Exploratory Data Analysis and Predicting the Existence of Convective Weather on NOAA Hourly Data

Tianhao Wu, Jaewoong Lee

May 6, 2022

CivEng 290-2: Data Science in Aviation

Professor: Dr. Banavar Sridhar

GSI: Ke Liu

Table of Contents

1. Introduction

Convective weather is the result of surface convection on the Earth that transports the excess heat energy from the surface of the earth to the upper atmosphere. When the surface convection is combined with other atmospheric phenomena such as humid air and certain thermal conditions, convective weather such as storms, tornadoes, and turbulence are developed that may cause delay or damage to aircraft. Correct prediction of convective weather can reduce risks and unnecessary operations in the aviation industry and minimize the operation cost.

In this project, massive weather condition data sets are obtained from National Oceanic and Atmospheric Administration (NOAA). The data are collected from weather observation stations, and mainly we will investigate the locations at airports considering flight planning purposes as well. The data have been collected since 1960, and after a careful investigation of the data sets, a decision was made to cover only the recent 20 years of data due to the inconsistency in the data set before 2000. To obtain strong insight into the data sets, the data are processed first and then visualized thoroughly from various aspects. The data are visualized to see the monthly and daily trends over 60 years for each type of convective weather, then statistical analysis is conducted corresponding to each convective weather type. After obtaining a good understanding of the data, various machine learning techniques are used to train models and predict each type of convective weather.

2. Literature Review

According to the National Transportation Safety Board (NTSB) [1], about 70 percent of flight delays and 23 percent of all aviation accidents are caused by weather, and the total national cost estimated caused by weather is about $3 billion per year. The main convective weather can be summarized as thunderstorms with severe turbulence, intense up and down drafts, lightning, hail, heavy precipitations, icing, wind shear, microbursts, strong low-level winds, and tornados.

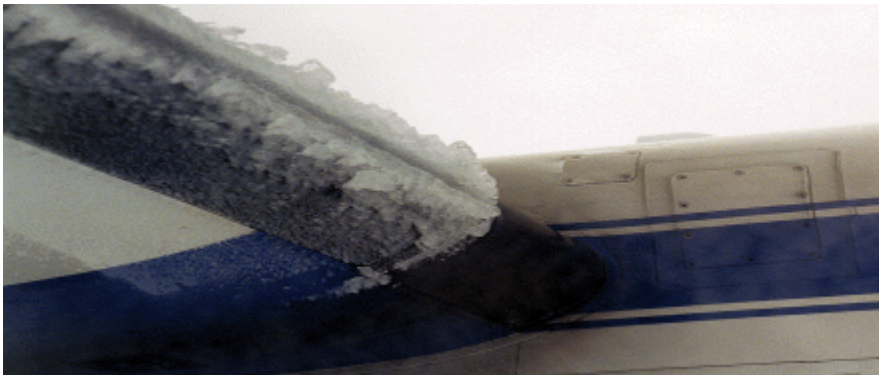

Figure 2.1. Structural icing on an aircraft's wing

Figure 2.2. A missing engine torn off by turbulence

Various studies and efforts have been made to avoid these unwanted convective weather issues. However, forecasting convective weather can be a challenging task. According to Richard [2], assessing the potential convective weather is one of the daily routines at the National Weather Service (NWS). A variety of data such as observation data, outputs from weather models, and satellite images are combined together to predict convective weather. However, perfect prediction of convective weather is still impossible with the current observation and forecasting concepts. New technologies such as Doppler radar, wind profiler, gridded model based data sets, and lightning detection systems can enhance the NWS' forecasting system to analyze the current atmosphere and predict convective weather more accurately.

Efforts are made to analyze the accuracy of convective weather forecasting at the center and sector levels by Yao Wang and Banavar Sridhar [3]. Convective Weather Avoidance Model (CWAM) is used to generate forecast data by combining data from weather radars, satellites, surface observations, and mathematical models. Then, the forecast data are statistically analyzed for accuracy assessment.

A deep learning approach was taken by Kanghui et al. to forecast different types of convective weather in China. [4] A Convolutional Neural Network (CNN) was trained with numerical and observation data, and the results are compared to Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MP). For simplicity, heavy rain, thunderstorm, hail, and convective gust were tested. 144 observations were taken over a (*L x L)* grid patch, so the input dimension was *L x L x 144*. Then, each grid is normalized to (*M x 7 x 7 x 144*). Data arrays were labeled by either 0 or 1, denoting the corresponding event NOT occurs or occurs, respectively. They successfully predicted convective weather up to 6 hours ahead. The average probability of success for forecasting thunderstorms and hail by CNN was 0.336.

From the literature review, it is obvious that convective weather forecasting is a very cumbersome task, and it is almost impossible to perfectly predict. Therefore, convective weather forecasting should be an ensemble of tools such as satellite, radar, visual observation, data, and mathematical models. Especially, for the correct prediction, real-time observations from the target location should be included in the model as well. In this project, fully data-driven Machine Learning techniques will be adopted to predict convective weather at certain locations, especially airports.

3. Project Problem Statement

In this project, there are two studies related to convective weather prediction.

For the first part, a thorough Exploratory Data Analysis (EDA) will be performed to better understand the data set we have. We would be able to learn data handling skills and intuition about how to interpret the data.

After understanding the data, a variety of machine learning techniques will be used to build convective weather predictors. The convective weather predictors will be based on fully the previous NOAA data, so it is not related to any real-time data.

4. Data

We are exploring “Global Hourly dataset” provided by NOAA. The database includes weather data observed through sensors located worldwide from the year 1901 to the present year 2022.

Data Format:

The data is recorded following the guidelines of Federal Climate Complex Data Documentation for Integrated Surface Data (ISD). The dataset is recorded in 3 parts:

1. Control data section (station information)
2. Mandatory data section (key weather information)
3. Additional data section

Each station generates an annual weather report as comma-separated values. The report records weather data at hourly intervals. A sample weather data looks as follows:

	STATION	DATE	SOURCE	LATITUDE	LONGITUDE	ELEVATION	NAME	REPORT_TYPE	CALL_SIGN	QUALITY_CONTROL	...	OC1	OD1
0	72494023234	2020-01-01T00:00:00	4	37.6197	-122.3647	2.4	SAN FRANCISCO INTERNATIONAL AIRPORT, CA US	FM-12	99999	V020	...	NaN	NaN
1	72494023234	2020-01-01T00:56:00	7	37.6197	-122.3647	2.4	SAN FRANCISCO INTERNATIONAL AIRPORT, CA US	FM-15	KSFO	V030	...	NaN	NaN
2	72494023234	2020-01-01T01:56:00	7	37.6197	-122.3647	2.4	SAN FRANCISCO INTERNATIONAL AIRPORT, CA US	FM-15	KSFO	V030	...	NaN	NaN
3	72494023234	2020-01-01T02:56:00	7	37.6197	-122.3647	2.4	SAN FRANCISCO INTERNATIONAL AIRPORT, CA US	FM-15	KSFO	V030	...	NaN	NaN
4	72494023234	2020-01-01T03:56:00	7	37.6197	-122.3647	2.4	SAN FRANCISCO INTERNATIONAL AIRPORT, CA US	FM-15	KSFO	V030	...	NaN	NaN

Table 4.1: Sample NOAA global hourly data

The number of stations varies as time progresses and more stations are built and put in use. Figure 4.1 is the geographical visualization of 12,981 weather stations actively reporting

weather information in the year 2022. Weather stations are clustered spatially on coasts, the United States and Europe.

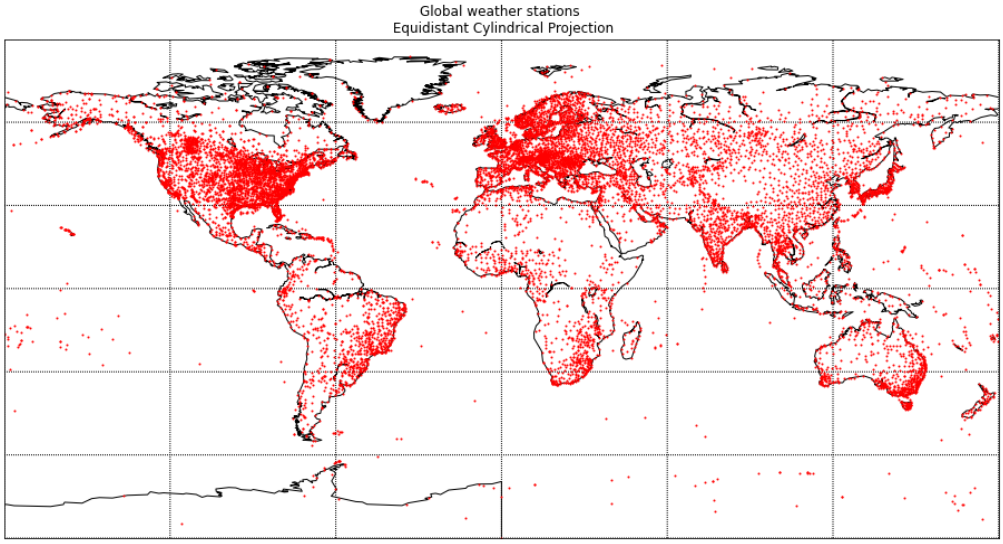

Figure 4.1: Weather stations worldwide

Figure 4.2 is the geographical visualization of 2,824 stations located in the United States. To provide a rough overview of the number of stations in each state, there are 173 stations in California and 217 stations in Texas.

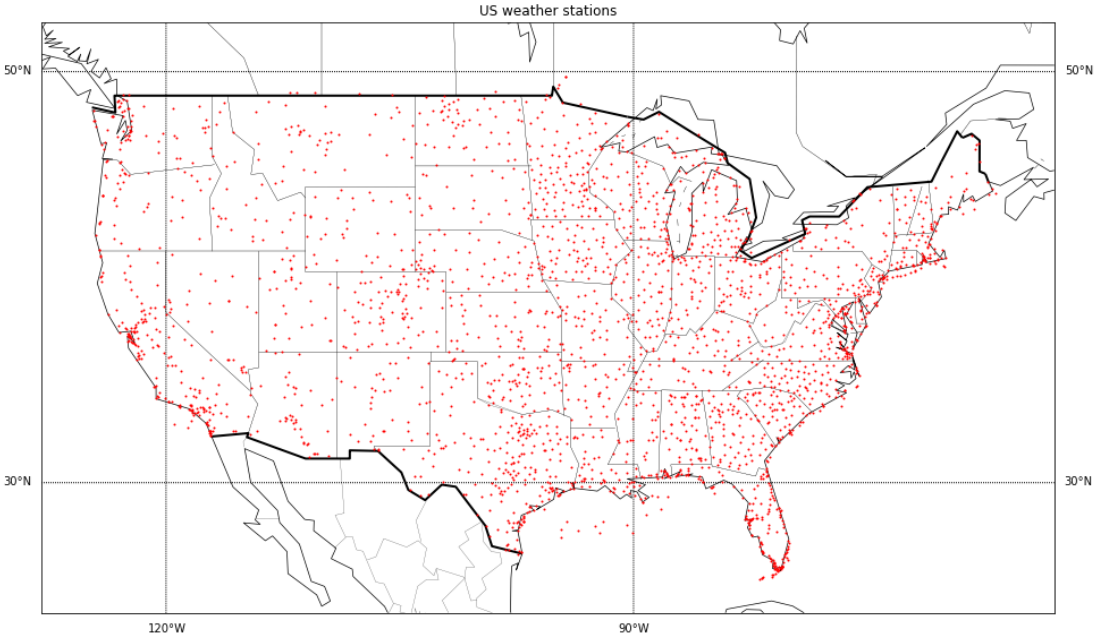

Figure 4.2: Weather stations in the United States

Data Selection:

Through our preliminary data inspection, we found that data in the additional data section is sparse and incomplete. Therefore, we decide to primarily utilize the control data section and mandatory data section for our data exploration Additionally, we will include “AT” columns in the additional data section, which correspond to weather occurrence data and serves as a classification target.

Regarding the timeframe, we will use worldwide stations’ weather data that occurred in the years 1980~2022 for our exploratory data analysis. For weather classifications, due to limited computing power and a huge dataset, we will use Houston Intercontinental Airport’s weather report from the year 2000 to the present.

Data Preprocessing:

The comma-separated values provided by NOAA are sparse and missing frequently. During our preprocessing, we extract each separated column, parsing values in it according to ISD documentation. For missing values, since the report is recorded in time series, we fill the missing values using previous neighbors’ data at every feature’s timesteps.

Additionally, we made an important assumption during data preprocessing. Since weather occurrence, “AT”, is recorded once per day, we generalize this daily feature into an hourly feature. Meaning if bad weather is recorded one day, all hourly data on that specific day will be marked as the same bad weather.

Data Features:

We manually selected features we think are useful, which include:

1. Date [year-month-day-hour-minites]
2. Geographic coordinates of the stations: latitude and longitude
3. Wind speed [meters/seconds] and wind angle [angular degree]
4. Ceiling height [meters]
5. Visibility distance [meters]
6. Dew point temperature [celsius]
7. Sea level pressure [hectopascal]
8. Observed weather occurrence in binary values (0,1)

5. Exploratory Data Analysis

5.1 Weathers temporal visualizations

5.1.1 Overal Insight into the Data Set

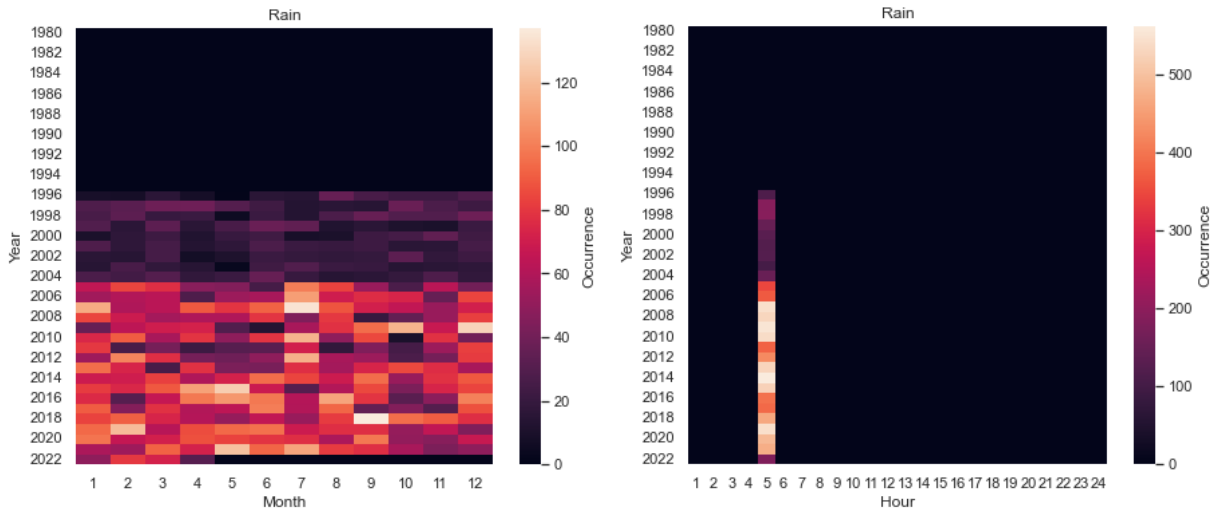

Figure 5.1.1: Heatmap of Rain from 1980 to 2022

Due to the huge volume of data, we select the Houston, TX, area only for visualization and data analysis. Two heatmaps are represented in Figure 5.1.1. The heatmaps are the summary of the data set rearranged monthly and hourly, respectively. As shown in the figures, there is no weather data recorded from any of the observation locations until 1995. Additionally, there are fewer data recorded between 1996 to 2004. This might imply that more weather observatories became online and started recording more data since 2005. So, the data set will be curtailed to include only from 2000 to 2022 for the rest of the work. Additionally, in the second figure, there are recordings only at 5 am. which means the convective weather observations are made only once a day in the morning at 5 am. Therefore, the hourly analysis will be omitted hereafter. The plotting idea is obtained from NOAA NWS. [6]

5.1.2 Monthly Analysis

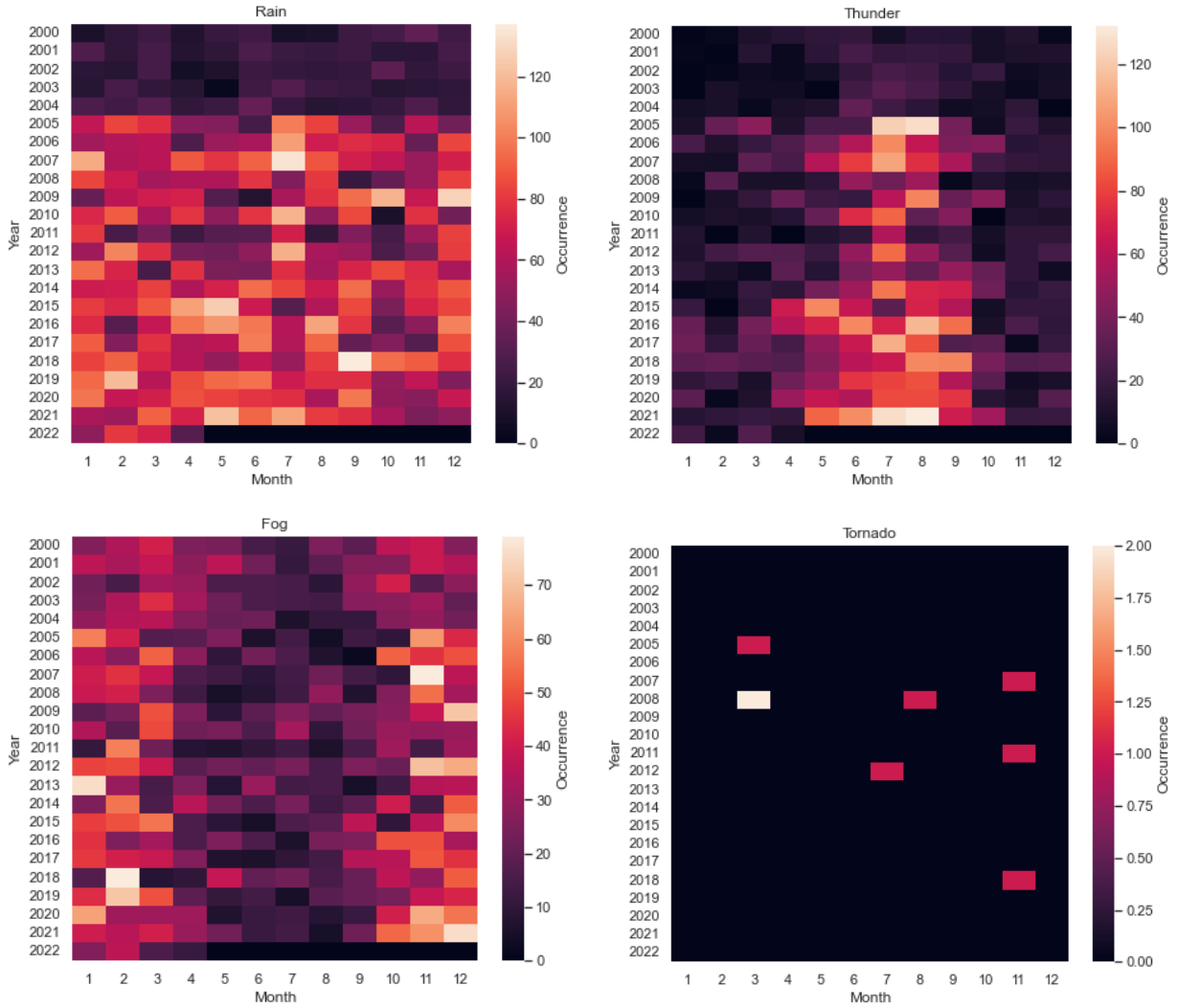

Figure 5.1.2: Monthly Trend of Selected Convective Weather

Monthly trends of Rain, Thunder, Fog, and Tornado are shown in Figure 5.1.2. There are still fewer data between 2000 and 2004. However, in the fog figure, the fewer data might not look clear because the observation of fog is fewer in general compared to rain and thunder. There is no clear monthly trend in the rain heatmap. However, there are slightly more observations between May and September. Thunder and fog show very clear monthly trends over the entire observation years. More thunders are observed between April and September, and even more during the summer season. Contrary to thunder, fog is more dominant during the winter season. Fog is condensed vapor to fine droplets of water suspended in the air and it is generally formed when the temperature becomes low and hit the dew point. Temperature is low during the winter season, so it makes sense that more fog is observed during winter. The occurrence of tornados is very low compared to other convective weather. Only a few numbers of tornadoes are observed in the Houston area. Also, there is no clear trend in tornadoes. There are no weather data after the April of 2022 because this project is conducted in April of 2022.

5.1.3 Daily Analysis

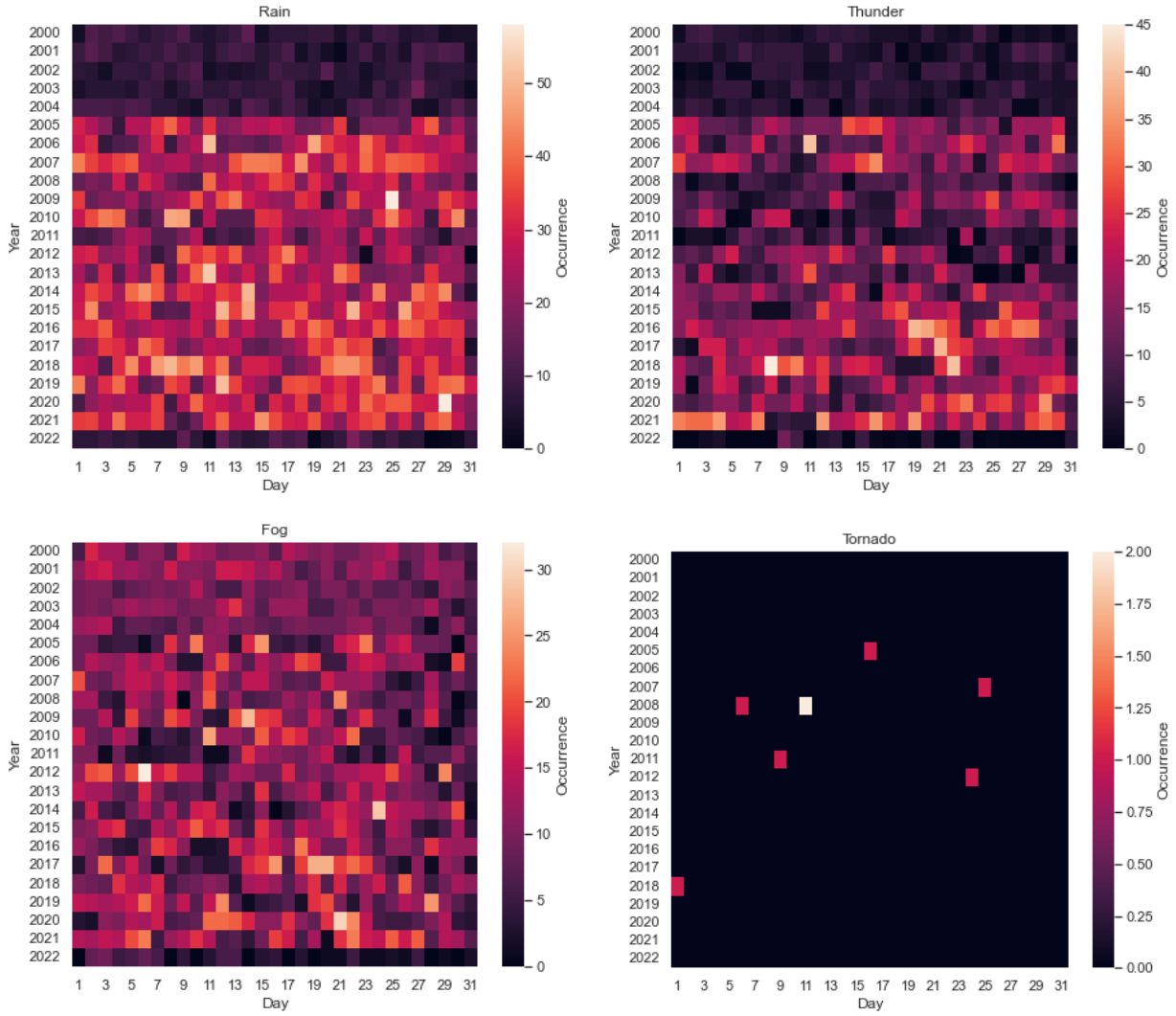

Figure 5.1.3: Daily Trend of Selected Convective Weather

Again, there are fewer data recordings between 2000 and 2004. No clear daily trends are detected in all of the selected convective weathers. However, now the x-axis is dates, not months. So, now we can see the daily trends in 2022 because even though the weather is recorded until April in 2022, the daily trend between January and April is shown in 2022 row. However, since it's only weather recordings between January and April, the total count of convective weather detection is low. So the color of 2022 row is darker compared to other years.

5.1.4 Statistics

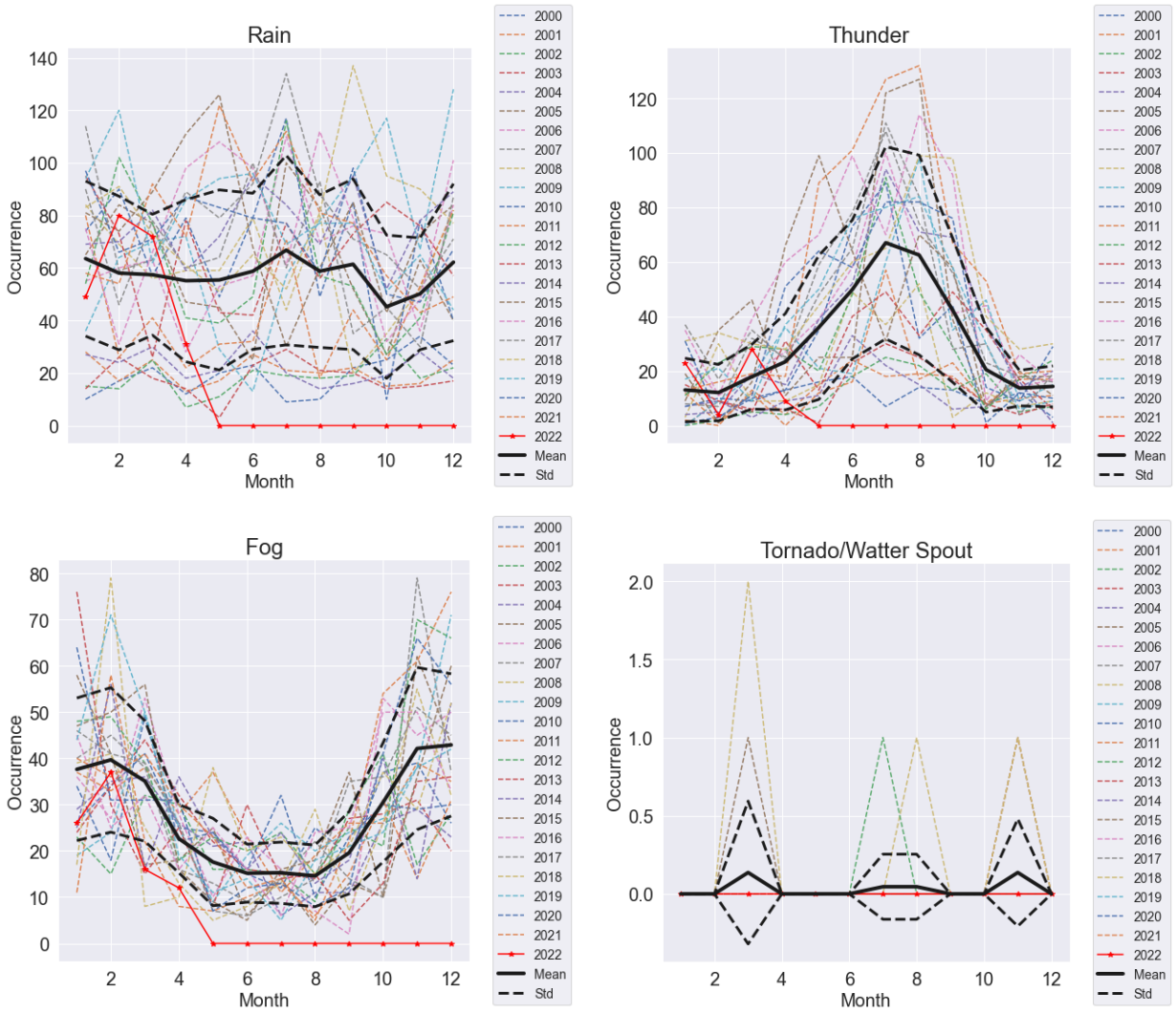

Figure 5.1.4: Statistics of Selected Convective Weather

The same data set is investigated in a different way. The monthly trend of each convective weather is plotted to see the annual statistics. In all figures in Figure 5.1.4, each line corresponds to each year from 2000 to 2022. As mentioned with Figure 5.1.2, there is no clear monthly trend in Rain, so the average has no steep change. Also, the variance is sort of uniform over all of the months. Thunder has a higher average during summer in a similar way Figure 5.1.2 with higher variation during summer and smaller variation during winter. Also, fog shows higher averages and variations during winter and smaller averages and variations during summer. Tornado has the same average in March and November, but the variance is higher in March. It means that there were three observations of tornados in total in both March and November. However, in March, there were two occurrences of tornado in March in 2008 and one observation in 2005 resulting higher variance in March. In September, there were three tornadoes observations in total, but only one observation per year resulting lower variance in November.

5.2 Weathers spatial visualizations

The following plots figure 5.2.1 are bubble plots regarding the number of days weather occurred in the United States in the year 2021.

As rains have more impact locally, the bubble is set to be a certain radius across all stations and uses color gradients to represent the frequency of the weather. Darker blue represents more number of days occurring the weather, and vice versa. Rains occurred more in coastal areas and easts of the United States.

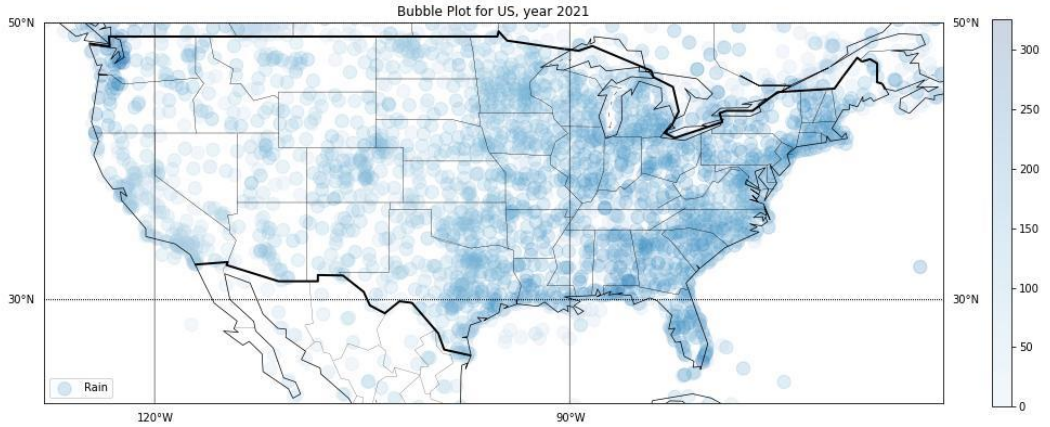

Figure 5.2.1: Bubble Plots for the US, Rainy days

Bubble plot figure 5.2.2 represents the frequency of the thunder. Where smaller bubble represents less frequent thunder and vice versa. From the plot, we can tell that thunder is frequently occurring in the southeast of the US.

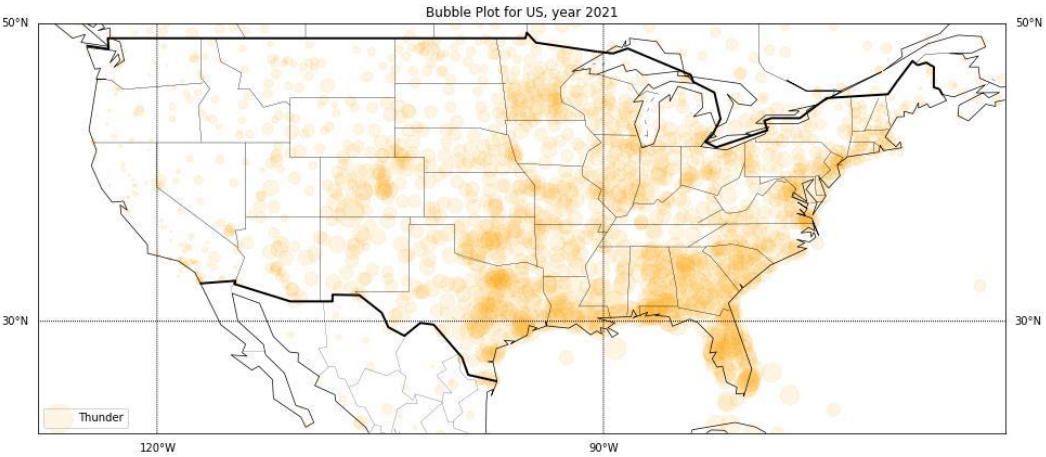

Figure 5.2.2: Bubble Plots for the US, Thunder days

Following bubble plot figure 5.2.3 zooms in Houston area, and plotting frequencies of weathers respect to the size of the bubble for comparability. The rains occur more frequently than thunder, and tornado rarely occurs even in the area damaged frequently, which is telling us that tornado is a rare event, and probably difficult to predict due to limited sample sizes. Therefore, we decide our sole task is to classify if thunders will occur on a day.

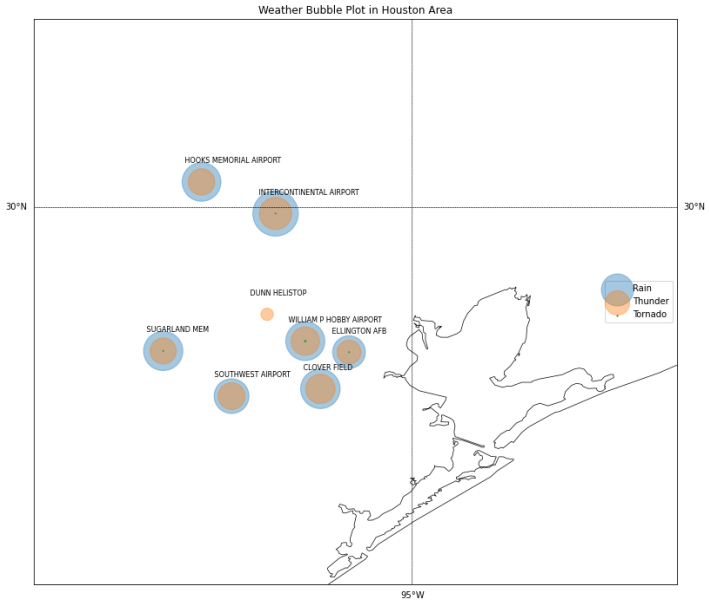

Figure 5.2.3: Bubble Plots for Houston Area, bad weather

6. Methodology

Binary Classification Task

Our task is a binary classification task on thunder occurring on a day given an hourly weather observation. We will train the model with hourly weather observation, and the model could predict whether or not a thunder occurs on that day. We exclude “DATE_year” data since we assume every year has almost the same weather patterns, this helps us reduce the features.

Classifiers Models

We use some classical machine learning classifiers models, which include:

1. Logistic regression
2. Decision Tree
3. Linear Discriminant Analysis
4. Quadratic Discriminant Analysis
5. Random Forest
6. K-Nearest Neighbors
7. Naive Bayes

We also include ensemble methods including the above classifiers with voting techniques of:

1. Hard voting
2. Soft voting

We also include the majority class as a baseline for the model. For our data, non-thunder weathers are the majority in the training dataset. Therefore, the majority class baseline model will classify any observations into non-thunder weather.

We perform fine-tuning on classifiers using grid search cross-validation methods in order to find the best hyperparameters for classifiers.

Data Splitting

We performed 70 - 15 - 15 proportion on dataset splits for training - development - test respectively. Since we include time data in our features, we treat the splits as a time series problem, where random shuffling is not performed when splitting training-development and test datasets. In another word, the test set split out later and recent 15% of total data. After that, the training and development set is split randomly since it does not affect the results of model testing.

Dealing Imbalanced Data

Figure 6.1 shows the histogram of 3 bad weather occurred in the years 2000-2022. The y-axis of the histogram represents the number of weather occurred on an hourly basis, based on our assumption stated in the data section. The plot indicates that there’s an extreme imbalance between weather occurring and weather not occurring, where bad weather is not

observed in the majority of hours. Therefore, we should treat the imbalanced data using techniques such as oversampling and undersampling.

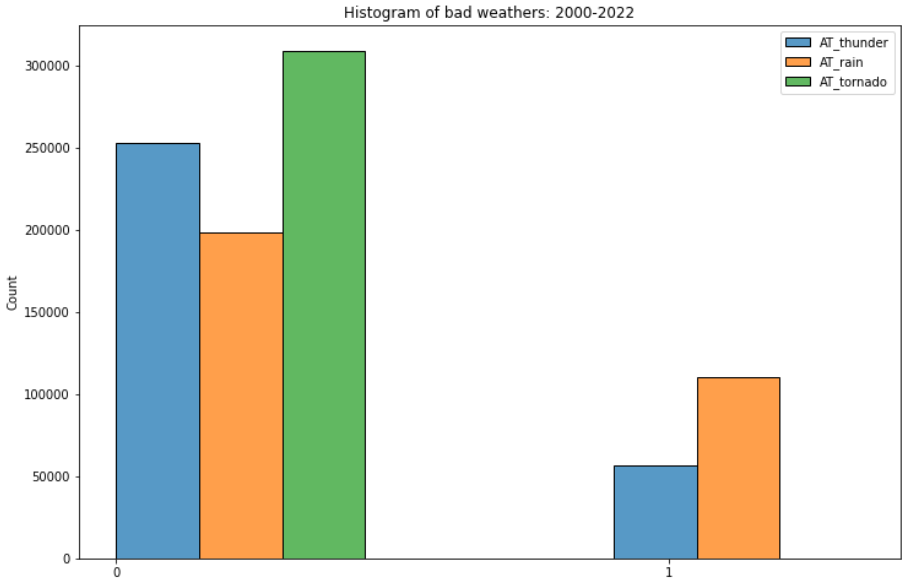

Figure 6.1: Histogram of bad weather

Here we utilize the technique of oversampling, where minority class (1) is oversampled. The histogram after oversampling is shown below, figure 6.2:

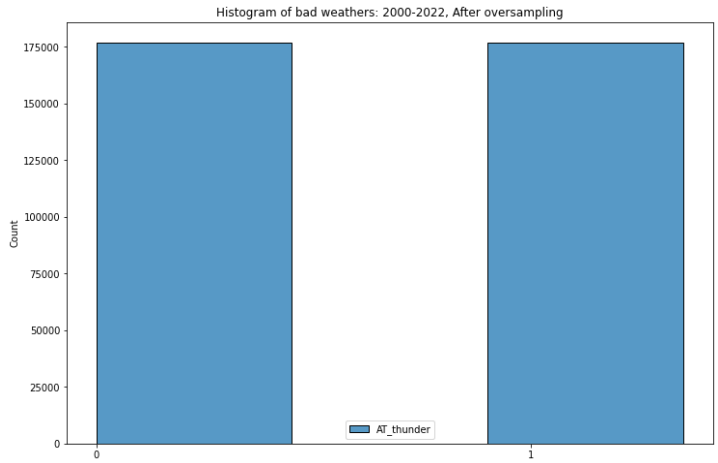

Figure 6.2: Oversampling thunder observation

We perform classification with and without using oversampling data and compare the metrics in the results section.

7. Results and Discussion

In this results section, we show classification results on the training set and test set, with and without oversampling, before and after fine tunings. We will provide visualized results, including tables, bar plots, and some confusion matrix between prediction values and true values.
We use several metrics to evaluate the performance of the training, the metrics include accuracy, precision, recall, F1-score, and AUC-ROC.

Default tuning without oversampling

Training Results:

Figure 7.1.1 is the bar plot comparing model results for 7 classifier models with default hyperparameters and without oversampling.

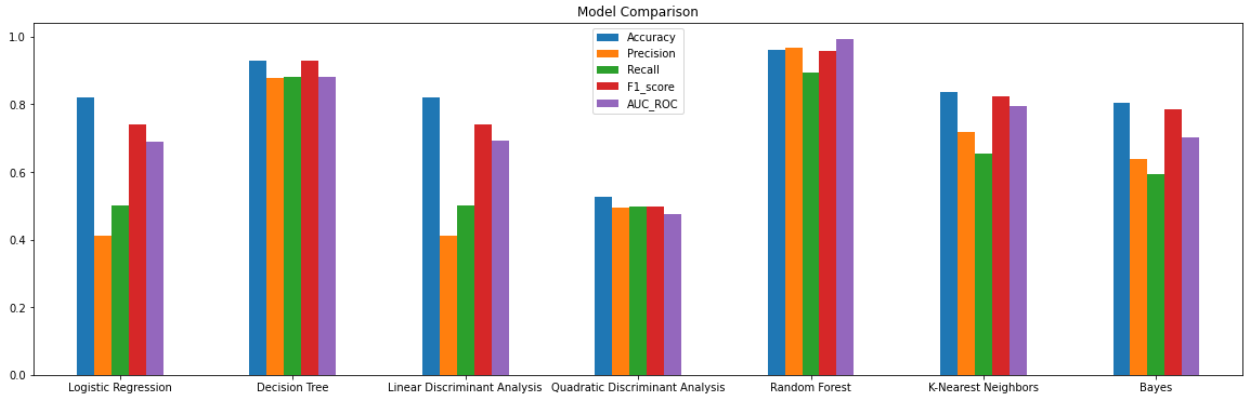

Figure 7.1.1: Model comparison for the training dataset, default tuning without oversampling

The training results indicate that the decision tree and the random forest is outperforming other classifiers. Random forest takes the longest time to compute but achieves the best accuracy score. Quadratic discriminant analysis is suffering from low scores on the training dataset.

The following figure 7.1.2 is the results of ensemble methods with hard and soft votings, using all classifiers above. The results of the ensemble method are not optimal, as they have low recall and F1-score, even with moderate accuracy.

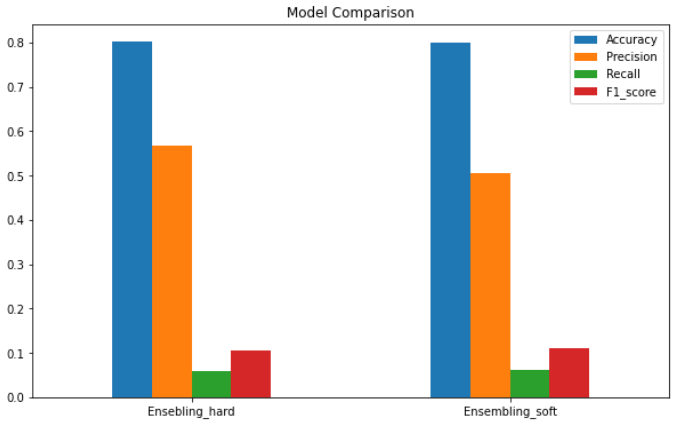

Figure 7.1.2: Model comparison for the training dataset, ensemble learning

We think the main reason for the poor performance could be interpreted from the low recall score. The classifier is voting every observation to the majority class and is very picky about voting observations to the non-majority class. Therefore, they achieve ~80% accuracy on training, which is the percentage of the majority class, and have low recall scores. Following is the confusion matrix of the logistic regression classifier on the test dataset. We find that the classifier is classifying all cases into the majority class, which verifies the observation.

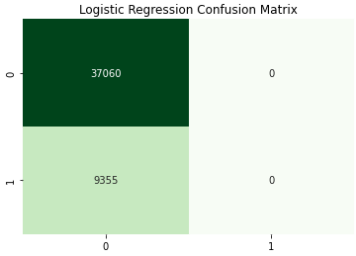

Figure 7.1.3: Confusion Matrix of Logistic Regression for test prediction and true label

Testing Results:

Figure 7.1.4 is the model comparison of accuracy for different classifiers. The furthest left is the baseline model, the majority class classifier. From the plot and the accuracies calculated, we found that the only classifier outperforming the baseline (0.798 test accuracy) is the random forest classifier (0.804 test accuracy).

Figure 7.1.4: Model comparison of test data set, default tuning without oversampling

Default tuning with oversampling

In this subsection, we will show the training and testing results of the oversampling training dataset to fix the imbalance dataset. As introduced previously in the methods section, we oversample the minority class, creating a larger training dataset with more balanced labels.

Training Results:

Following figure 7.2.1 is a bar plot of the metric results for different models. From the plot, we can tell that again decision tree and the random forest is outperforming other classifiers. However, results with such high accuracy often mean the model is overfitting the training dataset.

Figure 7.2.1: Model comparison for the training dataset, default tuning with oversampling

Testing Results:

Following plot, figure 7.2.2, is the accuracy comparison between 7 classifiers as well as the majority class baseline. The results trend is similar to the testing without oversampling, and for some classifiers, the accuracy is even worse. Therefore, we decide not to use oversampling data for further tunings.

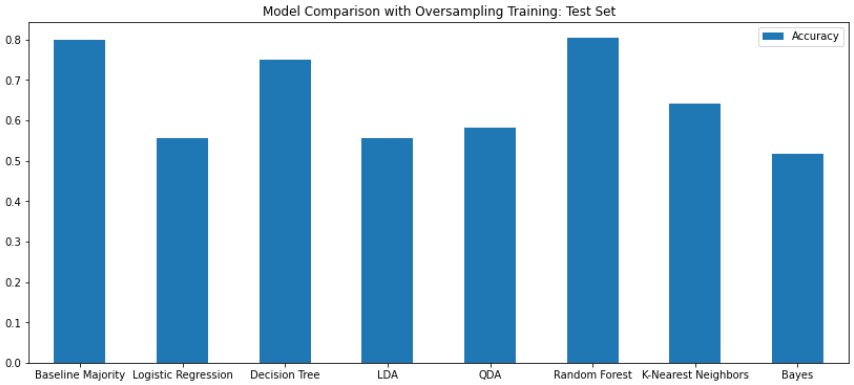

Figure 7.2.2: Accuracy for classifiers on the test dataset

However, with oversampling, prediction does make more sense by predicting both 0 and 1 classes, which are non-thunder and thunder day respectively. Figure 7.2.3 is the confusion matrix of the decision tree. The prediction is no longer predicting the majority class like one without oversampling.

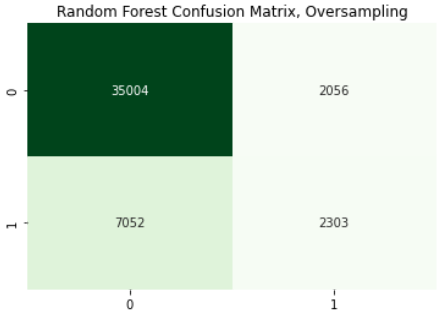

Figure 7.2.3: Confusion matrix for Random Forest Classifier

Fine-tuning random forest classifier

Since the random forest classifier is performing the best of all classifiers and beating the majority class baseline in both cases, we decide to fine-tune the random forest classifier with different hyperparameters.

We use a grid search method with cross-validation folds equal to 3 to train the random forest model. The hyperparameter has a max depth ranging from 2 to 8, minimum samples leaf ranging from 1 to 6. Following figures 7.3.1 and 7.3.2 are testing accuracies with varying hyperparameters.

From the results, we can conclude that tuning minimum samples leaf does not improve the results much, and higher max depth of the random forests may increase the results a little. However, the margin of the increase is almost ignorable.

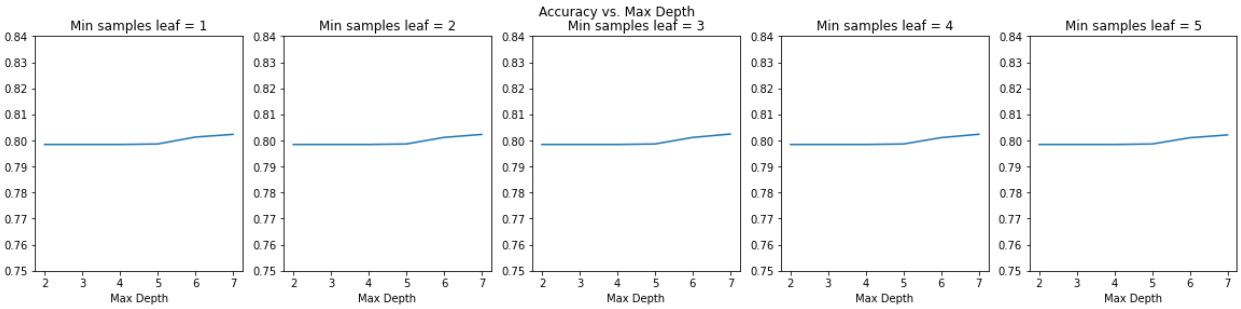

Figure 7.3.1: Accuracy vs. Max depth with different minimum samples leaf

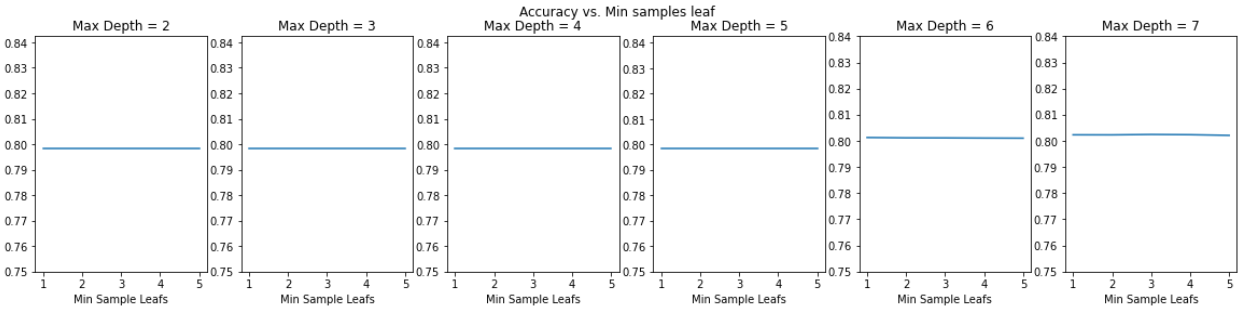

Figure 7.3.2: Accuracy vs. Minimum samples leaf with a different max depth

Discussion

From all classifiers we have experimented with, we found that the random forest classifier with a max depth of 6 generates the best results for the training and testing dataset. The model has 0.829 training accuracy and 0.802 test accuracy. However, the margin of improvement from the majority class baseline model (0.798 test accuracy) is too small and could be disregarded. We also noticed oversampling is not helping much in this task when training models. Generally, oversampling should not worsen the performance of models especially when we are dealing with imbalanced data with minority events.

We are not satisfied with the performance of the models we have. Theoretically, when we are provided with more information about weather, we suppose to have better predictability of the bad weather. However, it's not the case here. We suspect that the bad performance is caused by the data itself and the assumption we have made. Our assumption of generalizing daily observations into hourly observations could be flawed and misleading to model. Weather data with features we've selected also could be too few and too random to predict the weather. We should explore more data records provided by NOAA and not be limited to the ones we've extracted.

8. Conclusion

We have created a data preprocessing pipeline for bulk downloaded data files, where each file is processed individually and concatenated according to the parameter range we input, including time range and geographical coordinates. Based on the data we've processed, we performed several machine learning techniques, including classifiers and ensemble methods, with and without oversampling the training data.

From the modeling results, we find that a random forest classifier with a max depth of 6 without oversampling training data produces the best result. However, the margin of improvement from the majority baseline is so small that we conclude the classifier is not functioning well.

We suspect that the reason for poor modeling performance is due to the data itself. We may not include enough features or label the data incorrectly. For further exploration of the task, we suggest trying different data sources, for example, the NOAA "global summary of the day" dataset could be more suitable for the task of ours that is predicting daily bad weather.

For future improvement, a Convolutional Neural Network (CNN) approach may be used for modeling of a stronger classifier. For example, we may label each convective weather with a different label and train the model to define different convective weather, in a similar way to the famous MNIST hand-written number classifier. [5] Due to the enormous size of the data set, our laptops were not able to handle the CNN problem. High-performance PCs would be able to handle the long computation time and allow us to design a better prediction model. Also, it is clear that additional real-time or different types of data such as radar or satellite would be helpful to build more accurate prediction models like CWAM mentioned in the literature review.

9. Contributions

This project was carried out by Tianhao Wu and Jaewoong Lee. Data collection, data processing, weather spatial analysis, building and validating classification models, oversampling imbalanced dataset, comparing models: Tianhao Wu. Initial project conception, introduction, literature review, weather temporal analysis, data processing: Jaewoong Lee. Both authors contributed to the preparation of the final manuscript.

References

[1] Pace, D., “ATM-Weather Integration Plan Overview”, FAA AJP-B, Aviation Weather Office, May 27, 2009.

[2] McNulty, R. P. (1995). Severe and convective weather: A central region forecasting challenge. *Weather and Forecasting*, *10*(2), 187–202. https://doi.org/10.1175/1520-0434(1995)010<0187:sacwac>2.0.co;2

[3] Wang, Y., & Sridhar, B. (2010). Convective Weather Forecast Accuracy Analysis at center and Sector Levels. *29th Digital Avionics Systems Conference*. https://doi.org/10.1109/dasc.2010.5655494

[4] Zhou, K., Zheng, Y., Li, B., Dong, W., & Zhang, X. (2019). Forecasting different types of convective weather: A deep learning approach. *Journal of Meteorological Research*, *33*(5), 797–809. https://doi.org/10.1007/s13351-019-8162-6

[5] LeCun, Y., Cortes, C., & Burges, C. J. C. (n.d.). *The mnist database*. MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. Retrieved May 6, 2022, from http://yann.lecun.com/exdb/mnist/

[6] *Storm prediction center WCM page*. Storm Prediction Center. (n.d.). Retrieved May 6, 2022, from https://www.spc.noaa.gov/wcm/