Report on Traffic Volume Prediction: Linear Regression, Random Forest, and SVM Comparison

IT IS FOR YOU GUYS FOR REFERENCE

1. Introduction

This report compares three machine learning algorithms—**Linear Regression**, **Random Forest**, and **Support Vector Regression (SVR)**—on the task of predicting traffic volume. The dataset used is the **Metro Interstate Traffic Volume** dataset, available from the UCI Machine Learning Repository. The aim is to explore how different models perform in predicting traffic flow based on weather and time features.

2. Dataset Description

The dataset contains historical traffic volume data recorded hourly on the I-94 Interstate highway. Key features used in this study include:

- **Temperature** (`temp`): Temperature in Kelvin.
- **Rainfall** (`rain_1h`): Amount of rain in the past hour.
- **Snowfall** (`snow_1h`): Amount of snow in the past hour.
- **Cloud coverage** (`clouds_all`): Percentage of cloud coverage.
- **Hour** (`hour`) and **Day of the week** (`day_of_week`).

These features serve as predictors to estimate the target variable, **traffic volume**, which is the number of vehicles passing through the highway during a specific hour.

3. Evaluation Metrics Used

Since this is a **regression problem**, where the goal is to predict continuous traffic volume, the evaluation metrics chosen to assess model performance are:

- **MSE (Mean Squared Error)**: Measures the average squared differences between the predicted and actual traffic volumes, making it sensitive to large deviations.
- **R^2 (R-squared)**: Explains the proportion of variance in the traffic volume data that the model can capture. A higher R^2 value means better model performance in explaining the variability of the data.

Why MSE and R^2 are Used

- **MSE** and **R^2** are appropriate for this regression task because they directly quantify the accuracy of predictions for continuous numerical data.
- Metrics like **Precision**, **Recall**, and **mAP** are typically reserved for **classification tasks**, where the goal is to predict discrete categories. However, in this case, predicting traffic volume requires measuring how closely the predicted values align with actual values, and that's where MSE and R^2 excel.

4. Time Comparison for Model Training

A key aspect of this analysis is the comparison of **training times** for each model. Training time is a crucial factor in selecting models, especially for real-time applications where frequent retraining may be necessary.

- **Linear Regression**: The training time for Linear Regression is typically the shortest because it is a simple, direct method with a closed-form solution.
- **Random Forest**: Training Random Forest models is more computationally intensive as it involves constructing multiple decision trees. As a result, the training time is significantly longer.
- **SVR**: SVR models, especially with non-linear kernels like RBF, also require more training time as they involve optimization over a large number of support vectors.

The training time comparison was visualised, showcasing the differences between these models. This is important in scenarios where computational resources or time constraints are limiting factors.

5. Other Considerations (Not Based on Data)

In addition to accuracy and speed metrics, several other important factors were considered theoretically:

- **Resource Utilisation (CPU usage and memory consumption)**:
  - **Linear Regression** has minimal CPU and memory usage due to its simplicity.
  - **Random Forest** requires significantly more resources, particularly for large datasets.
  - **SVR** is also computationally expensive, particularly for large feature spaces or non-linear kernels.
- **Robustness**:
  - **Linear Regression** is less robust under varying conditions or when the relationships between features are non-linear.
  - **Random Forest** is highly robust, handling interactions between features and varying conditions well.
  - **SVR** is also robust, particularly with non-linear kernels like RBF, but may struggle with very large datasets.
- **Model Size and Hardware Requirements**:
  - **Linear Regression** models are typically small and lightweight.
  - **Random Forest** models grow significantly larger as the number of trees increases, potentially requiring more memory.
  - **SVR** models vary in size, but can become large if many support vectors are involved.

6. Dataset Source

The dataset used in this project is the **Metro Interstate Traffic Volume Dataset**, and it can be accessed via the following link: Metro Interstate Traffic Volume Dataset.

7. Model Performance on Accuracy Metrics

In the evaluation of the models using Mean Squared Error (MSE) and Mean Absolute Error (MAE), it was observed that Linear Regression exhibited the highest MSE and MAE values compared to Random Forest and Support Vector Regression (SVR).

Reasons for Higher MSE and MAE in Linear Regression

1. **Simplicity of the Model**:
   - Linear Regression assumes a linear relationship between the independent variables (features) and the dependent variable (traffic volume). If the actual relationship is non-linear, Linear Regression will struggle to capture that relationship accurately, leading to larger errors.
2. **Sensitivity to Outliers**:
   - Linear Regression is sensitive to outliers because it minimizes the squared differences between predicted and actual values. A few extreme values can disproportionately influence the model, resulting in higher MSE and MAE.
3. **Feature Interactions**:
   - Linear Regression cannot model interactions between features unless explicitly defined (e.g., by adding polynomial terms). In contrast, models like Random Forest and SVR can automatically capture complex relationships and interactions among features.
4. **Data Distribution**:
   - If the distribution of the traffic volume data is skewed or has heavy tails, Linear Regression may not perform well, leading to larger prediction errors.
5. **Multicollinearity**:
   - If the independent variables are highly correlated, it can lead to multicollinearity issues, which make the coefficients unstable and the model less reliable.
6. **Limited Capacity to Fit**:
   - Linear Regression has a high bias, meaning it may oversimplify the data. Other models like Random Forest and SVR can capture more complexity (lower bias), thus achieving better accuracy on diverse datasets.

Conclusion

Overall, while Linear Regression is a powerful and interpretable model for certain datasets, its limitations become apparent when the underlying data relationships are more complex or influenced by outliers. Therefore, in situations where the data exhibits non-linearity or complex interactions, more advanced models like Random Forest and SVR tend to outperform Linear Regression, resulting in lower MSE and MAE.