Gulf Coast Area Rainfall Prediction

by Jimmy Smart

Abstract

Our livelihood is so greatly affected by the weather. Depending on the region someone lives in usually determines the type of weather patterns that they'll see on a yearly, monthly or even daily basis. A large majority of industries greatly depend on their region's weather and climate. Using machine learning modeling, we tried our hand at being able to predict rainfall for the Gulf Coast area. With access to multiple features such as, seasonal temperatures and humidity percentages we set out to find their impacts at predicting rainfall. Overall we set out to answer the question: Can we predict rainfall for the Gulf Coast area?

1. Introduction

Farmers need a good mixture of rain and sunlight so that their crops may grow. Too much sun but not enough rain, may cause issues to their crop yields. Just as too much rain and not enough sun can affect their crop yields as well. Seasonal temperatures factor in as well, as a majority of crops grow based on the daily temperatures. Some may grow better in months that are typically known to have cooler temperatures, while other crops prefer to grow in the warmer months.

Commercial fishing also keeps a keen eye on weather patterns as life found in the Gulf usually have their own growing seasons as well as seasons for farming.

For farmers, commercial fishing and various other industries who depend on their regional climates, their best approach in co-existing with weather patterns is by using past trends to create future predictions.

History is known to repeat itself at times but predicting if/when history will repeat itself is an ongoing issue. The best approach is look for patterns of the past while comparing it to patterns of the present.

Weather has shown signs of yearly and seasonal patterns but it's not always set in stone. Just like other factors in our world, weather patterns are changing as well. Global warming has become a hot debate in our culture today as more and more people feel that the certain factors in our environments are raising temperatures around the world which in turn are also raising the temperatures of our oceans, seas and gulfs. This project isn't necessarily a global warming, project, but as we go through our historical weather data and try to predict future weather patterns, we will be able to see if there has been a general uptick in temps and see if there is a correlation with any effects it's having on our data over the years.

2. Data

This project will build a model that focuses on a dataset consisting of decades of daily weather from various locations within our Gulf Coast area. Where we're looking at consists of an area in the southern part of the United States that shares a coastline on the Gulf of Mexico. Our data is from https://ww.ncdc.noaa.gov/ and consists of 5 different datasets from each of our 5 chosen cities along the Gulf Coast. Each dataset consists of daily weather data and is saved as a CSV file.

We spent a good deal of time cleaning and wrangling our datasets before creating various dataframes to help us better visualize our dataset. Our initial exploratory data analysis (EDA) consisted of looking at the shape of our dataframe (699929, 141) and converting our datetime value was converted into a readable datetime object through to_datetime. We decided to focus on our daily data so we dropped hourly and monthly features along with any other features that we didn't feel would be necessary with the rest of our project.

We then set out to find any null or missing values along with correcting any of our values that were input incorrectly and would cause issues later on.

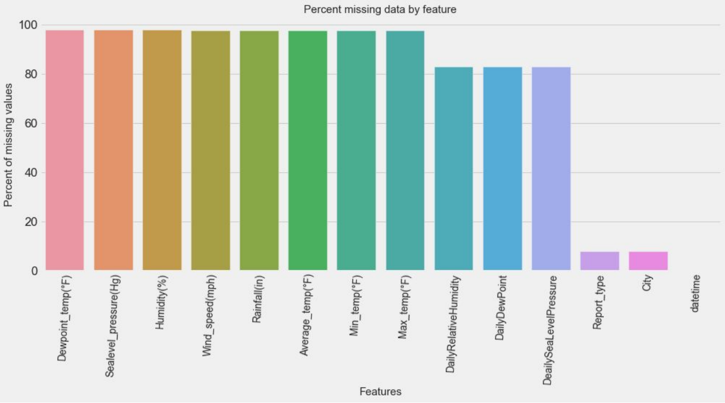

Fig. 1: Percentage of missing data by feature

3. Data Analysis

We created a heatmap so we could see how our features related to each other. What we found was that Dewpoint, Humidity and Sealevel positive correlations. While Average temperature and Windspeed had a negative correlation with our rainfall variable.

As we dove further into our dataset we used various groupby functions to get a better idea of what our data was telling us. What we found is were various trends in regards to rainfall and temperatures over our decade of historical data.

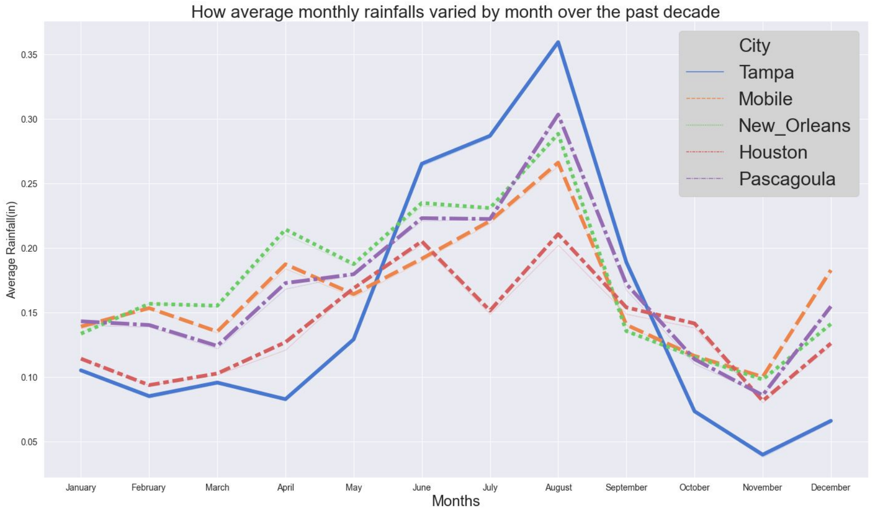

Fig. 2: Average monthly rainfalls varied by months

A look at how average monthly rainfalls varied by months for each of our cities. We can see that Tampa had the highest and lowest totals amongst our 5 cities. We can also see that the summer months have the highest average rainfalls of the year. Considering hurricanes and tropical storms are known to affect this area of the country and that hurricane season is known to be between early June and late November, this may also be a factor.

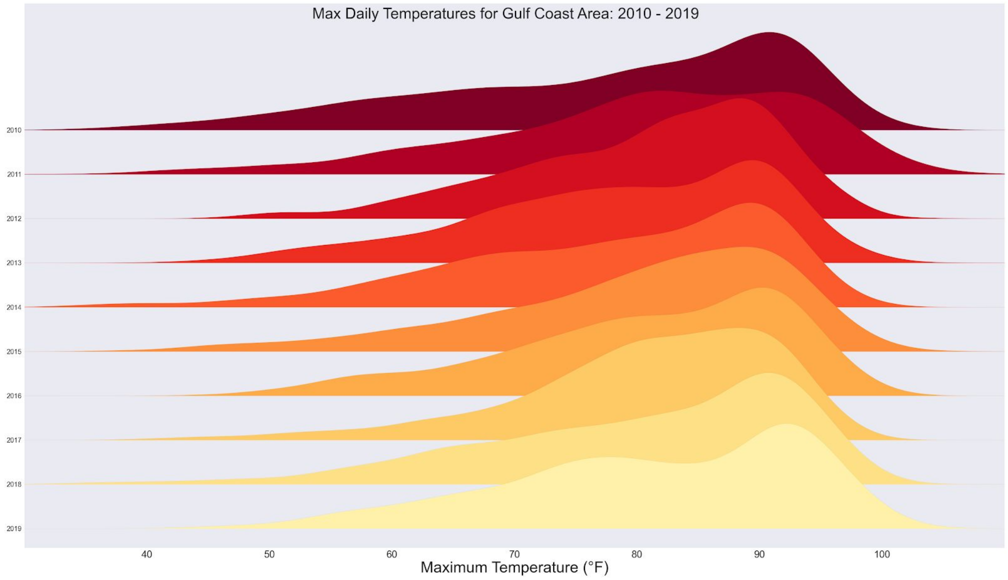

Fig. 3: Max daily temperatures

Here we can see how our average temperatures have fared over the months of our dataset. Summer months are typically known to be hot in this area with a bit of a break from the heat in the other months of the year. As we can see, there has been an uptick in average max temperatures in the Gulf Coast Area. Gaining almost 3 degrees hotter over the past decade. At least in the Gulf Coast area there is a warming trend occurring.

4. Machine Learning Modeling

For our modeling we ran 4 different regression models on 2 datasets. The 1st dataset we took out any rainfall outliers over 5.5 inches of rain in a day. For the other dataset we left the Rainfall outliers in.

FB Prophet

The 1st model we ran was Facebook's Prophet Time series. We were able to tune the models hyperparameters in our quest to see if we could use this model to properly predict rainfall for years to come.

Between our 2 versions of models (with/without outliers) the graphs showed that monthly volumes of rainfall dropped in 2011 but steadily increased until 2016 when it started to decrease slightly each year. This graph also showed that there may be a downward trend for the next following years.

Another trend graph showed an interest since it focused on monthly features. Traditionally weather is affected by seasonal habits. As we saw in our plot, the warmer months of May-September were higher than the rest of the year but there were dips within those months.

Our Prophet models provided nice visuals especially in regards to past trends and what the next couple of years may look like. Our MAE, MSE and RMSE scores are decent but could be better.

Our tables showing our yhat metrics are useful but can use more tuning so our predictions have less error percentages and can better predict future rainfalls.

Linear Regression

Our 2nd model focused on Linear Regression. We once again separated between datasets with and without outliers we then 1st created a simple linear regression with Humidity as our independent variable and Rainfall(in) as our dependent variable.

We conducted a 70% split of our dataset as we conducted **train, test, split()** before training with **LinearRegression()** and fitting our training. Our dependent was set as (Rainfall(in)), with our other features serving as our independent features.

Our Testing r2 score isn't as high as we would like it to be but we also see that our training score didn't come out very well either. Our training score is way off as well. The features that we used don't work as well with our Rainfall variable as well as we would have liked it too.

MAE and RMSE metrics for our models are good results for what we had to work with within our LR models.

XGBoost

Our 3rd model focused on XGBoost. We once again separated between datasets with and without outliers. Once again Rainfall(in) is our dependent variable.

We conducted a 70% split of our dataset as we conducted **train, test, split()** before training with **XGBRegressor()** and fitting our training.

For both of our model version (with and without outliers) we created and ran multiple XGB models to see which one produced the best metrics.

For our model with outliers excluded, our test and training metrics showed that our model wass a bit too overfit. Our model with outliers included had less overfitting and may be the model that we focus on to build a better model that helps us predict rainfall.

Our feature importance keeps in line with our other models and correlation heatmaps with rainfall. Humidity has usually been at or near the top. Dewpoint seems to be a bit lower than usual.

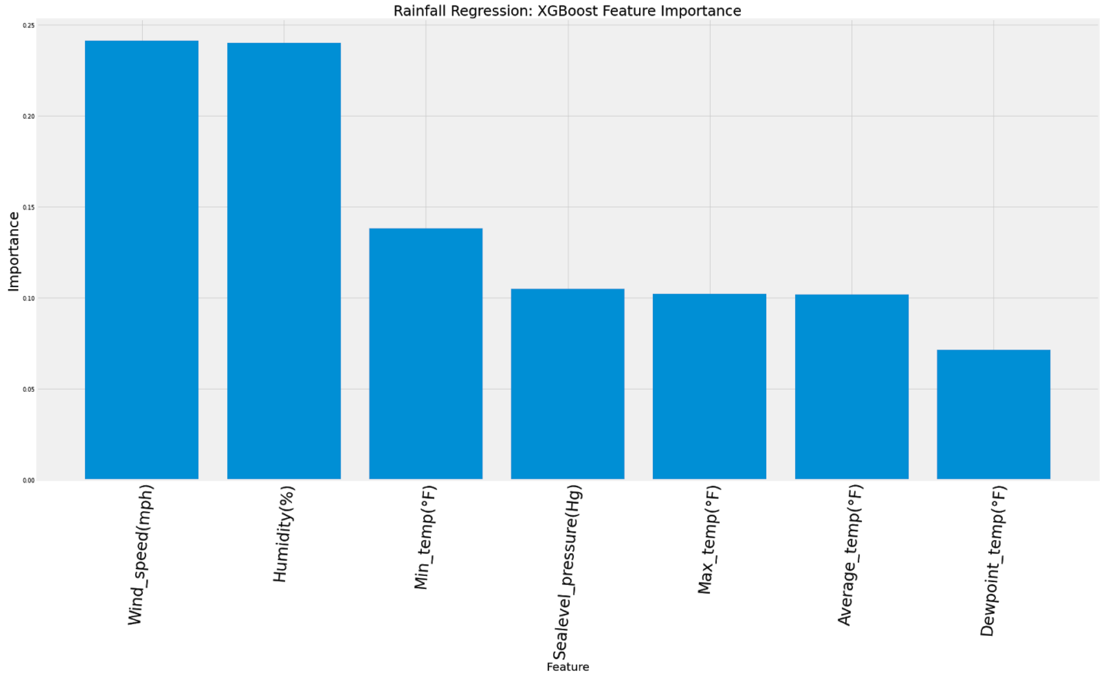

Fig. 4: XGBoost Feature Importance

Our score was a bit low. We could add other outside features to see if it would help increase our score. With more features, our efforts to tune our hyperparameter models so they aren't as overfitted as they currently are.

Random Forest

Our 4th model focused on Random Forest. We once again separated between datasets with and without outliers. Once again Rainfall(in) is our dependent variable.

We conducted a 70% split of our dataset as we conducted **train, test, split()** before training with **RandomForestRegressor()** and fitting our training.

For both of our model version (with and without outliers) we created and ran multiple RF models to see which one produced the best metrics.

Our models provided better MAE and RMSE metrics but as with our other models, our r2 test score isn't as high as we would have liked it to be.

The feature importance tables are consistent with our other models, at least in the 2 main features that have the highest importances amongst the other features. Humidity and

Windspeed leads the way but the difference between our RF models is that the minimum and maximum temperature features swap importances places between the 2 models.

Our RF isn't bad but to make it better and useful to use for rainfall predictions more will be needed to gain better r2 scores as well as better contain any overfitting between our test and training sets.

5. Comparing Models

With 4 different models and 2 versions of each we collected our various model metrics so that we could view them all in 2 tables.

Models without outliers:

TABLE 1: Comparing models without outliers

	name (without outliers)	R2 - Test	R2 - Train	MAE - Test	MAE - Train	RMSE - Test	RSME - Train
1	FB Prophet	-0.001	NaN	0.248	NaN	0.468	NaN
2	Linear Regression(multiple)	0.223	-2.600	0.214	0.220	0.396	0.406
3	XGBoost	0.255	0.650	0.161	0.220	0.373	0.217
4	Random Forest	0.292	0.847	0.167	0.065	0.382	0.142

As we have touched on throughout our report, our r2 scores are a bit lower than we would like and there is overfitting/underfitting issues with our models. XGBoost and RF would be the models that we would focus on to create a better prediction model. With more, or different features, we could rerun the models and use more hyperparemter tuning methods to achieve a better result.

Our MAE and RMSE metrics can be better but are not too far off of where we need to be. As mentioned, reworking our models a bit should also help us achieve better MAE and RMSE results.

Models with outliers:

TABLE 2: Comparing models with outliers

	name (with outliers)	R2 - Test	R2 - Train	MAE - Test	MAE - Train	RMSE - Test	RMSE - Train
1	FB Prophet	-0.003	NaN	0.261	NaN	0.544	NaN
2	Linear Regression(multiple)	0.174	-2.726	0.235	0.233	0.511	0.456
3	XGBoost	0.346	0.277	0.167	0.233	0.409	0.318
4	Random Forest	0.311	0.851	0.170	0.065	0.196	0.026

Our XGBoost model looks to be the model worth focusing on and trying to achieve better results with. The r2 score could be higher but unlike with our other models this model. The MAE and RMSE are good as well for what the model had to work with. The RMSE is a bit high considering the lower MAE.

RF model is respectable too. Working on the overfitting would be a major focus if this model was chosen. The MAE and RMSE metrics are low which is what we want to see.

Prophet and LR models didn't work well with the features and variable we ran. Adding features along with taking out low scoring features could help boost their results.

6. Conclusion

Along with the models that we showcased within our project and reports, we also ran various versions of our 4 models with different approaches. Such as models that only focused on only 1 or 2 of our cities, instead of all 5 cities. These models performed about the same as the results shown within this report. We also ran models trying to predict the simplicity of if it would rain or not, instead of trying to predict the amount of rain a future date may or may not have. These models also produced the same range of metrics as the models we showed in our report. This tells us that the features we used for our project will need to be revisited.

To better predict weather for the Gulf Coast Area, looking at other parts of the world might be a good option. Considering weather that eventually affects certain regions, in our case the Gulf Coast area, we know that storms and weather patterns are influenced in other parts of the world 1st.

XGBoost with outliers include looks to be our best performing model. It didn't produce the metrics that we would have liked but it does show promise that if more attention was given to this model, we could produce better results and predictions. As mentioned, the model could perform better with more features while also removing lower importance features.

Our RF model with outliers included didn't obtain a good r2 score but we were encouraged by the MAE and RMSE scores. Just as we mentioned with the XGBoost model that we consider our best performing model, this RF model could also be an option to further pursue to see if we can achieve better results with some reworking of our features and by tuning the hyperparameters of our RF model.

Our, lesser performing models Prophet and LR models showed some valuable information even if we didn't like the metics that they offered. For Prophet, the visualization of future trends can be usual for people and industries who rely on future rainfall predictions. The visuals easily showed past trends and future trends in regards to rainfall.

7. Future Steps

For our project we casted a pretty wide Gulf Coast region so depending on who needed the prediction info, we could turn our models' hyperparameters to have more cities and/or smaller radius to a given area. Instead of our cities ranging from Houston all the way to Tampa, we could solely focus on Houston, Tampa or any of the other cities we chose for our project. More than likely a big storm hitting the New Orleans area would also affect people nearby in Mississippi and lower Alabama. Also depending on the size and direction of a storm, it could pass through all 5 of our cities in its path.