Machine Learning 1 Project: BikeShare

Group Members: Agnes Nguenda, Emmanuel Asong, Nigel Nyajeka, Meakin Marange

1. Business Case

Executive Summary:

This report analyzes Bikeshare's operations at two stations in Washington D.C., '21st & I St NW' and '21st St & Pennsylvania Ave NW', with the goal of optimizing bike and dock allocations to meet daily demand. By applying machine learning techniques, we aim to provide strategic insights and operational recommendations to improve the efficiency of Bikeshare's services.

Introduction:

Bike-sharing services have gained popularity in recent years, offering Washingtonians a convenient and eco-friendly transportation alternative on the premise of traffic problems and rising fuel prices. To optimize its services, Bikeshare needs to allocate resources and meet customer demand, such as bikes and docks, effectively. This report investigates the use of machine learning models to predict daily demand at two stations and provides insights to make informed operational decisions. The ultimate goal is to leverage machine learning recommendations to allow customers to find a bike when they need one and also get a free dock.

Five models were trained and evaluated: Linear Regression, Lasso, KNN, Ridge Regression, and Elastic Net. Cross-validation was used to tune the hyperparameters. Model performance was assessed using comparing the Mean Squared Error (MSE) to determine the most efficient model to use.

2. Data collection

We combined historical data for both stations from January to April 2022, including bike pickups, drop-offs, which was derived from the number of bicycles that were used from each trip to and from a particular station. Weather data, such as temperature, humidity, and precipitation, were also incorporated and downloaded from Bikeshare's website. Features such as day of the week, time of day, and month were considered to capture temporal patterns in demand. The dataset was split into training and testing sets.

3. Data preparation

The data preparation phase is critical for ensuring the quality and consistency of the input data for the predictive models. In this section, we describe the steps taken to clean, preprocess, and transform the collected datasets for the optimization of bike and dock allocation at Bikeshare's two selected stations.

- ***Merging Data***

We began by combining the historical bike pickups and drop-offs data for both '21st & I St NW' and '21st St & Pennsylvania Ave NW' stations. The merged dataset includes information on the number of bikes picked up and dropped off. Next, we integrated the weather data with the bike usage dataset. The weather data contains information on temperature, humidity, precipitation, and other relevant conditions.

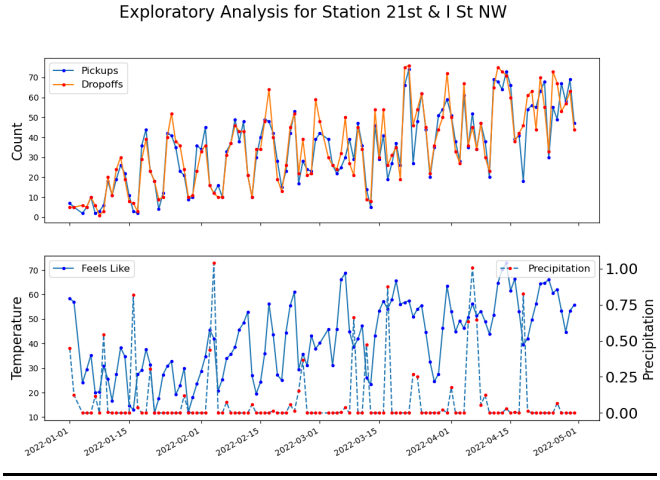
- ***Data Cleaning***

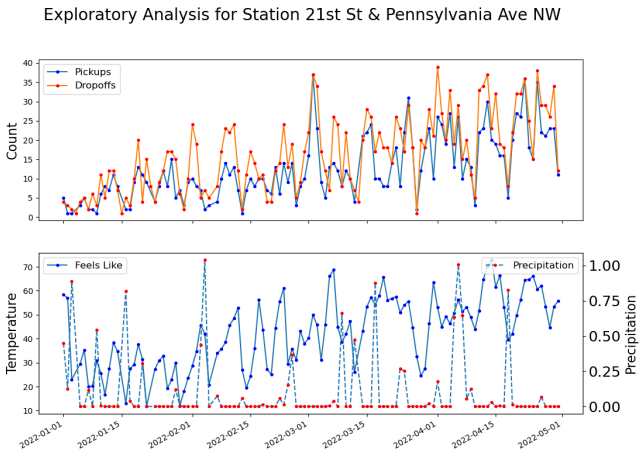
Data cleaning is vital for improving the quality of the input data and ensuring accurate predictions. We checked the weather dataset for missing values. In cases where missing values were found, we removed the affected records when necessary.

4. Assumptions

- Demand and Supply are aggregated at a daily level.
- Picking up and dropping off bikes is assumed to happen instantaneously and simultaneously on any given day.
- The supply of bikes is infinite to meet demand when rebalancing bikes is available.
- Only 2 locations were considered, 21st & I St NW' and '21st St & Pennsylvania Ave NW.
- Rebalancing bikes can only add a maximum to the maximum docks available at a station.

5. Exploratory analysis





1. **Docking Set up:**
    a. The realtime data from Bikeshare observed distinct differences in the available docks for both 21st and Ist and 21st and Pennyslvania Ave. Generally, the available docks at 21st and Ist (35 Docks) was higher, as compared to 21st and Pennyslvania Ave (19 Docks) likely due to the increased demand at the station.
2. **Weather Impact:**
    a. **Temperature:** We found a positive correlation between temperature and bike usage, with demand increasing as temperatures rose. This trend is disrupted by precipitation.
    b. **Precipitation:** There was a negative correlation between precipitation and bike usage, indicating that rainy or snowy days typically result in lower demand for bikes.
3. **Station-specific Patterns:**
    a. **Differences in Demand:** We observed varying levels of demand between the two stations, which may be attributable to differences in the surrounding areas, such as the presence of commercial buildings, residential areas, or popular destinations. Generally, 21st and Ist was observed to be a high demand spot with more pickups and drop offs as compared to 21st and Pennsylvania Ave.

b. **Imbalances in Pickups and Drop-offs:** An imbalance between the number of pickups and drop-offs at each station was noted. This finding indicates the need for a proactive approach to redistributing bikes and docks to ensure optimal service levels.

6. Model Performance evaluation

Table 1: Model Comparison

	MSE *Lowest = green*				**Out of Model Accuracy**	**Number of Features**	**Interpretability**
Station	**21st st & I ave**	**21st st & I ave**	**21st & Penn Ave**	**21st & Penn Ave**	*BEST=1 POOR =5*	*Less features desirable*	*Easy desirable*
Activity	**Pick up**	**Drop off**	**Pick up**	**Drop off**			
Linear Regression	213.572	233.961	47.308	55.297	4	26	EASY
Lasso	143.928	133.142	26.123	44.019	1 (BEST)	10	MODERATE
Ridge	146.89	130.41	31.147	45.02	3	26	MODERATE
Elastic Net	144.573	129.268	28.391	44.231	2	19	MODERATE
KNN	267.795	237.325	44.659	60.20	5 (POOR)	26	MODERATE

7. Model Selection

LASSO model was chosen as the preferred model. Our analysis revealed that LASSO outperformed the other models when evaluated by the Mean Squared Error(MSE) as shown in table 1 above. 3 out 4 MSEs from the LASSO model were found to be the lowest of the 5 models which were used and yielded the best performance on the validation set. Furthermore, LASSO had the least number of features after hyper

parameter tuning, which reduces the cost of running the model. It's interpretability was considered to be more difficult than linear regression but still understandable.

8. Predictive modeling

1. Daily Resource Allocation

Table 2: Predicted number of bikes for pick up and drop off

	Predicted			
Station	**21st st & I ave** 35 Dock Installed		**21st & Penn Ave** 19 Docks Installed	
Activity	Pick up	Drop off	Pick up	Drop off
Lasso Model	33	55	13	17
Rebalancing	22		4	

After analyzing a set of features from the test dataset, the Lasso Regression model predicted the number of pickups and drop-offs for the two stations, as displayed in Table 2.

9. Decision-making

On the given day, the '21st & I St NW' station is predicted to have 33 pickups and 55 drop-offs, resulting in a surplus of 22 bikes. As this station has 35 docks, there is no immediate need to rebalance the bikes. However, it is essential to note that this assumption is based on the premise that pickups and drop-offs occur instantaneously and simultaneously. Further analysis is required to balance bikes throughout the day effectively. Similarly, the '21st St & Pennsylvania Ave NW' station is also expected to have more drop-offs than pickups. Therefore, the same logic described above applies to this station.

In cases where the number of pickups exceeds drop-offs, Bikeshare must supply extra bikes from storage or other stations experiencing low demand. This action ensures that customers have access to bikes when needed. If the difference between drop-offs and pickups exceeds the combined available docks at the two stations, Bikeshare should consider installing additional docks. Doing so will accommodate the increased demand and prevent overcrowding at the stations.

10. Recommendations,

- Implement a system that automatically recommends the allocation of bikes and docks based on the model's real-time predictions.
- Monitor and update the model regularly to maintain its accuracy and adapt to changing demand patterns.
- Plan bike maintenance in months with high precipitation because that is when bikeshare has low demand.

- Major rebalancing may be needed in busy months, i.e. no precipitation and high-temperature months
- Evaluate station performance - Regularly assess the performance of each station based on the actual demand and compare it with the model's predictions. If a station consistently underperforms or faces low demand, consider relocating or closing it down to optimize resources.
- Use App data to predict demand and not actual bike pick-ups because demand may be missed when docks are empty.

11. Insights and limitations

Insights:

- The chosen model effectively predicts bike demand at the two stations.
- Weather conditions significantly impact bike demand.
- Resource allocation can be optimized by using the model's predictions to meet fluctuating demand.

Limitations:

- The model's performance may be limited by the quality and quantity of available data.
- The analysis was limited to daily pickups but it would be critical to understanding the hourly pickups from the data to have a comprehensive understanding of bikeshare's operations in an entire day.
- Unforeseen events (e.g., road closures, special events) might impact bike demand and are not accounted for in the model.
- The model may need to be retrained periodically to account for changes in patterns and trends.

12. Conclusion

This study demonstrates the potential of machine learning models in optimizing Bikeshare's operations by predicting bike demand at specific stations. After comparing the performance of various machine learning models, we found that Lasso Regression yielded the most accurate and reliable predictions for bike demand at the two stations: '21st & I St NW' and '21st St & Pennsylvania Ave NW'. In this report, we outlined some dynamic and impactful decision recommendations based on the insights gained from the Lasso Regression model to optimize bike and dock allocation for Bikeshare's operations. By leveraging data and the selected model, Bikeshare can make data-driven decisions to allocate resources more efficiently and improve its overall service quality.

13. Appendix

	Predicted			
Station	21st st & I ave	21st st & I ave	21st & Penn Ave	21st & Penn Ave
Activity	Pick up	Drop off	Pick up	Drop off
Linear Regression	50	51	26	15
Lasso	33	55	13	17
Ridge	33	18	13	
Elastic Net	33	35	13	17
KNN	38	42	7	21