DNSC 6315: Machine Learning 2

Assignment 3: Final Report

Assignment by
N M Emran Hussain
GWID: 24414095

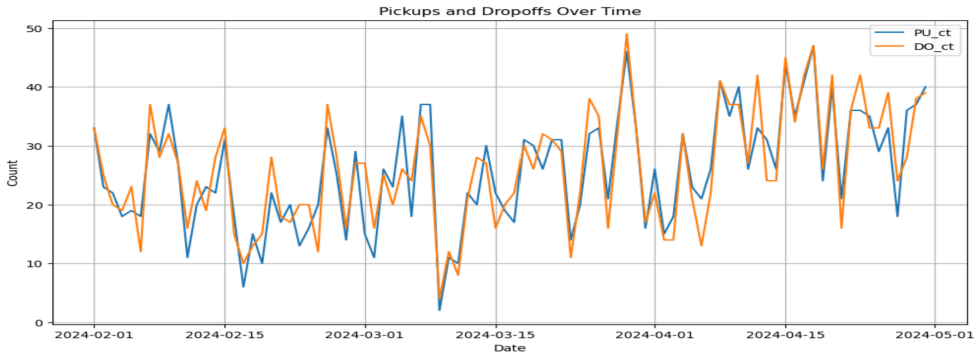
Business Understanding:

The business goal is to improve operational efficiency and user satisfaction by analyzing Capital Bikeshare ridership patterns from February to April 2024. By examining pickups and drop-offs, the aim is to identify high-demand areas, manage supply-demand imbalances, and support strategic decisions like fleet redistribution or station expansion. This insight-driven optimization reduces costs and improves service availability.

Exploratory Analysis:

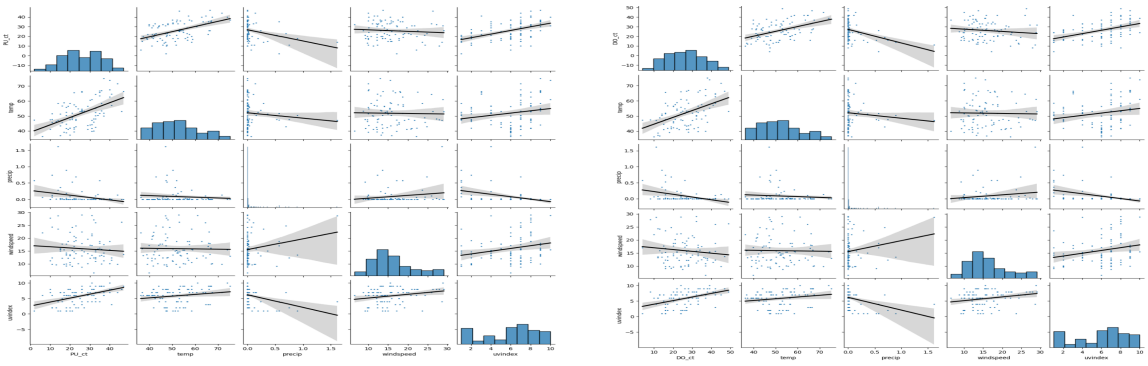
The project starts by importing and exploring bikeshare datasets, focusing on trip counts (pickups and drop-offs). Descriptive statistics and exploratory visualizations help detect station activity levels, temporal patterns, and potential predictors of high usage. The merged dataset includes features such as station location, time (hour, weekday), and usage counts, giving a detailed view of bike traffic flow. We can carry out the following key exploratory analyses to better understand Capital Bikeshare usage patterns before modeling.

- Pickups and Drop-offs over time



A line plot was created to visualize `PU_ct` (pickup count) and `DO_ct` (drop-off count) over time.

- Pairwise Relationships Between Predictors and Target: A Seaborn `pairplot` was used to explore how bike pickup counts (`PU_ct`) relate to various weather conditions like Temperature (`temp`), Precipitation (`precip`), Windspeed, UV Index.



Predictive Modeling:

In this project, predictive modeling aims to forecast the number of bike pickups available (PU_ct) at Capital Bikeshare stations for providing the service and number of drop-offs or number of available docks (DO_ct). Accurate predictive models like 'Linear Regression', 'Ridge Regression', 'LASSO', 'Elastic Net', 'KNN', 'Regression tree', 'Random Forest', 'Gradient Boosting' and 'Neural Network will help in real-time decision-making, such as bike redistribution and service expansion planning. We calculated hyperparameter tuning individually and separately suitable for each model and we will also calculate MSE for PU_ct and DO_ct separately for each of the models mentioned above and finally we will calculate Out-of-sample-cost for each model. Here MSE (Mean Square Error) of a model accurately predicts the number of available bike pickups and dock availibility. Here Out-of-sample-cost reflects the business-relevant metric based on operational performance. To capture various types of relationships in the data, a diverse set of linear, tree and non-linear models were trained and compared:

- **Linear Regression:** A baseline model assuming a straight-line relationship between features (e.g., hour of day, weather, station) and the target (bike pickups). While easy to interpret, it's often too simplistic for complex systems like bikeshare demand. Using the Linear regression model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Linear Regression	**54.8883**	**69.5465**	**76.6389**

- **Ridge, LASSO, and Elastic-Net Regression:** These are regularized linear models:Ridge (L2 penalty) shrinks coefficients to reduce overfitting. LASSO (L1 penalty) can shrink some coefficients to zero — helpful for feature selection. Elastic Net combines both penalties for balance. These models are useful when predictors are correlated or when the dataset contains many features. Using the Ridge, LASSO, and Elastic-Net Regression model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Ridge	**54.8891**	**69.5454**	**76.6389**
LASSO	**71.5580**	**79.7125**	**76.6389**
Elastic Net	**62.5541**	**72.4992**	**76.6389**

- **K-Nearest Neighbors (KNN):** KNN predicts demand at a station by averaging the outcomes of its nearest similar scenarios (based on feature distance). It’s non-parametric and flexible but can be sensitive to noise and scales poorly with data

size. Using the KNN model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
KNN	**64.9297**	**72.8519**	**76.6389**

- **Regression Tree:** A decision tree that splits data into regions based on rules learned from the features. It captures nonlinear relationships and interactions but can overfit unless pruned. Using the Decision tree model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Decision Tree	**86.3950**	**95.9265**	**76.6389**

- **Random Forest:** An ensemble of decision trees built on different random samples and subsets of features. It reduces overfitting and variance, making it highly effective for prediction. It's robust and typically one of the top performers in tabular data. Using the Random Forest model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Random Forest	**73.1887**	**85.3664**	**76.5833**

- **Gradient Boosting:** This model builds trees sequentially, each one improving upon the previous by correcting its errors. It tends to outperform Random Forest in many settings due to its fine-tuned learning, though it's more sensitive to hyperparameters. Using the Gradient Boosting model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Gradient Boosting	**92.4546**	**111.9453**	**76.4722**

- **Neural Network:** A feedforward neural network was used to model complex, nonlinear relationships. Neural nets can be powerful with large, rich datasets, but they require careful tuning and are less interpretable. Using the Neural Network model, MSE for Number of bikes available (PU_ct), Numbers of Docks available (DO_ct) and Out-of-sample-cost is given below:

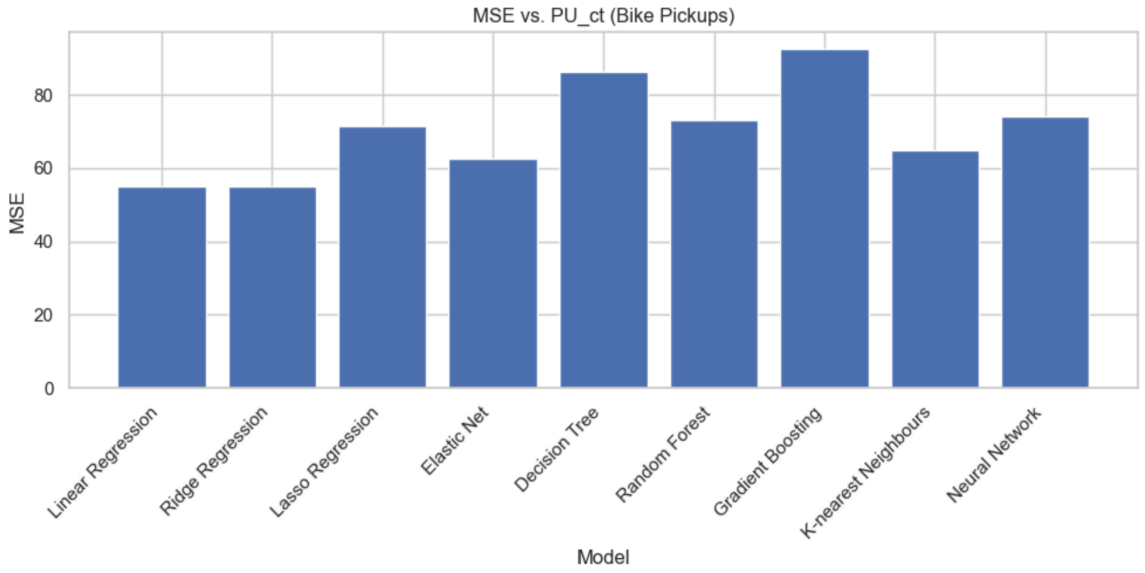
Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Neural Network	**74.2597**	**100.5584**	**76.5833**

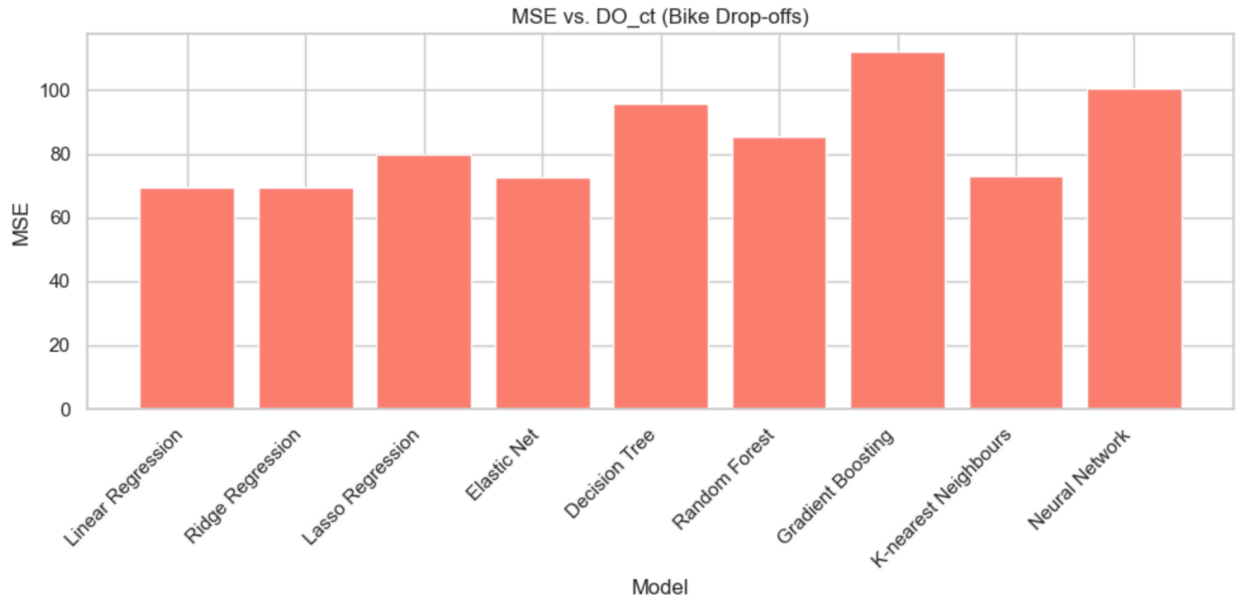
Performance Evaluation:

- **Prediction performance:** After summarizing the model performance for bike pickup (PU_ct) and drop-off (DO_ct) predictions, along with the out-of-sample decision cost for each model, the resulting table appears as follows:

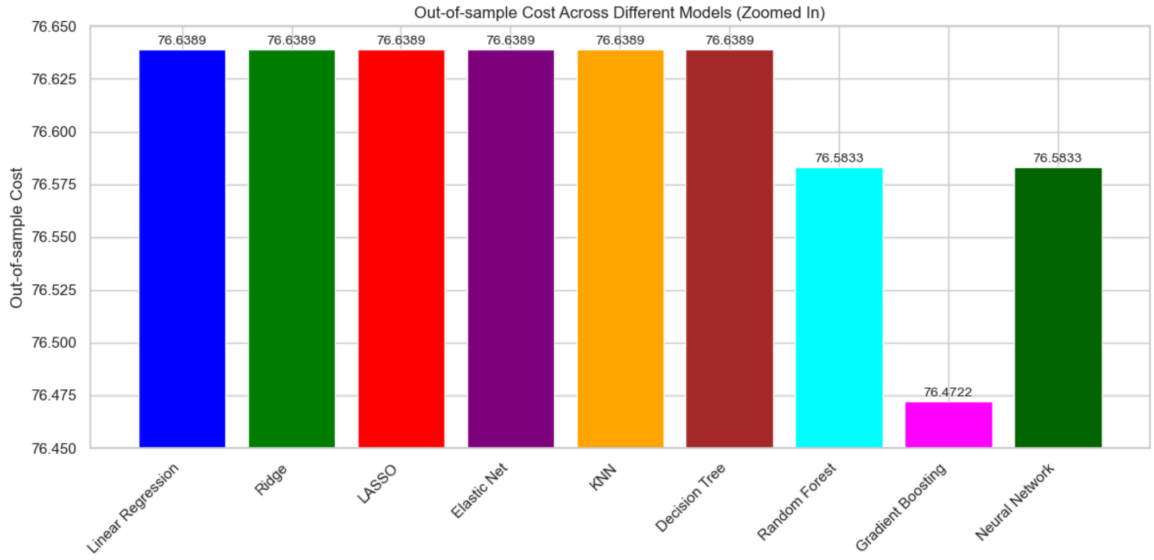
Model	MSE (PU_ct)	MSE (DO_ct)	Out-of-sample-cost
Linear Regression	54.8883	69.5465	76.6389
Ridge	54.8891	69.5454	76.6389
LASSO	71.5580	79.7125	76.6389
Elastic Net	62.5541	72.4992	76.6389
KNN	64.9297	72.8519	76.6389
Decision Tree	86.3950	95.9265	76.6389
Random Forest	73.1887	85.3664	76.5833
Gradient Boosting	92.4546	111.9453	76.4722
Neural Network	74.2597	100.5584	76.5833

- These charts given below help visually confirm the predictive accuracy of each model. However, the lower MSE doesn't always mean better operational performance — that's where your out-of-sample cost analysis comes in (see below).





- The zoomed-in plot given below that now clearly shows the small differences in out-of-sample cost across models. Each bar is uniquely colored. Values are labeled precisely. The y-axis is zoomed in (76.45–76.65) to make subtle cost differences visible. We can now clearly see that Gradient Boosting achieves the lowest cost, followed by Random Forest and Neural Network.



- **Decision Performance:** From the tables and plots given above, Linear & Ridge Regression show the lowest MSE for both PU_ct and DO_ct, indicating better numerical prediction accuracy. LASSO, Elastic Net, and KNN show higher MSEs but all models have the almost same out-of-sample-cost. This suggests the prediction errors may not translate to worse operational decisions. Decision Tree performs the worst across all metrics. Random Forest and Neural Network have higher MSEs, but produce the lowest

out-of-sample-cost (76.5833) meaning they make better allocation decisions. Gradient Boosting has the lowest out-of-sample-cost overall (76.4722) despite high MSEs, suggesting it's most effective in operational terms.

Conclusion:

This project analyzed Capital Bikeshare usage data from February to April 2024 to enhance operational decision-making through predictive modeling. The analysis revealed that while Linear and Ridge Regression achieved the lowest prediction errors (MSE) for bike pickups (PU_ct) and dock availability (DO_ct), they did not perform best in operational terms. Interestingly, Gradient Boosting, despite having higher MSEs, resulted in the lowest out-of-sample cost, indicating superior real-world decision impact. This was closely followed by Random Forest and Neural Networks, suggesting that more complex models may offer better cost-oriented outcomes even with lower predictive accuracy.

If the goal is to prioritize numerical prediction accuracy and the small differences in out-of-sample cost (within ~0.16 units) are operationally negligible, then Linear and Ridge Regression are strong choices. These models are simpler, more interpretable, and computationally efficient. However, when fine-tuned cost efficiency or real-time operational decisions are critical, ensemble models like Gradient Boosting or Random Forest may be better suited due to their slightly superior performance in decision-based outcomes. Therefore, the choice of model should depend on the business context—whether the priority is simplicity and prediction accuracy or complex, optimized allocation decisions.

This analysis has several limitations. The models may be sensitive to feature scaling and require more robust hyperparameter tuning. The use of static weather data and a limited historical window (only three months) restricts the generalizability of the results. Additionally, user behavior patterns and real-time dynamics were not incorporated. Future work should explore longer time periods, adaptive learning with real-time data, and behavioral features to build more dynamic and effective forecasting systems for bikeshare operations.