Project: Enviromental Impacts Prediction

CMPT 3510 – Machine Learning I

Introduction

This assignment accounts for 35% of your final grade in the course. The first phase, the classification phase of the project, is worth 15%, and the regression phase of the project is worth 20%. Moreover, there is a project reflection activity done after the first (classification) phase of the project that is worth a further 15% of the course mark.

This project runs in tandem with CMPT 2400: Data Analytics and Preparation. Both courses will use the National Pollutant Release Inventory (NPRI) dataset from Environment Climate Change Canada, which contains data on pollutant releases across Canada from 2000 to 2022. While the focus in CMPT 2400 is on exploratory data analysis and data pre-processing, CMPT 3510 emphasizes transforming this data into a time series format that supports machine learning, specifically for predicting environmental impacts through regression. You will have to be in the same team you are in for the CMPT 2400 course.

The project involves restructuring the NPRI data to make it compatible with time series analysis, performing feature engineering, and building machine learning models. The primary goal is to predict pollutant release trends, addressing various environmental impact questions. The project is split into two phases, each focusing on time series data classification and regression modeling, along with a Guided Reflection to assess your methods and approach.

These projects help you practice the concepts and skills you learned on real-world datasets. This will help you understand the kinds of issues that arise in datasets and how to handle them.

Phase 1: Classification Project (15%)

In this phase, you will:

- **Transform the NPRI dataset** into a structure that supports time series analysis, enabling future regression predictions. This initial transformation involves organizing data chronologically and ensuring compatibility with temporal modeling requirements.

- Although the main project focus is on regression, you will start by exploring a **classification task** by categorizing pollutant releases into broad classes. This serves as a preliminary step to familiarize yourself with the time series dataset and understand the patterns within the data.
- **Provide detailed explanations** of your dataset transformation choices, feature engineering decisions (if any made), and the reasoning behind the classification task. Visualizations can be included if they enhance your explanations but should be supplementary to your descriptive analysis.

Guided Reflection (15%)

After completing Phase 1, you will participate in a guided reflection exercise. This will help you:

- Reflect on your dataset transformation, feature engineering, and classification modeling process, assessing how well it aligns with the intended prediction goals.
- Critically evaluate whether your classification approach provided meaningful insights into pollutant release patterns and environmental impacts.
- Answer guided questions, such as whether the problem you framed aligns with the original regression focus and explore areas for improvement as you move into Phase 2.

Phase 2: Regression Project (20%)

Building on Phase 1, you will:

- **Further refine the dataset** for time series prediction. Using the feedback and lessons learned from Phase 1, you will enhance the NPRI dataset to ensure it is fully suitable for regression modeling, enabling accurate forecasting of environmental impacts over time.
- **Develop regression models** that predict pollutant release quantities based on various factors, such as economic conditions, regulatory changes, or industrial activity. These models will tackle the regression problems originally posed, allowing you to predict pollutant release trends into the future.
- **Engage in advanced feature engineering** to optimize your regression models. This involves identifying and creating new features that can improve model accuracy and interpretability.
- **Document the complete process** in a Jupyter Notebook, showing how you transformed the dataset, engineered features, and built regression models. You will aggregate multiple regression predictions to address the main environmental impact prediction questions posed by the project.

Learning Outcomes

By the end of this project, you will have gained hands-on experience with the following skills:

- **Time Series Data Transformation**: You will learn how to reformat and prepare datasets for time series analysis, adapting real-world data that is initially unsuitable for machine learning tasks.
- **Data Cleaning and Feature Engineering**: You will practice advanced data cleaning techniques and feature engineering tailored to time series analysis, focusing on improving model accuracy and interpretability.
- **Regression Modeling for Time Series**: You will develop regression models that address environmental impact prediction, gaining a deeper understanding of the unique challenges associated with time series prediction.
- **Critical Analysis and Justification**: You will enhance your ability to critically assess your data preparation and modeling decisions, justifying your approaches and reflecting on the effectiveness of your methods.
- **Effective Communication of Technical Work**: You will focus on providing clear, concise, and well-supported explanations of your data transformations, modeling steps, and results, using visualizations selectively to aid in your explanations.

Assignment Instructions

For this project, you will work in groups of up to four students. Each group will select a specific environmental impact prediction problem using the **National Pollutant Release Inventory (NPRI)** dataset. The assignment instructions for each phase are as follows:

Phase 1: Time Series Data Preparation and Classification

1. **Transform the Dataset**: Restructure the NPRI dataset into a time series format compatible with machine learning. Focus on making the data suitable for temporal analysis, addressing any initial issues related to time series compatibility.
2. **Develop a Classification Model**: Define a preliminary classification task by binning continuous pollutant release data into categorical classes. While the main emphasis is on time series preparation, this task will provide insights and set the stage for the regression work in Phase 2.
3. **Explain Your Approach**: Document each step in a Jupyter Notebook, explaining your data transformation choices, feature engineering techniques, and classification approach. Use visualizations only as needed to support your explanations.

Guided Reflection

- **Reflect on Your Process**: Answer specific questions about your data preparation and classification approach, evaluating how well your transformations align with the overall prediction goals.

- **Prepare for Phase 2**: Consider how your choices in Phase 1 will impact the regression modeling in Phase 2 and identify any improvements to apply.

Phase 2: Regression Modeling and Time Series Prediction

1. **Refine the Time Series Dataset**: Further adjust the dataset based on your Phase 1 experience to ensure it is well-prepared for regression analysis. Focus on enhancing the dataset's compatibility with time series modeling.
2. **Build Regression Models**: Develop and evaluate regression models to predict specific pollutant release trends over time. The focus here is on using the prepared time series data to forecast environmental impacts.
3. **Explain and Justify Your Decisions**: Provide a thorough explanation of your data preparation, feature engineering, and regression modeling steps in your Jupyter Notebook. Visualizations should be used sparingly, only where they support your explanations.

Submission: For both phases, submit your Jupyter Notebooks with well-documented code, explanations, and any visualizations on Moodle by the deadlines.

Proposed Environmental Impact Prediction Problem Using NPRI Dataset

Predictions using external factors

1. What are predicted releases of nitrogen oxides and carbon monoxide from the oil and gas extraction sector (NAICS code 211110) in 2023 if the price of oil goes up or down?
    a. Crude Oil pricing NRCan
    b. Crude oil prices Stats Can
2. What are the predicted number of facilities reporting to the program in 2023 under different economic growth scenarios.
    a. GDP by industry
3. Has the federal carbon pricing system (started in 2019) decreased emissions of nitrogen oxides and carbon monoxide (substances released from burning fossil fuels)? And if so, what are the predicted decreases in the release of these substances as the carbon pricing system gradually increases in price?
    a. Federal carbon pricing went into effect in 2019 at $20 a Tonnes, increasing by $10 each year to $50 per Tonnes in 2022. From 2023 to 2030, the pricing will increase by $15 per year.
        i. It is important to keep in mind that some provinces already have carbon pricing systems in place, more info can be found here.
4. Depending on provincial population trends, predict the amount of ammonia, nitrate ions and phosphorous released to water from wastewater treatment plants in 2023 (data)

NPRI data predictions

1. Which industry reported the most spills and predict the number of spills that will occur in 2023 and the total amount spilled.
2. Based on NPRI data, which industry is predicted to have the highest growth of releases in 2023? Which will have the largest decline?
3. Based on NPRI data, what is the predicted proportion of releases to disposals in 2023.
4. What are the predicted trends for criteria air contaminants (sulphur dioxide, nitrogen oxides, volatile organic matter, particulate matter and carbon monoxide) for 2023.
5. Based on NPRI data, which province is predicted to have the largest decrease of substance releases (air, water and/or land) in the next five years.
6. What are the expected proportions of substance sent to landfills compared to substances sent for treatment or recycling in 2023
7. What are the predicted trends for methanol and ethanol releases across Canada and predict their releases in 2023?
8. What are the predicted proportions of nitrate ions to ammonia releases from wastewater treatment plants in 2023?

Related Resources:

1. Carbon pollution pricing systems across Canada
2. Factsheet of facility fuel type distinction
3. GDP by Industry

Project Timeline and Milestones

The project follows a structured timeline, with specific deliverables due at each stage. The dates align with the 2400 course schedule, ensuring that students in both courses progress together. Here are the key milestones:

Date	Milestone	Description

September 17	**Team Formation**	Project introduction and team formation.
September 25	**Meet and Greet with ECCC**	Introduction to the project by Environment Climate Change Canada (ECCC) representatives.
October 29	**Feedback Session 1**	Submit your Phase 1 deliverables (Jupyter Notebook with explanations) and receive feedback during your class time, either 11:00 AM - 12:30 PM or 2:00 PM - 3:30 PM.
November 05	**Demo 1**	Present your Phase 1 results to the ECCC team. Demos will be held from 12:30 PM to 3:30 PM.
December 03	**Feedback Session 2**	Submit your Phase 2 deliverables (Jupyter Notebook with explanations) and receive feedback during your class time, either 11:00 AM - 12:30 PM or 2:00 PM - 3:30 PM.
December 11	**Final Demo**	Present your final project output, focusing on the regression modeling results and overall findings. Demos will be held from 12:30 PM to 3:30 PM.

Note: All feedback sessions are held during class times, while demos will be held from 12:30 PM to 3:30 PM on Tuesdays. Each deliverable should be submitted on Moodle by the specified deadline.

Project Rubric

The project is evaluated across three components: **Phase 1 (Classification Project)**, **Guided Reflection**, and **Phase 2 (Regression Project)**. Each component has a specific breakdown, and achieving **Exemplary** scores can allow students to reach the maximum score for each phase. Overall exceptional work may also earn additional bonus marks.

Phase 1: Classification Project (15%)

Criteria	Exemplary (6 Points)	Proficient (5 Points)	Basic (3-4 Points)	Needs Improvement (1-2 Points)
Dataset Transformation	Dataset is expertly organized for time series classification,	The dataset is well-prepared and meets requirements for time series classification.	Dataset is partially prepared, with some missing elements for time series.	Dataset is poorly prepared or not suitable for time series.

	with innovative enhancements.			
Classification Model	The model is highly appropriate with insightful classification label creation.	Model is suitable for the task, with clearly justified classification labels.	The model meets basic requirements but lacks depth or justification.	The model is inappropriate or lacks a clear classification approach.
Explanation & Justification	Explanations are detailed, insightful, and supported by effective visual aids.	Explanations are clear and comprehensive, with supporting visuals.	Explanations are present but limited, with minimal visual support.	Explanations are unclear or missing, with little or no visual aids.

Guided Reflection (15%)

Criteria	**Exemplary (6 Points)**	**Proficient (5 Points)**	**Basic (3-4 Points)**	**Needs Improvement (1-2 Points)**
Insight & Reflection	Provides deep insights with thoughtful analysis exceeding expectations.	Provides thoughtful reflection, adequately analyzing the process.	Reflection is present but lacks meaningful insights or depth.	Reflection is minimal or lacks detail.
Evaluation of Methods	Evaluation is thorough, with insightful connections to broader impacts.	Provides accurate and sufficient evaluation with well-reasoned analysis.	Evaluation is basic, overlooking significant aspects.	Evaluation is incomplete or lacks depth.
Connection to Phase 2	Detailed connections to Phase 2, with a forward-thinking plan.	Connections are mentioned with a reasonable plan for Phase 2.	Connections to Phase 2 are minimal or underdeveloped.	Does not connect Phase 1 outcomes to Phase 2 planning.

Phase 2: Regression Project (20%)

Criteria	**Exemplary (8 Points)**	**Proficient (7 Points)**	**Basic (4-6 Points)**	**Needs Improvement (1-3 Points)**
Refinement of Dataset	Dataset refinement is exceptional, enhancing usability for regression.	The dataset is well-prepared, meeting time series regression requirements.	Dataset is partially prepared, with missing elements for time series.	Dataset refinement is minimal or unsuitable for regression.

Regression Model Development	The model is well-developed with innovative feature engineering.	The model is appropriate for regression with clear feature engineering.	The model meets basic requirements but lacks depth in feature engineering.	The model is inadequate with minimal feature engineering.
Explanation & Justification	Explanations are comprehensive, with effective visual aids.	Explanations are clear, with adequate support from visual aids.	Explanations are basic, with minimal visual support.	Explanations lack clarity, detail, or necessary visual aids.

Grading and Bonus Points

- **Exemplary**: Reaching the highest scores within each category allows students to achieve the section maximum. Exceptional work across all sections may result in up to **5 additional bonus points** beyond the total project score.
- **Proficient**: Meeting all requirements at a high standard ensures full marks for each component.
- **Basic**: The submission meets minimum expectations but lacks detail or thorough justification.
- **Needs Improvement**: Submission falls short of basic requirements in detail or quality.

Demo Presentations

Throughout this project, you will have the opportunity to present your work to **Environment Climate Change Canada (ECCC)** representatives in **two demo sessions**. These sessions not only allow you to showcase your technical progress but also help you develop essential professional communication skills required in real-world machine learning projects.

Demo 1: Phase 1 Presentation

Your first demo will focus on the results of Phase 1. This presentation should highlight your dataset transformations and initial classification approach. During this session:

- **Provide a Clear Narrative**: Guide the ECCC representatives through your Phase 1 methodology, from data preparation to your classification task setup. Make sure your presentation is logically organized and easy to follow.
- **Highlight Practical Insights**: Explain how your approach helps in understanding pollutant release patterns, even at this preliminary stage. Make connections to potential real-world applications where possible.
- **Professional Presentation**: Aim for concise and professional delivery, just as you would when presenting to a client. Practice explaining technical concepts in an accessible way for both technical and non-technical stakeholders.

Final Demo: Phase 2 Presentation

The second demo will showcase your comprehensive results, including the regression modeling phase. This session is an opportunity to demonstrate the full scope of your work and how it addresses the project's goals. Key expectations for this demo include:

- **Well-Organized Insights**: Present your data transformation, feature engineering, and regression modeling steps with a clear focus on their contributions to understanding environmental impacts.
- **Data-Driven Recommendations**: Use the insights from your models to explain possible actions or implications. Show how your work supports data-driven decisions that ECCC could leverage.
- **Preparedness and Professionalism**: Treat this final presentation as a client-facing demo. Be prepared to answer questions, provide additional context, and justify your methodological choices. Demonstrating preparedness will give you valuable experience in managing stakeholder expectations.

In both demos, ECCC representatives will evaluate not only your technical proficiency but also your ability to deliver a clean, well-organized presentation. Clear communication is critical when presenting machine learning projects to clients, and these sessions will help you refine this skill for future professional settings.