An Ensemble Method Based Aggregated Model by Analyzing Data of Existing Precipitation Prediction Models

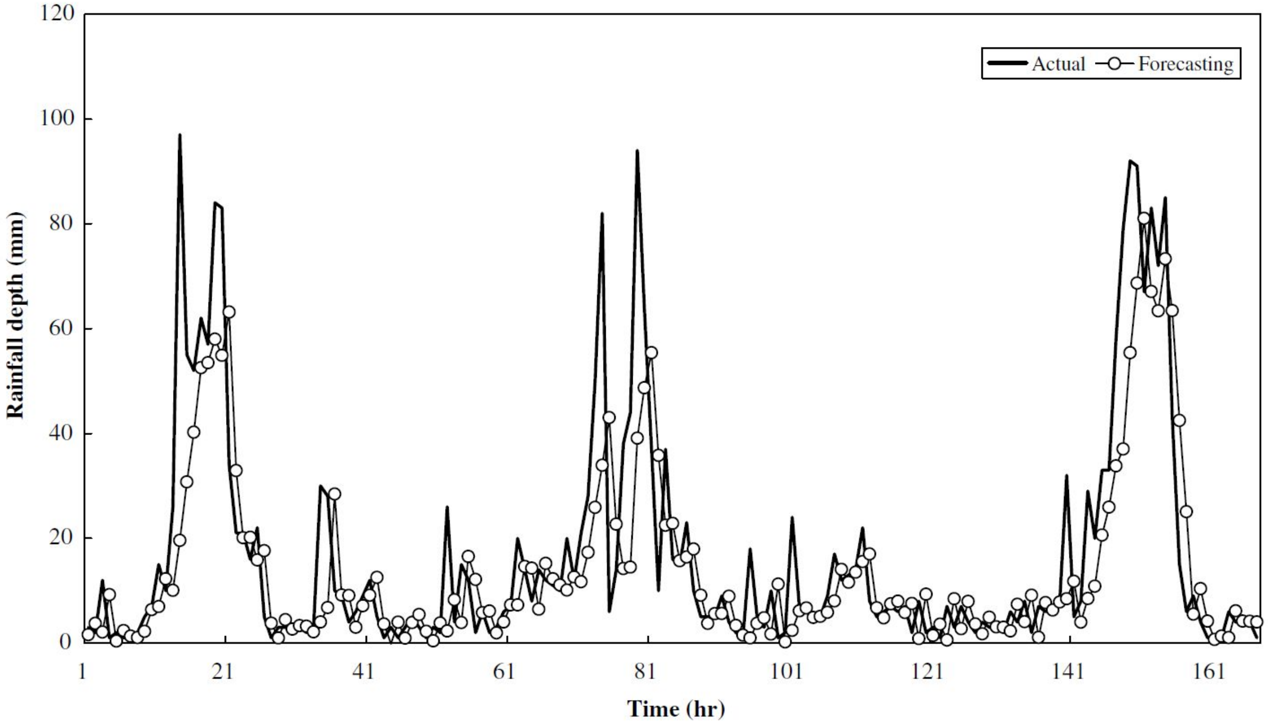
Kallol Das

Research Problem

Research Problem:

- ➔ Existing prediction models are not predicting rainfall accurately (Claim)
- ➔ Most of the cases they are overpredicting.
- ➔ Sometimes the rate of false positive is way high.
- ➔ These models predict good in some areas (e.g. mountains, deserts) and worse in other areas.

Evidence for the claim



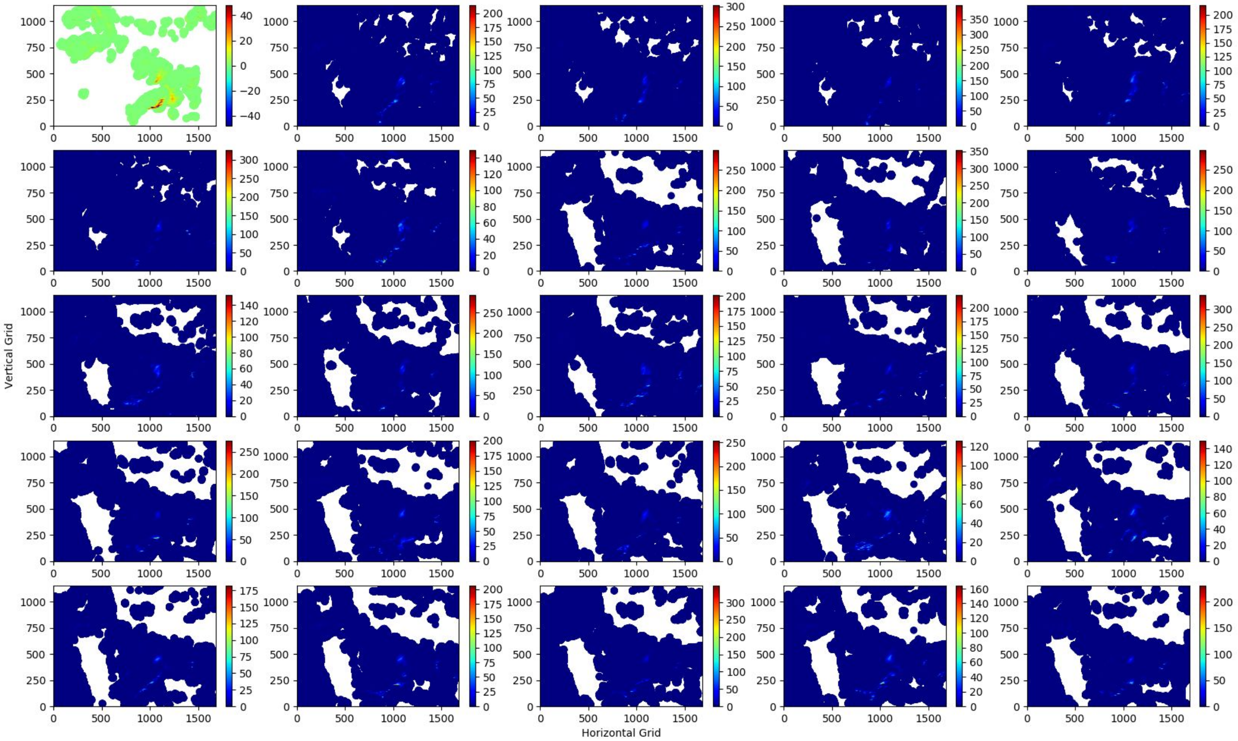
1. Wei-Chiang Hong, Rainfall forecasting by technological machine learning models, Applied Mathematics and Computation, Volume 200, Issue 1, 2008, Pages 41-57, ISSN 0096-3003

Dataset

Data Provider: Oklahoma University
Advisor: Dr. Ramyaa
Dataset:

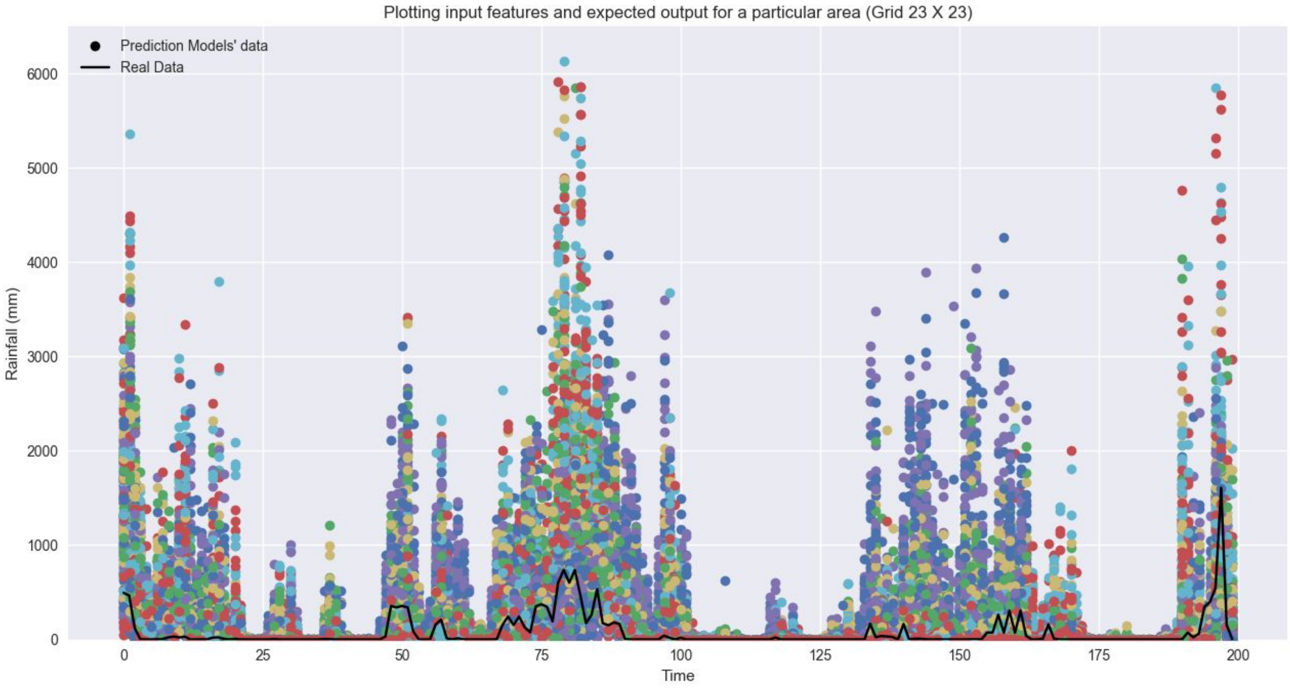
- ➔ Rainfall Data
- ➔ 39 Days data in total and each day has 20 times
- ➔ 39 **Prediction Models dataset** and 1 **Real verification dataset** for each day and time

Dataset



Visualization of precipitation real verification data and prediction models' data. The first image (image in the top left corner) is the real verification precipitation data and rest of the 24 images are visualizing precipitation prediction data from 24 prediction models

Comparing Prediction Models for a Particular Coordinate



Research Question

Research Question: Can we implement a new model from these existing models data which has comparatively lower overprediction and lower false positive rate?

Related Work

- ➔ Proposed new models from the real meteorological data by considering a good amount of features e.g., temperature, wind speed, humidity etc.
- ➔ The error rates for the prediction data of those new models are high
- ➔ New models are being proposed instead of trying to improve the existing models

Assumption

Hypothesis and Assumption:

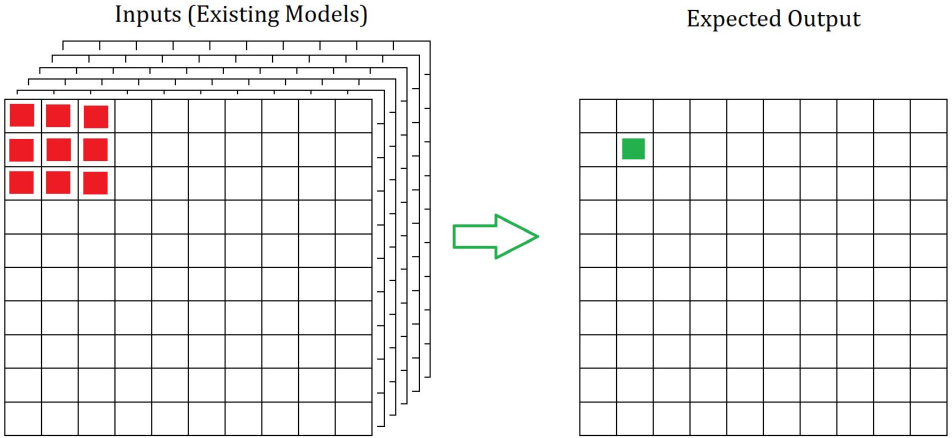
- ➔ Different models give wrong prediction in different area
- ➔ An aggregated model would perform better

Resolution Selection

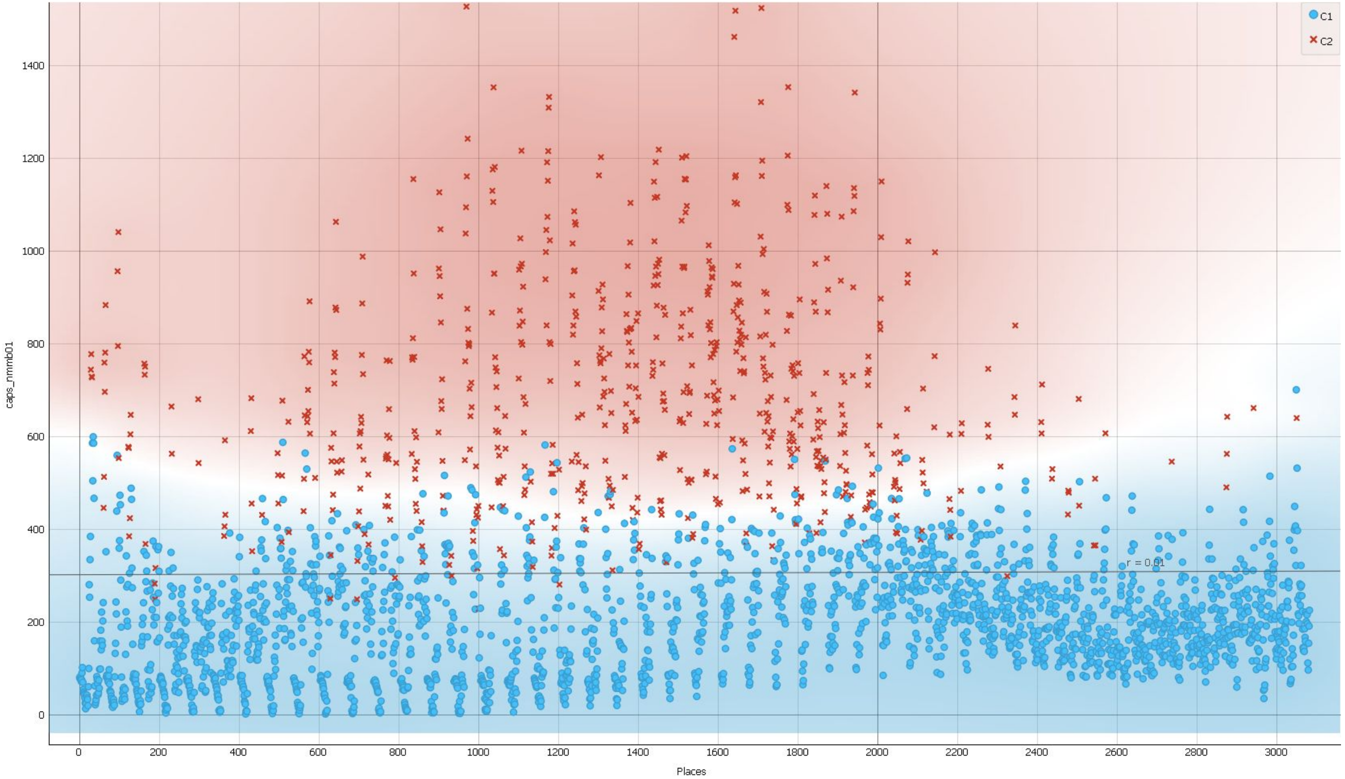
- ➔ Prediction data for 1155X1683 different geographical coordinates
- ➔ Every coordinates cover little more than 1 mile square
- ➔ 25X25 grid to start with
- ➔ Later tried with high resolution 15X15

Input Grid

- ➔ For a particular area which has high error rate, the neighbor grids reflect closer amount to the target value
- ➔ 3X3 grid has been given as a input to predict the middle value



Clustering MAE of an Existing Model

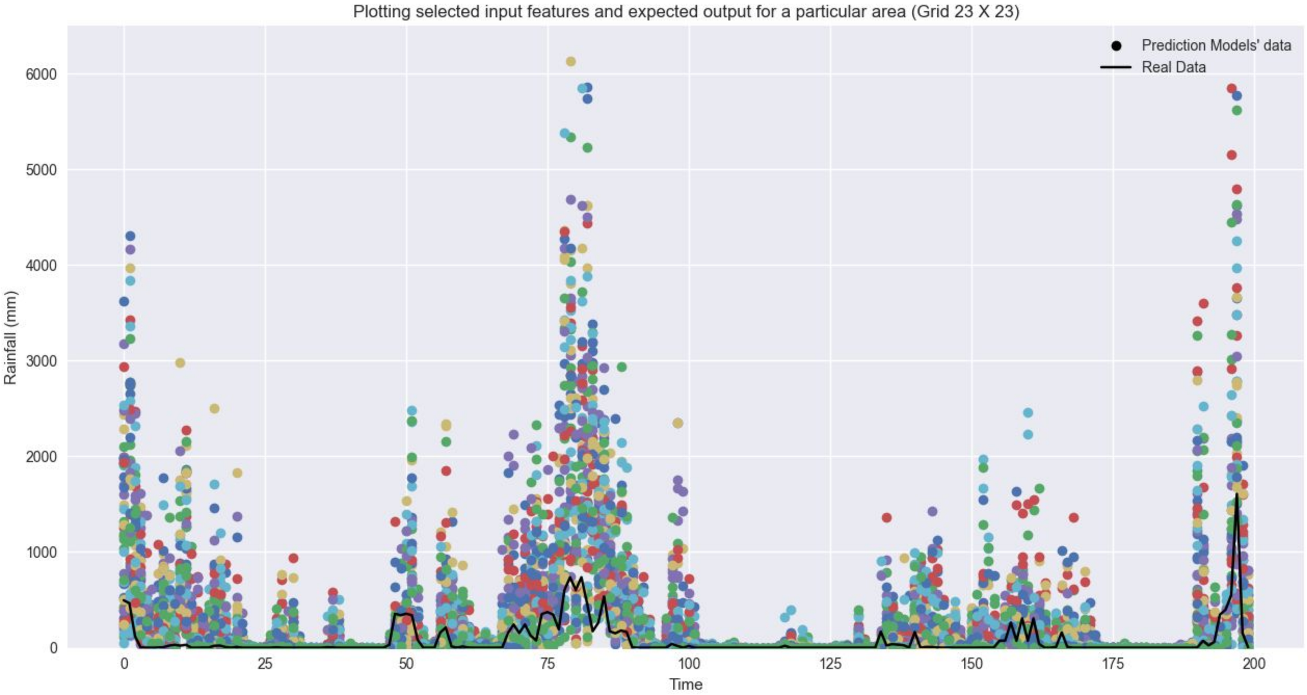


Shows the clustering of MAE for an existing prediction model data

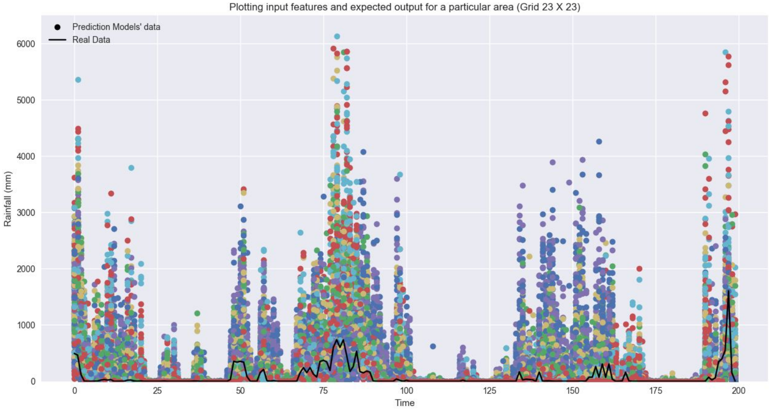
Univariate Feature Selection

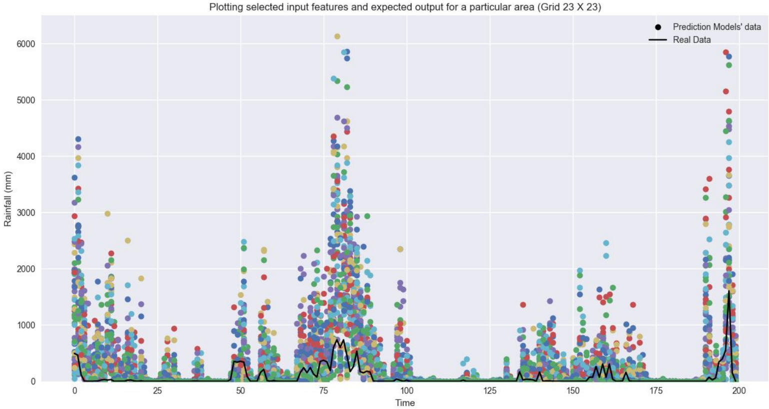
- Univariate linear regression shows the relationship between dependent and independent variable.
- Univariate feature selection is a linear regression technique to test individual importance of each independent variable.
- Selected 50 important features out of 216

After Selecting Features



After Selecting Features





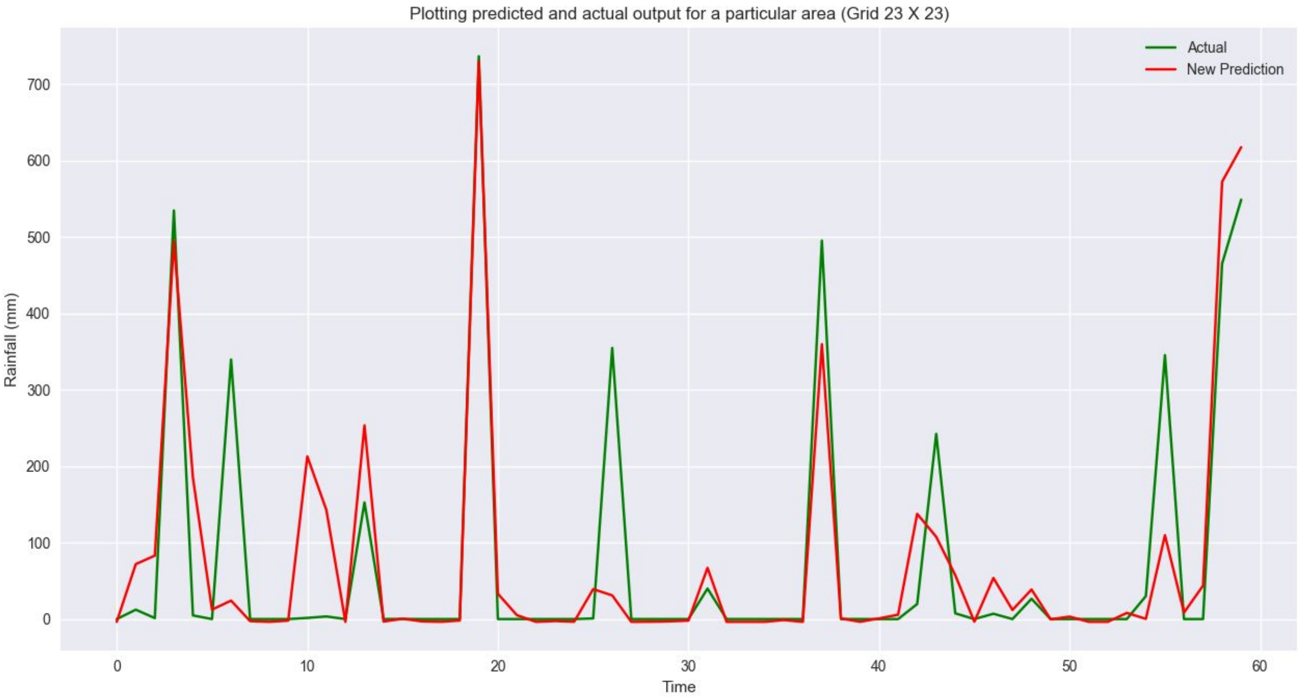
Dimensionality Reduction: PCA

- ➔ Since, the input dataset has only 200 samples, so it is important to reduce the dimension according to curse of dimensionality.
- ➔ Taking 5 PCA dimensions
- ➔ Principal Component Analysis (PCA) has been used to reduce dimension of our input vector.
- ➔ Principal component analysis convert the high dimensional data into lower dimension.

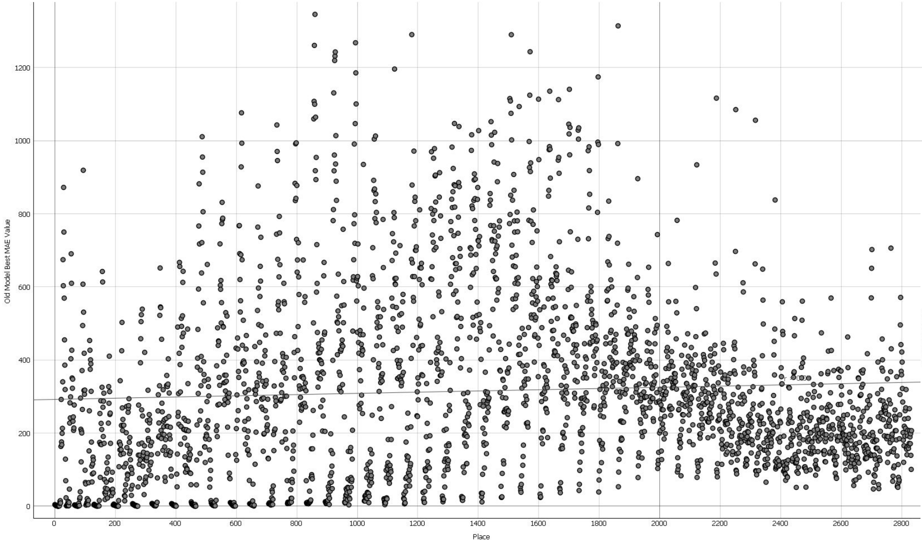
Machine Learning Models

- ➔ Used Linear Regression, Random Forest, Neural Network, Support Vector Machine, K-Nearest Neighbours
- ➔ Ensemble Technique

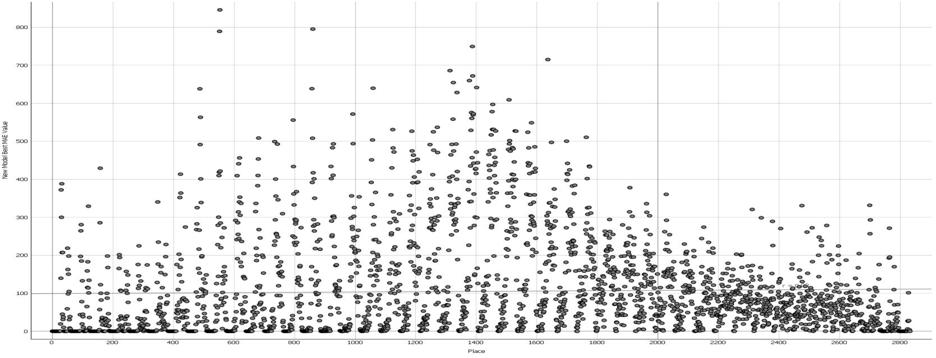
Prediction Graph | Random Forest



Compare Old and New Model



Old Models' best MAE



New Model's best MAE

Summary of the Results

Shows the percentages of the places where new prediction model's MAE and RMSE is better than existing prediction models

Grid	MAE	RMSE
25X25	98.44% cases	100.0% cases
15X15	99.06% cases	100.0% cases

Summary

- ➔ Proposed an ensemble approach to develop a New Aggregated Model to predict precipitation based on the dataset of some existing prediction models.
- ➔ Used feature selection and extraction techniques
- ➔ Used different machine learning models
- ➔ Promising results

References

1. Wei-Chiang Hong, Rainfall forecasting by technological machine learning models, Applied Mathematics and Computation, Volume 200, Issue 1, 2008, Pages 41-57, ISSN 0096-3003
2. THOMPSON, J.C., 1950: A NUMERICAL METHOD FOR FORECASTING RAINFALL IN THE LOS ANGELES AREA. Mon. Wea. Rev., 78, 113–124
3. Hernández E., Sanchez-Anguix V., Julian V., Palanca J., Duque N. (2016) Rainfall Prediction: A Deep Learning Approach. In: Martínez-Álvarez F., Troncoso A., Quintián H., Corchado E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2016. Lecture Notes in Computer Science, vol 9648. Springer, Cham
4. Beda Luitel, Gabriele Villarini, Gabriel A. Vecchi,Verification of the skill of numerical weather prediction models in forecasting rainfall from U.S. landfalling tropical cyclones, Journal of Hydrology, Volume 556, 2018, Pages 1026-1037, ISSN 0022-1694
5. Schneider, Astrid, Gerhard Hommel, and Maria Blettner. “Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications.” Deutsches Ärzteblatt International 107.44 (2010): 776–782. PMC. Web. 7 May 2018.
6. Jake Lever, Martin Krzywinski, Naomi Altman, Principal component analysis, Nature Methods, 2017/06/29/online, 14, 641, Nature Publishing Group, a division of Macmillan Publishers Limited.
7. Herve Abdi and Lynne J. Williams, Principal component analysis, 2010 John Wiley and Sons, Inc. WIREs Comp Stat 2010 2 433–459