Predictive Analysis of Temperature Using NOAA Station Data

Milestone: Project Proposal Second Submission

Student 1: Tim Terry

414-534-5552 (Tel of Student 1)

terry.ti@northeastern.edu

Percentage of Effort Contributed by Student 1: 100%

Signature of Student 1: Tim Terry

Submission Date: 02/06/2025

PROBLEM SETTING:

Accurate temperature predictions have significant implications for agriculture, energy, public health, natural disaster preparation, and public policy. By analyzing historical data, identifying patterns, and developing predictive models, we can improve our understanding of temperature variations and patterns while enhancing decision-making capabilities. But despite advancements in meteorological modeling, temperature sensor capabilities, increases in sensor density and geographic dispersal, achieving high accuracy in temperature predictions remains a challenge. Surface temperature is influenced by a number of factors including solar radiation, cloud cover, humidity, and wind patterns. Overall, weather itself is chaotic and small changes can lead to large forecasting errors. In terms of data collection, weather stations provide limited spatial coverage, despite increases in station density, and models often presume temperature readings and changes are positively correlated across stations within wide distance boundaries. In addition, station data is often incomplete and contain gaps in recorded observations that may adversely affect model accuracies. With these limitations, forecasting models themselves are limited and may not sufficiently capture all relevant processes accurately.

PROBLEM DEFINITION:

This project aims to determine the pre-processing activities required to develop a predictive model for temperatures across the United States using historical data from the National Oceanic and Atmospheric Administration (NOAA). It is acknowledged that NOAA data is comprehensive. However, the data requires extensive pre-processing before it can be effectively utilized for model training. This pre-processing is critical to properly handle data quality issues, geospatial variations, measurement inconsistencies, and preparing the data for optimal model performance.

Specifically, this project seeks to address the following challenges related to pre-processing NOAA temperature data:

- Data Acquisition and Integration: NOAA provides various datasets with varying spatial and temporal resolutions, data formats, and reporting frequencies. This project will focus on the Global Historical Climatology Network (GHCN) which is an integrated database of daily climate summaries from land surface stations across the globe. GHCN contains records from over 100,000 stations in 180 countries and territories. The GHCN datasets provide numerous daily variables, including maximum and minimum temperature, total daily precipitation, snowfall, and snow depth.
- Data Cleaning and Quality Control: NOAA data may contain missing values, outliers, erroneous readings, and inconsistencies due to instrument malfunctions, data entry errors, or other factors. This project will investigate appropriate techniques for identifying and handling these data quality issues. These techniques may include missing value imputation, outlier detection and removal, and data consistency checks.
- Feature Engineering and Selection: Daily raw temperature data may not be the most informative input for a predictive model. This project will explore the creation of new features derived from the raw data, such as temporal features, geospatial features, derived features (moving averages), and feature selection.

- Data Transformation and Normalization: Many machine learning models perform best when the input data is scaled or normalized. This project will investigate appropriate data transformation techniques, such as standardization, min-max scaling, or other methods, to ensure that the data is within a suitable range for the chosen model. The project will also consider transformations to address potential skewness or other distributional characteristics of the data.

DATA SOURCES:

NOAA Global Historical Climatology Network (GHCN): The Global Historical Climatology Network daily (GHCNd) is an integrated database of daily climate summaries from land surface stations across the globe. GHCNd is made up of daily climate records from numerous sources that have been integrated and subjected to a common suite of quality assurance reviews.

- ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/
- ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-countries.txt
- ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-inventory.txt
- ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt
- ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-states.txt

DATA DESCRIPTION:

- ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt

FORMAT OF "ghcnd-countries.txt"

Variable	Columns	Type
CODE	1-2	Character
NAME	4-50	Character

FORMAT OF "ghcnd-inventory.txt"

Variable	Columns	Type
ID	1-11	Character
LATITUDE	13-20	Real
LONGITUDE	22-30	Real
ELEMENT	32-35	Character
FIRSTYEAR	37-40	Integer
LASTYEAR	42-45	Integer

FORMAT OF "ghcnd-stations.txt"

Variable	Columns	Type
ID	1-11	Character
LATITUDE	13-20	Real
LONGITUDE	22-30	Real
ELEVATION	32-37	Real
STATE	39-40	Character
NAME	42-71	Character
GSN FLAG	73-75	Character
HCN FLAG	77-79	Character
WMO ID	81-85	Character

FORMAT OF "ghcnd-states.txt"

Variable	Columns	Type
CODE	1-2	Character
NAME	4-50	Character

FORMAT OF "yyyy.csv.gz" (Daily Station Observations)

Variable	Length	Type
ID	11	Character
YMD	8	Character
ELEMENT	4	Character
DATA VALUE	5	Character
M-FLAG	1	Character
Q-FLAG	1	Character
S-FLAG	1	Character
OBS-TIME	4	Character

METHODOLOGY:

This project will utilize the Cross-Industry Standard Process for Data Mining (CRISP-DM) as its framework due to its comprehensive and iterative nature, which is appropriate for the complex tasks required in the development of a temperature prediction model using the NOAA GHCN data. CRISP-DM's structured and iterative approach provides an excellent roadmap for navigating the project's various stages. Specifically, CRISP-DM's emphasis on data understanding and preparation is crucial for handling the inherent complexities and potential inconsistencies within NOAA's diverse datasets.

Data Mining Objectives

- **Data Acquisition and Preparation:**
    - Obtain daily historical temperature data from NOAA stations, including actual readings or observations related to temperature, precipitation, snow, and other measurement metrics.

    - Obtain NOAA station data and supporting tables which include geolocation information such as latitude, longitude, elevation, country and state, status, and other location specific information.

- **Data Cleaning and Preprocessing:**
    - Filter station data for United States. Stations for NOAA record observations across 180 countries and territories.
    - Filter station data for temperature recording. Stations for NOAA collect information on temperature, precipitation, snowfall, snow depth, and other variables unrelated to this analysis project.
    - Handle missing values using appropriate imputation techniques (e.g., mean imputation, interpolation).
    - Identify and address outliers using robust statistical methods. NOAA uses -9999 measurement values for non-reported data.
    - Transform variables as needed (e.g., normalization, standardization). The base data records temperatures in Celsius and elevation in meters. It may be reasonable to convert these values to Fahrenheit and feet for easier interpretation for a US based audience.
- **Exploratory Data Analysis:**
    - Conduct statistical analysis to assess data quality and characteristics:
        - **Normality:** Test for normality of temperature distributions using appropriate statistical tests (e.g., Shapiro-Wilk test).
        - **Variance:** Analyze the variance of temperature data across different regions, seasons, and time periods.
        - **Stationarity:** Investigate the stationarity of time series data to determine if statistical properties remain constant over time.
        - **Correlations:** Determine correlations of temperature observations across stations within specific distance parameters.
    - Visualize data using appropriate techniques (e.g., histograms, box plots, time series plots, correlation matrices) to identify trends, patterns, and anomalies.
- **Model Exploration:**
    - **Regression Models (Linear, Polynomial, Multiple Linear):** Identify relationships between temperature and various predictor variables (latitude, longitude, elevation, time of year, etc.).

- **ARIMA (Autoregressive Integrated Moving Average):** A time series model that accounts for trends, seasonality, and autocorrelation in the data. Considered suitable for capturing temporal patterns in temperature data at a specific location.
- **Spatial Statistics/Geostatistics (Kriging, Gaussian Processes):** These methods specifically account for spatial correlation. Kriging, for example, interpolates temperature values based on the values at nearby locations, considering the spatial structure of the data. This can be useful where station coverage is sparse and predictive power is decreased due to a reduced set of observations.

- **Model Evaluation:**
  - **Mean Absolute Error (MAE):** The average absolute difference between the predicted and actual temperatures. Lower MAE indicates better accuracy. The MAE is less sensitive to outliers than RMSE.
  - **Root Mean Squared Error (RMSE):** The square root of the average squared difference between predicted and actual temperatures. Lower RMSE indicates better accuracy. More sensitive to outliers than MAE. Penalizes larger errors more heavily.
  - **AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion):** These metrics are used for model selection, particularly in ARIMA models. They help balance model fit with model complexity. Lower AIC or BIC values indicate a better model. They penalize models with too many parameters.