The Weather Wizard

By:

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

GOAL:

Develop a precise and dependable **weather predictor** that delivers a confidence score with visualizations for the likelihood of rain on any given day.

01

Audience

- General Population
- Weather Channels
- News Agencies
- Transportation Companies

02

Purpose

Weather prediction is essential for modern-day life. Our model will be trained with a diverse dataset and will use binary classification to deliver clear, reliable, and accurate results.

03

Context

Weather, increasingly so due to climate change, tends to be pretty unpredictable. We'd like to produce a model that performs with higher accuracy *and* precision than current weather applications, making day-to-day decisions that are impacted by daily weather conditions far easier for users.

Dataset Details

Source

01 Obtained from WeatherUnderground.com, from the Austin KATT station

02 Posted on Kaggle

Attributes

Date, TempHighF, TempAvgF, TempLowF, DewPointHighF, DewPointAvgF, DewPointLowF, HumidityHighPercent, HumidityAvgPercent, HumidityLowPercent, SeaLevelPressureHighInches, SeaLevelPressureAvgInches, SeaLevelPressureLowInches, VisibilityHighMiles, VisibilityAvgMiles, VisibilityLowMiles, WindHighMPH, WingAvgMPH, WindGustMPH, PrecipitaionSumInches, Events

About

Historical temperature, precipitation, humidity, and wind-speed for Austin, Texas from February 2013 to July 2017

1,320 Entries

All entries are quantitative values, except for 'Date' and 'Events'. 'Events' column contains strings describing the day's weather event, if any occurred.

Method

Extract-Transform-Load

Implement the ETL Pipeline - Data Curation & Cleaning

Coding & Visualization

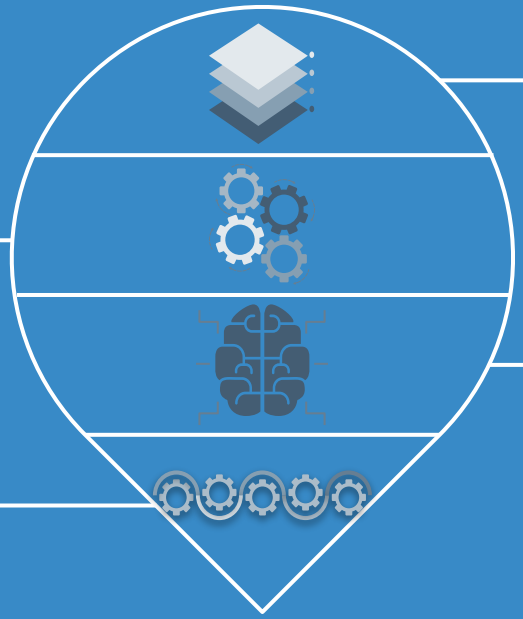
Use binary classification ML to predict if rain will occur.

Train & Evaluate

Classifiers (Logistic, MLP, RandomForest, etc) to determine which performs with the highest accuracy

Evaluation

Analyze results and viability of our model; explain prediction techniques using LIME, correlation, and bivariate analysis.



Training

With the training and testing datasets loaded, we imported sklearn and began to train the ML algorithm using Logistic and MLP classification.

- **Accuracy of MLP Classifier = 85.7%**
- **Accuracy of Logistic Classifier = 87.0%**

We predicted a single instance using our Logistic Regression model, using the values from one day in our dataset where it rained.

- **Result: it is likely to rain in the near future (classifier confidence = 94.5%)**

Hypothesis

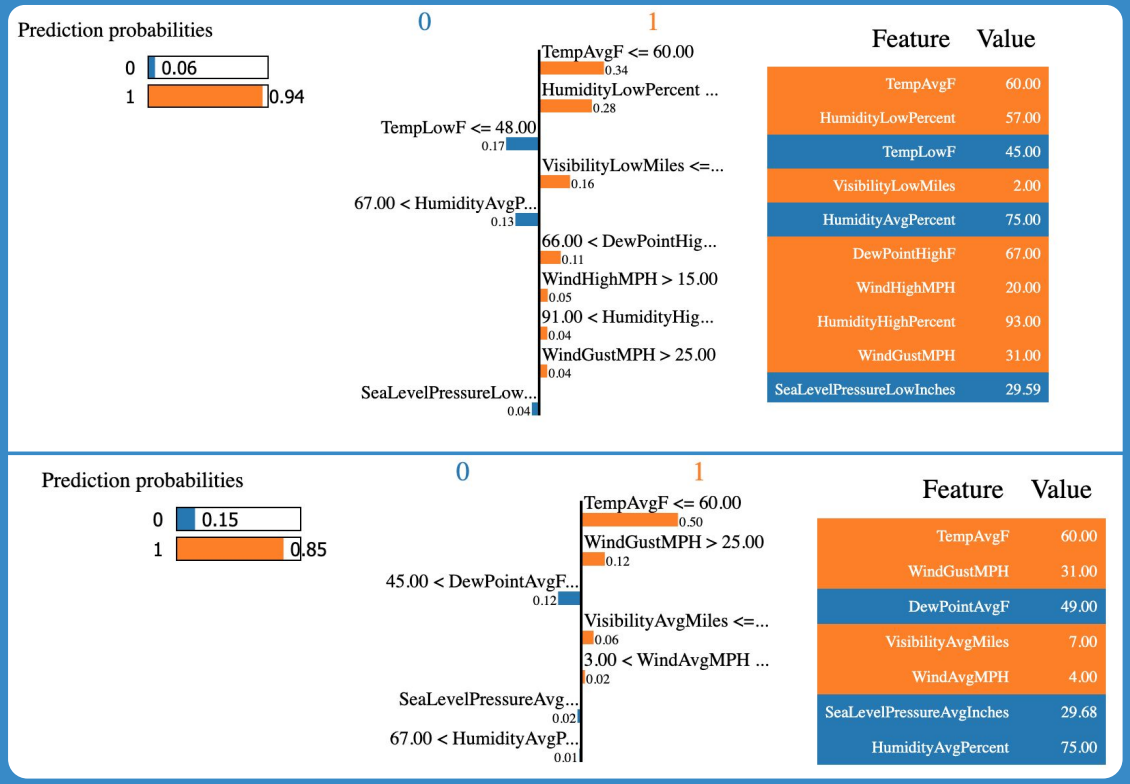
Question: which features play the largest role in the likelihood of rain?

Hypothesis: the averages (temperature, dew point, humidity, sea level pressure, visibility, and wind) and wind gust will have the largest impact on outcome.

Testing: train our model on these features, evaluate using LIME, correlation, and bivariate analysis

- **Accuracy of Logistic Classifier on smaller feature set = 82.9%**

LIME Visualizations

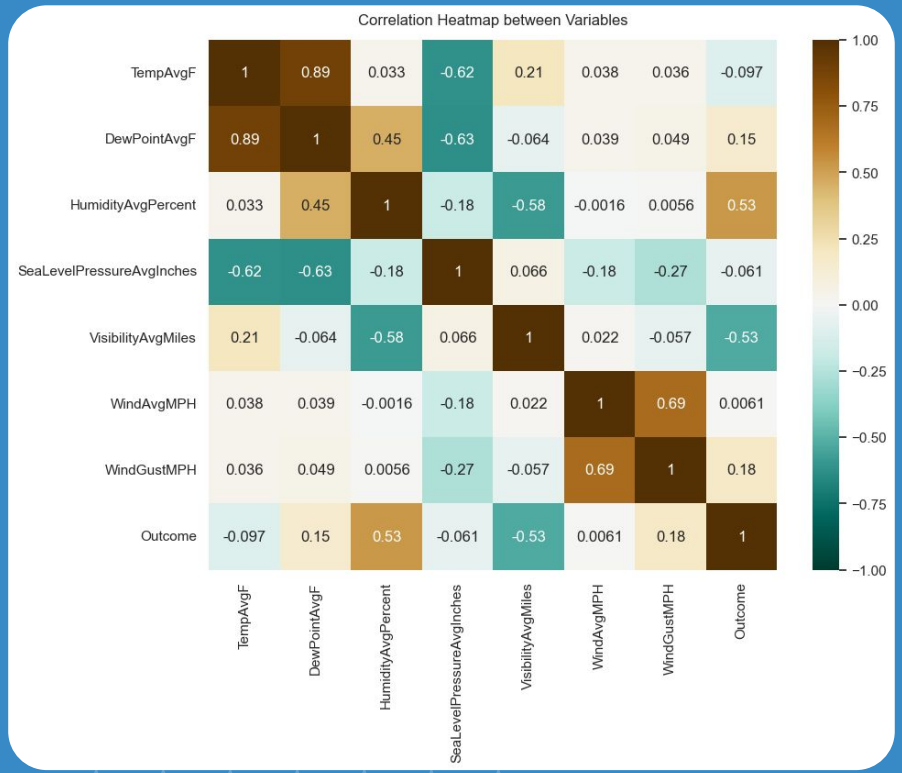


Both the larger and smaller feature set correctly predicted the outcome (94.5% vs 85.2% confidence).

From LIME, it appeared the smaller dataset performed with minimal deviation from the larger and did not have a significant impact on the accuracy of the Logistic Classifier or the classifier confidence.

The larger feature set also made “HumidityLowPercent” and “VisibilityLowMiles” decisive features, ignoring the actual average values, making the classifier falsely confident.

Correlation Coefficient between Different Categories



Strong positive correlations

- TempAvgF and DewPointAvgF
- WindAvgMPH and WindGustMPH

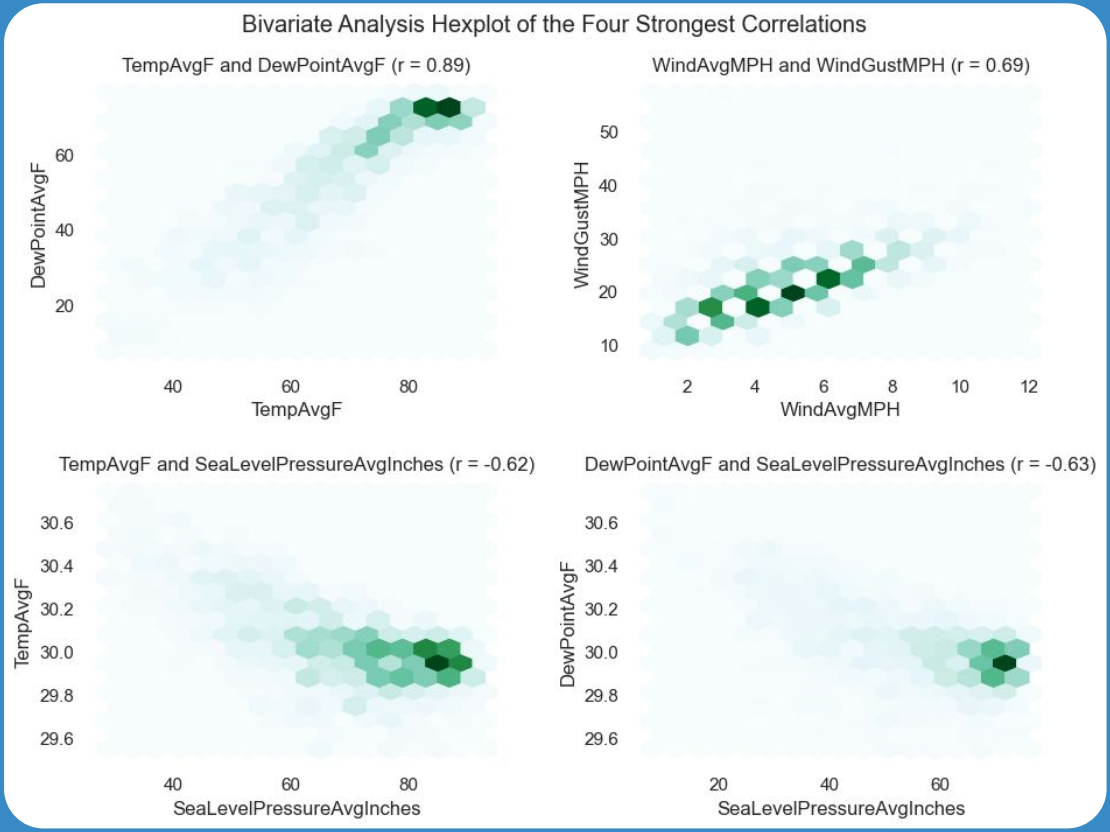
Strong negative correlations

- TempAvgF and DewPointAvgF with SeaLevelPressureAvgInches

Other interesting takeaways

- Correlation between Humidity and Visability with Outcome

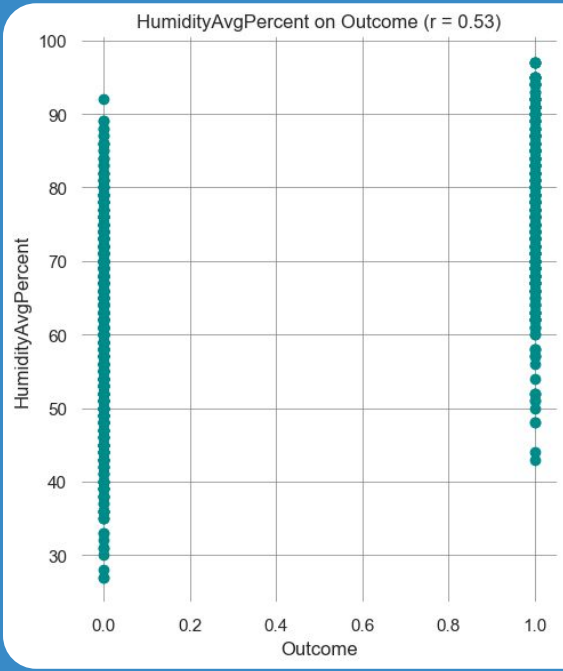
Bivariate Analysis

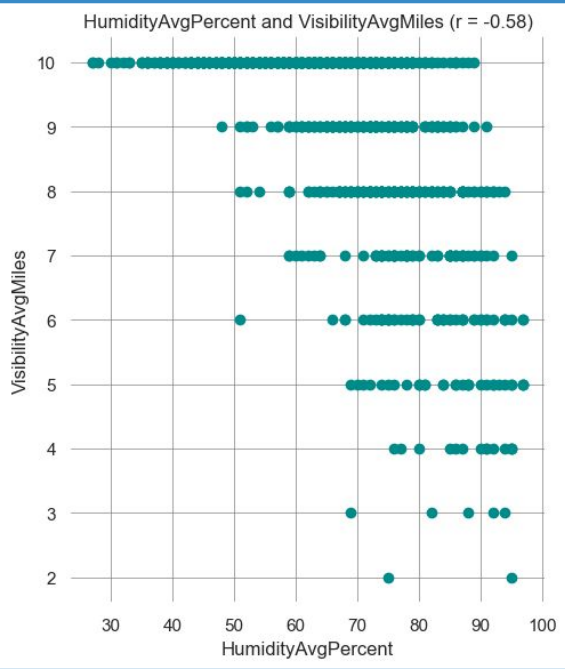


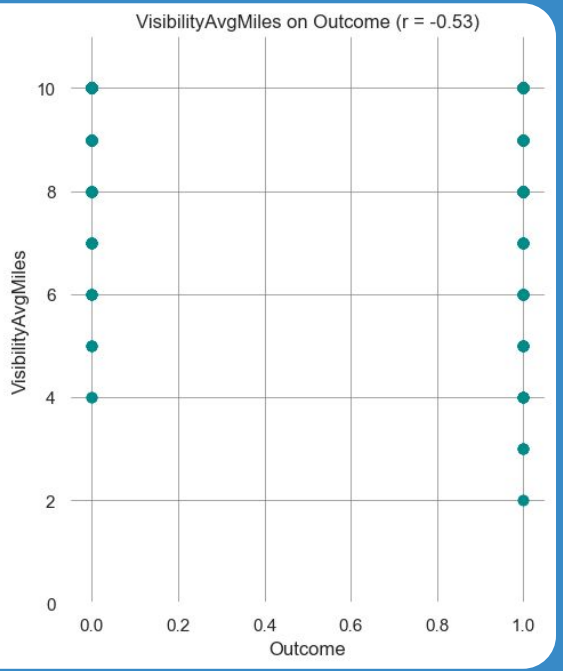
- The selected variable pairs have the highest correlation coefficients (r-values).
- TempAvgF and DewPointAvgF are closely correlated with each other and SeaLevelPressureAvgInches

Bivariate Analysis (cont.)

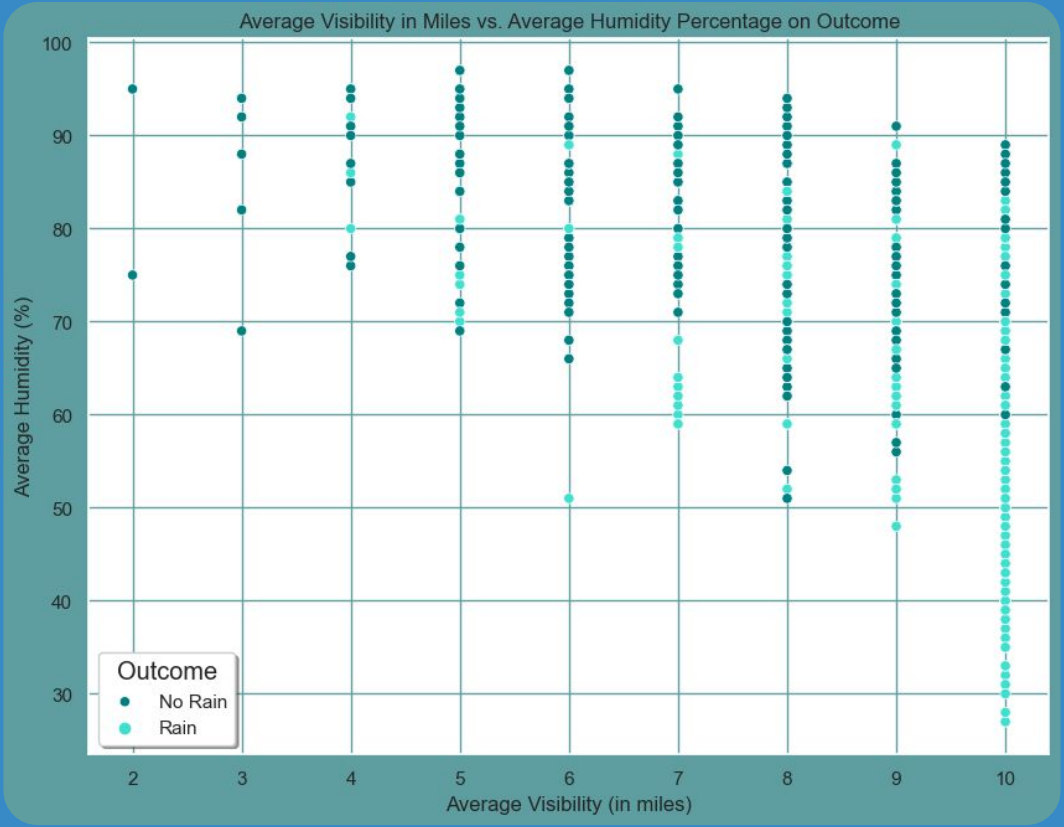
- **Humidity and visibility** have somewhat of a negative correlation, indicating that an increase or decrease of either of the variables dictates chances of rain and the increase or decrease of the other variable.







Bivariate Analysis (cont.)



Conclusion & Future Implications

Conclusion

Our model performed with relatively high confidence (85.2%) and revealed interesting insight into the interactions between different features

Potential Weaknesses

- Model is trained on data from Feb 2013 - July 2017
- Limited to Austin, TX

Future Implications

- Update dataset with July 2017 - present
- Expand prediction capabilities to lightning, flash floods, snow, etc.
- Include data from surrounding regions in TX, bordering states, etc.