DS 203: Programming for Data Science - Exercise 7

Swayam Saroj Patel (22B1816)

March 30, 2025

Data Analysis and Preprocessing:

So checking through the initial description of the data set we get this:

Statistic	id	recorded_day	air_pressure	max_temp	temperature	min_temp	dew_point	humidity	cloud_cover	wind_speed
count	2190	2190	2081	2180	2169	2180	2190	1971	1971	2169
mean	1094.50	183.00	1013.58	26.38	23.95	22.18	20.45	82.07	75.64	21.82
std	632.34	105.39	5.66	5.65	5.22	5.05	5.29	7.80	18.13	9.91
min	0.00	1.00	999.00	10.40	7.40	4.00	-0.30	39.00	2.00	4.40
25%	547.25	92.00	1008.60	21.30	19.30	17.70	16.80	77.00	69.00	14.20
50%	1094.50	183.00	1013.00	27.80	25.50	23.90	22.15	82.00	83.00	20.50
75%	1641.75	274.00	1017.80	31.20	28.40	26.43	25.00	88.00	88.00	28.00
max	2189.00	365.00	1034.60	36.00	31.50	29.80	26.70	98.00	100.00	59.50

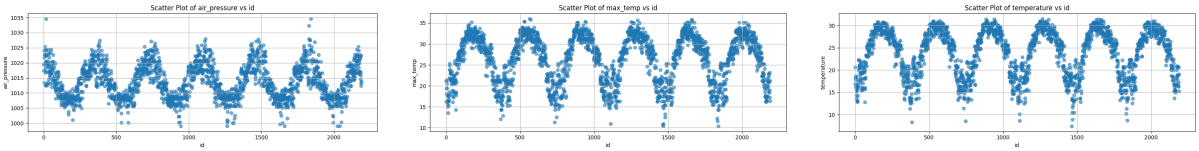
Table 1: Descriptive Statistics for the Rainfall Prediction Dataset

As one can see there are missing values in some columns and the exact number of missing values are:

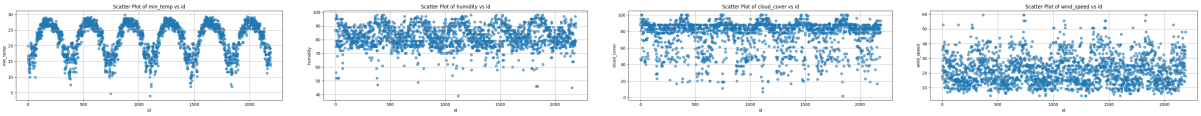
Column	**Missing Values**
id	0
recorded_day	0
air_pressure	109
max_temp	10
temperature	21
min_temp	10
dew_point	0
humidity	219
cloud_cover	219
wind_speed	21
raining	0

Table 2: Missing Values in Each Column

To decide how to deal with them we should take a look at their scatter plot and confirm if there are any trends present or not.



(a) Air Pressure (b) Max Temp (c) Temperature



(d) Min Temp (e) Humidity (f) Cloud Cover (g) Wind Speed

Figure 1: Scatter Plots for Selected Features

As we can see that more or less all of the features follow a trend so it would be a good idea to use techniques like interpolation or knns to fill in the missing point and move on. Here I have used KNN to fill the values.

PCA and T-SNE:

PCA:

PCA (Principal Component Analysis) is a linear dimensionality reduction technique. It projects your data onto directions (principal components) that explain the most variance.

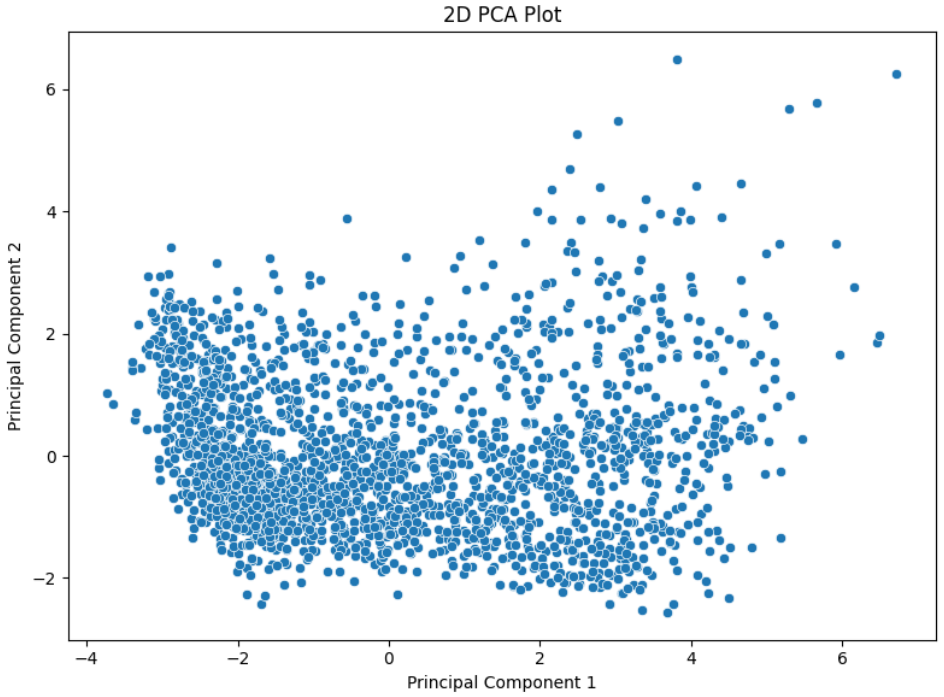


Figure 2: PCA plot

In the 2D PCA plot, the data points form a somewhat continuous area without any sharply defined clusters. This suggests:

- There is no obvious linear separation of any classes.
- The data variance is spread along a few principal components, but in 2D, the separation is not there.

T-SNE:

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a nonlinear dimensionality reduction technique designed primarily for visualization. It tries to preserve local neighborhoods, often revealing clusters or subgroups more distinctly than PCA.

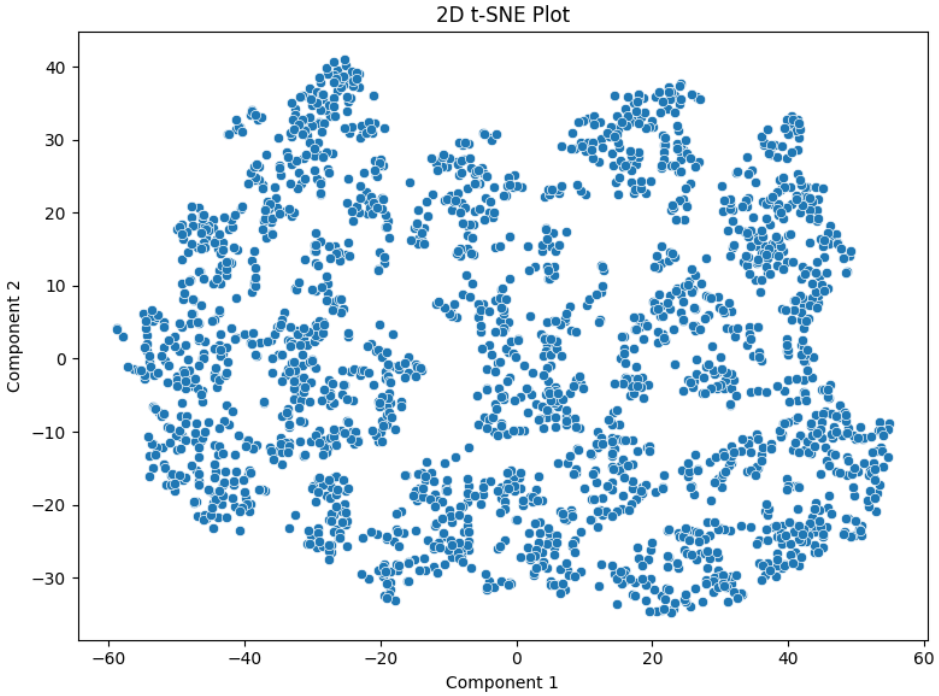


Figure 3: t-SNE plot

In the 2D t-SNE plot, we can see more pronounced clusters. This implies:

- The data may have nonlinear structure that the 2D PCA doesn't capture as clearly.
- There could be subpopulations or patterns in the data (e.g., different weather conditions) that do not appear as separate in a simple linear projection.

Thus t-SNE has better class separation in low dimensions.

VIF and PCA:

The **Variance Inflation Factor (VIF)** is a measure used to detect the degree of multicollinearity among predictor variables in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the presence of correlations among the predictors. In general:

- A VIF value of 1 indicates no correlation between a predictor and any other variables.
- A VIF between 1 and 5 suggests moderate correlation.
- A VIF of 10 or more is considered indicative of high multicollinearity, which can adversely affect the stability and interpretability of the model.

High multicollinearity means that the predictors share redundant information, leading to unstable coefficient estimates in the model. Removing or combining features with high VIF can help reduce overfitting, simplify the feature space, and improve the overall reliability of the model.

Feature Removal Based on VIF:

Feature	VIF
recorded_day	1.091347
air_pressure	3.551576
max_temp	32.710982
temperature	83.474037
min_temp	45.645627
dew_point	12.036188
humidity	1.867260
cloud_cover	1.785725
wind_speed	1.169384

Table 3: VIF for Each Feature

Based on the computed VIF values for our dataset, we observe that the features `max_temp`, `temperature`, and `min_temp` have the highest VIF values. Therefore, we consider three different strategies for feature removal:

- **Drop 3 Columns:** Remove the three most problematic features:
  - `max_temp`
  - `temperature`
  - `min_temp`
- **Drop 6 Columns:** In addition to the above, remove:
  - `dew_point`
  - `air_pressure`
  - `humidity`
- **Drop 8 Columns:** Further remove the following:
  - `cloud_cover`
  - `wind_speed`

The average ROC-AUC scores obtained for each model are as follows:

- **Full Feature Set:** 0.8800
- **Reduced (Drop 3 Columns):** 0.8787
- **Reduced (Drop 6 Columns):** 0.8614
- **Reduced (Drop 8 Columns):** 0.4532

The results indicate the following:

1. **Full Feature Set vs. Reduced (Drop 3 Columns):** The full feature set achieves an average ROC-AUC of 0.8800, while the model trained after removing the three most problematic features (`max_temp`, `temperature`, and `min_temp`) achieves an almost similar score of 0.8787. This suggests that these features may not be contributing significantly to the model's predictive performance, or that their information is largely redundant. Removing them reduces multicollinearity without sacrificing model accuracy.
2. **Reduced (Drop 6 Columns):** When we further drop additional features (`dew_point`, `air_pressure`, and `humidity`), the average ROC-AUC score drops to 0.8614. This indicates that while these additional features may be moderately correlated with other predictors.

3. **Reduced (Drop 8 Columns):** A reduction that drops eight columns (removing `max_temp`, `temperature`, `min_temp`, `dew_point`, `air_pressure`, `humidity`, `cloud_cover`, and `wind_speed`) results in a drastic performance decline, with an average ROC-AUC score of only 0.4532. This score is nearly equivalent to a random classifier (ROC-AUC ≈ 0.5), which demonstrates that an excessive reduction of features causes loss of critical predictive information.

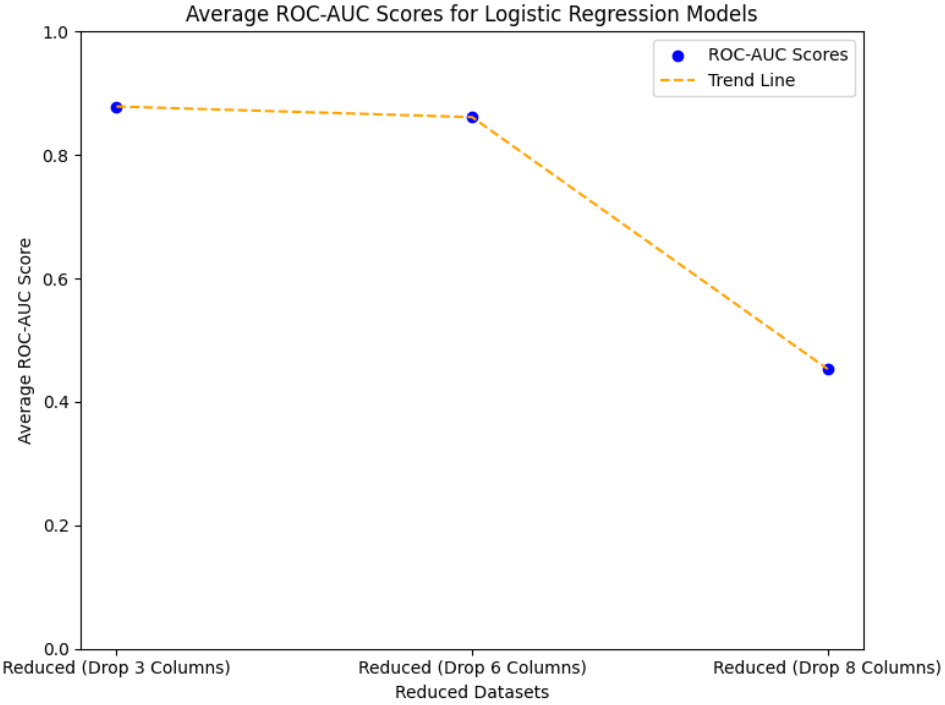


Figure 4: Plot of average ROC-AUC scores for different feature reduction strategies

PCA-Based Feature Reduction:

The average ROC-AUC scores obtained for each configuration are as follows:

- **PCA (1 Component):** 0.5610
- **PCA (3 Components):** 0.8717
- **PCA (6 Components):** 0.8792

1. **PCA (1 Component):** With only one principal component, the model achieves an average ROC-AUC of approximately 0.5610. This score is significantly lower than the full feature set performance.
2. **PCA (3 Components):** When retaining the top three principal components, the ROC-AUC increases substantially to about 0.8717. This indicates that these components collectively capture a large portion of the variability in the data that is relevant for classification, allowing the logistic regression model to perform nearly as well as with the full feature set.
3. **PCA (6 Components):** Further increasing the number of components to six results in a slightly higher ROC-AUC of approximately 0.8792. The incremental gain suggests that while additional components do capture more variance, much of the critical information for classification is already contained within the top three components.

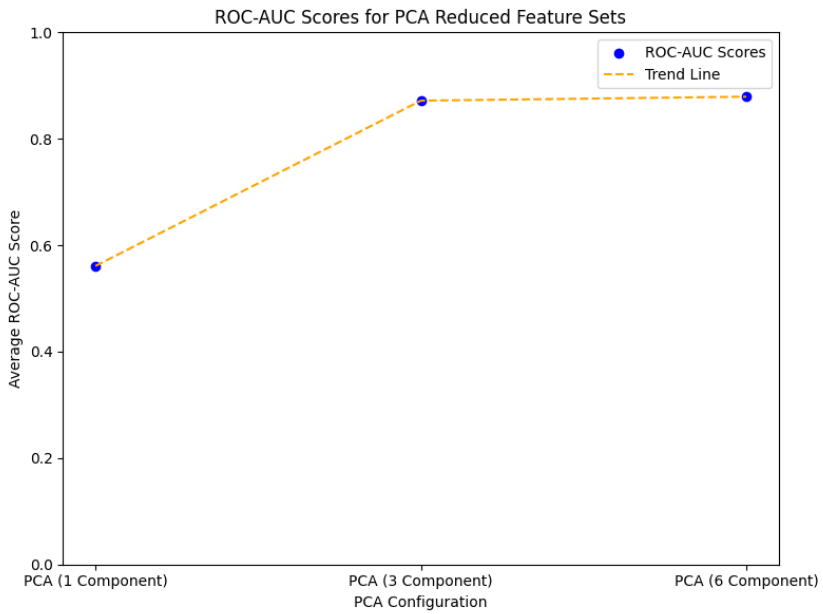


Figure 5: Plot of average ROC-AUC scores for PCA configurations: 1, 3, and 6 components.

Class Imbalance:

Based on the ROC-AUC comparisons and the need to balance model robustness with predictive power, the **Reduced (Drop 3 Columns)** dataset is the optimal choice for further implementation. It effectively addresses multicollinearity by removing the three most problematic features while preserving nearly the same predictive performance as the full feature set.

Choosing a Technique for Handling Class Imbalance

1. **Class Weighting:** This approach adjusts the cost function during model training by assigning a higher weight to the minority class. It makes the model more sensitive to misclassifications of the minority class without altering the actual data distribution.

2. **SMOTE (Synthetic Minority Over-sampling Technique):** This method creates synthetic examples for the minority class by interpolating between existing minority samples, thereby increasing the representation of the minority class.

Given that our baseline performance is strong and the imbalance does not appear to be extremely severe, the **class weighting** method is recommended for this assignment. It offers a simple yet effective way to enhance the sensitivity of our model towards the minority class without modifying the dataset. This approach not only helps in maintaining model stability but also avoids the potential risks associated with generating synthetic samples.

Comparison:

Observed ROC-AUC Scores

- **Before Handling Class Imbalance:** The best ROC-AUC obtained was approximately **0.8787**.
- **After Handling Class Imbalance:** With logistic regression using `class_weight='balanced'`, the ROC-AUC was **0.8781**.

Analysis and Interpretation

1. **Baseline Robustness:** The high ROC-AUC score of 0.8787 before balancing indicates that the model was already effective at distinguishing between the classes. The slight decrease to 0.8781 after balancing suggests that class imbalance was not severely affecting overall discrimination.

2. **Effect of Class Weighting:** By assigning greater weight to the minority class, the `class_weight='balanced'` parameter ensures that misclassifications in the minority class are penalized more heavily. Although the overall ROC-AUC did not improve, it rather decreased. But this technique may still enhance other performance metrics (e.g., recall or precision) for the minority class.

The near-identical ROC-AUC scores before and after applying class weighting suggest that the original model was robust. However, using class weighting is justified as it potentially improves the classifier's sensitivity to the minority class without sacrificing much of overall performance. This balanced approach ensures that both classes receive appropriate consideration during model training, which is particularly important in real-world applications where minority classes may represent critical outcomes.