Tianjia Chen
STAT 625 Case Study
December 6, 2015

Predict Weather in New Haven

1. Motivation

It is interesting to think about whether we can use some weather indications in some areas outside of New Haven to predict weather type of New Haven. Such weather types may include rain, fog, thunderstorm, snow, or any combination of them. The weather indicators may include temperature, dew point, humidity, sea level press, etc.

We need to take into consideration two problems when we construct the prediction model:

(1) What area we should use for weather indicators (predictors)?

(2) Is there any time lag between predictors and indicators? If there is, how long?

For the first problem, I considered three cities, Pittsburgh, Harrisburg and Philadelphia in Pennsylvania. The reasons are as follows:

(1) These three cities are spread out horizontally, and they locate exactly in the west part to New Haven.

(2) Weather data in main cities are easily obtained.

For the second problem, considering the weather of New Haven can be easily influenced by weather of PA, which usually takes one day or more, I set several time lag options and picked one through correlation analysis. It will be discussed further in the Method part.

2. Data

The authoritative source of weather data is National Oceanic and Atmosphere Administration (NOAA). However, since there is no data about weather type in NOAA, I chose another alternative data source: Weather Underground[1].

Here is a short summary of data used in this report:

Data Meta Source: Shared sources with NOAA (140,000+ weather stations in US)

Data Range: 1981.1.1-2015.10.31 (Daily observations)

---

[1] http://www.wunderground.com/

Training Set: 1981.1.1-2014.12.31 (observations, missing data excluded)

Testing Set: 2014.12.31-2015.10.31 (Daily observations)

Variables: temperatures (high, average, low), dew point (high, average, low), humidity (high, average, low), sea level press (high, average, low), visibility (high, average, low), wind (high, average, low), event (fog, rain, snow, thunderstorm, etc.).

3. Method

3.1 Exploratory Data Analysis

1. How similar were weathers in 4 cities?

As we can see from Table 1, the distributions of weather types are quite similar in 4 cities.

Weather	Harrisburgh	Philidelphia	Pittsburgh	PA	New Haven	Difference
Fog	0.10	0.11	0.10	0.10	0.11	0.01
Fog, Rain	0.12	0.12	0.11	0.12	0.11	-0.01
Fog, Rain, Snow	0.01	0.01	0.03	0.02	0.01	-0.01
Fog, Rain, Thunderstorm	0.03	0.02	0.04	0.03	0.02	-0.01
Fog, Snow	0.02	0.01	0.04	0.02	0.01	-0.01
Fog, Thunderstom	0.00	0.00	0.00	0.00	0.00	0.00
Good	0.49	0.51	0.40	0.47	0.55	0.08
Rain	0.16	0.17	0.17	0.17	0.12	-0.05
Rain, Snow	0.01	0.01	0.02	0.01	0.01	0.00
Rain, Snow, Thunderstorm	0.00	0.00	0.00	0.00	0.00	0.00
Rain, Thunderstorm	0.03	0.01	0.03	0.02	0.03	0.01
Snow	0.03	0.02	0.07	0.04	0.02	-0.02
Thunderstorm	0.00	0.00	0.00	0.00	0.00	0.00
Tornado	0.00	0.00	0.00	0.00	0.00	0.00

Table 1 Proportions of Weather Types in Four Cities.

2. Is there any correlation between weather indicators in PA and New Haven?

To answer this question, I looked at scatterplots between weather indicators in PA and New Haven, and then calculated their correlations. Several different time lag periods were considered. It turned out that one-day time lag has the highest correlations (see Figure 1).

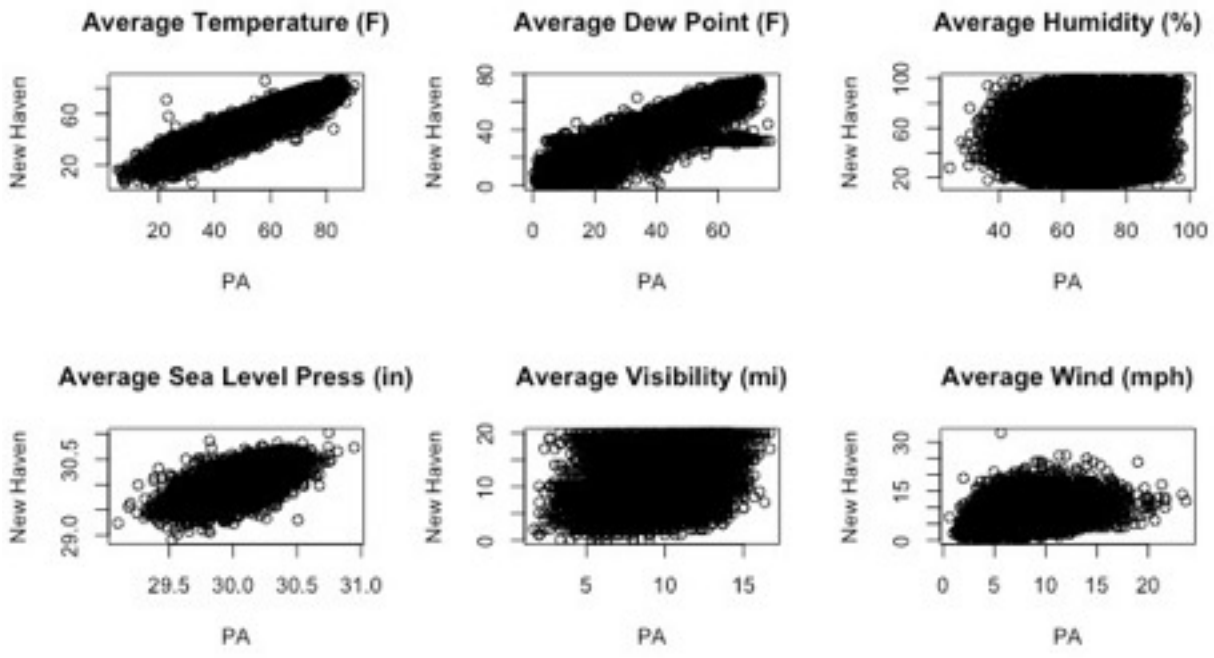

Figure 1 Scatterplots between Indicators in PA and New Haven (one day time lag)

From Figure 1 and the correlations, I used average temperature, average dew point and average sea level press as the final predictors.

3.2 Prediction Model

For classification problem, before constructing the prediction model, I looked at the positive/ negative rate for each weather event (Table 2).

weather	yes (1)	no (0)
rain	2917	6859
fog	2619	7157
snow	474	9302
tornado	1	9775
thunderstorm	558	9218

Table 2 Positive/negative rate for Weather Event

For events of rain and fog, the imbalanced problem is not so severe, and the naive prediction method without balance adjustment may work. For events of snow and thunderstorm, however, the imbalanced problem needs to be solved. For event of tornado, since there was only one event, I omitted this event.

I used logistic model as the basic model for binary classification. To solve the imbalance problem, I applied four methods: oversampling, undersampling, bagging and adaptive boosting. To measure the prediction results, I used four measures: accuracy, sensitivity, specificity and precision. Considering that people care about the cases where the weather is bad and they do the protection, I concentrated on the measure of sensitivity. The result for each event can be found in Table 3, 4, 5, 6.

Method	Accuracy	Sensitivity	Specificity	Precision
Naïve method	0.73	**1.00**	0.73	0.01
Oversampling	0.67	0.43	0.83	0.62
Undersampling	0.68	0.43	0.83	0.62
Bagging	0.73	**1.00**	0.73	0.01
Adaptive Boosting	0.77	0.76	0.77	0.20

Table 3 Prediction Performance for Rain

Method	Accuracy	Sensitivity	Specificity	Precision
Naïve method	0.90	NaN	0.90	0.00
Oversampling	0.58	0.12	0.91	0.37
Undersampling	0.58	0.12	0.91	0.50
Bagging	0.90	NaN	0.90	0.00
Adaptive Boosting	0.85	**0.14**	0.90	0.10

Table 4 Prediction Performance for Fog

Method	Accuracy	Sensitivity	Specificity	Precision
Naïve method	0.90	0.40	0.90	0.07
Oversampling	0.85	0.39	1.00	0.97
Undersampling	0.85	0.39	0.99	0.93
Bagging	0.90	**0.50**	0.91	0.07
Adaptive Boosting	0.85	0.35	0.95	0.60

Table 5 Prediction Performance for Snow

Method	Accuracy	Sensitivity	Specificity	Precision
Naïve method	0.95	NaN	0.95	0.00
Oversampling	0.58	0.11	1.00	1.00
Undersampling	0.56	0.10	1.00	1.00
Bagging	0.95	NaN	0.95	0.00
Adaptive Boosting	0.84	**0.20**	0.98	0.73

Table 6 Prediction Performance for Thunderstorm

The very low precision came from the fact that in some results there were very few, or even no true positive at all. The NaN in sensitivity came from the fact that in some results there were no true positives and false negatives (there were no positive events in testing set). As we can see from the above tables, the model provided relatively good performance in sensitivity for rain and storm, but performed poorly for fog and thunderstorm. Possible reason could be the predictors

we used, such as temperature, dew point, sea level press, can well explain rain and storm, but are not suitable for prediction of fog and thunderstorm.

Last, let's look at the precision of all events in testing set, with prediction results for four events with highest sensitivity, respectively (Table 7).

Weather	Accuracy
Fog	0.95
Fog, Rain	0.95
Fog, Rain, Snow	0.98
Fog, Rain, Thunderstorm	0.99
Fog, Snow	0.99
Fog, Thunderstom	0.96
Good	0.65
Rain	0.85
Rain, Snow	0.99
Rain, Snow, Thunderstorm	1.00
Rain, Thunderstorm	0.96
Snow	0.93
Thunderstorm	0.86

Table 7 Prediction Accuracy for Events

If we only look at the prediction accuracy for weather events, the results are quite satisfying. But we should notice that prediction accuracy only look at the proportion of true positives and true negatives. The high accuracy could come from the high proportion of true negatives, which is of course a good thing but we actually want to concentrate more on the relationship between true positives and false negatives for weather prediction.