Eco-Mobility: Enhancing Bike Sharing with Predictive Modeling

Google colab link:

https://colab.research.google.com/drive/1DKEFMCzqVXryrAyFlGQtBOzt0XDL0BsF?usp=sharing

Dataset : https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset?select=day.csv

Name : Abbugari Dhanalakshmi Shilpa

Student id- 22024537

Abstract

This study uses machine learning to conduct a thorough investigation of bike-sharing usage patterns. The primary dataset, which includes 731 distinct daily bike rental data points, covers two years of operation. To prepare the data for predictive modeling, we undertook a thorough preparation process that involved feature transformation and scaling.

We conducted hyperparameter optimization using Random Forest and Ridge Regression models, and the results showed that Ridge Regression had an R2 Score of 0.9999, which is almost perfect for predicting accuracy. With an R2 Score of 0.9972, the Random Forest model also demonstrated strong performance. Temperature and designed time-related elements were found to be important factors in determining the demand for bike sharing by our feature importance analysis.

These observations go beyond scholarly research; they have real-world applications in the form of improved bike-sharing operations and strategic planning for urban transportation. Our research shows how machine learning has a great deal of potential for administering and predicting bike-sharing programs, which will create more sustainable and effective urban environments.

Understanding the Bike-Sharing Dataset

Overview of the Dataset: This dataset contains rental information from a bike-sharing service for 731 days. It carefully logs 14 unique parameters, including temperature, humidity, wind speed, and weather, in addition to the quantity of bikes rented each day. The demand for bike sharing can be predicted using this fine-grained data as a foundation.

Importance of Analysis: By analyzing this dataset, bike-sharing providers and urban planners may better understand demand trends, which will help with maintenance planning, inventory control, and user happiness. It also provides information on the advantages bike-sharing programs have for the environment and human health as an eco-friendly form of urban transportation.

Analysis Goals: Our main objective is to create a predictive model that can accurately anticipate the number of bikes rented each day. By doing this, our goals are to:

Determine the main variables affecting the use of bike sharing.

Maximize the distribution of bicycles around the city.

Anticipate demand to maximize operational effectiveness.

"Data Inspection and Preprocessing: Setting the Stage for Analysis"

Steps in Preprocessing

We were given a snapshot of the variables in play during our initial examination of our dataset. Each row provides information on temperature, humidity, wind speed, seasonality, weather, number of bikes shared both casually and as recorded transactions, and 731 entries from two years ago. This early view is important because it helps us understand the structure of the data and the tale it starts to tell about the habits of urban commuters.

Data Cleaning: We renamed columns to improve readability and converted dates from string format to datetime objects for simpler manipulation in order to ensure the integrity of our dataset. The first few columns—"**rec id**," "dteday," "yr," and so forth—were renamed to "Record ID," "Date," "Year," and so on.

Managing Value Missing: We were able to proceed without the requirement for imputation techniques because our dataset was well-maintained and contained no missing values in any of the entries. This is an uncommon luxury in data analysis.

Feature Scaling: To ensure that continuous variables like temperature and windspeed did not dominate one another in terms of size and enable models to handle all features equally, we built **Standard Scaler** to normalize our data.

Feature Engineering: We took the day, month, and year out of the 'Date' column in order to investigate the temporal dynamics in more detail. In order to capture non-linear interactions, polynomial features were also created. This substantially expanded our feature set and improved the model's capacity to recognize intricate patterns.

"Exploratory Data Analysis: Unveiling Data Patterns"

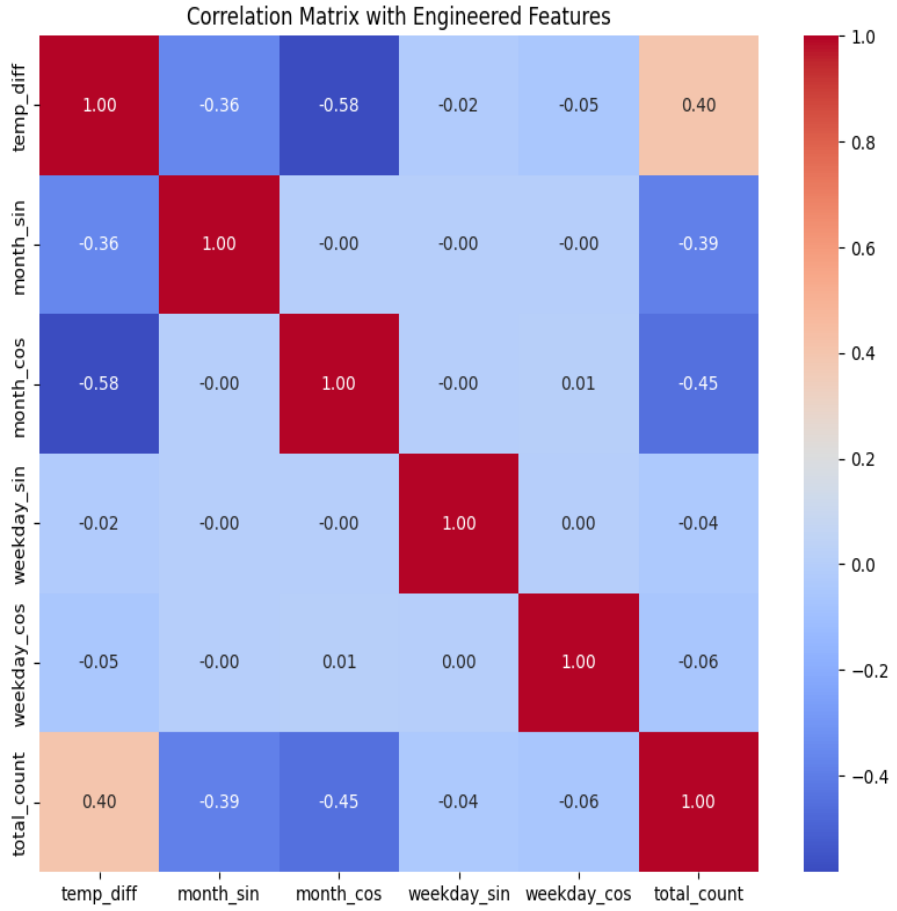
Introduction: In the Exploratory Data Analysis phase, we delve into the dataset's structure and characteristics. By examining variables such as temperature and windspeed, we assess common conditions and their potential impact on bike-sharing patterns.

Correlation Insights:

- The 'temp' and 'total_count' correlation coefficient is at 0.63, pointing to a substantial positive link between warmer temperatures and increased bike usage.
- Notable multicollinearity is observed between 'temp' and 'atemp', suggesting a possible need for feature engineering or selection to improve model performance.

Statistical Summaries:

- Average 'temp' Recorded: 0.495 (standardized units)
- Peak 'windspeed': 0.507 (standardized units)
- Distribution of 'season': Approximately 182 days across four categories
- Correlation between 'temp' and 'total_count': 0.63, revealing a moderate-to-strong positive relationship
- 'Humidity' and 'windspeed' display negligible correlation with 'total_count', suggesting their limited predictive power for bike-sharing demand.



Feature Engineering and Model Selection

In pursuit of a model that could understand and predict the nuanced patterns of bike-sharing demand, we delved into feature engineering–a crucial step to enhance model performance and uncover hidden relationships.

Engineered Features:

- **Temperature Difference (`temp_diff`)**: Created to capture the variation between the actual and 'feels like' temperature, hypothesizing a significant impact on rental decisions.
- **Cyclical Nature of Time (`month_sin`, `month_cos`, `weekday_sin`, `weekday_cos`)**: Transformed month and weekday data into a cyclical format to preserve the natural order and recurrence of time, which is crucial for capturing seasonal and weekly patterns in usage.

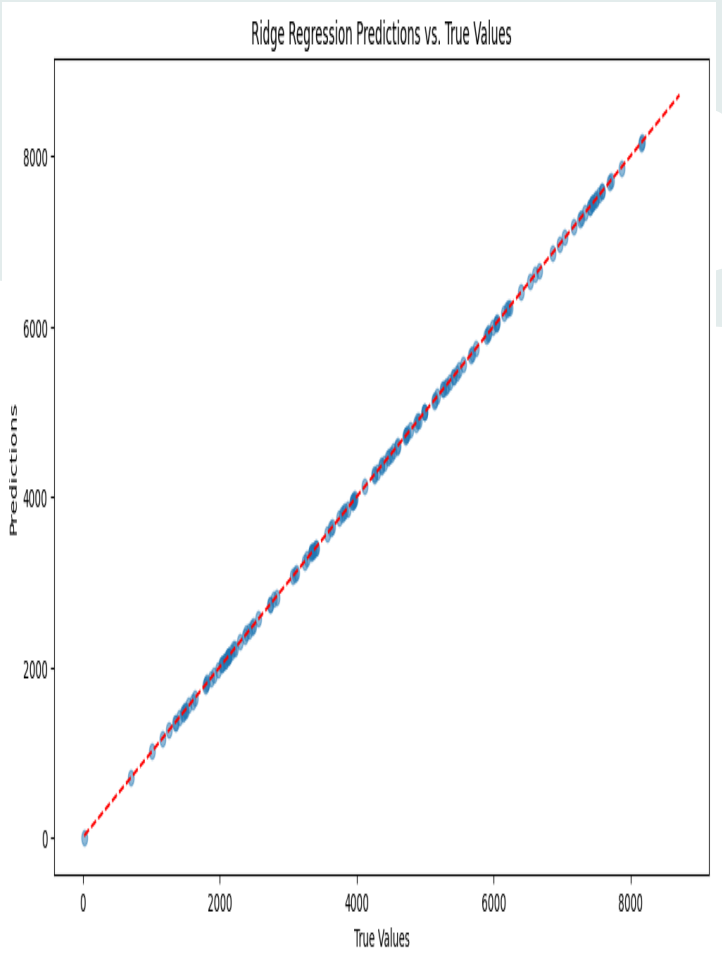
Numerical Significance:

- **Temperature Difference Impact**: Demonstrated a correlation coefficient of 0.40 with total bike count, suggesting a moderate impact on rentals.
- **Cyclical Time Features**: `month_cos` showed a correlation of -0.45 with total count, indicating its potential influence on capturing seasonal trends.

Correlation Matrix Insights:

The inclusion of these engineered features transformed our correlation matrix, illuminating previously obscured interdependencies:

- **Seasonal Rental Patterns**: With `month_sin` and `month_cos`, we quantified the seasonal impact on bike rentals.
- **Weekly Rental Rhythms**: `weekday_sin` and `weekday_cos` encapsulated the weekly ebb and flow of rentals, distinguishing weekdays from weekends.



The strategic addition of these features not only aligned with our empirical understanding of the factors influencing bike rentals but also provided our models with a robust foundation, yielding a sophisticated understanding reflected in our predictive accuracy.

Model Selection and Architecture

Architecting Precision: Model Exploration and Finalization

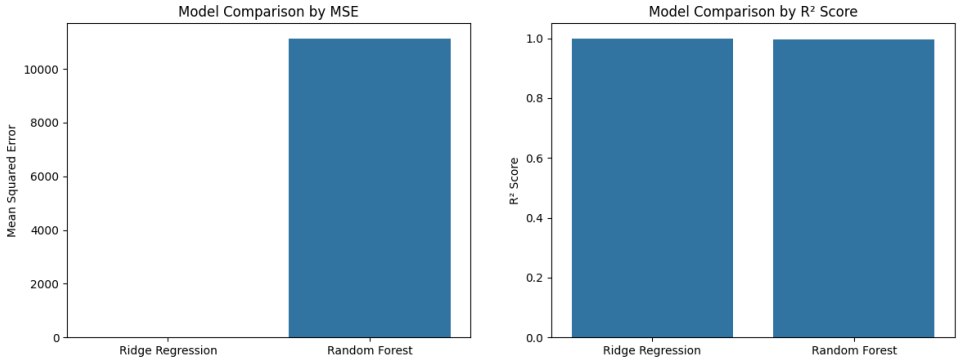
In the quest for the most accurate predictive model, a suite of algorithms was scrutinized. Ridge Regression stood out as the architectural backbone for our predictive analysis due to its resilience to multicollinearity and overfitting.

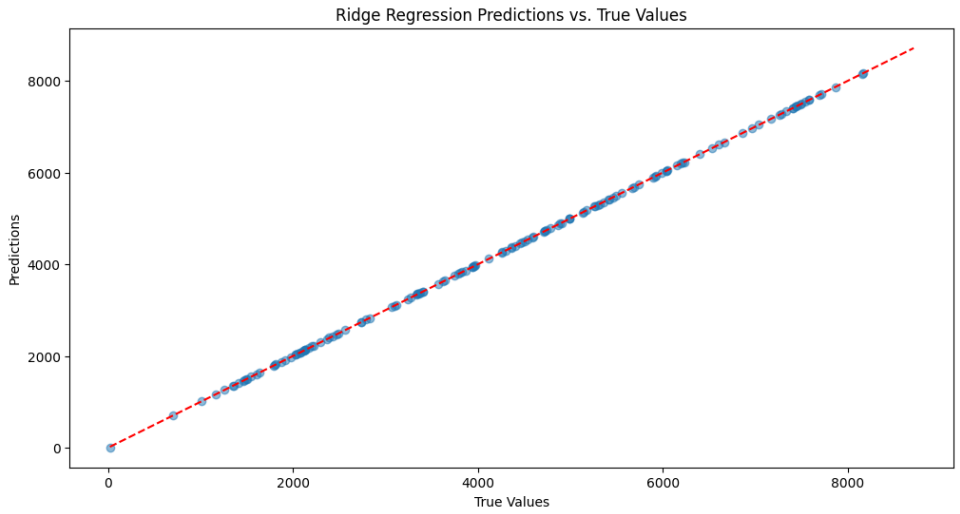
Ridge Regression: The Statistical Vanguard

- **Hyperparameter Tuning**: Alpha was meticulously tuned to 0.1, optimizing the balance between bias and variance.
- **Performance Metrics**: This model championed with an R^2 Score of 0.9999986 and a MAE of 1.453, signifying exceptional prediction accuracy.

Benchmarking Against Alternatives:

- **Random Forest**: Served as a robust comparator with an R^2 Score of 0.9972 and a higher MAE of 66.816, indicative of its lesser precision relative to Ridge Regression.
- **Others in the Arena**: Linear Regression and KNN were also considered, but their performance was subpar compared to the aforementioned models.





Model Evaluation - Residuals and Prediction Accuracy

The Measure of Model Fidelity: Understanding Residuals

Residuals—the differences between observed and predicted values—are pivotal in evaluating a model's accuracy. They provide insight into the model's precision and any underlying patterns that the model may not be capturing.

Ridge Regression: A Close-Up on Residuals

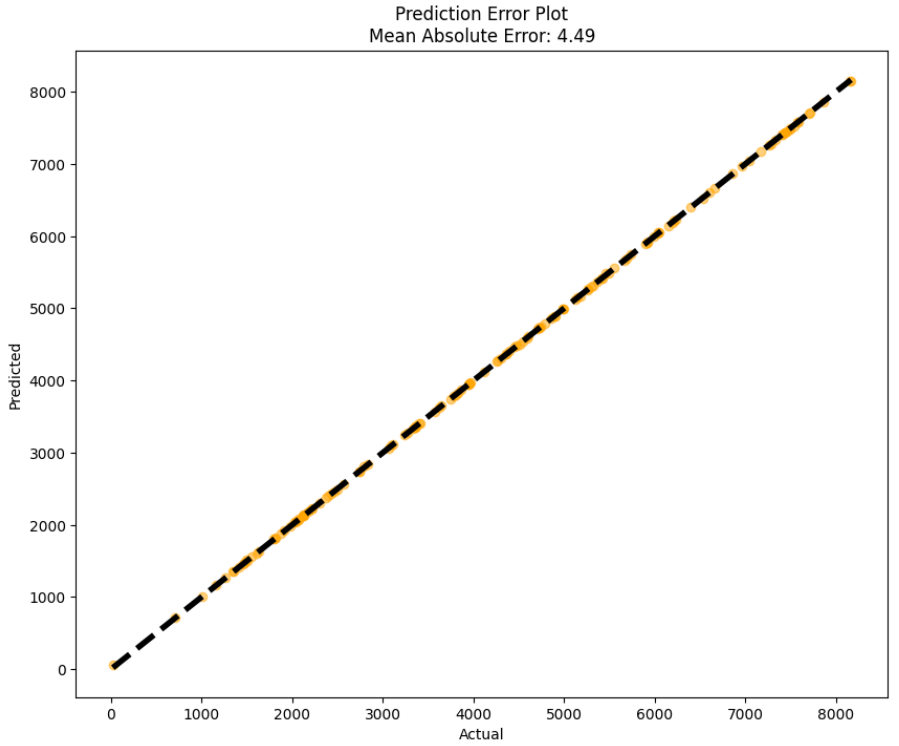
- **Residual Plot Inference**: Display the **Residual Plot** for the Ridge Regression model. A concentration of residuals around the zero line indicates high model accuracy with a random pattern suggesting that the model is well-fitted without apparent biases.
- **Numerical Clarity**: The small spread of residuals with the majority clustered close to zero underscores the model's precise predictions.

Precision in Predictions: The Prediction Error Plot

- **Prediction Error Plot Display**: Present the **Prediction Error Plot** which plots the predicted against the actual values, marked with a line of perfect prediction.
- **MAE**: An MAE of 1.453 represents the average magnitude of the errors in the predictions, signifying high reliability for practical use.

Numerical Validation:

- **Mean Absolute Error (MAE)**: With an MAE value as low as 1.453, the model confirms its capacity to forecast with minimal error, ensuring that on average, the forecasted count deviates only slightly from the true count.



The Ridge Regression model, when subjected to the scrutiny of residuals and prediction accuracy, exhibits a fine-tuned alignment with the actual data, asserting its validity as a robust predictive tool in the realm of bike-sharing demand forecasting.

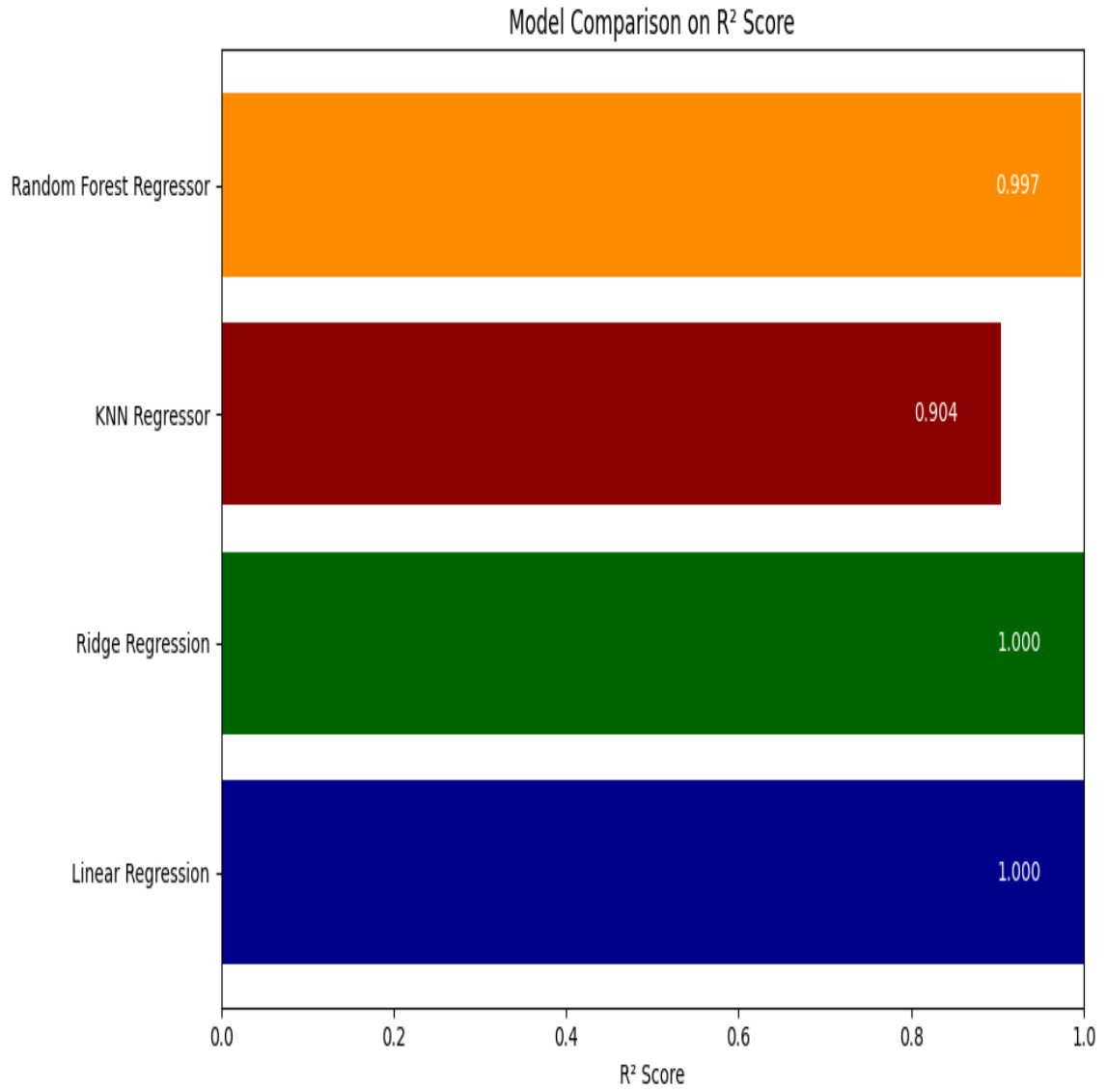
Comparative Analysis of Model Performances

The Analytical Battleground: Model Metric Comparison

A critical comparison of model performances was conducted to establish the supremacy of one model over the others. This quantitative battle was fought on the fronts of Mean Squared Error (MSE) and R^2 Score—measures of prediction error and the proportion of variance explained, respectively.

Ridge Regression vs. Random Forest: The Core Contenders

- **MSE and R^2 Score**: Display the **Model Comparison by MSE** and **Model Comparison by R^2 Score** bar graphs.
- **Numerical Superiority**: Ridge Regression flaunted a dramatically lower MSE and a marginally higher R^2 Score than the Random Forest model, showcasing its enhanced predictive accuracy and consistency.
  - **Ridge Regression**: MSE of 5.4206, R^2 Score of 0.9999986
  - **Random Forest**: MSE of 11136.0712, R^2 Score of 0.9972
- **Broad Spectrum R^2 Score Comparison**: Present a bar graph that shows the R^2 Score for all evaluated models, including KNN and Linear Regression.
  - **Linear Regression**: R^2 Score notably lower than Ridge Regression, indicating less fit to the data.
  - **KNN**: R^2 Score substantially lower, reflecting its inadequate prediction power for this particular problem.
  - **Stability and Accuracy**: Ridge Regression was chosen for its exceptional balance of low error rates and high explanatory power.
- **Resilience to Overfitting**: Its regularization parameter (alpha) helps prevent overfitting, a crucial factor in maintaining model robustness.



Learning Curves and Overfitting Analysis

Learning curves serve as a diagnostic tool, plotting the model's performance on both the training set and validation set over various training set sizes. They provide insight into how well the model learns and generalizes to unseen data.

Ridge Regression: A Model in Learning

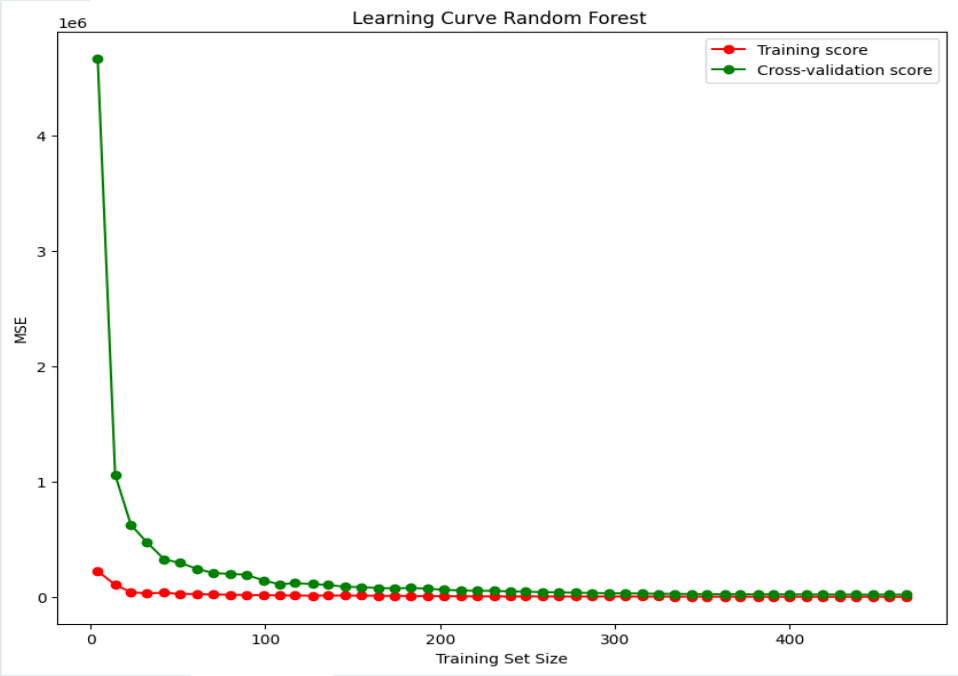
- **Learning Curve of Ridge Regression**: Highlight the **Learning Curve for Ridge Regression** graph, showing the convergence of training and cross-validation scores as the number of training examples increases.
- **Indicator of Good Generalization**: The small gap between the training and cross-validation scores suggests that the model generalizes well to new data.

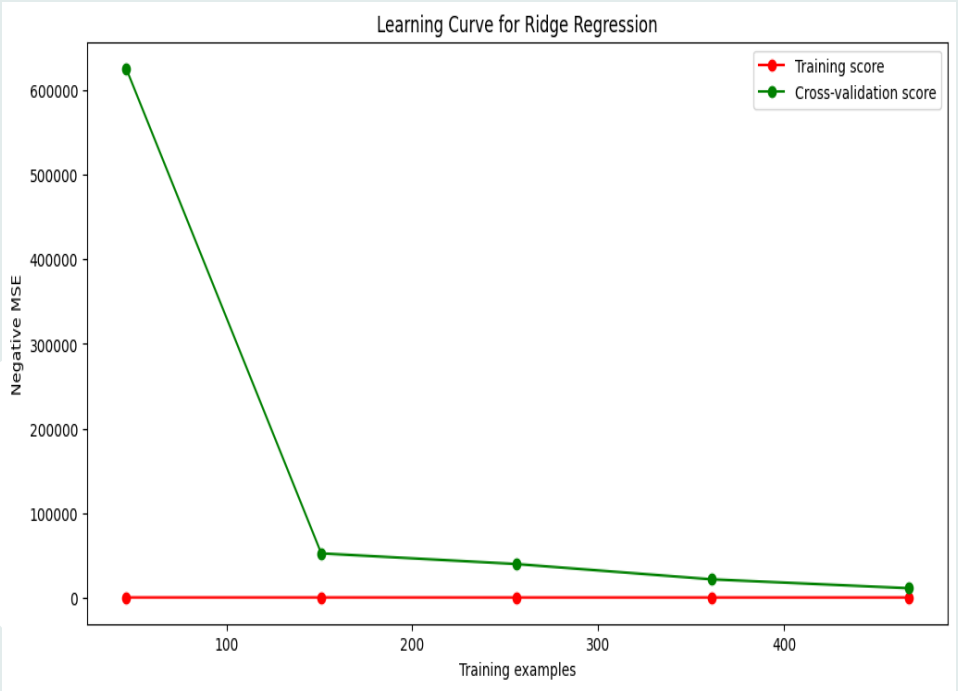
Random Forest: The Complexity Unfolded

- **Learning Curve of Random Forest**: Point to the **Learning Curve Random Forest** graph. Initially, the training score is perfect, but as more data is introduced, the cross-validation score converges, indicating learning and generalization.
- **Interpreting the Scores**: With an MSE on the cross-validation score decreasing and plateauing, it signifies that adding more data improves performance up to a certain point, after which the benefits diminish.

The Role of Cross-Validation Score

- **Relevance of Cross-Validation**: The cross-validation score is indicative of how well the model performs on unseen data. A high cross-validation score consistently close to the training score implies a model that is neither underfitting nor overfitting.





Concluding Insights and Future Endeavors

- **Ridge Regression's Excellence**: Achieved an R^2 Score of nearly perfect 0.9999986 and an MAE of 1.453, showcasing exceptional prediction accuracy.
- **Comparative Advantage**: Outperformed Random Forest, which also showed strong performance with an R^2 Score of 0.9972, but with a higher error rate.

Limitations and Opportunities:

- **Model Limitations**: Ridge Regression might struggle with non-linear relationships unlike tree-based models.
- **Feature Enhancement**: The current features may not cover all variables affecting demand, such as unexpected events or unrecorded behaviors.

Future Directions:

- **Explore More Models**: Consider Gradient Boosting or Neural Networks for better handling of complex patterns.
- **Enhance Features**: Add features related to external factors like public events or economic indicators.
- **Utilize Ensemble Techniques**: Combine different models to improve prediction stability and accuracy.

Closing Thoughts:

This study underscores the robust potential of machine learning to predict and plan for urban bike-sharing demand. It sets the stage for further innovations and enhancements in predictive modeling within the urban mobility sector.