Biking Company

Estimated Bike Users Report

presented by:

Joseph Linton

Copeland Carter

Camilo Hozman

Phillip Marvin

# I. Hyperparameters

To start off, we decided to train an XGBoost decision tree model to serve as a baseline to grade the neural network model, because the decision tree model is quicker and easier to train. It got R2 scores between 0.781 - 0.828, which serves as a sort of “upper bound” for comparison.

The final model we ended up using has 5 layers: 4 densely connected layers (the final one having 1 node representing the number of total bikes), and 1 35% dropout layer between the first and second layers. The final number of trainable parameters ended up being 42,497. Leaky RELU and RELU seemed to work best as the activation functions. The learning rate worked best at 0.0001. We decided to use early stopping (with a patience value of 7) in order to stop the model at the optimal epoch, instead of train the model for a very long time and retraining it at the previous model’s optimal epoch. This seems a more efficient solution.

For the loss function, we decided to use the mean squared logarithmic error function. We chose this because of the scale of our target feature, and the fact that the data was organized as a time series.

II. Insights

We discovered something interesting when trying to use an implementation of the [Hyperband Algorithm](https://jmlr.org/papers/volume18/16-558/16-558.pdf). Because the algorithm doesn’t take into account the test set, given the option to, it tends to choose very large models. We suspect this is because a bigger model will be able to better memorize the data. However, this leads to the model overfitting to the data and picking hyperparameters which are actually very bad once evaluated against the test set.

We determined that about 1 million parameters were required to completely overfit the data. A fairly good model came from approximately halving the number of nodes in each of the layers of that model. Intuitively, this makes a sort of sense. If you assume that a decent model is at least half as good as a perfect model, then halving the model’s ability should, in theory, descend to a decent model instead of an overfitted model.

Another thing we learned that was key to success, was where and how to split the data. Because the data is formatted in a time series, randomly sampling data was considered to be a poor choice. We tried splitting the data 80% from the front, (training on the first 80% and testing on the latter 20%) and kept getting bad results. We were unable to figure out why, until we tried splitting the data at 25% from the front (testing on the front 25% and training on the latter 75%) instead. This resulted in dramatically better results, particularly in the neural network model. With the data split at the back, the best model ran for only 114 epochs, whereas with the data split at the front, it ran for twice that amount of time, indicating more learning was able to be done. See the tables below for details.

XGBoost Data Prediction Errors

|  | Data split at the back | Data split at the front |
| --- | --- | --- |
| Test Set | R2: 0.781 | R2: 0.726 |
| Mini Holdout Set | R2: 0.828 | R2: 0.736 |

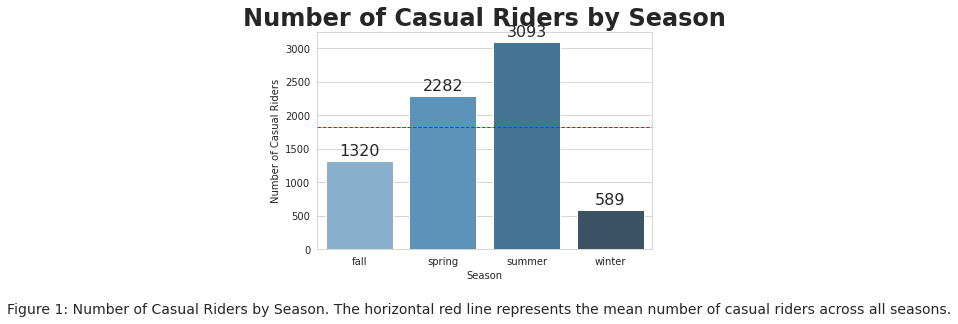
Table of the data of the final model with training data split at the back

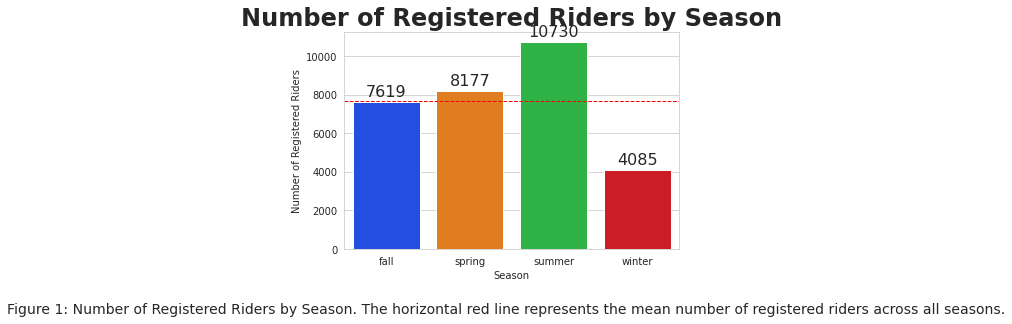
|  | Learning History | Test Set Errors | Mini Holdout Errors |
| --- | --- | --- | --- |
| Data split at the back |  | R2: 0.447 | R2: 0.884 |
| Data split at the front |  | R2: 0.447 | R2: 0.828 |

# 

# 

The correlation is significant between casual and registered riders. The p-value for the regression coefficient of the "casual" variable is very low (p < 0.0001), indicating that there is a significant linear relationship between the number of casual riders and the number of registered riders. The R-squared value of 0.524 also indicates that the linear regression model explains 52.4% of the variance in the number of registered riders.





Comparing the number of casual riders to the number of registered riders by season, we can see that summer has the highest number of both types of riders.

However, the difference between the number of casual and registered riders is much larger for summer compared to the other seasons.

This suggests that summer is a popular season for both casual and regular biking, but there may be more casual riders taking advantage of the warm weather and longer days in summer compared to the other seasons.

Additionally, winter has the lowest number of riders, both casual and registered, indicating that colder weather may deter people from biking during this season.

# 

# V. Python Notebooks

Below are Github Gist links to the notebooks we used during this case study:

(Note: It was decided that we’d each create our own Notebooks)

Joseph: [Notebook](https://colab.research.google.com/drive/1b9_sI8_dEZGHuhgvDPaJCZsMdwdjXs4j?usp=sharing),

Copeland: [Notebook](https://colab.research.google.com/gist/smartycope/1c6c724ccba63b34f1a4e0668049da69/bikerentalproject.ipynb)  
Philip: [Bikes](https://colab.research.google.com/drive/1embxCh1eAGmtkGGT0qV4cgfzCpax6BX1?usp=sharing)  
Camilo: [Notebook](https://colab.research.google.com/drive/1S17Lw9n8MtT1z6LAKuy9YkgQug0ngBvd#scrollTo=c_W2X7Tys8BI)

[Casual Riders by Hour of the Day](https://raw.githubusercontent.com/phi1ny3/Bike-CSVs/main/Casual%20riders%20by%20hour%20of%20day.png)

[Raw data correlations](https://raw.githubusercontent.com/phi1ny3/Bike-CSVs/main/Correlations.png)