Assignment Question 3

To predict bike-sharing counts per hour based on various features such as weather, day, time, humidity, wind speed, and season, I need to select a suitable machine learning (ML) model. In this scenario, I am dealing with a regression problem where we aim to predict a continuous target variable (bike-sharing counts). Here's my approach for selecting an appropriate ML model:

1. **Exploratory Data Analysis (EDA)**: Before selecting a model, it's crucial to conduct EDA to understand the relationships between the features and the target variable. This involves analyzing distributions, correlations, and trends in the data, as well as identifying any outliers or missing values.
2. **Feature Engineering**: Feature engineering plays a vital role in improving model performance. I will create new features or transform existing ones to better capture the relationships in the data. For example, i can extract features like hour of the day, day of the week, and month from the 'dtoday' column.
3. **Model Selection**: Considering the nature of the problem and the dataset characteristics, several ML models are suitable for regression tasks like predicting bike-sharing counts. Here are some potential options:
    - **Linear Regression**: This is a simple and interpretable model that works well when the relationship between the features and the target variable is linear. It's a good baseline model to start with.
    - **Random Forest Regression**: Random forest is an ensemble learning method that works well with both categorical and numerical features. It can capture complex nonlinear relationships in the data and handle interactions between features effectively.
    - **Gradient Boosting Regression**: Gradient boosting algorithms like XGBoost or LightGBM often performs well in regression tasks. They build multiple weak learners sequentially, each correcting the errors of its predecessor, and combine them to create a strong predictive model.
4. **Model Evaluation**: Once we select candidate models, we evaluate their performance using appropriate metrics such as mean squared error (MSE), root mean squared error (RMSE), or mean absolute error (MAE). We can use techniques like cross-validation to ensure the robustness of our model evaluations.
5. **Model Interpretability**: While predictive performance is essential, model interpretability is also valuable. Linear regression provides interpretable coefficients, while tree-based models offer feature importances.

Based on the above considerations, my initial choice would be to start with a Random Forest Regression model. Random forests are robust, versatile, and often perform well out of the box without extensive hyperparameter tuning. They can handle a mix of categorical and numerical features, as well as interactions between

them. Additionally, they provide feature importances, which can help in understanding which features are most influential in predicting bike-sharing counts.

However, I would also experiment with other models such as Gradient Boosting Regression and Neural Networks to see if they can offer better predictive performance.