BAIS 3500

Bike Rental Demand Forecasting

December 10, 2021

Lecture A and B Team 12

Reno Chen

Jimin Huh

Yuqing Luo

Yucong Ma

The University of Iowa, Tippie College of Business

Executive Summary

The bike-sharing market has been gradually increasing, especially over the past two years during the pandemic. In order to improve performance and satisfy the growing number of customers, our team analyzes historical bike-sharing data with data mining techniques and assists bike-sharing company, Capital Bike Share, accurately predict the demand for bikes.

Problem Description

Background

As pandemic arises, everything is changing to be non-contact. According to the New York Times, "social-distancing, sustainability, and accessibility helped accelerate e-biking during the pandemic, and the trend is showing up in urban bike-sharing programs" (Glusac, 2021). A bike-sharing system is a process where people can rent a bike and return it automatically. Through this system, a user is able to rent a bike from certain places and return it to another location. The demand for bike-sharing is increasing, and therefore, at the same time, bike rental companies such as Capital BikeShare have problems analyzing and predicting the demand of bike rental for their companies.

Business Goal

As the data analysis team of the company Analytopedia, our business goal is to assist our client, Capital BikeShare, a bike-sharing company, to forecast the demand for bike rental. The data we provide would allow the company to optimize its profit by increasing revenue and utilizing the optimal time when users use bikes the most, ultimately providing enough bikes for users.

Data Mining Goal

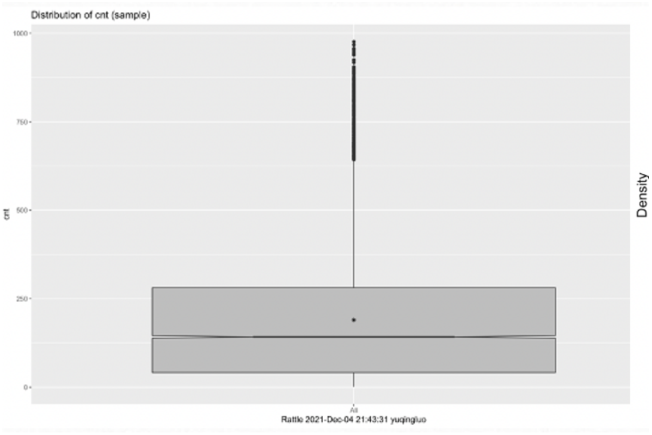
As a supervised learning and regression problem, our data mining goal is to use historical data to see how each feature affects the target variable, which is the number of bike rentals, and to predict the future demand. We would like to figure out which features have a strong impact on the target variable result and focus on these features to provide a systematic plan for companies to follow.

Data Description

Data

The project uses data from Kaggle, "Bike Sharing in Washington D.C." (https://www.kaggle.com/marklvl/bike-sharing-dataset). We would like to use this hourly data **from January 1, 2011, to December 31, 2012**, to determine which conditions can affect users using a bike-sharing system in the Washington, D.C., area. The dataset has 17,379 instances and 14 features, including 8 categorical features and 6 numeric features. Our target variable is 'cnt', which is the number of total rentals, and we used year, month, hour, weekday, holiday, working day, humidity, and temperature to analyze the models.

Exploratory Analysis



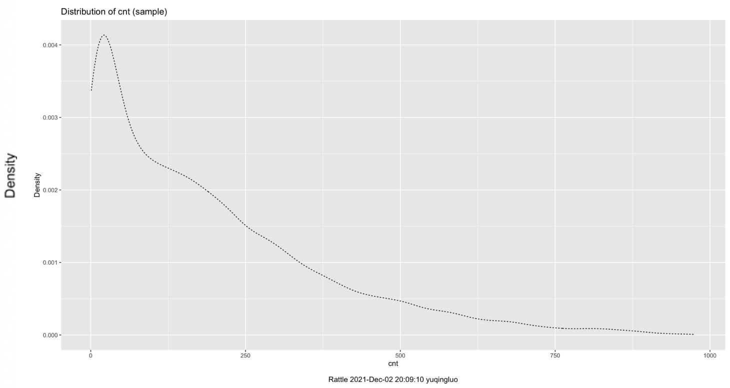


Figure 1. Box plot, Distribution of cnt (left)

Figure 2. Right-skewed distribution of cnt (right)

Figure 1 demonstrates that the target variable has lots of outliers, and this can be supported by *Figure 2* where the distribution of the histogram is skewed to the right side.

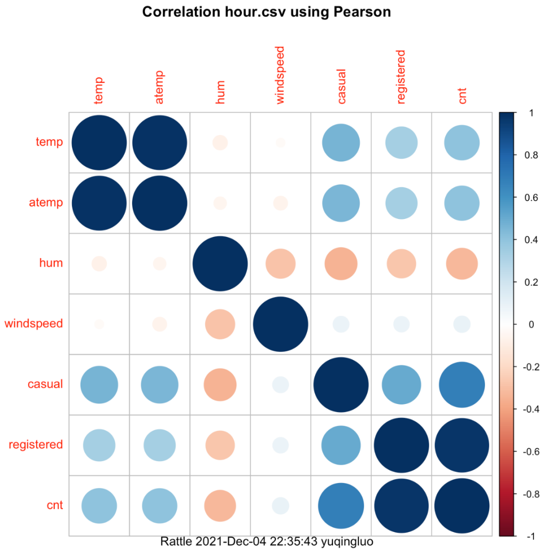


Figure 3. Correlation between each features

Figure 3 demonstrates the correlation plot. The feature “atemp” and “temp” have high similarities and are highly correlated with each other, which caused multicollinearity. The plot also shows that the feature “windspeed” weakly correlates with the target variable ‘cnt’. Further, the features “casual” and “registered” are leakage variables as the target variable is the sum of these two features.

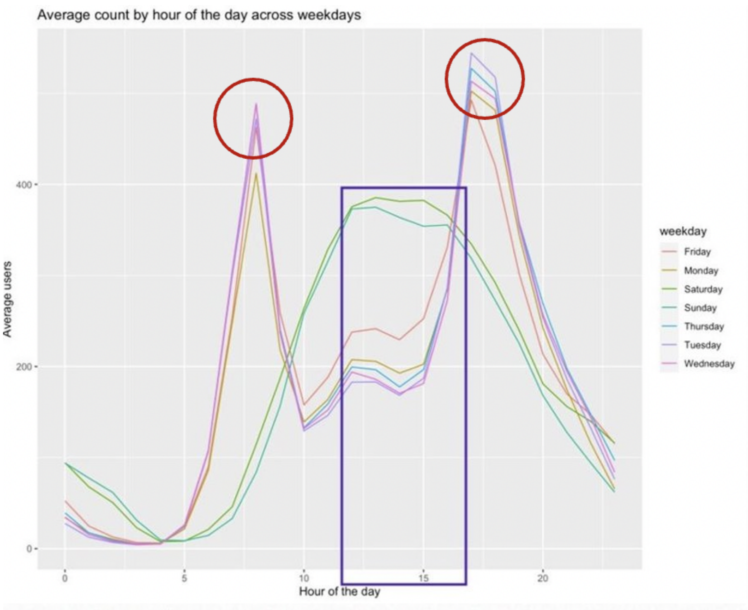


Figure 4. Average count by hour of the day across a week

Figure 4 clearly supports that the demand patterns for bikes during weekdays and weekends are different. There are two peaks on weekdays, around 8 am and 6 pm. However, the peak demand is between 11 am and 4 pm on weekends.

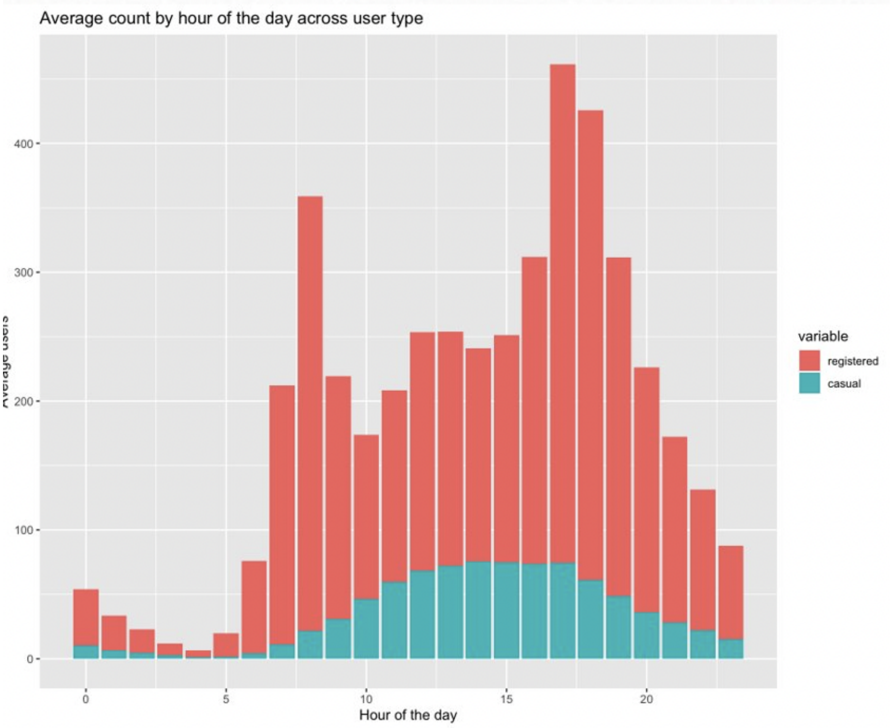


Figure 5. Average count by hour of the day across user types

Figure 5 indicates that most of the bikes are rented by registered customers. Moreover, the demand patterns for registered and casual users are also different. Registered users tend to rent bicycles around 7 am to 8 am and 4 pm to 6 pm, while casual users rent bikes around the afternoon.

Data Preprocessing

We used Rattle in the RStudio to preprocess the dataset. First, we remove unnecessary features, such as the date of time. The date of time is redundant because we already have year, month, and hour in the dataset. After that, we removed the leakage variables "casual" and "registered." Also, we dropped "atemp" because it strongly correlates with another independent variable, "temp." We want to avoid multicollinearity, which can affect the performance of our model. We also dropped the feature "windspeed," as the correlation plot above shows that its correlation with the target variable is very weak. Finally, we rescale the numeric features temperature and humidity into 0-1 and put all categorical features into indicator variables.

Data Mining Solution

Models and Performance Evaluation

Table 1. Models, Average MAE, and Standard Deviation

Model	Average MAE	Standard Deviation
Linear Regression	64.3167764	1.77796134
Decision Tree	41.6239292	0.57640458
Random Forest	29.2752004	0.50364927
Neural Net	**30.1724179**	**0.24649797**

We tried different models to find the optimal one: linear regression, decision tree, random forest, and neural net. We chose to use MAE to evaluate the models because it is more robust to data with outliers. Furthermore, compared to r-squared, MAE can convey more information to understand and interpret the result. Here in the chart, we noticed that both random forest and neural net performance are good. The average MAE score of the neural net was slightly higher than that of the random forest, but the standard deviation of the neural net is much lower than that of the random forest. A practical model should not only have a good performance on forecasting but also a uniform prediction on the whole dataset. Therefore, we determined the neural net model to be our best model, but the random forest can be an alternative model to choose.

Conclusion

Recommendations

In summary, we recommend our client, Capital BikeShare, to:

1. Serve more bikes around the office building areas for specific hours. According to *Figure 4*, bike rental usage increases from 7 am to 8 am and from 4 pm to 6 pm during weekdays, and those are when people

usually go to work or get off the job during this period. It also increases from 11 am to 4 pm during weekends, and therefore, to satisfy the higher demand for bicycle rental, Capital BikeShare should pay more attention to these periods.

2. Update the data frequently and periodically. The demand patterns will change over time due to unpredictable factors, such as COVID-19 and the increasing price of gasoline. We suggest that the company regularly update the data to avoid bias when using our model to predict.
3. Create a loyalty program. Capital BikeShare only has certain memberships such as daily pass, monthly pass, yearly pass, and student membership. We recommend the company have a reward program to create a stronger relationship with customers (e.g., one free pass after ten rides). As shown in *Figure 5*, registered customers tend to have a higher demand for bikes; a loyalty program would help the company to maintain a sustainable relationship with current customers and attract potential customers.
4. Create a campaign that aims at office workers. Based on *Figure 5*, we found that the demand patterns of registered and casual users are different. Registered users tend to rent bicycles around 7-8 am and 4-6 pm during workdays, which indicates that our current customer base might be regular office workers. We suggest the company focus on this market. Creating a campaign for our target market can more efficiently attract new customers to sign up for our service.

Limitations and Next Steps

The following are our limitations and the ways to improve our the our model performance in the future:

1. The data is outdated. As we have mentioned in the data description, we collected data from 2011 to 2012. Many conditions have changed in these years, so updated data is needed to analyze and forecast models more accurately. For instance, if we have recent two years' data, we can explore and predict models under the conditions of the Covid-19 pandemic.
2. We have a limited number of features to select to build our model. Some features, such as air quality, traffic condition, and location, can be valuable features to improve our model performance. In the future, we plan to collect and add more related information to our dataset and build more precise models.
3. The area is limited, which causes the data not to be generalizable. The data provided by Kaggle is represented only for the bike-sharing system in Washington D.C. Thus, therefore, it is not applicable to predict the demand in other cities because the lifestyle and culture could not be the same for various areas. If the company plans to expand its market in other cities, we need to do an initial survey in those cities to collect new data for our model.
4. We didn't utilize the information of casual and registered users when building the model as they are leakage variables. However, we think training two models for casual and registered users might improve performance.

References

Glusac, E. (2021, March 2). *Farther, Faster and No Sweat: Bike-Sharing and the E-Bike Boom*. NewYork Times. retrieved December 2, 2021, from https://www.nytimes.com/2021/03/02/travel/ebikes-bike-sharing-us.html