Project 1 Capital Bike Share Data Set

11/14/2024

Brian Zhang

**Introduction:**

Business Context and Problem:

Capital Bike Share operates a bike sharing system in Washington DC, providing a convenient and eco-friendly transportation solution for locals and tourists. It’s reported that Capital Bike Share has experienced significant growth in recent years. With an increased record setting annual ridership of around 31.3% through May 2024 and a total of 2,000,128 trips this calendar year (Greater Greater Washington, 2023). With the increase in demand, understanding the factors that drive bike rental demand is key for the company’s continued growth. Accurate demand predictions can help with bike allocation, availability, and kiosk placements, improving both the user experience and profitability.

Objective:

The objective of this analysis is to develop and use a predictive model for bike rentals and to provide actionable insights to its key drivers. By interpreting the relationships between rental demand and other independent variables—such as weather conditions, time of day, season variations etc.—the analysis aims to aid Capital Bike Share’s growth with data driven analysis to enhance their company.

Report Overview:

The dataset used for this analysis contains hourly records of bike rentals between the years 2011 and 2012. It includes both numerical variables (temp, hum, etc.) and categorical variables (season, weathersit, etc.). The dependent variable “cnt” represents the total number of bike rentals. This report presents a comprehensive analysis, starting with an exploratory data analysis, followed by a development of a multiple regression model to come to actionable recommendations. Both key statistical and visual insights are provided to ensure clarity and relevance for Capital Bike Share’s operational decision-making.

**Exploratory Data Analysis:**

To better understand factors that influence bike rental demand, an exploratory analysis was conducted to summarize the dataset, identify key trends, and uncover potential relationships between the dependent variable (cnt) and the independent variables.

Summary Statistics:

A summary (Figure 1) of the dataset reveals that bike rentals (cnt) has a range of 1 to 977, with the average being 189 rentals per hour. In addition, the summary table also reveals that other dependent variables “causal” and “registered” vastly differ, with registered having a range of 0 to 886 and a mean of 153.8 versus causal only having a range of 0 to 367 and a mean of 35.68 (Figure 2). Immediately from this table it seems that registered users out ride causal riders by almost 5 times.

Correlation Analysis:

A correlation matrix was created to highlight any significant relationship between “cnt” and other numerical predictors. Immediately it’s noticeable that the predictors registered, and casual are extremely high due to both registered and casual summing up to “cnt”. In addition, instant can be disregarded as it’s just a recording index. Disregarding these three, the next highest correlation would be “temp” and “atemp” having a correlation of 0.4 (Figure 3) indicating higher rentals during warmer weathers. “hr” being very close behind with a correlation of 0.39, indicating that hourly variations do influence bike rentals. The third strongest correlation would be “hum” with its value being -0.32, indicating that as humidity increases, the number of riders will most likely decrease.

Multicollinearity:

Multicollinearity among predictors was assessed using the VIF method. Multicollinearity indicates if you’re overfeeding a model and if there are redundant variables. The results indicated that there was a high multicollinearity between temp and atemp, yr and month, and instant. Moving on, the variables atemp, year, and instant will be excluded from the final regression model as they are redundant.

Key Visualizations

* Figure 4 shows a scatterplot that indicates a positive non-linear relationship between temperature and “cnt”, with diminishing returns as temperature increases.
* Figure 5 shows the bike rentals per hour versus each season. It’s clear that there are season preferences when it comes to biking. With Winter (1) being the clearly the lowest season for bike rentals
* Figure 6 shows a bell-shaped curve, indicating that people prefer moderate humidity levels while avoiding both high and low extremes.
* Figure 7 shows two distinct peaks, one during the morning hours (around 7 – 9 am) and the other in the evenings (around 4 to 7 pm). This pattern is particularly noticeable for registered riders.

**Regression Model**:

The goal of the regression model is to analyze bike demand (cnt) using predictors while addressing issues like multicollinearity and heteroskedasticity. The final model was selected through manual feature selection and automated regression.

Model Building:

The initial model included all predictors, both numerical and categorical. However, as mentioned previously, after running a multicollinearity analysis the variables atemp, year, instant, and month were removed. In addition, when creating a predictive model, it was key to understand that some variables weren’t numerical/continuous rather they are categorical. Therefore season, weathersit, hr, weekday, holiday, and workingday were converted into categorical variables for the regression model.

The next step was to split the data into three subsets: training, validation and test sets to utilize machine learning to build the predictive model. In total the data set was split into 60% training, 20% validation, and 20% testing. After splitting the dataset, the numeric predictors were standardized to ensure comparability between variables and coefficients. Based on the exploratory analysis, there was a need to square terms for temp, hum, and windspeed were added to capture both diminishing or accelerating effects. The first regression model was created with the equation being:

*cnt = β0  + β1\*(season) + β2\*(mnth) + β3\*(temp) + β4\*(hr) + β5\* (holiday) + β6\*(weekday) + β7\*(workingday) + β8\*(weathersit) + β9\*(windspeed) + β10\*(hum) + β11\*(temp2) + β12\*(hum2) + β13\*(windpseed2) + ε*

Model Violations:

When creating the model it is important to look for regression assumption violations. These include multicollinearity, heteroskedasticity, non-linearity, normality of residuals, influential observations, and autocorrelation. Multicollinearity was already addressed (the variables were removed) and running it again with the categorical variables still indicate that there isn’t any strong multicollinearity detected. Next, to determine heteroskedasticity a Breusch-Pagan test was conducted which indicated that there was strong heteroskedasticity. In addition, when looking at the residuals when testing for non-linearity, there is a clear cone shape (Figure 8) between the fitted values and the residuals, also indicating high heteroskedasticity. The reason why having strong heteroskedasticity is an indication of a poor model is because it means the variability of the residuals (errors) in the model are not constant across the range of predicted values (explanatory variables). which violates the assumption of homoscedasticity. The next test to look at is the linearity of the model. The Q-Q plot (Figure 9) shows that the model’s residuals do not follow a normal distribution along the red line (heavy tail and skewness), therefore suggesting non-normality. This would indicate inaccurate p-values and confidence intervals for coefficients and create an issue with prediction intervals. When looking at if there are any influential points that might bias the regression line, the residual versus leverage graph (Figure 10) indicates that there aren’t any points in cook’s distance, only that there are a couple of outliers in the dataset. Therefore, there isn’t a violation. Finally, when looking for autocorrelation, it can be tested using the Durbin Watson test. It’s to make sure there aren’t any errors that correlation to each other or else the coefficients will be inconsistent which would result in inaccurate predictions. Because the D-W Statistic on the Durbin Waston table is less than 2 (Figure 11) as well as a p-value of 0, the null hypothesis can be rejected, and it can be determined that there is no autocorrelation.

In summary there were three violations with the first regression model. To correct for heteroskedasticity, non-linearity, and non-normality of residuals, a combination of robust standard errors and variable transformations can be used. This includes removing mnth, workingday, and weathersit due to its low p-value from calculating the robust standard errors. Furthermore, the dependent variable cnt can be log-transformed to stabilize variance, address heteroskedasticity, and improve the normality of residuals, making the model assumptions more robust. The edited regression model ends up being:

*log(cnt + 1) = β0  + β1\*(season) + β2\*(temp) + β3\*(hr) + β4\*(holiday) + β5\*(weekday) + β6\*(windspeed) + β7\*(hum) + β8\*(temp2) + β9\*(hum2) + β10\*(windpseed2) + ε*

Model Refinement & Outputs:

With all model violations corrected, there can be the implementation of an automatic model selection procedure using stepwise selection. Stepwise selection was employed in two forms: backwards selection—which removes predictors sequentially based on their insignificance—and stepwise selection—which both ads and removes predictors to identify the strongest model. As a result, both backwards and stepwise selection removed the weekday variable.

When looking at the output of the two models the R^2 is identical, meaning both models explain about 34.5% of the variance of cnt, which could suggest that there might be additional factors not capture by the model (such as demographics). One key difference is the F-statistic, with f\_step having a higher value, meaning a better overall fit relative to the simpler model. Next when examining the RSME, both models have an small RSME of around 0.904 meaning that both models is performing well in terms of prediction accuracy, with the average prediction error of cnt being 0.904. Finally, when examining the p-values for both models both models have very small p-values meaning the values are statistically significant except for weekday, in the model f1. Meaning that when evaluating models f\_step is a better model overall due to better p-values and f-statistic.

**Interpretation of results**:

The regression model f\_step provides important insights into factors that would influence bike rental demand. These factors are season, temperature, hour of the day, holiday, windspeed and humidity.

There are two categorical variables that stand out from the rest: season and hr. The season variable is significant in predicting bike rentals as its coefficient is 0.132, and a low standard error of 0.139 as well as a low p value. This means bike rentals increase approximately 13.2% during warmer seasons. hr has an even high coefficient of 0.612, meaning that for every increase in hr there’s around 0.612 increase in the number of bikes rentals. Its p-value is also extremely low of about p < 2e – 16.

When examining key numerical variables such as temp, hum, and windspeed, it's important to note that windspeed has a p-value above 0.05, indicating that it is not statistically significant at the conventional level. This means that changes in windspeed do not have a strong enough effect on bike rentals to be confidently distinguished from random noise, unlike temp and hum, which have strong and significant relationships with the outcome variable.

The temperature (temp) predictor has not only a high coefficient (0.399) and low standard error and p-value but most importantly a strong negative marginal effect on bike rentals (Figure 12). It’s important to realize that the presence of temp^2 with an estimate of -0.107 shows that there are diminishing returns at high temperatures. Meaning that while warmer weather encourages more log bike rentals, it reaches to a point where it reduces the rate of an increase. The estimated standard value is around 0.008546102. Before that critical point cnt increases by an average of 0.5854, after that point it increases by only 0.2167.

Humidity (hum) is similar with its coefficient being -0.34580, with a similar low p-value and standard error. It has a negative sloping marginal effect curve (Figure 13), which would indicate that as humidity increases log of cnt decreases. The critical value is approximately **0.**0015. Before this point, the marginal effect is **-**0.2211, and after the critical value, it becomes more negative at -0.4692. Which would indicate after the critical point the negative impact on bike rentals intensifies.

**Limitations and Assumptions**:

Despite the insight from the model there are still several limitations and assumptions to address:

1. Standardization of the variables: The model uses standardized version of numerical variables. While it is useful in interpretation and model training, it can make the results less intuitive for stakeholders unless transformed back.
2. Omitted Variable Bias: As referenced before, because the model only explains about 34.5% of the variance in bike rentals it would suggest that there are other factors such as demographics, road conditions, or any external events that might have not been captured by the dataset.
3. Seasonal and Weather-Related Effects: The model relies heavily on weather related variables like temperature, humidity and windspeed. Although they are predictable to an extent, it is still very difficult to make real-time predictions and planning for sudden weather shifts, which might impact bike rental demand
4. Assumptions of Linearity: Although there are non-linear relationships addressed through polynomial terms, other interactions or non-linearities might still be present.
5. Normality of Residuals: The Q-Q plot showed small deviations from normality, particularly in the tails. Which would suggest that the residuals aren’t perfectly normally distributed, which would affect the confidence intervals and hypothesis testing.

Recommendations and Conclusions

The analysis identified that temperature, season, humidity, and the time of day were significant predictors of bike usage. These factors indicate clear patterns on how both environmental and temporal factors drive user behavior. To better capitalize on consumer, demand the following recommendations should be implemented:

1. Optimize bike availability by time of day. Both graphs and regression modeling support the idea that time is a key influencer of bike demand. The company should ensure there is an adequate supply of bikes during peak commute hours (morning and evenings) to match demand patterns.
2. Have seasonal adjustments in the Bike Fleet Management. Increase the number of available bikes during high demand seasons like Spring and Summer while better allocating the resources used during Winter.
3. Employ weather-responsive tactics. When there’s high humidity or if the temperature goes over the critical value, provide additional discounts or incentives to increase rentals demand.
4. Build or deploy proper infrastructure for bikes when being used during each season. For example, in the winter there are fewer bikes outside to prevent decay, while in the summer provide bike shelters from the sun.
5. Collect more data to better predict weather. By using predictive modeling, it can be possible to better forecast demand based on the weather (temperature, humidity, rain or no rain, etc.), allowing for better resource allocation and experience.

In summary, temperature, humidity, season, and time of day are key predictors in bike usage. Capital Bike Share should focus on trends in the environment and how temporal factory may influence demand. By better understanding these patterns, Capital Bike Share can enhance user satisfaction and drive demand.

Appendix:

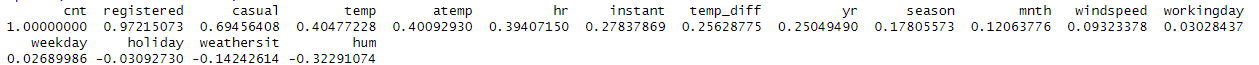
A close-up of a white background

Description automatically generatedFigure 1:

Figure 2:

A screenshot of a computer screen

Description automatically generated

Figure 3:

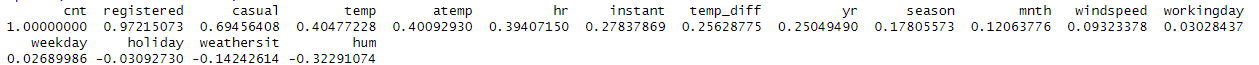
Figure 3 (continued):

Figure 4:

A graph of a graph

Description automatically generated

Figure 5:

A graph of a graph

Description automatically generated

Figure 6:

A graph with a line

Description automatically generated

Figure 7:

A graph of a number of people

Description automatically generated with medium confidence

Figure 8:

A graph with a line and dots

Description automatically generated with medium confidence

Figure 9:

A graph of a normal q-q plot

Description automatically generated

Figure 10:

A graph of a number of black dots

Description automatically generated with medium confidence

Figure 11:

A black and white text

Description automatically generated

Figure 12:

A graph with a line

Description automatically generated

Figure 13:

A graph with a line

Description automatically generated

References:

Greater Greater Washington. (2023, June 16). *Capital Bikeshare sets record for monthly ridership in May*. Retrieved from <https://ggwash.org/view/93867/bikeshare-beat-capital-bikeshare-sets-record-for-monthly-ridership-in-may>