# Project Reflection

## Team

Alma Soria, Angela Lekivetz, Crystal Blackburn, and Michaela Goud

## Business Problem Statement

**1. Business Problem Statement: State the problem you are trying to solve. Quote your problem verbatim.**

The problem that we are trying to solve is problem #5: Based on NPRI data, which industry is predicted to have the highest growth of releases in 5 years? Which will have the largest decline?

**2. ML Problem Definitions: Translate that into what it is exactly that you are trying to predict. Is it one value? Is it multiple values?**

In reality, this problem is best approached as a regression problem rather than a classification problem. Instead of simply assigning industries to categories like "Decline," "Stable," or "Growth," like we did in the Phase 1. Our goal is to predict a continuous numeric value that represents the change in pollutant releases for each industry over a 5-year period.

a) **Is there a column in the dataset you are predicting or are you constructing a new column? If it is an existing column, mention its name. If you are constructing a new column, explain exactly how you are constructing that new column.**

To achieve this, we construct a new column called "Growth_Value." For each industry, this value is calculated as follows:

Growth_Value = (Total Releases in 2022) – (Total Releases in 2017)

This "Growth_Value" is a single numeric value that indicates how much the releases have increased or decreased over the five-year period. A positive Growth_Value means that releases have grown, while a negative value means that releases have declined.

Our regression model will be trained to predict this Growth_Value. By doing so, we can determine with more precision which industry is expected to see the highest increase in releases and which one is projected to experience the largest decrease.

b) **Aggregation Strategy: If it is multiple values, do you integrate that back into a single value? If so, how exactly do you do that?**

The value that we are trying to predict will be a single value that will allow us to predict the highest growth of releases in 5 years and the largest in decline. However, for this we will also use of all the releases type, the final unit that we chose is kg because is not as small as g or big as tonnes:

The columns for releases are:

- Release to Air - Fugitive (kg)
- Release to Air - Other Non-Point (kg)
- Release to Air - Road dust (kg)
- Release to Air - Spills (kg)
- Release to Air - Stack / Point (kg)
- Release to Air - Storage / Handling (kg)
- Releases to Land - Leaks (kg)
- Releases to Land - Other (kg)
- Releases to Land - Spills (kg)
- Releases to Water Bodies - Direct Discharges (kg)
- Releases to Water Bodies - Leaks (kg)
- Releases to Water Bodies - Spills (kg)
- Sum of release to all media (<1tonne) (kg)

We did create columns by aggregating them to create the columns: 'Total_Air_Releases’, 'Total_Land_Releases', 'Total_Water_Releases, and 'Total_All_Releases'. Our problem is too general in the sense that we care about the total releases, there are no questions about specific type releases, therefore, we used these aggregations and then we performed a pivot approach to Create Years Columns for the 5 past years (the data that we do have) to be able to have what we need to perform the prediction for the next 5 years (the data we want to predict).

Additionally, we created the column ‘Industry_Sector’ where we classified all the NAICS_Title we had access to in our dataset. Using the NAICS codes that define industry sectors in Canada.

**3. Feature Requirements: Exactly what things that you have access to at the time of operation? How do they affect the value of the thing/things you are trying to predict? List all.**

For this project, our goal is to predict how an industry's pollutant releases will change over the next five years. To do this, we are using historical NPRI data and constructing a new target variable called “Growth_Value.” This variable represents the change in total releases over a five-year period and is calculated as:

Growth_Value = (Total Releases in 2022) – (Total Releases in 2017)

We don’t have a single column in the original dataset that tells us this information directly. Instead, we aggregate multiple release-related columns to get a complete picture of each facility’s releases.

List of all things we have access:

- Reporting_Year: The year in which the facility or company submitted its environmental data to the NPRI. This allows for tracking trends and changes in releases over time.
- NPRI_ID: A unique numerical identifier assigned to each reporting facility or company. This ID can be used to link to additional information about the facility, such as its location details, contact information, and historical reporting data.
- Number of employees: The total number of employees working at the reporting facility. This can be used to normalize release data by facility size, allowing for comparisons between facilities of different scales.
- Company_Name: The official name of the company responsible for the reported environmental data. This helps identify major polluters and track their performance over time.
- Facility_Name: The name of the specific facility where the pollutant releases occurred. This provides geographic context and allows for identifying pollution hotspots.
- NAICS: The North American Industry Classification System (NAICS) code assigned to the facility. NAICS is a hierarchical classification system (e.g., 2-digit sectors, 3-digit subsectors) that categorizes businesses based on their primary economic activity. This information is crucial for understanding which industries contribute most to pollution and for analyzing sector-specific trends.
- NAICS Title: The descriptive title corresponding to the NAICS code, providing a human-readable description of the industry sector.
- City: The city or municipality where the reporting facility is located. This allows for geographic analysis of pollution patterns and identification of urban areas with high emissions.
- Latitude: The latitude coordinate of the facility's location. This enables mapping and spatial analysis of pollution data.
- Longitude: The longitude coordinate of the facility's location. This, along with latitude, provides precise geographic coordinates.
- CAS_Number: The Chemical Abstracts Service (CAS) registry number for the substance being released. CAS numbers are unique identifiers assigned to chemical substances, allowing for standardized identification and tracking of specific pollutants across different datasets and studies.
- Substance Name: The common name or chemical name of the substance being released. This provides a human-readable identifier for the pollutant.
- Units: The units of measurement used to quantify the amount of substance released (e.g., tonnes, kilograms, grams). Standardizing these units is essential for accurate comparisons and analysis.
- Estimation_Method: The method used by the facility to estimate the quantity of the release or disposal of the substance. This might include direct monitoring, emission factors, mass balance calculations, or other estimation techniques. Understanding the estimation method provides insights into the accuracy and reliability of the reported data.
- Release to Air - Fugitive: The estimated amount of the substance released into the air from fugitive sources. These are typically uncontrolled or unintended releases, such as leaks from equipment, vents, or open storage.

- Release to Air - Other Non-Point: Releases to the air from other non-point sources, which are diffuse sources that cannot be easily attributed to a single point. This might include emissions from surface areas, evaporation from open containers, or windblown dust.
- Release to Air - Road Dust: The amount of the substance released into the air specifically due to road dust generated by vehicles or other activities on unpaved roads.
- Release to Air - Spills: The amount of the substance released into the air as a result of accidental spills or leaks, typically involving a sudden and uncontrolled release of a significant quantity.
- Release to Air - Stack / Point: The amount of the substance released into the air from stack or point sources. These are typically controlled releases from industrial processes, such as emissions from smokestacks, chimneys, or exhaust vents.
- Release to Air - Storage / Handling: The amount of the substance released into the air during the storage or handling of materials. This might include emissions from loading/unloading operations, material transfer, or storage tanks.
- Releases to Land - Leaks: The amount of the substance released onto land due to leaks from facilities, equipment, or storage containers. This can result in soil and groundwater contamination.
- Releases to Land - Other: Releases to land that do not fall under specific categories like leaks or spills. This might include land disposal of waste materials, application of pesticides, or releases from land treatment processes.
- Releases to Land - Spills: The amount of the substance released onto land from spills, similar to air spills but with the impact focused on soil and groundwater contamination.
- Releases to Water Bodies - Direct Discharges: The amount of the substance directly discharged into water bodies, such as rivers, lakes, or streams. This can include wastewater discharges from industrial processes or municipal treatment plants.
- Releases to Water Bodies - Leaks: The amount of the substance released into water bodies due to leaks from facilities, equipment, or pipelines.
- Releases to Water Bodies - Spills: The amount of the substance released into water bodies as a result of spills or accidental discharges.
- Sum of release to all media (<1tonne): A summary of all releases to air, land, and water, specifically for cases where the total release quantity is less than 1 tonne. This provides a simplified reporting category for minor releases.

For this regression problem, the key features available at the time of operation are the historical pollutant release values and some contextual attributes like Substance_Name Company_Name, City, Province, Geographical locations, Units. The geographical location could help us to have a broader context and understanding of our data, however, for our specific problem they are not relevant for the ML model, and by introducing them we could just have unnecessary noise. We do have a lot of information available, but we need to think which information or data will really help us to solve our problem. For example:

Historical Release Data (2017 to 2022):
We have yearly release values for each industry. These numeric values are the most direct indicators of how releases have changed over time. In our model, we use these values to

understand the trend, with a particular focus on the difference between 2022 and 2017. A consistent increase or decrease over these years is expected to strongly influence the predicted Growth_Value.

Industry Sector Previously ‘NAICS_Title’:
The category of industry (e.g., petroleum, manufacturing, chemical production) is available as a categorical feature. Different industry sectors have distinct operational processes, regulatory environments, and historical release patterns. This information helps the model adjust its predictions based on the typical behavior associated with each sector.

Release Type:
Although we aggregate release values from different media (such as air, land, and water), knowing the breakdown by release type can provide additional context. For example, some industries may have higher releases in one medium compared to others, which can indicate specific operational challenges or regulatory pressures.

Other Potential Operational Features:
As we previously mentioned, additional features such as the number of employees, facility size, or NAICS codes might be present in the dataset. These factors can indirectly influence release values—for example, larger facilities or those in more complex industrial sectors might experience higher release volumes due to increased production levels.

Each of these features contributes to a more comprehensive understanding of the industry's historical performance and operational context. By integrating these data points, the model aims to predict a continuous numeric value (Growth_Value) that represents the absolute change in total releases from 2017 to 2022. This predicted value will help identify which industries are likely to see the highest growth and which are expected to have the largest decline in pollutant releases over the next five years.

4. **Does it make sense for the target value to be affected by what you listed? Explain. If you answer with yes, argue why. (*One good practice is to put yourself in the shoes of the ML model: are you able to make a good prediction with the information provided in features you are supplying it with, then it is reasonable to assume the ML model may be able to do a good job. If you can’t see how you can make a good prediction with the information provided, don’t expect to get good results out of your ML model. Machine learning can do no magic*).**

Yes, it makes sense for our target value—Growth_Value, which represents the change in total releases between 2022 and 2017—to be influenced by the features we have available. Because our release data from 2017 to 2022 directly reflects the trends in pollutant outputs over time. Since our target value is calculated as the difference between the total releases in 2022 and those in 2017, this historical information is essential for capturing the underlying trend. Also, different industries operate under different regulatory environments and have varying operational practices. Knowing the industry sector helps the model learn patterns that are typical for each sector. For example, some sectors might consistently show higher growth in releases while others might exhibit a decline due to stricter controls. About the different release types, although we ultimately aggregate the data to form the total releases, the breakdown by release type (such as air, land, and water) provides additional context. Certain release types might be more heavily regulated or have specific operational challenges, and these details can influence the overall change in emissions.

If we talk about other operational features (for example Number of Employees, Facility Size), these features provide an indication of the scale and complexity of each facility. Larger facilities or those with more complex operations might naturally have higher release volumes, which can affect the magnitude of the change over time.

Also geographical and chemical Information can provide details such as the facility's location (City, Province, Latitude, Longitude) and the chemical specifics (CAS_Number, Substance Name, Units) that might not be directly used in the final predictive model, but they offer valuable context. This information can help explain regional differences or chemical-specific trends that may influence emission levels.

The features we have not only give us direct historical data on pollutant releases but also provide contextual details about each facility and industry. This combination of factors forms a foundation for predicting the continuous value of Growth_Value. If the model can learn from these historical trends and context, it is reasonable to expect that it will make accurate predictions regarding which industries will see the highest increase or the largest decline in pollutant releases over the next five years.

**5. Based on the answer to the last question, if some things do not make sense, how can you construct variables that do make sense in being used to make predictions about the (different) targets?**

For this we need to understand the data. Understand how the data that we have can be somehow transformed for example aggregate and normalizing the data which means to be more specific that instead of using raw values, we can aggregate related release data into summary metrics. For example, we have already combined multiple release measures into totals (such as Total_All_Releases). This provides a clearer picture of overall performance than dealing with individual release types separately.

We can continue transforming our data, creating new columns and variables, and making use of feature engineering to create new features. An example of this would be the ratio or per-unit measures. Another example would be calculating releases per employee or per unit of production that can help normalize for the size of the facility. This can assist our model to learn trends that are comparable across facilities of different scales.

If some features, like detailed geographical coordinates or chemical names, are not directly relevant to predicting changes in releases, we can either remove them or combine them into broader categories. For example, rather than using exact latitude and longitude, we might construct a variable that groups facilities by region and that can provide an additional insight that could help us to solve other problems where region is important.

Another way to do this, is applying mathematical transformations (such as logarithmic scaling) that can help adjust for skewed distributions and make the features more directly related to the target variable. This can be especially useful if some release values vary over several orders of magnitude.

By constructing variables in these ways, we ensure that the information fed into the model is directly relevant to predicting our target, specifically Growth_Value, which represents the change in total releases over a 5-year period. This art of crafting of features helps the machine learning model make more accurate predictions.

**6. How should you frame your problem then?**

The problem should be framed as a regression problem. Our goal is to predict a continuous numeric value—Growth_Value—which is defined as the difference between the total pollutant releases in 2022 and those in 2017 for each industry. This target value represents the absolute change in releases over a five-year period, with positive values indicating growth and negative values indicating a decline.

To predict Growth_Value, our model will use historical release data from 2017 to 2022 along with other contextual features (such as industry sector, number of employees, and facility characteristics). By concentrating on a continuous outcome, we will be able to more precisely quantify the change in releases, enabling us to determine which industries are likely to see the highest increase or the largest decrease in pollutant releases over the next five years.

**7. How are you planning to assess your results? Be minute, detailed and exact.**

For this regression problem, where our target variable is Growth_Value (the change in total pollutant releases from 2017 to 2022), we can make use of evaluation techniques like:

Evaluation Metrics:

Mean Absolute Error (MAE): Measures the average absolute difference between the predicted and actual Growth_Value.

Mean Squared Error (MSE): Calculates the average squared difference between predictions and actual values, which penalizes larger errors more heavily.

Root Mean Squared Error (RMSE): The square root of MSE, providing an error metric in the same units as Growth_Value.

R-squared (Coefficient of Determination): Indicates the proportion of the variance in Growth_Value that is predictable from the features.

Cross-Validation: We will perform k-fold cross-validation (e.g., with k=5) on our training set.

This will provide average error metrics (MAE, MSE, RMSE) and R-squared scores across folds, along with their standard deviations, giving us a sense of model stability and generalizability.

Test Set Evaluation: After finalizing the model and tuning its hyperparameters, we will assess its performance on a hold-out test set using the same metrics.

We will compare the predicted Growth_Value against the actual values to ensure that the model is not overfitting.

Residual Analysis: We will create residual plots (plots of errors versus predicted values) to check for any systematic patterns that might indicate biases or heteroscedasticity (unequal error variances).

Histograms or density plots of the residuals will help us confirm if they are approximately normally distributed.

Visualization: Scatter plots comparing actual versus predicted Growth_Value will be used to visually assess the model’s performance.

These visualizations help identify outliers and understand if the model is consistently under- or over-predicting in certain ranges.

**8. Does your assessment make sense?**

We are aware we still have material in class to cover that will help us refine our current ideas. But as of today, our assessment makes sense. We are using multiple, well-established methods to evaluate the performance of our regression model. Specifically, we're measuring error with MAE, MSE, RMSE, and R-squared, which provide a comprehensive view of the model's accuracy and reliability. Additionally, by performing cross-validation, we ensure that our model's performance is consistent across different data splits. Overall, this evaluation strategy is thorough and appropriate for our problem, giving us confidence in the model's predictive capabilities.

**9. If not, how can you update it?**

If our assessment isn't giving us the full picture, we can update it in a few simple ways:

1. Add More Metrics:
   Use additional measures like Mean Absolute Percentage Error (MAPE) or adjusted R-squared. These extra numbers can help us better understand how close our predictions are to the real values.
2. Improve Cross-Validation:
   Instead of just using one round of k-fold cross-validation, we could try nested cross-validation or increase the number of folds. This helps us be more confident that our model works well on different parts of our data.
3. Look Closer at the Errors:
   By examining plots of the prediction errors (residuals), we might see patterns that indicate our model is missing something. If we notice issues, we could try transforming our target value or using methods that handle these problems better.
4. Include Domain Knowledge:
   We could use what experts know about pollutant releases to set specific limits or thresholds. This ensures our evaluation measures are not only statistically good but also make sense in real-world terms.

These changes would give us a more complete and clear picture of how well our model predicts the change in releases over time.