Executive Summary

Bike Rental Predictions Machine Learning Analysis Report

## Presented By:

Becca Roeth

Ethan Houser

Kyle Bell

Kyle Parks

Makayla Whipple

Chantal Borchardt

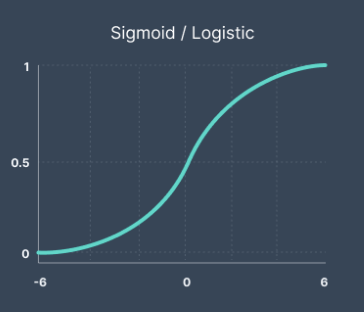
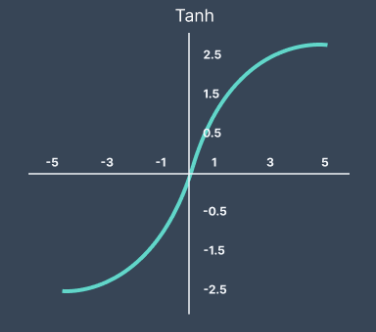
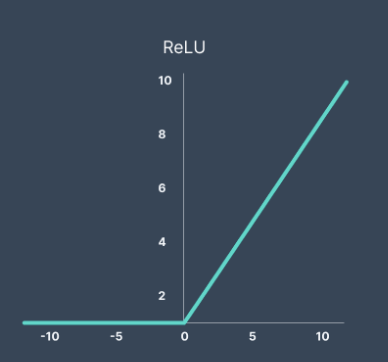
# I. Introduction Summary

# WelcomeBike Bike Rentals has recently been expanding and it’s important that as this company grows that we are able to understand the current state of the current branches to help understand the overall performance. When it comes to our **neural network model** we’ve been able to …

* Create different kinds of models out of different data sets (with or without COVID-19) with different **activation functions**

**Neural networks** are models that strive to identify patterns within data. It is a model that is constantly learning and adjusting based on the data. This allows it to continuously improve performance overall.

* **Activation functions** are derived from equations, they define the shape the model makes when identifying patterns. The model can only be a certain shape when categorizing values. However, by using multiple activation functions, we can better identify values that won’t fit the shape of just one modelWe use ReLU, Tanh, and Sigmoid functions



*Activation Functions (V7labs)*

[*https://www.v7labs.com/blog/neural-networks-activation-functions*](https://www.v7labs.com/blog/neural-networks-activation-functions)

The main concern for creating a neural network is overfitting. We don't want our model to learn the training data too well that it ends up performing poorly on unseen data. This can lead to poor performance when the model starts applying new features. This is why it’s important to make the model focus on generalization.

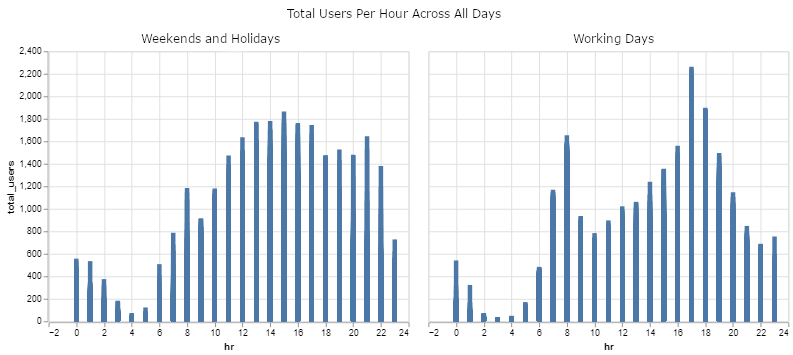
# 

# II. Addressing Questions

When it comes to the temperature features within the dataset, we decided to leave them as celsius. Changing the temperatures would not have any effect on our model and it’s important to stay consistent with the data that we’re given. For graphical purposes we chose to plot it as fahrenheit for comprehension but kept it celsius within the model.

We felt that the most important hyperparameters were batch size and number of epochs. We were able to apply an early stopping mechanism to help control the number of epochs and reduce overfitting. Within the early stopping we applied a patience value of 30 to monitor the validation loss which allowed it to stop the training of the model if no improvement was observed.

*The following figure is able to show the days and times of the week that are best for the company to clean or rotate bikes.*



*Figure 1. Total Users Per Hour Across All Days, Comparing* Weekends and Holidays vs Working Days

Overall, the number of actively rented bikes begins to increase at 8am, peaks at 11am, and begins to fall at 9pm. Both holidays and weekdays have an extremely low number of users between 3 am and 5 am. Individually, weekends and holidays have a sharp decline in usage around 11:00 pm while the weekdays have a sharp usage dropoff earlier around 9 pm. This means that there is only a need for longer operational hours during the holidays and weekends. The company can afford to rotate bikes at earlier times on the weekdays.

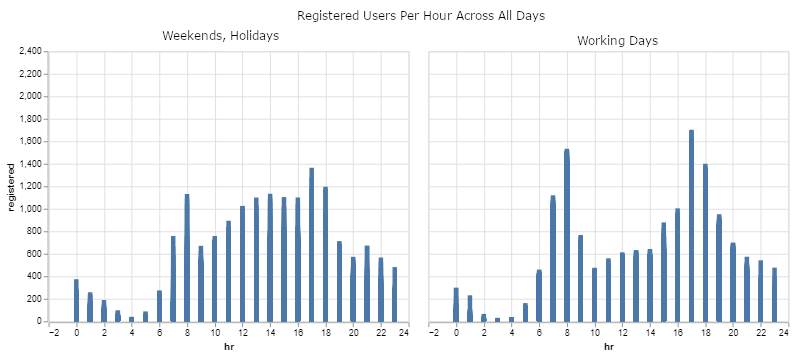
When it comes to using AI to predict the likelihood of damage based on user profile data, such as name, birthday, sex, or address in order to add insurance premiums we chose not to implement these premiums. We don’t want to participate in any unethical or illegal implications regarding altering or affecting the data. The best approach to adding premiums would be to apply premiums to riders that have proven to have a history of previous damages.

For our loss function for the training process we chose to use the mean squared error. This allows for the model to minimize the difference between its predicted values and the actual values. After the model was applied we were able to apply a r-squared that helped show how well our models predictions performed overall. With our high r-squared that we predicted helps indicate strong predictive power.

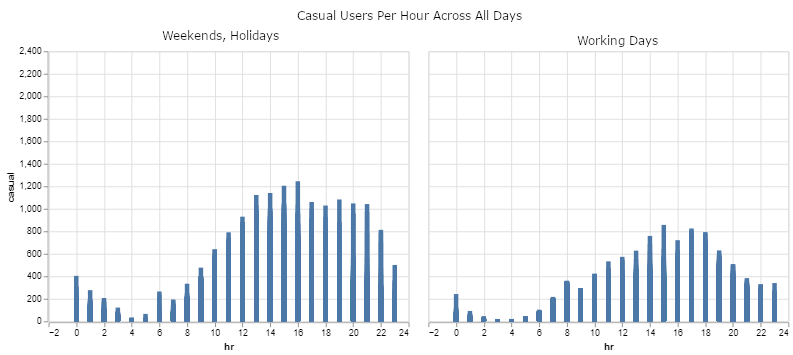
# 

# III. Data Exploration

## Notable Trends



*Figure 2. Registered Users Per Hour Across All Days*

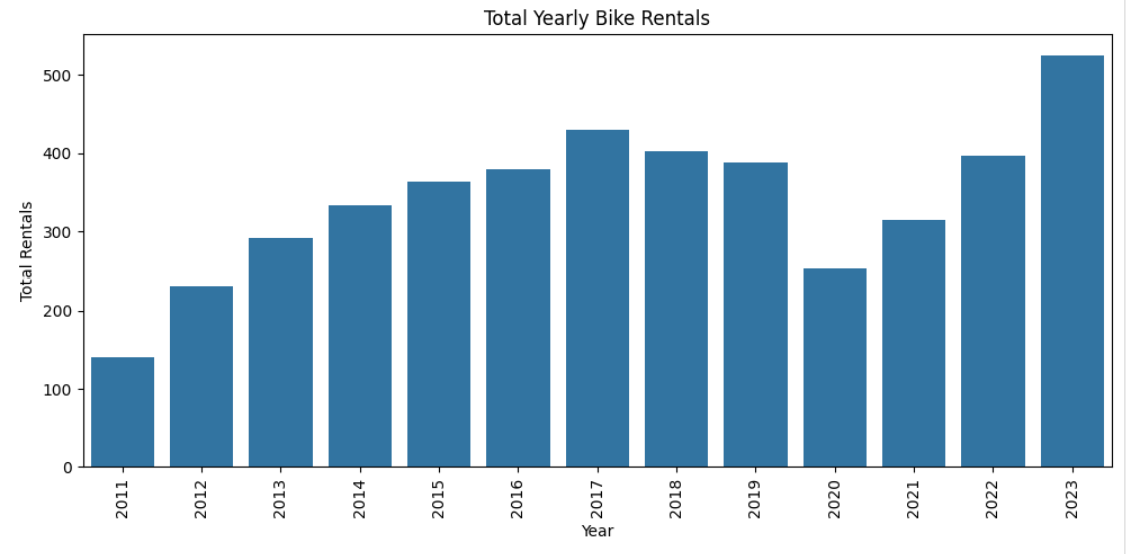


*Figure 3. Casual Users Per Hour Across All Days*

More registered users rent more bikes at 8am and 5pm- rush hour- even on days not considered working days. Casual users do not share this use pattern, but a ‘hill’ of 8am - 11am - 9pm is common to both types of users.

## Issues with the Data

We found that including data from 2020 when the COVID-19 pandemic started skewed the accuracy of the data. People stayed in their homes during Lockdown, so the bike usage statistics plummeted for about a year. The COVID-19 Pandemic was a major event that affected all activity and business across the world.



*Figure 4. Total Yearly Bike Rentals*

WelcomeBike Bike Rentals sales managed to significantly drop in 2020 due to COVID-19 but were able to bounce back within the next 2 years. This is able to show that the company is back on track and with the increase of sales in the years following 2020 it’s evident that sales are continuing to increase. Overall the sales are actually able to surpass their previous highest amount by the year of 2023.

# 

# IV. Machine Learning Model

## Feature Engineering

## While the Machine Learning module can figure out a good deal of information on its own (for example tourist seasons out of months, election years, etc.) even if it may not know what they mean, there are some features we engineered:

* dteday - Datetime, from days
  + This gives a chronological order to the data
* day\_of\_year - the day of the year (out of 365)
* day\_of\_four\_years - the day out of four years (out of 1461)
* day\_of\_week - derived from datetime’s day of the week

## Model Training

Our model training approach involved several strategies to improve the predictive accuracy:

1. Included Temporal Features:
   1. The model was trained on chronological data to understand trends over time. This approach leverages the sequential nature of the data, allowing the model to learn from historical patterns.
2. Feature Utilization:
   1. We utilized the majority of the provided data features, including humidity, weather conditions, and temperature. Additional features such as day, month, year, and day of the week were also included to provide temporal context.
3. Future Improvements:
   1. Due to time constraints and errors on our end, we were unable to include certain features that could further improve our model’s accuracy. Our plan includes implementing the following features:
      1. Is Weekend: A feature indicating if the day is a weekend.
      2. Is Rush Hour: There were trends showing that rush hour during weekdays had a significant increase in bike rentals.
      3. Is Freezing: A feature indicating if the temperature is below freezing.
      4. Days since COVID-19: Our model does not currently understand why the rentals during COVID-19 were low because we did not provide that information to it.
4. Data Expansion:
   1. To further improve the model, we plan to gather additional data on road conditions, such as whether the roads are icy or dry. This data could provide insights into rental patterns during different weather conditions.

# V. Results, Action Items, and Limitations

## Results

| **R^2** | 0.5613 |
| --- | --- |
| **Mean Absolute Error** | 198.69 |
| **Median Absolute Error** | 109.70 |
| **Mean Squared Error (MSE)** | 95104.35 |
| **Root Mean Squared Error (RMSE)** | 308.39 |

**R-Squared** - proportion of the variation in the number of active bikes that is predictable from the date, hour, and other data. It ranges 0-1 where 1 is an impossibly perfect prediction.

**Mean Absolute Error (MAE)** - MAE is the average of the absolute differences between the predicted values (the number of active bikes in a given hour) and the actual values. It considers magnitude, but does not consider whether an error is an overestimation or underestimation.

**Median Absolute Error (MedAE)** - MedAE is the median of the absolute differences between the predicted values and the actual values.

**Mean Squared Error (MSE)** - MSE is the average of the squared differences between the predicted values and the actual values.

**Root Mean Squared Error (RMSE)** - The RMSE value represents the average difference between the predicted values and the actual values.

## Action Items & Conclusions

* Consider including precipitation data in future data models. The data gives good general behavior statistics, but more specific information like the road conditions during intense weather or how much precipitation fell that day would increase context insight on customer behavior.
* Look into bike rental usage by location. Where do the most bikes get rented from? Where do the most bikes get returned to?
* The best time for bike rotation and cleaning is during the late hours of the night. Generally between 9 pm and 5 am.
* Cut bike rental usage around 1 am at the latest on holidays, and around 10:00 pm on working days. Any usage after those times is negligible.

## Limitations

* Neural networks take a long time to train even if they are small. Proper pacing and time management was key for this project’s success.
* Given the time constraints, we were not able to look into a model that excluded the data from lockdown as much as we would have liked to. The team considers the data from March 2020 to March 2021 to be a historical outlier. In the future, the company might want to consider excluding that outlier from the data.

# VI. Python Notebooks

Master:

<https://colab.research.google.com/drive/1ivqwjBB46SZ8p_mTKKQj1T1LuduA4c7x#scrollTo=fcrsPaqDkpYR>

Models:

<https://colab.research.google.com/drive/1e9XPk5LcGiWXryruFBLPYmB9aHkEMCQm#scrollTo=YQCclVUXylwJ>

<https://colab.research.google.com/drive/1Ot9EmI1PncDBFYFSZD21VzMECQJeM-Wm#scrollTo=D_ro7TrHwp8j>

<https://colab.research.google.com/drive/1P5knhNJXWWrEy3FL8sQx7RgarQYYKoxn>