**Report:**

* **Write in a suitable for non-technical audience**
* **Report findings (strengths/weaknesses of potential models for use with new instances)**
* **EXPLAIN DECISIONS**

**What we still need to do:**

* **Decide on what we will say we’ve done (can just say in accordance w/ contact)**
* **Select a model**
* **Make predictions (in valid format)**
* **Save predictions**
* **Submit:**
  + **.ipynb**
  + **PDF version of notebook (w/ output and markdown)**
  + **A clear concise PDF summary report (recommended 6 pages)**
  + **predictions.csv text file**
  + **Individual statements on individual contribution (maximum 1 page)**

**19-20: An exceptionally clearly presented notebook and summary report, demonstrating extensive successful implementation of all the basic ML process elements, very well explained decisions and analysis of results. Assignment conveys significant insight into data preprocessing, and important features in the success of the model.**

Introduction:

* The Big Picture: predicting temperature, why it’s useful
* The dataset: ERA5, size, common use

The ERA5 is a publicly available dataset from the Centre for Medium-Range Weather Forecasts (ECMWF). This dataset provides hourly estimates for a large range of atmospheric, ocean-wave and land-surface quantities recorded over many years. Data pertains to a specific geolocation which is identifiable by a given longitude and latitude.

Predicting the temperature based off of the many features included within this dataset is a problem which proves especially relevant to weather forecasting. Machine learning methods can be used to obtain a prediction of temperatures for data fed into the model

Goal: Create models for predicting temperature given observations about weather

Methods:

* Exploration of data structure
* Detail preprocessing/ data cleaning
* Considerations of making use of available data
* Feature engineering (methods of nearby & others) – why are nearby features useful
* Self-made validation set
* Cross validation for hyperparameter tuning
* Construction of several machine learning regression models
* Selection of one model to make predictions on the test data
* Evaluation methods

Methods

Data Analysis and Exploration

The initial exploration of the ERA5 dataset focused on understanding the structure, distribution and relationships within the weather data. A systematic analysis of all variables was conducted to identify patterns that could inform the modelling approach and preprocessing steps. First assessed the structure and quality of the data, checking for missing values, outliers, and inconsistencies that could impact model performance.

The temperature distribution (t2m) was examined first, it showed a roughly normal distribution ranging primarily between 270K and 300K (approximately -3°C to 27°C), with a peak around 282-285K (9-12°C). This aligned with expected UK climate patterns, giving confidence in the data validity. Also explored the distributions of all other meteorological variables to identify potential outliers or anomalies.

Geospatial analysis was particularly important, as weather patterns are inherently tied to location. Temperature maps across the UK revealed clear regional variations with generally warmer temperatures in southern England and cooler temperatures in northern Scotland. The seasonal analysis further showed how these patterns shift throughout the year, with more pronounced north-south temperature gradients in winter and summer compared to transitional seasons.

Correlation analysis identified meaningful relationships between temperature and other meteorological variables. Notably, surface pressure (sp) showed a moderate positive correlation with temperature, while wind components (u10, v10, u100, v100) displayed complex relationships that varied by location and season.

Data Preprocessing

Based on the exploration, a consistent preprocessing pipeline was created and applied to all models:

1. Temporal Feature Extraction: Decomposed the `valid\_time` datetime field into component features (year, month, day, hour) and derived a seasonal indicator to capture cyclical weather patterns.

2. Categorical Encoding: Applied one-hot encoding to categorical features including year, month, day, hour, season and precipitation type (ptype) to allow models to learn distinct patterns for each category.

3. Numerical Feature Processing: For continuous variables, applied median-based imputation to handle any missing values, followed by standardisation to ensure all features contributed equally to model training regardless of their original scales.

4. ID Removal: Removed the ID field as it contained no predictive information.

The preprocessing approach employed scikit-learn pipelines to ensure consistent treatment of features across training and testing datasets, preventing data leakage and enabling reproducible results.

Feature Engineering for Spatial Context

A key innovation in our approach was incorporating spatial context through 3 distinct methods of "nearby features":

1. Radius-Based Aggregation (Gradient Boosting): For each weather observation point, calculated statistical aggregations (mean, minimum, maximum) of key meteorological variables from surrounding locations within a 0.25-degree radius (approximately 25km). This approach allowed the model to capture local weather patterns beyond individual measurement points.

2. Pressure Gradient Calculation (Gradient Boosting): Computed pressure gradients along latitude and longitude directions to quantify how rapidly atmospheric pressure changes across space. These gradients provide crucial information about potential air movement that drives weather changes.

3. Distance and Bearing Features (Linear Regression): Calculated the distance and directional bearing from each observation point to a central reference point. This established a consistent spatial reference frame that helped capture how temperature varies with distance and direction from central UK.

These spatial features only slightly enhanced model performance by incorporating the contextual information that meteorologists traditionally use when analysing weather patterns.

Model Development and Validation Strategy

Employed a consistent model development and validation strategy across all models:

1. Dataset Splitting: Used a 80/20 train-test split with a consistent random seed (42) to ensure all models were evaluated on identical data subsets.

2. Cross-Validation: For hyperparameter tuning, implemented k-fold cross-validation to ensure models parameters are robust across different data subsets.

3. Final Model Training: After hyperparameter selection, final models were trained on the full training dataset to maximise learning from available data.

This methodical approach allowed for fair comparison between models while maintaining scientific rigor in the evaluation process.

Model Construction Techniques

Developed four distinct regression models, each with different strengths and characteristics:

Gradient Boosting Regression: This ensemble technique builds decision trees sequentially, with each tree correcting errors made by previous trees. Implemented extensive hyperparameter tuning using RandomizedSearchCV to optimise parameters like:

- Number of estimators (100-400)

- Learning rate (0.01-0.1)

- Maximum tree depth (3-7)

- Minimum samples for splitting nodes (2-6)

- Subsample ratios (0.75-0.9)

The tuning process systematically explored these parameter combinations to minimise prediction error while avoiding overfitting.

[Space for other models]

Evaluation Methodology

To comprehensively assess model performance, multiple evaluation metrics were calculated:

1. Root Mean Squared Error (RMSE): Measures the average magnitude of prediction errors in Kelvin, with lower values indicating better performance.

2. R-squared (R²): Quantifies the proportion of temperature variance explained by the model, with values closer to 1.0 indicating stronger predictive power.

3. Mean Absolute Error (MAE): Provides the average absolute difference between predictions and actual temperatures in Kelvin, offering an intuitive measure of prediction accuracy.

Conducted both numerical evaluation using these metrics and visual assessment through scatter plots of predicted versus actual temperatures, residual analysis and feature importance examination to gain comprehensive insights into model performance and behaviour.

Results:

* Training/validation scores of all models
* Improvements from nearby features
* Thorough evaluation on whether models are over- or under-fitting training data
* Thorough justification of hyperparameters (i.e. why grid/random search parameters are applicable)
* Why we’re selecting the main one (need to justify)
* Results from final model on test set
* What these mean in business terms

Discussion:

* Extra things we explored
* Other methods of nearby features and why we didn’t use them
* Issues we had with the models
* What could be added to make the project better
* More powerful machine for faster fine tuning
* More complex neural network architectures (regularisation methods such as dropout/early stopping with callbacks)

Conclusions

References