WelcomeBike Bike Rentals

Executive Summary

## Presented by

Jed Chang

Bridger Norman

Celeste Popoca

Luke Russell

Kelin Tang

Ella Yang

1. Summary

As data analysts for WelcomeBank Bike Rentals, we received bike rental data to determine whether we could accurately predict the number of bike rentals for a given day and hour for their Washington DC branch. Based on those features we make one-hot encoding on season and weathersit. We make two outputs casual and registered. Our calculations lead us to get an R2 correlation score of 0.87, which is a good positive correlation. This means our model fits the data’s pattern well. With all of this information, we should be able to make our strategy better, such as how many bikes to put in which area.

1. Methodology

Process:

1. Create a vanilla model to know
2. Convert the Date column to datetime format
3. Created graphs to visualize the data
4. Based on the visualizations, we chose which features would most effectively return the desired output
5. One hot encode datasets
6. Scale the data
7. Plot loss during training
8. Tested our model with new data
9. Determined how well the model predicted the outcome

We used the vanilla model to create a base to compare how well we increased the probability of guessing outcomes correctly. To increase the probability, we manipulated the way the model interpreted the data by adding new columns from existing columns, like the date and time columns. The categorical data were one-hot encoded or changed to numerical categories to help the model interpret the data better.

Once we shaped the data to be easily interpreted by the model, we scaled the data. In other words, the data’s features or columns were normalized between a range of 0 and 1. This prevents any data from standing out and causing overfitting.

After the scaling, we made multiple plots to show our progress with training the data and to see if the data was trained too much and overfitted. We also created graphs to help us visualize any patterns that were hard to see in the list. Once we trained the data enough times, we tested the model with new data to determine how well it predicted new outcomes. We measured the probability using the R2 correlation (see the next section for the definition).

1. Results, Action Items, and Limitations

The bike rental prediction model has a high R2 score of 0.901, which means it has roughly a 90% correlation with the variation in total bike rentals [see Table 1]. This parameter comes from testing the neural network on a set of data the model was not previously trained on.

| R-Squared ( R2) | 0.901 |
| --- | --- |
| Root Mean Squared Error (RMSE) | 49.15 |

(Table 1)

Note:

**R-squared (R2)**: “R2 score as a measure of how close the predicted values are to the actual values. It ranges from 0 to 1, where 0 means the model does a poor job of explaining the data, and 1 means the model perfectly predicts the data.” [2]

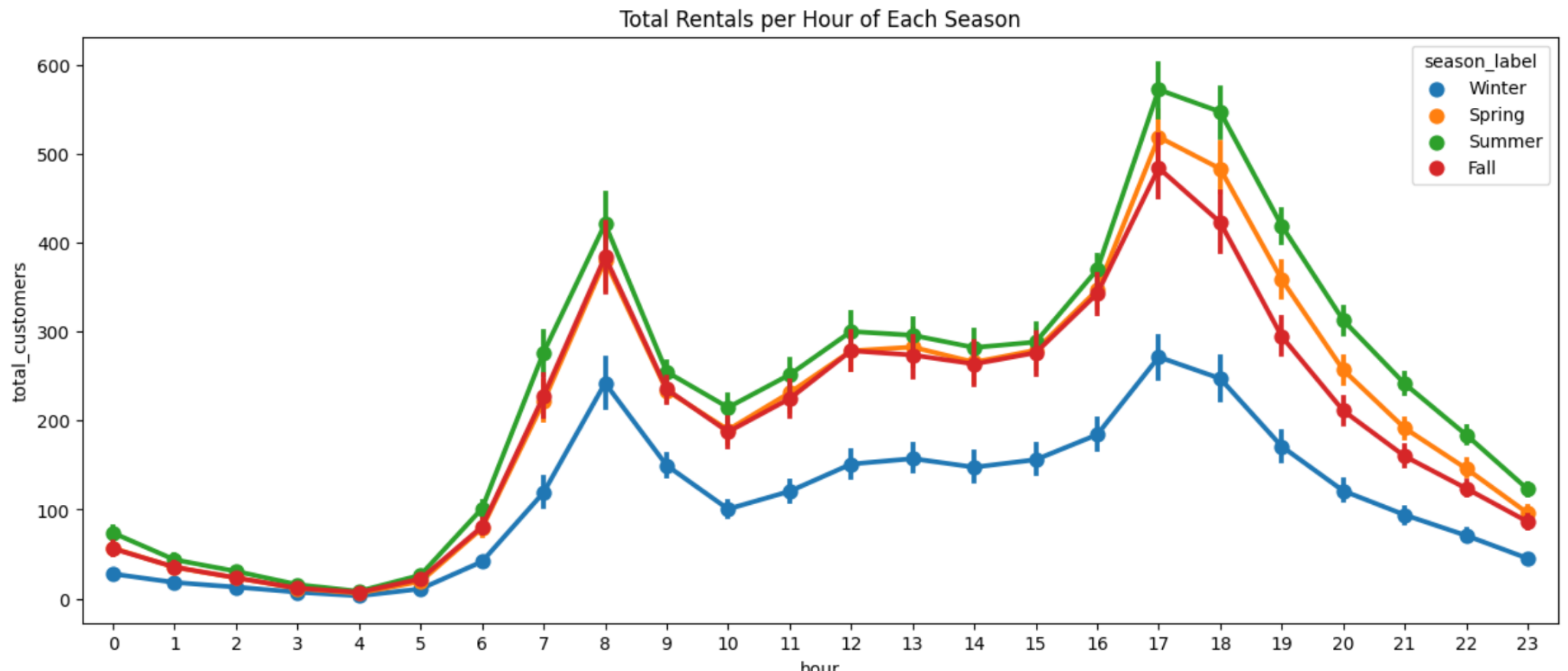
**Root Mean Squared Error (RMSE)**: A way to measure the average square root difference between predicted and actual values. It calculates the average of the squared errors, giving more weight to larger errors. Using RMSE allows us to have our error term in the same units as the original data.

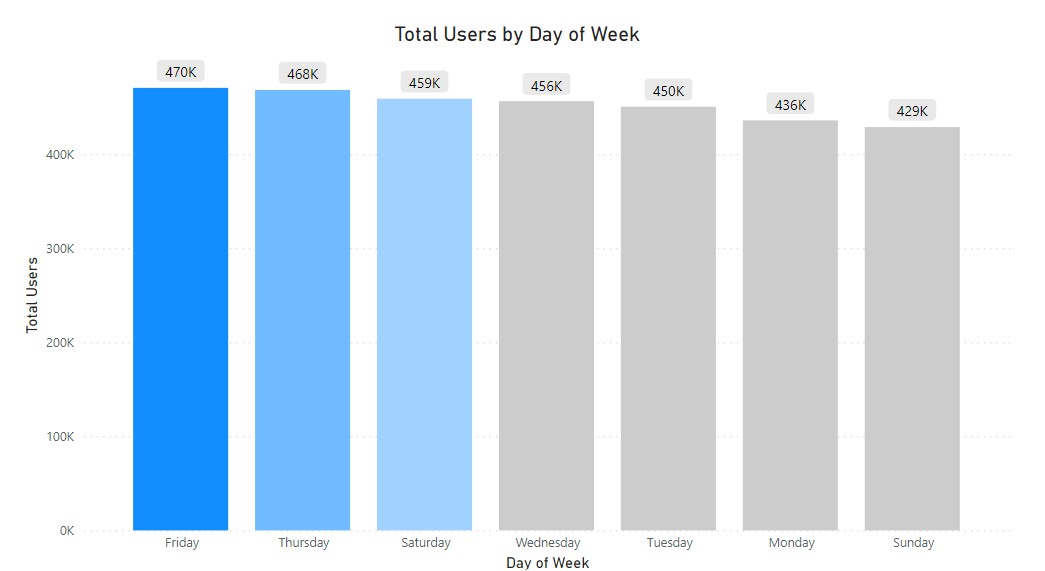
Based on these results, the bike rental prediction neural network model demonstrates promising performance for the company. The model can provide reasonably accurate estimates for the number of bikes needed for a certain day. It can assist the company in providing reliable forecasting estimates to meet demand, aiding in decision-making processes, and facilitating effective planning and development strategies.

1. Action Items

Use the neural network to forecast the number of bikes needed to meet demand on any given day.

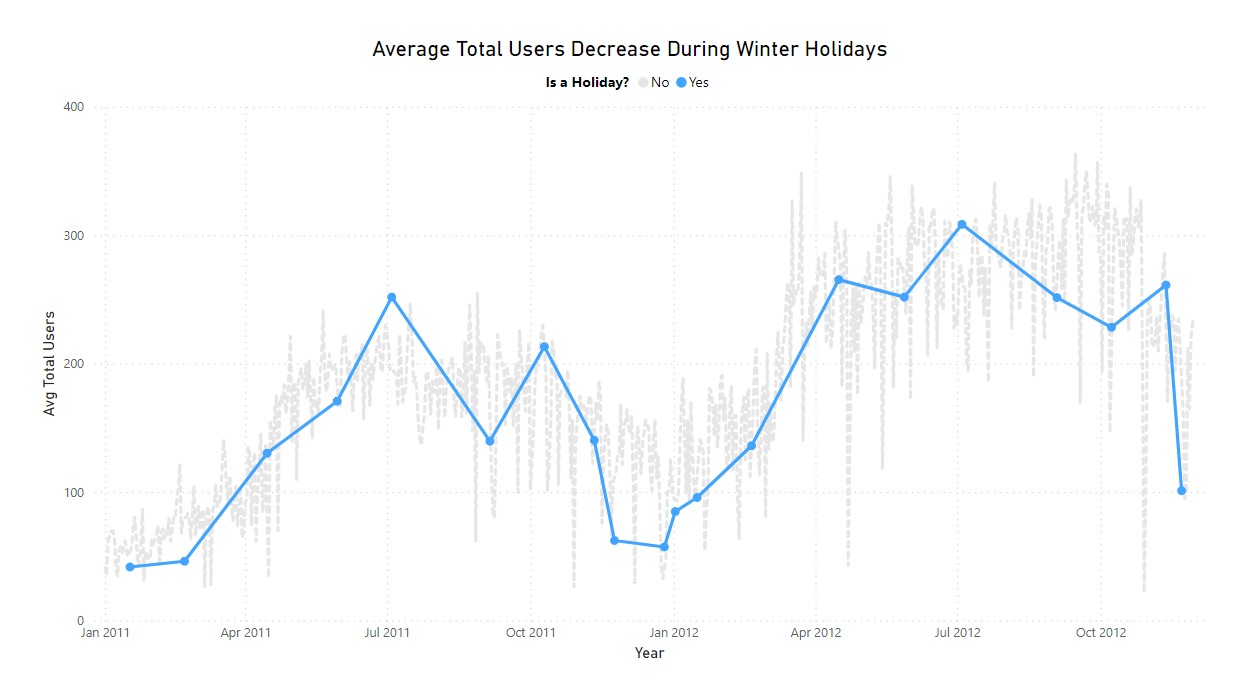
Make sure our bikes are in place and ready during the busy hours of 8 am and 4-6 pm while people are going to and from schools and workplaces. (as shown in Figure 1)

(Figure 1)

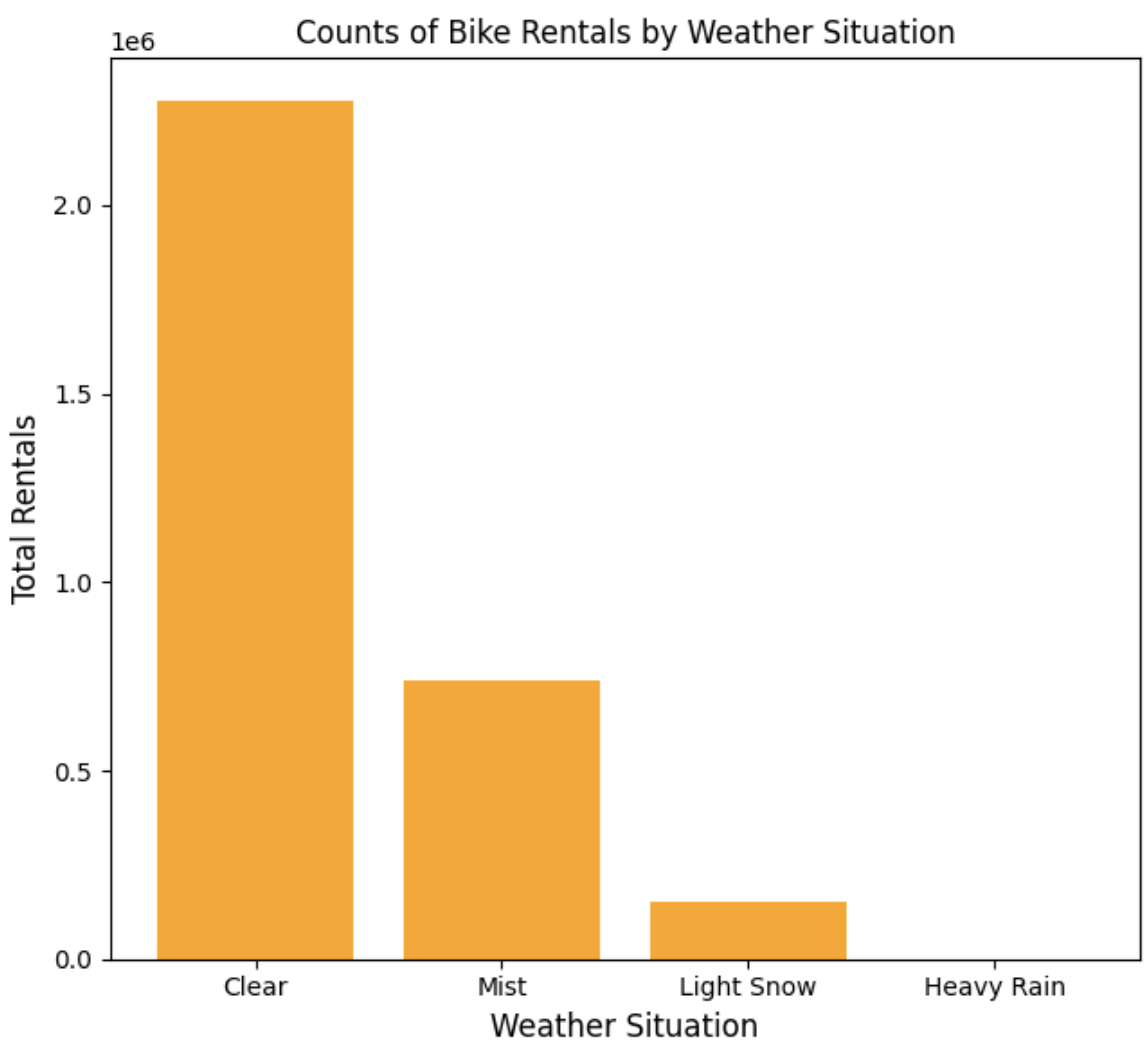
As the weekend approaches, we see an increase in bike rentals. With this information, we can prepare for this by making more bikes available during these weekdays. 

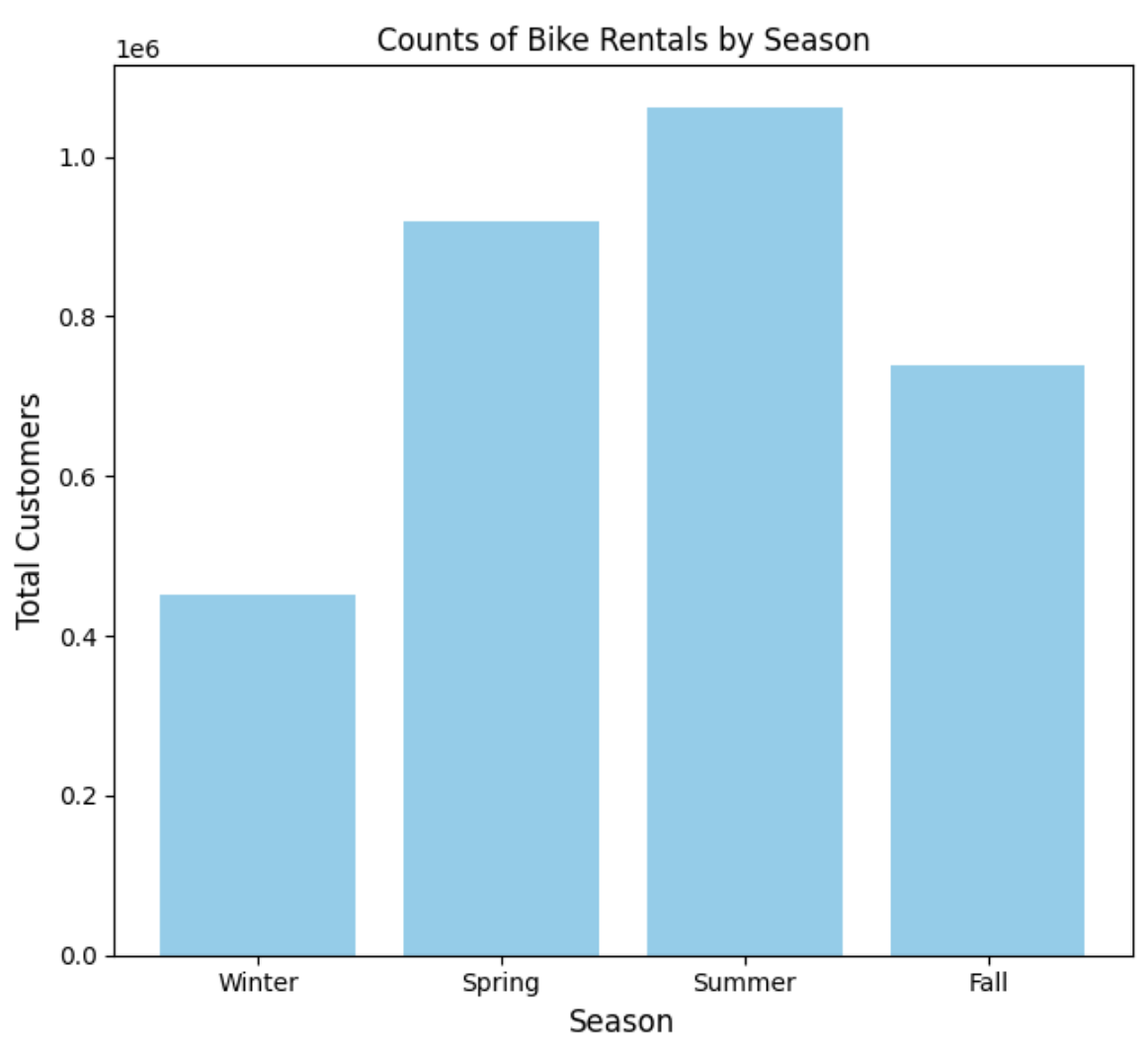
(Figure 2)

Fall and winter holidays have lower usage while spring and summer holidays perform better than normal. Raise prices during spring and summer (See Figure 3 below).

(Figure 3)

Have more bikes available during the summer season and clear weather when people are more active and likely to rent our bikes. (As shown in Figures 4 & 5)

(Figure 4)



(Figure 5)

1. Limitations

Lack of accurate temperature data.

Our holdout set to test the data only contained information on December making some columns obsolete for the model predictions.

The neural network itself can be seen as a limitation because how we arrived at our answer is not directly identifiable as there are hidden layers within the model. These hidden layers create new features using correlations within the data that assist in making the outputs and we do not have access to the logic behind those new features.

1. Q&A

**How many layers do you think the network should have?**

We decided that a simple model was best for our data and chose to have 3 layers: 1 input, 1 hidden, and 1 output.

**Which of the following hyperparameters do you feel has the most potential for model improvement?**

The hyperparameters part of our model that we believe have the most potential for improvement are the number of epochs (number of times the model was run) and the learning rate. We felt that we could’ve done better predicting the data without so many runs and less “jump”. The “jumps” in the graph of the learning rate were too big for us. It showed that the model was not learning as well as we hoped although it was still improving.

**How did you handle the temperature features?**

We created a new feature with the ‘temp\_c’ and ‘feels\_like\_c’ columns of the data since they were similar.

**How will we know if our model has strong predictive power?**

To check the predictive power of our model, we used R2 correlation to give us a decimal number between 0 and 1. If the number is more significant than 0.87, we can assume our data predict outcomes correctly 87% of the time.

**What are you planning to use for the loss function?**

We used the MSE or Mean Squared Error as our loss function for the improvement of the model.

**We would like to add an insurance premium to the rental cost based on the user’s profile information to predict any damage. We are concerned that there may be ethical/legal implications here, what would you recommend?**

We could use the above information if the user is registered because we were given some of their personal information. However, we should limit the personal data we take from the user. For example, we should not accept or request any medical information from the user. We can use basic personal information like name and age. Another feature would be to keep track of the average speed and frequency of use in a week or month. This could be measured using GPS tracking and things like Google Maps which predicts when streets are busy and determine if the user was going too fast in a busy street or time of day.

1. Python Notebooks

Below is Github Gist link to the notebook we used during this case study:

<https://colab.research.google.com/gist/kelintang/4ba84da085c2fa433a4f59ee565fdd77/bike_casual_registered_minmaxscalerhodloutmini.ipynb?authuser=1>