TERM PROJECT REPORT

CLIMATE IMPACT PREDICTIONS

Problem Statement:

Climate Change and Agricultural Crop Production:

Climate change presents a profound and multifaceted threat to global agricultural systems, posing significant challenges for sustainable food production. The rising average temperatures, increasingly unpredictable rainfall patterns, and heightened frequency of extreme weather events are putting immense pressure on crop yields. This disruption is particularly acute for staple crops such as maize, wheat, soybean, and rice, which form the backbone of food systems in many regions of the world.

Rising temperatures can induce heat stress in crops, particularly during their critical flowering and fruiting stages, leading to a direct reduction in yields. For example, maize and wheat are highly sensitive to temperature fluctuations, with even moderate increases in heat during their growing periods drastically lowering their productivity. Moreover, regions that were once suitable for these crops are becoming too hot or dry to sustain them, forcing shifts in agricultural practices or crop choices.

In tandem, the availability of water is becoming a major concern. More frequent and severe droughts reduce the water available for irrigation, stunting crop growth or even leading to total crop failure in regions reliant on consistent water supply. On the other hand, unpredictable rainfall patterns make it difficult for farmers to plan planting and harvesting schedules, further exacerbating the stress on agricultural systems.

Additionally, warmer temperatures and altered precipitation patterns are creating favorable conditions for pests and diseases to thrive. Many pests that previously were contained within specific regions are now spreading to new areas, affecting crops that previously had limited exposure to these threats. As pests and diseases migrate, they increase the likelihood of crop damage and loss, further jeopardizing food security.

The impacts of climate change are not only a challenge to current agricultural systems but also pose a serious risk to achieving sustainable food production. Understanding how these environmental factors interact with crop yields is crucial for developing adaptive strategies that ensure global food security in the face of an evolving climate.

Objective:

The goal of this project is to **predict the impact of environmental changes on crop yields**. Specifically, the aim is to predict the percentage change in crop yields (relative to baseline periods) based on various environmental parameters such as temperature, precipitation, and CO_2 concentration, etc. In doing so, we will identify the most critical environmental factors driving changes in crop productivity and explore the relationships between these factors and yield projections. By applying machine learning models, we seek to gain a deeper understanding of how complex interactions between climate variables impact agricultural systems and how these insights can be used to inform climate-resilient agricultural practices.

Collection and Description of Dataset:

This project uses a comprehensive dataset that compiles 8703 simulations from 202 peer-reviewed studies published between 1984 and 2020. These simulations provide yield projections for four major crops: maize, soybean, rice, and wheat, which are critical for global food systems. The dataset spans 91 countries and captures a broad range of environmental and socio-economic conditions.

The dataset is structured to include projections of crop yields under different climate change scenarios, with and without adaptation measures. Each simulation includes:

- **Geographical coordinates** (latitude and longitude) to represent location-specific data.
- **Current temperature and precipitation levels** to set a baseline for existing climate conditions.
- **Projected changes in temperature and precipitation** for the 21st century based on different greenhouse gas emission scenarios.
- **Crop species** (maize, soybean, rice, wheat) and **CO2 concentrations**.
- **RCP (Representative Concentration Pathways) scenarios**: These represent different potential trajectories of greenhouse gas emissions, which influence global climate outcomes. The dataset includes:
  - **RCP2.6**: A scenario with low GHG emissions and significant mitigation efforts.
  - **RCP4.5**: A stabilization scenario where emissions peak by mid-century and decline afterward.
  - **RCP6.0**: A stabilization scenario where emissions peak later, around 2080.
  - **RCP8.5**: A high-emission scenario with no significant mitigation and a large increase in GHG emissions.

The dataset also includes **relative changes in yield** expressed as a percentage deviation from a baseline period. These yield projections are modeled both with and without the effects of increased CO2 concentrations and with or without **adaptation strategies** such as modified irrigation practices, planting dates, or the use of more resilient crop varieties.

Key Features:

- **Crop species**: Maize, soybean, rice, wheat.
- **Geographical location**: Latitude and longitude for each simulation.
- **CO2 emission scenarios**: RCP2.6, RCP4.5, RCP6.0, RCP8.5.
- **Current climate data**: Baseline temperature and precipitation levels.
- **Projected climate data**: Changes in temperature and precipitation over time.
- **Local and global warming**: Measured as degrees of temperature increase.
- **Yield impact**: Projected relative yield changes, expressed as a percentage change from the baseline period.
- **Adaptation measures**: Indications of whether adaptation strategies were applied in the simulation.

Data Preprocessing:

The original raw dataset had 8703 rows and 52 feature columns. However, 14 of these columns did not provide any relevant information and to use for prediction purposes. These 14 columns included features like ID, Reference number, Reference, doi, Publication year, etc. Moreover, some of these columns had just a few hundred observations. So, during the first step, we removed these irrelevant columns from the dataset, i.e., columns 0 to 2 and 41 to 51. The picture below represents the information about the original dataset.

```
RangeIndex: 8703 entries, 0 to 8702
Data columns (total 52 columns):
 #   Column                                                       Non-Null Count  Dtype
---  ------                                                       --------------  -----
 0   ID                                                           8703 non-null   int64
 1   Ref No                                                       8703 non-null   object
 2    Methods                                                     8703 non-null   object
 3   Scale                                                        8703 non-null   object
 4   Crop                                                         8703 non-null   object
 5   Country                                                      8703 non-null   object
 6   Site(location)                                               2694 non-null   object
 7   Region                                                       8703 non-null   object
 8   latitude                                                     8666 non-null   float64
 9   longitude                                                    8666 non-null   float64
 10  Current Average Temperature (dC)_area_weighted               8666 non-null   float64
 11  Current Average Temperature_point_coordinate (dC)            8666 non-null   float64
 12  Current Annual Precipitation (mm) _area_weighted             8666 non-null   float64
 13  Current Annual Precipitation  (mm) _point_coordinate         8666 non-null   float64
 14  Future_Mid-point                                             8703 non-null   int64
 15  Baseline_Mid-point                                           8702 non-null   float64
 16  Time slice                                                   8703 non-null   object
 17  Climate scenario                                             8703 non-null   object
 18  Scenario source                                              8703 non-null   object
 19  Local delta T                                                4392 non-null   float64
 20  Local delta T from 2005                                      8666 non-null   float64
 21  Annual Precipitation change each study  (mm)                 3554 non-null   float64
 22   Annual Precipitation change  from 2005 (mm)                 8666 non-null   float64
 23  Global delta T from pre-industrial period                    8703 non-null   float64
 24  Global delta T from 2005                                     8703 non-null   float64
```

```
26  Climate impacts (%)                                            8703 non-null   float64
27  Climate impacts relative to 2005                               8703 non-null   float64
28  Climate impacts per dC (%)                                     8703 non-null   float64
29  Climate impacts per decade (%)                                 8703 non-null   float64
30  CO2                                                            8703 non-null   object
31  CO2 ppm                                                        8538 non-null   float64
32  Fertiliser                                                     8703 non-null   object
33  Irrigation                                                     8703 non-null   object
34  Cultivar                                                       8703 non-null   object
35  Soil organic matter management                                 8703 non-null   object
36  Planting time                                                  8703 non-null   object
37  Tillage                                                        8703 non-null   object
38  Others                                                         8703 non-null   object
39  Adaptation                                                     8703 non-null   object
40  Adaptation type                                                8703 non-null   object
41  Reference                                                      8703 non-null   object
42  doi                                                            8493 non-null   object
43  Publication year                                               8703 non-null   int64
44  Note1
(* = corrected by HW)                                  333 non-null    object
45  Note2
(* = Local temperature is  estimated )                 4274 non-null   object
46  Note3
(* = Local delta Pr is  estimated )                    5399 non-null   object
47  Note4
(* = Global temperature is  estimated )                594 non-null    object
48  Seasonal Precipitation change (mm) each study (local baseperiod)  386 non-null    float64
49  Base precipitation (annual) (mm) (local base period)           1917 non-null   float64
50  Annual Preciptation change (%) (relative to local base)        366 non-null    float64

51  Base precipitation (seasonal) (mm) (local base period)         189 non-null    float64
```

After removing irrelevant columns, we were left with 38 features. In the next step, we found certain columns with too many missing values. For these features removing the rows with missing values was not possible because of the small dataset and imputing so many missing values would also have introduced noise in the dataset. So, we decided to remove the columns with >50% missing values.

Hence, we end up removing 5 more columns from the dataset, resulting in a total column number of 33.

Next, we observed that there were four feature columns that could be used as the response variable. These features were- Climate impacts (%), Climate impacts relative to 2005, Climate impacts per dC (%), Climate impacts per decade (%). These four features tell the same thing but from four different perspectives. We decided to keep the feature “Climate impacts (%)” as the sole response variable in our dataset and removed the other three redundant variables. As a result, we had 30 columns in total at this step, including our response variable. Below is the description of our final dataset after removing the irrelevant, redundant, and sparse variables.

```
RangeIndex: 8703 entries, 0 to 8702
Data columns (total 30 columns):
 #   Column                                                 Non-Null Count  Dtype
---  ------                                                 --------------  -----
 0   Scale                                                  8703 non-null   object
 1   Crop                                                   8703 non-null   object
 2   Country                                                8703 non-null   object
 3   Region                                                 8703 non-null   object
 4   latitude                                               8666 non-null   float64
 5   longitude                                              8666 non-null   float64
 6   Current Average Temperature (dC)_area_weighted         8666 non-null   float64
 7   Current Average Temperature_point_coordinate (dC)      8666 non-null   float64
 8   Current Annual Precipitation (mm) _area_weighted       8666 non-null   float64
 9   Current Annual Precipitation  (mm) _point_coordinate   8666 non-null   float64
 10  Future_Mid-point                                       8703 non-null   int64
 11  Baseline_Mid-point                                     8702 non-null   float64
 12  Time slice                                             8703 non-null   object
 13  Climate scenario                                       8703 non-null   object
 14  Local delta T from 2005                                8666 non-null   float64
```

```
 15   Annual Precipitation change  from 2005 (mm)           8666 non-null   float64
 16  Global delta T from pre-industrial period              8703 non-null   float64
 17  Global delta T from 2005                               8703 non-null   float64
 18  Climate impacts (%)                                    8703 non-null   float64
 19  CO2                                                    8703 non-null   object
 20  CO2 ppm                                                8538 non-null   float64
 21  Fertiliser                                             8703 non-null   object
 22  Irrigation                                             8703 non-null   object
 23  Cultivar                                               8703 non-null   object
 24  Soil organic matter management                         8703 non-null   object
 25  Planting time                                          8703 non-null   object
 26  Tillage                                                8703 non-null   object
 27  Others                                                 8703 non-null   object
 28  Adaptation                                             8703 non-null   object
 29  Adaptation type                                        8703 non-null   object
```

Imputing missing values:

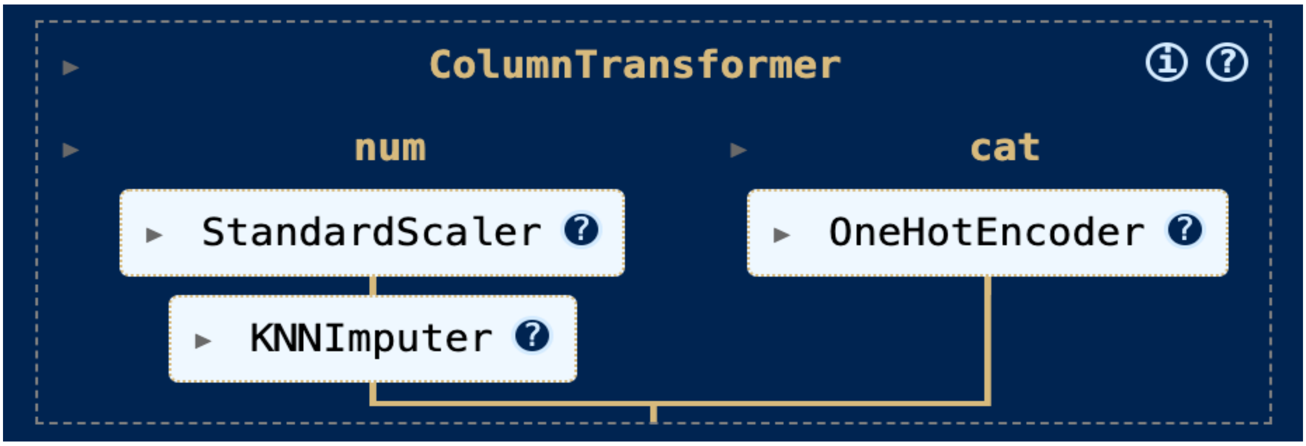
Next step was to deal with the missing values in the final dataset. If you notice in the description of our final dataset in the picture above, there were 8 columns with the same number of missing values. These features were- latitude, longitude, Current Average Temperature (dC)_area_weighted, Current Average Temperature_point_coordinate (dC), Current Annual Precipitation (mm)_area_weighted, Current Annual Precipitation (mm) _point_coordinate, Local delta T from 2005, and Annual Precipitation change from 2005 (mm). They all had 37 missing observations. After seeing this interesting pattern, we decided to look deeper into this and found that they all had missing observations corresponding to the level "Global" of the categorical variable "Scale". Hence, we decided to impute the missing values in each of these 8 variables with the mean value of these 8 variables corresponding to the level “Global” of the categorical variable “Scale”.

After we impute these 8 variables, we were left with just two columns with missing values. Both were numerical, namely, Baseline_Mid-point (1 missing value) and CO2 ppm (165 missing values). We chose to impute these missing values using KNNImputer from scikitlearn library.

Feature scaling and encoding:

Our final dataset had 30 columns, out of which 16 were categorical and remaining numerical. The numerical features were scaled using StandardScaler class from scikitlearn and the categorical features were encoded using One-Hot encoder from scikitlearn library. Scaling numerical features is very important especially for gradient based machine learning models like linear regression models, and neural networks.

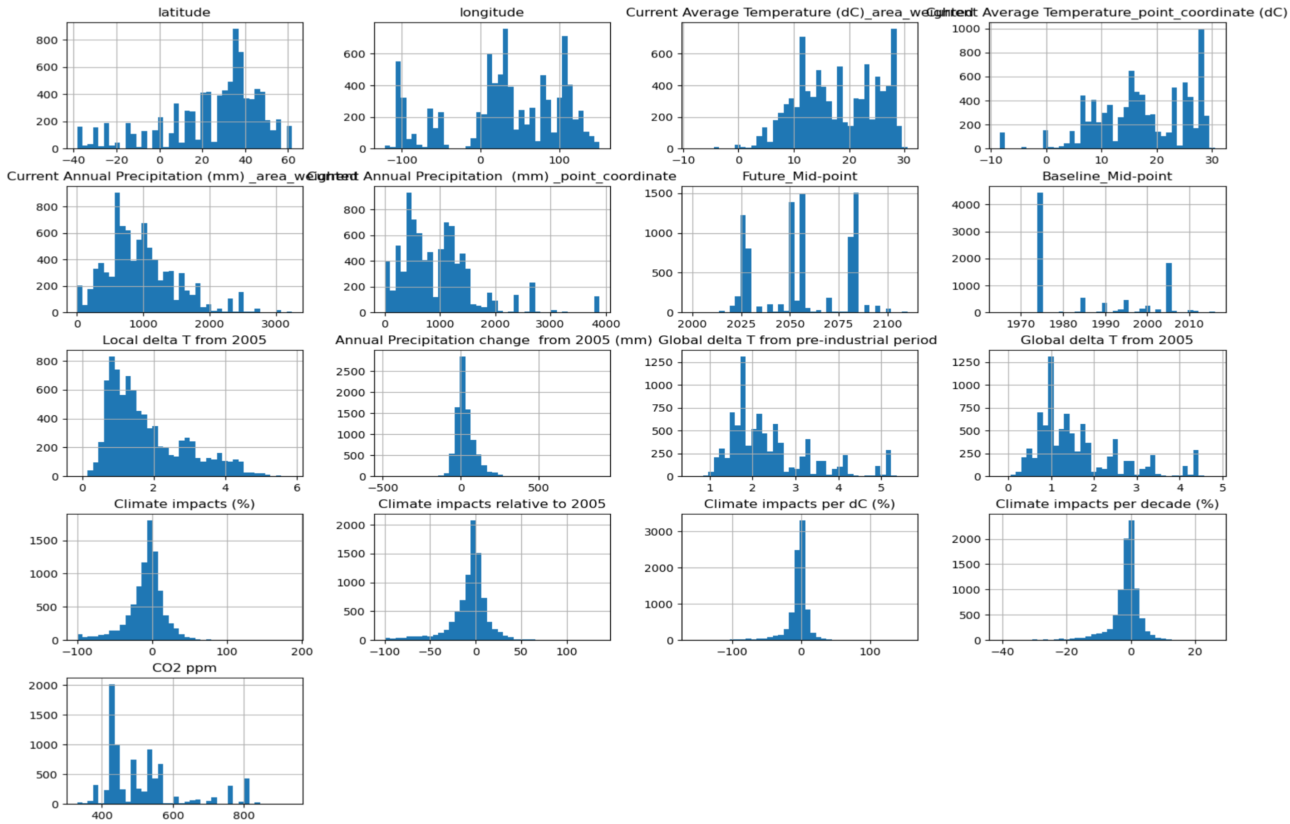
Below is the representation of our preprocessing pipeline:



Since we had 16 categorical features with several of them having multiple levels, we ended up having 172 features in total after one-hot encoding from preprocessing step. It is important to note that several of these 172 features had sparse entries, hence feature importance analysis becomes very important here.

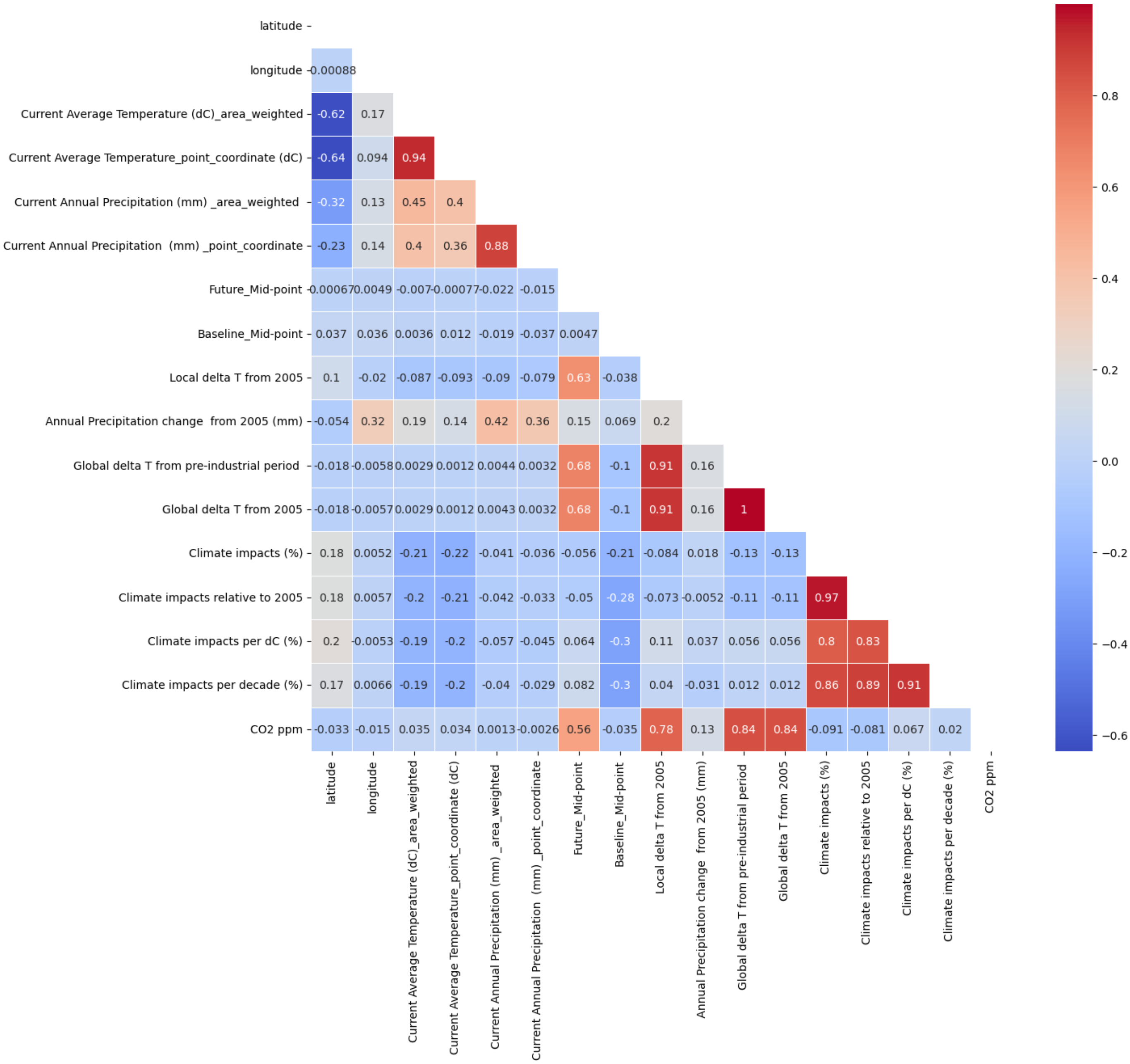
Data Visualization:

To get an idea of the distribution of each of the numerical features, we looked at the distribution plots. We noticed that although several of the features had continuous numerical type distributions, but two of them, namely, Future_Mid-point and Baseline_Mid-point had discrete numerical type distributions, where few of the observations are repeated majority of the times. Under different scenarios, these types of variables can be tackled in different ways, but we decided to treat them as numerical type only.

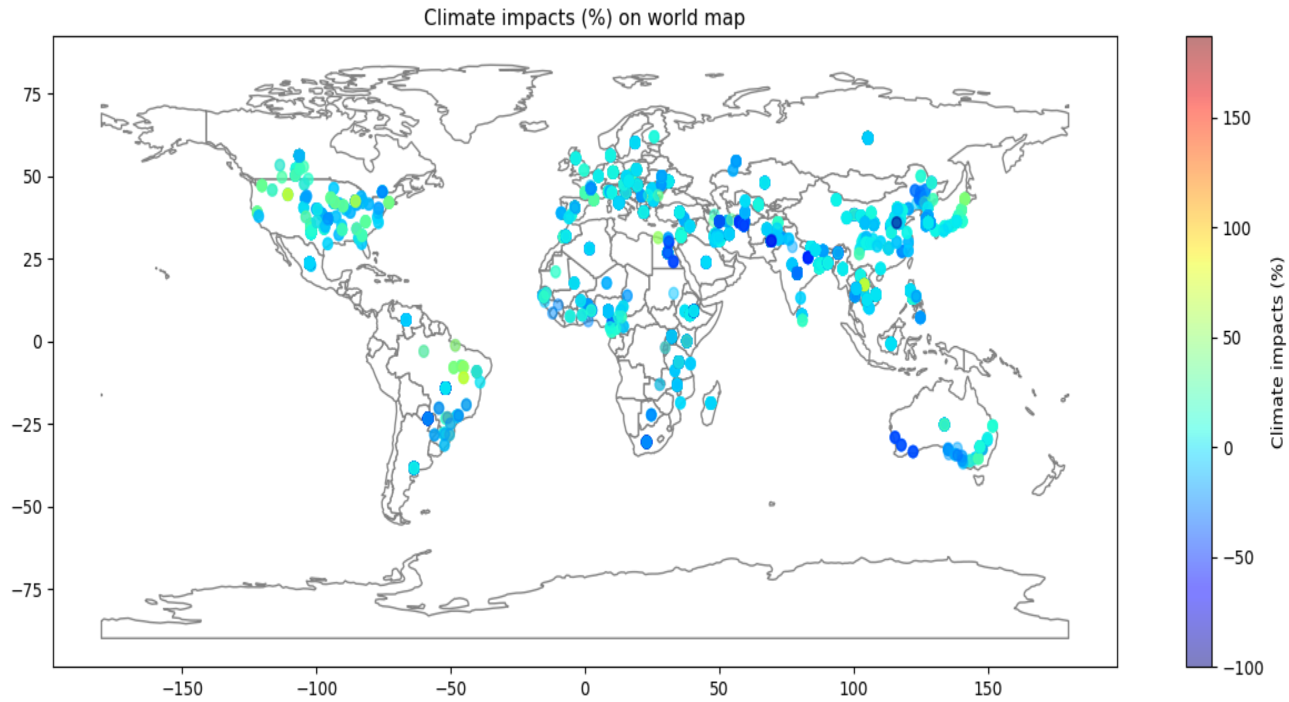


Next, we looked at the correlation between the numerical variables. Several of the input variables showed high correlation among themselves, whereas the response variable itself did not come out to be highly correlated to any of the single numerical variables. Again, it became more evident after this that feature importance will be important to remove multi-collinearity among the input features and reduce over-fitting of the models.

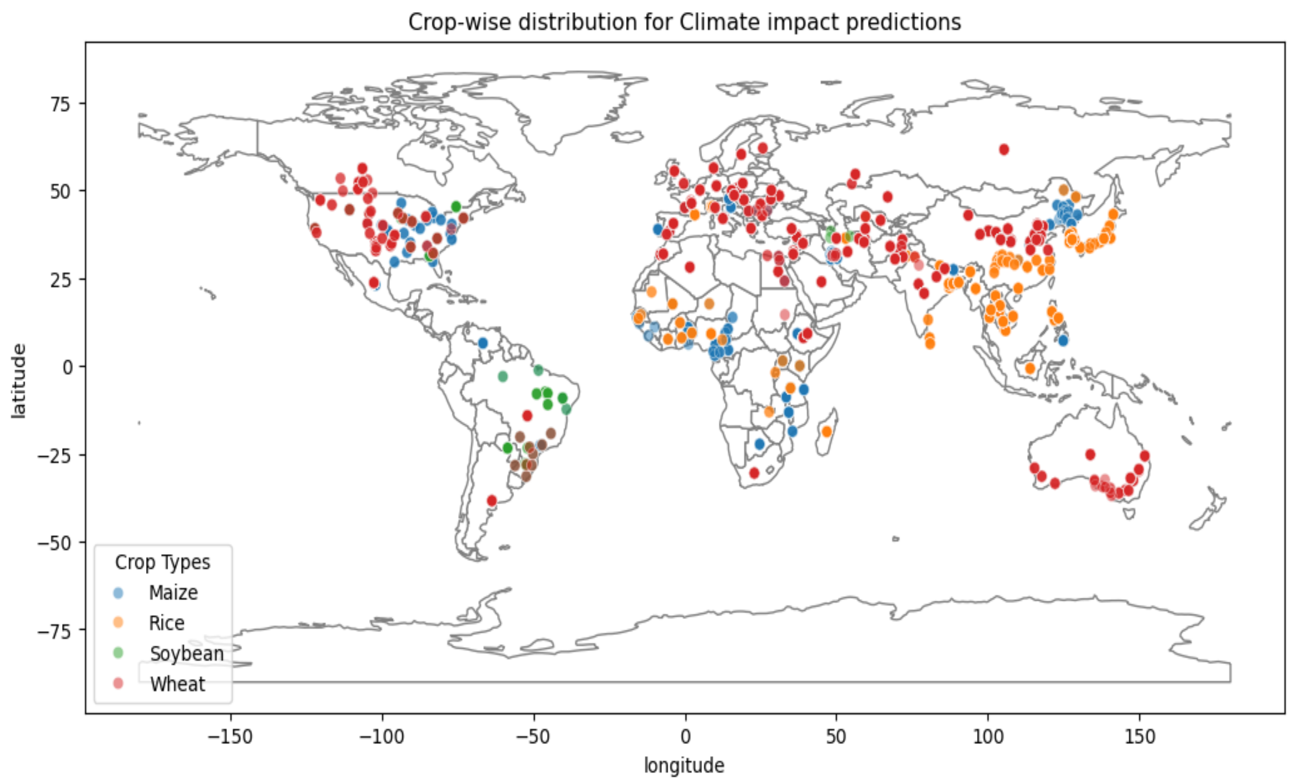
Below is the correlation matrix for the reference:



Next, we were interested to look at the geographical distribution of our response variable on the world map. This helped us understand the severity of the climate change impacts on crop yields around different parts of the world. In the map below, the blue dots indicate negative climate impacts and darker the color more severe the impacts.



Similarly, we were interested to see the crop-wise distribution of the response variable around the world. We observed that the majority of the climate impact predictions for rice came out from Asia and then Africa. Whereas, for wheat, the predictions were more spread out in the world. Predictions for soybeans mostly came from the US, and South America. Following is the map for further elaboration:



Modeling Techniques:

We performed modeling using the following models:

(a) Regression Models- Linear, Ridge, Lasso, Symbolic, Polynomial

(b) Decision Tree Regressor

(c) Random Forest Regressor

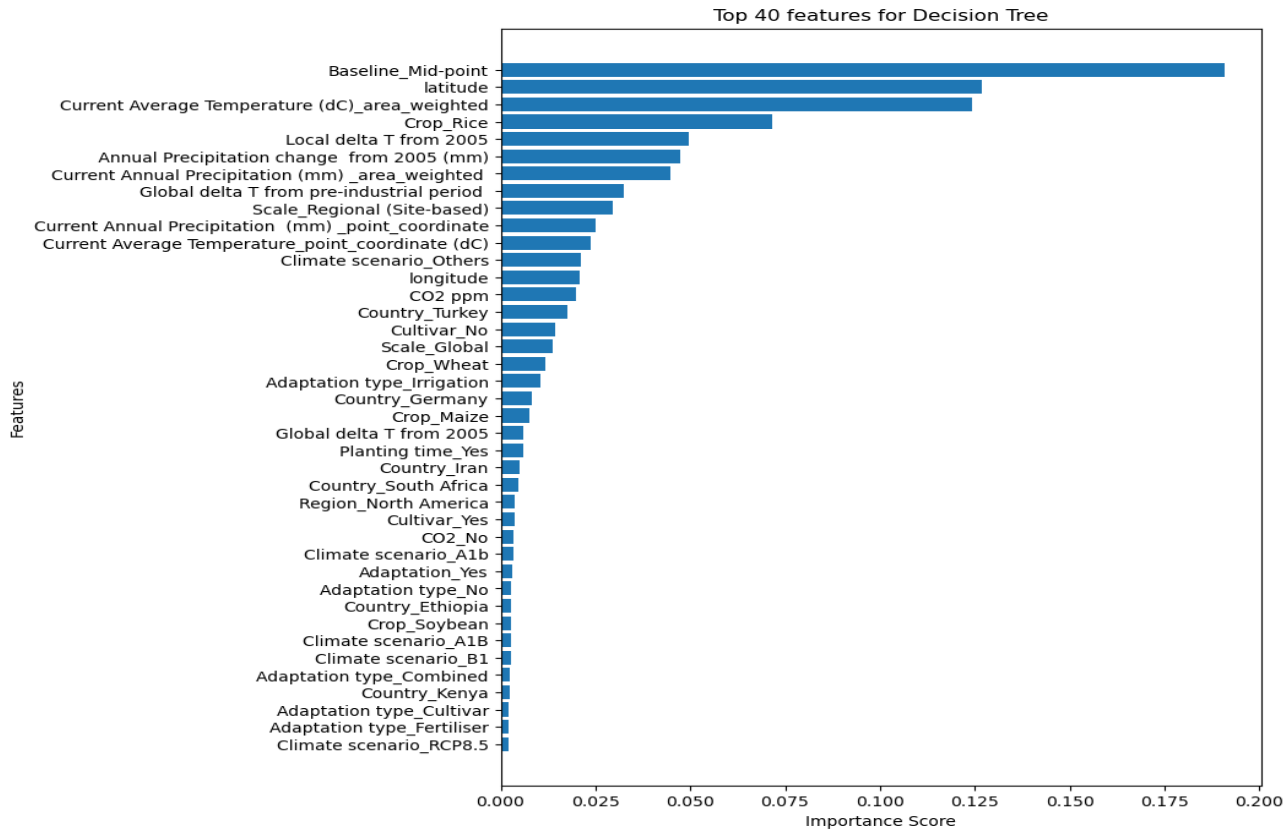
(d) XG Boost Regressor

(e) Neural Network Models- 1-layer, 3-layer, 4-layer, progressive, wide-deep net

The thought process behind choosing the modeling techniques was to have a diverse group of models with varying complexity levels to capture variability in the data. The regression models span from a simple linear regression to a complex polynomial regression. The decision tree regressor model is a very common supervised machine learning technique, so we decided to use it. The idea behind using random forest regressor and XGBoost was to improve upon the results of linear regressions and decision tree models. Then, to capture the maximum variability in the data, we trained different kinds of neural networks as well. Our assumption was that the neural nets will outperform all other modeling techniques. However, the results indicate something else. Their efficacy turned out to be like that of the random forest and XGBoost models.

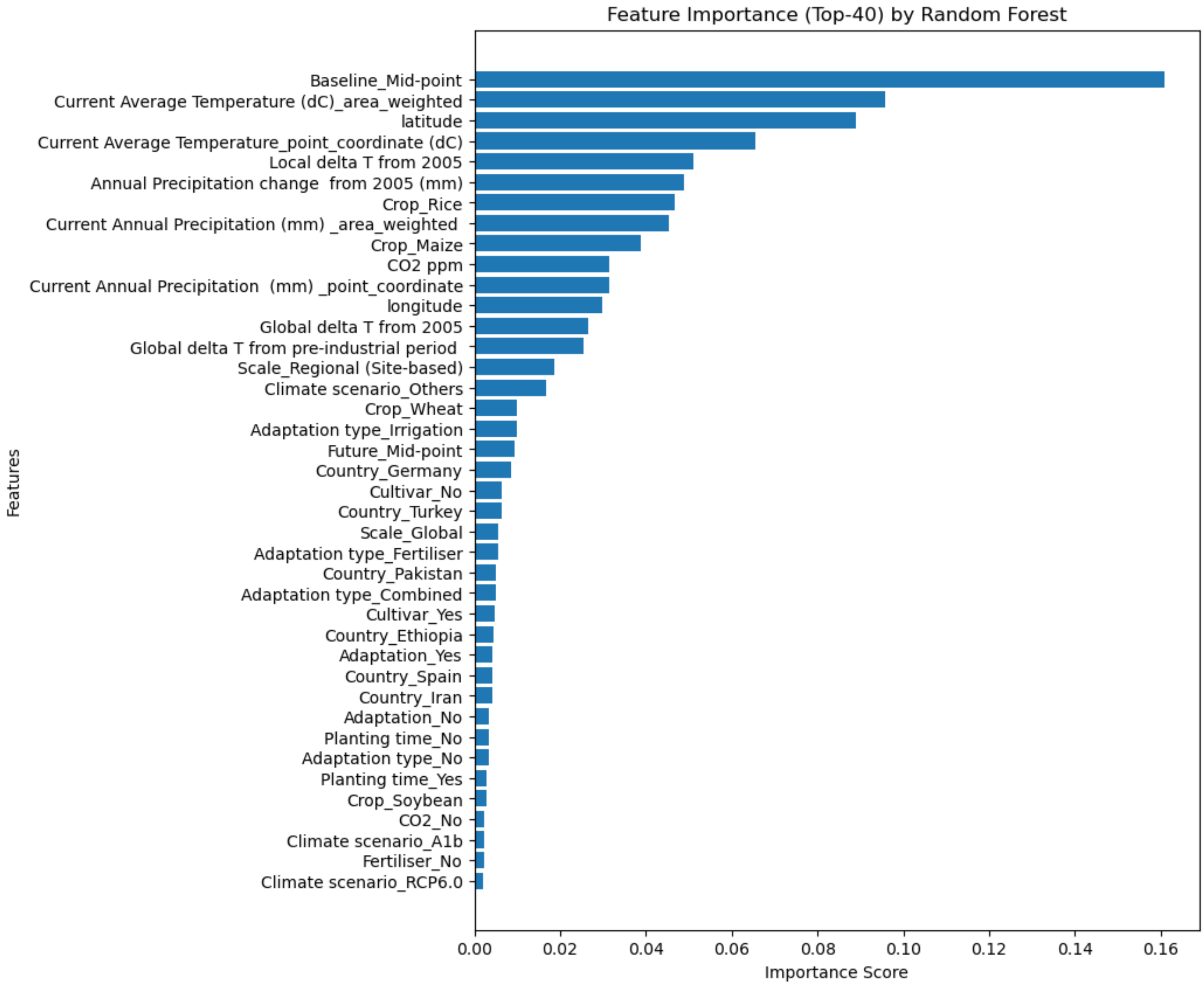
Feature Selection:

The feature selection analysis used in this project was inherently part of the decision tree regressor and random forest regressor models. We did not compute separate feature selection procedures in this project. Below are the top 40 features (out of 172) by decision tree and random forest regressor models:



The top 6 features according to the decision tree model are:
(1). Baseline_Mid-point, (2). Latitude, (3). Current Average Temperature (dC)_area_weighted,

(4). Crop_Rice, (5). Local delta T from 2005, (6). Annual Precipitation change from 2005 (mm)



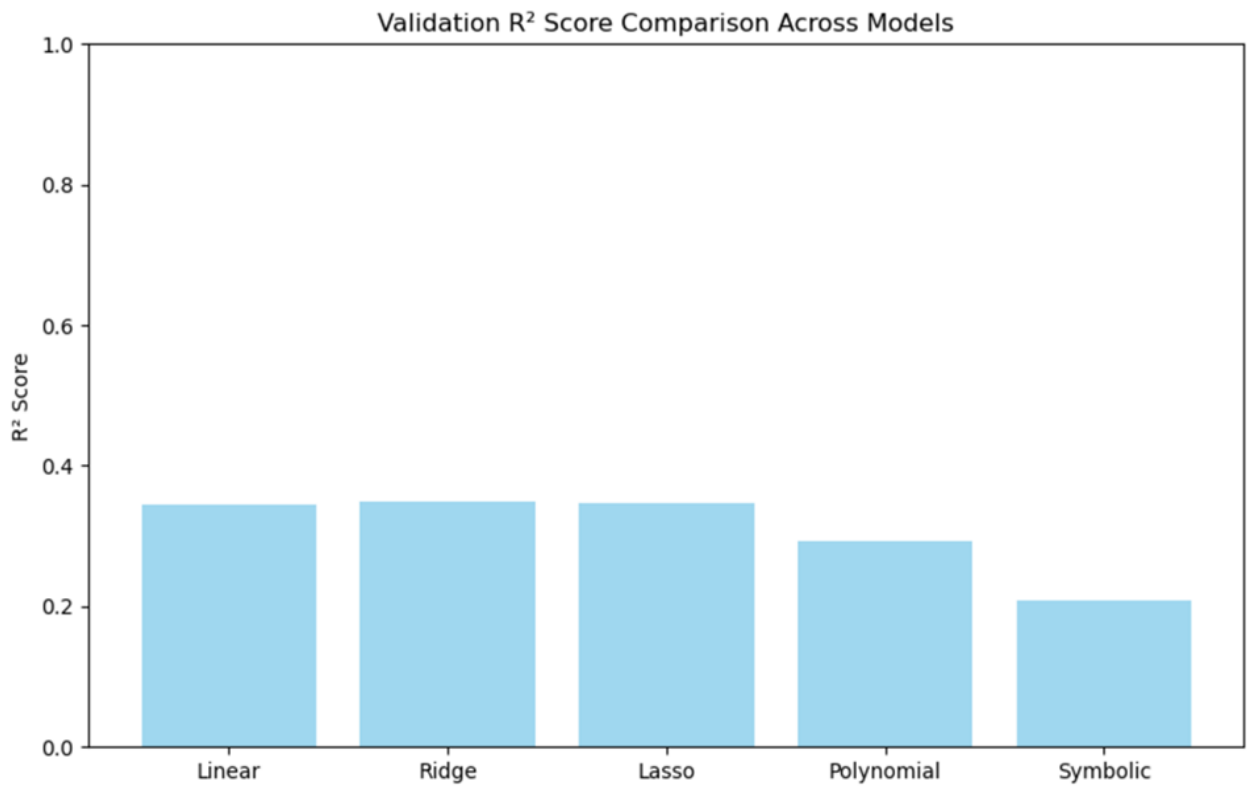
The top 6 features according to the random forest model are:
(1). Baseline_Mid-point, (2). Current Average Temperature (dC)_area_weighted, (3). Latitude,

(4). Current Average Temperature_point_coordinate, (5). Local delta T from 2005, (6). Annual Precipitation change from 2005 (mm)

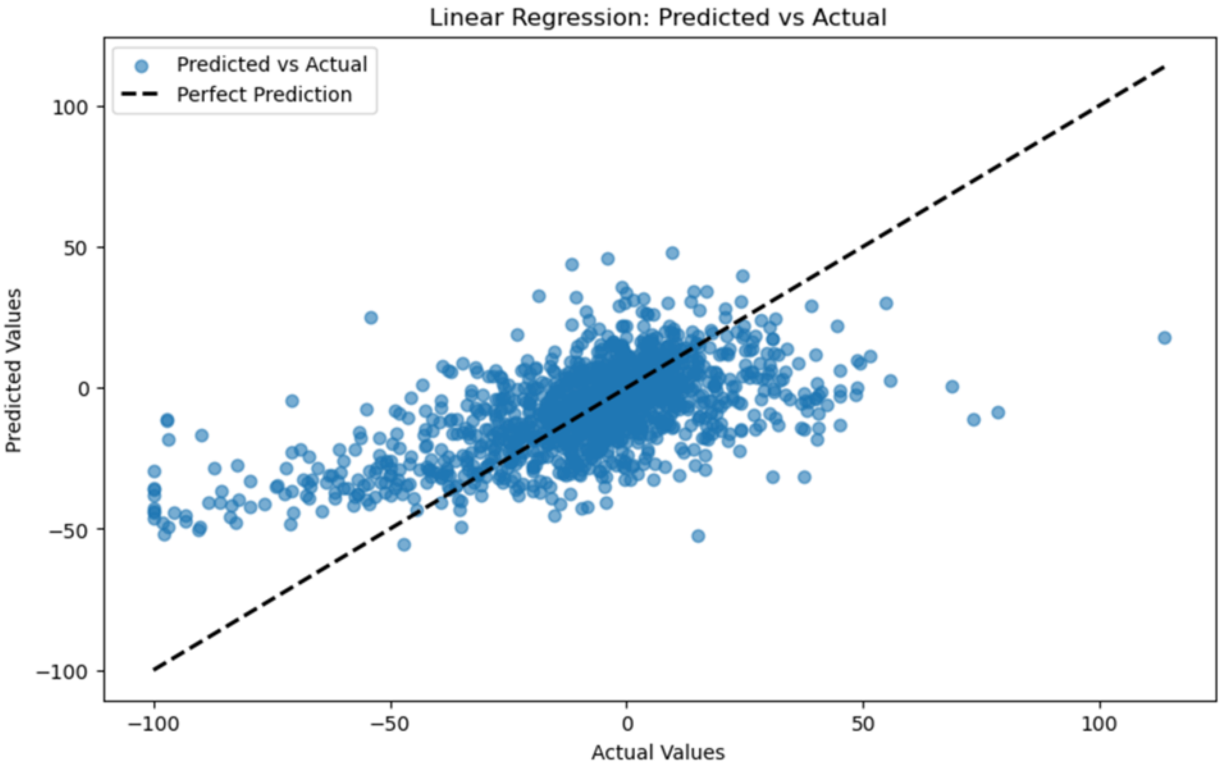
Results:

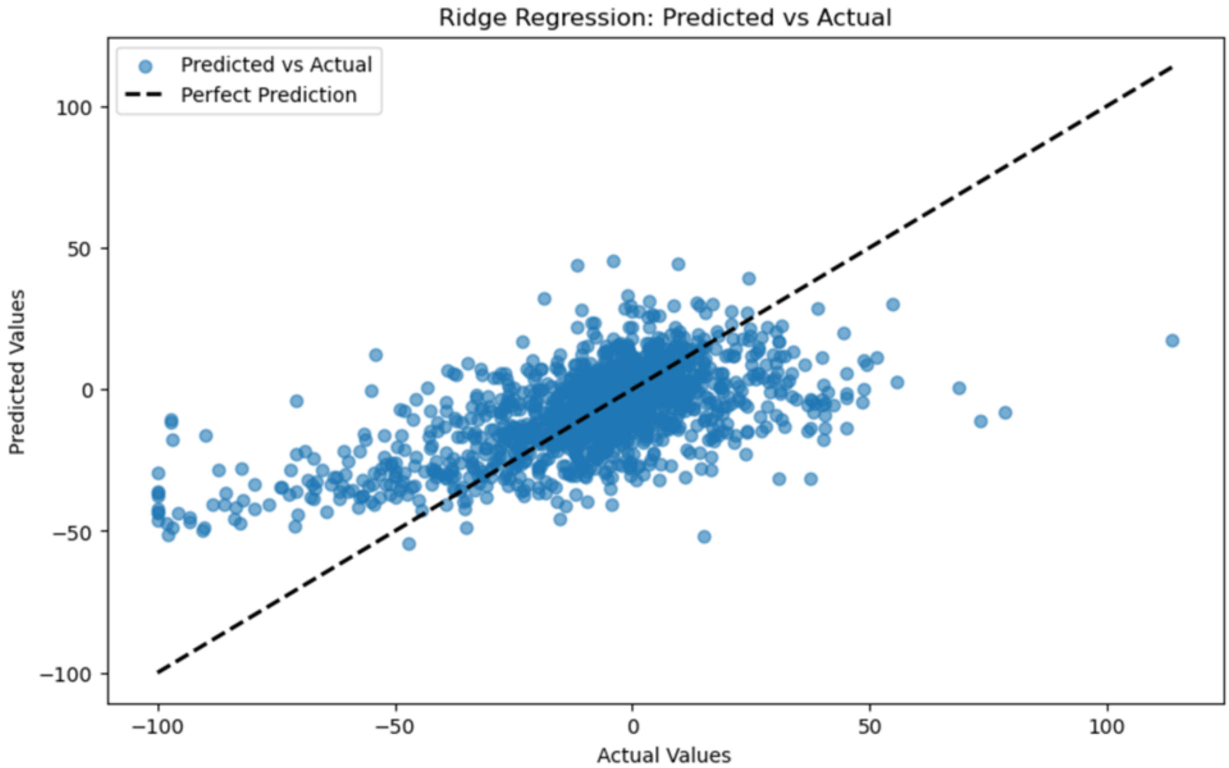
Regression models -

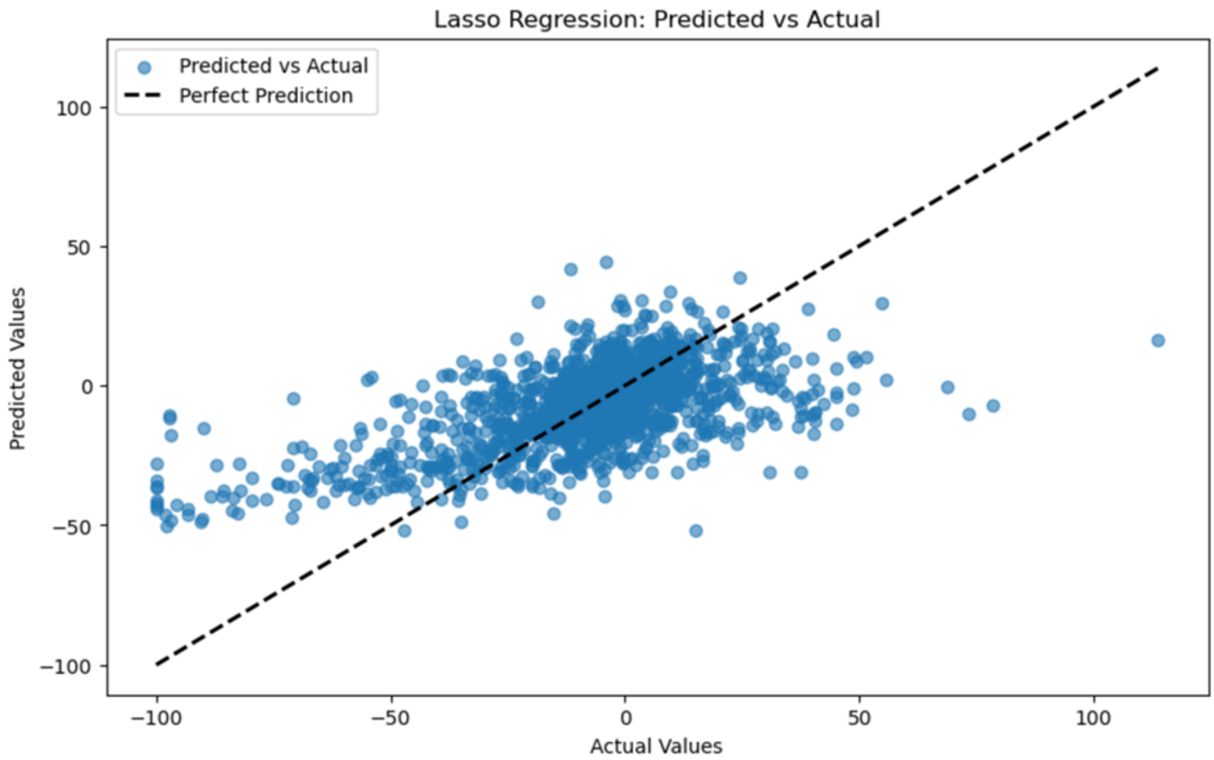
Model	In-Sample MSE	In-Sample R^2	Validation MSE	Validation R^2	Mean R^2 CV Score (5-Fold)
Linear Regression	381.178	0.386	400.603	0.344	0.288
Ridge Regression	381.754	0.385	397.635	0.349	0.299
Lasso Regression	386.176	0.378	399.291	0.346	0.298
Polynomial Regression	81.239	0.869	431.313	0.293	N/A
Symbolic Regression	487.901	0.214	483.315	0.208	N/A

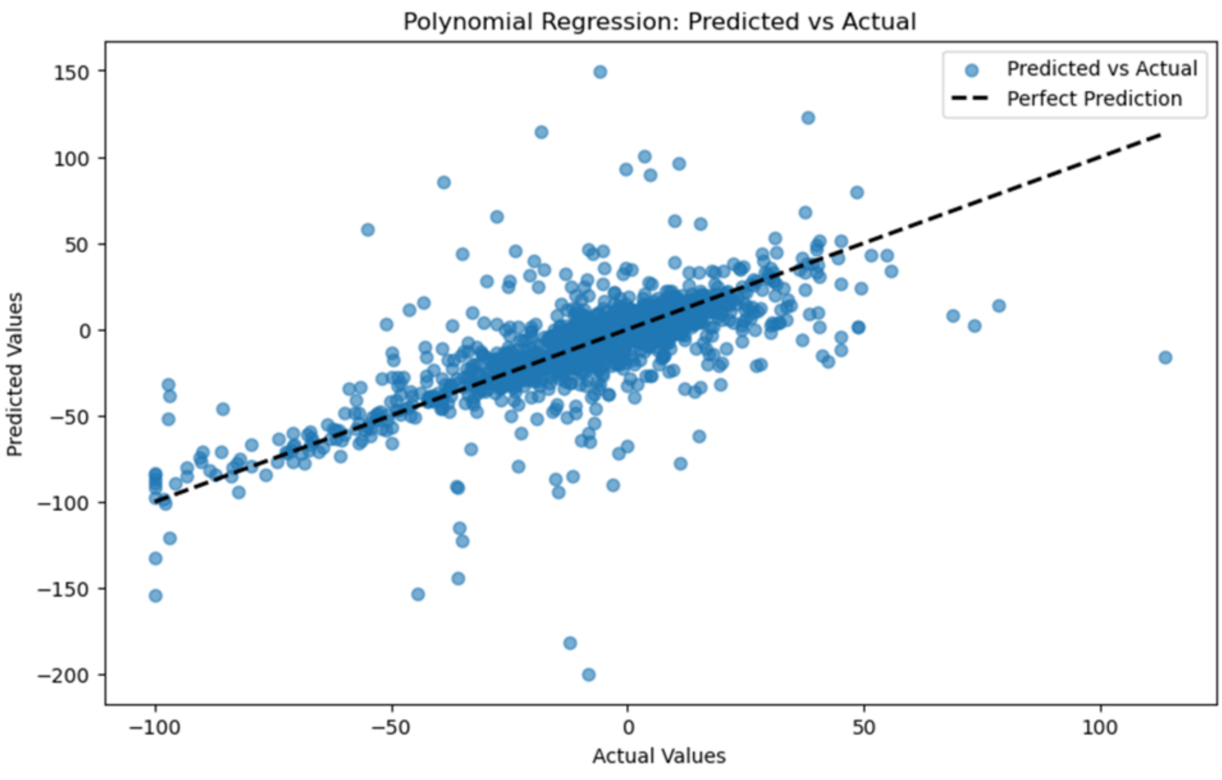


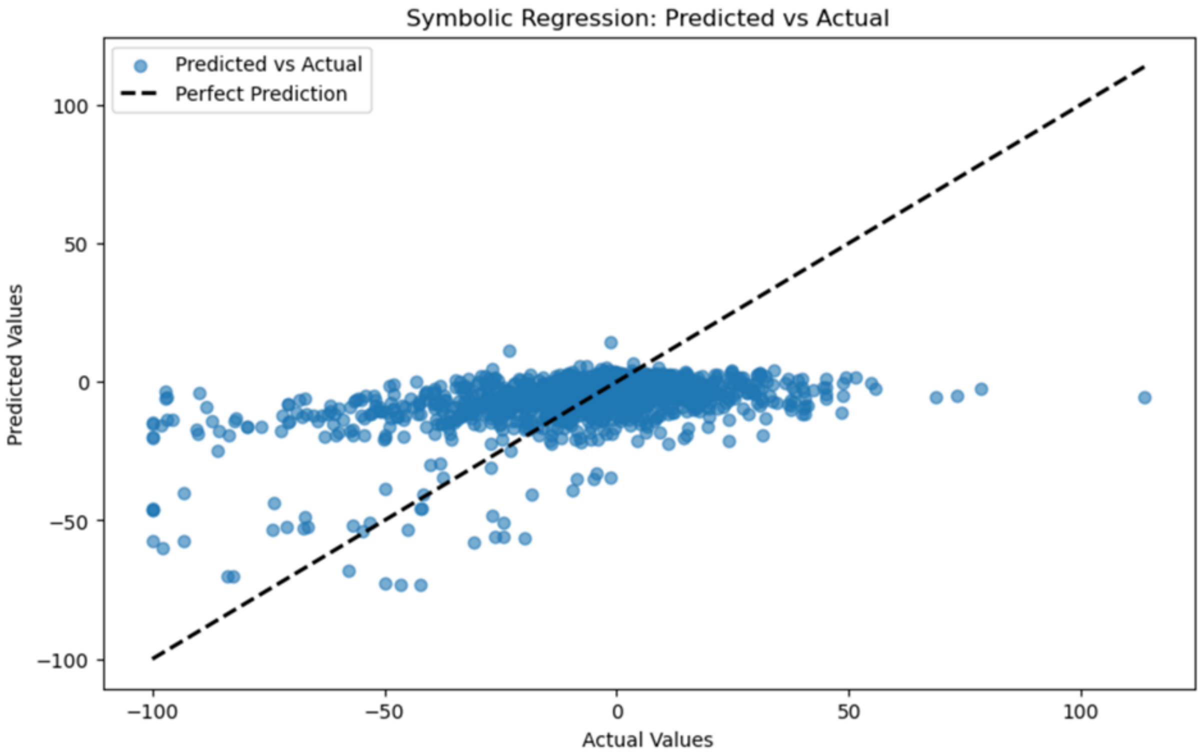
By looking at this graph, Linear, Ridge and Lasso validation R^2 values are similar. However, polynomial and symbolic regression tends to be lower than those of three regression models.


Linear Regression: Predicted vs Actual
Predicted vs Actual
Perfect Prediction
Predicted Values
Actual Values
100
50
0
−50
−100
−100
−50
0
50
100


Ridge Regression: Predicted vs Actual
Predicted vs Actual
Perfect Prediction
Predicted Values
Actual Values
100
50
0
−50
−100
−100
−50
0
50
100


Lasso Regression: Predicted vs Actual
Predicted vs Actual
Perfect Prediction
100
50
0
−50
−100
Predicted Values
−100
−50
0
50
100
Actual Values


Polynomial Regression: Predicted vs Actual
Predicted vs Actual
Perfect Prediction
150
100
50
0
−50
−100
−150
−200
Predicted Values
−100
−50
0
50
100
Actual Values



Decision Tree, Random Forest and XGBoost Regressor models -

Models with default parameters:

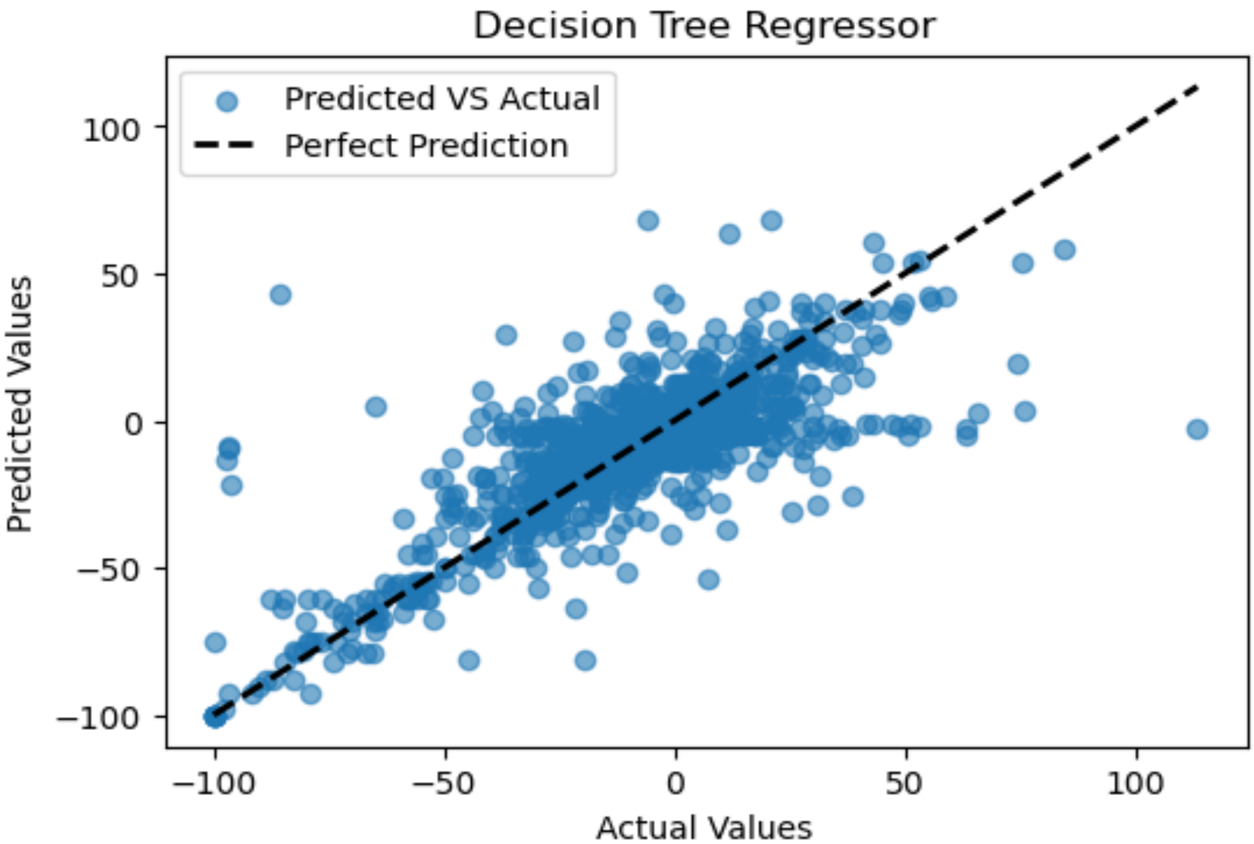
Model	In-Sample MSE	In-Sample R^2	Validation MSE	Validation R^2	Mean R^2 CV Score (5-Fold)
Decision Tree Regressor	35.4	0.942	267.637	0.568	NA
Random Forest Regressor	50.267	0.918	182.998	0.704	NA
XGBoost	35.449	0.942	208.205	0.664	NA

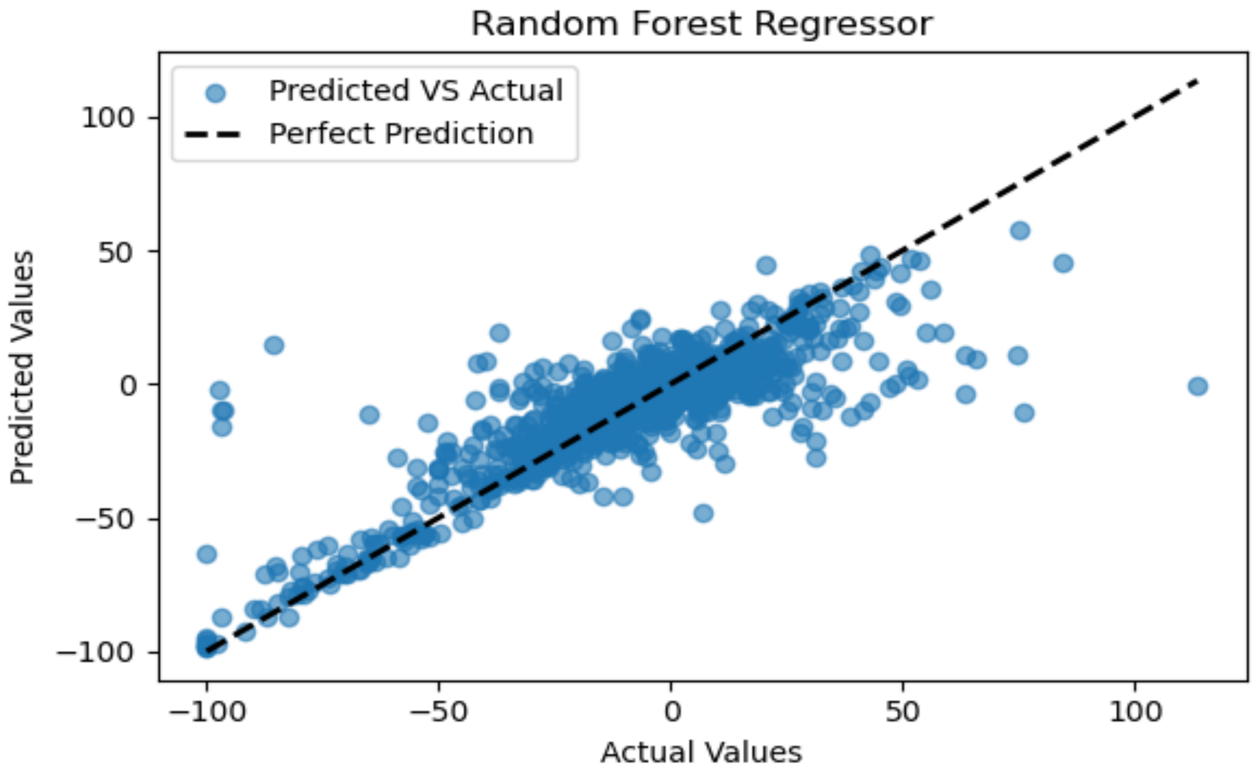
Models with hyper-paramters tuned:

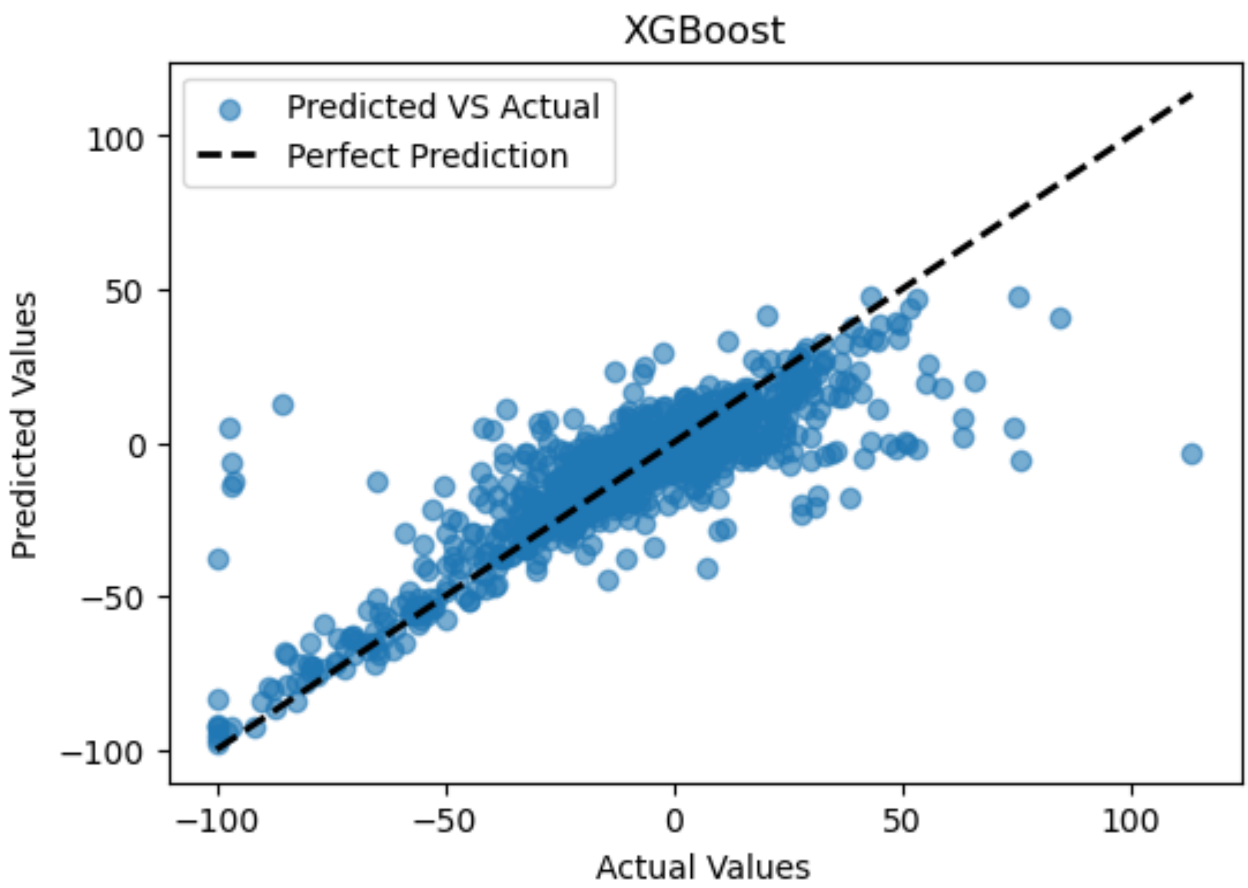
Model	In-Sample MSE	In-Sample R^2	Validation MSE	Validation R^2	Mean R^2 CV Score (5-Fold)
Decision Tree Regressor	114.217	0.814	235.625	0.620	NA
Random Forest Regressor	74.629	0.878	179.521	0.710	NA
XGBoost	114.217	0.942	178.342	0.712	NA

Models with selected features (top 40):

Model	In-Sample MSE	In-Sample R^2	Validation MSE	Validation R^2	Mean R^2 CV Score (5-Fold)
Decision Tree Regressor	114.157	0.814	242.342	0.609	NA
Random Forest Regressor	73.761	0.880	181.391	0.707	NA
XGBoost	NA	NA	NA	NA	NA




Random Forest Regressor
Predicted VS Actual
Perfect Prediction
Predicted Values
Actual Values
100
50
0
−50
−100
−100
−50
0
50
100


XGBoost
Predicted VS Actual
Perfect Prediction
Predicted Values
Actual Values
100
50
0
−50
−100
−100
−50
0
50
100

Neural Network Models -

Neural Network Architectures Evaluated:

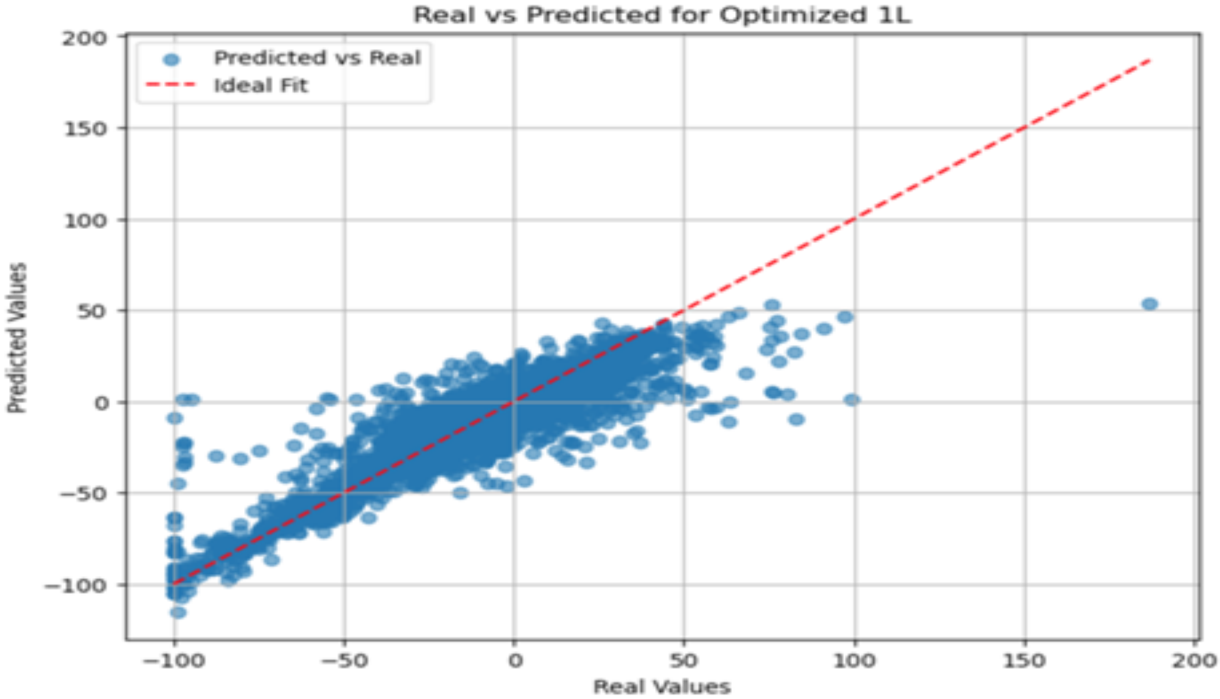
1. **1L (Single-Layer Network)**: A baseline architecture with one dense layer for linear patterns.
2. **3L (Three-Layer Network)**: A moderately deep architecture designed to capture intermediate complexity.
3. **4L (Four-Layer Network)**: A deeper architecture to model more intricate feature interactions.
4. **Progressive Network**: Features increasing neurons per layer, allowing for hierarchical learning of complex patterns.
5. **Wide & Deep Network**: Combines shallow and deep learning components to simultaneously capture linear and nonlinear patterns.
6. **Convolutional Neural Network (CNN)**: Utilizes convolutional layers adapted for structured data.

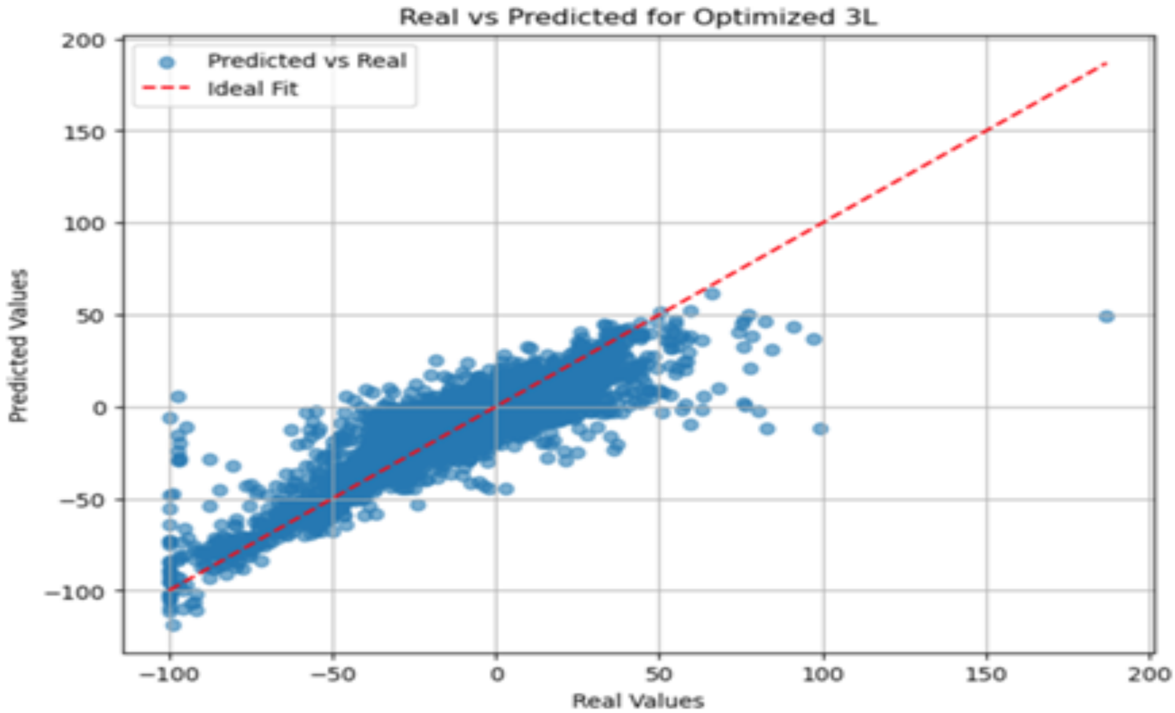
Non-Optimized Models

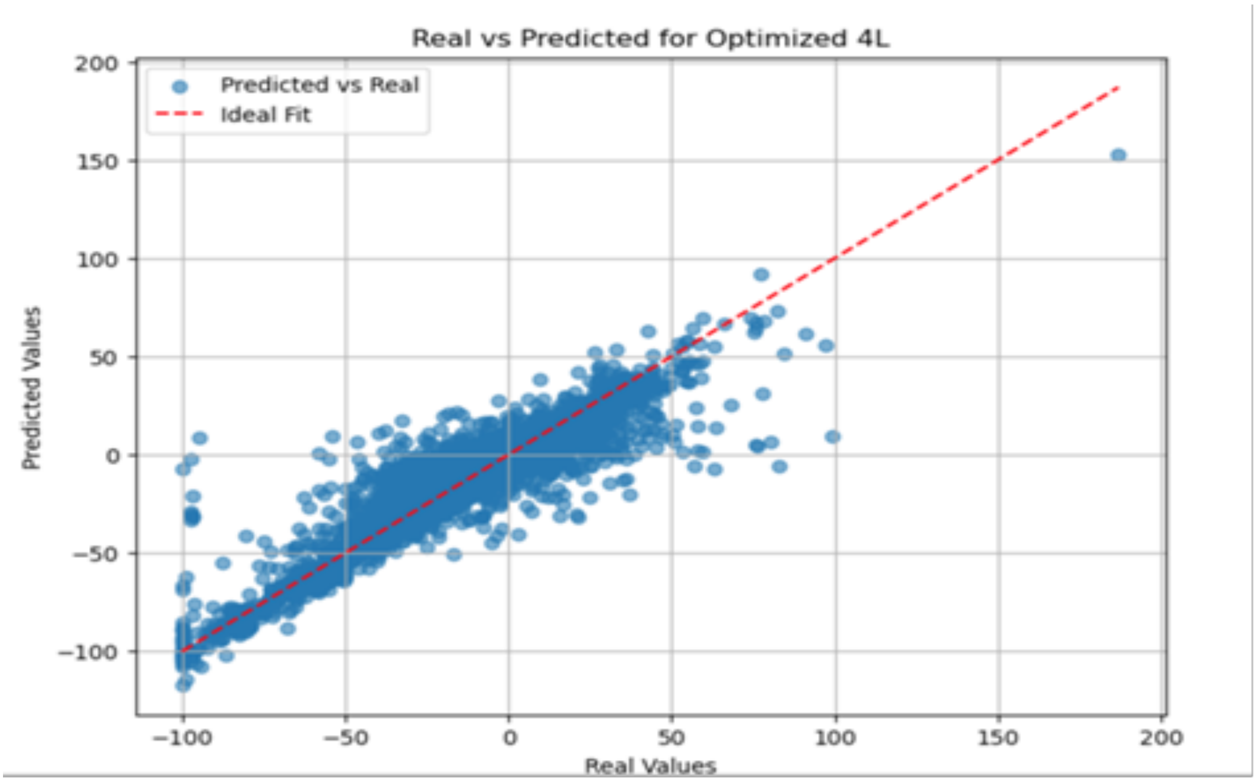
Model	Test MSE	Test R^2	Validation R^2	CV MSE	CV R^2
4L Model	165.72	0.725	0.710	185.46	0.701
3L Model	166.55	0.723	0.710	183.04	0.705
Progressive Model	172.32	0.714	0.702	180.28	0.709
Wide & Deep Network	184.91	0.693	0.690	185.27	0.701
CNN	206.17	0.657	0.651	198.42	0.680
1L Model	235.26	0.609	0.605	195.06	0.686

Optimized Models

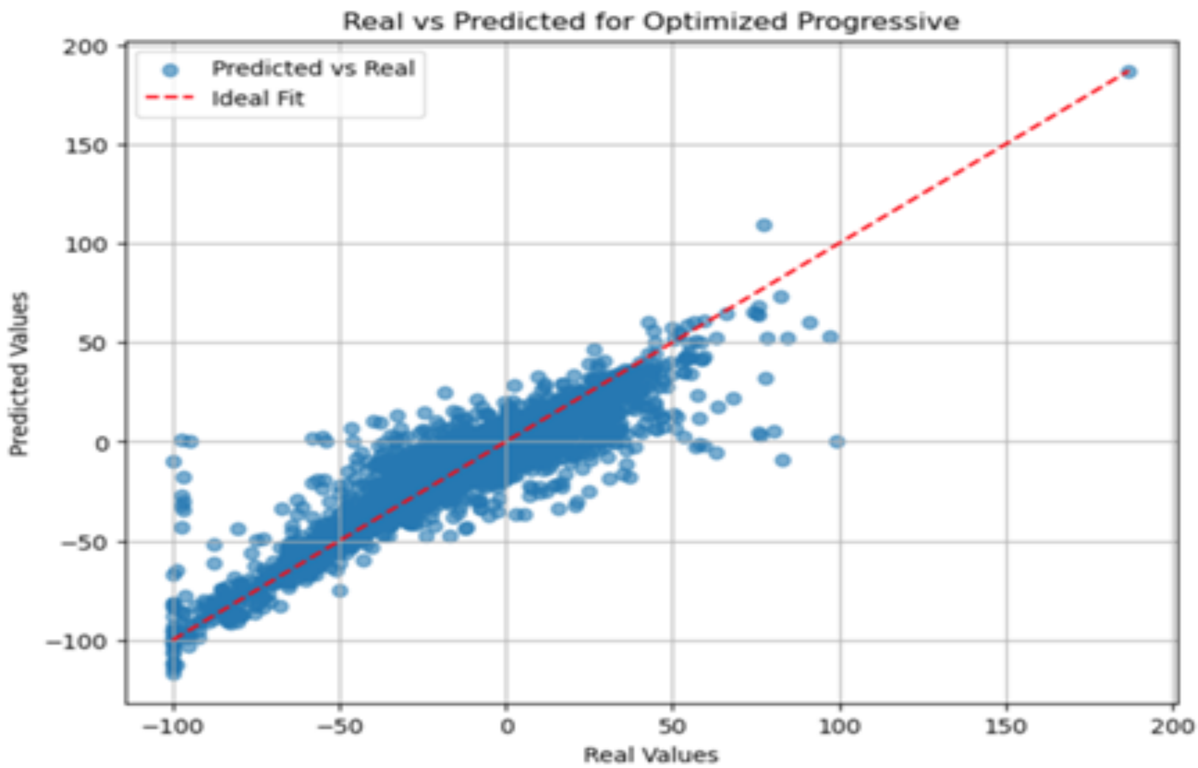
Model	Test MSE	Test R^2	Validation R^2	CV MSE	CV R^2
4L Model	161.76	0.731	0.713	182.99	0.705
3L Model	174.04	0.711	0.706	182.63	0.705
Progressive Model	165.02	0.726	0.724	181.14	0.708
Wide & Deep Network	172.71	0.713	0.711	185.75	0.701
CNN	183.86	0.695	0.687	201.17	0.675
1L Model	170.04	0.718	0.703	193.50	0.688


Real vs Predicted for Optimized 1L
Predicted vs Real
Ideal Fit
Predicted Values
Real Values
200
150
100
50
0
−50
−100
−100
−50
0
50
100
150
200

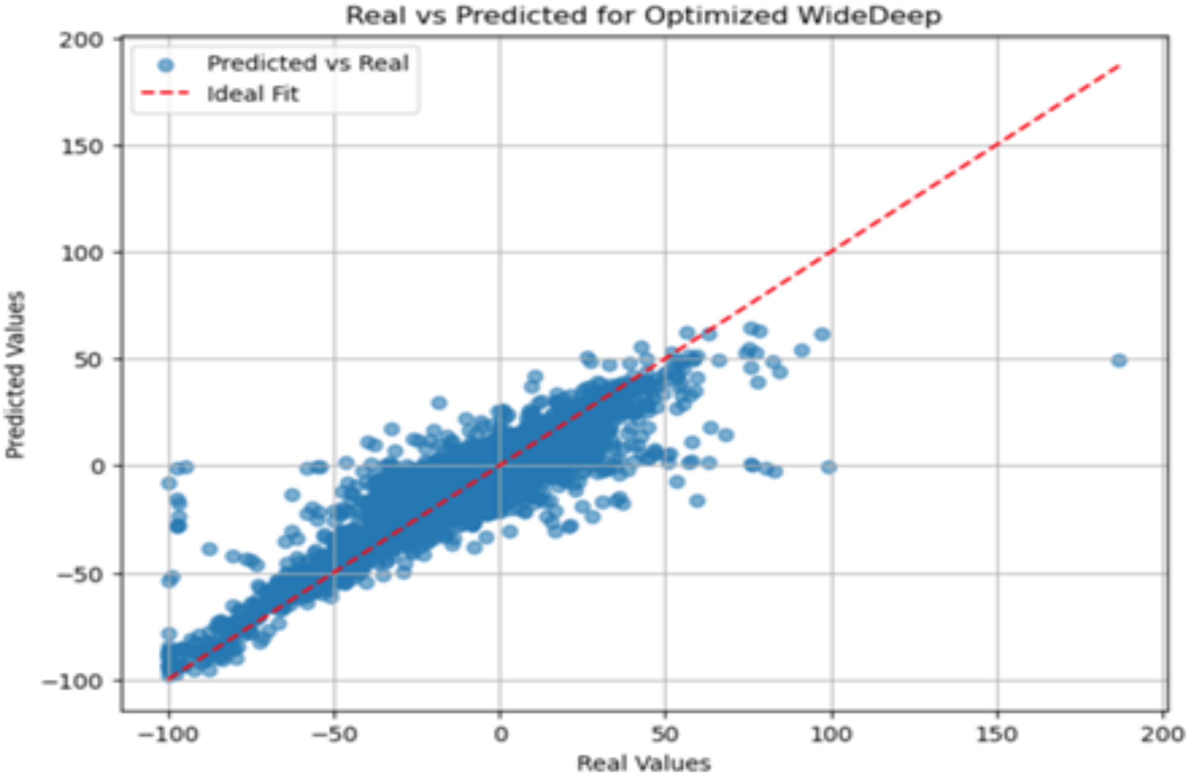

Real vs Predicted for Optimized 3L
Predicted vs Real
Ideal Fit
Predicted Values
Real Values
200
150
100
50
0
−50
−100
−100
−50
0
50
100
150
200

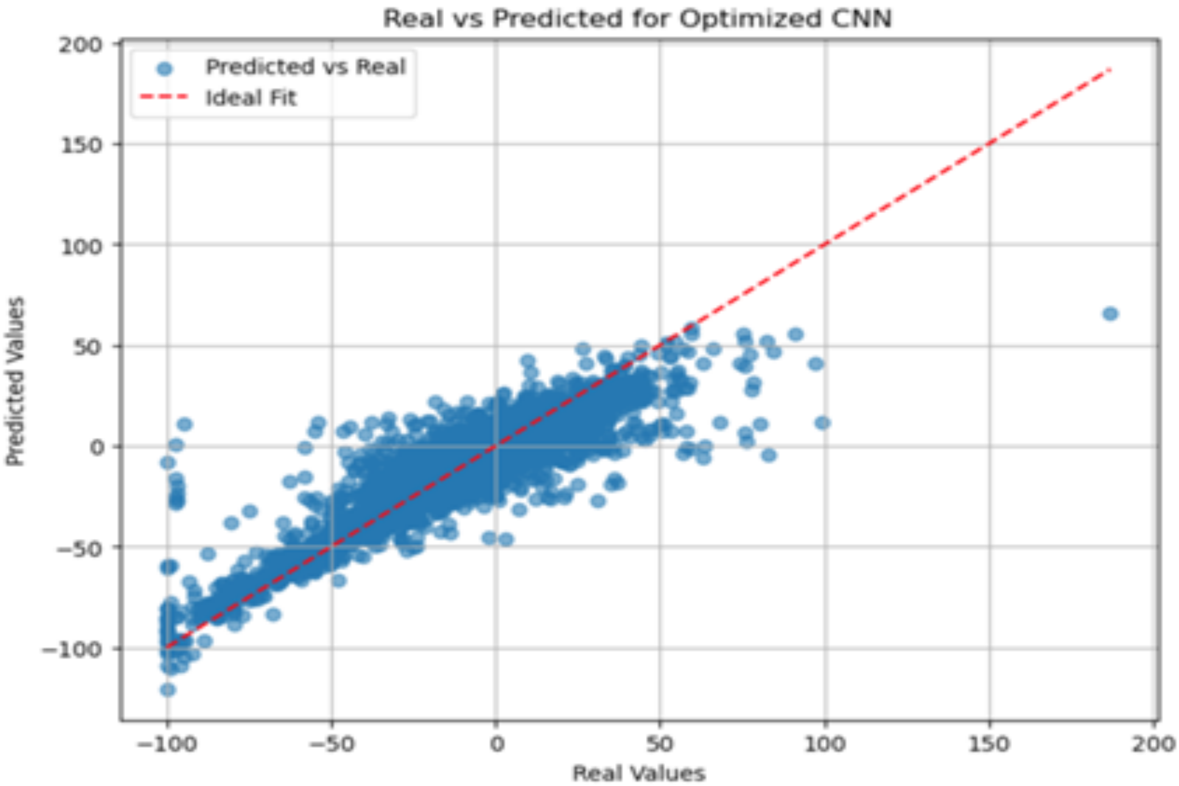


Real vs Predicted for Optimized 4L
Predicted vs Real
Ideal Fit
Predicted Values
Real Values
200
150
100
50
0
−50
−100
−100
−50
0
50
100
150
200



Real vs Predicted for Optimized Progressive
Predicted vs Real
Ideal Fit
Predicted Values
Real Values
200
150
100
50
0
−50
−100
−100
−50
0
50
100
150
200


Real vs Predicted for Optimized WideDeep
Predicted vs Real
Ideal Fit
Predicted Values
Real Values
200
150
100
50
0
−50
−100
−100
−50
0
50
100
150
200


Real vs Predicted for Optimized CNN
Predicted vs Real
Ideal Fit
Predicted Values
Real Values
200
150
100
50
0
−50
−100
−100
−50
0
50
100
150
200

Conclusions:

1. **Best Performing Model**:

- Among the neural net models, the 4L Model achieved the best overall test performance, while the Progressive Model was the most robust in cross-validation.

- The performance of Random Forest Regressor and XGBoost models was similar to the best performing neural net model, especially after hyper-parameter tuning.

2. **Impact of Optimization**:

- Hyperparameter tuning significantly improved performance on validation set, especially for the complex models like polynomial regression, random forest regressor, XGBoost, and deep neural networks.

- Moreover, hyperparameter tuning resulted in reduced over-fitting on the training data for the tree-based models like decision tree regressor, random forest regressor and XGBoost.

3. **Model Observations**:

- Simpler models like linear regression and its variants failed to capture complexity in the dataset and performed poorly on the training data as well as validation data.

- The Decision tree regressor model provided decent improvement over the linear regression models.

- The ensemble models like random forest regressor and XGBoost were better able to capture variability and provided significant improvement over both linear regression models and decision tree regressor model.

- Among the neural nets, the 1L Model struggled to capture dataset complexity. Wide & Deep Networks were consistent, but less performant compared to the Progressive Model. CNNs demonstrated limitations for structured datasets.

4. **Recommendations**:

- In our case, we saw similar performances for ensemble models and deep neural network models on the validation data. This could be an indication of an issue with the dataset itself, i.e., not having enough data to train the high-end deep neural networks. Although this may not always true, but the size of the dataset could be crucial. So, for the future work we recommend amending the dataset with more instances if possible or trying data augmentation techniques.

- Although we did perform cross-validation techniques for hyper-parameter tuning, it is still possible to further gain some improvement in model performance by further increasing the depth of cross validation search through the possible values of hyper-parameters. In our project, we were limited by the compute power available in hand. By increasing compute power, model optimization could be further improved.

- Another way to achieve higher model performance is by tapping into the feature engineering side of the preprocessing step. Try producing newer features that may be more correlated with the response compared to the original.

- Apart from these, following are in general recommendations regarding using neural nets:
    - Use the **Progressive Model** for robust and generalizable predictions.
    - Deploying the **4L Model** when optimized test performance is critical.
    - Further refine CNNs or explore hybrid architectures for future datasets with inherent spatial structures.

Future Work:

1. Explore advanced augmentation techniques to enhance data diversity.

2. Investigate hybrid models combining CNNs with dense networks.

3. Develop explainable AI methods to better understand feature importance and relationships.