Intel oneAPI:Predict the quality of freshwater

Build a machine model to predict whether the freshwater is safe to drink or not.Based on the measures like pH, TDS, etc.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
    - Missing value imputation
        - Mean Imputation
            - Odor
            - Total Dissolved Solids
        - Median Impuation
            - pH
            - Iron
            - Nitrate
            - Chloride
            - Lead
            - Zinc
            - Turbidity
            - Fluoride
            - Copper
            - Odor
            - Sulfate

        - Conductivity
        - Water Temperature
        - Air Temperature
    - Color and Source features don't have any interaction with other numerical measures.
    - Month, Day, Time of day these features don't have any relevant information for determining the quality of freshwater.

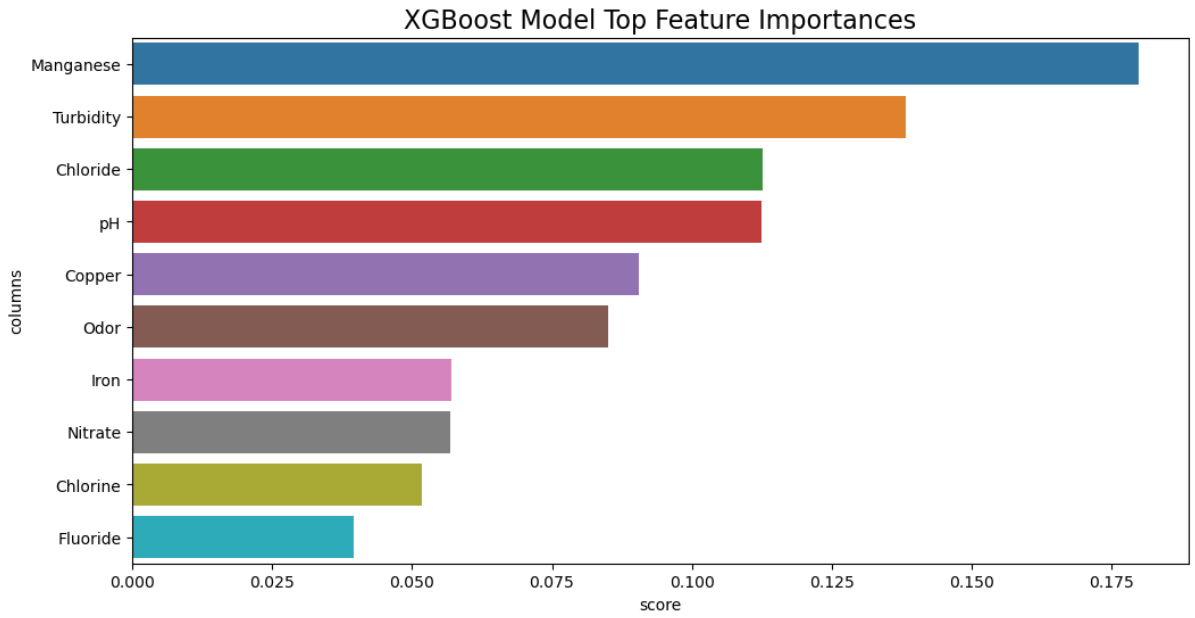
- The final features for the model
    - pH
    - Iron
    - Nitrate
    - Chloride
    - Lead
    - Zinc
    - Turbidity
    - Fluoride
    - Copper
    - Odor
    - Sulfate
    - Conductivity
    - Chlorine
    - Manganese
    - Total Dissolved Solids

- Water Temperature
- Air Temperature

- Created stratified train and test dataset from the entire dataset.
- Xgboost, lightgbm, catboost trained and evaluated with F1 score.

1. Trained XGBoost model with the parameter 500 estimators.

- Feature Importance of XGBoost model



- For faster inference XGBoost model is converted into oneDAL model. Then, the oneDAL model was used to predict the test accuarcy and train accuracy.

- Train classification report

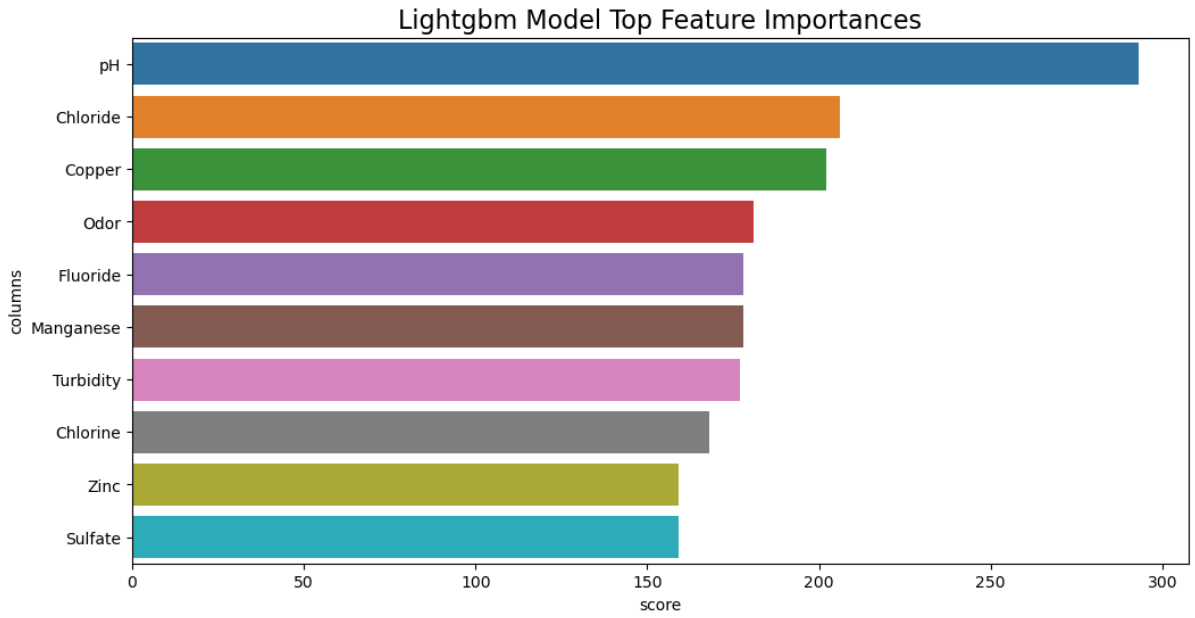
	precision	recall	f1-score	support
Safe to drink	0.94	0.86	0.90	2781565
Not safe to drink	0.73	0.87	0.79	1209519
accuracy			0.86	3991084
macro avg	0.83	0.86	0.85	3991084
weighted avg	0.88	0.86	0.87	3991084

- Test Classification report

	precision	recall	f1-score	support
Safe to drink	0.94	0.86	0.90	1370025
Not safe to drink	0.73	0.87	0.79	595733
accuracy			0.86	1965758
macro avg	0.83	0.87	0.85	1965758
weighted avg	0.88	0.86	0.87	1965758

2. Trained Lightgbm model with default parameters.

- Feature Importance of Lightgbm model



- For faster inference Lightgbm model is converted into oneDAL model. Then, the oneDAL model was used to predict the test accuarcy and train accuracy.

- Train classification report

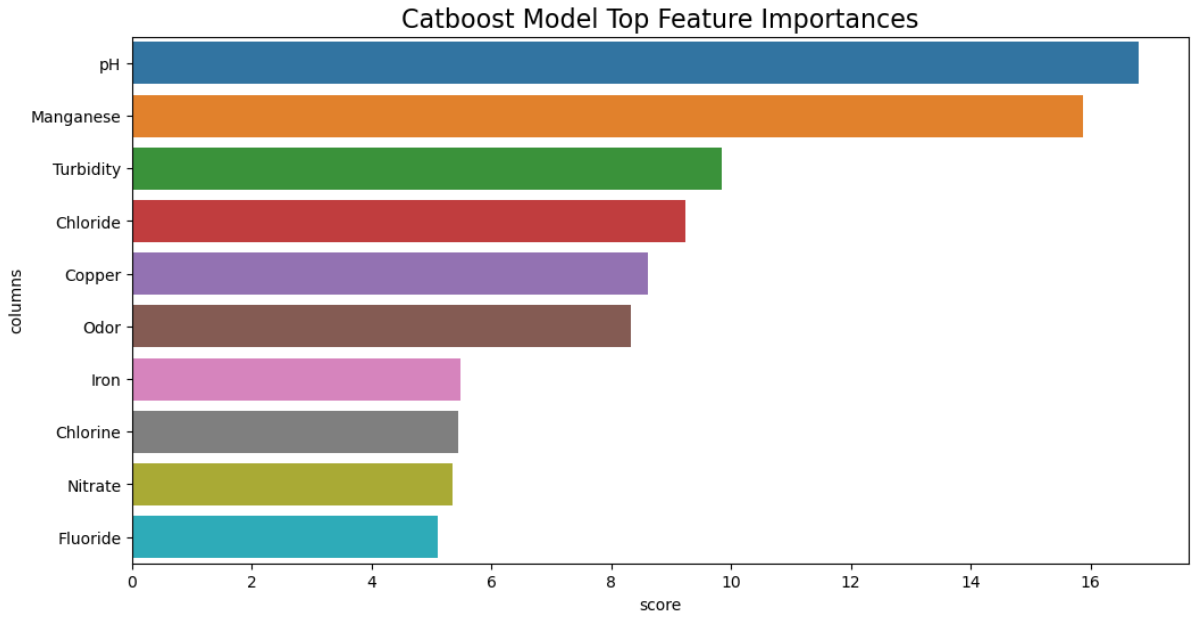
	precision	recall	f1-score	support
Safe to drink	0.99	0.84	0.91	2781565
Not safe to drink	0.73	0.97	0.83	1209519
accuracy			0.88	3991084
macro avg	0.86	0.91	0.87	3991084
weighted avg	0.91	0.88	0.89	3991084

- Test Classification report

	precision	recall	f1-score	support
Safe to drink	0.99	0.84	0.91	1370025
Not safe to drink	0.73	0.97	0.83	595733
accuracy			0.88	1965758
macro avg	0.86	0.91	0.87	1965758
weighted avg	0.91	0.88	0.89	1965758

3. Trained Catboost model with default parameter.

- Feature Importance of Catboost model



- For faster inference Catboost model is converted into oneDAL model. Then, the oneDAL model was used to predict the test accuarcy and train accuracy.

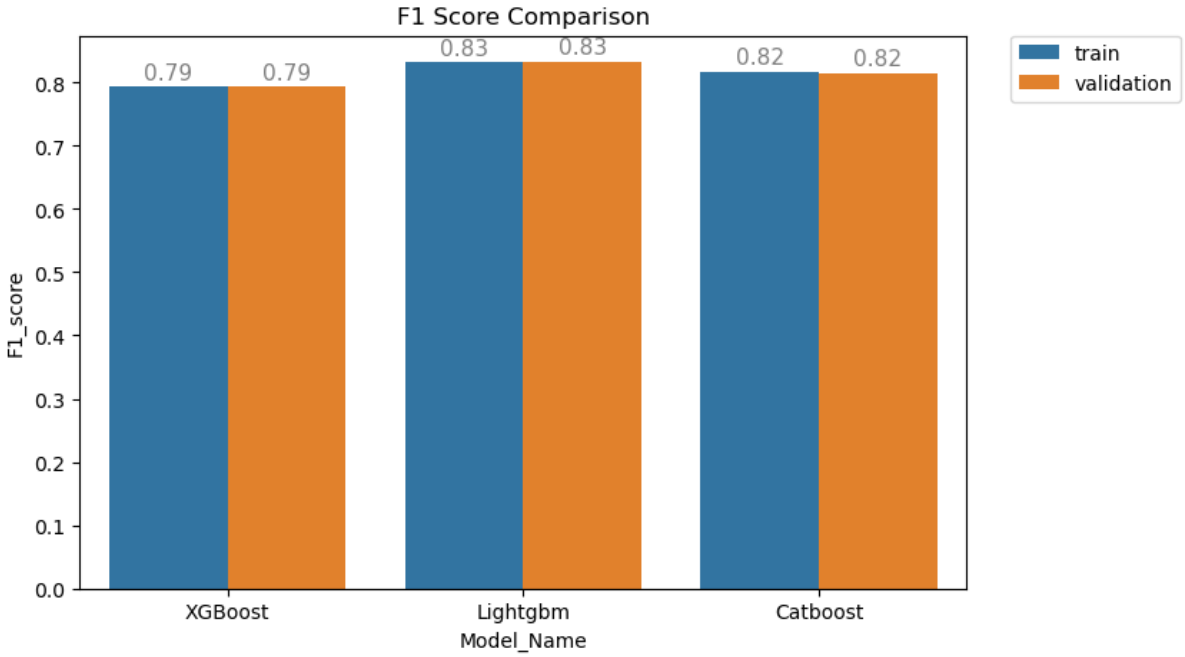
- Train classification report

	precision	recall	f1-score	support
Safe to drink	0.94	0.86	0.90	2781565
Not safe to drink	0.73	0.87	0.79	1209519
accuracy			0.86	3991084
macro avg	0.83	0.86	0.85	3991084
weighted avg	0.88	0.86	0.87	3991084

- Test Classification report

	precision	recall	f1-score	support
Safe to drink	0.96	0.85	0.90	1370025
Not safe to drink	0.73	0.92	0.82	595733
accuracy			0.87	1965758
macro avg	0.85	0.89	0.86	1965758
weighted avg	0.89	0.87	0.88	1965758

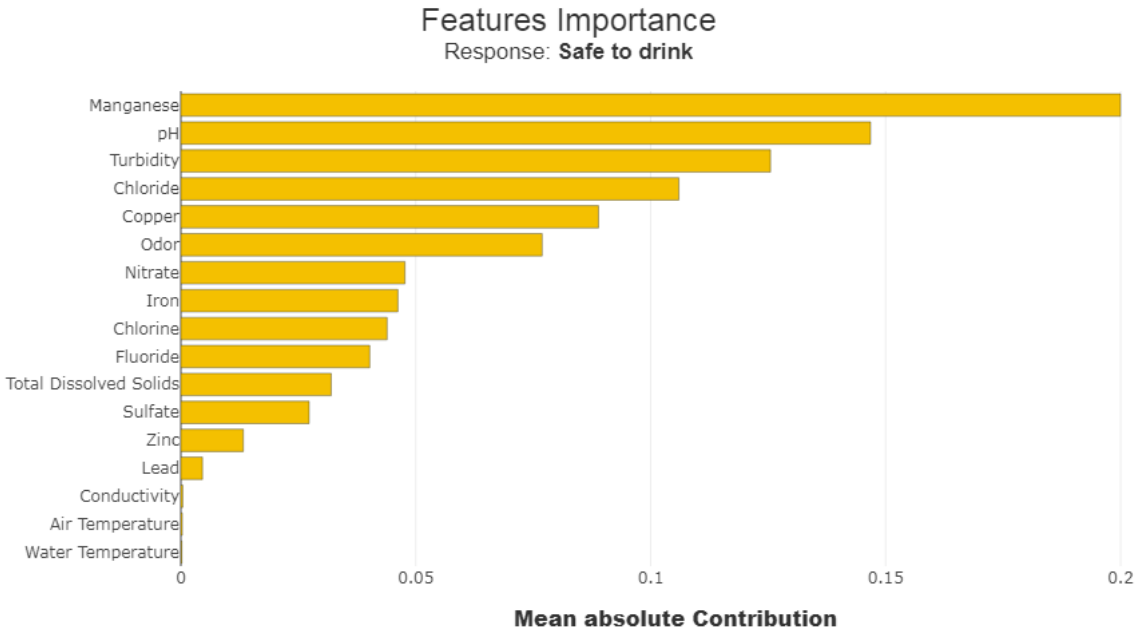
- Model Comparison



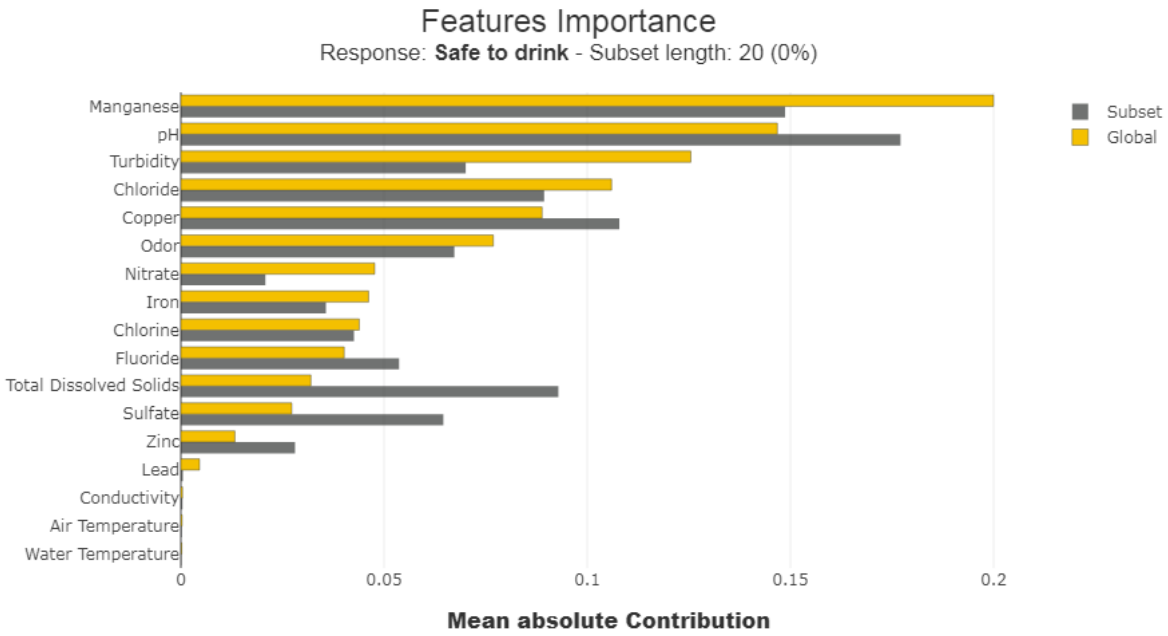
	Model_Name	Type_of_Data	F1_score
0	XGBoost	train	0.793467
1	XGBoost	validation	0.793757
2	Lightgbm	train	0.832525
3	Lightgbm	validation	0.832876
4	Catboost	train	0.817200
5	Catboost	validation	0.815234

- The lightgbm model works better on both train and test data compared to other models.

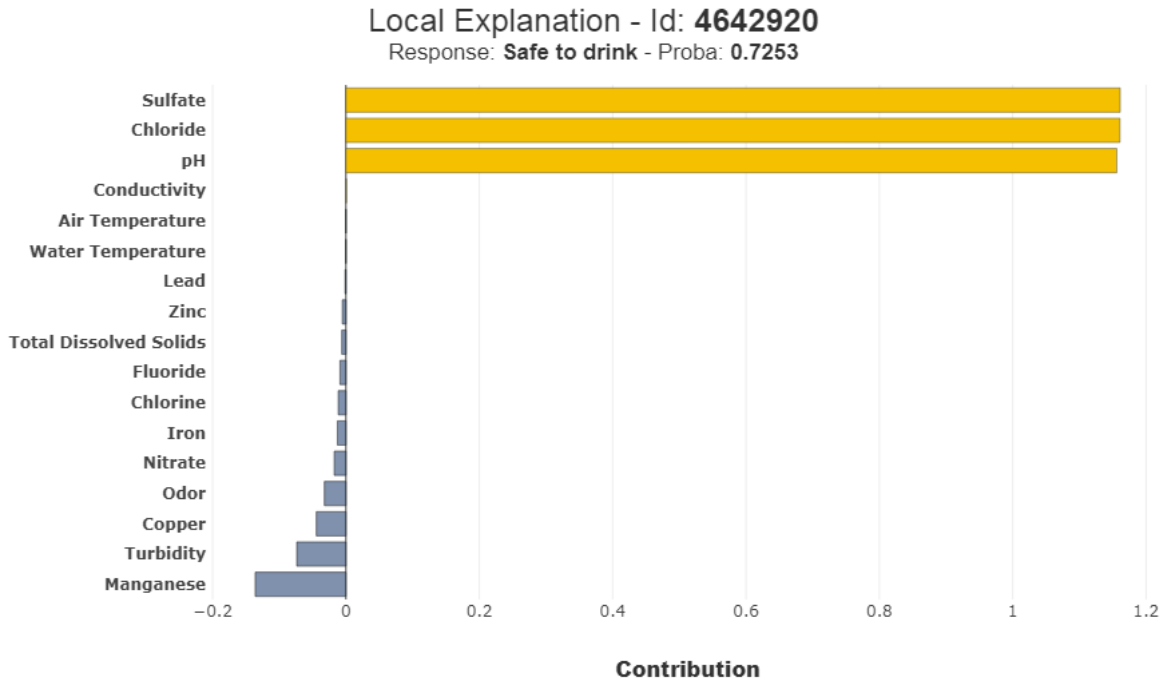
- Lightgbm Model Explanation with shapash-SHAP
  - Lightgbm model Feature Importances



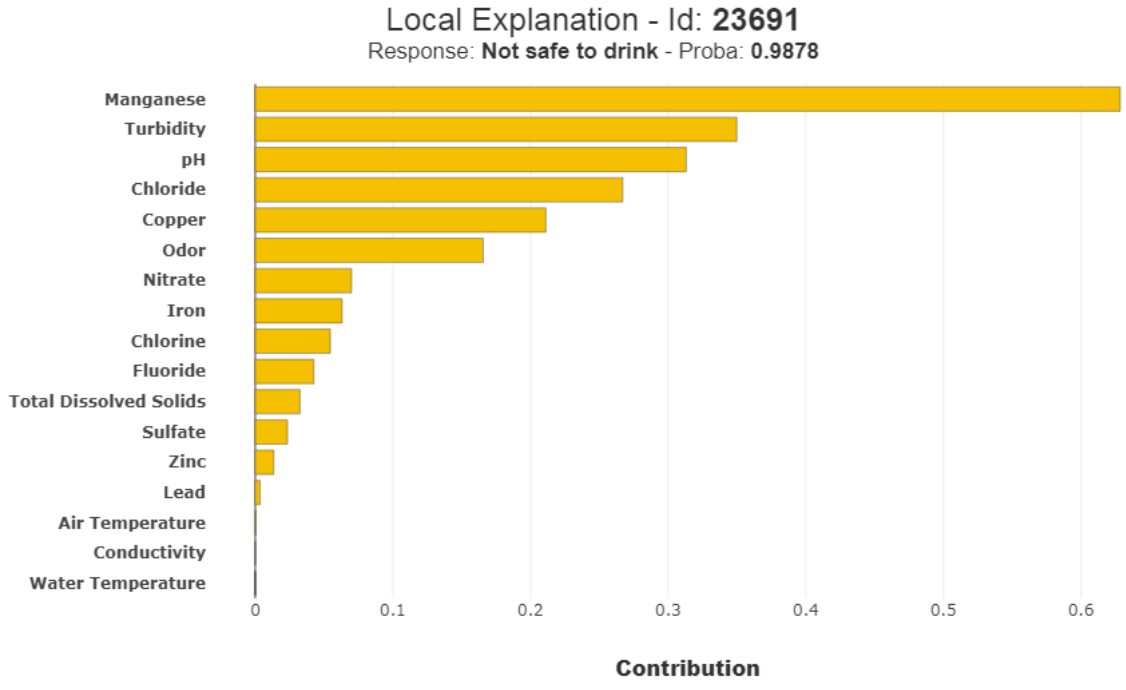
  - Lightgbm model subset Feature Importances



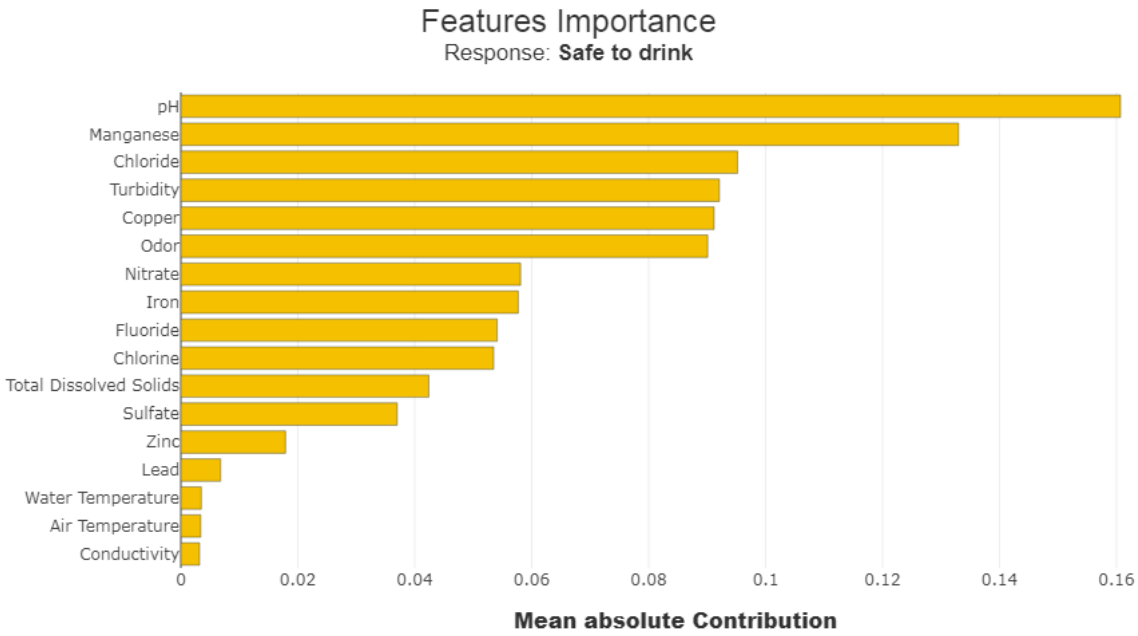
- lightgbm model Local explantion for class 1(safe to drink)



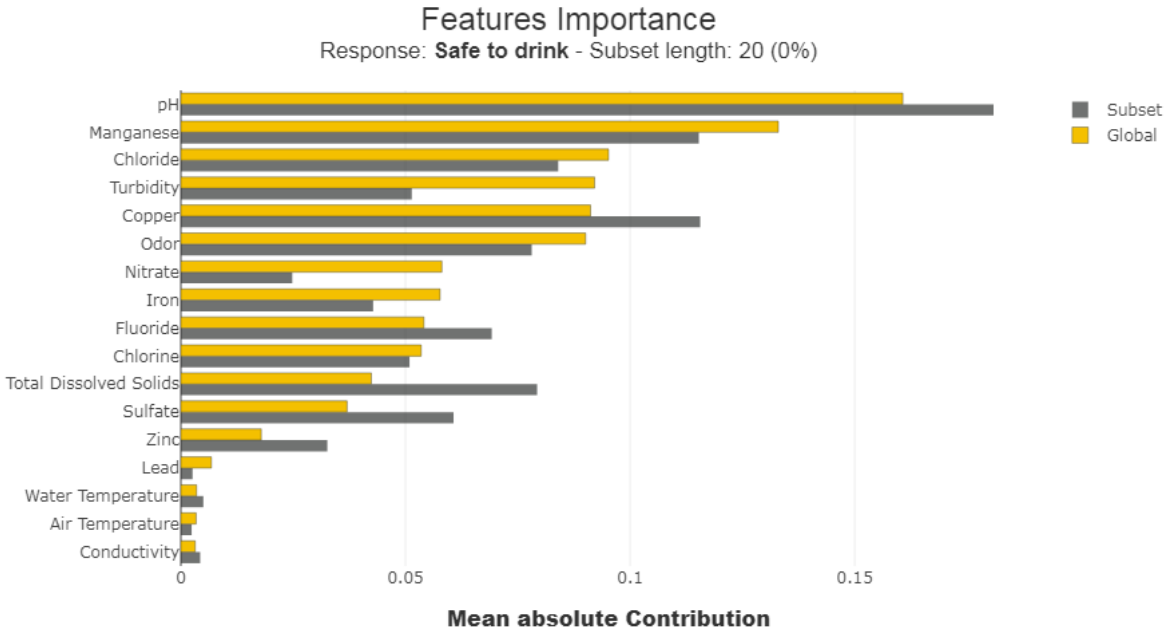
- lightgbm model Local explantion for class 0(not safe to drink)



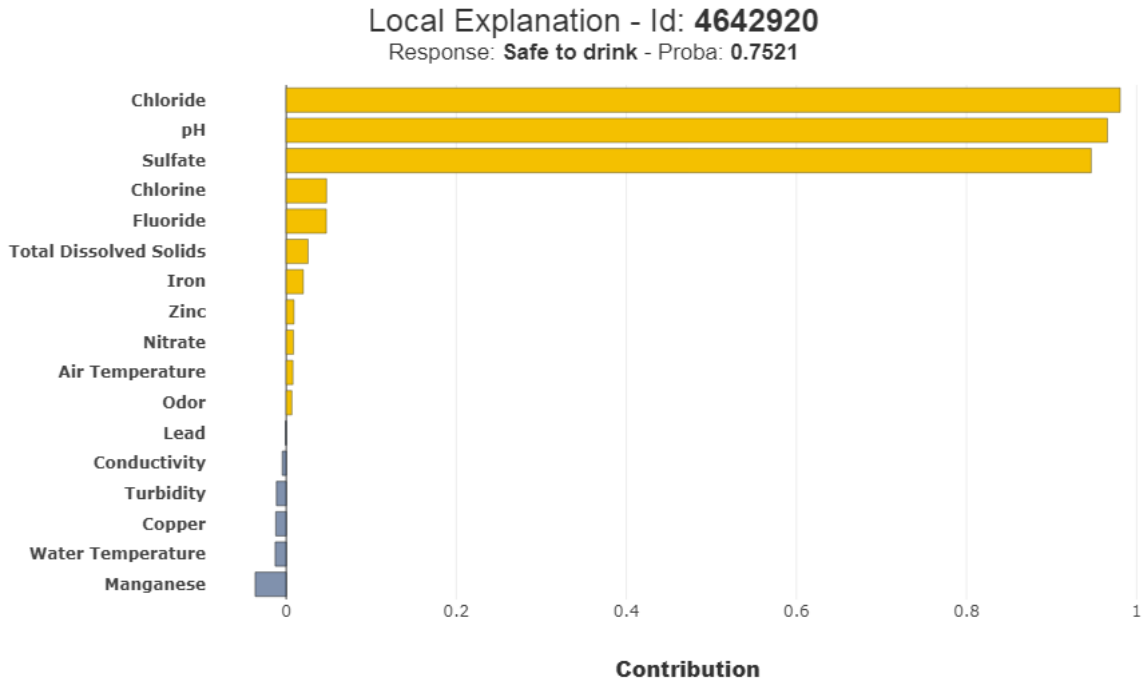
- Catboost Model Explanation with shapash-SHAP
  - Catboost model Feature Importances



  - Catboost model subset Feature Importances



- Catboost model Local explantion for class 1(safe to drink)



- Catboost model Local explantion for class 0(not safe to drink