Machine Learning with Python

Meghan McGrath

June 5th, 2024

Objective

Use machine learning to help predict the consequences of climate change while working as a data analyst at ClimateWins, a European nonprofit organization.

?

Can machine learning be used to predict whether weather conditions will be favorable on a certain day? (If so, it could also be possible to predict danger.)

?

Historically, what have the maximums and minimums in temperature been?

?

ClimateWins has heard of ethical concerns surrounding machine learning and AI. Are there any concerns specific to this project?

Hypotheses

- Machine learning models can accurately predict future weather patterns in Europe by analyzing historical weather data from the past century.
    - Training models on historical data to identify patterns/trends which allows for better forecasting of future conditions.
- Machine learning can identify significant changes in weather patterns, correlating these changes with rising global temperatures & other indicators of climate change.
    - Comparing recent data with historical trends so machine learning can highlight deviations and provide insights into the impact of climate change.
- Machine learning models can accurately predict daily weather conditions, including temperature, wind speed, and precipitation
    - Training on historical weather data so machine learning can develop models to provide reliable daily forecasts.
- ClimateWins can use existing weather data and machine learning tools to make significant advancements in climate prediction and fight climate change.

Data Source

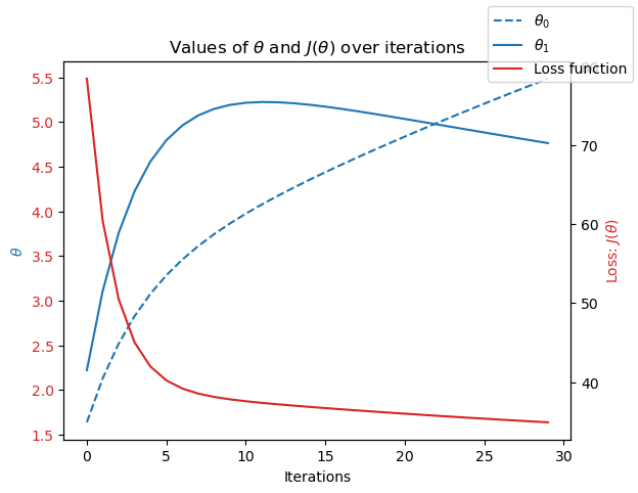
- Data set based on weather observations from 18 different European weather stations from the late 1800s to 2022.
- Includes daily recordings for temperature, wind speed, snow, global radiation, and more.
- Data set collected by: European Climate Assessment & Data Set Project

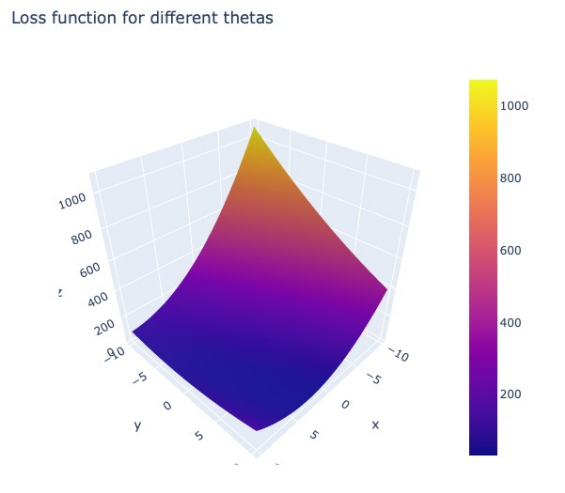
Bias & Accuracy

- **Measurement bias** – no confirmation of measurement standardization across each station (i.e. varying equipment) & more accuracy in current vs. past data as technology has improved
- **Human bias** – stations interpreting data in different ways
- **Selection bias** - are the 18 stations a good representation of Europe/the world?
- **Cultural bias** – cultural beliefs/values can shape how people perceive and respond to climate change. A machine learning algorithm that reflects certain European cultural perspectives may not be a good representation of the rest of the world with different cultural norms.

Gradient Descent Optimizations

Next step: Run optimizations (for 3 stations across 3 different years) to get the maximum and minimums for each station and measure how well a model's predictions match the actual data.





Station: Madrid
Year: 2018

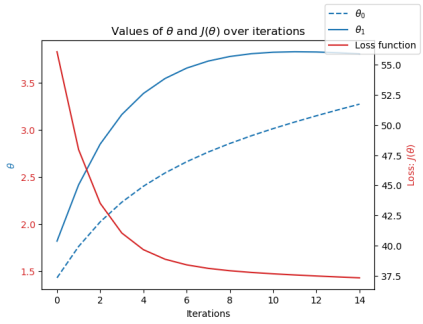
1st Optimization:
num_iterations=30
theta_init=np.array([[1],[1]])
alpha=0.05

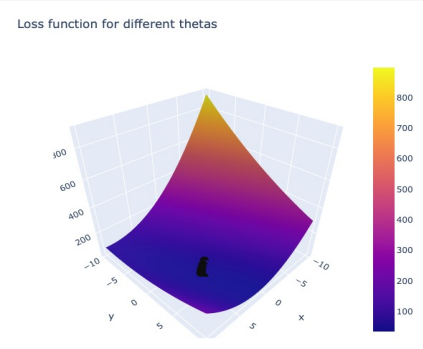
2nd Optimization:
num_iterations=30
theta_init=np.array([[4],[6]])
alpha= 0.05

Goal: For the loss profile, move the red line to trend toward 0 to get the smallest possible error and improve accuracy. The 3D "canyon" shows the path to the lowest loss per the optimization inputs.

Gradient Descent Optimizations

Station: Budapest
Year: 2017



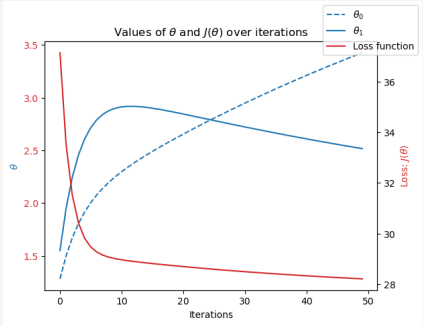


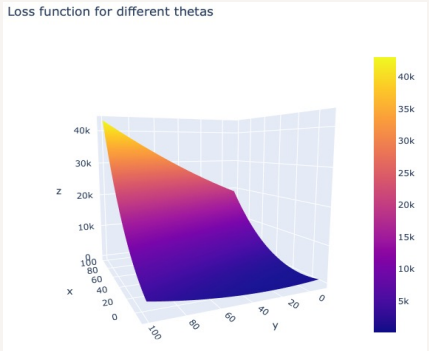
1st Optimization:
num_iterations=15
theta_init=np.array([[1],[1]])
alpha=0.05

2nd Optimization:
num_iterations=50
theta_init=np.array([[-1],[-1]])
alpha= 0.05

3rd Optimization:
num_iterations=50
theta_init=np.array([[-1],[-1]]
alpha= 0.01

Station: Budapest
Year: 1997





1st Optimization:
num_iterations=50
theta_init=np.array([[1],[1]])
alpha=0.05

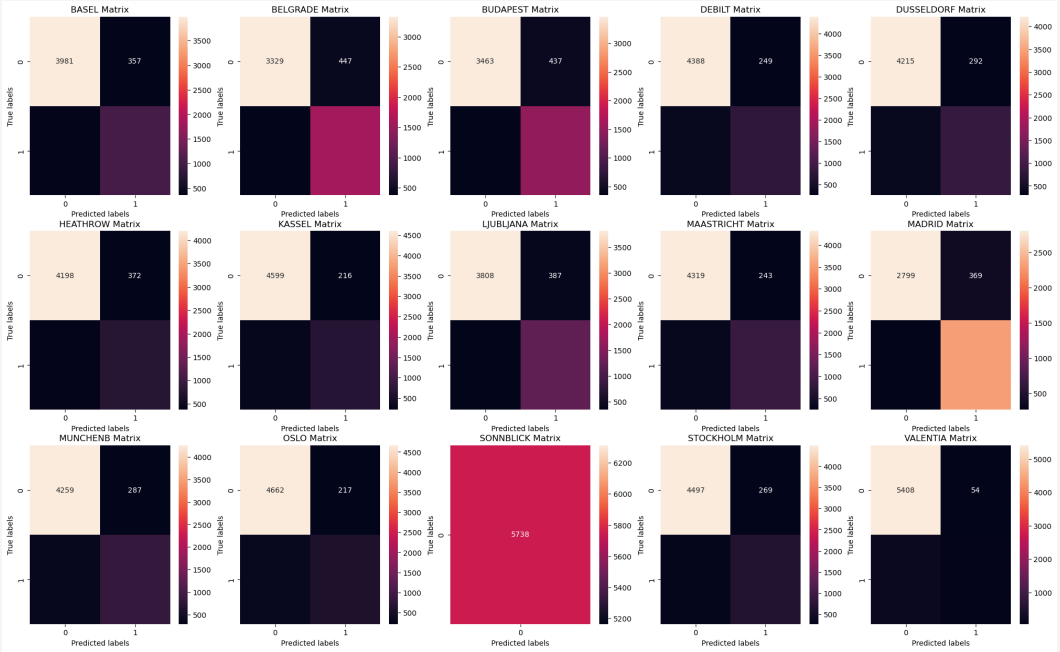
2nd Optimization:
num_iterations=50
theta_init=np.array([[0],[0]])
alpha= 0.05

KNN Model – Confusion Matrix

Accuracy score: 52%

What is a confusion matrix?

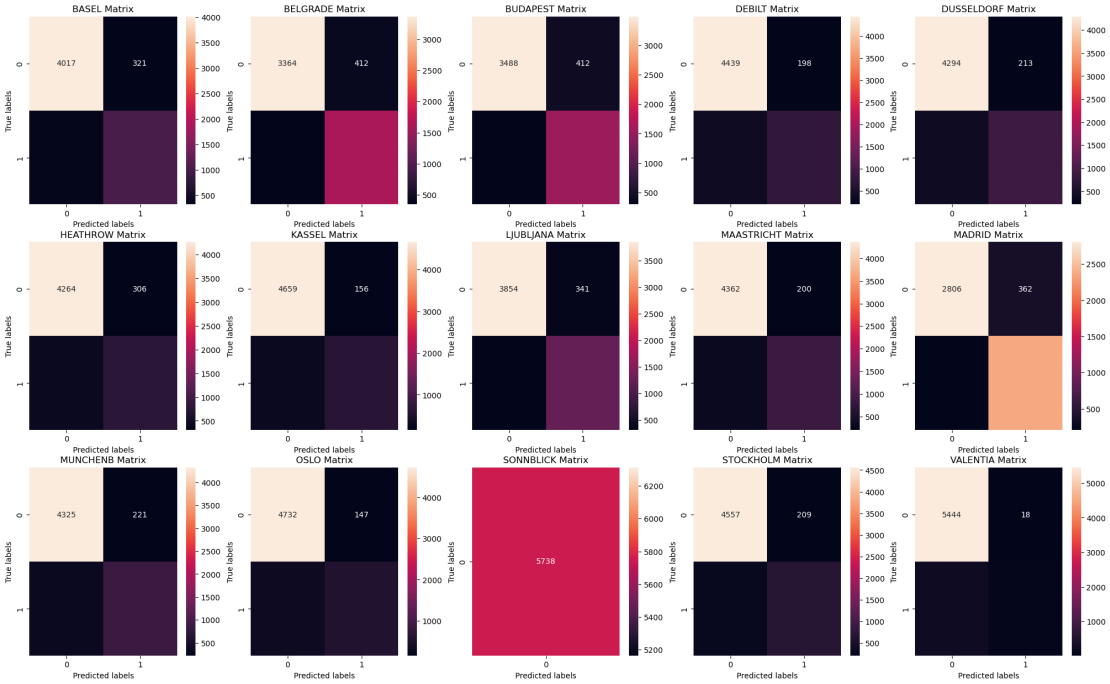
A confusion matrix is a tool used in machine learning to represent the accuracy of a model – is this example, the accuracy of the algorithm for each station.. It shows the **actual vs. predicted values** and how many of the variables have been classified or misclassified by the model in the four quadrants (i.,e. true positives, true negatives, false positives, and false negatives).



#Run the model with neighbors equal to 1 to 4, test the accuracy
k_range = np.arange(1,4)

KNN Model – Confusion Matrix

Accuracy score: 58%



#Run the model with neighbors equal to 1 to 10, test the accuracy
k_range = np.arange(1,10)

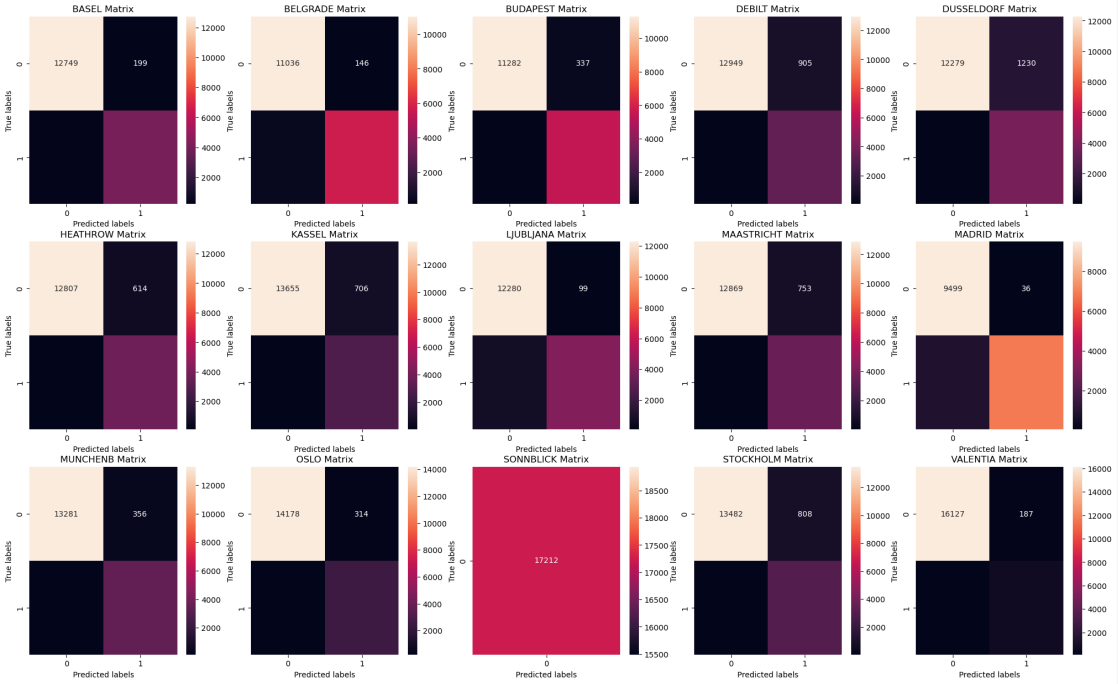
The KNN model is **not** an accurate predictor as most quadrants in the confusion matrix show **no relationship.**
For each station, only the top left-hand corner (the variables 0 and 0) shows the occurrences when the variables were labeled correctly.

Features that may contribute to overall inaccuracy:

- Overfitting – the algorithm may conform to one aspect and not represent the entire data set.
- Some weather stations may have more unique/volatile weather conditions that could impact the “outliers”.
- Stations more exposed to environmental factors = more susceptible to inaccurately collected data (ex. Stockholm is situated on islands & Oslo is surrounded by forests/mountains)

ANN Model – Confusion Matrix

Accuracy score: 82%



The ANN model is the better predictor in comparison to the KNN model. The number of correct labels is much higher across the stations.

Key Takeaways

- Neither the confusion matrixes for the KNN or ANN model are fully accurate by station.
    - Example of overfitting in Sonnblick - overfitting may be caused by the model learning not only the underlying patterns in the training data but also the noise/random fluctuations. Sonnblick is a mountainous area in Austria that most likely incurs large weather fluctuations.
- Accuracy of a model is influenced by the quality (accuracy and completeness) and relevance of the datasets. Relevant features are going to provide stronger predictive power and reduce noise. It's important to use features that are strongly correlated with the target variable of the model – in this example, we have access to each stations cloud cover, wind speed, humidity and precipitation to name a few.
    - Next steps: Trim down these features, as making the model simpler could lower the risk of overfitting. It's important to use relevant features so the model can learn the underlying patterns rather than false correlations, leading to greater accuracy in real-world applications.
- Overall, the **ANN model should be recommended** over the KNN model to ClimateWins as it is more successful in predicting weather station data.