Prediction of Precipitation in London

Users

Meteorologist Agencies

Weather Forecasters

Energy Companies

BC Hydro

Urban Planners

Infrastructure

Agriculture

Farmers

WorkFlow

01

EDA

02

Model Selection

03

Model Evaluation

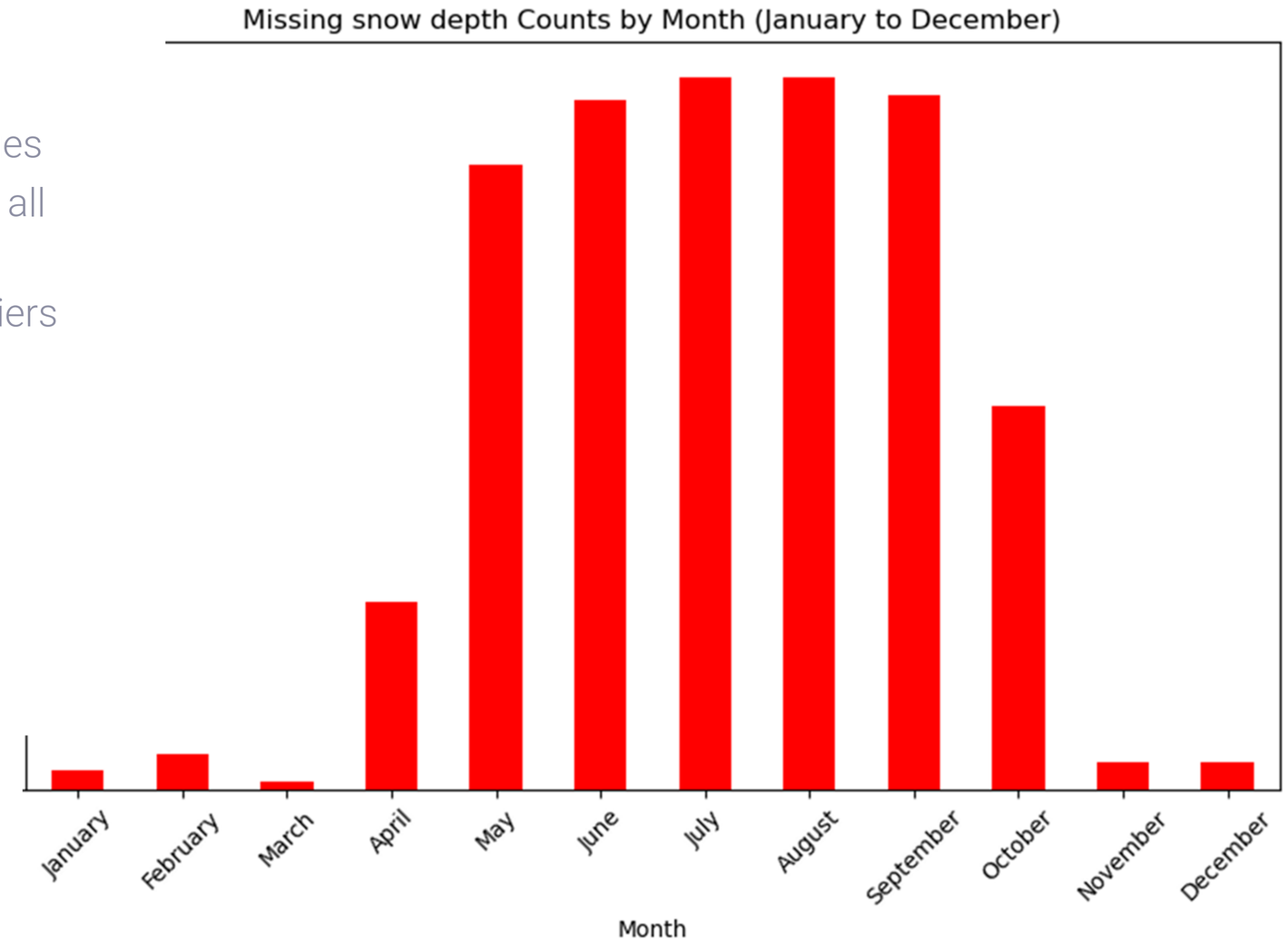
04

Recommendation

Pre-Processing

EDA

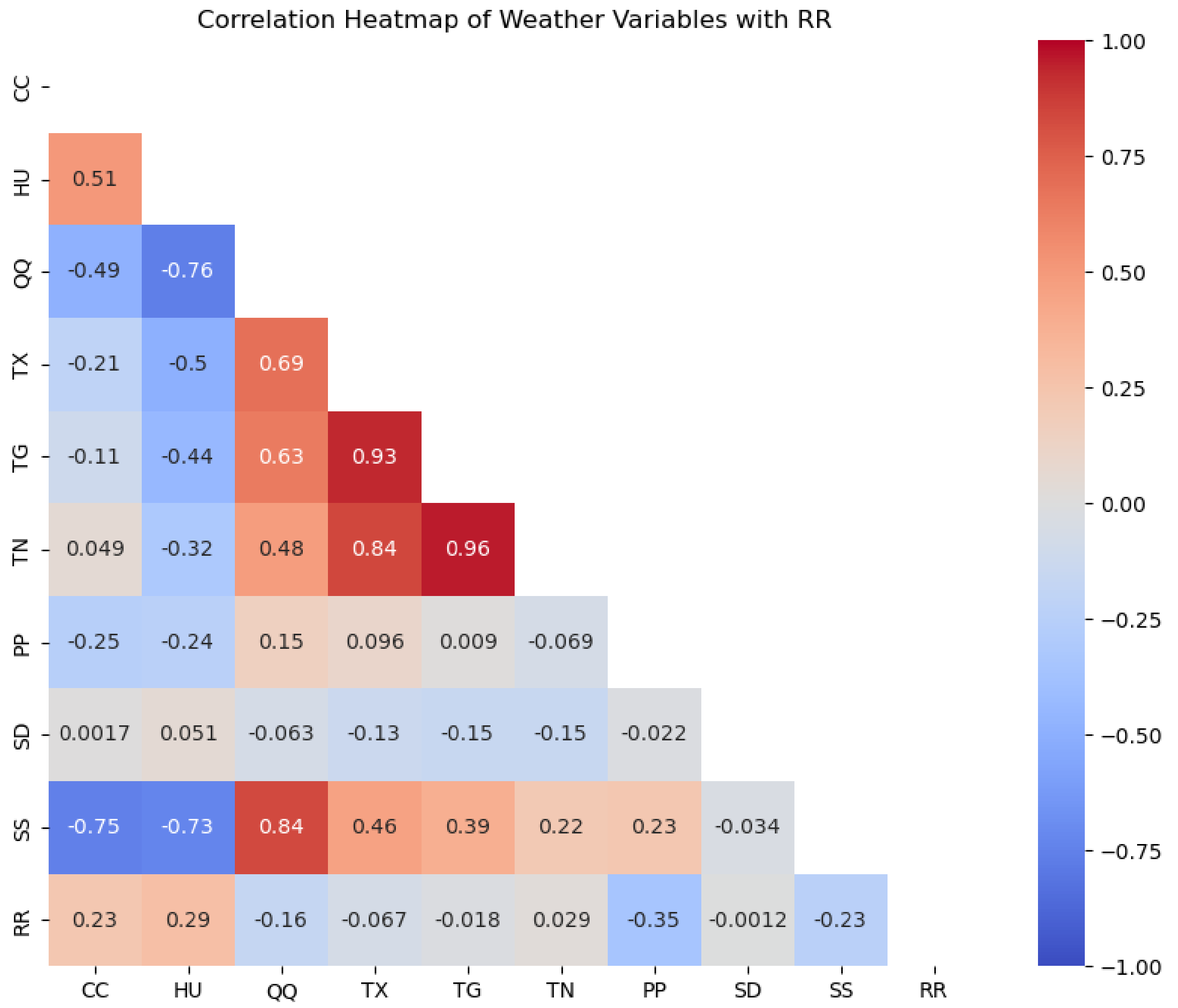
- Removed rows with any null values
- SD made the decision to convert all null values to 0
- Removed rows with strange outliers such as TX < TN



Pre-Processing

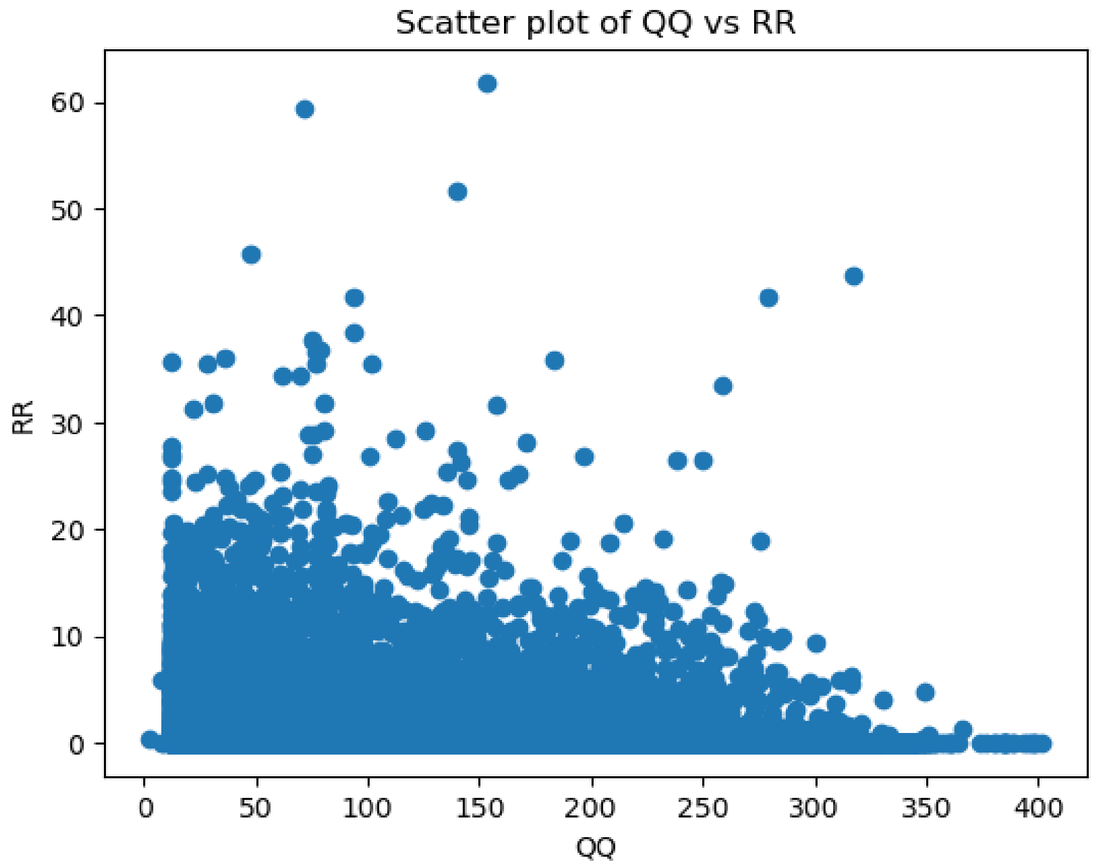
EDA

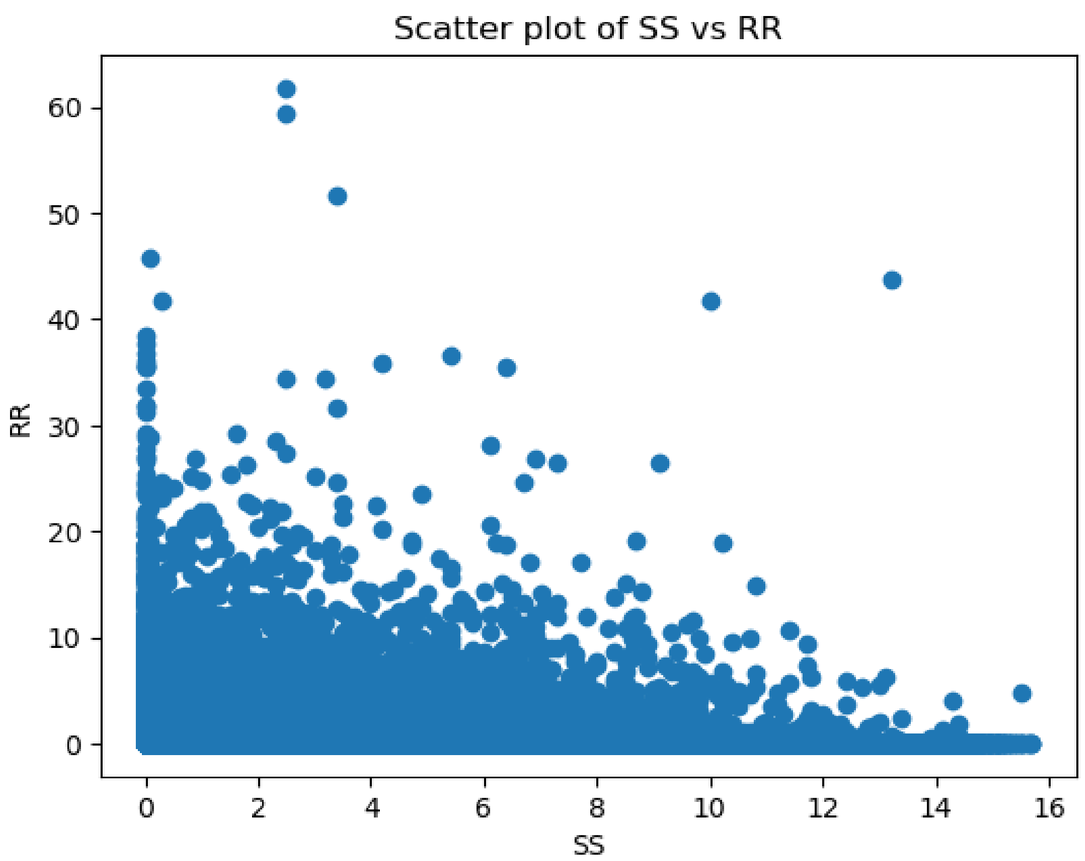
- Strongest positive correlation is with HU (Humidity)
- Strongest negative correlation is with PP (Pressure)



Pair-Plot with Precipitation

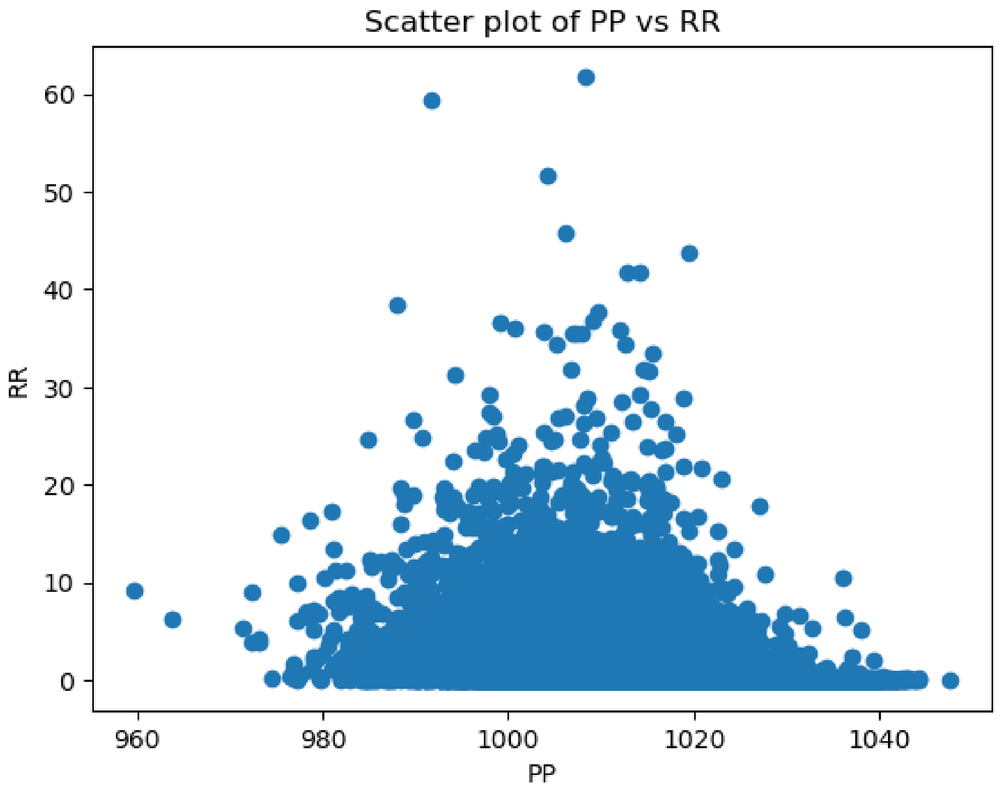
EDA

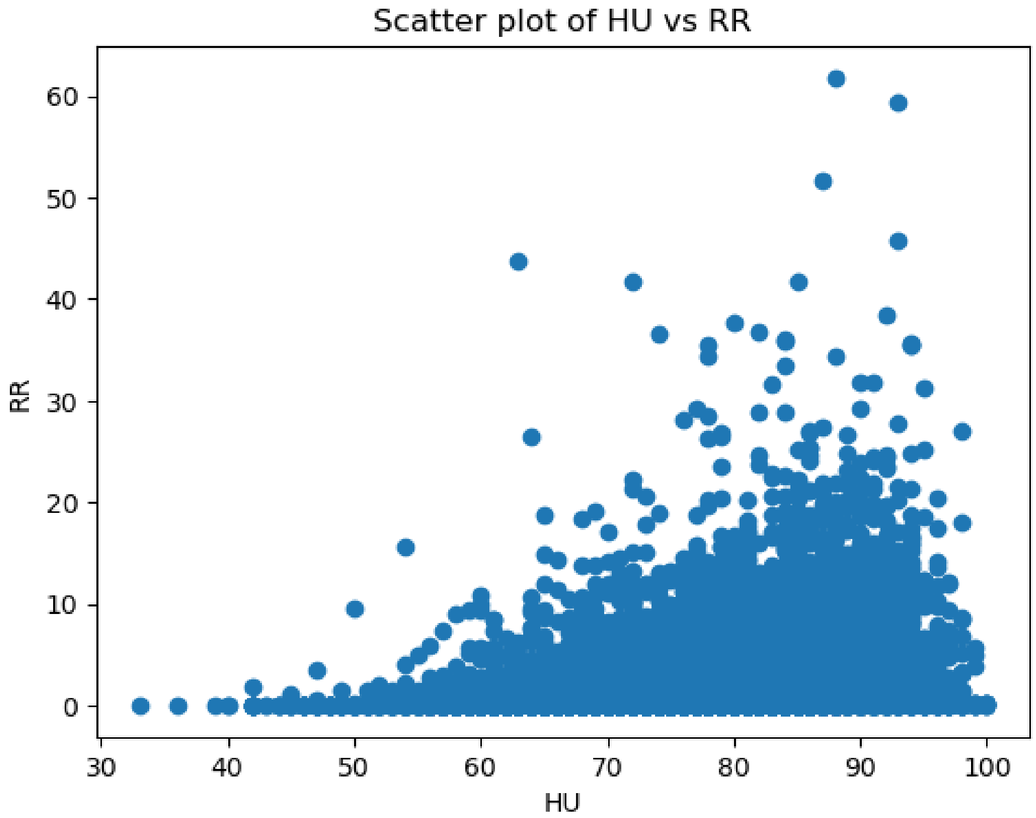




Pair-Plot with Precipitation

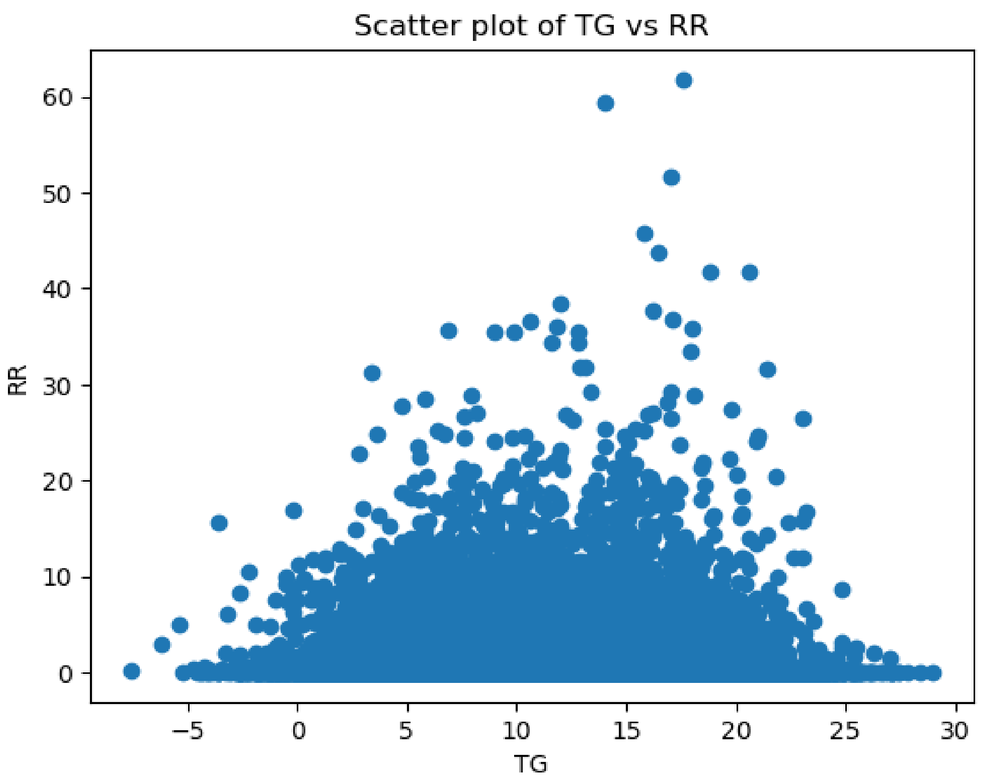
EDA

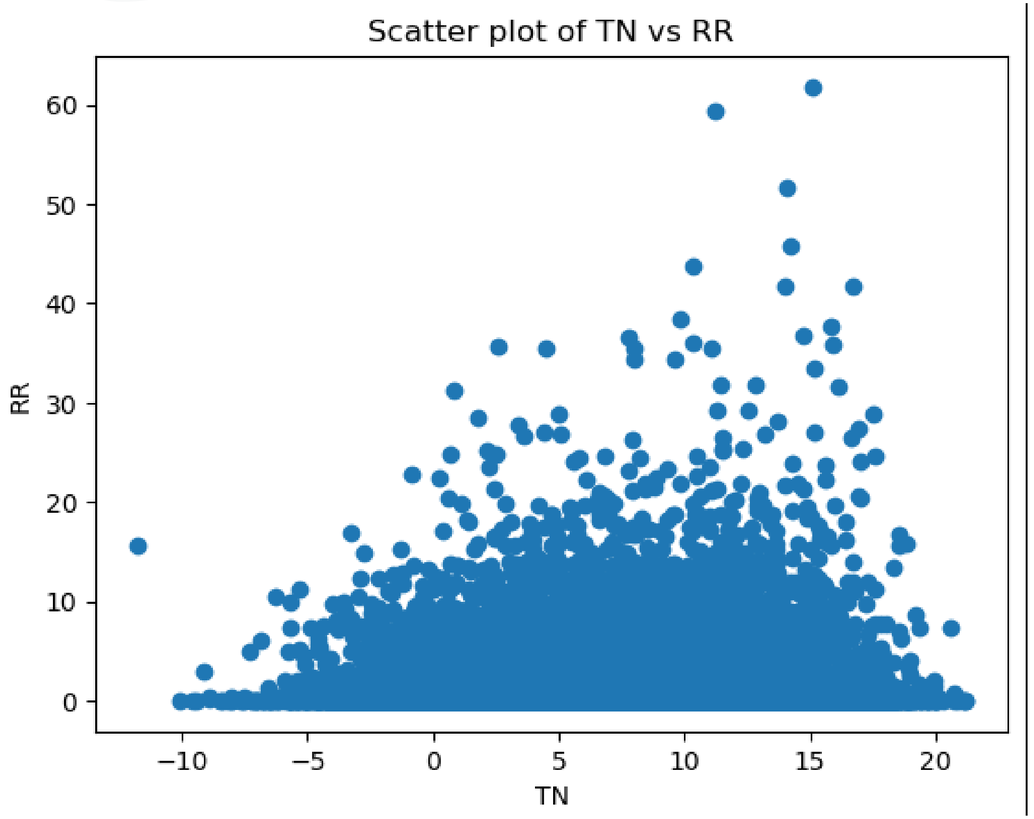




Pair-Plot with Precipitation

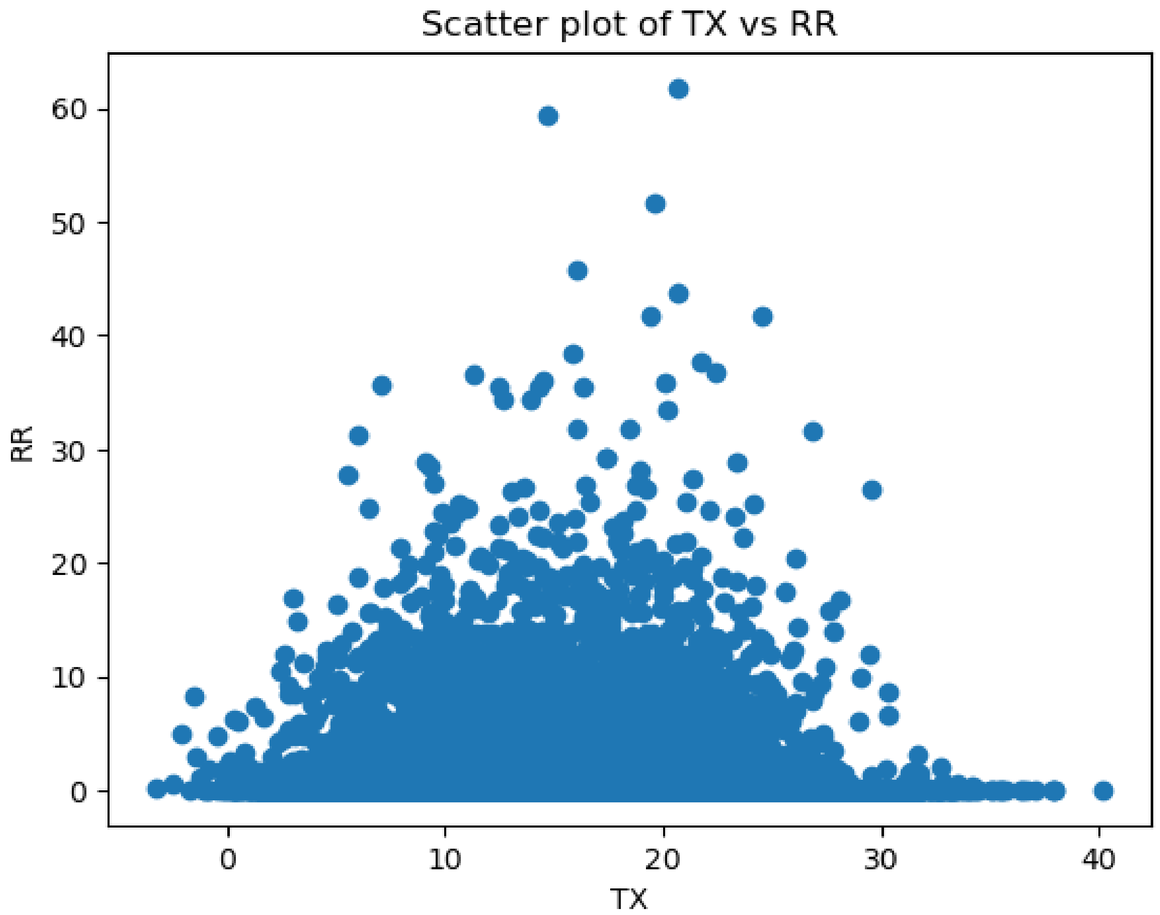
EDA





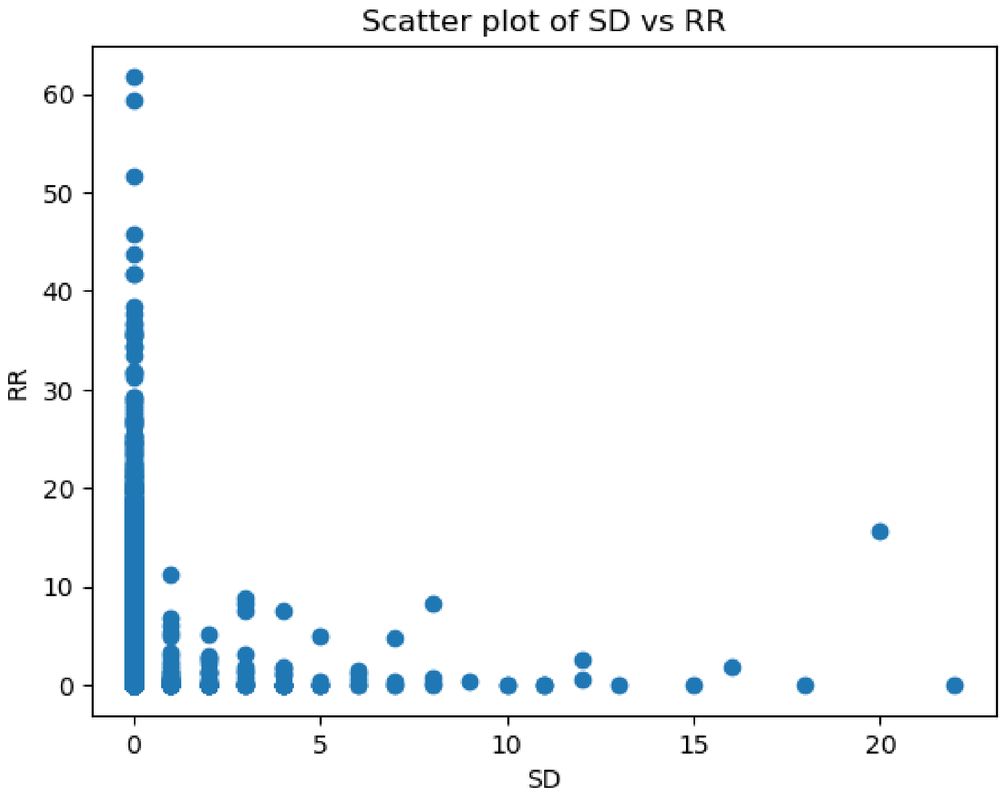
Pair-Plot with Precipitation

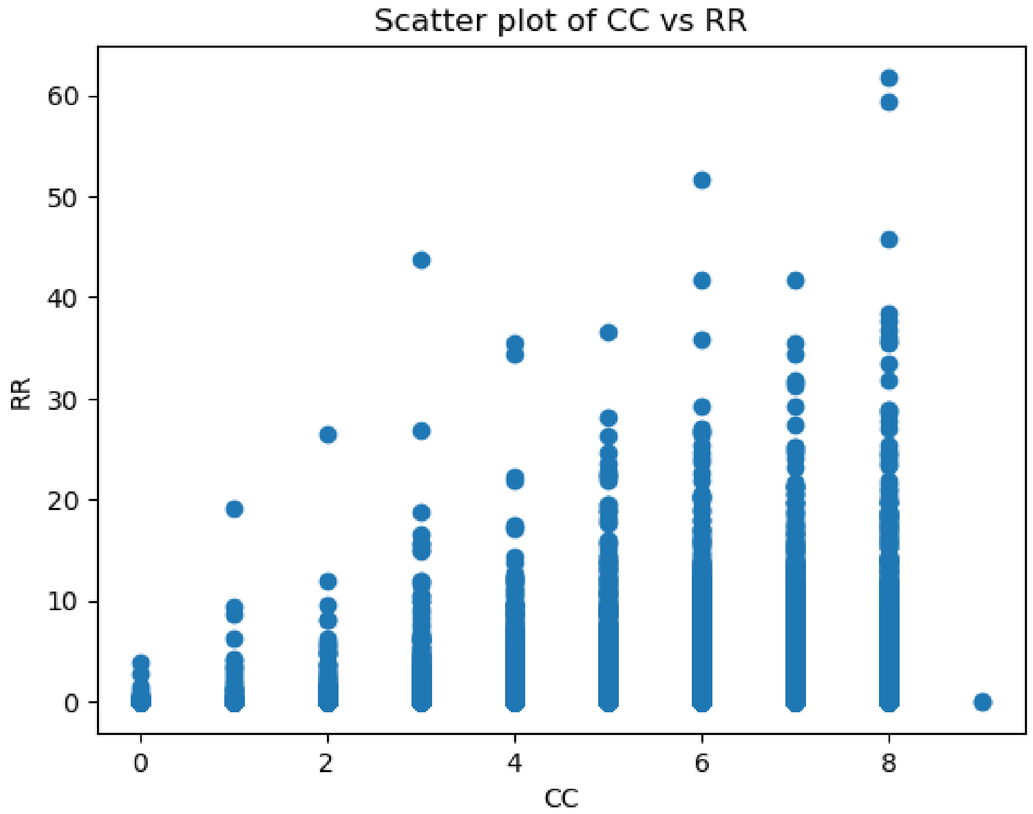
EDA



Pair-Plot with Precipitation

EDA





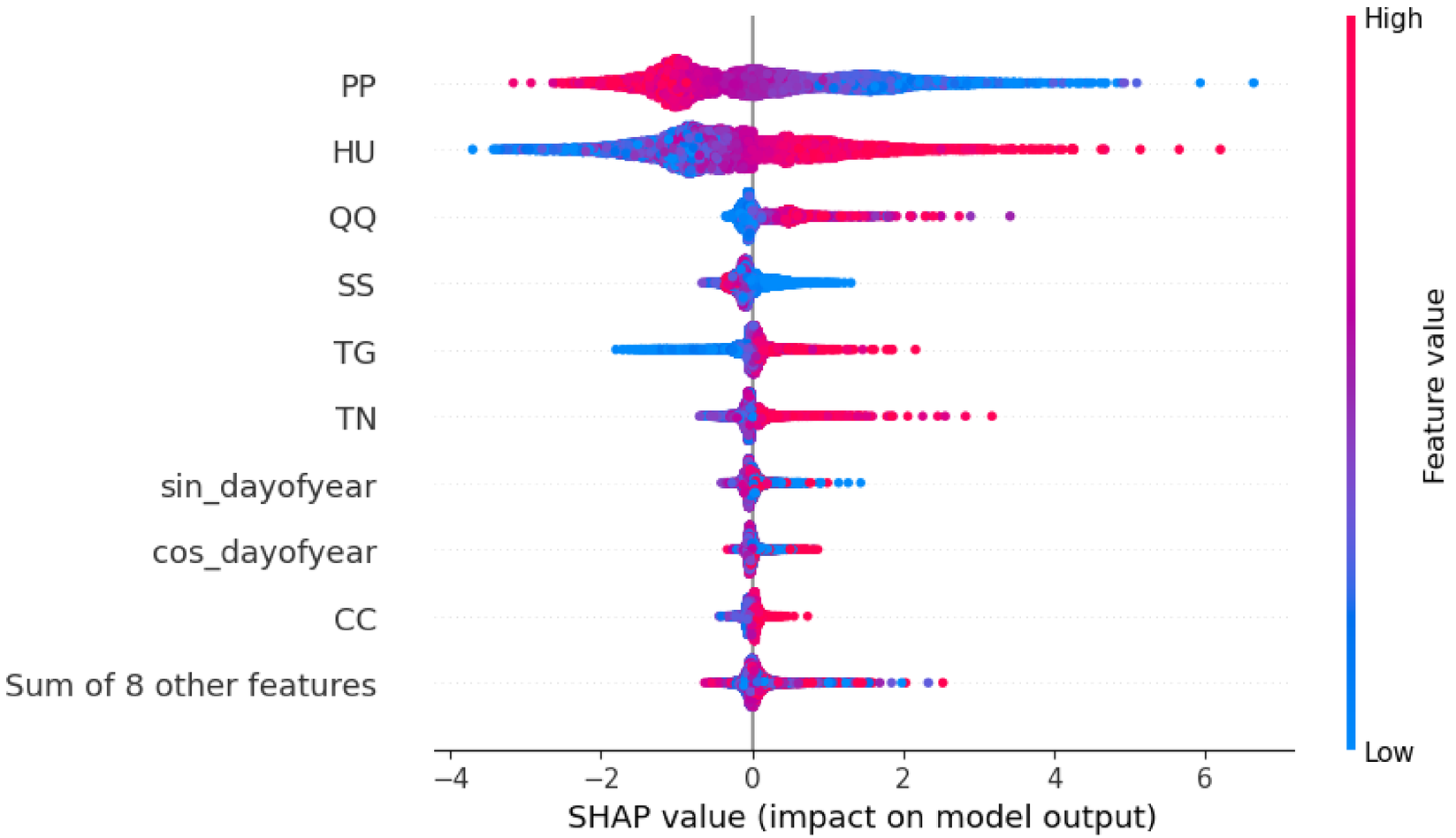
LGBM

MODELS

- Fast training speed and high efficiency
- Use of SHAP to determine top features
- Run model again feature selection and then with hyperparameters

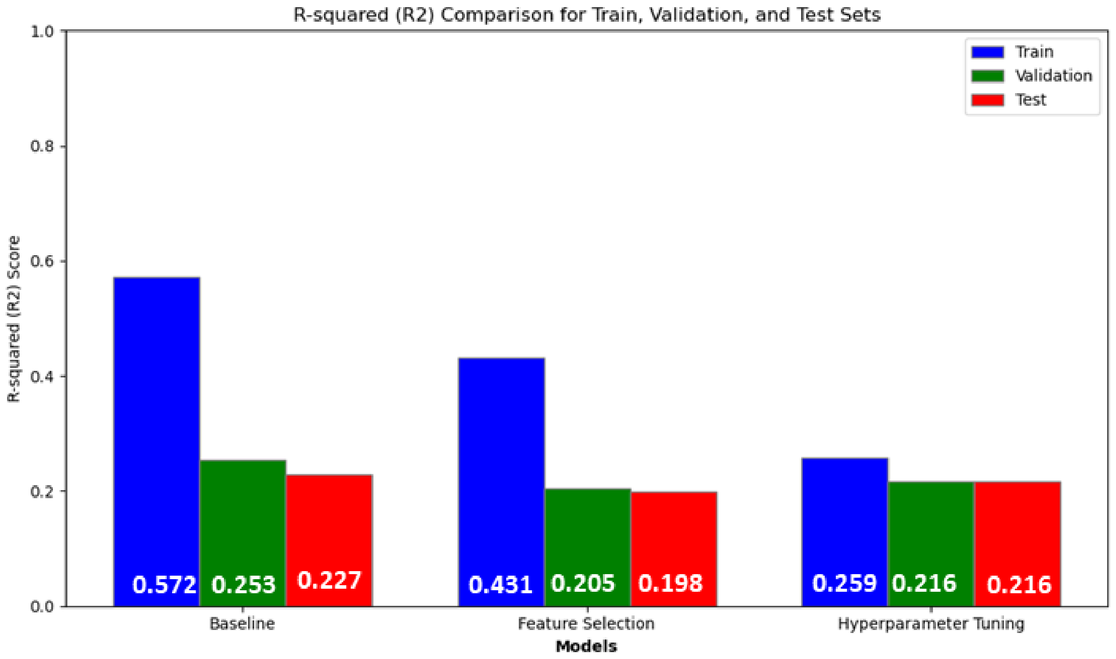
EVAL

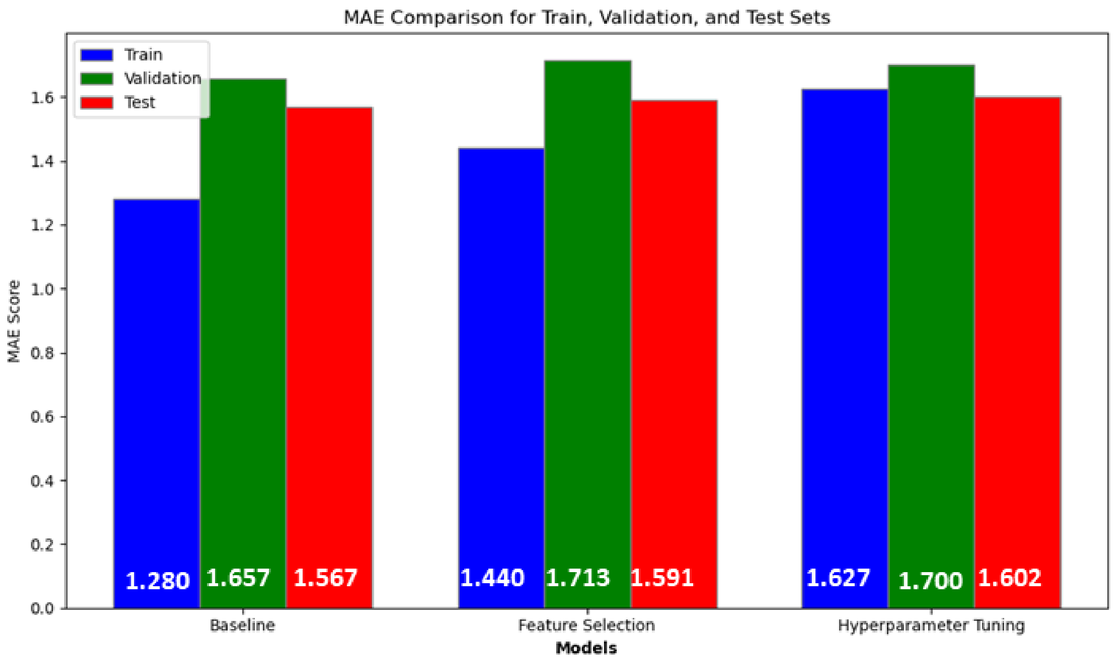
SHAP Feature Selection



EVAL

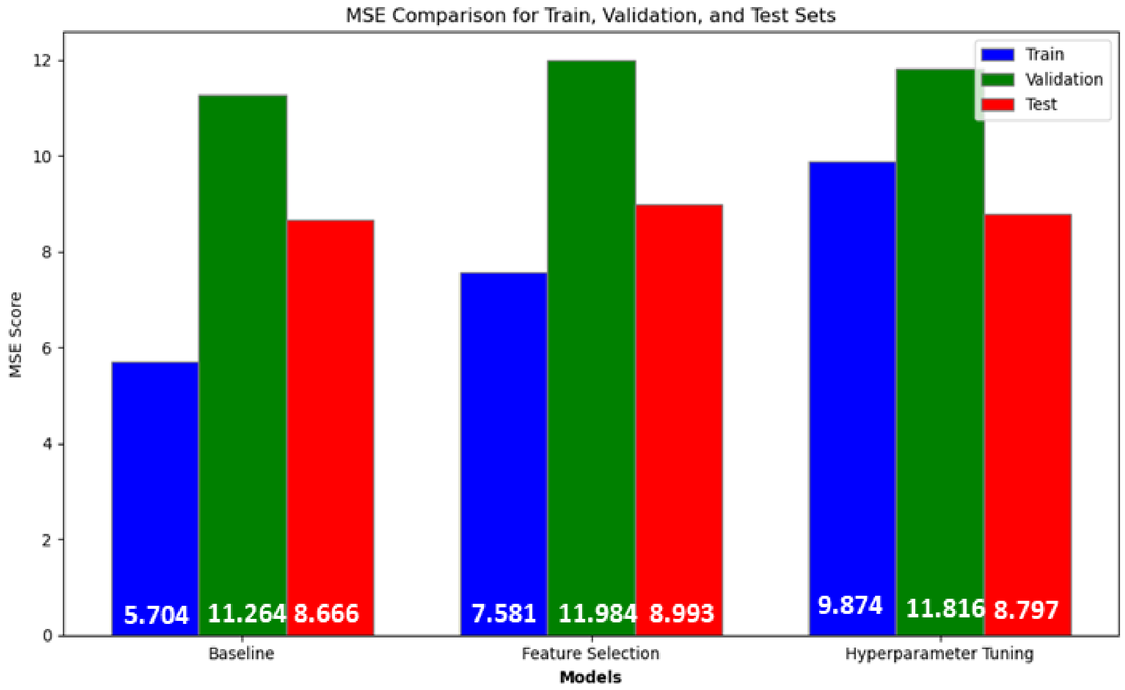
Performance Metrics

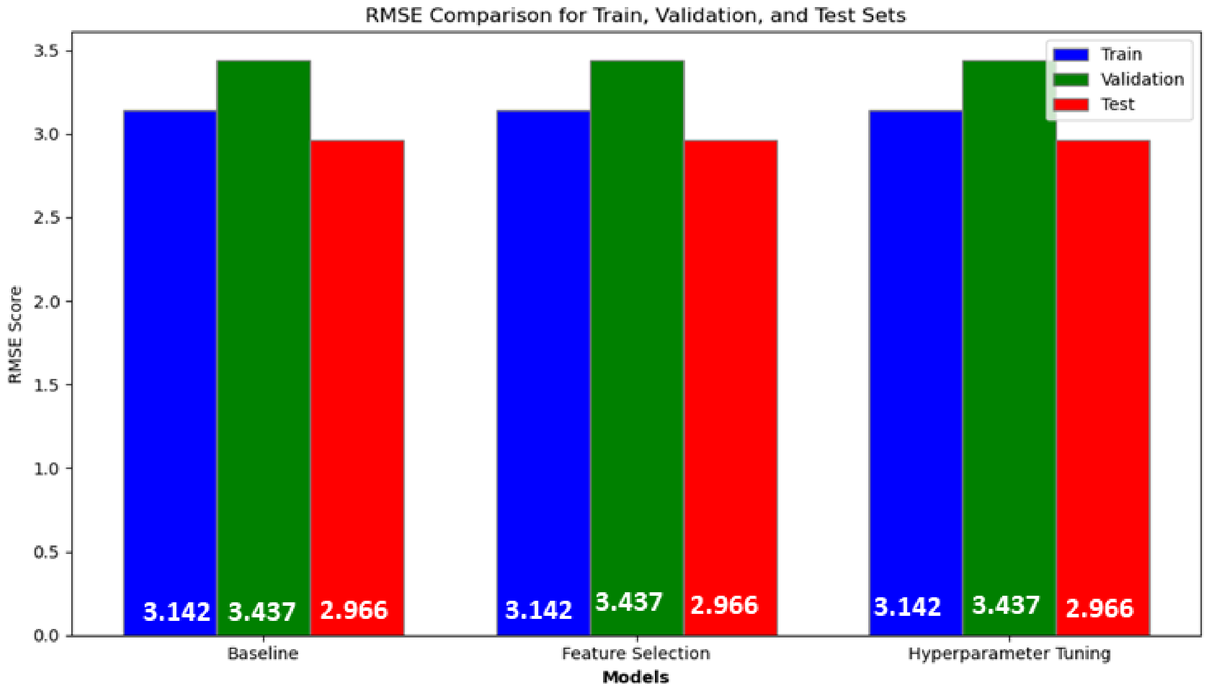




EVAL

Performance Metrics





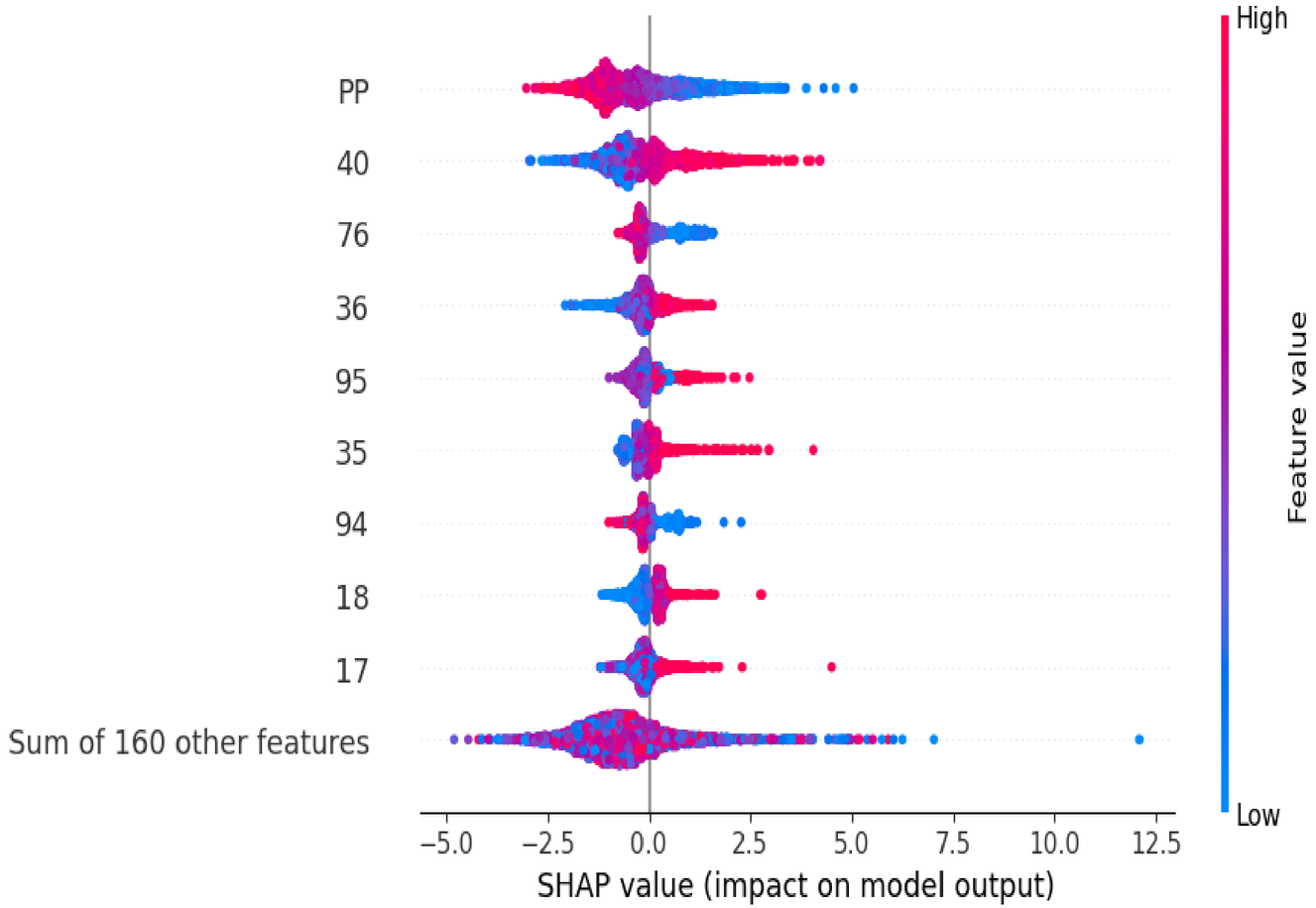
XGBoost

MODELS

- Helpful for imbalanced data
- Possible better predictiveness
- Use Polynomial transformation to generate polynomial and interaction features then SHAP to determine top features
- Run model again feature selection and then with hyperparameters

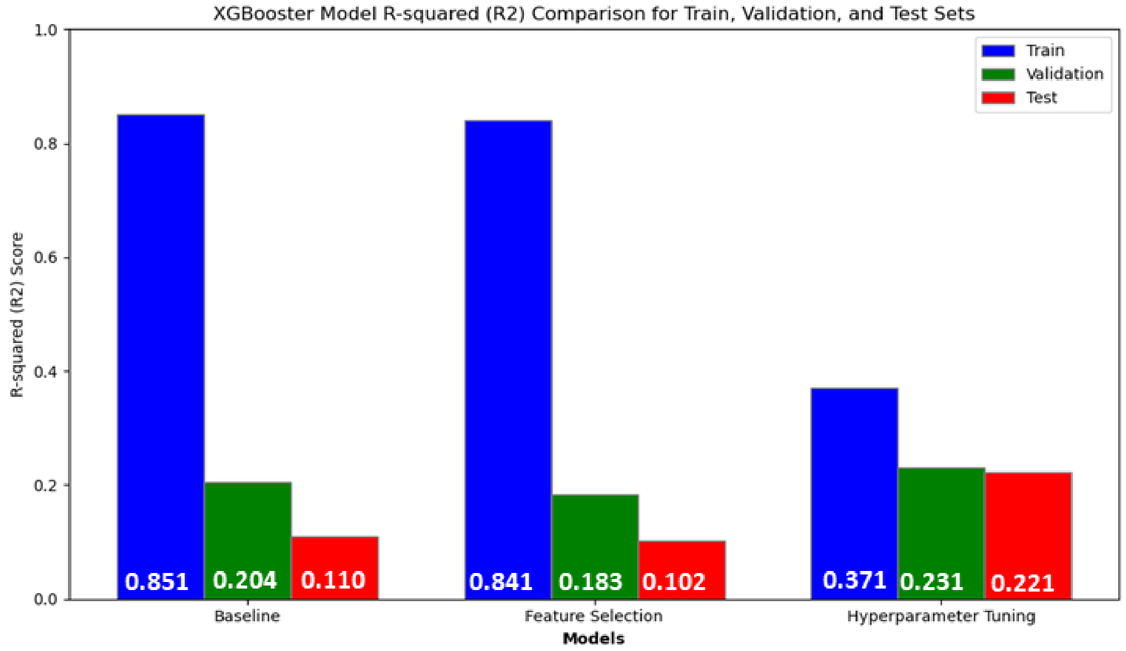
EVAL

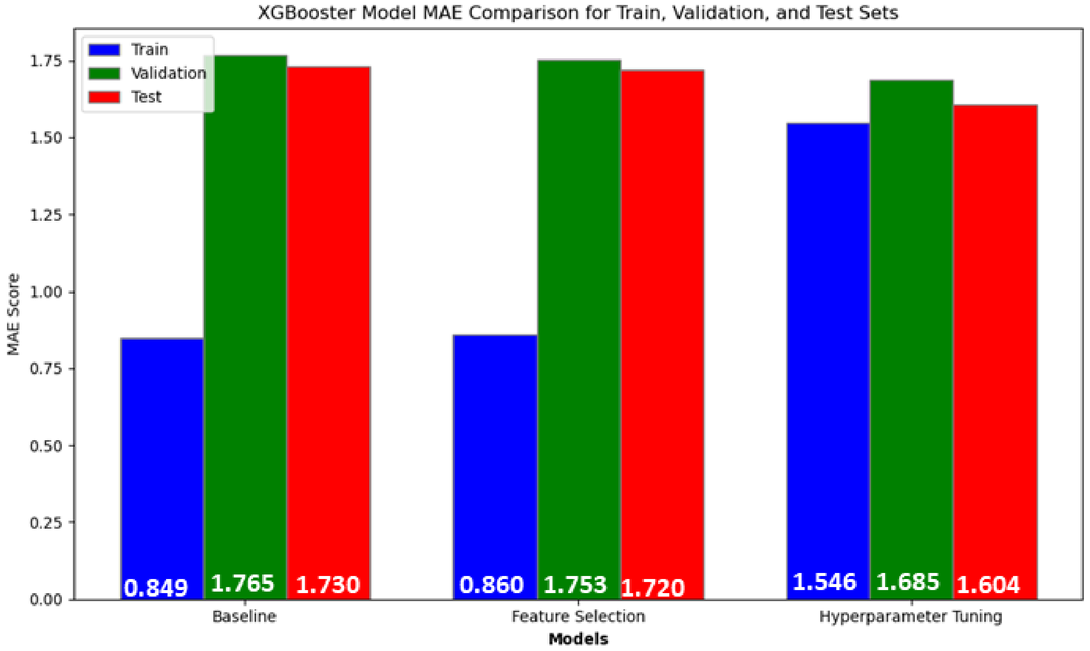
SHAP Feature Selection



EVAL

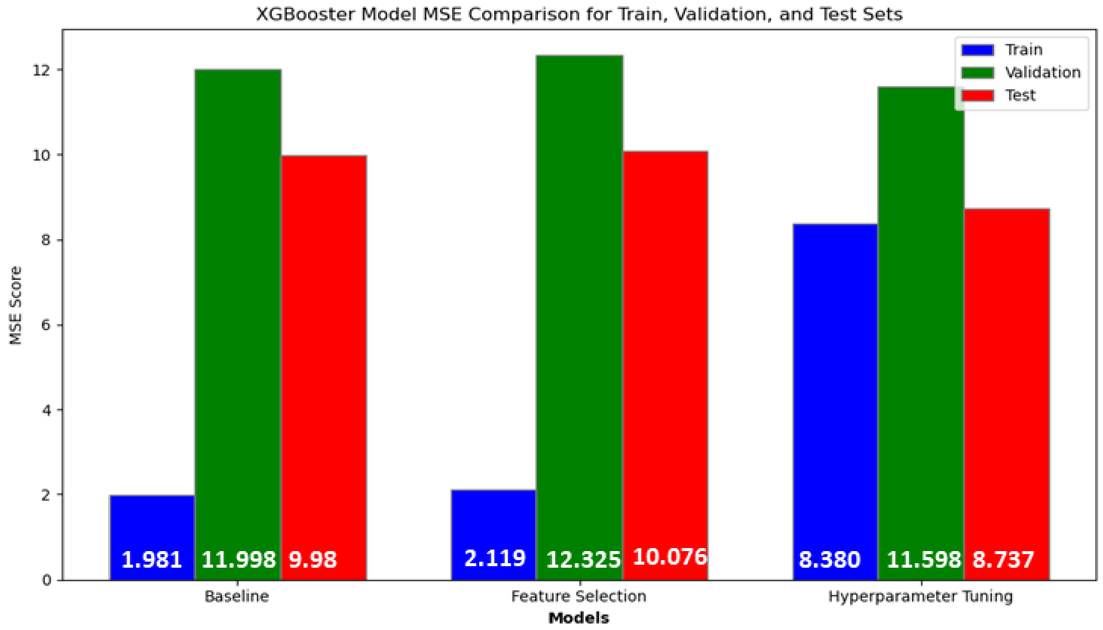
Performance Metrics

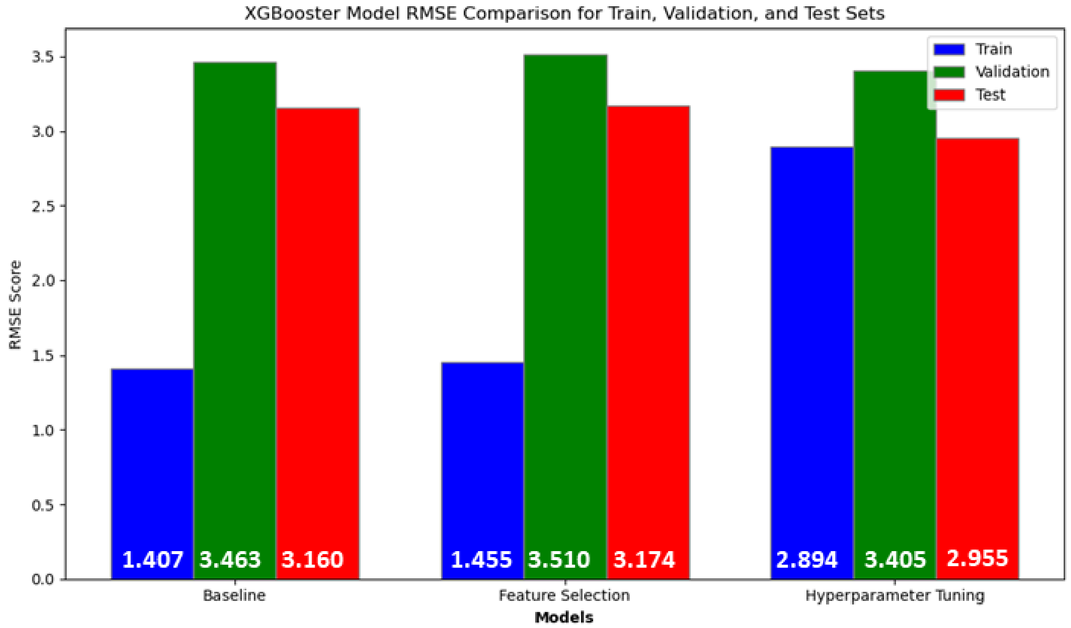




EVAL

Performance Metrics





SELECT

Comparisons between Models

Performance Metric (Test)	Tuned XGBooster Model	Tuned LGBM
R2	0.2240	0.2256
MAE	1.6187	1.5902
MSE	8.7101	8.6927
RMSE	2.9513	2.9483

SELECT

Comparisons between Models

- XGBooster likely needs to incorporate more of features
- Tuned LGBM better for real world usability
- Faster and more efficient
- Better for larger datasets

CONCLUSIONS

Conclusions

- Better models that incorporate more interacting features
- Precipitation highly variable therefore gather more information to use
- Missing features:

Prevailing winds, Topography (elevation), Pollution, Water bodies (distance from them eg: River Thames)

- Improve optimization and use forecast methods such as VARMA, DeepAR

Credits Slide

Data was attained from European Climate Assessment & Dataset

Reference

- https://www.ecad.eu/dailydata/index.php