Thunder Prediction

Bad weather classification

Tianhao Wu,
Jaewoong Lee

May 2, 2022
CivEng 290-002

Berkeley
UNIVERSITY OF CALIFORNIA

Outline

1. Motivation
2. Data Preparation
3. Exploratory Data Analysis
4. Classification Methods
5. Future Improvements
6. Challenges

Motivation

Predict bad weathers to prepare for flight planning / trajectory optimization

Data Preparation

Raw Data

FEDERAL CLIMATE COMPLEX
DATA DOCUMENTATION
FOR
INTEGRATED SURFACE DATA
(ISD)

January 12, 2018

NOAA - National Centers for Environmental Information
US Air Force - 14th Weather Squadron
151 Patton Avenue
Asheville, NC 28801-5001 USA

Time Range: 1901 - 2022
Granularity: Hourly
24*365 = 8760 observations / station / year
Variables: 79 Features, Sparse

Global Hourly Data

	STATION	DATE	SOURCE	LATITUDE	LONGITUDE	ELEVATION	NAME	REPORT_TYPE	CALL_SIGN	QUALITY_CONTROL	...	MW2	MW3	MW4	MW5
0	72243012960	2000-01-01T00:00:00	3	29.98	-95.36	29.0	HOUSTON INTERCONTINENTAL AIRPORT, TX US	SY-MT	IAH	V020	...	NaN	NaN	NaN	NaN
1	72243012960	2000-01-01T00:53:00	C	29.98	-95.36	29.0	HOUSTON INTERCONTINENTAL AIRPORT, TX US	FM-15	IAH	V020	...	NaN	NaN	NaN	NaN

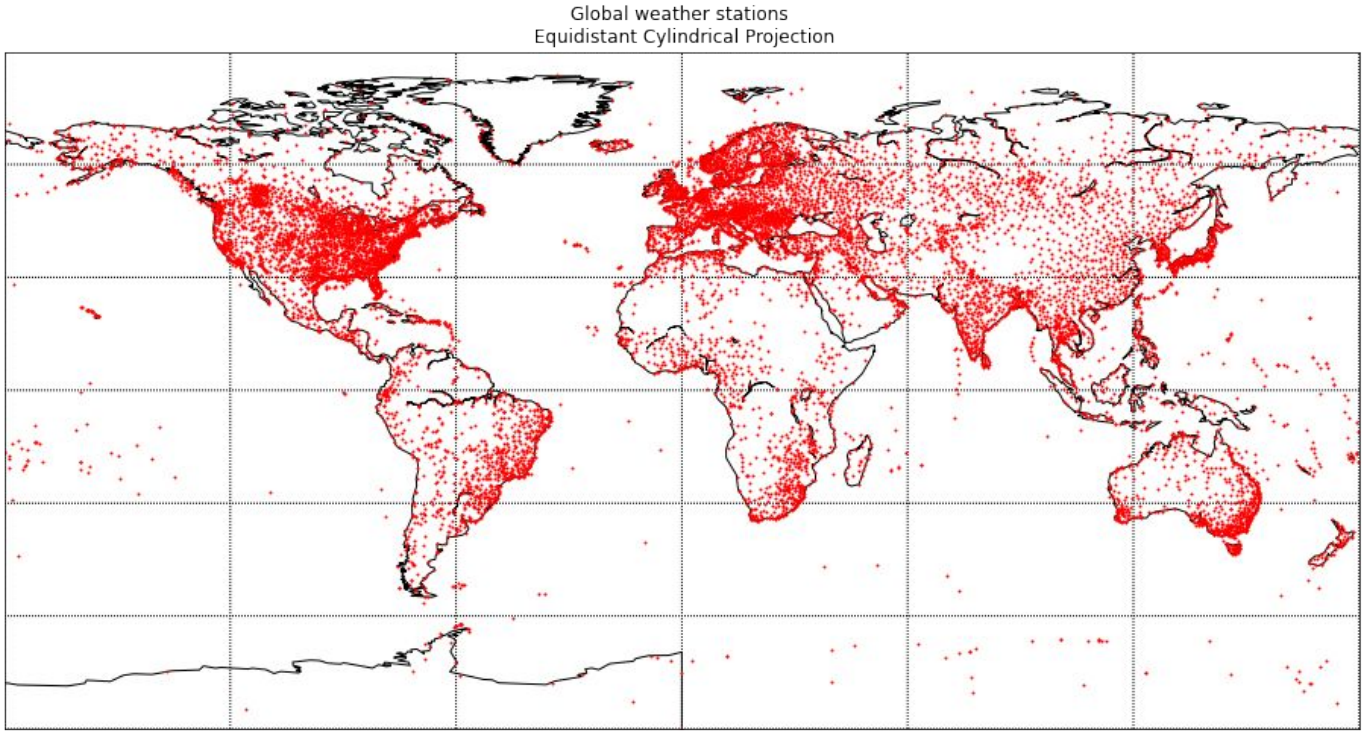
Data Cleaning

1. Select and Parse columns
   ['DATE','WND','CIG','VIS','TMP','DEW''AT1','AT2','AT3']
   Single column contain multiple information
   Examples:
      'WND': 999,9,C,0000,5
      'AT1': MW,01,FG ,5
2. Replace "999" as NaN
3. Fill NaN forward & backward
4. Get binary values of AT (daily weather observation)
5. Generalize weather observation to 1 day (Assumption!)

Loop for selected stations over selected time,
Concatenate Data Frames

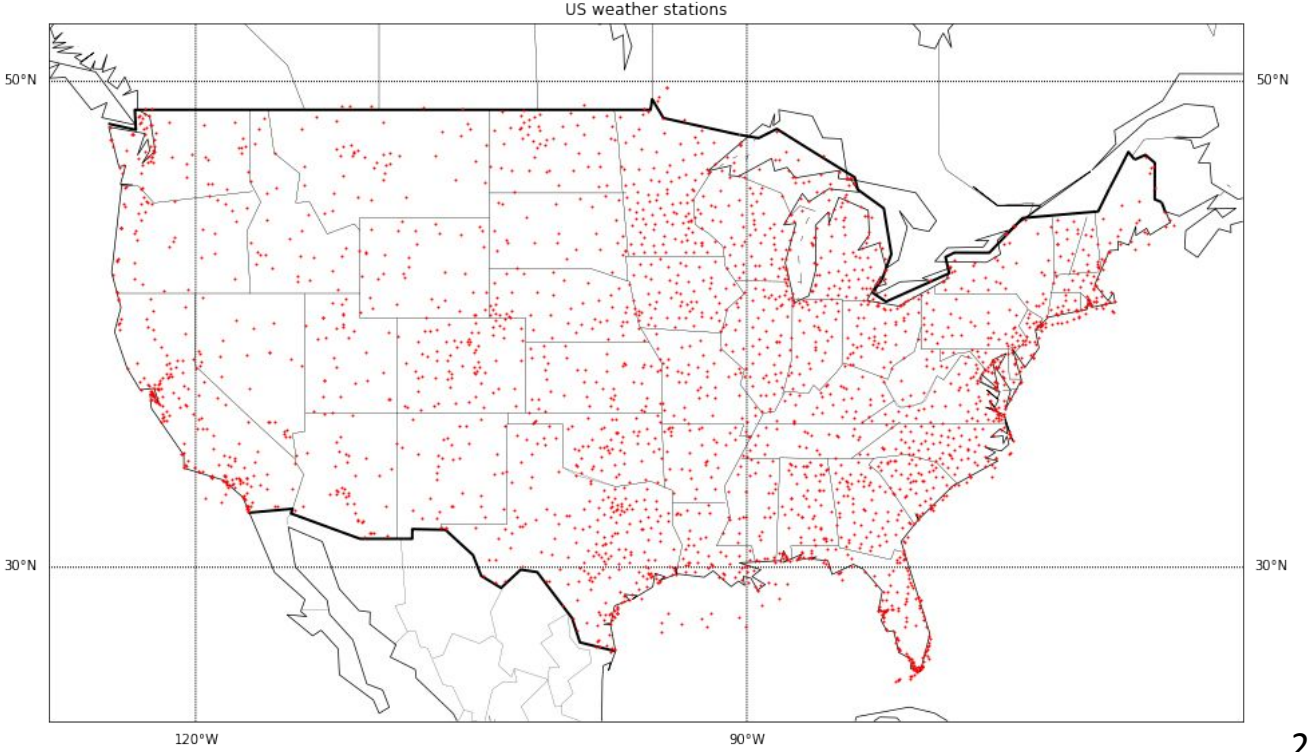
Exploratory Data Analysis

Stations Overview - World



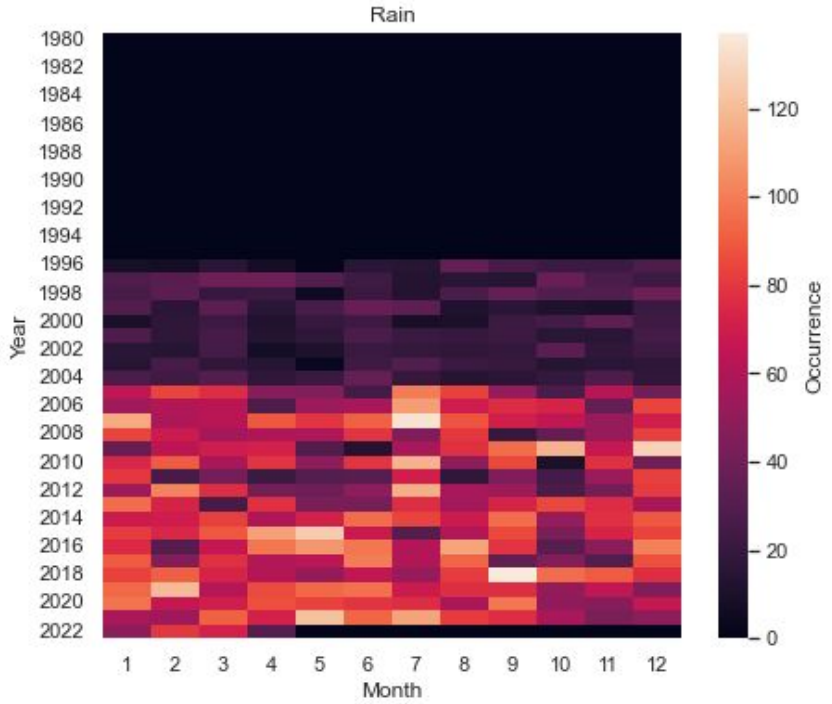
NOAA - Global Hourly
12981 Stations Worldwide

Stations Overview - US



2824 Stations in US
173 Stations in California
217 Stations in Texas

Monthly Overview - Raw Data Heat Map

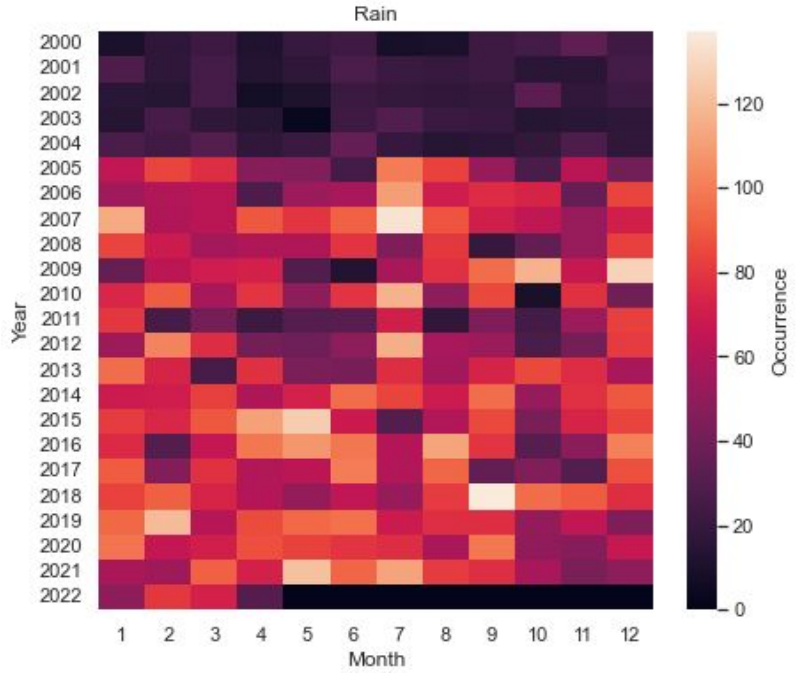


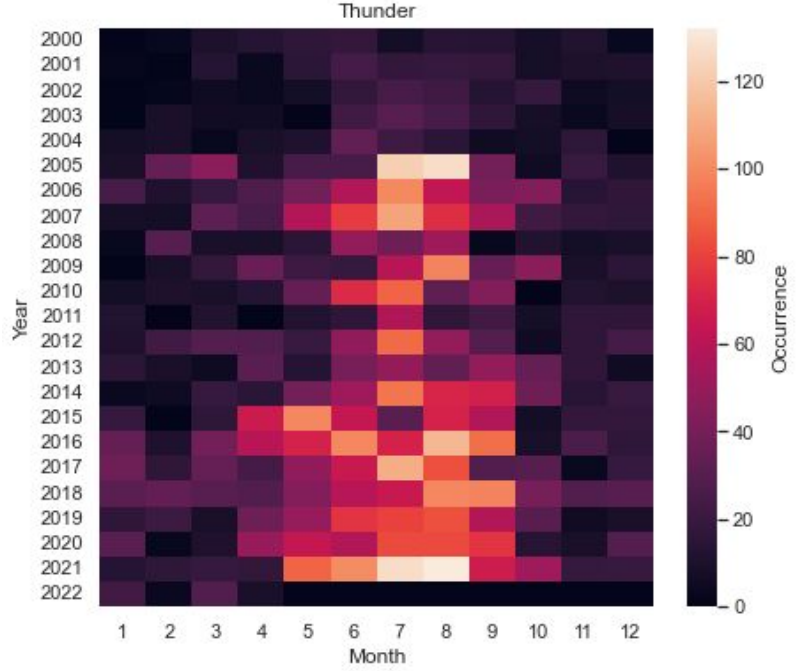
- Houston, TX area (~ 4.8 million samples)
- Missing Data until ~1995
- Less Data between 1995 ~ 2003

- Cleaned irrelevant and faulty data
- Removed repeating samples
- Data transformation
- Rebuild the data set for analysis

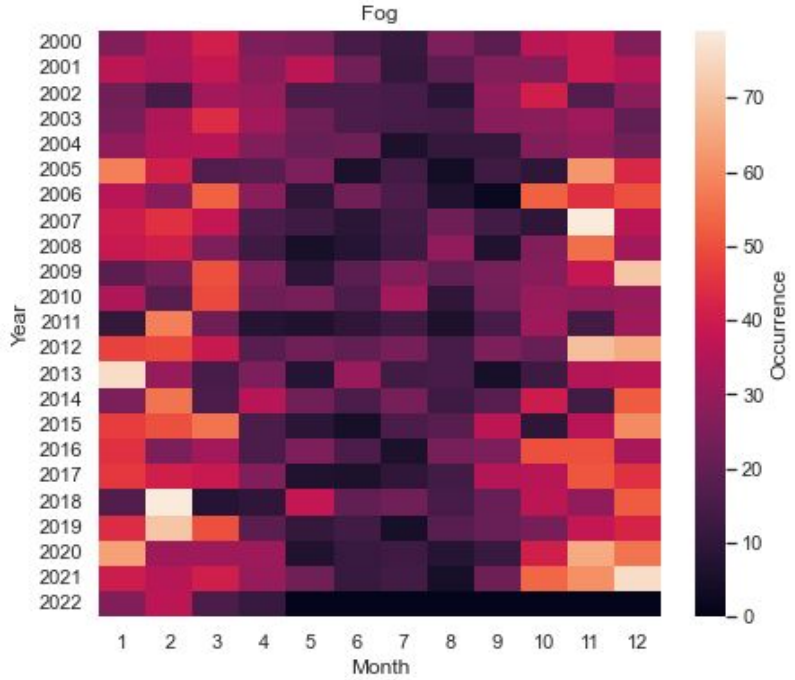
Berkeley
UNIVERSITY OF CALIFORNIA

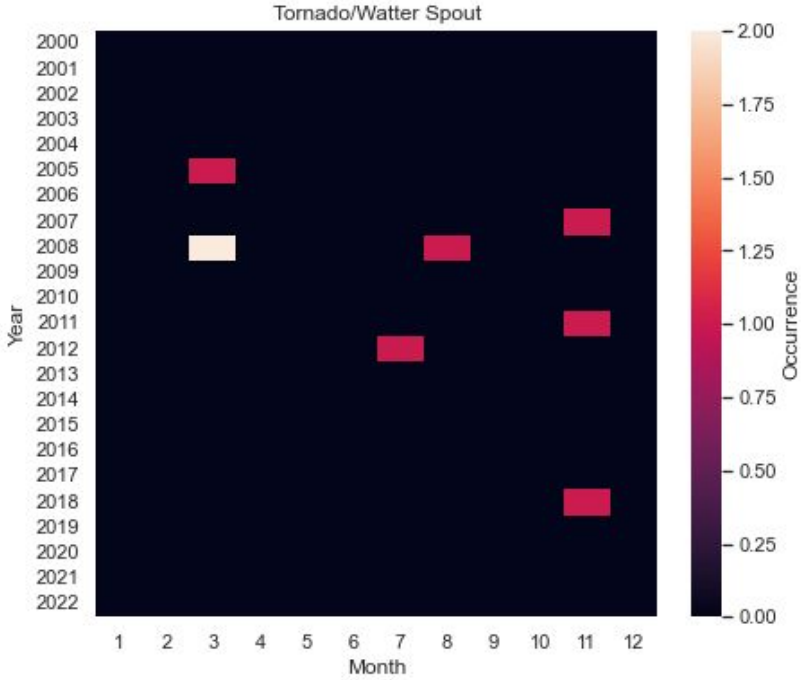
Monthly Overview - RAIN / THUNDER





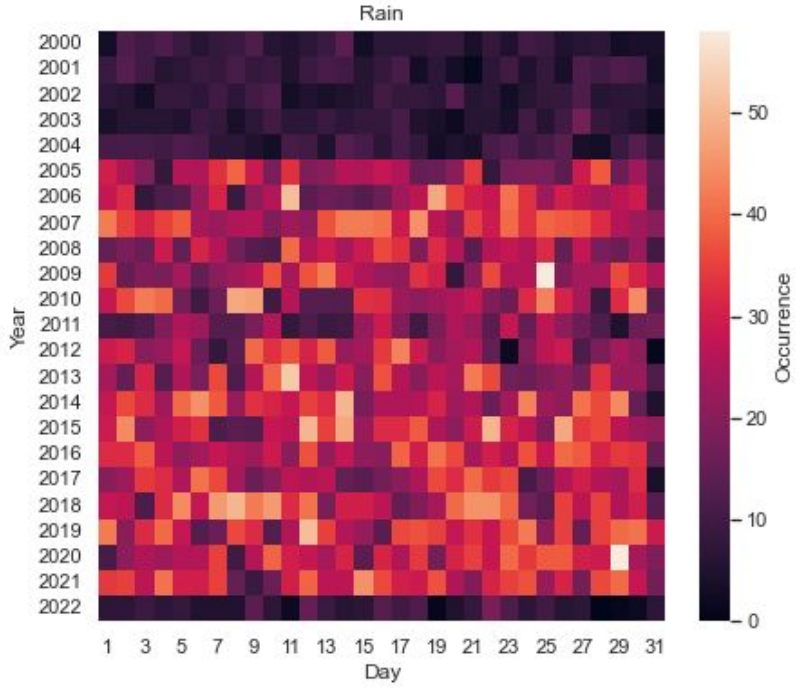
Monthly Overview - FOG / TORNADO

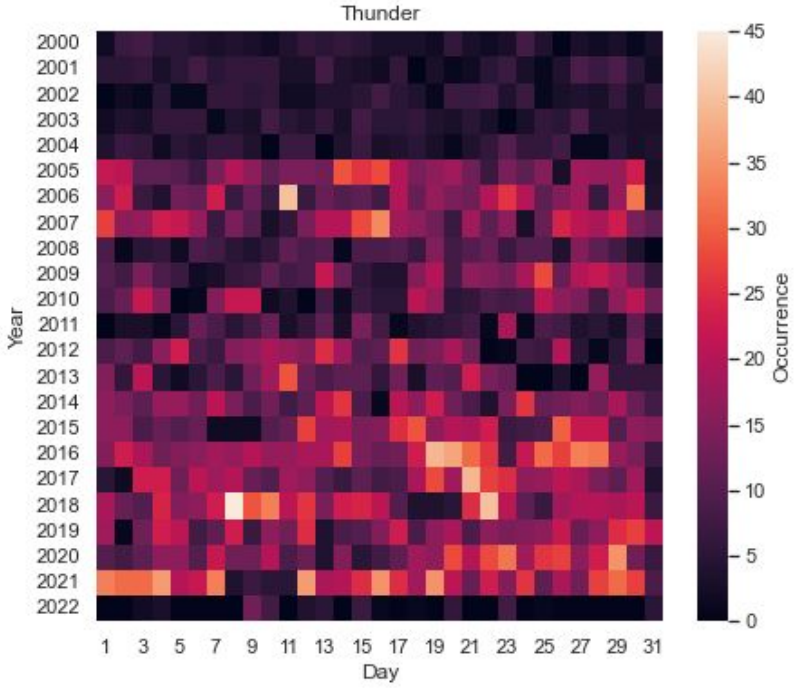




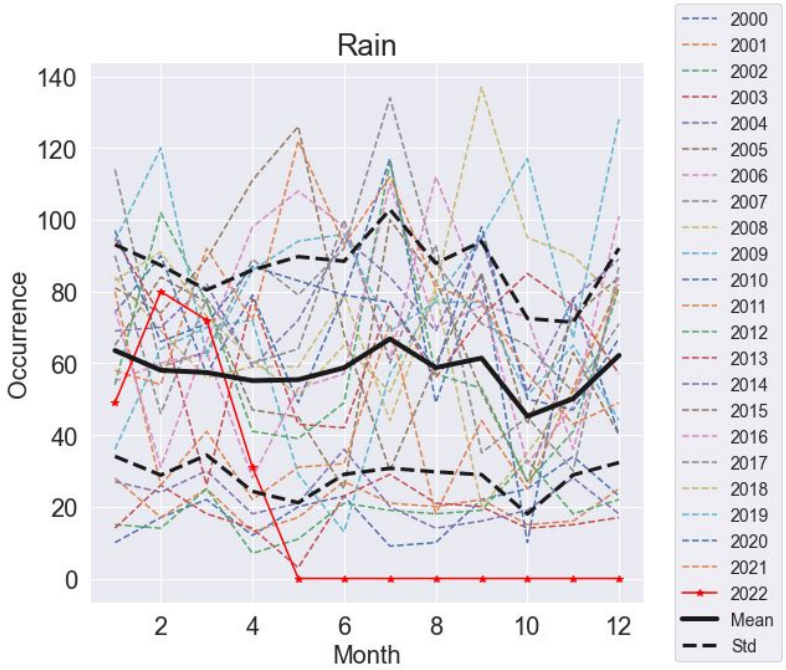
Berkeley
UNIVERSITY OF CALIFORNIA

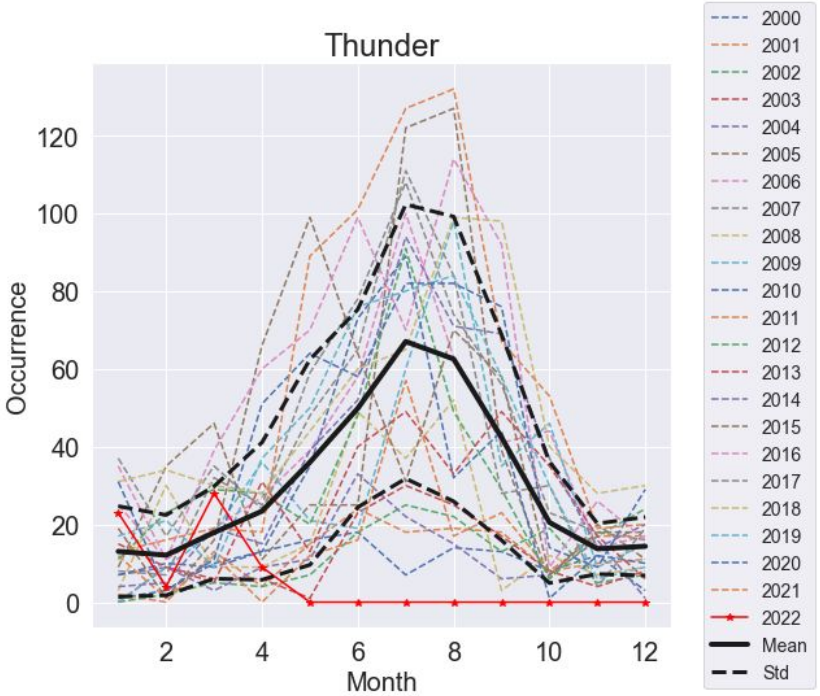
Daily Overview - RAIN / THUNDER



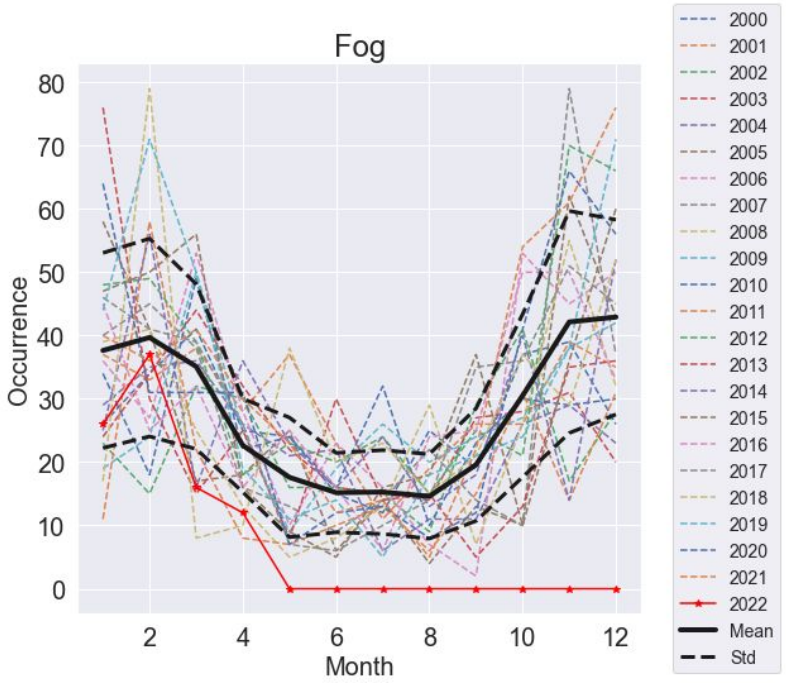


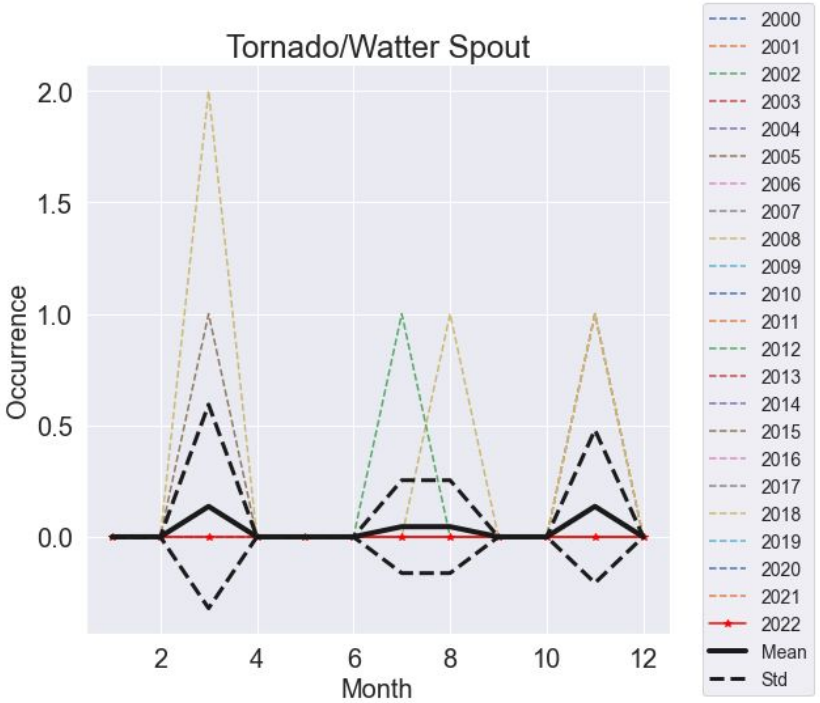
Statistics over Yearly & Monthly Date





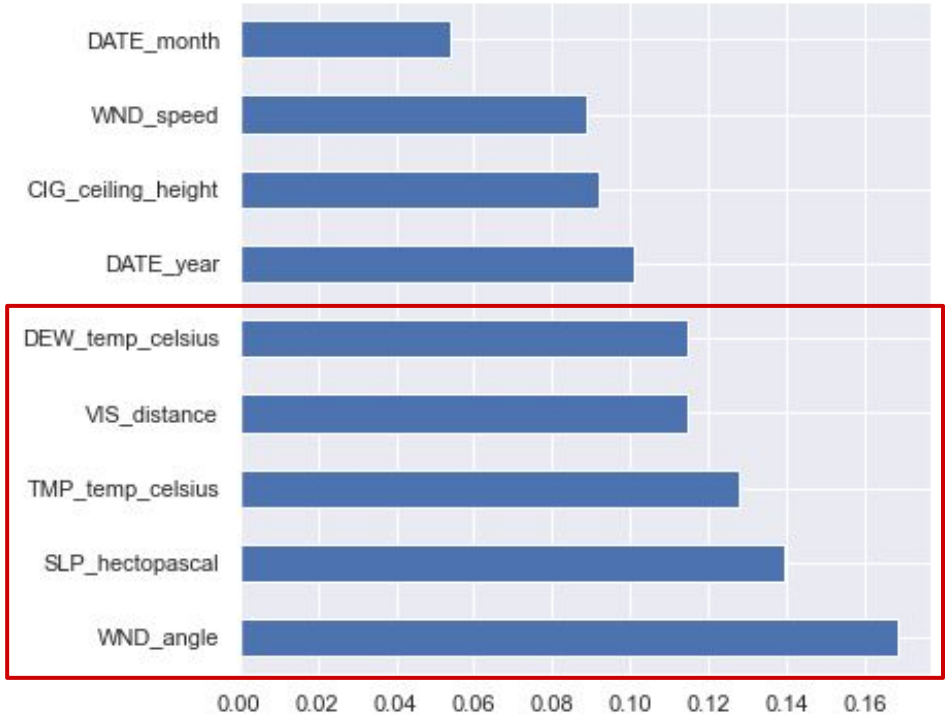
Statistics over Yearly & Monthly Date





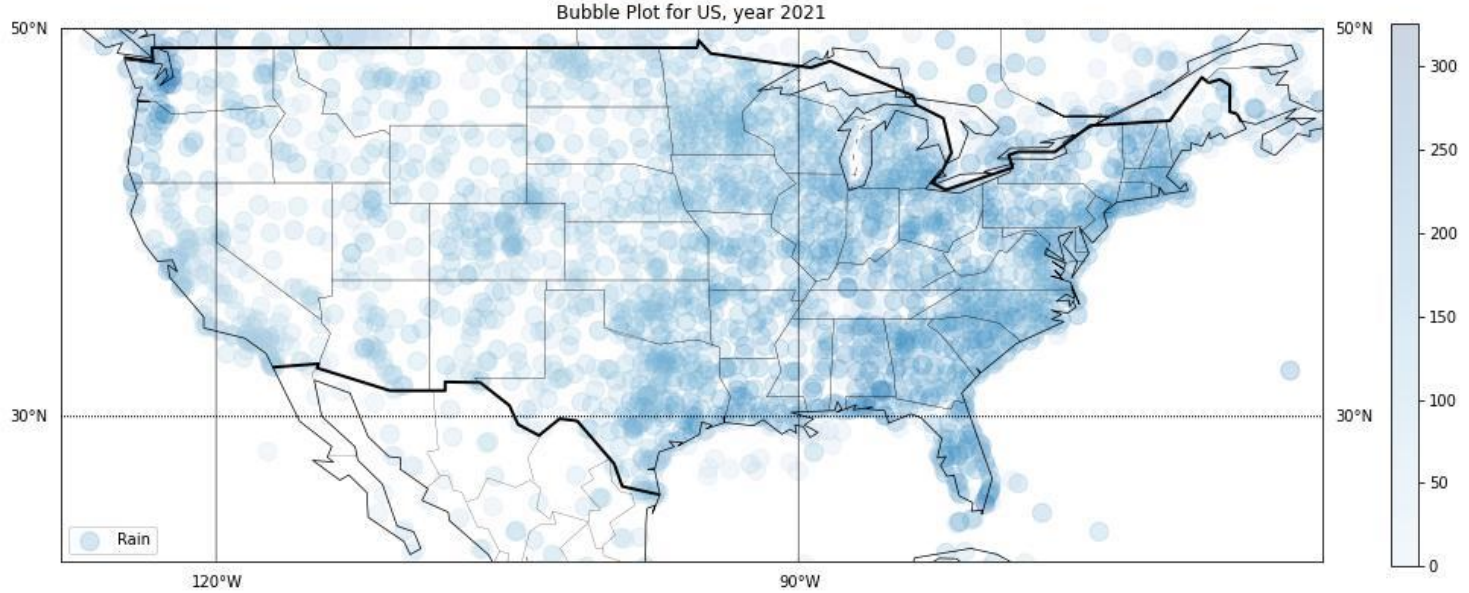
Berkeley
UNIVERSITY OF CALIFORNIA

Feature Importance in Bad Weather Report



- SLP: Sea Level Pressure
- TMP: Temperature
- VIS: Visible Distance
- CIG: Ceiling Height

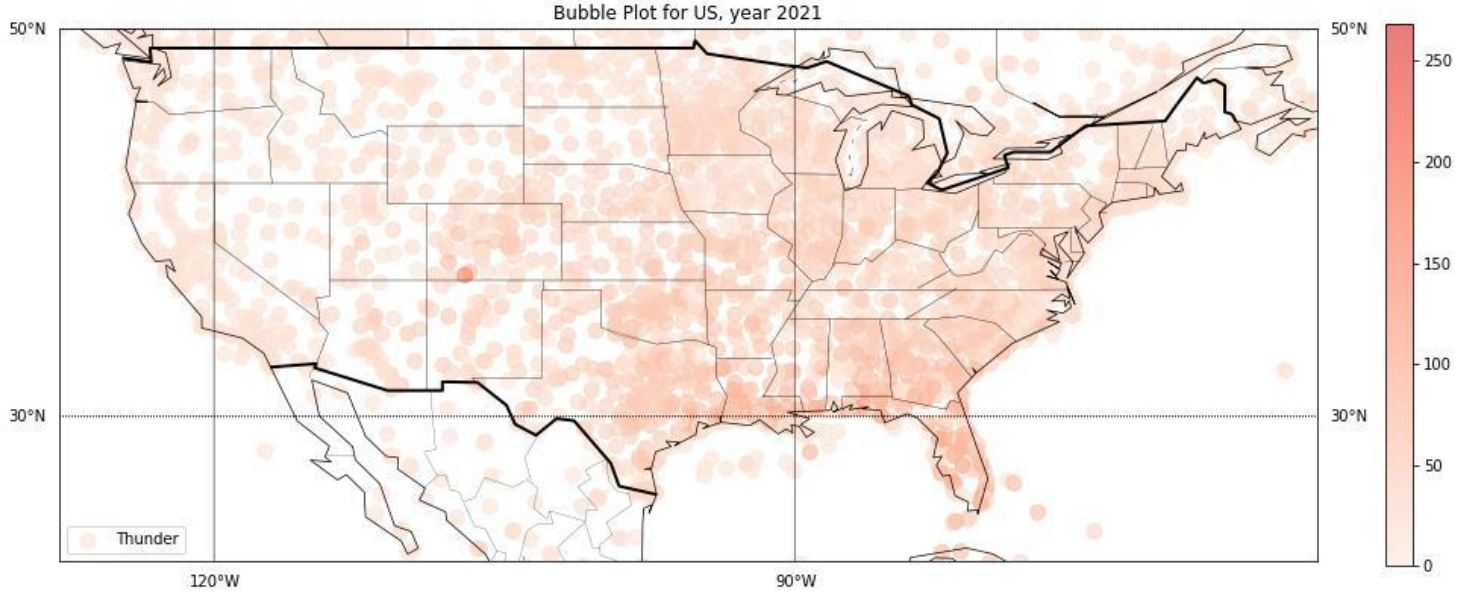
Bubble Map of Weathers (RAIN) - US



Color Gradient ~ Number of weathers occured

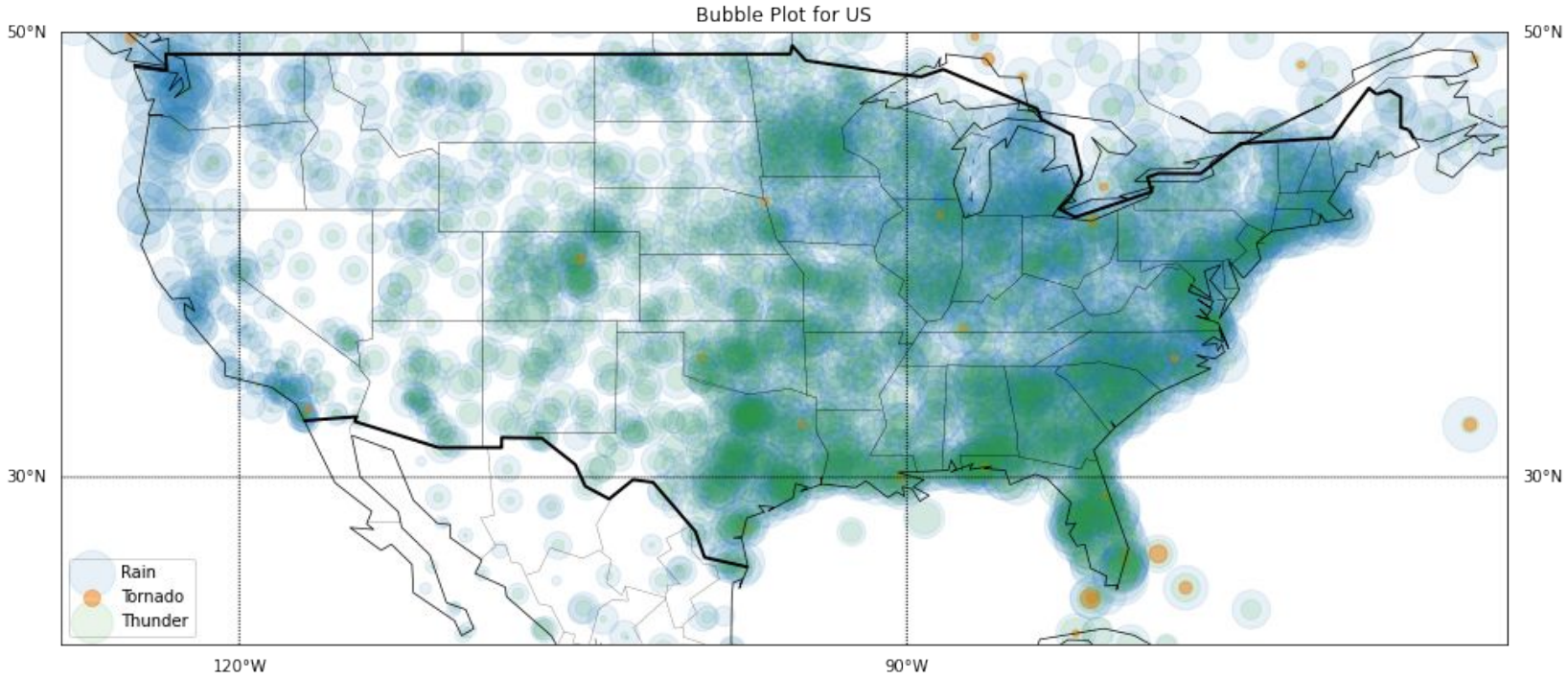
Bubble Map of Weathers (THUNDER) - US

TORNADO?



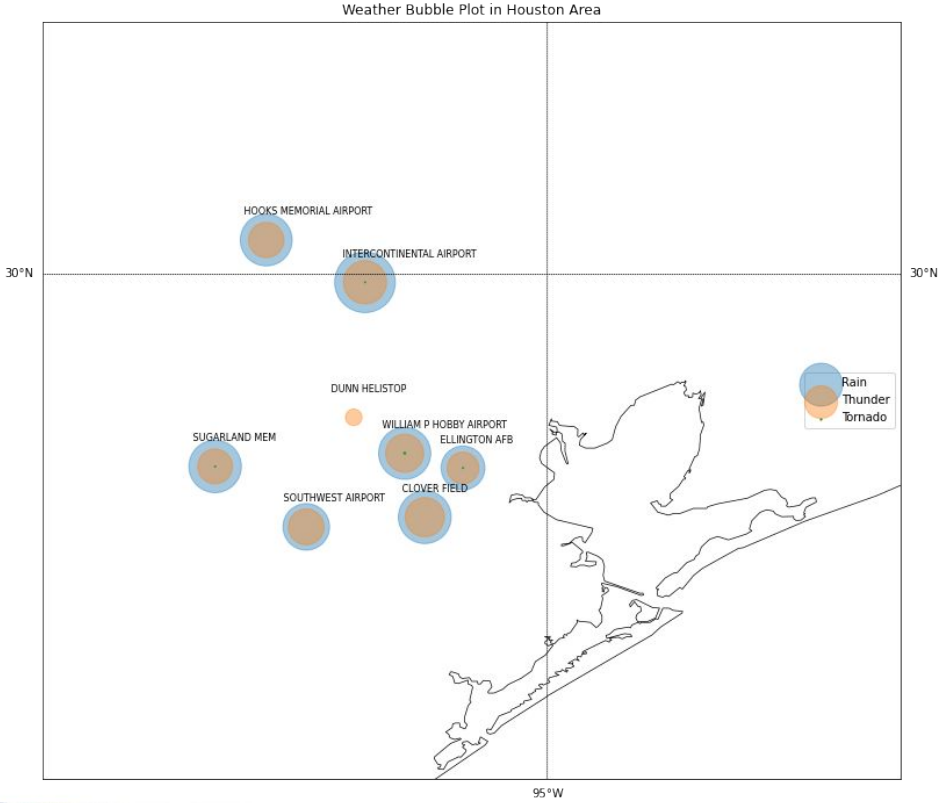
Color Gradient ~ Number of weathers occured

Bubble Map of Weathers - US



Marker Size ~ Number of weathers occured

Bubble Map of Weathers - Houston Area



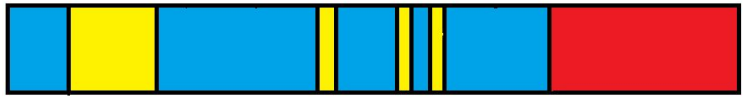
Marker Size ~ Number of weathers occured

Classification Methods

Data

We select one station: Houston Airport
Time range: Year 2000 - 2022
Dataset shape: 309,428 observations

Train-Validation-Test Split: 70-15-15
- Train+Val & Test Splits according to time (non-shuffle)
- Train & Val split with shuffle



Supervised Learning

Baseline: Majority rule
- 18% thunders -> Predict all data as non-thunder
- Baseline accuracy for test set is 0.798

Training data shape:
X: (215670, 13)
Y: (215670, 1)

X:
DATE: year,month,day,hour
Latitude, Longitude
Wind angle
Wind speed
Ceiling height
Visibility
Temperature
Dew point temperature
Sea level pressure

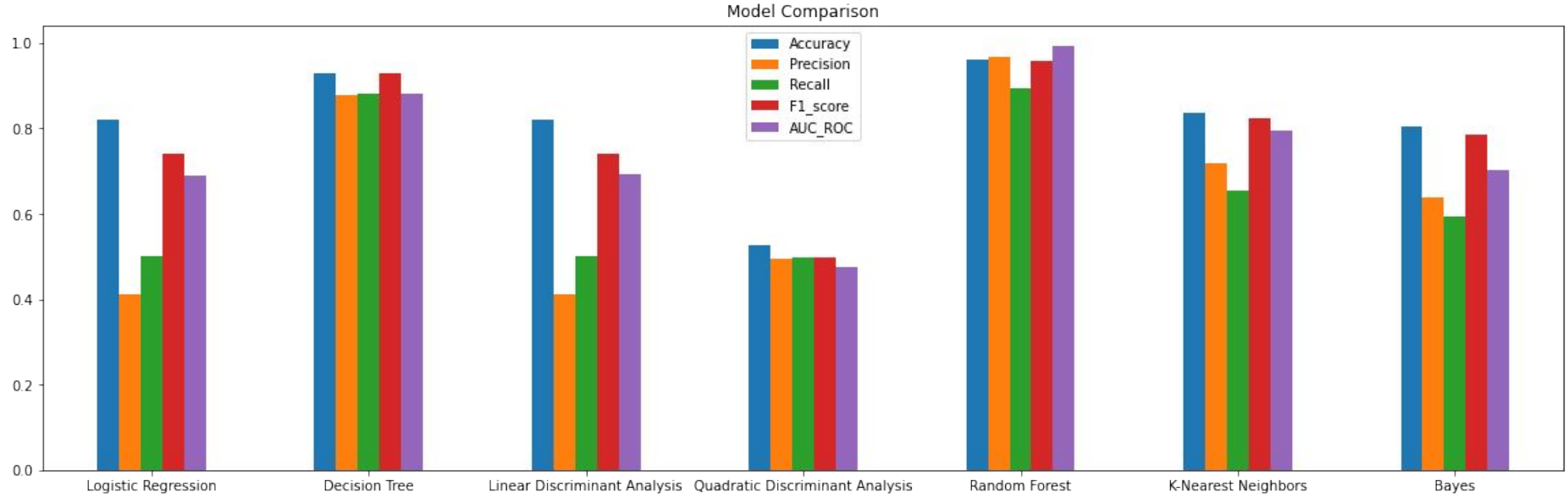
Y:
Thunder_binary

DATE_year	DATE_month	DATE_day	DATE_hour	LATITUDE	LONGITUDE	WND_angle	WND_speed	CIG_ceiling_height	VIS_distance	TMP_temp_celsius	DEW_temp_celsius	SLP_hectopascal
2005	7	27	14	29.98	-95.36	230.0	3.1	7620.0	16093.0	28.9	23.3	1016.6
2004	8	28	5	29.98	-95.36	190.0	1.5	22000.0	16093.0	26.0	24.0	1015.5
2006	10	18	23	29.98	-95.36	50.0	4.1	671.0	2414.0	25.0	23.0	1006.5

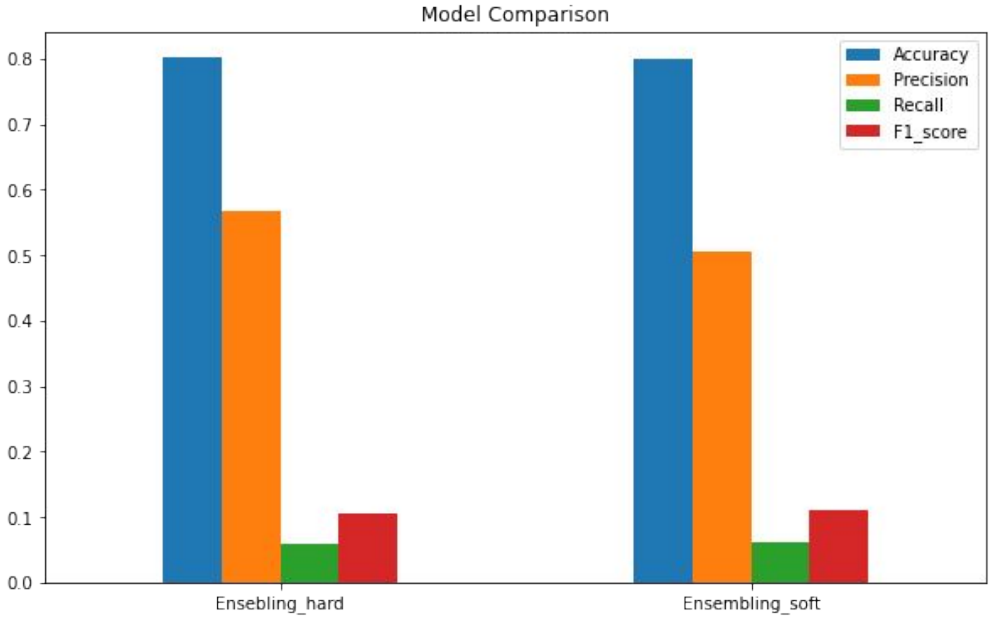
Model Comparison

	Model	Fitting time	Scoring time	Accuracy	Precision	Recall	F1_score	AUC_ROC
4	Random Forest	23.583221	0.611752	0.959531	0.966304	0.894381	0.957838	0.992388
1	Decision Tree	1.281843	0.026187	0.929693	0.879450	0.881624	0.929790	0.881624
5	K-Nearest Neighbors	0.408562	3.324306	0.836866	0.719312	0.655562	0.823084	0.795989
0	Logistic Regression	1.314030	0.020501	0.821422	0.410711	0.500000	0.740887	0.688961
2	Linear Discriminant Analysis	0.433447	0.031723	0.821422	0.410711	0.500000	0.740887	0.692926
6	Bayes	0.101537	0.028895	0.804331	0.638136	0.593777	0.785782	0.700893
3	Quadratic Discriminant Analysis	0.189265	0.034565	0.525873	0.493154	0.496627	0.497315	0.476229

Model Comparison - Training Set

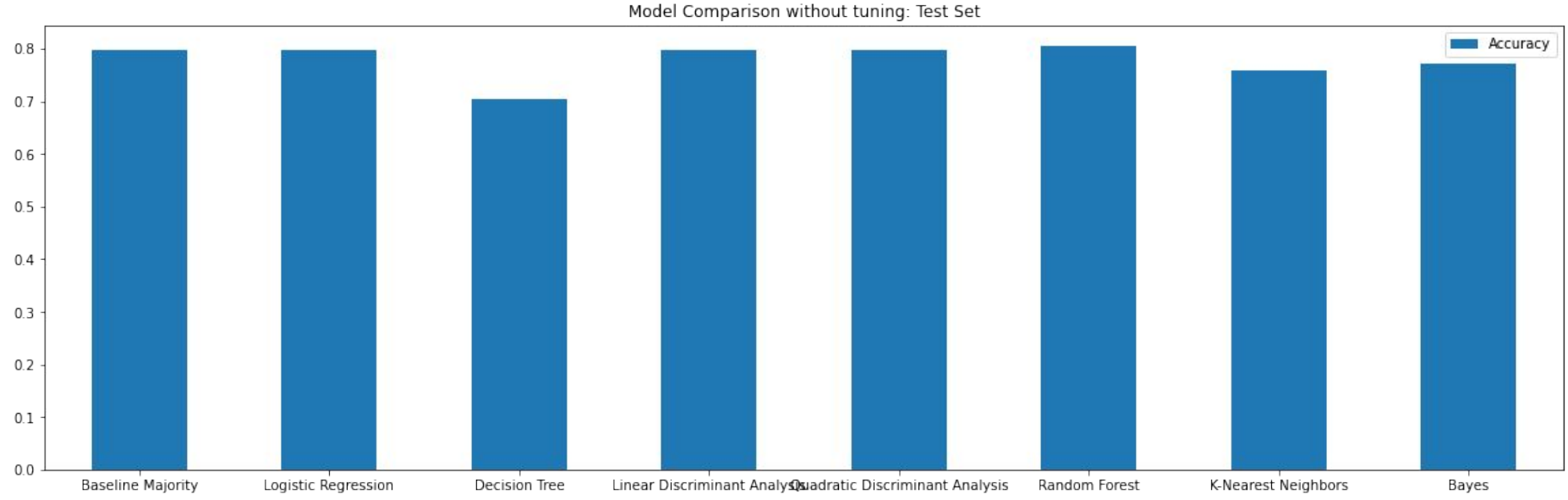


Ensemble Method - Training Set



	Model	Accuracy	Precision	Recall	F1_score	AUC_ROC
0	Ensebling_hard	0.80125	0.567992	0.058044	0.105324	not applicable
1	Ensembling_soft	0.79875	0.506173	0.061358	0.109448	0.686952

Model Comparison - Without tuning



	Model	Accuracy
5	Random Forest	0.804481
0	Baseline Majority	0.798449
1	Logistic Regression	0.798449
3	Linear Discriminant Analysis	0.798449
4	Quadratic Discriminant Analysis	0.798039
7	Bayes	0.771453
6	K-Nearest Neighbors	0.758268
2	Decision Tree	0.703889

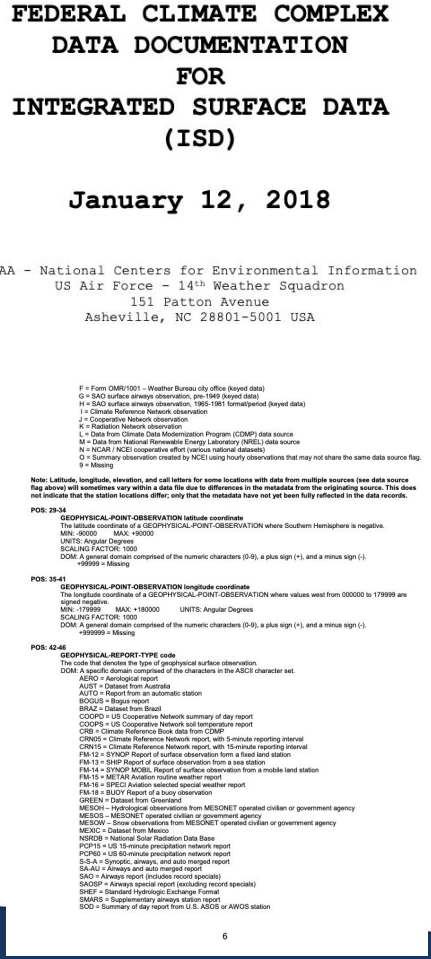
Future improvements

- Fine tunings according to validation set
  - Features selection
  - Parameters selection
- Detailed analysis
  - Confusion matrix visualize
- Other models
  - NN
  - Time series
- Multiclass- MultiOutput classifiers
- Oversampling
  - Imbalanced data
- Model using more stations
  - Sampling, PCA
  - Computing power ?
- Various weather prediction/classification
  - Improve model to predict more various types of bad weather
- Different data source/format
  - Using global daily summary dataset

Challenges & Lessons learned

- Large file size when bulk downloading
- Data formats are messy, missing, sparse
  - Parse & extract according to documentation
- Data preprocessing eating up RAM
  - Preprocess each .csv then combine
- Data storing taking up enormous disk space
- SVM & NN takes long time to train
- Feasibility of the classification models

2013.tar.gz	2021-02-05 13:50	4.2G
2014.tar.gz	2021-02-04 19:30	4.4G
2015.tar.gz	2021-02-04 19:13	4.5G
2016.tar.gz	2021-02-04 19:51	4.5G
2017.tar.gz	2020-12-07 00:21	4.6G
2018.tar.gz	2020-12-06 08:51	4.6G
2019.tar.gz	2020-10-24 01:16	4.7G
2020.tar.gz	2021-02-28 07:25	4.6G
2021.tar.gz	2022-02-09 00:35	4.6G



FEDERAL CLIMATE COMPLEX
DATA DOCUMENTATION
FOR
INTEGRATED SURFACE DATA
(ISD)

January 12, 2018

NOAA - National Centers for Environmental Information
US Air Force - 14th Weather Squadron
151 Patton Avenue
Asheville, NC 28801-5001 USA

F = Form OMR/1001 – Weather Bureau city office (keyed data)
G = SAO surface airways observation, pre-1949 (keyed data)
H = SAO surface airways observation, 1965-1981 format/period (keyed data)
I = Climate Reference Network observation
J = Cooperative Network observation
K = Radiation Network observation
L = Data from Climate Data Modernization Program (CDMP) data source
M = Data from National Renewable Energy Laboratory (NREL) data source
N = NCAR / NCEI cooperative effort (various national datasets)
O = Summary observation created by NCEI using hourly observations that may not share the same data source flag.
9 = Missing

Note: Latitude, longitude, elevation, and call letters for some locations with data from multiple sources (see data source flag above) will sometimes vary within a data file due to differences in the metadata from the originating source. This does not indicate that the station locations differ; only that the metadata have not yet been fully reflected in the data records.

POS: 29-34
GEOPHYSICAL-POINT-OBSERVATION latitude coordinate
The latitude coordinate of a GEOPHYSICAL-POINT-OBSERVATION where Southern Hemisphere is negative.
MIN: -90000 MAX: +90000
UNITS: Angular Degrees
SCALING FACTOR: 1000
DOM: A general domain comprised of the numeric characters (0-9), a plus sign (+), and a minus sign (-).
+99999 = Missing

POS: 35-41
GEOPHYSICAL-POINT-OBSERVATION longitude coordinate
The longitude coordinate of a GEOPHYSICAL-POINT-OBSERVATION where values west from 000000 to 179999 are signed negative.
MIN: -179999 MAX: +180000 UNITS: Angular Degrees
SCALING FACTOR: 1000
DOM: A general domain comprised of the numeric characters (0-9), a plus sign (+), and a minus sign (-).
+999999 = Missing

POS: 42-46
GEOPHYSICAL-REPORT-TYPE code
The code that denotes the type of geophysical surface observation.
DOM: A specific domain comprised of the characters in the ASCII character set.
AERO = Aerological report
AUST = Dataset from Australia
AUTO = Report from an automatic station
BOGUS = Bogus report
BRAZ = Dataset from Brazil
COOPD = US Cooperative Network summary of day report
COOPS = US Cooperative Network soil temperature report
CRB = Climate Reference Book data from CDMP
CRN05 = Climate Reference Network report, with 5-minute reporting interval
CRN15 = Climate Reference Network report, with 15-minute reporting interval
FM-12 = SYNOP Report of surface observation form a fixed land station
FM-13 = SHIP Report of surface observation from a sea station
FM-14 = SYNOP MOBIL Report of surface observation from a mobile land station
FM-15 = METAR Aviation routine weather report
FM-16 = SPECI Aviation selected special weather report
FM-18 = BUOY Report of a buoy observation
GREEN = Dataset from Greenland
MESOH = Hydrological observations from MESONET operated civilian or government agency
MESOS = MESONET operated civilian or government agency
MESOW = Snow observations from MESONET operated civilian or government agency
MEXIC = Dataset from Mexico
NSRDB = National Solar Radiation Data Base
PCP15 = US 15-minute precipitation network report
PCP60 = US 60-minute precipitation network report
S-S-A = Synoptic, airways, and auto merged report
SA-AU = Airways and auto merged report
SAO = Airways report (includes record specials)
SAOSP = Airways special report (excluding record specials)
SHEF = Standard Hydrologic Exchange Format
SMARS = Supplementary airways station report
SOD = Summary of day report from U.S. ASOS or AWOS station

6

Berkeley
UNIVERSITY OF CALIFORNIA

Thank you!

Any Questions?