Abstract

This study explores and develops new methodologies for predicting agricultural outcomes, such as crop yields, in microclimates characterized by sparse meteorological data. Specifically, it focuses on reducing the dimensionality in time series data as a preprocessing step to generate more straightforward and more explainable forecast models. Dimensionality reduction helps manage large data sets by simplifying the information into more manageable forms without significant loss of information. We explore and utilize various 'similarity' metrics, including Kullback-Leibler Divergence, Euclidean Distance, Manhattan Distance, Cosine Similarity, Pearson Correlation, and Spearman Rank Correlation. These metrics help identify patterns and relationships across different microclimate features for locations and seasons (time and space dimensions). We analyze continuous, temporally aligned data streams from two distinct geographic locations to assess the similarity of weather features like temperature, humidity, and cloud cover. Our similarity scoring method involves comparing long-term weather patterns to identify common traits that might influence crop yields. Additionally, we examine seasonal blocks of meteorological data across different seasons within the same and separate geographic regions. Analyzing data in blocks helps in understanding how seasonal variations impact agricultural outcomes. This information gets used as input for multiple machine learning techniques, ranging from small classical models to advanced approaches like Long Short-Term Memory networks. The study employs high-dimensional temporal datasets from four geographic regions in New Mexico (Otero, Sierra, Doña Ana, and Chaves) as input data, with pecan crop yields as the outcome of interest. Our models suggest a predictive link between similar microclimates and agricultural outcomes like crop yield. Moving forward, we propose future avenues of research to refine these predictive models. We also propose to explore a new Mixture of Experts architecture, which combines insights from various specialized models to provide more precise and localized predictions across different agricultural regions. This research, still in its early stages, holds promise for improving forecasting practices in agriculture.

Keywords

Agricultural forecasting, Crop yield prediction, Microclimates, Dimensionality reduction, Time series data, Similarity metrics, Machine learning, Long Short-Term Memory (LSTM), Kullback-Leibler Divergence, Euclidean Distance, Manhattan Distance, Cosine Similarity, Pearson Correlation, Spearman Rank Correlation, Meteorological data, Seasonal variations, Pecan yields, New Mexico, Predictive modeling, Mixture of Experts architecture, High-dimensional data, Environmental data analysis, Precision agriculture, Data preprocessing, Sequential data analysis, Recurrent neural networks (RNNs), Climate variability, Agricultural productivity, Food security, Sustainable agriculture

Introduction

Agriculture contributes to 6 percent of the global economy, providing essential resources such as food, fiber, and fuel to sustain human life. The success of agricultural endeavors is intrinsically linked to many factors, among which climate plays a pivotal role. Climate variables such as temperature, humidity, and precipitation significantly influence crop growth, development, and yields. Therefore, understanding and accurately predicting the relationship between climate and agricultural outcomes is vital for optimizing farming practices, enhancing productivity, and ensuring food security.

One of the challenges in agricultural forecasting is the variability of microclimates—localized climate conditions that can differ significantly from the broader regional climate. Identifying and understanding these microclimates can provide valuable insights into crop performance under diverse environmental conditions. This study uses similarities between geographically distinct microclimates to improve predictive models for agricultural outcomes, specifically crop yields.

Dimensionality reduction in time series data is a preprocessing step to simplify complex datasets, making them more manageable and interpretable without substantial information loss. Various similarity metrics, including Kullback-Leibler Divergence, Euclidean Distance, Manhattan Distance, Cosine Similarity, Pearson Correlation, and Spearman Rank Correlation, are utilized to identify patterns and relationships across different microclimate features over both spatial and temporal dimensions.

In this study, we analyze continuous, temporally aligned meteorological data streams from four distinct geographic locations to assess the similarity of weather features such as temperature, humidity, and cloud cover. We aim to identify common traits that may influence crop yields by comparing long-term weather patterns.

We apply these insights as inputs for various machine learning models, ranging from traditional statistical methods to advanced techniques like Long Short-Term Memory (LSTM) neural networks. These models can handle complex, high-dimensional data and generate meaningful predictions. The study utilizes high-dimensional temporal datasets from four geographic regions in New Mexico (Otero, Sierra, Doña Ana, and Chaves), with pecan crop yields as the primary outcome of interest. Despite the limited dataset—22 entries spanning six years across the top four locations in New Mexico—preliminary models demonstrate a predictive relationship between similar microclimates and agricultural outcomes.

This research is still in its early stages, yet it shows promise for enhancing forecasting practices in agriculture. Future work will focus on refining these predictive models, defining new methods for deploying similarity metrics as new dimensions to data and exploring a new Mixture of Experts architecture, which integrates insights from specialized models to deliver more accurate and localized predictions for different agricultural regions. By advancing our understanding of microclimate similarities and their impact on crop yields, this study aims to contribute to the development of more reliable and precise agricultural forecasting methods.

Methods

Dataset Preparation

Data Source and Description

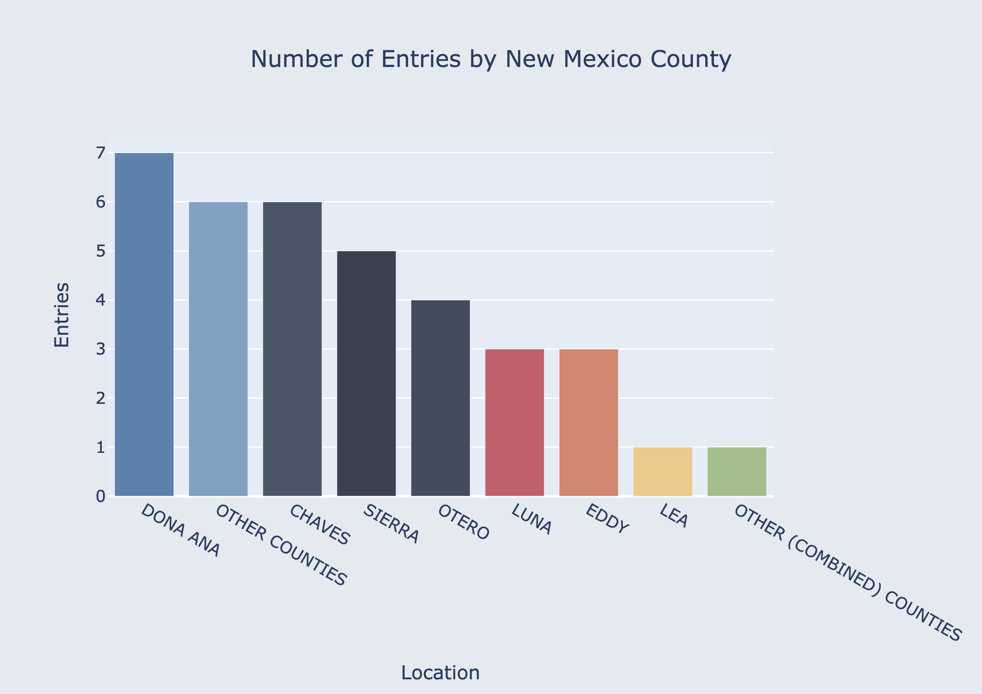
This study's primary datasets are from two distinct sources, each offering unique insights into different aspects of agricultural forecasting. The crop yield data, specifically for pecans, were sourced from the United States Department of Agriculture's National Agricultural Statistics Service (NASS) via their Quick Stats database (https://quickstats.nass.usda.gov/). This repository provides detailed, county-level agricultural yield data across various crops, recorded annually. The sparse amounts of available data motivated the choice of pecan crop yield data, making it a prime candidate for finding new methodologies that can open access to farmers with less available data for their crops.

We received the meteorological data from the Open Weather API (https://openweathermap.org/api), a service that offers extensive meteorological observations updated hourly. This dataset includes various weather variables crucial for agricultural analysis, such as latitude, longitude, dew point, perceived temperature, minimum and maximum temperatures, humidity, wind speed, wind direction, cloud cover, and a general weather identification code. These features were selected for their direct impact on crop growth and yield outcomes, providing a comprehensive environmental snapshot essential for modeling microclimatic effects on agricultural productivity.

The study focused on four New Mexico regions: Dona Ana, Chaves, Sierra, and Otero. These regions were selected based on the availability of longer-spanning yield data. While still sparse, these locations had the most yield data available for pecans. The choice of these locations also allows for a diverse examination of microclimatic impacts across different geographic profiles, enhancing the generalizability of the forecasting models developed.

The graph below illustrates the distribution of data entries available for each county, highlighting the extent of data coverage and the basis for comparative analysis across the regions. Dona Ana leads with seven entries, Chaves with six, Sierra with five, and Otero with four, culminating in 22 data points. This distribution supports a comprehensive analysis and underscores the variability and richness of the dataset used in this study.

This detailed approach to data selection ensures that the models developed are grounded in reliable datasets, facilitating accurate predictions and insights into the intricate relationship between weather conditions and agricultural outputs. Test\_font\_size



Data Preparation

Temporal Alignment and Feature Selection

Precise temporal alignment of the four meteorological data streams was critical due to the sensitivity of similarity scores to timing discrepancies. Even a single day’s misalignment could significantly distort the perceived similarities between the locations. Additionally, some weather features exhibited substantial data gaps, with less than 60% data availability. Features with insufficient data, including visibility, wind gust, and various precipitation measurements (rain one hour, rain three hours, snow one hour, snow three hours), were excluded from the analysis to maintain robustness and reliability.

Yield Data Preparation

The yield data required minimal cleaning, as the county and the yield measured in pounds per acre were the primary variables of interest. This straightforward dataset allowed us to focus on analyzing the meteorological influences without the complication of extraneous data points.

Dataset Construction for Model Building

To explore the optimal combination of features for predictive accuracy, we constructed several datasets to fit the needs of the different predictive models:

1. Dataset 1 (D1): This dataset paired the similarity scores from the various similarity calculations with the corresponding pecan yields between the two counties involved in each score. This approach aimed to link microclimate similarities to yield outcomes directly.

2. Dataset 2 (D2): Here, we matched the yield for each year with meteorological data instances spanning that same year for each geographical location. This dataset aimed to assess year-long climatic impacts on yields.

3. Dataset 3 (D3): We aggregated the meteorological data by computing the average of each feature across the 24-hour instances within each day, significantly reducing the dataset size from 192,840 instances to 2,557. The target variable was the annual yield for each location, reflecting pecans' nearly year-long growth and harvest cycle. Normalization was applied to ensure uniformity in data scale, using the Min-Max scaling method. This normalization ensures that all feature values are proportionally scaled between 0 and 1, facilitating more effective learning by the models.

4. Dataset 4 (D4): The final dataset flattened the entire year's meteorological

data into a single array for each location and then paired it with the corresponding annual yield. Dataset 4 resulted in a final shape of (22, 122,976) for the inputs and (22,) for the outputs. Min-max scaling was similarly applied to normalize the data, enhancing model performance by equalizing the feature scales.

Implications for Model Efficacy

We designed each dataset configuration to test different hypotheses about the relationships between climate data and crop yields, from direct comparisons of similar microclimates (D1) to more comprehensive, integrative approaches (D3, D4). The diverse dataset construction and preparation approaches ensure a thorough exploration of data dimensions and relationships, which is crucial for developing robust predictive models in agricultural forecasting.

Similarity Scoring

Seven different similarity scoring metrics were employed to assess the relationships between meteorological data across diverse geographic locations in New Mexico. These metrics fall into three categories: vector-based similarities, correlation-based similarities, and information-theoretic similarity. Each category and metric was chosen based on its potential to uncover meaningful patterns within the data that are relevant to predicting agricultural outcomes like crop yields.

1. Vector-Based Similarities:

Cosine Similarity: This metric measures the cosine of the angle between two vectors, emphasizing their orientation rather than magnitude. It is instrumental in environments where the scale of data varies significantly, but the direction of data points is more critical for similarity. In the context of weather data, where absolute values might differ but trends over time align, cosine similarity can effectively show these parallel patterns.

Euclidean Distance: This metric computes the straight-line distance between two points in Euclidean space. It is sensitive to the magnitude of changes, making it suitable for detecting fine-grained differences in weather parameters like temperature or humidity levels between locations. However, it can be overly sensitive to outliers or anomalies in the data.

Manhattan Distance: This sum of the absolute differences between points in a vector can be advantageous in high-dimensional data due to its emphasis on variability across multiple dimensions rather than the diagonal path. Significant differences have less influence in a single dimension, making it useful for datasets where each dimension contributes equally to the overall similarity, such as varied meteorological factors.

2. Correlation-Based Similarities:

Pearson Correlation Coefficient: Measures the linear relationship between datasets. It is highly effective in cases where the data distribution is expected to be normal; however, it may be less reliable with meteorological data that often exhibits non-normal distributions or is affected by extreme weather events.

Spearman Rank Correlation: This non-parametric measure assesses how well the relationship between two variables can be described using a monotonic function. It is especially suitable for meteorological data as it does not assume a normal distribution and is less sensitive to outliers, thus providing a more robust analysis of ordinal or ranked data.

Kendall Tau Rank Correlation: Like Spearman, Kendall Tau is effective for data with ties or ordinal nature. It is advantageous when the data set is small or has many tied ranks, which can be expected in meteorological data grouped by categories like wind speed ranges.

3. Information Theoretic Similarity:

Kullback-Leibler Divergence: This measure quantifies the divergence from one probability distribution to another. It is particularly insightful for measuring the 'distance' between different distributions of meteorological phenomena. However, its asymmetric nature and susceptibility to producing infinite values when probabilities are zero make it less practical for some comparative analyses in this context.

Applicability and Observations

Our study utilized these metrics across 546 observations, calculated by evaluating each of the seven metrics across 13 features for 6 location pairs. Our comprehensive exploration aimed to identify the most predictive features and similarity measures, enhancing the model's ability to forecast crop yields while reducing data dimensionality.

Correlation-based metrics (Pearson, Spearman, and Kendall Tau) generally proved the most effective, capturing the linear and ordinal relationships necessary for predicting agricultural outcomes. These metrics help to reveal how tightly linked meteorological variables are across different regions, which aids in understanding how similar weather patterns could predict similar crop yields.

Conversely, while applicable in specific contexts, vector-based metrics showed variable effectiveness. Euclidean and Manhattan distances were particularly good at highlighting absolute differences in climate data, which could be critical when precise measurements affect crop outcomes. However, their sensitivity to outliers and the measurement scale sometimes could have improved their utility in our analysis.

Kullback-Leibler Divergence, an information-theoretic measure, offered unique insights into the distributional differences between data sets but needed to be more practical due to its tendency to produce infinite values when encountering zeros in the data, which could occur with certain weather variables like

Visual Representation and Further Analysis

To visually represent the effectiveness and distinctions between these metrics, we included a heatmap that illustrates the degree of similarity each method found when analyzing each geographic location against each other. This visualization helps quickly identify metrics most informative for specific features and regions, providing a clear path for model refinement and hypothesis generation.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph of different types of type of type of type of type of type of type of type of type of type of type of type of type of type of type of type of type of type

Description automatically generated

By integrating these similarity scores into our predictive models, we aim to enhance our understanding of how microclimate characteristics influence pecan yields. This approach not only aids in agricultural planning and management but also contributes to more robust and resilient food production systems in the face of climate variability.

Data Pipeline

In developing our predictive models, we implemented a strategic approach to data partitioning to optimize model training, validation, and testing. The datasets were systematically divided to ensure robust training and accurate evaluation while preventing overfitting.

Data Splitting Strategy

Standard Split for Model Development: For most models, the data was partitioned into three segments: 80% for training, 16% for validation, and 4% for testing. This distribution was chosen based on the following considerations:

Training Data (80%): This substantial portion allows the models to learn as comprehensively as possible from various examples, covering a wider range of scenarios and variability within the data. A more extensive training set is essential for complex models, especially in cases involving nuanced meteorological data where the patterns might take time to appear.

- Validation Data (16%): Validation is an intermediate check during model training. It is used to fine-tune model parameters, adjust learning rates, and make decisions about model iterations. This relatively large validation set ensures that the model generalizes well over different data sets and helps mitigate the overfitting problem.

- Testing Data (4%): The testing set is smaller, as it is used solely to evaluate the model's performance after the model parameters have been finalized. This phase is critical as it objectively evaluates a model's predictive power on unseen data, reflecting its potential real-world performance.

Adjusted Split for Specific Models: Some models, notably simpler, did not require a separate validation set. For these models, we split the data into 80% for training and 20% for testing. The rationale for this adjusted split includes Increased Testing Data (20%): Providing more data for testing allows for a more thorough evaluation of the model's performance across a broader set of unseen examples. This is particularly important for simpler models to ensure their simpler hypotheses remain under varied conditions. The validation step can be omitted in models where hyperparameter tuning is minimal or not required. The omission of the validation step typically applies to models that are less complex or where the risk of overfitting is lower due to the algorithm's nature or the dataset's scale.

Implications of Data Splitting

The chosen data partitioning strategies were pivotal in balancing the need for thorough training and the necessity of unbiased model evaluation. This approach maximizes learning from the available data and maintains the integrity and reliability of the model evaluation process. By strategically allocating data to training, validation, and testing phases, we can ensure that our models are well-trained, generalizable, and robust against overfitting, which is vital for deploying these models in real-world agricultural forecasting.

Models

Baseline Regressor

As an initial step in our model evaluation process, we employed a Baseline Regression Model. This model establishes a baseline for comparison with more complex predictive models. By design, the Baseline Regression predicts the mean of the training target values, providing a simple benchmark against which the performance of other models can be assessed.

Purpose and Utility

The primary purpose of the Baseline Regressor is to set a baseline level of performance that any other sophisticated model must surpass to prove its utility. In predictive modeling, particularly in complex fields like agricultural forecasting, it is necessary to distinguish improvements driven by intelligent data handling and algorithmic sophistication from what can be achieved by simple average-based predictions.

Methodology

Normalization: Before training, the dataset underwent normalization, where the mean value from each feature in the training set was subtracted from the corresponding feature values. This process centers the distribution around a zero mean, a standard practice to ensure that the model is not biased by the natural variance in the features. Normalization helps compare features on a similar scale and improves the stability and performance of the learning algorithms.

Prediction Strategy: The Baseline Regressor's strategy is straightforward—it calculates and uses the mean of the target values from the training dataset as the predicted value for all instances. This approach, while simplistic, is instrumental in establishing the minimal expected performance level for any predictive model.

Performance Metrics

Training Time: The Baseline Regressor's simplicity is reflected in its exceptionally short training time, recorded at merely 0.0005 seconds. This efficiency shows the model's role as a fundamental benchmark rather than a sophisticated predictive tool.

Root Mean Squared Error (RMSE): The performance of the Baseline Regressor was quantitatively assessed using the RMSE, which measures the average magnitude of the prediction errors. The RMSE for the Baseline Regressor was 484.535, indicating the average error between the predicted and actual values. This metric provides a reference point for evaluating the effectiveness of more advanced models. A significant reduction in RMSE in subsequent models would demonstrate the value of incorporating complex algorithms and data preprocessing techniques.

Implications

Using a Baseline Regression is a standard methodological approach in model development. It offers a clear benchmark for the predictive accuracy that more sophisticated models must exceed. By comparing the performance of advanced models against this simple baseline, we can more effectively measure the incremental benefits of employing complex machine learning techniques and various data preprocessing methods.

XG Boost (Extreme Gradient Boosting)

Overview and Implementation

Extreme Gradient Boosting (XG Boost) is a machine learning algorithm with a gradient boosting framework. Our study employed XG Boost for regression tasks to predict agricultural yields based on meteorological data. The model operates by sequentially building decision trees, with each tree attempting to correct the errors made by its predecessors. The final prediction is the weighted sum of the predictions from all trees in the ensemble, which enhances the model's ability to generalize across complex datasets.

Model Configuration and Training

For the control model, we utilized Dataset 4 (D4), which includes the entire year's meteorological data flattened into an array and matched with corresponding yield data. We optimized the model using the following parameters:

• Max Depth: 1024, allowing the model to capture intricate data patterns at a substantial computational cost.

• Learning Rate: 0.001, to ensure gradual and steady improvements, reducing the risk of overfitting.

• Subsample: 0.8, using 80% of the data for building each tree, balancing bias and variance.

• Colsample by Tree: 0.8, using 80% of features for each tree, which helps prevent overfitting and ensures diverse trees in the model.

• Number of Estimators: 10,000, a high number to provide comprehensive learning opportunities.

• Regularization Terms (Alpha and Lambda): Both are set to 2 to reduce overfitting by penalizing significant coefficients.

The training of this control setup required 477.335 seconds, resulting in a Root Mean Squared Error (RMSE) of 646.427. Surprisingly, this RMSE was worse than the simple baseline regressor, suggesting that the model might need to be more complex or better suited to the relatively small dataset size (only 22 entries), leading to overfitting despite the regularization and subsampling strategies

Model Optimization Using Similarity Metrics

To explore the potential of using similarity metrics to enhance model performance, we employed Dataset 1. To optimize XG Boost's settings for this dataset, a grid search over 32,400 parameter combinations was conducted across 5-fold cross-validation, totaling 162,000 fits. This exhaustive search identified the most effective parameters:

• Alpha: 2

• Colsample by Tree: 1

• Learning Rate: 0.01

• Lambda: 2

• Max Depth: 2

• Number of Estimators: 100

• Subsample: 1

The optimized model trained with these parameters achieved an RMSE of 477.903. The training process was remarkably faster, taking only 0.073 seconds, 6538 times quicker than the control model. This significant reduction in training time, combined with a notable improvement in RMSE compared to the control (26.07% better than the control model and 1.368% improvement over the baseline regressor), suggests that leveraging similarity metrics can be highly beneficial, especially in contexts where data is sparse but pattern rich.

MLP (Multi-Layer Perceptron)

Overview and Implementation

The Multi-Layer Perceptron (MLP) is a fully connected neural network, a foundational architecture in modern machine learning. Each layer in an MLP consists of neurons fully connected to all activations in the previous layer, and its outputs are calculated by applying a nonlinear activation function to the weighted sums of its inputs. This design allows MLPs to learn complex, nonlinear relationships in the data, making them suitable for various predictive modeling tasks.

However, MLPs can struggle with sparse datasets, where the number of data points is limited relative to the complexity of the model. Sparse data-trained MLPs often lead to overfitting, where the model learns the noise in the training data instead of generalizing from it.

Model Training and Architecture

Control Model: The control MLP we trained using Dataset 4 over 1,000 epochs, allowing sufficient time for the network to adjust its weights and biases to minimize the prediction error. The training was executed on a Nvidia 4090 GPU, resulting in a training time of 45.84 seconds. The architecture of the MLP, as outlined in the referenced figure, was designed to balance complexity with performance to prevent overfitting while maintaining the capacity to capture significant patterns in the data. The control model's final Root Mean Squared Error (RMSE) was 382.851.

Similarity Score Model: An alternative MLP configuration we trained using Dataset 2. This dataset was presumed to provide a richer feature set by emphasizing the relationships and patterns across different locations, potentially enhancing predictive accuracy. The model followed the same architecture as the control but was subjected to a more extended training duration of 759.42 seconds due to the nuanced nature of the data it processed. The extended training period was necessary to adequately learn from the enriched yet complex input data provided by the similarity scores. This model achieved an RMSE of 366.67, indicating a 24.32% improvement in accuracy over the baseline regressor and a 4.23% improvement over the control MLP.

Performance Analysis and Implications

The improvement in RMSE by the similarity score model over both the baseline regressor and the control MLP shows the potential benefits of incorporating processed similarity metrics into the training data. This approach enhances the model's ability to discern and generalize from underlying patterns in complex meteorological datasets, suggesting a promising direction for further research. The results advocate for continued exploration into optimized architectures and training strategies that leverage similarity scores to boost predictive performance in sparse data scenarios. By refining the MLP's architecture and training on datasets enriched with similarity metrics, we aim to harness further the network's capacity to model complex interactions effectively and deliver more accurate predictions for agricultural outcomes.

LSTM (Long Short-Term Memory Networks)

Overview and Implementation

Long-short-term memory Networks (LSTMs) are specialized Recurrent Neural Networks (RNNs) designed to address the vanishing gradient problem commonly encountered in traditional RNNs, especially when processing sequences with long-term dependencies. By incorporating memory cells that can maintain information in memory for long periods, LSTMs are ideally suited for analyzing time-series data or any sequential data where the timing and order of events are critical.

Model Configurations and Training

Control Model Configuration:

Dataset: We trained the control LSTM using Dataset 4, which comprises comprehensive year-long meteorological data transformed into sequences to reflect temporal dynamics.

Architecture: The network architecture included 20 layers with a high hidden state dimension of 1280, designed to capture complex patterns and dependencies in the data. This extensive network aimed to learn the detailed features presented in the sequential data.

Training Details: The model training was executed on the same Nvidia 4090 GPU, running for 5325 epochs, which took 525 seconds. Such extensive training was necessary to refine the model's weights across numerous layers and data points.

Performance: The final model achieved an RMSE of 421.519, which serves as a baseline for evaluating the effectiveness of LSTM in handling complex time-series data.

Similarity Scored Model Configuration:

Dataset: The Similarity Scored LSTM was trained using Dataset 3.

Optimization Approach: For the similarity-scored LSTM, we implemented a custom grid search to determine the optimal set of hyperparameters, focusing on a model configuration that balances performance with computational efficiency.

Model Training: This model was configured with a more streamlined architecture, consisting of only five layers and a hidden size of 128, reflecting a design prioritizing generalization and speed. Over 846 models were evaluated during the grid search, with the best-performing model chosen based on its RMSE and computational efficiency.

Training Details: Training took 59.925 seconds, dramatically faster than the control model, showcasing the benefits of a more optimized model structure in speed without compromising the learning capability.

Performance: This optimized model yielded an RMSE of 145.141, demonstrating a substantial improvement over the control model with a 70.05% reduction in RMSE compared to the baseline regressor and a 65.58% improvement over the control LSTM model.

Implications and Future Directions

The significantly improved performance of the similarity-scored LSTM model proves the potential of similarity-scoring metrics in meteorological data. The dramatic reduction in training time combined with superior prediction accuracy suggests that LSTMs, when properly tuned and optimized for specific datasets like those involving similarity metrics, can offer powerful tools for predictive analytics in fields requiring sequential data analysis.

Results

The results of this study demonstrate advancements in the prediction of agricultural outcomes through the integration of advanced machine learning models and novel data preprocessing techniques. By analyzing meteorological data from distinct geographic regions in New Mexico and leveraging similarity metrics, we have developed a robust methodology that improves the predictability of crop yields in microclimates with sparse data. Below, we detail the findings from each model employed in the study, showing their effectiveness in forecasting pecan yields.

Baseline and Baseline Regressor Performance:

Our baseline model, the baseline regressor, yielded an RMSE of 484.535. This initial metric set a fundamental benchmark for subsequent models, ensuring that any improvement in prediction accuracy could be attributed to more sophisticated data handling and modeling techniques rather than random chance.

XG Boost Model Insights:

Optimized for complex patterns in high-dimensional temporal data, the XG Boost model initially performed below expectations with an RMSE of 646.427 on dataset 4. This outcome suggested potential overfitting despite the extensive data and regularization efforts. However, when retrained with similarity metrics, the model achieved a more promising RMSE of 477.903. This improvement reveals the value of incorporating similarity scores to enhance model responsiveness to subtle patterns in the data.

MLP Model Achievements:

The MLP models demonstrated a marked improvement in predictive accuracy. The control MLP model achieved an RMSE of 382.851, while the model utilizing similarity scores from dataset 2 further reduced the RMSE to 366.67. The similarity-scored MLP represents a 24.32% improvement over the baseline regressor.

LSTM Model Performance:

The LSTM models showcased the most significant advancements. The control LSTM, with a deep architecture of 20 layers, initially recorded an RMSE of 421.519. Optimization through a custom grid search tailored for similarity metrics led to a dramatic improvement, with the similarity-scored model (5 layers, hidden size of 128) achieving an RMSE of 145.141. This result is a 70.05% improvement over our baseline and a 65.58% improvement over the control LSTM model, demonstrating the LSTM's capability to handle long-term dependencies in sequential data effectively.

Conclusion

This study aimed to enhance the predictive accuracy of agricultural outcomes in microclimates characterized by sparse meteorological data. By employing advanced machine learning techniques and innovative data preprocessing methods, including various similarity metrics, we have developed models that improve the forecast of crop yields, specifically pecan yields in New Mexico.

Our approach centered on integrating dimensionality reduction and similarity scoring to effectively handle and interpret high-dimensional temporal data streams from distinct geographic locations. Through analysis and modeling, we demonstrated that machine learning models, particularly those equipped with mechanisms to capture and analyze complex, sequential data patterns, such as LSTM networks, are exceptionally potent in this context.

The results were compelling. The LSTM models optimized with similarity metrics improved remarkably in predictive accuracy, with a reduced RMSE by over 70% compared to a simple baseline regressor. This improvement shows the efficacy of the tailored data preprocessing techniques when dealing with sparse data.

Moreover, the study revealed the importance of model choice and parameter optimization in achieving high accuracy in predictions. While the XG Boost and MLP models showed notable improvements with the integration of similarity scores, the LSTM models displayed superior capability in leveraging the temporal dynamics of the data, suggesting a promising direction for future research in agricultural forecasting and other related fields.

Introducing a Mixture of Experts architecture could refine the predictions by combining the strengths of various specialized models, potentially leading to even more robust and localized predictive insights. Additionally, expanding this methodology to other regions and crop types could help generalize the findings and enhance the models' applicability and impact.

In conclusion, this research further advances our understanding of the predictive relationships between microclimate variables and agricultural outputs. It also contributes to the broader field of precision agriculture. By improving forecast accuracy, this work aids in optimizing farming practices, enhancing productivity, and ensuring food security in the face of climatic variability, thus supporting sustainable agricultural development globally.