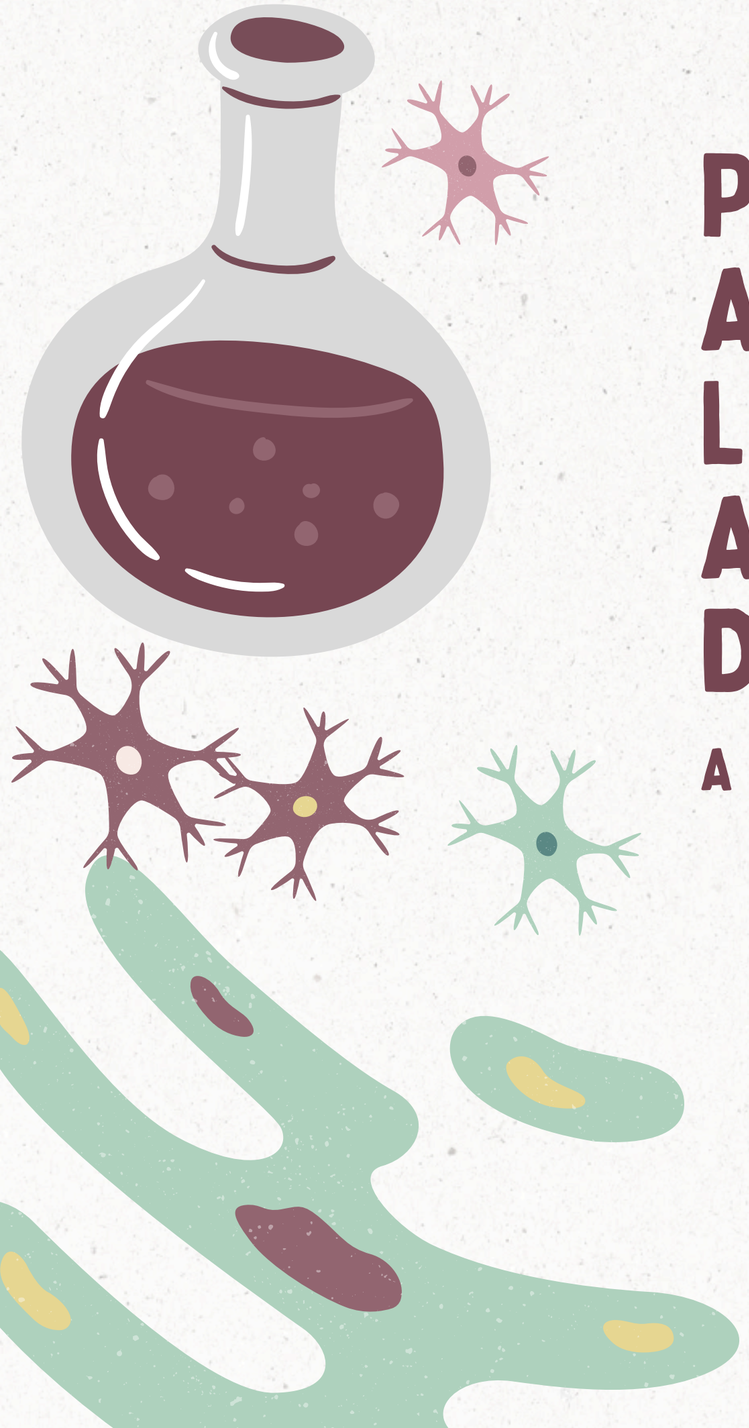


PREDICTING RAIN: A STATISTICAL AND MACHINE LEARNING APPROACH ANALYZING THE WEATHER-AUS DATASET

A STATISTICAL APPROACH USING LEVERAGE SCORE SAMPLING

PRATHAM BANSAL 2023392
PRIYANSHU SHARMA 2023408
RISHABH DWIVEDI 2023434
SIDHARTH KUMAR 2023526
UMANG AGGARWAL 2023567
VANSH TYAGI 2023582

PROBLEM STATEMENT

Context: Predicting RainTomorrow (Binary Classification) using the weatherAUS dataset (142k+ rows and 22 features)

Models Selected:

1. **Logistic Regression**
2. **Random Forest**

BASELINE PERFORMANCE

TRAINING DATA SIZE: N = 113,754

ASYMPTOTIC COMPLEXITY:

Logistic Regression: O (N.d) per iteration (where d = 22).

Random Forest: O (T.N.K.logN) (where T = no. of trees = 100, K = √d, where d=22)

EMPIRICAL RESULTS (FULL DATASEST):

Logistic Regression:

- Time: 13.22 seconds
- Accuracy: 83.90%

Random Forest

- Time: 3.69 seconds
- Accuracy: 83.47%

Scenario	Samples	Time (s)	Accuracy	F1 Score	Recall
Small Subsample (100)	100	0.0416	0.8007	0.7977	0.8007
Theoretical (499)	499	0.0799	0.8302	0.8235	0.8302
5% Scale	5687	1.2501	0.8390	0.8291	0.8390
10% Scale	11375	1.3160	0.8376	0.8269	0.8376
100% Scale	113754	13.2246	0.8390	0.8287	0.8390

Table 1: Logistic Regression Results

Scenario	Samples	Time (s)	Accuracy	F1 Score	Recall
Small Subsample (100)	100	0.0146	0.7156	0.7206	0.7156
Theoretical (753)	753	0.0194	0.7881	0.7815	0.7881
5% Scale	5687	0.1106	0.8045	0.7971	0.8045
10% Scale	11375	0.1743	0.8111	0.8033	0.8111
100% Scale	113754	3.6903	0.8347	0.8233	0.8347

Table 2: Random Forest Results

LEVERAGE SCORE SAMPLING TECHNIQUE

Weighted Randomized Sampling using Statistical Leverage Scores

WORKING

- We estimate the importance of each data point using sensitivity scores.
- Sensitivity upper bound:

$$s_i = a_i (A^\top A)^\dagger a_i^\top$$

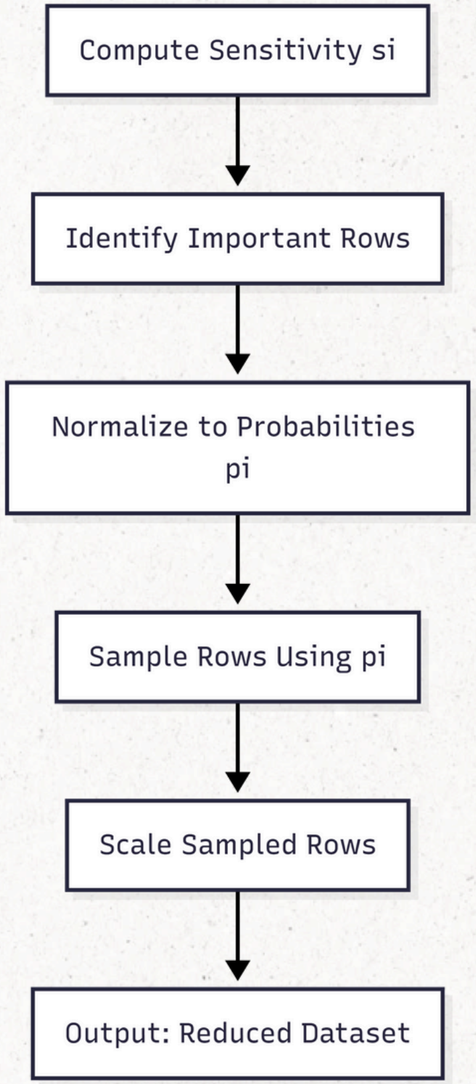
- Higher s_i indicates more influential row

  Lower s_i indicates less important row
- Convert sensitivities into sampling probabilities:

$$p_i = \frac{s_i}{\sum_j s_j}$$

- Sample rows according to p_i and scale them to preserve their contribution.

This identifies those cases that are unique (high leverage) and ensures they are picked, preserving the model's accuracy even with small dataset.



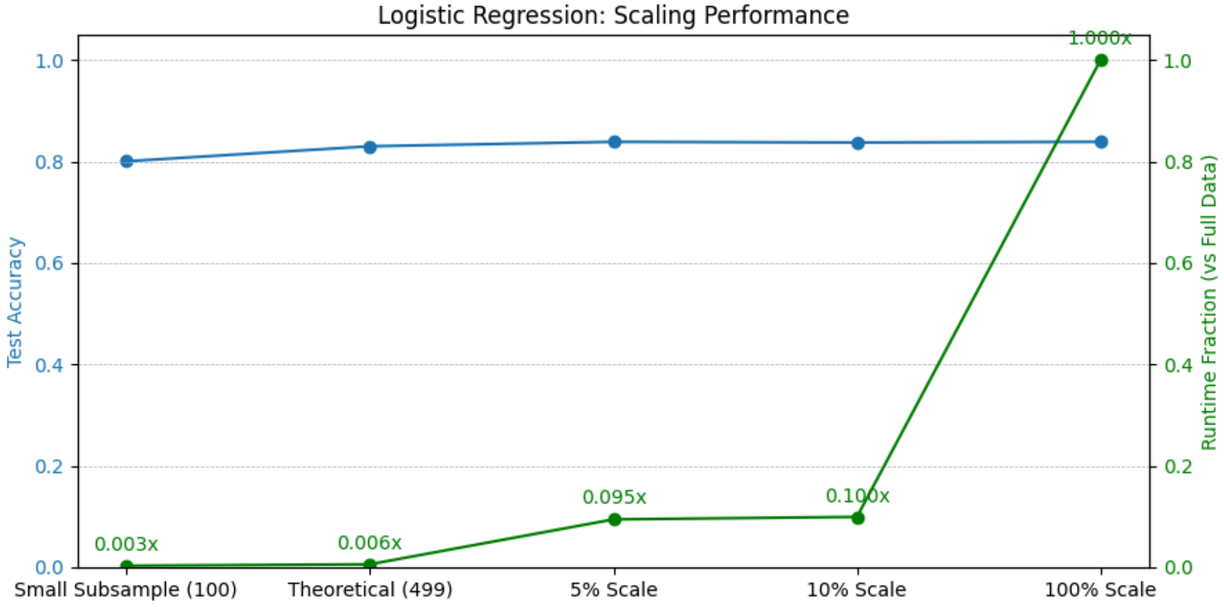
LOGISTIC REGRESSION ACCURACY COMPARISON AND TRAINING TIME IMPROVEMENT

Accuracy Comparison

- Theoretical Min (499 samples): 83.02% Accuracy
- Full Data (113,754 samples): 83.90% Accuracy
- Less than 0.9% drop in accuracy using only 0.4% of the data

Training Time Improvement

- Full Time: 13.22s to Scaled Time: 0.08s
- Improvement: ~165x faster



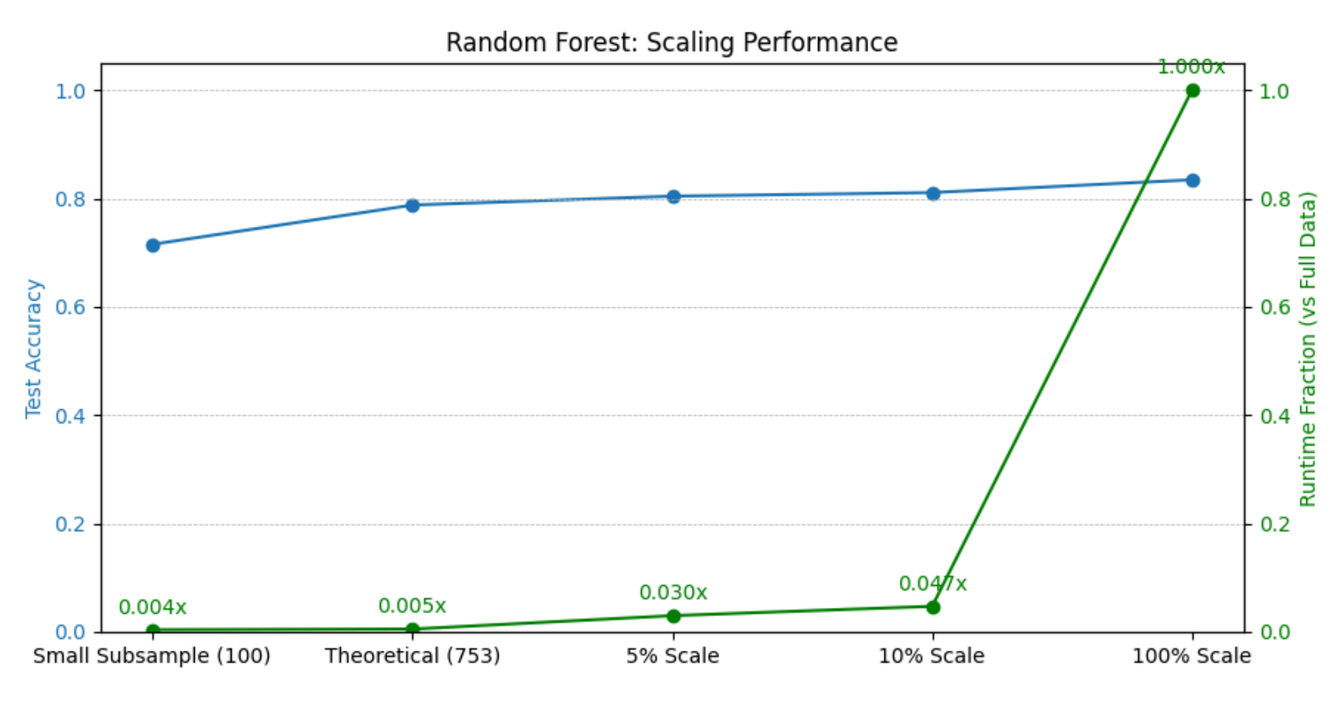
RANDOM FOREST ACCURACY COMPARISON AND TRAINING TIME IMPROVEMENT

Accuracy Comparison

- Theoretical Min (753 samples): 78.81% Accuracy
- Full Data (113,754 samples): 83.47% Accuracy
- Random forest require slightly more data (5% scale / 5687 samples) to hit 80%+ accuracy, showing they are more sensitive to data variance than linear models (Logistic Regression)

Training Time Improvement

- Full Time: 3.69s to Scaled Time: 0.02s
- Improvement: ~185x faster



ASYMPTOTIC COMPARISON

LOGISTIC REGRESSION

- Original: O(N.d)
- Scaled: O(n.d) where n very less than N (specifically n is approx 500)

Scenario	Samples	Time (s)
Small Subsample (100)	100	0.0416
Theoretical (499)	499	0.0799
5% Scale	5687	1.2501
10% Scale	11375	1.3160
100% Scale	113754	13.2246

Table 1: Logistic Regression

RANDOM FOREST

- Original: O (T.N.K.logN)
- Scaled: O(T.n.K.logn) where n very very less than N (specifically n is approx 750)

Scenario	Samples	Time (s)
Small Subsample (100)	100	0.0146
Theoretical (753)	753	0.0194
5% Scale	5687	0.1106
10% Scale	11375	0.1743
100% Scale	113754	3.6903

Table 2: Random Forest

CONCLUSION & FUTURE WORK

Conclusion:

1. Leverage Sampling works: We successfully reduced the dataset by 99.6% (from 113k to 499 rows) while maintaining 99% of the original accuracy for Logistic Regression.
2. Theoretical Bounds hold: The PAC learning formula accurately predicted the point where the model convergence stabilizes.
3. Efficiency: Training time became negligible, making this approach ideal for streaming data or real-time systems.

Future Work:

1. Apply Deep Learning Algorithms in the future to improve model performance.
2. Research more datasets in this space to be able to move from a “Yes/No” prediction to a more comprehensive probabilistic prediction.