Final Process

Katie Yi, [katieyiowa@gmail.com](mailto:katieyiowa@gmai.com), July 11, 2023

Final Model: Extreme Gradient Boosting (XGB)

# Contents

[**Contents 1**](#_qrckzyqlvnt)

[**Folder Set Up with Generated Files 2**](#_fh0y7exj8rgo)

[**Final Processing Steps 3**](#_pjwq5benoqkb)

[Step 0: Merging 3](#_c6l1au6nifif)

[Step 1: Derive Velocity Magnitude Feature 3](#_var5z56low4m)

[Step 2: Modeling 3](#_f9kve1gu9vvd)

[**Final Model 4**](#_lfiwpitbf5gw)

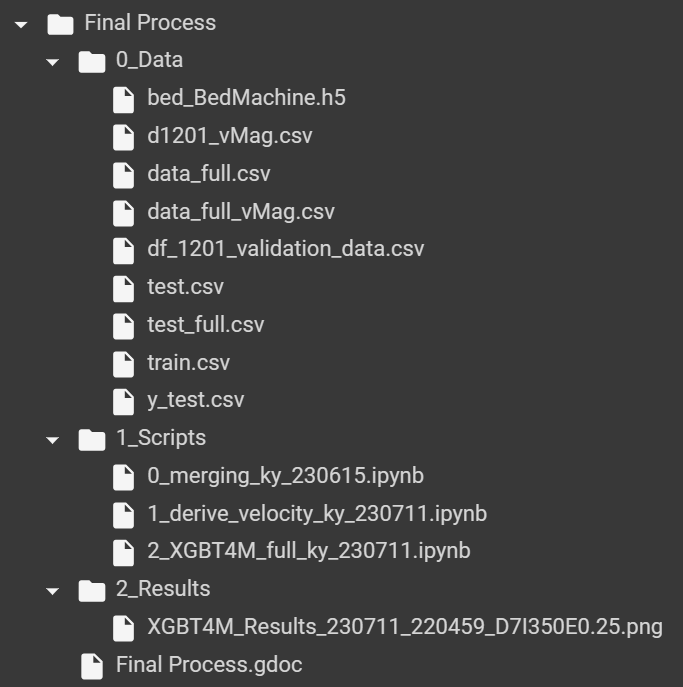
[Parameters 4](#_pmjec6bv2t50)

[Statistics 4](#_3tldvftjasxr)

[Visualizations 5](#_4l4z4ygfqozr)

[Maps 6](#_vgoz6hkjmgye)

# Folder Set Up with Generated Files



# 

# Final Processing Steps

Note: Change the main file path (“mainPath” variable) as needed in the "paths” cell; ensure packages are installed

Raw Starter Files: (currently uses the Nearest Neighbors interpolation)

* Adjust data inputs as needed for testing
* train.csv
* test.csv
* y\_test.csv

## Step 0: Merging

File: [0\_merging\_ky\_230615.ipynb](https://colab.research.google.com/drive/1KuOLbvd-U_1w7alLSONvqrBQgFPTJAMA?usp=drive_link)

Input: raw starter files

Output: (to data folder)

* Test\_full.csv
  + This is the merged test only, this file is not actually used in any scripts
* Data\_full.csv
  + This is the merged train, test, y\_test files.
  + This is the most important file in the scripts.

## Step 1: Derive Velocity Magnitude Feature

File: [1\_derive\_velocity\_ky\_230711.ipynb](https://colab.research.google.com/drive/1PkyyGw7ZU1OCpV6O-yo_dlwsASufvTO3?usp=drive_link)

Purpose: calculate the velocity magnitude of iceflow at (x,y)

Input:

* data\_full.csv
* df\_1201\_validation\_data.csv

Output:

* data\_full\_vMag.csv
* d1201\_vMag.csv

## Step 2: Modeling

File:  [2\_XGBT4M\_full\_ky\_230711.ipynb](https://colab.research.google.com/drive/1jkmRaVhnQruiuW-Gq8GRJdlTj8QOygCp?usp=drive_link)

Input:

* data\_full\_vMag.csv
* d1201\_vMag.csv
* bed\_BedMachine.h5 #1201 physics model

Output: (Maps to 2\_Results folder)

* 2\_Results/XGBT4M\_Results\_<datetime>\_D<depth of tree>I<number of iterations>E<assigned eta>.png

# Final Model

## Parameters

* Standard Scaling
* Generated seed = 168
* 60-40-20 train-test-validation split
* max\_depth= 7
* n\_estimators= 350
* min\_child\_weight= 0.25
* subsample= 0.8
* eta=.25

## Statistics

Training time:

CPU times: user 3min 36s, sys: 323 ms, total: 3min 36s

Wall time: 2min 8s

Model Prediction Beginning

Model predicted.

Transform data back to original scale.

CPU times: user 3.84 s, sys: 6.45 ms, total: 3.84 s

Wall time: 2.73 s

Validation stats statements.

RMSE: 32.284464365975296

RMSE Percentage: 12894.338512283366

Mean Absolute Error: 22.13960268173217

Mean Absolute Percentage Error: 1.0216433060319823

R^2 Score: 0.9673897122924211

CPU times: user 1.21 s, sys: 0 ns, total: 1.21 s

Wall time: 659 ms

Predicting 1201

Time taken: 12.964ms

Predicted 1201 compared to the Physics Model

Euclidean Distance: 96433.46875

Cosine Similarity: 97.688%

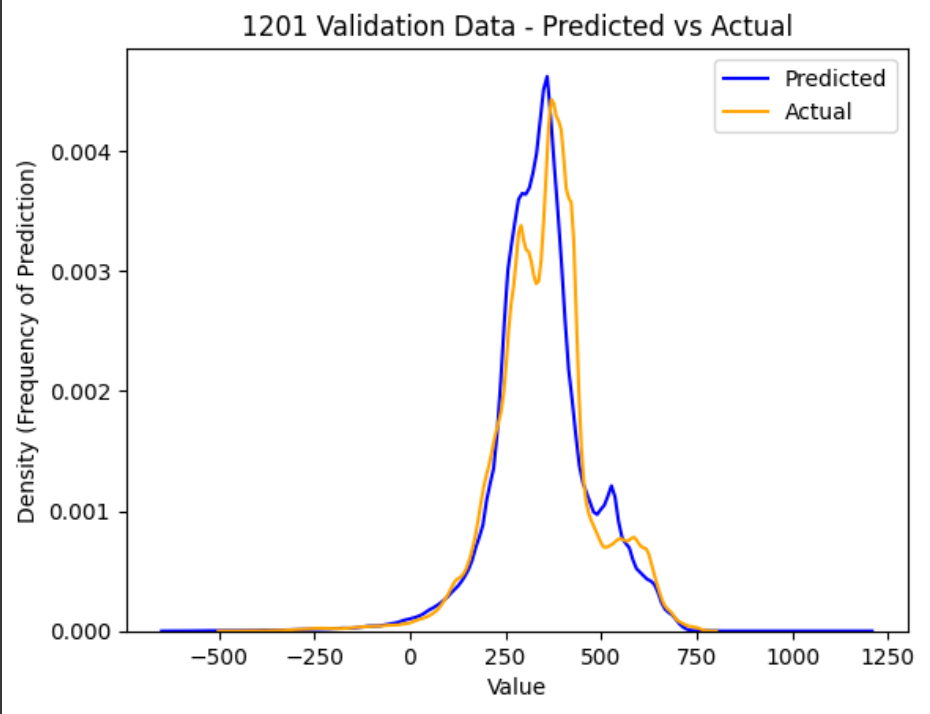
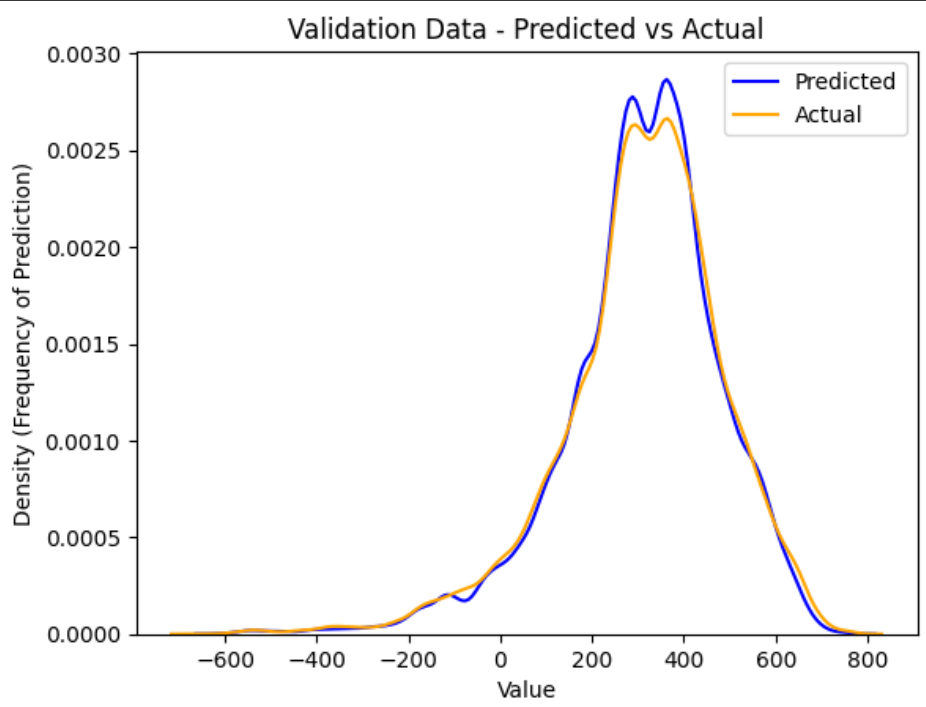
Pearson Correlation Coefficient: 80.611%

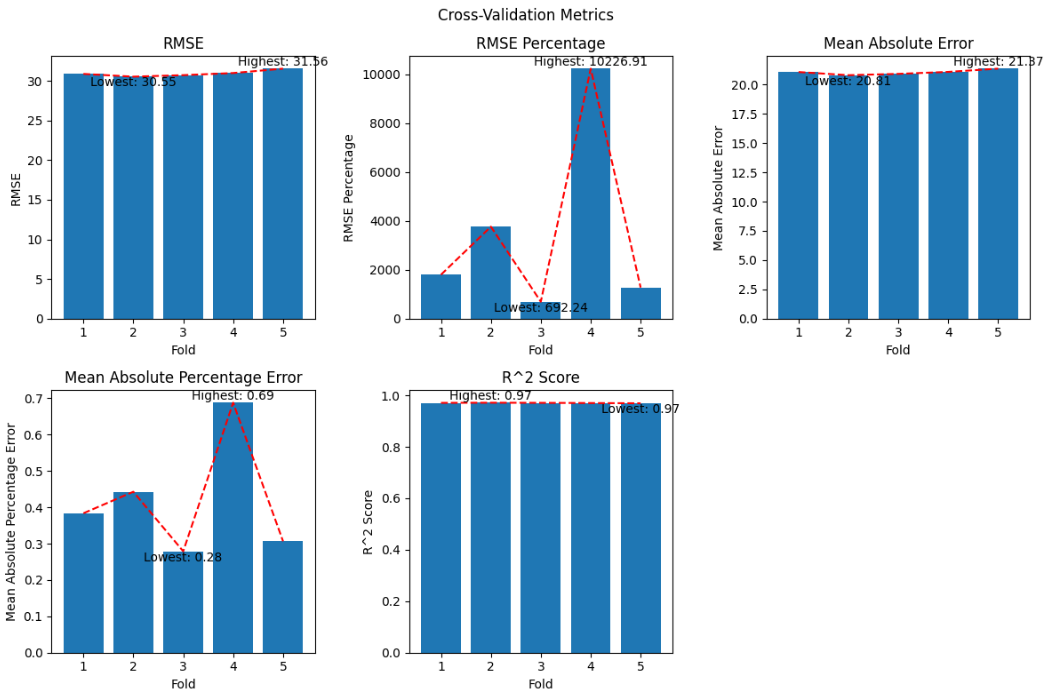
KFold Cross Validation; k = 5

CPU times: user 24min 51s, sys: 1.91 s, total: 24min 52s

Wall time: 15min 11s

## Visualizations





RMSE: [30.55, 31.56]

MAE: [20.81, 21.37]

R^2: .97

Jumps in MAPE & RMSE are due to the large range we are working with - this is common.

## Maps

Solution for dataset 1-2 assignment of prediction and physics model:

Data set 1 is the predicted data and data set 2 is the physics data.

