DS 740

Final Project

Introduction:

Over the past 20 years, the U.S. Army Corps of Engineers Rock Island District Navigation and Operations section has been tracking dredging events within the Mississippi River and Illinois Waterway. Dredging and the maintenance of a navigable channel is one of the key missions of the Rock Island District. One of the challenges with meeting the demand for dredging, is that it is often a reactive activity that occurs following a survey that identifies an area of shoaling. To best allocate resources towards the efficient operation of dredge plants, a better understanding of dredge material volumes along the river is crucial. The purpose of this project is to see if river gages can be used to predict dredge volumes within the river.

The data used includes USACE Rock Island Districts internal dredging and survey datasets, coupled with historic river gauge readings (Rivergages.com) at the date of the survey as well as 7 days prior and 14 days prior to a survey. After cleaning, the dataset for this project contained 6,533 observations and 159 variables. Due to the highly dimensional nature of the dataset, PCA was used to first explore the variables followed by XGBoost to develop a model to predict dredging quantity volume from river gage forecasts. These forecast windows are often the best data available for dredge coordinators to base dredging and survey decisions prior to a grounding.

Data Preparation and Cleaning:

To prepare the data for analysis, missing values were interpolated for the river gauge readings. In order to keep the interpolation window small, an average interpolation was only completed if an empty reading was between two observed readings. Any other NAs were omitted since it was harder to interpolate between multiple-day gaps as the median would not have been reflective of the state of the river. In addition to interpolation, the survey and dredge data was filtered for the intended window of analysis from April 1999 through September 2021 to match the resolution of the gage readings. Additional filters were used to ensure the data was for the Illinois River and the Mississippi River as well as had pools within Rock Island District, the data also was filtered for routine surveys as well as routine channel dredging as that was the focus of the project.

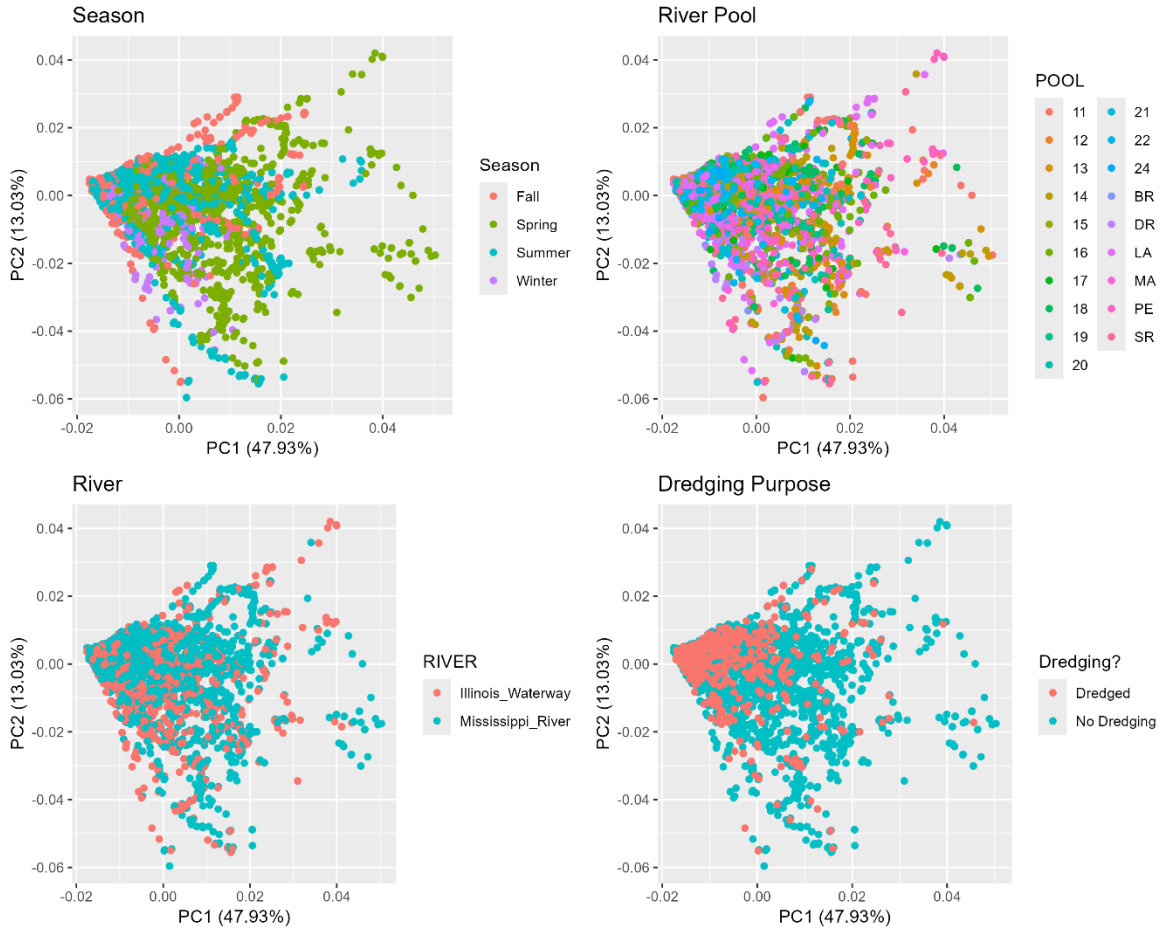
Once the data was cleaned, gage data was joined to the survey and dredging data. An additional join was completed for observations from 7 days prior to the survey and 14 days prior to the survey to represent a forecast window. While methods like XGBoost can handle gaps in the data, PCA cannot, which is why the values were interpolated. After doing

some final checks for NAs, a “Season” function was created to classify the data into seasons to support interpretation.

Principal Components Analysis

PCA was used to see if there were any discernable clusters within the variables of the data. With over 100 gages across two river systems, we hypothesized that there would be a potential cluster between the two river systems. There were no obvious clusters between the observations based on the full dataset and when visualized across the different metrics of “Season”, “Pool”, “River”, and “Dredge Type” (Fig. 1). There does seem to be some form of grouping along PC1 of “Dredge” to “No Dredging” points, however the observations appear to be randomly distributed as there still is some overlap within the PCA space. The loadings do seem to have a split between the Mississippi River Gages and the Illinois Water Way Gages which could explain PC2, however; since PC2 only captures 13% of the cumulative variation, this isn’t a very strong signal within the data.

Fig 1.



After seeing the split between the Mississippi River and Illinois Water Way loadings (Fig 2), I decided to split my data along those river systems and see if we get a better separation of the clusters (Appendix A).

Fig. 2

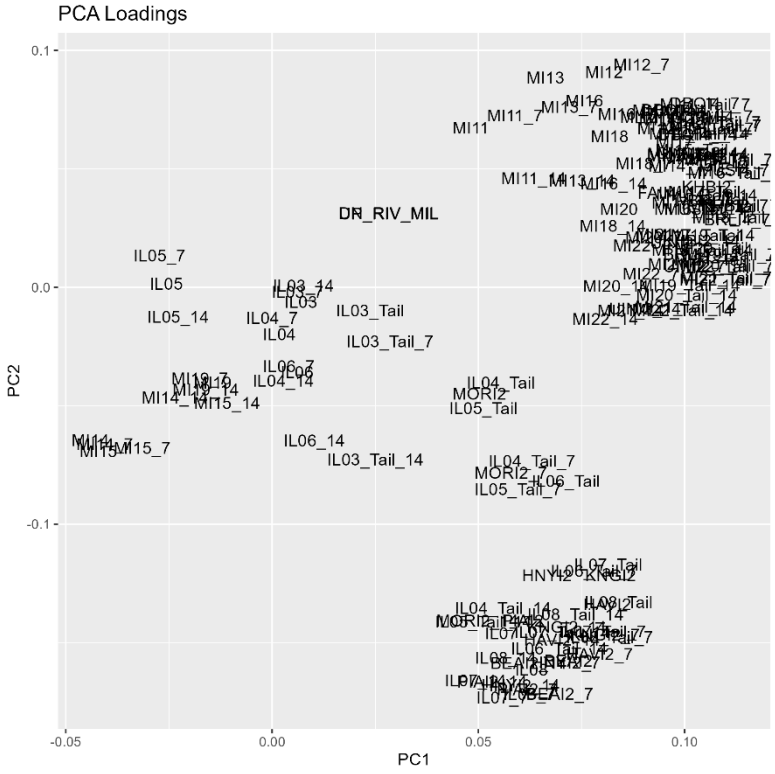
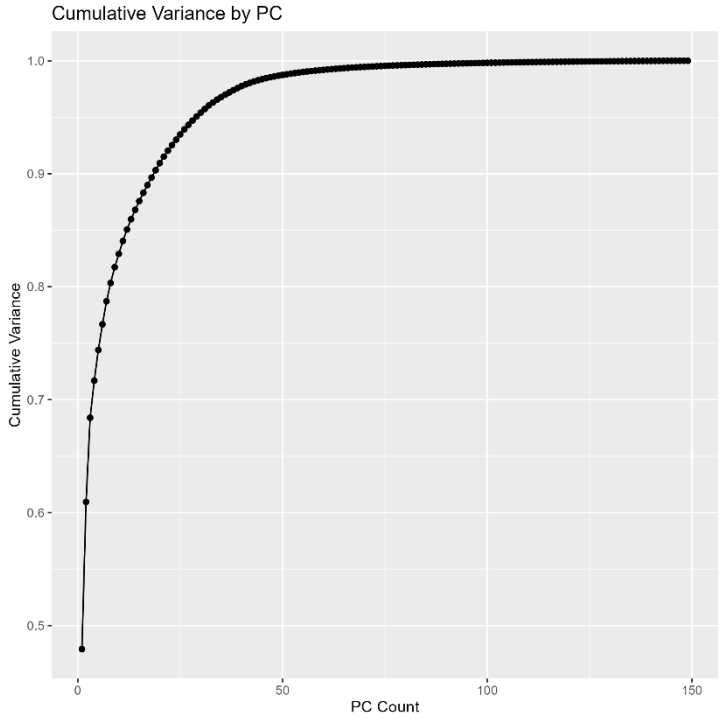


Fig. 3

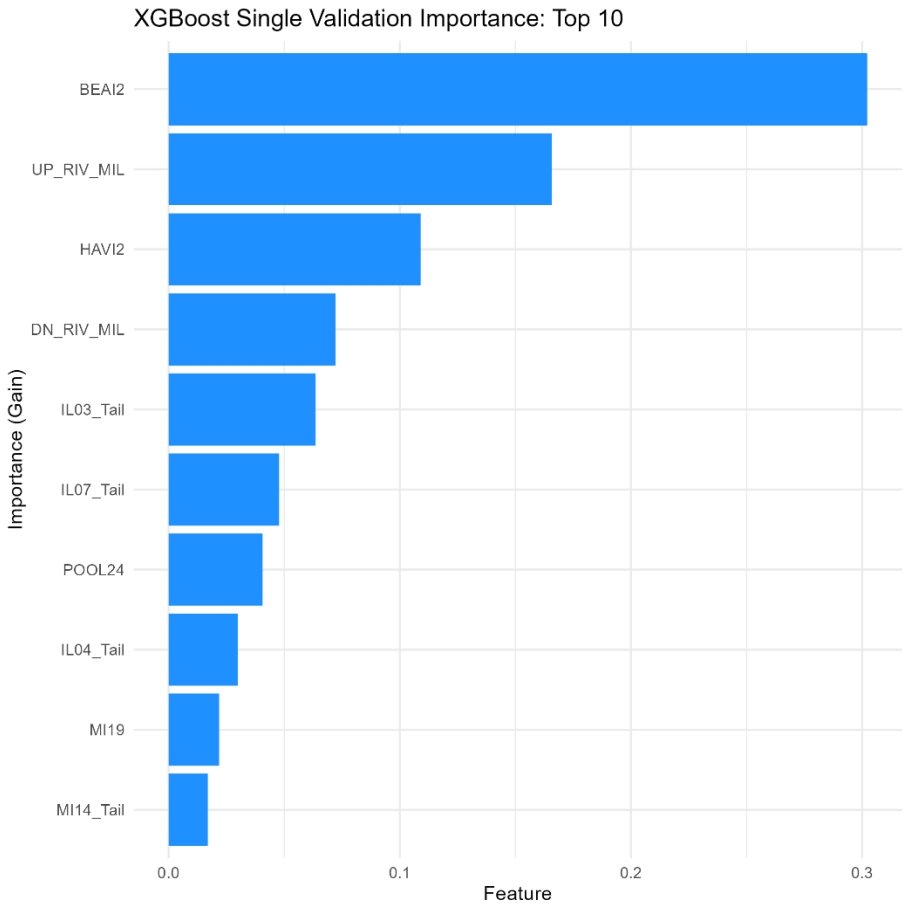


When looking at the cumulative variance plot for the PCA (Fig 3), we can see that the optimal number of principal components is ~40 which again shows that PCA is having a hard time fitting the data. This could also be due to having 150 different covariates.

XGBoost

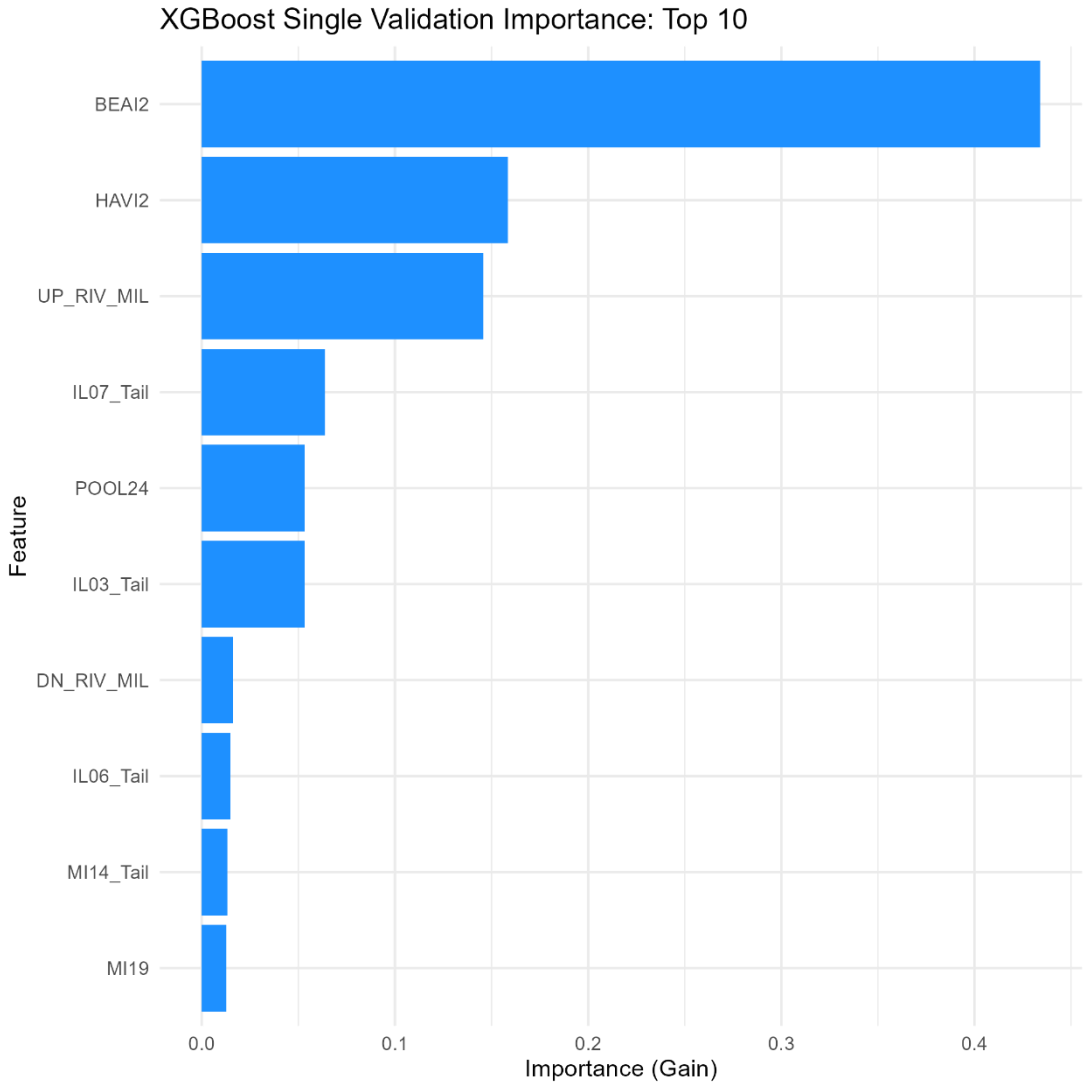
For XGBoost, the model was run on the non-interpolated data. This is due to the fact that XGBoost is able to handle missing data and outliers. To prepare the data for XGBoost, character fields that were to be used for prediction were factored and fields that were not useful for the analysis were omitted. Using a single 5-fold cross validation, the initial XGBTree model was run using caret. Nrounds and the max depth were tuned to cover 15,25,50, and 100 rounds and depths of 1:10. The number of rounds were increased as well as the depth because, as learned through the initial PCA exploration, there weren't many clear linkages between the gage data and the volume dredged response variable. Allowing the model to run over more depths and for more rounds allowed it to have better flexibility in fitting.

Fig.4



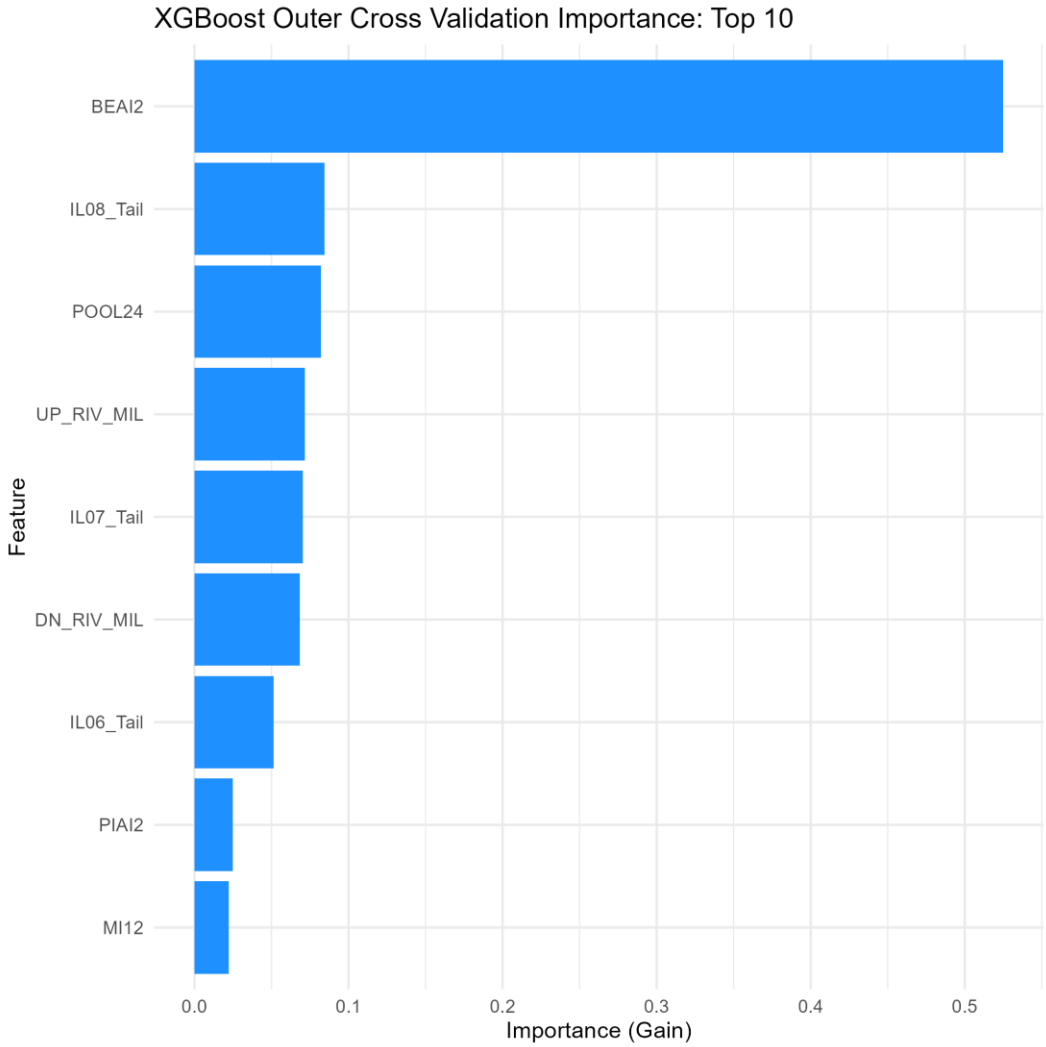
For the single layer of validation, the final values values used for the model were nrounds = 25, max_depth= 1, eta = 0.3, gamma = 0, colsample_bytree = 1,min_child_weight = 1 and subsample = 1 and RMSE, which we will be using to assess model performance, was 10539.48. R^2 was also low, which indicates a poor model fit. When comparing model predictions to the data, RMSE was 10932.81. The top three important variables (Fig 4) were river gage BEAI2 observations, river gage HAVI2 observations, and the upper river mile of the cut. Both BEAI2 and HAVI2 are on the Illinois Waterway and while they may be effective on that system, they may not be successful at predicting volumes on the Mississippi River which could have resulted in the poor model performance of XGBoost.

Fig. 4



After completing an outer layer of validation using 5 folds, the best model was the model with the values nrounds = 15, max_depth = 1, eta = 0.3, gamma = 0,

colsample_bytree = 1, min_child_weight = 1 and subsample = 1 with an RMSE of 10679.81. When fitting the model to the data, the RMSE was10988.37 which is even slightly worse than the single validation model. When looking at the important variables (Fig 5), the number one variable remained BEAI2 which actually had a higher gain than the single CV. A lot of the other important variables shifted, indicates the unpredictability of the model as there is little consistency between single and outer cross validation.



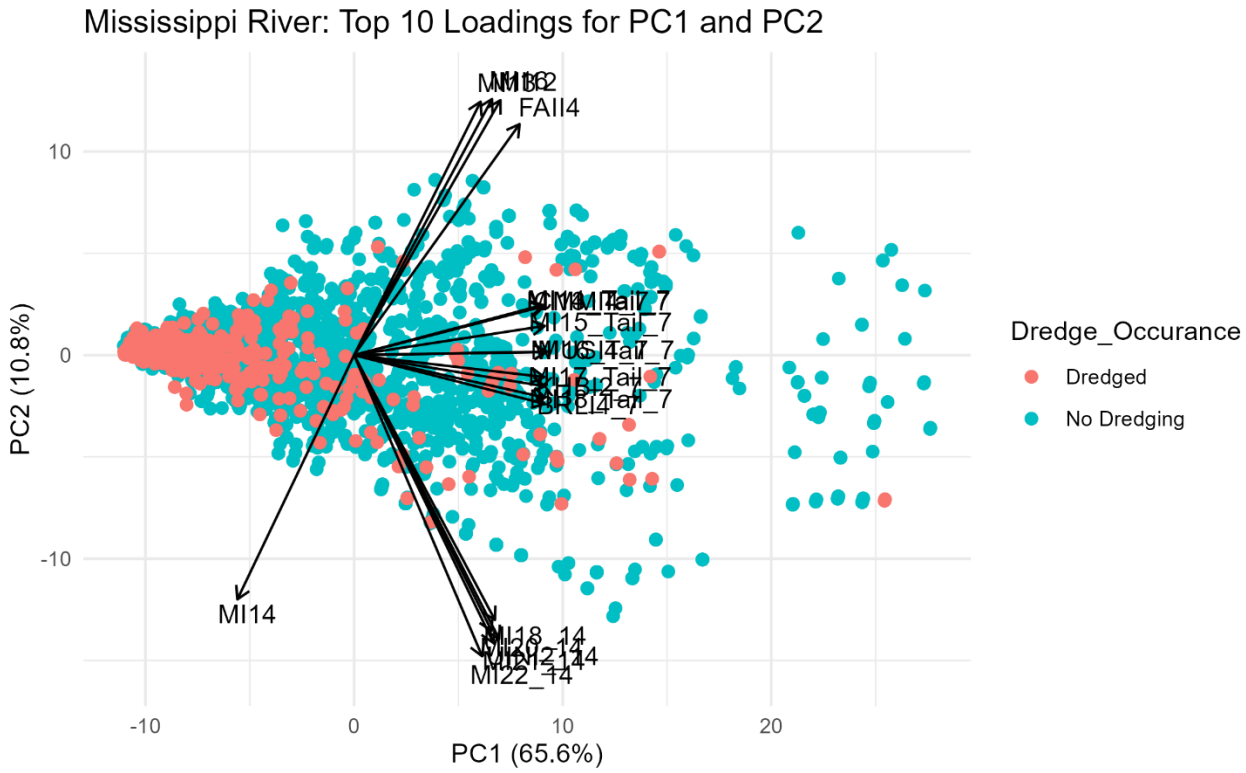
When fitting the best model from the single layer of validation to the data, the final RMSE was 10904.96 for PCA, the final RMSE was 10887.04. When looking at the average volume dredged in the dataset, the average is 14,531. The RMSE is high compared to an average channel dredging job which indicates that the models do a poor job of modeling dredge volume based on observed, 7 day, and 14-day gage readings.

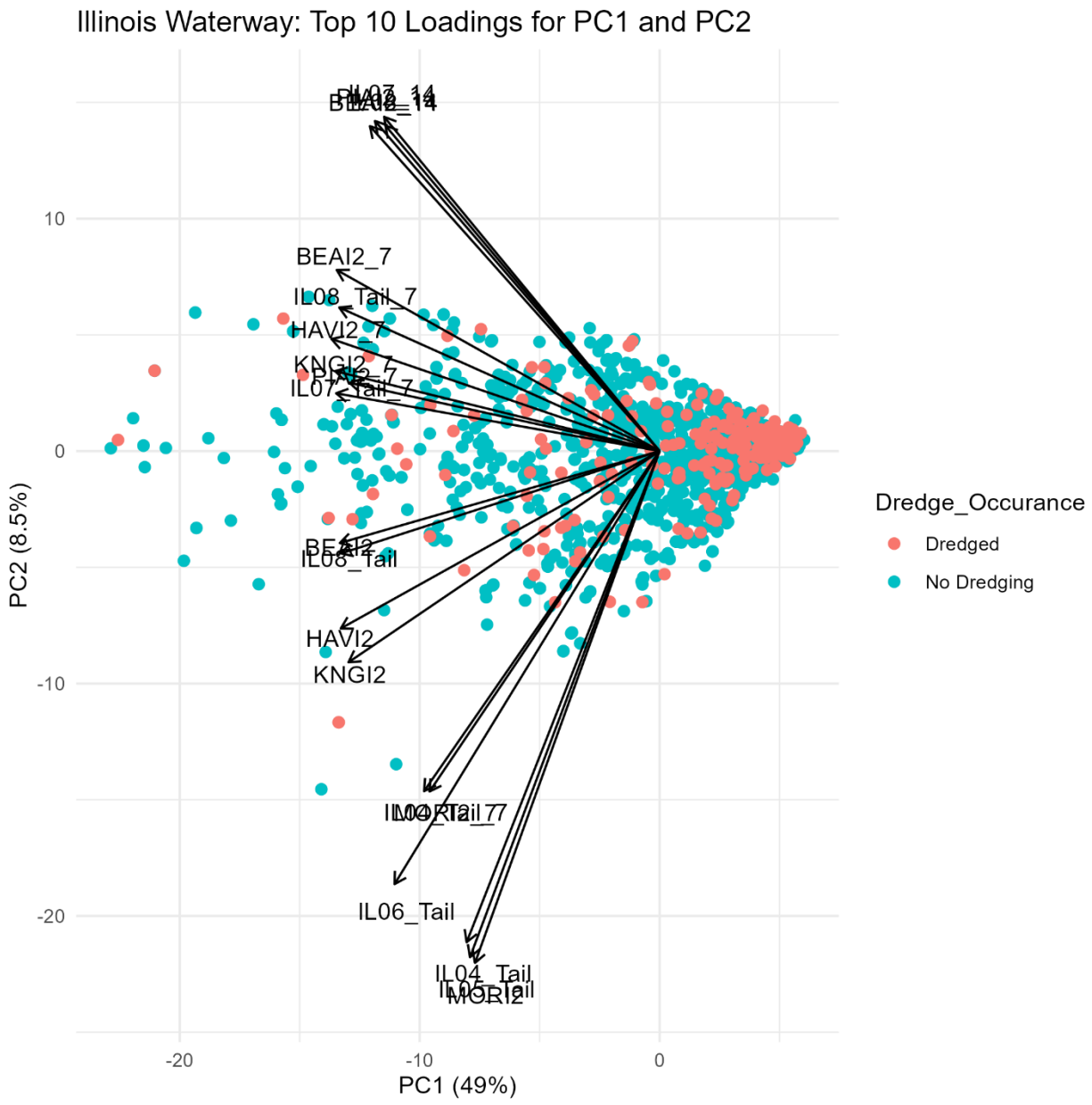
Conclusion:

Main-channel river gage observations, 7 day, and 14 day forecasts are poor predictors of dredging volume on their own. This approach was the first of many in an ongoing project to use machine learning and deep learning models to predict dredge material volume within the Mississippi River and Illinois Waterway. Future work will focus on bringing in additional predictors, like tributary river gages, which can capture incoming sediment loads to the system, and measures of river geomorphology to better constrain and develop a predictive model.

Appendix

PCAs for Mississippi and Illinois Waterways




Illinois Waterway: Top 10 Loadings for PC1 and PC2
BEAI2_7
IL08_Tail_7
HAVI2_7
KNGI2_7
IL07_Tail_7
BEAI2
IL08_Tail
HAVI2
KNGI2
IL06_Tail
IL04_Tail
MORI2
10
0
-10
-20
-20
-10
0
PC1 (49%)
PC2 (8.5%)
Dredge_Occurance
Dredged
No Dredging

XGBVarImp plots by River