**ENEC490.003.SP24 Spring 2024: Assignment 4**

**Due: 5/6/2024 by 11:59pm (no grace periods or late assignments)**

**Summary:** For this assignment, we will examine how classification methods can be used to predict the probability of different outcomes and help us make complex decisions involving economics and decarbonization.

You will assume the role of Director of ESG for a large Fortune 500 company that has set aggressive net-zero goals. It is your job to help your company decarbonize as quickly and cost effectively as possible. But you can’t do this through converting all your company’s fleet, infrastructure, and facilities to renewables fast enough to meet your goals. So, to supplement those conversions, you are going to buy carbon credits by investing in energy projects in Seattle, where you are headquartered. Your budget for the first year of investments is $500,000. To be successful, you need to acquire carbon savings at a cost of $150/MTCO2e or less, the total simple payback for the projects you invest in needs to be 5 years or less, and the projects you invest in need to have a predicted probability of 80% or higher that they will achieve the estimated carbon savings. Which projects will you invest in to achieve these goals?

**Objective:** The objective of this assignment is to find a group of energy conservation measure (ECM) projects in Seattle that meet the specified budget and performance criteria. To accomplish this, first you will build a classification model on historical data for 3,000 ECM projects in Seattle that uses building characteristics, energy data, and ECM data to predict when a project’s actual carbon savings will be greater than or equal to the expected carbon savings. Your classification model should have an accuracy of 0.7 of higher using the accuracy\_score metric from sklearn.metrics and your test data set. Once you have built a classification model that meets this performance metric, you will use that model to predict the probability that a set of 300 proposed ECM projects will meet or exceed the expected carbon savings.

Using this information, you will select a group of projects that fit within the budget and achieve the desired carbon savings rate and simple payback criteria. You do not need to use multi-criteria decision analysis (MCDA) to select your group of projects but can if you choose to or you can use another method to find a group of projects that meet those criteria.

**Data Files:** There are two data files for this assignment:

1. **2022\_Building\_Energy\_ECM\_and\_Benchmarking\_20240407.xlsx:** this is the data for historical ECM projects in Seattle including building characteristics, energy data, and ECM project data. You can go to [this page](https://data.seattle.gov/Permitting/2022-Building-Energy-Benchmarking/5sxi-iyiy/about_data) to see descriptions of the building and energy data columns from City of Seattle. For the ECM/project data, here are some column interpretations to help:
   1. ECM\_Cost: this is the total project cost for the listed ECM
   2. EUL: this is expected useful life (EUL), which is measured in years and estimates how lifetime of the project.
   3. Expected\_Annual\_Carbon\_Savings: the estimated annual carbon savings in MTCO2e for the project.
   4. Expected\_Annual\_Cost\_Savings: the estimated annual cost savings for the project
   5. Actual\_Annual\_Carbon\_Savings: the actual annual carbon savings for the project in MTCO2e
   6. Actual\_Annual\_Cost\_Savings: the actual annual cost savings for the project
2. **2024\_Building\_ECM\_Estimates.xlsx:** this is the data for the set of 300 proposed ECM projects. It has the same building characteristic, energy, and ECM project columns as the historical data except we no longer have the actual annual carbon and cost savings, since these are proposed projects that you are considering investing in. This file also includes two new columns, which you can use to analyze the carbon and financial impacts:
   1. Expected\_Total\_Carbon\_Savings: this is the total expected carbon savings in MTCO2e for the project and is calculated by multiplying EUL (the project’s lifetime in years) by the expected annual carbon savings.
   2. Expected\_Total\_Cost\_Savings: this is the total expected cost savings for the project and is calculated by multiplying the EUL (the project’s lifetime in years) by the expected annual cost savings.

**Steps to include in your analysis:**

1. Load the data from the 2022\_Building\_Energy\_ECM\_and\_Benchmarking\_20240407.xlsx file
2. Add a column called Actual\_Carbon\_Savings\_Greater\_Than\_Expected, which you can calculate as a 1 when Actual\_Annual\_Carbon\_Savings >= Expected\_Annual\_Carbon\_Savings and a 0 if not. This will be your target variable for your classification model.
3. Clean and feature engineer the data: you might choose to do this iteratively and see how it effects results, but you will at minimum need to drop rows that have NA values from your predictor variables. This data set is (relatively) clean, so of the 3,000 rows of data you should have at least 1,500 or more rows remaining before you do your train/test split to build your classifier.
4. Build your model(s): choose your X features and use train/test split with test\_size=0.2 to build your classifier using Actual\_Carbon\_Savings\_Greater\_Than\_Expected as your y variable. You can try any of the classifiers we looked at in class and are also free to tune hyperparameters on the model(s) to improve accuracy.
5. Use sklearn.metrics accuracy\_score to ensure your chosen classifier model has an accuracy score of 0.7 or higher. (note: you shouldn’t need to do tons of feature engineering/cleaning to get above this threshold so please let me know if your models are coming up far short of this so we can troubleshoot them together)
6. Once you have chosen a classifier that meets the accuracy score threshold, load the 2024\_Building\_ECM\_Estimates.xlsx file.
7. Set your X features from the ECM Estimates files, be sure to perform the same feature engineering steps on these columns as you did when you fit your classifier.
8. Predict the probability that each ECM project will achieve actual carbon savings above the estimated carbon savings using the .predict\_proba function. You only want to invest in projects with a high probability of achieving carbon savings at or above the estimated values, so you should filter to select projects where the predicted probability is greater than or equal to 0.8.
9. Now that you have your list of projects that you are highly confident will attain the estimated carbon savings, you need to find a group of projects that meet these criteria:
   1. Fit within your $500K budget but use at least $400K. So, your total ECM\_Cost for the selected projects should be between $400K and $500K.
   2. The cost of your carbon savings should be $150/MTCO2e or less. So, your total ECM\_Cost divided by the sum of your Expected\_Total\_Carbon\_Savings for your selected projects should be $150 or less.
   3. The total simple payback for your selected projects should be less than 5 years. So, your total ECM\_Cost divided by the sum of your Expected\_Annual\_Cost\_Savings for your selected projects should be 5 or less.

**Deliverable:** submit a jupyter notebook with your python code and comments showing your approach to the assignment, how you built your classification model, predicted the probabilities for each ECM project, and selected your group of projects. You should also clearly list the group of projects you selected by their ID and the total ECM\_Cost, cost of carbon per MTCO2e, and total simple payback for your selected projects.