CODEJAM

DATA CRUNCH

The Future of Harveston: Predicting Nature's Shifts

By Xforce
Data_Crunch_079
University Of Moratuwa

Table Of Contents

Problem Understanding & Dataset Analysis

- ## Clearly Defined Objective
    - Our project aimed to develop accurate forecasting models for five critical environmental variables (Temperature, Radiation, Rain Amount, Wind Speed, and Wind Direction) to support Harveston's agricultural planning. These predictions enable farmers to adapt to climate variations, optimise resource allocation, and implement effective risk management. By delivering forecasts in standardised meteorological units, our models directly support agricultural sustainability goals across the kingdom.
- ## Key Findings from Data Analysis
    - Our analysis of 84,960 training records across multiple kingdoms revealed distinct patterns essential for model development. Temperature data exhibited consistent seasonal trends with kingdom-specific variations, suggesting simplified modeling approaches could be practical. Radiation measurements showed a strong correlation with seasonal cycles and geographic positioning. Rainfall patterns demonstrated pronounced geographical variation with a notable right-skewed distribution requiring specialised transformation. Wind variables presented unique challenges, particularly Wind Direction, which necessitated circular statistical methods to address its inherent periodicity. We identified strong seasonal dependencies in all variables that directly impact agricultural planning.
- ## Preprocessing Methodology
    - We implemented comprehensive preprocessing to enhance data quality and predictive accuracy. This included temperature standardisation through Kelvin-to-Celsius conversion, geographic coordinate processing with derived spatial relationship features, and temporal feature engineering using cyclical sine/cosine transformations. These transformations adequately represented the circular nature of calendar variables, effectively capturing seasonality without creating artificial discontinuities. Kingdom-specific statistical aggregations provided contextual baseline features that significantly improved prediction accuracy by capturing regional climate characteristics. This approach enabled us to address the complex interactions between location, season, and meteorological patterns that define Harveston's diverse climate systems.

Feature Engineering & Data Preparation

- ## Feature Creation Techniques
  - Our feature engineering approach systematically captured temporal, spatial, and domain-specific patterns within Harveston's climate data, resulting in 88 carefully constructed features. To represent cyclical time patterns, we transformed calendar variables using sine and cosine functions, enabling our models to recognise appropriate relationships between adjacent periods (e.g., December and January). This cyclical encoding was simultaneously applied to multiple time scales—day, month, and season—creating a rich temporal representation.
  - We enhanced the kingdom data with direct and derived spatial features for geographical representation. Beyond basic latitude and longitude coordinates, we implemented trigonometric transformations (latitude/longitude sine/cosine components) and created interaction terms that captured the relationship between geographic position and seasonal effects. These geographic-seasonal interactions proved crucial for modeling how climate patterns vary by location throughout the year.
- ## Feature Selection Strategy
  - Our feature selection process combined domain knowledge with empirical validation through cross-validation testing. Kingdom-based features demonstrated exceptional value, confirming our hypothesis that regional climate variations significantly impact prediction accuracy. Statistical aggregations by kingdom for each target variable created powerful baseline predictors that captured local climate behaviors.
  - The optimal feature set included 88 variables across several categories: baseline temporal features (12), kingdom one-hot encodings (variable based on dataset), geographical coordinates and derivatives (9), kingdom-specific statistics (multiple per target variable), and strategic interaction terms. Feature importance analysis from our gradient boosting models revealed that kingdom-based features and seasonal indicators consistently ranked highest across target variables, validating our domain-driven approach.
- ## Data Transformation Approach
  - Each target variable required specific transformation strategies to address its unique statistical properties. For Rain_Amount, which exhibited positive skewness, we implemented a specialized approach combining direct modeling with non-linear dampening for values exceeding 5mm, followed by kingdom-specific adjustment factors ranging from 0.85 to 0.95 based on historical precipitation patterns.
  - Wind Direction presented a fundamental challenge due to its circular nature—standard regression cannot appropriately model the discontinuity between 359° and 0°. Our solution decomposed directions into sine and cosine components, built separate models for each component using ElasticNet and Ridge regression, respectively, then recombined predictions using arctangent transformation and appropriate range adjustment.
  - We applied a 1.02 calibration factor for Radiation to account for systematic measurement characteristics while maintaining physical consistency with meteorological principles. All models underwent rigorous validation to ensure transformations preserved the signal while improving prediction stability.

Model Selection & Justification

- ## Modeling Approach and Algorithm Selection
    - Our modeling strategy employed a variable-specific approach, recognising that different environmental factors exhibit unique statistical properties requiring tailored methodologies. After thoroughly evaluating multiple algorithms, we selected an ensemble-based framework customized for each target variable.
    - For Average Temperature prediction, our analysis revealed remarkable consistency across temporal and spatial dimensions. Despite testing complex modeling approaches, we determined that a fixed-value prediction (24.8°C) with minor seasonal adjustments produced optimal results. This approach outperformed more complex alternatives in cross-validation testing, demonstrating that constrained models can capture fundamental relationships more effectively than flexible ones when underlying patterns exhibit high stability.
    - For Radiation forecasting, we implemented a weighted ensemble combining XGBoost and LightGBM models (45%/55% weighting, respectively). The XGBoost component utilised 250 estimators with depth-5 trees and careful regularisation parameters to prevent overfitting, while the LightGBM implementation employed 300 estimators with leaf-wise growth. This ensemble captured global radiation patterns and local variations, with a custom 1.02 adjustment factor applied to account for systematic measurement characteristics.
    - Rain Amount prediction presented specific challenges due to its right-skewed distribution and high geographic variability. Our solution utilised LightGBM with L1 regularisation, optimized for mean absolute error to reduce sensitivity to outliers. For values exceeding 5mm, we applied non-linear dampening through logarithmic transformation, preserving prediction fidelity across the full range of possible precipitation levels. Kingdom-specific calibration factors (ranging from 0.85 to 0.95) were implemented to account for regional microclimate effects.
    - Wind variables required specialised approaches. For Wind Speed, a weighted ensemble combining XGBoost (70%) and LightGBM (30%) with optimised hyperparameters provided superior performance. Wind Direction employed our trigonometric decomposition method with ElasticNet for sine component prediction and Ridge regression for cosine component prediction, followed by arctangent transformation to reconstruct directional forecasts.
- ## Cross-Validation Framework
    - Our validation strategy employed 8-fold cross-validation with stratified sampling to ensure the representative distribution of geographic and temporal factors across folds. This approach provided robust performance estimates while maintaining sufficient training data in each fold to capture complex patterns. The consistent random seed (42) ensured reproducibility across all modeling steps.
    - For each target variable, model selection decisions were based on comprehensive evaluation metrics derived from cross-validation, emphasising sMAPE to align with the competition evaluation criteria. This rigorous validation framework prevented overfitting and ensured our models would generalise effectively to future periods, providing Harveston's agricultural community with reliable forecasting capabilities.
    - Our final ensemble implementation achieved a score of 36.8, representing substantial improvement over baseline approaches and demonstrating the effectiveness of our specialised modeling strategy.

Performance Evaluation & Error Analysis

- ## Evaluation Metrics and Methodology
  - Our forecasting models were evaluated using the Symmetric Mean Absolute Percentage Error (SMAPE), which effectively normalizes prediction accuracy across variables with different units and scales. This metric proved ideal for Harveston's diverse climate variables by providing a balanced assessment that normalizes errors relative to both actual and predicted values. We supplemented this with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) analyses, confirming our models' consistent performance across different error formulations.
- ## Model Performance Analysis
  - Our ensemble approach achieved a final score of 36.8 across all target variables, significantly outperforming single-model solutions. The fixed-value approach for temperature prediction maintained consistent accuracy across kingdoms and seasons, demonstrating that constrained models often capture fundamental patterns more reliably than complex alternatives. For Radiation and Wind Speed, our weighted ensembles reduced error by approximately 12% compared to individual algorithms by leveraging the complementary strengths of XGBoost and LightGBM. Rain Amount prediction benefited substantially from our kingdom-specific calibration framework, with differentiated adjustment factors (0.85-0.95) effectively addressing regional precipitation patterns while preserving model sensitivity across rainfall scenarios.
- ## Diagnostic Analysis
  - Residual analysis confirmed our models' statistical validity through multiple diagnostics. Autocorrelation tests revealed minimal temporal dependence in residuals, with Durbin-Watson statistics consistently near the optimal 1.8-2.2 range, indicating our features successfully captured time-based patterns. Residual distributions demonstrated approximately normal patterns for most variables, with expected minor deviations for Rain Amount due to its inherent skewness. Heteroscedasticity examinations showed consistent error variance across prediction ranges, with our ensemble approach effectively mitigating any mild heteroscedasticity detected.
- ## Limitations and Improvement Areas
  - Despite strong performance, we identified several enhancement opportunities. Our temperature predictions could benefit from more nuanced geographical adjustments for microclimates in kingdoms with unique topography. Extreme rainfall events remain challenging to forecast due to their rare and stochastic nature, suggesting potential for specialized high-precipitation models. Additionally, incorporating external climate change projections could enhance long-term forecasting capabilities beyond the current prediction window. Wind pattern predictions showed slightly elevated error rates across kingdom boundaries, indicating that more sophisticated spatial interpolation methods could improve transition zones between regions.

Interpretability & Business Insights

- ## Applied Value of Forecasting Results
  - Our forecasting models deliver substantial practical value to Harveston's agricultural ecosystem by providing accurate predictions of critical environmental variables that directly inform farming decisions. Temperature forecasts enable optimal planting schedule determination, allowing farmers to confidently select appropriate crop varieties based on anticipated thermal conditions. This is particularly valuable for temperature-sensitive crops requiring specific growing degree day accumulations.
  - Radiation predictions support critical decisions for crops where solar intensity directly impacts yield and quality. Vineyard managers can anticipate sugar development in grapes, while grain farmers can optimize fertilizer application timing based on expected photosynthetic activity. The implemented 1.02 radiation adjustment factor ensures predictions accurately reflect the full spectrum of solar intensity relevant to plant development.
  - Rainfall forecasts provide the foundation for water resource management. Our kingdom-specific adjustment approach acknowledges regional variation in precipitation patterns, enabling localized planning. These predictions allow farmers to prepare irrigation infrastructure before dry periods and reduce water usage during adequate rainfall, optimizing crop health and resource utilization.
  - Wind predictions (speed and direction) enable strategic scheduling of critical agricultural activities, including pesticide application, pollination management, and frost protection. These forecasts also help livestock managers implement appropriate shelter during potentially stressful weather conditions.
- ## Implementation Strategy
  - To maximize impact, we recommend a comprehensive implementation strategy with four key components:
  - First, establish a centralized prediction system with multi-channel distribution through mobile applications, agricultural extension offices, and community radio broadcasts, ensuring accessibility regardless of technological resources.
  - Second, predictions should be translated into agriculturally relevant guidance rather than presented with technical meteorological values. For example, provide actionable insights such as "adequate soil moisture expected for germination" instead of raw rainfall amounts.
  - Third, implement a tiered prediction schedule with different horizons: 1-3 day forecasts for immediate operations, 7-10 day outlooks for short-term planning, and monthly trends for strategic resource allocation.
  - Finally, a continuous improvement framework comparing predictions against actual conditions, with model parameters automatically adjusted based on performance, must be established. This adaptive approach will progressively enhance prediction accuracy while building farmer trust, ultimately supporting Harveston's agricultural sustainability through improved climate adaptation.

Innovation & Technical Depth

- ## Novel Modeling Approaches
  - Our solution implemented several innovative techniques significantly enhancing prediction accuracy across all target variables. We developed a specialized approach to wind direction forecasting using trigonometric decomposition. By decomposing directional data into sine and cosine components and modeling each separately (ElasticNet for sine, Ridge regression for cosine), we effectively addressed the circular discontinuity problem that traditional regression models cannot handle.
  - For temperature prediction, we adopted an unconventional fixed-value approach (24.8°C) with minor seasonal adjustments after discovering that complex models consistently converged toward similar predictions. This insight enabled a highly efficient solution that outperformed more computationally intensive approaches while providing stable forecasts.
  - Our ensemble methodology implemented strategically weighted combinations tailored to each target variable. For radiation, a 45%/55% weighting between XGBoost and LightGBM models captured complementary aspects of the underlying patterns. For rainfall prediction, we incorporated kingdom-specific adjustment factors (0.85-0.95) based on historical precipitation patterns, effectively capturing regional microclimate effects.
- ## Advanced Feature Engineering
  - We developed 88 carefully engineered features through a structured temporal, spatial, and domain-specific data representation approach. Temporal features utilized cyclical encodings with sine and cosine transformations at multiple time scales simultaneously (day, month, season), capturing both short-term fluctuations and long-term seasonal trends.
  - Geographic feature engineering extended beyond basic coordinates to include derived features such as absolute latitude, trigonometric transformations, and specialized interaction terms. These geographic-seasonal interactions captured the complex relationship between location and temporal patterns, addressing how seasonal effects vary by geography.
  - We implemented a specialized 1.02 adjustment factor for radiation modelling based on observed solar radiation patterns, maintaining physical consistency with meteorological principles.
  - These innovations collectively contributed to our final score of 36.8, representing a substantial improvement over baseline approaches and demonstrating the effectiveness of our technically sophisticated yet carefully calibrated modeling strategy.

Conclusion

Our comprehensive approach to forecasting Harveston's critical environmental variables successfully balanced sophisticated modeling techniques with practical agricultural applications. By developing specialized solutions for each target variable—temperature, radiation, rainfall, wind speed, and wind direction—we created a forecasting system that achieves high accuracy while remaining interpretable and actionable for farming communities.

The final model's score of 36.8 reflects the effectiveness of our feature engineering, algorithm selection, and ensemble integration strategies. Most importantly, these predictions provide valuable guidance for agricultural planning, resource management, and risk mitigation across Harveston's diverse kingdoms.

Key technical innovations in our approach included the trigonometric transformation method for wind direction forecasting, strategic ensemble weighting for radiation and wind speed prediction, and kingdom-specific calibration factors for rainfall forecasting. The fixed-value temperature prediction with seasonal adjustments demonstrated that well-validated simplified approaches can sometimes outperform more complex alternatives.

Throughout our modeling process, we maintained a strong focus on agricultural relevance, ensuring that the forecasting system addresses the specific needs of Harveston's farming communities. The proposed implementation recommendations would enable widespread access to these predictions through multiple channels, with appropriate translation into actionable agricultural insights.

The forecasting framework we developed could be further enhanced by incorporating climate change projections, expanding the temporal prediction horizon, and creating more sophisticated spatial interpolation methods for cross-kingdom boundary areas. As Harveston adapts to changing environmental conditions, data-driven prediction systems will remain essential for maintaining agricultural productivity and food securit