Predicting Precipitation with Machine Learning

Kaitlyn Lane, Patrick Ledoit, Jessica Ling, Conor Luppnow

Project Background

Why Rainfall Prediction?

- Agricultural, economic impacts
- Disaster mitigation: floods, landslides

Also:

- Chaotic, not “easily” predictable
- Breadth and type of meteorological data

Photo by Dibakar Roy from Pexels: https://www.pexels.com/photo/people-walking-in-the-rain-near-metropolitan-building-kolkata-india-18158842/

Prior Work

Two open-source projects on the same Kaggle dataset

- Gaurav, V. (2018):
  - Investigated 7 initial models without hyperparameter tuning and reported mean absolute error

- Sudharsan, D. et al. (2021)
  - Random Forest, Lasso, Elastic Net models on 1 meteorological subdivision of India

Their Data: Monthly precipitation for 36 meteorological regions in India from 1901-2015

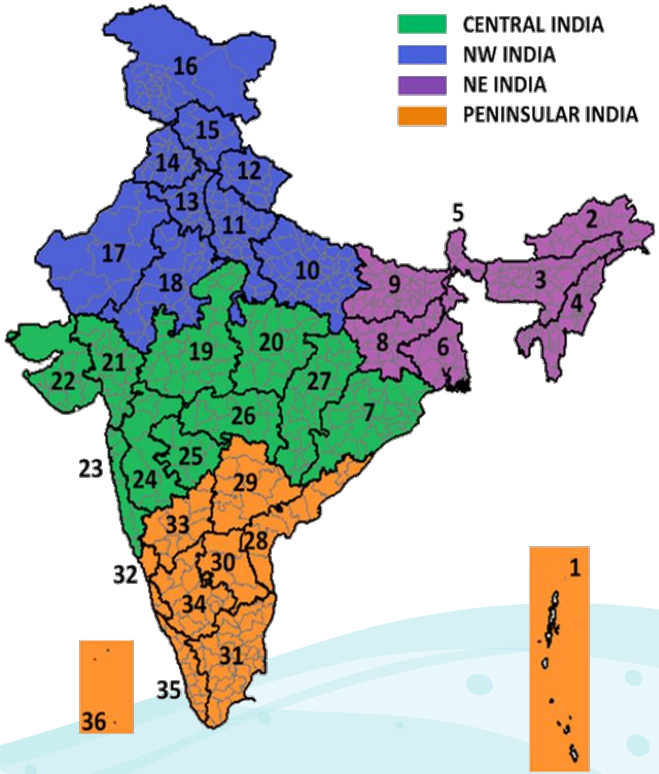
Prior Work - Major Failings

Their Data:

- Extremely low dimensionality
- Predict next month's rainfall only on previous 3 months'

Their Models:

- Little to no hyperparameter tuning
- Limited Scope



Our Contribution Goals

1. Expand Data: acquire new, relevant features

2. Improve Models: test a wide range of models and tune hyperparameters

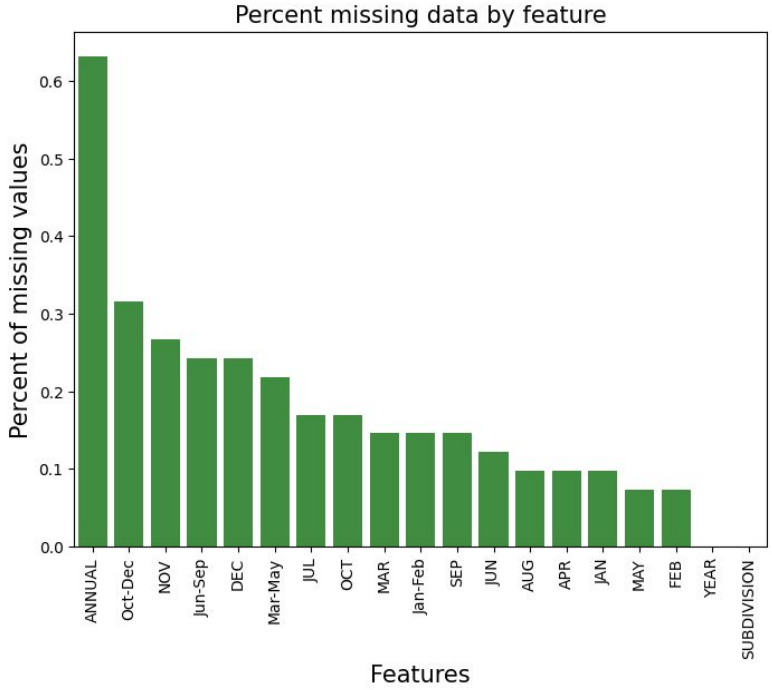
Contribution 1: Data

Initial Kaggle Dataset:

- Data from 1901-2015
- Monthly granularity
- **Rainfall Only**
- All 36 Meteorological Subdivisions
- Missing data, primarily pre-1950

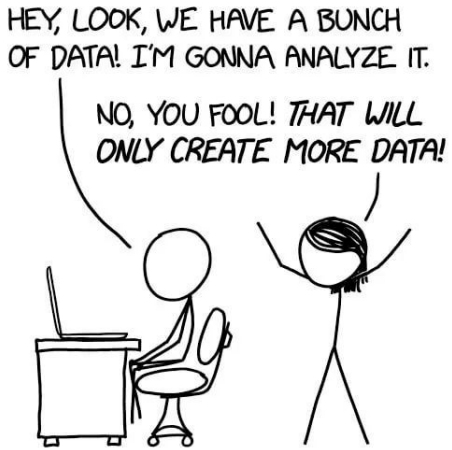
Weather and Climate IN

- Data from 2010-2020
- Monthly granularity
- **Temperature, Dew Point, Precipitation, Wind, Sea Level Pressure**
- All 36 Meteorological Subdivisions **+ Delhi**
- No missing data



Data Collection Challenges

- Weather data for the United States is readily available
  - For India the same is not true
- Precipitation data readily available, not much else
- Data was hidden behind paywalls or was only available in the form of graphs/charts
- Sacrifice long time scale for detailed data over
- Finally found detailed data from https://weatherandclimate.com/
  - Slightly obscure website that was difficult to find



Chhattisgarh Weather In September 2012

Historical Data for September 2012 in Chhattisgarh, India

Temperature	Max	Average	Min	
Max Temperature	34.0°C (93.2°F)	31.23°C (88.21°F)	26.0°C (78.8°F)	
Avg Temperature	31.0°C (87.8°F)	28.27°C (82.89°F)	25.0°C (77.0°F)	
Min Temperature	25.0°C (77.0°F)	23.33°C (73.99°F)	20.0°C (68.0°F)	
Dew Point	Max	Average	Min	
Dew Point	26.0°C (78.8°F)	22.83°C (73.09°F)	17.0°C (62.6°F)	
Precipitation	Max	Average	Min	Sum
Precipitation	29.2mm | 1.15in	4.12mm | 0.16in	0.0mm | 0in	123.7mm | 4.87in
Snowdepth	0.0mm | 0in	0.0mm | 0in	0.0mm | 0in	0.0mm | 0in
Wind	Max	Average	Min	
Wind	12.0kmh | 7.46mph	7.23kmh | 4.49mph	4.0kmh | 2.49mph	
Gust Wind	19.0kmh | 11.81mph	11.6kmh | 7.21mph	6.0kmh | 3.73mph	
Sea Level Pressure	Max	Average	Min	
Sea Level Pressure	29.2mb	4.12mb	0.0mb	

daily

Web Scraping

Month Year (2010-2020)

https://weatherandclimate.com/chhattisgarh/september-2012

Meteorological Region:

- Andaman and Nicobar Islands
- Andhra Pradesh
- Arunachal Pradesh
- Assam
- Bihar
- Chandigarh
- Chhattisgarh
- Dadra and Nagar Haveli
- Daman and Diu
- Delhi
- Goa
- Gujarat
- Haryana
- Himachal Pradesh
- Jammu and Kashmir
- Jharkhand
- Karnataka
- Kerala
- Ladakh
- Lakshadweep
- Madhya Pradesh
- Maharashtra
- Manipur
- Meghalaya
- Mizoram
- Nagaland
- Odisha
- Puducherry
- Punjab
- Rajasthan
- Sikkim
- Tamil Nadu
- Telangana
- Tripura
- Uttar Pradesh
- Uttarakhand
- West Bengal

Contribution 2: Improving Models

- We used the following Models:
  - random forest regressor, support vector regression, elastic net, ridge, multilayer perceptron regressor, and linear regression.
  - Used sklearn's gridsearchCV to obtain best hyperparameters, using three fold cross validation.
- Comparing test results of the default and best parameters after hyperparameter tuning we saw very marginal improvements. (Outlier: SVR)
- The issue with hyperparameter tuning is that the more parameters chosen to test, the longer the computation time.

Contribution 2: Optimal Hyperparameters

- Out of the parameters sampled, listed below are the best sets obtained from gridsearchCV.

Random Forest Regressor	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
SVR	{'C': 1, 'coef0': 0.01, 'degree': 3, 'kernel': 'linear'}
Elastic Net	{'alpha': 1, 'l1_ratio': 1, 'max_iter': 1000}
Ridge Regression	{'alpha': 3, 'max_iter': 200}
MLP	{'activation': 'relu', 'hidden_layer_sizes': (100,), 'learning_rate': 'constant', 'max_iter': 800}
Linear Regression	{'features': ‘sea_level_preassure_max_mb ’, ‘sea_level_preassure_avg_mb ’, ‘sea_level_preassure_min_mb ’, ‘wind_min_kmh’}

Overall Conclusions

- Most of our performance improvements came from data collection
  - Feature selection was very important
- Best-performance initially was Random Forest
- Best-performing models after tuning: **Linear Regression** and **Ridge Regression**

Model	MAE (Original Dataset)[mm]	MAE (new dataset)[mm]	MAE (tuned params)[mm]
Random Forest Regressor	85.69	2.252	2.091
SVR	127.70	**81.96**	**1.778**
Elastic Net	94.99	3.205	1.539
Ridge Regression	94.90	1.55	1.50
MLP (New)	87.19	4.322	3.211
Linear Regression (New)	93.64	1.595	1.468 (1.7% error)

Future Works

1. Collect daily weather data across multiple years
2. Kaggle dataset maintained by Nidula Elgiryewithana started this by recording daily temperature beginning in August 2023 and stopped in April 2024
3. Create a dataset that contains daily weather predictors such as humidity, temperature, dew-point, wind speed, etc

1. Enhanced feature selection for all other models
2. Try a larger set of hyperparameters and a larger range leaning towards underfitting (Bias-variance tradeoff)
3. Rolling window, non-linear

p-values by feature:

sea_level_preassure_avg_mb	0.000000e+00
sea_level_preassure_min_mb	2.966913e-09
sea_level_preassure_max_mb	1.192575e-07
dew_pt_min_C	2.280218e-02
temp_max_C	2.309954e-02
gust_avg_kmh	2.914913e-02
temp_avg_daily_max_C	5.894258e-02
wind_avg_kmh	1.111400e-01
wind_min_kmh	1.176298e-01
temp_max_daily_min_C	2.160140e-01