Optimizing Bike Rental Predictions Using Advanced Regression Techniques

Detailed Introduction

The project's main goal is to use the Bike Sharing Dataset, which includes hourly historical data gathered from the Capital Bikeshare system in Washington, D.C., in 2011 and 2012, to analyze and forecast bike rental trends. This dataset is a useful resource for comprehending the dynamics of bike-sharing systems since it offers a wide range of variables, such as seasonal fluctuations, environmental conditions, and temporal parameters.

The main objective is to create a prediction model that can use these variables to forecast demand for bike rentals. The project's goal is to identify the main trends and factors influencing renting behavior by using regression techniques. This will make it easier to comprehend how different factors, such the weather, temperature, time of day, and user type (casual vs. registered), affect the utilization of bike sharing.

Beyond forecasting, the research aims to offer practical insights that can assist urban planners and legislators in improving the administration of bike-sharing programs. Decisions about fleet distribution, system scalability, and adjustment to peak and off-peak demands can be guided by these insights.

Objectives of the Project

Key Objectives

1. **Identify Influential Factors**: Analyze how variables like temperature, humidity, season, and time of day impact bike rental demand.
2. **Address Variability**: Detect unusual patterns in rentals, such as spikes during holidays or dips caused by adverse weather conditions.
3. **Develop Regression Models**: Build and validate predictive models that accurately estimate the total number of bike rentals at hourly and daily levels.
4. **Support Urban Planning**: Deliver insights that can improve the operational efficiency and user satisfaction of bike-sharing systems.

Dataset Overview

- **Source**: UCI Machine Learning Repository.
- **Samples**: 17,379 hourly observations from 2011 to 2012.
- **Features**: Environmental, temporal, and user-related factors, including:
    - Temperature
    - Feels-like temperature
    - Humidity
    - Windspeed
    - Season
    - Year
    - Month
    - Hour
    - Weekday
    - Holiday
    - Working day
    - Weather situation
    - Casual and registered users
- **Categorical Variables**: Season, year, month, hour, weekday, working day, holiday, weather situation.
- **Target Variable**: Total bike rental count (cnt, continuous) – sum of casual and registered users.
- **Characteristics**:
    - Temporal variability with hourly and seasonal fluctuations.
    - Environmental factors show varying skewness and outliers.
    - Moderate correlation among certain features like temperature and "feels-like" temperature.
- **Objective**: Use these features to predict total bike rentals and analyze their impact on rental patterns.

Data Preprocessing Steps

1. **Checking for Missing Values**:
    - Verified if any columns had missing values using .isnull().sum().
    - Outcome: No missing values were found in the dataset, indicating it is complete and ready for analysis.
2. **Duplicate Row Removal**:
    - Identified and removed any duplicate rows using .drop_duplicates().
    - Outcome: No duplicate rows were found, ensuring all records in the dataset are unique.
3. **Validating Categorical Data**:
    - Checked the uniqueness of values in categorical columns (season, yr, mnth, hr, weekday, workingday, holiday, and weathersit) to ensure consistency and correctness.
    - Outcome: All values in categorical columns were valid and aligned with expected ranges.
4. **Inspecting Numerical Columns**:
    - Summary statistics (mean, median, min, max) were calculated using .describe() to check for unusual ranges or anomalies in numerical columns (temp, atemp, hum, windspeed, casual, registered, cnt).
    - Negative or non-numeric values were also checked, but none were found.
    - Outcome: The data appeared clean, with no irregularities in numerical features.
5. **Dropping Irrelevant Columns**:

    Columns Removed:
    - instant: A record identifier that does not contribute meaningful information to modelling.
        - dteday: Redundant, as date-related trends are captured in mnth, hr, and weekday.
        - Outcome: Reduced dataset to relevant features for modeling.
6. **One-Hot Encoding for Categorical Variables**:
    - Applied **one-hot encoding** to categorical columns like season, hr, weathersit, weekday, yr, and mnth to convert them into numerical features for machine learning.

- Dropped the first category of each encoded variable to avoid multicollinearity (drop_first=True).
- Outcome: Created additional columns representing categories, enabling machine learning algorithms to process categorical data effectively.

7. **Renaming Columns for Clarity**:
    - Renamed one-hot encoded columns for better readability (e.g., season_1 to season_spring, weather_2 to weather_mist).
    - Outcome: Improved column clarity, enhancing interpretability of the dataset.
8. **Train-Test Splitting**:
    - Split the dataset into **training** and **testing** sets using train_test_split() with a **70-30 split** to ensure the model is evaluated on unseen data.
    - Outcome: Training set size: 12,165 rows; Testing set size: 5,214 rows.
9. **Feature Scaling**:
    - Applied **standardization** using StandardScaler to normalize numerical features like temp, atemp, hum, windspeed, and registered.
    - Standardization ensures features are on a similar scale, preventing large-valued features from dominating smaller-valued ones in machine learning models.
    - Outcome: Scaled numerical features for improved model performance.

Exploratory Data Analysis (EDA)

Univariate Analysis

- Temperature and feels-like temperature exhibit symmetric, bell-shaped distributions, with most observations centered around moderate values and no significant outliers.
- Humidity is left-skewed, with most data in the moderate-to-high range and a few outliers at low humidity levels.
- Windspeed is heavily right-skewed, with most values at low speeds and a few high-value outliers.
- Registered users show a strongly right-skewed distribution, with frequent low counts and high-value outliers during peak rental hours.

Bivariate Analysis

- Positive correlations are observed between temperature, feels-like temperature, and total rentals, with rentals peaking at moderate weather conditions.
- Humidity and windspeed have negative correlations with total rentals, as high humidity and strong winds deter biking activity.
- Registered users show a near-linear positive relationship with total rentals, indicating their predictable contribution to demand.

Multivariate Analysis

- Temperature's impact on bike rentals varies by season:
    - **Summer**: Strong positive correlation, with rentals peaking at higher temperatures.
    - **Fall**: Moderate correlation, clustering around moderate temperatures.
    - **Winter**: Weak correlation, with lower rentals at lower temperatures.
- Environmental factors like high humidity and windspeed interact to reduce rentals, while favorable conditions encourage usage.
- Time-based patterns reveal that registered users dominate weekday commuting hours, whereas casual users are more active during weekends and midday.

Outcomes from Assumptions Validation

1. Linearity

- Scatter plots between predictors (e.g., temperature, atemp, humidity) and the target variable (cnt) show linear trends for most continuous variables, particularly temperature and atemp.
- Categorical variables like season and workingday display distinct clusters, supporting their categorical influence on bike rentals.

2. Independence

- **Durbin-Watson Test**: The statistic is approximately 2, indicating minimal autocorrelation in the residuals.
- **Outcome**: Residual independence is confirmed, satisfying this assumption for regression analysis.

3. Homoscedasticity

- Residual plots against fitted values reveal consistent spread, particularly for linear and ridge regression models.
- Minor deviations from homoscedasticity are observed in features with outliers, like windspeed.

4. Normality of Residuals

- Histogram and Q-Q plots of residuals demonstrate near-normal distribution with slight tails.
- **Outcome**: The assumption of normality is approximately satisfied, suitable for models relying on normality.

5. Multicollinearity

- **Variance Inflation Factor (VIF)**:
    - Key predictors (temperature, atemp) show low to moderate VIF values, indicating manageable multicollinearity.
    - registered and casual variables are strongly correlated with cnt but are managed using regularization techniques.

Key Takeaways from Assumption Validation

1. **Linearity**:
    - Predictors like temperature and atemp show linear relationships with the target variable (cnt), validating their suitability for regression modeling.
2. **Independence**:
    - Residuals display minimal autocorrelation, confirmed by the Durbin-Watson statistic (~2), ensuring that observations are independent.
3. **Homoscedasticity**:
    - Residual plots indicate a consistent variance of errors for most predictors, satisfying the homoscedasticity assumption.
4. **Normality of Residuals**:
    - Residuals are approximately normally distributed, with minor deviations in tails, acceptable for regression models.
5. **Multicollinearity**:

- Variance Inflation Factor (VIF) values for predictors are low to moderate, indicating minimal multicollinearity, particularly for temperature and atemp.

Impact on Modeling

- **Improved Model Accuracy**: Validating assumptions like linearity and independence ensures reliable relationships between predictors and the target, enhancing predictive performance.
- **Stability and Interpretability**: Addressing multicollinearity and outliers stabilizes the model coefficients, making them interpretable and meaningful.
- **Valid Error Estimation**: Satisfying homoscedasticity and normality assumptions allows for accurate confidence intervals and hypothesis testing in regression models.
- **Better Handling of Data Characteristics**: Adjustments for overdispersion or non-normality ensure the model aligns with the dataset's true distribution, preventing biased results.

Regression Analysis Results and Interpretation:

Interpretation of Simple Linear Regression: Temperature vs. Bike Rentals

1. Training Set

- **Scatter Plot Insights**:
    - The data points show a clear upward trend, indicating a positive linear relationship between temperature and bike rentals.
    - The red regression line has a positive slope, suggesting that as the temperature increases, bike rentals also increase.
    - Despite the positive trend, there is significant variability around the regression line, indicating that other factors not captured by temperature influence bike rentals.
- **Metrics**:
    - **RMSE**: 167.9418
        - Indicates the average prediction error in terms of the dependent variable (bike rentals). The high RMSE suggests substantial prediction error.
    - **R-squared (R^2)**: 0.1606

        - Shows that only 16.06% of the variance in bike rentals is explained by temperature. This relatively low R^2 indicates that temperature alone is a weak predictor.

2. Test Set

- **Scatter Plot Insights**:
    - Similar to the training set, the test set shows a positive relationship between temperature and bike rentals.
    - The blue regression line reflects the same weak positive linear trend observed in the training set.
    - A wide spread of data points around the line indicates the model's limited predictive capability.
- **Metrics**:
    - **RMSE**: 162.7146
        - Slightly lower than the training set, but still high, confirming poor prediction accuracy.
    - **R-squared (R^2)**: 0.1645
        - Marginally higher than the training set, indicating 16.45% of variance is explained by temperature. This minimal improvement reaffirms the weak relationship.

3. Summary

- The linear regression model captures a weak positive relationship between temperature and bike rentals.
- The low R^2 values highlight that temperature explains only a small portion of the variability in bike rentals, suggesting other factors have a significant impact.
- High RMSE values in both sets indicate that the model has limited predictive power.
- The trends observed in the scatter plots align with the metrics, confirming that while temperature has a positive influence, it is insufficient as a standalone predictor.

Multiple Linear Regression: Actual vs. Predicted Bike Rentals

1. Training Set

- **Scatter Plot Insights**:
    - The predicted values closely align with the actual bike rental counts, following the red dashed "perfect fit" line.
    - Minor deviations are observed, but the high alignment indicates the model fits the training data well.
- **Residuals vs. Fitted Values**:
    - Residuals are scattered with a slight funnel shape, indicating some heteroscedasticity at higher fitted values.
    - Most residuals are clustered around the zero line, showing minimal error for the majority of predictions.
- **Metrics**:
    - **R^2**: 0.9744
        - Indicates that 97.44% of the variance in bike rentals is explained by the model, reflecting an excellent fit.
    - **RMSE**: 29.3186
        - Low RMSE suggests a small average prediction error in terms of bike rental counts.

2. Test Set

- **Scatter Plot Insights**:
    - Predicted values closely match the actual bike rentals, with points aligning well along the blue dashed "perfect fit" line.
    - Slightly larger deviations at higher values indicate minor errors in extreme predictions.
- **Residuals**:
    - Similar to the training set, residuals remain small and centered around zero, indicating consistent prediction accuracy across data splits.
- **Metrics**:
    - **R^2**: 0.9734
        - High R^2 shows that 97.34% of the variance in bike rentals is captured by the model, consistent with the training set performance.

- **RMSE**: 29.0292
    - Slightly lower than the training set, reinforcing the model's robust performance on unseen data.

3. Summary

- **Model Performance**:
    - The multiple linear regression model performs exceptionally well, with high R^2 values (~97%) and low RMSE (~29) for both training and test sets.
- **Residual Analysis**:
    - Residuals are generally small and centered around zero, indicating minimal prediction errors.
    - Slight heteroscedasticity suggests the model might benefit from further refinement for extreme values.
- **Conclusion**:
    - The model is highly effective at predicting bike rentals, capturing most of the variability in the data with minimal error. It generalizes well to unseen data, as evidenced by the consistent test set performance.

Polynomial Regression: Bike Rentals Prediction

1. Training Set

- **Scatter Plot Insights**:
    - Predicted values align closely with the actual bike rental counts, forming a nearly perfect fit along the red dashed "perfect fit" line.
    - The model captures nonlinear patterns effectively, resulting in minimal deviations.
- **Residuals vs. Predicted Values**:
    - Residuals are tightly clustered around the zero line, with minor variations indicating a good fit.
    - No significant pattern in residuals, suggesting that the polynomial regression model captures most of the variability in the training data.
- **Metrics**:
    - **R^2**: 0.9929
        - Indicates that 99.29% of the variance in bike rentals is explained by the polynomial regression model, signifying an excellent fit.

    - **RMSE**: 15.4509
        - A very low RMSE demonstrates high prediction accuracy in the training data.

2. Test Set

- **Scatter Plot Insights**:
    - The predicted values deviate significantly from the actual bike rentals, leading to extreme discrepancies in the predictions.
    - The near-flat regression line and extreme outliers indicate a severe overfitting issue with the polynomial model.
- **Residuals**:
    - Extreme deviations in residuals highlight the model's failure to generalize to unseen data.
- **Metrics**:
    - **R^2**: -358118273874.7134
        - The highly negative R^2 indicates that the model performs worse than a simple mean-based prediction, highlighting extreme overfitting.
    - **RMSE**: 106526449.1616
        - The enormous RMSE demonstrates the model's inability to make meaningful predictions on the test set.

3. Summary

- **Model Performance**:
    - The polynomial regression model performs exceptionally well on the training set, capturing nearly all variability with minimal error.
    - However, it fails catastrophically on the test set, producing extreme errors and invalid predictions, likely due to overfitting.
- **Residual Analysis**:
    - Training residuals indicate a good fit, but the test residuals show that the model fails to generalize to unseen data.
- **Conclusion**:
    - While polynomial regression captures complex patterns in the training data, it overfits significantly, making it unsuitable for this task without

further regularization or simplification. Regularization techniques like Ridge or Lasso regression might mitigate overfitting.

Regularization Techniques: LASSO, Ridge, and Elastic Net Regression

1. LASSO Regression

- **Training Set**:
  - **Scatter Plot Insights**:
    - Predicted values align closely with the actual bike rentals, forming a near-perfect fit along the ideal fit line.
    - Minor deviations indicate slightly higher prediction error compared to Ridge Regression.
  - **Residuals vs. Fitted Values**:
    - Residuals are scattered around the zero line but show slight variability at higher fitted values.
    - Most residuals remain small, suggesting the model captures the majority of the variability.
  - **Metrics**:
    - **R^2**: 0.9742 : Explains 97.42% of the variance in bike rentals. This is an excellent score, slightly lower than Ridge Regression, indicating a robust model fit.
    - **RMSE**: 29.4703: The average prediction error for the training set. While low and indicative of high accuracy, it's marginally higher than Ridge Regression (29.3188), implying LASSO's predictions are slightly less precise in the training data.
- **Testing Set**:
  - **Scatter Plot Insights**:
    - The predicted values align well with actual bike rentals, closely following the ideal fit line.
    - Slight deviations from perfect fit are observed at higher bike counts.
  - **Metrics**:
    - **R^2**: 0.9733: Explains 97.33% of the variance in the testing set, confirming that the model generalizes well to unseen data.

        - **RMSE**: 29.0967: The average prediction error on the testing set. This is a low value, showing that LASSO maintains predictive accuracy on unseen data. It's slightly better than the training RMSE, indicating a good balance between bias and variance.

- **Comparison**:
    - Slightly higher RMSE compared to Ridge Regression, indicating marginally weaker predictive performance. Comparable to multiple linear regression but better at managing multicollinearity.

2. Ridge Regression

- **Training Set**:
    - **Scatter Plot Insights**:
        - Predicted values closely match actual bike rentals, aligning almost perfectly with the ideal fit line.
        - The model shows the smallest deviations among regularization techniques.
    - **Residuals vs. Fitted Values**:
        - Residuals are minimal and evenly distributed around zero, indicating strong model performance.
    - **Metrics**:
        - **R^2**: 0.9744

          Indicates that 97.44% of the variance in bike rentals is explained by the model. This is an excellent score, demonstrating a very high level of predictive accuracy.
        - **RMSE**: 29.3188

          Reflects the average error in predicting bike rentals. The low RMSE value signifies that the model's predictions are very close to the actual values, further validating the model's accuracy and reliability.
- **Testing Set**:
    - **Scatter Plot Insights**:
        - The predicted values align tightly with actual rentals, closely following the ideal fit line.
        - Very small deviations at higher values suggest excellent generalization.

    - **Metrics**:
        - **R^2**: 0.9734
        - **RMSE**: 29.0272
- **Comparison**:
    - Best overall performance among all models, with the lowest RMSE and high R^2 values. Slightly better than LASSO Regression and equivalent to multiple linear regression in accuracy but more robust due to regularization.

3. Elastic Net Regression

- **Training Set**:
    - **Scatter Plot Insights**:
        - Predicted values align with the actual bike rentals but show slightly larger deviations compared to Ridge and LASSO Regression.
    - **Residuals vs. Fitted Values**:
        - Residuals are centered around zero but display slightly more variability, indicating higher prediction errors.
    - **Metrics**:
        - **R^2**: 0.9683
        - **RMSE**: 32.6237
- **Testing Set**:
    - **Scatter Plot Insights**:
        - Predicted values show alignment with actual rentals but with noticeable deviations, especially at extreme values.
    - **Metrics**:
        - **R^2**: 0.9674
        - **RMSE**: 32.1381
- **Comparison**:
    - Performs worse than Ridge and LASSO Regression, with higher RMSE and slightly lower R^2 values. While it manages multicollinearity, it sacrifices precision compared to other models.

Summary

- **Ridge Regression** emerges as the best-performing model with the lowest RMSE and highest R^2, both for training and testing data, making it superior to LASSO and Elastic Net.
- LASSO Regression performs closely but shows slightly higher RMSE, while Elastic Net has reduced accuracy and generalization. Multiple linear regression offers comparable performance but lacks the robustness of Ridge Regression. Polynomial regression suffers from severe overfitting, making it unsuitable for generalization.

Quantile Regression: Bike Rentals Prediction (Quantile = 0.5)

1. Training Set

- **Scatter Plot Insights**:
  - Predicted values closely align with actual bike rental counts, forming a strong fit along the red dashed "perfect fit" line.
  - Minor deviations are observed, but the alignment indicates that quantile regression effectively captures the central trend.
- **Residuals vs. Predicted Values**:
  - Residuals are well distributed around the zero line, with slight variability at higher predicted values.
  - Most residuals remain small, showing the model captures the majority of the variability in the training data.
- **Metrics**:
  - **R^2**: 0.9674
    - Indicates that 96.74% of the variance in bike rentals is explained by the quantile regression model, reflecting strong performance.
  - **RMSE**: 33.0785
    - Slightly higher than Ridge Regression, showing marginally larger prediction error on the training set.

2. Test Set

- **Scatter Plot Insights**:
  - Predicted values align well with actual bike rentals, following the ideal fit line with minimal deviations.

    - Performance is consistent with the training set, showing the model's robustness and generalization capability.
- **Residuals vs. Predicted Values**:
    - Residuals are centered around zero and display a similar pattern to the training set, indicating consistent prediction accuracy.
- **Metrics**:
    - **R^2**: 0.9663
        - Indicates that 96.63% of the variance is explained on the test set, slightly lower than the training set but still strong.
    - **RMSE**: 32.6593
        - Lower than the training set, confirming the model's ability to generalize to unseen data.

Comparison with Other Models

- Quantile regression performs slightly worse than Ridge Regression in terms of R^2 and RMSE but offers a robust alternative for capturing the central tendency of the data.
- Unlike polynomial regression, it avoids overfitting and generalizes well to unseen data.
- Performance is comparable to Elastic Net Regression but less precise than Ridge and LASSO Regression, which have slightly better RMSE and R^2 values.

Summary

- **Model Performance**:
    - Quantile regression captures the central trend effectively, showing strong alignment with actual values and consistent residual patterns.
- **Residual Analysis**:
    - Residuals indicate good model performance with minimal error and no significant patterns.
- **Conclusion**:
    - Quantile regression is a robust choice for capturing central trends, with performance slightly below Ridge Regression but superior to Elastic

Net. It generalizes well and avoids overfitting, making it suitable for datasets with high variability.

Poisson Regression: Bike Rentals Prediction

1. Training Set

- **Scatter Plot Insights**:
  - Predicted values generally follow the actual bike rental counts but show a tendency to overestimate at higher rental counts.
  - The regression line deviates from the ideal fit line at extreme values, suggesting that the model struggles with higher counts.
- **Residuals vs. Predicted Values**:
  - Residuals are centered around zero for lower predicted values but show a systematic pattern, with increasing residuals as the predicted counts grow.
  - The spread widens for higher predicted values, indicating potential heteroscedasticity.
- **Metrics**:
  - **R^2**: 0.8957
    - Indicates that 89.57% of the variance in bike rentals is explained by the model, a strong but not optimal fit compared to Ridge or Quantile Regression.
  - **RMSE**: 59.1883
    - Higher than other regression models, reflecting larger average prediction errors.

2. Test Set

- **Scatter Plot Insights**:
  - Predicted values follow the actual bike rentals with deviations at higher rental counts.
  - Performance is consistent with the training set, showing reasonable generalization.
- **Residuals vs. Predicted Values**:
  - Residuals exhibit a similar pattern to the training set, with increasing residuals as predicted values grow.

    - Consistency in the residual pattern suggests that the model behaves similarly on unseen data.
- **Metrics**:
    - **R^2**: 0.8992
        - Slightly better than the training set, showing good generalization but still lower than models like Ridge or Quantile Regression.
    - **RMSE**: 56.5248
        - Lower than the training set, indicating reasonable predictive performance but higher error compared to Ridge or LASSO Regression.

Comparison with Other Models

- **Ridge Regression**:
    - Outperforms Poisson Regression with higher R^2 (~97%) and significantly lower RMSE (~29), demonstrating better accuracy and generalization.
- **Quantile Regression**:
    - Quantile Regression has similar R^2 (~96%) but lower RMSE (~32), making it more precise for central trends.
- **Polynomial Regression**:
    - Poisson Regression avoids the extreme overfitting seen in Polynomial Regression, making it a more robust alternative.
- **Elastic Net Regression**:
    - Poisson Regression has higher RMSE and lower R^2 compared to Elastic Net, making it less precise.

Summary

- **Model Performance**:
    - Poisson Regression explains a significant portion of the variance but struggles with higher rental counts, leading to larger prediction errors.
- **Residual Analysis**:
    - Residuals highlight systematic errors and widening spread at higher predicted values, indicating heteroscedasticity.
- **Conclusion**:

- Poisson Regression is a robust model for count-based predictions but falls short in accuracy compared to Ridge or Quantile Regression. It avoids overfitting but lacks the precision of regularized models.

Negative Binomial Regression: Bike Rentals Prediction

1. Training Set

- **Scatter Plot Insights**:
  - Predicted values deviate significantly from actual bike rental counts, especially for higher rental values.
  - The regression line does not align with the ideal fit line, indicating poor model fit.
- **Residuals vs. Predicted Values**:
  - Residuals show a systematic pattern, with errors increasing as predicted values grow.
  - A distinct widening pattern in residuals reflects poor handling of higher rental counts, suggesting significant heteroscedasticity.
- **Metrics**:
  - **R^2**: 0.0212
    - Indicates that only 2.12% of the variance in bike rentals is explained by the model, showing very poor fit.
  - **RMSE**: 181.3498
    - High RMSE reflects substantial prediction errors on the training data.

2. Test Set

- **Scatter Plot Insights**:
  - Predicted values again fail to align with actual bike rentals, particularly for higher values, where deviations are more pronounced.
  - The model struggles to generalize beyond lower rental counts.
- **Residuals vs. Predicted Values**:
  - Similar to the training set, residuals show a consistent pattern of increasing errors with higher predicted values.
  - The systematic nature of errors suggests the model is fundamentally unable to capture the underlying distribution.

- **Metrics**:
    - **R^2**: 0.1589
        - Indicates that 15.89% of the variance in bike rentals is explained by the model, a slight improvement over the training set but still poor.
    - **RMSE**: 163.2537
        - Slightly lower than the training RMSE but remains high, reflecting poor predictive performance.

Comparison with Other Models

- **Ridge Regression**:
    - Outperforms Negative Binomial Regression significantly, with R^2 (~97%) and RMSE (~29), demonstrating superior accuracy and generalization.
- **Quantile Regression**:
    - Quantile Regression offers much better performance, with R^2 (~96%) and lower RMSE (~32), effectively capturing central trends.
- **Poisson Regression**:
    - Performs better than Negative Binomial Regression, with higher R^2 (~90%) and lower RMSE (~56), suggesting it is a more suitable count-based model.
- **Polynomial Regression**:
    - Negative Binomial Regression avoids overfitting but is significantly less accurate than Polynomial Regression in training performance.

Summary

- **Model Performance**:
    - Negative Binomial Regression shows poor performance, with extremely low R^2 values and high RMSE, making it unsuitable for predicting bike rentals.
- **Residual Analysis**:
    - Residual patterns reflect fundamental issues in capturing variability, especially for higher rental counts.

- **Conclusion**:
    - Negative Binomial Regression fails to model the underlying data distribution effectively. It is outperformed by Ridge, LASSO, Quantile, and even Poisson Regression. This model is not recommended for this dataset.

Partial Least Squares Regression

Training Results

1. Scatter Plot :

- The scatter plot shows predicted vs. actual bike rentals.
- Most points align closely with the red "perfect fit" line, indicating accurate predictions on the training set.

2. Residuals vs. Predicted :

- Residuals are centered around zero, with no clear patterns or trends.
- Minor deviations at higher predicted values indicate a well-fitted model, with no significant bias or heteroscedasticity.

- **Metrics**

  Mean R^2: 0.9686 (±0.0002)
    - Indicates that 96.86% of the variance in bike rentals is explained by the model during training.
    - The high R^2 value shows strong predictive power for training data.

  Mean RMSE: 32.5070 (±0.0609)
    - A low RMSE reflects a small average error in predictions, confirming that the model performs well on the training data.

Testing Results

Scatter Plot:

- The scatter plot of test predictions also shows a strong alignment with the ideal fit line, confirming the model's ability to generalize well to unseen data.

Residuals vs. Predicted:

- Similar to the training set, residuals are evenly distributed around zero in the test data.

- The absence of strong patterns suggests that the model assumptions are satisfied, and predictions are unbiased.

Metrics

- **Mean R^2**: 0.9674 (±0.0000)
    - The high R^2 value confirms that the model captures most of the variability in the testing data, maintaining strong predictive accuracy.
- **Mean RMSE**: 32.1368 (±0.0000)
    - A low RMSE indicates that the model predictions are accurate for the test dataset, with minimal average prediction error.

Conclusion:

- PLS Regression demonstrates excellent predictive accuracy and generalization, as shown by high R^2 and low RMSE across training, validation, and testing datasets.
- Residual plots confirm that the model satisfies linear regression assumptions, with minimal bias and no heteroscedasticity.
- The strong alignment between predicted and actual values makes PLS Regression a reliable choice for this dataset.

Principal Component Regression (PCR): Bike Rentals Prediction

1. Training Set

- **Scatter Plot Insights**:
    - Predicted values align well with actual bike rentals, forming a close fit along the red dashed "perfect fit" line.
    - Deviations are minimal, indicating that PCR effectively captures the main variance in the data.
- **Residuals vs. Predicted Values**:
    - Residuals are centered around the zero line, with increased spread at higher predicted values.
    - No major systematic patterns, but the variance at larger counts suggests potential heteroscedasticity.
- **Metrics**:
    - **R^2**: 0.9377

        - Indicates that 93.77% of the variance in bike rentals is explained by the model, reflecting strong performance.
    - **RMSE**: 45.7467
        - A slightly higher RMSE compared to Ridge Regression and Partial Least Squares (PLS), suggesting moderate prediction errors.

2. Test Set

- **Scatter Plot Insights**:
    - Predicted values maintain alignment with actual bike rentals, following the blue dashed "perfect fit" line.
    - Deviations remain consistent with the training set, showing good generalization.
- **Residuals vs. Predicted Values**:
    - Residuals are distributed around zero, with a similar pattern to the training set, indicating no overfitting.
- **Metrics**:
    - **R^2**: 0.9359
        - Indicates that 93.59% of the variance in bike rentals is captured on unseen data, consistent with the training performance.
    - **RMSE**: 45.0806
        - Slightly lower than the training RMSE, demonstrating robust predictive performance.

Comparison with Other Models

- **Ridge Regression**:
    - Ridge Regression outperforms PCR, with higher R^2 (~97%) and lower RMSE (~29), making it more precise and accurate.
- **PLS Regression**:
    - PLS has a slight edge over PCR, with R^2 (~96%) and RMSE (~32), making it better at capturing data variability.
- **Quantile Regression**:
    - Quantile Regression is similarly robust but shows slightly better performance with lower RMSE (~32) and higher R^2 (~96%).
- **Negative Binomial Regression**:

- PCR significantly outperforms Negative Binomial Regression, which has much lower R^2 (~15%) and far higher RMSE (~163).

Summary

- **Model Performance**:
  - PCR performs well, explaining ~94% of the variance with consistent results across training and test sets. However, it is less precise than Ridge and PLS Regression.
- **Residual Analysis**:
  - Residuals suggest good fit, with minor variability at higher predicted values, indicating some heteroscedasticity.
- **Conclusion**:
  - PCR is a strong choice for dimensionality reduction and robust predictions, but Ridge and PLS Regression provide better accuracy and generalization for this dataset.

Model Improvement Steps:

1. **Feature Transformation**:
   - Applied **Yeo-Johnson transformation** on continuous features like humidity, windspeed, registered, and atemp to handle skewness and normalize data without requiring positive values.
   - Replaced transformed features with normalized versions using **Quantile Transformation** (output distribution set to "normal") for better handling of outliers and non-linear distributions.
2. **Feature Selection**:
   - Identified important features using **Random Forest Regressor** based on feature importance scores.
   - Retained only features with an importance score greater than 0.01, simplifying the dataset and focusing on impactful predictors.
3. **LASSO Regularization**:
   - Performed hyperparameter tuning for LASSO Regression using **GridSearchCV** with parameters like alpha (regularization strength) and fit_intercept.
   - Selected the optimal hyperparameters (alpha: 0.01, fit_intercept: True) based on the highest cross-validated R^2 score.

    - Evaluated the LASSO model, achieving a training R^2 of 0.8373 and a test R^2 of 0.8241, confirming robust generalization.
4. **Cross-Validation**:
    - Utilized **K-Fold Cross-Validation (5 folds)** to ensure model stability and mitigate the risk of overfitting during hyperparameter tuning.
5. **Feature Coefficients Analysis**:
    - Extracted feature coefficients from the best LASSO model.
    - Filtered non-zero coefficients to identify impactful predictors, and sorted them by absolute value to highlight the most significant features influencing predictions.
6. **Dimensionality Reduction**:
    - The combination of feature selection (using Random Forest) and LASSO regularization reduced dataset complexity while maintaining model performance.

Best Model

The LASSO Regression model stood out as the best-performing model, demonstrating a strong balance between accuracy, robustness, and interpretability. Before model improvement, it achieved a training RMSE of 29.47 and R^2 of 0.9742, with corresponding testing metrics of RMSE 29.10 and R^2 0.9733, indicating high accuracy and minimal overfitting. However, feature transformations and selection were not fully optimized, leaving opportunities for enhancing model robustness and interpretability.

After implementing improvements, the model underwent significant refinement. Key transformations, including Yeo-Johnson and Quantile transformations, effectively handled skewness and normalized the data, improving the handling of outliers and feature scaling. Additionally, hyperparameter tuning optimized regularization strength (alpha = 0.01) and the intercept setting, ensuring a more robust model configuration. Feature selection further streamlined the dataset, retaining only impactful predictors and reducing noise, complexity, and overfitting risks.

Post-improvement metrics included a training R^2 of 0.8373 and testing R^2 of 0.8241, reflecting a slight trade-off in accuracy for improved generalization and reduced model complexity. The identification of non-zero coefficients allowed for clear insights into the most significant predictors, enhancing the model's interpretability. While the marginal reduction in R^2 suggests some compromise in accuracy, the enhanced robustness and simplicity make the model more suitable for practical applications. Overall, the LASSO Regression model effectively balances performance and usability, making it a reliable choice for real-world deployment.