The Future of Harveston: Predicting Nature's Shifts

By NeuroNix - Data_Crunch_049
University of Moratuwa

1. Problem Understanding & Dataset Analysis

Land of Harveston, where agriculture is the backbone. There, farmers have long relied on traditional knowledge and instinct to guide their planting and harvesting decisions. But lately, unpredictable weather conditions have made decision-making complicated. With over a decade of climate data recorded across 30 kingdoms, this solution aims to develop a predictive model to forecast future weather patterns, particularly rainfall trends. By leveraging machine learning, this model will help farmers optimize planting cycles, allocate resources efficiently, and prepare for extreme weather, ensuring sustainable agriculture in an evolving climate.

Expected outcomes include:

1. Identifying key patterns and trends in the dataset to improve forecasting accuracy.
2. Developing a robust model that generalizes well to unseen data while minimizing prediction errors.
3. Providing actionable insights that stakeholders can use for strategic planning and operational efficiency.
4. Implementing an interpretable forecasting framework that can be adapted for real-world applications.

Key Findings and Data Analytics Techniques Used

Through the histogram plot, FFT plot, box plot, data distribution analytics(mean, max and min), and missing value-based analysis approaches, several key insights were uncovered in the dataset:

- The dataset does not contain any missing values, eliminating the need for imputation or handling of missing data.
- The dataset includes separate columns for Day, Month, and Year.
- Temperature values in the dataset were recorded in both Celsius and Kelvin.
- Rainfall data exhibited an annual cyclic trend, but a few extreme values were identified as outliers, potentially distorting forecasting accuracy.
- Data values have inconsistent decimal precision.

Preprocessing Steps and Justification

- Date Transformation: A new column was created by merging Day, Month, and Year to represent a complete Date field, enabling time-based analysis.
- Unit Standardization: Temperature values were converted to celsius to ensure consistency across the dataset.
- Outlier Removal: Extreme outliers in rainfall data were identified and removed to prevent skewing of model training and ensure reliable trend capture.

These preprocessing steps were crucial in ensuring the dataset was well-structured, consistent, and ready for feature engineering and model training.

2. Feature Engineering & Data Preparation

Selected Features & Justification

Temporal Features:

- **Year, Month, Day:** Essential for capturing seasonality and yearly trends.
- **dayOfYear:** Represents the day's position in the year (1-365), helping the model recognize seasonal cycles.
- **weekday:** Day of the week, for short-term weather trends.
- **days_in_month:** helps the model understand variations between months.

Cyclical Encoding for Time Variables:

- **month_sin, month_cos:** Encodes the month in a cyclical manner (e.g., January and December are close in the cycle).
- **day_sin, day_cos:** Represents daily changes in a cyclic format, ensuring continuity in the model's learning.

Justification: Time-based variables (months, days) are cyclical; using sine and cosine transforms prevents the model from misinterpreting December (20) and January (1) as distant.

Geographical Features/ Categorical Feature:

- **latitude, longitude/ kingdom:** Helps capture regional weather variations based on geographic location.
- **Justification:** Different kingdoms may experience different climatic conditions due to their geographical positioning.Allows the model to learn regional climate differences.

Impact on Model Performance

- **Using temporal and cyclical encodings** improved the model's ability to capture seasonal weather patterns.
- **Geographical features (latitude, longitude)** helped differentiate climate behaviors between regions.

Transformations for Data Stationarity

- Applied **StandardScaler** to numerical features for uniform feature distributions.
  - **Justification:** Ensures features with different scales do not disproportionately affect model performance.
- Applied **One-Hot Encoding** to categorical feature, `kingdom`.
  - **Justification:** Allows models to differentiate between categorical values without imposing an ordinal relationship.

3. Model Selection & Justification

Baseline & Advanced Models Considered:

- **Baseline:** Linear Regression.
- **Advanced:** Decision Tree Regressor (Chosen Model), Random Forest, XGBoost,LSTM, prophet.

Models approached and opinion,

- **Linear Regression** - Assumes a linear relationship ,can't capture nonlinear dependencies in weather patterns.
- **Decision Tree Regressor (Chosen Model)-** Handles both linear and nonlinear relationships effectively, Works well with missing or unstructured data,Computationally efficient for a given medium-sized dataset, Can overfit, but pruning techniques mitigate this.
- **Random Forest -** An ensemble of multiple decision trees, reducing overfitting, robust predictions and improves generalization, Handles feature importance analysis well,Higher computational cost comparatively.
- **XGBoost -** builds trees sequentially, Strong performance with feature-rich data and handles missing values well, Requires careful hyperparameter tuning to avoid overfitting.
- **LSTM -** Suitable for sequential and time-dependent data, Can learn long-term dependencies, Computationally expensive and requires large datasets for optimal performance,Not chosen due to the relatively small dataset size and interpretability concerns.
- **Prophet -** forecasting model developed by Facebook, Handles seasonality and trend components effectively, Simple to implement and interpret,Works well with structured time series data but struggles with highly nonlinear dependencies.

Why Decision Tree Regressor?

- Interpretable: Unlike black-box models (e.g., LSTM), decision trees provide clear insights into feature importance.
- Nonlinearity Handling: Can capture complex relationships in the weather data
- Computational Efficiency: Faster training and inference times.
- Good Performance: lower error rates, less prone to overfitting compared to other models.

Hyperparameter Optimization

- **Decision Tree Regressor**
  - Grid Search: Optimized parameters such as `max_depth`, `min_samples_split`, and `max_features`.
  - Cross-Validation**:** Ensured model generalization across different time periods.
- **Random Forest & XGBoost:** Used Grid Search and Bayesian Optimization to tune `n_estimators`, `max_depth`, and learning rate.
- **LSTM:** Experimented with different numbers of hidden layers and dropout rates to prevent overfitting.
- **Prophet:** Tuned seasonality parameters and growth rate constraints to match weather trends.

Validation Approach

- **k-Fold Cross-Validation:** Splitting the dataset into sequential folds while ensuring past data is always used to predict the future.

4. Performance Evaluation & Error Analysis

Evaluation Metrics used

- **Root Mean Squared Error (RMSE)-** Measures the standard deviation of residuals (prediction errors). Penalizes large errors more than small errors, making it useful for evaluating overall model accuracy.
- **Mean Absolute Error (MAE) -** Represents the average absolute difference between predicted and actual values. Provides a more interpretable measure of typical prediction error.
- **Symmetric Mean Absolute Percentage Error (SMAPE) -** A scale-independent metric that normalizes absolute errors based on the magnitude of the actual values. Useful for comparing performance across datasets with different scales.
- **R^2 Score -** Represents the proportion of variance in the target variable explained by the model.It will explain how much variance in the weather data is explained by the model(>0.7 is good).

Model Performance Comparison

Model	Description
Decision Tree	**Chosen for its high accuracy (highest SMAPE)**. Decision Trees are simple, interpretable, and effective for structured datasets, making them ideal for the given weather prediction task. Their decision-making process is transparent, allowing for easy identification of patterns in the data. The model was selected due to its superior performance and ease of understanding.
Random Forest	**Second-best performance**. Random Forest, an ensemble method, improves accuracy by reducing overfitting and variance compared to a single Decision Tree. Although it performed slightly less than the Decision Tree, it still produced robust results and was considered a strong contender.
XGBoost	**Third-best performance**. XGBoost provides high accuracy through gradient boosting, but it did not outperform Decision Tree and Random Forest for this specific dataset. Despite its efficiency and powerful capabilities, it may have struggled with the dataset's specific characteristics, such as feature dependencies or data volume.
LSTM	**Lower performance**. LSTM, a deep learning model, is typically effective for sequence-based data but was not ideal for this task. The weather prediction task may not have had enough sequence-related patterns for LSTM to excel, leading to its lower accuracy in comparison.
Prophet	**Less accurate for this dataset**. While Prophet is great for time series forecasting, it may not have performed well with the given features (which are not purely time series in nature). This could explain its relatively lower accuracy in comparison to the tree-based models.

Residual Analysis

Avg_Temperature →0.6067 → Strong positive autocorrelation

Radiation → 0.4639 → Very strong positive autocorrelation

Rain_ Amount → 0.9023 → Moderate positive autocorrelation

Speed_x → 1.0873 → Mild positive autocorrelation

Speed_y → 0.7965 → Moderate positive autocorrelation

Residuals exhibit significant positive autocorrelation, meaning successive residuals are correlated, which is problematic for time series forecasting.

While Residuals of Avg_Temperature is close to a normal distribution, it is still not normally distributed. For other target variables, residuals are not normally distributed.

All the target variables exhibit a funnel-shaped pattern in the residuals vs. predicted values plot, indicating the presence of heteroscedasticity, where the variance of residuals increases with larger predicted values.

Model Limitation: The model struggled with predicting extreme weather conditions accurately, such as sudden temperature spikes ,and heavy rainfall. Decision Trees, while powerful, are prone to overfitting, especially with limited data. This challenge was mitigated by tuning hyperparameters like max depth and using cross-validation, but continuous monitoring is necessary to ensure generalizability.

Potential Future Improvements:

- **Ensemble Methods:** Incorporating ensemble learning techniques such as **Random Forest** and **XGBoost** could help further reduce overfitting and increase model robustness.
- **External Data Sources:** The inclusion of additional data sources, such as historical weather data from nearby regions, or atmospheric pressure readings, could significantly enhance forecasting accuracy.
- **Real-Time Model Updates:** Implementing a **feedback loop** to retrain the model periodically with new data would help the model stay up-to-date with evolving climate patterns.
- **Model stacking:** Stacking two best performing models can improve overall accuracy.

5. Interpretability & Business Insights

Real-World Applications of Forecasting Results

- **Agricultural Planning**: The weather forecasts can assist farmers in planning their crop cycles, irrigation schedules, and resource allocation inorder to reduce the risk of crop damage due to unforeseen weather extremes.
- **Climate Monitoring**: Policymakers and environmental agencies can use the forecasting model to monitor climate trends and prepare for potential weather-related challenges such as floods, droughts, or temperature extremes.
- **Disaster Preparedness**: The model's predictions can be used for early warning systems to notify communities and government agencies of potential weather-related disasters.
- **Energy Management**: With accurate temperature, wind speed and radiation forecasts, energy companies can better manage the production of renewable energy sources like solar and wind.
- **Business Management**: By predicting the future trends in the market and customer demand, a business can scale and prepare for the unforeseen circumstances

Improvements in Forecasting Strategy and Model Deployment

- **External Data Incorporation**: Inclusion of additional features, such as atmospheric pressure, humidity, and historical weather data from nearby regions. These could offer a more comprehensive understanding of weather patterns.
- **Stacking Models**: To improve robustness and performance, stacking models could be explored. Combining multiple models may help reduce the variance and bias present in the predictions of individual models.
- **Real-Time Model Deployment**: In operational settings, deploying the model in real-time through an API could allow stakeholders (farmers, policymakers, etc.) to access weather forecasts as they become available. The model could be connected to real-time weather data sources (e.g., weather stations) for continuous learning and updates.
- **Model Explainability**: For greater trust in the model's predictions, implementing model interpretability tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) could provide insight into how different features influence the model's decision-making process.
- **Continuous Model Updates**: As weather patterns evolve, the model should be updated periodically to ensure it remains accurate. Incorporating a feedback loop where the model can be retrained with new data would help maintain its relevance and accuracy.

6. Innovation & Technical Depth

To enhance forecasting accuracy and efficiency, several innovative techniques were implemented:

- Seasonal Trend Pattern Analysis: The dataset exhibited strong annual cyclic trends, particularly in variables like rainfall. Seasonal decomposition techniques such as STL (Seasonal-Trend Decomposition using Loess) and Fourier transformations were used to identify recurring trends and periodic variations in the data.
- Abnormality Detection and Correction: Deviations from the seasonal patterns were analyzed to detect abnormal spikes or drops in values. By comparing actual data points to expected seasonal trends, outliers were identified and addressed to improve model stability and forecasting precision.
- Handling of Data Irregularities: Custom outlier detection and correction strategies were used to prevent extreme deviations from distorting model performance, ensuring robust and reliable predictions.
- Cyclic Feature Encoding: To capture the periodic nature of time-related variables, sin-cos transformations were applied to Month and Day values. Additionally, the longitude and latitude of the kingdom were transformed into sin-cos values to better represent spatial cyclicality, allowing the model to learn geographical influences on weather patterns effectively.
- Automated Hyperparameter Tuning with Grid Search: Grid Search was employed to systematically explore different combinations of hyperparameters for models such as ARIMA, LSTM, and XGBoost. This process helped in selecting the optimal parameters, enhancing model accuracy, and reducing the risk of underfitting or overfitting.

7. Conclusion

Key Findings

The Decision Tree Regressor model outperformed baseline models, including Linear Regression, in terms of key evaluation metrics . This indicates that the model is capturing important relationships between environmental variables and their forecasts.
Feature engineering techniques such as Temporal Features, Geographical Features, and seasonality encoding significantly improved the model's predictive capabilities.Data preprocessing, including standardization and outlier removal, was critical in improving model stability and performance.

Best-Performing Model

The **Decision Tree Regressor** was the best-performing model within experimented ones. It was particularly effective in handling both linear and nonlinear relationships, and it performed well without extensive feature scaling. The model's ability to capture complex interactions between variables.

Challenges Faced

Handling Inconsistent Data -Though the dataset was mostly clean, occasional inconsistencies could still affect model training. Ensuring complete data and handling any gaps in real-time scenarios would be an ongoing challenge.

Extreme Weather Events -The model struggled with predicting extreme weather conditions accurately, such as sudden temperature spikes or heavy rainfall.

Overfitting Risk - Decision Trees, while powerful, are prone to overfitting, especially with limited data.

Potential Future Improvements - Incorporating Ensemble learning Methods, Inclusion of additional data sources, implementing real-time model update, stacking two best performing models, and improved model interpretability could further improve the model performance and adaptability.

In summary, the forecasting model has the potential to provide valuable insights for agricultural planning and climate monitoring. While the Decision Tree Regressor performed well, there is room for improvement, particularly in handling extreme weather events and integrating more data sources to enhance accuracy and robustness.