The Weather Wizard: Leveraging Supervising Machine Learning to Build a Viable and Reliable Weather Predictor

Courtenay-Dee O'Brien, Utsav Nimavat, Zoe Toy, and Jaiyoung Lee

School of Information, University of Texas at Austin

I310D: Foundations of Human-Centered Data Science

Dr. Abhijit Mishra

May 1, 2023

The Weather Wizard: Leveraging Supervising Machine Learning to Build a Viable and Reliable Weather Predictor

Introduction

As climate change continues to become a prominent problem, weather applications and traditional forecasting methods have become less certain and accurate (Garthwaite, 2021). Meteorologists predict changes in weather patterns by considering a number of atmospheric variables such as temperature, air pressure, cloud patterns, precipitation, and wind's speed, direction, and moisture level. In an effort to acquire a complete picture of conditions, they also thoroughly review data gathered from a myriad of technical tools as well as radar, satellite, ground-based and airborne instruments ("Forecast Process," n.d.). This process is not only time-consuming and inefficient, it is antiquated. With the number of technological applications, devices, and daily innovations, the weather forecasting is not aware, nor are they taking advantage of, all the more novel and capable tools at their disposal (Schultz et al., 2021). The purpose of our model, the Weather Wizard, is to modernize *and* streamline the forecasting process to deliver forecasts with great precision and little error, eliminating the need for time-intensive human analysis.

Dataset Description

Raw Data & Attributes

By relying on supervised machine learning methods to deliver detailed, reliable predictions, the Weather Wizard viably improves weather forecasts. This model was trained with a detailed, historical dataset from the Austin KATT Station. It contained 1,320 entries from February 2013 to July 2017. The raw dataset was comprised of the following 21 attributes:

Date	PrecipitationSumInches	Events
TempHighF	TempAvgF	TempLowF

DewPointHighF	DewPointAvgF	DewPointLowF
HumidityHighPercent	HumidityAvgPercent	HumidityLowPercent
SeaLevelPressureHighInches	SeaLevelPressureAvgInches	SeaLevelPressureLowInches
VisibilityHighMiles	VisibilityAvgMiles	VisibilityLowMiles
WindHighMPH	WindAvgMPH	WindLowMPH

Collecting & Cleaning Data Processes

While implementing the Extract-Transform-Load (ETL) Pipeline, we reduced the dataset to 18 feature columns eliminating the ‘PrecipitationSumInches’ and ‘Events’ column. Given that our model’s goal was to predict whether or not rain will occur based on input from the dataset’s other features, ‘PrecipitationSumInches’ was an unnecessary attribute to have as it revealed the answer to our query immediately. To utilize binary classification algorithm capabilities, we transformed the information on rain from the ‘Events’ column into binary numbers. They were stored in the new ‘Outcome’ column we created where a value of ‘0’ indicated rainfall did not occur and a value of ‘1’ indicated rainfall did occur on the given day. Additionally, all rows with '-' representing incomplete data were dropped. We then split the dataset into a training set by which the ML algorithm would be trained on, and a testing set which would be the benchmark for how accurate our data was. The training set included the first 1000 rows of data, and the testing set included the remaining 218 rows. They were then exported to .csv files.

Methods

Classifier Selection & Algorithms

With the training and testing datasets loaded, we imported sklearn and trained the machine learning algorithm using Logistic Regression (LR) and Multilayer Perceptron (MLP) classification. After

comparing their accuracy results on the test data, we decided to move forward with the LR classifier as it returned an accuracy of 87.0% compared to the MLP score of 85.7%. We confirmed the validity of our model by testing it with the exact attribute values from a day in our initial data set where rain did occur with the LR model. Exceeding our expectations, the model not only delivered the correct result but also returned with a classifier confidence of 94.5%.

While we were satisfied with these results, we hypothesized that the 18 feature columns from the initial dataset felt repetitive and unnecessary, so we selected seven attributes we hypothesized would have the largest impact on the possibility of rain's occurrence to test again. Using this smaller feature set, our classifier returned a confidence of 85.2%. Both predicted the correct outcome: it *will* rain on the given day with the weather measurements provided. Though the larger returned with higher accuracy, we chose to limit our feature set after viewing the LIME analysis as shown in Figure 1. We learned the smaller dataset performed with minimal deviation from the larger and thus, did not have a significant impact on the accuracy of the Logistic Classifier or the classifier confidence. (Figure 1) This made our analysis to determine which variables possess the strongest influence on the probability of rain more digestible, *and* the variables' interaction with one another clearer.

Project Analysis and Results

By creating a Seaborn heatmap to illustrate the correlation between weather variables, as shown in Figure 2, we observed interesting relationships and dependencies on variables that significantly influence whether or not it will rain.

The following variables were found to be the two strongest, positive correlations:

- Average Temperature ℉ and Average Dew Point ℉, $r(216) = 0.89$, $p < .05$.
- Average Wind MPH and Wind Gust MPH, $r(216) = 0.69$, $p < .05$.

As our bivariate analysis hexplot in Figure 3 shows, as the average temperature on a given day increases, the average dew point will increase as well. Similarly, as the average MPH of wind increases, the MPH of wind gust tends to positively follow suit.

Alternatively, the following variables were found to be the strongest, negative correlations:

- Average Sea Level Pressure (in inches) and Average Temperature ℉, $r(216) = -0.62$, $p < .05$.
- Average Sea Level Pressure (in inches) and Average Dew Point ℉, $r(216) = -0.63$, $p < .05$.

By observing the strength and direction of these values, it is appropriate to generally assume that as the average sea level pressure decreases, average temperature and average dew point will reduce as well.

The model also revealed an apparent insignificance on rain between many attributes and outcome within our dataset. Viewing the right-most column on the heatmap, the following variables on outcome were found to be weakly correlated:

Positive, Weak Correlations	***Negative, Weak Correlations***
Wind Gust MPH, $r(216) = 0.18$, $p < .05$.	Sea Level Pressure (in inches), $r(216) = -0.061$, $p < .05$.
Average Dew Point ℉, $r(216) = 0.15$, $p < .05$.	Average Temperature ℉, $r(216) = -0.097$, $p < .05$.
Average Wind MPH, $r(216) = 0.0061$, $p < .05$.	

Lastly, a final interesting insight derived from our visualizations were the following medium-strength, opposing correlations to outcome:

- Average Humidity Percentage was **positively correlated** with Outcome, $r(216) = 0.53$, $p < .05$.
- Average Visibility (in miles) was **negatively correlated** with Outcome, $r(216) = -0.53$, $p < .05$.

Given that Average Humidity Percentage and Average Visibility (in miles) were both found to have moderately-strong, negative correlations with one another, $r(216) = -0.58$, $p < .05$, it is implied that they are connected and have the strongest influence on determining whether it will rain or not.

Conclusion and Future Implications

While our model performed with high confidence and revealed interesting insights into the interaction between weather variables, it would be ignorant to not point out areas of concern or weakness currently within the Weather Wizard. Our model's performance remains undoubtedly determined and linked by the quality of the data it was trained on. The raw dataset used for training and testing from the Austin KATT station only took weather conditions within the city of Austin, TX from February 2013 to July 2017 into account. In order to deliver a model for present-day use in Austin, TX, it needs to be trained with more recent data. To create a model generalizable to many contexts, the dataset should contain larger samples of data from various locations and timeframes. At this time, the current algorithm written, trained, and tested can be used for weather predictions so long as the data inputted is complementary to the dataset we trained it on. However, in order to leverage its predictive insights in a different location or region with the highest accuracy it is capable of, one must re-train the model on historical and present data from the specified area of interest (AOI). Additionally, another area of the Weather Wizard's vulnerability is in the algorithm we opted to use. We chose to use a binary classification algorithm because it is one of the most efficient machine learning classifiers for binary classification problems and is very efficient to train. However, one disadvantage to this method is that it is often found to be sensitive to outliers (Kharwal, 2021). With the variability of weather patterns due to climate change, it is appropriate to assume more outliers will be found in data from the last decade or so than before. A way we attempted to address this was in limiting our model to seven features instead of the former 18 columns. By putting emphasis on the averages of those attributes instead of also considering their minimum and maximum values, we hoped that the algorithm would be less susceptible to inaccurate results due to outliers. While this did address and resolve our outlier concerns, this will still prove to be a weakness for future implications.

For future project applications, we would build on the Weather Wizard's current foundation. First, we would like to update the dataset with Austin, TX weather data July 2017-onward up to present-day. In doing so, this will enhance its ability to predict rain in Austin city limits with greater precision and reliability. Next, we would like to move past the model's predictive capabilities exclusive to rain and

program it to predict other weather events such as thunderstorms, lightning, flash flooding, and snow. After achieving this goal and confirming the model performance operates at the highest possible standard, we would like to iteratively expand the training dataset in phases to include data from surrounding regions in Texas, then bordering Southern states, the rest of the country, North America, Northern Hemisphere, et cetera. In time, as the dataset domain expands, the model will apply the data to build an interactive map or visualization, as shown on television or the Weather Channel, that illustrates predicted weather patterns across the world. However, rather than being built by traditional or existing meteorology tools and weather applications, it is created by machine learning and artificial intelligence. The Weather Wizard has promising potential for future implications in determining weather, as do many other areas within our current technological landscape.

Figure 1

LIME Analysis Results

Figure 2A

Correlations Between all Attributes using Seaborn Heatmap

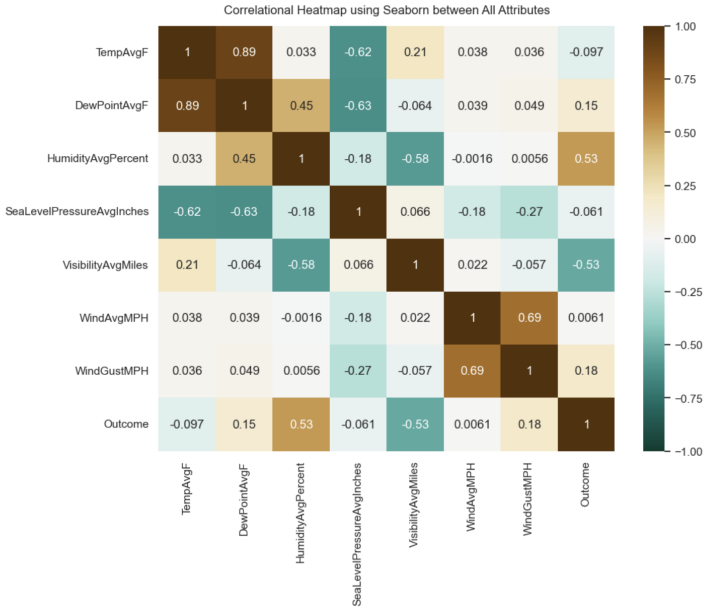
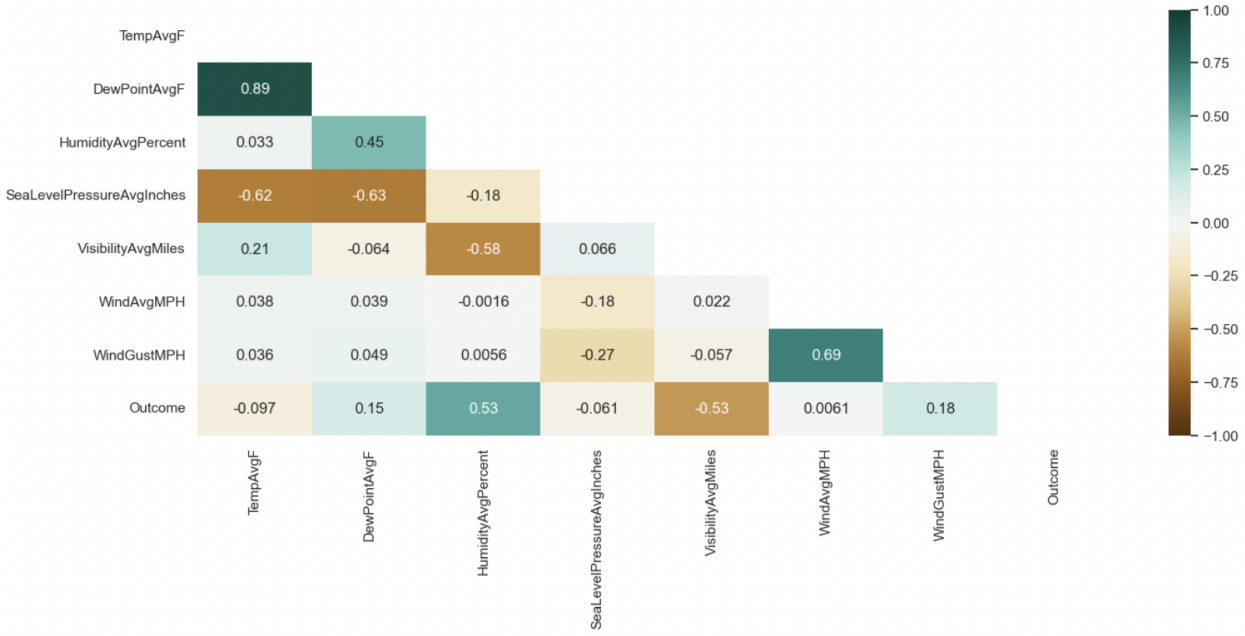


Figure 2B

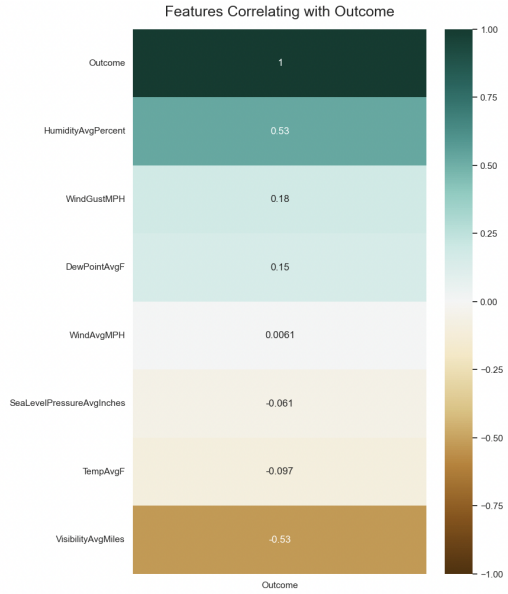
Triangular Heatmap of Correlations Between All Attributes



Note. This figure does not contain any differing data from Figure 2A, it is only to offer a different perspective between the attribute correlations.

Figure 2C

Single-Column Correlational Seaborn Heatmap Between All Attributes and Outcome



Note. This figure does not contain any differing data from Figure 2A or 2B, it isolates the rightmost column of data for closer inspection.

Figure 3

Bivariate Analysis of Features with Strongest Correlation

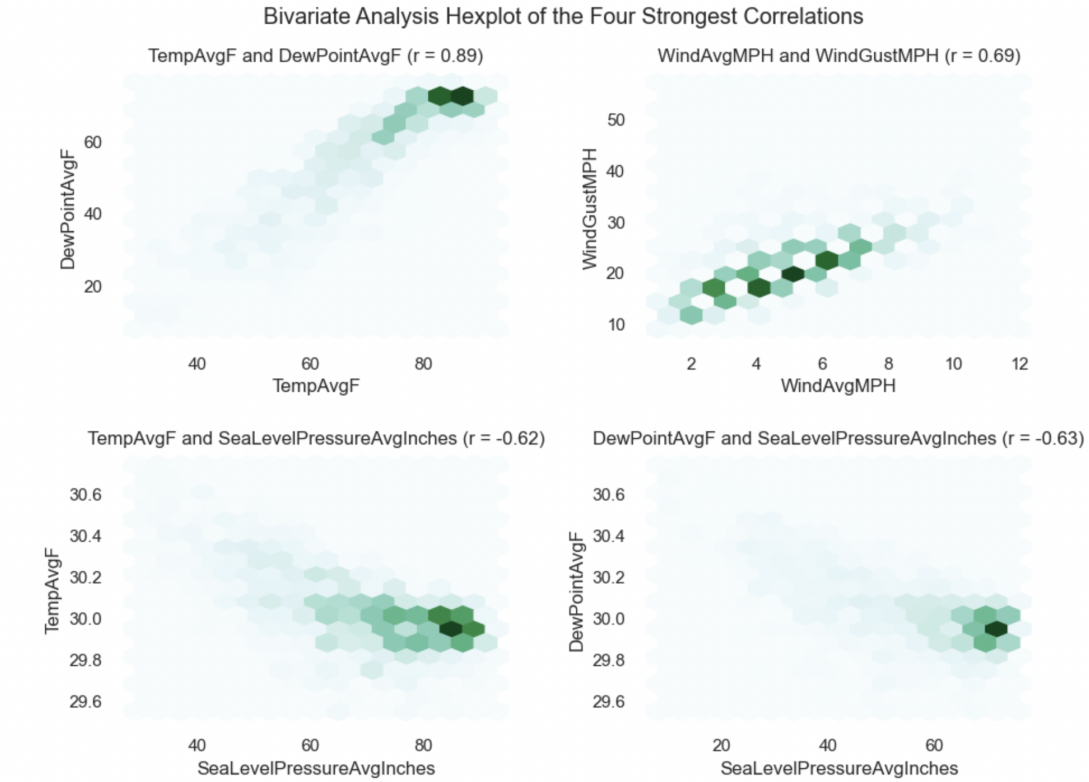


Figure 4

Scatterplot Insight into Humidity, Visibility, and Outcome: Correlation Direction, Interaction, & Significance

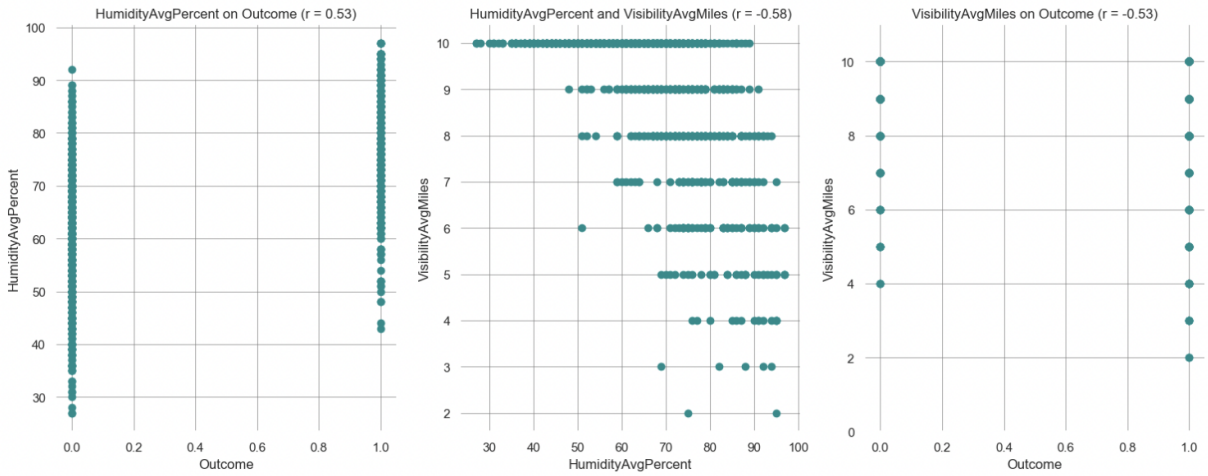
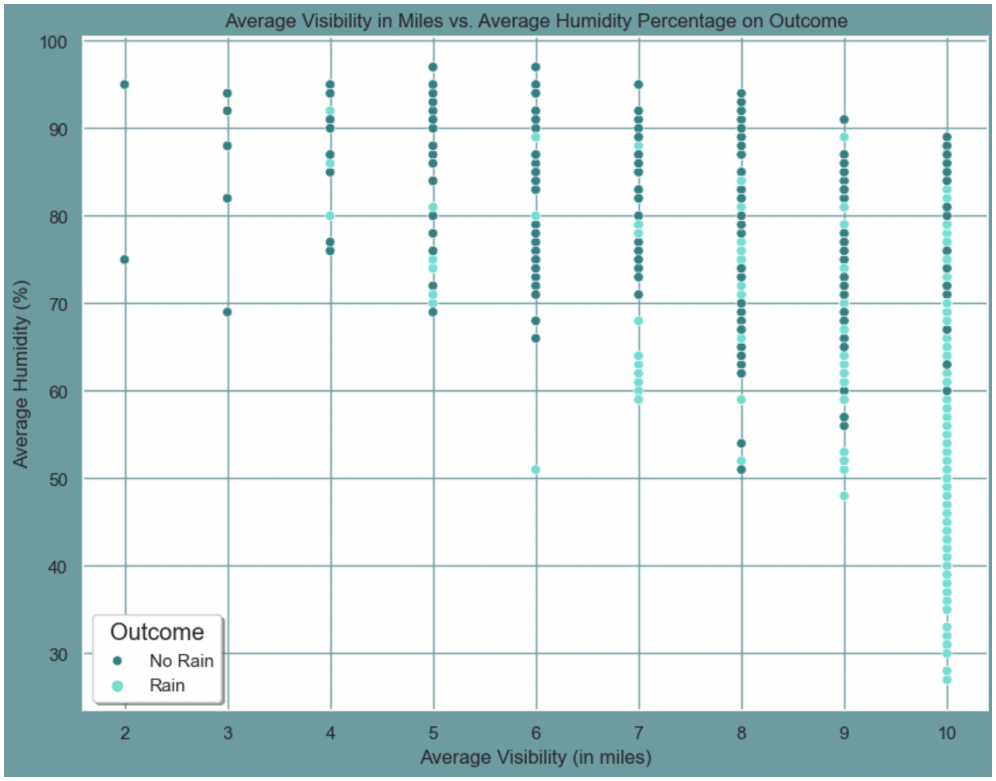


Figure 5

Interaction Between Average Visibility vs. Average Humidity on Outcome



References

Garthwaite, J. (2021, December 14). *Climate of Chaos: Stanford researchers show why heat may make weather less predictable*. Stanford News. Retrieved April 19, 2023, from https://news.stanford.edu/2021/12/14/warming-makes-weather-less-predictable/

Kharwal, A. (2021, November 12). *Binary Classification Algorithms in Machine Learning*. TheCleverProgrammer. Retrieved April 19, 2023, from https://thecleverprogrammer.com/2021/12binaryclassificationalgorithminmachinelearning/

National Weather Service. (n.d.). *Forecast Process*. Weather.gov. Retrieved April 19, 2023, from https://www.weather.gov/about/forecast-process#:~:text=The%20forecast%20process%20is%20oughly,complete%20picture%20of%20current%20conditions.

Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A., & Stadtler, S. (2021, April 5). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A, 379*(2194). https://doi.org/10.1098/rsta.2020.0097