Barley Yield in Germany

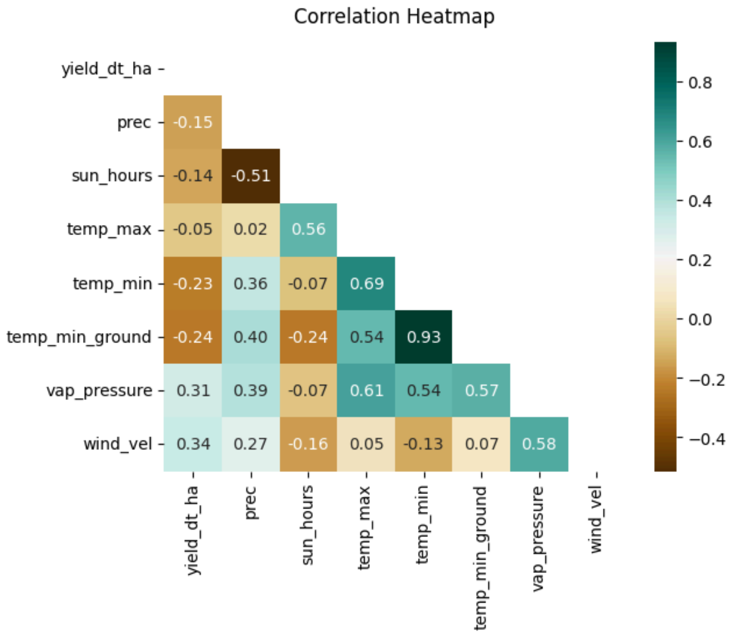
Introduction

The objective of this project is to develop a predictive regression model for barley yield in Germany using climate data spanning from 1990 to 2009. The dataset includes variables related to weather conditions such as precipitation, sun hours, temperature, vapor pressure, and wind velocity. The aim is to select three climate variables that significantly influence barley yield, build a regression model using these variables, and evaluate the model's performance.

Correlation Analysis

The correlation matrix reveals that the correlation between the predictors and the yield variable is generally weak, with the highest absolute correlation being 0.34. The most correlated variable with yield is wind velocity, followed by vapor pressure. The analysis also shows strong correlations between some predictor variables, indicating potential multicollinearity issues.

Scatterplots are utilized to detect any non-linear relationships between yield and the predictors, which may not be captured well by linear correlation. Additionally, they help in identifying outliers. In this case, scatterplots confirm the results of the correlation matrix, showing weak associations between yield and the predictors. There do not appear to be significant non-linear relationships in the data. The dataset does not exhibit clear outliers, which is understandable given that all variables are yearly aggregates for the entire nation. This aggregation tends to smooth out any extreme values that might otherwise appear in more granular or localized datasets.



First Approach: Forward Selection

First Loop: The variable with the lowest p-value is vap_pressure. Additionally, it has the lowest RMSE and is the only one with a positive R-squared.

Second Loop: We proceed by selecting wind_vel as the first variable and attempt to add a second variable. The best performing second variable is temp_min_ground. The table shows that it has the lowest p-value, and it increases the R-squared and reduces the RMSE when added to the model.

Third Loop: Once vap_pressure and temp_min_ground are part of the model, none of the remaining variables show a significant p-value. sun_hours has the lowest p-value among them and would slightly improve the model's prediction ability, as indicated by a lower RMSE and higher R-squared.

So far I would select a two-variable model with vap_pressure and temp_min_ground. I would avoid selecting a third variable.

Test for Multicollinearity

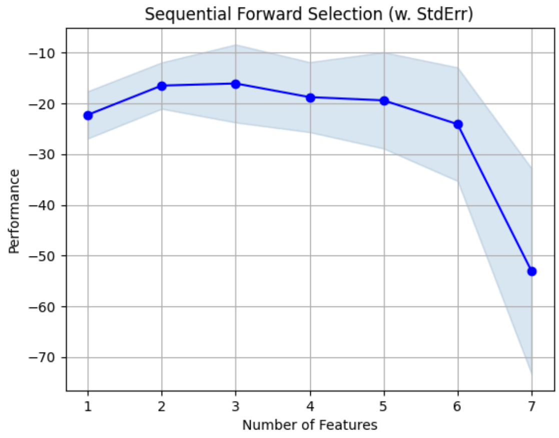
The test in the two variables model shows that there is a high multicollinearity. This is a problem, as it can cause significant issues in the regression model, such as unstable Coefficients, reduced interpretability, inflated Standard Errors. In conclusion, a model with one variable is the only one statistically sound. That variable is `vap_pressure`. Nevertheless, the prediction ability of the model is low, with an R squared=0.06 and an RMSE=4.33

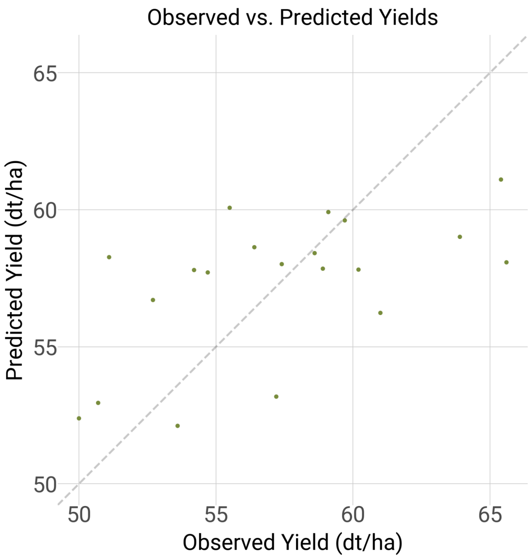
Second Approach: Focus on Prediction Ability

From a statistical perspective, we aim to determine whether a parameter is significantly different from zero to ascertain if it genuinely influences yields. From a prediction standpoint, we focus on maximizing prediction metrics, such as MSE, R² and RMSE. In this approach, we will select the set of parameters that optimize prediction performance using cross-validation techniques. While the individual parameters may not always be statistically significant, and we shouldn't make any claims about the individual effect of the parameters on the predicted variable.

The cross-validation process splits the data into folds and evaluates the model's performance on each fold, ensuring that the selected predictors generalize well to unseen data.

Results show the the model performs similarly with 2 or 3 features. The 3 features have a larger standard deviation - that is the deviation of the scores in the different cross validation samples. That suggests that it will be better to use only 2 features. The selected two features are ['temp_min', 'vap_pressure']. The Obs. vs. Predicted plot shows the fit of the final model on the full dataset. The final perfomance metrics are an R-squared of 0.32 and an RMSE of 3.7.





Conclusion

Two approaches for variable selection were tested: one focusing on statistical significance, and another focusing on accuracy in unseen data. The dataset is limited for statistical analysis due to the low number of observations and high correlation among the predictors. The statistical approach shows that we can only build a prediction model with one variable that is statistically significant. Adding a second variable introduces the issue of multicollinearity.

If the objective is to make predictions for a new, unseen year, I would use the second approach, which selected temp_min and vap_pressure, as they maximized performance on unseen data. In this approach, we should avoid making statistical claims about the effect of each factor due to multicollinearity. The goal is solely to make accurate predictions on unseen data.

Suggestions for Improving Model Performance:

- Perform PCA: Perform Principal Component Analysis (PCA) to reduce the dimensionality of the predictors and use the principal components for regression. This helps mitigate multicollinearity and retains most of the variability in the data.
- Increase Number of Observations: Increasing the number of observations and improving temporal and spatial granularity can provide a more robust dataset for analysis. Aggregating data over specific stages of the crop cycle or specific geographic areas where barley is planted can improve model accuracy.
- Crop Modeling: Crop modeling can be used to generate additional variables, aggregated for particular stages of the crop cycle, potentially increasing model performance. It can also be used to expand the dataset and simulate how yield would change under different conditions.

By addressing these areas, we can build a more accurate and robust model for predicting barley yield.