Documentation

This is the documentation for Evaluating Machine Learning and Statistical Models for Greenland Subglacial Bed Topography.

Table of Contents

Getting Started

Welcome to the documentation for Evaluating Machine Learning and Statistical Models for Greenland Subglacial Bed Topography. Whether you're new to the project or a returning explorer, this guide is designed to help you navigate the complexities of this exciting research.

Before delving into the technicalities, familiarize yourself with the foundational materials:

Understand the background, steps by reviewing published materials:

1. Read the published paper: Evaluating Machine Learning and Statistical Models for Greenland Subglacial Bed Topography
    a. May2023_to_March2024 > Copy_official_paper.pdf
2. Read the ICMLA slide deck:
    a. May2023_to_March2024 > Copy_ICMLA_presentation.pdf

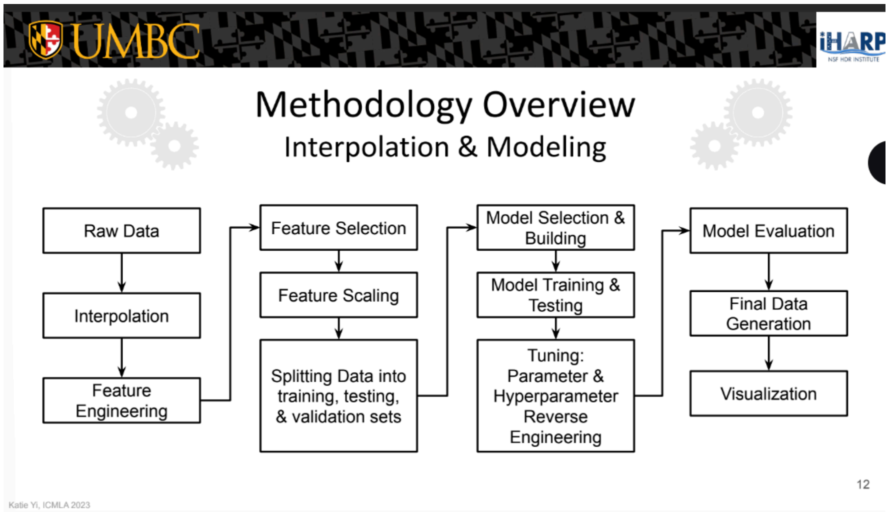
For additional background information read:

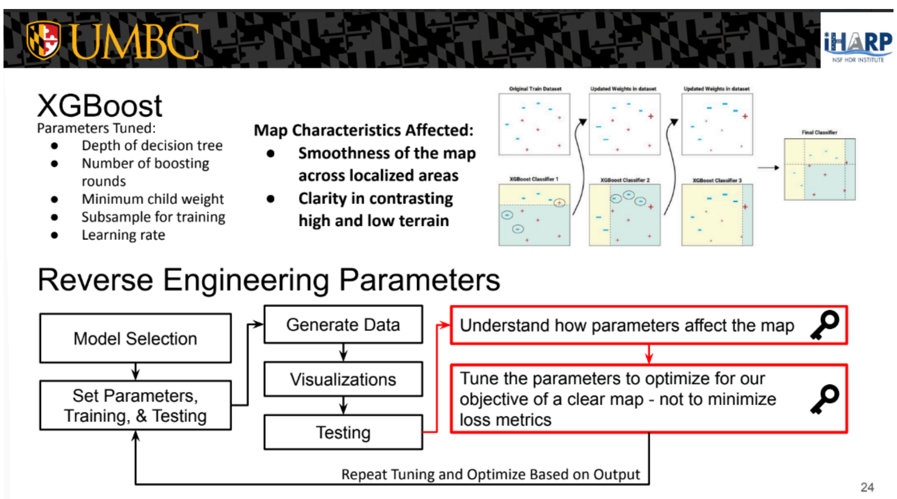
1. May2023_to_March2024 > Final_XGB_Writeup.pdf
2. Reference, but do not fully read: May2023_to_March2024 > Documentation.pdf

Code for this project can be found on github:
[big-data-reu/2023-projects/team-1 at main · big-data-lab-umbc/big-data-reu (github.com)]

Pipeline

Below are the two most important pipelines of this project. They describe the process of the overall project as well as the parameter tuning process.





Measuring Success

There is a lot that goes into modeling. How do you actually measure success? We opted to use the RMSE, MAE, and Coefficient of Determination. An additional metric is the TRI score. Since the objective of this project is to build a tool, another metric that is collected is Quality, specifically usability of the graphs.

I did not use the RMSPE because this gives more weight to larger errors. I did not want to skew my overall results due to a small region not having sufficient input. I preferred a wider net to capture smoothness. I also feel it did not accurately reflect the points we were able to capture as seen by the density plots.

In addition to just metric testing, I consulted our expert Dr. Mathieu Morlinghem to determine the performance of our models. I created a blind test and asked for preferences between the two graphs. Questions I asked included but are not limited to the following:

1. What stands out to you about graph one?
2. What stands out to you about graph two?
3. What characteristics do you prefer across the graphs? This question is focused on collecting the expert knowledge on the region. For example: "I prefer the dark trough areas appearing more connected versus broken up by the edges of the terrain"

There is no formal list to capture the questions asked of our expert. Instead I remained unbiased throughout the collection process taking notes on the feedback and preferences. This feedback would later come in handy during the tuning process to accentuate the characteristics to improve usability, a measure of quality. There is opportunity to formalize this process and involve more user feedback.

Cautions

1. Do not overfit the dataset just to make the model look better.
2. Understand the metrics you are using and how they work with topography and change.
3. Ensure you are checking data assumptions. For example, Greenland topography is not a perfect normal curve.

Review: Understanding the Past

After reading up on the background you should be able to answer the following questions:

1. What is interpolation and why do we need it?
2. Where does our data come from?
3. How big is our data?
4. What variables make up our data?
5. How supervised is our data?
6. How do you measure success in this project?
7. What models have been tested? What models are the most successful?
8. What can you do next?

Warnings

I have decided to call this section Warnings, but truely it comes from my words of wisdom working with modeling on this project. Below is a list of tips and events to be aware of as you begin this journey of exploration.

1. Metamorphic testing: do not get caught up in the metamorphic testing cycle for too long. This will take over your whole project if you let it. Accept good results and continue to build forward.
2. Documentation: write everything you test down and use version controlling if you can. I found myself wondering if the model I ran yesterday is better than my new models today; it was easy to look up without having to retest 30 pairwise values.
3. Big picture thinking: as you build out the pipeline to success you will get focused on each step. As you are working consider two things: (1) how will this affect the next step? and (2) does this make sense for our overall objective? Sometimes you just need to step back from the terminal and think about the big picture.
4. KISS: Keep it simple. Only do as much as you need to do for success. This does not mean be lazy about testing; instead it means only use what you need. If the data does not call for a super fancy complex model, do not force it.

Next Steps

March 2024

Literature Review*

Models

1. What are we looking for?
   We are trying to identify if any other models will fit our data requirements. If yes, we would like to consider implementing them.
   Note: not all additional models tested will be included in the final chapter due to space limitations. Judgments will be made based on model relevancy and model performance.

Metrics

1. What are we looking for?
   We are trying to identify existing literature on topography prediction accuracy. Specifically, we are trying to find a tool to quantify the 1201 dataset predictions where we have no true-known values.
2. What should you search for?
   Please look for formal articles and publications. Potential search prompts include: How to quantify "topography" prediction "accuracy."

Code Review

Code review of extended work by the University of Wisconsin.
Steps:

1. Review the code and understand what is happening
2. Create a new jupyter notebook file
3. Recreate the uncertainty quantification code
   a. Format the file to fit into our pipeline
   b. Add comments to every line so a 4th grader could follow what is happening.
4. Plug the quantification metrics into the pipeline
5. Run partial data for a test
6. Run the full dataset for results

Model Additions

We will look to implement random forest as an additional model for our research.
[Update: this has already been started in the REU 2023 folder]
Find the code in Github > Model Used > Machine Learning > randomForest_full_ky_240206.ipynb
Steps:

1. Tune the parameters and hyperparameters based on the tuning guidelines

Metric Implementation

We will look to implement metrics identified in the literature review.