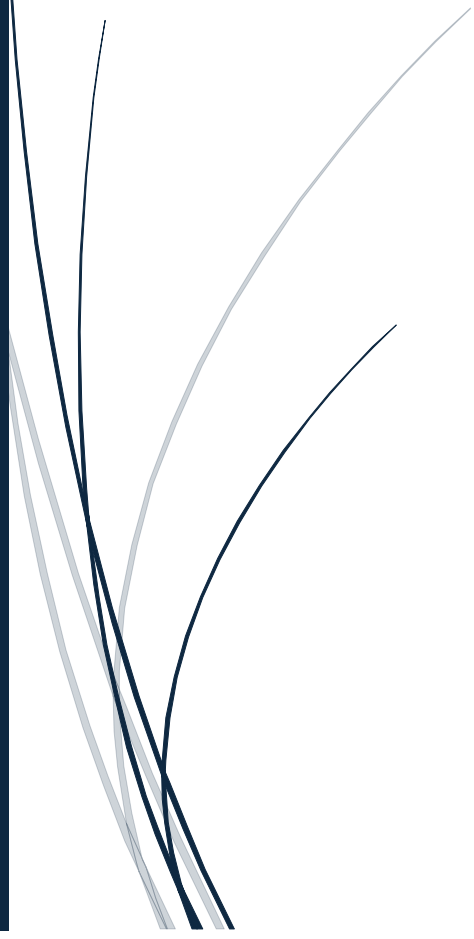
08.07.2025

ML models

For Weather Conditions and Climate Change prediction



Denis Kleptsov
CAREER FOUNDRY DATA ANALYTICS COURSE

Contents

This document outlines the scope, objectives, hypotheses, and plan for the ClimateWins weather prediction project. Its purpose is to establish a shared understanding of the project's goals and to serve as a foundational guide for the analytical work to be performed.

---

1. Project Management Plan

This section details the project's structure, schedule, key outputs, and intended audience.

1.1. Schedule and Milestones

The project is planned for a **4-week duration**, broken down into the following high-level phases:

- **Phase 1: Scoping & Hypothesis (1 Week)**
    - **Milestone:** Finalize and receive sign-off on this Project Scoping Document.
    - **Activities:** Thoroughly analyze the project brief, assess data sources, and develop testable hypotheses.
- **Phase 2: Data Preparation & Exploratory Data Analysis (EDA) (1 Week)**
    - **Milestone:** Deliver a cleaned, analysis-ready dataset and an EDA report.
    - **Activities:** Merge datasets, handle missing values, correct data types, and perform initial statistical and visual analysis to understand trends, distributions, and relationships.
- **Phase 3: Modeling & Validation (2-3 Week)**
    - **Milestone:** Develop and validate at least two predictive models (one simple, one complex).
    - **Activities:** Engineer features, split data for training and testing, build and train supervised learning models, evaluate performance using appropriate metrics, and iterate on model improvements.
- **Phase 4: Final Reporting & Presentation (4 Week)**
    - **Milestone:** Deliver the final presentation deck and technical report.
    - **Activities:** Synthesize all findings, create visualizations, and prepare tailored deliverables for both executive and technical audiences.

1.2. Project Deliverables

1. **Project Scoping Document (This Document):** The foundational plan for the project.
2. **Cleaned & Prepared Dataset:** A single, merged, and analysis-ready dataset in .csv format.
3. **Exploratory Data Analysis (EDA) Report:** A Jupyter Notebook detailing the data cleaning process, key statistical findings, and visualizations of historical weather trends.
4. **Final Presentation Deck:** A concise, high-level presentation for the ClimateWins management team, summarizing the project's key findings, model performance, and strategic recommendations.

5. **Technical Model Report:** A detailed document or notebook appendix containing the final model code, performance metrics (e.g., confusion matrix, precision, recall), feature importance, and methodology for reproducibility.

1.3. Audience Definition

- **Primary Audience: ClimateWins Management & Strategic Team.**
    - **Profile:** Semi-technical, with deep domain knowledge in climate but not necessarily in machine learning algorithms.
    - **Tailoring:** Communication will focus on the "so what" of the analysis. The final presentation will prioritize clear, actionable insights, business value, and strategic implications over complex technical jargon. Model performance will be explained in intuitive terms (e.g., "The model correctly identifies 90% of pleasant days").
- **Secondary Audience: Project Mentor & Future Data Scientists.**
    - **Profile:** Highly technical.
    - **Tailoring:** The Jupyter Notebooks and Technical Model Report will serve this audience, providing complete transparency into the code, methodology, and detailed performance metrics to allow for validation, reproduction, and future extension of the work.

---

2. Hypothesis Development

The following hypotheses have been formulated to guide the research and modeling phases of this project.

1. **Hypothesis (Climate Trend):** Analyzing European temperature data from the last century reveals a statistically significant increasing trend in both annual mean and maximum temperatures, providing quantitative evidence of regional climate change.
2. **Hypothesis (Model Complexity):** If we compare a simple classification model (e.g., Logistic Regression) with a more complex ensemble model (e.g., Random Forest), then the complex model will achieve a higher accuracy in predicting "pleasant" weather days.
3. **Hypothesis (Feature Importance):** If we analyze the features used by a trained predictive model, then temperature and pressure will be the most important predictors for classifying a day as "pleasant."
4. **Primary Research Hypothesis (Predictive Power): If a supervised classification model is trained on historical weather features (e.g., temperature, wind speed, sunshine duration), it can predict whether a future day will be classified as "pleasant" with significantly greater accuracy than a random baseline.**
    - **Independent Variables (IVs):** Historical daily weather metrics for a given station (e.g., BASEL_temp_mean, BASEL_wind_speed, BASEL_sunshine).
    - **Dependent Variable (DV):** The binary classification of whether a day is "pleasant" (e.g., BASEL_pleasant_weather, a value of 1 or 0).

3. Data Foundation & Assessment

This project will be built upon two primary datasets provided by the European Climate Assessment & Data Set project.

3.1. Data Source Summary

Dataset 1: Feature Data

- **File Name:** Dataset-weather-prediction-data.csv
- **Source:** External data from the European Climate Assessment & Data Set project. Trustworthiness is considered **High** due to its origin from a reputable scientific body.
- **Collection Method:** Automated sensor readings from 18 weather stations across Europe.
- **Data Contents:** Contains daily weather observations from as early as the late 1800s to 2022. Variables include DATE, MONTH, and station-specific metrics such as cloud_cover, wind_speed, humidity, pressure, global_radiation, precipitation, snow_depth, sunshine, and temp_mean, temp_min, and temp_max.

Dataset 2: Target Variable Data

- **File Name:** Dataset-Answers-Weather_Predict.csv
- **Source:** External, likely derived from the feature dataset by the course provider (CareerFoundry). Trustworthiness for the project's purpose is considered **Medium**, as the criteria for what constitutes "pleasant" are not defined.
- **Collection Method:** Likely derived programmatically from the feature dataset.
- **Data Contents:** Contains a binary (0/1) indicator, [STATION]_pleasant_weather, for each station for each day, starting from 1960.

3.2. Population Summary

The data represents the daily weather conditions recorded at 18 specific, fixed-point weather stations across Europe. The temporal scope is extensive, but the geographic scope is limited to these points, with a concentration in Western and Central Europe. This is a sample, not a complete representation of all weather across the entire European continent.

3.3. Data Limitations & Bias

- **Geographic Bias:** The station locations are not evenly distributed across Europe, with fewer stations in Southern and Eastern Europe. Model performance may be biased towards the regions with more data.
- **Measurement Changes:** Over 100 years, measurement technology and standards have likely evolved, which could introduce inconsistencies in the data that are not immediately apparent.
- **Missing Data:** A preliminary review reveals missing values across multiple variables and periods, necessitating robust imputation or handling strategies.

- **Undefined Target Variable:** The primary limitation is that the pleasant_weather target variable is a "black box." The business logic used to create this label is unknown, which prevents us from validating its relevance to ClimateWins' specific strategic goals of identifying "favorable" or "dangerous" conditions.

3.4. Relevance and Justification

- **Relevance:** The datasets are highly relevant as they directly provide the necessary features (independent variables) and a pre-labeled outcome (dependent variable) to test the Primary Research Hypothesis.
- **Justification:** Despite the undefined nature of the "pleasant" label, these datasets are the designated starting point for this project. They enable an immediate proof of concept for supervised learning without the initial overhead of defining and engineering a target variable from scratch. The project will proceed by using pleasant_weather as a proxy for "favorable" conditions, with the explicit goal of refining this definition later.

---

4. Data Wishlist

To enhance this analysis and better align it with ClimateWins' mission, the following data would be highly valuable:

1. **Business Rules for "Pleasant Weather":** The specific quantitative thresholds (e.g., temp > X, sunshine > Y) used to create the pleasant_weather label.
    - **Value:** This would allow us to validate the existing target variable and customize it to create a more relevant "dangerous weather" label.
2. **Humidity Data:** Comprehensive relative humidity data for all stations across the entire time period.
    - **Value:** Humidity is a critical factor in how temperature is perceived (heat index). Its inclusion would significantly enhance the model's ability to identify truly hazardous heat conditions.
3. **Labeled Extreme Weather Events:** A historical log of officially declared extreme weather events (e.g., heatwaves, floods, major storms) tied to specific dates and locations in the dataset.
    - **Value:** This would provide an objective, externally validated "ground truth" for training a model to predict dangerous conditions, which is a core long-term goal for ClimateWins.
4. **Socio-Economic Data:** Location-based data on population density, critical infrastructure (hospitals, power grids), and vulnerable populations.
    - **Value:** This would enable a risk-based analysis, helping to prioritize predictions for areas where extreme weather would have the most severe human and economic impacts.

5. Limitations and Ethical Considerations

Limitations

- **Proof-of-Concept Scope:** This project will produce a proof-of-concept model and not a production-ready, real-time weather forecasting system. The model's predictions are based on historical correlations, not a dynamic simulation of atmospheric physics.
- **Resource Constraints:** The project is limited to a single analyst and a 4-week timeline, which restricts the number of models and feature engineering techniques that can be thoroughly explored.

Ethical Considerations

- **Risk of Misinterpretation:**
    - **Issue:** The model's predictions ("pleasant" vs. "not pleasant") could be misinterpreted as safety guarantees. A **false negative** (predicting a "pleasant" day that turns out to be dangerous) poses a significant risk if the results are misused.
    - **Mitigation:** All deliverables will include clear disclaimers stating the experimental nature of the model and that it is **not suitable for public safety decisions**. Communication will focus on the model's probabilities and confidence levels, not deterministic predictions.
- **Algorithmic Bias:**
    - **Issue:** The model may be more accurate for regions with more data, potentially leading to a disparity in predictive quality across Europe. Using these results to direct resources could inadvertently disadvantage underrepresented regions.
    - **Mitigation:** Model performance will be evaluated on a per-station/per-region basis to identify and transparently report any significant disparities. This assessment will be a key component of the final report.
- **Data Privacy:**
    - **Issue:** The provided datasets contain no Personally Identifiable Information (PII).
    - **Mitigation:** The risk is **Low**. No specific mitigation is required for the current data.