MGT 6203 Group Progress Report

Optimizing Bike Sharing Sustainability: Demand Forecasting and Inventory Management

Team 83:

Amynah Reimoo, Sidarth Sudhakar, Raashid Salih, Xinxing Ren, Andrew Ramirez

Github Repo: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-83

BACKGROUND INFO

Bike Sharing operates in a very dynamic environment, and without a mechanism in place to better gauge demand, it may lead to losses to the company incurred from overstocking, understocking, rebalancing of resources and suboptimal pricing. This has further consequences downstream in terms of customer satisfaction, customer retention, and the environment which greatly impacts the sustainability of such a business. We are trying to identify the elements that play a significant role in affecting demand and how we can predict the demand as accurately as possible based on this information.

Literature Review

The demand for bikes not only fluctuates by location, but also the time of day, the month, whether or not it's a weekday, or a holiday (Ramesh et al. 2021). Other factors that may affect demand include the temperature, land-use and points of interest in the city as well as the population density (Xu et al, 2018). There are other important things to factor in when trying to provide the demand forecast, which is the actual real world implementation of the redistribution, and whether it actually improves user satisfaction (Gammelli et al. 2022). Ultimately, there are two things that need to be true: firstly, a client would need to have a bike available when they arrive at a specific pick up location and secondly, there should be an empty slot available when they arrive at a station to return their bikes.

Despite the presence of a large body of literature on this topic, very few studies have taken a feature engineering approach to forecast demand (Yang et.al, 2020). This approach can not only detect patterns in data that would otherwise go unnoticed but is also likely to provide more favorable statistical measures, such as R-squared and adjusted R-squared.

Another approach to predict demand involves the use of graph structures (Yang et al. 2020). A cluster of locations as a node, and the volume of trips taken between the nodes generates the edges. Graph features, such as Out-strength, In-strength, Out-degree, In-degree and Page-Rank were then used as inputs to other models to predict short-time demand (Yang et-al), achieving better accuracy. While it is ideal to predict demand for individual stations, using clusters has also proved to be valuable as it provides useful insights into the relationship between stations.

CURRENT PROGRESS

Data Collection

The datasets spanning the years 2015-2019 were obtained from Capital BikeShare's official repository, and were consolidated together to obtain a single usable file. This resulted in a file which is 2.15 GBs in size. **After that, the file was augmented with data from external sources**, namely: open-meteo for weather data via REST API and holidays package for public holidays/event data via Python. This increased the file size to 3 GBs, and is the final dataset that all members of the team are working with. However, working with the dataset directly was cumbersome owing to its large size which slowed down the pace at which members could develop their respective solutions. **A stratified sample (grouped by date) which consists of 5% of the final dataset was thus made.** At 157 MBs, this greatly improved the capacity of the team to rapidly experiment and prototype the code they were coming up with.

Data Clean Up

Our data cleaning process started with visual inspection of the dataset, during which we shortlisted columns and fields that needed to be cleaned. **The first step was to ensure that spelling, punctuation and data types were consistent throughout the dataset.** For example, the column 'Member type' was checked for inconsistent capitalization so we changed all entries to lowercase. We also checked for correct data types across all columns to ensure seamless analysis for later. **Next, we examined the dataset for missing entries** and it was observed that 'Holiday' had 97.33% missing values. However, upon further inspection, it was concluded that the missing values indicated that the day was not a holiday. Therefore, the NaN values were replaced with the string 'Not a holiday' to accurately reflect this information. **The data was also checked and cleaned for miscellaneous discrepancies.** For example, there might be rows with a negative trip duration. As this is impossible, these rows were also removed from the dataset. The **next stage of our data cleaning involved outlier removal**, using the Tukey's fences technique. By equating k=1.5 and using the equation [Q1-k(Q3-Q1), Q3+k(Q3-Q1)], we were able to remove 1,281,322 rows from the dataset by using the duration column. *This is understandably a crude method of identifying and dealing with outliers, but it is the best one on hand to quickly generate results that we could build upon.* A more thorough examination of outliers could be done for the next run of the project.

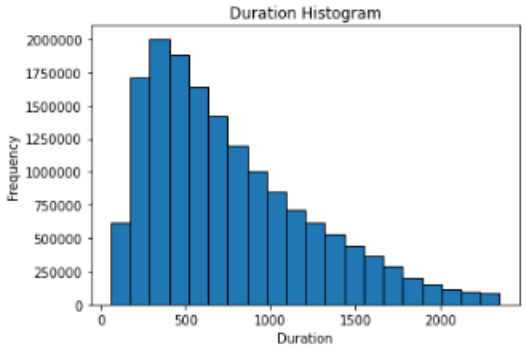
Exploratory Data Analysis (EDA)

At this stage, three columns were shortlisted for EDA: duration, start station and end station as they can provide useful insights into popular routes and potentially aid our demand forecasting. We'll be focusing on external data sources like weather and holidays for the next iteration.

Duration:

As it can be seen in Figure 1 below, the mean duration of all rides is 759.6 seconds (12.7 minutes). The maximum and minimum ride durations are 2352 seconds (39.2 minutes) and 60 seconds (1 minute) respectively. Therefore, we can conclude that the majority of customers are using these bikes for shorter trips, which is further confirmed by the right-skewed histogram derived from the data.

```
count    1.593746e+07
mean     7.595961e+02
std      4.788025e+02
min      6.000000e+01
25%      3.840000e+02
50%      6.410000e+02
75%      1.037000e+03
max      2.352000e+03
Name: Duration, dtype: float64
```



(left) Figure 1: count, mean, standard deviation, and five-number summary of duration of rides

(right) Figure 2: Histogram displaying the duration of bike rides

Stations:

The most popular station in both 'start station' and 'end station' is Columbus Circle/Union Station. This is logical as the station is the transportation hub of the city. On the other hand, Fort Stanton Rec Center is the least popular start station and end station.

Members:

If we compare between casual and registered members, we can see that registered members account for 13,370,527 datapoints (which is roughly 83.89%) while casual riders account for 2,566,929 (which is roughly 16.11%).

Relationship between ride duration and membership:

A combined boxplot was derived from our data to see if there were any differences between the ride durations of casual riders versus members. We can observe that typically, casual riders use the bikes for longer duration compared to registered members. This may be because registered members have much to gain with routine, short duration trips which might be why they registered in the first place. We can also see that casual riders

have a larger range when it comes to duration, which indicates higher variation. This also highlights what was pointed out earlier that registered members are more predictable in terms of their ride duration. Furthermore, the graph shows that casual riders have more outliers, further reinforcing this notion. Finally, outliers for casual riders tend to be larger, indicating "unusually" long trips.

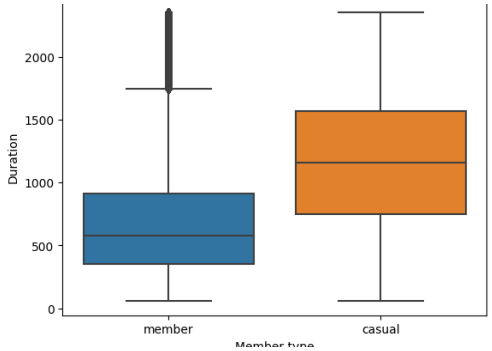


Figure 9: Box-plot of ride durations by member type

Demand Forecasting Modeling

We began univariate time series forecasting by employing auto-regression models to forecast the number of rides per hour. This of course assumes the hypothesis that the data has an autoregressive component (that the current number of rides depends on the past number of rides), which we think is not the case, rationally speaking. However, the steps taken along the way will determine for certain if that is the case, and will inform how we tackle modeling after the fact.

Before building the model, several preprocessing techniques and statistical tests were made to analyze the characteristics of the underlying time series. **The dataset has timestamp information which includes both a date, and a time. This information was split into its more granular components**, which will be extremely useful for our analysis going forward.

Although the natural frequency of the data is in the seconds level, the lowest granularity in which the models are being built is in hours. **To support future work, the data is sampled at other high level intervals** like day, week, and month with variables of interest being summed up.

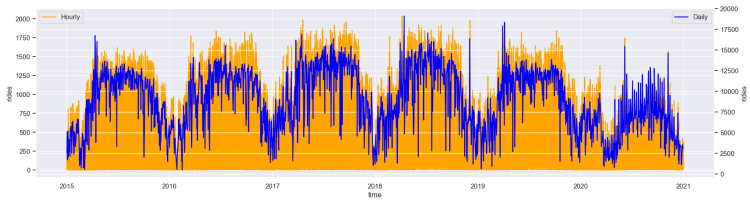


Figure 3: Number of rides per hour, and per day, between 2015-2020

By analyzing the initial plots, it is evident that the time series exhibits seasonal behavior. **Decomposed time series clearly shows a constant process (trend) with a heavy stationary component.** This indicates the use of a Seasonal AR Model.

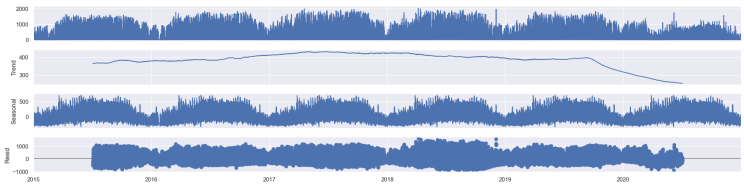


Figure 4: Decomposition of the time series data

1. **Test for Stationarity - With a high ADF statistic and low p value, it is very evident that the time series in its natural form is itself stationary.** This complements the findings from Decomposition as well thus eliminating the need for differencing the series.

```
ADF Statistic: -12.570608
p-value: 0.000000
Critical Values:
        1%: -3.430
        5%: -2.862
        10%: -2.567
```

Figure 5: ADF test results

2. **Test for Auto-Correlation** - It is conclusive from PACF that the influence dampens after 3-4 lags which indicate the AR Component. The ACF on the other hand again exhibits seasonal behavior at various orders (24, 168…), but the effect diminishes around 7th lag.

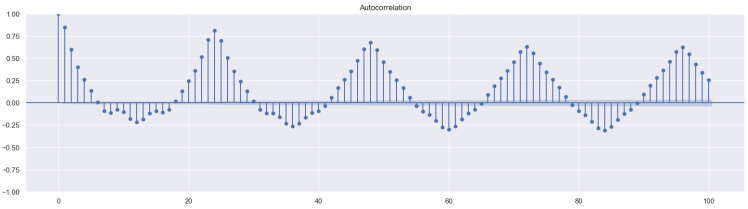


Figure 6: Autocorrelation function (ACF)

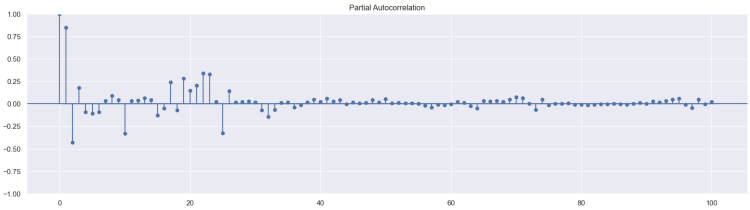


Figure 7: Partial autocorrelation function (PACF)

SARIMA Model - Taking all the previous findings into account, a Seasonal ARIMA model is built using several lag parameters with no integral component to test the best performing model. The best model resulted from an AR lag of 1 and MA component of 2 with 24 (1 day) seasonal component. This model still has a very high MAPE of **44%** and RMSE of **135.03**

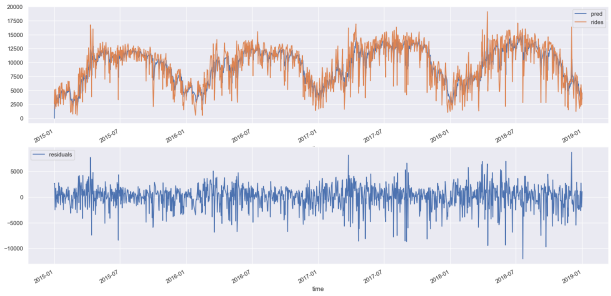


Figure 8: SARIMA model predictions vs actual, and residuals

a. This indicates that although the time series is stationary and exhibits clear seasonal behavior, the AR and MA components of the timeseries are statistically significant over many lag values which might cause overfitting.
b. Another potential approach is to use models like GARCH or VAR to deal with the high variability, but the efficacy is tenuous based on the results from the primary AR models.
c. Therefore, we accept our hypothesis that the data does not display a strong autoregressive component and building a regression model by engineering several features including information like time, weather and holidays seems to be a better method.

CHALLENGES AND NEXT STEPS

One of the major challenges that we faced was the size of the data we were dealing with, which pushed the limits of what we were used to and hindered the rate at which we could develop our solution. However, the stratified sample from the data served us nicely in that regard, and got us used to the process such that we don't have trouble completing the project.

For next steps, we'd want to continue EDA with data from our external sources (weather and holidays), and use that in conjunction with our original data to perform multivariate analysis and answer some questions. We'd also like to augment said analysis by using explainable models (like linear regression and perhaps decision trees) to understand the different features and their significance. On the demand forecasting front, we want to eschew a time series based approach and instead treat each observation on its own based on our results. This entails engineering the timestamp feature to something more usable,

and also investigating the effects of our external data sources (weather and holidays) as mentioned in our literature review. An interesting component to all of this is the geographic features and how to handle it, as the data being used for demand forecasting will have to be aggregated by the hour (or day) which said features are not able to be conformed to. A potential solution is pointed out in the literature review by utilizing a clustering approach. However, its efficacy on our use case remains to be seen.

WORKS CITED

A. A. Ramesh, S. P. Nagisetti, N. Sridhar, K. Avery and D. Bein, "Station-level Demand Prediction for Bike-Sharing System," *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, NV, USA, 2021, pp. 0916-0921, doi: 10.1109/CCWC51732.2021.9375958. https://ieeexplore.ieee.org/document/9375958

Daniele Gammelli, Yihua Wang, Dennis Prak, Filipe Rodrigues, Stefan Minner, Francisco Camara Pereira, Predictive and prescriptive performance of bike-sharing demand forecasts for inventory management, Transportation Research Part C: Emerging Technologies, Volume 138, 2022, 103571, ISSN 0968-090X, https://doi.org/10.1016/j.trc.2022.103571.

Chengcheng Xu, Junyi Ji, Pan Liu, The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets, Transportation Research Part C: Emerging Technologies, Volume 95, 2018, Pages 47-60, ISSN 0968-090X, https://doi.org/10.1016/j.trc.2018.07.013.

Yuanxuan Yang, Alison Heppenstall, Andy Turner, Alexis Comber, Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems, Computers, Environment and Urban Systems, Volume 83, 2020, 101521, ISSN 0198-9715, https://doi.org/10.1016/j.compenvurbsys.2020.101521.