

Influenza Outbreak Forecast in New York City with Weather Data

Zhengxing Li, Ziying Tao, Guangjun Xue, Yichen Zhang
 zh1028@eng.ucsd.edu, zit037@eng.ucsd.edu, guxue@eng.ucsd.edu, yiz037@eng.ucsd.edu

Predicting

- Influenza is one of the most significant diseases in humans.
- Better epidemic predictions would set up more appropriate public health prevention and intervention strategies in temperate cities.
- Epidemics occur mainly during the season months with abnormal changing of temperature, **precipitation**, **UV radiation**, and wind speed.
- Random Forest, **Linear Regression**, Gradient Boosting, K-Nearest Neighbors Algorithm.
- Based on the future weather data to predict the Influenza.

Data

Datasource:

- Daily weather data from NOAA(2011-2018)
- Weekly influenza infections data from Centers for Disease Control and Prevention(2011-2018)

Data preprocess:

- Data Cleaning
- Aggregate daily weather data into weekly data
- Merge weather data and influenza data

year	week	temp_weeklyMin	temp_weeklyMax	temp_weeklyMean	slp_weeklyMean	wdsp_weeklyMedian	mxdsp_weeklyMax	prcp_weeklyMedian	uv_time	
0	2011	1	24.1	46.0	31.371429	1007.542857	5.6	15.0	0.00	0.39
1	2011	2	18.0	37.9	27.471429	1018.271429	6.4	15.0	0.00	0.39
2	2011	3	14.0	41.0	27.500000	1013.442857	6.4	15.0	0.01	0.40
3	2011	4	6.1	37.9	28.728571	1016.542857	4.1	15.9	0.01	0.41
4	2011	5	21.9	44.1	31.014286	1020.042857	6.3	18.1	0.06	0.42
5	2011	6	15.1	45.0	31.371429	1015.857143	6.5	21.0	0.00	0.43

Figure 1. Preprocessed Dataset

Features

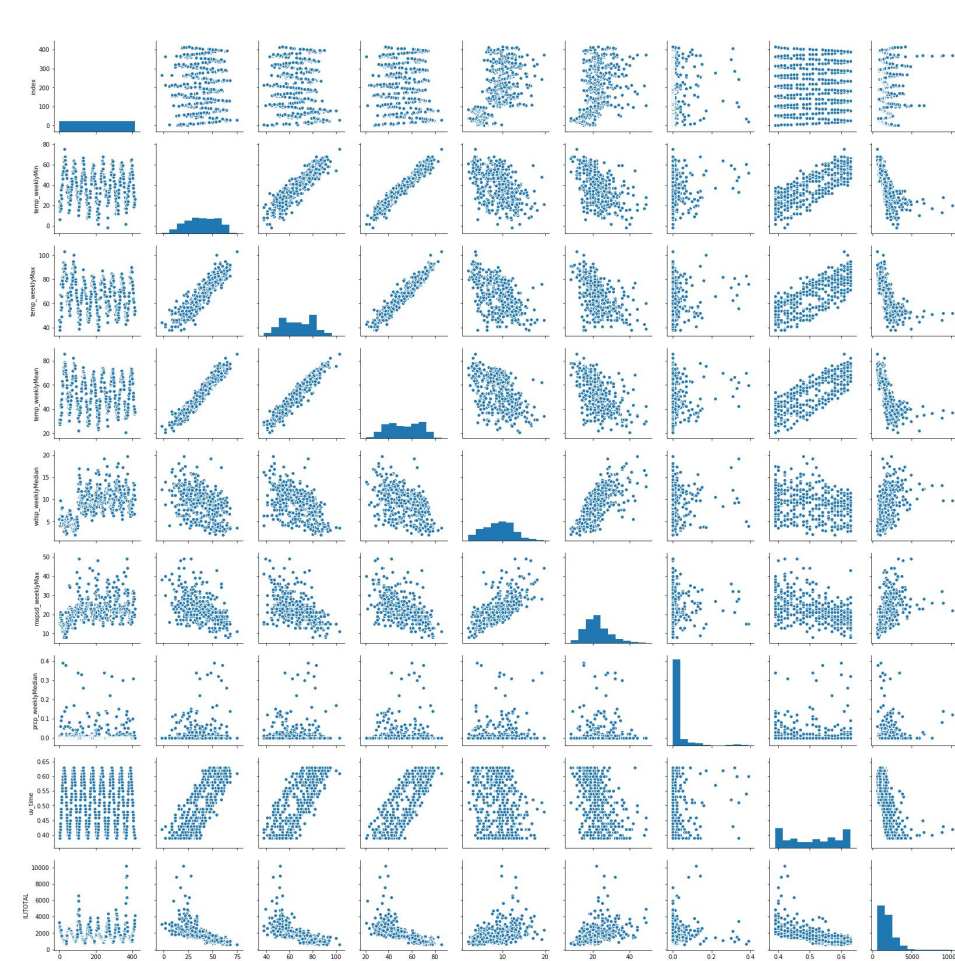


Figure 2. Pairplot of all features

The raw data includes features: weekly highest temperature, weekly lowest temperature, weekly average temperature, weekly sea level pressure, weekly max wind speed, weekly minimum wind speed, weekly average wind speed, weekly average uv time, weekly average precipitation.

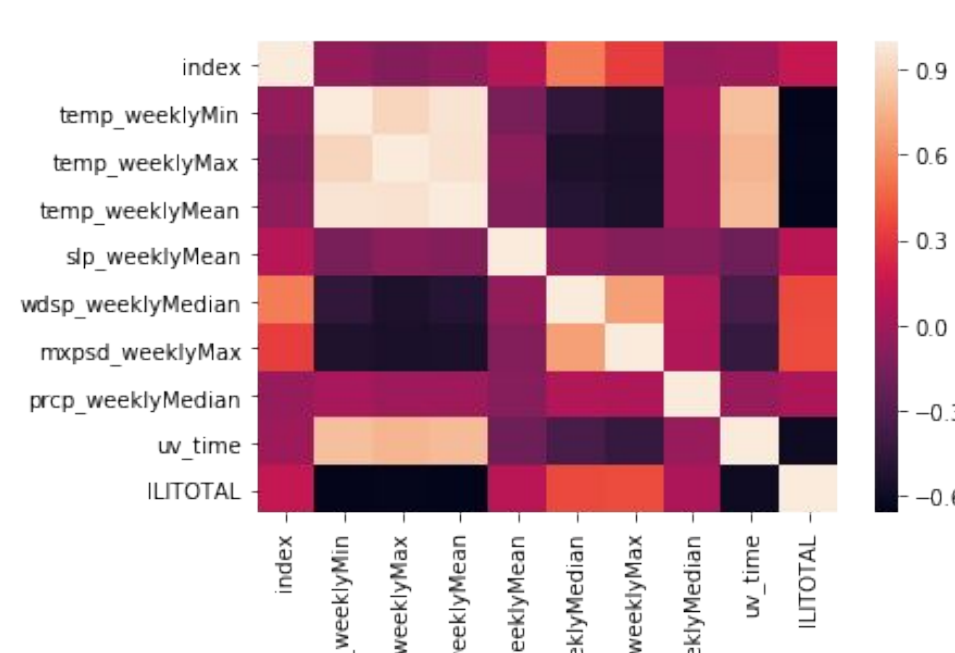


Figure 3. Correlation heatmap of features and influenza infections

Before we get started to develop models, we created visualizations of the features to explore the potential relationship between them. From the heatmap of correlation between features, we can pick the features of more importance.

Models

We adopt five models to do the prediction.

1. **Linear Regression:** $\hat{y} = h_{\theta}(x) = \theta^T \cdot x$.

2. **Random Forest Regression:** $C_{bag}(x) = \arg\max_m \{C(S^m, x)\}_{m=1}^M$

We set bootstrap=True and n_estimators=100.

3. **Gradient Boost:**

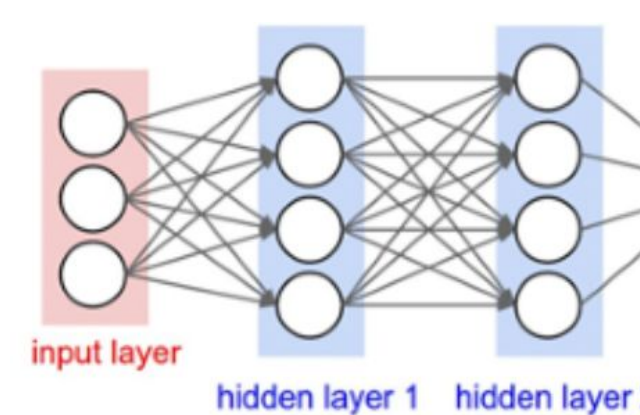
$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}R_{jm}(x) \quad \gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Models

4. **KNN:** $P(y = j | X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j)$

Given a point x, we will choose K nearest points from dataset and have a vote among them to decide the prediction.

5. **DNN:**



Our network consists of one input layer and two hidden layer and the output layer. We use relu as activation function. And epoch=20000.

Results

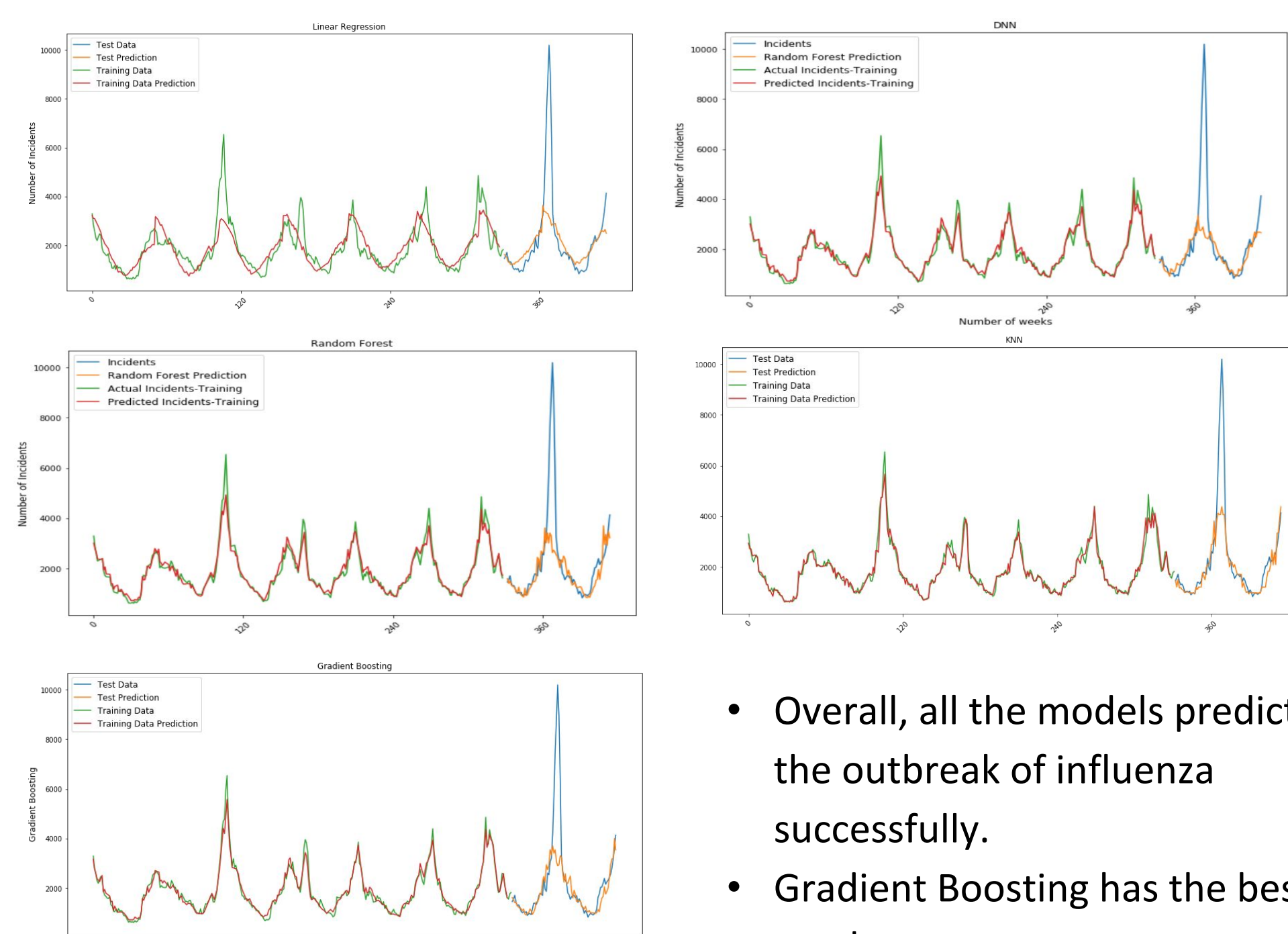


Figure 4. Predictions for models

- Overall, all the models predict the outbreak of influenza successfully.
- Gradient Boosting has the best result on testset.

	Training Data(332 weeks)	Testing Data(83 weeks)
Linear Regression	558.823	1403.405
Random Forest	222.476	1461.625
Gradient Boosting	169.714	1392.070
KNN(k=2)	601.175	1500.434
DNN	527.139	1470.413

Table 1. RMSE of training data and testing data

Discussion

Linear Regression model, using the combination of all the features, is a simple but effective way in predicting. For that flu infection is the result of interaction of weather features, KNN model, which is based more on similarity, is not very suitable for this case. Compare with the result of random boosting, we can find that random forest is much easier to tune and harder to overfit.

Our weather data are regularly changing with seasons changing, which likes the data are combined by several lines just with positive or negative slopes and linearly at every small locality. Then, during reading references, we find the New York City is a typical temperature city where the UV radiation and precipitation play the significant role. So we use the models to test and train the data without UV radiation and precipitation respectively, the consequence shows that the RMSE without UV/Precip is four times of the RMSE with UV/Precip, which agrees with the reference and we approve these by using the machine learning.

Future

We will use universal weather data, like finding different meteorological stations' data, and using the longer span of year, or do more literature research to find a better way, to do test. Besides, we will also improve our model and try several new models to see the consequence which is more reasonable.

Reference

- [1] World Health Organization: Influenza (Seasonal). 2014
- [2] Viboud C, Flahault A. Influenza Epidemics in the United States, France, and Australia.
- [3] Tamerius JD, Viboud C. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates.
- [4] Finkelmann BS, Grenfell BT. Global Patterns in Seasonal Activity of Influenza A/H3N2, A/H1N1, and B from 1997 to 2005: Viral Coexistence and Latitudinal Gradients.
- [5] Moura FEA, Perdigão ACB, Siqueira MM. Seasonality of influenza in the tropics: a distinct pattern in Northeastern Brazil.
- [6] Rao BL, Banerjee K. Influenza surveillance in Pune, India, 1978–90.
- [7] Rao BL, Yeolekar LR, Kadam SS, Pawar MS, Kulkarni PB, More BA, Khude MR. Influenza surveillance in Pune, India, 2003.
- [8] Dosseh A, Ndiaye K, Spiegel A, Sagna M, Mathiot C. Epidemiological and virological influenza survey in Dakar, Senegal: 1996–1998.
- [9] Fuhrmann C. The effects of weather and climate on the seasonality of influenza: what we know and what we need to know.
- [10] Lowen AC, Steel J, Mubareka S, Palese P. High temperature (30 °C) blocks aerosol but not contact transmission of influenza virus. J Virol.
- [11] Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering. 2009 Apr; 1(1):1793–8201.
- [12] Han J, Kamber M. Data mining: Concepts and techniques. Morgan and Kaufmann; 2000.
- [13] Smith BA, McClendon RW, Hoogenboom G. Improving air temperature prediction with artificial neural networks. International Journal of Computational Intelligence. 2007; 3(3):179–86.