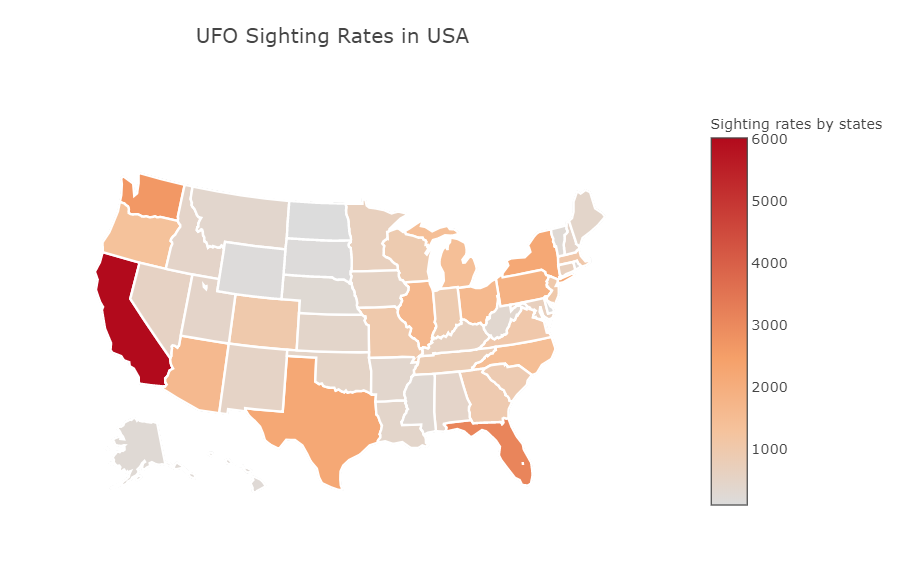
In about 14,000 of 51,871 reports, people said the shape of the UFO they saw was light. The second most common shape is circular, there are about 7000 reports said that they saw circular shape UFO.



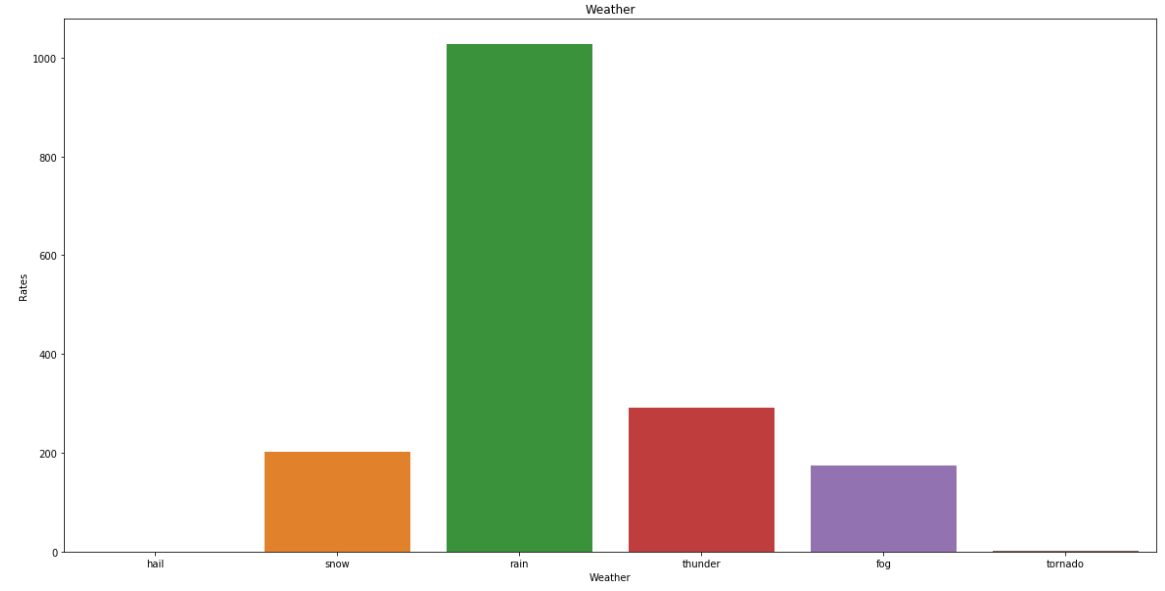
Next, I want to see how long did the UFO sighting usually last. Over 300 reports of UFO sighting only last zero to fifteen seconds. The second highest rates of UFO sighting last for over 240 seconds.



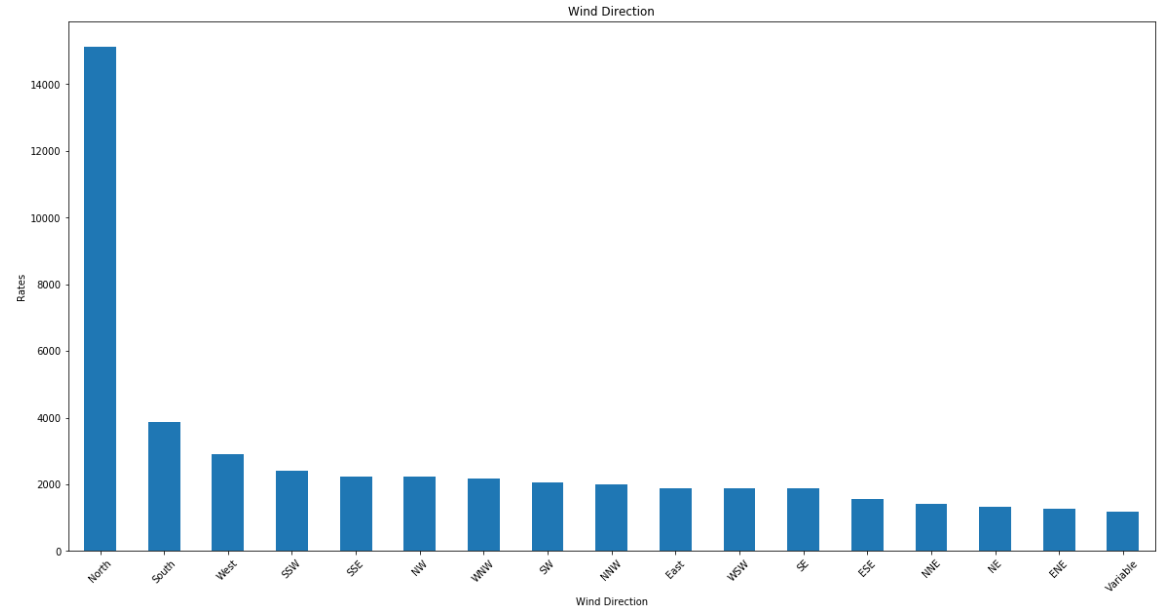
Next, I wanted to know which States has the highest chance of a UFO sighting. Over 6,000 of the UFO sighting happened in California, the next highest is in Florida, which there are about 4,000 reports.



Furthermore, other than sunny weather, UFO sighting mostly occurred during raining weather.

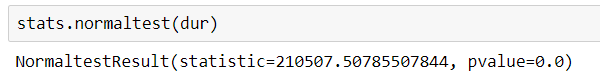


In addition, surprisingly, UFO sighting happened heavily during north wind weather.

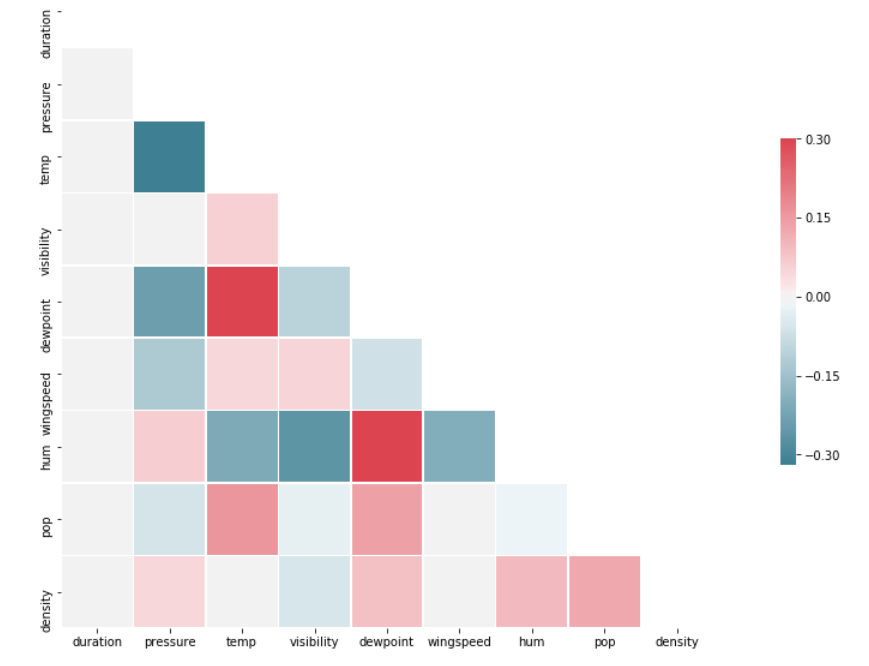


Inferential Statistics

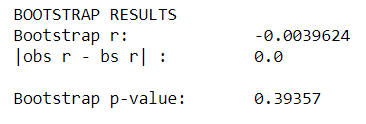
For inferential statistics part, first, I test if the duration of UFO sighting is normally distributed, since the p-value is equal to 0 which we can reject the null hypothesis and conclude that the distribution of duration is not normal.



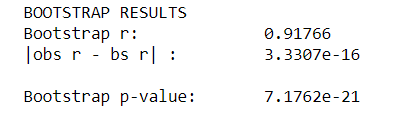
Next, I want to find out are there strong correlations between the dependent variable, duration of UFO sighting, and the independent variables, weather and population variables.. I had done a correlation heat map first to check out the correlation. From the map we can see that there is no correlation relationship between duration and the weather and population variables.



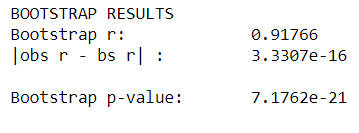
I tested one of the correlation relationship between duration and population using the bootstrap method, and since p-value is greater than 0.05, we cannot reject the null hypothesis and conclude that there is no correlation between populations and the sighting duration. Therefore, the test tells the same result as the map.



Next, I also tested if there is a correlation relationship between sighting rates and population/population density. I used the bootstrap method on both test. For the correlation relationship between sighting rates and population test, since the p-value is less than 0.05, we can reject the null hypothesis and conclude that there is correlation between populations and the sighting rates.



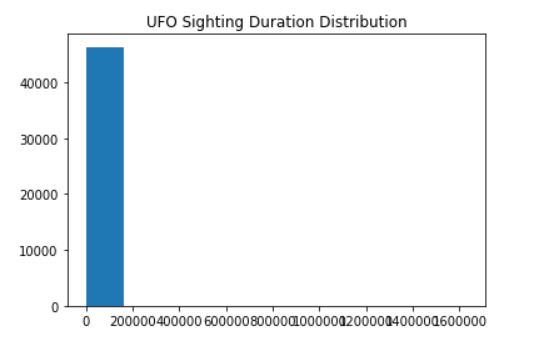
For the correlation relationship between sighting rates and population density test, since the p-value is greater than 0.05, we cannot reject the null hypothesis and conclude that there is no correlation between populations with density and the sighting rates.



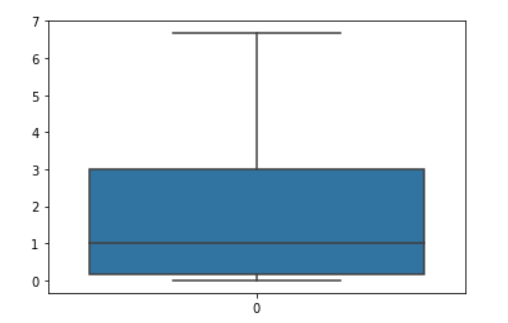
**Machine Learning:**

The goal of this project is building a predictive model to predict the month of UFO appearance base on weather, geographical and population information.

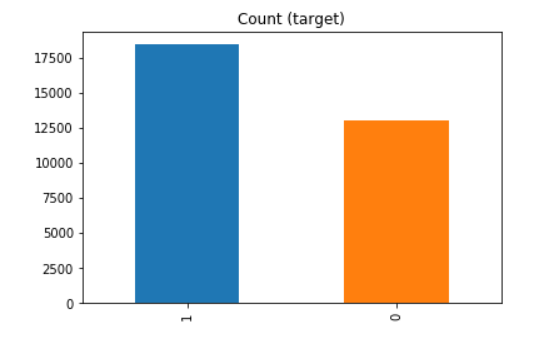
First, since the duration is recorded in seconds, I changed to minute. In the inferential statistics part, we know that the duration of the UFO sighting is not normally distributed. Looking at the distributed below, we can see that the distributed is heavily toward zero. As a result, I had to normalized the duration of UFO sighting so that the distribution makes more sense and hopefully it has a more normally distribution.



I cut off the duration of UFO sighting in the range between greater than 0 and less than 7 and it left 31418 rows of data. Looking at the box plot, it does not have outliers anymore.



Since I want do a prediction on the moth of the UFO appearance, I built a binary classification model base on UFO shape, State, hour, day, year, duration, pressure, temperature, visibility, dew point, wind speed, wind direction, weather condition, humanity, population, and population density. If the UFOs appear on January to June, it will be classified to be 1, if the UFOs appear on July to December, it will be classified to be 0. Looking at the plot below, we can see that class 1 has 18403 rows of data, and class 0 has 13015 rows of data. Two classes are close to balance which the model will have a better prediction result.



Algorithms Selection

I split the data into 70% of the data for training the model, and 30% of the data for testing the model performance. Then, I applied logistic regression, K-nearest neighbors, support vector machine, naive bayes, Adaboost, XGboost, and random forest classifier to build the classification model. We look at both the accuracy score and the AUC score for all the algorithms and find the best one of them.

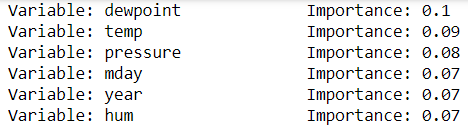
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | KNN | SVM | Naïve Bayes | AdaBoost | XGBoost | Random Forest |
| Accuracy | **0.5856** | **0.6005** | **0.6366** | **0.5856** | **0.6367** | **0.6346** | **0.7028** |
| AUC | **0.5157** | **0.6048** | **0.6770** | **0.5175** | **0.6703** | **0.6784** | **0.7620** |

Looking at chart above, we can see that random forest performed the best. However, I used the model to do prediction on the train and test data. The accuracy shows that the random forest is over fitted since the training data accuracy score is much higher than the test data accuracy score. This might due to I did not set the depth of the decision trees. I had to do features selection and parameter tuning to try to get the more accuracy model.

|  |  |
| --- | --- |
|  | Accuracy |
| Training | 1.0000 |
| Test | 0.7026 |

Features Selection

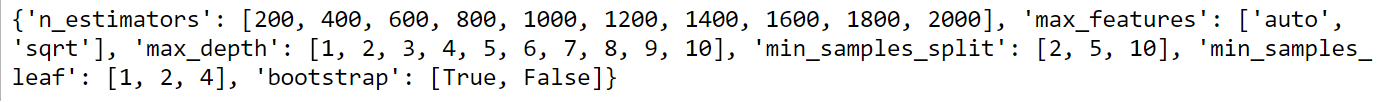
**Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of model.** Irrelevant or partially relevant features can negatively impact model performance. Therefore I tried to find the features are significant to the model. Looking at the features importance, I picked the first six most important variables to train the classification model using random forest again. The six most important variables are dew point, temperature, pressure, humanity, day of the month, and year. After features selections, although the accuracy and AUC score got lower, but it has better ability to do prediction on new data.



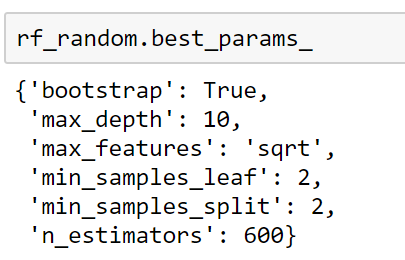
|  |  |  |
| --- | --- | --- |
|  | Random Forest(After Features Selection) | Random Forest (Original) |
| Accuracy | **0.6579** | **0.7028** |
| AUC | **0.7027** | **0.7620** |

Hyperparameters Tuning

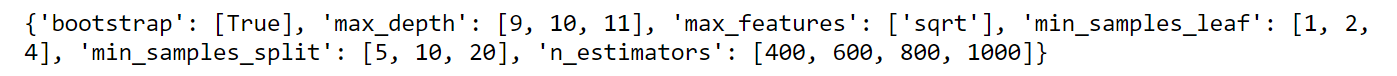
Since random forest tends to over fit the model if we did not set up the depth of the decision tree. We need to do parameter tuning. Usually, we only have a vague idea of the best hyperparameters and thus the best approach to narrow our search is to evaluate a wide range of values for each hyperparameter. Using Scikit-Learn’s RandomizedSearchCV method, we can define a grid of hyperparameter ranges, and randomly sample from the grid, performing K-Fold CV with each combination of values. I used below parameters to do random search to fit the model and tried to find the best result of among them. There are total 500 of combinations.



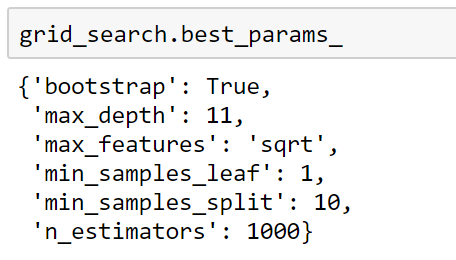
I got the best parameters from fitting the random search is:



Random search allowed us to narrow down the range for each hyperparameter. Now that we know where to concentrate our search, we can explicitly specify every combination of settings to try. We do this with GridSearchCV, a method that, instead of sampling randomly from a distribution, evaluates all combinations we define. I used below parameters to do grid search to fit the model and tried to find the best result of among them. There are total 540 of combinations.



I got the best parameters from fitting the grid search is:



The accuracy score and AUC score did not change much after parameter tuning.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest (After Features Selection and Parameter Tuning) | Random Forest (After Features Selection) | Random Forest (Original) |
| Accuracy | **0.6616** | **0.6579** | **0.7028** |
| AUC | **0.7009** | **0.7027** | **0.7620** |

I tried to use the model to do prediction on the train and test data again. Although we got lower accuracy and AUC score after features selection and parameter tuning. However, we can see that the accuracy and AUC score of doing prediction on the train and test data got closer which means the model is not over fitted the test data, and will have better prediction result on future new data.

|  |  |  |
| --- | --- | --- |
|  | Accuracy (After Features Selection and Parameter Tuning) | Accuracy(Before Features Selection and Parameter Tuning) |
| Training | 0.7119 | 1.0000 |
| Test | 0.6547 | 0.7026 |

**Conclusion:**

By looking at the confusion matrix of the model, we can see that the model does not perform well for the true negative. It has low recall score for the 0 class. One reason it is because we have less class 0 data. Adding more class 0 data may actually help the model to have a better prediction result. This model has better prediction result for UFO appearance on July to December.

