08.07.2025

Initial Analysis

of project brief

Denis Kleptsov
CAREER FOUNDRY DATA ANALYTICS COURSE

Contents

1. Deconstruction of the Business Requirements Document (BRD)

Project Motivation:
ClimateWins, a European non-profit, is concerned with the increasing frequency and severity of extreme weather events due to climate change. They lack a dedicated data science team and struggle to synthesize vast amounts of climate data (from sources like NOAA, JMA, and European stations) into actionable insights. The core motivation is to leverage machine learning to better understand and predict the consequences of climate change, specifically focusing on extreme weather, to inform their strategy and potentially create a predictive model for the future.

Project Objective:
The primary objective is to assess the applicability of machine learning for weather prediction and deliver a proposal on the best approach. A key, measurable sub-objective is to build a proof-of-concept supervised learning model that can predict whether a given day's weather conditions will be "favorable" or "dangerous" in mainland Europe, using historical data. The ultimate deliverable is a presentation of these findings to ClimateWins.

Project Scope:

- Analysis of historical weather data from 18 European weather stations (late 1800s to 2022).
- Using Python for all data preparation and modeling.
- Researching and explaining the differences between AI/ML, supervised/unsupervised learning, and linear/nonlinear models.
- Developing a supervised learning model (classification) to predict a binary outcome (e.g., "pleasant" vs. "not pleasant").
- Identifying historical temperature maximums and minimums in the data.
- Assessing ethical considerations specific to the project.
- Creating and delivering a final presentation summarizing the assessment of ML tools for ClimateWins.

Out-of-Scope:

- Analysis of data from outside the provided European Climate Assessment & Data Set (e.g., the mentioned NOAA hurricane and JMA typhoon data).
- Implementation of unsupervised learning models (though the concept must be understood).
- Building a real-time prediction pipeline.
- Developing a public-facing application or API.

Key Stakeholders:

- **ClimateWins:** The non-profit organization and primary project sponsor. They are the main consumer of the final recommendations.

- **Project Mentor:** Responsible for reviewing deliverables, providing feedback, and guiding the project's analytical direction.
- **Data Analyst (The Project Lead):** Responsible for executing all analytical tasks, from data preparation to modeling and presentation.

Known Assumptions

- It is assumed that historical weather patterns in the dataset are sufficiently indicative of future conditions to build a useful predictive model.
- It is assumed that a binary classification of a day's weather (e.g., "favorable" vs. "dangerous") can be meaningfully defined and is useful for the organization's strategic goals.
- The provided dataset is of sufficient quality and completeness for machine learning.

Constraints:

- **Financial/Personnel:** ClimateWins is a non-profit with limited funds and no dedicated data science or engineering team. The project must be executed by a single analyst.
- **Technical:** The project is to be completed using Python and its associated libraries.
- **Data:** The analysis is constrained to the provided dataset from 18 European weather stations.

2. Foundational Questions for Project Design

2.1. **Clarifying Questions:**

- What is the specific, operational definition of a "favorable" or "pleasant" day versus a "dangerous" one?
- Who is the ultimate end-user for these predictions (e.g., internal strategists, policymakers, the general public), and what specific decisions will these predictions inform?
- What are the key performance indicators for a successful model (e.g., prediction accuracy, minimizing false negatives, interpretability)?

2.2. **Funneling Questions (based on Clarifying Question #1):**

- **Which specific features** in the dataset (e.g., maximum temperature, wind speed, snow depth, global radiation) should be combined to create the "favorable"/"dangerous" label?
- Are there **pre-defined, quantitative thresholds** that ClimateWins considers dangerous (e.g., temperature > 35°C, wind speed > 60 km/h), or do we need to establish these thresholds as part of the analysis?
- Should this definition be **static across all of Europe**, or should it be dynamic and adapt to different geographic locations and seasons (e.g., 25°C might be pleasant in Sweden but cool in Spain)?

2.3. **Privacy and Ethics Questions:**

- What is the ethical responsibility of ClimateWins if the model produces a "false negative" (i.e., predicts a safe day that turns out to be dangerous), and how should the model's decision threshold be tuned to mitigate this specific risk?
- How will we address and communicate the inherent uncertainty and potential for error in our predictions to ensure they are not misinterpreted as infallible forecasts?
- Could the model inadvertently create biased outcomes, for example, by performing poorly in regions with historically sparser or less reliable data, leading to a disparity in protection for different communities?

3. Initial Hypothesis and Data Wishlist

A. Initial Hypothesis Formulation:
If we engineer a target variable that classifies each day as "pleasant" or "not pleasant" based on a combination of temperature, wind, and precipitation thresholds, **then** a supervised machine learning model (such as a Logistic Regression or Random Forest) trained on historical weather features can predict this classification for a future day with an accuracy significantly greater than a random baseline.

B. Preliminary Data Wishlist:
This list includes features available in the dataset and ideal features to enhance the analysis.

- **Core Data (from provided dataset):**
    - Date: For time-series analysis and tracking seasonal patterns.
    - Station_ID/Location: To account for regional climate differences.
    - TG: Mean Temperature (°C).
    - TX: Maximum Temperature (°C).
    - TN: Minimum Temperature (°C).
    - FG: Mean Wind Speed.
    - RR: Precipitation amount.
    - SD: Snow Depth.
    - Q: Global Radiation (sunshine).

- **Ideal Data (Wishlist for model improvement):**
    - **Humidity Data:** Relative humidity is a critical factor in how temperature is perceived (e.g., heat index).
    - **Atmospheric Pressure Data:** Changes in pressure are strong indicators of upcoming weather shifts.
    - **Labeled Extreme Events:** A historical log of confirmed extreme weather events (e.g., officially declared heatwaves, floods, major storms) for specific dates and locations to validate our definition of a "dangerous" day.

4. Preliminary Audience and Deliverable Assessment

A. Audience Definition:
The primary audience is the **management and strategic team at ClimateWins**, as well as the **project mentor**. This audience is intelligent and domain-knowledgeable (in climate) but should be considered **semi-technical**. They need to understand the project's conclusions, the model's capabilities and limitations, and the strategic implications, without necessarily needing a deep dive into the complex algorithms. The presentation must prioritize clarity, business value, and actionable recommendations over raw technical details.

B. Suggested Deliverable:
A **formal presentation (e.g., PowerPoint or Google Slides) combined with a supplementary, well-documented Jupyter Notebook.**

Justification: This two-part deliverable serves both the primary and secondary needs of the audience.

- The **presentation** will convey the high-level story, including the problem, methodology, key findings (e.g., visualizations of temperature trends), model performance in layman's terms, and a clear recommendation on the best path forward for ClimateWins.
- The **Jupyter Notebook** will act as a technical appendix, providing complete transparency of the project's workflow, code, and detailed analysis for the mentor or any future technical person who wants to validate, reproduce, or build upon the work.