Relationships across Chemical Exposure Concentrations

Chloe Warren, Anna Burkholder

Advisor: Daniel Q. Naiman

Per- and polyfluoroalkyl substances (PFAS) are a group of synthetic chemicals that have been heavily used for manufacturing purposes. There are significant records of human exposure to this chemical group through air, house dust, drinking water, and food. Such exposure has been a pressing concern surrounding the PFAS group given that the majority of health implications associated are widely inconclusive as of today. The lack of historical data on concentration levels, amongst other missing properties, such as undocumented chemicals and missing lab data, creates difficulty for scientists trying to research these health concerns. Our goal was to see if we could find a promising method to fill in missing concentration levels from the 2013-2014 NHANES (National Health and Nutrition Examination Survey) laboratory data on a chemical(s) within the PFAS chemical group. We were able to achieve predictions with accuracies as high as 87.86% in our results for one chemical, leading us to believe that there are promising options for filling data gaps in other PFAS-related research opportunities. But many other predictions led to poor results, so while we see an opportunity here, it may be very niche in scope.

Lasso regression

Figure 1.1. Lasso coefficients as a function of alpha

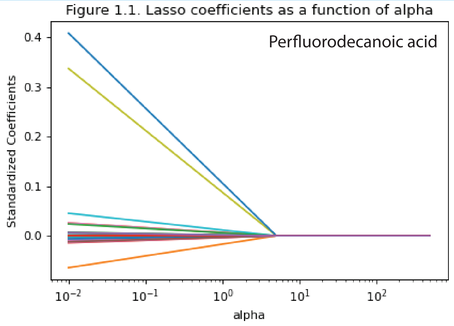


Figure 1.3. Mean square error on each fold

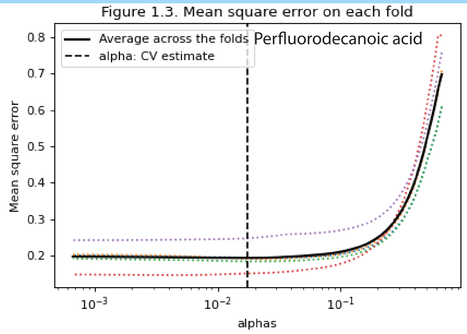


Figure 2.1. Lasso Combo vs Outside

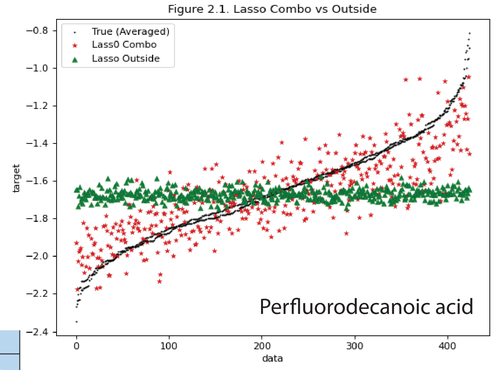


Figure 6.3. Lasso Combo vs PFAS

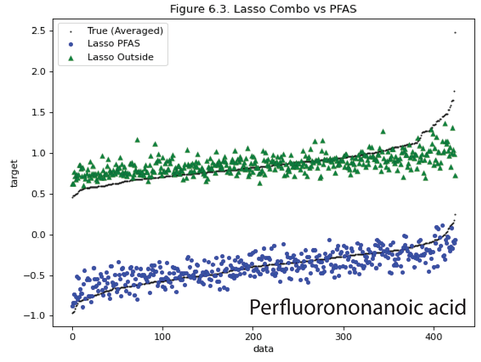
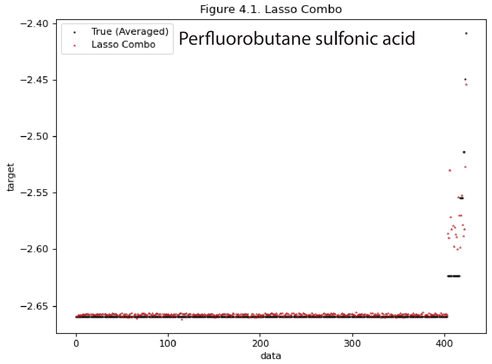


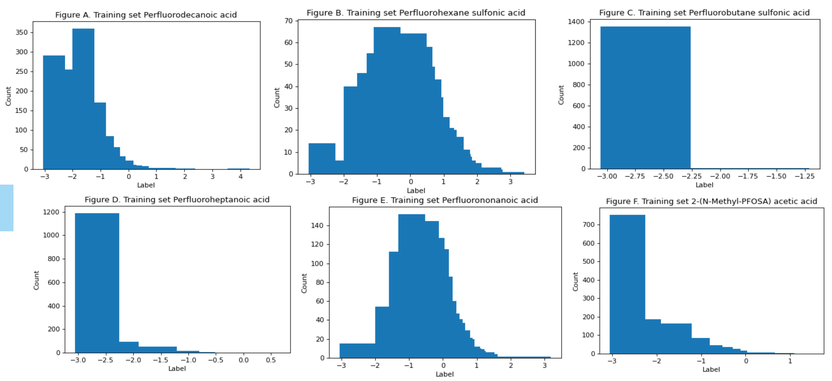
Figure 4.1. Lasso Combo



Lasso	Average difference from actual			R-Squared			MSE			RMSE		
Chemical	Combination	PFAS	Outside	Combination	PFAS	Outside	Combination	PFAS	Outside	Combination	PFAS	Outside
Perfluorodecanoic acid	0.118	0.118	0.225	0.684	0.689	0.010	0.022	0.023	0.075	0.150	0.151	0.274
Perfluorohexane sulfonic acid	0.187	0.187	0.228	0.323	0.291	0.291	0.055	0.060	0.060	0.234	0.244	0.244
Perfluorobutane sulfonic acid	0.004	0.004	0.001	0.737	0.001	0.000	0.000	0.001	0.00001	0.011	0.024	0.003
Perfluoroheptanoic acid (ng/mL)	0.078	0.078	0.011	0.008	-0.008	0.015	0.009	0.009	0.000	0.093	0.095	0.015
Perfluorononanoic acid (ng/mL)	0.108	0.108	0.137	0.639	0.627	0.313	0.019	0.018	0.040	0.137	0.133	0.200

Feature label distribution

These histograms shows a beakdown of the training set label distributions for combination dataset per chemical. (The reason why some of these values show as negative is that we took the natural log of the concentration levels during the data preprocessing phase)



Classifiers: Stepwise logstic regression and XGBoost

Chemical	Accuracy of Predictions via Stepwise Logistic Regression		
	Combination Dataset	PFAS Dataset	Outside Dataset
Perfluorodecanoic acid	87.859	83.548	70.165
Perfluoroundecanoic acid	78.118	79.839	50.753
2-(N-methyl-PFOSA) acetate	59.106	58.687	50.682
Perfluoroheptanoic acid	87.482	87.304	78.408
Perfluorononanoic acid	99.576	99.078	89.035

Stepwise logistic regression with three features gave us our best classification results for the most chemicals. This particular model had to be built from the ground up, rather than imported from a package, and we postulate that this may have contributed to its success. Furthermore, the use of the BIC as a metric for inclusion in the model proved effective for our prediction models, more so than the maximum likelihood estimate, which our alternative stepwise regression classifier uses as its inclusion metric.

Figure 1.2. Stepwise Logistic Regression 3

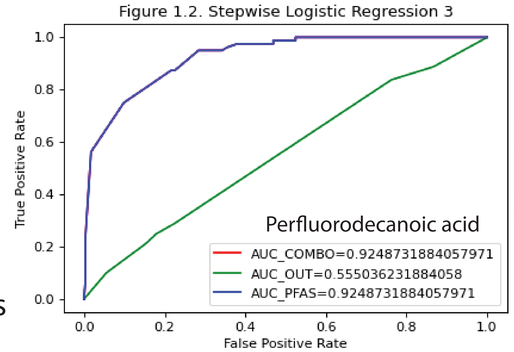


Figure 1.3. Stepwise Logistic Regression

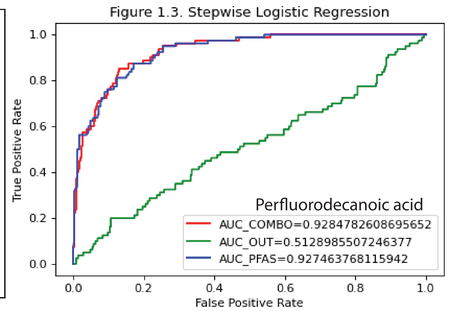


Figure 1.1. Logistic Regression

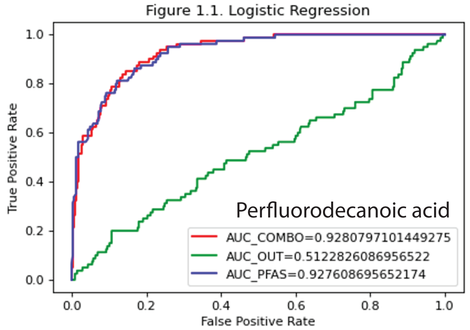
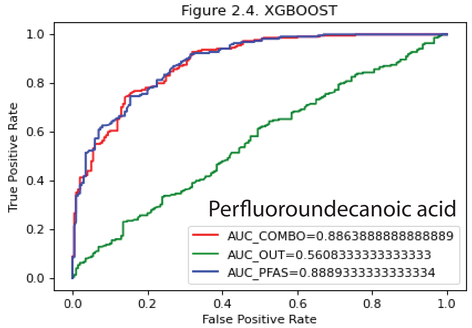


Figure 2.4. XGBOOST



Of the eight chemicals included in the PFAS group, five were selected to assess our classification models. Most of these chemicals had a relatively even distribution of binary values, however one chemical, Perfluorononanoic acid, contained a high number of of 0 values and a much lower amount of 1 values. Because of this our classification models performed exceedingly well, however this is attributed to overfitting. More interesting information can be gained by assessing our classification models against other chemicals, which we have displayed here.