CIS 545 Project Proposal

Spring 2023

Group members & Duties
Hongliang Chen - EDA
Mingxin Xue - Visualization for EDA
Yinuo Zhao - Modeling

**Note: this distribution of responsibilities may change in the future as the project progresses in more detail.

Data Source
Source link:
https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data
Quick description:
Our chosen dataset is the earth surface temperature data, which contains general temperature changes on the earth, including land and ocean, from January 1750 to December 2015. The measurements include the average temperature, minimum temperature, and maximum temperature. They also contain monthly average temperature data for various places around the world, which includes information such as location, date, and temperature measurements, as well as the average temperature anomaly and temperature measurement uncertainty.

Project Plans
Our research will focus on exploring the long-term temperature trends, identifying regions that are especially vulnerable to climate change, and assessing the impact of climate change on various parts of the world based on the geographic coordinates. Our ultimate goal is to develop a model for predicting future temperature trends.
Currently we are considering using the time series modeling moving average as our baseline model and then apply we might try some simple models like linear regression and ensemble models like boosting tree. The time series modeling will be particularly useful when analyzing trends and patterns in climate change and Earth surface temperature data over time. Also spatial modeling can be useful here, as it is used to represent data in a geographic context.

Value Propositions
We are motivated by the facts that this model can tell us whether the climate will be a threat to human beings and can help people to develop effective strategies to mitigate the impacts of climate change, especially based on different locations on Earth. Also, policymakers can use this data to develop policies that reduce greenhouse gas emissions and promote sustainable development practices.

Anticipated Obstacles & Challenges

The earth's surface temperature data may contain gaps or missing values, which could make it challenging to identify trends over time. We need to think about proper methods to deal with this missing data for time series modeling. Also for time series data, when doing cross validation, we need to carefully split the training and test sets by time slices, not by a random split. We may also meet data ethics problems like information leakage if we decide to create some additional lagging and rolling features. Model selection will involve the use of complex statistical models, such as regression models or machine learning models. Due to the complexity of this dataset, selecting the most suitable model could be a challenge.