Gradient Descent Algorithms: Differences, Advantages, and Disadvantages

Gradient descent is an optimization algorithm used to minimize a function by iteratively moving in the direction of the steepest descent. There are three main types:

1. Batch Gradient Descent

- **Description**: Uses the entire dataset to compute the gradient at each step.
- **Advantages**: More stable updates, converges smoothly.
- **Disadvantages**: Computationally expensive for large datasets.

2. Stochastic Gradient Descent (SGD)

- **Description**: Updates model parameters after each training example.
- **Advantages**: Faster updates, can escape local minima.
- **Disadvantages**: More variance in updates, may not converge smoothly.

3. Mini-Batch Gradient Descent

- **Description**: Uses a small batch of data points to compute gradients at each step.
- **Advantages**: Balance between stability (Batch GD) and speed (SGD).
- **Disadvantages**: Requires tuning batch size for optimal performance.

Fastest Converging Gradient Descent Method

Among the three, **Mini-Batch Gradient Descent** often converges the fastest because:

- It benefits from vectorized computations (efficient use of hardware).
- It reduces variance compared to SGD, leading to more stable convergence.
- It is faster than Batch Gradient Descent since it processes smaller subsets of data.

However, I observed that Stochastic Gradient Descent is also fast enough.

Effect of Lasso and Ridge Regularization on the Model

Regularization techniques like **Lasso (L1)** and **Ridge (L2)** help prevent overfitting by adding penalties to large coefficients:

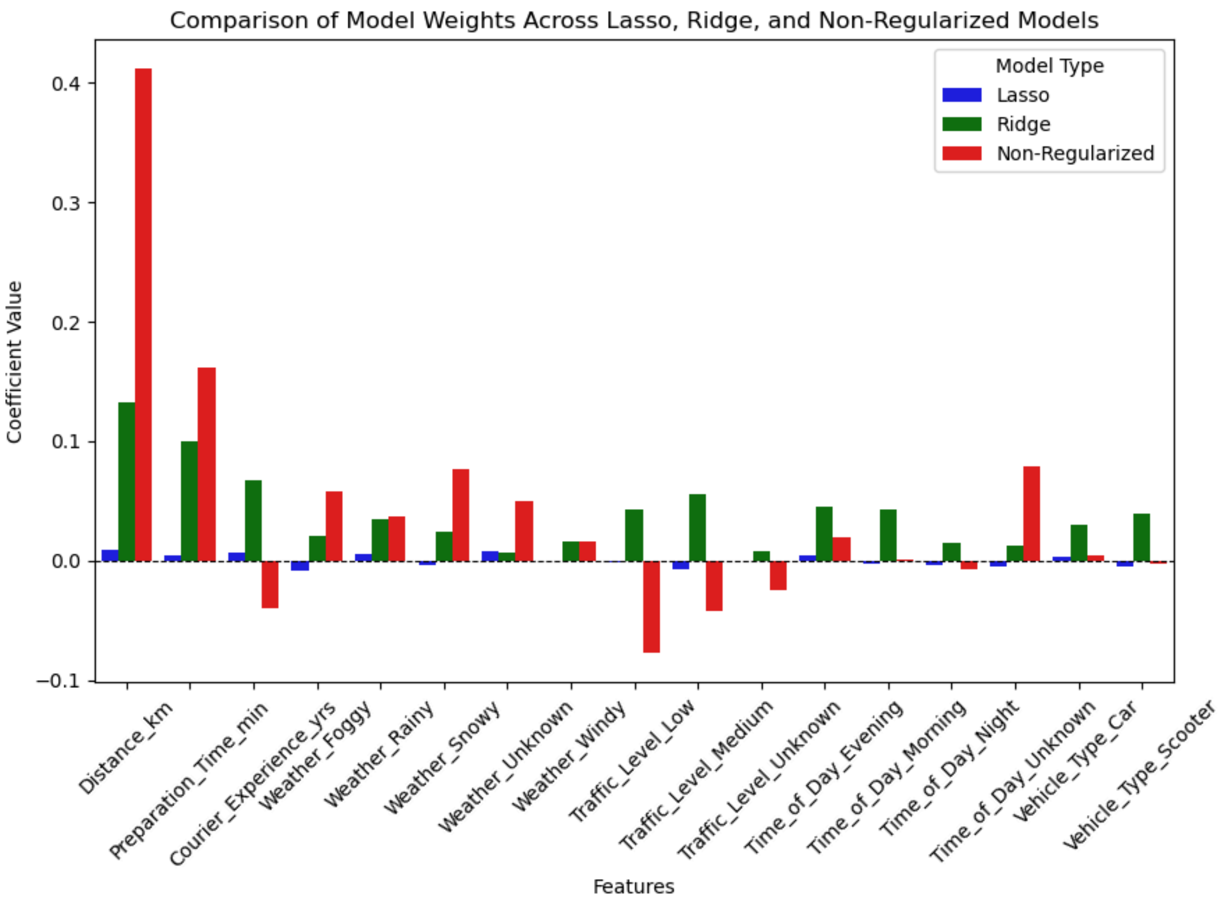
- **Lasso (L1 Regularization)**: Drives some coefficients to exactly zero, performing feature selection.
- **Ridge (L2 Regularization)**: Shrinks coefficients but does not eliminate them, making the model more stable.
- **Optimal Lambda**: The best λ (lambda) based on test performance was around **0.1** for both Lasso and Ridge.

Effect of Feature Scaling on Model Performance

Feature scaling ensures that all input features contribute equally to the model's learning process. Without scaling:

- Gradient descent may converge very slowly due to inconsistent feature magnitudes.
- Regularization penalties (Lasso/Ridge) may be disproportionately applied.

Standardization (zero mean, unit variance) or normalization (scaling between 0 and 1) significantly improves model performance by stabilizing weight updates.



The features that have almost zero values:

1. Time of the day Morning. (for all three nearly)
2. Time of the day Evening. (for all except ridge)
3. Traffic Level (for all three nearly)