Will Guangpu Wu
Apr. 7, 2024

Prediction on Precipitation with GLM
What Factors Could be Helpful in Predicting Rainfall

1. INTRODUCTION

Weather forecasting has been important in various fields, from agricultural production to daily life activities. Historically, people around the world have developed various methods to forecast weather. However, the traditional techniques are often regional or not precise. This underscores the importance of constructing prediction models upon empirical data. Many studies perused methods such as generalized linear model (GLM), time series forecasts, and stochastic simulations, highlighting key factors for predicting precipitation such as temperature, humidity, and ratio of snow falls. (Barooti, Esmaili, & Ghahraman, 2020) (Zhang, Vincent, Hogg, & Niitsoo, 2000).

This study aims to build a user-friendly GLM for practical everyday applications, producing binary prediction on whether it will rain, with easily-obtainable measures like humidity and temperature. The model is built upon a dataset collected on Toronto, with a 10-year range from 2009 to 2019. (Vivas, 2020).

Will Guangpu Wu
Apr. 7, 2024

2. METHODS

2.1 Data Preparation

The dataset used consists of a variety of atmospheric conditions. It is imported and canned to remove missing values. Variables derived from this dataset include: "is_winter," which flags days occurring during the winter months; "snowed," indicating days where snowfall was recorded, as inspired by the factor of snow fall ratio mentioned above. And most importantly, the binary response variable of interest "rained", marking days on which rainfall was observed, is derived from a variable measuring precipitation in mm.

2.2 Choice of Method

Logistic regression is used since the outcome variable, whether it rains on a day, is binary. The GLM is in the form of:

$$\log\left(\frac{\pi}{1-\pi}\right) = X\mathrm{B}$$

Where π indicates the chance of rain; X denotes the predictor variables and B indicates their corresponding coefficients.

2.3 Model Selection

For the selection of the most predictive variables for the model, this study employed multiple approaches, including stepwise selection based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Lasso regression methods.

AIC and BIC were both used to balance the model's complexity with predictive power. AIC focuses on the goodness of fit and BIC penalizes complex model to prevent overfitting. Their expressions are as follows, where $\hat{L}$ maximum value of the likelihood function for the model; n is the sample size; and p is the number of predictors in the model.

$$AIC = 2p - 2\log(\hat{L})$$
$$BIC = \log(n)\,p - 2\log(\hat{L})$$

Lasso regression, on the other hand, is a shrinkage method, it enhances model's interpretability and accuracy by penalizing less influential predictors, reducing their coefficients to zero.

The process started with the full model, in which all covariates are fitted. The full model is then proceeded to stepwise selections and shrinkage selection, keeping only a subset of the full model. The three candidate models, selected respectively by these three methods, are proceeded

to comparison and refine. Statistically insignificant variables (p-value <0.05) are removed because they are less influential to the prediction.

2.3 Model Diagnostics and Validation

The assumptions of logistic regression are checked. Firstly, the binarity of response variable has already been assured during data preparation. Secondly, to address multicollinearity, the Variance Inflation Factor (VIF) was assessed for each predictor in the model, and those with high VIF values were removed, ensuring the predictors were independent.

Although the Logistic models do not require perfect normality, the Normal Q-Q Plots is plotted to helps identify outliers. The outliers identified in this step are removed to improve prediction accuracy.

Predictive performance was further evaluated through a 10-fold cross validation (CV), where the data is split into 10 sets. For each set, the model was trained on 9 sets and tested on the remaining set, ensuring that every data point had a chance to be validated. It uses two key diagnostic tools: the Receiver Operating Characteristic (ROC) curve and the calibration curve. The ROC curve illustrates the diagnostic ability of a binary classifier. It plots true positive rate (sensitivity) against the false positive rate (1 - specificity). An area under the ROC curve (AUC) closer to 1 indicates excellent model performance to differentiate between the binary outcomes. The calibration curve plots the predicted probability of rainfall against the observed outcomes. An ideal calibration curve closely aligns with the 45-degree line, in which the predicted probabilities match the observed outcomes.

Will Guangpu Wu
Apr. 7, 2024

3. RESULTS

3.1 Key variables in the Dataset

The cleaned dataset contains 3,954 observations and 6 predictors after removing outliers. The following are a summary of key variable from the data set. Visualizations of these variables are presented in *Figure 4* and *Figure 5* in the appendix.

Table 1: Summary of the Key Variables in the Data Set

	Mean	Standard Deviation	Median	Minimum	Maximum	Range
Median of temperature Celsius. **"medTempC"**	8.6062	10.0816	8.5	-21.5	29	50.5
Cloud Coverage Rate (%) **"cloudcover"**	44.5706	27.9090	43	0	100	100
Log of visibility in km **"log_visibility"**	2.2395	0.1384	2.3026	0.6931	2.9444	2.2513
Atmospheric pressure in millibars **"pressure"**	1016.1297	7.2538	1016	990	1041	51
Humidity (%) **"humidity"**	77.5781	10.0534	78	37	100	63
Dummy variable, when =1 indicates the day is in winter months **"is_winter"**	0.3340	NA	NA	0	1	1
The response variable. Dummy variable, when =1 indicates a rain on the day. **"rained"**	0.5429	NA	NA	0	1	1

Will Guangpu Wu
Apr. 7, 2024

3.2 The Models

Table 2: Summary of All Candidate Models, including the Final Model

	Models			
	Final Model	AIC Selected	BIC Selected	Lasso Selected
	(1)	(2)	(3)	(4)
medTempC	0.0947***	0.1083**	0.0322***	0.0381***
	(0.0078)	(0.0466)	(0.0105)	(0.0110)
cloudcover	0.0521***	0.0506***	0.0531***	0.0518***
	(0.0026)	(0.0027)	(0.0025)	(0.0029)
windspeedKmph		-0.0419**	-0.0598***	-0.0409**
		(0.0192)	(0.0170)	(0.0198)
log_visibility	-14.8581***	-5.3166***	-5.5069***	-5.3299***
	(0.9855)	(0.5388)	(0.5323)	(0.5398)
pressure	-0.0335***	-0.0173**	-0.0206***	-0.0166**
	(0.0076)	(0.0071)	(0.0069)	(0.0073)
humidity	0.0292***	0.0258**		0.0110**
	(0.0054)	(0.0110)		(0.0054)
is_winter1	-0.6015***			
	(0.1532)			
DewPointC		-0.0735		
		(0.0470)		
sunHour				0.0066
				(0.0209)
uvIndex		0.3319***	0.3558***	0.3284***
		(0.0646)	(0.0638)	(0.0667)
WindGustKmph		0.0300**	0.0419***	0.0309**
		(0.0124)	(0.0113)	(0.0126)
winddirDegree				0.0002
				(0.0006)
moon_illumination_percent				0.0005
				(0.0012)
Intercept	62.8152***	23.7562***	29.6681***	24.2718***
	(8. 2961)	(7.6060)	(7.2180)	(7.7808)
N	3,954	4,048	4,048	4,048
Log Likelihood	-1,660.1030	-2,017.5950	-2,020.7940	-2,018.5990

VIF of Final Model:

medTempC	cloudcover	log_visibility	pressure	humidity	is_winter1
3.329855	2.105633	1.147315	1.216120	1.235277	2.545367

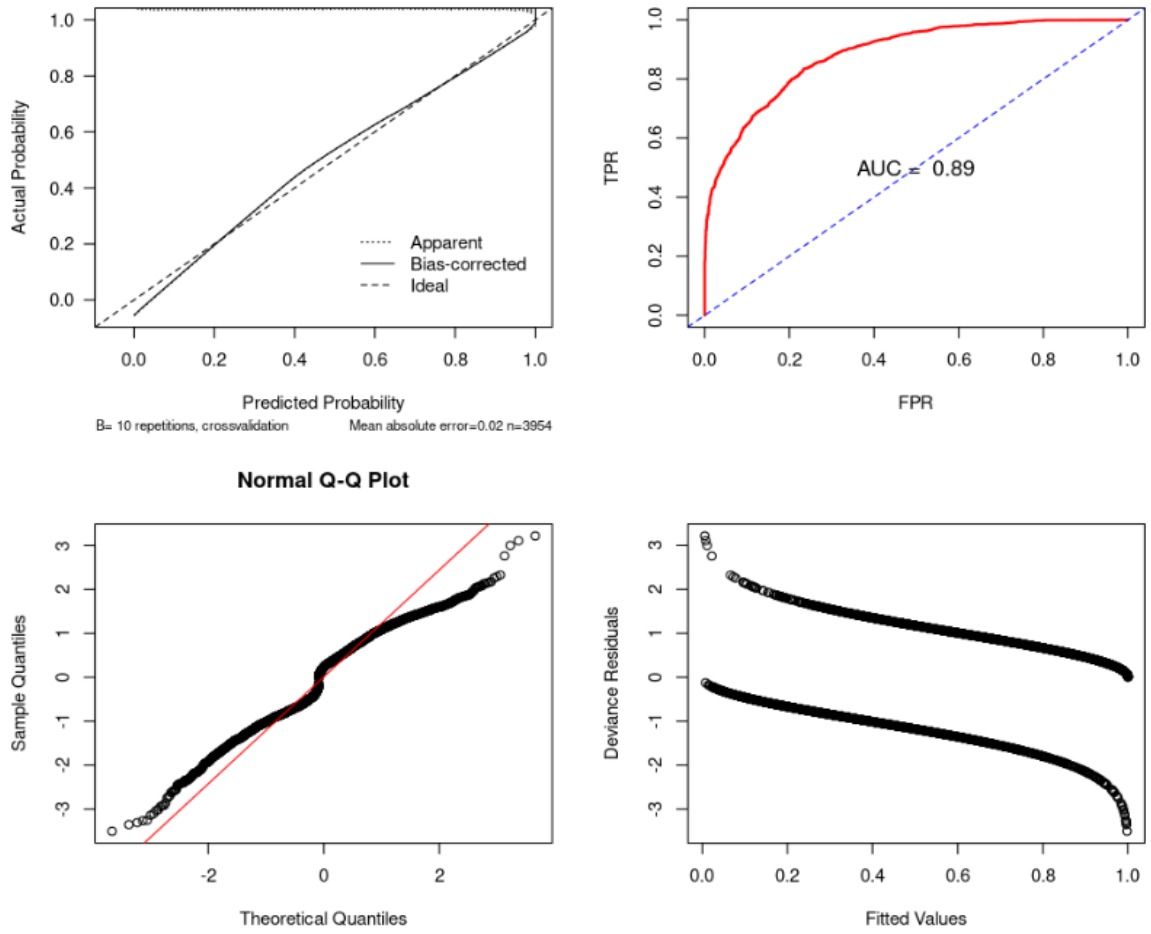
Will Guangpu Wu
Apr. 7, 2024

```
================================================================================
Notes:                                  ***Significant at the 1 percent level.
                                         **Significant at the 5 percent level.
                                          *Significant at the 10 percent level.
```

Each candidate model is diagnosed and compared. The final model demonstrated the least multicollinearity among all candidates, as indicated by very low VIFs (all less than 5). The dummy variable indicating snow ("snowed") is removed from the model because of insignificance (p-value>0.05), although the literacy research suggests it as a key factor in prediction for precipitation.

3.3 Diagnostics and Validation of Final Model

Figure 3: Calibration Plot, ROC, Noraml-QQ plot, and Deviance Residuals of the Final Model



The normal QQ Plot are not aligned with the 45-degree line, which is acceptable in the case of logistic model. The validations demonstrate a good quality of fit, and a high ability of the model to discriminate between the binary outcomes of rain and not rain. The calibration plot in *Figure 3* indicates that the ideal model and the bias-corrected model are relatively close to the

perfect calibration, which represents a good match between predicted probabilities and actual outcomes. This suggests that the predicted probabilities of rainfall are consistent with the observed frequencies. The ROC curve demonstrates a strong model performance with an AUC of 0.89, significantly higher than 0.5. This indicates the model's effectiveness in classifying the positive (rain) outcomes correctly.

Will Guangpu Wu
Apr. 7, 2024

4. DISCUSSION

4.1 Model Interpretation

In the final model displayed above, the positive coefficient for median temperature (medTempC), is 0.0947, which suggests that for each $1°C$ increased in median temperature, there is a 9.47% increase in the odds of rainfall occurring, holding other variables constant. Such interpretation is similar for other numerical variables. The coefficient for categorical variable 'is_winter' being -0.6015 indicates that holding other variables constant, the odds of rainfall are $\exp[-0.6015] - 1 = 45.2\%$ lower during winter. These findings align with the understanding that certain temperature thresholds can contribute to precipitation, and that rainfall is significantly less during winter. Given the model's strong predictive performance using readily obtainable factors, it can be reasonably expected to be useful for everyday applications.

4.2 Limitations

The model's limitations include potential unobserved effects not captured by the predictors used. Incorporating Generalized Linear Mixed Models (GLMM) could address this by accommodating random effects and unmeasured variables. Additionally, the dataset which is Toronto-specific may limit generalizability. To enhance the model's applicability, future studies should consider using a more geographically diverse dataset.

(word count: approximately 1280 words excluding references, table and figure captions)

Will Guangpu Wu
Apr. 7, 2024

REFERNCES

1. Barooti, H., Esmaili, K., & Ghahraman, B. (2020). Stochastic simulation of daily rainfall using generalized linear models in semiarid simulation of rainfall. Journal of Climate Research, 1398(37), 1-20. https://clima.irimo.ir/article_103401_en.html

2. Vivas, L. (2020, April 27). Weather North America. Kaggle. https://www.kaggle.com/datasets/luisvivas/weather-north-america?select=toronto.csv

3. Xuebin Zhang, Lucie A. Vincent, W.D. Hogg & Ain Niitsoo (2000) Temperature and precipitation trends in Canada during the 20th century, Atmosphere-Ocean, 38:3, 395-429, DOI: 10.1080/07055900.2000.9649654 https://www.tandt0online.com/doi/abs/10.1080/07055900.2000.9649654

Will Guangpu Wu
Apr. 7, 2024

APPENDIX

Figure 4: Histogram for numerical predictors, and bar-plot for "snowed"

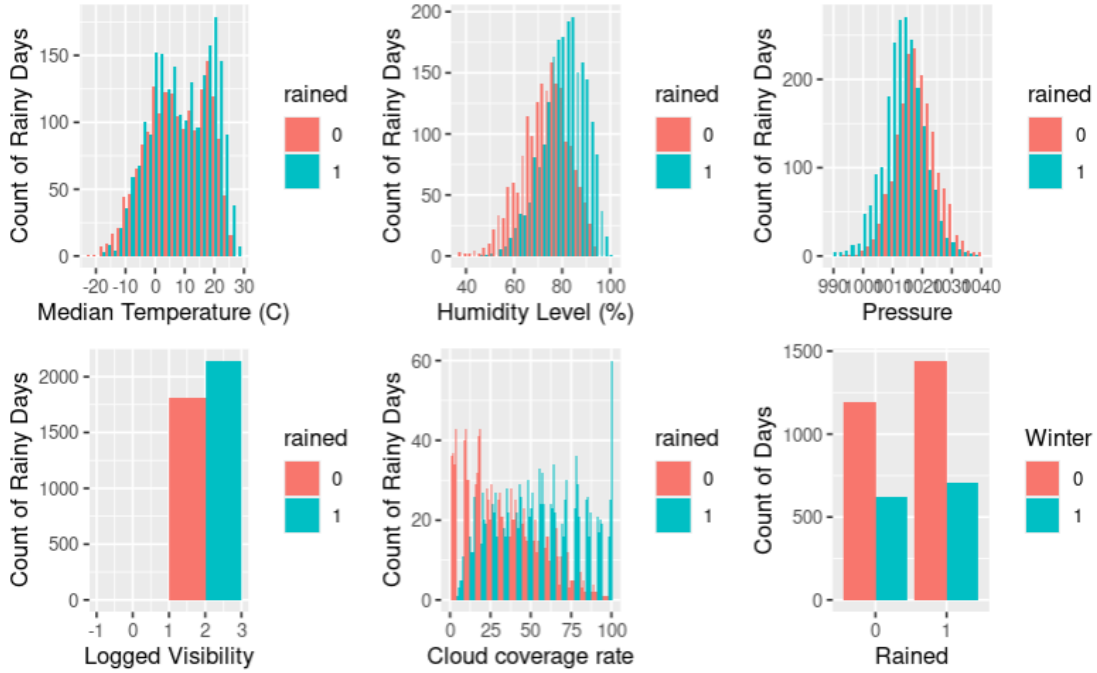


Figure 5: Boxplot for numerical predictors