City D Temperature Prediction

September 11, 2024

1 Team Details:

1.1 Team Name:

Edgerunners

1.2 Team Members:

- **V**endhan U - Team Leader
- **N**aveen Kumar S
- **P**avendhan D
- **P**aulin A

1.3 Github Link

https://github.com/PavendhanD/Shaastra-ML-Challenge.git

2 Problem Statement

The objective is to predict the 'TAVG' of CITY D using the data of the other three cities (CITY A, CITY B, CITY C). The dataset contains the geographical and meteorological data of three cities, including longitude, latitude, elevation, temperature, precipitation, and snow depth. There will always be errors in finding the accurate prediction of temperature, weather forecast, climatic changes and environment changes, since these are naturally occurring events, they can't be predicted accurately, but by finding the patterns beneath these events we can predict the events at certain level. The goal is to build and Machine learning model fitted with this dataset to find the TAVG of the CITY D efficiently.

3 Understanding the Problem

Prediction of temperature of City D here includes a dynamically changing variable influenced the multiple environmental and geographical factors. The dataset consists of this factors from three different Cities (A, B, C). The geographic features refers to latitude, longitude and elevations of temperature affected by various phenomenon like solar radiation and atmospheric pressure variations. For instance, the amount of solar energy is dependent on the latitude of the specific location. Other factors like precipitation and snow depth also influences the temperature by affecting surface albedo and latent heat release. To predict any city's temperature accurately, It's crucial to analyze the relationship between all the features from other relevant cities and also to find the influence on the temperature patterns. By leveraging this datapoint, we can supposedly build a prediction model that generalizes to City D based on other environmental and geographical factors.

4 Understanding the Data Structure

This dataset comprises of the information of three cities, they provide the date , geographical and meterological data of CITY A, CITY B, CITY C.

4.1 Date Information

The date is provided for each and every recording of data for all the three cities. It includes the day, month and year of the recorded data.

4.2 Geographical Information

- **Latitude**: The latitudes coordinates of three cities are provided which affects the temperature by having varying intensity of solar radiation.
- **Longitude**: The longitude coordinates of the three cities are provided which determines coastal lines and other geographical features.
- **Elevation**: This provides the altitudes of the three cities which determines the temperature being cooler at higher altitudes

4.3 Meteorological Features

- **Precipitation**:This provides the recorded precipitation level for all the three cities which is responsible for the cloud cover and latent heat exhaustion.
- **Snowfall**: This provides the recorded snowfall level for the three cities which affects the temperature.
- **Temperature**: The Maximum, minimum and Average temperature is provided for the three cities.

TAVG is the target variable for our model, which can be predicted by building a model and training the model with the data of other three cities.

5 Model Architecture

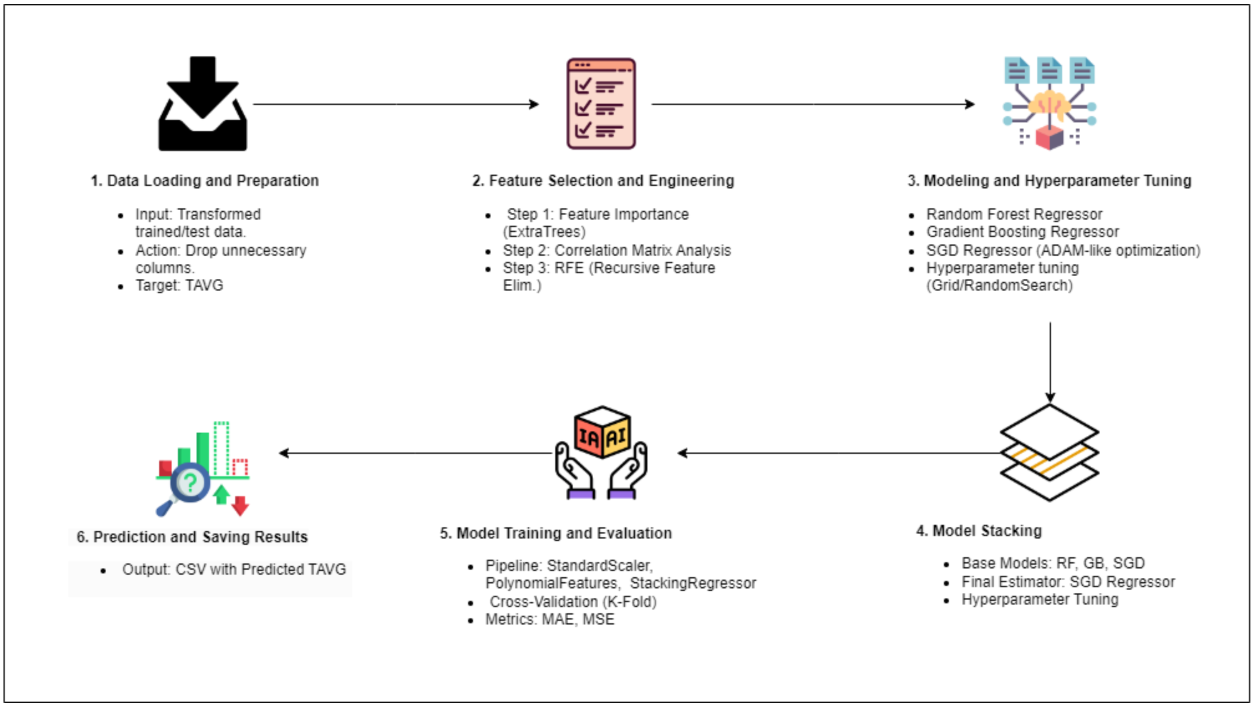
1. Data Loading and Preparation
Input: Transformed trained/test data.
Action: Drop unnecessary columns.
Target: TAVG
2. Feature Selection and Engineering
Step 1: Feature Importance (ExtraTrees)
Step 2: Correlation Matrix Analysis
Step 3: RFE (Recursive Feature Elim.)
3. Modeling and Hyperparameter Tuning
Random Forest Regressor
Gradient Boosting Regressor
SGD Regressor (ADAM-like optimization)
Hyperparameter tuning (Grid/RandomSearch)
IA AI
6. Prediction and Saving Results
Output: CSV with Predicted TAVG
5. Model Training and Evaluation
Pipeline: StandardScaler, PolynomialFeatures, StackingRegressor
Cross-Validation (K-Fold)
Metrics: MAE, MSE
4. Model Stacking
Base Models: RF, GB, SGD
Final Estimator: SGD Regressor
Hyperparameter Tuning

Figure 1: Architecture Diagram

6 Feature Engineering

This is one of the most important aspect of machine learning as it is necessary to bring out the maximum from a dataset to be used on the model. New features can be engineered based on the limited data this helps the model to better understand the correlation between data. Multiple feature engineering techniques are used in this model for the betterment of the dataset they are,

- **Date Feature Engineering**: The date column is in the format of day/moth/year which cannot be used for modelling, so new features are modelled from the date data. They are engineered to different features such as day of the week, month, year, day of the year and sine, cosine for cyclic features.
- **Distance Calculations**: Geographical distance and elevation differences are calculated between the cities by finding the differences between the geographical features.
- **Weather-Based Features**: : new features are calculated like temperature differences, cummulative temperature and temperature anomalies to understand the patterns in the data.
- **Interaction Features**: engineering new features by interacting the variables together which results in finding the interaction between the various features of the dataset
- **Statistical Summary Features**: Mean, variance and median of temperature and precipitation are calculated for understanding the trends in the dataset.
- **Lagged Features**:By adding the lagged values of temperature and precipitaion to find the temporal dependencies in the dataset
- **Location-Based Clustering**: Clustering techniques like Kmeans clustering are used to group the data of same and similar locations.
- **Polynomial Features**:To transform features to polynomial combinations, we use polynomial features which enhances the model's ability to better understand the complexity between the features
- **Data Scaling**:A Standard scaler is used to standardize the data. It is very crucial for the data have equal contribution to the model, SGD and gradient boosting will require scaling of data for better performance.

7 Feature Selection

We have used the following various feature selection methods sequentially to simplify the models complexity:

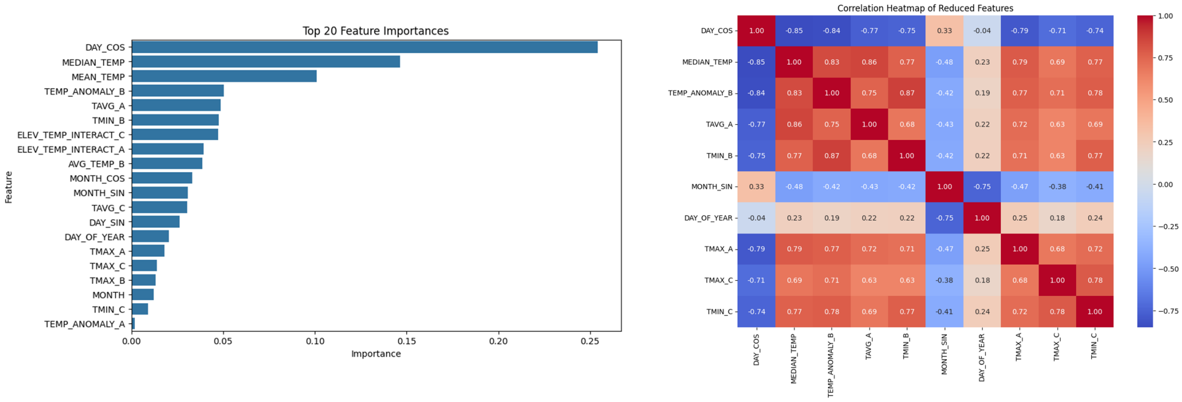

Figure 2: Feature Importance Visualization

7.1 Feature Importance via ExtraTreesRegressor

We have use ExtraTressregressor to rank features based on their feature importance in the prediction of TAVG and scrutinized the top 20 features.

7.2 Correlation Matrix

Identification and removal of highly correlated features for less redundancy and more model generalization.

7.3 Recursive Feature Elimination

Recursive elimination of significantly less important features using a simple Linear regression model to select the final top 10 features.

8 Optimization Techniques

- **Nesterov Accelerated Gradient (NAG)**:We used NAG to improve the convergence speed and its efficiency by find the next gradient position in prior in turn reducing the oscillations and settling on local minima.
- **Randomized Search for Initial Hyperparameter Tuning**:We used RandomizedSearchCV sampled parameter combinations to efficiently search the parameter space, enabling the model for further fine-tuning.
- **Grid Search for Precise Hyperparameter Optimization**: After narrowing down parameter ranges, GridSearchCV was used then to find the parameters combinations for ensured optimal performance.

9 Base Models

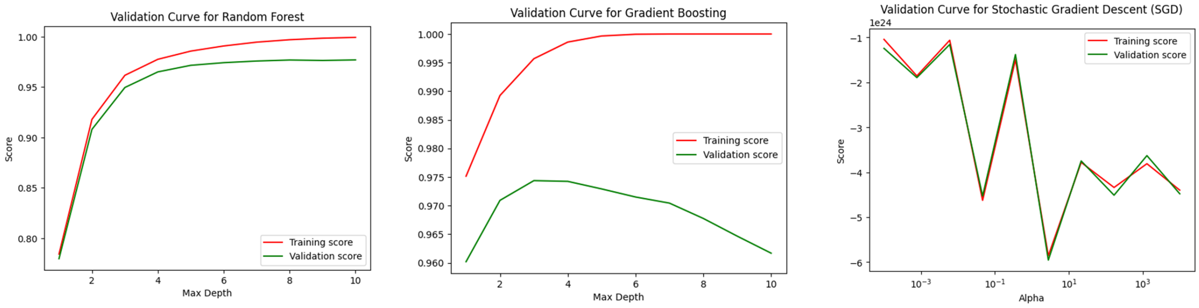

Figure 3: Model Performance Visualization

9.1 Random Forest Regressor

: Ensemble Model: Very useful due to its capability to work with high dimensional data while capturing linear and non-linear relationships. Tuned parameters n estimators and max depth for overfitting and performance measures.

9.2 Gradient Boosting Regressor

: Iterative Learning: Used due to that is characterized by its iteratively correcting its own mistakes and trying to learn every subtle detail. Tuned learning rate and n estimators for the reason of learning rate and overfitting.

9.3 Stochastic Gradient Descent (SGD) with Elastic Net Regularization

: Efficient Optimization: Great for large datasets, and effective with resistant problems when fine-tuned using Elastic net regularization, which encourages sparsity while maintaining a low possibility of overfitting.

10 Stacking Regressor for Model Integration

Stacking model is used to combine multiple models to bring the best out of the the base models. Random forest, Gradient Boosting and SGD models are stacked up together to predict the target variable of the model, This increase diverse capabilities of the model. The final model is incorporated with a meta-model for better prediction

10.1 Hyperparameter Tuning for Stacking Model

GridsearchCV was used to tune the parameters of the stacking model to identify the best fit stacking model, parameters like alpha and l1 ratio for elastic net regularisation to enhance model's performance

11 Evaluation of the Model

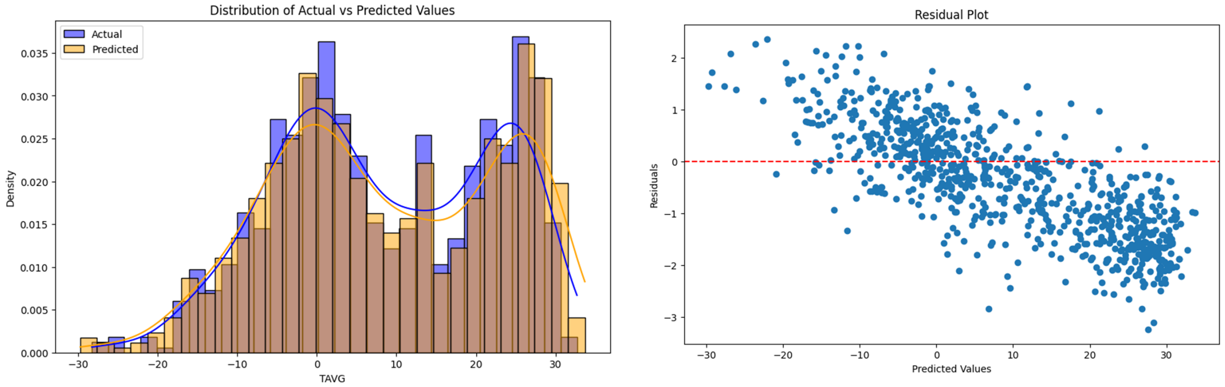

Figure 4: Final Evaluation

K Fold Cross Validation was applied next with 5 Splits and its effectiveness for prediction of the model performance on unseen data was assessed. Precision and Magnitude of Error were assessed against Mean Absolute Error (MAE) and Mean Squared Error (MSE). Training of the Model and Efficiency Assessment The last model was established and assessed through MAE and MSE, which reveals prediction's correctness and its errors distributions.

- **Final Prediction and Submission** For City D, forecasting was done and finalised to the last model saving its predictions in Csv format for submission or for further analysis.

12 Impact of the Model

The ability to forecast the average temperature of City D accurately has a number of benefits that come with it. It will improve meteorological service when predicting extreme weather events like cold snaps and heat waves; hence, assisting disaster management. In climate studies, understanding regional temperature change will improve mitigation and adaptation policies. Precise forecasting in agriculture helps one to plan better and minimize the risks involved. Energy management-for example-allows for highly accurate predictions; thereby optimizing the grid operations for better integration of renewable energies. This also relates to urban planning concerning urban heat island effect. On the whole, this project is aimed at promoting environmental modeling towards data-driven decisions for a more sustainable world.