The Data Open - Team 15

What actionable insights can we make to improve water quality in California?

Kabith Mahendran, Pavle Mihajlovic, Vaibhav Khaitan, Zhe Yang

September 29th 2018

1 Topic Question

Human life is impossible without water. With access to safe water, children demonstrate much better health outcomes, enabling them to focus on education and achieve more in life. While populations with the least access to good water sources are largely clustered in developing nations, recent events such as the Flint water crisis in Michigan and the California drought remind us that even in economically advanced countries like the USA, access to clean and safe drinking water supplies can be threatened.

With the topic question in mind, we produce our own question. **What actionable insights can we make to improve water quality in California?** To determine water quality we must understand water quality across the entire state. Therefore we aim to exploit a plethora of different data to ascertain water quality in the state. Utilizing the given data we will be harnessing the power environmental and societal factors to better understand how this reflects water contaminant levels. Through the use of this data we plan on creating actionable insights for the state government of California to use to reduce levels of contamination within water on a regional basis.

2 Executive Summary

We have produced a important set of actionable insights for the government of California. Through the use of drought data and economic indicators we have assembled important determinations about what factors contribute in what manner to water contamination levels.

Our first recommendation is to analyze counties and regions by level of specific chemical water contamination. At an intuitive level similar levels of specific contaminant at a county level should have important causal factors that are contributing in similar ways. This should be apparent as water ways are shared between bodies of water within state. To compare this a clustering algorithm should be utilized to determine which counties are similar to each other.

Secondly, while California is a highly drought prone state, drought seems to have effects on specific chemical-geographic combinations. It has shown that high contamination level in Uranium is more highly related with those county regions experiencing droughts. With this in mind it should be of importance for the government to put resources into detecting uranium levels in drought-prone regions.

Thirdly, while one might imagine that droughts and other environmental factors might contribute to water quality, that conclusion is wrong. Many counties within California have water problems that are economically related, much more so than drought. Those with better socio-economic status tend to have better water supplies. Placing importance on developing poorer or impoverished regions should decrease water contamination levels.

Finally, we decide to suggest a 2-tiered system to deal with water contamination results based on understanding both environmental factors and socio-economic factors to ascertain water quality results. One tier of the system should be distinctly related to environmental results with key factors like droughts being placed into concern to ascertain specific chemical contamination levels. The second tier of the system should be related to developing socio-economic factors to benefit impoverished and down trodden regions of the state that seem to be lacking investment in water quality iniatives.

3 Technical Exposition

3.1 Introductory Exploration

To better understand the state of California's water quality we want to granularize our analysis. A obvious point of exploration is a county based exploration. For each county we can take a look at the levels of various chemicals within California to better understand geographic abnormalities. We take the average contamination by chemical over the years 2000-2016 to present a stable level baseline level of contamination for each county. Between the 4 chemicals surveyed in the chemical dataset, an excess amount of any of the 4 chemicals present pose a water safety risk. With the upcoming plots we wish to reveal characteristics about the county level concentration of chemicals in the water supply.

Figure 1: Average concentration levels of Arsenic

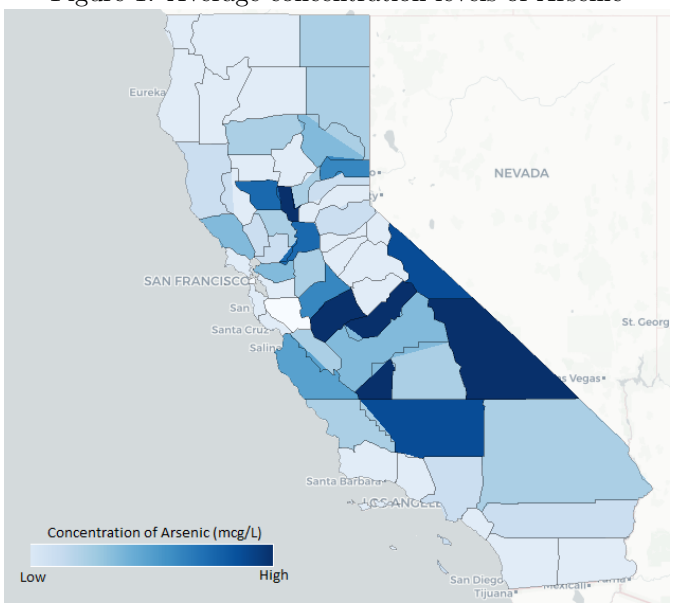


Figure 2: Average concentration levels of DEPH

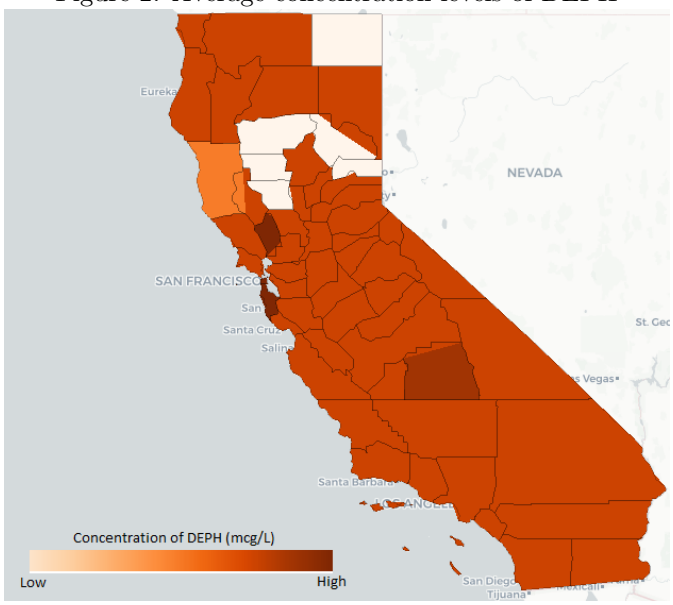


Figure 3: Average concentration levels of Nitrates

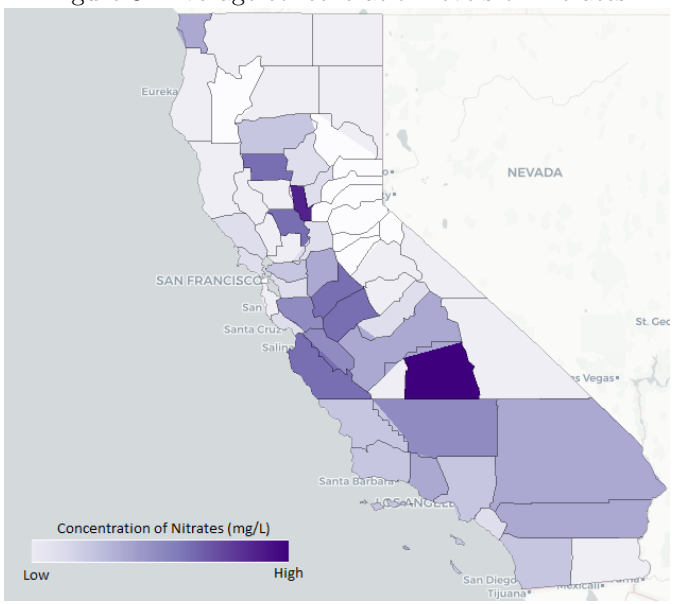
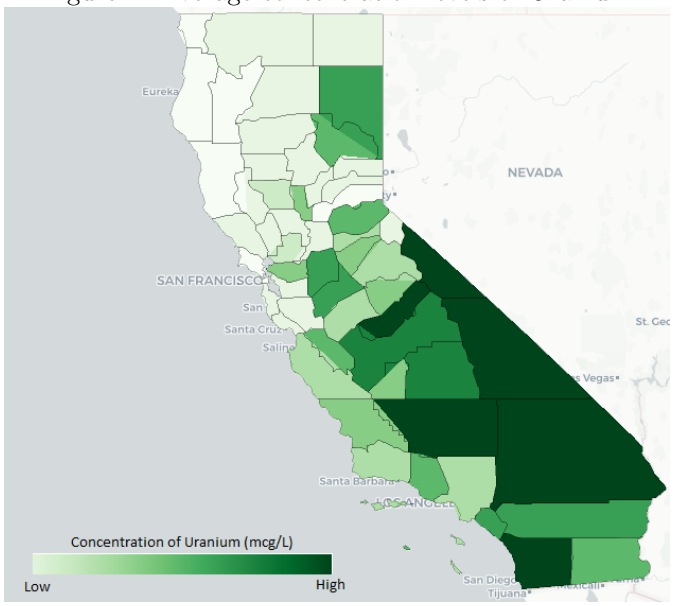


Figure 4: Average concentration levels of Uranium



3.2 Clustering Counties

From the previous visualizations we infer that we should cluster on a county level basis to extrapolate counties that exhibit similar patterns of contamination. We clustered the counties based on the average levels of chemical concentrations by unique chemical. The goal from this clustering was to identify counties that exhibit similar patterns in terms of chemical concentrations.

Based on the chemical values provided from the datasets, all of the counties in California are divided using clustering algorithm, specifically the fuzzy C-means clustering algorithm. Compared to the hard clustering algorithm such as the k-means clustering, the fuzzy C-means provides the possibility for one county having partial membership

in more than one cluster, which has the advantage to form the soft cluster for non-distinctive data with vague boundaries. Fuzzy C-means clustering is conducted through the minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, \quad 1 \le m < \infty$$

where N is the number of counties; C is the number of clusters; $u_{ij}(k)$ is the membership value of county i in the cluster j ; x_i is the data value at county i ; c_j is the center value of the cluster j; and m is the weighting exponent in the fuzzy C-means cluster. Ideally, all the chemical values in each county at all the years can be used be used as the vectors for the distance measurement for the membership calculation. However, the data records at each county for different chemicals are not even, which makes the approach of using time series as the calculating vectors impossible. In this approach, the average annual values of the four chemicals from 2010 to 2016 is used as the similarity indicators to divide all the counties into different clusters based on their chemical values. All values at each chemical are scaled to unit length before used as the input for the clustering.

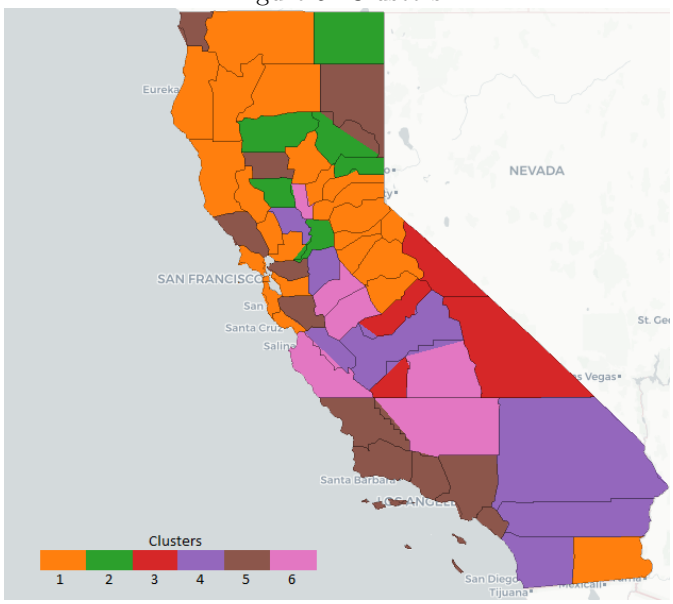
To determine the number of optimal cluster number, the following clustering validity index is used: Partition Entropy (PE), Partition Coefficient (PC) and Modified Partition Coefficient (MPC). Usually, the optimal cluster number can be selected by choosing the one that generate the largest PE values, the smallest PC and MPC values.

Table 1: the clustering validity index values for different clusters

Cluster Number	PE	PC	MPC
2	0.432	0.721	0.442
3	0.630	0.642	0.463
4	0.813	0.572	0.429
5	0.908	0.558	0.448
6	1.141	0.450	0.340
7	1.083	0.499	0.415
8	1.158	0.475	0.400
9	1.286	0.431	0.360
10	1.272	0.459	0.399
11	1.369	0.443	0.388
12	1.381	0.436	0.385
13	1.403	0.447	0.401

Based on the change of the values listed in table 1, the PC and MPC values reached the lowest when the number of clusters reach 6, while the PE values at cluster number 6 is reasonably large and certain change can be detected when number equals to 6. As a conclusion, the optimal number is 6 for the classification of all the counties in California based on four different chemicals. The results from the clustering algorithm can be viewed in Figure 5 below. These cluster numbers will be referred to throughout the remaining of the exposition.

Figure 5: Clusters



3.3 Principal Component Analysis

After the clustering is finalized, the Principal Component Analysis (PCA) is used to determine which variable can be used to interpret most of the information in each cluster i.e. the most significant chemicals among the four included.

Principal component analysis (PCA) is a dimensionality reduction technique that uses an orthogonal transformation to extract main principle component from a set of correlated variables. PCA is preferably used among the linear correlated variables. Nonlinear dimensionality reduction techniques such as ISOMAP or local linear embedding (LLF), especially the later one is required a large dataset for the input. Considered the limited record length we have in the dataset and the high correlation detected among the four included chemical variables, PCA is used for the principle component extraction.The PCA plots for the first two extracted principle components are shown in Figure 6.

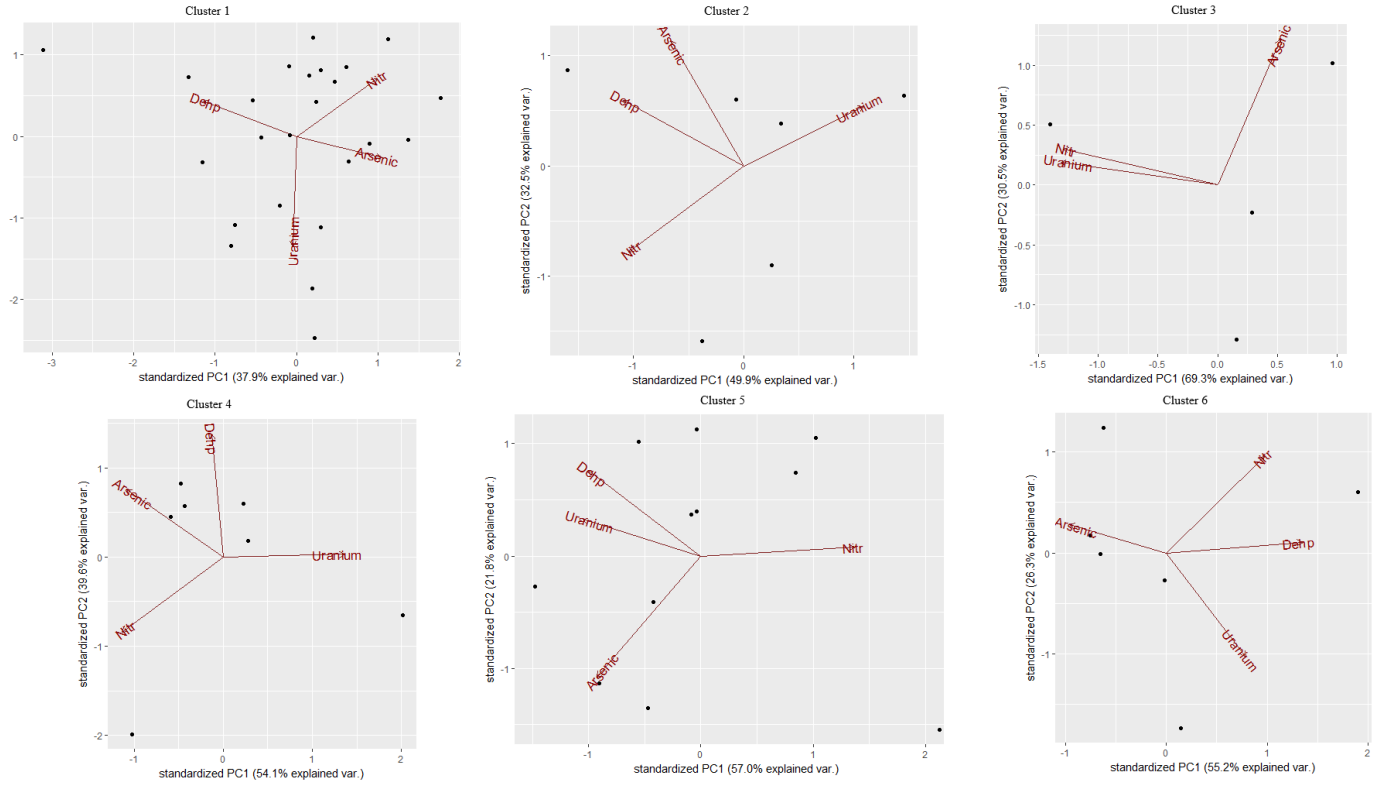
Based on the information provided from Figure 6, following conclusion can be obtained:

1. In the cluster 2, 4, 5 and 6, the first principle component explain more than or close to 50 percentage of the included information, the relevant variables related most to the first principle component is used as the most relevant indicators in these clusters: uranium for cluster 2 and 4, nitrate for cluster 5, DEHP for cluster 6.

2. In the cluster 1, the first principle component explain 37.9 percentage of the information, thus the second principle component also need to be considered in the analysis. The two factors that most relate to first and second principle components are uranium and arsenic, which will be used as the most important factors in Cluster 1.

3. In Cluster 3, the DEHP for all the counties are the same values. Despite what has been illustrate in the PCA in cluster 3, the most relevant factor for this cluster is DEHP.

Figure 6: PCA



3.4 Data Wrangling For Feature Relevance

With the clusters from above, the goal is now to be able to attain actionable insight to improve water quality. Since this analysis concentrates on the state of California, we thought that droughts would have predictive power of the water quality. Additionally, we considered features such as economic status and occupations within the country to help provide some features that would illustrate the resources for each county.

When creating the data set, it was developed on a yearly and county level basis. The previous clusters were used for each year in the respective county. The chemical concentrations were averaged yearly by county. This data set was then joined with the drought data. For each county, the average percentage of drought severity per year was calculated. Additionally, the previous years drought data was joined as another possible indicator. The occupation and industry data sets were also joined on an aggregate yearly level to get a representative picture of the economic status and resources available within the county. For missing features, the median was used as a placeholder to alleviate any pains from averages that incorporated outliers.

3.5 Factors in Water Quality

Model organization: Our analysis of water quality on a county basis has produced 6 groups of clusters on counties that are highly similar on a water contamination level. Through these 6 distinct groups we wish to determine what factors are the most important in determining water quality. Since our clustering algorithm reveals which counties are similar to each other on a water quality basis we decide attempt to determine factors related that differentiate the water quality among each cluster.

To analyze each feature we will want to attempt to understand how much a feature will contribute to a model's decision making. With this in mind we want to predict the amount of contamination within the water supply only utilizing our drought and economic factors. We decided to use a tree-boosting algorithm: extreme gradient boosting (aka XGBoost). The reason behind this choice is that it allows us to use a complex and computationally

intensive framework to determine feature importance. As well as its relevance in many Kaggle competitions many competitors have won data science competitions through the use of tree ensembling methods like XGBoost. Through the use of this algorithm we wish to use a baseline model to ascertain feature importance through the use of SHAP (SHapley Additive exPlanations). While native feature importance metrics in XGBoost are highly sensitive and sometimes produce misleading results, SHAP attempts to bring a unified approach to explain the output of any machine learning model. SHAP connects game theory with local explanations, uniting several previous methods and representing the only possible consistent and locally accurate additive feature attribution method based on expectations. Through the use of this new (2018) feature importance indicator and a simple baseline model we wish to determine which features are influencing the outcomes of water contamination.

3.6 Results

For the results, we have presented visualizations that can be interpreted by the colours. The red features are seen to have a negative effect on the mean value of the model while the the blue features seem to have a positive effect on the mean value. The plots below shows features each contributing to push the model output from the base value (the average model output over the training dataset we passed) to the model output. Features pushing the prediction higher are shown in red, those pushing the prediction lower are in blue.
For cluster 1, we tried to see which features would have the most significant impact to the levels of both uranium and arsenic. From Figure 7, we can see that having a small percentage of the population affected by a moderate drought in the previous year exhibits a relationship with an decrease in uranium in the water. Figure 8 shows that the farming and agriculture industry is indicative of a better water quality (represented by the level of arsenic).

Figure 7: Cluster 1 Uranium

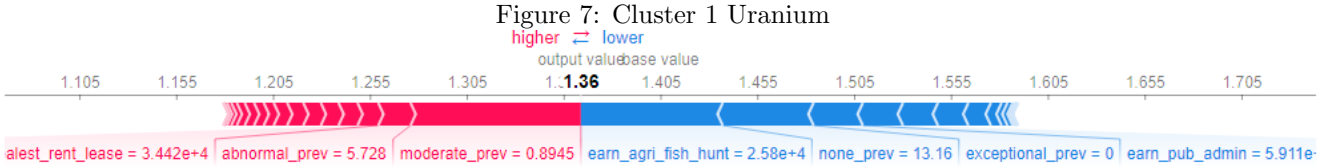
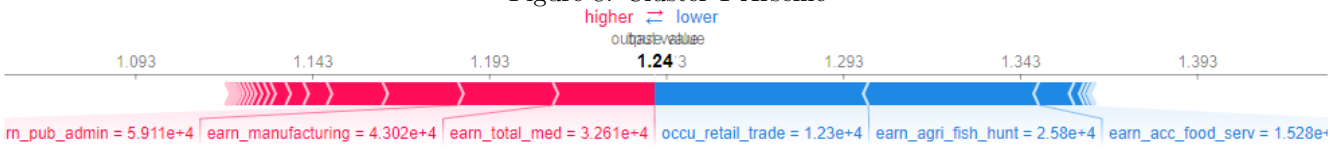
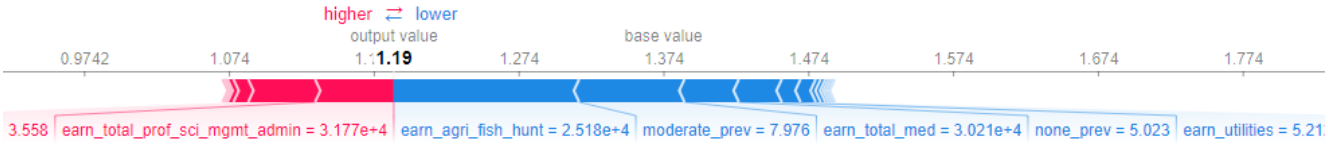


Figure 8: Cluster 1 Arsenic



Similar to the previous cluster, cluster 2 in Figure 9 shows that with an increase in agriculture in the current year and a very small drought percentage in the previous year, the water should have a lower level of uranium.

Figure 9: Cluster 2 Uranium



Cluster 3 uses features from the current year to model levels of DEHP in Figure 10. The non-existence of droughts positively impact the water quality (represented by the level of DEHP).

Figure 10: Cluster 3 DEHP

Cluster 4 5 interestingly models an economic factor of relevance to water quality as can be seen in Figure 11 and 12. One can understand this as the more resources being pooled in to these counties, provide a smaller value of Uranium and Nitrate in the water.

Figure 11: Cluster 4 Uranium

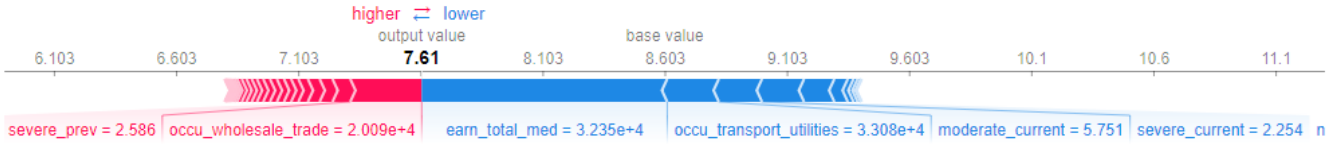
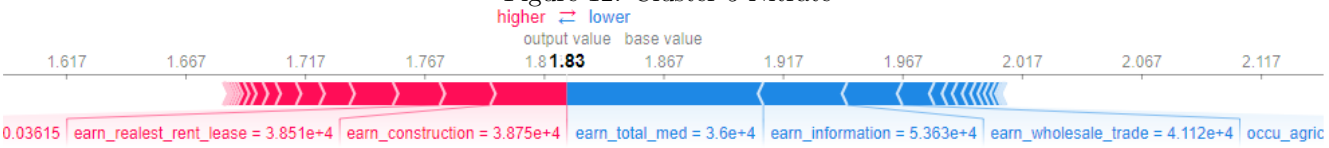


Figure 12: Cluster 5 Nitrate



Cluster 6 presents results similar to cluster 1 and 2 in Figure 13. Previous years with low or no droughts have a positive impact on water quality.

Figure 13: Cluster 6 DEHP

Although these above, figures depict more than the features discussed, we have discussed the most relevant ones and the illustrations should help in understanding any other important features. In the appendix, the features are also depicted with the largest bars representing the most important features.

4 Appendix

Figure 14: Cluster 1 Uranium

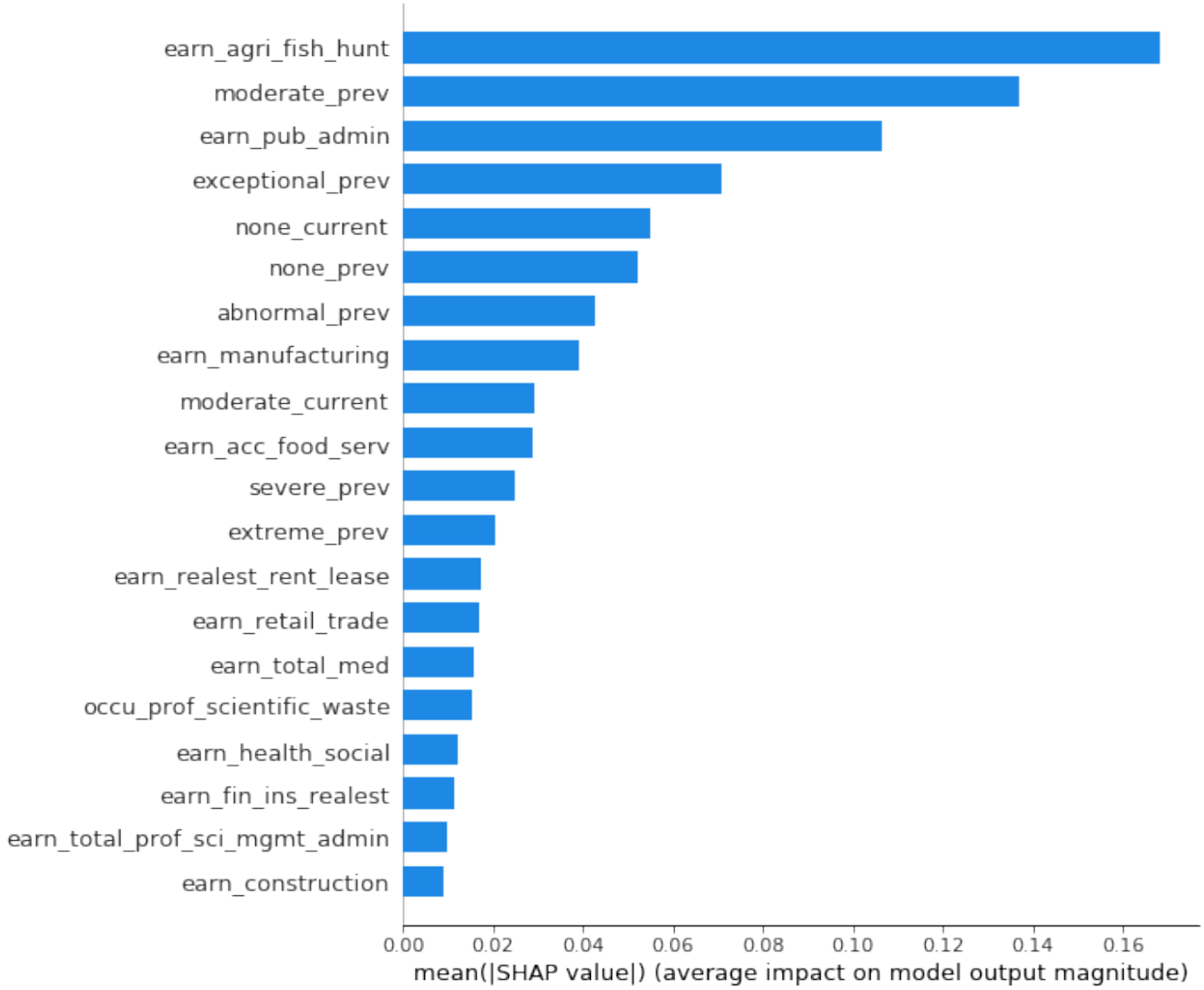


Figure 15: Cluster 1 Arsenic

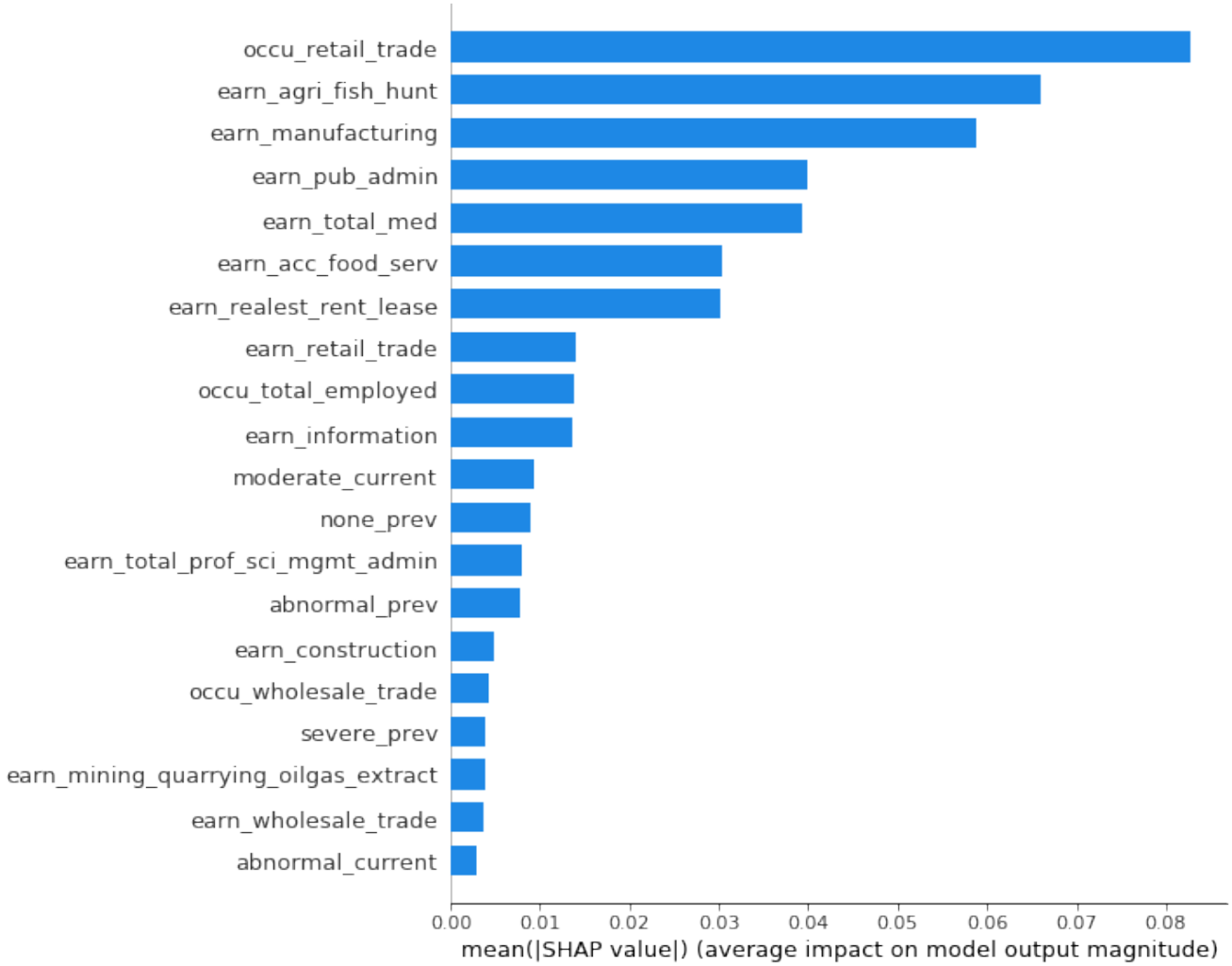


Figure 16: Cluster 2 Uranium

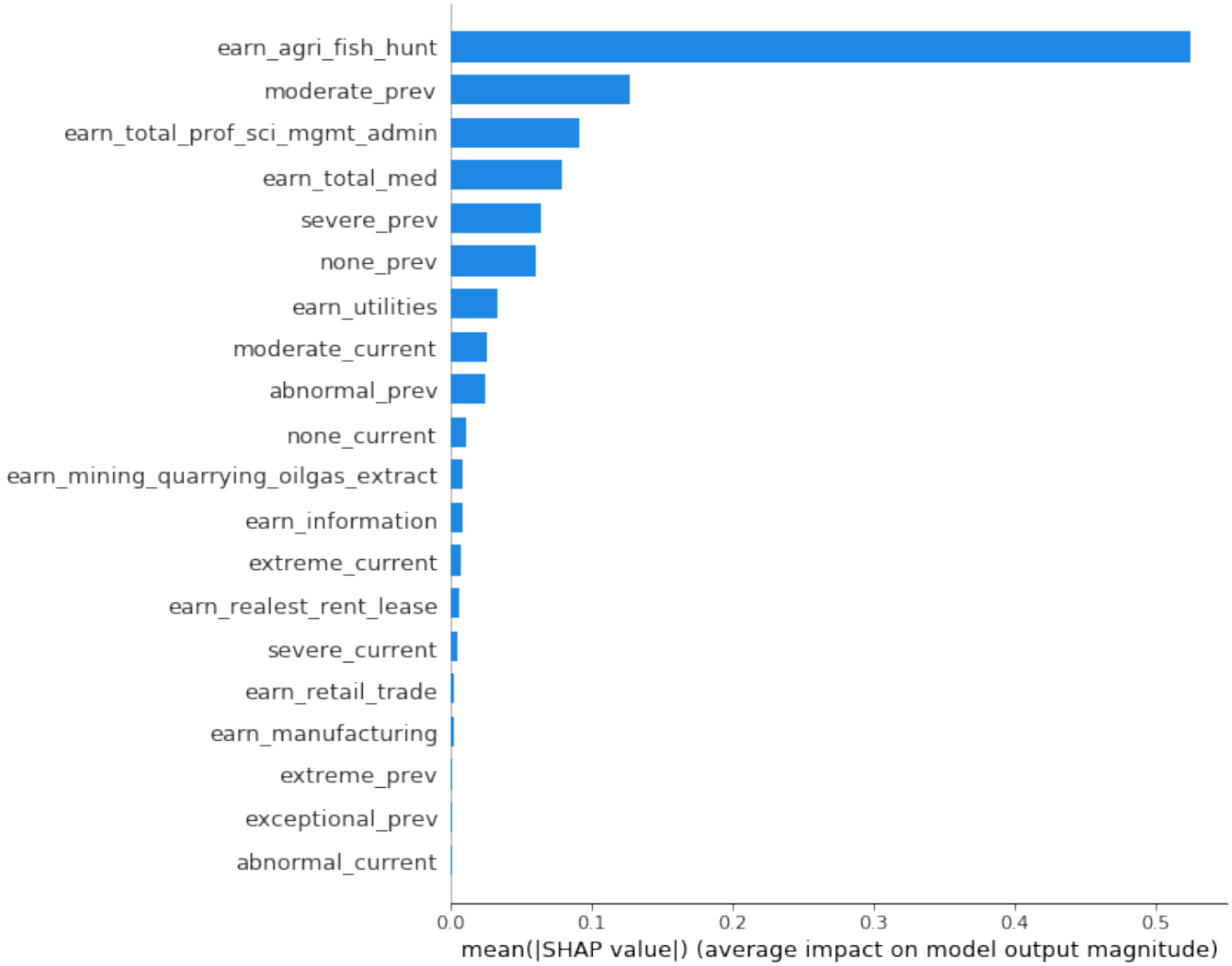


Figure 17: Cluster 3 DEHP

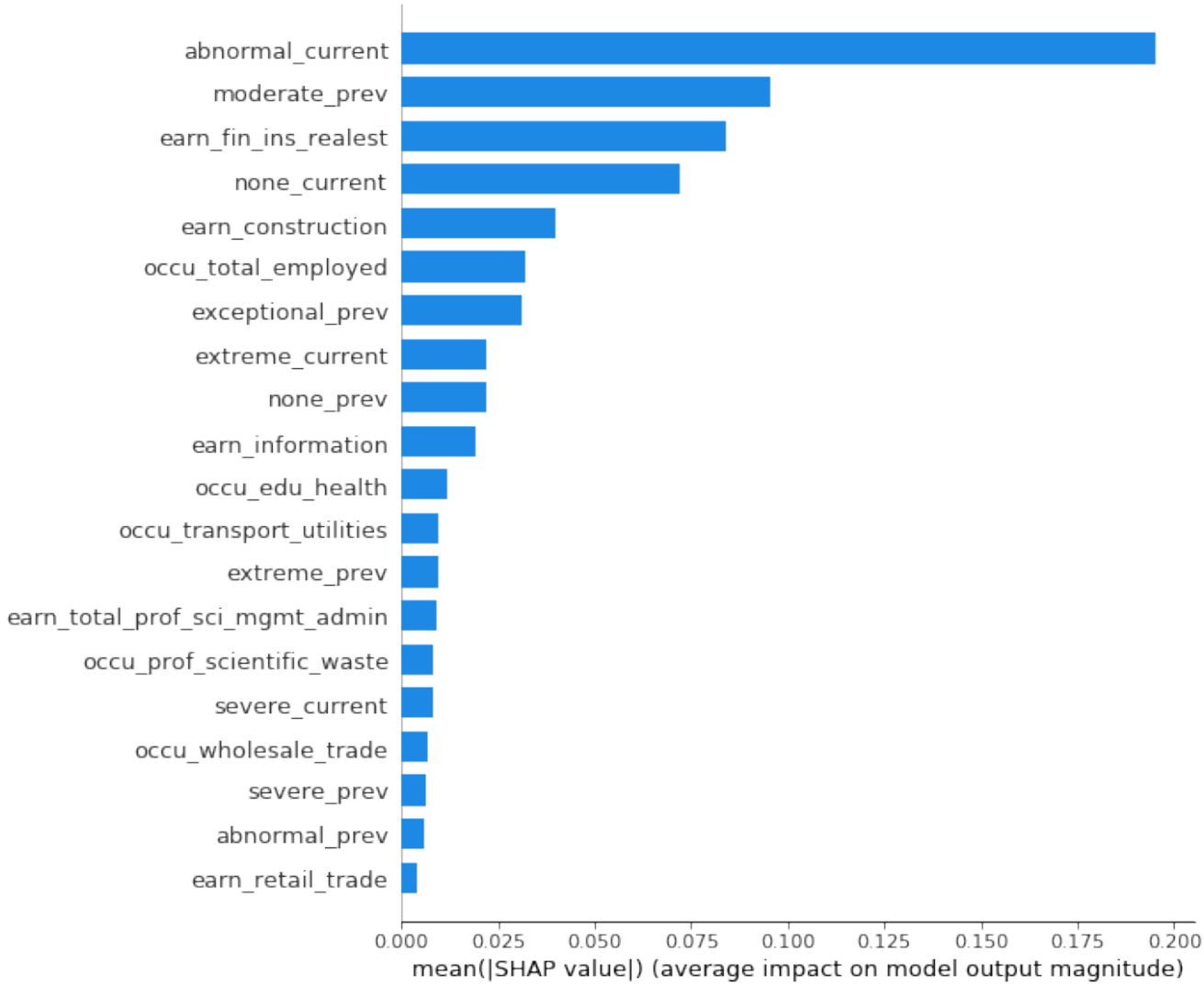


Figure 18: Cluster 4 Uranium

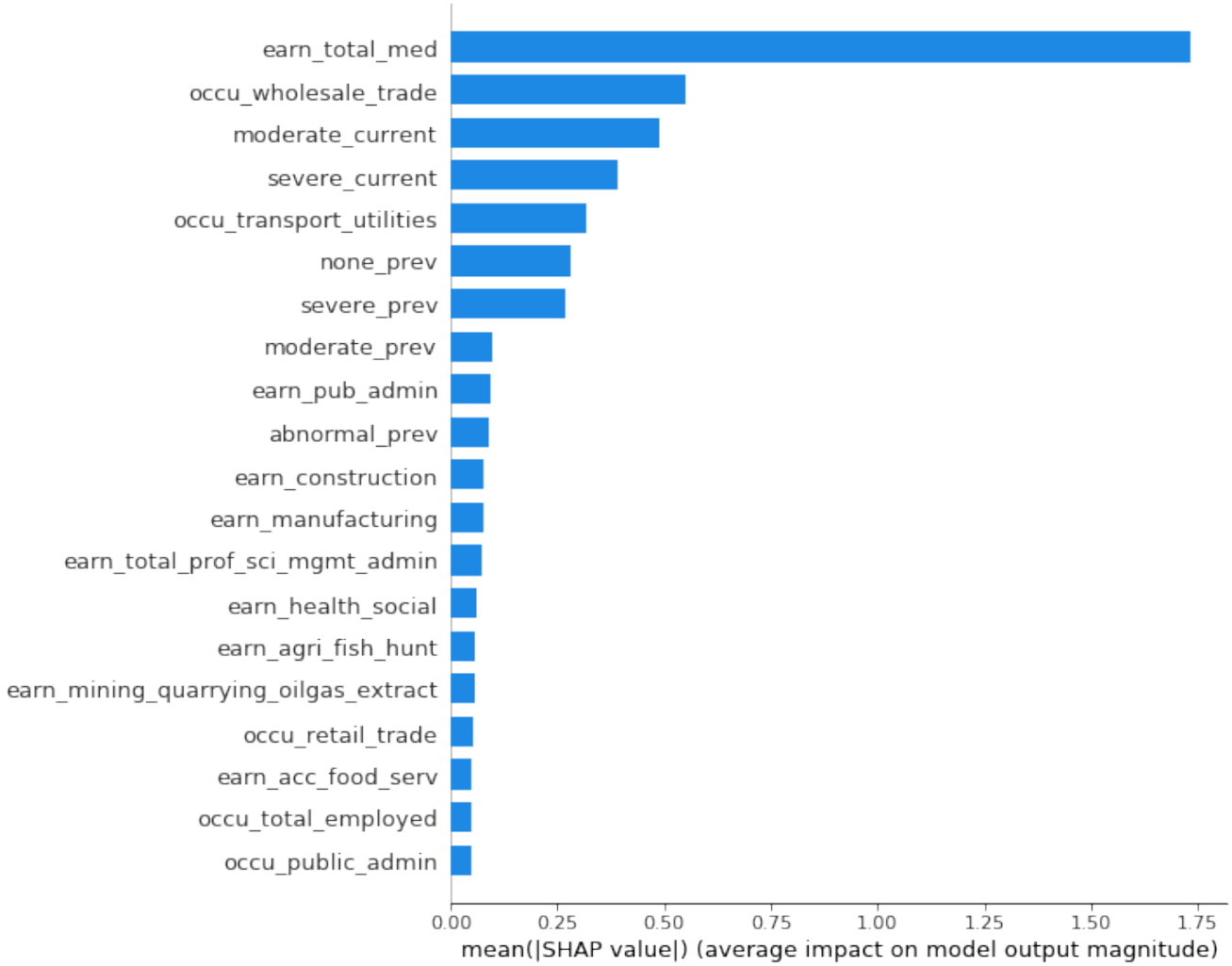


Figure 19: Cluster 5 Nitrate

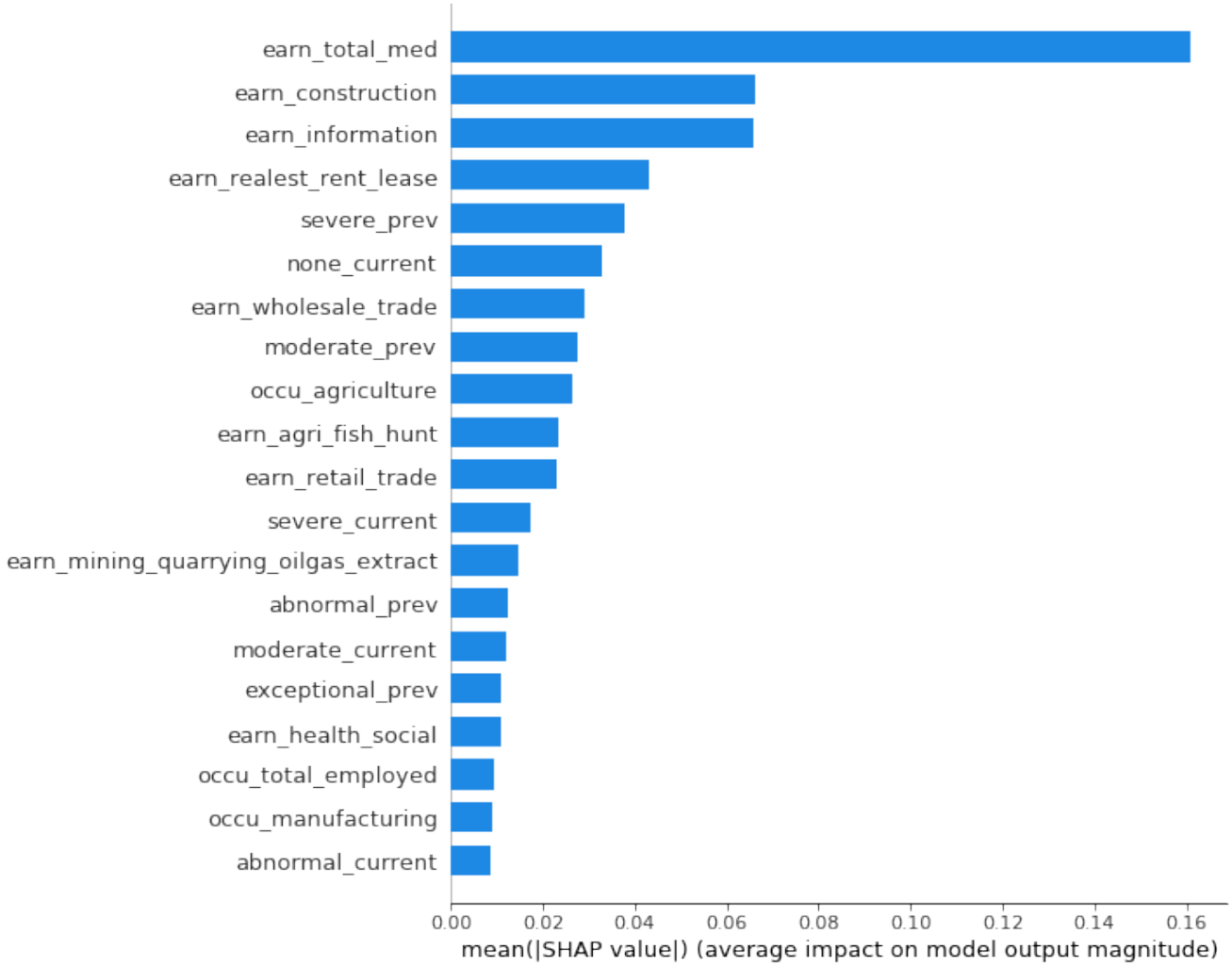


Figure 20: Cluster 6 DEHP