STAT4600 – Big Data Statistics I
Final Project Report
Group 1
Instructor: Haitao Liu

Climate Change in Worcester

Jade Zhang, Yan Shi, Jiamei Wang, Aditya Katheeth, Vishvendra Hudda

STAT4600 – Big Data Statistics I
Final Project Report
Group 1
Instructor: Haitao Liu

Abstract

This project analyses whether climate change exists in Worcester in terms of temperature and extreme weather, and how the insurance companies should change their prices and policies for weather insurance due to climate change in the predictable future. The project uses the weather data of Worcester for each day from Jan 1950- Dec 2018. Data visualization, linear regression, and ARIMA and SARIMA are applied for temperature analysis, and machine learning methods are used for extreme weather analysis. There was no significant change in weather over time but a slight increase in temperature has been observed over the years, and our models performed well in both temperature and extreme weather prediction.

1. Introduction

The global climatic shift has already had evident effects on the environment. Effects that scientists had predicted in the past would result from global climate change are now occurring as loss of sea ice, accelerated sea-level rise, and longer, more intense heat waves.

Scientists have concluded that global temperatures will incline for decades, mainly due to greenhouse gases produced by human activities. The Intergovernmental Panel on Climate Change, which includes more than 1,300 scientists from the United States and other countries, forecasts a temperature rise of 2.5 to 10 degrees Fahrenheit over the next century.

Abnormal weather is nothing but the difference between observed weather and its "normal value," which is typically calculated using the 30-year average. With climate change, the frequency and the intensity of abnormal weather patterns have dramatically increased, and the shift to warmer temperatures will aggregate this phenomenon.

The motivation to choose this topic is to create awareness, and we are galvanized to help businesses focus on damage control by putting their best bet on insurance companies. Business problems faced by the masses due to climatic shifts and extreme weather conditions push towards the emergence of weather insurance.

There are mainly two types of weather insurance:

Conditional weather insurance

Conditional weather insurance is more of a retail type of insurance, where a company stands behind a product guaranteeing its functional expectancy or quality during the complete duration of extreme weather conditions and will reimburse if the product fails during that time.

Weather cancellation insurance

Weather cancellation insurance is more event-based. Where an event planned had to be canceled due to harsh weather and are affected financially. Weather cancellation insurance reduces an organization's risk of planning an outdoor event. Weather cancellation insurance ensures that if weather inclination does occur. The organization does not take a financial hit. Instead, the insurance company will cover the cost based on the size and type of insurance purchased.

This study is the response to how should insurance company adjust their prices and policy for weather insurance due to climate change in the predictable future. The questions we would like research are as follows:

- Whether there has been a trend of change in temperature in Worcester over the years?
- Is it possible for us to predict future temperature?
- Has the occurrence of extreme weather become more frequent over the years?
- Is it possible for us to predict the extreme weather?

To better understand the problem, we have analyzed the change of temperature in Worcester from Jan 1950 to Dec 2018 to observe if there is a trend and predicted the temperature based on linear regression, ARIMA (AutoRegressive Integrated Moving Average), and SARIMA (Seasonal Autoregressive Integrated Moving Average). Then, we applied machine learning classifiers to predict extreme weathers. The details are illustrated in the following sections.

2. Method and Analysis

Our data is obtained from a public resource released by the National Oceanic and Atmospheric Administration (NOAA), which provides us two separate files for the different research questions. For the question about temperature change, the file contains columns including date, daily maximum temperature, and minimum temperature. For the question about extreme weather, in addition to the information in the previous file, the file consists of many more columns such as daily snowfall, daily precipitation, and the weather conditions as dummy variables corresponding to every day. For example, if February 7th, 1995 is with thunder, then the thunder column will be 1 in the 1995/02/07 row, and other weather condition columns will be 0.

The variables we use are also different for two questions. We treat the first question as a time-series prediction problem, and the independent variable and dependent variable are date and temperature correspondingly. To analyze and predict the extreme weather, we use temperature, precipitation, and snowfall as independent variables, and the extreme weather as the dependent variable.

In order to perform further process, we have separated the date as the year, month, and day in different columns, marked the season and the number of the week in every month, and defined the extreme weather variable based on the weather condition columns. The index was assigned, as well. The columns that would not be used are dropped. For the time-series prediction question, data is resampled by the mean of months to perform the ARIMA and SARIMA. Both daily maximum temperature and minimum temperature are predicted in our model, but since they are highly correlated, the results are similar in terms of accuracy.

To apply the machine learning methods, we have checked the correlation of independent variables. According to Figure 1, which visualizes the correlation between the variables while yellow dots represent extreme weather and black dots stand for non-extreme weather, it is evident that the daily maximum temperature is strongly correlated to the minimum temperature. Therefore, we have dropped the maximum daily temperatures and used the minimum daily temperatures instead.

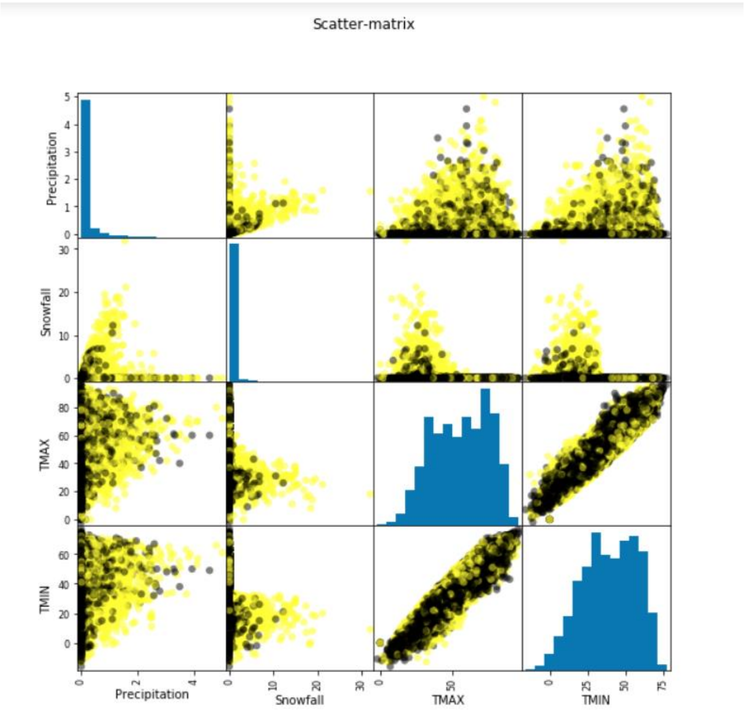
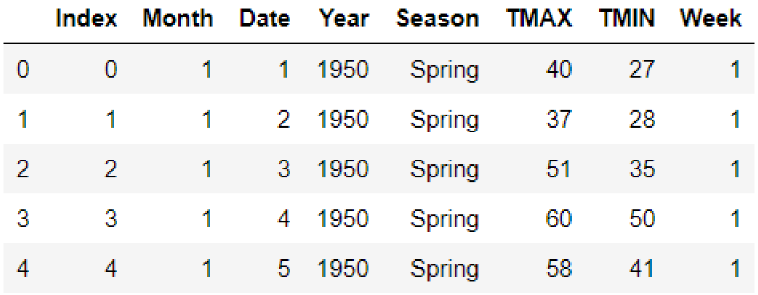


Figure 1. scatter matrix for correlation check

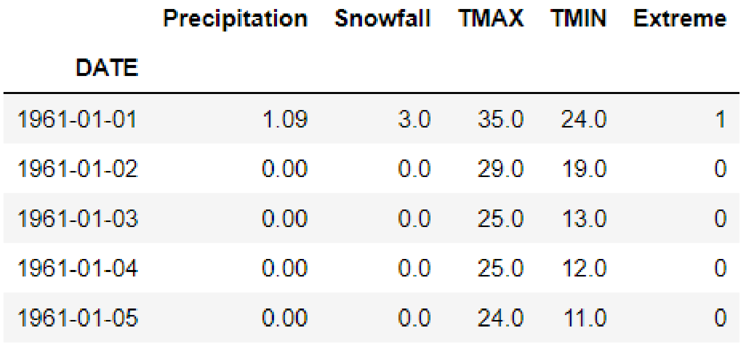
The missing values are either replaced or deleted due to their properties. For instance, if only one-day value is missing, we replace it with the next-day value, and if the values of the whole month are missing, we delete them without replacement.

The final cleaned version of our data for both research aspects is shown as follows (Figures 1 and 2).



	Index	Month	Date	Year	Season	TMAX	TMIN	Week
0	0	1	1	1950	Spring	40	27	1
1	1	1	2	1950	Spring	37	28	1
2	2	1	3	1950	Spring	51	35	1
3	3	1	4	1950	Spring	60	50	1
4	4	1	5	1950	Spring	58	41	1

Figure 2. clean data for the question about temperature



	Precipitation	Snowfall	TMAX	TMIN	Extreme
DATE					
1961-01-01	1.09	3.0	35.0	24.0	1
1961-01-02	0.00	0.0	29.0	19.0	0
1961-01-03	0.00	0.0	25.0	13.0	0
1961-01-04	0.00	0.0	25.0	12.0	0
1961-01-05	0.00	0.0	24.0	11.0	0

Figure 3. clean data for the question about extreme weather

3. Results

3.1. Does the trend of temperature change over the years?

In order to observe the trend of temperature over the years, we first visualized the temperatures by grouping it by days, months, and years. However, as Figure 4, 5, 6, and 7 imply, it is difficult to obtain the obvious trend through the statistical graph.

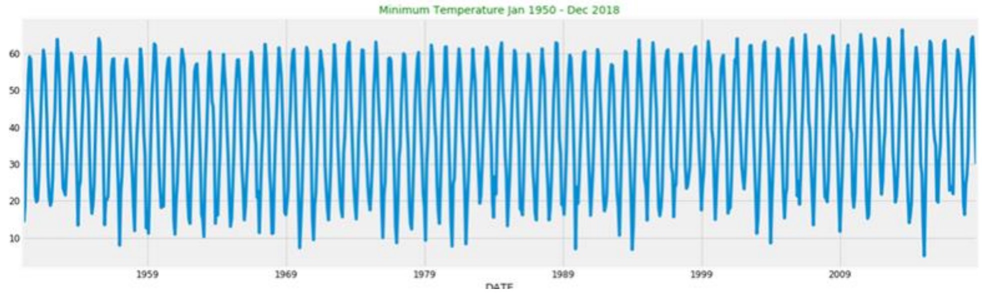


Figure 4. The minimum temperature from Jan 1950 to Dec 2018

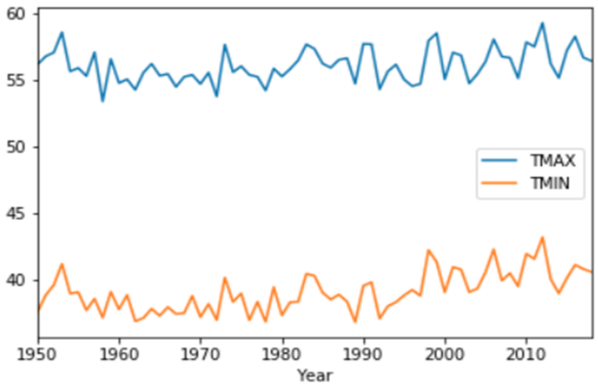


Figure 5. The mean temperature from Jan 1950 to Dec 2018 grouped by years

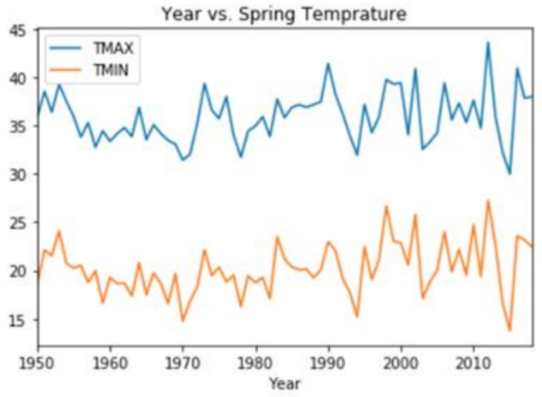


Figure 6. The mean spring temperature from Jan 1950 to Dec 2018

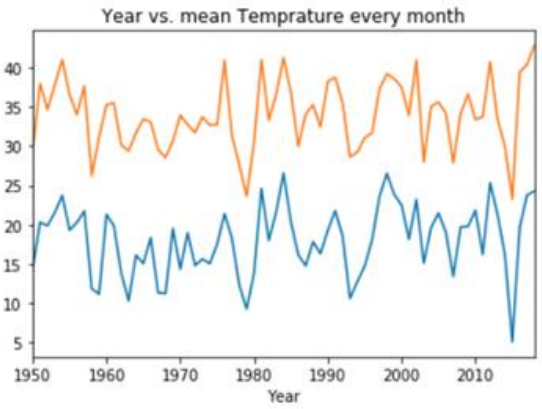


Figure 7. The mean January temperature from Jan 1950 to Dec 2018

Therefore, we decomposed the temperature over years as shown in Figure 8. Based on the result, we can observe that there is an upward trending over time, showing the temperature is slowly moving upward over the years.

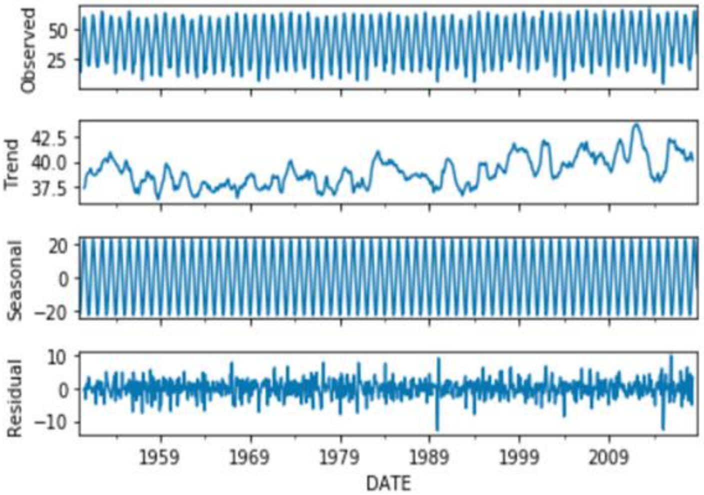
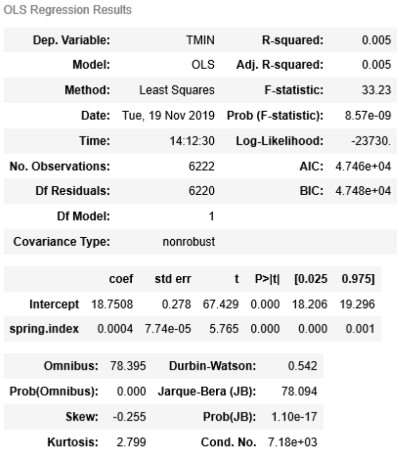


Figure 8. Data visualization for temperature and time

3.2. Is it possible for us to predict future temperature?

To tackle this question, we implement several models to find the best fit. Our first model is a linear regression model. However, since our data is seasonal and highly fluctuated, linear regression cannot explain our data. In Figure 9, the coefficient correlation is only 0.005, which means our model can explain only 0.5% of data. Meanwhile, Prob (F-statistic) is quite small, indicating that our model is not significant. We then switched the model and used one designed for non-stationary time-series. In this case, it is ARIMA. It turns out that the ARIMA model did a better job than linear regression. We used the Root Mean Square Error (RMSE) as of measurement. It is the standard deviation of the residuals and measures how residuals spread out. So, less RMSE, better prediction. From Figure 10, we found out that our RMSE is 3.796. To further optimize our result, we decide to implement SARIMA. Compare to ARIMA, SARIMA ended up with a better result as the model supports the direct modeling of the seasonal component. As shown in Figure 11, with the SARIMA model, our RMSE is 3.24, which is lower than ARIMA's.



OLS Regression Results

Dep. Variable:	TMIN	R-squared:	0.005
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	33.23
Date:	Tue, 19 Nov 2019	Prob (F-statistic):	8.57e-09
Time:	14:12:30	Log-Likelihood:	-23730.
No. Observations:	6222	AIC:	4.746e+04
Df Residuals:	6220	BIC:	4.748e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	18.7508	0.278	67.429	0.000	18.206	19.296
spring.index	0.0004	7.74e-05	5.765	0.000	0.000	0.001

Omnibus:	78.395	Durbin-Watson:	0.542
Prob(Omnibus):	0.000	Jarque-Bera (JB):	78.094
Skew:	-0.255	Prob(JB):	1.10e-17
Kurtosis:	2.799	Cond. No.	7.18e+03

Figure 9. Linear Regression summary table

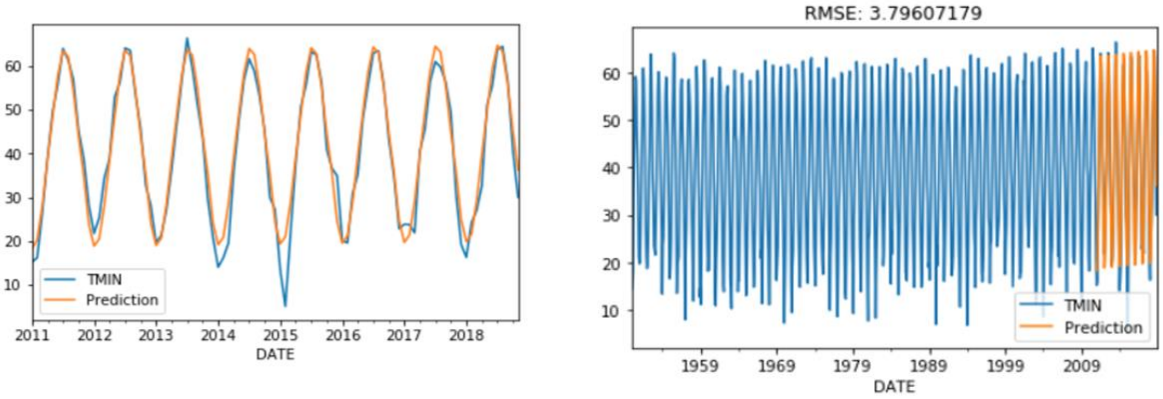
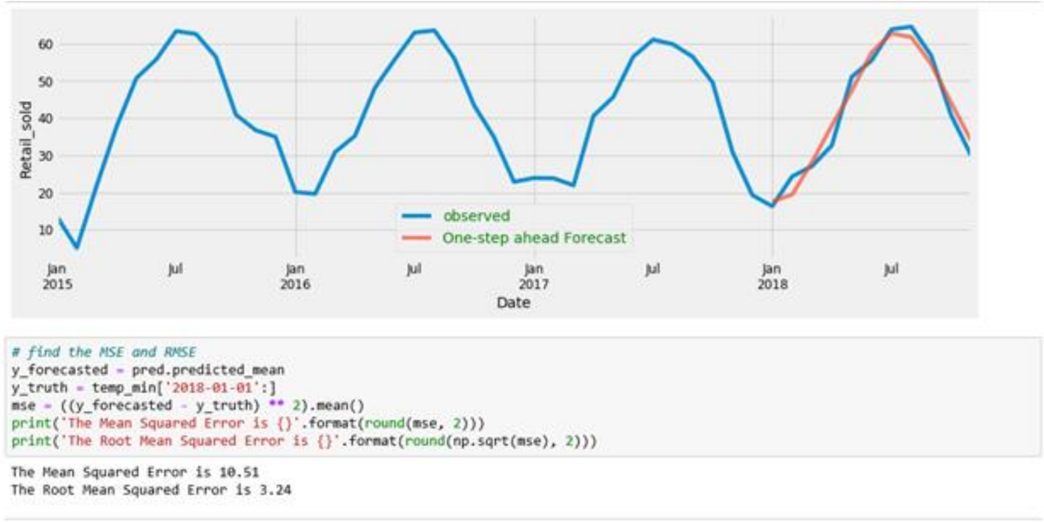


Figure 10. ARIMA model prediction and RMSE



```
# find the MSE and RMSE
y_forecasted = pred.predicted_mean
y_truth = temp_min['2018-01-01':]
mse = ((y_forecasted - y_truth) ** 2).mean()
print('The Mean Squared Error is {}'.format(round(mse, 2)))
print('The Root Mean Squared Error is {}'.format(round(np.sqrt(mse), 2)))
```

```
The Mean Squared Error is 10.51
The Root Mean Squared Error is 3.24
```

Figure 11. SARIMA model prediction and RMSE

3.3. Does extreme weather become more frequent?

Using data visualization as Figure 12 shown below, we found out that the number of extreme weathers varies over the years. There is no clear sign indicating the relationship between time and the number of extreme weathers.

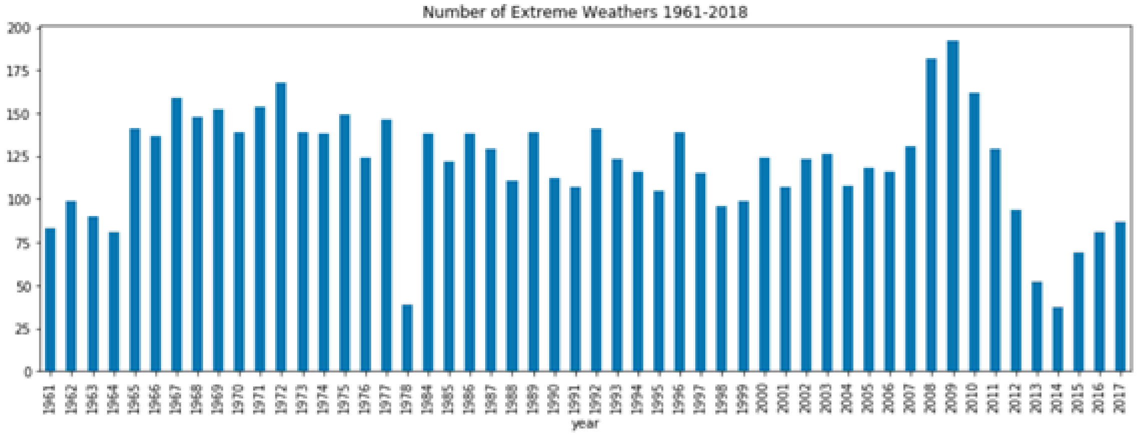


Figure 12. Extreme weathers' histogram over the years

3.4. Is it possible for us to predict extreme weather?

Appling different machine learning algorithms, our model's accuracy is as the table shows below. As shown in Table 1, our selected model accuracy ranges from 73% to 86%. SVM has 74% training accuracy and 73% testing accuracy. There are no overfitting or underfitting problems, as accuracy with cross-validation is 75%. Logistic Regression has 75% training accuracy and 76% testing accuracy. Accuracy with cross-validation is 76%, so no overfitting or underfitting for logistic regression. K-Nearest Neighbors had 82% training accuracy and 77% testing accuracy. However, accuracy with cross-validation is only 74%, indicating potential overfitting problems. Gaussian Naïve Bayes has 73% training accuracy and 74% testing accuracy with a 73% cross-validation accuracy, suggesting model fit data properly. Random Forest has the highest training accuracy and testing accuracy, but there is a gap between training accuracy and cross-validation accuracy, suggesting flows in model.

Model	**Accuracy on Training Set**	**Accuracy on Test Set**	**Accuracy with Cross-Validation (5 folds)**
Support Vector Machine	0.74	0.73	0.75
Logistic Regression	0.75	0.76	0.76
K-Nearest Neighbors	0.82	0.77	0.74
Gaussian Naïve Bayes	0.73	0.74	0.73
Random Forest	0.86	0.79	0.78

Table 1. Model performance for predicting extreme weather

4. Conclusion

Based on the modeling analysis and predictions of temperature and extreme weather in Worcester area in the past 50 years, as well as the evaluated performance of the different models, this report further conducts some conclusions about the weather conditions in Worcester.

From the analysis of temperature, we come up with the first conclusion that even though there is no significant change in the temperature, the average temperature has slightly increased over 50 years in Worcester.

The global surface temperature has risen for decades, the trend line of average temperature in Massachusetts is similar to this general trend, especially from 1993 to 2016 ("Climate Change | MEPHT," 2019), the annual average temperature in Massachusetts exceeded the average value in the 20th century. Meanwhile, based on the analysis of the data about Worcester we have analyzed, the city has been affected by the overall change of the state.

Meanwhile, we also solved the second research question using two models. Based on the ARIMA and SARIMA models, we are able to conduct temperature predicting. SARIMA has a more satisfying performance in our research, and we assume that it is because of the property of SARIMA that it is more fit for the seasonal data than ARIMA.

Analyzing numbers of extreme weather during the time given in the dataset, we found that there is no significant change in the number of extreme weathers due to time over 50 years in Worcester. In other words, there is no correlation between time and the number of extreme weathers.

Since we are predicting whether the extreme weather will happen nor not, we choose several classification models with great performance. Our selected models are as follows: Support Vector Machine, Logistic Regression, and Gaussian Naïve Bayes. All these models end up with satisfying accuracy scores, which are over 73%. So, we conclude that we can use machine learning to predict extreme weather.

In summary, based on our project, we conclude that it is possible for us to predict the future temperature using classic models. Moreover, it is possible to predict the future extreme weather by temperature, precipitation, and snowfall using machine learning models. The further recommendations focused on business problems will be provided based on the results from our project.

For the insured and potential customers, it is necessary to recognize the local climate change and be able to predict it. In the past few decades, the facts that some buildings and communities in Massachusetts were damaged by the historic snowstorm that might be caused by climate change has resulted in significant economic losses. According to a new study from the University of Waterloo, these losses could be avoided if the insurance industry considered climate change, and homeowners do not have to suffer financial loss in this case.

To recover the losses caused by the extreme weather, the insurance companies will have to raise interest rates or divest from high-risk areas, and when that happens, thousands of people will lose coverage or fail to afford it. On the contrary, some insurance companies understand climate change better than others, and these organizations will survive and may sell climate services to their peers who are trying to understand the problem. For example, to gain advantages in price competition and better calculate costs and gains and losses, insurance companies should protect consumers by looking forward to the future and protect consumers from accidents. Therefore, weather analysis and prediction in this area can make both the insured and the insurance companies better cope with and adapt to the financial risks related to climate change.

5. Limitations

Although from the visualization of model performance, we see that accuracy is considerable and the model works well, there is still a large room for improvements. At first, the unpredictability of weather causes that even with a perfect model and an understanding of the initial conditions, relatively accurate forecasts can still have deviations. In this research, we focus on the short-term forecast, but the long-term prediction can be conducted in the future.

Other classic models and Neural Networks can be applied for time-series problem to see if it is possible to get a more accurate prediction. Using daily data in a relatively short period rather than using monthly data can be considered as well. As for the classification problem in modeling, the larger dataset would be helpful for models to improve accuracy and applying Neural Networks could be one of the solutions for this issue. Also, the parameters can be adjusted to improve the performance.

Due to the tight schedule, there are only four independent variables involved in the project; we could find more independent variables for further research to conduct more comprehensive results.

6. References

Climate Change | MEPHT. (2019). Retrieved 4 December 2019, from https://matracking.ehs.state.ma.us/Climate-Change/index.html

National Climate Assessment. (2019). Retrieved 4 December 2019, from https://nca2014.globalchange.gov/highlights/report-findings/extreme-weather

Predictability limit: Scientists find bounds of weather forecasting. (2019). Retrieved 4 December 2019, from https://www.sciencedaily.com/releases/2019/04/190415154722.htm

New climate-change study focuses on extreme weather. (2015). *Physics Today*. doi: 10.1063/pt.5.028986

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L., & Hegerl, G. et al. (2017). Understanding, modeling, and predicting weather and climate extremes: Challenges and opportunities. Weather And Climate Extremes, 18, 65-74. doi: 10.1016/j.wace.2017.10.003