Report: Weather Data Analysis and Ridge Regression Prediction for Weather parameters

Introduction

An analysis was conducted on weather data from New Delhi, focusing on daily climate variables such as temperature, humidity, wind speed, and pressure. The primary aim was to explore these variables over time and to build a predictive model for future mean temperatures, humidity, wind speed and mean pressure using Ridge Regression.

Data Overview

The dataset DailyDelhiClimateTrain.csv was imported, containing the following columns:

- *date*: The recorded weather observation date.
- *meantemp*: The mean temperature recorded for the day.
- *humidity*: The day's humidity level.
- *wind_speed*: The wind speed recorded on the day.
- *meanpressure*: Atmospheric pressure on the day.

Initial data exploration was performed using *df.head(), df,tail(), df.describe(),* and *df.info()*. Weather observations from 01-01-2013 to 01-01-2017 were included, with 5 main columns and 1462 entries.

	date	meantemp	humidity	wind_speed	meanpressure
0	2013-01-01	10.000000	84.500000	0.000000	1015.666667
1	2013-01-02	7.400000	92.000000	2.980000	1017.800000
2	2013-01-03	7.166667	87.000000	4.633333	1018.666667
3	2013-01-04	8.666667	71.333333	1.233333	1017.166667
4	2013-01-05	6.000000	86.833333	3.700000	1016.500000
	date	meantemp	humidity	wind_speed	meanpressure
1457	2016-12-28	17.217391	68.043478	3.547826	1015.565217
1458	2016-12-29	15.238095	87.857143	6.000000	1016.904762
1459	2016-12-30	14.095238	89.666667	6.266667	1017.904762
1460	2016-12-31	15.052632	87.000000	7.325000	1016.100000
1461	2017-01-01	10.000000	100.000000	0.000000	1016.000000

Fig: head and tail of dataset table

	meantemp	humidity	wind_speed	meanpressure
count	1462.000000	1462.000000	1462.000000	1462.000000
mean	25.495521	60.771702	6.802209	1011.104548
std	7.348103	16.769652	4.561602	180.231668
min	6.000000	13.428571	0.000000	-3.041667
25%	18.857143	50.375000	3.475000	1001.580357
50%	27.714286	62.625000	6.221667	1008.563492
75%	31.305804	72.218750	9.238235	1014.944901
max	38.714286	100.000000	42.220000	7679.333333

Fig. number of values and range of values in table

```
RangeIndex: 1462 entries, 0 to 1461
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   date          1462 non-null   object
 1   meantemp      1462 non-null   float64
 2   humidity      1462 non-null   float64
 3   wind_speed    1462 non-null   float64
 4   meanpressure  1462 non-null   float64
dtypes: float64(4), object(1)
memory usage: 57.2+ KB
```

Fig. check for null values and number of entries

Exploratory Data Analysis (EDA)

Line Plots

Plotly was used to visualize trends in weather variables:

1. Mean Temperature: A line plot showed the daily variation in mean temperature over the years. There is a similar pattern over the year, so prediction of weather is possible.

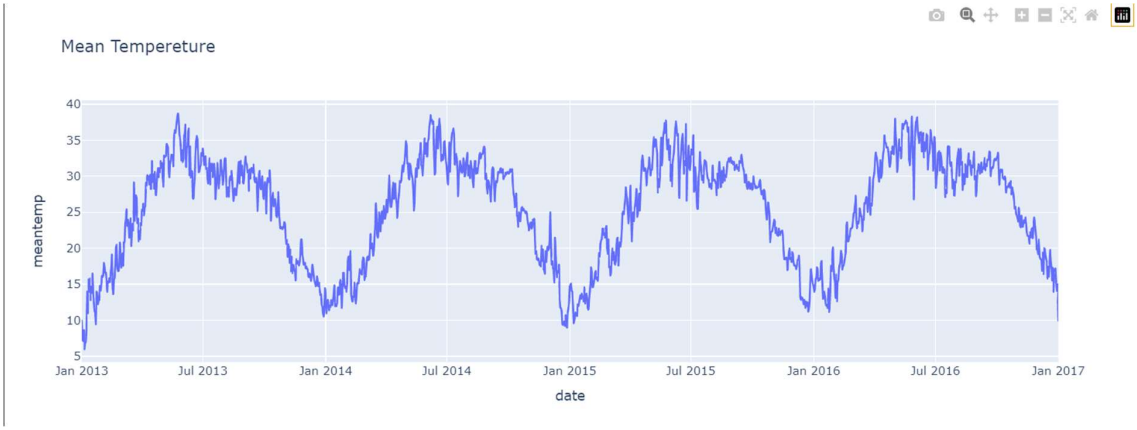

Fig: line plot of mean temperature with respect to date

2. Humidity: A similar line plot displayed humidity fluctuations over time. It also had repetitive pattern over the years.

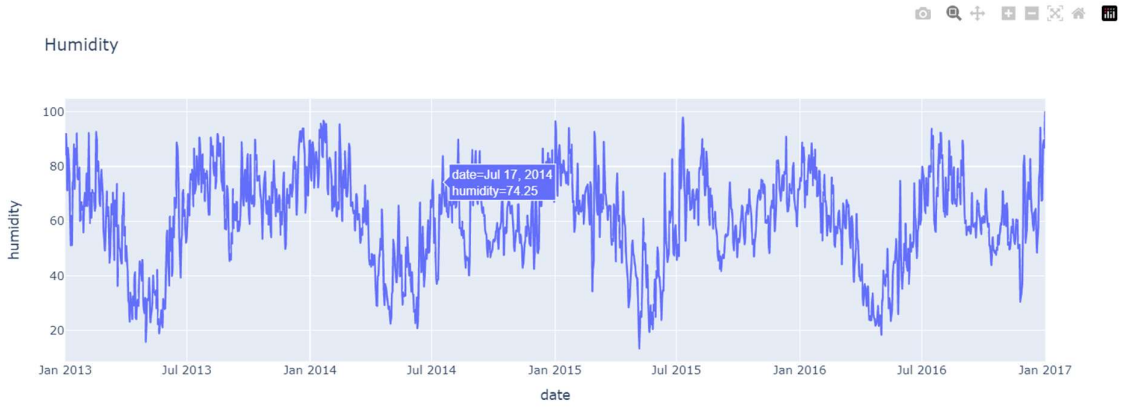

Fig: line plot of humidity with respect to date

3. Wind Speed: A similar line plot displayed wind speed fluctuations over time. It also had repetitive pattern over the years.

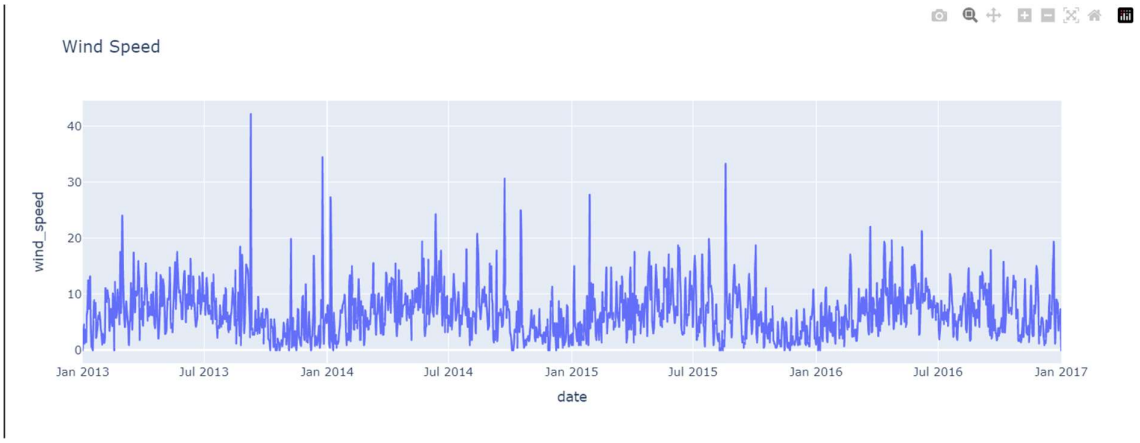

Fig: line plot of wind speed with respect to date

4. Mean Pressure: A similar line plot displayed mean pressure fluctuations over time. It also had repetitive pattern over the years. There were some abnormality which are difficult to predict but for most part it had same pattern.

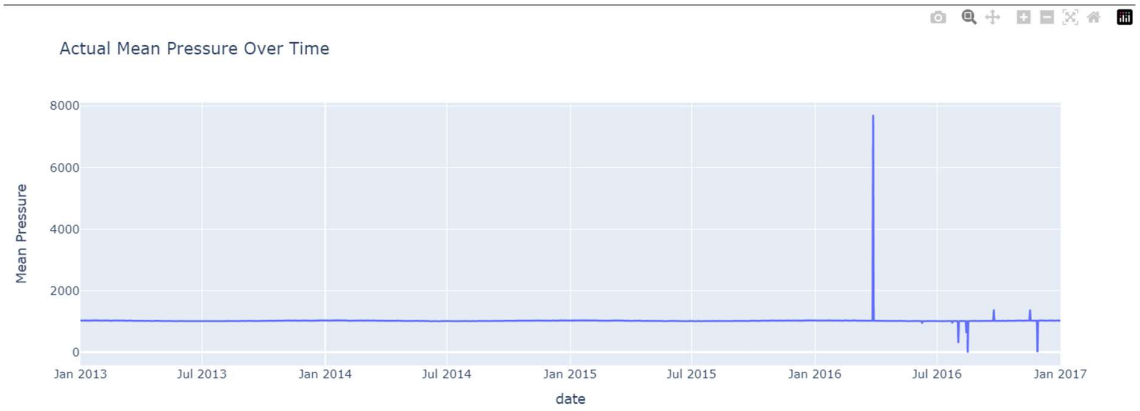

Fig: line plot of mean pressure with respect to date

Expanding Date to month and year

Since temperature, wind speed, pressure and humidity vary with month and year so it will be great to add month and year for feature vector.

Correlation Check

Correlation check is important to choose the feature vectors non-correlated vector does not add value to the predictions and will only increase training time and predictions time which should be avoided.

	date	meantemp	humidity	wind_speed	meanpressure	year	month	target
date	1.000000	0.130454	-0.050036	-0.024733	0.013823	0.968247	0.245594	0.125456
meantemp	0.130454	1.000000	-0.571951	0.306468	-0.038818	0.103803	0.122667	0.974146
humidity	-0.050036	-0.571951	1.000000	-0.373972	0.001734	-0.071381	0.074950	-0.543233
wind_speed	-0.024733	0.306468	-0.373972	1.000000	-0.020670	0.015642	-0.160668	0.287907
meanpressure	0.013823	-0.038818	0.001734	-0.020670	1.000000	0.022501	-0.035055	-0.034325
year	0.968247	0.103803	-0.071381	0.015642	0.022501	1.000000	-0.003642	0.102736
month	0.245594	0.122667	0.074950	-0.160668	-0.035055	-0.003642	1.000000	0.106894
target	0.125456	0.974146	-0.543233	0.287907	-0.034325	0.102736	0.106894	1.000000

Fig: correlation values all feature with respect to each other

Scatter Plot

The correlation values can be visualized using scatter plot. For temperature and humidity this can be done for all other features.

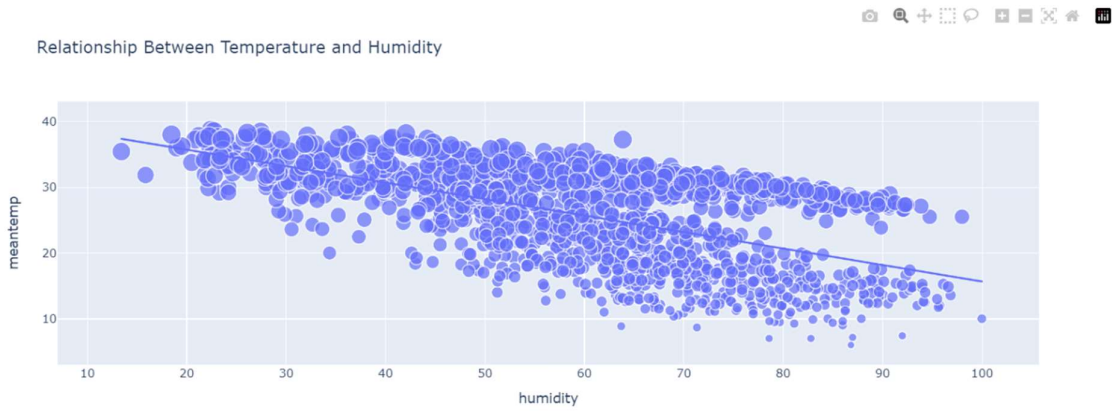

Fig: scatter plot for temperature and humidity

A scatter plot with a trendline was generated to explore the relationship between mean temperature and humidity. The plot suggests a negative correlation, with higher temperatures typically associated with lower humidity.

Seasonal Temperature Change

Using **Seaborn**, a line plot was generated to observe temperature changes across months for different years. It was shown that temperatures were highest during the summer months and lowest during winter, following a yearly pattern.

Feature Engineering

- **Date Conversion**: The date column was converted into the datetime format. Additional features such as **year** and **month** were extracted to capture annual and seasonal trends.
- **Target Variable** A new target column for the model was created by shifting the meantemp column by one day, allowing the prediction of the next day's mean temperature based on the current weather conditions.

Ridge Regression Model

Data Preparation

The features used for training the model include:

- meantemp
- humidity
- wind_speed
- meanpressure

- year
- month

The target variable was the next day's mean temperature. Missing values were handled using forward-fill.

The date column was converted to datetime format to enable the extraction of additional time-based features such as **year** and **month**. The next day's values for mean temperature, humidity, wind speed, and pressure were shifted into new columns (meantemp_next, humidity_next, wind_speed_next, meanpressure_next) to serve as target variables for prediction.

TimeSeriesSplit & GridSearchCV

Since the dataset involves time-series data, we used TimeSeriesSplit with 5 splits to ensure that the training set is always earlier in time than the test set. We performed hyperparameter tuning for the Ridge Regression model using GridSearchCV, evaluating it with negative mean squared error as the scoring metric. The best alpha value selected through cross-validation was 1.0.

Feature Selection and Target Variables

The feature matrix X included:

- meantemp: Mean temperature of the day.
- humidity: Humidity for the day.
- wind_speed: Wind speed for the day.
- meanpressure: Atmospheric pressure for the day.
- year: The year of the observation.
- month: The month of the observation.

The target matrix Y was constructed using the next day's values of the weather variables:

- meantemp_next: Mean temperature of the next day.
- humidity_next: Humidity of the next day.
- wind_speed_next: Wind speed of the next day.

- meanpressure_next: Mean pressure of the next day.

Model Training and Evaluation

Ridge Regression was applied to predict each of the target variables. A range of alpha values for regularization was tested using **GridSearchCV** with **TimeSeriesSplit** cross-validation. This ensures that the model does not use future data points during training. The best alpha values were selected for each target variable based on the performance of the model.

The best alpha values found were:

- For **meantemp_next**: 1.0
- For **humidity_next**: 100
- For **wind_speed_next**: 0.1
- For **meanpressure_next**: 100

```
Best alpha for meantemp_next: 0.1
RMSE for meantemp_next: 1.6487913437751103
Best alpha for humidity_next: 100.0
RMSE for humidity_next: 7.963394902536819
Best alpha for wind_speed_next: 0.1
RMSE for wind_speed_next: 3.9551060636492696
Best alpha for meanpressure_next: 100.0
RMSE for meanpressure_next: 179.86405343965092
```

Visualization of Predictions

Using **Plotly**, actual vs predicted values were visualized for each of the weather parameters:

1. **Mean Temperature**: The actual and predicted mean temperatures closely followed each other with some deviations.
2. **Humidity**: Predictions aligned well with the actual humidity, though some fluctuations were noted.
3. **Wind Speed**: Predictions were in line with the actual wind speed, with lower variation.
4. **Mean Pressure**: Predicted values for pressure matched the actual trends. There is abnormality which were not predictable with these features.

Fig: predicted values and actual values of temperature

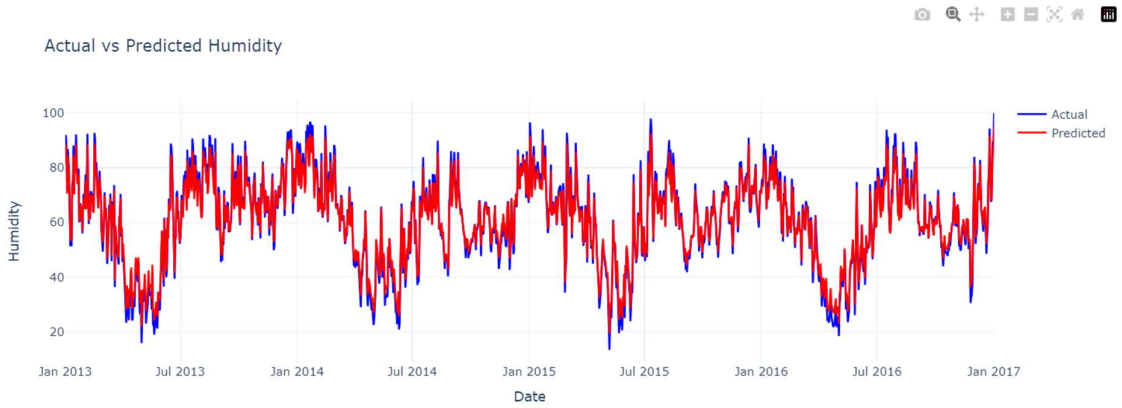

Fig: predicted values and actual values of humidity

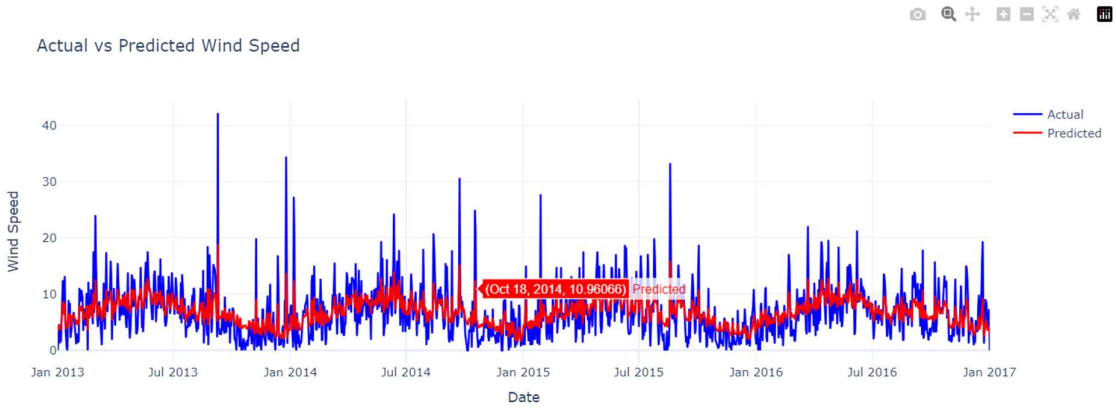

Fig: predicted values and actual values of speed

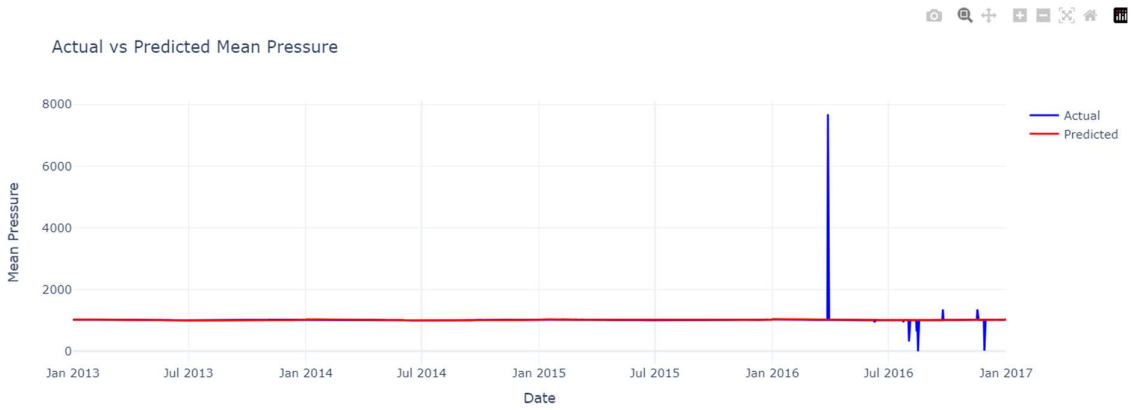

Fig: predicted values and actual values of mean pressure

Detailed Metrics

1. **Mean Temperature**
    - Mean Squared Error (MSE): 0.07
    - Root Mean Squared Error (RMSE): 0.27
    - Mean Absolute Error (MAE): 0.21
    - R-squared (R^2): 1.00
    - Explained Variance Score: 1.00
    - Mean Absolute Percentage Error (MAPE): 1.02%

The model performed exceptionally well for mean temperature, achieving near-perfect results, with an R^2 of 1.00, indicating that the model explains almost all the variance in the temperature data. The MAPE of 1.02% indicates minimal error between the predicted and actual values.

2. **Humidity**
    - Mean Squared Error (MSE): 4.65
    - Root Mean Squared Error (RMSE): 2.16
    - Mean Absolute Error (MAE): 1.74
    - R-squared (R^2): 0.98
    - Explained Variance Score: 0.98

- Mean Absolute Percentage Error (MAPE): 3.57%

The model performed well in predicting humidity with a high R^2 of 0.98. The RMSE and MAE indicate slight deviations in predictions, but overall, the model effectively captures the humidity trend.

3. **Wind Speed**
    - Mean Squared Error (MSE): 7.76
    - Root Mean Squared Error (RMSE): 2.79
    - Mean Absolute Error (MAE): 2.04
    - R-squared (R^2): 0.63
    - Explained Variance Score: 0.63
    - Mean Absolute Percentage Error (MAPE): inf%

The wind speed predictions show a moderate R^2 of 0.63, suggesting that the model captures some of the variance, but its performance could be improved. The infinite MAPE value suggests some significant discrepancies in the predictions, potentially due to outliers or small actual values causing divisions by zero.

4. **Mean Pressure**
    - R-squared (R^2): 0.00
    - Explained Variance Score: 0.00
    - Mean Absolute Percentage Error (MAPE): 29.32%

The model failed to accurately predict mean pressure, with an R^2 of 0.00, meaning it does not explain any variance in the data. The MAPE of 29.32% indicates a large prediction error. Further analysis of the pressure data or alternative models may be needed for improvement.

Conclusion

The Ridge regression model performed well for predicting temperature and humidity, with near-perfect accuracy for mean temperature. However, the model struggled with wind speed and mean pressure predictions, especially mean pressure, which showed no correlation between actual and predicted values. The infinite MAPE for wind speed also signals an issue with extreme values. Future work could involve refining the model, addressing outliers, or using more advanced machine learning techniques for wind speed and pressure prediction.