PREDICTING DAILY TEMPERATURE AND PRECIPITATION IN MUMBAI

PROJECT CONTENTS

1. Business Goal
2. Data Transformation
3. Multivariate Regression
4. Neural Network Model
5. Conclusions

BACKGROUND

Understanding and predicting temperature fluctuations and likelihood of precipitation are crucial for various industries and sectors, particularly in regions like Mumbai, where weather patterns can significantly impact daily life and economic activities.

BUSINESS QUESTION

Can we predict temperature and likelihood of precipitation in Mumbai based on other weather conditions?

DATA

Daily Mumbai temperature data with humidity, dew, sea level pressure, precipitation and wind speed and direction

- **Dependent Variable**: Temperature
- **Independent Variables:**
  - Sea level Pressure
  - Humidity
  - Dew
  - Wind Speed
  - Wind direction
  - Precipitation (Categorical)

DATA TRANSFORMATION

The categorical precipitation variables with two classes – Yes and No have been transformed to 1 (If “Yes”) and 0 (if “No”) through encoding
For neural networks, to ensure all independent variables are of the same range, we used the formula: (value – min)/(max – min) for range standardization

MULTIVARIATE LINEAR REGRESSION

REGRESSION METHODOLOGY

1

INITIAL MODEL

We use the variables which are present in the dataset with a lagged time frame of 1 which gives an **R-Squared of 83.4%**

2

MODEL ENHANCEMENT

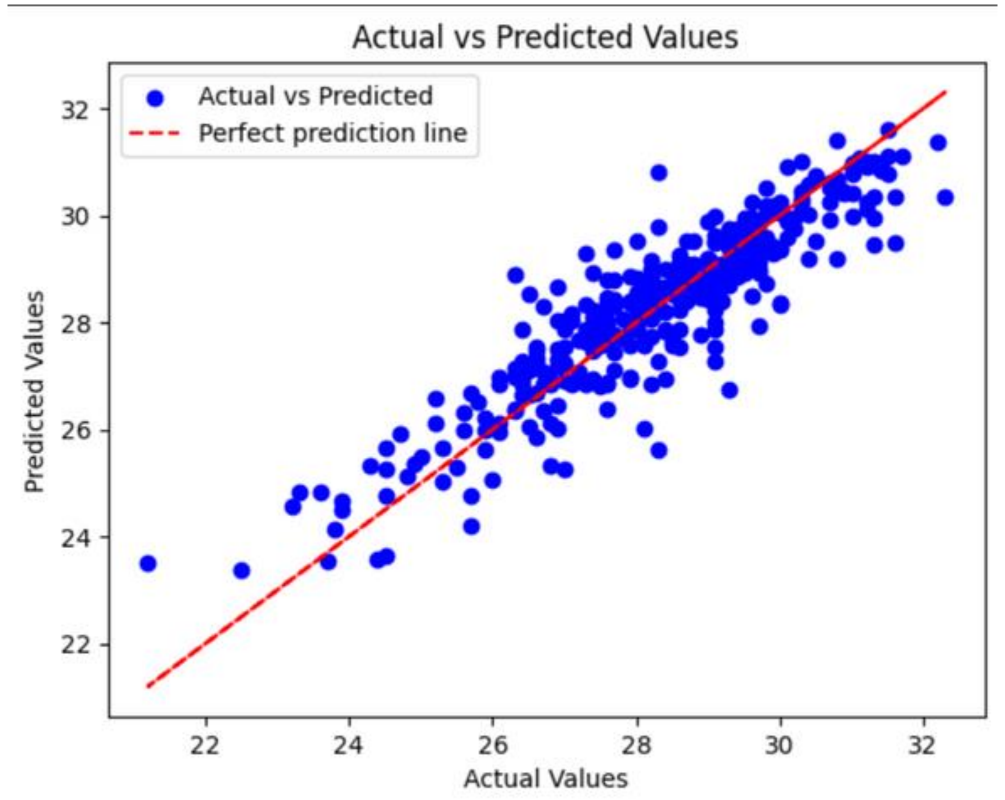
We then create variables used in STEP 1 with a lagged time frame of 2 and use them as independent variables along with the variables used in STEP 1 which gives us **an R-Squared of 83.8%**

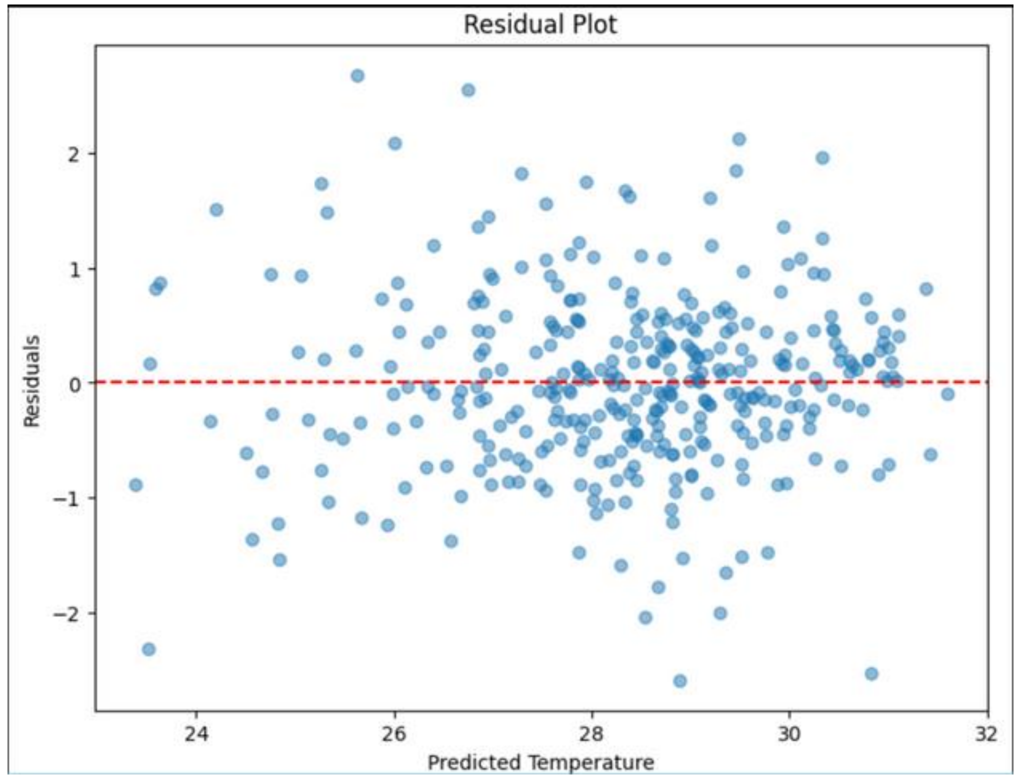
3

REFINEMENT & FINALIZATION

We then remove the variables which are having p-values lower than our alpha threshold and **re-run the model. Now our R-Squared is still 83.8%** and we decide to stop and this is our final model

LINEAR REGRESSION - MODEL 1



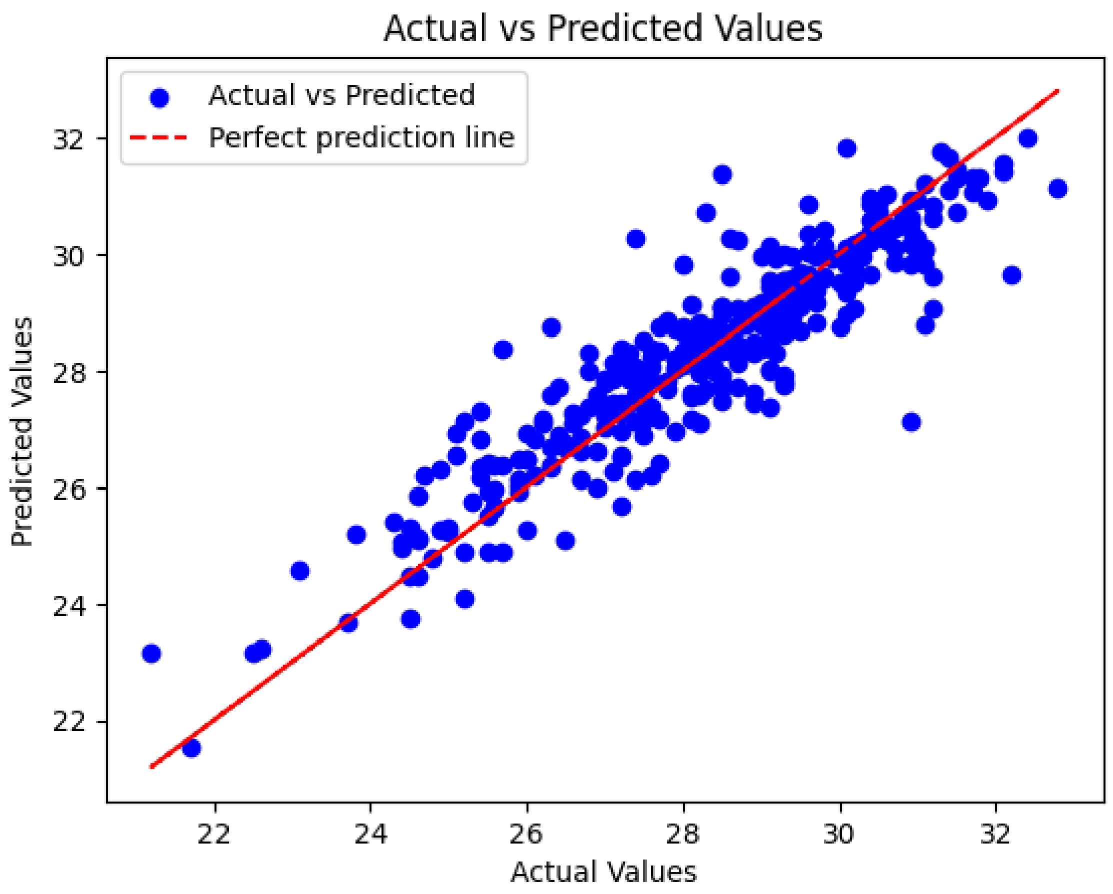


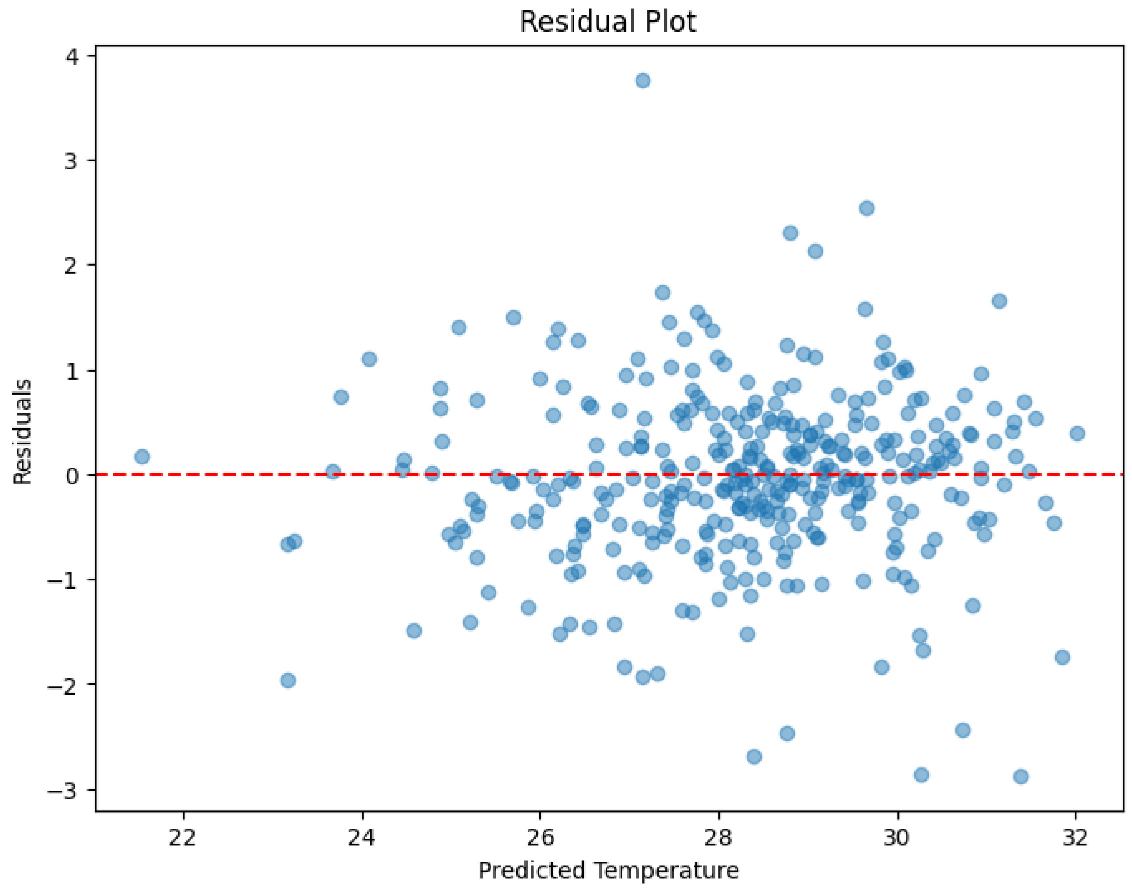
Regression Equation:

temp= -8.85+0.89 temp_lagged1+0.013 humidity_lagged1+0.01 sealevelpressure_lagged1-0.001 winddir_lagged1+0.002 solarradiation_lagged1-0.002 windspeed_lagged1

- **Dependent variable**: Temperature
- **Independent variables**: temp_lagged1, humidity_lagged1, sealevelpressure_lagged1, winddir_lagged1, solarradiation_lagged1, windspeed_lagged1
- **Positive Coefficients** (templagged1, humiditylagged1,sealevelpressure_lagged1, solarradiation_lagged1) A unit increase in any one of these variables is associated with a (1 unit* coefficient) increase in temperature, holding other factors constant.
- **Negative Coefficients** (winddir_lagged1,windspeed_lagged1) A unit increase in any one these variables is associated with a (1 unit* var. coefficient) decrease in temperature, holding other factors constant.
- **R-squared:** The model has an r-squared of 0.834 which indicates that the prediction is very accurate, explaining 83.4% of the variability in the model

LINEAR REGRESSION - MODEL 2





Regression Equation:

temp=-11.87 + 1.05 temp_lagged1 + 0.02 humidity_lagged1 + 0.05 sealevelpressure_lagged1 - 0.003 winddir_lagged1 + 0.0017 solarradiation_lagged1 - 0.167 temp_lagged2 - 0.0131 humditiy_lagged2 -0.041 sealevelpressure_lagged2 + 0.0026 * winddir_lagged2

- **Dependent variable**: Temperature
- **Independent variables**: 'temp_lagged1', 'humidity_lagged1', 'sealevelpressure_lagged1', 'winddir_lagged1', solarradiation_lagged1', 'windspeed_lagged1', 'temp_lagged2', 'humidity_lagged2', 'sealevelpressure_lagged2', 'winddir_lagged2', 'solarradiation_lagged2', windspeed_lagged2'
- **Positive Coefficients** (templagged1, humiditylagged1,sealevelpressure_lagged1, solarradiation_lagged1,winddir_lagged2) A unit increase in any one of these variables is associated with a (1 unit* var.coefficient) increase in temperature, holding other factors constant.
- **Negative Coefficients** (winddir_lagged1,temp_lagged2, humditiy_lagged2 ,sealevelpressure_lagged2) A unit increase in any one these variables is associated with a (1 unit* var.coefficient) decrease in temperature, holding other factors constant.
- **R-squared:** The model has an r-squared of 0.834 which indicates that the prediction is very accurate, explaining 83.4% of the variability in the model

FINAL MODEL

DEPENDANT VARIABLE

Temperature

REGRESSION EQUATION

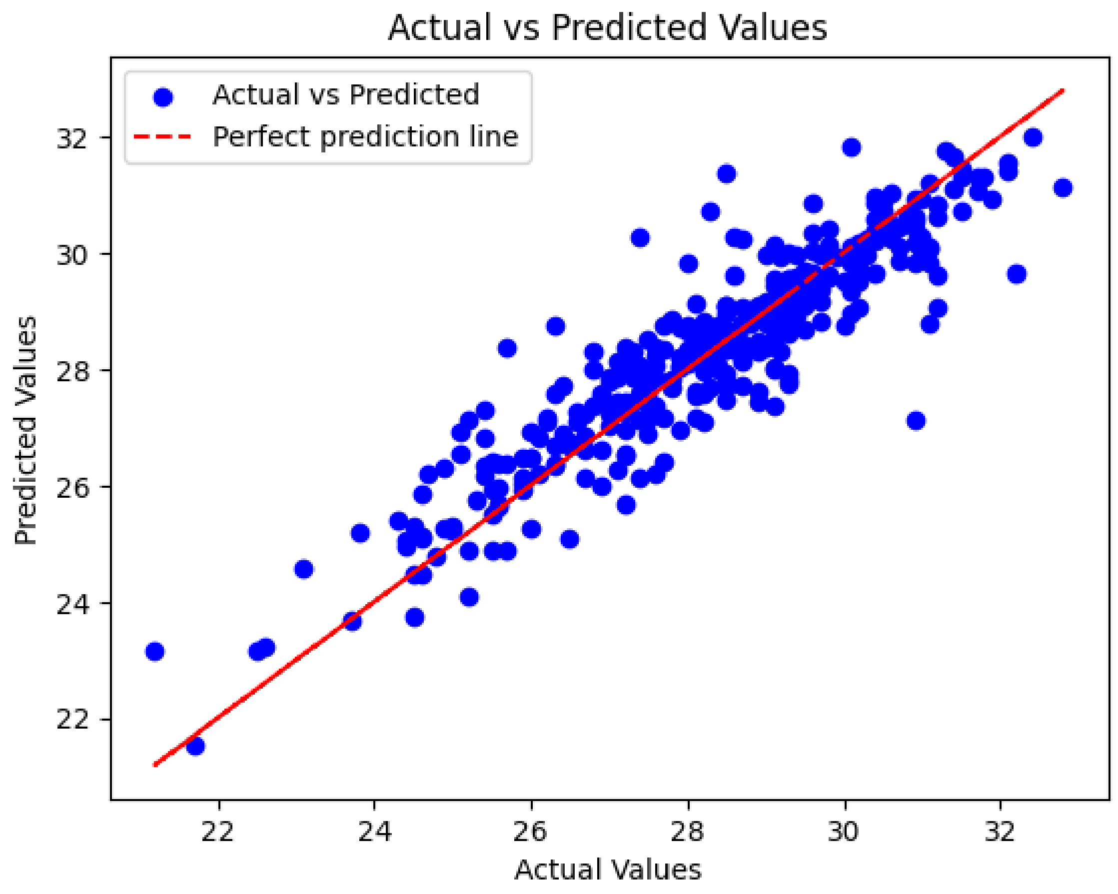
-11.87 + 1.05 temp_lagged1 + 0.02 humidity_lagged1 + 0.05 sealevelpressure_lagged1 - 0.003 winddir_lagged1 + + 0.0017 solarradiation_lagged1 - 0.167 temp_lagged2 - 0.0131 humditiy_lagged2 -0.041 sealevelpressure_lagged2 + 0.0026 * winddir_lagged2

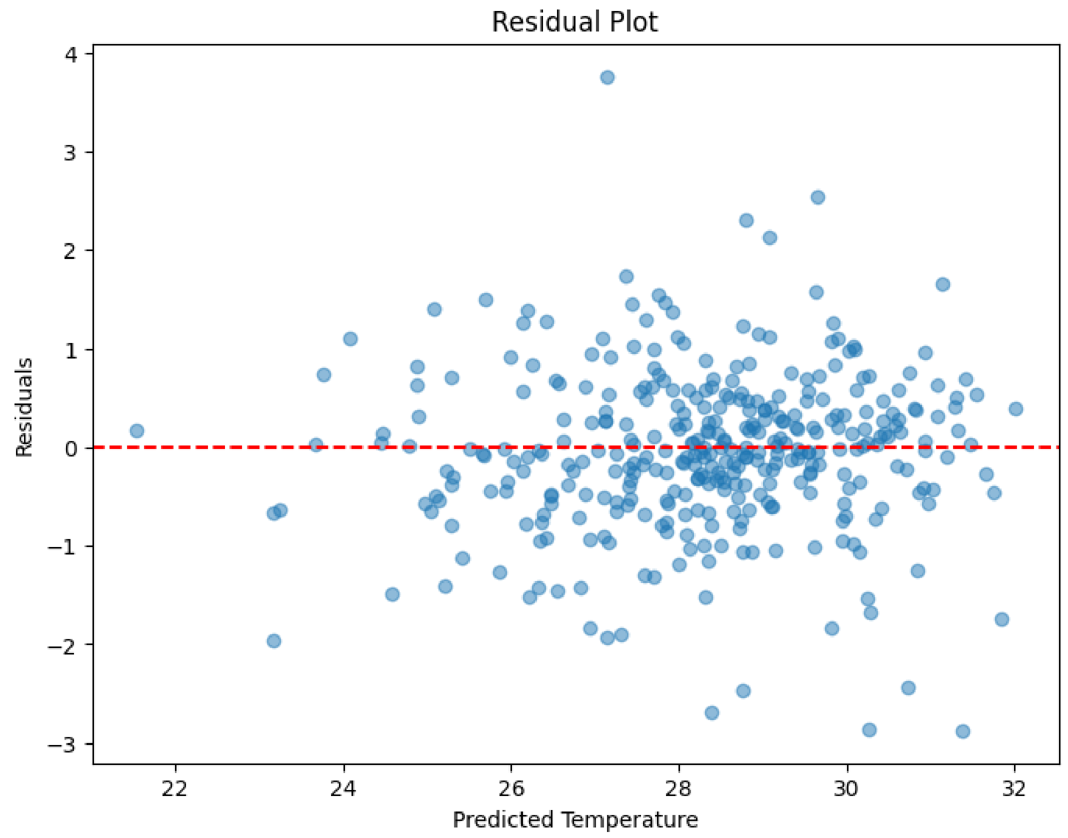
R - SQUARED VALUE

0.838

INDEPENDANT VARIABLES

temp_lagged1,humidity_lagged1, sealevelpressure_lagged1,winddir_lagged1, solarradiation_lagged1, temp_lagged2,humidity_lagged2, sealevelpressure_lagged2, winddir_lagged2





FINAL MODEL SUMMARY

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   temp   R-squared:                       0.838
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                     812.0
Date:                Mon, 08 Apr 2024   Prob (F-statistic):               0.00
Time:                        06:51:00   Log-Likelihood:                -1680.9
No. Observations:                1423   AIC:                             3382.
Df Residuals:                    1413   BIC:                             3434.
Df Model:                           9
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     -11.8737     10.872     -1.092      0.275     -33.200       9.453
temp_lagged1                1.0530      0.031     34.062      0.000       0.992       1.114
humidity_lagged1            0.0238      0.005      5.208      0.000       0.015       0.033
sealevelpressure_lagged1    0.0551      0.022      2.529      0.012       0.012       0.098
winddir_lagged1            -0.0028      0.001     -3.881      0.000      -0.004      -0.001
solarradiation_lagged1      0.0017      0.000      3.762      0.000       0.001       0.003
temp_lagged2               -0.1678      0.031     -5.503      0.000      -0.228      -0.108
humidity_lagged2           -0.0131      0.004     -2.945      0.003      -0.022      -0.004
sealevelpressure_lagged2   -0.0412      0.022     -1.864      0.063      -0.085       0.002
winddir_lagged2             0.0026      0.001      3.691      0.000       0.001       0.004
==============================================================================
Omnibus:                      122.567   Durbin-Watson:                   1.864
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              364.573
Skew:                          -0.431   Prob(JB):                     6.82e-80
Kurtosis:                       5.325   Cond. No.                     7.65e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.65e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Daily temperature in Mumbai can be predicted with an **accuracy of 83.8%** using the **significant positive predictors**: temp_lagged1, humidity_lagged1, sealevelpressure_lagged1, solarradiation_lagged1, winddir_lagged2 and the **significant negative predictors**: winddir_lagged1, temp_lagged2, humidity_lagged2

NEURAL NETWORK MODEL

METHODOLOGY

1

FEATURE SELECTION

Identified dependent (Precipitation: Rain/No Rain) and independent features for precipitation prediction. Standardized range of the independent variables

2

MODEL SELECTION

Selected the Neural network model using keras for the binary classification task with 2 outputs: Rain and No rain
Type of Layer: Dense with relu and softmax activation.
independent variables: 7
(Temperature, Dew, Humidity, Sealevelpressure, wind direction, wind speed, solar radiation)

3

MODEL TRAINING

The model was compiled using adam optimizer and sparse categorical cross-entropy as the loss function for binary classification. We set # epochs as 20. The intermediate Dense layers had 16 and 8 nodes respectively

4

MODEL EVALUATION

The model gave us daily predictions of rain with the associated probability of the outcome.
Accuracy, Sensitivity,Specificity, TPV and NPV were the measures used for evaluation of the model.

CLASSIFICATION MATRIX

Comprehensive summary of the performance of the model

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	711 **True Positive** **results correctly classified as rain.**	80 **False Positive** **wrongly classified as rain when it didn't rain.**
Negative (0)	102 **False Negative** **wrongly classified as no rain when it rained.**	888 **True Negative** **results correctly classified as no rain.**

PERFORMANCE METRICS OF THE MODEL

Provides an overall measure of how well the model is performing across the classes.

90%

Accuracy

Measures the proportion of correctly classified instances out of the total instances, reflecting the **overall correctness of the model's predictions.**

.

87%

Sensitivity

Measures the proportion of true positives out of all actual positive instances, showcasing the model's ability to capture all positive instances. Also known as **true positive rate (TPR) or recall**

90%

NPV

Measures the **accuracy of negative predictions** made by the model. It indicates the probability that a negative prediction made by the model is correct

90%

PPV

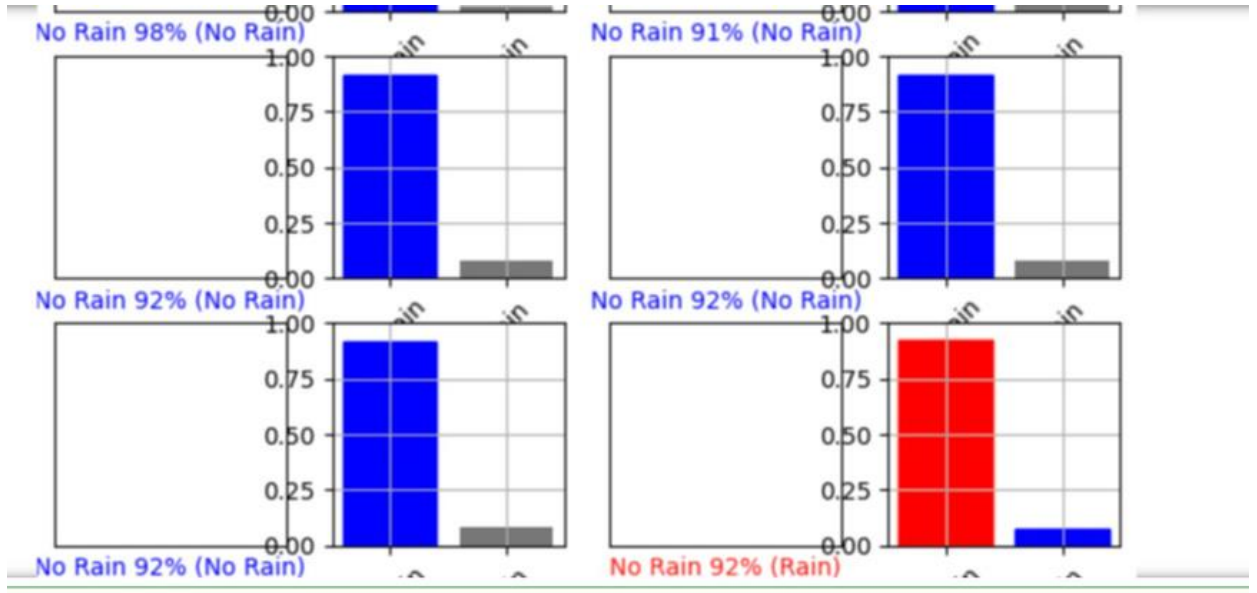
Measures the **accuracy of positive predictions** made by the model. It indicates the probability that a positive prediction made by the model is correct.

92%

Specificity

Measures the model's ability to **correctly identify negative instances**. It indicates the proportion of true negatives correctly identified by the model out of all actual negative instances.

Summary Plot and Prediction Output



Summary Plot

```
Out[26]: [array([[ 0.05661966,  0.19406575,  0.31341434, -0.21595106, -0.11777759,
            0.44053796, -0.48489684, -0.476045  ,  0.37950838, -0.00542914,
           -0.46397725,  0.17772977,  0.28990638,  0.57076347, -0.34182316,
           -0.15546471],
          [ 0.35191903,  0.20172974,  0.29135275, -0.0673306 ,  0.21605326,
            0.70223963,  0.1547805 , -0.35383573, -0.44025642,  0.16522676,
            0.41677958, -0.4250972 , -0.31299657,  0.6182369 , -0.49791718,
           -0.35874403],
          [ 0.15879838,  0.28434837,  0.17425701, -0.14034912, -0.3726715 ,
            0.3826853 ,  0.4794284 ,  0.29984975,  0.12450009, -0.57907873,
           -0.21322943,  0.14854357, -0.37074196, -0.00448209,  0.23851055,
           -0.2751124 ],
          [ 0.26681754,  0.34573275, -0.24059965, -0.3509642 ,  0.05811617,
           -0.27316594, -0.03696785,  0.12492216,  0.15875266,  0.25787106,
           -0.2579129 ,  0.5694015 ,  0.6558902 , -0.21839185,  0.30902234,
            0.356906  ],
          [ 0.14096926, -0.33020997, -0.17294882, -0.05801302,  0.27587947,
           -0.28831962,  0.21632318, -0.4686675 , -0.4042958 , -0.3197844 ,
            0.19362909,  0.10532106,  0.13529491, -0.11723153,  0.0792993 ,
           -0.32922792],
```

Layer 0 Weights

A	B	C	D	E	F	G	H	I	J	K	L
TempNorm	dewNorm	humidityNorm	sealevelpressureNorm	winddirNorm	solarradiationNorm	windspeedNorm	ClassLabelActual	ClassLabelPrediction	ClassLabelAccurate	NoRainProbability	RainProbability
0.650793651	0.306666667	0.132183908	0.957081545	0.327076677	0.587198849	0.060657119	0	0	1	0.996135473	0.003864544
0.523809524	0.355555556	0.232758621	0.991416309	0.178514377	0.585760518	0.060657119	0	0	1	0.994264424	0.005735585
0.420634921	0.426666667	0.347701149	0.927038627	0.318290735	0.605177994	0.075821398	0	0	1	0.987675667	0.012324288
0.492063492	0.36	0.25862069	0.909871245	0.244808307	0.587558432	0.060657119	0	0	1	0.991232932	0.008767067
0.547619048	0.377777778	0.234195402	0.871244635	0.239217252	0.558432219	0.060657119	0	0	1	0.986673713	0.013326288
0.531746032	0.413333333	0.281609195	0.862660944	0.178115016	0.532182668	0.045492839	0	0	1	0.978660285	0.021339711
0.468253968	0.533333333	0.422413793	0.901287554	0.334664537	0.492988134	0.053917439	0	0	1	0.95113951	0.048860501
0.507936508	0.511111111	0.382183908	0.91416309	0.367811502	0.475368572	0.030328559	0	0	1	0.963838458	0.036161542
0.484126984	0.377777778	0.272988506	0.845493562	0.207268371	0.540453074	0.060657119	0	0	1	0.981587887	0.018412132
0.46031746	0.333333333	0.239942529	0.862660944	0.236022364	0.542610572	0.060657119	0	0	1	0.98841244	0.011587594
0.468253968	0.426666667	0.313218391	0.86695279	0.351837061	0.574613448	0.123841618	0	0	1	0.977314055	0.022685969

Prediction File Output

FINDINGS AND RECOMMENDATIONS

1. **Temporal Correlation:** The regression analysis demonstrates a strong temporal correlation in temperature, with significant influences from its own past values. The positive coefficient for temp_lagged1 indicates that current temperature is positively influenced by its immediate past, while the negative coefficient for temp_lagged2 suggests a cooling effect following a period of higher temperatures.

2. **Positive Influences:** The positive coefficients for humidity, sea level pressure, and solar radiation at previous time points suggest that higher levels of these environmental factors are associated with increased temperatures. This indicates that changes in humidity, atmospheric pressure, and solar radiation contribute to temperature fluctuations over time.

3. **Business Value:** The high R-squared value of 0.834 suggests that the regression model provides a good fit to the data and effectively explains approximately 83.4% of the variability in temperature. The neural network model for rain also gave us high accuracy (90%), sensitivity (87%) and specificity (92%). These models help in accurate temperature and precipitation predictions which drive informed decision-making, reduce operational costs, and enhance customer satisfaction across tourism sector.

4. **Long term planning:** The identified temporal correlation highlights the lasting impact of temperature's past values, enabling businesses to anticipate future fluctuations. This insight along with the high accuracy in precipitation prediction through neural networks, supports the development of resilient long-term strategies, allowing organizations to adapt operations and invest in measures to withstand environmental changes effectively.

THANK YOU