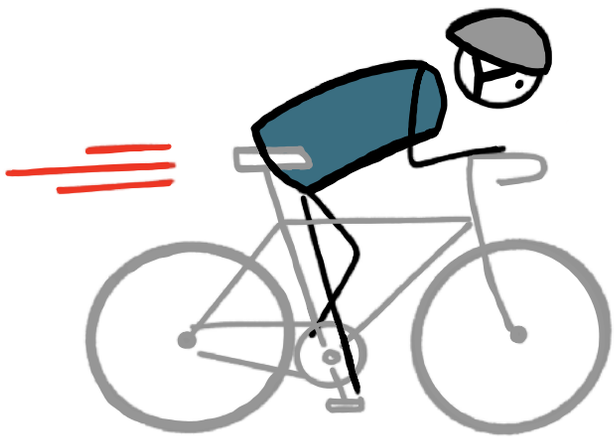
Team One

A Data Analytics Approach: Leveraging Real-Time Data for Optimizing Bike-Sharing Operations in Los Angeles

Yusuf Shehadeh (Student ID: 7395116)
Mohammed Sabaawi (Student ID: 7381108)
Kemal Yazgan (Student ID: 7396285)
Assia Benmimouni (Student ID: 7381347)
Samuel Rizzo (Student ID: 7395422)
Ivan Kamal Mehieddin (Student ID: 7396805)



Contents

Chapter 1

Excecutive Summary

Business Problem The Los Angeles bike-share system faces challenges to overcome traditional urban mobility issues, including pollution, safety concerns, and inefficiencies. Our analysis aims to improve system performance and promote a sustainable and efficient transportation alternative.

Data Our analysis shows that demand for bicycles varies throughout the day and week, with peak periods in the evening and on Fridays. In addition, we examined the impact of weather conditions, such as temperature and precipitation, on the use of the system

The Analytics Solution To address the business problem, predictive analytics models were developed to predict idle times at bike stations. Feature engineering was used to derive relevant features from the data to improve the accuracy of the models. Two main models, DecisionTreeRegressor and LightGBMRegressor, were selected to perform the predictions. A stacking regressor was used to further improve the prediction performance.

Implications The analysis revealed important findings about usage patterns, weather conditions, and the impact on bicycle use in Los Angeles. Peak periods were identified, and weather conditions such as temperature and precipitation were shown to affect bicycle use. In addition, potential expansion and closure locations for bicycle stations were identified to increase system efficiency and improve service quality.

Recommendations The analysis results provide important insights to optimize the bicycle rental system in Los Angeles and promote sustainable urban mobility. Based on these findings, we recommend the following actions:

- **Optimize the station network:** Identify potential locations for expansion or closure of bike stations to increase system efficiency. Targeted station placement can better manage peak times and locations.
- **Targeted marketing campaigns:** Use identified peak times for bicycle use to implement targeted marketing campaigns. By increasing demand at specific times, you can increase system utilization.

- **Account for weather conditions:** Adjust bicycle operations scheduling based on weather conditions. Incentives or offers can be provided during inclement weather to encourage usage.

Chapter 2

Detailed Report

2.1 Problem Description

Urban transport has always relied heavily on hybrid vehicles, raising environmental and safety concerns. A major shift in mobility is required to solve this and meet the decarbonization goals. This change includes the growth of transportation services such as bike sharing. The project aims to help bike-share companies use real-time data to simplify operations, increase profits and improve service quality. The main goal of the business is to increase the efficiency of bike sharing by gaining a deeper understanding of the docking station network and to accurately predict when the bike is not working.

2.2 Business Goal

The main aim of the project is to increase the efficiency and profitability of bike sharing companies in smart competition. This includes improving network awareness, optimizing bike sharing and improving customer service. In addition, the second goal is to make these companies important in the city's security plans. By improving efficiency, these companies can see how alternative transport can help reduce carbon emissions, traffic congestion and health problems associated with urban transport models.

2.3 Data Science Goal

The project aims to conduct a thorough analysis of the docking station network and bike usage patterns, interpreting existing data to understand user behavior, peak usage times, popular routes, and key stations. This information will help operators shape their operational and marketing strategies.
Additionally, the project plans to develop a predictive model to accurately forecast bike idle times at different docking stations. This model will use historical usage data, time series analysis, and possibly real-time factors like weather or day of the week. A precise prediction model will allow operators to optimize maintenance schedules, improve bike redistribution strategies, and ensure bike availability at all docking stations, thereby improving service quality.

2.4 Data Description

As part of our project, we obtained access to data on bikesharing in Los Angeles as well as weather data for the corresponding time period. This data includes information such as temperature, cloud cover, cloud description, pressure, wind speed, precipitation, and felt temperature.

2.4.1 Bikesharing Data

The bikesharing data includes information on about 1.5 million transactions in Los Angeles. Each transaction includes the start and end time of a bike booking. There are also another 4 columns containing the geographic coordinates of the start and end stations (`start_station_lat`, `start_station_lon`, `end_station_lat`, and `end_station_lon`). In addition, the IDs of the start and end stations (`start_station_id` and `end_station_id`) and the ID of the bike used (`bike_id`) are recorded.

2.4.2 Impact of Weather on Bikesharing

Combining the two datasets data allows us to more closely examine the impact of weather on the use of the bikesharing system in Los Angeles. We can analyze how different weather conditions affect bike demand, trip duration, and other relevant parameters.

2.4.3 Advantages of Using Idle Time at Stations

project emphasizes bike idle time at specific stations for several reasons. It enables bike sharing and reduces environmental impact by helping to identify high demand areas and accurately predict their demand. It can also improve infrastructure, such as expanding bike stations. Monitoring the use of stations prevents service interruptions and maintains high service levels. More importantly, it will help plan maintenance and repair to ensure the bike's reliability.

2.4.4 Reasons for Not Using Individual Bike Idle Time

The reasons why we have not elected for the idle time of individual bikes are as follows. First, this method may introduce distortions if certain bikes are inactive for long periods of time due to maintenance or repair work, unnaturally extending the idle time. This could lead to difficulties in distinguishing between actual idle time and technical problems.

2.4.5 Brief Data Preparation Details

The weather and metro data were cleaned and standardized. "Idletime" was calculated by comparing trip start and end times. Before visualization, additional work was done.
Thirteen new columns were added to the Metro file based on "idletime", indicating the time of day and day of the week when "idletime" occurred. A new table, "Result", was created with all end stations as indices, their geographic coordinates, average "idletime", and the sum of "idletimes" at different times of the day and days of the week.
The metro and weather datasets were merged based on the closest timestamp to create a comprehensive dataset for further visualization and predictive modeling.

2.5 Data Analytics

2.5.1 Descriptive Analytics

Station-Level-Insights

Our main task is to provide relevant insights into the current state and performance of the docking station network.

1. **Creating Comparative Histograms of the Top 20 Highest and Lowest Average Idletimes** For the visualization on a station basis, two histograms were first created to provide an overview of the twenty stations with the highest and lowest average idle times.

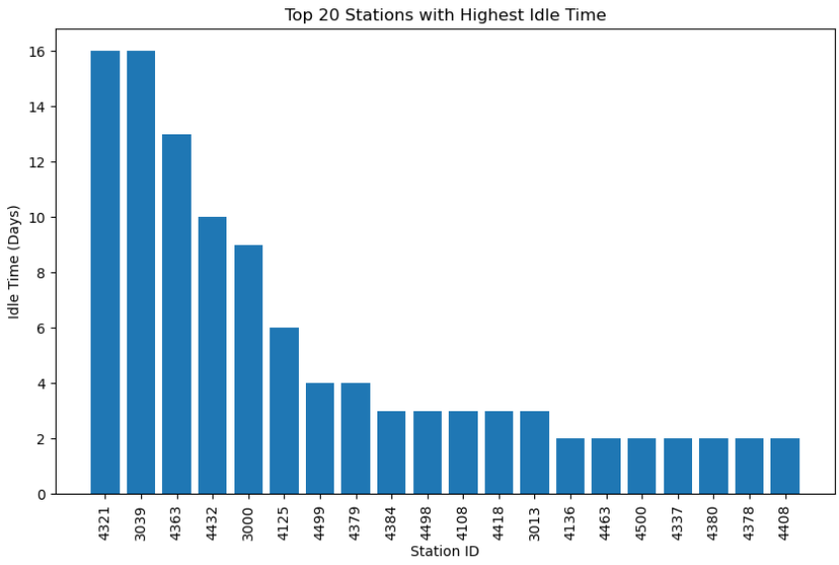


Figure 2.1: Top 20 Stations with Highest Idle Time

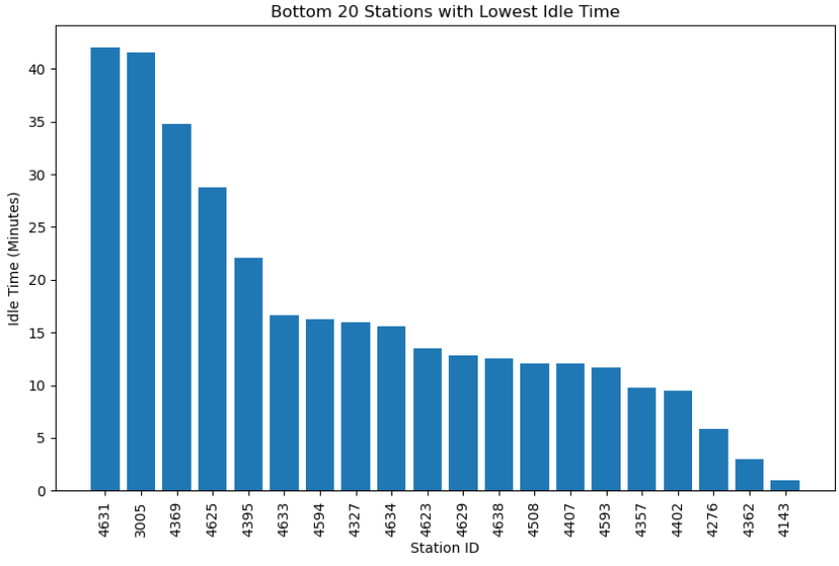


Figure 2.2: Bottom 20 Stations with Lowest Idle Time

2. **Identifying Expansion and Closure Candidates Based on Average Idle Time** The end stations were sorted by their average idletime and divided into two halves (upper 50% and lower 50%). For each half, the average "idletime" was calculated. Stations in the upper half, whose "idletime" is

above the average, are considered potential candidates for expansion. Stations in the bottom half, whose "idletime" is below average, are considered potential closure candidates.

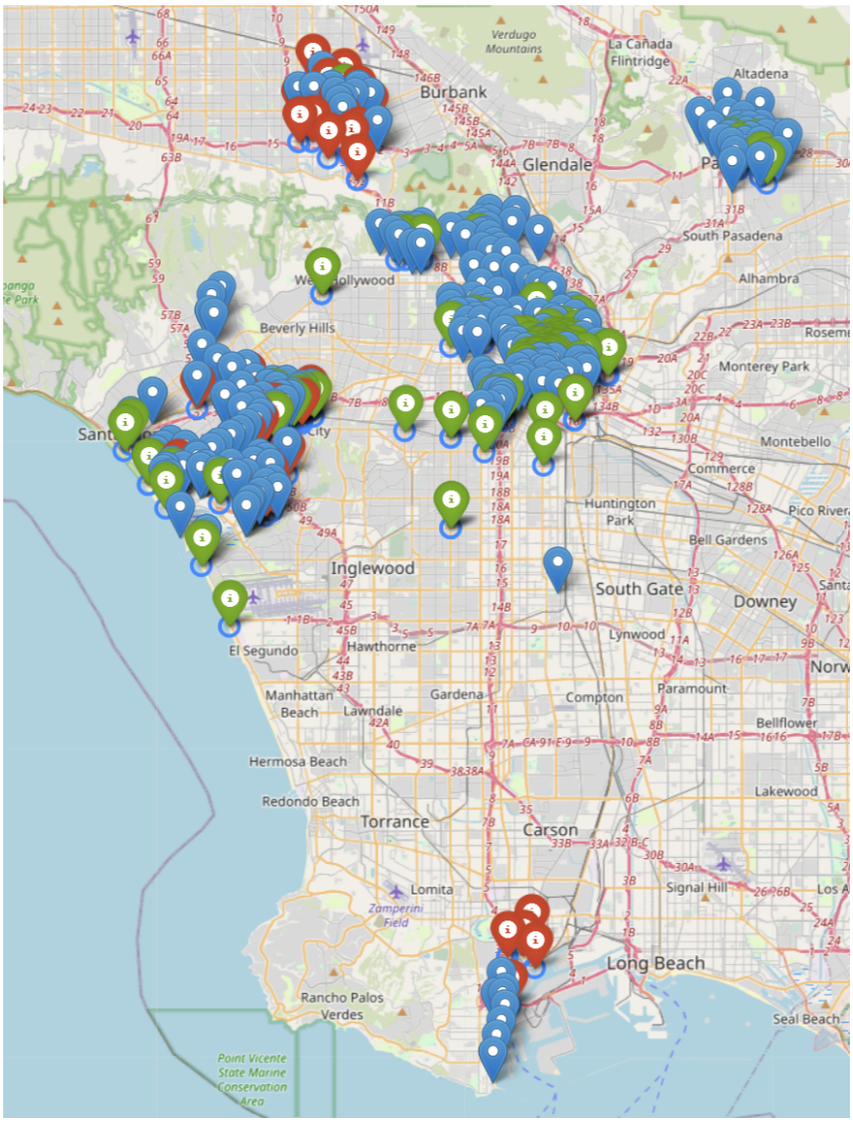


Figure 2.3: Map with 500m radius marker around candidates for closure or expansion

3. **Identifying Green Expansion Zones Based on Idle Time and Neighbor Analysis** On a map, all end stations are marked, and potential expansion candidates are indicated by a 500-meter radius ring. An expansion is suggested if the station's idletime is less than the average idletime of all neighboring stations within the 500-meter radius, and it has fewer than two

neighbors. All suggested expansions are listed and highlighted in green on the map.

To be expanded:

$$\begin{bmatrix} 4587 & 4632 & 4626 & 3014 & 4594 \\ 4327 & 4623 & 4629 & 4276 & 4362 \end{bmatrix}$$

4. **Identifiying Red Closure Zones Based on Idle Time and Neighbor Analysis** On the same map, all end stations are marked, and potential closure candidates are indicated by a 500-meter radius ring. A closure is suggested if the station's idletime is greater than the average idletime of all neighboring stations within the 500-meter radius, and it has more than two neighbors. All suggested closures are listed and highlighted in red on the map.

   To be removed:

$$\begin{bmatrix} 3039 & 4363 & 3000 & 4379 & 4384 \\ 4418 & 4463 & 4500 & 4337 & 4380 \\ 4408 & 4394 & 4338 & 4332 & 4373 \\ 4213 & 4374 & 4385 & 4341 & 4417 \\ 4413 & 4598 & 4462 & & \end{bmatrix}$$

   As you can see above 23 Station are to be removed and 10 Stations are to be expanded.

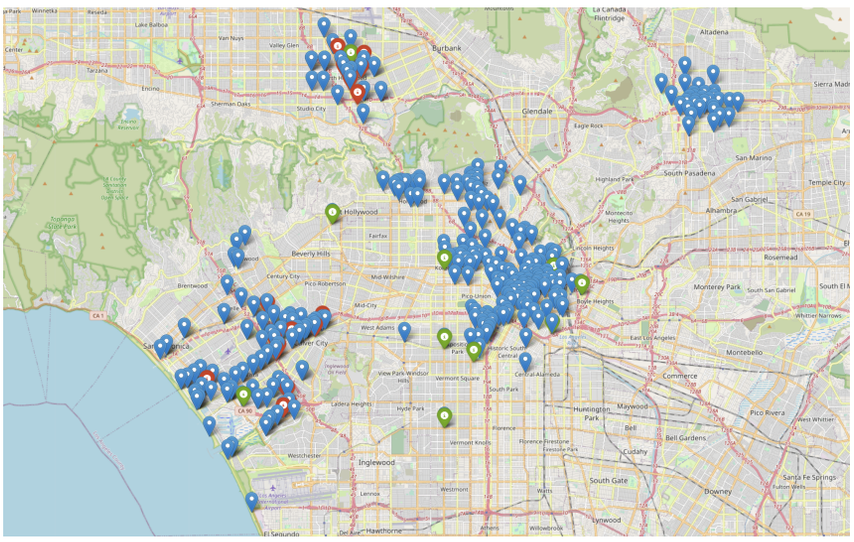


Figure 2.4: End result with stations to be expanded or closed

Overall System Performance

In preparation for the upcoming social media campaign promoting bikesharing in Los Angeles, valuable data insights are crucial to highlight the system's success and gain valuable user behavior patterns.

1. **Usage by Day** The analysis of usage by different hours of the day reveals interesting patterns. Between midnight and 5 AM, the number of bike rentals is relatively low, but as the day progresses, the demand steadily increases. The peak usage is observed around 5 PM (see Figure 2.5).

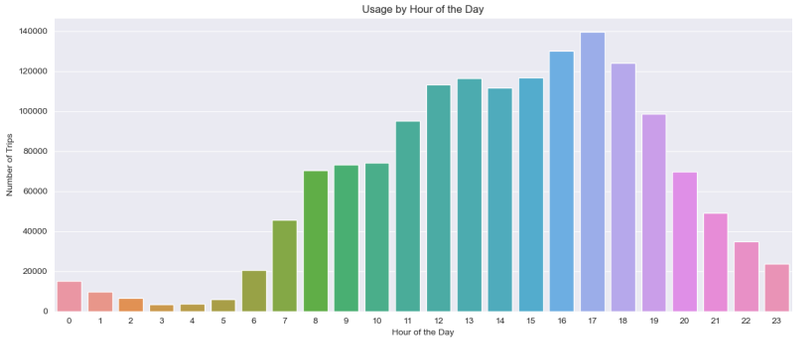


Figure 2.5: Hourly Distribution of Bike Rentals throughout the Day

2. **Usage by Day of the Week** Weekly usage shows a consistent trend. Trips are moderate on Mondays and Wednesdays, slightly higher on Tuesdays and Thursdays, peak on Fridays, and slightly decrease over the weekend. (see Figure 2.6).

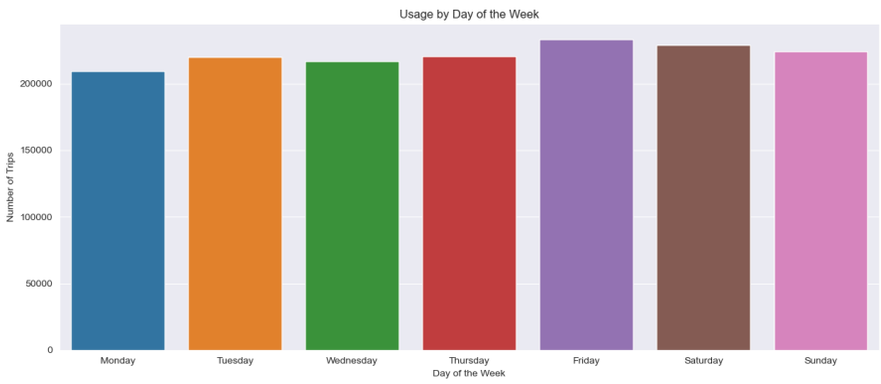


Figure 2.6: Weekly Distribution of Bike Rentals by Day

3. **Usage by Month** When examining monthly usage, bike rentals tend to be lower during the winter and early spring months. However, usage begins to rise steadily from late spring, peaking during the fall. The decrease observed as winter approaches could be due to colder weather conditions. (see Figure 2.7).

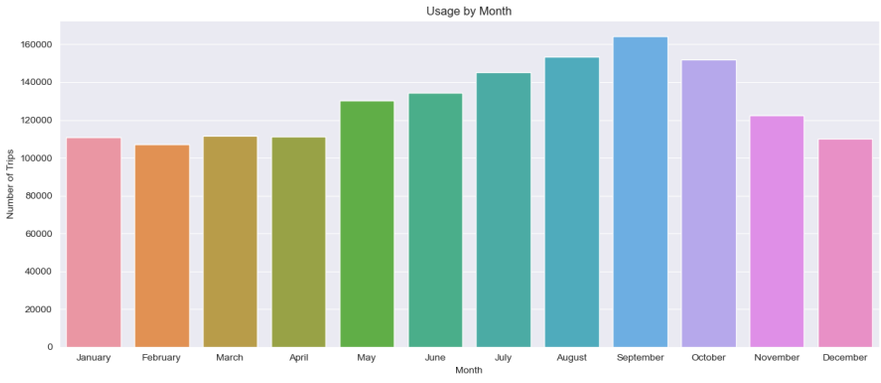


Figure 2.7: Monthly Distribution of Bike Rentals by Month

4. **Usage by Year** Bike-sharing usage has seen variations over the years. There was an increase from 2017 to 2018, a slight decline in 2019, and a significant drop in 2020 due to the COVID-19 pandemic. As conditions improved in 2021, usage slightly increased, and with the easing of restrictions in 2022, usage rose again. (see Figure 2.8).

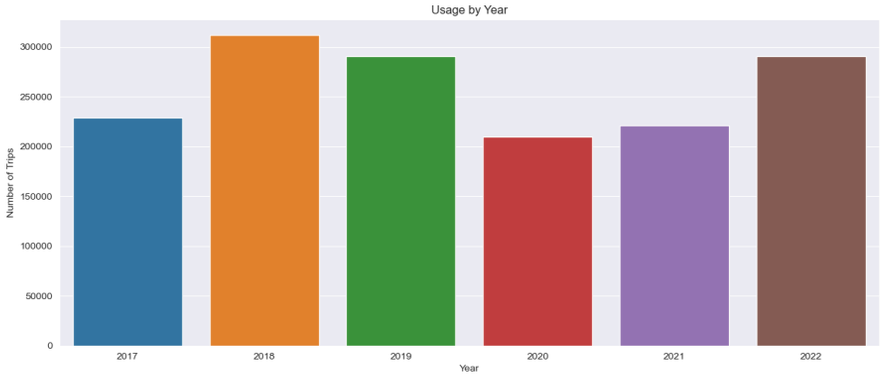


Figure 2.8: Yearly Distribution of Bike Rentals by Year

5. **Usage by Temperature** Temperature significantly impacts bike usage. Usage peaks in moderate temperatures, but as temperatures rise to higher levels, bike usage slightly decreases, possibly due to discomfort from extreme heat. (see Figure 2.9).

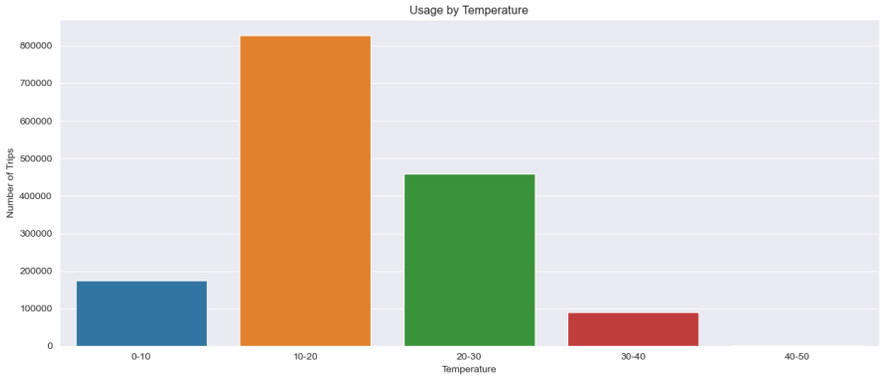


Figure 2.9: Relationship between Bike Usage and Temperature

6. **Usage by Precipitation** Bike rentals virtually cease during periods of "Very light", "Light", "Moderate", and "Heavy" precipitation as is evident in Figure 2.10. However, upon closer inspection (Figure 2.11), it's evident that the degree of precipitation, albeit insignificant on a larger scale, does impact bike rentals.

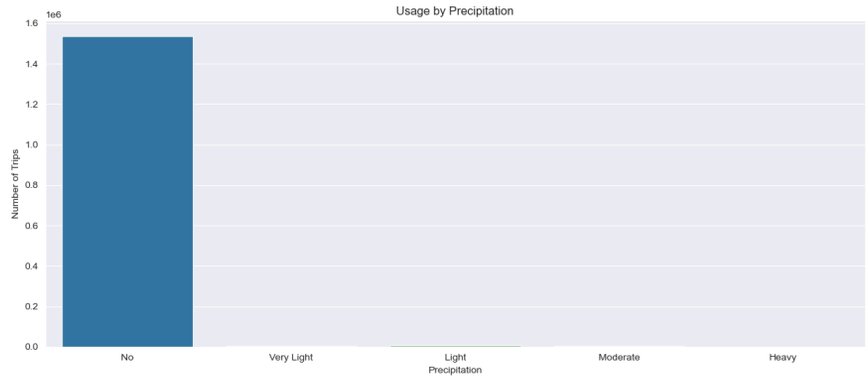


Figure 2.10: Impact of Precipitation on Bike Rentals

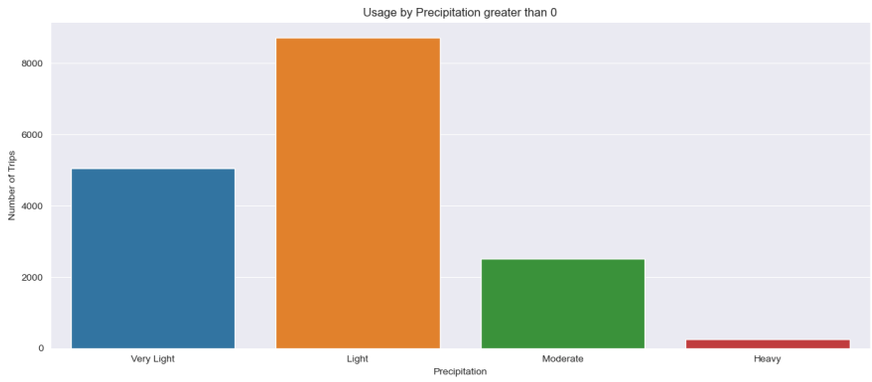


Figure 2.11: Impact of Precipitation on Bike Rentals (zoomed in on Very Light, Light, Moderate and Heavy)

7. **Usage by Cloud Cover Description** Fair weather conditions lead to the highest bike rentals, with cloudy conditions also being favorable. This data analysis provides key insights into user behavior, demonstrating the success of the bike-sharing system in Los Angeles. These insights can be used to craft

a persuasive social media campaign promoting bike-sharing and to optimize the docking station network, enhancing the overall user experience. (see Figure 2.12).

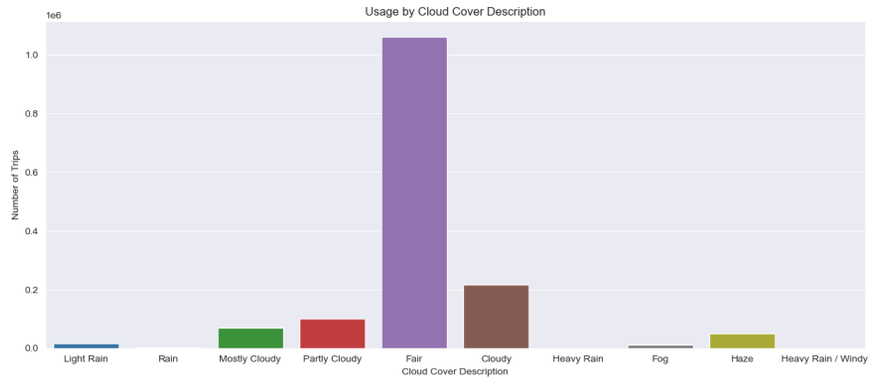


Figure 2.12: Impact of Cloud Cover on Bike Rentals with "Fair"

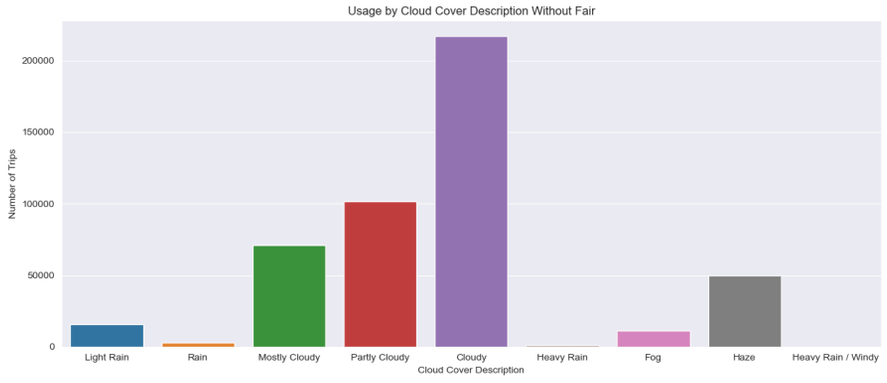


Figure 2.13: Impact of Cloud Cover on Bike Rentals without "Fair".

2.5.2 Predictive Analytics

Feature Engineering

During our feature engineering process, we aimed to create a robust set of features that would exhibit a correlation with our target variable. We carefully considered various aspects to achieve this goal.

Firstly, we explored using numerous features to enhance the accuracy of our model. However, upon closer examination, we identified certain features that lacked meaningful relevance for our specific task. These included features like "start_time," "end_time," "bike_id," "start_station_id," "start_station_lat," and "start_station_lon."

To streamline the feature set and improve model performance, we made some informed decisions. Instead of using both "start_time" and "end_time," we opted to introduce a new feature called "end_hour," which captures the hour of drop-off. This new feature provides a concise representation of the time aspect without redundant information.

Furthermore, we found that the identity of the bike a person uses is not crucial for our prediction task. As such, we decided to exclude the "bike_id" feature from our final set.

For predictions related to idle time, it was pertinent to consider the end station's identity rather than the start station's. Consequently, we made the decision to drop the "start_station_id" feature and retain the "end_station_id" feature to better inform the model.

By carefully justifying our selection of features and making these adjustments, we expect our model's performance to improve and yield more meaningful insights for the target variable.

Model Building

Initial Selection: Based on our preliminary understanding and objectives, we initially decided on Radial Basis Function (RBF) and Decision Tree models. These models were selected due to their promising performance in handling non-linear data and flexibility in modeling decisions respectively.

Evaluation and Adjustment: However, upon testing, we observed that the RBF model took an excessive amount of time to process our large dataset.

While RBF can be a powerful tool, it often struggles with scalability for large datasets, which was a major drawback in our case. On the other hand, the Decision Tree model executed swiftly, proving to be more efficient.

Final Selection: To address the limitations of RBF, we explored other models. Through this explorative process, we found LightGBM to be a standout performer. LightGBM is a gradient boosting framework that uses tree-based algorithms and is designed to be distributed and efficient. It can handle large-sized data while maintaining high efficiency and model accuracy, hence making it a suitable alternative.

In conclusion, we selected DecisionTreeRegressor and LightGBMRegressor as our final models. While the Decision Tree model offers simplicity, interpretability, and fast execution, LightGBM brings to the table its high efficiency and excellent performance with large-sized datasets, which we believe will be beneficial for our predictive analytics task.

Model Evaluation

We evaluated the Decision Tree and LightGBM models individually, using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) as performance metrics.

The Decision Tree model has the following values: MSE: 2712.67, MAE: 2.97, R2: 0.90. These results show that the Decision Tree model was able to explain a large amount of the variance in the data (90%), with reasonably small error magnitudes. The MSE suggests that the model has made a few large errors in its predictions.

The LightGBM model performed better in terms of MSE (2231.54) and R2 (0.917), but had a higher MAE (4.02) than the Decision Tree. This indicates that while the LightGBM model made predictions that were, on average, closer to the actual values, it may have made more small errors.

Based on these results, both models show potential and perform well on the dataset. However, the LightGBM model seems to perform slightly better overall, due to its lower MSE and higher R2 score.

In the next phase, we trained a Stacking Regressor model using the best Decision Tree and LightGBM models as base models and Lasso as a meta-model. The performance of the Stacking Regressor model was then compared

to the individual base models.

The Stacking Regressor showed general improvement on all metrics on the validation set: MSE dropped to 2142.99, MAE to 3.85, and R2 increased to 0.921. These results indicate that the Stacking Regressor, by effectively combining the predictions of the base models, was able to achieve better performance than the individual models.

Based on these results, the Stacking Regressor model shows promise and seems to have successfully leveraged the strengths of the individual models to make more accurate predictions.

Finally, we evaluated the performance of the Stacking Regressor on the test set. The Stacking Regressor maintained its strong performance, with an MSE of 2067.94, MAE of 3.94, and R2 of 0.920. These results confirm that the Stacking Regressor is not only effective at making accurate predictions, but is also able to generalize well to unseen data.

Based on this evaluation, the Stacking Regressor demonstrates robust performance and good generalization, making it a strong candidate for deployment.

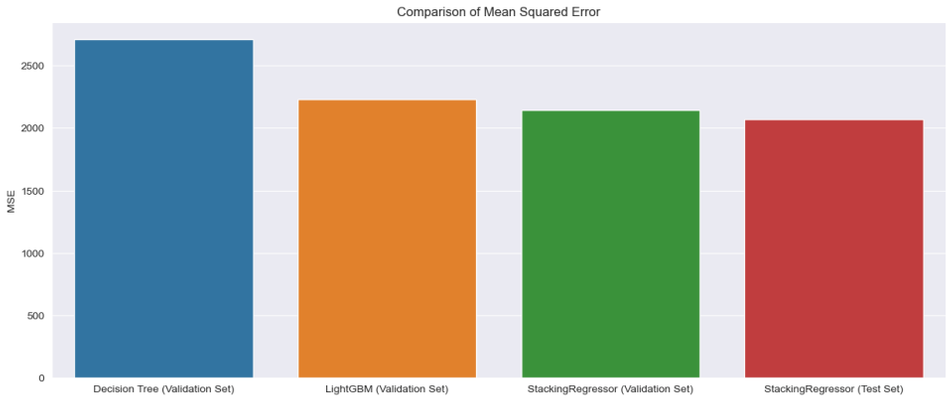


Figure 2.14: Mean Squarred Error

Figure 2.15: Mean Absolute Error

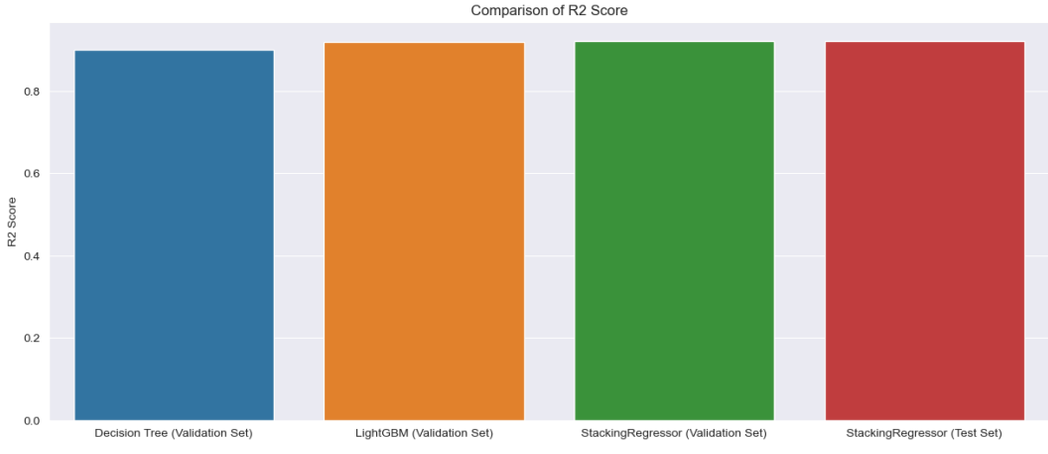


Figure 2.16: R2-Score

2.6 Conclusions

The proposed business recommendations are based on a thorough data analysis of the bicycle rental system in Los Angeles. They offer numerous advantages to optimize the system and promote sustainable urban mobility. At the same time, however, there are some limitations that should be considered during implementation.

2.6.1 Benefits of the business recommendation

1. Efficient station network By optimizing the station network, bike stations can be targeted for placement to increase utilization while reducing idle time. This improves the efficiency of the system and better meets the needs of users.

2. Targeted marketing campaigns By identifying peak times for bicycle use, targeted marketing campaigns can be implemented to increase demand at those times. This helps to spread usage evenly throughout the day and avoid bottlenecks.

3. Adjusting for weather conditions Accommodating weather conditions allows the bicycle rental system to respond flexibly to inclement weather. Appropriate incentives or offers can encourage users to use despite adverse conditions.

2.6.2 Limitations of the business recommendation

1. Dependence on external factors Business recommendations are based on historical data and weather conditions. Future changes, such as unforeseen weather events or changing user behavior, may affect the effectiveness of the recommendations.

2. Implementation Costs Implementing the recommendations may require investments in the bike-share system infrastructure, adoption of new technologies, and marketing campaigns. These costs must be carefully weighed to ensure that the benefits justify the investments.

3. Data accuracy and privacy The quality and accuracy of the data used are critical to the effectiveness of the recommendations. At the same time, user privacy must be protected as personal data is included in the analysis.

Overall, the proposed business recommendations offer a promising opportunity to optimize the bicycle rental system in Los Angeles and promote sustainable urban mobility. However, it is important to consider the limitations and challenges, and to continually review and adjust the recommendations to ensure the long-term success of the system. With careful implementation and continuous improvements, the bicycle rental system can become an important component of an environmentally friendly and efficient urban mobility strategy.

Chapter 3

References

- Seaborn: http://seaborn.pydata.org
- Python: https://python-visualization.github.io/folium/modules.html
- Numpy: https://numpy.org/doc/1.25/user/index.html
- Pandas: https://pandas.pydata.org/docs/
- LightGBM: https://lightgbm.readthedocs.io/en/stable/index.html
- Geopy: https://geopy.readthedocs.io/en/stable/
- Scikit: https://scikit-learn.org/stable/
- LightGBM Research: https://www.geeksforgeeks.org/lightgbm-light-gradient-boo
- Miscellaneous: The entire material from the lectures and the workshops.
- GitHub Repository: https://github.com/Bjufen/DSML_Team01.git

Chapter 4

Supplementary document

4.1 Protocol

LaTex	Ivan
Cover Sheet	Ivan
Summary	Ivan, Samuel, Kemal
Problem Description	Samuel
Business Goal	Samuel
Data Science Goal	Samuel
Data Description	Assia
Bike Sharing Data	Assia
Impact of Weather on Bike Sharing	Assia
Benefits of using idle time at stations	Assia
Reasons for not using individual idle times	Assia
Brief Data Preparation	Assia
Descriptive Analytics (Station-Level-Insights)	Assia
Descriptive Analytics (Overall System Performance)	Ivan
Predictive Analytics (Discussion)	Kemal
Predictive Analytics (Selection of Regression Models)	Mohammed
Predictive Analytics (Feature Engineering)	Mohammed
Model Building	Yusuf
Model Evaluation	Yusuf
Conclusion	Kemal, Ivan, Samuel
Protokollieren (Abläufe, Aufgabenverteilung etc.)	Samuel

4.2 Code

Data Clean Up (Metro)	Yusuf, Mohammed
Data Clean Up (Weather)	Yusuf
Data Clean Up (Points of Interest)	Kemal
Idle Time	Mohammed, Assia, Yusuf
Day Times and Weekdays	Mohammed, Kemal
Visualization (Station Level Insights)	Mohammed, Kemal, Assia
Visualization (Overall System Performance)	Yusuf, Samuel, Ivan
Merge	Mohammed, Yusuf
Feature Engineering	Mohammed
Model Building	Yusuf
Model Evaluation	Yusuf
For further details	See GitHub Repository (See here)