TRAVEL RECOMMENDATION

CMPT 353: Computational Data Science

Authors:
Arvin Bayat Manesh, aba191
Keishi Hazel Allam, kha125
Jonathan Yang, sya171

GitHub: https://github.sfu.ca/aba191/CMPT-353—Group-Project.git

December 4, 2023

Contents

1 Introduction

1.1 Problem

In a world where wanderlust meets data science, we present the TravelWeather Recommendation project. Leveraging extensive weather data from the Global Historical Climatology Network (GHCN), our team has embarked on a journey to transform the way we explore and plan our adventures.

At the heart of our project lies a robust framework of modeling and data training. We've analyzed and harnessed the wealth of information within GHCN, employing techniques to extract meaningful patterns and correlations. This amalgation of weather and machine learning forms the backbone of our travel recomendation system, ensuring that your next destination aligns seamlessly with your preferences.

As avid travellers ourselves, we understand the important role weather plays in shaping travel experiences. Whether you seek sun-drenched beaches, crisp mountain air, or the charm of a rainy day in a quaint town, our TravelWeather Recommendation is poised to be your trusted companion. Gone are the days of unpredictable weather surprises- our system empowers you to make informed decisions, allowing you to curate your idea travel iterinary.

Welcome to TravelWeather Recommendation, where very journey begins with the perfect forecast.

1.2 Background

The K-Nearest Neighbours model predicted the city name of given GHCN weather data (accuracy of around 66%) in the previous course exercise. The model can return different cities that are kilometres apart with slightly different data. To visualize this, please check out Figure 1. (The interactive and animated dashboard made by Tableau can be found here: Temperature Comparison and Precipitation Comparison) We can conclude from this that there are patterns in weather features in different cities. TravelWeather Recommendation is a method of applying and implementing the concept of recommending locations to visit based on a given month and weather feature preferences.

1.3 Data Acquisition

The Global Historical Climatology Network data (GHCNd), is a comprehensive database, encompasses daily variables such as maximum and minimum temperature, total daily precipitation, snowfall, and snow depth. Notably, this data set comprises records from over 100,000 stations across 180 countries and territories, spanning from 1832 to 2022, reflecting a global perspective on climate patterns.

In establishing the foundation for our TravelWeather Recommendation project, we honed in on the extensive dataset from the Global Historical Climatology Network (GHCN). Our analysis centered on data spanning from 1997 to 2022, primarily focusing on climate records from Canada and the United States. To manage the sheer volume of information, we strategically extracted data in two-year intervals, adopting a strategic approach to data extraction. Subsequently, our focus shifted to data cleaning, a critical step in ensuring the integrity of the data set for subsequent training and modeling phases.

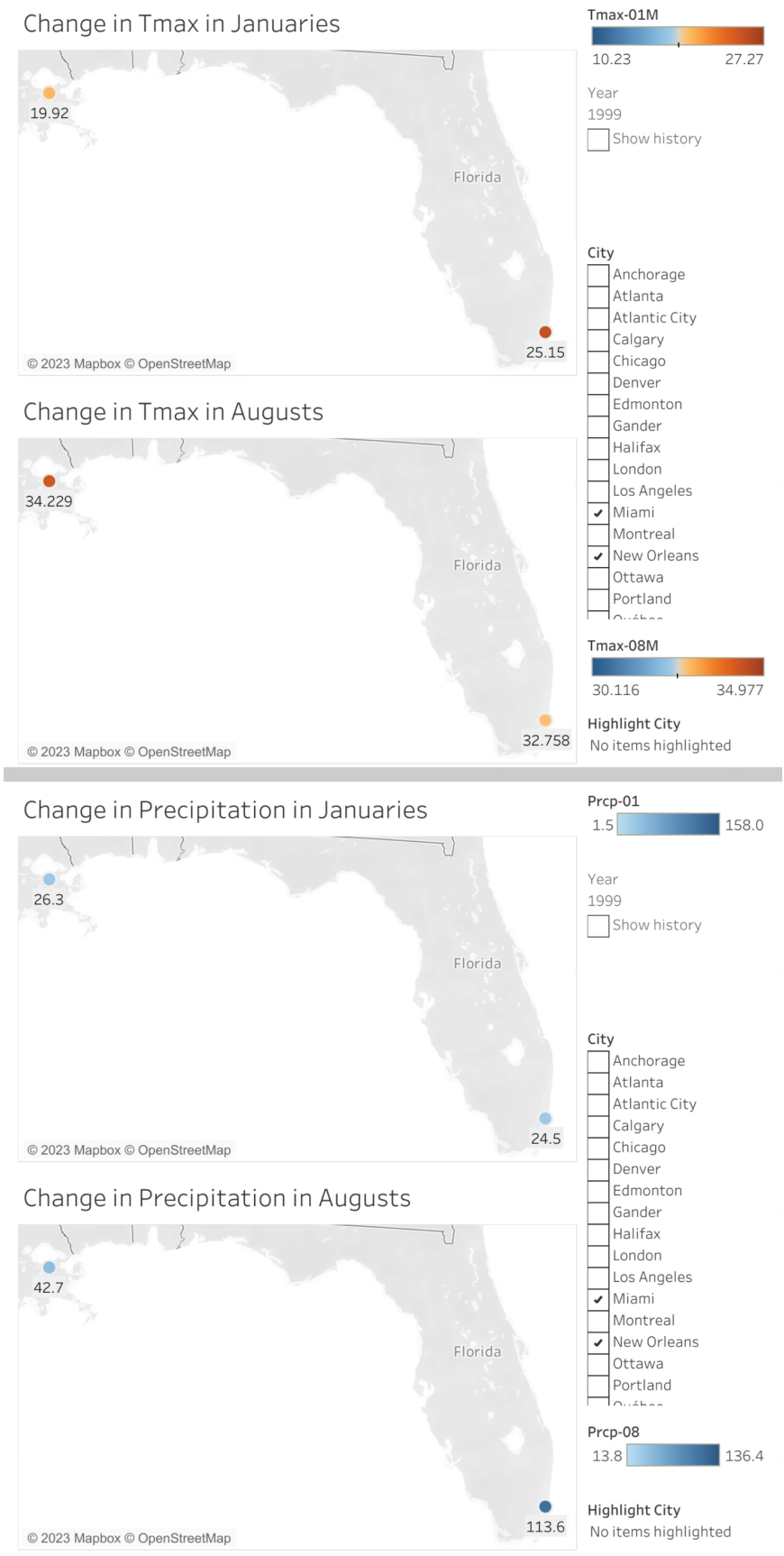

Figure 1: Comparison of 2 cities with similar weather features

2 Methodology

2.1 Data Cleaning

2.1.1 Data Extraction and Aggregation

In the initial phase of data cleaning, the primary objective is to extract GHCN data from the SFU cluster. This process results in the generation of 32 JSON file that has been compressed using gzip, with each set of 32 encapsulating 2 years of data. The next step involves aggregating these individual files into a cohesive and comprehensive CSV file. This streamlined representation simplifies the handling of the extracted information, forming the foundational dataset for subsequent cleaning and analysis.

The extraction script, adapted from the provided instructions on the WeatherData page in Coursys, leverages Spark for efficient processing on the SFU cluster. The output comprises 32 json.gz files, each containing a subset of the GHCN data, facilitating the storage and transfer of information. Subsequently, the data is transferred to the local computer for efficient cleaning and aggregation along with the data extracted from other years. This step is necessary as we extract data in the SFU cluster in two-year intervals, considering the time-intensive nature of extracting data from the entire range of 1997 to 2022 at once.

2.1.2 Cleaning and Formatting

Following data extraction, the second phase involves a cleaning and formatting process for the combined data. This phase is crucial for ensuring the integrity and consistency of the dataset, preparing it for subsequent analyses.

The cleaning script filters the data, handling various variables such as TMAX, TMIN, PRCP, SNOW, and SNWD. It systematically drops rows with blank 'state' entries since we are focusing on Canada and the US only, arranging the data by 'state', 'year', and 'month'. Moreover, the cleaning script includes pivotal transformations to achieve the desired format. The script pivots the data to organize it by 'state' and 'year', with columns representing each month for each variable (TMAX, TMIN, PRCP, SNOW, and SNWD). Subsequently, the multi-level columns are flattened for ease of interpretation and analysis. The resulting output serves as a refined representation, ready for further processing in subsequent stages of the project.

As a post-cleaning step, it is recommended to rename the output CSV file to reflect the corresponding year range (e.g., '2015-2016.csv') and store it in the 'combined clean data' folder for comprehensive organization. The output for each year range is available in the 'combined clean data' folder on GitHub for reference.

2.1.3 Finalization and Data Combination

The concluding stage of the data processing pipeline focuses on the finalization and combination of all cleaned data files from different year ranges. As the extraction process occurs biennially, the script produces an aggregated CSV file, encompassing a comprehensive dataset spanning from 1997 to 2022.

In this step, the script integrates state names corresponding to state codes, further enriching the dataset. The subsequent script adds a new column 'state name' to the CSV

dataframe, utilizing information from the 'ghcnd-states.txt' file obtained from the GHCN website. This file contains a list of state codes and corresponding state names.

As part of the final touch-ups, the script removes rows where 'state' is equal to 'UM,' ensuring data integrity for subsequent analyses. The resulting dataset, stands as the fully cleaned, formatted, and enriched dataset, ready for advanced analyses and utilization in subsequent stages of the project. This file serves as the foundational resource for generating travel weather recommendations based on historical climate patterns.

2.2 Model and Training Data

2.2.1 Command Line Interface(CLI)

After extracting and cleaning the data from the computer cluster, data is read from 01-Data-Cleaning/weather-final.csv to make a recommendation for the user's next holiday location.

To do so, there is a Command Line Interface (CLI) which prompts the user for their desired season (spring, summer, fall, winter), temperature unit (Celsius, Fahrenheit), high temperature range, low temperature range, precipitation in mm, snowfall in mm, and snow depth in mm, user's location longitude, user's location latitude. All of these inputs are validated, and the user is asked to input a correct value if they provide an input that does not make sense (e.g., if the provided input is a string that is not convertible to an integer for a certain feature which requires an integer).

The current user's location based on longitude and latitude is used to let the user know the capital of the recommended province/state as well as the code of that region.

2.2.2 Training Data

After gathering the inputs from the user, one of four models (spring, summer, fall, winter) is trained based on the user's desired season. We increased the accuracy of our models by making a VotingClassifier from the sickit-learn library. This VotingClassifier is composed of three distinct classifiers: GaussianNB, SVC (Support Vector Classifier), and a RandomForestClassifier. In order to predict different states/provinces, the values for the maximum temperature and the minimum temperature given to the corresponding model are randomized. These random numbers for minimum and maximum temperatures are chosen from the range which the user has specified.

After predicting the state name value, the user is informed of the three recommended states/provinces based on the provided answers, the region code and the capital of that region.

2.3 Location Feature

Our program includes a location feature that sorts recommendation results from closest distance to farthest distance to the user. As parameters, the following information was passed: state1, state2, state3, and the user's latitude/longitude. Using a helper function modified from the haversine formula, the feature function will return a dataframe of states ranked by their distance from the user.

3 Result

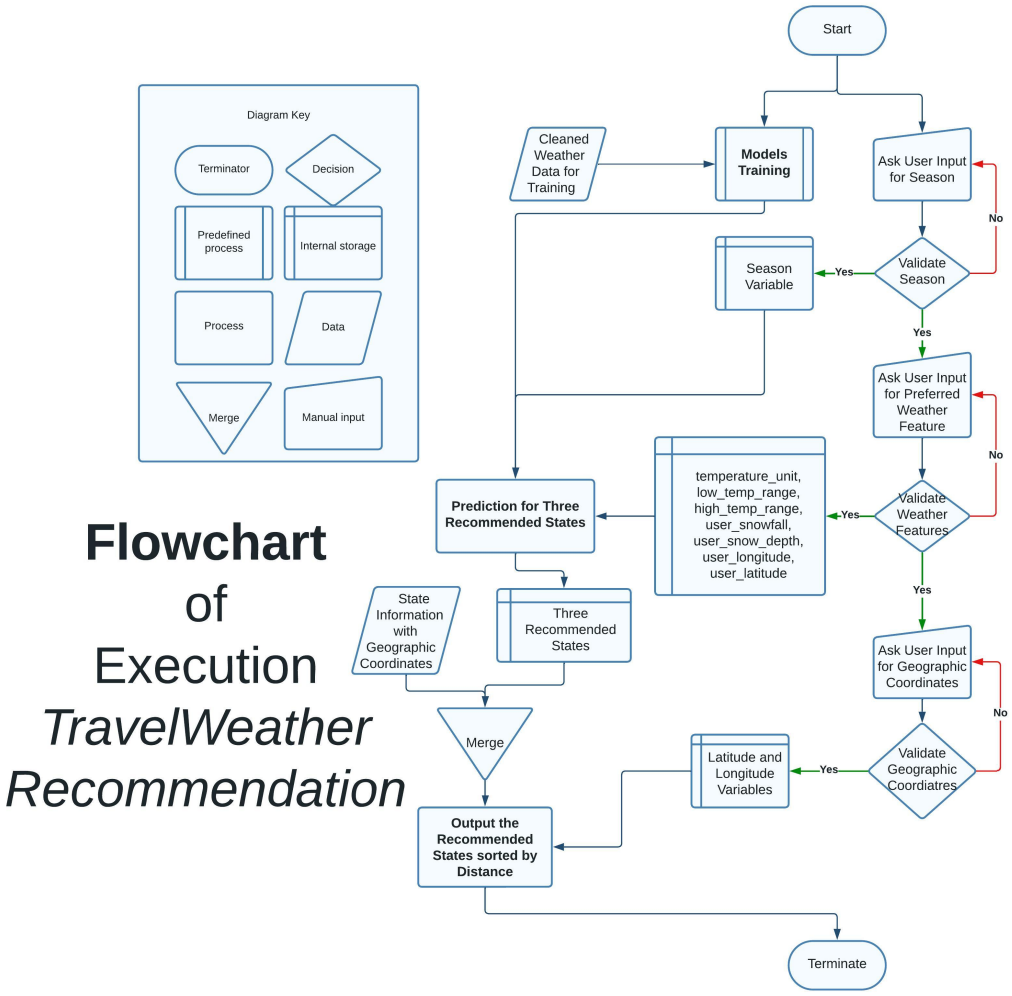

Figure 2: Flowchart of Execution, TravelWeather Recommendation

The overall execution of our TravelWeather Recommendation program was illustrated by the flowchart above. The flowchart's program side is on the left, and the user side is on the right. It follows these three crucial steps: from **model fitting** to **prediction** and **providing a tailored output**, to offer users accurate and personalized insights for their travel plans.

4 Recommendations

In light of the comprehensive scope of our Travel Recommendation project, we have identified key areas for further enhancement and development. Building upon the foundation laid by our utilization of the Global Historical (GHCN) weather data, we recognize the potential to refine and extend the capabilities of our application. Here are expanded recommendations based on our project:

4.1 Enhance Travel Recommendations

Elevate the travel recommendation system significantly by extending its scope beyond weather forecasts. Introduce personalized suggestions for food, activities, and itineraries based on individual user preferences. Furthermore, integrate more machine learning models to take into account a broader spectrum of user interests. This approach ensures a finely curated and customized travel experience, elevating the overall quality of recommendations.

4.2 Global Expansion with Predictive Analytics

Recognize the potential of our project to transcend geographical boundaries by implementing predictive analytics models. Incorporate additional data sources and employ predictive modeling techniques to enhance the accuracy and relevance of travel recommendations for an extensive array of international destinations. Utilize data science methodologies to adapt the application to diverse climates, cultural preferences, and travel dynamics, positioning it as a versatile and indispensable tool for users worldwide.

4.3 Implement Advanced User Interface

Acknowledge the significance of a visually appealing and user-friendly interface by employing data-driven design principles. Enhance user experience through data visualization, intuitive navigation, and interactive elements. By incorporating data-driven insights, we aim to optimize user engagement and satisfaction.

4.4 Optimizing Model Training: Minimizing User Waiting Time

In the current iteration of TravelWeather Recommendation, users are required to wait for the models to undergo training each time the program is initiated. This waiting period can be efficiently minimized by segregating the training process. By doing so, the models will undergo training only once, allowing for repeated utilization in subsequent predictions without the need for redundant training.

4.5 Explore More Features of Tourist Attraction

This system can be enriched by incorporating diverse tourist attractions. Specifically, we propose augmenting the recommendation engine to include historical landmarks, culinary hotspots, and unique geographical features. This ensures that users receive suggestions aligned not only with their weather preferences but also with their broader travel interests. By integrating historical attractions, gastronomic delights, and distinctive landscapes into our recommendation system, this will provide a more comprehensive and personalized travel experience. This will cater to the varied interests of our users, making their journeys not just weather-perfect but culturally and experientially fulfilling.

4.6 Randomize X Values for Distinct Province/State Predictions

In the current iteration of the project, we cannot guarantee that a certain model is going to predict three distinct location. Even though we are trying to achieve this by randomizing the minimum and maximum temperature inputs, there are instances where

the three predictions are not distinct. By randomizing other features such as precipitation, snowfall, and snow depth we might have been able to guarantee different location outputs.

5 Limitations

- **Recommending Cities, Instead of States:** We couldn't map the prediction to cities due to the time constraints.
- **Impact of Visualization:** By adding a larger dataset, we could have improved the visualization and allowed for a more in-depth examination of weather features. A more comprehensive methodology could have been used to explore further details for illustrating similarities in weather features between different citie/areas.
- **Extend Training Dataset For Earlier Observations:** We could have utilized a larger dataset from earlier years to enhance our predictive models. By expanding the dataset, we could have improved the training process by lowering the variance, potentially leading to more accurate and robust predictions.
- **Pipeline Difficulties - Training and Predicting:** We were unable to find an effective method of separating the training process from the prediction program. If we could find a solution, we would only need one training process, and the user would receive their recommendations faster whenever they launched the program.
- **Dealing With Null Values:** Due to the time constraint, we simply removed all raws with null values from the program. As a result, we lost some of the valuable data. This is a limitation of the GHCN dataset, as some weather stations can only record a few weather features. However, we should have thought of a better way to deal with Null values.

6 Accomplishment Statements

6.1 Arvin Bayat Manesh

- Implemented the user command line interface, with validation. All the inputs from the user is validated for correct format. As an example, the user should type in an integer in the 0-100 range for precipitation, otherwise the program is going to ask the user again for the correct format.
- Increased the accuracy of our models by making a VotingClassifier from the sickit-learn library. This VotingClassifier consists of GaussianNB, SVC, and a RandomForestClassifier.
- Randomized the input values for the models to recommend three distinct holiday locations to the user. These randomized numbers are chosen from the range which the user chose for their desired maximum and minimum temperature range.
- Using the trained models and the user inputs predicted three holiday locations, and displayed them to the user.
- Minimized the user waiting time by only training the model which corresponds to their desired season for their trip.
- Conducted thorough testing on the holiday recommendation location script to ensure the prediction is accurately made from the raw data, and that the recommneded locations are ready to be used for the location feature.

6.2 Keishi Hazel Allam

- Implemented and optimized data extraction process using Spark on a cluster, generating compressed JSON files for efficient storage and facilitating seamless data transfer
- Automated data aggregation precedures, developing a python script that combines individual JSON files into cohesive CSV formats, streamlining data handling and accessibility.
- Performed data cleaning and formatting process, ensuring the integrity and consistency of diverse datasets through strategic filtering and systematic handling of various variables
- Designed cleaning script with pivotal transformations, optimizing data organization and readability for advanced analytical tasks, including muti-level column flattening

6.3 Jonathan Yang

- Maintained comprehensive meeting notes for all group meetings, enhancing team communication and accountability.
- Coordinated the development of a random forest data analysis model for four-season analysis to provide a foundational draft for team refinement.

- Produced effective visualizations using LucidChart and Tableau, including a flowchart of execution and exploratory data analysis visualizations, to ensure the audience comprehensively understands the purpose and complete operation of the program.
- Implemented the location feature into the program and conducted thorough testing to ensure that the additional feature seamlessly integrates with the existing system.

References

Adapted code to extract the Weather Data in SFU Cluster:

- https://coursys.sfu.ca/2023fa-cmpt-353-d1/pages/WeatherData

Weather data set and ghcnd-states.txt:

- https://www.drought.gov/data-maps-tools/global-historical-climatology-network-ghcn