Machine Learning Model Development for Crop Yield Prediction: Ridge Regression Analysis and Recommendations

Introduction

In this project, we aimed to develop a machine-learning model capable of predicting crop yields using both climatic and agricultural data. The primary goal was to predict the yield of crops in hectares, a critical task for ensuring sustainable agricultural practices and better planning for food production.

To achieve this, we used two yield target variables from different sources: "yield_fao_1000ha" (Food and Agriculture Organization) and "yield_usda_1000ha" (United States Department of Agriculture). Both datasets included various independent features, such as climatic conditions (temperature, precipitation, etc.), soil metrics, and agricultural statistics (harvested area, production).

The machine learning phase of this project focused on determining which target variable would provide the most reliable predictions and identifying the most important features influencing crop yield. Additionally, we explored several regression models to find the one that would best fit the data and generalize well to unseen data. The ultimate goal was to select a robust model that could serve as a foundation for further agricultural predictions and decision-making processes.

During this phase, we also addressed challenges like multicollinearity between the two target variables and highly correlated independent features. Rigorous model evaluation and feature selection were employed to ensure that the final model not only delivered good performance but also maintained interpretability and avoided issues like data leakage.

Data Preprocessing

1. **Data Cleaning**: During the earlier data preprocessing phase, the dataset underwent thorough cleaning to ensure that it was free of missing values, outliers, and inconsistencies. This step ensured that the data was in optimal shape for modeling.

2. **Feature Engineering**: No additional feature engineering was applied in this phase. The focus was on leveraging the existing features without creating new ones, as the provided features were already sufficient for the model's performance.

3. **Standardization:** To ensure that all features were on the same scale and to improve the performance of the Ridge Regressor model, MinMaxScaler was used to scale the dataset. This scaling technique was crucial to handle varying units and ranges in the dataset by transforming

the features to a specified range (0, 1 in this case), resulting in more stable model performance.

4. **Train-Test Split**: The dataset was split into training and testing sets, with 30% of the data used for testing and 70% used for training the model. This ensured a robust evaluation of the model's performance on unseen data.

Model Selection

In the model selection process, we focused on choosing a regression model that could accurately predict crop yield. The below table shows the performance metrics of multiple regression models.

	Model	R2 Score	MAE	MSE	RMSE
0	LinearRegression	0.963437	0.027953	0.001551	0.039386
2	RidgeRegressor	0.625583	0.094541	0.015886	0.126038
3	ElasticNet	0.566502	0.109686	0.018392	0.135618
9	XGBoost	0.479866	0.121877	0.022068	0.148553
6	RFR	0.369523	0.130235	0.026750	0.163553
8	ADABoost	0.368233	0.131261	0.026804	0.163721
7	ExtraTrees	0.293891	0.134067	0.029959	0.173086
10	CatBoost	0.250151	0.141939	0.031814	0.178366
5	SVR	0.122568	0.161754	0.037227	0.192944
4	DTR	0.067523	0.165430	0.039563	0.198904
1	LassoRegressor	-0.000356	0.176869	0.042443	0.206017
11	LightGBM	-0.000356	0.176869	0.042443	0.206017

After considering multiple models, we opted for the Ridge Regressor due to its consistent performance across multiple metrics and its ability to handle multicollinearity.

Ridge Regression was selected for the following reasons:

- **Multicollinearity Management:** Ridge Regression applies L2 regularization, which helps to manage multicollinearity by penalizing large coefficients, ensuring that no one feature dominates the model. This was crucial since the dataset contained features that were highly correlated, which could otherwise skew results.

- **Model Stability**: Ridge Regression provides stable results when compared to other models, such as Random Forest and Decision Tree, which were prone to overfitting or underfitting in some cases. The smooth penalization of coefficients ensures that the model generalizes well on unseen data.

- **Feature Selection**: The model was further optimized using forward feature selection, which helped in narrowing down the most important features that contributed significantly to yield prediction. This step reduced noise and improved the predictive power of the model.

Final Features for the Ridge Regression Model:

After feature selection, the following features were identified as the most important for the final model:

- `area_harvested_usda_1000ha`
- `production_usda_1000ha`
- `area_harvested_fao_1000ha`
- `production_fao_1000ha`
- `soil_temp_L1_C`
- `soil_temp_L2_C`
- `soil_temp_L3_C`
- `soil_temp_L4_C`
- `temp_C`
- `precipitation_era5_mm`
- `wind_northward_m_s`
- `soil_water_L2_fraction`
- `soil_water_L4_fraction`
- `precipitation_chirps_mm`

These features were selected based on their impact on predicting the target variable, `**yield_usda_1000ha**`. Ridge Regression's L2 regularization was used to handle multicollinearity among the features.

Model Evaluation

Hyperparameter tuning was conducted to improve the accuracy of the Ridge Regressor model, resulting in enhanced performance metrics as below results:

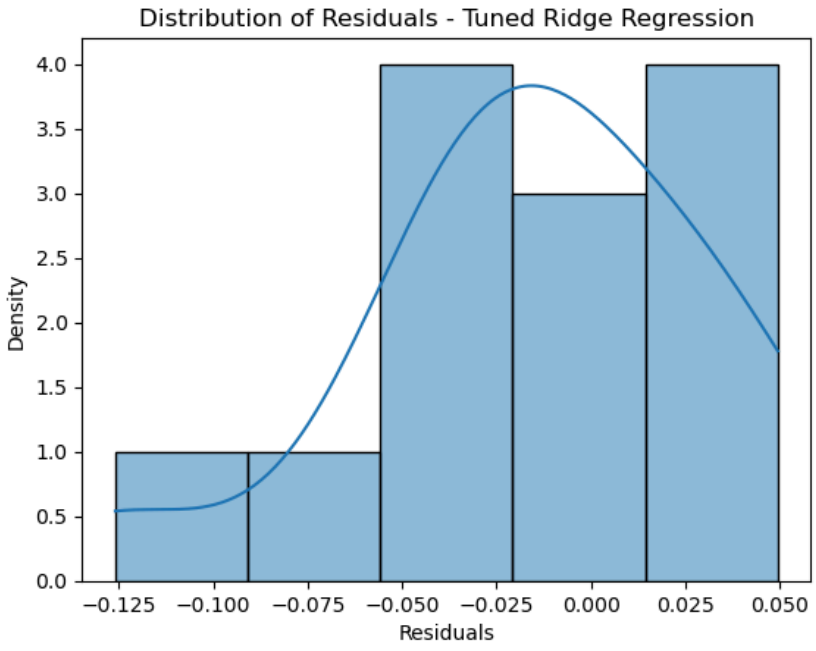
R2 Score: 0.9487
MAE: 0.0022
MSE: 0.0357
RMSE: 0.0467

The performance of the Tuned Ridge Regressor model was evaluated using multiple regression metrics on the target variable (yield_usda_1000ha). The evaluation metrics included:

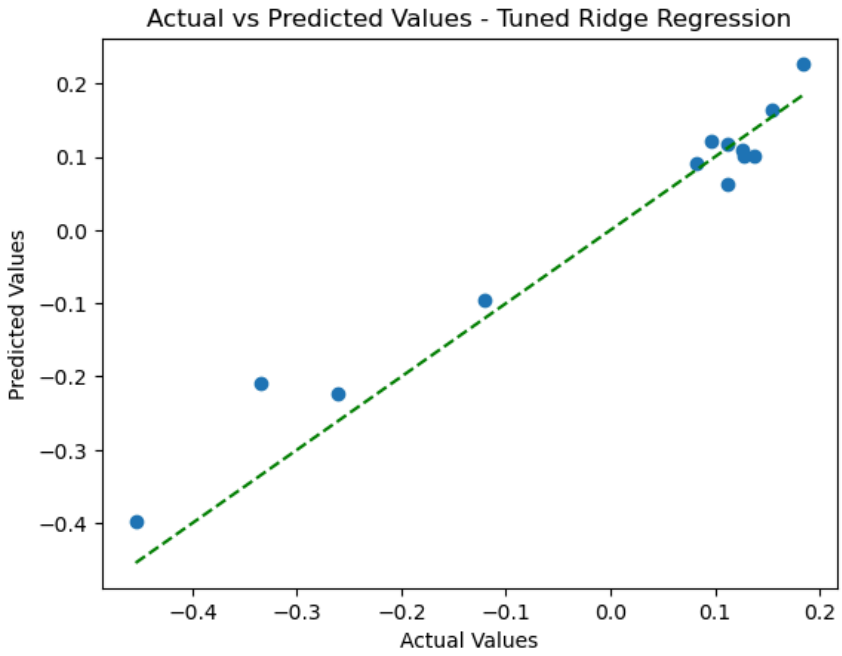
- **Root Mean Squared Error (RMSE)**: RMSE was used to measure the difference between predicted and actual values. For the target variable yield_usda_1000ha, the model achieved a low RMSE of **0.0467**, indicating high prediction accuracy.
- **R-squared (R^2):** The R^2 score quantifies how well the model explains the variance in the target variable. The Ridge Regressor for yield_usda_1000ha achieved an R^2 score of **0.9487**, demonstrating that the model explains approximately **95%** of the variability in the data, which is a strong indicator of a good fit.
- **MSE**: The Mean Squared Error is also quite low **(0.0357)**, showing that the squared differences between the predicted and actual values are small.
- **MAE**: The Mean Absolute Error is very low **(0.0022)**, indicating that the average absolute difference between predicted and actual values is small.

In addition to numerical evaluations, we used visualizations to further assess the model's performance:

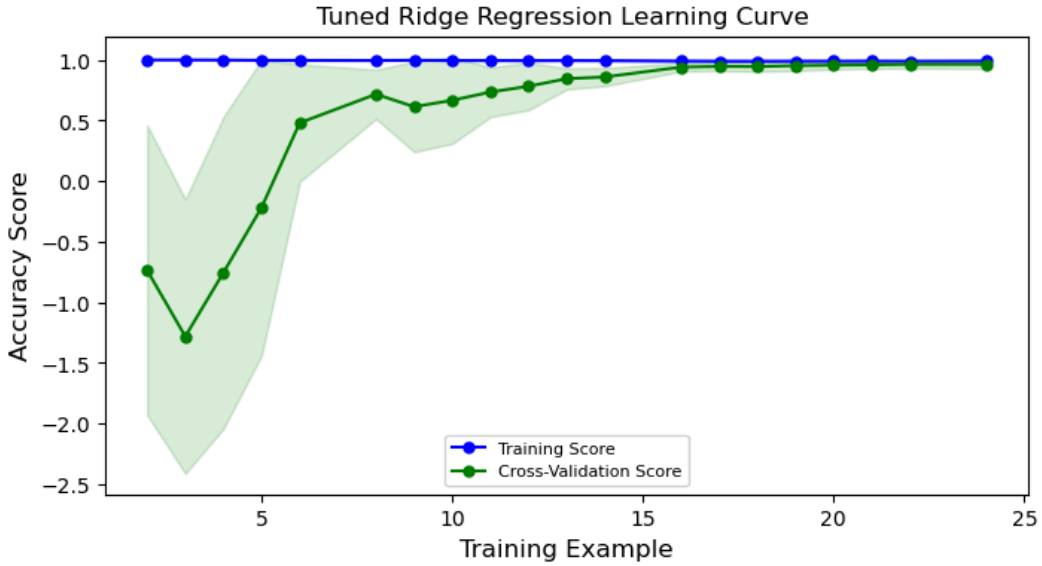
Distribution of Residuals—Tuned Ridge Regression: This plot shows how the residuals are distributed, indicating whether the model errors are normally distributed.



Actual vs. Predicted Values—Tuned Ridge Regression: This scatter plot compares actual crop yields with predicted yields, providing a visual representation of model accuracy.



Tuned Ridge Regression Learning Curve: The learning curve visualizes how the model performance evolves with increasing training data, showing whether the model benefits from more data.



Conclusion

In this modeling phase, we explored multiple machine-learning models to predict crop yields using two target variables: yield_usda_1000ha and yield_fao_1000ha. After conducting thorough testing and evaluation of various models, including Random Forest, Decision Tree, and Ridge Regressor, the Ridge Regressor emerged as the most suitable model for this task.

Key factors leading to the selection of the Ridge Regressor included its ability to handle multicollinearity effectively, its stable performance across different metrics, and the feature selection process that improved model efficiency. Specifically, for the target variable yield_usda_1000ha, the Ridge Regressor achieved an excellent RMSE of 0.0467 and an R^2 score of 0.9487, indicating high prediction accuracy and a strong model fit.

In conclusion, the Ridge Regressor model with optimized feature selection is ready for deployment, and it is expected to provide reliable predictions in the real-world application of crop yield forecasting. This model's strong performance across various metrics highlights its robustness and suitability for future use in agricultural decision-making processes.

Recommendations

1. **Inclusion of More Data**: One of the major limitations of the project was the small dataset used for model training and evaluation. To enhance the model's performance and generalizability, we recommend incorporating more data from diverse sources. Additional historical data on crop yields, climate patterns, and regional conditions will help improve prediction accuracy and reduce potential biases in the model.

2. **Domain Expertise**: While our technical expertise allowed us to build a functional model, we acknowledge that our knowledge of agriculture and crop dynamics was limited. Collaboration with domain experts in agriculture, climatology, and crop science will provide valuable insights that can further refine the feature selection process and model interpretation.

3. **Testing on More Diverse Locations:** The model was primarily trained and tested on specific data ranges. Expanding the testing scope to include different geographic locations with diverse climatic and agricultural conditions will help assess the model's robustness and adaptability.

4. **Continuous Model Updates:** Since crop yield is influenced by ever-changing environmental and market factors, it’s essential to regularly update the model with new data and retrain it to adapt to evolving conditions, ensuring its predictions remain relevant over time.