Enhancing Topography Prediction of the Greenland Ice Sheet Using Extreme Gradient Boosting Model

Katherine L. Yi

Department of Computer Science, Purdue University

Big Data REU, UMBC

Dr. Jianwu Wang

July 13, 2023

Abstract

This paper presents the development and evaluation of an Extreme Gradient Boosting (XGB) model for predicting the topography of the Greenland Ice Sheet (GrIS) in the context of rising sea levels caused by melting ice. Accurate predictions of sea level rise require a precise understanding of the bed topography beneath the ice sheet, which is currently incomplete. To address this, satellite and radar imagery datasets were leveraged, and machine learning techniques were employed. The XGB model was chosen for its effectiveness in regression tasks and scalability. Preprocessing steps involved feature selection, scaling, and data splitting. The XGB model exhibited exceptional performance, surpassing alternative models, with a Root Mean Square Error (RMSE) of 32.680, Mean Absolute Error (MAE) of 22.273, and coefficient of determination (R^2) of 0.96. The model's accuracy and precision were validated using multiple metrics. Additionally, the XGB model generated a detailed topography map, highlighting its ability to capture complex spatial patterns. The results contribute to a better understanding of topography prediction and have implications for environmental planning, infrastructure development, and natural disaster management. Future studies can explore additional features and advanced techniques to further improve the model.

Introduction

This paper presents our work on the development and evaluation of an Extreme Gradient Boosting (XGB) model for predicting topography in the context of rising sea levels caused by the melting Greenland Ice Sheet (GrIS). Accurate predictions of sea level rise require a precise understanding of the bed topography beneath the ice sheet, which is currently incomplete and hinders our ability to model and predict the ice sheet's behavior. The bed topography, which encompasses the shape and structure of the underlying bedrock, plays a vital role in the stability and response of the ice sheet to climate change, influencing ice flow and its interaction with warm ocean water.

To address these data gaps, we leverage satellite and radar imagery datasets and employ machine learning and deep learning models. By combining data on ice flow speed, surface height, and ice thinning rates, our aim is to create a comprehensive view of the ice bed topography. This research significantly advances our knowledge of ice sheet dynamics, improves predictions of sea level rise, and provides valuable insights for effective decision-making in mitigating the impacts of climate change.

By bridging the data gaps and refining our predictive models, we can enhance our understanding of the behavior of the Greenland Ice Sheet and enable the development of targeted strategies by policymakers, scientists, and stakeholders. These strategies will be instrumental in mitigating the impacts of sea level rise and facilitating adaptation to the challenges posed by climate change.

Background: Understanding the Data

This section provides a background on the data used in our study, including the variables, their sources, and their meanings. The dataset consists of various measurements related to the topography of the Greenland Ice Sheet (GrIS), which is crucial for estimating the bed topography beneath the ice and improving mapping accuracy.

The variables in our dataset include surf_x and surf_y, representing the coordinates of cell centers in meters. The variables surf_vx and surf_vy denote the ice flow speed in the x-direction and y-direction, respectively, measured in meters per year. Surf_elv represents the surface height above sea level in meters. The variables surf_dhdt and surf_SMB provide information on ice thinning rates and snow accumulation, respectively, both measured in meters per year and representing the depth or difference in "ice equivalent." The variables track_bed_x and track_bed_y specify the specific x-coordinate and y-coordinate inside a cell, measured in meters. Finally, the derived variable v_mag represents the velocity magnitude of ice flow at each (x, y) coordinate.

The target variable in our study is track_bed_target, which represents the bed height beneath the ice. Our goal is to estimate the topography of Greenland's ice beds by predicting the missing data points in the topographic maps. By improving the mapping variation and filling in the data gaps, we aim to enhance the accuracy of bed mapping and obtain a more comprehensive understanding of the topography of the Greenland Ice Sheet.

The dataset contains a total of 1,442,401 points, with 632,706 (43.86%) of the data points known. Our analysis focuses on utilizing the available data to predict the missing values and improve the representation of the topography of the Greenland ice beds.

Methodology

In this study, the methodology pipeline for developing and evaluating the Extreme Gradient Boosting (XGB) model for topography prediction consisted of several key steps. Firstly, the interpolated dataset using K-nearest neighbors (KNN) was preprocessed to handle missing data and ensure compatibility, including the derivation of additional features and the selection of relevant variables in collaboration with domain experts, such as the velocity magnitude of ice flow. This step aimed to enhance the model's performance by including important indicators of topography. Subsequently, feature selection was performed to identify influential features, such as ice flow speed, surface height, ice thinning rates, snow accumulation, and the derived velocity magnitude. This process ensured that only relevant variables were included in the model. After feature selection, the data was standardized using the StandardScaler to ensure consistent scaling across different variables. Finally, the dataset was randomly split into training, testing, and validation sets, allowing for a comprehensive evaluation of the model's performance. This systematic preprocessing pipeline enabled the development of an accurate and robust XGB model for topography prediction based on the KNN-interpolated dataset.

Preprocessing

The provided nearest neighbor interpolated data requires preprocessing before modeling. Preprocessing involves deriving additional features, feature selection, scaling values, and randomized splitting.

To incorporate real-world characteristics into our dataset, we collaborated with Dr. Mathieu Morlighem, an Evans Family Professor of Earth Sciences at Dartmouth University and a domain expert in ice-sheet and sea-level systems. Dr. Morlighem reviewed the features present in our nearest neighbors dataset and emphasized the importance of the velocity magnitude of ice flow as an indicator of topography. Based on his feedback, we derived the ice flow velocity magnitude at each (x,y) coordinate by calculating the standard magnitude equation applied to the scalar values of ice flow in the x direction (vx) and y direction (vy). This derived feature, denoted as "v_mag," was included in our modeling.

After deriving the velocity magnitude feature, we performed feature selection to identify relevant and insightful features that contribute to a clear final topography map. The selected features for the final model include ice flow in the x-direction (meters per year), ice flow in the y-direction (m/y), surface height (m), ice thinning rates (m/y), snow accumulation (m/y), and velocity magnitude of ice flow at each coordinate (m/y).

Features such as cell center coordinates used for data interpolation were not selected because they caused poor topography map predictions and did not improve the model significantly. Therefore, these features were excluded from the final feature set.

Next, we scaled the target values using StandardScaler from the sklearn preprocessing module. All data was scaled together, considering the presence of outliers in real-world data and the need for stability when working with diverse datasets.

The final step in preprocessing involved randomly splitting the data into training, testing, and validation datasets. We used the train_test_split function from the sklearn model selection module for this purpose. The randomization seed was set to 168 to ensure reproducibility. After experimenting with different ratio splits, we found that the best results were obtained with a 60% training, 40% testing, and 20% validation data split. The training data was further split to allocate the validation data, maintaining the same seed.

In summary, the preprocessing steps included deriving the velocity magnitude feature, selecting relevant features, scaling the data using StandardScaler, and randomizing the data into training, testing, and validation sets.

Modeling

In our pursuit of an effective topographic prediction model, we explored several alternative methods, including Gaussian Process Regression (GPR), Variational Autoencoder (VAE), and Space-Time Gaussian Process (STGP). However, these methods did not meet the desired level of performance, providing valuable insights into different modeling approaches.

Gaussian Process Regression (GPR) is a non-parametric probabilistic model that estimates the relationship between input features and target variables. Despite its potential, the GPR model did not achieve the desired level of predictive accuracy, with an RMSE of 136.930 and an R^2 of 0.113, indicating limitations in capturing the underlying patterns of the data.

Variational Autoencoder (VAE) is a deep generative model that learns to encode and decode data distributions. Similarly, the VAE model did not perform up to the desired level, resulting in an RMSE of 147.980 and an R^2 of -0.026, suggesting challenges in accurately capturing the complexity of the topographic data.

Space-Time Gaussian Process (STGP), which incorporates both spatial and temporal dependencies, aimed to capture the evolving nature of topographic patterns over time. However, the STGP model fell short in terms of predictive accuracy, achieving an R^2 of 0.265, indicating limited explanatory power.

We also evaluated previous year models, including Gradient Boosting, XGB, Lasso Regression, Dense, LSTM, and Dense+LSTM, based on their performance in predicting topography. While some models exhibited moderate predictive accuracy, others demonstrated limited performance in capturing the complex relationships within the data.

Among the previous year models, the Dense+LSTM model emerged as the most promising approach, with an RMSE of 55.040 and an R^2 of 0.840. Building upon the insights gained from the Dense+LSTM model, we directed our efforts towards the Extreme Gradient Boosting (XGB) model, which demonstrated superior performance and scalability. By surpassing the achievements of the Dense+LSTM model, the XGB model provided a significant advancement in topographic prediction (Faruque et al.).

Considering the limitations observed in the alternative methods and previous year models, we further explored the capabilities of the XGB model and VAE models. After careful evaluation, we found that the XGB model exhibited superior performance and predictive accuracy, making it the preferred choice for our topographic prediction task. Its ability to capture complex spatial patterns and provide accurate predictions solidifies its position as the leading model in our study.

In the following sections, we will delve into the XGB model, discussing its methodology, fine-tuning process, and comprehensive evaluation. The XGB model emerged as a superior choice, surpassing the performance of the alternative methods and previous year models, and providing valuable insights into topographic prediction. Additionally, we will provide insights into the limitations and challenges encountered with the alternative methods and previous year models, further emphasizing the significance of the XGB model's development.

By acknowledging the limitations of the alternative methods and previous year models, we can highlight the need for a more robust and efficient approach. The subsequent discussion will focus on the development of the XGB model, showcasing its superior performance and providing a comprehensive understanding of its implementation and evaluation.

XGBoosting

After testing other models, we decided to pursue the Extreme Gradient Boosting (XGB) supervised machine learning algorithm, renowned for its effectiveness in regression tasks and its impressive speed, making it highly efficient for processing large datasets. The XGB model combines multiple simple decision trees that learn by minimizing errors, enhancing flexibility and enabling its suitability for diverse topographic data. Moreover, XGB incorporates techniques to prevent overfitting, ensuring reliable and accurate predictions, while its ability to aggregate predictions from individual trees into a strong final model solidifies its standing as the preferred choice for our objectives.

In justifying our choice to adopt the Extreme Gradient Boosting (XGB) algorithm for our project, we carefully considered the importance of finding the right balance among several key parameters. These parameters greatly impact the performance and adaptability of the model, and striking the right balance is crucial for achieving optimal results.

The first parameter, the depth of the decision trees, is set at a value of 7, striking a balance between capturing intricate patterns and avoiding overfitting. A higher depth may risk overfitting, while a lower depth may sacrifice capturing power.

The number of boosting rounds and number of XGBoost trees is set at 350, striking a balance between a comprehensive learning process and training efficiency. Increasing this value too much may lead to overfitting, while too few rounds may result in an underfit model.

The minimum child weight, set at 0.25, helps control overfitting by setting a threshold for the minimum amount of samples required to split a node. This value is low enough to allow the model to generalize and reduce overfitting risk.

The subsample parameter, set at 0.8, determines the fraction of data used for each tree. This value strikes a balance between incorporating diverse samples to prevent overfitting and providing sufficient coverage of the dataset.

The learning rate (eta), set at 0.25, influences the speed of learning and convergence. This value strikes a balance between faster convergence and stable performance without overshooting.

Finding the right balance among these parameters is crucial. If the depth is too high or the number of boosting rounds is excessive, the model may become too complex and prone to overfitting. Conversely, setting these parameters too low may result in an oversimplified model that fails to capture important patterns. Similarly, adjusting the minimum child weight and subsample parameters too drastically may skew the model's generalization ability or coverage of the dataset. Finally, a learning rate that is too high or too low may hinder the model's ability to converge to an optimal solution. By carefully fine-tuning these parameter values, we can strike the optimal balance, maximizing the model's performance, capturing important patterns, and avoiding overfitting. This enables us to achieve reliable and accurate predictions tailored to our project's needs.

The performance of the reduced XGB model was evaluated using various metrics. The model exhibited exceptional performance, surpassing all other models in our analysis. The achieved results, including a Root Mean Square Error (RMSE) of 32.680, a Mean Absolute Error (MAE) of 22.273, and an impressive coefficient of determination (R^2) value of 0.96, highlight the effectiveness of the XGB model.

It is worth noting that these results far exceeded our expectations and surpassed the performance of alternative models that were considered. The high accuracy and precision of the XGB model provide valuable insights into the underlying data, enabling us to make informed decisions.

In addition to the remarkable predictive accuracy, the XGB model also contributed to the creation of an exceptional topography map. The map derived from the XGB model showcases clear and detailed features, further illustrating the model's capability to capture and represent complex spatial patterns.

Overall, the results obtained from the reduced XGB model not only establish it as a leading contender among alternative models but also contribute significantly to our understanding of the underlying phenomenon. The outstanding performance of the XGB model, coupled with its ability to generate visually appealing topography maps, presents an invaluable resource for further analysis and decision-making processes.

Metric Testing

To validate the results, we considered a range of metrics to comprehensively assess the performance and accuracy of the model. While all metrics provide valuable insights, three key metrics were given equal importance: the coefficient of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE). These metrics play a critical role in evaluating the performance and drawing meaningful conclusions.

R^2 is a crucial metric as it measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A high R^2 value indicates a strong relationship between the predicted and actual values, providing a measure of how well the model fits the observed data. Focusing on R^2 allows for an understanding of the achieved predictability and an assessment of the goodness of fit.

Similarly, RMSE provides an essential measure of the average magnitude of the residuals or errors between the predicted values and the actual values. By calculating the square root of the average squared differences, RMSE offers a comprehensive evaluation of the model's predictive accuracy. Emphasizing RMSE allows for a focus on the precision of predictions and an understanding of the typical magnitude of errors.

Furthermore, MAE is given equal importance as it provides insights into the average magnitude of errors, independent of their direction. By calculating the average absolute difference between the predicted and actual values, MAE offers a robust measure of the model's accuracy. It allows for a clear assessment of the typical error magnitude and helps quantify the overall performance of the model.

While R^2, RMSE, and MAE are of equal importance in this evaluation, additional metrics such as cosine similarity and the Pearson correlation coefficient have also been considered. Cosine similarity assesses the similarity between predicted and actual values, regardless of their magnitude, while the Pearson correlation coefficient measures the linear relationship between two variables. These metrics contribute to the comprehensive evaluation of the model's performance and provide valuable insights into its predictive capabilities.

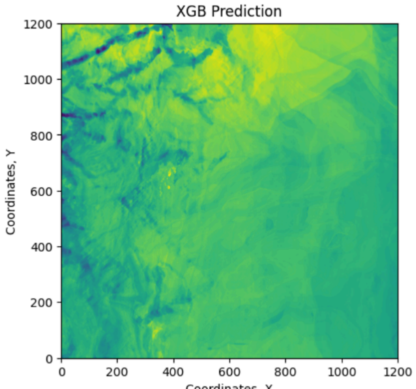
By placing equal emphasis on R^2, RMSE, and MAE, the evaluation ensures a thorough assessment of the model's goodness of fit, precision of predictions, and average magnitude of errors. This approach acknowledges the importance of MAE in quantifying the model's accuracy and its impact on the overall evaluation. Alongside R^2 and RMSE, MAE serves as a primary benchmark for accurately evaluating the model's performance. This comprehensive analysis allows for informed decisions and reliable conclusions based on the core aspects of model evaluation.

Results

The performance of the reduced XGB model was evaluated using various metrics. The model exhibited exceptional performance, surpassing all other models in our analysis. The achieved results, including a Root Mean Square Error (RMSE) of 32.680, a Mean Absolute Error (MAE) of 22.273, and an impressive coefficient of determination (R^2) value of 0.96, highlight the effectiveness of the XGB model.

It is worth noting that these results far exceeded our expectations and surpassed the performance of alternative models that were considered. The high accuracy and precision of the XGB model provide valuable insights into the underlying data, enabling us to make informed decisions.

In addition to the remarkable predictive accuracy, the XGB model also contributed to the creation of an exceptional topography map. The map derived from the XGB model showcases clear and detailed features, further illustrating the model's capability to capture and represent complex spatial patterns.



Overall, the results obtained from the reduced XGB model not only establish it as a leading contender among alternative models but also contribute significantly to our understanding of the underlying phenomenon. The outstanding performance of the XGB model, coupled with its ability to generate visually appealing topography maps, presents an invaluable resource for further analysis and decision-making processes.

Conclusion

In conclusion, this write-up presented the process of developing and evaluating an Extreme Gradient Boosting (XGB) model for predicting topography. The preprocessing steps included feature selection, scaling, and data splitting, while the modeling phase focused on optimizing the XGB algorithm by carefully tuning key parameters. The evaluation of the model's performance emphasized the importance of metrics such as R^2, RMSE, and MAE, which collectively demonstrated the exceptional accuracy and precision of the XGB model.

The results obtained from the XGB model exceeded expectations, outperforming alternative models and providing valuable insights into the underlying data. The creation of a detailed and visually appealing topography map further highlighted the model's effectiveness in capturing complex spatial patterns. These findings solidify the XGB model as a superior choice for predicting topography and contribute to a better understanding of the phenomenon.

Moving forward, the insights gained from this research can be applied to various fields that require accurate topographic predictions, such as environmental planning, infrastructure development, and natural disaster management. The success of the XGB model showcases the potential of machine learning algorithms in tackling complex geospatial problems. Future studies can explore further improvements to the model by considering additional features and incorporating more advanced techniques.

References

1. Faruque, O., Alam, H., Hossain, E., & Hussein, A. (m.d.). From Data to Topography: Deep Learning Approach to Predict Ice Bed. ms, University Maryland Baltimore County.