Feeling the heat: Prediction of rise in average surface temperature using human-economic activity alternative data

Project status Report #001

Deepti Saravanan (ds6812), Jahnavi Swetha Pothineni (jp5867)

Goal

To predict anomalies in average surface temperature in different countries using various human activity stats as factors.

Context: Greenhouse gas emissions are the most significant contributor to the climate change. Understanding the human factors that cause greenhouse gas emissions is imperative to take future action to combat climate change.

CRISP-DM Ongoing Steps

Business Requirements

- The objective of the project is finalized.
- Literature review was performed to understand the domain and gather background information.
- Decided on what resources and features to focus on – Primary focus on **human activities** that affect climate change. Data would be collected from multiple sources such as Our World in Data, Atlas big, Berkeley Earth.
- A project plan was drafted.
- Decided on what tools and platforms to use – **Python and relevant statistical and visualization tools in Python**
- Criteria for success set – **Establish and explain relationships between different human activities and climate change on a global scale**.

Data Assessment

- Data is extracted from multiple sources such as Our World in Data, Atlas big, Berkeley Earth. [*Ongoing: This is an evolving list where more human activities would be added and analyzed*]. [Issue faced]
- Ensured extracting only high-quality data by being very selective on data sources.
- Data exploration was performed to understand the structure of the data, the distribution of values, and its relevance to our problem statement.

- Understanding of the data was done with a detailed note on concerns and issues to tackle – Extracted data was **not well-structured to use, had a lot of missing values, and needs heavy data processing and feature selection**.
- Data definitions were created explaining what the data schema is and what each feature means.

Data Preparation

- Extracted data were **converted into a useable structured table format and saved**.
- **Data Cleaning** was performed – Missing values treatment, converting features to the right data types, selecting the timeframe relevant to our work, splitting and renaming columns, and grouping features by year wherever necessary.
- **Data Fusion** was performed year-wise and country-wise.
- **Feature Selection** was performed – Features with **very low variance** (up to 95% similarity in the values) were dropped. **Pearson Correlation test was performed and visualized using a heatmap** to identify highly correlated features to avoid data redundancy.

Next Steps

Modeling

- **Model selection** would be performed keeping in mind the inferences of previous work from the literature review part. **Assumptions for the model** would be drafted if required.
- **Selected models would be built** in Python. Models would be **run and assessed**.

Questions

- We have selected a timeframe of 1950-2020 to scrape data related to human activities of interest.
- For many such activities, historical data is only available from the 1990s.
- Imputation of data works well when there are missing values in between. But in our case, there is a whole chunk of data missing.
- There was a thought of training models to learn with recent data to predict the past. But there is a high risk of error propagation that might heavily affect the performance of the main predictive model.
- There is also an important aspect to consider – the pattern of human activities back in the 1950s is very different from the recent years. Given that we focus on predicting which human activities highly affect climate change now, considering data from the 1950s might become irrelevant. Hence, there is a trade-off between the amount of data and data relevance.
- Suggestions and thoughts on this, and also other reliable data sources to mine from would be of immense help.