Modeling Bike Rentals with ML

## presented by

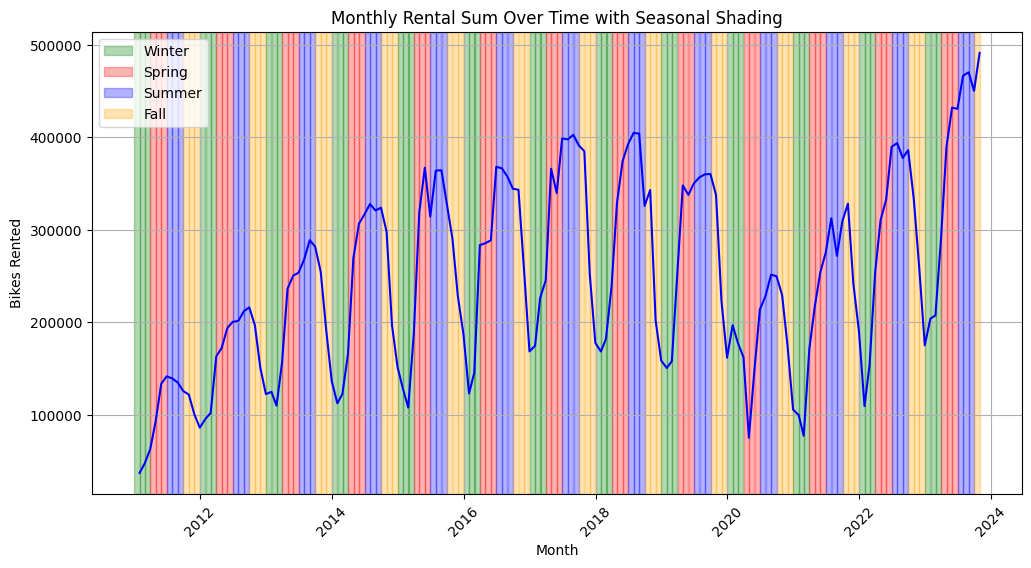
Kepler Ridge

Marshall Potts

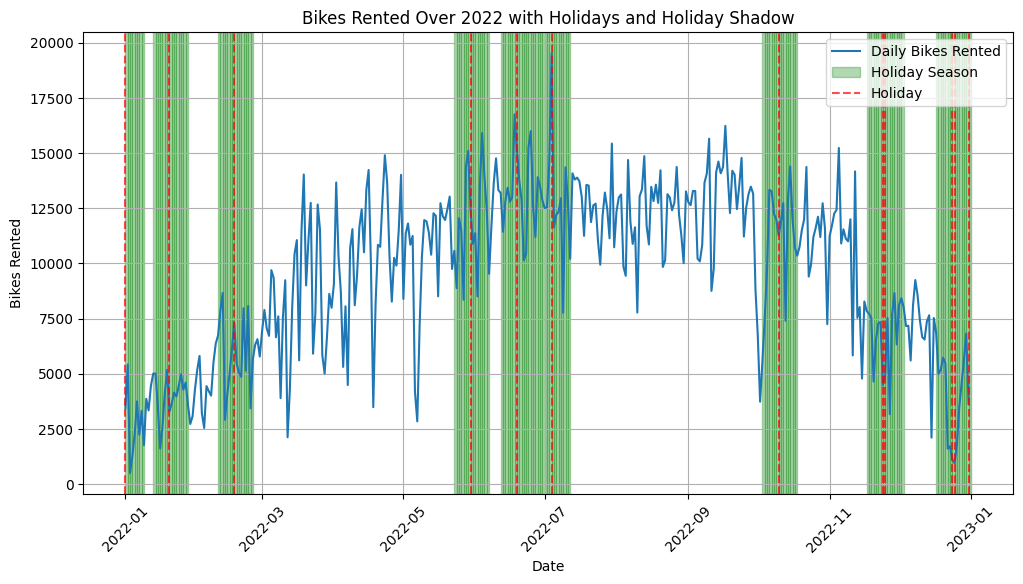
Taein Oh

1. Discovery and Methodologies

Identifying meaningful relationships in the data was key to building the model. The first observation was the significant volatility in bike rentals over time, driven by the season. The data clearly shows that rentals drop in winter and fall, while spring and summer see a marked increase.



On a more detailed level, it was observed that holidays, as well as the days surrounding them, can also impact bike rentals. This is highlighted in the following chart.



A fter identifying the impact of holidays and seasons, features were engineered to enhance model performance. The key improvement was correctly labeling holidays, as some data points were misidentified or not labeled. A new feature, "holiday shadow," was created for days within a week of a holiday, and Fridays following a holiday were also marked as holidays.

An external weather API was tested to include precipitation levels, but this added complexity without improving performance, so it was excluded. Continuous data was normalized, and date information was converted to an epoch timestamp.

Once the data was cleaned, initial models were built, with challenges in determining the correct layers, learning rate, optimizer, batch size, and number of epochs. Using keras-tuner and running hundreds of tests, the optimal parameters were found: two Dense layers with 32 and 256 perceptrons, ReLU activation, followed by a 20% dropout and output layer. The best learning rate, found through grid search, was 0.01.

With these parameters, the model performed well on the training set, achieving a Mean Absolute Error of 58.86, Mean Squared Error of 8568.46, Root Mean Squared Error of 92.57, and an R2 of 0.93. The most impactful hyperparameters were the learning rate and batch size, which significantly affected model performance.

1. Python Notebooks

https://gist.github.com/keplerridge/6e4bd4f05c4a39d13d8de484af866600

1. Discussion Responses
2. How many layers do you think the network should have?
   1. It was detemined through a grid search methodology mentioned above that 2 dense layers and one output layer would be best with one dropout between the second dense layer and the output layer.
3. If you run and the model and the results are lower than expected which hyperparameters, do you feel have the most potential to improve?
   1. The hyperparameters that seemed most impactful were the learning rate and the batch size. These two causes tremendous changes to the performance on our training model.
4. Talk about feature engineering
   1. Described above in the methodologies section. A couple more that didn’t seem to be as impactful but did improve the others but still slightly improved the model was marking every sample in 2020 and 2021 as positive for COVID as well as specifying commuting hours from 7-8AM and 4-5PM.
5. How to find an optimal learning rate
   1. This was also found using the grid search method mentioned in methodologies.
6. How will we know if the model has strong predictive power - What is our loss function?
   1. To assess our model's predictive power, we used multiple metrics. We checked the R2 value to evaluate how well the predictions fit the data and used Mean Absolute Error (MAE) to measure accuracy, as it is resistant to outliers, which is important for Neural Nets. MAE provided an easy way to track the model's performance during each epoch.
7. Could we use personal data to add insurance premium and what are the implications? Is it ethical?
   1. A limitation of the data is the lack of bike damage or repair records. Without labeled data for damaged bikes, training a reliable model is difficult. Additionally, using personal user information could introduce bias. Registered users typically rent bikes between 7-8 AM and 4-5 PM for commuting, while casual users rent more in the evening, between 6-8 PM. This time-based usage pattern provides a better foundation for risk assessment than relying on static user profiles.
8. What days and times could we pull bikes for maintenance?
   1. Registered and casual bike usage is very low between 1 AM and 4 AM, suggesting this is the optimal time for maintenance and disinfection with minimal impact on revenue. Even up to 11 PM and midnight, there are still over 200 users. The next graph shows that Sunday and Monday have the fewest users, making these the best days for these services. A graph of different colored lines

      AI-generated content may be incorrect.

The number of registered and casual users is very low between 1 AM and 4 AM, making this the ideal time for bike maintenance and disinfection with minimal revenue impact. Even up to 11 PM and midnight, there are still over 200 users. As for the best days for these services, the next graph shows that Sunday and Monday have the fewest users. So removing bikes between 1AM and 4AM on Sundays and Mondays would be best. A graph with orange lines

AI-generated content may be incorrect.

1. Are we back on track from COVID or still recovering?
   1. From 2021 onward, there has been a gradual recovery, with numbers increasing throughout the year. Although rentals haven't fully returned to 2019 levels, the upward trend suggests a steady rebound as mobility and commuting normalize.