Capital Bikeshare Case Analysis

Zecheng Hu

Akshay Thirumal

Nilgun Aslanova

Department of Decision Science

George Washington University

Washington DC 20052

Author contributions: **Zecheng** wrote the introduction,conducted and wrote exploratory analysis, and trained, evaluated, and wrote the ML models for linear regression and Ridge regression.Zecheng also wrote the decision performance and the conclusion part. **Akshay** trained, evaluated, and wrote the ML models for LASSO, Elastic Net, and KNN regressor, he also evaluated and wrote the prediction performance. Akshay also performed the decision performance evaluation and picked the best models with respect to the different strategies.

I. Introduction

In the bicycle-sharing system such as Capital BikeShare, the availability of shared bikes and docking stations plays a pivotal role in ensuring smooth, efficient commuting experiences. However, maintaining an optimal balance between bike and dock availability poses a significant operational challenge for service providers. To address this challenge, utilizing analytics and predictive modeling techniques can offer valuable insights into the dynamic nature of bike usage patterns, facilitating informed decision-making in daily bike rebalancing efforts.

This paper delves into the utilization of various regression models to tackle the operational challenge of ensuring consistent bike and dock availability on a daily basis. By employing techniques such as linear regression, ridge regression, LASSO, Elastic Net, and KNN regressor, we aim to harness the power of data analytics to forecast demand, anticipate supply needs, and optimize the distribution of bikes across docking stations.

In this study, a series of weather data spanning from February of 2023 to June of 2023 was used. In combination with bike share system data from the Foggy Bottom location at the same time frame, we wanted to investigate the relationship between bike movement (pickup and dropoff) at Foggy Bottom and various weather conditions such as temperature and wind speed.

Furthermore, this study incorporates cross-validation methodologies to evaluate the performance of each regression model. Through comparative analysis and hyperparameter tuning, we seek to identify the most effective model for accurately predicting bike and dock availability in real-world scenarios.

This paper aims to train a machine learning model that best captures the complexities of bike usage dynamics, ultimately offering actionable insights to improve the operational efficiency of bike-sharing systems. By reporting the final model's performance using test data, we want to create a final model that would minimize the cost of unsuccessful dropoff and pickup. We will also create a final model that would maximize quality of service.

The paper consists of seven sections which are organized as follows. Section II presents the ETL process and exploratory analysis. Section III trains and evaluates multiple prediction models. Section IV conducts prediction performance evaluation and trains a final model Section V analyzes and evaluates the decision strategy. Finally, section VI summarizes the findings and provides some recommendations and concludes.

II. ETL process and exploratory analysis

This research utilized weather data collected from February 2023 to June 2023, alongside bike share system data from the Foggy Bottom location during the same period. The objective was to explore the correlation between bike movement (both pickups and dropoffs) at Foggy Bottom and different weather parameters, such as temperature and wind speed. Utilizing Python, we have extracted and merged daily Bikeshare system data with the corresponding Weather data to create the data frame we needed for our prediction models.

We picked 8 variables to show the relationship between weather and Capital Bikeshare pickup and dropoff count. The variables are temperature, feels-like, dew, windspeed, visibility, UV index, precipitation, and the icon of the weather. These 8 variables are based on our own experiences and intuitions. We think that they would be the most related to whether or not people would use Capital Bikeshare. Both temperature and feels like would influence a biker's decision because if the weather is too

cold or too hot, people are unlikely to bike outside. The dew also would make people feel uncomfortable outside if it's too high or too low. The same logic would apply to wind speed and visibility since biking at high wind speed and low visibility would have safety concerns. If an area is more likely to rain, people would be less likely to bike outside. We also chose the icon because we think that most people would only look at the weather icon to determine if it is good weather to go outside.

Both the scatter plots (See Appendix for all scatterplots) against the pickup count and the dropoff count showed similar trends. We can see from the scatter plot that the higher the temperature and feels like temperature, the higher the pickup and dropoff counts are, which means that people are utilizing the Capital Bikeshare system. There are also positive relationships between dew point and pickup and dropoff counts. For the UV index, there is a positive trend. The underlying cause could be that the higher the UV index, the sunnier it is and bike tends to bike when there is good weather. The relationship between visibility and pickup and dropoff counts seems unclear as the majority of the data is between 8 to 10. The same thing would be applied to wind speed, precipitation, and icons as there is not a clear trend.

III. Predictive modeling (All regression results are in the code)

A. Linear regression

First, we wanted to conduct a linear regression model to hopefully capture the effect of weather parameters on bike movement at Foggy Bottom. The best linear regression model is the model with the lowest testing MSE. To find out which coefficients are best for predicting bike system data, we utilized a for loop to conduct multiple linear regressions to find out which one has the lowest test MSE. Each time a for loop is run, one additional independent variable is added to the regression model.

When we conduct the multiple linear regression, there are a few assumptions that we will make. First, we assume that the relationship between the target variable and independent variables is linear. We can see from the scatter plot that they seem to be linear. Second, we would assume that the error term has a constant variance (homoscedasticity). Third, we would assume that the underlying residuals are normally distributed. Last, we would assume that all independent variables are not correlated with each other. Due to the difficulty of isolating each weather attribute to be independent of each other, we would have to assume that it is true to conduct a multiple regression model.

The best linear regression model is the model with the lowest testing MSE. As we can see from the graph (See Appendix), for pickup, the best prediction model is with 5 independent variables when the testing MSE was the lowest. The variables are temperature, feels-like, precipitation, dew point, and wind speed. As for dropoff, the best prediction model has 4 variables, they are temperature, feels-like, precipitation, dew point. The coefficients from both regression models are as follows. And the final regression equation as shown below.

$$pickup\ count = \beta_1 * temp + \beta_2 * feelslike + \beta_3 * dew + \beta_4 * precip + \beta_5 * windspeed$$

$$dropoff\ count = \beta_1 * temp + \beta_2 * feelslike + \beta_3 * dew + \beta_4 * precip$$

B. Ridge regression

As mentioned in the previous section, we have a few assumptions for the linear regression that we don’t know if it's violated or not. Therefore, we have decided to use ridge regression to deal with issues such as multicollinearity, overfitting, or high variance. Regularization enhances the conditioning of the problem and decreases the variance of the estimates. Larger values indicate more regularization. Ridge

regression effectively handles multicollinearity, where predictors are highly correlated. By shrinking the coefficients, it mitigates the problem of unstable estimates caused by multicollinearity. We will also be performing hyperparameter tuning to find the best alpha for our ridge regression. By utilizing the RidgeCV() function in python, we have automated the regression to find the best alpha with the lowest MSE. The MSE for pickup data is 63.42 and 68.44 for dropoff, with an r-squared value of 0.4431 and 0.3805 respectively. Which indicates that the Ridge model is a moderate fit.

C. LASSO

Lasso regression employs L1 regularization. This technique not only helps in reducing overfitting but also performs feature selection by driving some coefficients to zero, hence simplifying the model if it has unnecessary complexity. We identified optimal coefficients for predictors such as temperature, feels like,, dew point, precipitation, and wind speed. These coefficients were determined using cross-validation to find the best alpha parameter, which balances model complexity and prediction accuracy.

The Mean Squared Error (MSE) and R-squared values calculated from the test data provide a measure of the model's prediction error and the proportion of variance explained by the model, respectively. The R-squared values suggest that the model explains approximately 44.16% & 38.01% of the variability in the dependent variable for pickup and dropoff respectively, along with Mean Squared Error (MSE) value of 63.59 & 68.49 indicate a moderate fit to the data.
Lasso regression, which inherently incorporates feature selection, the model avoids overfitting and becomes more interpretable.

D. Elastic Net

Elastic Net combines L1 and L2 penalties from Lasso and Ridge methods. This approach helps in addressing regression problems where predictors are correlated, thus assisting in feature selection and preventing overfitting, which can be common in traditional least squares models especially when predictors outnumber samples.

ElasticNetCV is used to determine the best alpha parameter via cross-validation, which ensures that the model's complexity is appropriately penalized. The Mean Squared Error (MSE) and R-squared statistics are then calculated to evaluate the model's predictive performance. With an MSE of approximately 65.62 for pickup and 69.65 for dropoff along with R-squared values of around 0.42 & 0.36 respectively, the model indicates a moderate fit to the test data.. Elastic Net is particularly useful in situations where some predictors are expected to have a small or zero effect on the outcome, thereby streamlining the model to focus on the most influential factors.

E. KNN regressor

` The K-Nearest Neighbors (KNN) algorithm predicts the value of a new data point based on the values of the K nearest points in the training data. Choosing the best K involves balancing bias and variance. A small value of K means that noise will have a higher influence on the result, and a large value makes the algorithm computationally expensive and may lead to over-smoothing, where the model fails to capture the underlying structure of the data.

To find the optimal K, we used cross-validation. Plotting the cross-validated MSE against different K values. The K at which the MSE is minimized is typically chosen as the optimal value.
The MSE for pickup data is 87.02 and 88.63 for dropoff, with an r-squared value of 0.23 and 0.19 respectively. Which indicates that the KNN model is not essentially a good fit.

IV. Prediction performance

Model	Pickup MSE	Pickup R2	Dropoff MSE	Dropoff R2
Linear	71.9816	0.274	72.1160	0.307
Ridge	63.4259	0.4431	68.4437	0.3805
Lasso	63.5934	0.4416	68.4906	0.3801
ElasticNet	65.6210	0.4238	69.6593	0.3695
KNN	87.0260	0.2178	88.6310	0.2021

After comparing the final model performances using the Test data, Ridge regression and Lasso models appear to be the best for both pickup and dropoff predictions. They have the lowest Mean Squared Error (MSE) and the highest R-squared (R2) values among the models for the pickup and dropoff data, indicating it predicts with the least error and explains the most variance of the dependent variable. Therefore, considering both mean error and variance, Ridge regression is the best balance for this dataset.

V. Decision performance

Not only do we need to consider the technical performances of our models, we also need to consider how useful it is to the decision maker as well. Therefore, we decide to evaluate our models by implementing two strategies: most minimization and quality maximization. We think that by evaluating our models from different perspectives, we can provide the decision makers options to choose the best strategy for the company.

In the cost minimization scenario, we're examining the predictive capabilities of a trained model to minimize cost. By using the model to predict pickup and dropoff, we aim to minimize the total cost by strategically allocating bikes(x) and open docks(y). The total cost is calculated based on penalties for unsuccessful pickups (α) and drop-offs (β), with the constraint that total bike space is 17. We would set the pentlay for unsuccessful pickup as 2 while unsuccessful drop off as 3. We think that customers would be more likely to be displeased when they cannot drop off a bike at this location because students need to drop off the bike here to go to class. Then we would allocate bikes and open docks using our trained model to minimize the total cost under predicted pickup and drop off count. Lastly, we would use the actual pickup and drop off count to compute the actual total cost. By evaluating the out-of-sample average cost, we can determine which model is more suitable for practical application.

In the quality maximization scenario, we want to maximize our quality of service by measuring the weighted average service level. We aim to allocate bikes(x) and open docks(y) so that our service level is maximized. The weighted average service level is calculated based on the area weights for each service level for pickups (α) and drop-offs (β),with the constraint that total bike space is 17. We would set the service level to be 0.5 for both pickup and drop off because we think it’s important to prioritize both drop off and pickup service. Then we would allocate bikes and open docks using our trained model to maximize the service level under predicted pickup and drop off count. Lastly, we would use the actual pickup and drop off count to compute the actual service level. By evaluating the out-of-sample average cost, we can determine which model is more suitable for practical application.

Our results showed that **ElasticNet and Lasso** are the best models for the cost minimization strategies with a total average cost of **91.40 and 91.46** .

The best model to maximize our service level is **ElasticNet and Lasso**, it has a weighted average service level of **35.55% and 35.18%**

VI. Conclusion - Recommendation & Limitations

In conclusion, this paper tried to address the operational challenge of ensuring consistent bike and dock availability in bicycle-sharing systems through the application of data analytics and predictive modeling techniques. By utilizing a variety of regression models including linear regression, ridge regression, LASSO, Elastic Net, and KNN regressor, combined with weather data and bike share system data, we aimed to understand the complex relationship between bike movement and various weather conditions.

Through various analysis and cross-validation, we have evaluated the performance of each machine learning model, and identified the best-performing models for our respective strategies. Without considering the real-life implication, **Ridge regression** emerges as the optimal choice for predicting both pickup and dropoff counts. It has the lowest Mean Squared Error (MSE) and the highest R-squared (R2) values compared to other models. From the Ridge regression model, temperature, dew, and precipitation have a negative correlation with both pickup and dropoff count while feels like and wind speed has a positive correlation.

However, we also evaluate our final models by different company strategies. Our findings indicate that the model with the lowest total cost, which are **ElasticNet and Lasso** , outperforms others in minimizing unsuccessful pickup or dropoff cost. It has a total average cost of **91.40 and 91.46** , making them the optimal choice for cost minimization strategies. Conversely, for maximizing service level, **ElasticNet and Lasso** are the top-performing models. With a weighted average service level of **35.55% and 35.18%** , these models would allocate a bike-dock ratio to maximize service quality and ensure a satisfactory experience for all users.

Management can derive substantial benefits from this study's findings and predictive models. By utilizing the insights from the analysis, management can make data-driven decisions to optimize the allocation of bikes and docks, leading to enhanced quality of service and cost savings. The implementation of cost minimization strategies, based on accurate predictions of bike demand and distribution, can result in significant reductions in unsuccessful pickups and dropoffs. By prioritizing quality maximization strategies, management can improve the overall service level for customers, leading to higher satisfaction and retention rates. Additionally, having a better understanding of bike usage patterns and demand fluctuations enables management to anticipate and mitigate potential risks, ensuring smooth operations and fewer disruptions. Overall, the study empowers management with valuable tools and insights to drive strategic planning efforts and improve the performance and sustainability of the bicycle-sharing system.

A limitation of this paper is that we made an assumption that all pickups and dropoffs happen at the same time. However, these events happen sequentially. This analysis opens doors to further studies that optimize the data every hour based on the real-time systems status and the predicted pickups and dropoffs for the next hour.