CS 760: Machine Learning Fall 2020

Machine learning algorithms, snowstorm, and climate change

AUTHORS: Zihao Zheng

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED OR MULTIPAGE.

ABSTRACT

Climate change has an increasing impact on people's production and life. Among many extreme meteorological conditions associated with climate change, snowstorm is typical and more influential among cities with lake regions. Even though climate experts have been making efforts in understanding the cause of snowstorm, the study involving making predictions, though critical, is limited. In this paper, we proposed four popular machine learning algorithms (logistic regression, support vector machine, random forest and naive Bayes) to predict the occurrence of snowstorm using a data set containing 13244 observations. Random forest, among those models, yield higher predictive performance, in no matter prediction accuracy and through ROC analysis. Limitations and potential future work is discussed in the last section.

1 Introduction and related work

With global climate change, many extreme weathers have become more common. One of the typical examples among those extreme weathers is snowstorm [Kunkel et al., 2002]. The extreme climate conditions result in what is called a snowstorm or sometimes even blizzard. Blizzards can kill, cause traffic accidents and bring cities to a halt, especially among cities with lake regions. The climatic conditions of the lakeside city (like heat and moisture fluxes from lake surfaces which result in precipitation [Niziol, 1987] and [Niziol et al., 1995]) make it more likely to have a snowstorm. On the one hand, the influential effect of snowstorm has been widely studied since early years [Kunkel et al., 2002, Braham Jr and Dungey, 1984, Schmidlin et al., 1992]. Also, there are many meteorological explanations and interpretations about the cause of the snowstorm [Cohen and Allsopp, 1988]. However, many literature suggests that the predictive study of snowstorm, and also the study of causal effect between snowstorm and other meteorological conditions is limited. Therefore, in this paper, we collect monthly meteorological data from different stations between 2000 to 2018 and use machine learning models to predict the chance of extreme snowstorms and also evaluate the prediction accuracy using appropriate methods.

2 Data set

2.1 Data availability

Following from many researches that target to analyze climate change, the Dane publiczne IMGW-PIB (https://dane.imgw.pl/ and https://dane.imgw.pl/data/) website provides rich data resources related to meteorological conditions. Fortunately, the R package `climate` [Czernecki et al., 2020] provides an interface to download those data sets.

We collect and summarize monthly data from 2000 to 2019. Each data was observed from climate meteorological stations, such as "BALIGRÓD-MCHAWA", "BARWINEK", etc. The data has 13244 observations and 7 different meteorological variables (besides those geographic information variables). One of the variables that we consider as response (y) is "snowstorm", indicating that whether the surveying station has had a snowstorm in a period of time. The other 6 independent variables are listed as the following:

1. Insolation: total monthly insolation [hours].
2. Sleet: total monthly days of sleet [days].
3. Hail: total monthly days of hail [days].
4. Fog: total monthly days of fog [days].
5. Pressure: monthly mean pressure at station level [hPa].
6. Seapressure: monthly mean sea level pressure [hPa].

In the following chapter, we will use those 6 meteorological conditions to predict the dummy variable indicating the presence of snowstorm.

2.2 Data summary

The following table Table 2.2 shows summary statistics for those six variables used to predict snowstorm. For the response variable snowsnorm (y), it is a dummy variable where 23.2% of observations are 1. Since we will evaluate the prediction accuracy in the following chapter therefore it is worth mentioning that this is a heavily unbalanced response variable. Specifically, trivially predicting all observations as 0 will yield a prediction accuracy larger than 76.7 % even though the sensitivity is low. We will discuss further in the next section.

3 Statistical approach

In this section, we describe the statistical approach that we apply for predicting the snowstorm. In order to evaluate the result and better avoid over fitting issue, we randomly splitted the data (13244 observations) into training data set (75 %

	mean	standard deviation
insolation	130.59	99.74
sleet	1.10	1.88
hail	0.22	0.59
fog	4.82	5.50
pressure	976.61	115.63
seapressure	998.77	182.41

Table 1: Summary of variables used to predict snowstorm

of the observations, 9933) and the test (validation) data set (25 % of the observations, 3311). For training the data, we consider the following four different models.

1. Logistic regression: the logistic regression for binary response (Bernoulli likelihood) is applied on the training data set. LASSO and elastic net (using R package `glmnet`) are also considered to penalized the model complexity. Since the number of variables is limited, the benefit of shrinkaged algorithm is not apparent.
2. Support vector machine.
3. Random Forest: a tree based algorithm that is more robust than a single decision tree (for example, the `CART` algorithm).
4. Naive Bayes: assuming independency among predictors, the Naive Bayes algorithm calculates the posterior prediction based on each variable. Since all variables are numerical, the naive Bayes algorithm relies on the Gaussian sampling model. It is worth mentioning here that the assumption for the naive Bayes algorithm might be challenged here which will result in the unreliable interpretation.

The above algorithm could be implemented by the codes in the homework during this semester. Also, `Rstudio` also provides packages (for example, `rminer` [Cortez, 2020]) to implement those machine learning algorithms. The next section summaries key results of model fitting and comparison among those machine learning algorithms.

4 Results

First, we will report the prediction accuracy as well as some detailed statistics (false negative and false positive as metrics to evaluate specificity and sensitivity) as the following table 4.

we can make several remarks regarding the predicting accuracy:

1. All models yield higher prediction accuracy compared to a trivial guess (predicting all y's as 0).
2. Among those four machine learning models, random forest has the highest prediction accuracy where the naive Bayes predicts relatively poorly compared to others.
3. The sensitivity of support vector machine is the highest however its specificity is not satisfied compared to others.

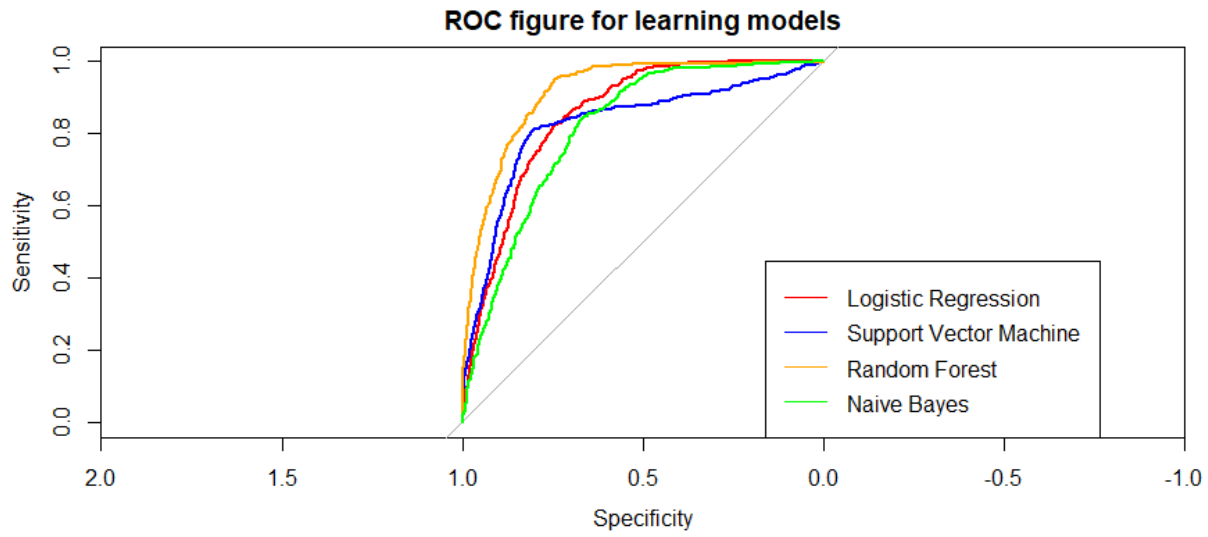

Figure 1: ROC curve of four machine learning models

4. The false negative rate of random forest is smaller than the other three models, which is the key reason resulting in the highest prediction accuracy.

learning model	accuracy	false positive	false negative
Logistic regression	82.97 %	4.02 %	13.02 %
Support vector machine	83.60 %	2.27 %	14.13 %
Random forest	87.07 %	4.05 %	8.85 %
Naive Bayes	80.43 %	7.82 %	11.75 %

Table 2: Comparison statistics among four machine learning models

The above comparison strategy is less informative since for all the prediction algorithm, we could report the probability for the snowstorm, rather than a single $0-1$ prediction. As an alternative, ROC analysis provides more information for the performance of those models. Figure 1 shows the ROC curve, sensitivity vs specificity among all four different models.

It is clear that the orange line, representing random forest, is above than the other three lines, which agrees with our previous observations that random forest provides the highest prediction accuracy than the others.

Also, a key statistic (metric) for ROC analysis is the AUC value (the size of area under the ROC curve, the larger the better). Table 4 shows the estimated value of AUC and its confidence interval (based on DeLong test) under the level of 5%. Those calculation could be implemented through `Rstudio` using package `pROC` [Robin et al., 2011]. It is clear that random forest is significantly better, in terms of AUC score, than all the other three competing models, under the level of 5% based on DeLong nonparametric test.

	lower bound	predicted value	upper bound
Logistic regression	0.84	0.85	0.87
Support vector machine	0.80	0.82	0.84
Random forest	0.90	0.92	0.93
Naive Bayes	0.80	0.81	0.83

Table 3: AUC prediction and its confidence interval under the level of 5%

5 Conclusions and future work

Through this paper, we discussed machine learning models using meteorological data to predict snowstorm, an important but extreme weather condition that affects heavily on our daily economy and life. Among all four models (Logistic regression, Support vector machine, Random forest and Naive bayes), we found random forest is better significantly than the others. Using random forest model, we can achieve prediction accuracy around 87 % where the false positive rate is only 4%. Also, the key features (important variable) selected by random forest model, agrees with the literature findings [Kunkel et al., 2002, Braham Jr and Dungey, 1984].

Besides the obvious benefit of machine learning models in predicting the occurrence of snowstorm and other climate meteorological conditions, we would also like to discuss on the limit of the current work, which might motivates potential future directions.

First, compared to the number of observations, the number of variables is small. This does not necessarily result in an severe issue but leave us more space to dig further, if a more comprehensive data set containing more variables is available.

Second, even though we have used strategies to avoid over fitting issue, such as penalizing model complexity and using cross validation, the bias-variance trade-off is still a big concern. Also, the dependency among predictors would also be worthwhile to consider further, with more knowledge of climate of meteorological condition.

Finally, the prediction accuracy of machine learning models is impressive, but the interpretability of them is limited. Motivated by large sample size, using random forest, or even neural network, might be popular in predicting benefit. However, people still also have to admit that those model is hard to interpret the mechanism. This viewpoint, sometimes names as accuracy-interpretability trade-off, is always highlighted in some machine learning literature [James et al., 2013, Murdoch et al., 2019].

References

R. R. Braham Jr and M. J. Dungey. Quantitative estimates of the effect of lake michigan on snowfall. *Journal of Climate and Applied Meteorology*, 23(6):940–949, 1984.

S. Cohen and T. Allsopp. The potential impacts of a scenario of c02-induced climatic change on ontafio, canada. *Journal of Climate*, 1(7):669–681, 1988.

P. Cortez. *rminer: Data Mining Classification and Regression Methods*, 2020. URL `https://CRAN.R-project.org/package=rminer`. R package version 1.4.5.

B. Czernecki, A. Glogowski, and J. Nowosad. *Climate: An R Package to Access Free In-Situ Meteorological and Hydrological Datasets For Environmental Assessment*, 2020. URL `https://github.com/bczernecki/climate/`. R package version 0.9.1.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

K. E. Kunkel, N. E. Westcott, and D. A. Kristovich. Assessment of potential effects of climate change on heavy lake-effect snowstorms near lake erie. *Journal of Great Lakes Research*, 28(4):521–536, 2002.

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

T. A. Niziol. Operational forecasting of lake effect snowfall in western and central new york. *Weather and Forecasting*, 2(4):310–321, 1987.

T. A. Niziol, W. R. Snyder, and J. S. Waldstreicher. Winter weather forecasting throughout the eastern united states. part iv: Lake effect snow. *Weather and Forecasting*, 10(1):61–77, 1995.

X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

T. W. Schmidlin, D. J. Edgell, and M. A. Delaney. Design ground snow loads for ohio. *Journal of Applied Meteorology*, 31(6):622–627, 1992.