**Predictive Modelling of Toronto Ferry Redemption and Sales Counts**

**By: Bill Wang**

Non-Technical Description

This exercise uses three machine learning models to predict redemption count and sales count at the Toronto Ferry terminal. The primary inputs of the model are historical redemption count and sales count data at the Toronto Ferry. “New Model” also additionally use historical weather data collected by Environment Canada. Below is a summary of the basic features as well as the performance of each model.

1. Base Model

* This model predicts redemption count of a given calendar day of the year by averaging a set of historical redemption counts that occurred on the same calendar day
* The average prediction error is about 70% of the average daily redemption count

1. Improved Model

* This model also predicts by averaging a set of historical redemption counts that occurred on the same calendar day, but it improves upon the base model by placing more importance on more recent data points
* The average prediction error is about 69% of the average daily redemption count

1. New Model

* This model predicts sales count by considering the weather (average temperature) of a given day and whether the day lies on a weekend
* The average prediction error is about 56% of the average daily redemption count

Technical Description

Data Sources:

1. Toronto Island Ferry Ticket Counts Data: January 1, 2015 – June 6, 2025
2. Daily Climate Data, Toronto, Environment Canada: January 1, 2015 – June 6, 2025

Base Model

This is a time-series model with one component of seasonality at the level of day of the year. The machine learning component trains the model 4 times using different splits of training and testing samples. The model predicts the redemption count in the testing samples by averaging redemption counts in the training sample that occurred on the same calendar day. The identifying assumption of this model is that some time-invariant factor about a specific calendar day can explain variation in Toronto Ferry redemption counts. This is a very restrictive assumption.

Improved Model

This model adds in a time-variant component of seasonality. The model is still predicting the redemption count in the testing samples by averaging redemption counts in the training sample that occurred on the same calendar day. However, more recent observations in the training sample are weighted more heavily. The assumption here is that more recent data is more predictive than less recent data. I think this makes sense because the longer time has elapsed, the less probable the time-invariant factor (or factors) still exists. As we can see from the model predictions, this model is indeed an improvement to the base model. A possible refinement to my improved model is to optimize the weighting matrix. Due to time constraints, I used a simple linear function to weight observations. However, if we were to optimize weights using a similar training-testing split, we could further improve the performance of the model.

New Model

In the previous two models, we tried to make predictions by capturing elements of seasonality. However, we do not attempt to understand the underlying mechanisms driving seasonality. What’s actually going on in the data generating process? What are the causal factors explaining traffic flow at Toronto Island on a given day? Here I suggest two causal factors: 1) the day of the week, and 2) weather conditions. In the first case, we would expect higher traffic on weekends compared to weekdays, when many people are at work or school. In the second case, we would expect higher traffic on days where the temperature is warmer. The model is a linear model with these two explanatory variables. I create a dummy variable that indicates whether a day falls on a weekend, and a second variable indicating the mean temperature on a given day (merged from Environment Canada data). The dependent variable is sales count. Like the previous two models, I create 4 training-testing splits, where the regression is done using the training sets, and validated with the testing sets. As a side-exercise, I also ran this model using redemption count as the dependent variable (for comparability), and it performed significantly better than the previous two models.