PROJECT DELIVERABLES

Annual Fall Foliage Prediction Machine Learning Model

By:

Khanh Van - kxv230013

Spring 2025

University of Texas at Dallas

Supervised By:

Professor Brian Wescott Ricks

Professor of Instruction - Computer Science

University of Texas at Dallas

TABLE OF CONTENTS

Executive Summary

Project Overview

The Annual Fall Foliage Prediction Machine Learning Model aims to provide accurate and timely predictions of fall foliage patterns across the United States. By utilizing historical foliage data, year-to-date weather records, and advanced machine learning techniques, the project seeks to address the growing demand for precise and user-friendly tools to plan autumn-related activities, tourism, and environmental analysis.

Objectives

1. Develop a machine learning model that predicts the timing and intensity of fall foliage at a granular geographical level.
2. Create a reliable pipeline for collecting, preprocessing, and integrating data from multiple sources, including NOAA weather data and Explore Fall fall foliage archived data.
3. Evaluate the model's performance by using different metrics including accuracy, mean squared error (MSE), or F1 scores to ensure reliability and precision.
4. Integrate model's predictions and results into a database using PostgreSQL for visualization and model-training purposes.
5. Visualize the prediction results in an intuitive and interactive format such as heatmaps or intensity map with time included.

This project not only highlights the intersection of data science, machine learning, and environmental science. But also provides practical value for stakeholders, including tourism agencies, environmental researchers, and individual travelers.

Introduction

Background & Motivation

As a computer scientist, countless hours are spent in front of monitors, navigating complex algorithms, analyzing numbers, and solving challenging problems. While this work is intellectually stimulating, it can often lead to mental fatigue and a diminished capacity for creativity and productivity. To maintain peak performance and balance, it's essential to engage in hobbies and activities that allow the mind to reset and recharge.

One such activity is exploring nature, which has been shown to reduce stress, improve focus, and spark creativity. Fall, in particular, offers a unique opportunity to connect with nature through its vibrant foliage and breathtaking scenery. The annual transformation of leaves into shades of red, orange, and yellow draws millions of people outdoors, whether for hiking, photography, or simply appreciating the beauty of the season.

Problem Statement

The timing and intensity of fall foliage vary annually due to environmental factors including temperature, precipitation, and daylight, making it difficult for travelers, photographers, and tourism agencies to plan effectively. Existing prediction methods are often too generalized, and climate change adds further uncertainty.

To address these challenges, this project leverages machine learning to create a precise and adaptable model for predicting fall foliage patterns. By integrating historical data and environmental variables, it aims to provide a reliable, user-friendly solution for planning and optimizing fall-related activities.

This project aims to develop a machine learning model to predict the timing and intensity of fall foliage across the United States. It involves collecting and integrating historical foliage and weather data, training and optimizing a predictive model, and presenting results through interactive visualizations like maps and charts.

Data Collection Process

Source of Data

This Machine Learning Model will rely solely on three different datasets:

- Archived Fall Foliage Data - [Explore Fall]
- Year-to-Date Weather Data - [NOAA]
- Average Day-Light Data - [U.S. Naval Observatory]

Archived Fall Foliage Data

Data Selection

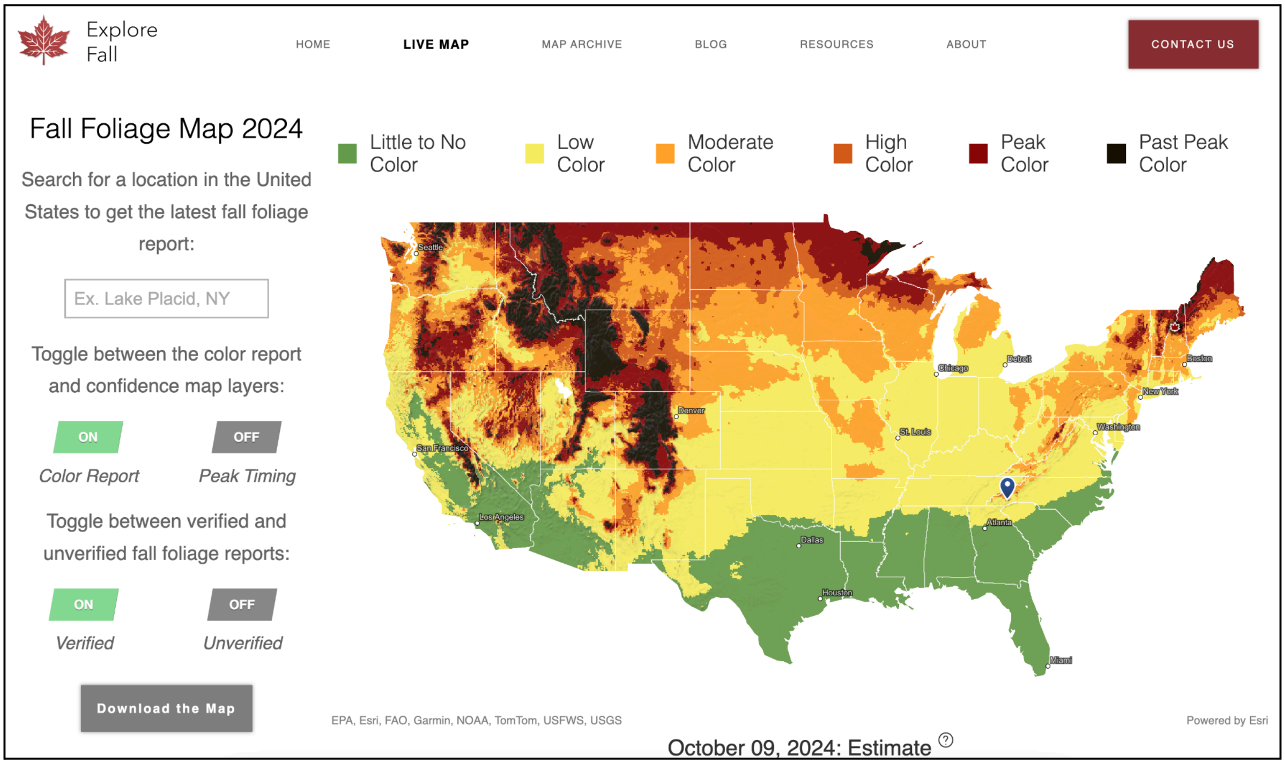
With a time constraint of roughly 3 months, this project will be focusing on only a few key and major destinations rather than attempting to cover the entire U.S. This approach allows for more efficient travel planning, reduces the stress of overextending, and ensures a richer tourism experience at iconic spots known for their stunning autumn colors.

The following destinations are the most scenic and popular choices for fall foliage sightseeing according to *Thrillist*, *US News Travel*, *Culture Trekking*, and *Lonely Planet*:

(Selected locations for the project)

Data Pre-Processing

The process of data preprocessing for Archived Fall Foliage Data will be done manually since there is no existing formatted dataset for Annual Fall Foliage. Instead, there will be an interactive map from Explore Fall that shows the intensity of Fall colors with dates indicator.



[Explore Fall Interactive Fall Intensity Map]

This project will be utilizing Fall Intensity Data from the past 10 years from Explore Fall. The Processed Data will includes:

- Locations - Aspen, Traverse City, Stowe, etc.
- The Time - Early (1st - 9th), Mid (10th - 19th), Late (20th - 31st).
- Months - September, October, November
- Intensity - Little to No Color, Low - Moderate, High - Peak, Past-Peak

Data will be recorded into Structured Query Language (SQL) file format. Updating and modifying using PostgreSQL for better efficient use of the Machine Learning process.

Year-to-Date Weather / Temperature Data

Data Selection

To support our machine learning model, historical climate data will be collected from the **NOAA Weather Forecast Office Climate Portal** (weather.gov/wrh/climate). The data is accessed by selecting a **specific state**, then a **local area or city within that state**.

For each selected location:

- We will retrieve monthly climate summaries provided in table format.
- The two main parameters collected will be:
  - Average temperature
  - Average precipitation
- These summaries include average values for all twelve months. We will manually extract data for the months of June through September, as they are most relevant to our prediction goals.
- Data will be exported to Excel, where we will calculate the average values across these months to form consistent climate features per location.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
1991	20.2	37.1	40.7	49.2	60.5	71.9	75.1	73.9	64.6	50.2	32.1	33.5	50.8
1992	33.9	37.2	41.9	49.1	60.8	65.5	67.6	67.3	64.3	51.9	31.7	23.7	49.6
1993	19.7	19.0	36.6	45.4	58.9	65.2	70.4	70.4	59.7	50.9	31.4	31.0	46.6
1994	19.9	20.4	41.7	48.4	63.1	70.0	71.0	72.0	64.6	52.5	38.7	29.7	49.3
1995	27.2	33.8	36.7	41.9	53.5	67.1	75.0	76.7	62.4	50.6	36.9	27.6	49.1
1996	16.8	29.3	29.6	47.0	55.9	69.3	71.8	71.5	60.0	50.9	27.4	20.4	45.8
1997	19.5	30.6	40.2	41.8	56.0	70.0	73.8	71.0	65.9	52.9	35.6	31.4	49.1
1998	27.3	35.4	29.8	48.5	61.8	64.8	75.7	73.8	70.2	51.4	41.3	31.5	51.0
1999	26.5	37.3	39.5	47.0	59.1	67.2	77.2	73.9	61.7	52.8	47.1	33.9	51.9
2000	28.5	36.5	42.7	49.5	62.1	67.5	75.5	76.4	66.3	53.3	28.1	17.9	50.4
2001	28.6	20.5	35.1	50.2	59.9	68.3	76.7	75.3	65.1	53.0	44.8	30.6	50.7
2002	30.8	35.2	28.4	50.5	56.5	74.2	79.8	74.2	65.2	46.1	38.7	33.4	51.1
2003	26.3	24.8	39.3	50.4	58.2	66.7	76.0	74.8	63.4	56.5	33.6	30.2	50.0
2004	24.5	26.4	42.4	51.0	60.6	63.5	72.7	69.0	66.7	52.1	39.8	32.3	50.1
2005	22.5	34.4	38.9	49.3	56.6	69.2	76.0	73.2	68.6	51.9	41.4	25.3	50.6
2006	39.6	26.7	34.9	53.6	61.2	71.3	78.9	72.8	58.9	48.7	38.0	30.2	51.2
2007	23.9	20.4	44.9	45.8	61.5	68.5	75.7	73.9	65.9	53.9	40.2	24.2	49.9
2008	23.3	26.5	37.5	44.8	55.4	66.9	75.1	71.8	63.9	50.5	37.5	21.3	47.9
2009	26.4	31.0	37.2	46.8	59.6	66.1	70.1	69.4	63.0	42.2	43.3	18.4	47.8
2010	22.4	23.5	40.9	51.7	57.3	68.9	76.1	75.9	63.6	55.4	36.4	24.8	49.7
2011	19.9	24.8	34.3	46.8	56.1	68.1	77.5	73.5	61.6	54.1	40.4	31.3	49.0
2012	32.2	29.6	52.7	53.7	62.2	74.4	81.8	75.0	66.8	50.4	39.8	29.0	54.0
2013	26.9	28.5	33.6	41.5	58.3	67.8	73.7	74.0	69.7	49.0	37.4	21.1	48.5
2014	25.8	22.0	36.3	48.6	58.5	67.3	72.1	71.6	64.3	55.2	32.3	27.5	48.5
2015	27.8	27.1	45.5	51.3	56.2	69.4	74.4	72.1	69.4	55.2	40.4	29.0	51.5
2016	26.9	36.9	45.0	49.5	58.6	73.7	74.2	72.7	65.4	57.1	45.3	23.8	52.4
2017	24.8	33.9	41.1	50.2	57.9	71.9	80.0	70.2	66.0	52.3	43.0	27.3	51.5
2018	25.8	20.7	37.4	41.9	64.9	72.7	74.9	73.1	66.0	48.3	35.6	28.4	49.1
2019	26.3	15.7	31.3	49.1	54.2	69.3	75.7	72.4	69.4	45.5	37.1	32.0	48.2
2020	26.9	31.2	42.6	48.3	58.1	74.6	76.4	76.2	66.3	48.3	44.7	33.7	52.3
2021	30.9	17.6	43.8	48.0	59.3	74.4	M	M	67.3	52.0	42.8	31.2	46.7
2022	26.0	25.4	36.3	44.4	59.0	71.5	78.0	75.9	67.1	52.1	31.5	20.0	48.9
2023	22.8	26.0	29.9	46.6	62.5	70.6	73.2	74.8	66.8	48.3	40.7	33.4	49.6
2024	18.6	35.9	37.4	48.5	59.1	69.4	74.4	73.8	70.0	55.3	37.2	31.1	50.9
2025	19.9	19.3	42.3	47.1	M	M	M	M	M	M	M	M	32.2
Mean	25.4	28.0	38.5	47.9	58.9	69.3	75.0	73.1	65.3	51.5	38.0	27.9	49.3
Max	39.6 2006	37.3 1999	52.7 2012	53.7 2012	64.9 2018	74.6 2020	81.8 2012	76.7 1995	70.2 1998	57.1 2016	47.1 1999	33.9 1999	54.0
Min	16.8 1996	15.7 2019	28.4 2002	41.5 2013	53.5 1995	63.5 2004	67.6 1992	67.3 1992	58.9 2006	42.2 2009	27.4 1996	17.9 2000	32.2

Average Precipitation/Temperature Generated Table (1991-2025)

Year-to-Date Day-Light Data

Data Selection

To collect average daylight duration per location, we used the **U.S. Naval Observatory** (**USNO**) **Daylight Duration Tool** (https://aa.usno.navy.mil/data/Dur_OneYear).

- For each selected city and year, we manually entered the location into the online form to generate a **daylight duration table** for the entire year.
- The output is a structured table listing daily **sunrise and sunset times**, along with the **calculated daylight duration** for each day.
- Using **Excel**, we filtered or isolated the data for the months of **June through September** and then calculated the **average daylight duration** across this four-month period.
- This average daylight value is then recorded for that location and used as one of the input features in our predictive model.

This manual method allows precise control over location-specific daylight data, complementing the temperature and precipitation data to form a complete environmental profile for each region.

Day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1	09:29	10:15	11:21	12:40	13:52	14:45	14:52	14:11	13:01	11:46	10:30	09:37
2	09:30	10:17	11:23	12:43	13:54	14:46	14:52	14:09	12:58	11:43	10:28	09:36
3	09:30	10:19	11:26	12:45	13:57	14:47	14:51	14:07	12:56	11:40	10:26	09:35
4	09:31	10:21	11:28	12:48	13:59	14:48	14:50	14:05	12:54	11:38	10:24	09:33
5	09:32	10:24	11:31	12:50	14:01	14:49	14:49	14:03	12:51	11:35	10:22	09:32
6	09:33	10:26	11:34	12:53	14:03	14:50	14:49	14:01	12:49	11:33	10:19	09:32
7	09:34	10:28	11:36	12:55	14:05	14:51	14:48	13:59	12:46	11:30	10:17	09:31
8	09:35	10:30	11:39	12:58	14:07	14:51	14:47	13:56	12:44	11:28	10:15	09:30
9	09:36	10:32	11:41	13:00	14:09	14:52	14:46	13:54	12:41	11:25	10:13	09:29
10	09:38	10:35	11:44	13:03	14:11	14:53	14:45	13:52	12:39	11:23	10:11	09:28
11	09:39	10:37	11:46	13:05	14:13	14:53	14:44	13:50	12:36	11:20	10:09	09:28
12	09:40	10:39	11:49	13:08	14:15	14:54	14:42	13:48	12:34	11:18	10:07	09:27
13	09:41	10:42	11:52	13:10	14:17	14:54	14:41	13:46	12:31	11:15	10:05	09:27
14	09:43	10:44	11:54	13:13	14:18	14:54	14:40	13:43	12:29	11:13	10:03	09:26
15	09:44	10:46	11:57	13:15	14:20	14:55	14:39	13:41	12:26	11:11	10:01	09:26
16	09:46	10:49	11:59	13:18	14:22	14:55	14:37	13:39	12:23	11:08	10:00	09:25
17	09:47	10:51	12:02	13:20	14:24	14:55	14:36	13:37	12:21	11:06	09:58	09:25
18	09:49	10:54	12:04	13:22	14:25	14:55	14:35	13:34	12:18	11:03	09:56	09:25
19	09:50	10:56	12:07	13:25	14:27	14:56	14:33	13:32	12:16	11:01	09:54	09:25
20	09:52	10:58	12:10	13:27	14:29	14:56	14:32	13:30	12:13	10:58	09:53	09:25
21	09:54	11:01	12:12	13:30	14:30	14:56	14:30	13:27	12:11	10:56	09:51	09:25
22	09:56	11:03	12:15	13:32	14:32	14:56	14:28	13:25	12:08	10:54	09:49	09:25
23	09:57	11:06	12:17	13:34	14:33	14:55	14:27	13:23	12:06	10:51	09:48	09:25
24	09:59	11:08	12:20	13:37	14:35	14:55	14:25	13:20	12:03	10:49	09:46	09:25
25	10:01	11:11	12:22	13:39	14:36	14:55	14:23	13:18	12:01	10:46	09:45	09:25

Average Day-Light Generated Table (Aspen, CO 2024)

Model Training

Random Forest Regressor

For training the prediction model, I utilized the Random Forest Regressor. I've tried out various models available in Scikit-Learn like K-means or clustering. However, with the labeled data I've collected, Random Forest Regressor is the most suitable model that can be used for training.

Throughout the training process, I experimented with various hyperparameters such as the number of estimators (trees), maximum tree depth, and minimum samples required to split an internal node. I discovered that different locations' data will behave differently so I would have to manually modify the hyperparameters accordingly for each location. The predicted results are desirable with a (+/-1) error range.

No.	Pred. Low	Pred. Peak	Pred. Past	Low	Peak	Past
1	2	4	5	2	3	4
2	1	4	4	2	3	4
3	2	3	5	2	3	5`
4	2	4	5	2	3	5

Prediction Results for Eastern Sierra (Results will varies based on locations)

Each location will have its own .pkl file due to difference inconsistency in data between locations, each location will have a different pattern from the others.

Limitations and Future Direction

Limitations in Data Collection

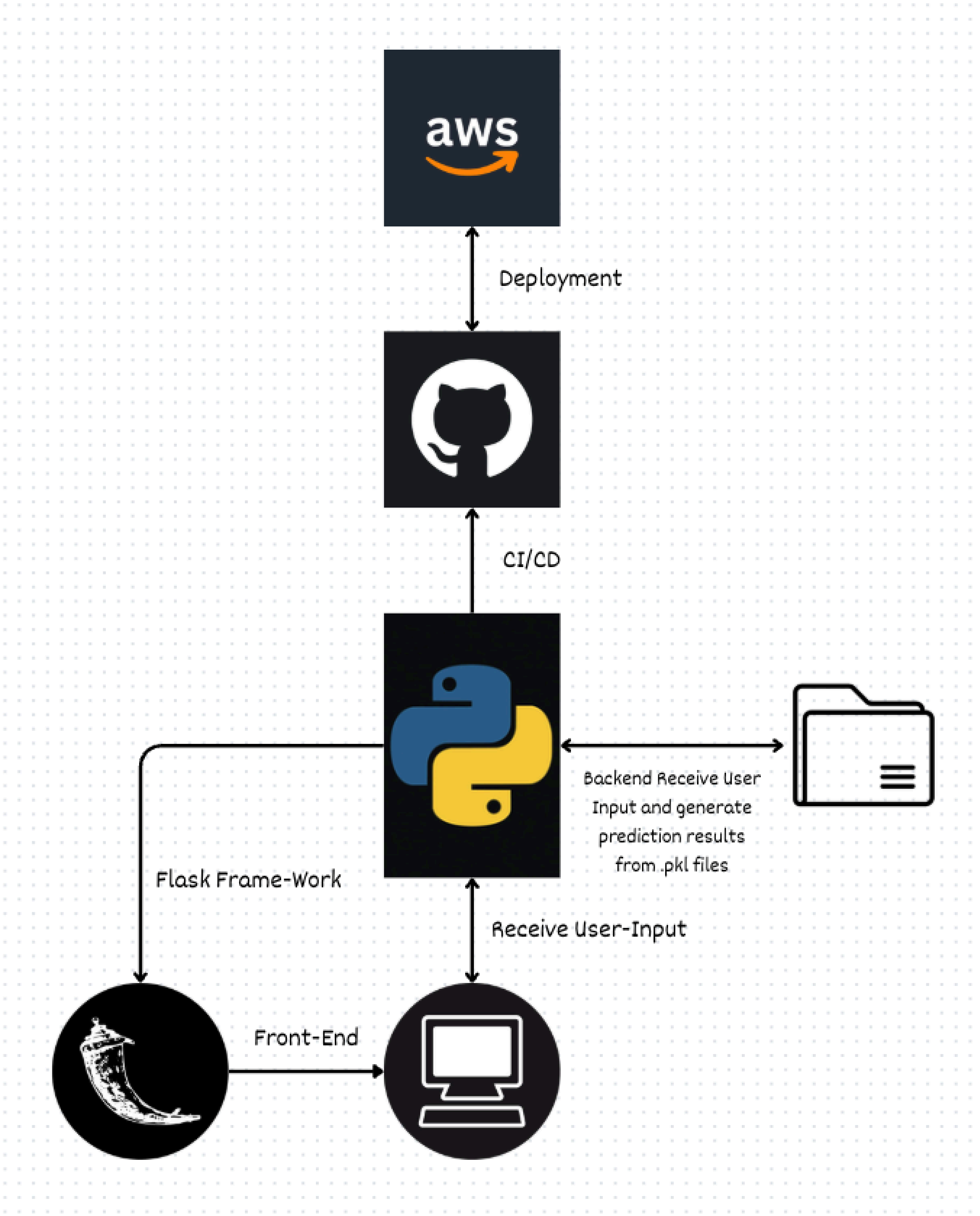
While building the prediction model, one of the major challenges I encountered was the lack of comprehensive and consistent historical foliage data across multiple regions and year. The data only spans from 1991-2025 (less than 40 years) which is not good enough data to train a model, it would cause a lot of noise throughout the data. Due to the limited dataset, the machine learning model could not be trained as effectively or accurately as desired. As a result, no further training or refinement of the model was feasible within the project timeline. Recognizing the constraint, I decided to pivot my focus toward strengthening my skills in full-stack web development so I can share my project with everybody and everyone can play around with it.

Full-Stack Web Application

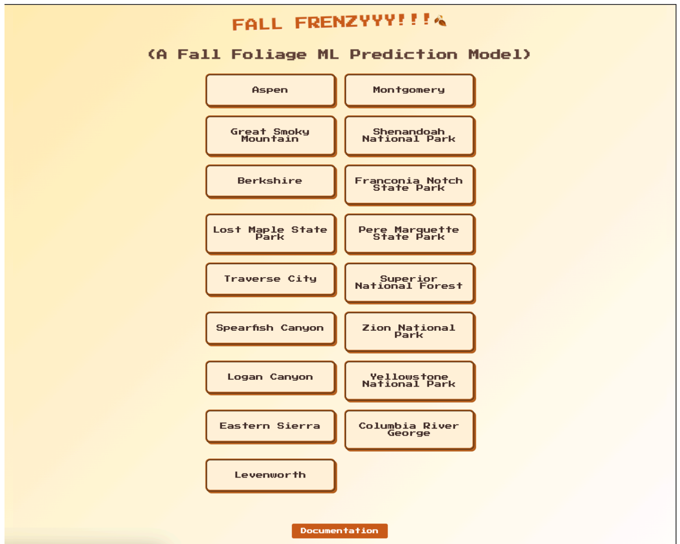
I focused on strengthening my skills in full-stack web development by building several applications using Python and Flask web framework. Through these projects, I gained hands-on experience in designing backend services, integrating third-party APIs, and creating responsive front-end interfaces. This process deepened my understanding of web architecture, improved my ability to manage server-client interactions, and enhanced my proficiency in deploying and maintaining scalable web applications.

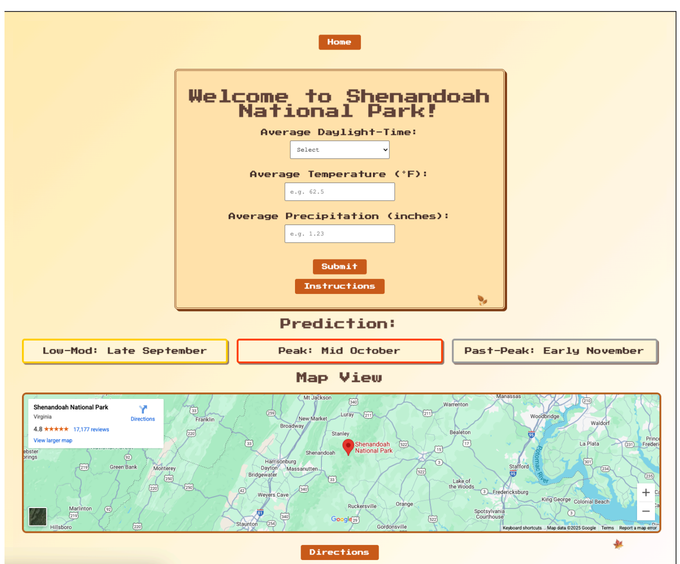
Web Application Layout:

- Main Page
    - Location(s)
    - Documentation of the Project
- Location Pages
    - Instruction Documentation (How to get data for input)
    - Data Input Fields
    - Prediction Output
    - Interactive Map to Location



Design Document of Full-Stack Web Application





[LINK TO WEB APPLICATION]

[Github Repository]

I would like to sincerely thank Professor Brian Ricks for his support and accommodations throughout the semester, which greatly contributed to the successful completion of my project.