Final Report: Rainfall Prediction

BIOS 635

Kevin Sullivan

Executive Summary

The objective of this project was to identify an accurate yet parsimonious model to predict rainfall using easily obtainable weather metrics. Accurate weather forecasting is crucial for emergency preparedness, public health resource allocation, and individual daily lives. To measure effectiveness, the model with a low misclassification error and sparse number of predictors was preferred. Over 800,000 models were fit from 8 different machine learning methods. The final preferred model—logistic regression with 5-fold cross-validation—used just three predictors and achieved a 10.5% error rate with a strong Kaggle score of 0.896. This model demonstrates how machine learning can support timely, effective public health and emergency planning decisions.

Problem, Dataset Overview, Feature Engineering

According to the National Oceanic and Atmospheric Administration, 90% of American counties have experienced a weather disaster in the past decade.[1] One of the most dangerous weather events is flash flooding.[2] Flash flooding can occur quickly and can have significant public health effects.[3] These effects can be mitigated through effective rainfall prediction. Correct weather forecasts can give people time to seek shelter, time that can save lives. Furthermore, weather events such as heavy rainfall are linked to numerous diseases and can have a significant public health impact.[4] Inaccurate weather models can directly cost lives and burden the public health system. Creating accurate and timely models that can predict rainfall on a given day should be a priority for both meteorologists and the public health community.

Through this project, I evaluate the efficacy of different machine learning methods in predicting the rainfall status of a given day based on certain weather metrics. Specifically, the challenge was to predict the binary variable *rainfall* given historical weather data. This project was chosen because it uniquely blends public health and meteorology. Every day, millions of people across the world make decisions based on a weather forecast. Whether someone brings an umbrella to work, whether someone has their birthday party outside, or whether a hospital approves life flights for a given day all depend on the weather and thus weather predictions. Inaccurate predictions can have significant economic, health, and well-being effects. Creating an accurate and easy-to-implement model that can guide decisions ranging from routine to critical can have significant public health and emergency management benefits.

All coding for this project was done using R version 4.4.3. The project idea and data are from the Kaggle competition titled **Binary Prediction with a Rainfall Dataset,** which was found on the list of exciting Kaggle competitions shared with the class and was started on March 25th.[5]

The dataset used for this report was generated by Kaggle using a deep-learning model to be used in a competition setting. There were 2,190 observations and 13 base features. Three additional features were engineered for a total of 16. There was no missing data or outliers present. The key outcome feature was *rainfall,* a binary variable that indicates whether or not it rained on a given day. The other 15 features were either categorical or continuous weather measures such as cloud coverage, wind speed, and temperature. A complete breakdown of all features can be found in **Table 1.** The goal of this project was to train an effective and pragmatic model that can accurately predict whether or not it will rain on a given day based on certain weather metrics. For this project, effectiveness was defined as a low misclassification error rate, and pragmatic was defined as having the lowest number of features.

Feature engineering was undertaken to provide a more nuanced and interesting approach to this binary classification problem. Three features were created. The first was a measure of dewpoint depression. This was calculated by subtracting the dewpoint for each day from the temperature of the same day. This provides a more nuanced metric than just temperature or dewpoint alone. Dewpoint depression specifically measures air saturation which is a key predictor of rainfall.[6] The second feature that was created was temperature range. This was calculated by subtracting each day's minimum temperature from its maximum temperature. This provides a more robust measure and scholarly research indicates that rainy days tend to have a smaller temperature range than non rainy days.[7] Lastly, a variable that measures both humidity and cloud coverage was created by multiplying the two features together. Days with rainfall tend to be both humid and cloudy.[8] This variable was created to capture this. All feature engineering was grounded in scholarly research, and a thorough review of these decisions can be found in the **appendix.**

Methods

Predicting the rainfall status of a given day based on weather metrics is an example of a binary classification problem. Because of this, various machine learning methods were evaluated, including logistic regression, k-nearest neighbors, generative classification models (QDA, LDA, Naive Bayes), methods for model optimization such as 5-fold and 10-fold cross-validation, model subset selection, model shrinkage methods, random forest, and XG Boost. The methods were split into two groups: iterative methods and full feature set methods.

To ensure a robust and thorough approach, four methods utilized an iterative function written in R by the author to fit every possible combination of features. Since there were thirteen predictor variables used (since ID and Day were dropped and Rainfall was the response variable), this amounted to $\sum_{i=1}^{13} \binom{13}{i} = 8,191$ models. The four methods that utilized this iterative approach were logistic regression, kNN, generative classification models (LDA, QDA, Naive Bayes), and resampling methods (logistic regression with 5 and 10-fold CV). The original training dataset was split 70/30 into training and validation sets, with the training data used to fit the models and the validation set used to test their accuracy. For the kNN method, k varied from 1 to 100. The misclassification error was computed for every model, and the model with the lowest error for each method was considered the best.

Other techniques did not require the same level of computational complexity. For these models, the full training set with 13 predictors was used. These methods include stepwise selection, model shrinkage methods such as Lasso and Ridge, random forest, and XG Boost. For model shrinkage methods, $\boldsymbol{\lambda}$ varied from 10^{-4} to 10^{10} and 100 trees were used for the random forest model. The misclassification error was then calculated for each model, and the model with the lowest error was considered the best. A breakdown of all methods used can be found in **Table 2.**

The model and associated relevant parameters with the lowest misclassification error rate and the fewest number of predictors were considered the final preferred model. Misclassification error was chosen as the main unit of analysis because it is the simplest to explain and captures an important requirement of an effective machine learning model: accuracy. Furthermore, models with a low number of predictors were preferred to make implementation easier for governmental and public health specialists. An accurate model that requires fewer inputs is both practical and efficient, which in turn can lead to more effective public health interventions. Lastly, rainfall status was predicted using the test data for each of the “best” models identified for each ML method, as defined by the criteria outlined above. These predictions were submitted to Kaggle, and the associated score (area under the ROC curve between the predicted probability and the observed target) was computed and served as a further evaluative criterion.

Results

A total of 868,251 models across 8 different machine learning methods were run. The 12 best models, defined as having the lowest misclassification rate for each method, were then used to predict rainfall status from the test set (**Table 3)**. The models ranged from 3-16 predictors, with misclassification errors ranging from 0.102 to 0.355. Most models had a misclassification error between 11% and 14%. The 5 and 10-fold CV methods yielded the models with the lowest misclassification error, while the logistic regression, kNN, and 5-fold CV models performed strongly on Kaggle.

The logistic regression model using 5-fold CV with three predictors (*pressure, sunshine, humid_cloud)* was chosen as the preferred model. This model had a misclassification rate of 0.105, which was the second lowest of the other methods. It also scored reasonably well on Kaggle, accruing a score of 0.896, which was not far from the competition leaders’ score of 0.906. The 10-fold CV model with six predictors was the most accurate, with a misclassification rate of 0.102. This model also performed well on Kaggle but was slightly behind the 5-fold CV model. Although this model had a lower misclassification error, it had three more predictors than the 5-fold CV model. Since the misclassification rate and Kaggle score for the two models were similar, and the 5-fold CV model was more parsimonious, it was ultimately chosen as the preferred final model.

Evaluation and Discussion

Predicting the existence of rainfall on a given day with weather data is a critical issue in the field of public health because it supports crucial planning initiatives. Out of the 8 ML methods, 4 utilized a systematic iteration process to fit all combinations of predictors, while the other 4 used the full feature set. The logistic regression model using 5-fold CV with the predictors *pressure, sunshine,* and *humid_cloud* was the second most accurate model and had the smallest number of features. It also scored the best on Kaggle. Because of this, it was chosen as the preferred final model. These features are all commonly collected in a weather setting and are minimally invasive to meteorologists, governmental officials, and public health officials. This makes this model pragmatic. It can be run without placing a burden on weather experts, emergency management officials, or the average person, and thus has the potential to lead to accurate, effective, and actionable predictions.

Some ML methods, such as kNN, LDA, QDA, and Naive Bayes, were relatively effective, while others such as Ridge, XG Boost, and Random Forest were not effective. This could be due to some confounding effect as well as the inherent complexity and randomness that accompany rainfall prediction. All of the models that were run had between a 10% and 35% misclassification error. This is a wide range of error rates and may be due to the complexity of the task as well as the techniques implemented. While systematic and grounded in public health and meteorological literature, it may not have captured the full nuances present in the data. Similarly, numerous factors influence whether or not it rains that go far beyond the scope of this data. Broad weather data alone can only tell specialists so much, and this is reflected in the error rates found. As meteorologists like to say, it may not rain, but it does not hurt to bring an umbrella.

Future Work

Predicting the rainfall status of a day based solely on weather metrics is a complex and imperfect task. No human can control the weather, and no model will always get it right. However, it is an important and beneficial task to continue to try and build more accurate rainfall prediction models. Novel approaches, such as neural networks or support vector machines, could be implemented in future work. Predicting the weather is an inherently messy and complex task that is subject to randomness. Because of this, any future work should combine ideas from many fields, including public health, meteorology, and data science. This interdisciplinary approach is needed not only to create more accurate ML models but to provide a more nuanced approach that can benefit people from all walks of life.

Appendix

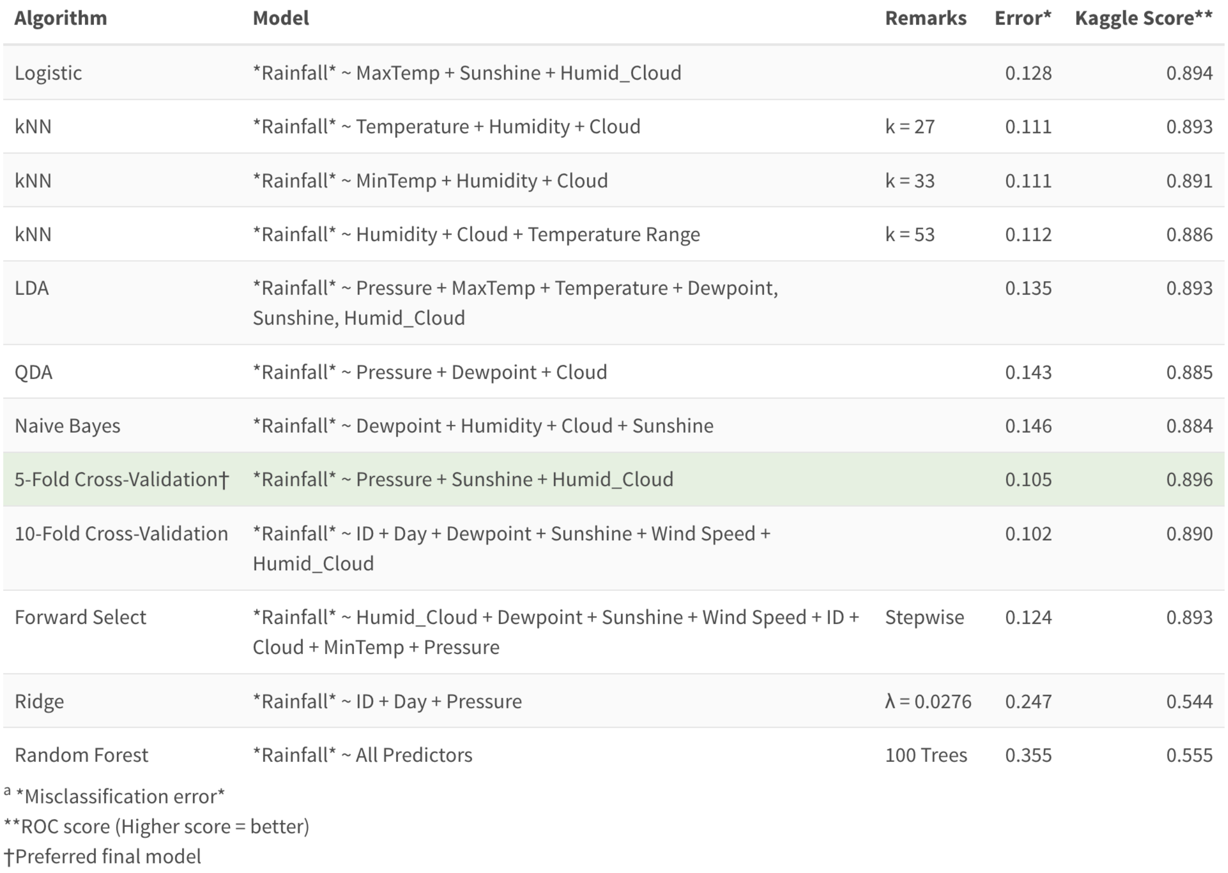
Table 1: Summary of Features

Feature	**Type**	**Notes**
Rainfall	Binary (Response)	**Kaggle**
ID	Categorical	**Kaggle**
Day	Cateogrical	**Kaggle**
Pressure	Continuous	**Kaggle**
Max Temperature	Continuous	**Kaggle**
Temperature	Continuous	**Kaggle**
Min Temperature	Continuous	**Kaggle**
Dew Point	Continuous	**Kaggle**
Humidity	Continuous	**Kaggle**
Cloud	Continuous	**Kaggle**
Sunshine	Continuous	**Kaggle**
Wind Direction	Continuous	**Kaggle**
Wind Speed	Continuous	**Kaggle**
Dew Point Depression	Continuous	**Engineered:** Temperature - Dewpoint
Temperature Range	Continuous	**Engineered:** Max Temp - Min Temp
Humidity X Cloud	Continuous	**Engineered:** Humidity * Cloud

Table 2: Methods of Machine Learning Evaluated

Method	Package/Function	Total Models Evaluated	Notes and Remarks
Logistic Regression	Base R, 4.4.3 *glm,* *family = "binomial"*	$\sum_{i=1}^{13}\binom{13}{i} = 8,191$	Training and Validation sets used
K-Nearest Neighbor	Caret *knn3*	$\sum_{i=1}^{13}\binom{13}{i} * 100 = 819,100$	Training and Validation sets used K varied from 1 to 100
Generative Classification Models: LDA, QDA, Naïve Bayes	MASS, e1071 *lda, qda, naiveBayes*	$\sum_{i=1}^{13}\binom{13}{i} * 2 = 16,382$	Training and Validation sets used
Resampling Methods: 5-Fold and 10-Fold Cross Validation	Boot *cv.glm*	$\sum_{i=1}^{13}\binom{13}{i} * 3 = 24,571$	Training set used Errors obtained from cross-validation
Model Subset Selection Methods: Forward, Backward, and Forward-Backward Stepwise Selection	MASS *StepAIC*	3	Training set used
Model Shrinkage Methods: Lasso and Ridge Regression	Glmnet *glmnet*	2	λ varied from 10^{-4} to 10^{10} Training set used
Random Forest	randomForest *randomForest*	1	100 Trees
XG Boost	xgboost *xgboost*	1	All Features

Table 3: Final Model Performance Evaluation



Algorithm	Model	Remarks	Error*	Kaggle Score**
Logistic	*Rainfall* ~ MaxTemp + Sunshine + Humid_Cloud		0.128	0.894
kNN	*Rainfall* ~ Temperature + Humidity + Cloud	k = 27	0.111	0.893
kNN	*Rainfall* ~ MinTemp + Humidity + Cloud	k = 33	0.111	0.891
kNN	*Rainfall* ~ Humidity + Cloud + Temperature Range	k = 53	0.112	0.886
LDA	*Rainfall* ~ Pressure + MaxTemp + Temperature + Dewpoint, Sunshine, Humid_Cloud		0.135	0.893
QDA	*Rainfall* ~ Pressure + Dewpoint + Cloud		0.143	0.885
Naive Bayes	*Rainfall* ~ Dewpoint + Humidity + Cloud + Sunshine		0.146	0.884
5-Fold Cross-Validation†	*Rainfall* ~ Pressure + Sunshine + Humid_Cloud		0.105	0.896
10-Fold Cross-Validation	*Rainfall* ~ ID + Day + Dewpoint + Sunshine + Wind Speed + Humid_Cloud		0.102	0.890
Forward Select	*Rainfall* ~ Humid_Cloud + Dewpoint + Sunshine + Wind Speed + ID + Cloud + MinTemp + Pressure	Stepwise	0.124	0.893
Ridge	*Rainfall* ~ ID + Day + Pressure	λ = 0.0276	0.247	0.544
Random Forest	*Rainfall* ~ All Predictors	100 Trees	0.355	0.555

[a] *Misclassification error*
**ROC score (Higher score = better)
†Preferred final model

Table 4: Key Findings from Literature Review – Factors that influence rainfall – Used to guide feature engineering

Variable	**Formula**	**Rationale**	**Source**
Dew Point Depression	Temperature minus Dewpoint	Lower values indicate the air is near saturation, which is a good indicator of rainfall.	Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches: Latif et al.
Temperature Range	Max Temperature minus Min Temperature	Rainy days tend to have smaller temperature ranges.	Atmospheric Science, Chapter 10: Wallace & Hobbs
Humidity and Cloud Cover (Interaction term)	Humidity * Cloud Coverage	Rain often comes from high humidity and thick cloud cover	Cloud Climatology: NASA

References

Report

1. https://research.noaa.gov/three-ways-noaa-research-works-to-improve-our-weather-forecasts/
2. https://www.fema.gov/fact-sheet/flash-flooding-be-ready-act#:~:text=Flash%20floods%20can%20sweep%20away,and%20deadly%20floods%20are%20coming.
3. Ibid.
4. The Effects of Changing Weather on Public Health: Patz et al.
5. https://www.kaggle.com/competitions/playground-series-s5e3/overview
6. Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches: Latif et al.
7. Atmospheric Science, Chapter 10: Wallace & Hobbs
8. Cloud Climatology: NASA

Appendix

- Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. Latif et al.
- Atmospheric Science, Chapter 10: Wallace & Hobbs
- Cloud Climatology: NASA