Classification of PNW Wetlands by Carbon Sink Potential

CPTS 315 Project

Author: Pat McCornack

Date: 12/12/2022

Pat McCornack, 2020

Introduction

Nature based climate solutions are an approach to drawing down atmospheric CO2 levels to mitigate climate change. While their implementation has been limited in the United States, there has been increasing interest in this approach as one pathway towards a net zero emissions future Carbon sequestration in coastal wetland soils, a form of "blue carbon", is of particular interest because of their significant potential for long-term carbon storage (Novick et al., 2022) and their role in flood mitigation (Li et al., 2018). Around 85% of the west coast's tidal wetlands have been lost to activities such as development (Poppe and Rybczyk, 2021). The potential for wetland restoration as pathway for carbon sequestration is high, but the true effects of coastal wetland restoration are poorly understood due to a paucity of data (Kauffman et al., 2020). Quantifying the carbon sequestration potential for coastal wetlands is an important precursor to restoration as this information will inform which areas to prioritize. The goal of this project is to attempt to classify wetlands by their carbon sequestration potential using data on environmental conditions of the wetland. My personal motivation for this project is an interest in using data mining tools to pursue a career in data-intensive ecology with a particular focus on blue carbon.

I chose to use a K-Nearest Neighbors approach to this problem using the dataset gathered by Kauffman et al., 2020. My goal was to test multiple K Nearest Neighbor approaches with the number of neighbors determined by cross-validation and determine which provided the highest classification accuracy. The simplest K Nearest Neighbor model performed the best with an classification accuracy rate of 73.2%.

Data Mining Task

K Nearest Neighbors is a classification technique in which the entire training set is used as the model. In order to classify a new observation a K Nearest Neighbor approach searches for the k nearest neighbors in the model. These neighbors vote on what classification to assign the new observation with the result being the observation is classified as whichever classification is it "closest" to (Leskovic et al., 2014). Euclidean distance was used to measure "closeness" in this project.

The dataset on wetland environmental parameters from Kauffman et al., 2020 includes continuous variables like pore-water salinity, pH, and dry bulk density as well as categorical variables such as the vegetation class. The K Nearest Neighbors approach to classification is best used with continuous data, but it can be applied to categorical variables through the use of dummy variables. My ultimate goal was to attempt to create a model that could consistently predict whether a wetland had low, medium, high, or very high potential for carbon sequestration using these parameters. In pursuit of this, I compared various approaches to implementing K Nearest Neighbors.

Figure 1 shows the distribution of classifications of the data. This dataset is already very balanced so there was no risk of bias being introduced from training the model on an unbalanced dataset.

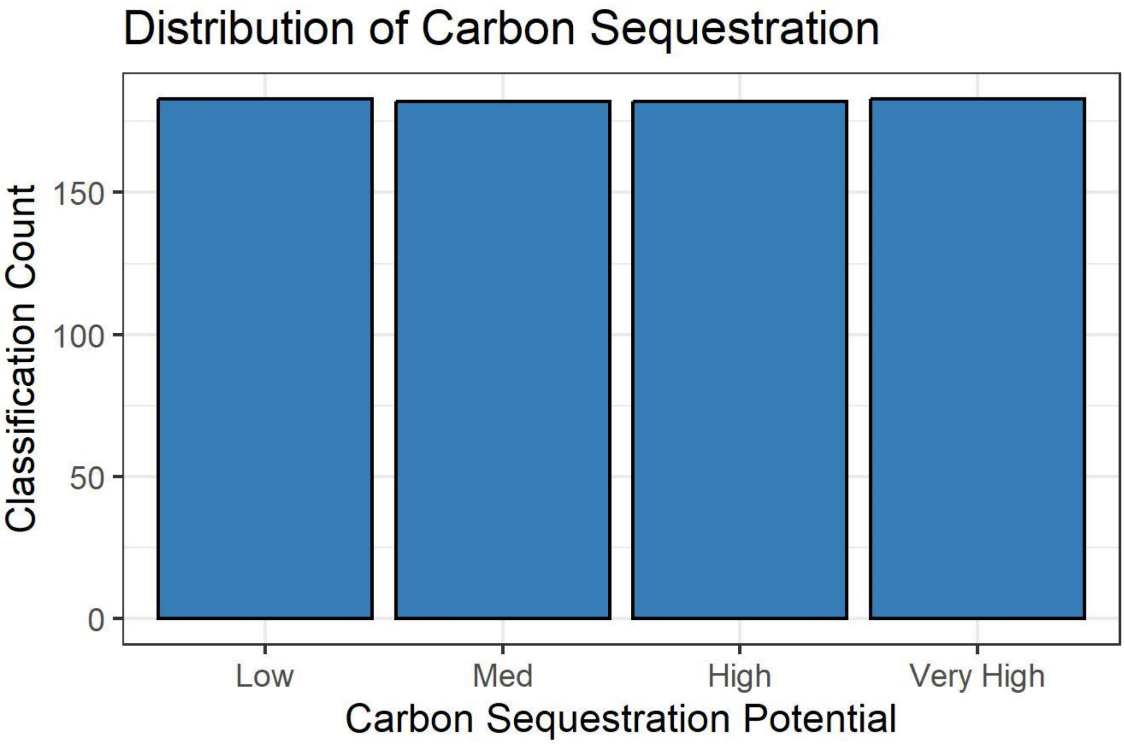

Figure 1: Distribution of classes for carbon sequestration potential.

Technical Approach

In order to answer the question of whether carbon sequestration potential could be predicted using data on the environmental conditions the dataset had to be preprocessed. The full dataset came as a set of csv files that needed to be leaned, subset, reclassified, and merged. The result was a dataframe of categorical and continuous data consisting of observations of the environmental conditions of wetlands with 730 observations of 12 variables. While more data would have been desirable, I had difficulty finding further datasets with similar parameters. The data was collected by taking soil cores at various depths from each wetland. Each soil core had a measurement of the fraction of carbon which was used to create the carbon sequestration potential classes. Each class corresponds to a quartile of the range of the data. After reshaping the dataframe, I normalized the continuous variables using min-max normalization to center and scale the data. Since the variables have different units and therefore different ranges normalization is an important step to avoid introducing bias. Additionally, I converted the categorical variables to dummy variable form in order to model them. This resulted in a dataset with 15 variables rather than 8.

After reshaping the data, I implemented three version of K Nearest Neighbors classification. One made the classification based on just the continuous variables. One used the full set of both categorical and continuous variables. The last used principal component analysis to attempt dimensionality reduction of the continuous variables before implementing K Nearest Neighbors. The motivation behind principal component analysis was that it seemed there may be multi-collinearity between some of the predictors. Principal component analysis would allow me to mitigate any multi-collinearity without dropping variables and losing information. Principal component analysis was performed using the prcomp built-in R function.

K Nearest Neighbors was implemented using the knn function from the 'class' package in R. Each implementation of K Nearest Neighbors was evaluated using 5 fold-cross validation, meaning that the full set was split into training and testing data and used to evaluate the model 5 times for each implementation.

Evaluation Methodology

For each implementation cross-validation was used to evaluate the accuracy as well as to select the optimal number of neighbors from a range of 1 to 50. Each observation was randomly assigned an index of 1 to 5 that corresponded to a fold to randomize the folds. The number of correct classifications was recorded for each fold and averaged for each number of neighbors. The average accuracy across folds was then compared across neighbors to find the number of neighbors that produced the highest accuracy for that implementation. After finding the optimal number of neighbors, the model was run with using that number to produce predictions. The accuracy of this model was recorded along with the number of neighbors for inter-model comparison. The predictions were also saved to compare their distribution against the distribution of the true labels.

Results and Discussion

Model Comparison

Table 1 provides an overview of the performance of each model. The K Nearest Neighbors implementation using only the continuous variables performed the best with an accuracy rate of 73.2%. The model with additional information incorporated as categorical variables performed slightly worse with an accuracy rate of 70.3%. The implementation where principal component analysis was performed using the continuous variables before running the classification had the lowest accuracy rate at 67.1%.

I find it surprising that the implementation using principal component analysis performed the worst. I expected that the use of principal component analysis before running the classification would reduce bias introduced to the model through multi-collinearity. Specifically, I expected there to be collinearity between the core elevation and porewater salinity variables because one would expect that cores from lower elevations would be constantly inundated by salt water while high elevation cores would have more filtration occurring.

The result that the model without the categorical variables performed better than the model with categorical variables was expected. This is not due to the categorical variables lacking relevant information. The vegetation class in particular is likely a strong productor of carbon sequestration potential (Kauffman et al., 2020). Rather, K Nearest Neighbors using a Euclidean distance metric is poorly suited to categorical variables. The number of dimensions of the categorical dataset is also expanded by the use of dummy variables, and the performance of K Nearest Neighbors degrades in higher dimensions. I included the use of categorical data in the analysis as an illustration of the limitations of K Nearest Neighbor classification. One way to address this limitation would be to consider other statistical distances outside of the Euclidean distance.

	KNN Continuous	**KNN Continuous and Categorical**	**KNN Continuous with PCA**
Maximum Accuracy	.732	.703	.671
Optimal Number of Neighbors	9	3	3

Table 1: Inter-model comparison of maximum accuracy and optimal number of neighbors.

KNN with Continuous Variables

Figure 2 show that the optimal number of neighbors for the model using only continuous variables is 9. The performance of the model very quickly drops off with additional neighbors beyond that. This model accurately predicted wetland carbon sequestration potential 73.2% of the time. Figure 3 shows the distribution of the predicted labels against the actual labels. My expectation was that bias would be introduced from multi-collinearity among variables, but this expectation is not expected here. The distribution of predicted variable is balanced. Due to the random assignment of folds, this testing dataset had a slightly larger number of "High" classifications than would be expected from the full dataset. This model does not appear to suffer from issues related to bias.

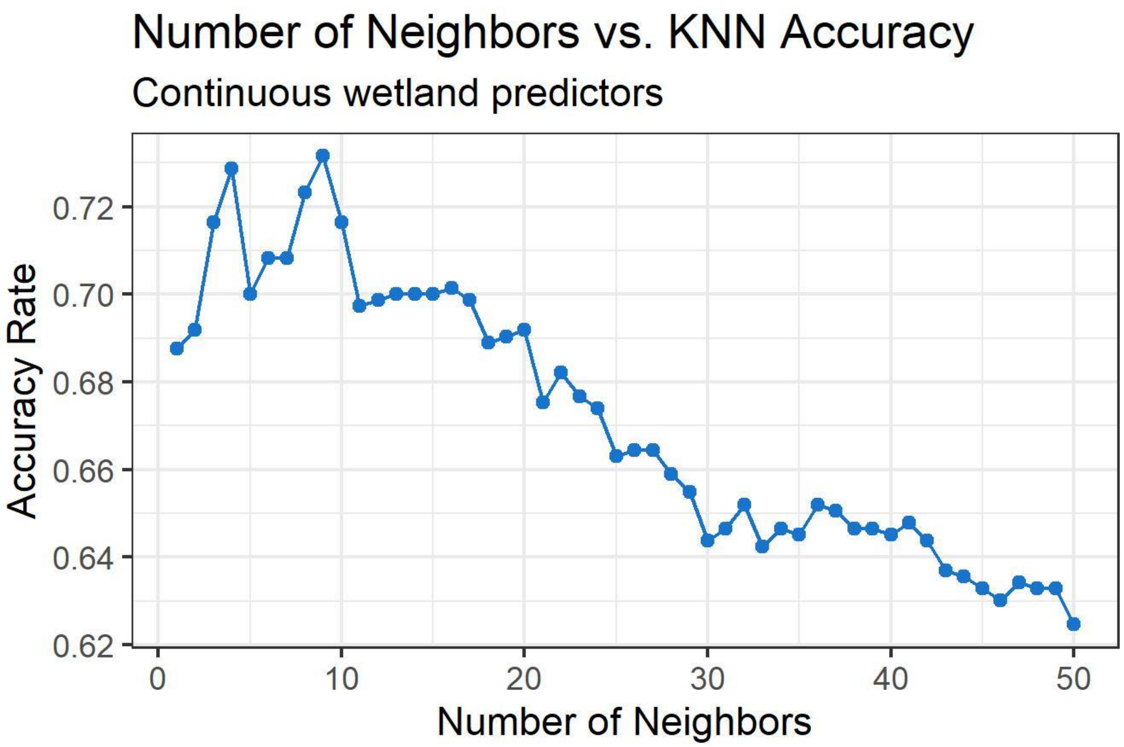

Figure 2: The error for the continuous variables only model for each number of neighbors.

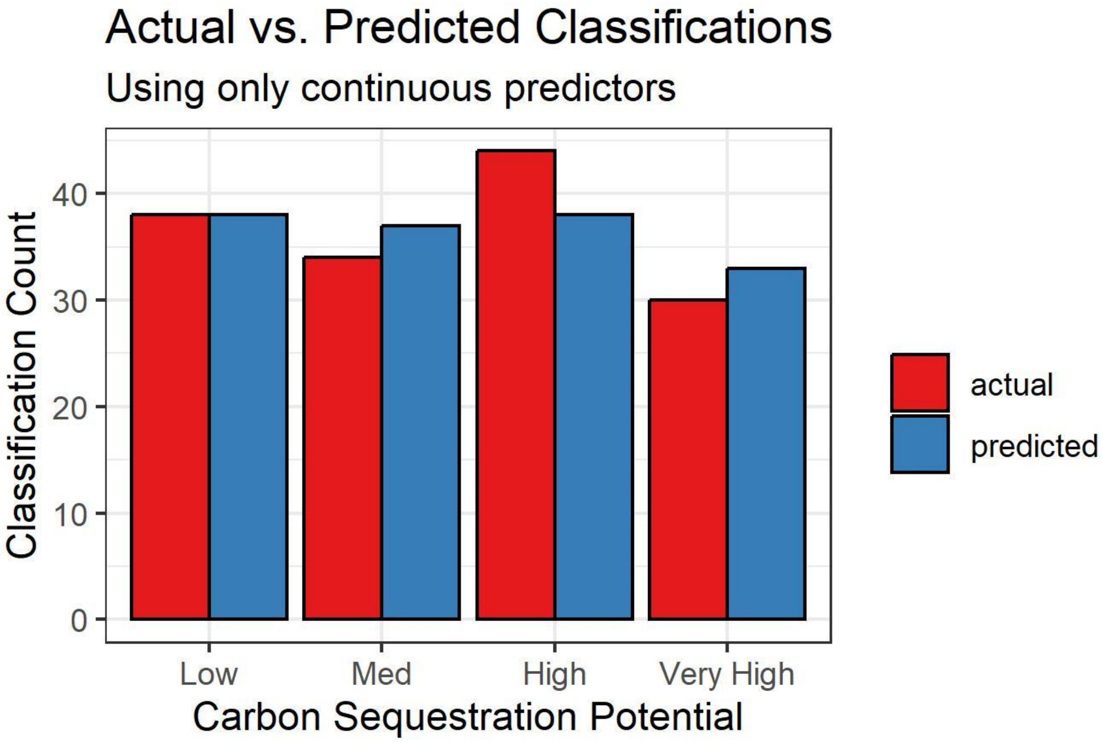

Figure 3: Predicted classifications against actual classifications using the only continuous predictors model.

KNN with Categorical and Continuous Variables

Figure 4 shows that the K Nearest Neighbors model using both continuous and categorical variables had an optimal number of neighbors of 3, after which performance sharply declines. The testing set had 146 observations of 15 variables, so I find this low number of neighbors surprising. This shows that adding the categorical variables to the analysis contributed more noise than information which is shown when the overall accuracy is compared to the continuous variables only model. Once again this is because Euclidean distance is not the proper distance metric for working with categorical variables. The poorer performance of this model is demonstrative of the fact that more data is not necessarily better, and modeling techniques need to be carefully selected to fit the data.

Figure 5 shows the distribution of predicted vs. actual labels for the continuous and categorical model. Once again, no obvious bias is observed. The increase in error is due to the increase in dimensions, not bias.

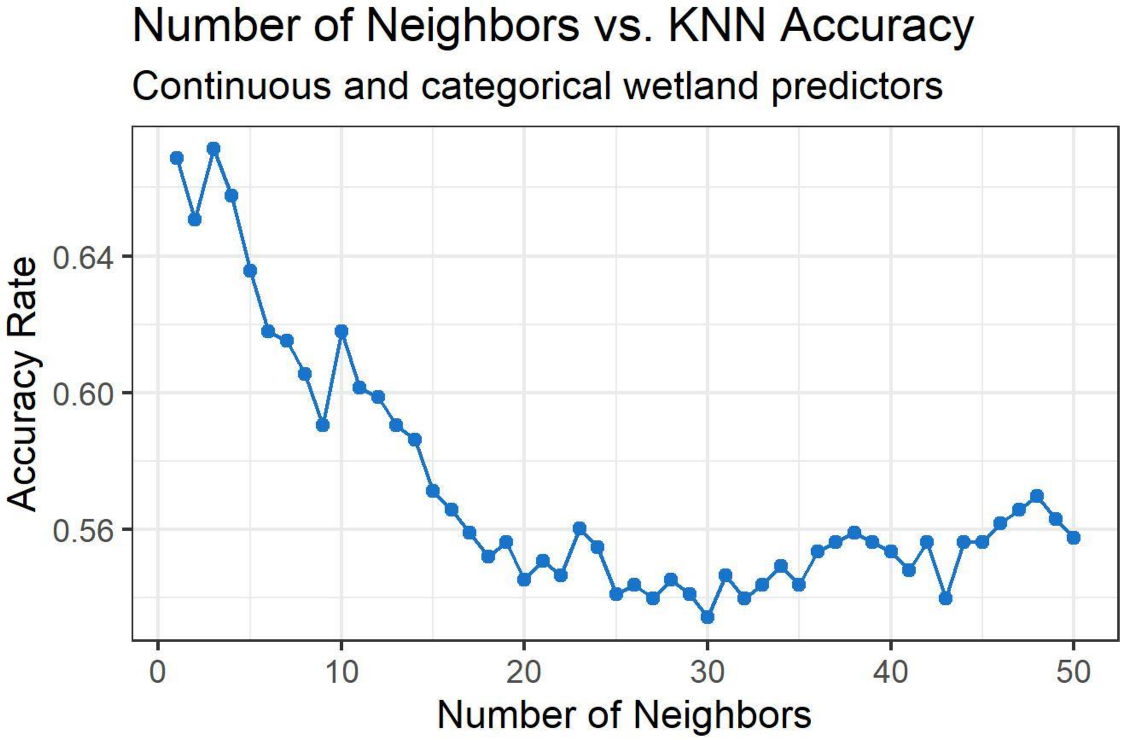

Figure 4: The error for the continuous and categorical variables model for each number of neighbors.

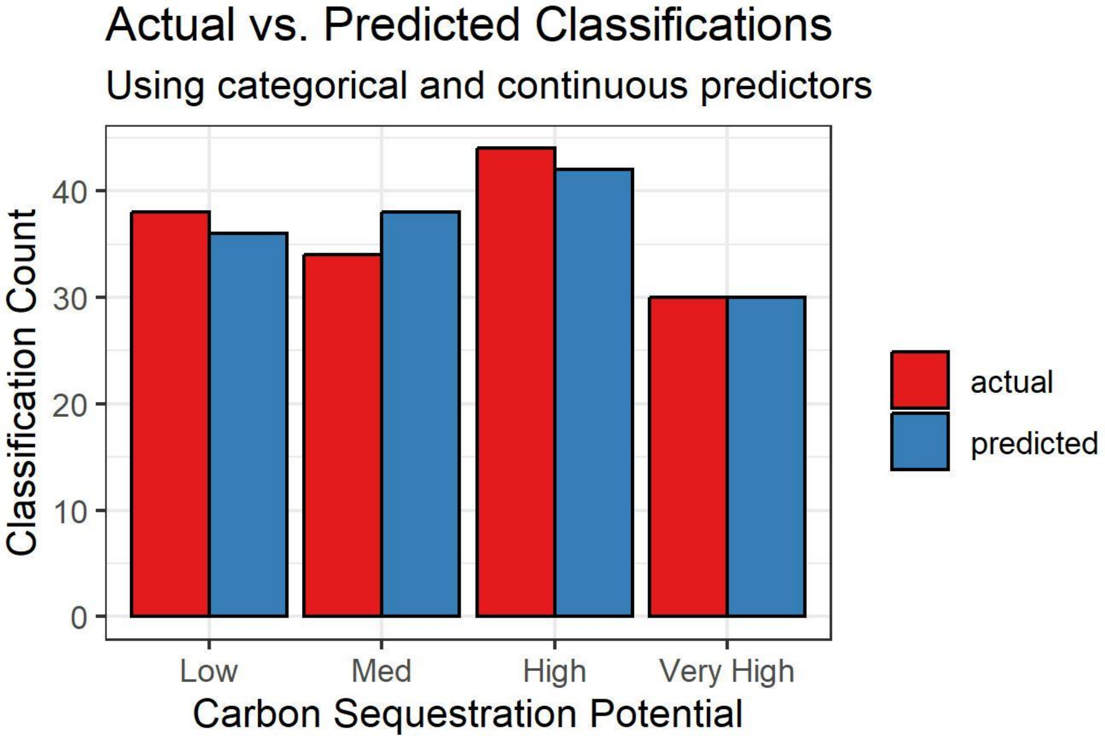

Figure 5: Predicted classifications against actual classifications using the continuous and categorical predictors model.

KNN using PCA of Continuous Variables

Figure 6 shows a similar trend to the continuous and categorical model where the model utilizing principal component analysis uses 3 optimal neighbors after which the performance quickly degrades. For this model, I used principal component analysis to reduce the dimensionality from 5 continuous variables to 3 principal components. These three principal components explained slightly over 95% of the variance in the dataset. While this model performed the worst out of the three, there is a trade-off in runtime. K Nearest Neighbors suffers from being computationally expensive with many dimensions. If this dataset had more observations, there could be benefit to be gained by reducing the dimensionality with principal component analysis to reduce runtime.

Figure 7 shows the distribution of predicted vs actual labels for model using principal component analysis on the continuous variables. Once again no obvious bias is seen in this distribution.

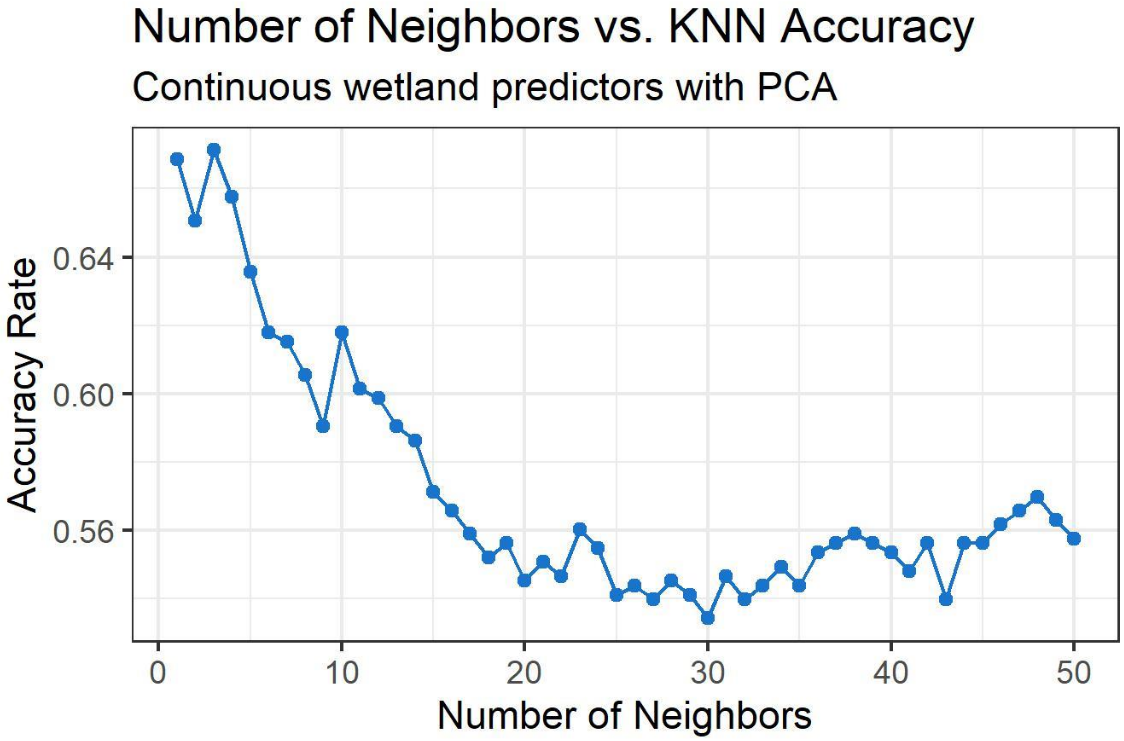

Figure 6: The error for the continuous variables model using PCA for each number of neighbors.

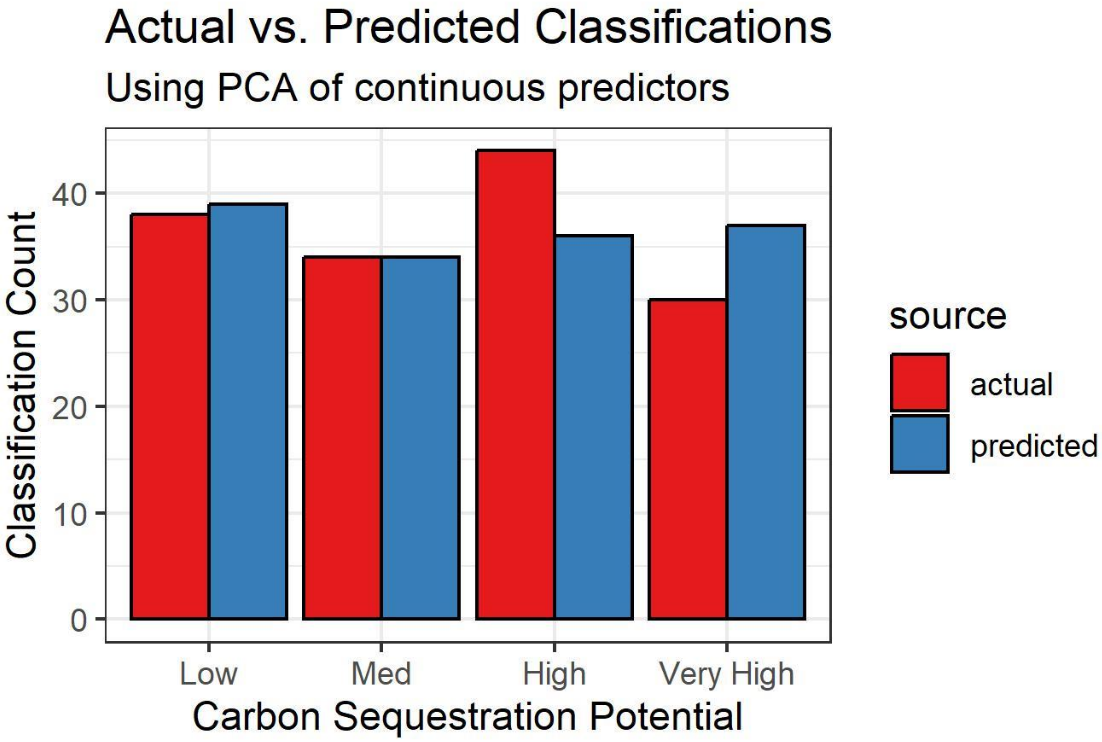

Figure 7: Predicted classifications against actual classifications using the continuous predictors with PCA model.

Conclusions

This work represents an initial effort at utilizing environmental parameters to predict the potential of wetlands for carbon sequestration. The aim of this is to provide information to land managers that allows for prioritization of resource allocation towards restoration work.

There are limitations to this work. All data used to train the models came from west coast tidal wetlands and the model should not be applied outside of this scope. Additionally, the environmental parameters used as predictors are resource intensive to collect. They require lab equipment and trained personnel to obtain. While this was meant as a proof of concept and demonstration of tools that can be used for this purpose, I plan to attempt to recreate the model using readily available data. This includes remote sensing products and eddy-covariance data collected by flux towers.

In this analysis the model that performed best was the simplest model that only utilized the continuous variables which achieved an accuracy rate of 73.2%. I believe that this could be improved by including the categorical variables using another distance metric or another classification model entirely. However, the loss of accuracy associated with the categorical variables was demonstrative of a common pitfall – more data isn't always better.

Lessons Learned

I was somewhat surprised to achieve a 73.2% accuracy rate. For a problem with 4 classification categories that is a non-trivial rate. I was also surprised by how much the model using principal component analysis suffered from the dimensionality reduction. I had actually expected it to perform better. These unexpected results are a good lesson that intuition is often wrong and the best approach is to just try things, especially when it comes to high dimensions.

In hindsight, I would have liked to have time to test other distance metrics like the Hamming distance using the K Nearest Neighbors approach. I would also have liked to implement the classification using other methods such as a support vector machine or other techniques outside of the scope of this class. This is likely a project that I will expand on as a demonstration piece for my portfolio.

Sources and Acknowledgements

Novick, K. A., Metzger, S., Anderegg, W. R. L., Barnes, M., Cala, D. S., Guan, K., Hemes, K. S., Hollinger, D. Y., Kumar, J., Litvak, M., Lombardozzi, D., Normile, C. P., Oikawa, P., Runkle, B. R. K., Torn, M., & Wiesner, S. (2022). Informing Nature-based Climate Solutions for the United States with the best-available science. Global Change Biology, 28, 3778–3794. https://doi.org/10.1111/gcb.16156

Kauffman JB, Giovanonni L, Kelly J, et al. Total ecosystem carbon stocks at the marine-terrestrial interface: Blue carbon of the Pacific Northwest Coast, United States (2020). *Global Change Biology*, 26:5679–5692. https://doi.org/10.1111/gcb.15248

Poppe KL, Rybczyk JM (2021) Tidal marsh restoration enhances sediment accretion and carbon accumulation in the Stillaguamish River estuary, Washington. *PLoS ONE* 16(9): e0257244. https://doi.org/10.1371/journal.pone.0257244

Li, Xiuzhen et al. "Coastal wetland loss, consequences, and challenges for restoration." *Anthropocene Coasts* 1.1 (2018): 1-15. http://dx.doi.org/10.1139/anc-2017-0001

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets* (2nd ed.). Cambridge: Cambridge University Press.

Code for cross validation was based on a STAT 435 lab run by Dr. Abhishek Kaul.