Assessing the Predictive Capacity of Machine Learning Models on Crop Yield

Author: Ryan Dorrington

Abstract

This study presents the development and evaluation of machine learning models for predicting crop yields per hectare based on Aglytica's client dataset. Preliminary results suggest the current dataset may be insufficient to achieve the desired accuracy for operational use. This paper advocates for the acquisition of additional data to enhance predictive performance.

1. Introduction

Accurate crop yield predictions are pivotal in modern agriculture, aiding in informed decision-making and resource allocation. With the rising global food demand, the ability to forecast yields is increasingly essential. Traditional empirical models in agriculture will soon be replaced by machine learning (ML) techniques, promising deeper insights by discerning patterns in vast datasets. Leveraging Aglytica's agricultural data, I aimed to test a handful of ML models for predicting crop yields (tonnes per hectare). This paper presents my efforts, results, and the emerging conclusion that a more extensive dataset is imperative for enhancing prediction accuracy.

2. Materials and Methods

Data Collection: For this endeavour, data was sourced from Aglytica's comprehensive agricultural dataset, which was collected from Aglytica's Australian farming clients. Several variables were taken into account to form a holistic understanding of the factors influencing crop yield. These included metrics related to precipitation, initial soil conditions, and cumulative effective rainfall over the season. Additionally, factors like the size of the farm and its geographical placement in specific agricultural zones were considered, which can considerably sway yield outputs.

Furthermore, it's noteworthy that the dataset was not confined to a singular crop type. Instead, it encompassed a broad spectrum of crops, adding another layer of complexity and versatility to the data.

Pre-processing: The raw data was sourced directly from an SQL database where it is stored on a per-farm, per-year basis. This section outlines the pre-processing steps taken to ensure the data was formatted optimally for analysis:

Data Correlation: Initial assessments included computing the correlation coefficient between the "yield per hectare" variable and other variables. This was an attempt to discern potentially significant variables.

Grouping and Filtering: The data was then parsed into groups based on the farm identifier and the type of crop. To ensure the integrity of data combinations, only farms with unique yearly records and at least a two-year data history was retained.

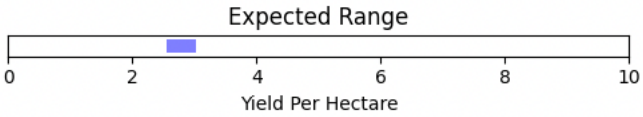
Data Transformation: In my analysis, data from each farm was grouped based on various trailing year combinations. Different combinations were trialled with each model to strike the right balance between feature vector size and data point count. For example, the MLP utilized a feature vector derived from a trailing two-year period. From each consolidated group, the yield per hectare from the most recent year was identified as the target variable. During this process, non-essential attributes were omitted to keep the focus on the most relevant data and to increase the number of data points.

One-Hot Encoding: Certain variables, specifically the name of the crop type and the agricultural zone that the farm could be found in, underwent one-hot encoding. This transformed these categorical variables into a format suitable for the classification branch of the MLP.

Dataset Splitting: Subsequent to these transformations, the data was partitioned into training, validation, and test sets, ensuring diverse data samples in each subset for a comprehensive evaluation.

Model Development: During the analysis with Aglytica's agricultural dataset, Multi-Layer Perceptrons (MLPs) were first tested, some having a combined regression and categorisation branch. However, as the research progressed, simpler linear layer models and Random Forest models showed promise in handling the dataset's complexities. An iterative approach was consistently applied to optimize hyper-parameters for each model type.

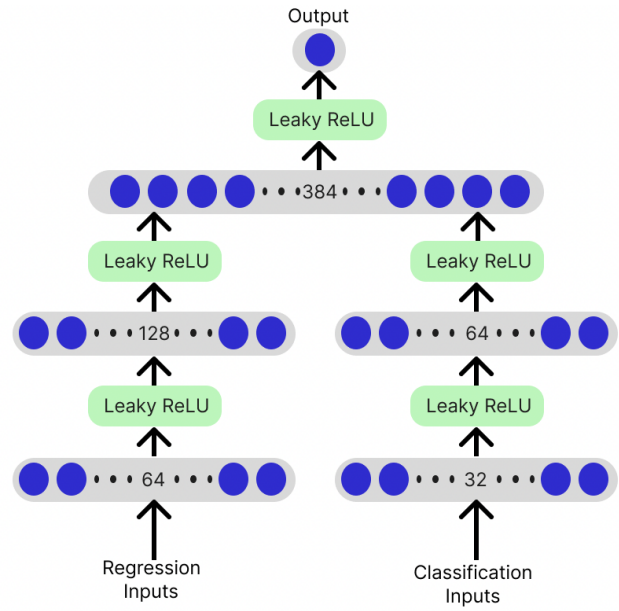
Evaluation Metrics: During the models' training phase, the Mean Squared Error (MSE) loss function was used as the primary evaluation metric. However, when assessing the final predictions, Mean Absolute Error (MAE) was employed. The decision to utilize MAE for the final evaluation stemmed from its suitability in providing a more user-friendly range of yield prediction.



3. Results

Model Performance: Three distinct models were analysed for their performance metrics. Here's a summary of their configurations and performance:

MLP with a Combined Regression and Classification Branch:



Input: Data from the preceding 2 years.
Training Data Dimensions:
Regression matrix: [1325, 14]
Classification matrix: [1325, 54]

Model Details:
Regression branch: 2 linear layers (with sizes of 64 and 128 nodes) and LeakyReLU activations.
Classification branch: 2 linear layers (with sizes of 32 and 64 nodes) and LeakyReLU activations.
Merged layer: 1 linear layer with 384 nodes and a LeakyReLU activation.
Output layer: Single node for regression output.

Hyperparameters:
Learning Rate (LR): 0.000001
Regularization (Lambda): 0.01
Epochs: 100,000

Performance:
Mean Absolute Error (MAE): 0.779

Single Linear Regression Layer:

Input: Data covering the previous 4 years.
Training Data Dimensions: [244, 84] (inclusive of one_hot datapoints)

Hyperparameters:

Learning Rate (LR): 0.00001
Regularization (Lambda): 0.001
Epochs: 38,000

Performance:

Mean Absolute Error (MAE): 0.616

Random Forest Model:

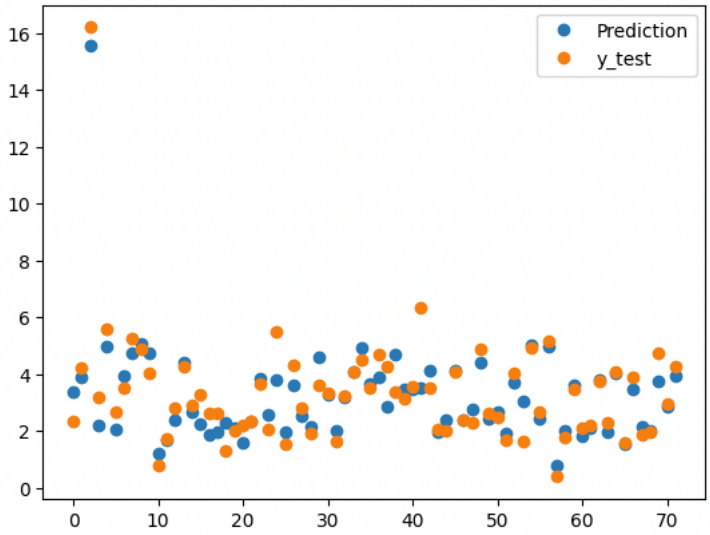
Input: Data harnessing the past 4 years.
Training Data Dimensions: [284, 84] (encompassing the one_hot datapoints)

Hyperparameters:

n_estimators: 80
max_depth: 9

Performance:

Mean Absolute Error (MAE): 0.459



For a detailed exploration or to delve deeper into the nuances of each model, please refer to the attached notebooks.

Data Adequacy Assessment: The transition from a 2-year to a 4-year trailing dataset significantly reduced the available data points, as many clients lacked extensive four-year records. Additionally, I opted to exclude certain months' rainfall data to maintain a larger dataset, sacrificing potential insights. The absence of soil type data, a key factor in many agricultural models, may have also influenced the predictions. For future endeavors, a comprehensive weather dataset and the inclusion of soil type data would be invaluable for improving accuracy and model robustness. In short, while the current data offered valuable insights, there's clear room for enrichment to maximize predictive capabilities.

4. Discussion

Analyzing the Results: The MLP might have been underserved by the limited dataset at hand. This notion is supported by the superior performance of the Linear Regression and Random Forest models, which typically excel over neural networks when data is sparse. Business Implications: Although the Random Forest model notably outperformed the other two models, its level of accuracy may still fall short of what's needed for reliable client insights. The current performance of all models suggests caution when considering their utility in real-world business scenarios.

Conclusion: My exploration into predicting crop yields using Aglytica's dataset highlighted the paramount importance of data quality and depth. Despite the notable performance of the Random Forest model, the models' current accuracy levels may not meet the requirements for practical client applications.

Recommendations:
Data Depth: Prioritise acquiring long-term datasets from clients for a richer historical view. Incorporate New Variables: Add variables like specific soil types and comprehensive weather data to boost accuracy.

External Data Sources: Partner with other agricultural data sources or expand to datasets from different regions.

Supplementary Material: For readers interested in a detailed exploration of the models and the nuances of each implementation, direct links to the Jupyter notebooks for each model are provided below:

Classification & Regression Hybrid MLP: https://github.com/ryandorrington/crop-yield-models/blob/main/Classification_Regression_Hybrid_MLP.ipynb

Single Linear Layer Model: https://github.com/ryandorrington/crop-yield-models/blob/main/Linear_Layer.ipynb

Random Forest Model: https://github.com/ryandorrington/crop-yield-models/blob/main/Random_Forest.ipynb

By accessing these notebooks, one can delve into the code, methodologies, and finer details associated with each model.