**1. Overview of the Algorithm**

* **Algorithm:** Linear Regression
* **Dataset Used:** Bicycle counts from Fremont Bridge and weather data
* **Purpose:** Predict daily bicycle counts using weather and time-related features

**2. Adjustments/Alterations/Additions**

**2.1. Feature Engineering Adjustments**

|  |  |  |
| --- | --- | --- |
| **Change Made** | **Reason for Change** | **Expected Impact** |
| Added **day of the week** as a one-hot encoded feature | Weekdays vs weekends might affect cycling behavior | Improve model accuracy for weekday cycling trends |
| Added **holiday feature** | Bicycle usage might decrease on holidays | Capture reduced activity on holidays |
| Computed **hours of daylight** | More daylight could encourage more cycling | Identify seasonal effects |
| Converted **temperature from tenths of degrees to Celsius** | Ensure proper scaling for regression model | Improve model interpretability |
| Converted **precipitation from tenths of mm to inches** | Standardize weather features | Improve numerical stability |

**2.2. Data Visualization Enhancements**

|  |  |
| --- | --- |
| **Visualization Added** | **Purpose** |
| **Correlation Heatmap** | Understand feature relationships |
| **Histogram of Total Bicycle Counts** | Analyze data distribution |
| **Scatter Plots (Temp, Daylight, Precipitation vs. Total Count)** | Identify relationships between numerical variables |
| **Box Plot (Day of the Week vs. Total Count)** | Compare distributions across days |
| **Feature Importance Bar Plot** | Interpret the model coefficients |

**2.3. Model Adjustments**

|  |  |  |
| --- | --- | --- |
| Change Made | Reason for Change | Expected Impact |
| **Train-test split (80%-20%)** | Ensure model generalizability | Reduce overfitting |
| **Cross-validation (5-fold)** | Evaluate model robustness | Provide a more stable performance estimate |
| **Bootstrap Error Estimation** | Assess uncertainty in feature importance | Improve model interpretation |

**Key Observations**

* The **inclusion of holidays, daylight hours, and weather variables significantly improved model performance**.
* **Cross-validation results showed that the model generalized well to unseen data**.
* **The correlation heatmap confirmed key relationships**, such as **negative correlation between precipitation and bicycle usage**.
* **Feature importance visualization showed that daylight, temperature, and dry days positively influence bicycle counts**.

**4. Analysis of Adjustments**

**4.1. Model Performance Analysis**

* The adjusted model **demonstrates better predictive power** as evidenced by a lower MSE and higher R².
* The most influential features were **daylight hours, temperature, and dry day**, suggesting that **weather conditions significantly affect cycling activity**.
* The introduction of the **holiday variable** improved predictions by accounting for lower cycling activity on public holidays.

**4.2. Limitations and Next Steps**

* **Precipitation effects may not be fully captured**—future models could incorporate **intensity of rain** or wind speed.
* **Feature interactions (e.g., effect of temperature on rainy days)** might improve the model.
* **A non-linear model (e.g., decision trees or neural networks) could be tested to capture complex relationships**.

**5. Summary & Future Work**

|  |  |
| --- | --- |
| Adjustment Type | Impact Summary |
| **Feature Engineering** | Improved model explanatory power |
| **Data Visualization** | Helped interpret key relationships |
| **Model Refinements** | Reduced error and improved generalizability |
| **Evaluation Metrics** | Provided robust performance assessment |

**Next Steps:**

1. Experiment with **polynomial regression or interaction terms**.
2. Explore **ensemble models like Random Forests** to capture non-linear trends.
3. Include **additional external data sources** (e.g., traffic events, holidays, or special events).