MGT 6203 Group Final 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

Table of Contents

Overview of Project

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. Therefore, we are trying to identify the elements that play a significant role in affecting demand and how we can model the demand as accurately as possible for the purpose of prediction based on this information. Our hypotheses generally revolve around the impact that temporal factors (like hour of the day, or holidays vs non-holidays) and weather (like temperature, precipitation) have on both casual riders and members. Briefly, We believe that certain time periods (like the 9-5) are overrepresented and that perhaps poor weather conditions negatively affect demand.

Overview of Data

Our data was sourced from Capital Bike Share who generously provides first party data based on their Data License Agreement. The data contained features corresponding to individual rides, so the start and end stations, the start and end times and ride duration. It was further 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.

The tidiness of some data is relative, but in our case it was rather clean as the data was well maintained. For our analysis, we had to confer appropriate data types to the features and did not have to deal with missing data in the traditional sense. Any non-holiday rows had NaN values which we promptly replaced with the string “Not a holiday”. We placed checks for consistent capitalization of categorical values, and also for impossible values (like zero/negative ride distance, for example). Finally, we utilized Tukey’s fences as a general approach to outlier removal with k=1.5 (although we admit that it is 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).

Our key variables were ride duration and user membership. These were primary in our analysis as they are correlated with profit and hence were focused on during EDA and also modeling. On the demand forecasting front, the number of rides per hour was our key variable. Feature engineering was imperative for analysis and especially modeling efforts when breaking down timestamps into constituent datetime elements like hour, day of the week, and day of the month, for instance. We also used boolean values to identify weekends and holidays which further streamlined said analysis.

Overview of Modeling

We used a variety of models based on the task at hand. They include:

1. **Linear Regression:** A tried and tested baseline, especially where explainability is concerned. We had to eschew it in favor of more sophisticated models due to its inability to explain the variation in the data for reliably predicting duration and demand.
2. **Logistic Regression:** A close cousin of Linear Regression. Used in the case of determining the factors that affect membership.
3. **Decision Trees:** Used as an alternative to linear regression to explain the factors that affect ride duration. It performed significantly better, as we will see later.
4. **SARIMA:** Our first time series model utilized towards demand prediction based on strong seasonal behavior and stationary attributes of the data. However, we forwent the use of time series based models for more conventional models due to the lack of the autoregressive component of the data which resulted in rather unsatisfactory MAPE and RMSE scores.
5. **Random Forest:** Our final model for the demand prediction use case which performed remarkably well. Explainability took a hit for a more complex model, but we prioritized performance in line with our objective to determine the feasibility of accurately predicting demand.

Simpler models were chosen first as a reliable baseline (linear regression, SARIMA), especially where interpretability was a concern. We graduated to the use of more complex models to achieve our objectives, in some cases preserving the benefits of greater explainability (decision trees) and in other cases not (random forest). The final models were useful in different ways. Explainable models (in the case of duration and membership modeling) gave us more insights and confirmed some of the findings made via EDA, while black box models (in the case of demand forecasting) validated our objective that we could indeed obtain accurate predictions. The final models have (what is conventionally understood to be) great performance.

Methodology

Our methodology followed a two prong approach in reference to our use cases. The first objective was to conduct general analysis on the data via EDA and modeling, especially with regards to our key variables, to obtain valuable insights. The second objective was to determine the feasibility of developing an accurate demand forecasting model, which involved its own EDA and modeling tasks. There was a smaller, more separate objective that straddled the line between the two with regards to dealing with the geospatial features present in our dataset. This required different techniques and was hence provided special attention.

Results

Univariate Analysis

As it can be seen in the Figure 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
```

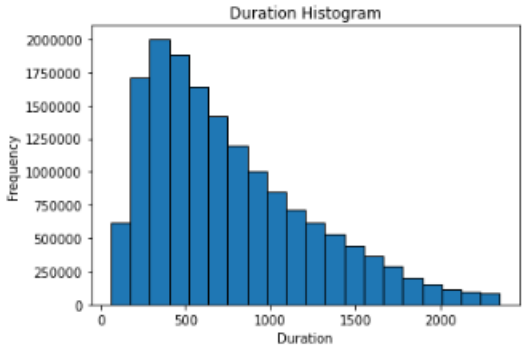


Figure: count, mean, standard deviation, and five-number summary of duration of rides

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. 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%).

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 durations 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.

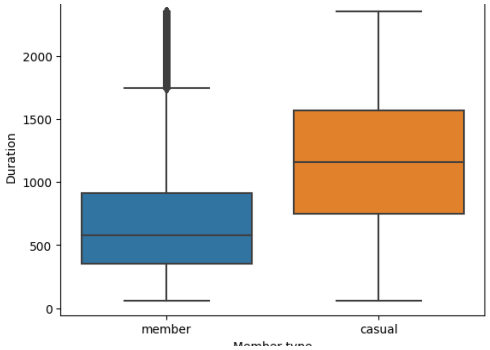


Figure: Box-plot of ride durations by member type

Bivariate Analysis

Firstly, there is a notable fluctuation in average ride durations throughout the day, while the average duration is high around 10 AM to 3 PM, while it is relatively low around 5 AM.

Furthermore, it's noticeable that ride durations also have variance based on the day of the week. Weekends, specifically days 5 and 6, which represent Saturdays and Sundays, have longer average ride times in contrast to weekdays. Additionally, monthly analysis indicates that the average ride durations are notably higher in the months of April, May, June, and July compared to other months.

A binary column, 'isWeekend' was introduced to the analysis, having 0 represents Mondays, 6 represents Sundays. Based on the analysis, isWeekend has a substantial difference in average ride times. Weekdays have an average ride time of approximately 997 units, whereas weekends have a significant increase, reaching an average of approximately 1457 units. This highlights a noteworthy impact of the day of the week on ride durations, with weekends consistently featuring longer rides.

Similarly, the classification of holidays and non-holidays (isHoliday) reveals a substantial effect on ride durations. Non-holidays register an average ride time of around 1111.58 units, while holidays significantly surpass this, reaching an average of approximately 1406.41 units. This indicates that holidays tend to be associated with longer average ride durations compared to regular days.

In essence, the analysis demonstrates that ride durations are influenced by various temporal factors, including the time of day, day of the week, and whether it's a holiday. These insights could be valuable for optimizing ride-sharing services and anticipating user demand based on temporal trends.

Pearson Correlation Coefficient Between Weather Variables and Riding Time

In order to explore the correlation between weather conditions and riding duration, Pearson correlation coefficient was used to understand how variations in temperature may be associated with changes in the duration of rides. In this case, the Pearson correlation coefficient was calculated between the 'temperature_2m' variable (representing the temperature) and the 'Duration' variable (representing riding time), having a result of statistic=0.048 and p-value = 0.0 meaning there is a weak correlation between temperature and the riding duration. As the temperature increases, there is a slight tendency for riding time to also increase.

Furthermore, analysis was conducted to find correlation between precipitation (rainfall) and the duration of rides to find out whether there is a meaningful connection between these variables. As a result of the Pearson correlation coefficient being -0.0098, the p-value = 0.0, which is significantly lower than the significance level of 0.05, suggesting that when there's rainfall, riders' duration tends to be shorter.

More analysis was performed to see if there's a correlation between humidity and the duration of rides. As a result of the Pearson correlation coefficient, the coefficient is 0.000945 and the p-value is 8.798e-05, which indicates that it is statistically significant, but the coefficient is low, so there is only a little impact of humidity on riders' duration.

As for the relationship between wind speed and duration, column 'windspeed_10m' was used to perform the Pearson correlation coefficient. Having the Pearson correlation coefficient to be -0.028 and the p-value to be 0.0, it suggests that as wind speed increases and the duration tends to be slightly shorter.

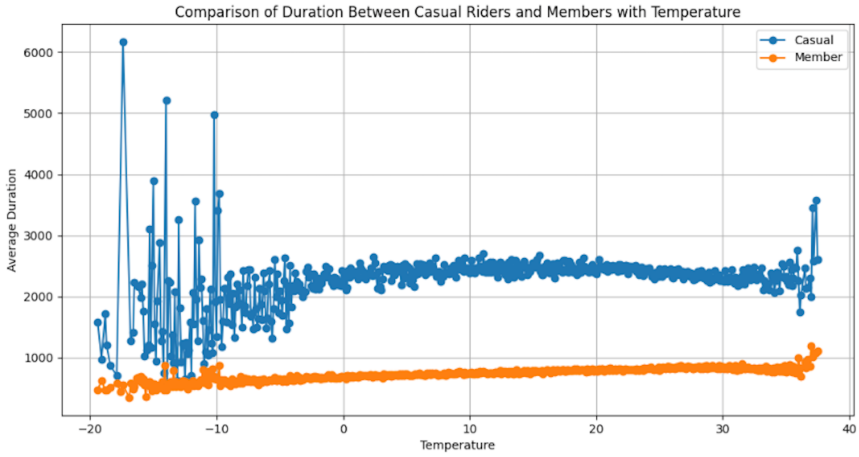
Multivariate Analysis

In this multivariate analysis, we aim to understand the variations in ride durations between casual and member riders across different weather conditions and time periods. The methodology involves distinguishing between casual and member riders, and then comparing their average ride times under various conditions.

First of all, a new binary column, 'isWeekend,' is created to distinguish between weekend (x>=5) and weekdays (x<5) riders based on the 'w_day' column. Secondly, a function called "avg_duration_by_group" is created to understand the average riders' duration under different weather conditions.

A chart was made to compare duration for different types of riders with regards to precipitation. For casual riders under precipitation, the ride durations varied across precipitation levels. When precipitation is at 0.0, the average duration for casual riders is 2430 units and as the precipitation becomes higher, the average duration for casual riders decreases. For members, the trend is similar, the average duration is 780 at 0 precipitation and the duration decreases to 555 units when precipitation at 16.1.

To compare duration between casual riders and members with temperature, the duration increases slightly as temperature increases.



By comparing the duration between casual riders and members with humidity, there's no noticeable trend in the average ride duration with changes in humidity levels. It suggests that humidity may not be a significant factor affecting the duration of bike rides in this dataset.

Compare duration between casual riders and members with wind speed,for both casual riders and members, there's a slight decrease in riders' duration as wind speed increases. For example, at the lowest wind speed of 0.0, the average duration is high (2995.41)

When comparing duration between casual riders and members during weekdays vs weekends, the data is segmented into two categories: weekdays (0) and weekends (1).For casual riders during weekdays (0), the average ride duration is approximately 2311.41 units, while the duration increases to about 2552.63 units during weekends

(1). This suggests that casual riders tend to engage in longer rides during weekends, possibly indicating a higher leisurely or recreational usage pattern during these days.

To compare duration between casual riders and members during holidays vs not a holiday, we can see a clear increase of duration in holidays for both casual riders and members. For casual riders during non-holidays (0), the average ride duration is approximately 2411.90 units. On holidays (1), there is a noticeable increase in the average ride duration, reaching around 2575.49 units. For members, the average ride duration is 776 units for non-holidays and 839 for holidays. This suggests that both casual riders and members tend to extend their ride durations during holidays.

Logistic Regression Model

We first built a logistic regression model with 'Member type' as the target variable and 'Duration', 'isHoliday', 'month', 'm_day', 'hour', 'w_day', 'temperature_2m', 'relativehumidity_2m', 'precipitation', 'windspeed_10m' as the independent variables. The goal was to predict the likelihood of a rider being a member based on temporal and meteorological factors. The model performed reasonably well with an accuracy of 0.85, precision of 0.82 and sensitivity of 0.97, as can be seen in the Figure.

The model also had a high ROC score of 0.78, as shown in the Figure. This shows that the model does a good job predicting whether a rider will be a member or a casual rider. This insight can be extremely useful to the business when planning out marketing campaigns and resource allocation. These statistics can also be beneficial to assess current membership plans and potentially introduce new membership or subscription models in the future.

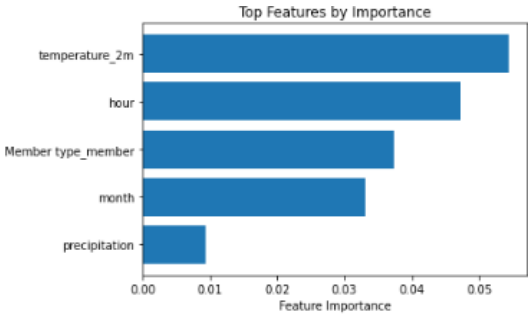
The top five variables with the largest coefficients are isHoliday, w_day, precipitation, hour and month which are representative of members. While the p-values reject the null hypothesis that the coefficient of the variables have no effect on the target variable, further feature selection might be required to carry out additional analysis. At this point, it is worth it to note that holding all other variables constant, a one unit increase in duration leads to a decrease in the log-odds of the rider being a member, rather than casual.

Linear Regression

We also fit a linear regression model with 'Duration' as the target variable and the rest of the variables as the dependent variables. The goal was to understand the dependency of trip durations on other factors. This could potentially help the business with inventory management and efficient resource allocation. However, the linear regression model did not perform very well with an R-square value of 0.02 and MSE of 224923.4815. Choosing a subset of different combinations of features did not improve the model performance either. Therefore, we decided to go a different way and fit a decision tree model to explore 'Duration'.

Decision Tree Model

In order to explore the relationship of trip durations with other variables, we fit a decision tree model. After experimenting with different subset of variables and conducting a feature importance test, the final the dependent variables that were chosen for this model are 'Start station number', 'End station number', 'Member type', 'temperature_2m', 'relativehumidity_2m', 'precipitation', 'windspeed_10m', 'holiday', 'isHoliday', 'year', 'month', 'm_day', 'hour', 'w_day'. Columns that contained redundant information or caused multicollinearity were removed e.g. 'Start Station number' was kept instead of 'Start Station Name'. The model performed reasonably well, with an R-square value of 0.72.



The top five features according to a features importance test are temperature, hour, member type, month and precipitation. Therefore, we can conclude that the trip duration is significantly impacted by these features.

Geospatial Data

Looking at the departure data, there are a few things we want to check. Firstly, if there is a general trend with the demand for each station, as well as if there is a difference in demand in each station based on whether it's a weekend, or a holiday. The following figure shows the top 50 stations in terms of usage across the 5-year period, along with their usage specifically during weekends, and holidays.

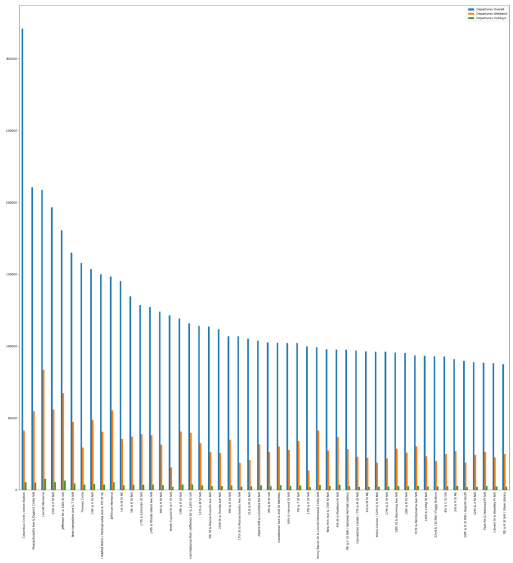


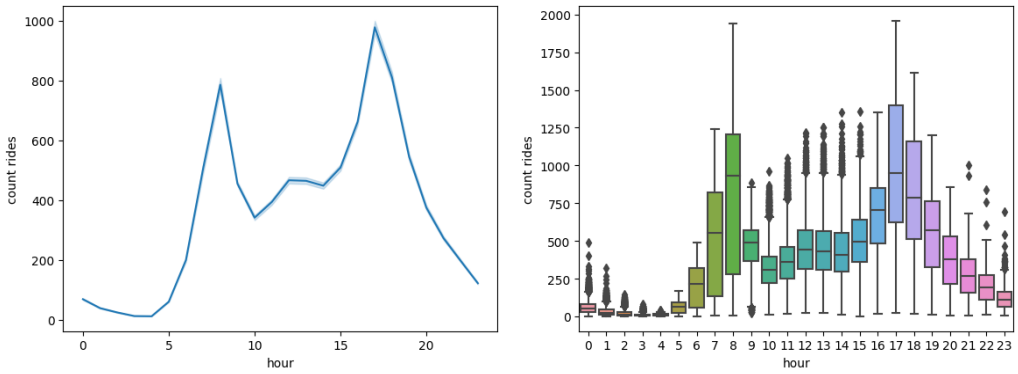
Figure: Top 50 stations ranked and ordered based on total departures.

Though the holiday information (green bar) isn't as clear to see, there does seem to be a distinct difference in popularity for the stations during the weekend, as opposed to their overall usage, which can be seen from the fact that the orange columns are not sorted in the same way. In fact, looking at the proportions, the most popular station overall Columbus Circle / Union Station has an extreme drop off in usage, to the point where other stations are more popular than it, despite it's great lead overall.

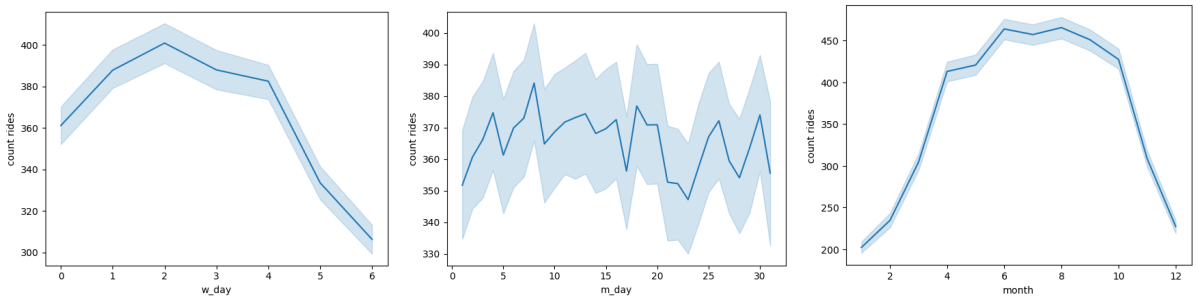
Taking a closer look just the weekend and holiday information, we get the following image. Though still not perfectly clear, it seems there are more similarities on each stations popularity during weekends and holidays, so they likely have slightly similar effects on overall usage. In general, these may be due to different stations being more popular for commutes, or general errand runs as opposed to what they may be used for during weekends/holidays. Taking a closer look at the top 3 stations and their daily departures, over the time period there is a fairly clear general trend that repeats each year, which suggests there is a seasonal trend that affects the demand.

Demand Forecasting EDA

When using the number of rides per hour as a key variable, charting it against the hours of each day produces the strongest relationship in this use case. We can clearly observe that there is a peak around 8 AM, followed by a sharp dip, which is once again countered by another peak around 5 PM. This corresponds rather neatly with the 9-5 workday, and fits in nicely with previous findings.

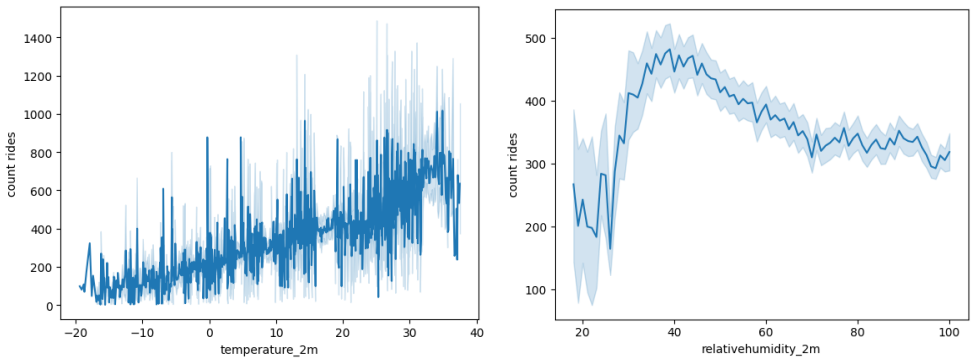


Given the amount of variation caused by holidays, weekends, and people who use bikes for purposes other than work, it is rather pertinent how strongly this particular 9-5 signal stands out from the noise, highlighting its role as the status quo. This is further reinforced by the fact that any number of rides outside said peak hours are considered outliers, as seen in the boxplot especially between the 9-5.

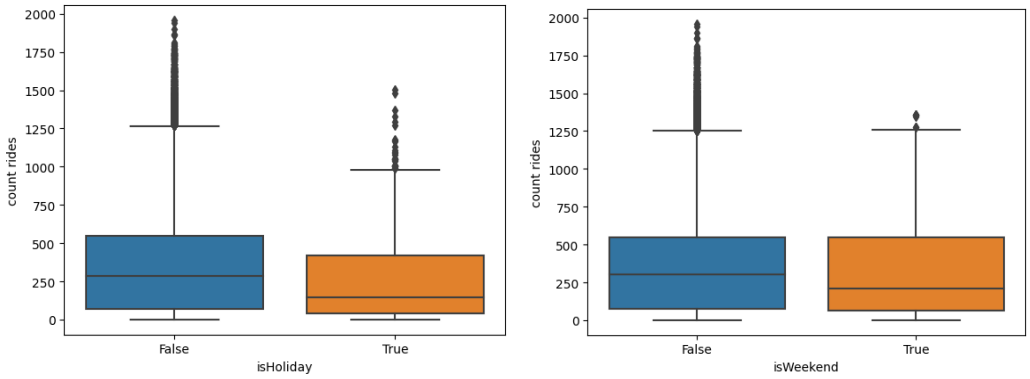


We can also observe a sharp dip in the number of rides during the weekend (0 representing Mondays), and when charted against the days of the month follows a cyclic pattern based on the weekly distribution. Curiously, when gauging the number of rides per month, we observe a consistent, distinct decrease at the beginning and ends of the year and peaks around the middle of the year. This may be because more riders use the service more in the summer than the winter, but that does not account for the significant work population who do their jobs (mostly) year round. This should also correspond to why temperature might be a contributing factor in the number of rides (and perhaps duration as well) since it's warmer in the middle of the year. We are not exactly sure why this is the case, but the findings show a clear effect of temporal variables on the number of rides.

As for the effect of weather on the number of rides, we can see that there is a positive correlation with temperature, although it's not that strong. An even weaker negative correlation exists with regards to humidity. No appreciable patterns can be gauged by charting against precipitation and wind speed.



Finally, we can observe some difference in the number of rides between holidays and non-holidays, probably owing to the great number of people utilizing the service for their commute to work. The same cannot be said for the distinction between weekends and weekdays purely from the boxplot, but weekdays depict more outliers. However it must be kept in mind that both the number of weekends and holidays are fewer than the number of weekdays and non-holidays.



Demand Forecasting Modeling

By analyzing the initial sampled data, it was evident that the time series exhibits a seasonal component and a very clear trend on both hourly and monthly levels. After testing for stationarity and auto-correlation, a Seasonal ARIMA model was built on hourly data using AR Lag of 1 and MA component of 2 with 1 day seasonal component.

This model had a very high MAPE of **44%** and RMSE of **135.03.** 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. Therefore, we accepted 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.

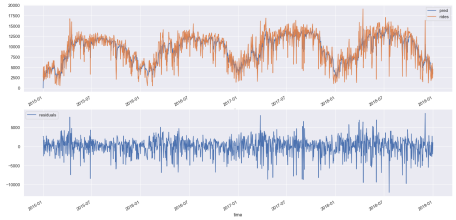


Figure: SARIMA model predictions vs actual, and residuals

In order to predict the demand, several features are engineered as per the initial hypothesis from the underlying ride dataset and station dataset. These include features like hour of the day, day of the week, month of the year which are more ordinal and other continuous variables like weather attributes are averaged over these intervals. Along with this, few binary variables like holiday/event and weekend are engineered. Since the categorical features like hour, day and month are cyclical in nature, instead of encoding them ordinary, these features are encoded with a sin function to mimic its periodic behavior.

From the Ride dataset, the number of rides is calculated for each of this group by summation, curating a training dataset of 2806 data points in total. This training dataset is passed through a linear model and Random Forest Regression model with Number of Rides as the target with 3 categorical temporal features, 4 weather features and 2 binary features. In order to avoid overfitting, the parameters are kept to the default settings with 32 trees in the forest. The same experiment is also repeated by using cyclical-encoded categorical features to understand the temporal feature effects.

It is evident that Random Forest performs significantly better than linear models in capturing the relationships at near 99% R2. By analyzing the feature importance based on gini impurity, it is clear that Hour of the day and temperature along with whether the day is a holiday or not makes a lot of difference. This neatly ties up with our initial hypothesis and preliminary EDA analysis where it was clear that the demand always peaked during certain hours of the day irrespective of other temporal features. A general positive trend is observed between Temperature and number of rides and a reduced demand during holidays.

Model	r2	mae	rmse
LinearModel_Ordinal	0.647577853	4769.068178	6285.311402
LinearModel_Cyc	0.572865081	4725.839719	6229.675506
RandomForest_Ord	0.993875091	321.3069873	612.8643186
RandomForest_Cycl	0.966541151	797.1974898	1365.927409

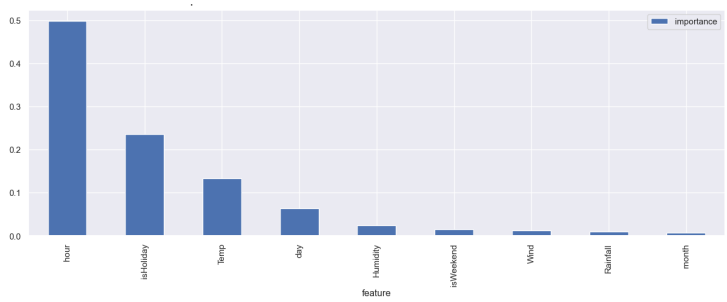


Figure: Feature importance for Random Forest model

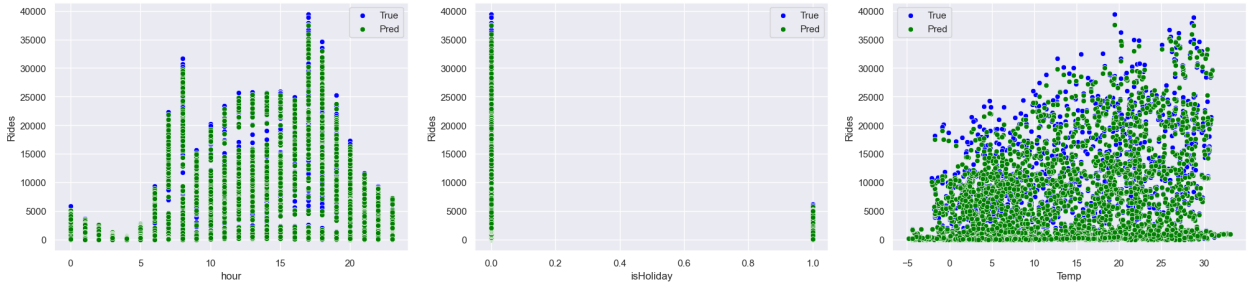


Figure: True vs Predicted values for Hour, isHoliday, and Temp variables

While a holistic view is rather vital to understand the macro part of Demand in the DC Area, station level analysis is necessary to understand certain nuances of demand and predict the accurate factors influencing them. In order to predict station level demand, the stations are grouped into clusters based on the Geographical location and analyzed individually. This stems from the belief that stations which are located closer to one another might experience similar riding patterns due to proximity to certain prominent destinations. K means clustering is used to group the stations into 25 and 45 clusters respectively and the ride data pertaining to all the stations in a specific cluster is analyzed individually.

Upon analyzing the individual models for each cluster, it was clear that the top features remain the same and match with the temporal analysis for most clusters. However, there are few clusters like the ones near the neighborhood of Fairfax County and Montgomery County that have slightly different top features where weather features like Humidity and Rainfall are given more importance than other temporal variables. The analysis also highlighted that the clusters which are located in the extreme outer regions of the DC Area have poor models (Having very low R2). This could be a result of low number of rides from these clusters resulting in a small training dataset which in turn produces a poor model.

To address the problems of inventory management, the clusters were analyzed by their average ride prediction on an hourly basis and the total capacity that they own. It was evident that most of the demand arises from the central region of the DC Area and the stations in these neighborhoods suffer from supply problems. The capacity of these stations are slightly lesser than the average hourly ride demand thus impacting the revenue. On the other hand, certain stations towards the southern and eastern borders of the DC Area have a capacity of more than 7 times of the maximum hourly demand for different weather conditions. This calls for a reshuffling or movement of bikes from surplus-demand stations to one with insufficient-demands.

Furthermore, analyzing these clusters with disproportionate demand to supply with a temporal aspect, it is evident that the high demand clusters are only under supplied during peak hours that is from 8am to 8pm whilst the stations with surplus capacity have no visible temporal effects. As a result, if increasing station capacity is limited due to other external factors, the bikes could be moved around from one cluster to another during non peak hours thus ensuring all demands are met optimally.

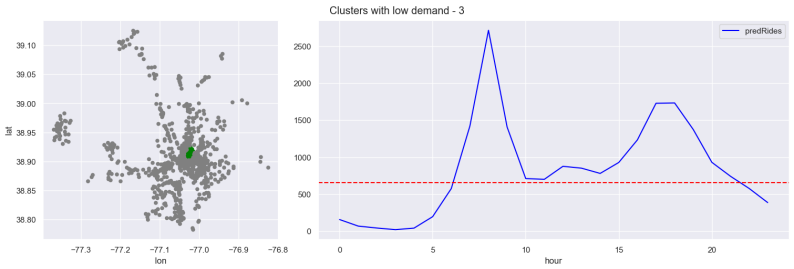


Figure: (right) Demand per hour for the low demand clusters (highlighted in left figure)

Conclusion, Takeaways, and Future Work

In conclusion, we analyzed and determined the significant effect temporal features have on key variables, especially the number of rides. We also observed the relative insignificance of weather effects on said variables as well. Also, event data, especially holidays, proved to be useful information in terms of modeling demand. Finally, we were able to develop a remarkably performant demand forecasting model after just a few iterations based on the collated data.

In terms of future works, there are plenty of potential (and definitely interesting) paths to take. For instance, more could be done with the geospatial data in terms of analyzing common routes and gaining new insights. On the demand modeling front, we opted to cluster the stations just based on geographical distance which served us well as an important preliminary implementation. However, stations far away from each other can have more in common based on proximity to certain key areas (like schools, offices, etc.) so investigating and utilizing metrics other than geographic distance could yield accurate results at a station level. This is what has also been referenced in our literature review.

Although we have explored how our demand forecasting model could work with regards to inventory management, efforts could be made on defining a more comprehensive framework (unfortunately, inventory management is the final module in the course). Also, given access to financial information (like membership costs throughout the years and operational costs), we could analyze and provide some understanding about the sustainability of the business. Furthermore, thorough investigations could be done on outliers, instead of using Tukey's fences as a blunt instrument (which served us well regardless, and was imperative to the timely completion of the analysis). Finally statistical tests (like the t-test or Mann-Whitney test) could be performed to obtain robust results with regards to the impact of boolean variables like holidays and weekends. However, we were limited by the sheer sample size of our data, which would "overpower" the results and result in misleading conclusions. These are but some avenues for possible future work.

Bibliography

- *System Data | Capital BikeShare. (n.d.).* https://capitalbikeshare.com/system-data
- *How to Check for Outliers — datatest 0.11.1 documentation. (n.d.).*

  https://datatest.readthedocs.io/en/stable/how-to/outliers.html
- *Feature Engineering - Handling Cyclical Features*. (n.d.). David Kaleko.

  https://blog.davidkaleko.com/feature-engineering-cyclical-features.html
- 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

Appendix

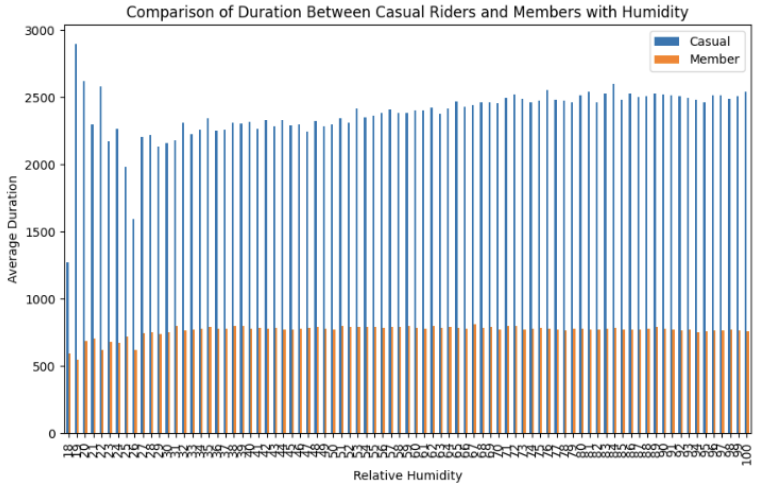


Figure: Duration vs Humidity

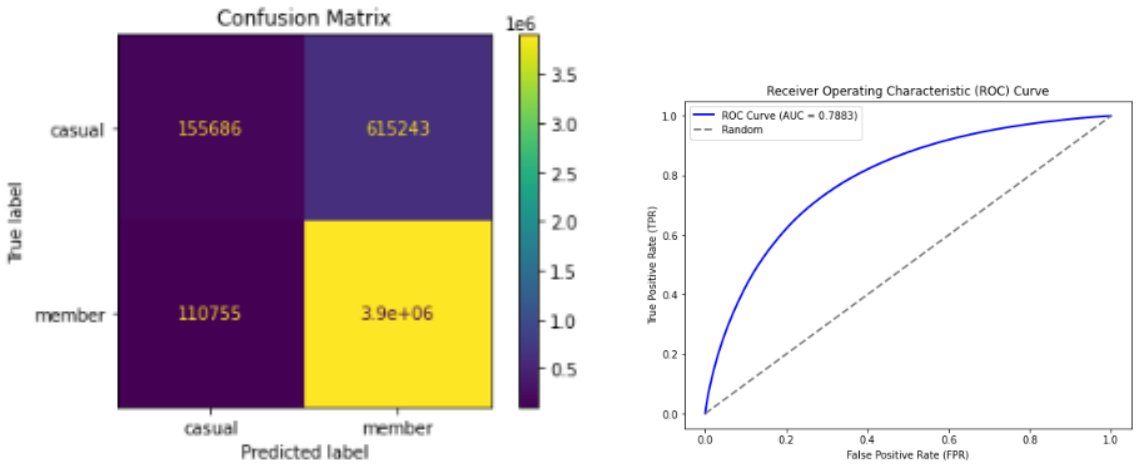


Figure: Confusion matrix results from logistic regression model and ROC curve

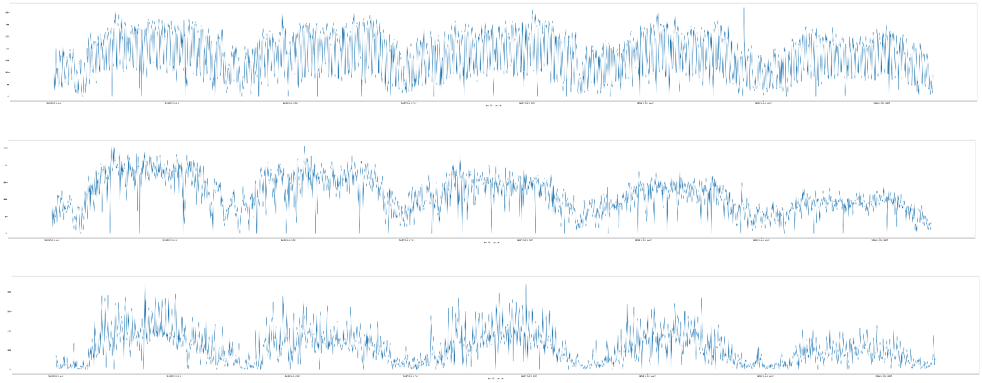


Figure (from top to bottom): Daily departures over the 5-year period for Columbus Circle / Union Station, Massachusetts Ave & Dupont Circle NW, and Lincoln Memorial

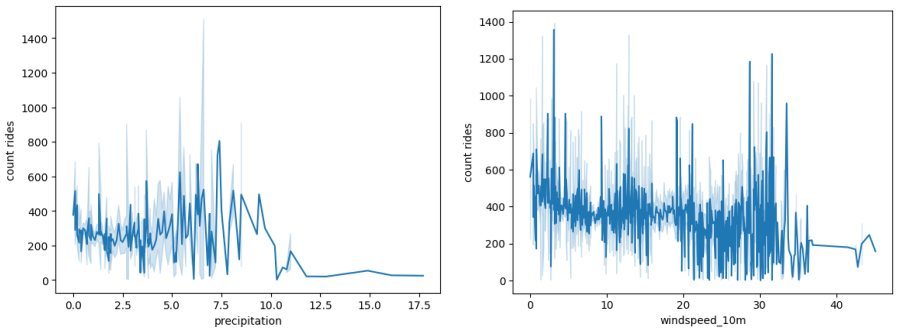


Figure: Number of Rides against Precipitation and Windspeed

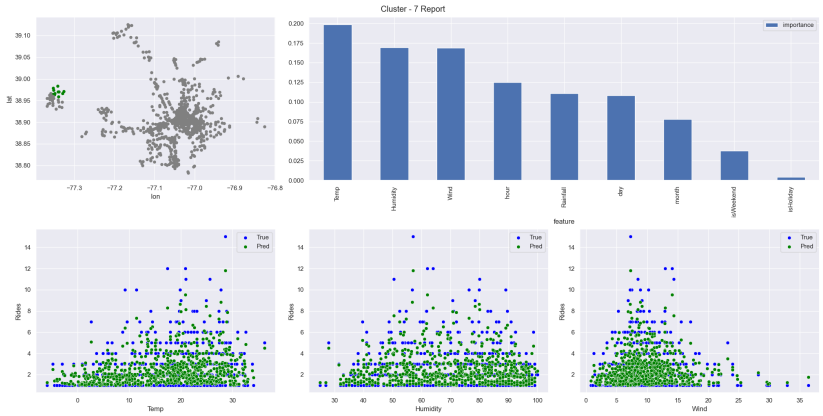


Figure: Fairfax County Cluster Report

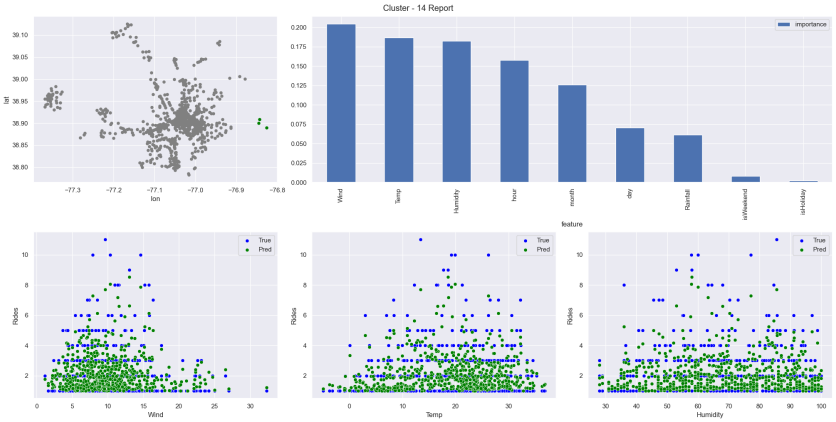


Figure: Montgomery County Cluster Report

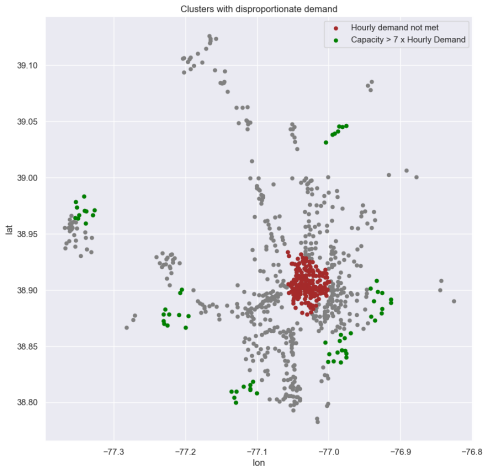


Figure: Highlighting clusters with exceptionally high (red) and low (green) demands

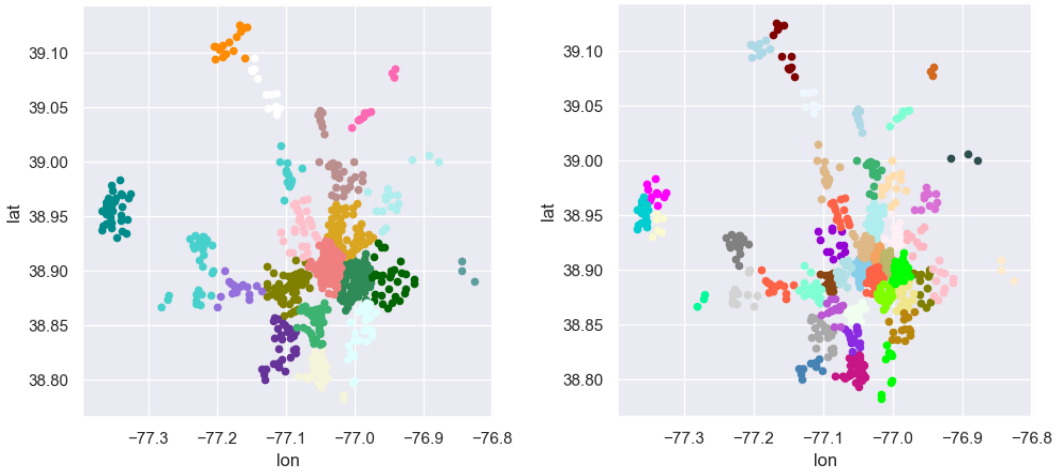
```
Optimization terminated successfully.
         Current function value: 0.364149
         Iterations 7
const                    0.000000e+00
Duration                 0.000000e+00
isHoliday                0.000000e+00
month                   1.183155e-101
m_day                    5.165257e-01
hour                    1.929518e-232
w_day                    0.000000e+00
temperature_2m           0.000000e+00
relativehumidity_2m      6.619614e-60
precipitation            1.199388e-33
windspeed_10m            1.982665e-82
dtype: float64
```

```
Accuracy: 0.8498307368491529
Precision: 0.8658
Sensitivity: 0.9718859898183847

isHoliday: -0.8666306737051184
w_day: -0.1923602445889657
precipitation: 0.12464081971369148
hour: -0.027753822339380764
month: 0.027713620641210412
temperature_2m: -0.021594059769123725
windspeed_10m: 0.0139360760205398
relativehumidity_2m: -0.003153240579171547
Duration: -0.0019013256813678158
m_day: 0.0003287213728424389

Intercept: 4.648964842631088
```

Figure: Logistic Regression model results



Figures: Grouping the stations into 25 and 45 clusters

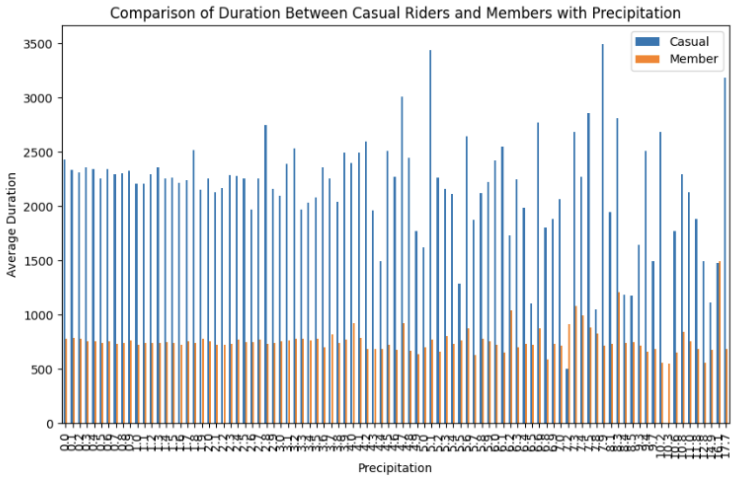


Figure: Duration versus precipitation for casual riders and members

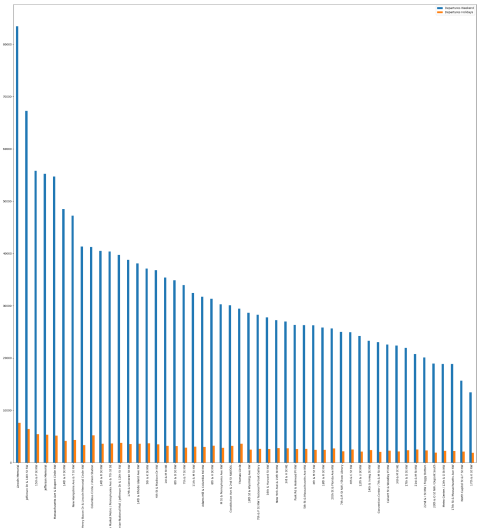


Figure: same top 50 stations, this time ranked in order of popularity during weekends.