MLOPS Assignment 2

1. Dataset Overview

- **Total Entries**: 17,379
- **Total Columns**: 17

2. Columns and Data Types

The dataset includes the following columns:

1. `instant` - Unique identifier
2. `dteday` - Date of record
3. `season` - Season of the year
4. `yr` - Year (0 for 2011, 1 for 2012)
5. `mnth` - Month of the year
6. `hr` - Hour of the day
7. `holiday` - Whether it is a holiday
8. `weekday` - Day of the week
9. `workingday` - Whether it is a working day
10. `weathersit` - Weather situation
11. `temp` - Temperature
12. `atemp` - Feels-like temperature
13. `hum` - Humidity
14. `windspeed` - Wind speed
15. `casual` - Number of casual users
16. `registered` - Number of registered users
17. `cnt` - Total number of bike rentals (target variable)

3. Preprocessing Steps

Feature Engineering and Transformation:

1. **New Features Created:**
    - **`avg_temp`**: Average of `atemp` and `temp` to capture overall temperature experience.
    - **`heat_index`**: A combined measure of temperature and humidity to estimate perceived heat.
    - **`day_night`**: Indicator for whether the hour is part of the day (6 AM to 6 PM) or night.

2. **Dropped Features:**
    - **`atemp`**: Removed as its information is now represented by `avg_temp`.
    - **`temp`**: Removed as its information is now represented by `avg_temp`.
    - **`hum`**: Removed as its information is now represented by `heat_index`.
    - **`instant`**: Dropped as it is a unique identifier and not needed for modeling.
    - **`casual`**: Dropped to avoid redundancy; `cnt` captures the total number of rentals.
    - **`registered`**: Dropped for the same reason as `casual`.
    - **`dteday`**: Dropped after converting to a datetime object which is not needed for direct modeling.

3. **Data Type Conversion:**
    - **`dteday`**: Converted to `datetime` format for potential time-based analysis.
    - Categorical columns were converted to the `category` type to facilitate easier handling and processing.

Updated Features:

- **Numerical Features**: `avg_temp`, `heat_index`, `windspeed`
- **Categorical Features**: `season`, `weathersit`, `day_night`

4. Data Preprocessing Pipeline

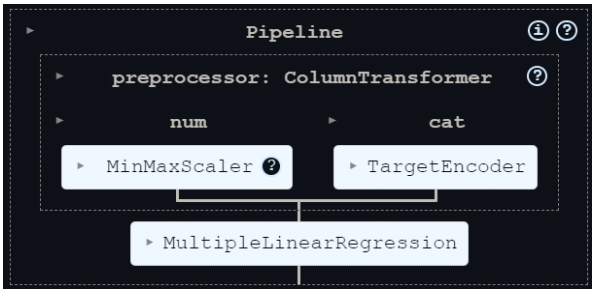
A comprehensive pipeline was created for preprocessing and modeling:

1. **Feature Scaling and Encoding:**
    - **Numerical Features**:
        - **Scaler**: Applied Min-Max scaling to normalize numerical features (`avg_temp`, `heat_index`, `windspeed`) to the range [0, 1].
    - **Categorical Features**:
        - **Encoder**: Used Target Encoding to transform categorical features (`season`, `weathersit`, `day_night`) based on their relationship with the target variable `cnt`.

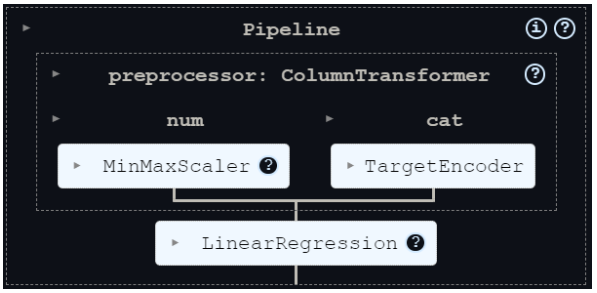
2. **Pipeline Construction:**
    - **ColumnTransformer**: Combined numerical scaling and categorical encoding into a single preprocessing step to streamline the process.
    - **Pipeline**: Integrated preprocessing with model training into a unified pipeline for efficient and reproducible model fitting.

Pipeline Definition:

- **Numerical Features**: Processed with **Min-Max Scaling**
- **Categorical Features**: Processed with **Target Encoding**
- **Combined Pipeline**:
    - **Preprocessor**: `ColumnTransformer` applying both scaling and encoding.
    - **Model Training**: Integrated into the pipeline for streamlined training and evaluation.



(From Scratch)



(Scikit-Learn)

5. Model Training

Two models were utilised to predict the number of bike rentals (`cnt`):

1. **Linear Regression (scikit-learn)**:
    - A standard Linear Regression model was used to predict bike rentals based on the processed features. This model was chosen for its simplicity and effectiveness in capturing linear relationships between the features and the target variable.
2. **Multiple Linear Regression (from scratch)**:
    - A Multiple Linear Regression model was implemented manually to gain deeper insights into the regression mechanics and to provide a comparison with the standard scikit-learn model. This model was designed to replicate the behavior of a linear regression but was constructed manually to better understand the underlying calculations.

6. Model Evaluation

The models were evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics:

Metric	Linear Regression (scikit-learn)	Multiple Linear Regression (from scratch)
MSE	19,588.3	19,588.3
R^2	0.381	0.381

- **MSE (Mean Squared Error)**: The MSE is identical for both models, indicating that both approaches have the same average squared error in their predictions.
- **R^2 (R-squared)**: The R^2 value is also the same for both models, suggesting that both models explain the same proportion of the variance in the target variable.

7. Summary

- **Target Variable**: `cnt` (total number of bike rentals)
- **Features Engineered**: Created new features `avg_temp`, `heat_index`, and `day_night`.
- **Dropped Columns**: Removed redundant or unnecessary features.
- **Preprocessing Pipeline**: Combined scaling and encoding into a `ColumnTransformer` and `Pipeline` for efficient processing.
- **Models Used**:
    - **Linear Regression** from `scikit-learn`: Applied for its effectiveness in capturing linear relationships.
    - **Multiple Linear Regression**: Implemented from scratch to compare results and understand regression mechanics.
- **Evaluation**: Both models achieved the same MSE and R^2 values, indicating comparable performance.