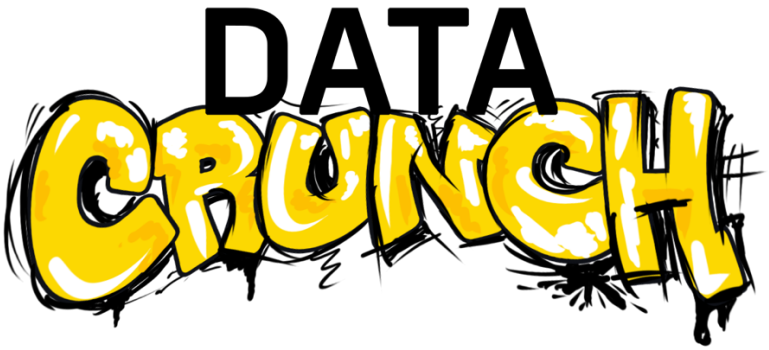

— Preliminary Round —

Organized By

Department of Computer Science & Engineering
University of Moratuwa

March 29, 2025

1. Background to the Problem

The Future of Harveston: Predicting Nature's Shifts

Overview

In the heart of Harveston, a self-sufficient land of sprawling fields and winding rivers, the people have lived in harmony with nature for generations. Their survival depends on the land's ability to provide, and every harvest is a delicate balance between the forces of nature.

But lately, the balance has begun to shift. The rains are becoming unpredictable, the sun's intensity varies, the winds change course without warning, and the once-reliable climate patterns seem to be evolving. The farmers of Harveston can no longer depend

solely on tradition and instinct to guide them—they need a new way to understand their environment.

For over a decade, scholars and record-keepers of Harveston have carefully documented the land's changing conditions, recording fluctuations in temperature, rainfall, solar energy, and wind patterns. However, the measuring units for some features vary between kingdoms. This invaluable collection of environmental data holds the answers, but its complexity is beyond human intuition. Now, they turn to data scientists and machine learning experts from across the world to help decipher the patterns hidden within these records.

2. Background to the Problem

Agriculture is the lifeline of Harveston, and the ability to anticipate changes in weather conditions is crucial for its survival. A well-crafted predictive model could mean better planning for planting and harvesting, improved resource management, and protection against unforeseen weather extremes. Your work could help ensure food security and economic stability for generations to come.

Can you build a model that predicts the shifts in Harveston's climate and helps its people secure their future?

Accurate predictions will help Harveston's farmers make informed decisions about planting cycles, resource allocation, and preparation for weather extremes, ensuring food security and economic stability.

3. Datasets and Variable Description

The dataset contains historical environmental measurements from multiple kingdoms across Harveston, including:

- ID - Unique identifier for each record.
- Year - The year of the recorded data.

- Month - The month of the recorded data.
- Day - The day of the recorded data.
- kingdom - The geographical region where the data was recorded.
- latitude (°) - Geographic latitude coordinate.
- longitude (°) - Geographic longitude coordinate.
- Avg_Temperature (°C or K) - Average recorded temperature.
- Avg_Feels_Like_Temperature (°C or K) - Average perceived temperature, combining wind chill, humidity, and solar radiation.
- Temperature_Range - Difference between maximum and minimum temperature.
- Feels_Like_Temperature_Range - Difference between maximum and minimum feels-like temperature.
- Radiation (W/m^2) - Shortwave solar radiation received.
- Rain_Amount (mm) - Total precipitation recorded, including rain and snow.
- Rain_Duration (hours) - Number of hours with precipitation.
- Wind_Speed (km/h) - Average wind speed recorded.
- Wind_Direction (°) - Dominant wind direction.
- Evapotranspiration (mm) - Total water loss due to evaporation and plant transpiration.

4. Problem Statement

Your mission is to develop time series forecasting models that accurately predict five critical environmental variables in the given units that will impact Harveston's agricultural future:

1. Average Temperature (°C)
2. Radiation (W/m^2)
3. Rain Amount (mm)
4. Wind Speed (km/h)
5. Wind Direction (°)

Participants must create models to forecast the above variables for future dates provided in the test.csv file.

5. Evaluation Criteris

Your submissions will be evaluated based on the following:

1. **Prediction Accuracy (85% of total score)**
    - A private leaderboard will be calculated based on the score from the entire test dataset according to the evaluation metrics given.
    - Note that the public leaderboard shows results on only a portion of the test data.
2. **Code Quality and Technical Report (15% of total Score)**
    - Your report will be evaluated according to the provided rubric covering:
        - Problem understanding and dataset analysis
        - Feature engineering and data preparation
        - Model selection and justification
        - Performance evaluation and error analysis
        - Interpretability and business insights
        - Innovation and technical depth

Table 1 - Evaluation Rubric for Report

Category	Criteria	Points
1. Problem Understanding & Dataset Analysis (15 pts)	Clearly define the problem, objectives, and expected outcomes	5
	Justifies preprocessing steps (handling missing values, scaling, smoothing, outlier removal)	10
2. Feature Engineering & Data Preparation (20 pts)	Creates relevant features (lags, moving averages, external variables, etc.)	10

	Explains feature selection and its impact on prediction performance	5
	Addresses data stationarity, transformations (log, differencing, normalization), if necessary	5
3. Model Selection & Justification (25 pts)	Uses baseline and advanced models (ARIMA, LSTM, Prophet, XGBoost, etc.)	5
	Justifies model selection based on dataset characteristics and problem requirements	10
	Optimizes hyperparameters effectively (grid search, Bayesian, autoML, etc.)	5
	Implements appropriate time series validation methods	5
4. Performance Evaluation & Error Analysis (20 pts)	Uses appropriate metrics (Given metric in the Kaggle and additional metrics)	5
	Compares model performances and provides reasoning for the best model choice	5
	Performs residual analysis (autocorrelation, normality, heteroscedasticity checks)	5
	Discusses limitations, potential biases, and areas of improvement	5
5. Interpretability & Business Insights (10 pts)	Explains how the predictions can be used in real-world applications	5
	Suggests ways to improve model deployment and forecasting strategy	5
6. Innovation & Technical Depth (10 pts)	Implements novel approaches (custom architectures, advanced feature engineering, ensemble learning, etc.)	10
Total Score		**100 pts**

6. Deliverables

- Predictions file matching the format in sample_submission.csv
- Jupyter notebook or script files containing your complete solution
- Technical report not more than 10 pages (PDF format) following the evaluation rubric guidelines

You must submit a **GitHub repository link** containing all the above deliverables in the given specified repo structure.
Rename your GitHub repository to match the team ID provided.
The repository must be public after the submission deadline; otherwise, your submission will be disqualified.

Repository structure

/your-repo-name (Data_Crunch_001)

├— Notebooks_and_Scripts/

├— your_notebooks

├— your_scripts

├— final_submission.csv

├— technical_report.pdf

7. Evaluation Metrics

The following metrics will evaluate the submissions:

- sMAPE – Symmetric Mean Absolute Percentage Error

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_{\text{true},i} - y_{\text{pred},i}|}{\frac{|y_{\text{true},i}| + |y_{\text{pred},i}|}{2}}$$

8. Rules

Teams and collaboration

- All team members must create a Kaggle team after joining the competition and submit entries as that team.
- The team's name should match the Team ID provided and must be displayed on the leaderboard.
- Submitting individually is not allowed.
- Code must not be shared privately outside of a team.
- Violation on one or more condition may cause to disqualified from the competition.

Submissions and winning

- One team can only make a maximum of 8 submission per day
- Before the end of the competition, you need to choose 2 submissions to be judged on for the private leaderboard. If you do not make a selection your 2 best public leaderboard submissions will be used to score on the private leaderboard.
- During the competition, your best public score will be displayed regardless of the submissions you have selected. When the competition closes your best private score out of the 2 selected submission will be displayed
- We maintains a **public leaderboard** and **private leaderboard** for the competition. The Public Leaderboard includes approximately 19% of the test dataset. While the competition is open, the Public Leaderboard will rank the submitted solutions by the accuracy score they achieve. Upon close of the competition, the Private Leaderboard, which covers the other 81% of the test dataset, will be made public and will constitute the final ranking for the competition.
- Note that to count, your submission must first pass processing. If your submission fails during the processing step, it will not be counted and not receive a score; nor will it count against your daily submission limit. If you encounter problems with your submission file, your best course of action is to ask for advice on the Competition's discussion forum.

- If two solutions earn identical scores on the leaderboard, the tiebreaker will be the date and time in which the submission was made (the earlier solution will win).
- You acknowledge and agree that Data Crunch may, without any obligation to do so, remove or disqualify an individual or team, if Data Crunch believes that such individual or team, is in violation of these rules. Entry into this competition constitutes your acceptance of these official competition rules.
- Data Crunch also reserves the right to disqualify you and/or your submissions if we believe that you violated the rules or violated the spirit of the competition or the platform in any other way. The disqualifications are irrespective of your position on the leaderboard and completely at the discretion of Data Crunch.

Reproducibility of submitted code

- If your submitted code does not reproduce your score on the leaderboard, we are reserve the right to adjust your rank to the score generated by the code you submitted.
- If your code does not run you will not be getting any marks. Please make sure your code runs before submitting your solution.
- **Always set the seed.** Rerunning your model should always place you at the same position on the leaderboard. When running your solution, if randomness shifts you down the leaderboard we reserve the right to adjust your rank to the closest score that your submission reproduces.

Join us in helping the people of Harveston adapt to their changing environment and secure their agricultural future!