Weather analysis and prediction for Phoenix Region

Gnanadeep Settykara

Abstract—Weather forecasting, which is an important part of daily living, is a great benefit of science & technology. Weather forecasting may help individuals preserve their possessions and lives. Phoenix is already well-known for its intense heat. Temperatures have been quickly altering as a result of global warming. As a result, forecasting weather can assist the government in providing assistance measures to poor people during intense heat waves and other natural disasters. Weather data near the Phoenix airport is evaluated in this effort to determine how temperatures are growing adversely each year. The reason for selecting data near Phoenix Airport is that airports have advanced sensors that can correctly detect weather. To forecast temperature change, the study introduces several regression techniques such as linear regression, ridge regression, LASSO, and Elasticnet regression. The key features focus on the maximum or lowest temperature and precipitation based on the previous day's measurements.

Index Terms— Weather forecasting, Regression, weather analysis, supervised learning.

1 Introduction

The prediction of meteorological conditions like temperature, dew point, humidity, wind speed, and rainfall in a specific area is known as weather forecasting. Current weather patterns, weather situations, tracking the circulation of air and clouds, and past weather patterns are all collected using instruments such as barometers, radars, and thermometers. Weather forecasting systems are one example of sophisticated equation systems that computers must solve. The prediction of weather-forecasting systems assists airports, or the navy if there is an abrupt change in climate, farmers in increasing agricultural yield production, and even mining personnel require climatic conditions to continually monitor the Earth's crust. Droughts & floods can be avoided with proper weather forecasting [1].

The latest Fifth Assessment Report of the Intergovernmental Panel on Climate Change confirmed that our climate and its extremes are changing (IPCC 2013). The Intergovernmental Panel on Climate Change (IPCC) is made up of scientists from all across the world. These researchers discovered that between 1900 and 2020, the world's surface air temperature increased by an average of 1.1° Celsius (almost 2°F) owing to the combustion of fossil fuels, which emits CO_2 and other greenhouse gases into the atmosphere. This may not appear to be a significant alteration, but it is unprecedented in the over 2000 years of data. Even a single degree may have a significant influence on the globe. Climate models project that if greenhouse gas levels continue to grow at their current rates, the Earth's average temperature globally will climb another 4° C (7.2° F) over the twenty-first century. To limit possible hazards and losses caused by weather and climate extremes, reliable forecasts of extremes on short and long time periods are required (IPCC, 2012; Seneviratne et al., 2012) [2]. In this research, meteorological data near the Phoenix airport is evaluated to see how temperatures are growing adversely each year. Phoenix is already well-known for its scorching temperatures. Temperatures are quickly altering as a result of global warming. As a result, forecasting weather can assist the government in providing relief measures to impoverished people during intense heat waves and other events.

From this research, we might be able to find answers to some interesting questions related to the weather data, (even though not all these questions are answered in this study, this study provides scope to know these answers) like-

a) How are the daily maximum temperature and weather on a given day related?
b) How are the daily minimum temperature and weather on a given day related?
c) How does the heat index get impacted by the daily maximum temperature?
d) Identify the trends within data about the daily max. and daily min. temperatures over the given period in the Phoenix region.
e) What do increase in daily minimum temperatures suggest?
f) How can the increase in daily temperatures in the phoenix region be correlated with issues like droughts, poor air quality, more greenhouse gas emissions, etc. to know the root cause?

The research objectives (RO) of this paper therefore include the following:

-> RO1: To describe the trends within data about the maximum and minimum temperatures recorded in the Phoenix region.

-> RO2: To predict the value of the temperature recorded on a particular date in future by using the past weather data in the Phoenix Region.

-> RO3: To defend the model for performing the prediction of weather in RO2

-> RO4: To defend the model for performing the prediction of weather in RO2 and to evaluate causal relationships between the weather and issues like droughts, poor air quality, green house emissions...etc implied by the RO2 model.

1.1 Related Work

Mulubrhan Balehegn et al. [3] "Many indigenous communities throughout the world utilize traditional weather and climate forecasting as guidance in making key decisions that allow them to survive and adapt to climate change-induced severe weather fluctuation. Traditional weather and climate forecasting is the most accessible and inexpensive source of weather & climate information in many African pastoral communities. Individual interviews and focused group discussions were used in this study to systematically record indigenous climate & weather forecasting knowledge among Afar pastoralists, to make such knowledge available, enhancing the use of this knowledge in climate change adaptation, and exploring synergies with modern-weather-forecasting systems." It shows the importance of the accessibility of weather forecasting to all people so that they can be prepared to face weather changes. But this study did not try to find the causal relationships underlying weather changes or the relationship between weather attributes and how those are responsible for weather fluctuations.

TanviPatil et al. [4] "The goal of this research is to forecast the weather over a specific period. Various types of variables are used to characterize the weather at any given time. Only the most important criteria are employed in the process of weather prediction. The selection of such traits is heavily influenced by the location you have chosen. The existing weather-condition attributes are utilized to fit a model, and future fluctuations in the attributes are analyzed using machine learning techniques and extrapolating the information." It tells which attributes or weather variables need to be chosen based on location, but it did not try to identify the trends in the weather data and the relationship between weather attributes.

All the studies on weather forecasting only focused on what attributes are needed to predict the weather and did not focus on the underlying relationships of weather variables with the weather. In this paper, I tried to identify the trends of weather variables along with their relationship with the weather and how they can be the reason for the increasing weather fluctuations like heat waves in the Phoenix region, Arizona.

2 Exploratory Data Analysis

The data source for this project is retrieved from the data collected and maintained by the "GHCN (Global Historical Climatology Network)-Daily", and "NCEI (National Centers for Environmental Information)", the link for this site is "https://www.ncei.noaa.gov/cdo-web/", this is one of the most reliable government data sources for any type of environmental data. This dataset consists of historical data of the Phoenix region starting from June 1933 till date. This dataset is prepared by the measurements of weather stations near phoenix airport, managed by the (NOAA) National Oceanic and Atmospheric-Administration. The dataset consists of nearly 30753 records and 53 fields, but out of all these fields, only 5 fields are core elements according to GHCN-Daily. These fields are Precipitation, Snowfall, Snow depth, Maximum temperature, and Minimum temperature.

Now looking at the interpretable records in the data, In row 18919, the Date is '22-05-1990', the Precipitation and snow are '0', the maximum temperature is '102', the minimum temperature is '73' and the Weather is 'Sweltering'. The interpretation of this row can be inferred by looking throughout the data from 1933 to 2022, we can observe that the maximum temperature is recorded on this given date which is 102 F. The record is self-explanatory as it can be observed that the date falls in the month of May, which might be the peak time of the summer. Also, a week before and after this date recorded similar high temperatures. Similarly in row 30467, the Date is '02-01-2022', the Precipitation and snow are '0', the maximum temperature is '59', the minimum temperature is '36' and the Weather is 'cool'. The interpretation of this row can be inferred by looking throughout the data from 1933 to 2022, this record is also self-explanatory as we can observe that the given date falls in the month of January, which is during the winter, and it can be observed that similar temperatures were recorded on the nearby days.

Both the above records suggest that the weather of a given day depends on maximum temperature and minimum temperature. For example, if the maximum temperature increases then the weather of the day is hot, and if the maximum temperature decreases then the weather of the day is cool. However, by looking at the whole dataset, we can notice a trend that the daily maximum and daily minimum temperatures of a day are increasing continuously from 1933 to 2022. So, even if the maximum temperature of a day is increased by a few Fahrenheits, the weather is changing from hot to sweltering.

The following figures show the scatterplots between the various weather variables for finding the relationships between them.

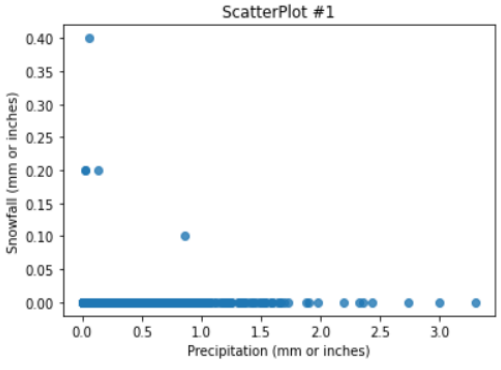

Fig-1. Scatterplot between Precipitation and Snowfall, which shows almost no correlation between them.

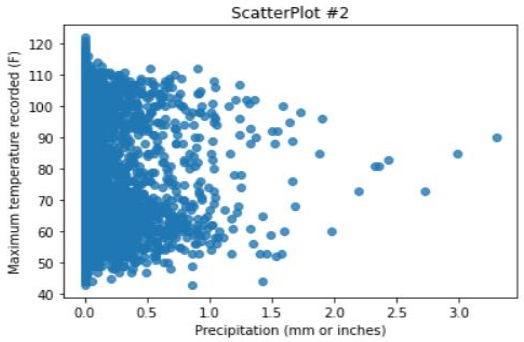

Fig-2. Scatterplot between Precipitation and Maximum temperature, which shows negative correlation between the variables.

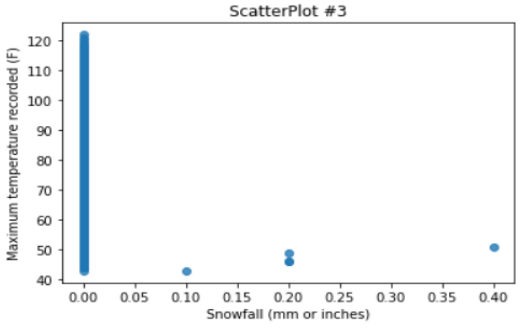

Fig-3. Scatterplot between snowfall and maximum temperature. It shows a negative correlation between the variables.

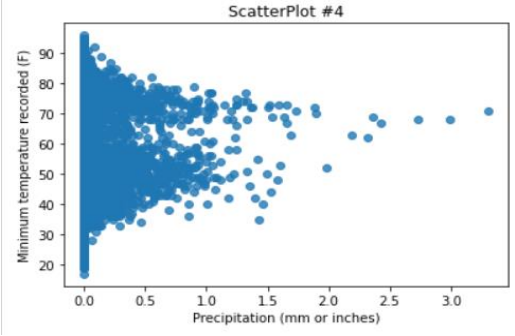

Fig-4. Scatterplot between precipitation and minimum temperature. It shows a negative correlation between the variables.

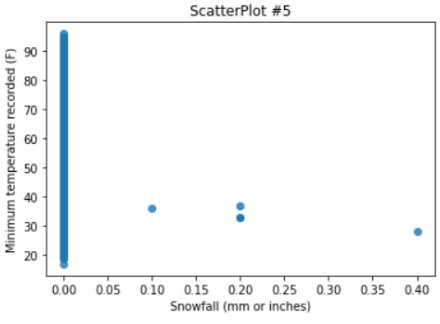

Fig-5. Scatterplot between snowfall and minimum temperature. It shows a negative correlation between the variables.

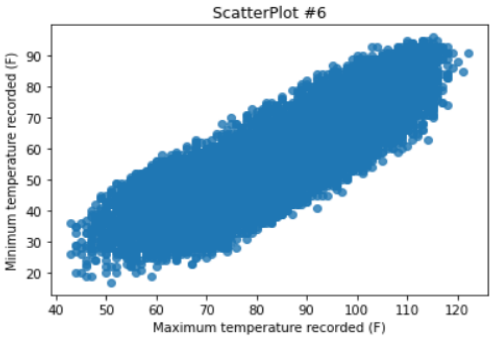

Fig-6. Scatterplot between Maximum and minimum temperature. It shows a high correlation between the two variables.

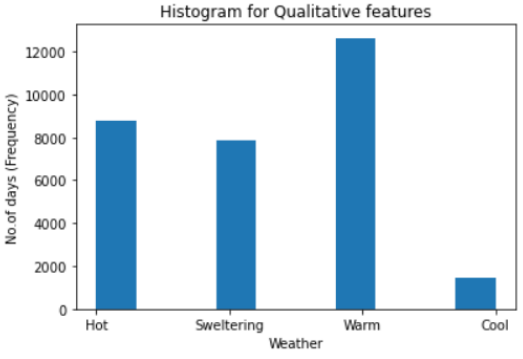

Fig-7. Histogram between the frequency and weather. It shows how many hot,sweltering, warm and cool days occurred over the given period from 1933 to 2022.

The summary statistics of the Quantitative data and Qualitative data present in the data set are given below in table-1 and table-2 respectively.

	Precipitation	Snow	Temp_ max	Temp_ min
Count	30753	30753	30753	30753
Mean	0.020119	0.00003	86.3354	60.0791
Std	0.108943	0.00307	16.0884	15.6935
Min	0	0	43	17
25%	0	0	73	47
50%	0	0	87	59
75%	0	0	101	74
Max	3.3	0.4	122	96

Table-1. Summary statistics of Quantitative data

Warm	12637
Hot	8801
Sweltering	7870
Cool	1445

Table-2. Summary statistics of Qualitative data describing the occurrence of each label

3 METHODOLOGY

In this paper, four types of regression algorithms are used. They are linear regression, LASSO, ridge, and ElasticNet regression. The simplest type of regression is linear-regression, which assumes-that the predictors have-a linear-relationship with the target-variable. After acquiring the dataset, the dataset is preprocessed by filling in missing values and removing null values. etc. Then the data set is split into train and test data, where 80% of the data is split into train data, and 20% into test data. Then, the regression algorithms are applied after training.

Ridge regression is a linear regression extension in which the loss-function is adjusted to reduce model complexity. This is accomplished by including a penalty parameter equal to the square of the coefficient magnitudes. Alpha is the parameter we must choose for the loss function. A high alpha value can cause under-fitting whereas a low alpha value might cause over-fitting. I used the alpha value of 0.05 for this model.

LASSO modifies the loss function to reduce model complexity by restricting the sum of the absolute values of the model coefficients. I used the alpha value of 0.05 for this model.

ElasticNet combines the characteristics of both the Ridge and Lasso regression methods. The ElasticNet class is used in scikit-learn to build an ElasticNet regression model. I used the alpha value of 0.05 in this model as well.

Then various evaluation metrics for each of these models are calculated to find out the best model among them.

4 RESULTS AND DISCUSSION

Linear Regression is the best model among all the models, because it gives the least error values (MSE, MAE, RMSE), and among the alternative models to linear Regression, Ridge is the best performing model by looking at the values of the metrics. Ridge model metrics are almost similar to those of the linear regression model, varying slightly.

Regression	MSE	MAE	RMSE
Linear	22.122105461	3.5081608922	4.703414234
Ridge	22.122111485	3.5081621797	4.703414874
LASSO	22.180968067	3.5165656563	4.709667511
Elastic-Net	22.180260738	3.5163317251	4.709592417

Table-3. Evaluation metrics of various models used in this paper.

For validating the results, I considered the row with the Date '21-08-1933', the Precipitation is '0.01' and snow is '0', snow_depth is '0', the maximum temperature is '102', the minimum temperature is '77' and the Weather is 'Sweltering'. For this record, the predicted value obtained through Linear regression is 101.4961215 F, which is pretty close to the actual temperature of the day 102 F, this output seems valid because, on a day with maximum temperature of 102F, and minimum temperature of 77F, it seems pretty legit that the day's temperature will be somewhere between those values and also the date falls in the month of june, which is summer and it sees the most hot and sweltering days in Phoenix, Arizona region.

Similarly, I considered the record with the Date '03-03-1983', the Precipitation is '1.98' and snow is '0', snow_depth is '0', the maximum temperature is '60', the minimum temperature is '52' and the Weather is 'Cool'. For this record, the predicted value obtained through Linear regression is 59.00156678 F, which is pretty close to the actual temperature of day 58 F, this output seems valid because, on a day with a maximum temperature of 60F, and minimum temperature of 52F, it seems pretty legit because, with precipitation of 1.98 inches, and temperatures between 60 and 52F, the day seems cool. Also, the date falls in the month of March, which remains cool in the Phoenix, Arizona region. Both the above results suggest that the results of the prediction are reliable.

In the dataset, I found two interesting features (A & B) which affect the predictions. They are 'precipitation' and 'Temp_max'. The 'Precipitation' (A) feature is the rainfall that occurred on a day. It affects the temperature/weather of the day in such a way that if the precipitation increases, the temperature of the day decreases and vice-versa. This seems pretty straightforward as rainfall usually cools the weather. The 'Temp_max' (B) feature is the maximum temperature recorded on a day. Weather generally depends on the maximum and minimum temperatures of the day. If the max temperature of the day is high, then the probability that the weather/ overall temperature of the day is hot/high.

If 'precipitation'/A is increased, then the predicted temperature decreases. Example: I took precipitation value '2' for the #record 2 explained above (Temp_max=60, Temp_min=52), and the output became 58.9721 F from 59.0015 F. Similarly, if the 'precipitation' value decreases, then the predicted temperature increases. Ex: If 'A'=1.5 for record 2 as mentioned above, the predicted temperature became 59.7065 F. If 'Temp_max'/ B is increased, then the predicted temperature increased. Example: I took Temp_max value '65' for the #record 2 (precipitation=1.98, Temp_min=52) as mentioned above, the output became 63.5751 F from 59.0015 F. Similarly, if the 'Temp_max' value decreases, then the predicted temperature decreases. Ex: If 'B'=55 for record 2 as mentioned above, the predicted temperature became 54.4280 F. Similarly, if both A and B features are increased, then the predicted value increased. If A and B are varied inversely, then the output depends on the B (Temp_max) feature.

5 CONCLUSION

From the results, it can be concluded that the weather is greatly dependent on the Temp_max feature and identified the trends and relationship of various weather variables. From the data, we can identify that the daily minimum temperatures and daily maximum temperatures are increasing from 1933 to 2022. The increasing daily minimum temperatures- which usually occur at night are warming up quicker than daily max temperatures. This shift will

provide less time to cool down and recuperate from the heat of the daily max temperature or day . This will result in an increase in the heat index, which ultimately causes heat waves. Due to this increase in daily max and min temperatures continuously, heat waves are projected to grow more frequent and stronger as the climate warms generally. This study can provide some scope to correlate the increasing temperatures with global warming issues like drought, green house emissions, poor air quality. Etc and necessary measures can be taken by the Phoenix government to protect the underprivileged people from heat waves.

ACKNOWLEDGMENT

I would like to express my gratitude to Professor Ruben Acuna for his continuous support and for providing me with the learning opportunity through this project.

REFERENCES

[1] Deepti Mishra, Pratibha Joshi "A Comprehensive Study on Weather Forecasting using Machine Learning (DOI-10.1109/ICRITO51393.2021.9596117)"

[2] Alexander, L.V., Zhang, X., Hegerl, G., Seneviratne, S.I., 2016. "Implementation Plan for WCRP Grand Challenge on Understanding and Predicting Weather and Climate Extremes "– the "Extremes Grand Challenge".

[3] Mulubrhan Balehegn , Selam Balehey, Chao Fu and Wu Liang. "Indigenous weather and climate forecasting knowledge among Afar pastoralists of northeastern Ethiopia: Role in adaptation to weather and climate variability"

[4] TanviPatil, Kamal shah (2021). "Weather Forecasting Analysis using Linear and Logistic Regression Algorithm. International Research Journal of Engineering and Technology (IRJET)"