Report Draft

Andrew H, Grace C, Nick H., Emily H.

2024-12-02

Introduction

While pondering a topic, we discovered that our group shared a mutual distaste for the random scattering of E-Scooters and how seemingly popular the pay-to-ride service was becoming in Fort Collins. This idea sparked our interest - we wanted to know just how popular these scooters have become over the last few years. Our team reached to reach out of Fort Collins Transportation Planner, Rachel Ruhlen, whom gave us the data that SPIN had been gathering since the service's launch in 2021.

Abstract

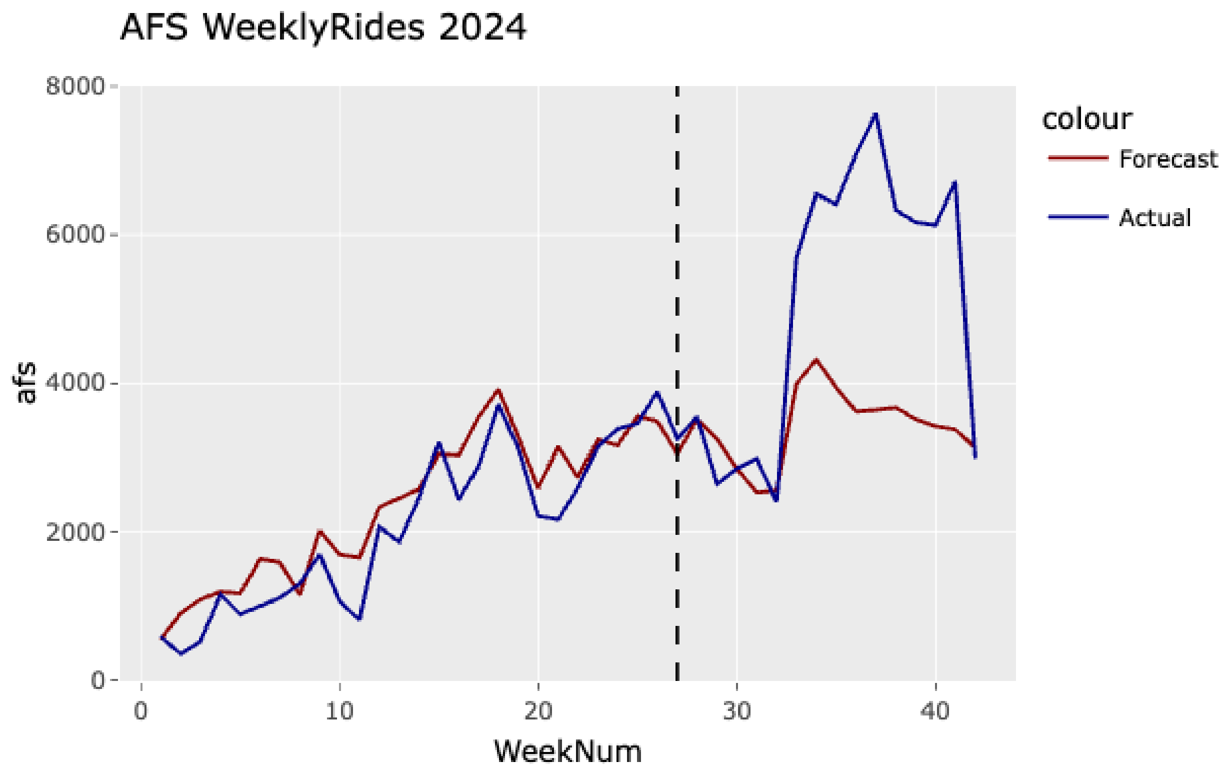
Fort Collins is widely considered one of most commuter friendly cities. As the push for Eco-friendly transportation methods rise, local E-Scooter/Bike companies like SPIN, are starting to become more widely available. This report aims to use statistical machine learning methods - Autoregressive integrated moving average (ARIMA), K Nearest Neighbors (KNN), Seasonal-Trend Decomposition (STLM), Boosting, etc… - and spatial geographical data to understand and forecast the longevity of ridership.

Methods

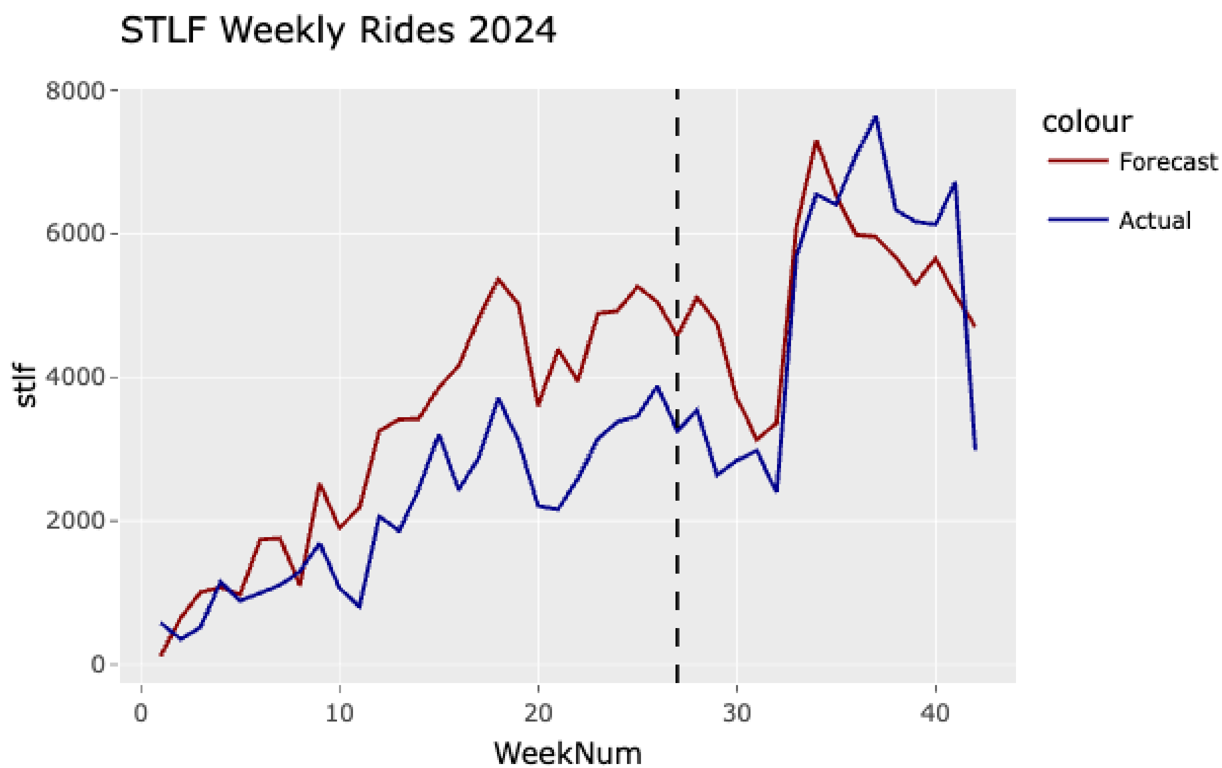
Using the "Date" column, we created month, year, and week number columns. We used separate data frames for weekly and monthly data. In each of these data frames, we grouped by the time frame of interest to get the sum of scooter rides in each interval. We then created two new columns for the normal and college prices. In both analyses data from 2024 was used as testing data, and previous data was used as the training data in the form of a time series. The accuracy of the models in both analyses was measured using the mean absolute percentage error (MAPE). In both analyses, several models were tested, but only the most accurate models will be included in this paper. The college price decrease initiated in July 2024 caused a huge spike in ridership that was not captured by most models. Thus, in both analyses, the final model is a combination of two models: one that was accurate before the price change and one that was accurate after the price change.

Weekly Rides

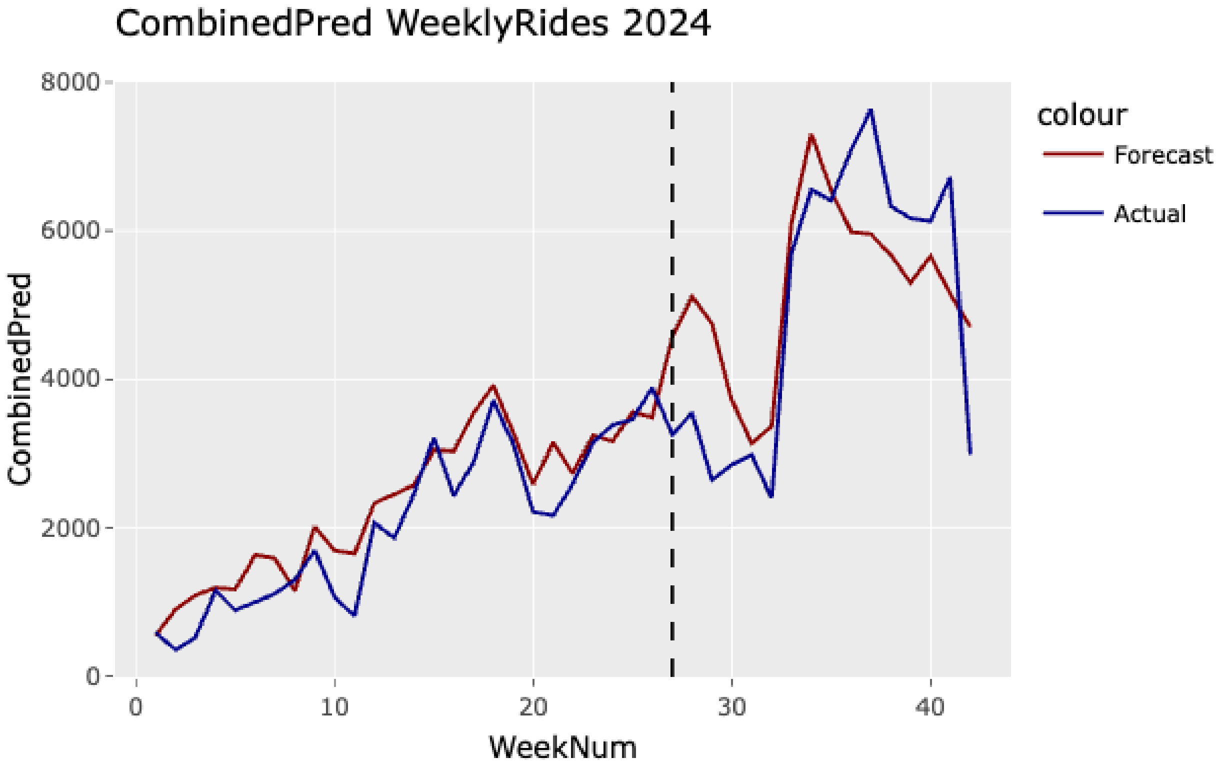
Before the price change, a hybrid model composed of an ARIMA, STLM, and THETAM model (weighted equally) most accurately predicted ridership when fit on the testing data, with a MAPE of 29.69. The plot below compares the actual ridership values to the forecasted values; the date of the price change is marked by a dotted black line.



An STLF model most accurately captured the spike in ridership when fit on the testing data, with an overall MAPE of 47.12.

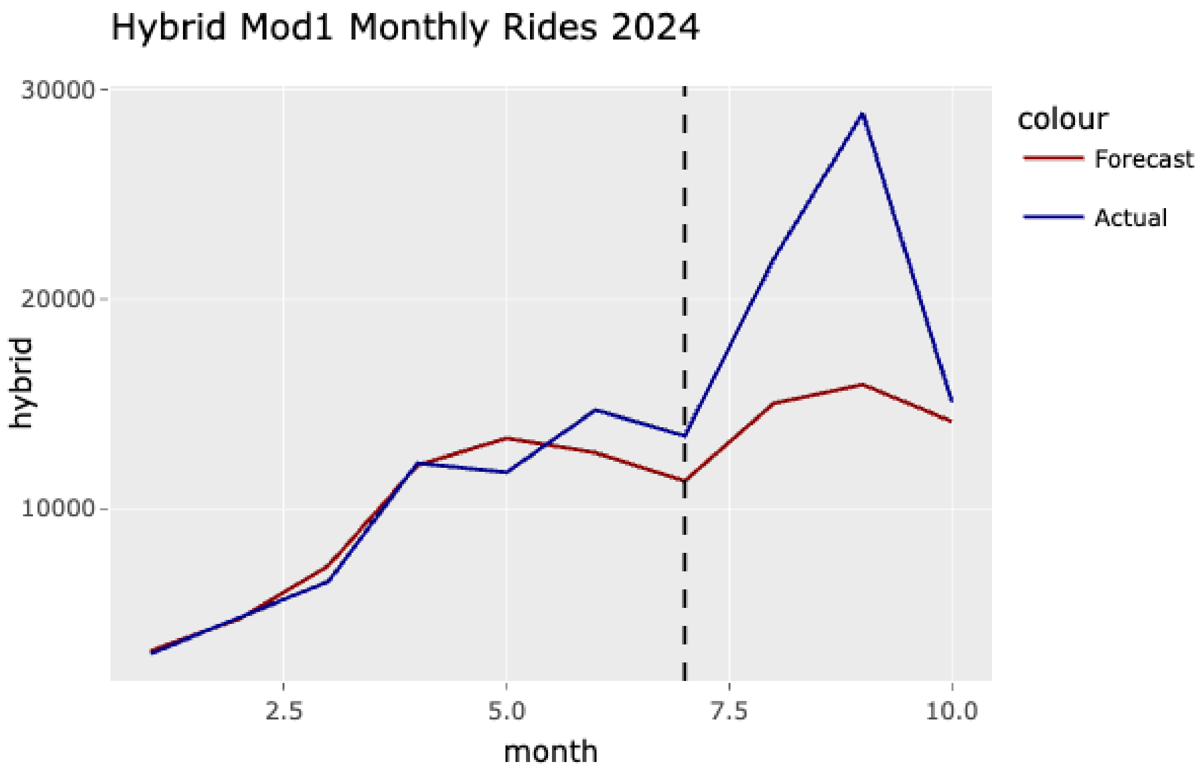


The final model for predicting weekly ridership uses the hybrid model before the price change and the STLF model after the price change. This model accurately captures the trend before the price change and the spike after the price change and has a MAPE of 29.04 on testing data.

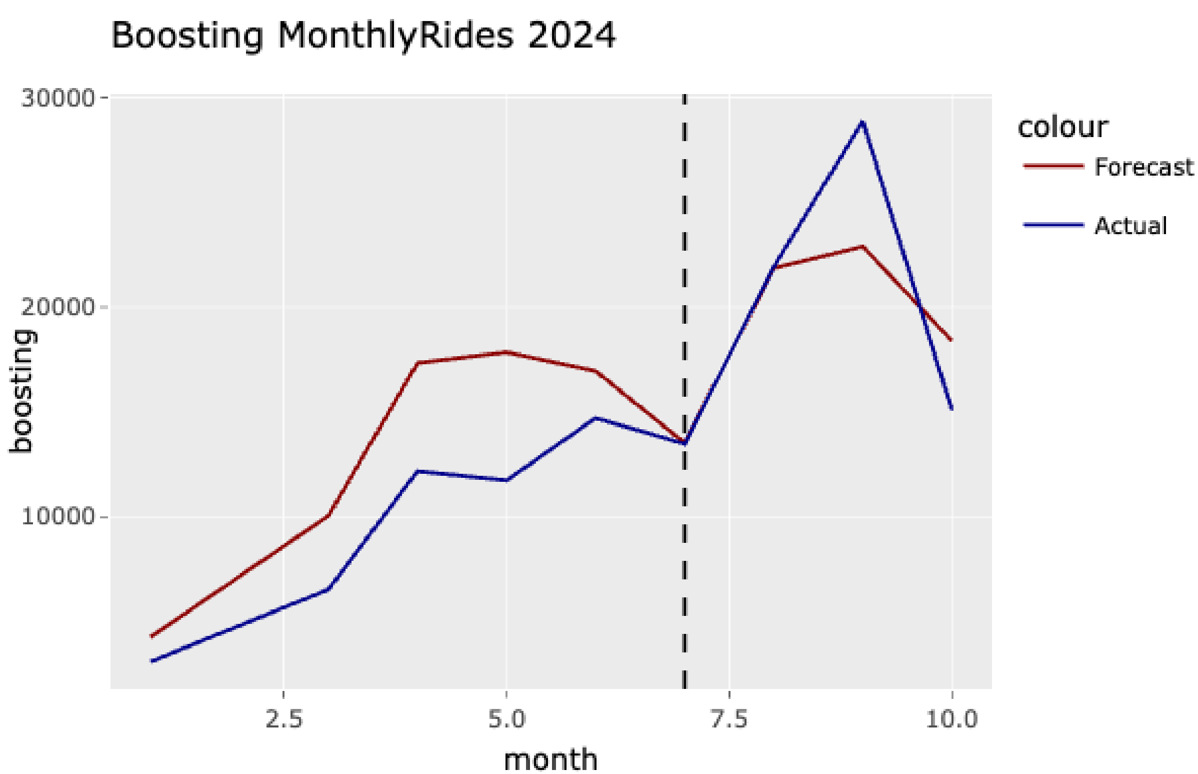


Monthly Rides

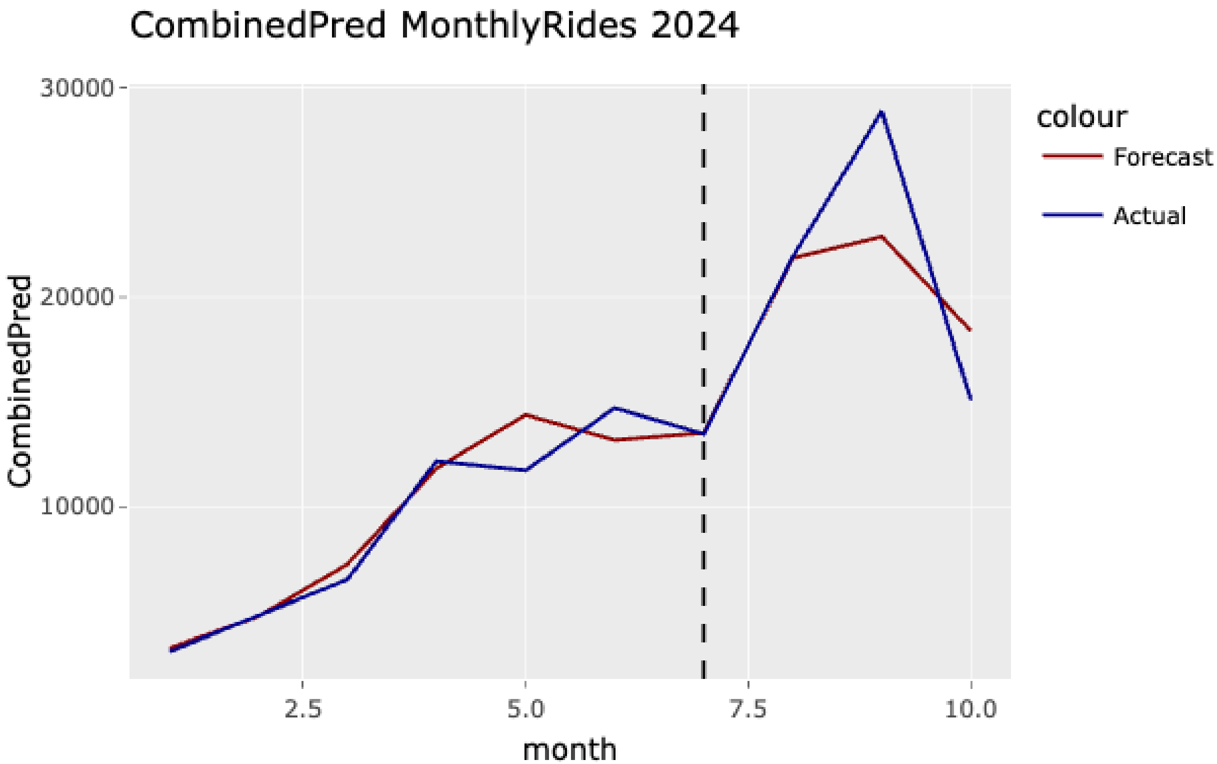
Before the price change, a hybrid model composed of an ARIMA, STLM, THETAM, NNETAR, ETS, and TBATS model (weighted equally) most accurately predicted ridership when fit on the testing data, with a MAPE of 14.39.



A boosting model most accurately captured the spike in ridership when fit on the testing data, with an overall MAPE of 29.39.



The final model for predicting weekly ridership uses the hybrid model before the price change and the boosting model after the price change. This model accurately captures the trend before the price change and the spike after the price change and has a MAPE of 8.86 on testing data.



KNN - Geospacial

SPIN provided our group with three separate "curb event" files containing geographic coordinates for the deployment, start, and end locations for spin scooters across Fort Collins. The rows included purely spatial data (latitude, longitude) that were unconnected to each other file-wise, as-in there was no way to tell which scooter starts at the pinged location. The data also included miscalculations like missing negatives that created large outliers in the data where certain scooters reported their start in places like Shanghai.

We mapped the data and compared the density score of the predicted results on a prediction grid of Fort Collins to determine if deployments are being distributed in an optimal matter. The model should help us determine if there are too many scooters being deployed in an area, and find areas where scooters are not being deployed enough to fit scooter demand.

A 20-fold cross validation was used to divide the map Fort Collins into 20 grid squares to compare our predictions. 20 was chosen as a tradeoff between complexity and accuracy. Both models selected an epanechnikov kernel and best k = 15. The deployment KNN model reported a means squared error of 0.04891901 and the start KNN model reported a means squared error of 0.04380498.

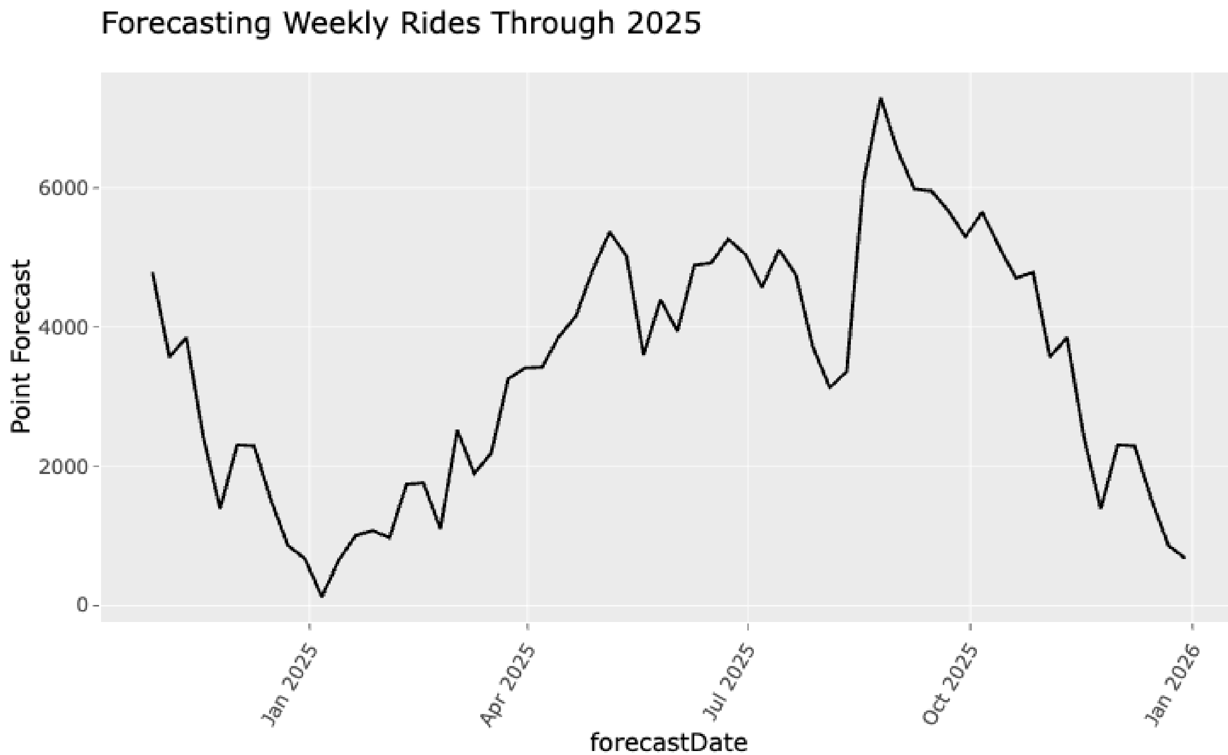
We made comparisons of the model predictions from our KNN regression to create a mesh of points across the map of Fort Collins. The deployment model predicts what deployment density should be at that location, and the start model predicts what the usage density should be. These represent the "expected" levels of deployment vs. usage, smoothed out to account for day − to − day variations.

Results

The predictions for future weekly and monthly ridership will be made assuming that the normal and college prices remain the same.

Weekly

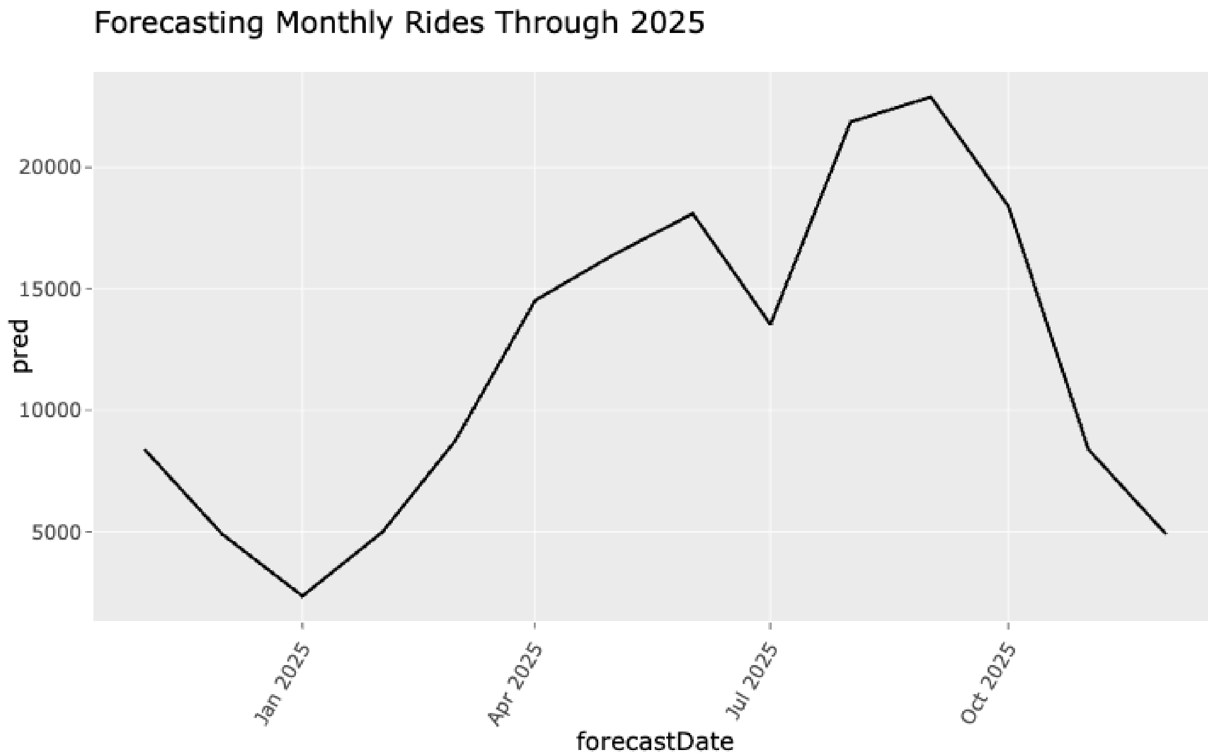
Since the STLF model is used for predicting after the price change, this model was used to predict ridership through the end of 2025. The plot is shown below, beginning at the end of the testing data.



In 2025, the week with the lowest ridership (around 122) will be the first week of January 2025 and the week with the highest ridership (around 7300) will be the week of August 25. The January prediction seems slightly low. The August prediction seems accurate, since it is slightly higher than that week in the previous two years.

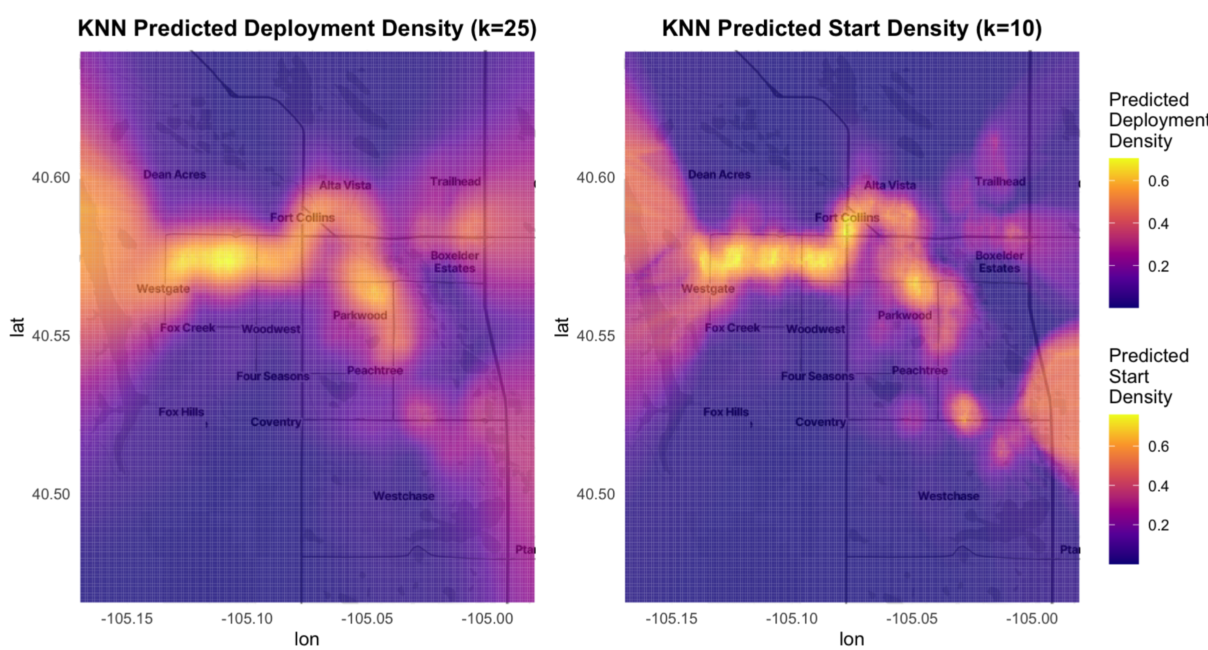
Monthly

Since the boosting model is used for predicting after the price change, this model was used to predict ridership through the end of 2025. The plot is shown below, beginning at the end of the testing data. *Note: this will not track a trend, will now produce the same predictions for each January, February, etc.



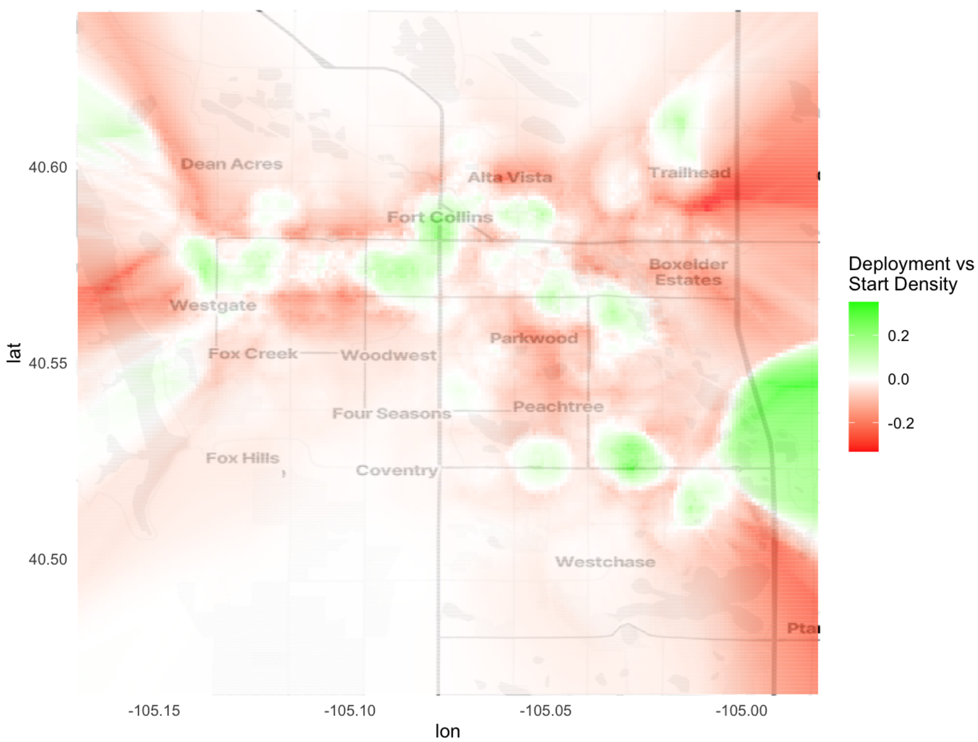
In 2025, SPIN can expect January to be the month with the lowest ridership (around 2350) and September to be the month with the highest ridership (around 22900). The January prediction is slightly lower than all previous January values. The September prediction is about the same as the September 2022 value.

KNN Deployment & Start Heatmaps



Compared side by side, the two models bear a strong resemblance to each other with the deployment model showing a more gradual gradient compared to the start model. This matches with the observations when comparing the raw data. There are notable edge effects in areas where the model is predicting past the SPIN service area in Fort Collins.

KNN Deployment Optimization Map



This map highlights where SPIN can improve their deployment strategies. Smooth transitions between colors on the map represent gradual changes in the relationship between deployment and usage patterns, which helps understand not only which areas to make changes, but just how significant those changes should be.

The model is less effective reaching toward the outer bounds of the service area in the data. As it reaches the outer bounds an edge effect takes hold and spreads the last noticed pattern past the service area of the data into spaces where no scooters are used.